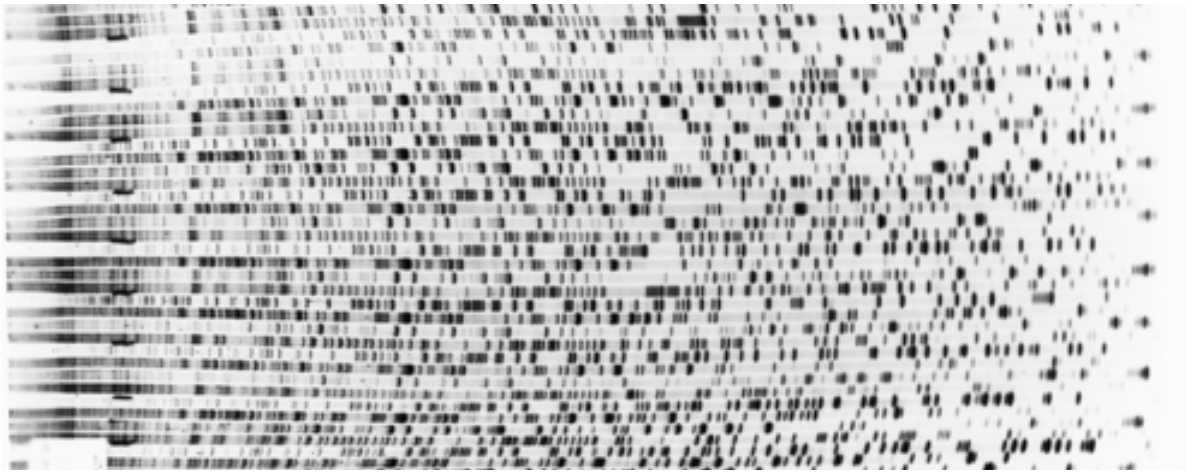
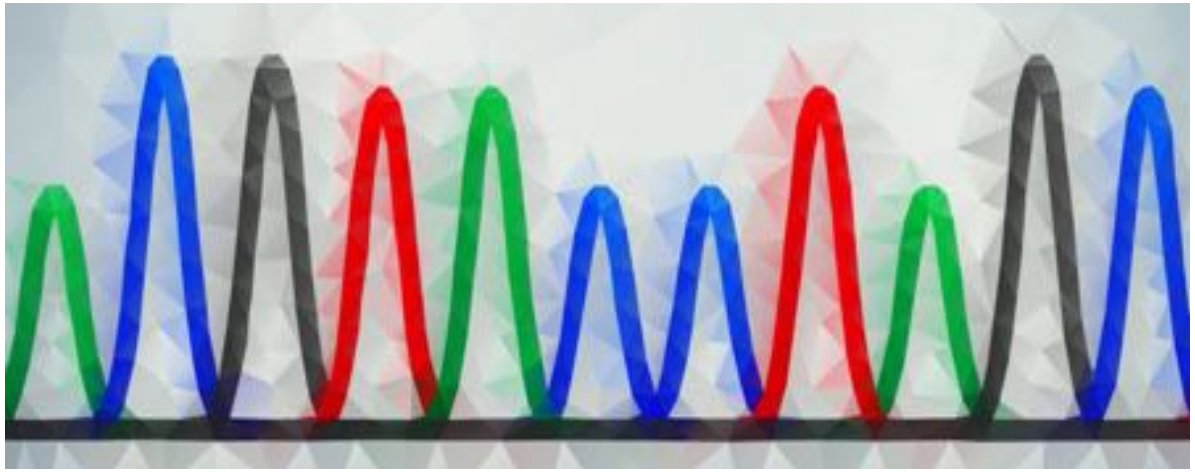
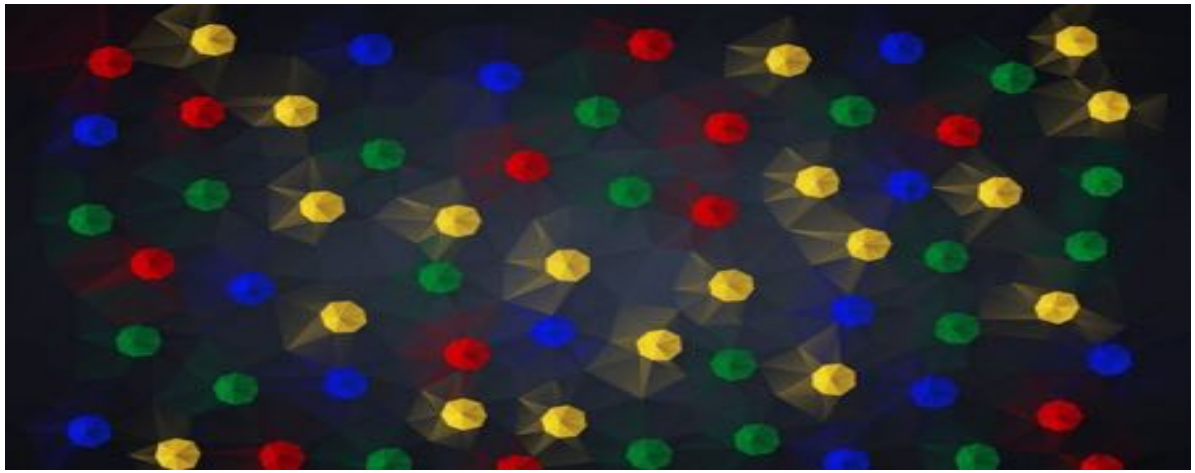
Next Generation Sequencing

