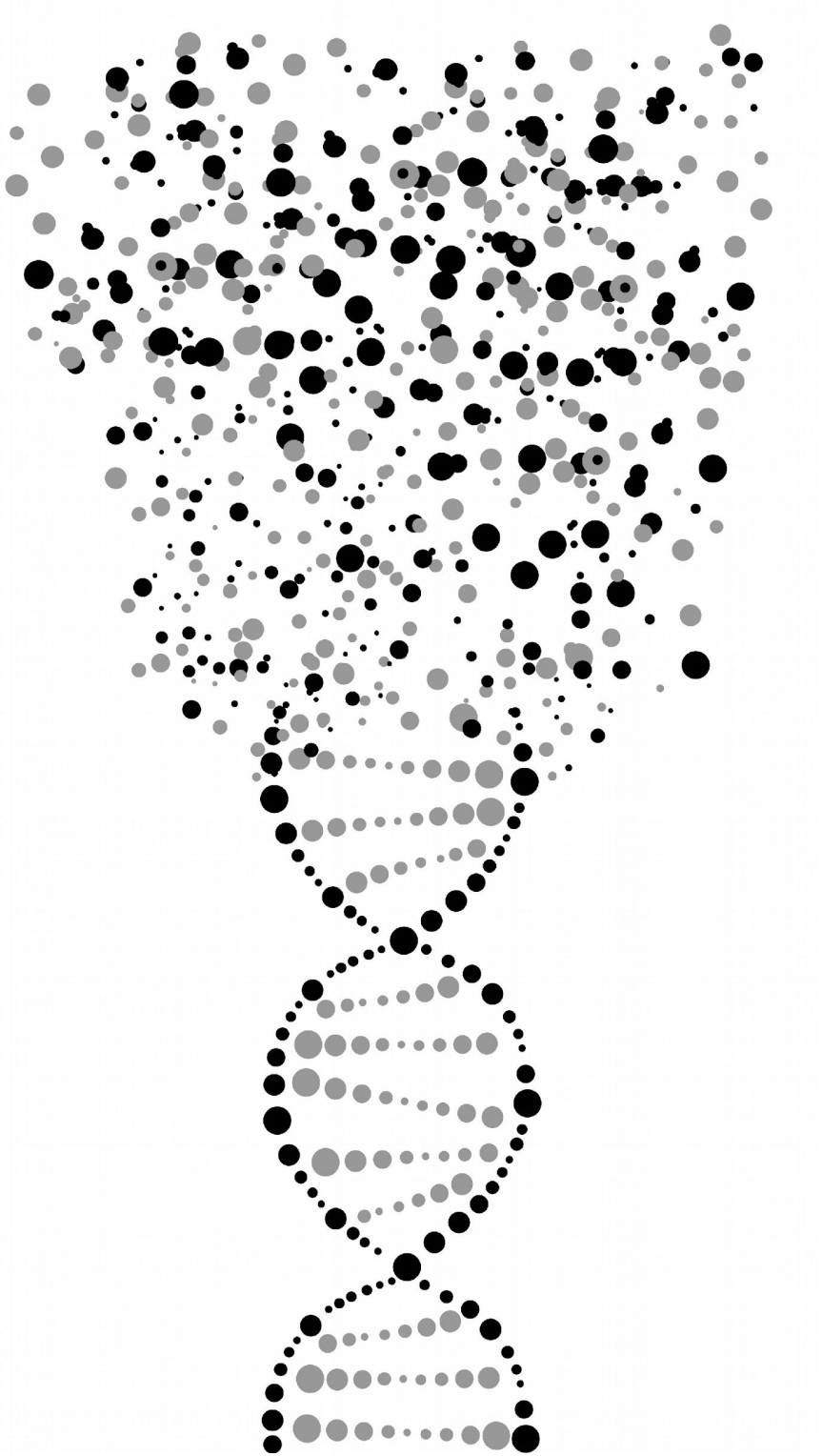
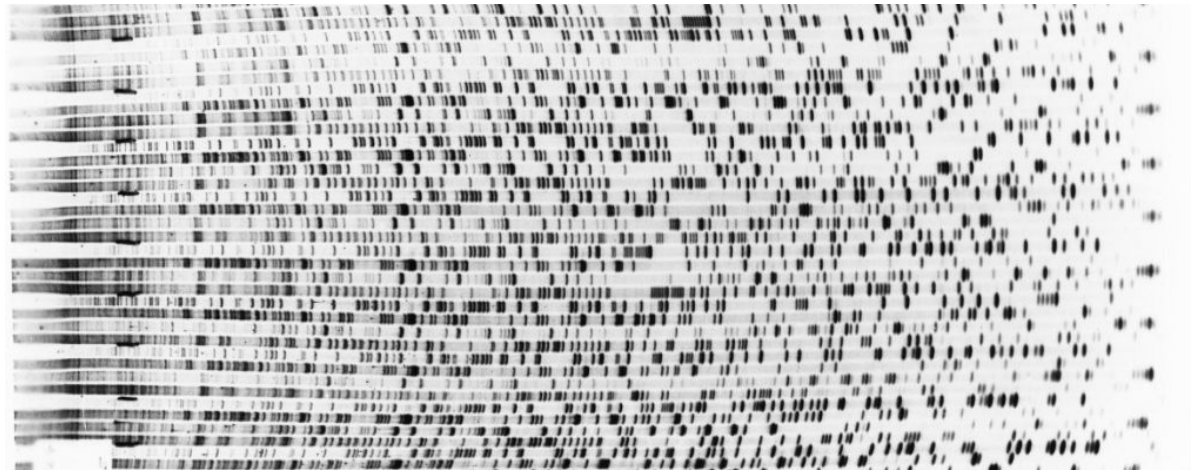
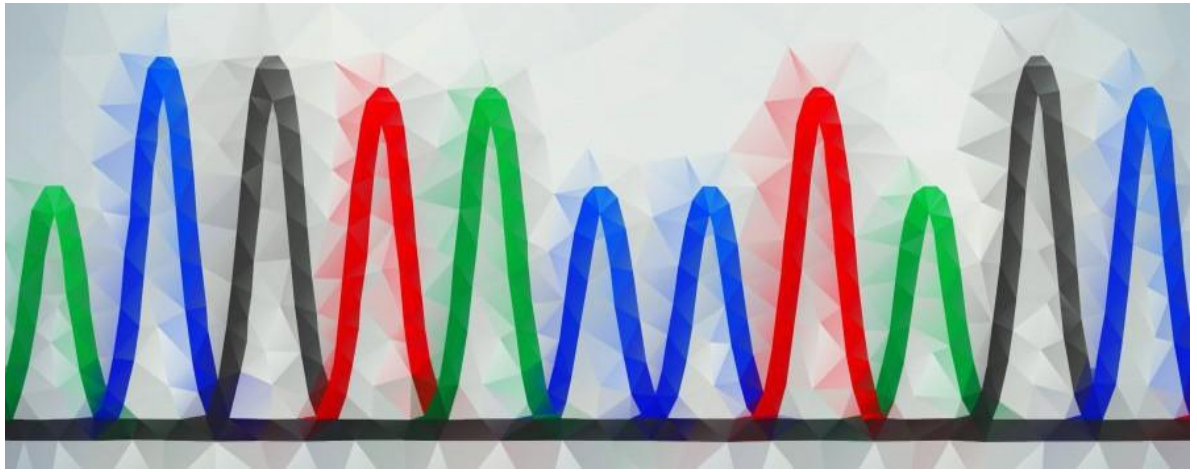
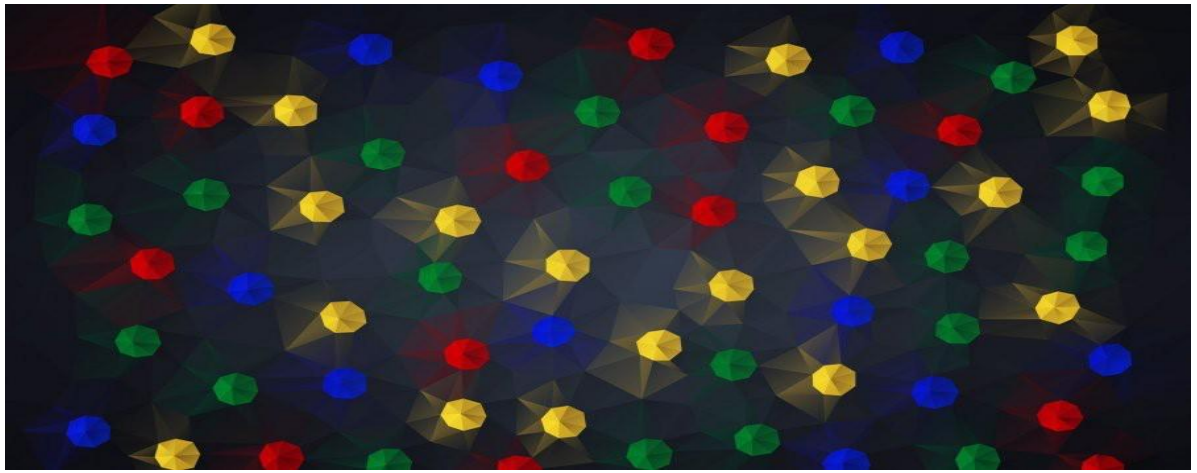


