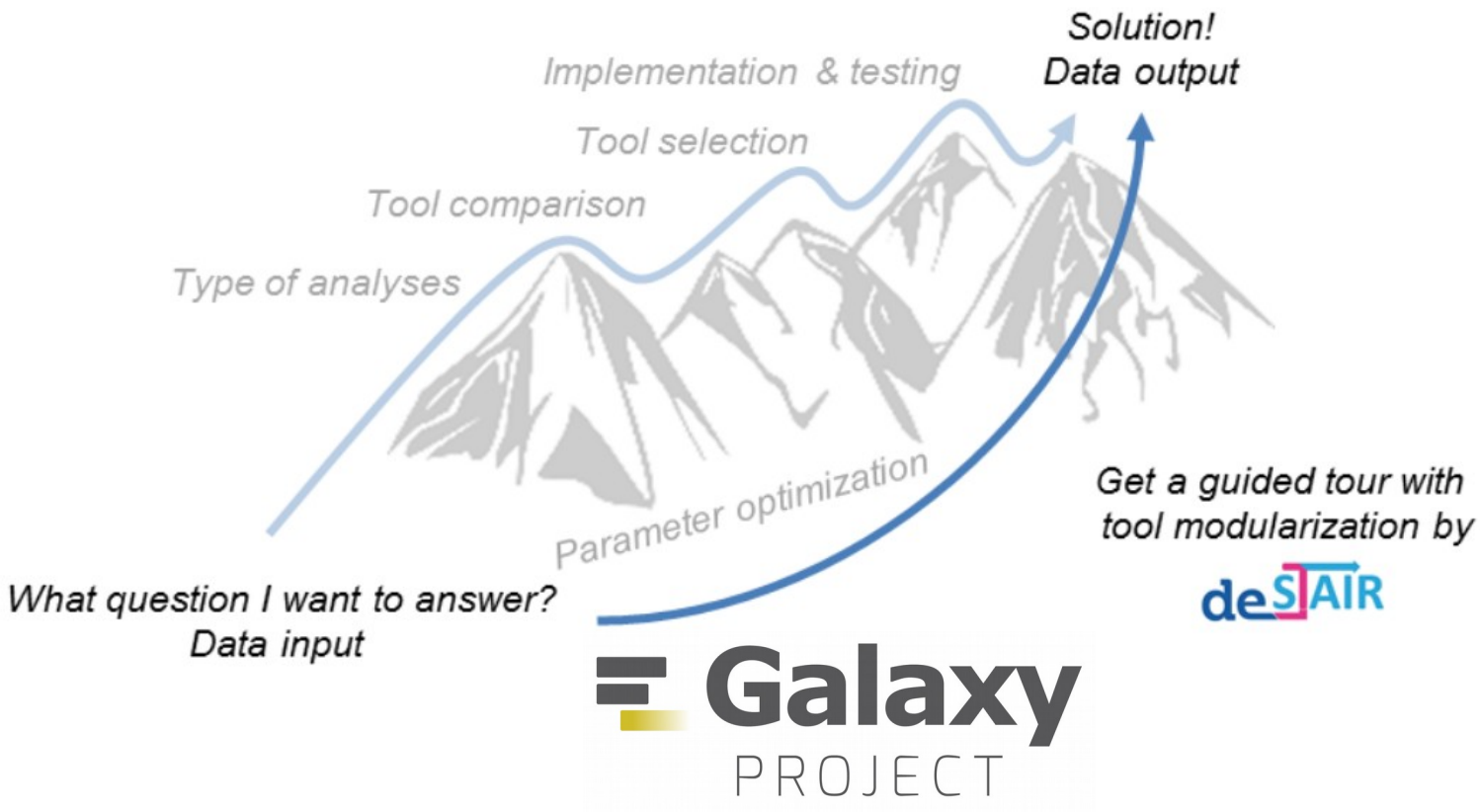
Leibniz Institute on Aging – Fritz Lipmann Institute

ECSEQ BIOINFORMATICS

de.NBI/de.STAIR Training: A primer for RNA-Seq processing

Konstantin Riege
Steve Hoffmann

Bioinformatics Services for Structured Analysis and Integration of RNA-Seq experiments (de.STAIR)