an Introduction







The future of SEQUENCING

ever more
MASSIVE

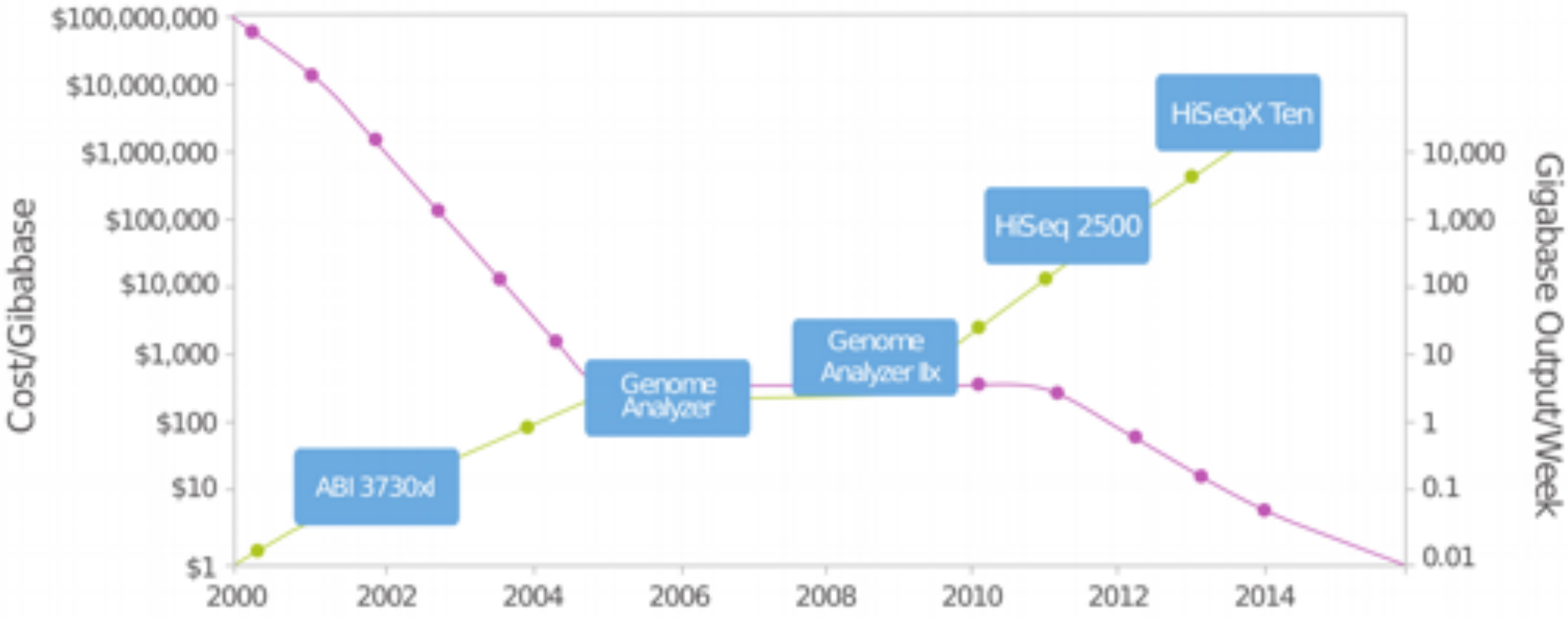
ever more
PARALLEL

ever more