Next Generation Sequencing

An Introduction







The future of SEQUENCING

evermore
MASSIVE

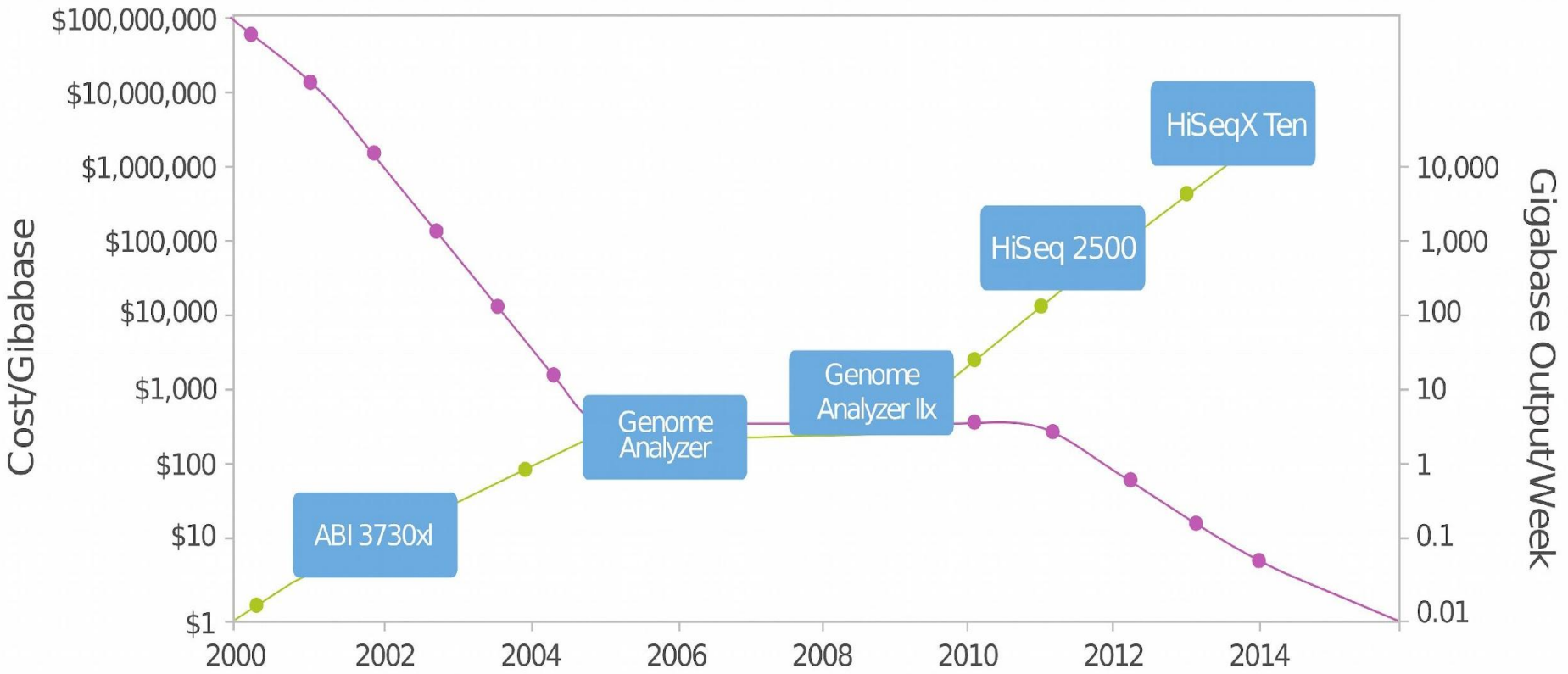
evermore
PARALLEL

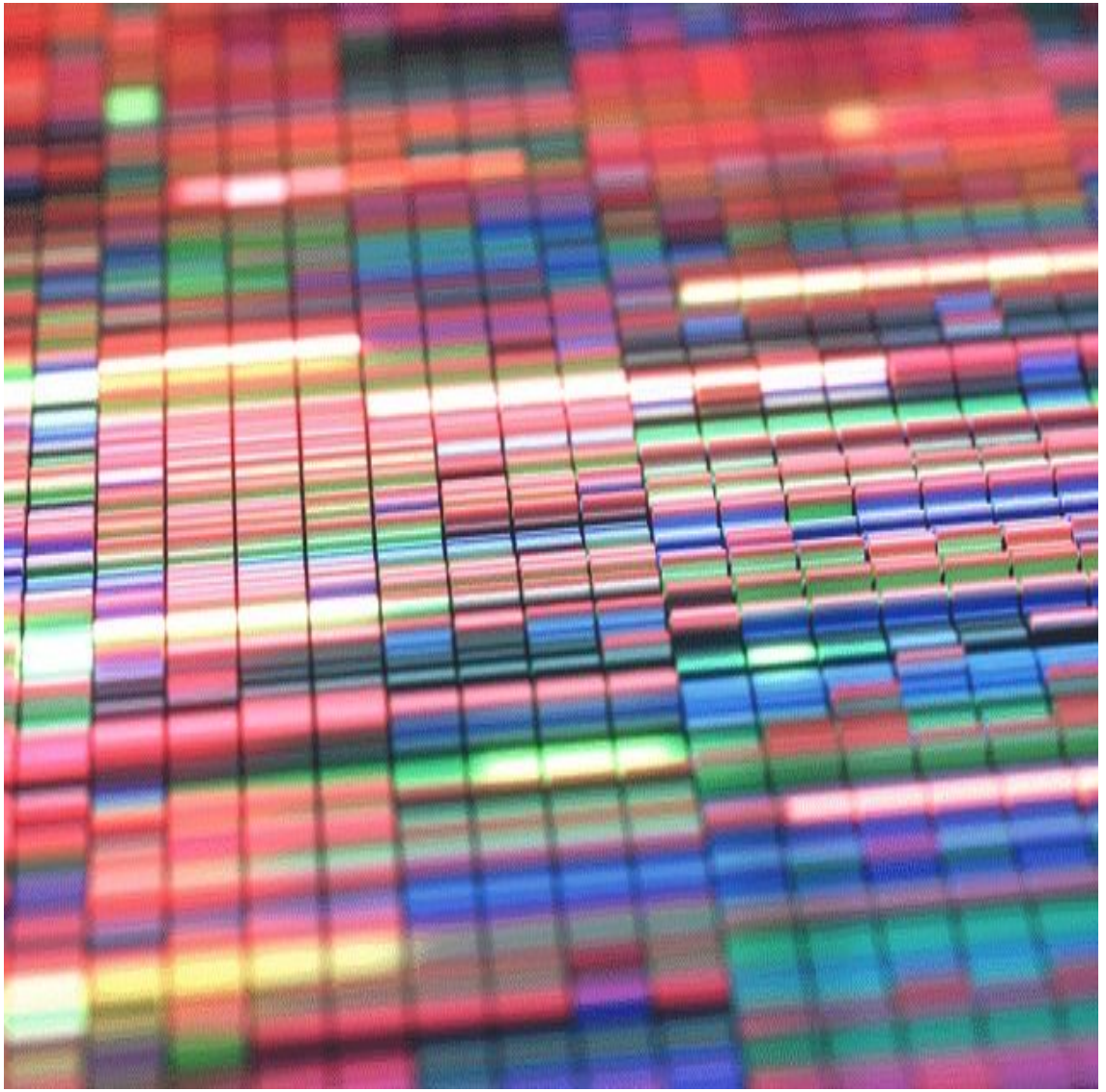
evermore
DATA

Next generation sequencing has outpaced

MOORE'S LAW:

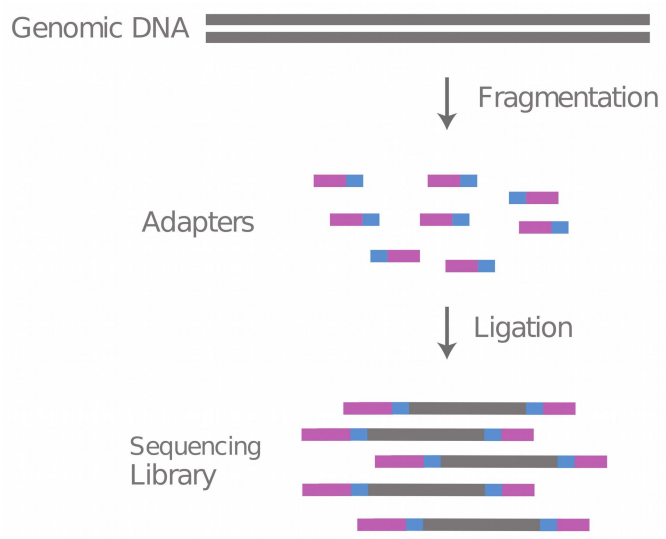
"overall processing power of computers will double every two years"