Sequencing techniques

Illumina sequencing platforms

miSeq

NextSeq 500

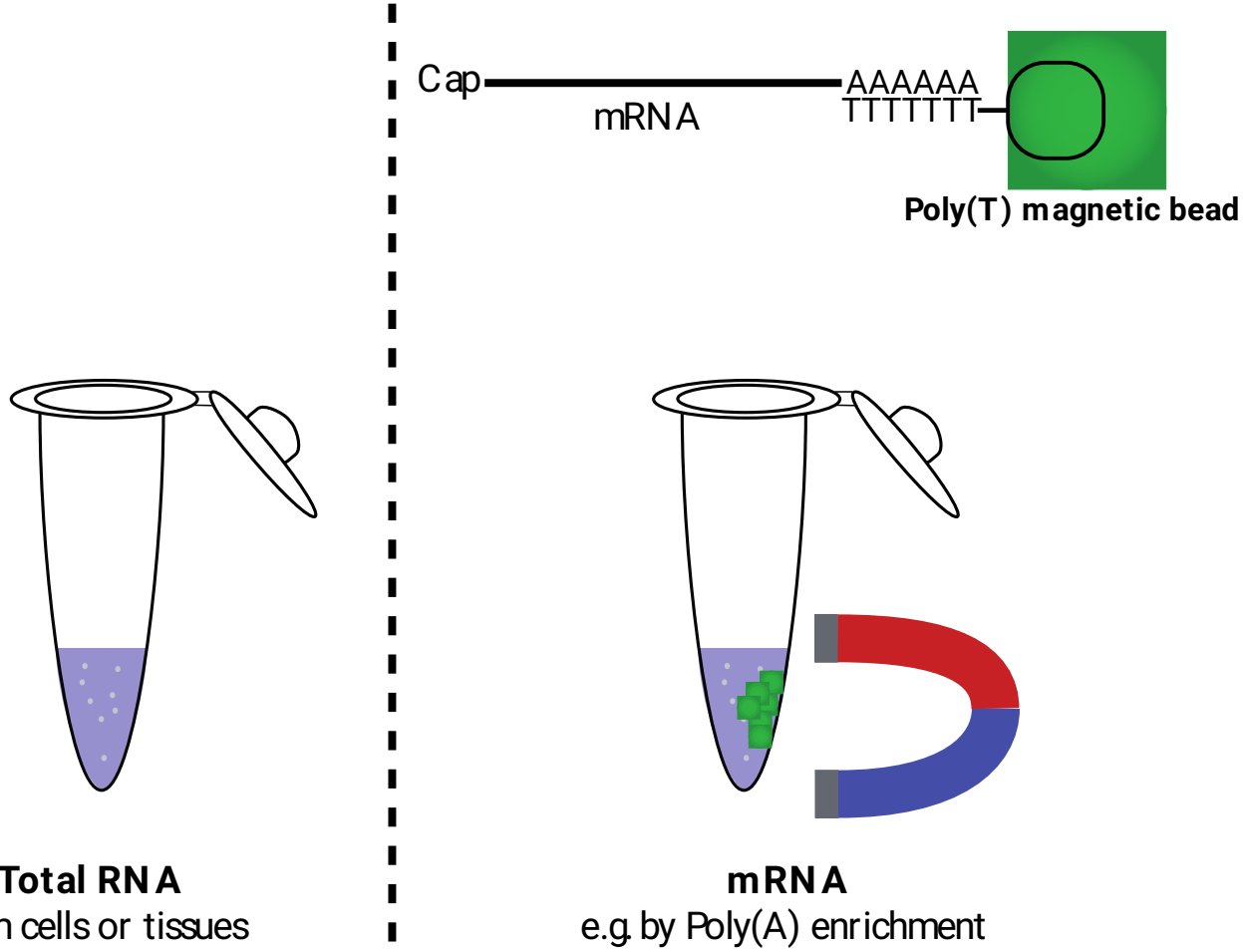
HiSeq 2500

HiSeq X Ten

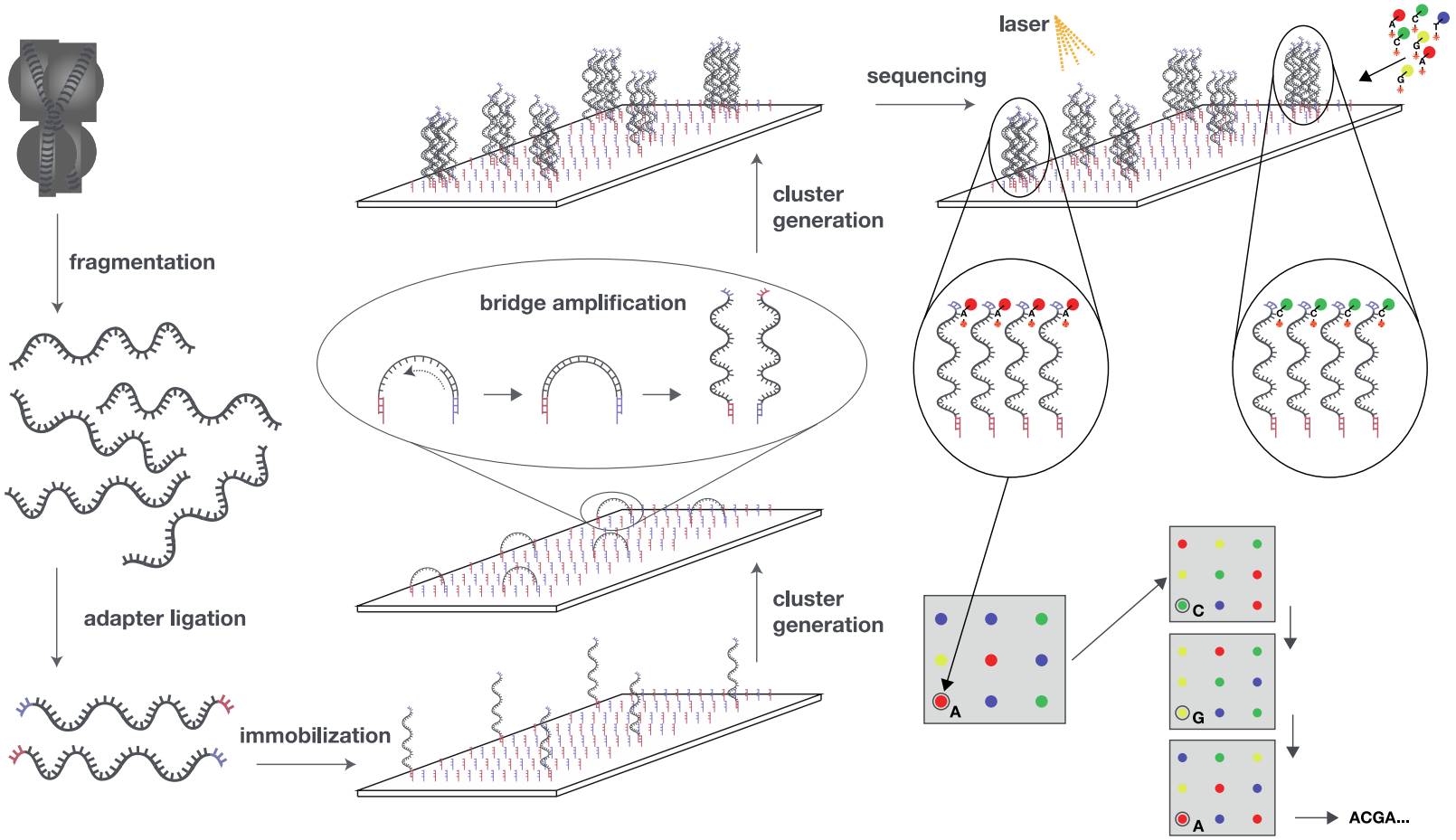
Illumina sequencing platforms

	Run time	Read length	Throughput		Cost	
	(hrs)	(bp)	# reads	bases/run	machine	per Gb
miSeq	65	2 x 300	25M	15Gb	$125k	$93
NextSeq 500	29	2 x 150	400M	129Gb	$250k	$33
HiSeq 2500	144	2 x 125	4B	1Tb	$740k	$29
HiSeq X Ten	72	2 x 150	6B	1.8Tb	$1M	$7

RNA-Seq library preparation

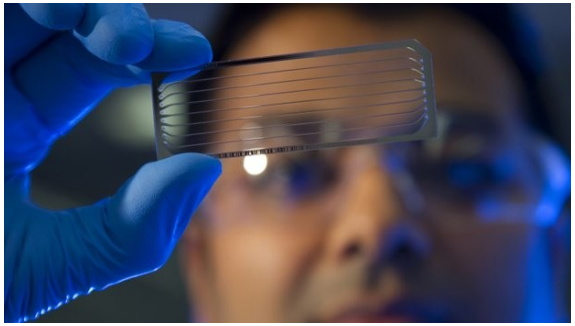


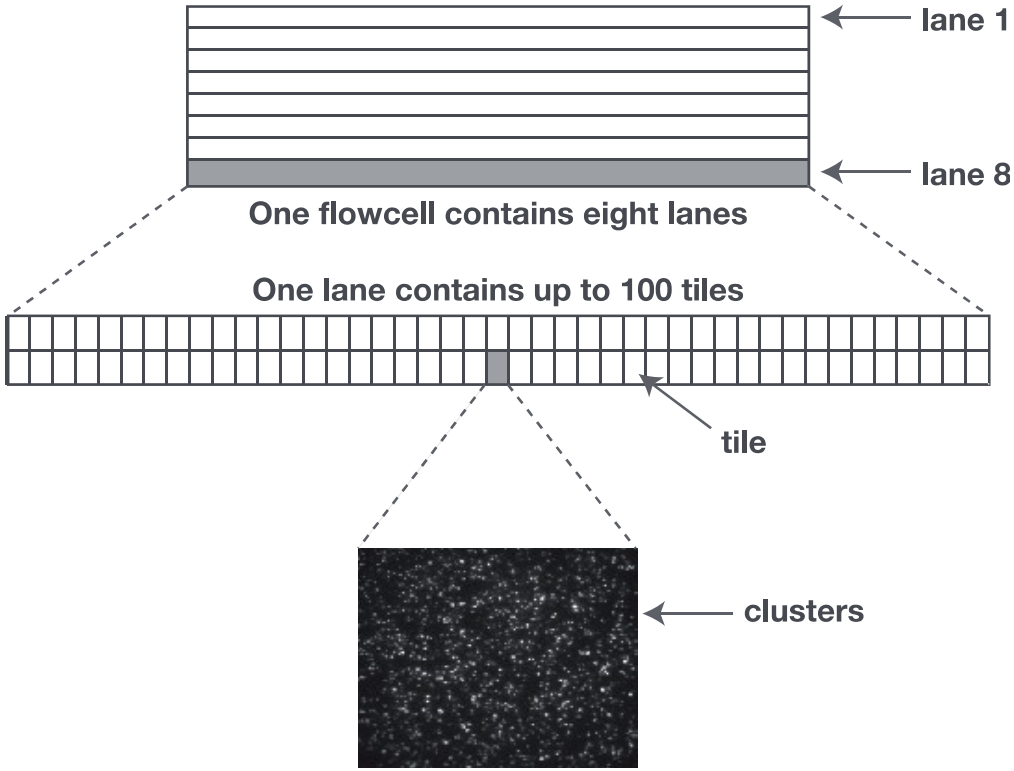
Sequencing workflow



Base-calling

- Bases are called from clustered intensities of emitted fluorescence
- Cluster density influences the base-calling quality





FASTQ format (1st line)

```
@SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGCCATGTTAGACAAGGCGCAGATATA
GGTGATGCTGATGCAGAAAAACGATT
+SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFDDEDCCDCCCCDDDDDBDBD>CDEE
>C@CDDDDDDCACAACCDCDBDBB<1
@SRR359063.2 D042KACXX:3:1101:5202:2193 length=101
CTCTGGTACAGAACACGTGGATTATAAGAGTTGCCGCTTCGCACAGAAGTCGGAGTTCTCTCACCACTTTTGAGC
TCTTCCTCGGCTTCTTCTTCCTCTTT
```

SRR359063.1	run ID
D042KACXX	flowcell ID
3	flowcell lane
1101	tile number within the flowcell lane
2690	'x'-coordinate of the cluster within the tile
2160	'y'-coordinate of the cluster within the tile

FASTQ format (2nd line)

```
@SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGCCATGTTAGACAAGGCGCAGATATA
GGTGATGCTGATGCAGAAAAACGATT
+SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFDDEDCCDCCCCDDDDDBDBD>CDEE
>C@CDDDDDDCACAACCDCDBDBB<1
@SRR359063.2 D042KACXX:3:1101:5202:2193 length=101
CTCTGGTACAGAACACGTGGATTATAAGAGTTGCCGCTTCGCACAGAAGTCGGAGTTCTCTCACCACTTTTGAGC
TCTTCCTCGGCTTCTTCTTCCTCTTT
```

The raw sequence letters

FASTQ format (3rd line)

```
@SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGCCATGTTAGACAAGGCGCAGATATA
GGTGATGCTGATGCAGAAAAACGATT
+SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFDDEDCCDCCCCDDDDDBDBD>CDEE
>C@CDDDDDDCACAACCDCDBDBB<1
@SRR359063.2 D042KACXX:3:1101:5202:2193 length=101
CTCTGGTACAGAACACGTGGATTATAAGAGTTGCCGCTTCGCACAGAAGTCGGAGTTCTCTCACCACTTTTGAGC
TCTTCCTCGGCTTCTTCTTCCTCTTT
```

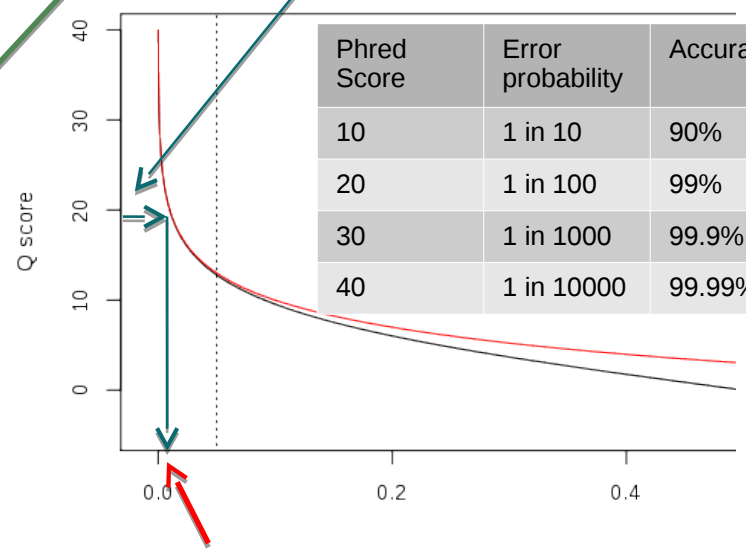
Begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

FASTQ format (4th line)

