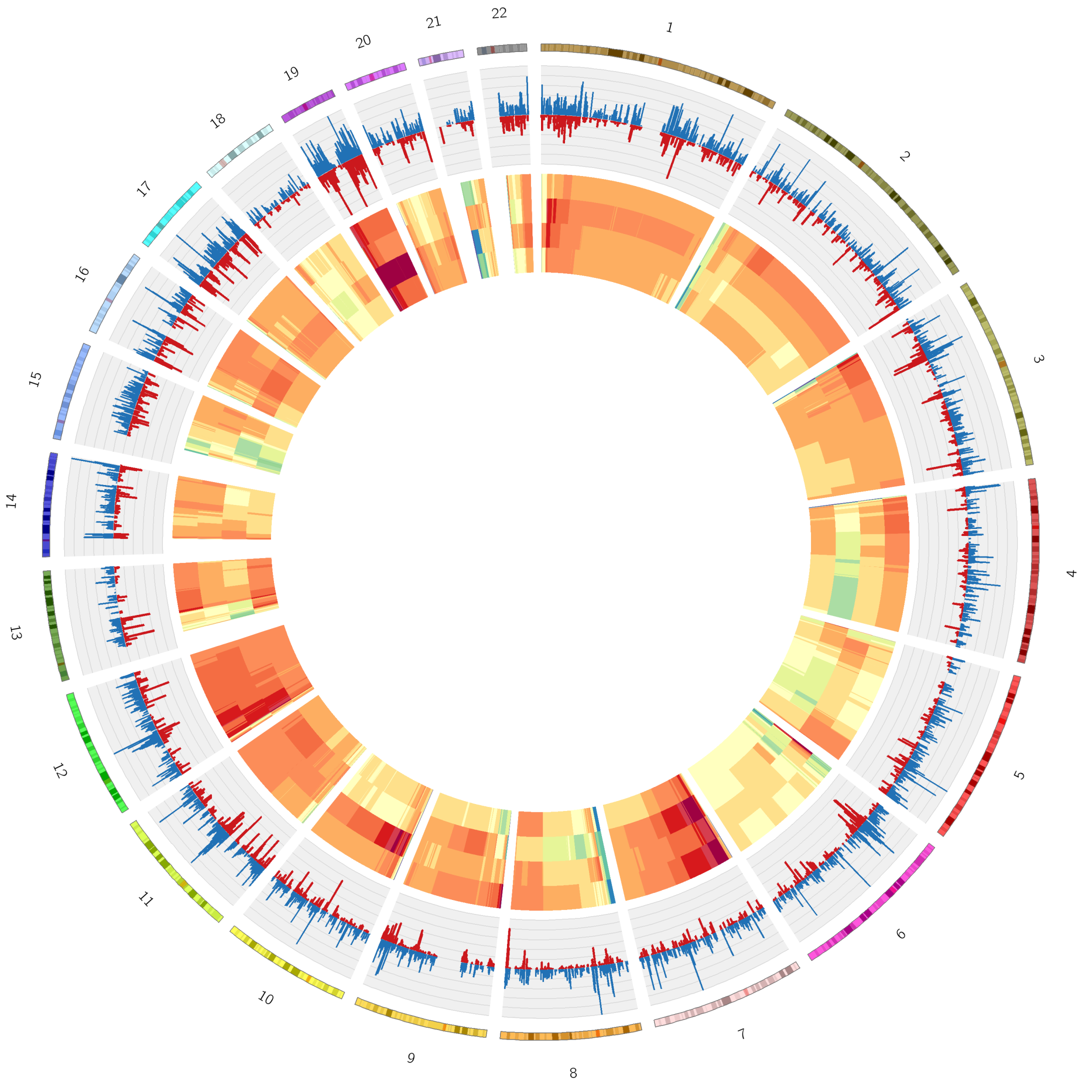


Introduction to Next-Generation Sequencing

Joanna Krupka

CRUK Summer School in Bioinformatics

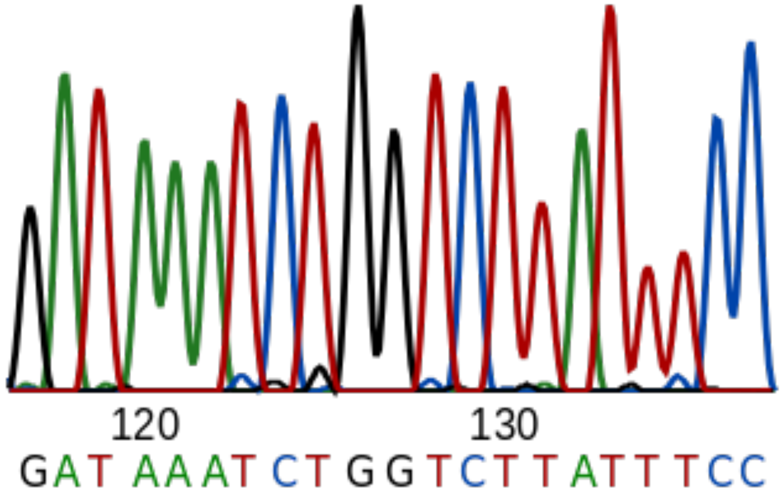
Cambridge, July 2020

CANCER RESEARCH UK

MRC Cancer Unit

UNIVERSITY OF CAMBRIDGE

Brave New World of Next Generation Sequencing



Sanger sequencing (1977)

Human Genome Project

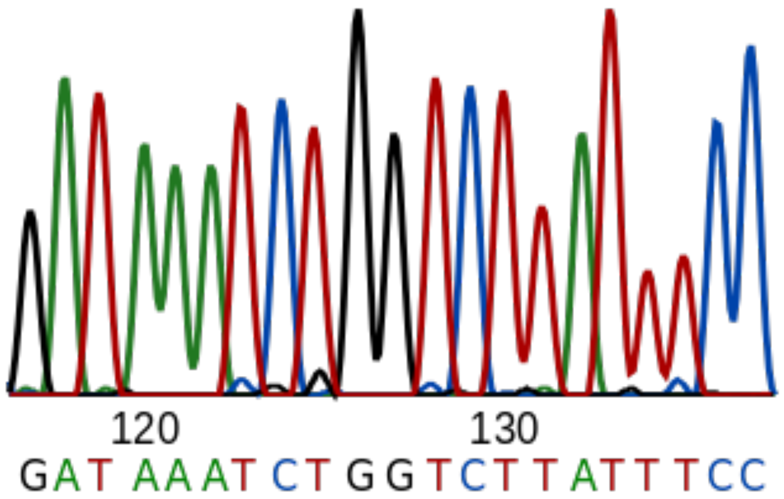
1990 - 2006

DNA Sequencing Technologies Key to the Human Genome Project

By: Heidi Chial, Ph.D. (*Write Science Right*) © 2008 Nature Education

Citation: Chial, H. (2008) DNA sequencing technologies key to the Human Genome Project. *Nature Education* 1(1):219

Brave New World of Next Generation Sequencing



Sanger sequencing (1977)

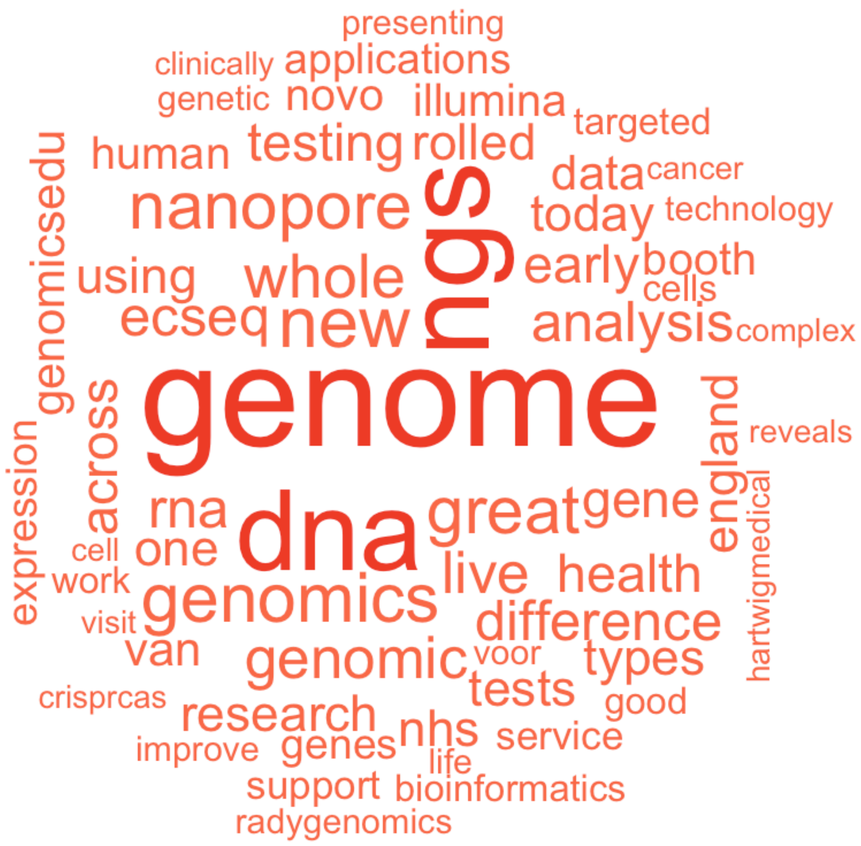
Human Genome Project

1990 - 2006

DNA Sequencing Technologies Key to the Human Genome Project

By: Heidi Chial, Ph.D. (*Write Science Right*) © 2008 Nature Education

Citation: Chial, H. (2008) DNA sequencing technologies key to the Human Genome Project. *Nature Education* 1(1):219



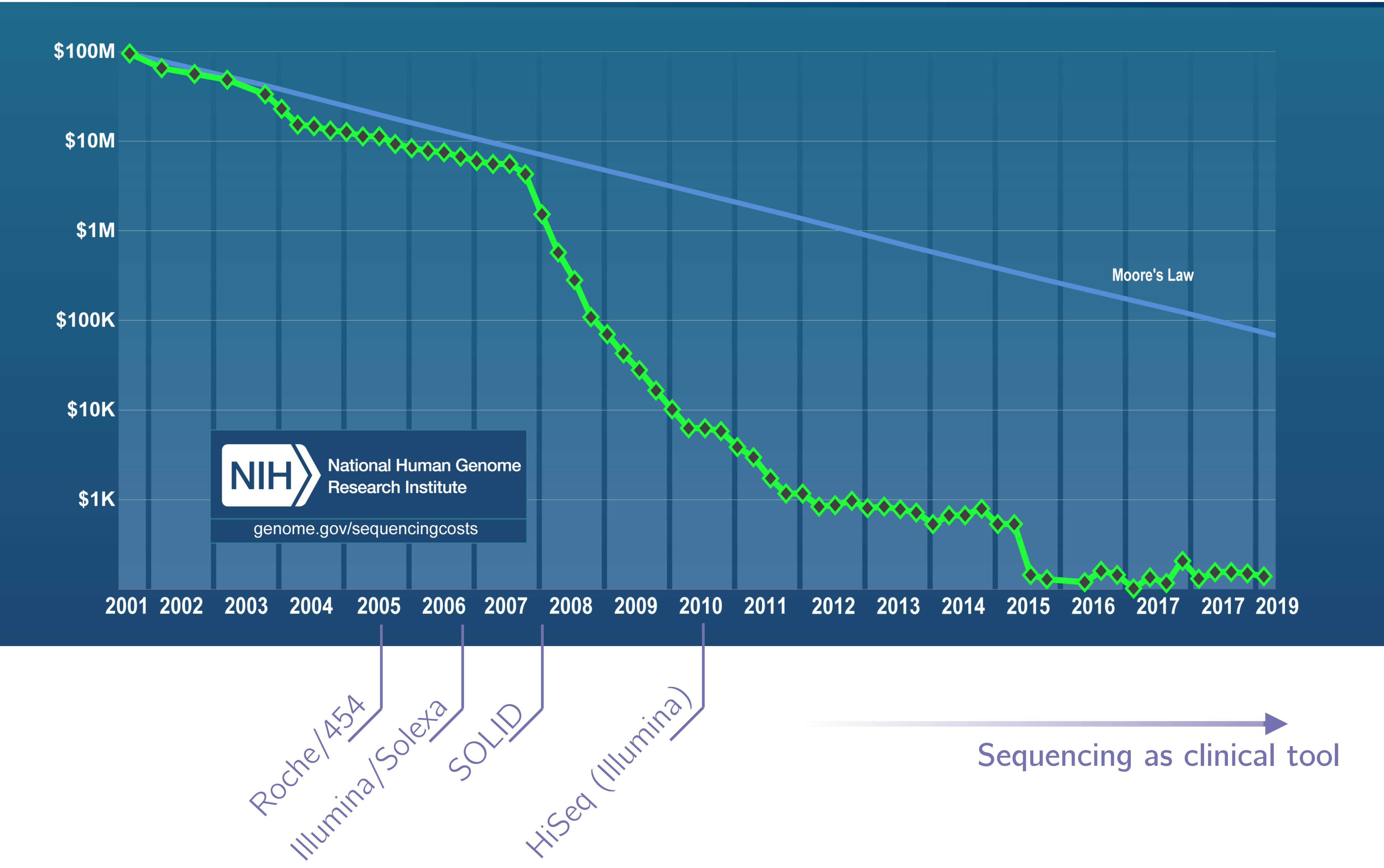
Next Generation Sequencing

mid 2000–present

= high-throughput sequencing

quicker and cheaper parallel sequencing of DNA and RNA

Cost of sequencing of human genome



Sequencing as clinical tool

Next generation sequencing technologies and limitations

Next generation sequencing

Short-read NGS

“Second-generation sequencing”