DATA

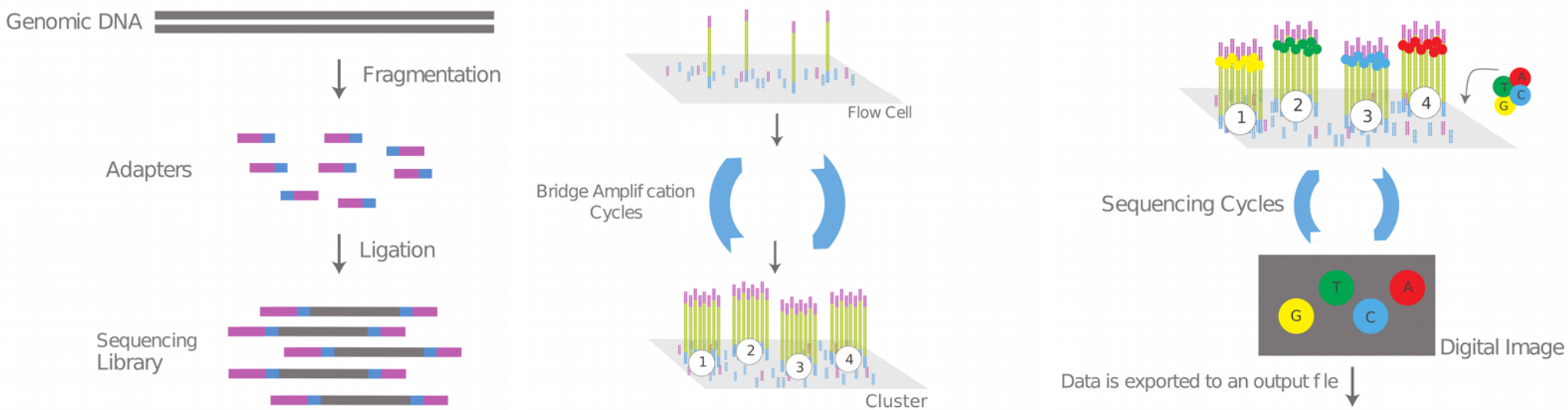
Next generation sequencing
has outpaced

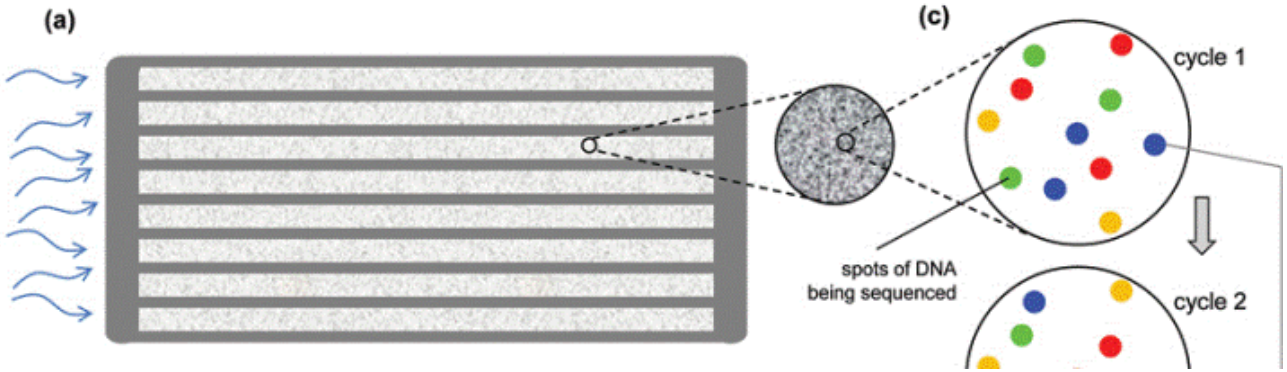
MOORE'S LAW:

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



A quick look at 2nd Generation Sequencing

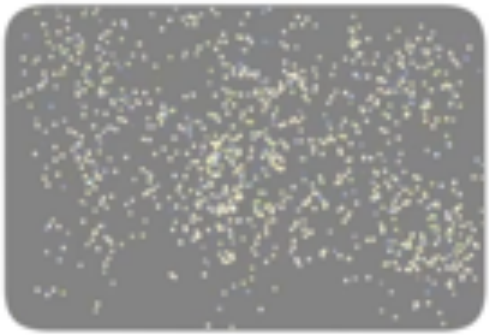




2nd Generation Sequencing

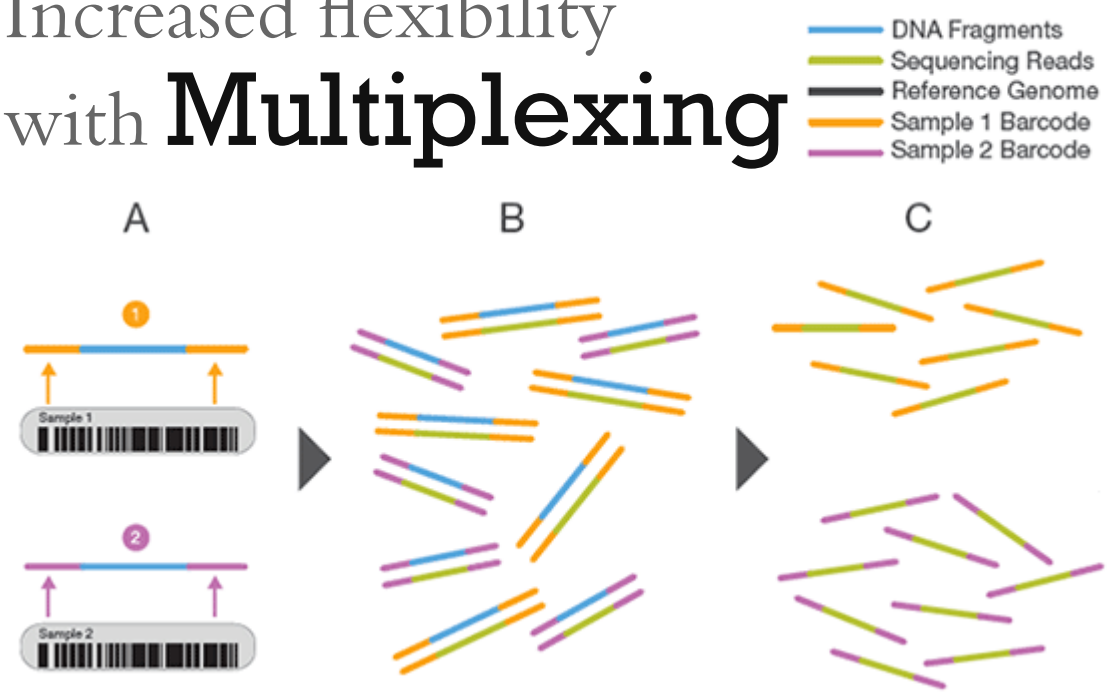
results in massively parallel sequencing of tens of gigabases ≈ 45 human genomes per day!

Sequencing



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 6 bln. reads!! **HOW DO WE PROCESS THIS DATA?**