#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFDDEDCCDCCCCDDDDDBDBD>CDEE >C@CDDDDDDCACAACCDCDBDBB<1

Dec	Hx	Oct	Char		Dec	Hx	Oct	Html	Chr
0	0	000	NUL	(null)	32	20	040		Space
1	1	001	SOH	(start of heading)	33	21	041	!	!
2	2	002	STX	(start of text)	34	22	042	"	"
3	3	003	ETX	(end of text)	35	23	043	#	#
4	4	004	EOT	(end of transmission)	36	24	044	$	$
5	5	005	ENQ	(enquiry)	37	25	045	%	%
6	6	006	ACK	(acknowledge)	38	26	046	&	&
7	7	007	BEL	(bell)	39	27	047	'	'
8	8	010	BS	(backspace)	40	28	050	(	(
9	9	011	TAB	(horizontal tab)	41	29	051	)	)
10	A	012	LF	(NL line feed, new line)	42	2A	052	*	*
11	B	013	VT	(vertical tab)	43	2B	053	+	+
12	C	014	FF	(NP form feed, new page)	44	2C	054	,	,
13	D	015	CR	(carriage return)	45	2D	055	-	-
14	E	016	SO	(shift out)	46	2E	056	.	.
15	F	017	SI	(shift in)	47	2F	057	/	/
16	10	020	DLE	(data link escape)	48	30	060	0	0
17	11	021	DC1	(device control 1)	49	31	061	1	1
18	12	022	DC2	(device control 2)	50	32	062	2	2
19	13	023	DC3	(device control 3)	51	33	063	3	3
20	14	024	DC4	(device control 4)	52	34	064	4	4
21	15	025	NAK	(negative acknowledge)	53	35	065	5	5
22	16	026	SYN	(synchronous idle)	54	36	066	6	6
23	17	027	ETB	(end of trans. block)	55	37	067	7	7
24	18	030	CAN	(cancel)	56	38	070	8	8

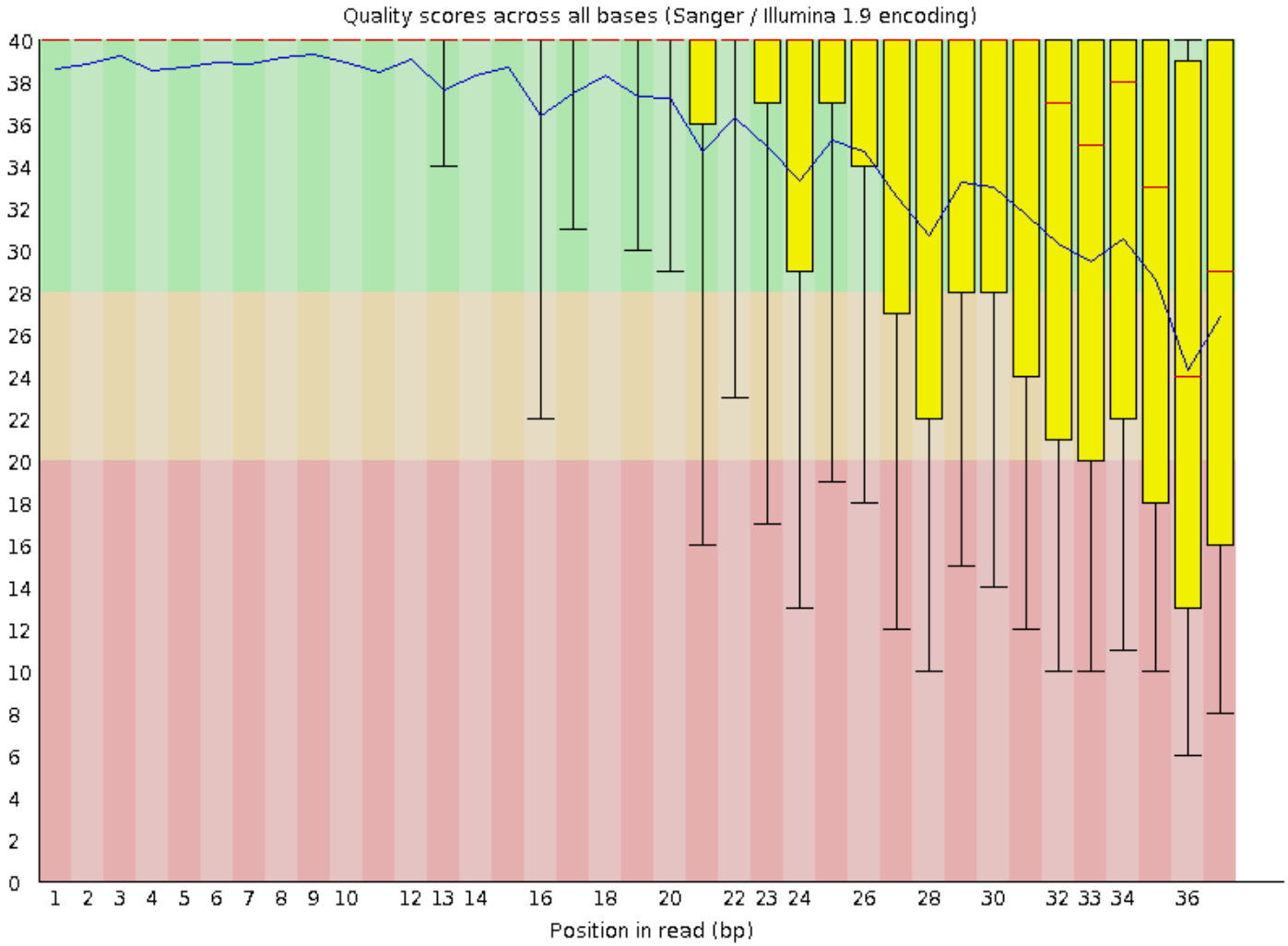
52 – 33 = 19 = Phred-(33)-Score



Phred Score	Error probability	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

the probability that the corresponding base call is incorrect

Quality analysis



Quality control and FASTQ processing tools

- FastQC
- MultiQC

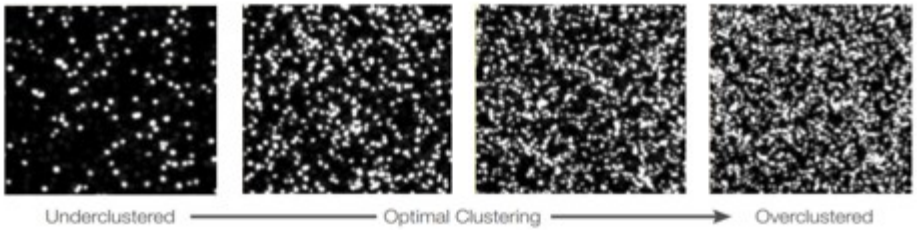
- Seqtk

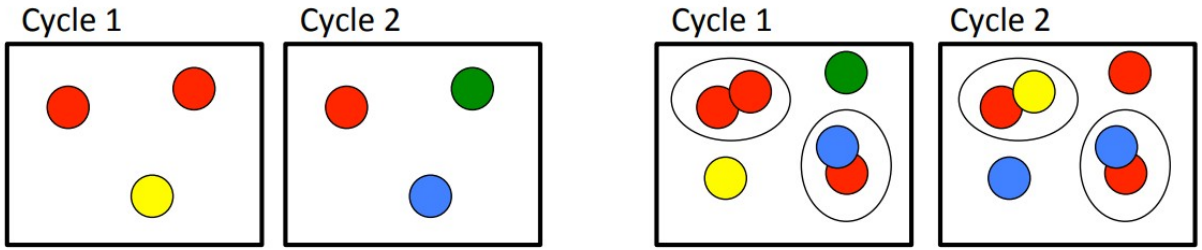
Batch effects and errors

Batch effects and errors

- Library preparation
  - Cluster density
  - Nucleotide diversity
  - Fragment size vs. read length
  - G/C bias
- Sequencing errors