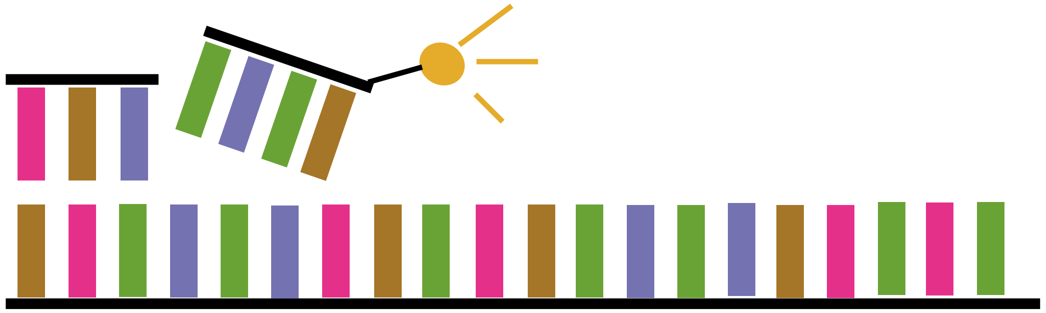
- error rates (0.1–15%)
- read lengths (35–700 bp)

Long-read NGS

“Third-generation sequencing”

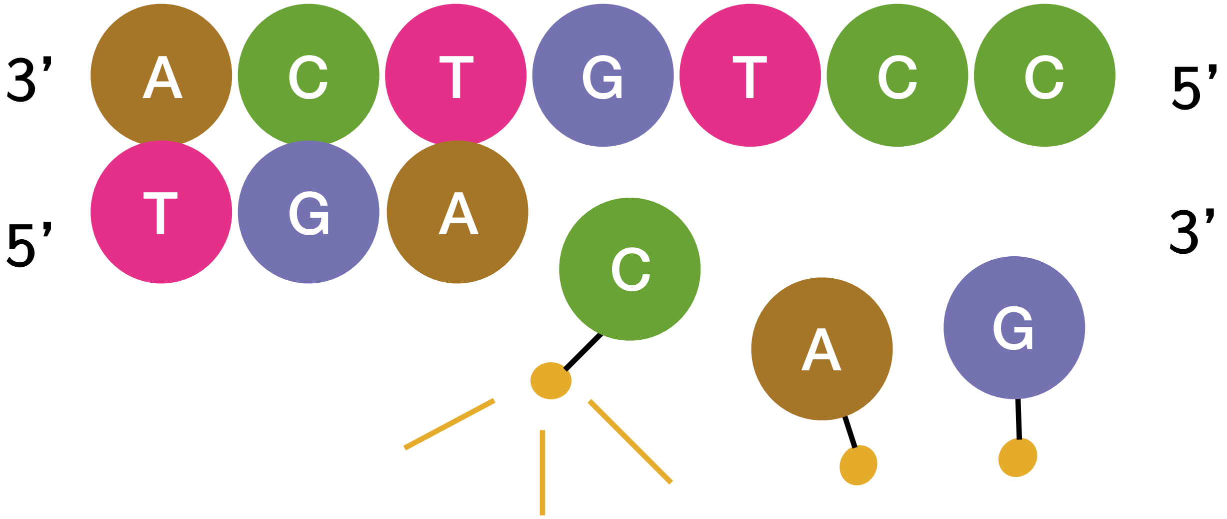
Sequencing by ligation

SOLiD



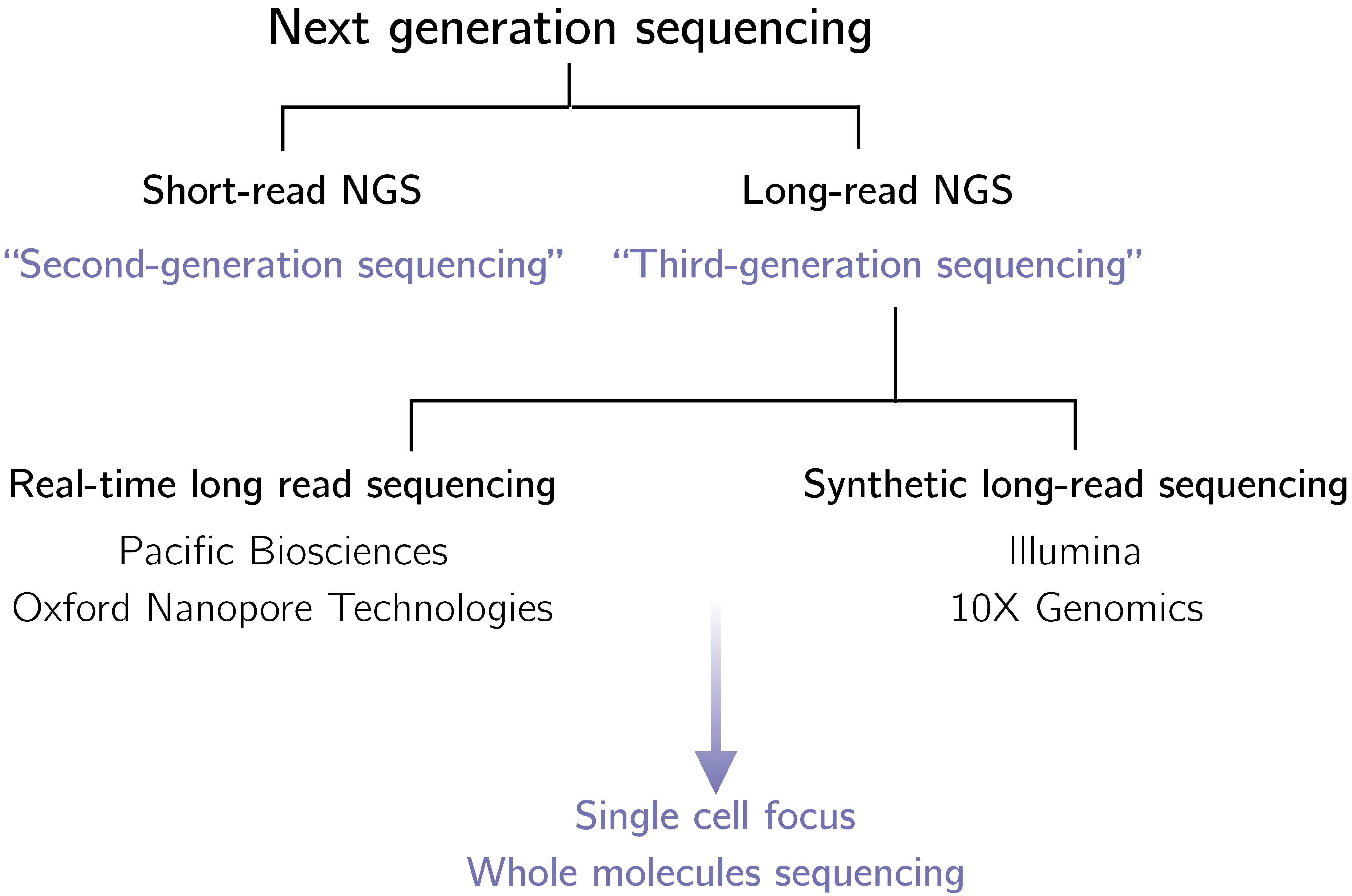
Sequencing by synthesis

Illumina/Solexa



Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Next generation sequencing technologies and limitations



Next generation sequencing

Short-read NGS

“Second-generation sequencing”

Long-read NGS

“Third-generation sequencing”

Real-time long read sequencing

Pacific Biosciences

Oxford Nanopore Technologies

Synthetic long-read sequencing

Illumina

10X Genomics

Single cell focus

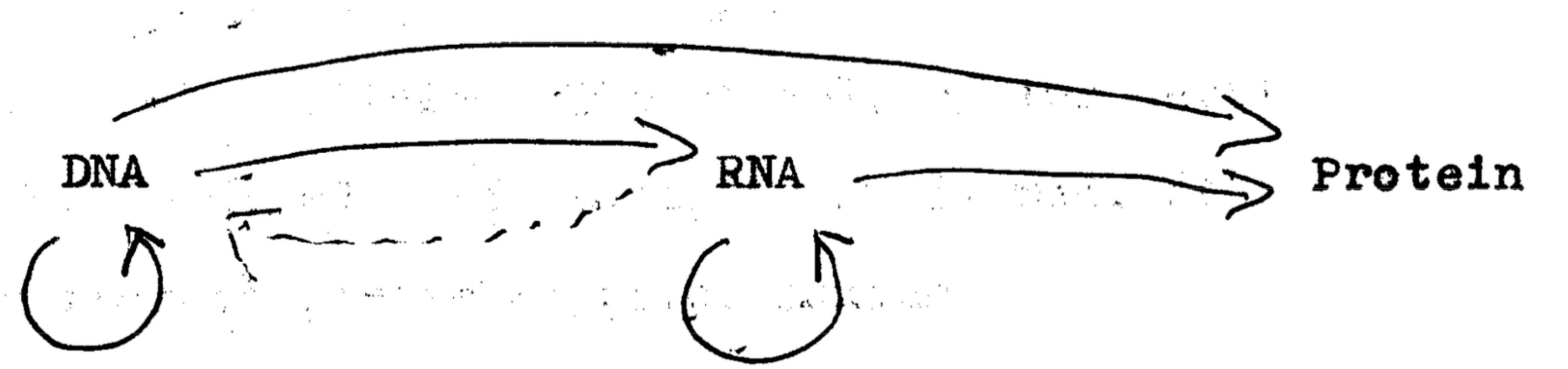
Whole molecules sequencing

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Sequencing techniques

Central dogma of molecular biology (Crick F. 1958)

Information flow



Transcription

Translation

Whole genome sequencing

Whole exome sequencing

HiC-Seq

ChIP-Seq

ATAC-Seq

...

scRNA-Seq

RNA-Seq

Ribo-Seq

SLAM-Seq

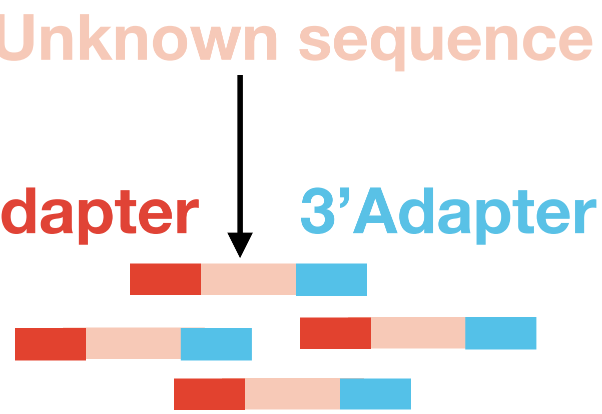
...