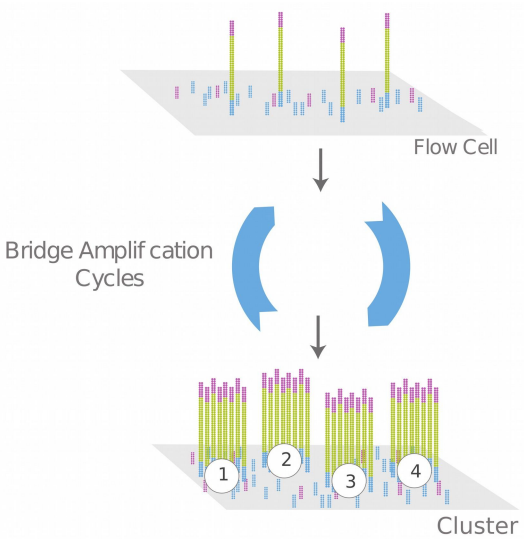


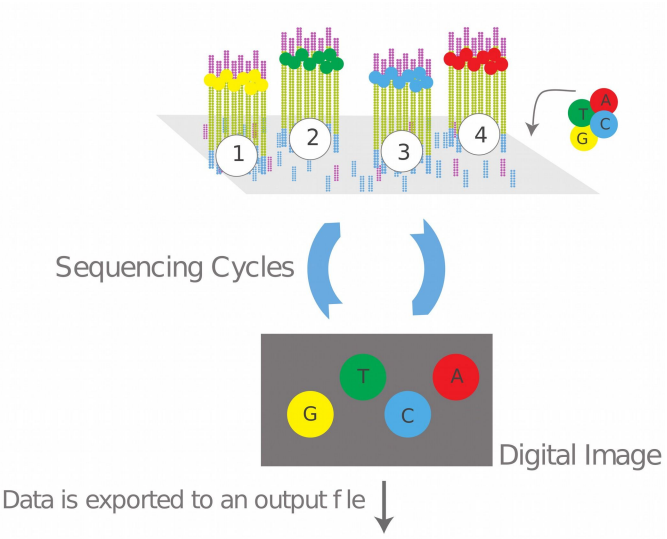


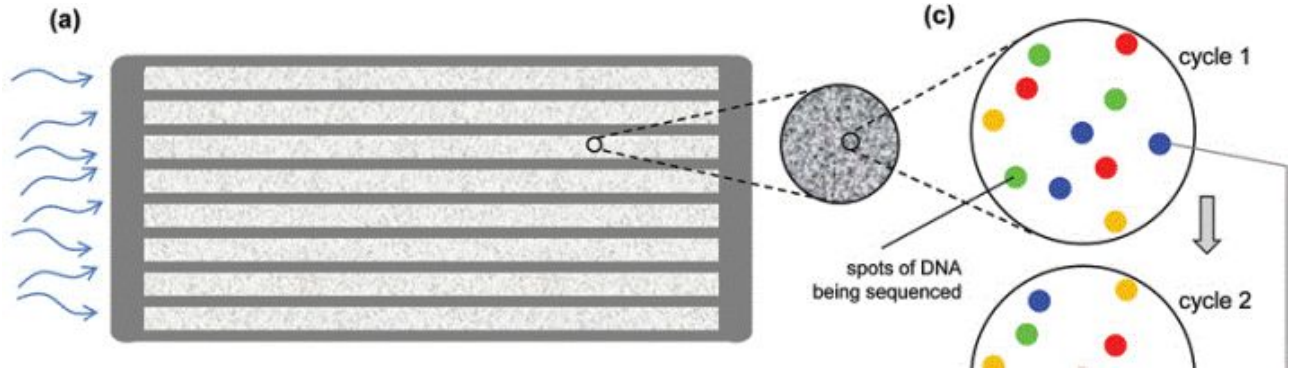
2nd Generation Sequencing

A quick look at 2nd Generation Sequencing

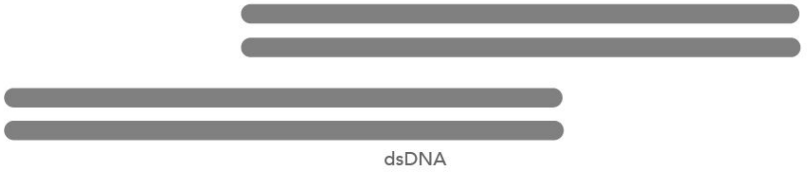








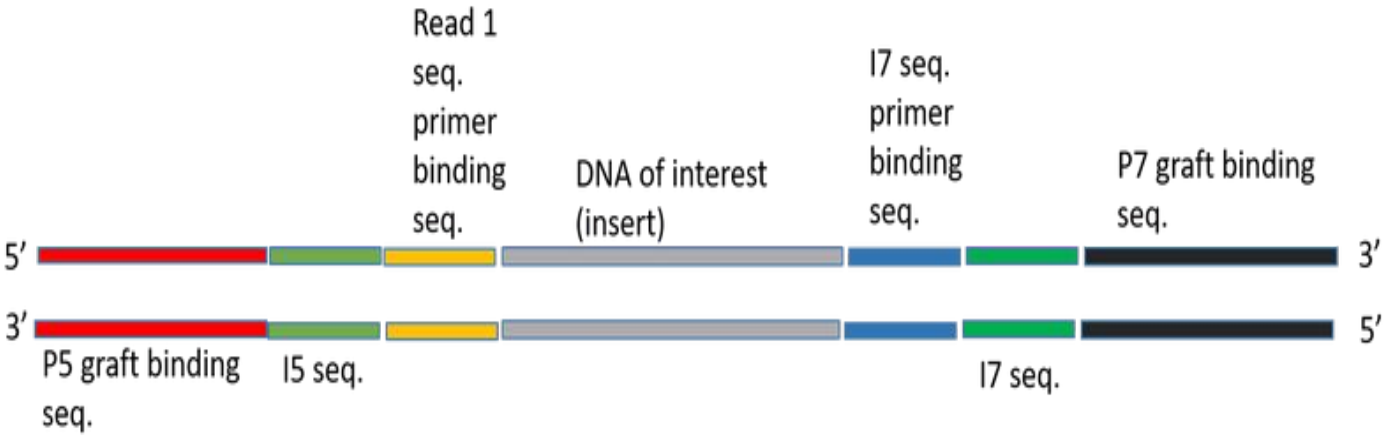
Fragmentation

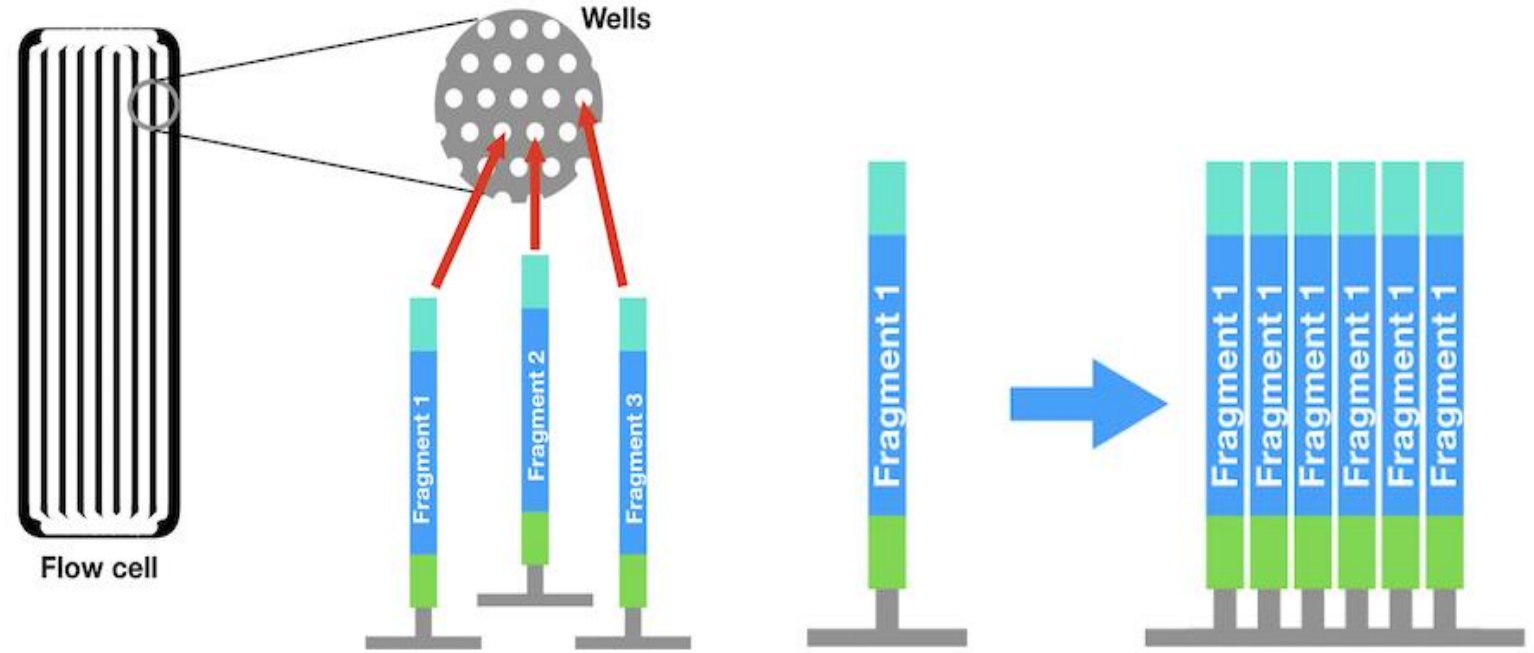


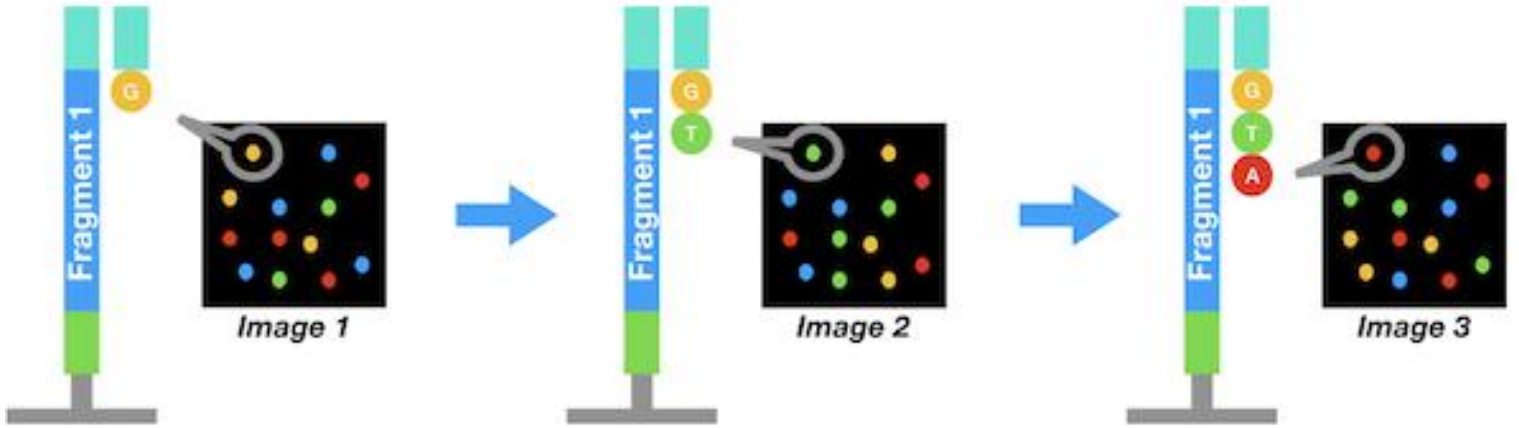
End repair and A-tailing

Ligation

PCR amplification







Base 1 = G

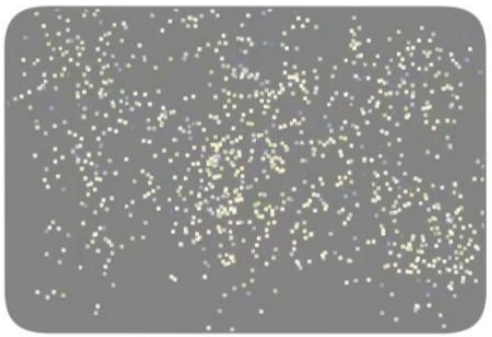
Base 2 = T

Base 3 = A

2nd Generation Sequencing

results in massively parallel sequencing of tens of gigabases ≈ 48 human genomes per day!