Batch effects and errors

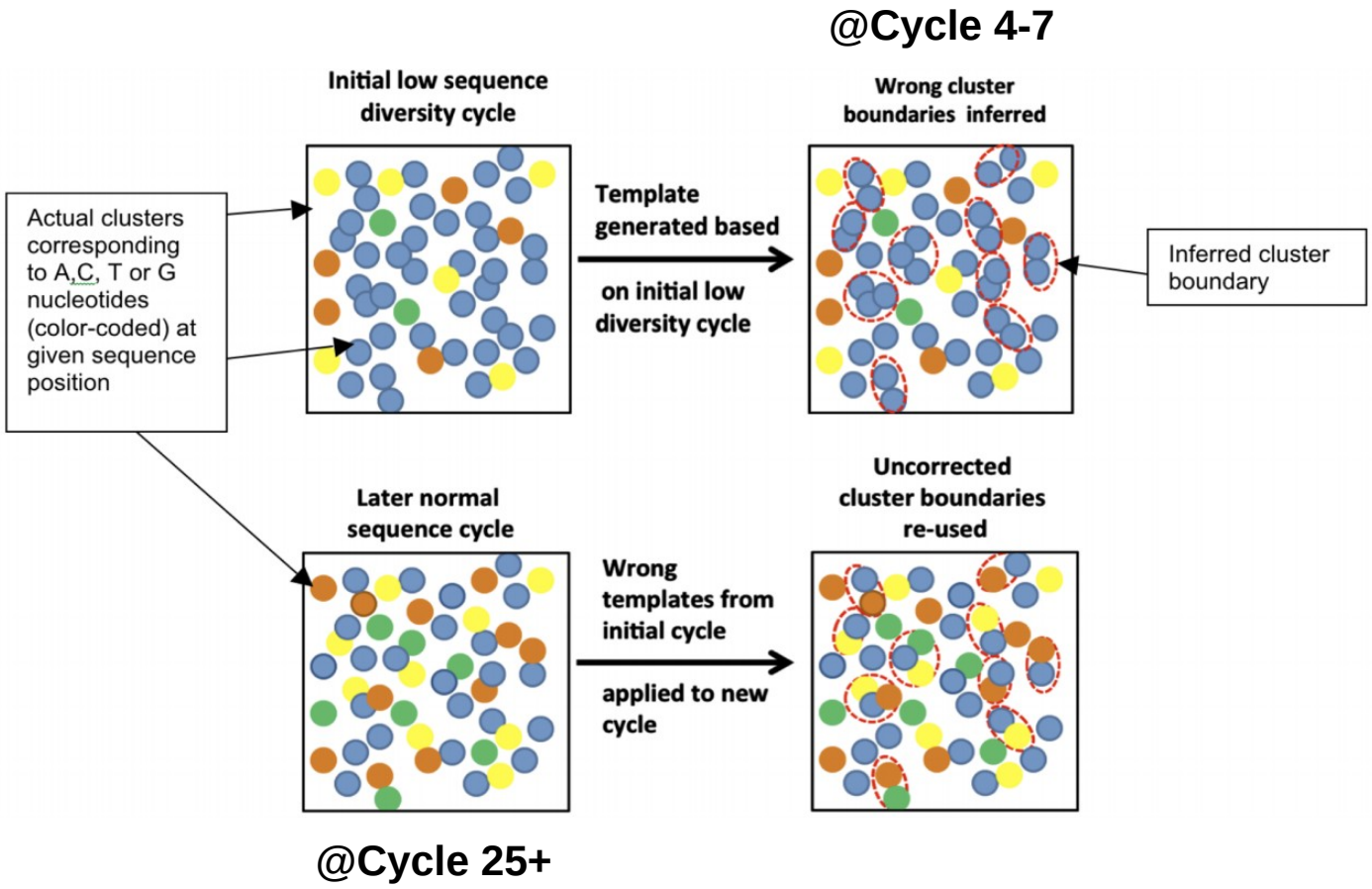




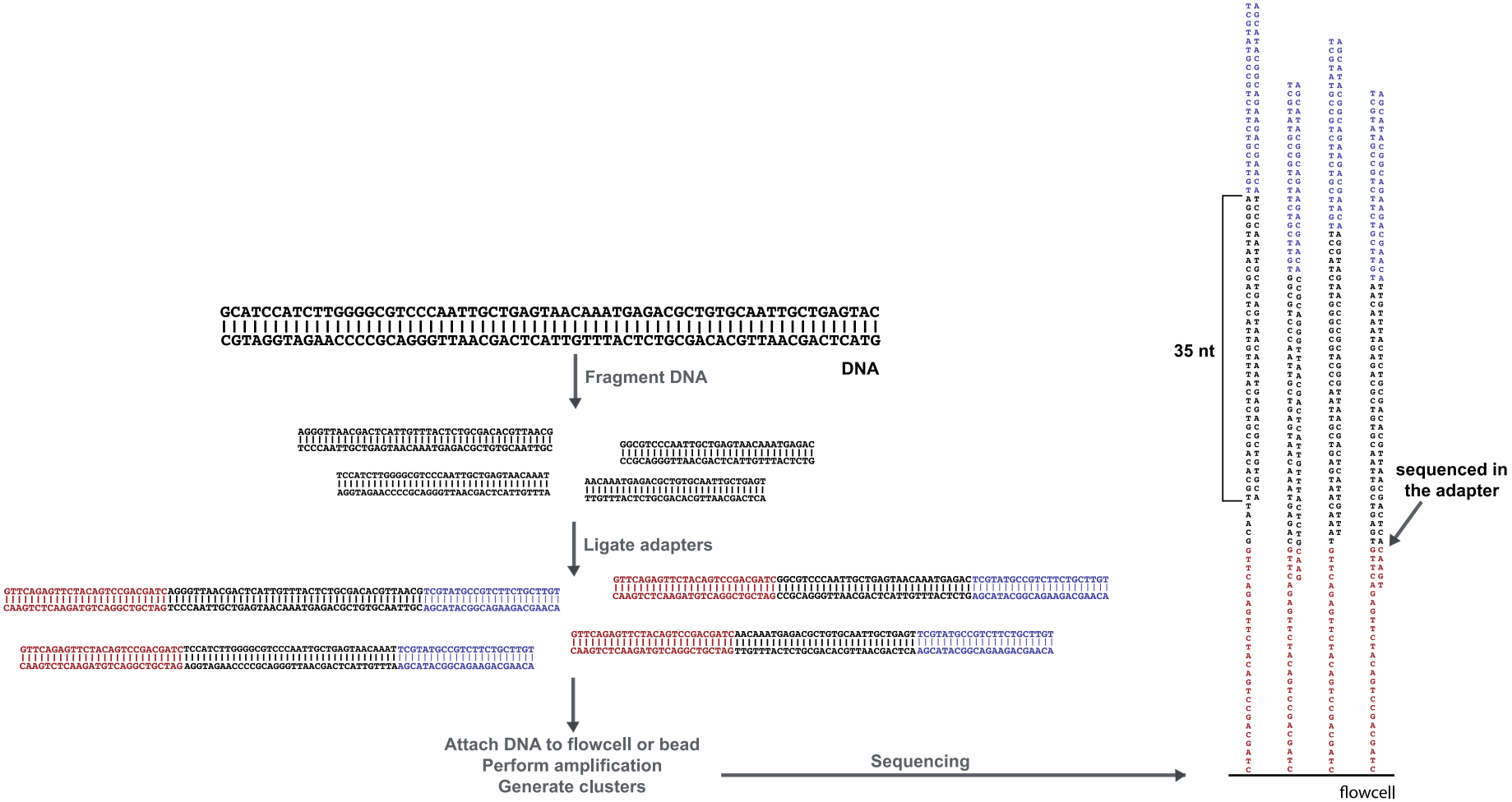
Well-spaced clusters easier to call

Densely-packed clusters difficult to call

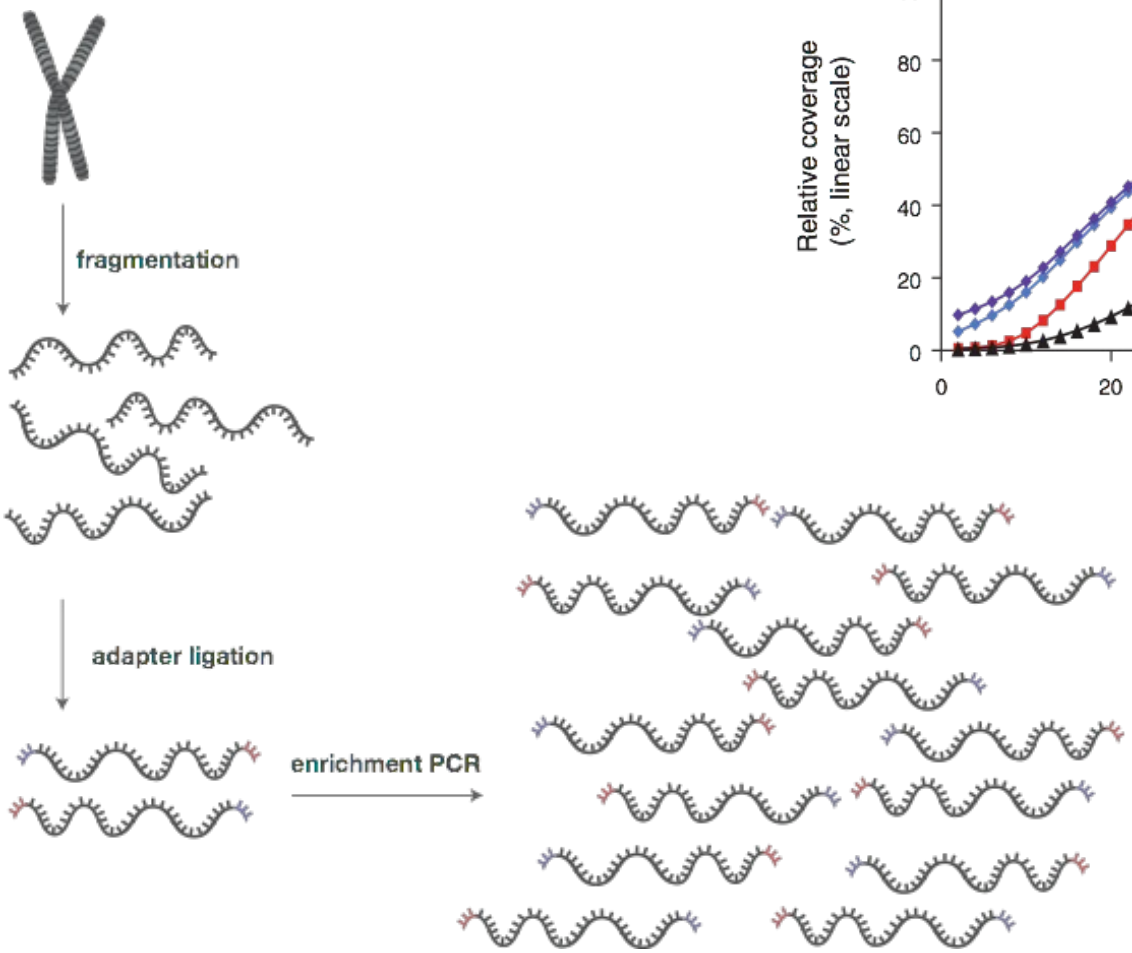
Batch effects and errors

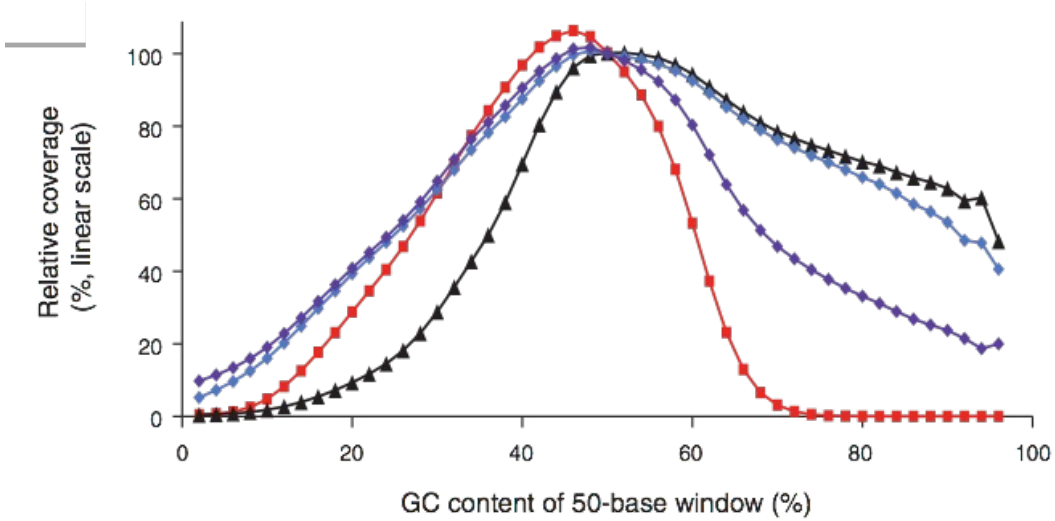


Batch effects and errors

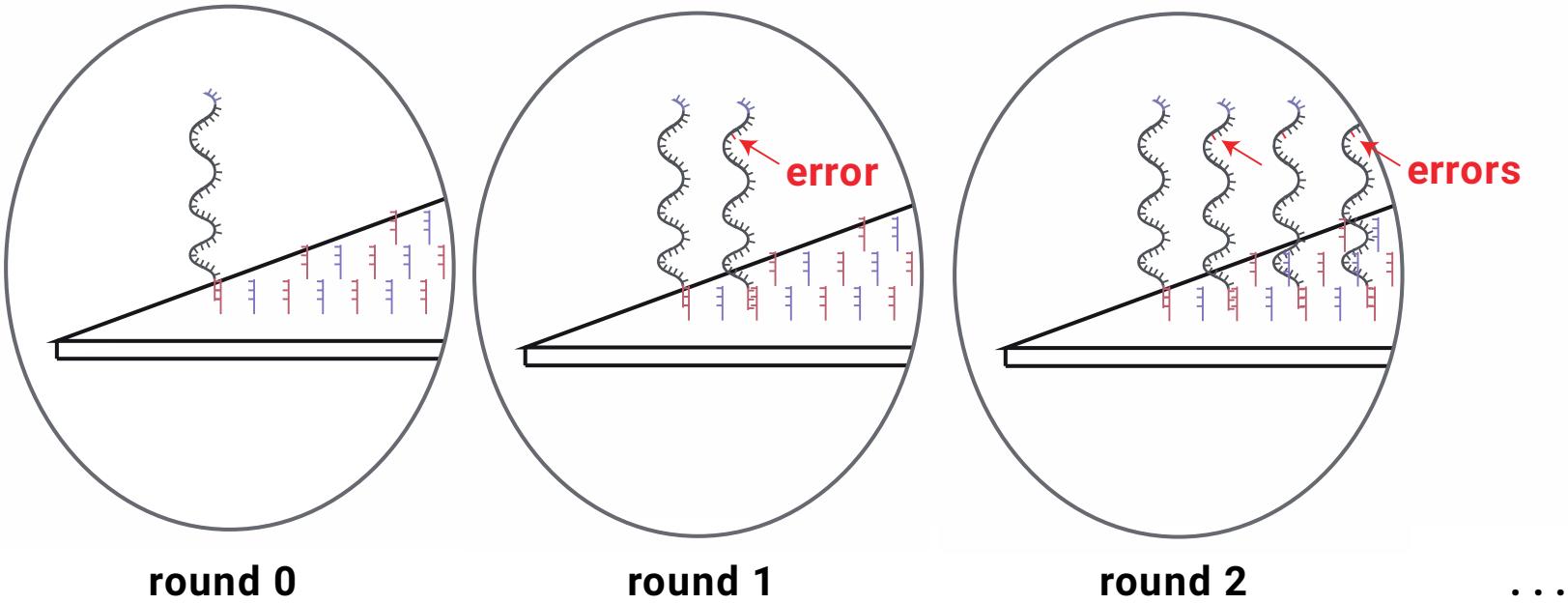