DNA

RNA

Illumina sequencing by synthesis

1 Library preparation



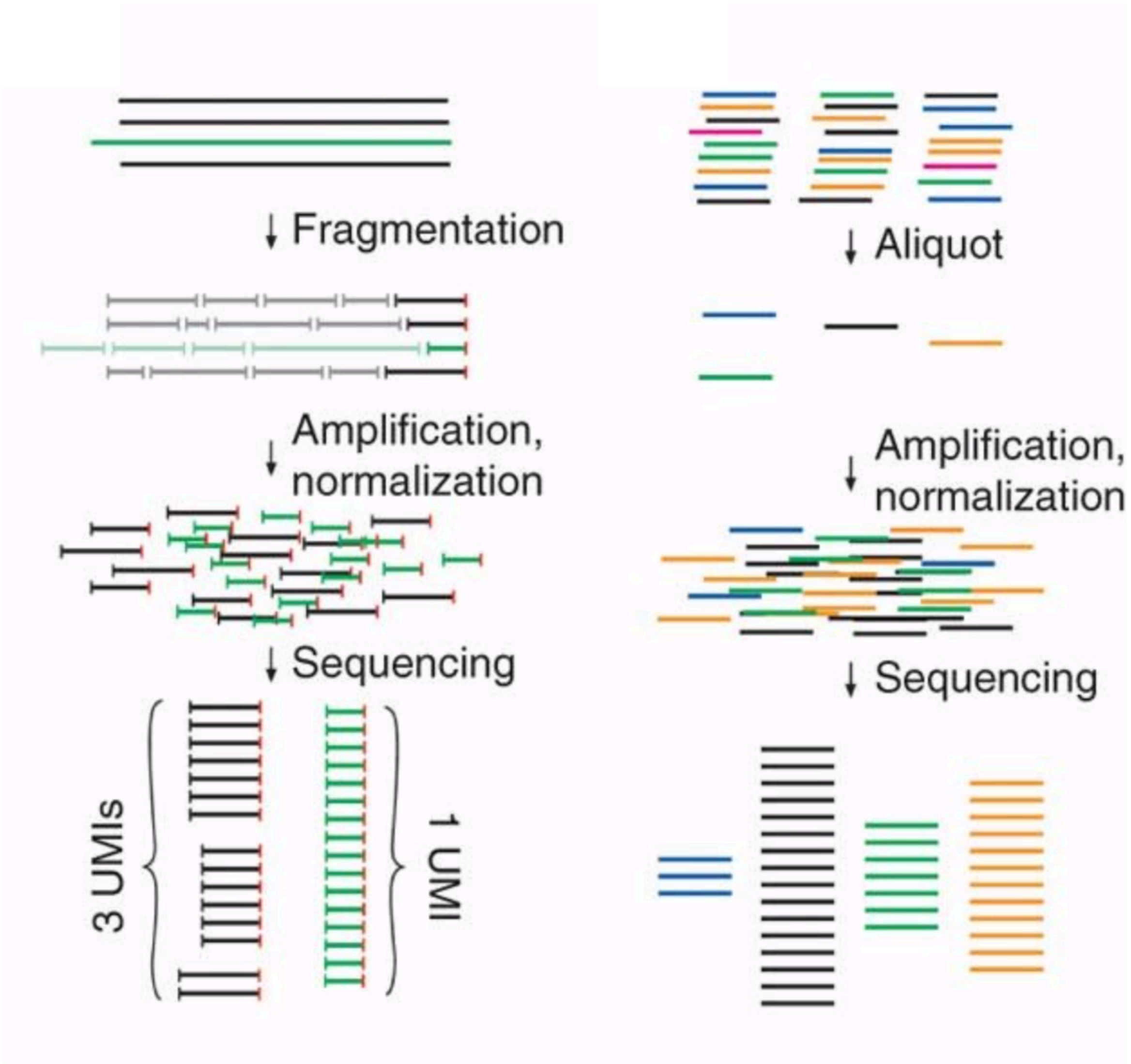
NOTE 1: High quality material needed for high quality experiment!

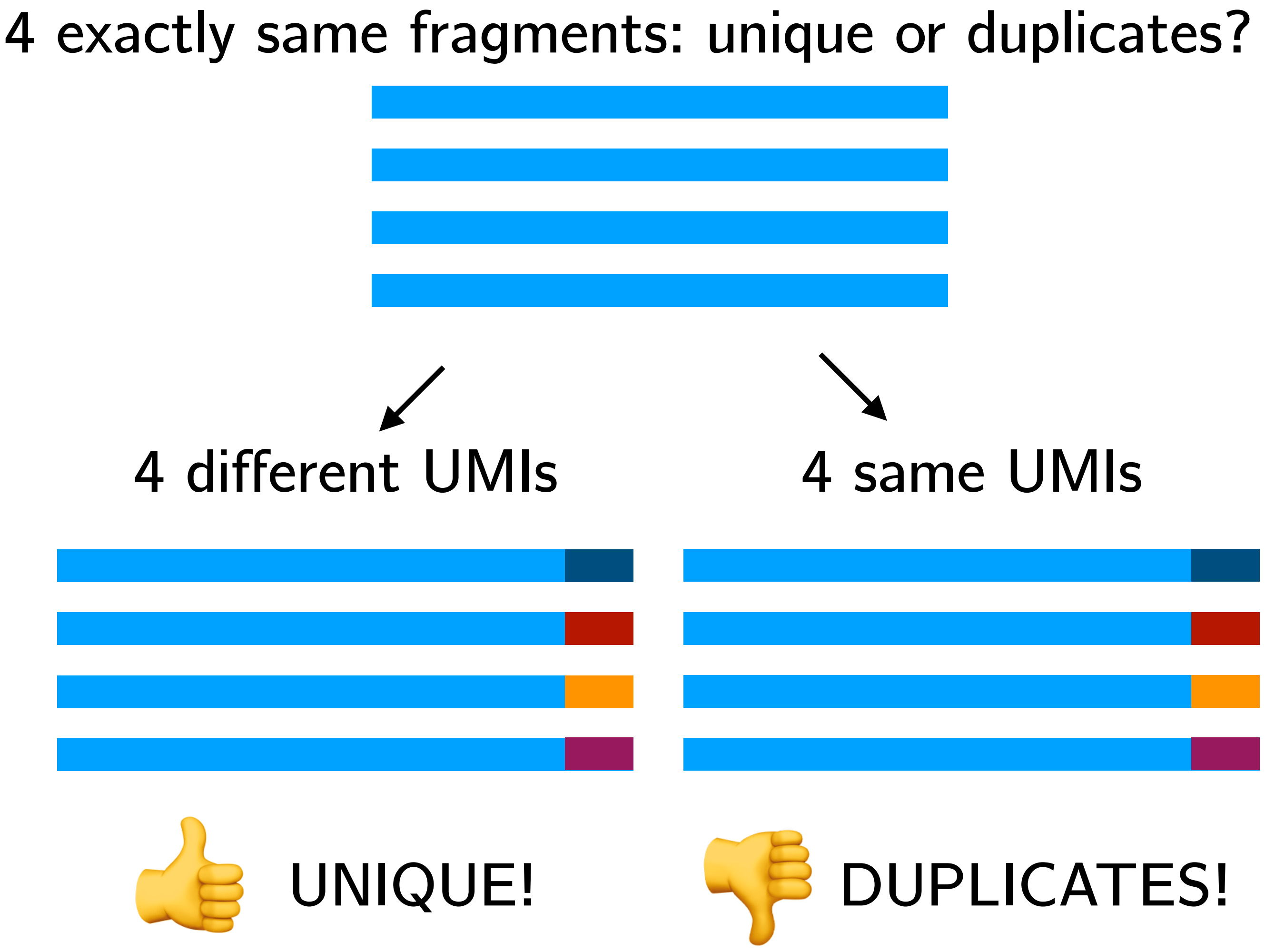
NOTE 2: Final step of library preparation is amplification. Some products are preferentially amplified, which introduces **library amplification bias.**

- Fewer cycles - fewer bias
- **Unique molecular identifiers**: oligonucleotides labels to identify duplicated fragments

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Unique molecular identifiers (UMIs)





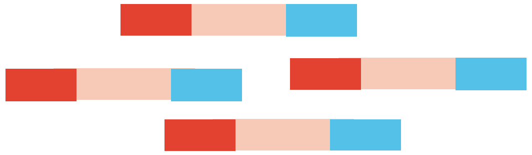
UMIs help to identify library amplification bias and quantify unique fragments (identical fragments with the same UMIs are likely to be duplicates)

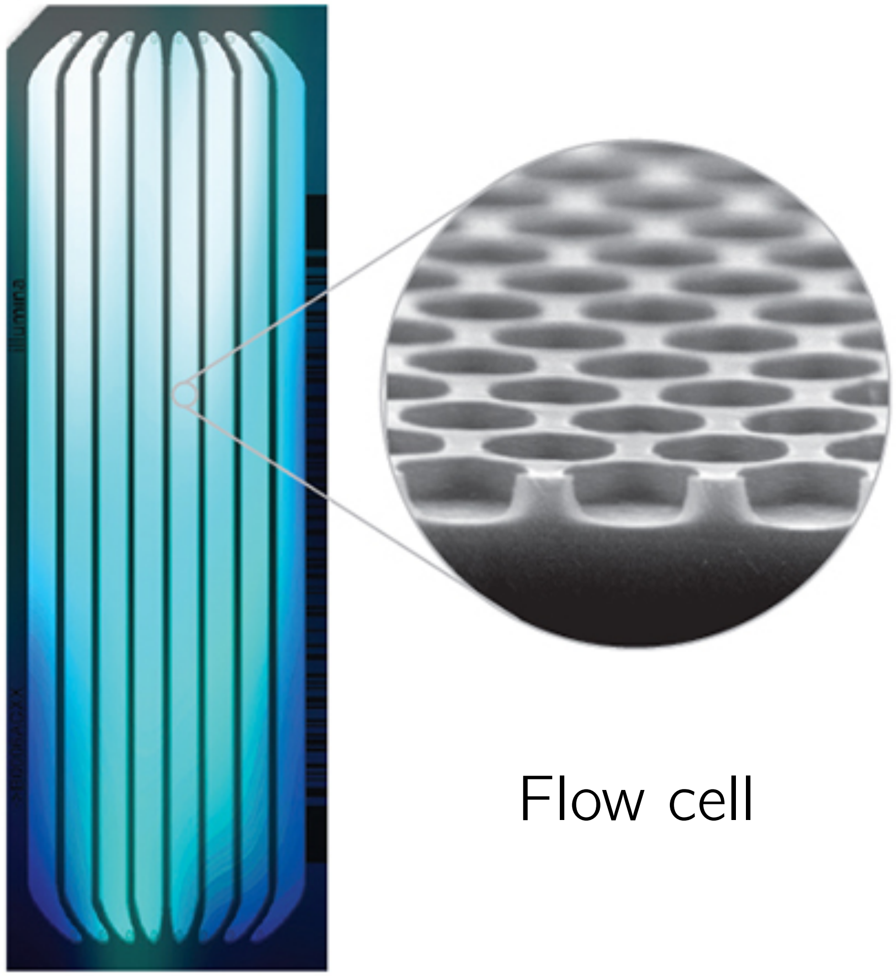
Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nature Methods, 9(1), 72–74.