Sequencing

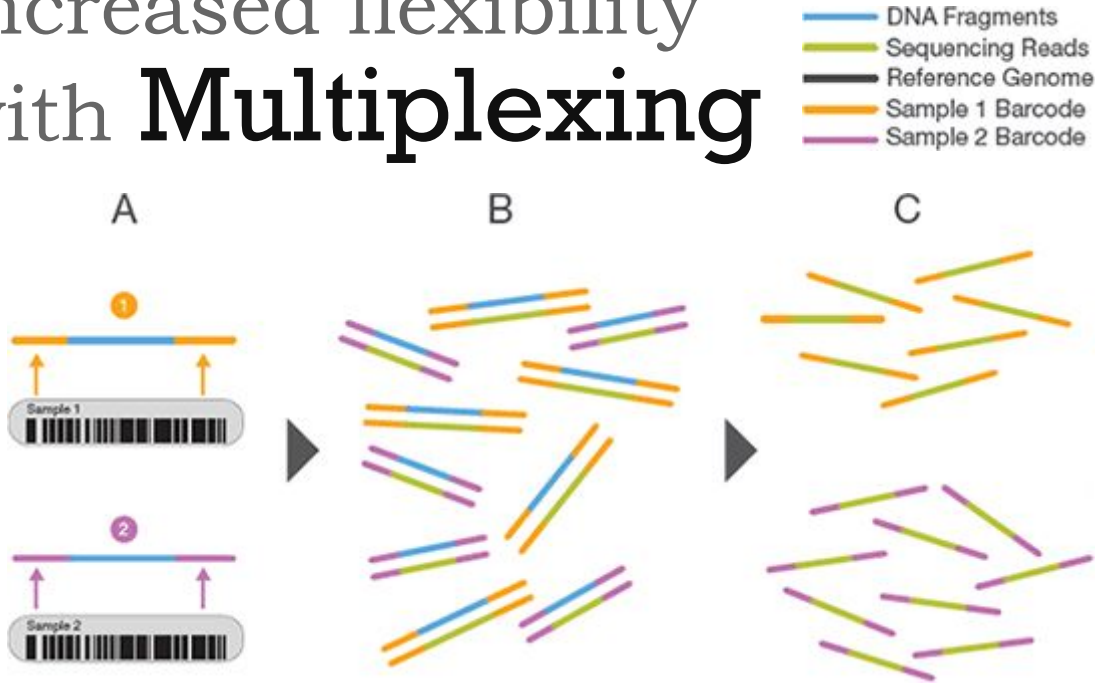


Flow cell

GAAACAAAAGCAATTGACA
CTTACGCCGTACTACCTCA
AGTAAGAAACAAAAGCAAT
ACGCCGTACTACCTCAGCA
CCTCAGCAGTAGTAAGAAA
GAAACAAAAGCAATTGACA
CTTACGCCGTACTACCTCA
AGTAAGAAACAAAAGCAAT
ACGCCGTACTACCTCAGCA
CCTCAGCAGTAGTAAGAAA
GAAACAAAAGCAATTGACA
CTTACGCCGTACTACCTCA
AGTAAGAAACAAAAGCAAT
ACGCCGTACTACCTCAGCA
CCTCAGCAGTAGTAAGAAA
GAAACAAAAGCAATTGACA
CTTACGCCGTACTACCTCA
AGTAAGAAACAAAAGCAAT
ACGCCGTACTACCTCAGCA

Increased coverage with **Paired-end sequencing**

Increased flexibility with **Multiplexing**



Sequence the first 35 – 400 base pairs call them: "READS"

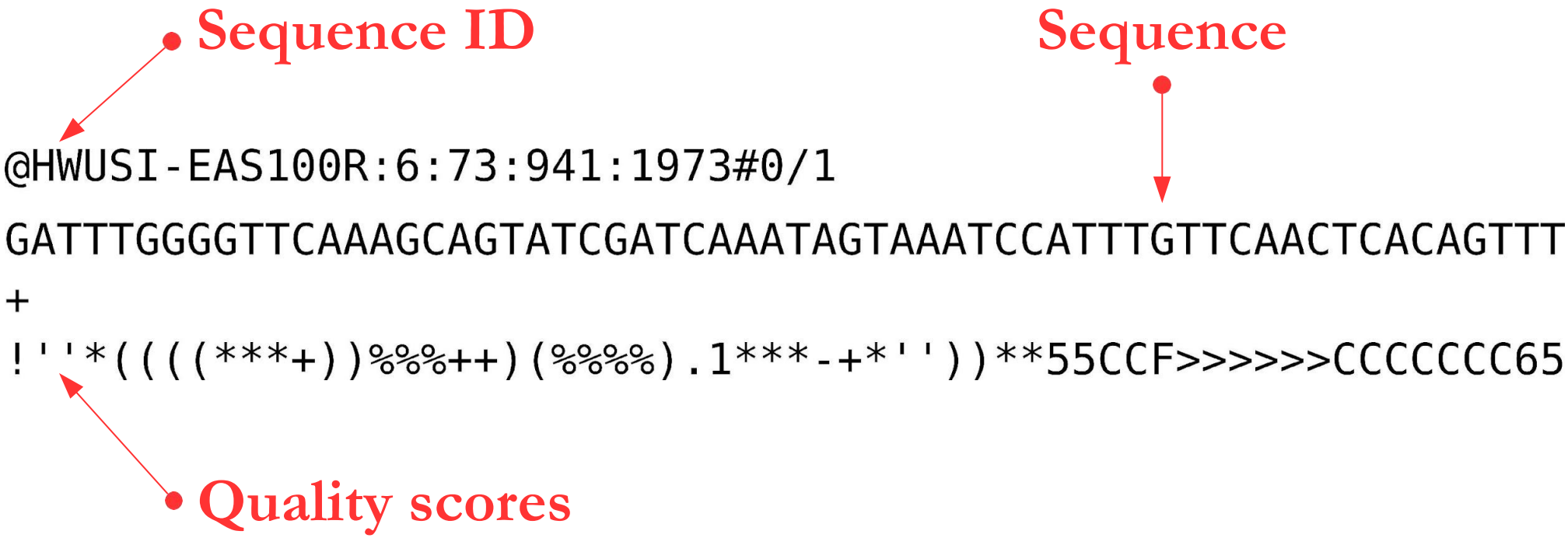
GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC

A typical run can have up to 6bln. reads!! HOW DO WE PROCESS THIS DATA?