Batch effects and errors



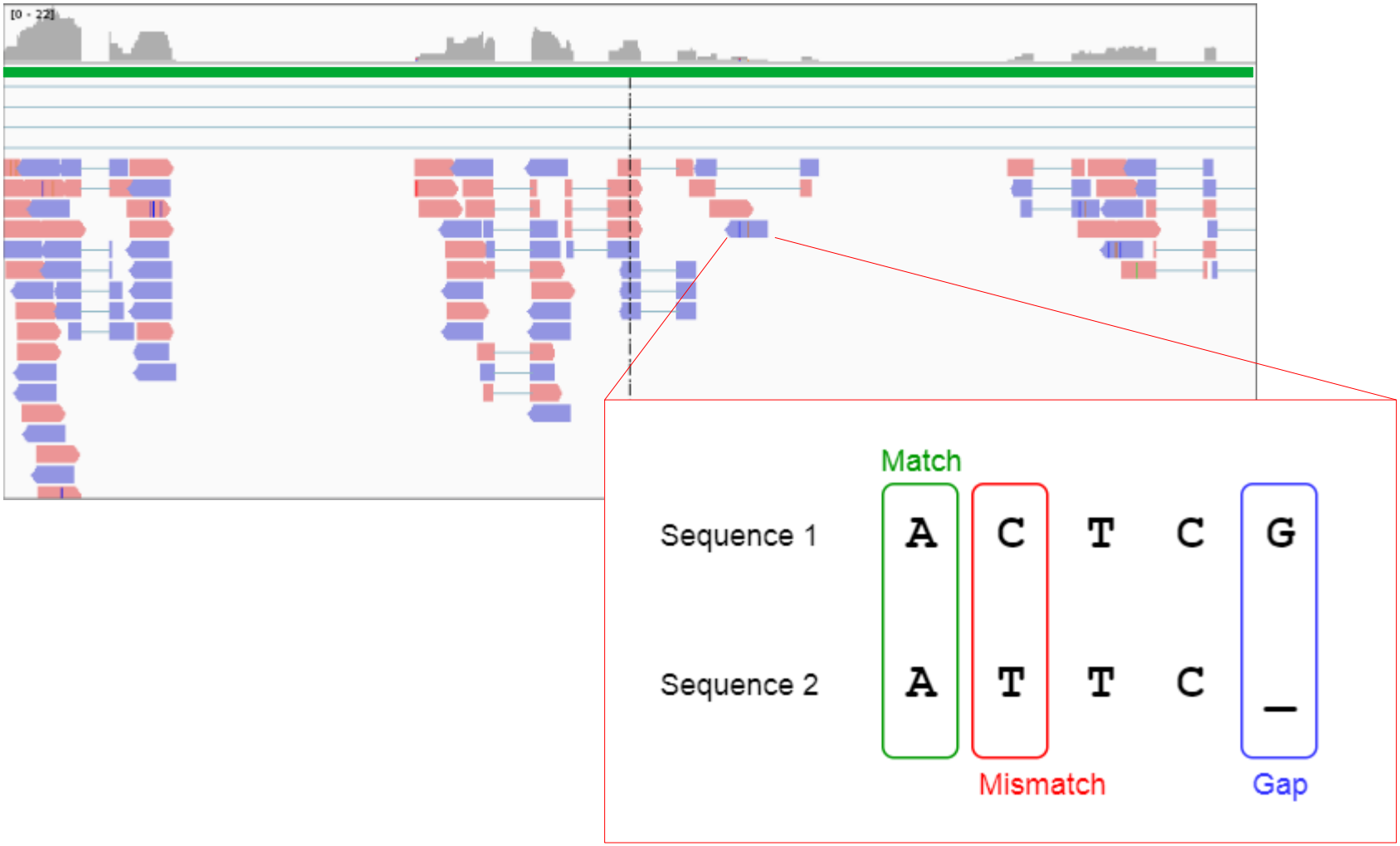


Batch effects and errors

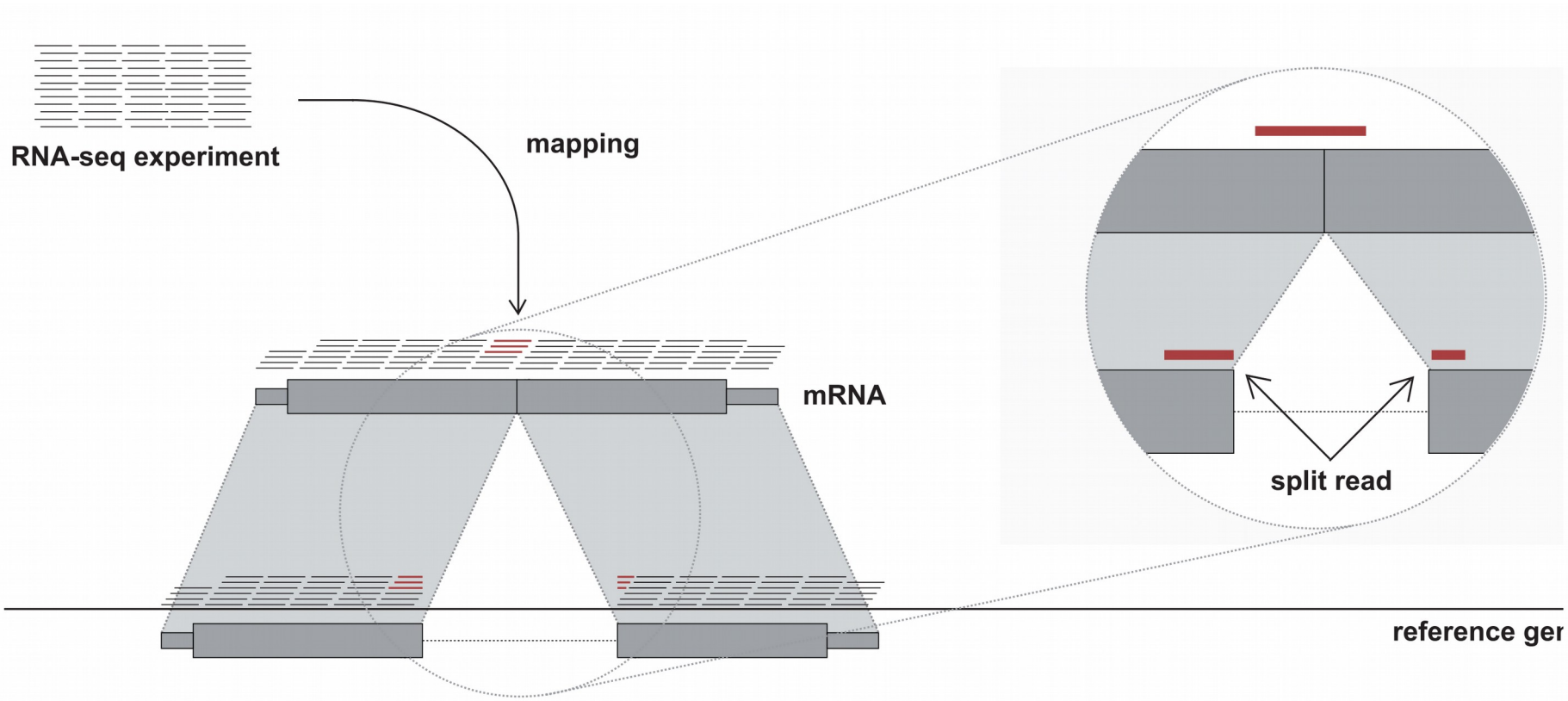


Reference alignments / Mapping

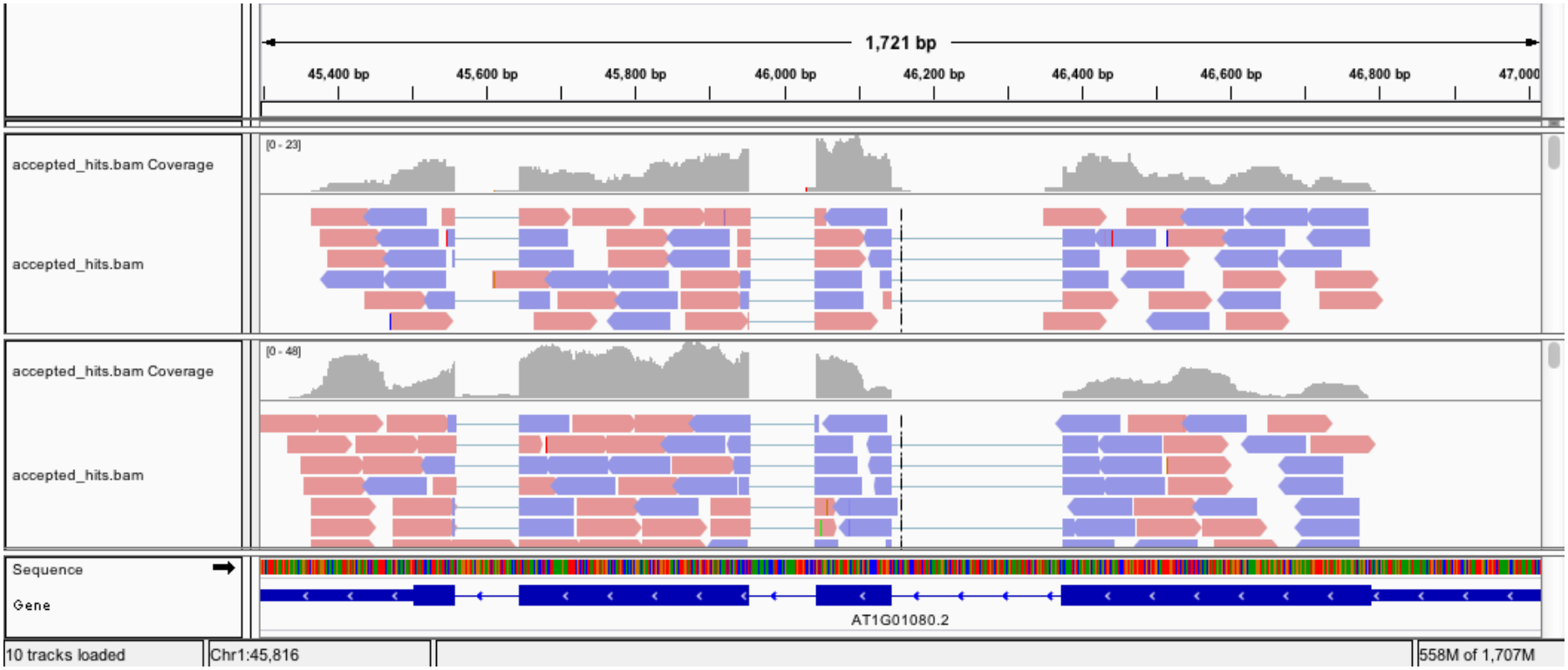
Reference alignments



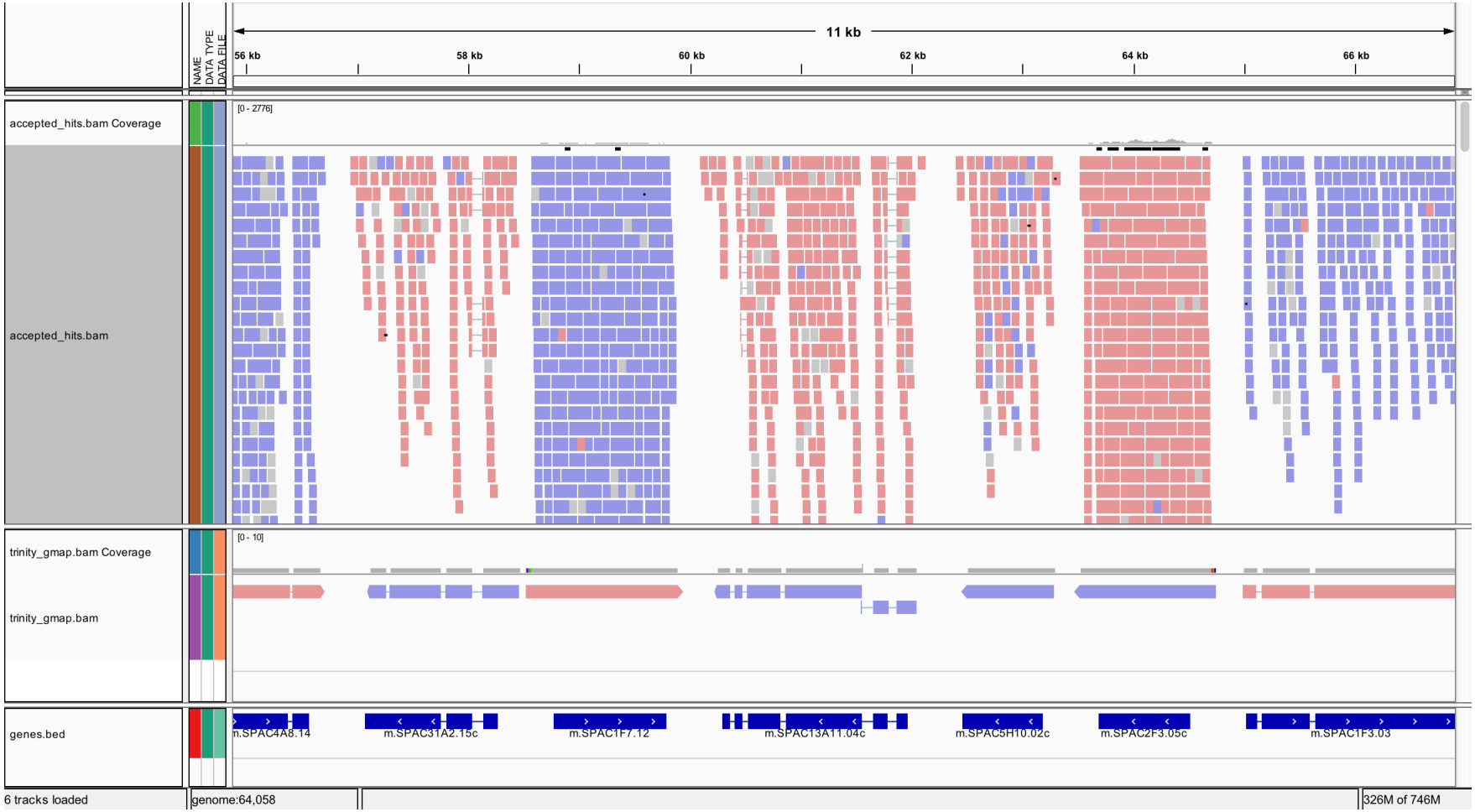
Reference alignments



Reference alignments – Non-Strand specific library



Reference alignments – Strand specific library



Mapping problems

- Difference „optimal alignment“ and „true alignment“
  - Quality of the reference genome and read
  - Scoring scheme of matches, mismatches and InDels

```
AT___GC        A_TG__C
ATTGAGC        ATTGAGC
```

This is more likely. Explained by one event

Normal scoring would give the same score for both alignments

This is less likely. Requires 2 events.