Illumina sequencing by synthesis

Based on the Solexa technology developed by **Shankar Balasubramanian** and **David Klenerman** at the University of Cambridge (1998)

1 **Library preparation**



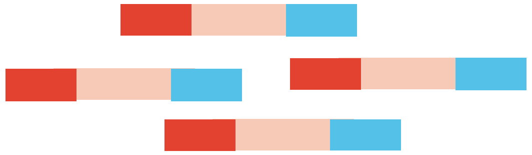


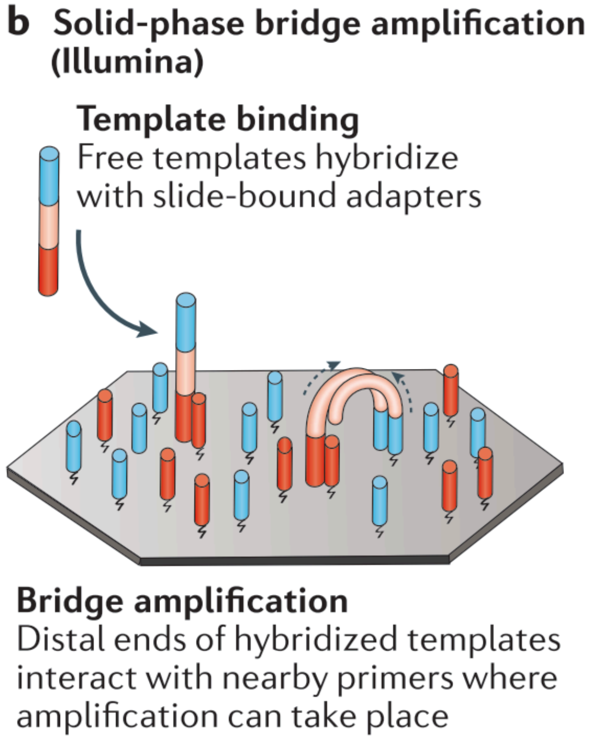
Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Illumina sequencing by synthesis

Based on the Solexa technology developed by **Shankar Balasubramanian** and **David Klenerman** at the University of Cambridge (1998)

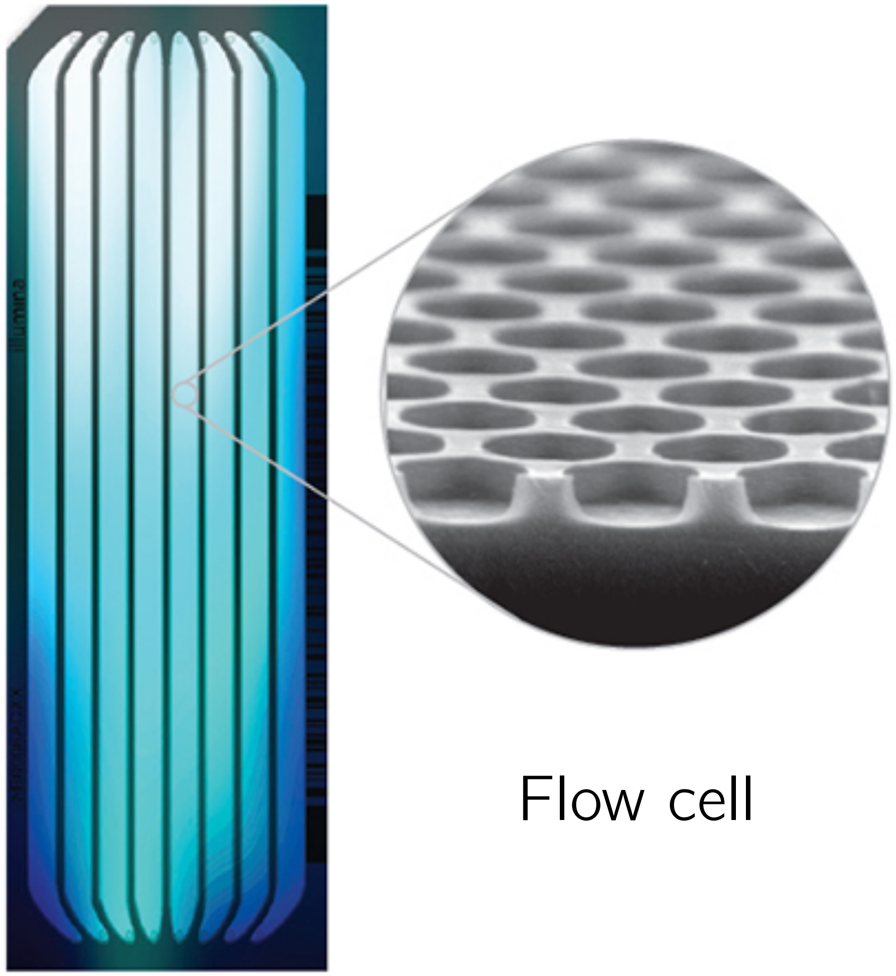
1 **Library preparation**




b Solid-phase bridge amplification (Illumina)
Template binding
Free templates hybridize with slide-bound adapters
Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place

2

Sequencing by synthesis



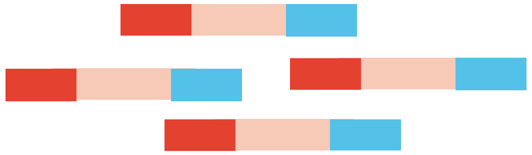
Flow cell

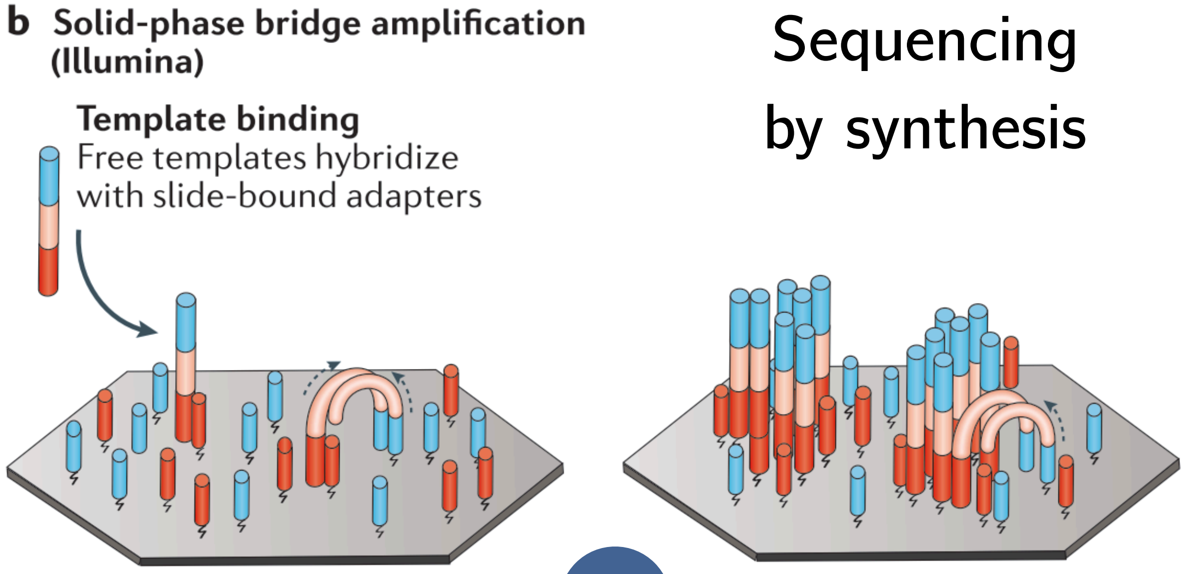
Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Illumina sequencing by synthesis

Based on the Solexa technology developed by **Shankar Balasubramanian** and **David Klenerman** at the University of Cambridge (1998)

1 Library preparation





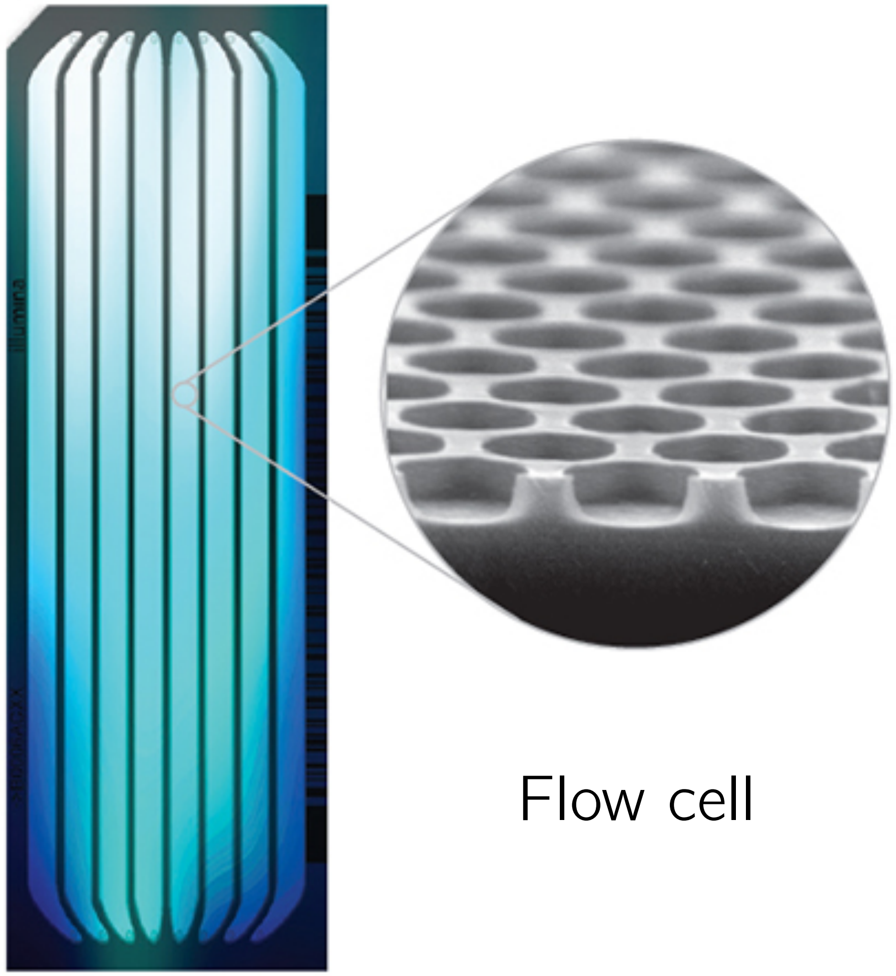
b Solid-phase bridge amplification (Illumina)

Template binding
Free templates hybridize with slide-bound adapters

2 **Bridge amplification**
Distal ends of hybridized templates interact with nearby primers where amplification can take place

3 **Cluster generation**
After several rounds of amplification, 100–200 million clonal clusters are formed

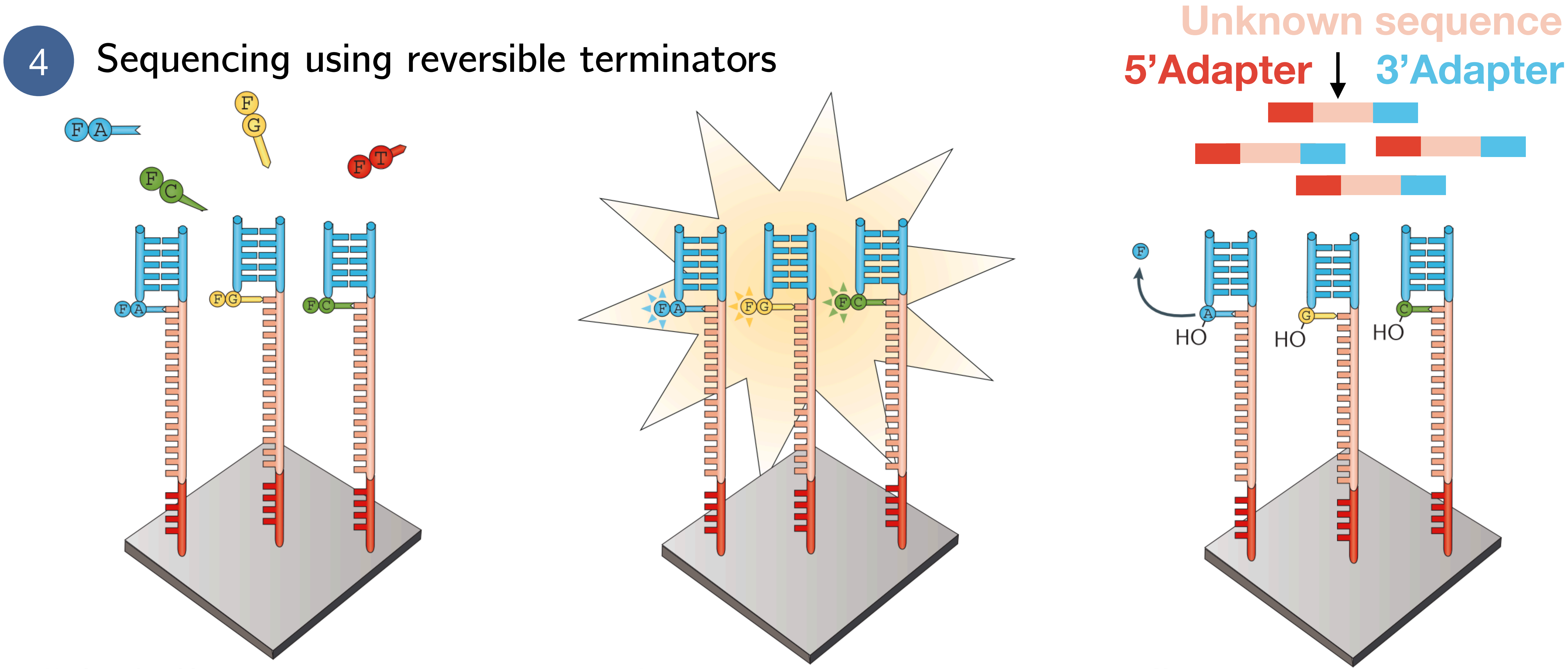
Sequencing by synthesis



Flow cell

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Illumina sequencing by synthesis