The

FASTQ FORMAT

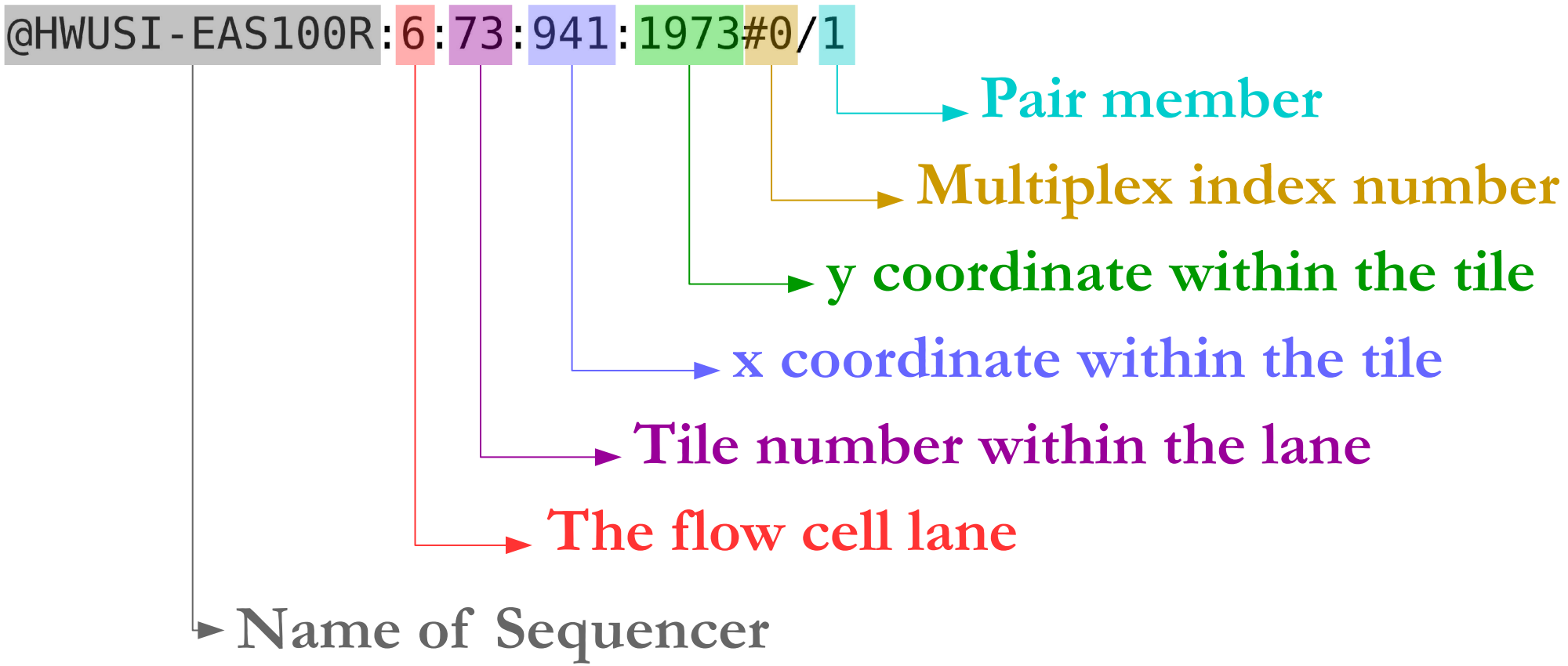
for efficient storage & information



The

FASTQ FORMAT

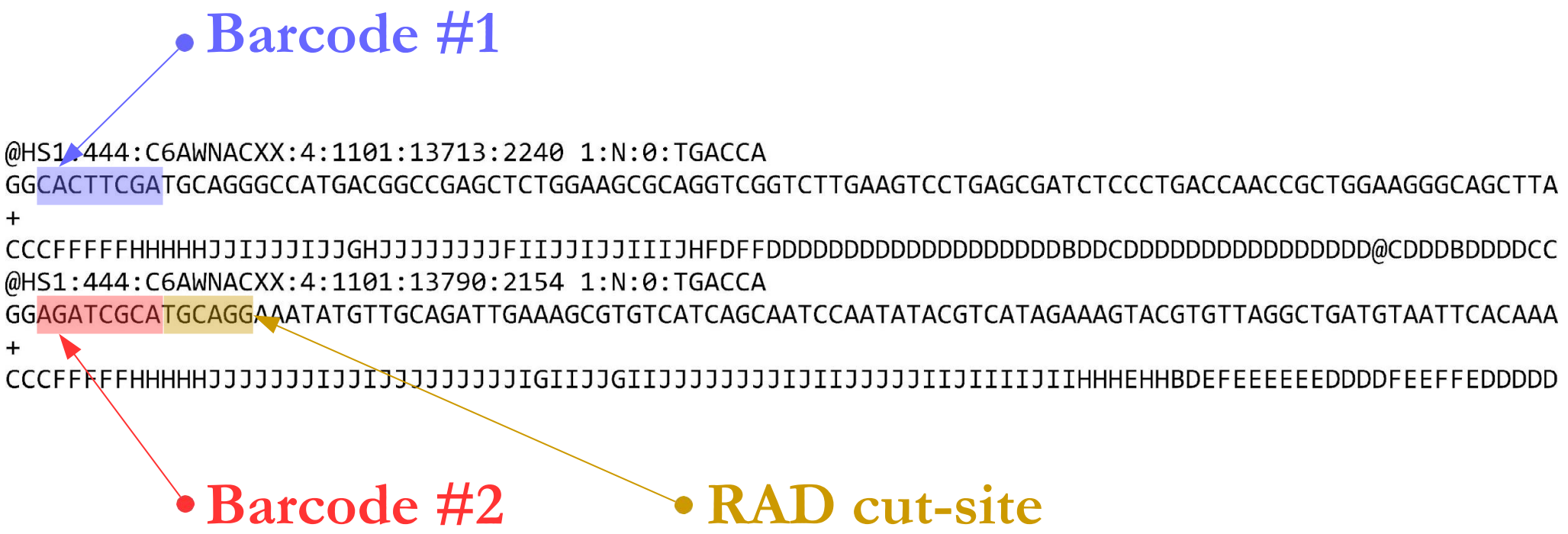
Sequence ID: Headers



The

FASTQ FORMAT

Sequences, barcodes, cut sites



The

FASTQ FORMAT

Phred Quality Scores

logarithmically related to the base-calling error probabilities

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

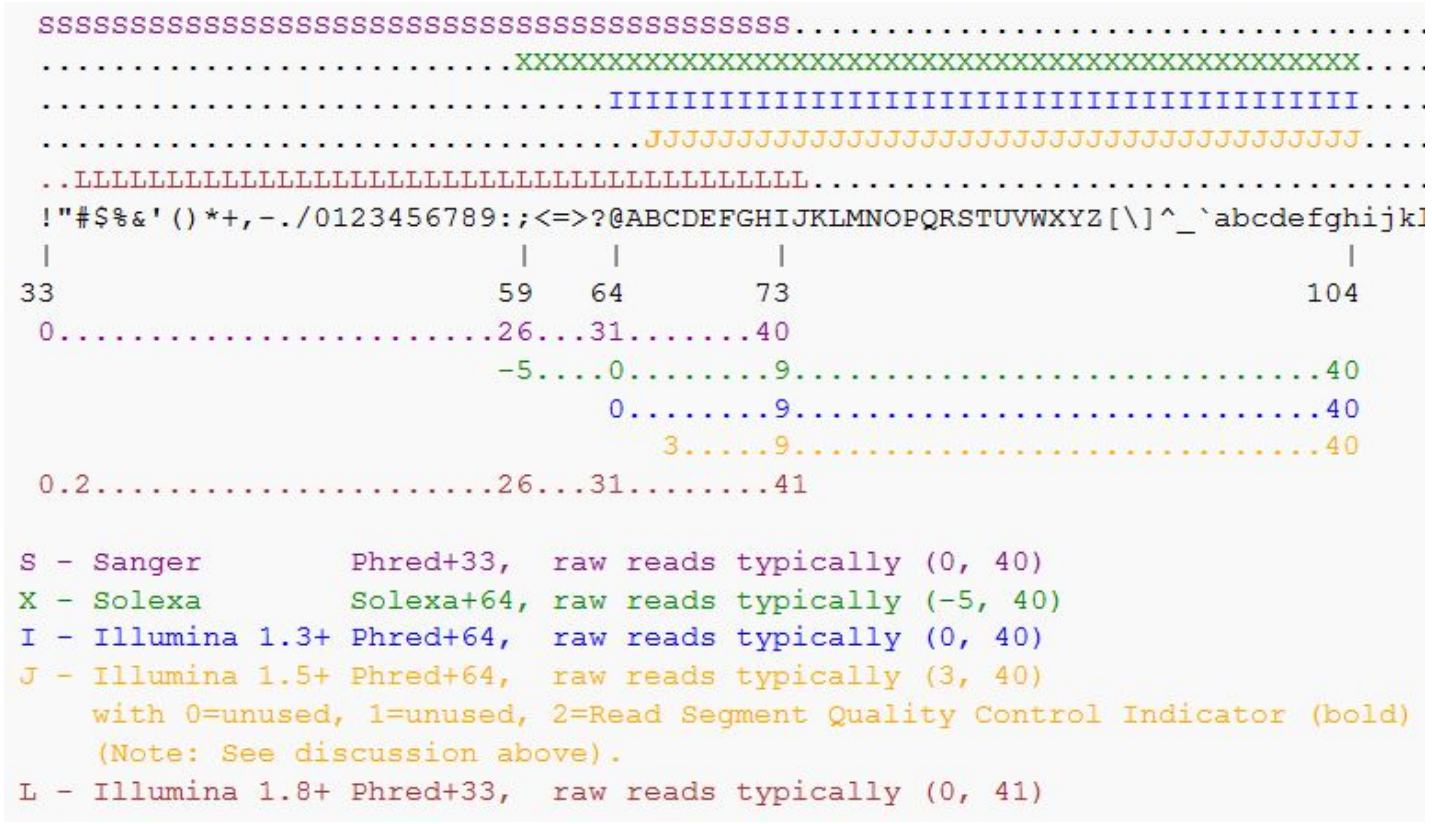
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