Mapping problems

- Difference „optimal alignment“ and „true alignment“
  - Quality of the reference genome and read
  - Scoring scheme of matches, mismatches and InDels
  - Information content of reads - Multiple mapping

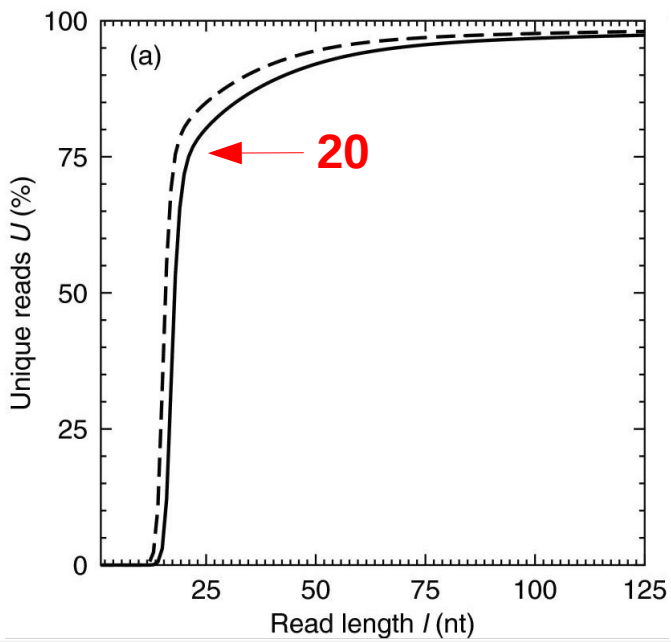
Expectation of an {ATCG}-word in a random {ATCG}-Sequence

$$E = p^m * n$$

$$E = 0.25^l * 3.2 * 10^9$$

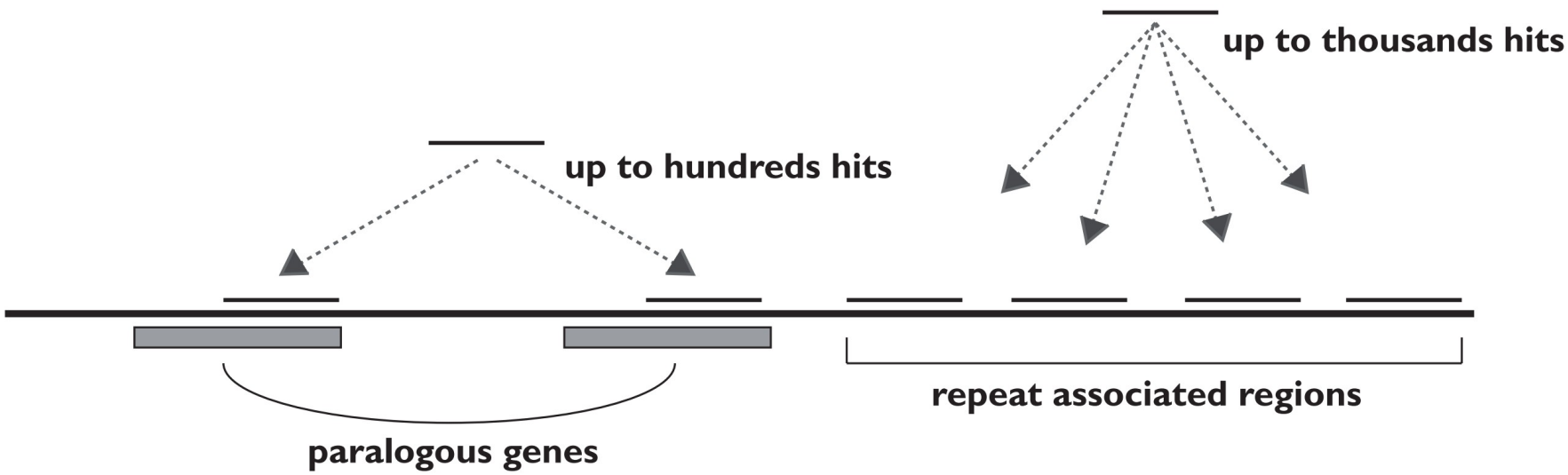
$$1 = 0.25^l * 3.2 * 10^9$$

$$l \sim 16$$

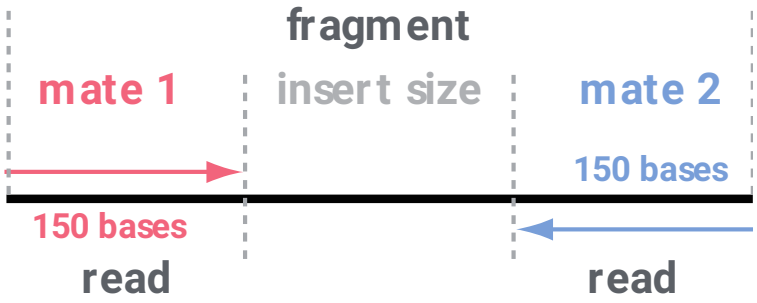


Mapping problems

- Difference „optimal alignment“ and „true alignment“
  - Quality of the reference genome
  - Scoring scheme of matches, mismatches and InDels
  - Information content of reads - Multiple mapping



Paired-End sequencing



fragment
molecule to be sequenced

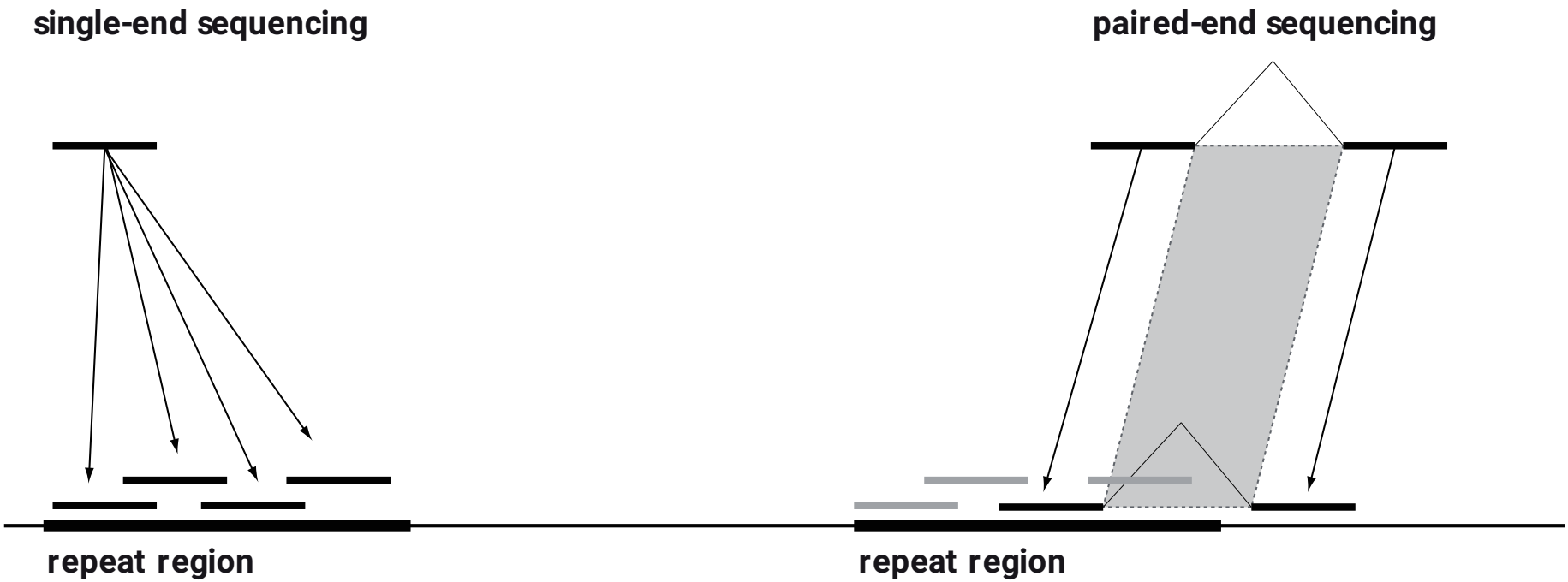
read
One sequenced part of a biological fragment (mate1 and/or mate2)

mate 1
sequence of the 5’ end of paired-end sequencing

mate 2
sequence of the 3’ end of paired-end sequencing

sequencing depth aka library size
The total number of all the sequences, reads or bases represented in a single sequencing experiment

Paired-End sequencing



SAM format (binary compressed: BAM)

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

SAM format (binary compressed: BAM)

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

SAM format (binary compressed: BAM)

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Often (NOT ALWAYS!) a mapping quality value 0 indicates a multiple mapped read – otherwise, if available grep for NH:i:1 tag

Summary

- Minimum read length 20
- Remove potential adapters
- Perform quality trimming (~PHRED 20)
- Use split-read mapping algorithms
- Adjust expected insert size
- Constrain scoring scheme
  - E.g. BWA default settings according to its manual
    minOUTscore:30
    MM/indelpenalty:4/6

    i.e. for read length 100: (100-30)/4~17% errors
- Filter SAM file for uniquely mapped reads