4 Sequencing using reversible terminators

Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

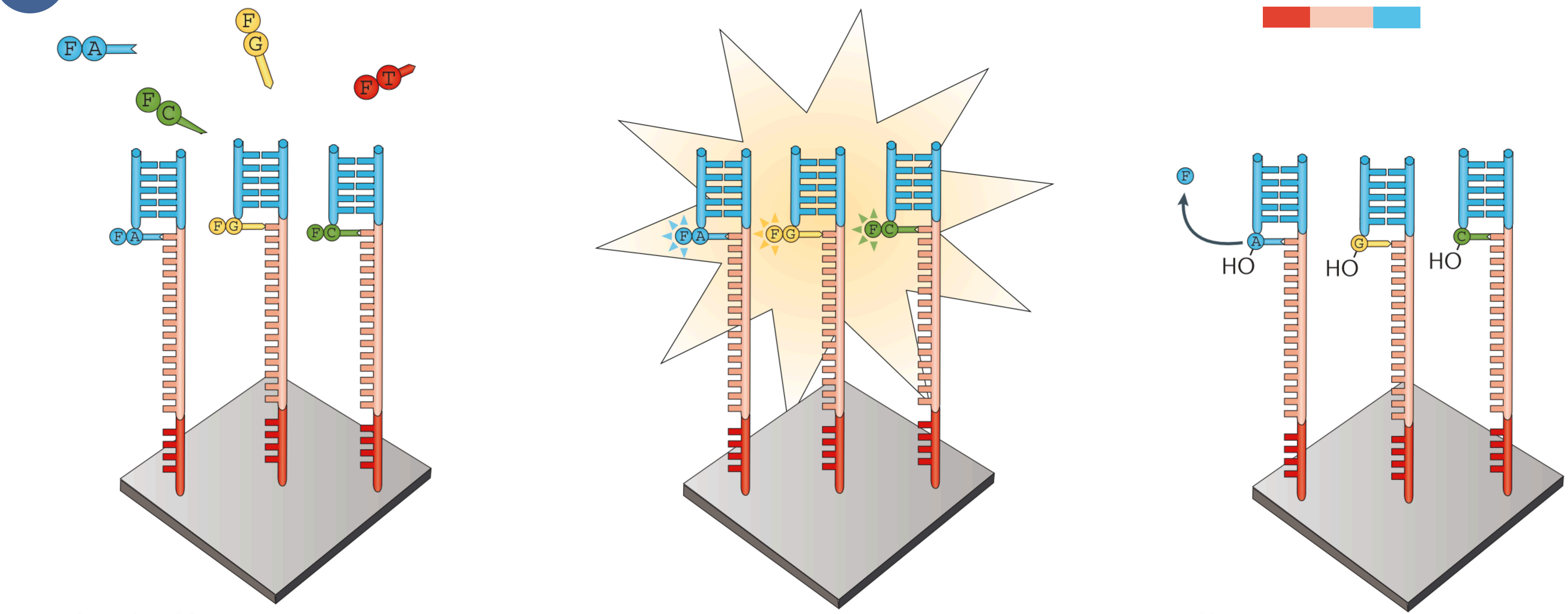
Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Illumina sequencing by synthesis

4 Sequencing using reversible terminators



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

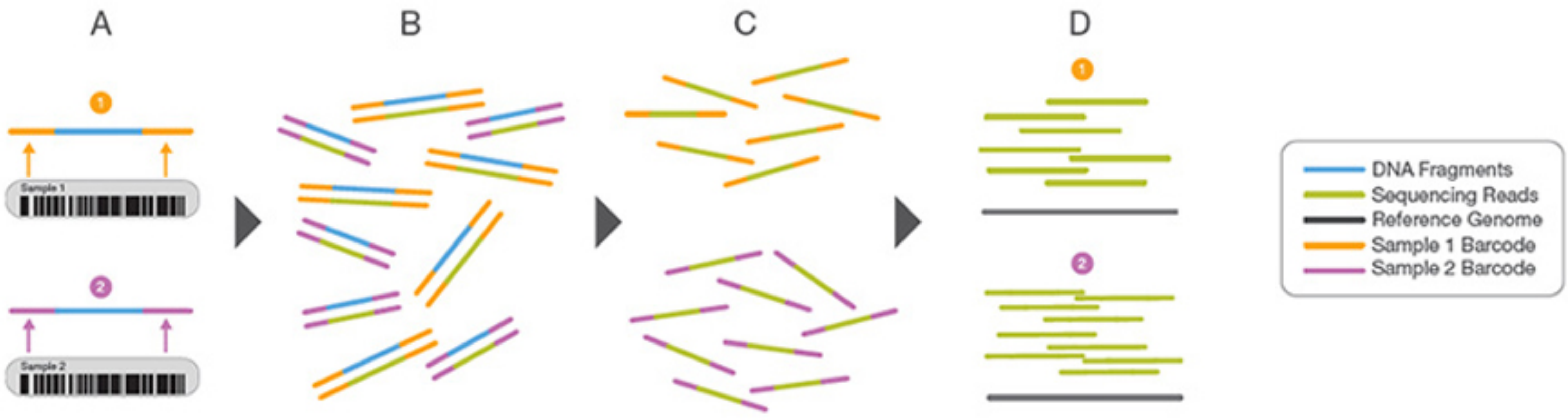
5 Output: sequence saved in FASTQ format

6 Bioinformatic analysis: quality check, alignment and data analysis

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Multiplexing

- Multiplexing gives the ability to sequence multiple samples at the same time
- Blocks against possible technical bias caused by differences between flow cell lanes
- Useful when sequencing small genomes or specific genomic regions.

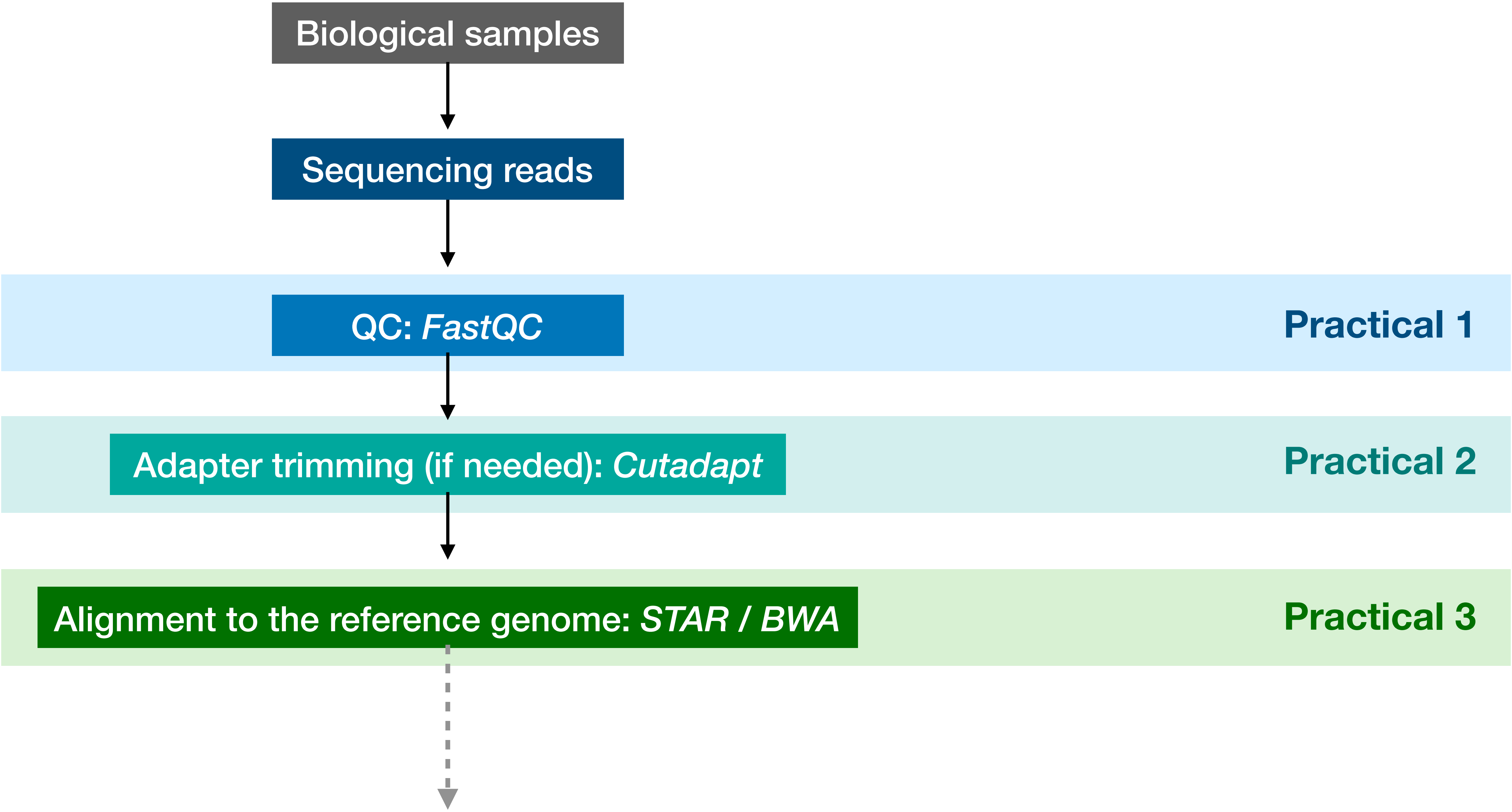


Different barcode adaptors are ligated to different samples.

Reads de-multiplexed after sequencing.

Source: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/multiplex-sequencing.html