The

FASTQ FORMAT

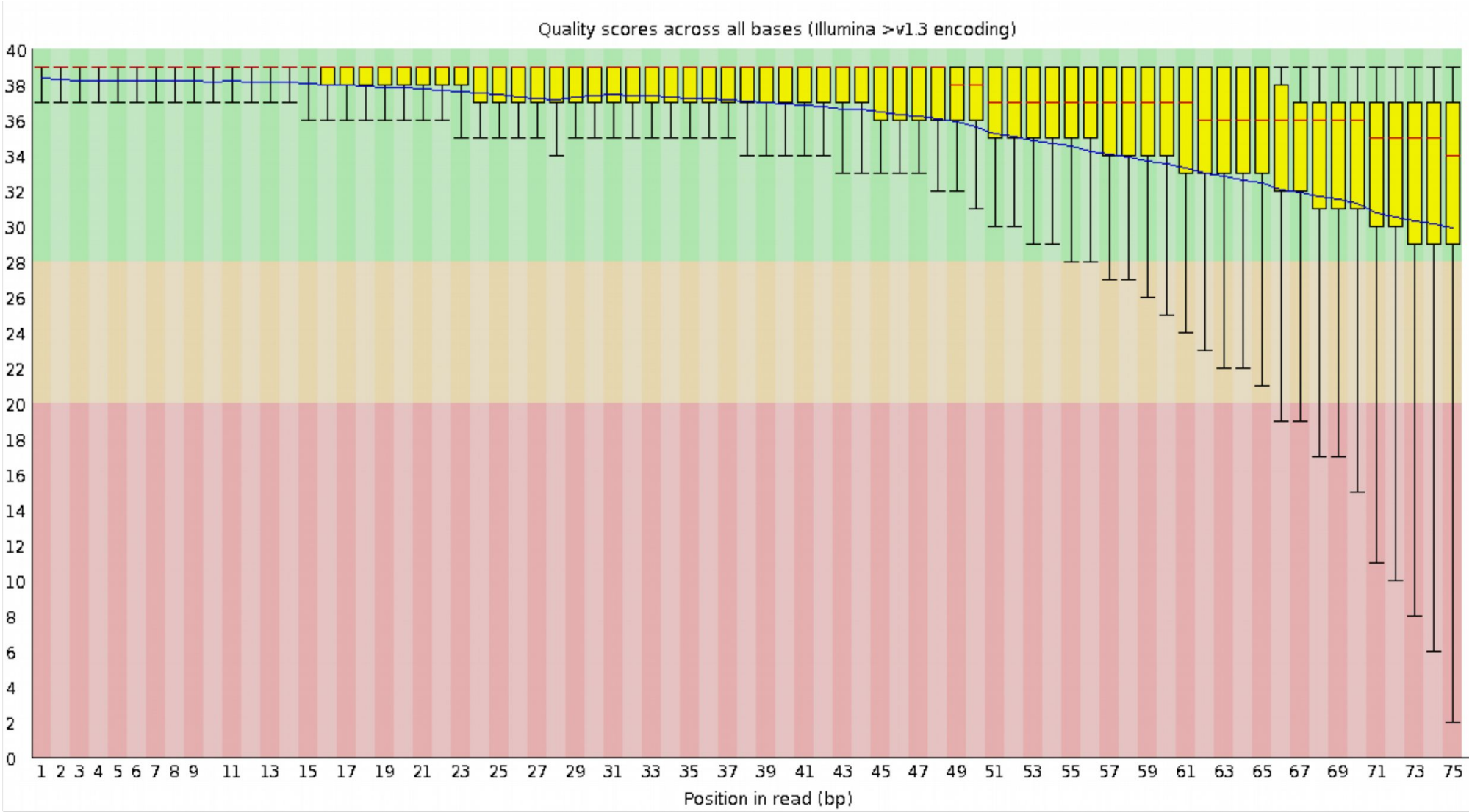
Phred Quality Scores

```
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```



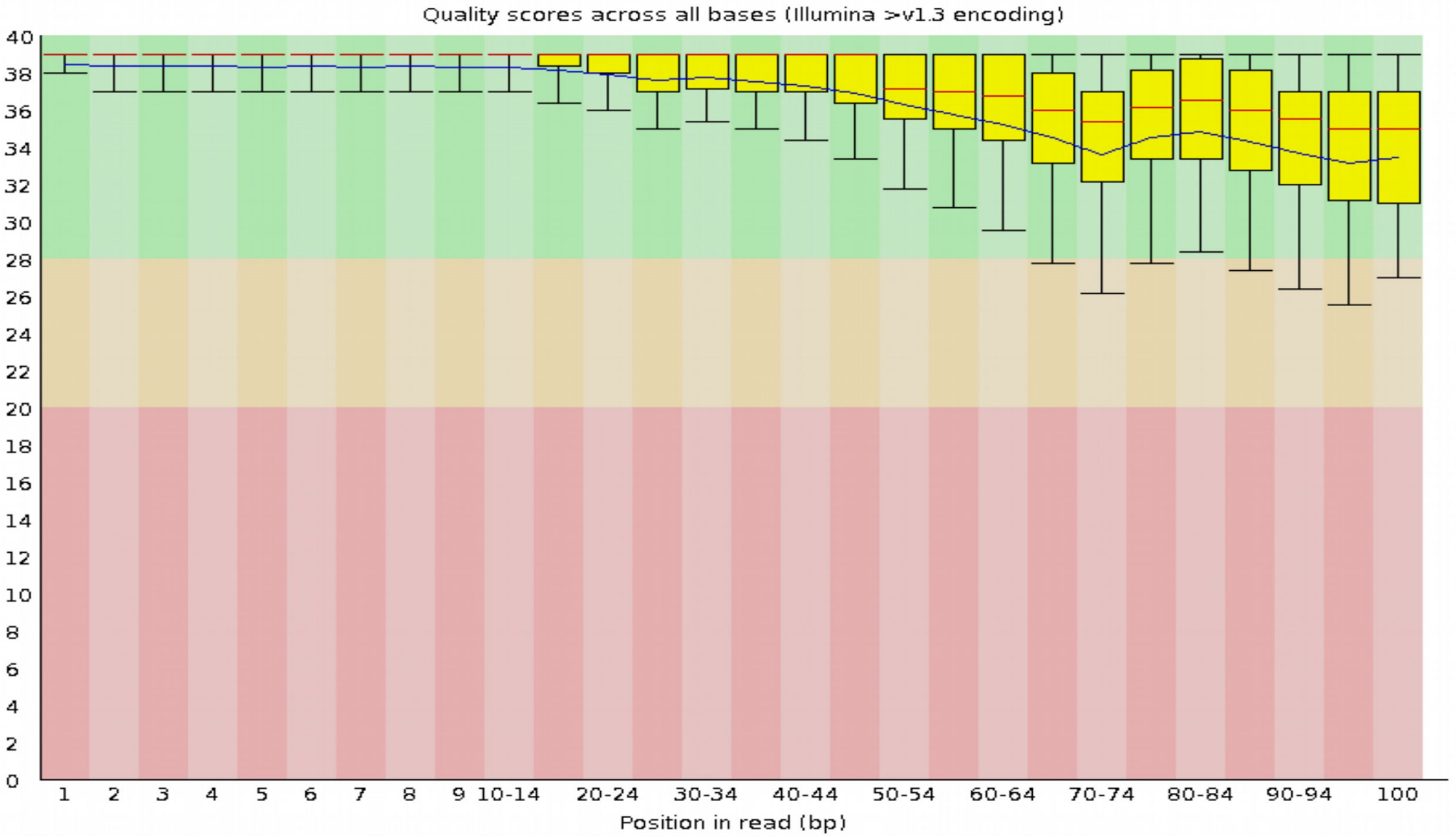
We can use quality scores to

Remove bad reads



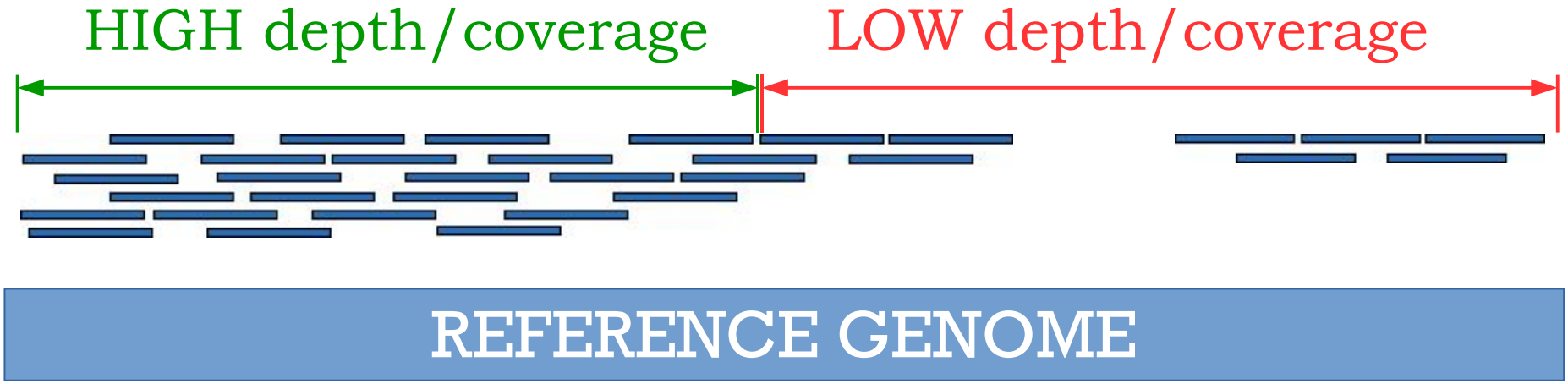
We can use quality scores to

Remove bad reads



Matching reads to a reference

MAPPING



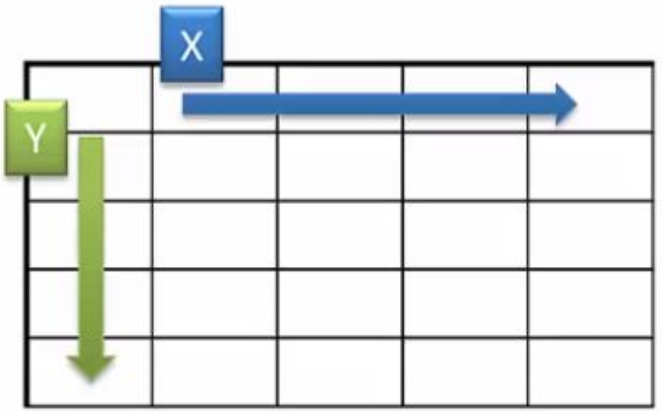
BWA BOWTIE SOAP NOVOALIGN

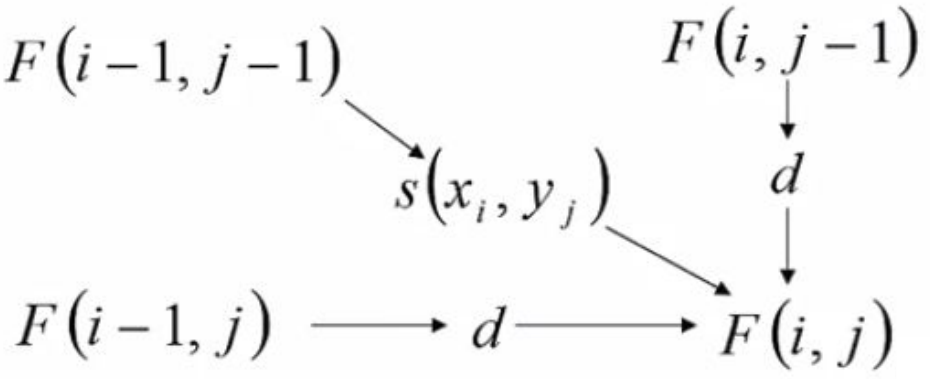
Dynamic Programming

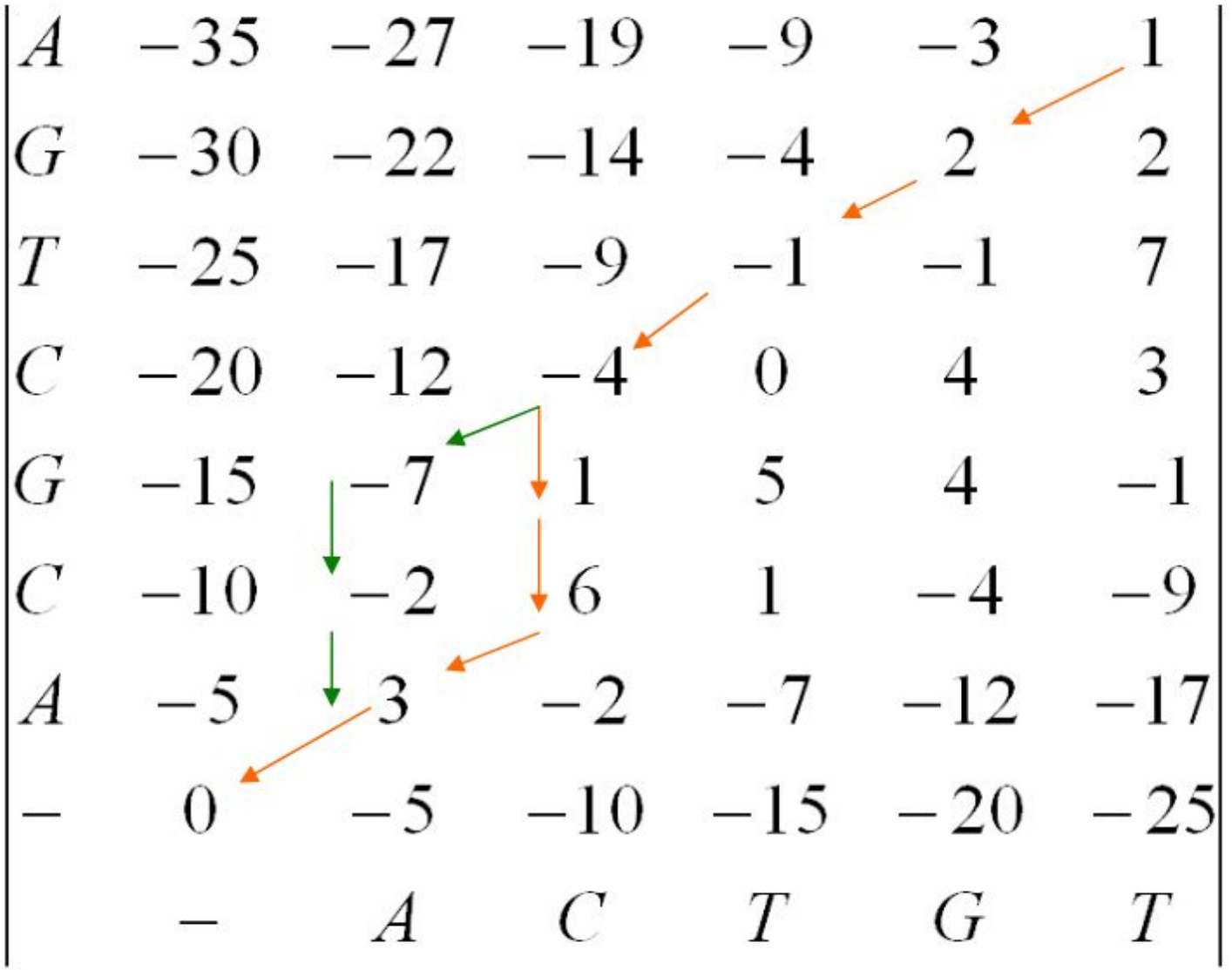
Needleman–Wunsch algorithm

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) & \text{x_i aligned to y_j} \\ F(i-1, j) + d & \text{x_i aligned to a gap} \\ F(i, j-1) + d & \text{y_j aligned to a gap} \end{cases}$$







ACGCTGA

A--CTGT

ACGCTGA

AC--TGT

How can we solve the **read alignment** problem?

Read

CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTAGGCCTTC x million

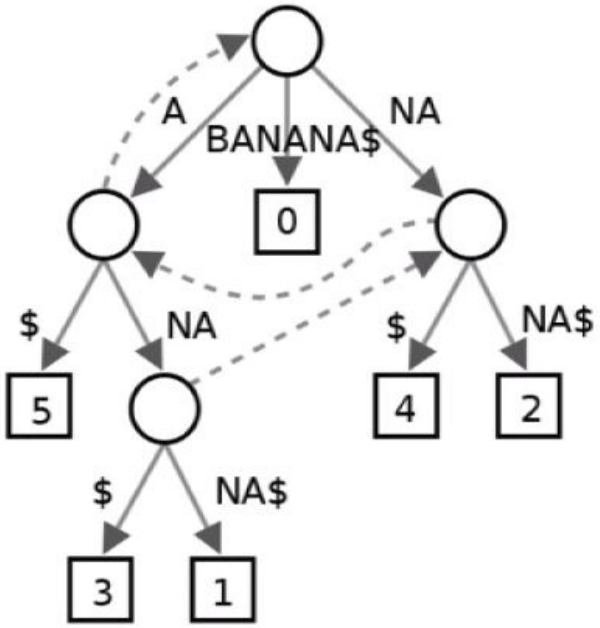
Reference

x million

—Index—

–A–

–B–



6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

$ B A N A N A
A $ B A N A N
A N A $ B A N
A N A N A $ B
B A N A N A $
N A $ B A N A
N A N A $ B A

Suffix tree

≥ 45 GB

Suffix array

≥ 12 GB

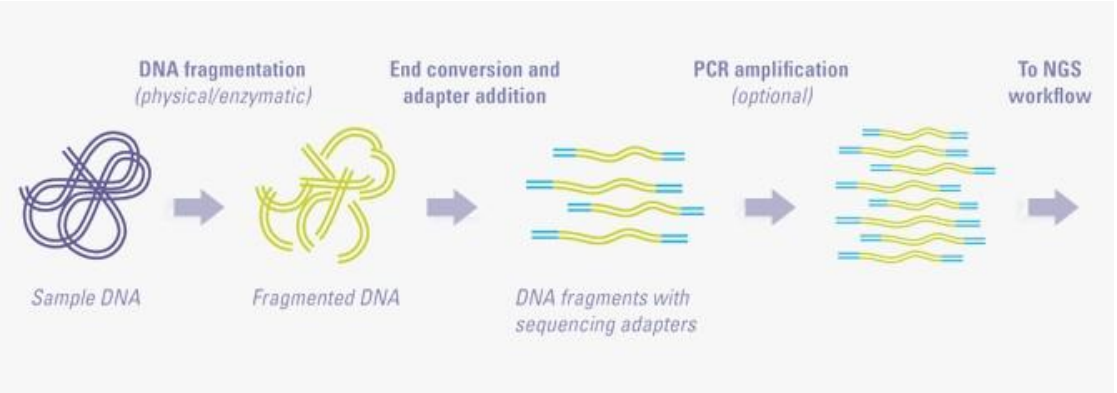
FM Index

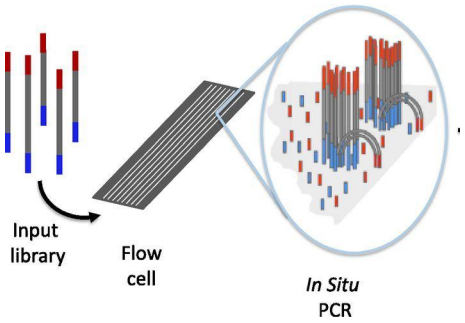
~ 1 GB

Remove

CLONES

that can artificially bias depth





1. Shater genomic DNA
2. Ligate adaptors to both ends & PCR amplify
3. Spread DNA molecules across flow cells
4. Goal: exactly one DNA molecule per flow cell lawn
5. Amplify the single molecule on each lawn

Remove

CLONES

that can artificially bias depth