Trimming, mapping and SAM post-processing tools

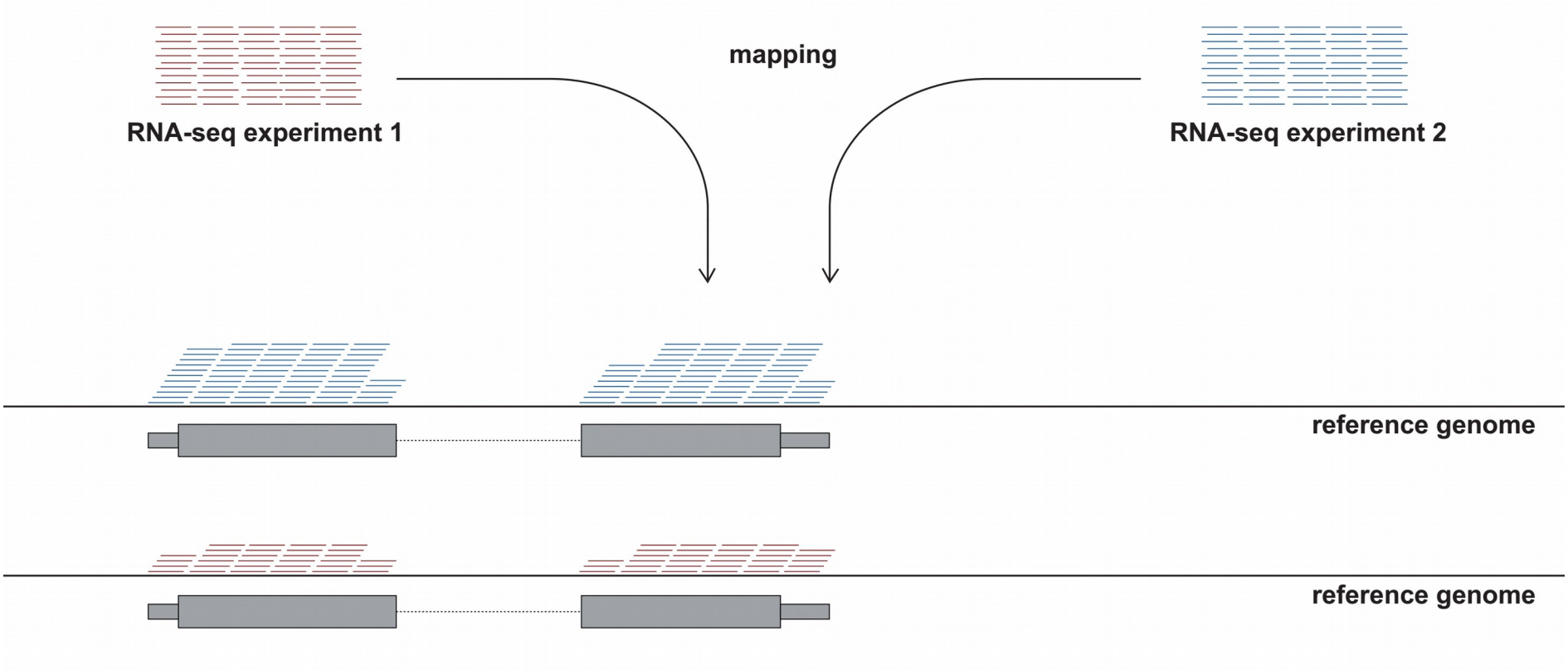
- Cutadapt
- PRINSEQ
- Trimmomatic

- Segemehl
- STAR
- BWA
- HISAT2

- Samtools
- Picard

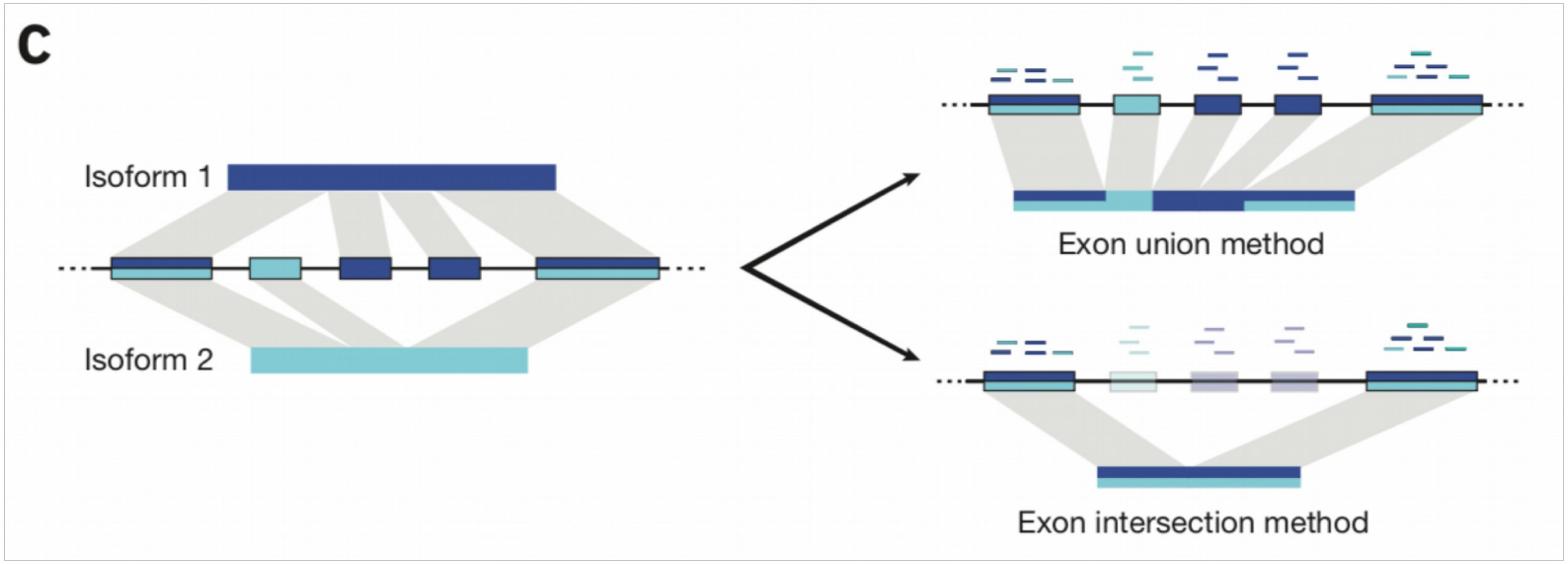
Differential expression analysis

Differential expression analysis



Problems in expression analyses

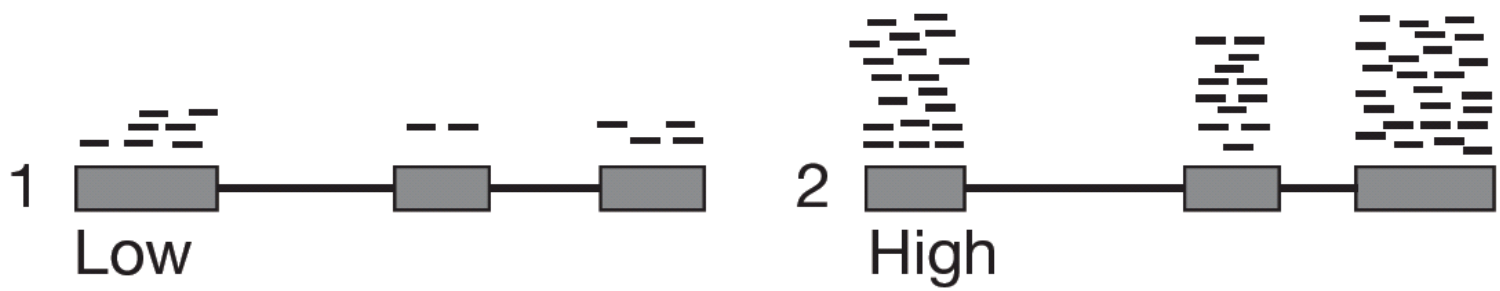
- Multiple isoforms and gene models



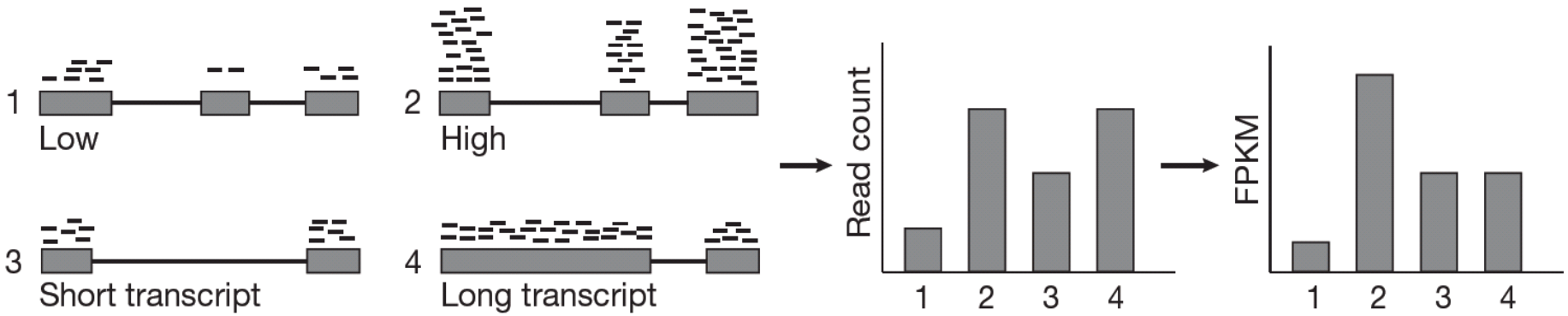
Problems in expression analyses

- Differences in transcript length

- Differences in sequencing depth



Normalization between samples



https://www.youtube.com/watch?v=TTUrtCY2k-w

Summary

- Infer strandness for quantification if not known
- Count only uniquely mapped reads (thumb-value: 70%)
- Take care of mapper specific mapping qualities
- Count fragments instead of reads for reasons of comparability of single-End and paired-End sequencing
- Count on unified exon level
- Normalize via FPKM

Quantification and diff. ex. analysis tools

- RseQC: infer_experiment.py
- HTSeq-count
- Featurecounts

- DESeq2
- edgeR

GO enrichment

GO enrichment

- Gene Set Enrichment Analysis
  - Using gene symbols or gene IDs
  - Ranked by p-value or expression ratio

http://www.webgestalt.org
http://wego.genomics.org.cn

Galaxy Training Network

https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html