Workflow for today



Biological samples

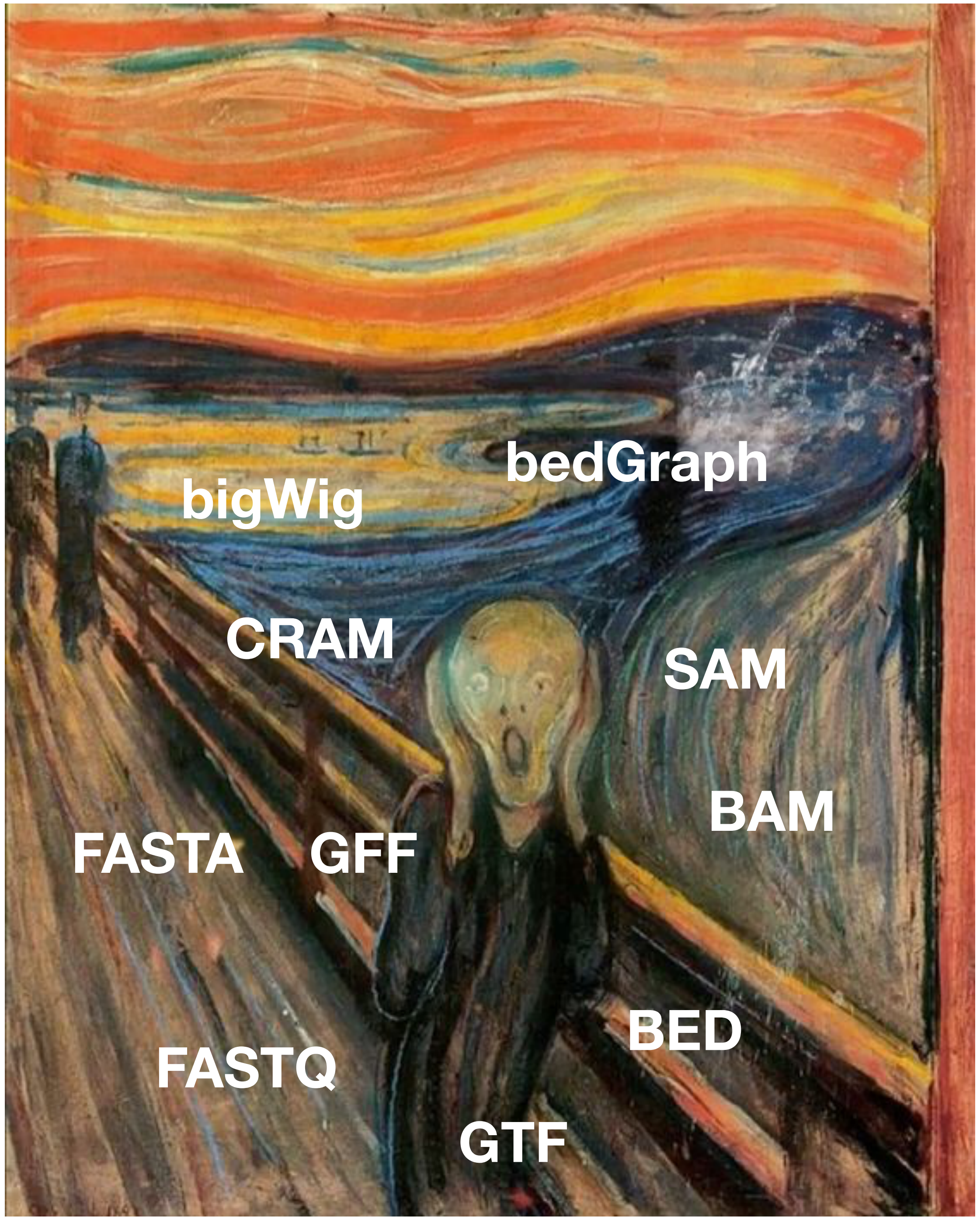
Sequencing reads

QC: *FastQC* — Practical 1

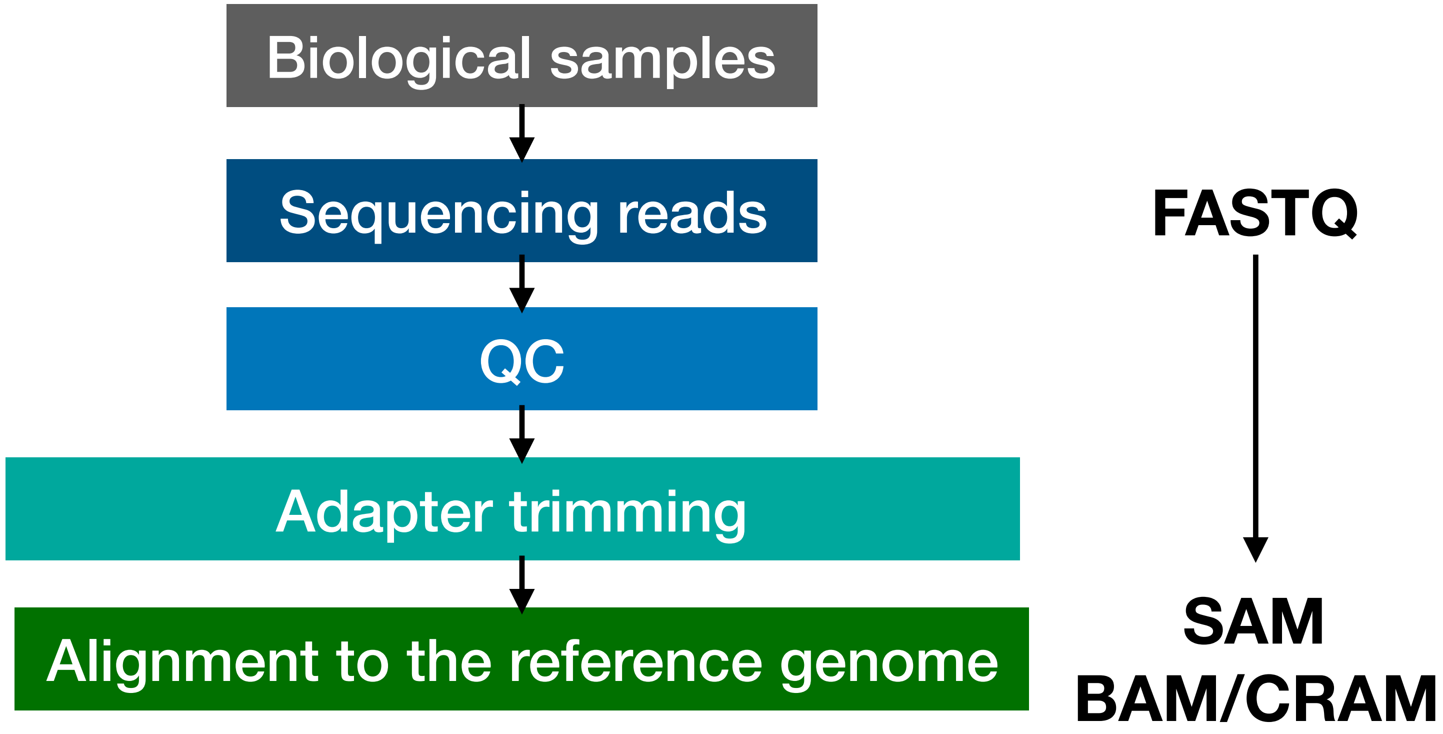
Adapter trimming (if needed): *Cutadapt* — Practical 2

Alignment to the reference genome: *STAR / BWA* — Practical 3

Common file formats: why so many?



Different formats - different informations

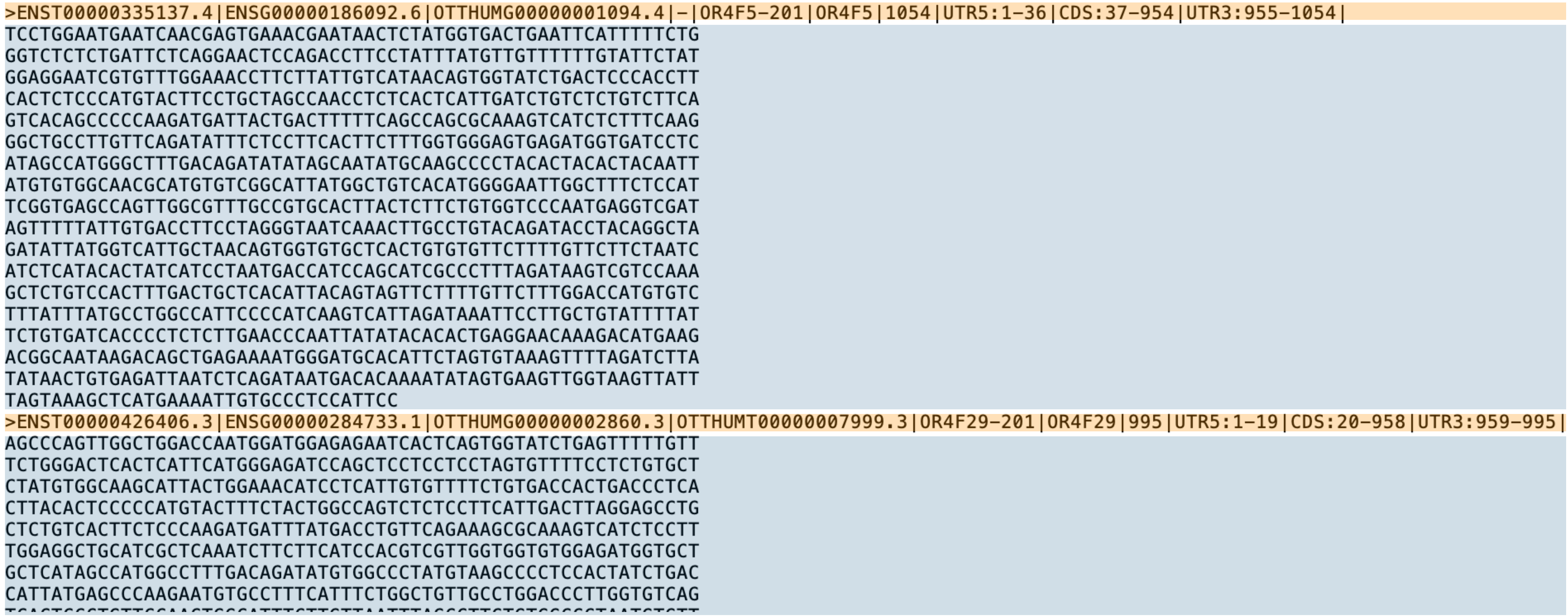


Nucleotide/peptide sequences: FASTA

A sequence in FASTA format consists of:

1st line starting with “>” followed by the sequence name

2nd line with the sequence itself



```
>ENST00000335137.4|ENSG00000186092.6|OTTHUMG00000001094.4|-|OR4F5-201|OR4F5|1054|UTR5:1-36|CDS:37-954|UTR3:955-1054|
TCCTGGAATGAATCAACGAGTGAAACGAATAACTCTATGGTGACTGAATTCATTTTTCTG
GGTCTCTCTGATTCTCAGGAACTCCAGACCTTCCTATTTATGTTGTTTTTTGTATTCTAT
GGAGGAATCGTGTTTGGAAACCTTCTTATTGTCATAACAGTGGTATCTGACTCCCACCTT
CACTCTCCCATGTACTTCCTGCTAGCCAACCTCTCACTCATTGATCTGTCTCTGTCTTCA
GTCACAGCCCCCAAGATGATTACTGACTTTTTCAGCCAGCGCAAAGTCATCTCTTTCAAG
GGCTGCCTTGTTCAGATATTTCTCCTTCACTTCTTTGGTGGGAGTGAGATGGTGATCCTC
ATAGCCATGGGCTTTGACAGATATATAGCAATATGCAAGCCCCTACACTACACTACAATT
ATGTGTGGCAACGCATGTGTCGGCATTATGGCTGTCACATGGGGAATTGGCTTTCTCCAT
TCGGTGAGCCAGTTGGCGTTTGCCGTGCACTTACTCTTCTGTGGTCCCAATGAGGTCGAT
AGTTTTTATTGTGACCTTCCTAGGGTAATCAAACTTGCCTGTACAGATACCTACAGGCTA
GATATTATGGTCATTGCTAACAGTGGTGTGCTCACTGTGTGTTCTTTTGTTCTTCTAATC
ATCTCATACACTATCATCCTAATGACCATCCAGCATCGCCCTTTAGATAAGTCGTCCAAA
GCTCTGTCCACTTTGACTGCTCACATTACAGTAGTTCTTTTGTTCTTTGGACCATGTGTC
TTTATTTATGCCTGGCCATTCCCCATCAAGTCATTAGATAAATTCCTTGCTGTATTTTAT
TCTGTGATCACCCCTCTCTTGAACCCAATTATATACACACTGAGGAACAAAGACATGAAG
ACGGCAATAAGACAGCTGAGAAAATGGGATGCACATTCTAGTGTAAAGTTTTAGATCTTA
TATAACTGTGAGATTAATCTCAGATAATGACACAAAATATAGTGAAGTTGGTAAGTTATT
TAGTAAAGCTCATGAAAATTGTGCCCTCCATTCC
>ENST00000426406.3|ENSG00000284733.1|OTTHUMG00000002860.3|OTTHUMT00000007999.3|OR4F29-201|OR4F29|995|UTR5:1-19|CDS:20-958|UTR3:959-995|
AGCCCAGTTGGCTGGACCAATGGATGGAGAGAATCACTCAGTGGTATCTGAGTTTTTGTT
TCTGGGACTCACTCATTCATGGGAGATCCAGCTCCTCCTCCTAGTGTTTTCCTCTGTGCT
CTATGTGGCAAGCATTACTGGAAACATCCTCATTGTGTTTTCTGTGACCACTGACCCTCA
CTTACACTCCCCCATGTACTTTCTACTGGCCAGTCTCTCCTTCATTGACTTAGGAGCCTG
CTCTGTCACTTCTCCCAAGATGATTTATGACCTGTTCAGAAAGCGCAAAGTCATCTCCTT
TGGAGGCTGCATCGCTCAAATCTTCTTCATCCACGTCGTTGGTGGTGTGGAGATGGTGCT
GCTCATAGCCATGGCCTTTGACAGATATGTGGCCCTATGTAAGCCCCTCCACTATCTGAC
CATTATGAGCCCAAGAATGTGCCTTTCATTTCTGGCTGTTGCCTGGACCCTTGGTGTCAG
```