```
      TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
        TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
          GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
           TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
            GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
               CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
                      GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
                        GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                       TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                        ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                                        ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
     CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC
```

Remove

CLONES

that can artificially bias depth

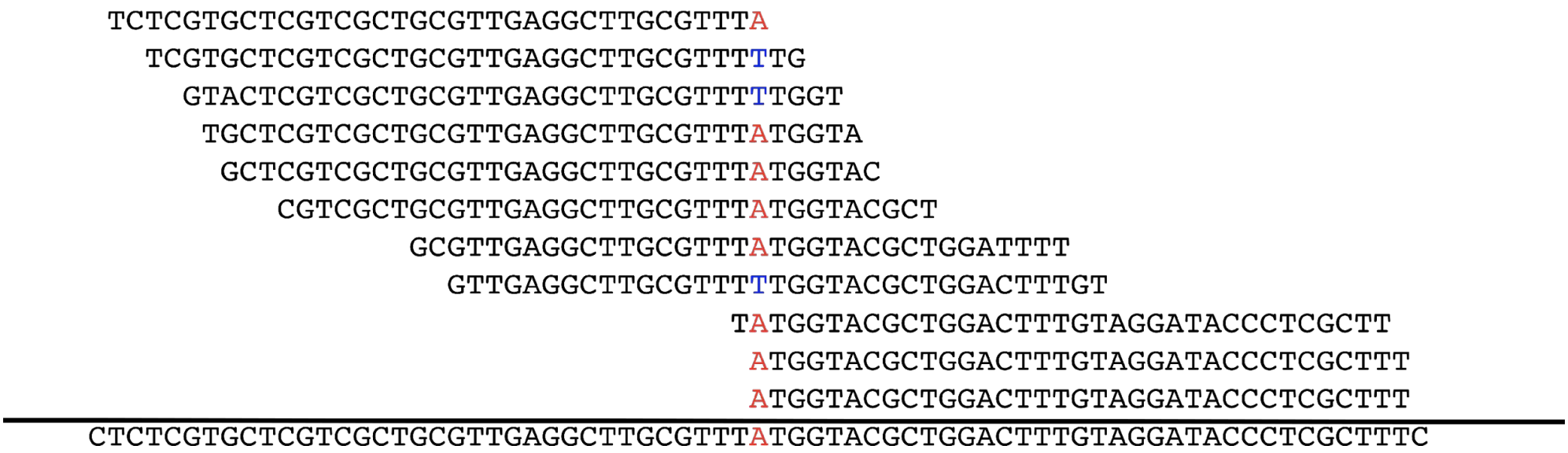
```
 TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
   TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
     GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
      TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
       GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
          CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
                 GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
                   GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC
```

Possible PCR clones

Remove

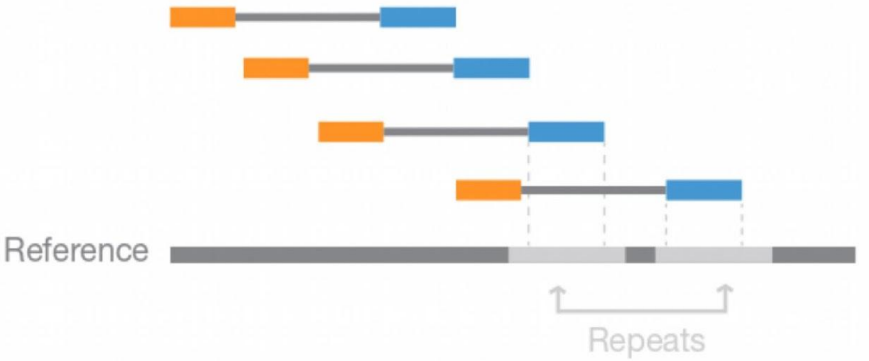
CLONES

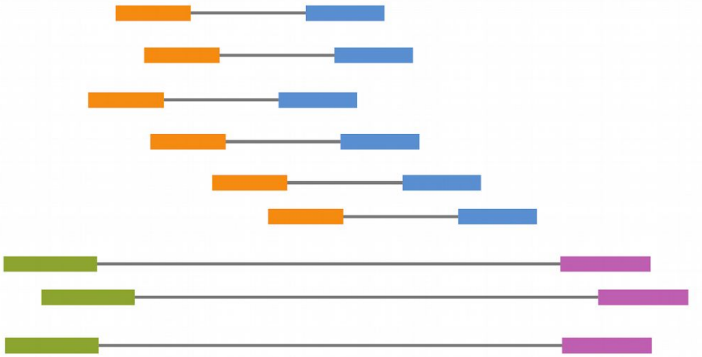
that can artificially bias depth



Mapping is more effective with

PAIRED-END DATA

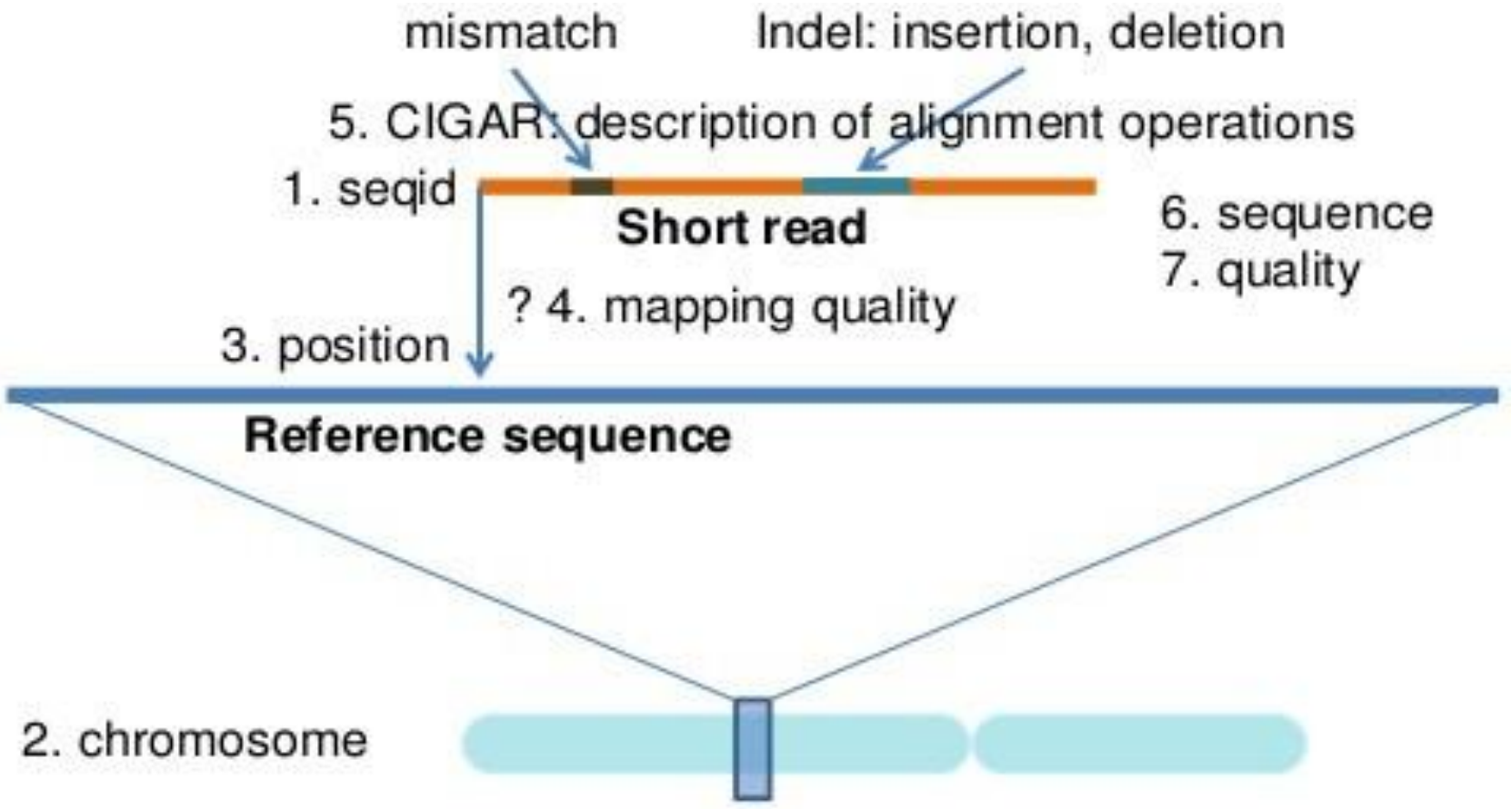




The

SAM FORMAT

Information rich storage of read alignments



The

SAM HEADER

Information about the files origin and content