A single FASTA file may contain > 1 sequence

Unaligned sequence: FASTQ

Unaligned sequence (reads) files generated from NGS machines

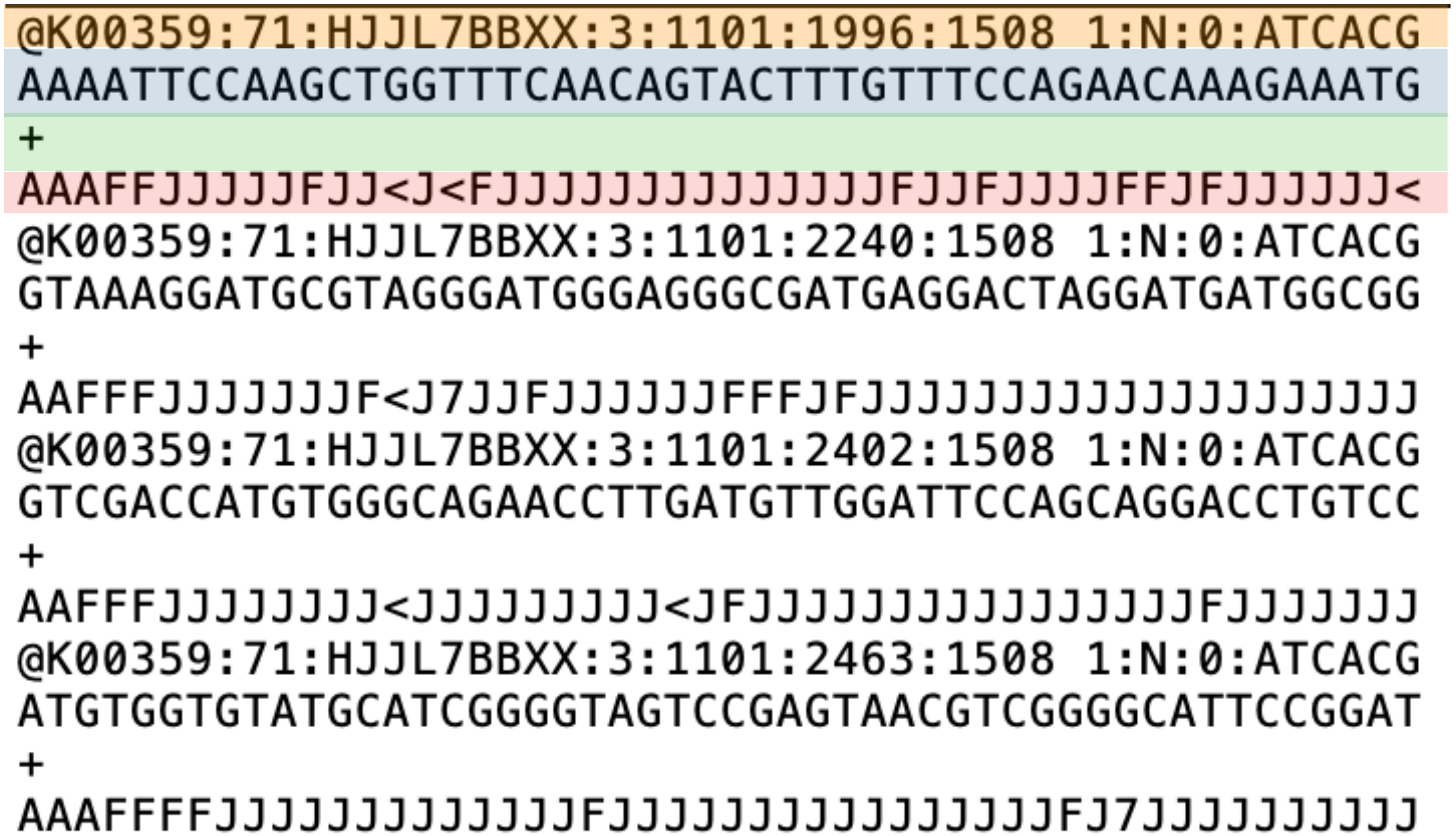
A sequence in FASTQ format consists of:

1st line starting with “@” followed by the read identifier.

2nd line with the sequence itself.

3rd line “+”

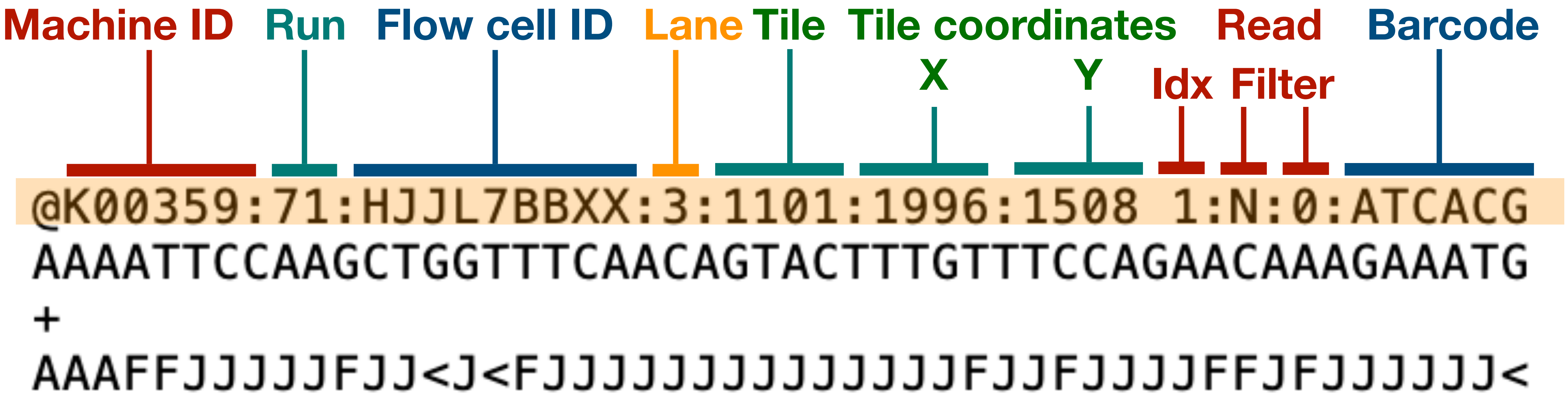
4th line Quality scores encoded as ASCII characters



```
@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG
AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAAGAAATG
+
AAAFFJJJJJFJJ<J<FJJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<
@K00359:71:HJJL7BBXX:3:1101:2240:1508 1:N:0:ATCACG
GTAAAGGATGCGTAGGGATGGGAGGGCGATGAGGACTAGGATGATGGCGG
+
AAFFFJJJJJJJF<J7JJFJJJJJJFFFJFJJJJJJJJJJJJJJJJJJJJ
@K00359:71:HJJL7BBXX:3:1101:2402:1508 1:N:0:ATCACG
GTCGACCATGTGGGCAGAACCTTGATGTTGGATTCCAGCAGGACCTGTCC
+
AAFFFJJJJJJJJ<JJJJJJJJJ<JFJJJJJJJJJJJJJJJJFJJJJJJJ
@K00359:71:HJJL7BBXX:3:1101:2463:1508 1:N:0:ATCACG
ATGTGGTGTATGCATCGGGGTAGTCCGAGTAACGTCGGGGCATTCCGGAT
+
AAAFFFFJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJFJ7JJJJJJJJJJ
```

Unaligned sequence: FASTQ

FASTQ header decoded (Illumina example):



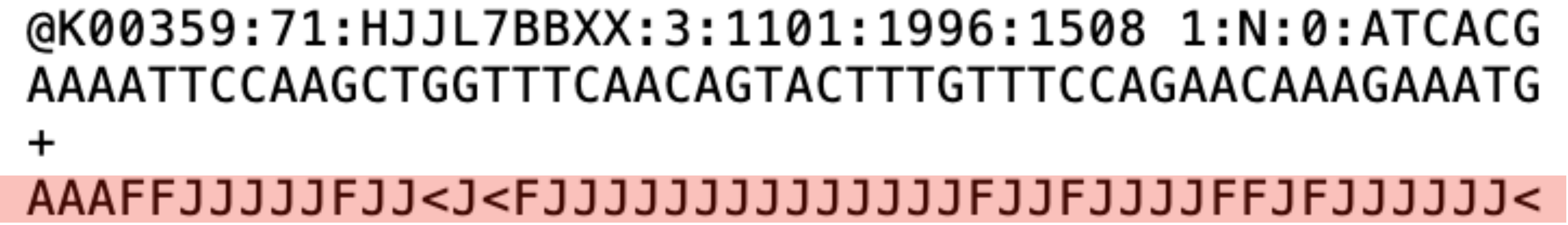
```
@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG
AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAAGAAATG
+
AAAFFJJJJJFJJ<J<FJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<
```

Unaligned sequence: FASTQ

Quality scores come after the "+" line

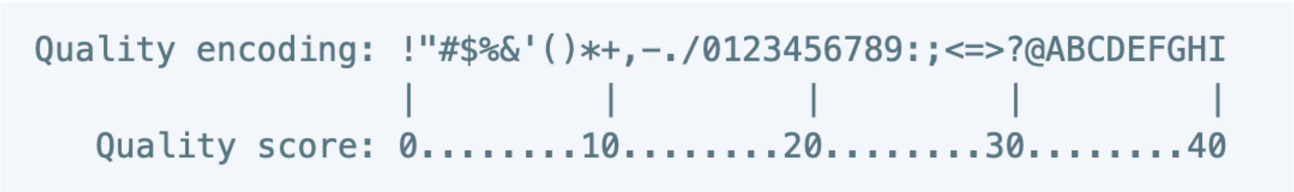
Quality Q is proportional to -log10 probability of sequence base being wrong e

$$Q = -10 \cdot log_{10}(e)$$



```
@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG
AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAAGAAATG
+
AAAFFJJJJJFJJ<J<FJJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<
```

Encoded in ASCII to save space:



```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0........10........20........30........40
```

Used in quality assessment and downstream analysis

SAM - Sequence Alignment Map

Unaligned sequence files generated from NGS machines are mapped to a reference genome to produce aligned sequence:

FASTQ(unaligned sequences) → SAM (aligned sequences)

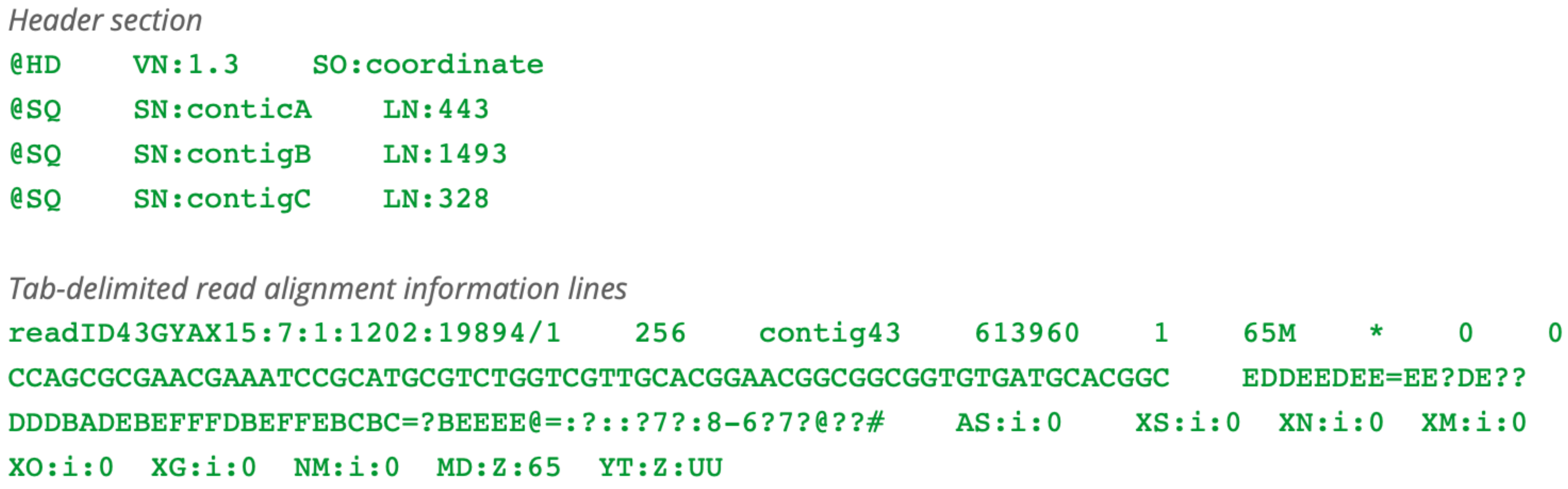
FASTA + quality FASTQ + location

SAM:

– Standard format for aligned sequence data

– Recognised by majority of software and browsers

– Starts with a header section followed by alignment information as tab separated lines for each read.