```
@HD VN:1.0  SO:coordinate
@SQ SN:1  LN:249250621 AS:human_v37.fasta
@PG ID:bwa  VN:0.5.4
@RG ID:UM0098:1 PL:ILLUMINA PU:HWUSI-L001
        LB:80DT:2010-05-05T20-40 SM:SD374
        CN:UMCORE
```

The

SAM HEADER

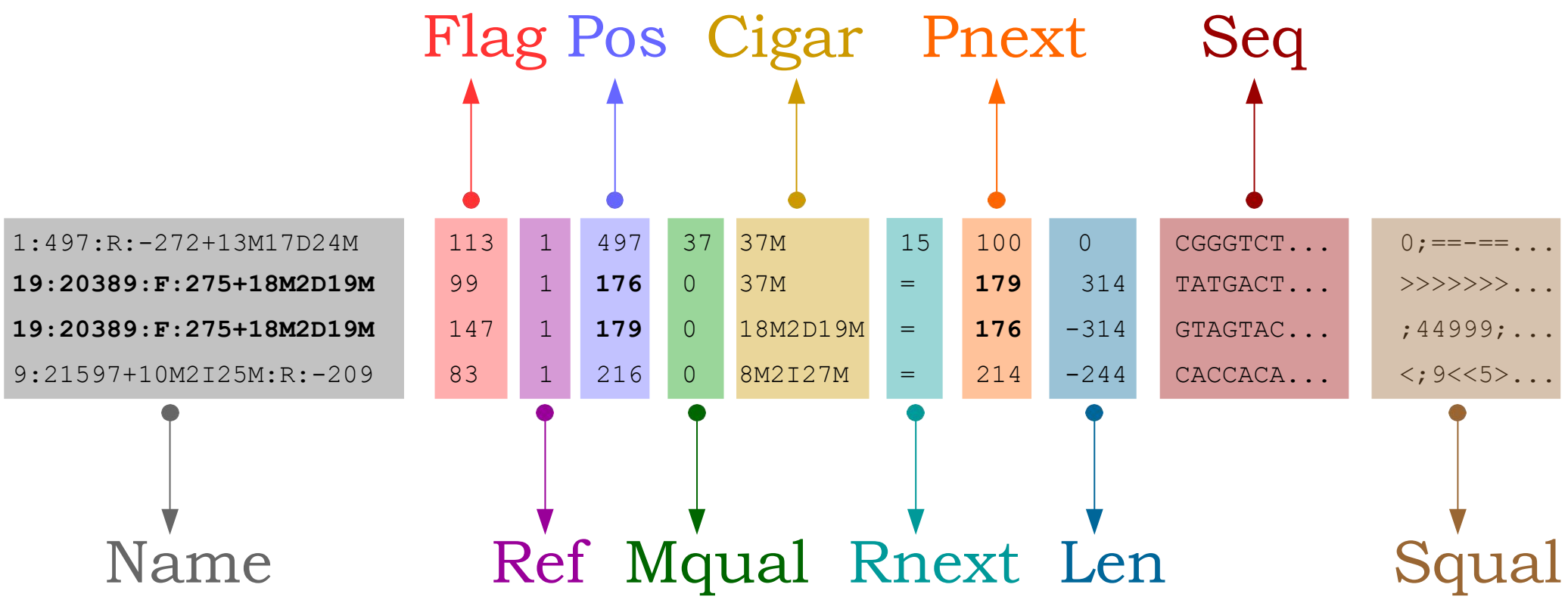
Information about individual read alignments

#	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-based leftmost Mate POSition
9	ISIZE	inferred Insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33=Phred base quality)

The

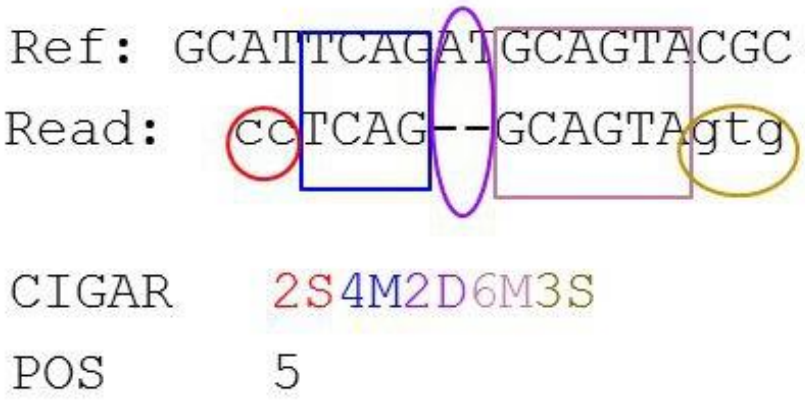
SAM HEADER

Information about individual read alignments



SAM CIGAR STRING

M: match/mismatch
I: insertion
D: deletion
P: padding
N: skip
S: soft-clip
H: hard-clip




```
REF:  CACGATCA**GACCGATACGTCCGA
READ1:  CGATCAGAGACCGATA
READ2:    ATCA*AGACCGATAC
READ3:   GATCA**GACCG
```

The padded CIGAR are different:

```
READ1: 6M2I8M
READ2: 4M1P1I9M
READ3: 5M2P5M
```

SAM FLAG : 99 000001100011

#	Binary	Decimal	Hexadecimal	Description
1	1	1	0x1	Read paired
2	10	2	0x2	Read mapped in proper pair
3	100	4	0x4	Read unmapped
4	1000	8	0x8	Mate unmapped
5	10000	16	0x10	Read reverse strand
6	100000	32	0x20	Mate reverse strand
7	1000000	64	0x40	First in pair
8	10000000	128	0x80	Second in pair
9	100000000	256	0x100	Not primary alignment
10	1000000000	512	0x200	Read fails platform/vendor quality checks
11	10000000000	1024	0x400	Read is PCR or optical duplicate
12	100000000000	2048	0x800	Supplementary alignment
SUM:	**000001100011**	**99**		

http://www.samformat.info/sam-format-flag

SAM FLAG : 113 000001110001

#	Binary	Decimal	Hexadecimal	Description
1	1	1	0x1	Read paired
2	10	2	0x2	Read mapped in proper pair
3	100	4	0x4	Read unmapped
4	1000	8	0x8	Mate unmapped
5	10000	16	0x10	Read reverse strand
6	100000	32	0x20	Mate reverse strand
7	1000000	64	0x40	First in pair
8	10000000	128	0x80	Second in pair
9	100000000	256	0x100	Not primary alignment
10	1000000000	512	0x200	Read fails platform/vendor quality checks
11	10000000000	1024	0x400	Read is PCR or optical duplicate
12	100000000000	2048	0x800	Supplementary alignment
SUM:	**000001110011**	**113**		

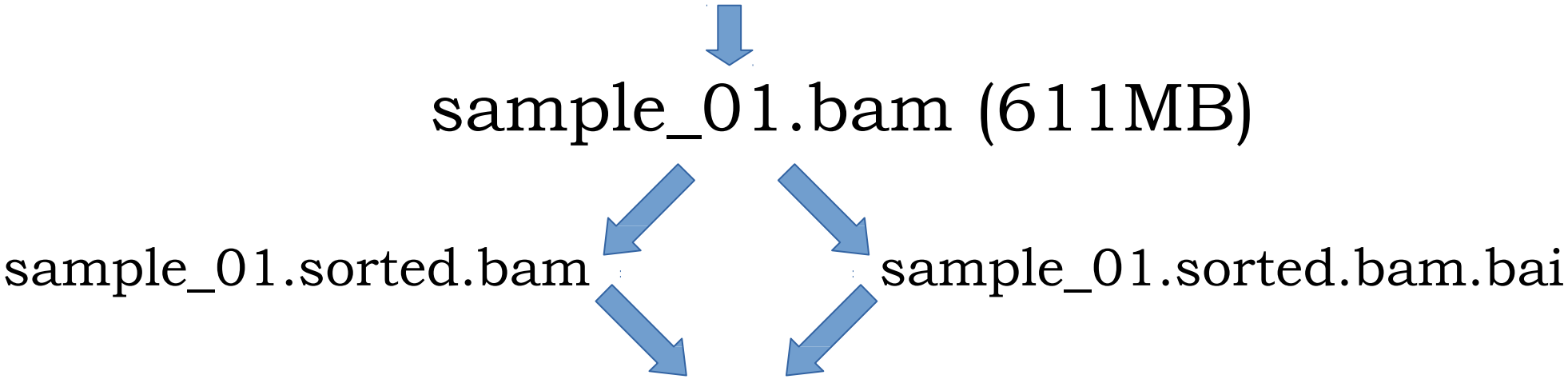
http://www.samformat.info/sam-format-flag

The

BAM FORMAT

Compressed, binary, indexed version of SAM

1:497:R:-272+13M17D24M	113	1	497	37	37M	15	100	0	CGGGTCT...	0;==-==...
19:20389:F:275+18M2D19M	99	1	**176**	0	37M	=	**179**	314	TATGACT...	>>>>>>>...
19:20389:F:275+18M2D19M	147	1	**179**	0	18M2D19M	=	**176**	-314	GTAGTAC...	;44999;...
9:21597+10M2I25M:R:-209	83	1	216	0	8M2I27M	=	214	-244	CACCACA...	<;9<<5>...



NGS FLOW CHART	FILE FORMAT	PROGRAMS
Raw sequence reads	Fastq	
Demultiplex and remove low quality reads	Fastq	Customscripts Fastqc/Fastx-toolkit
Map reads to reference genome	SAM/BAM	BWA/Bowtie Soap/Novoalign
Filter unpaired, unmapped and duplicate reads	SAM/BAM	SAMtools/Picard

DOWNSTREAM ANALYSIS

USEFUL LINKS

SAMtools: http://www.htslib.org

Picardtools:https://broadinstitute.github.io/picard/

BWA:http://bio-bwa.sourceforge.net

Bowtie:http://bowtie-bio.sourceforge.net/index.shtml

SOAP:http://soap.genomics.org.cn/index.html

Novoalign:http://www.novocraft.com/products/novoalign/

FASTX-toolkitt:http://hannonlab.cshl.edu/fastx_toolkit/

FastQC:http://www.bioinformatics.babraham.ac.uk/projects/fastqc/