Header section

```
@HD     VN:1.3    SO:coordinate
@SQ     SN:conticA      LN:443
@SQ     SN:contigB      LN:1493
@SQ     SN:contigC      LN:328
```

Tab-delimited read alignment information lines

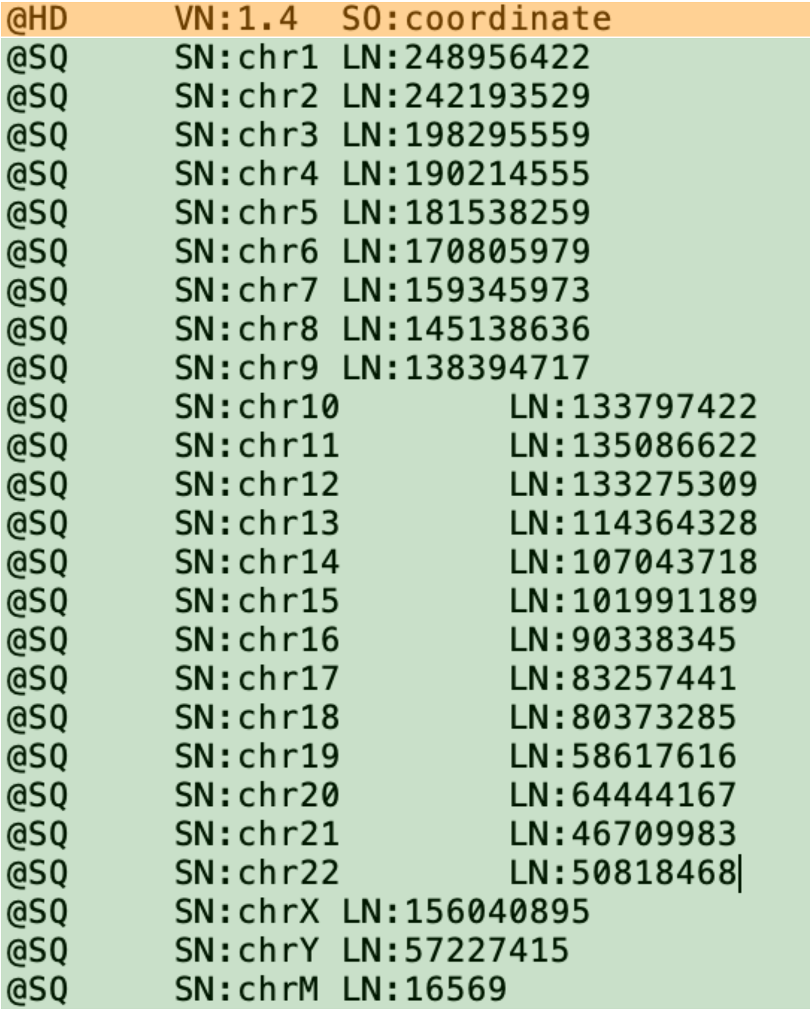
```
readID43GYAX15:7:1:1202:19894/1    256    contig43    613960    1    65M    *    0    0
CCAGCGCGAACGAAATCCGCATGCGTCTGGTCGTTGCACGGAACGGCGGCGGTGTGATGCACGGC    EDDEEDEE=EE?DE??
DDDBADEBEFFFDBEFFEBCBC=?BEEEE@=:?::?7?:8-6?7?@??#    AS:i:0    XS:i:0   XN:i:0   XM:i:0
XO:i:0   XG:i:0   NM:i:0   MD:Z:65   YT:Z:UU
```

http://www.metagenomics.wiki/tools/samtools/bam-sam-file-format

SAM - Sequence Alignment Map

SAM header

- Header lines start with ‘@’



```
@HD	VN:1.4	SO:coordinate
@SQ	SN:chr1	LN:248956422
@SQ	SN:chr2	LN:242193529
@SQ	SN:chr3	LN:198295559
@SQ	SN:chr4	LN:190214555
@SQ	SN:chr5	LN:181538259
@SQ	SN:chr6	LN:170805979
@SQ	SN:chr7	LN:159345973
@SQ	SN:chr8	LN:145138636
@SQ	SN:chr9	LN:138394717
@SQ	SN:chr10		LN:133797422
@SQ	SN:chr11		LN:135086622
@SQ	SN:chr12		LN:133275309
@SQ	SN:chr13		LN:114364328
@SQ	SN:chr14		LN:107043718
@SQ	SN:chr15		LN:101991189
@SQ	SN:chr16		LN:90338345
@SQ	SN:chr17		LN:83257441
@SQ	SN:chr18		LN:80373285
@SQ	SN:chr19		LN:58617616
@SQ	SN:chr20		LN:64444167
@SQ	SN:chr21		LN:46709983
@SQ	SN:chr22		LN:50818468
@SQ	SN:chrX	LN:156040895
@SQ	SN:chrY	LN:57227415
@SQ	SN:chrM	LN:16569
```

File-level metadata

VN: format version, SO: sorting order

Reference sequence dictionary

SN : name (eg. chr1), LN : length

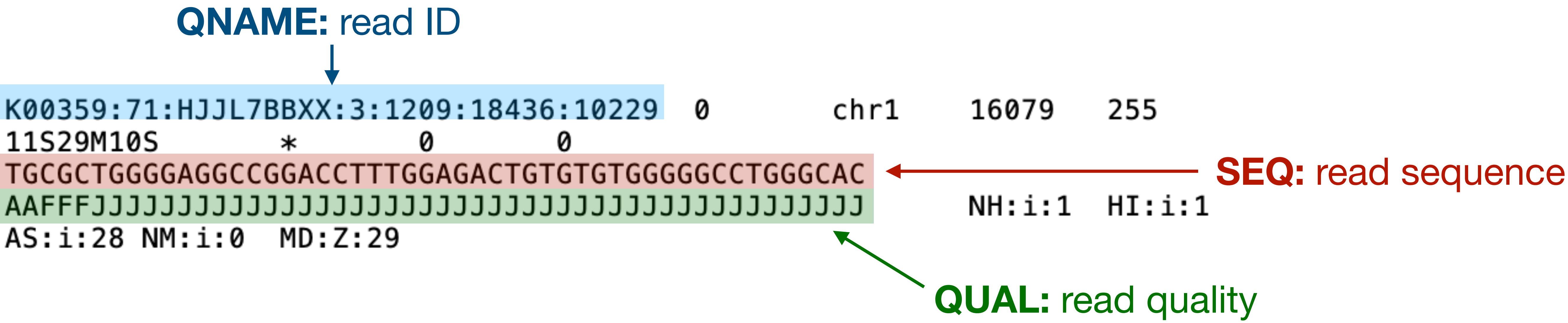
Full format specification:
https://samtools.github.io/hts-specs/SAMv1.pdf

SAM - Sequence Alignment Map

Aligned reads

- Organised as tab-delimited text
- Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

Read informations (as in FASTQ):



QNAME: read ID

```
K00359:71:HJJL7BBXX:3:1209:18436:10229  0       chr1    16079   255
11S29M10S       *       0       0
TGCGCTGGGGAGGCCGGACCTTTGGAGACTGTGTGTGGGGGCCTGGGCAC
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ      NH:i:1  HI:i:1
AS:i:28 NM:i:0  MD:Z:29
```

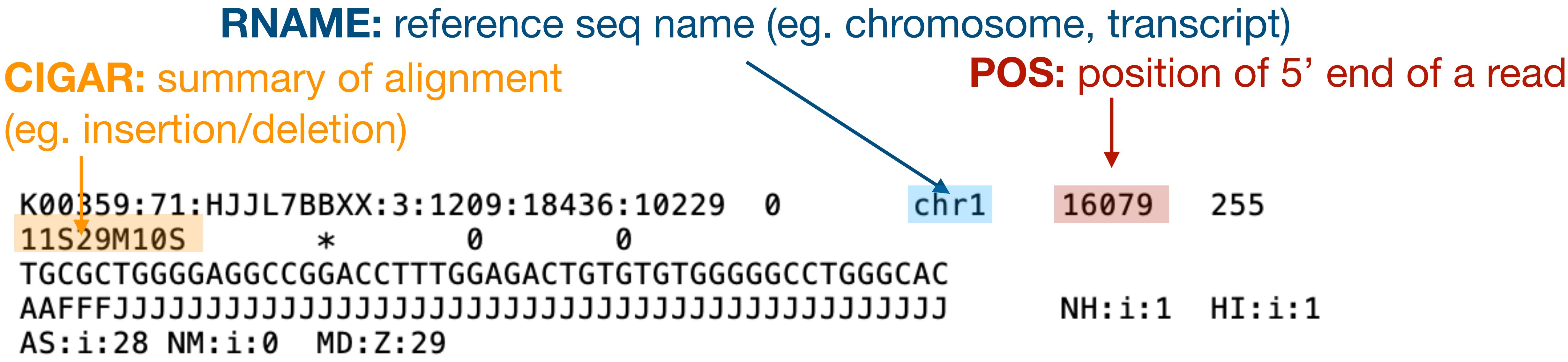
SEQ: read sequence

QUAL: read quality

SAM - Sequence Alignment Map

Aligned reads

- Organised as tab-delimited text
- Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.



RNAME: reference seq name (eg. chromosome, transcript)

CIGAR: summary of alignment (eg. insertion/deletion)

POS: position of 5’ end of a read

```
K00359:71:HJJL7BBXX:3:1209:18436:10229  0       chr1    16079   255
11S29M10S       *       0       0
TGCGCTGGGGAGGCCGGACCTTTGGAGACTGTGTGGGGGCCTGGGCAC
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ      NH:i:1  HI:i:1
AS:i:28 NM:i:0  MD:Z:29
```

CIGAR string encoding:

50M - continuous match of 50 bases

28M1D72M - 28 bases continuously match, 1 deletion from reference, 72 base match

Full format specification:
https://samtools.github.io/hts-specs/SAMv1.pdf

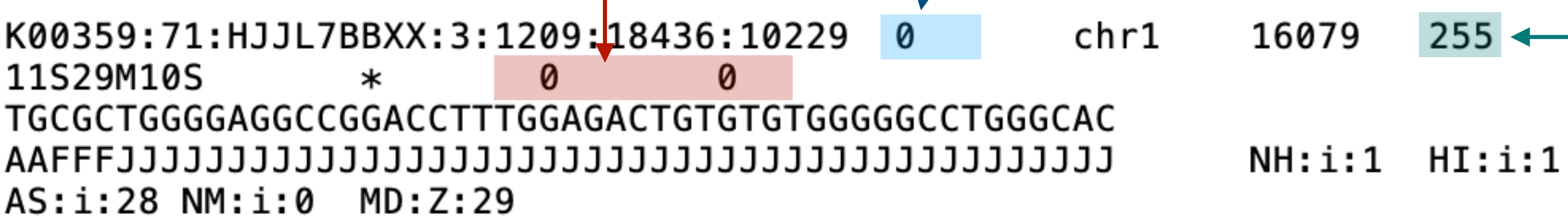
SAM - Sequence Alignment Map

Aligned reads

- Organised as tab-delimited text
- Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

Bit flag - TRUE/FALSE for pre-defined read criteria, like: is it paired? duplicate?

Paired read position and insert size



```
K00359:71:HJJL7BBXX:3:1209:18436:10229  0       chr1    16079   255
11S29M10S       *       0       0
TGCGCTGGGGAGGCCGGACCTTTGGAGACTGTGTGTGGGGGCCTGGGCAC
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ      NH:i:1  HI:i:1
AS:i:28 NM:i:0  MD:Z:29
```

Mapping quality

Flags explained:

https://broadinstitute.github.io/picard/explain-flags.html

Compressed aligned sequences - BAM and CRAM format

SAM files can be large, so to save space people usually store some compressed versions of them instead:

BAM

- Binary SAM file
- You also need to store an index file

CRAM

- Another way to compress alignment files
- The compression is driven by the reference the sequence data is aligned to, so it is very important that the exact same reference sequence is used for compression and decompression
- Typically 40-50% space saving compared to BAM files
- Full compatibility with BAM files
- For further information: http://samtools.github.io/hts-specs/

10 min break!