The

FASTQ FORMAT

for efficient storage & information

Sequence ID

Sequence

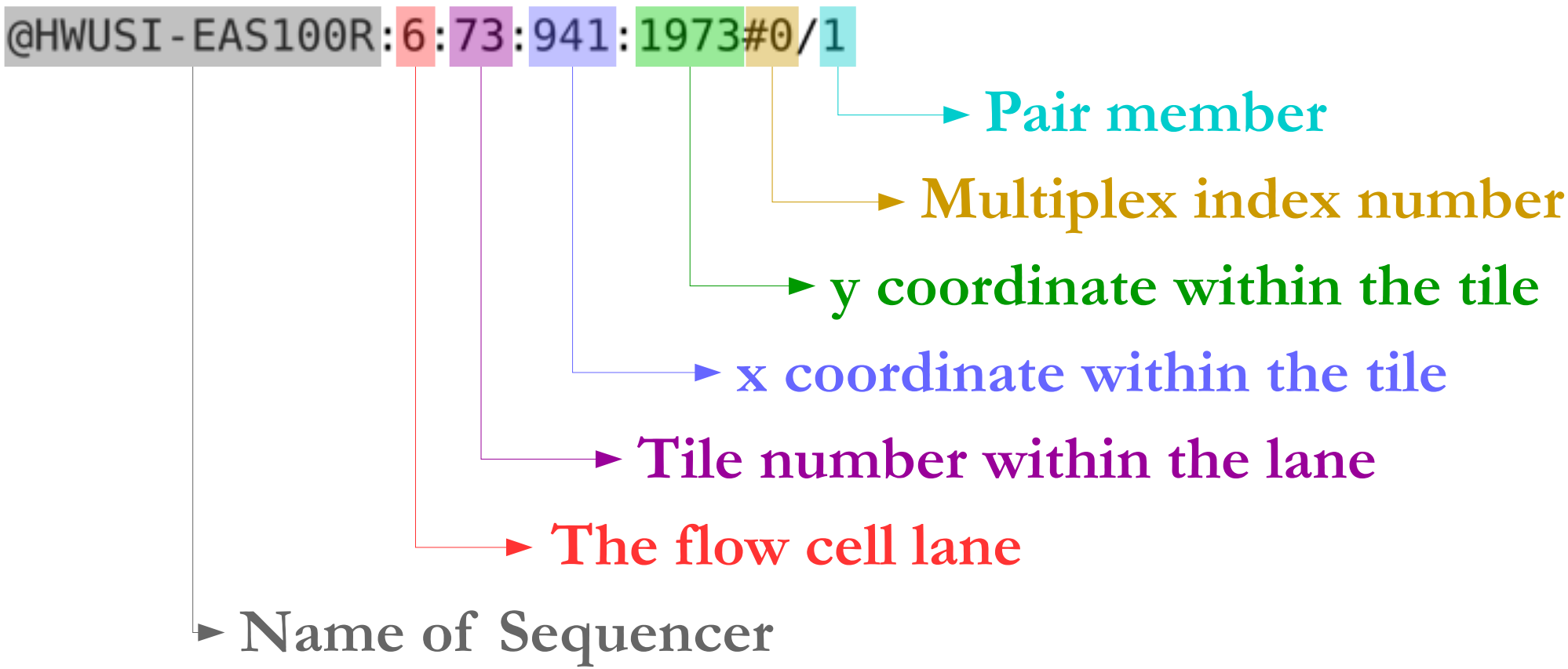
```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Quality scores

The

FASTQ FORMAT

Sequence ID: Headers



The

FASTQ FORMAT

Sequences, barcodes & cut-sites

• Barcode #1

```
@HS1:444:C6AWNACXX:4:1101:13713:2240 1:N:0:TGACCA
GGCACTTCGATGCAGGGCCATGACGGCCGAGCTCTGGAAGCGCAGGTCGGTCTTGAAGTCCTGAGCGATCTCCCTGACCAACCGCTGGAAGGGCAGCTTA
+
CCCFFFFFHHHHHJJIJJJIJJGHJJJJJJJJFIIJJIJJIIIJHFDFFDDDDDDDDDDDDDDDDDDBDDCDDDDDDDDDDDDDDDD@CDDDBDDDDCC
@HS1:444:C6AWNACXX:4:1101:13790:2154 1:N:0:TGACCA
GGAGATCGCATGCAGGAAATATGTTGCAGATTGAAAGCGTGTCATCAGCAATCCAATATACGTCATAGAAAGTACGTGTTAGGCTGATGTAATTCACAAA
+
CCCFFFFFHHHHHJJJJJJJJIJJIJJJJJJJJJIGIIJJGIIJJJJJJJJIJIIJJJJJIIJIIIIJIIHHHEHHBDEFEEEEEEDDDDFEEFFEDDDDD
```

• Barcode #2

• RAD cut-site

The FASTQ FORMAT

Quality Scores

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

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS..................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII...
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ...
  ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijk
  |                         |    |        |                              |
 33                        59   64       73                            104
  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)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

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)

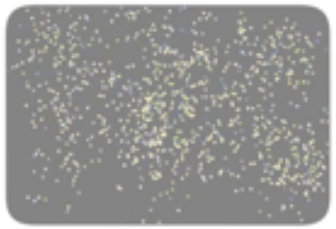
The

FASTQ FORMAT

Quality Scores

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

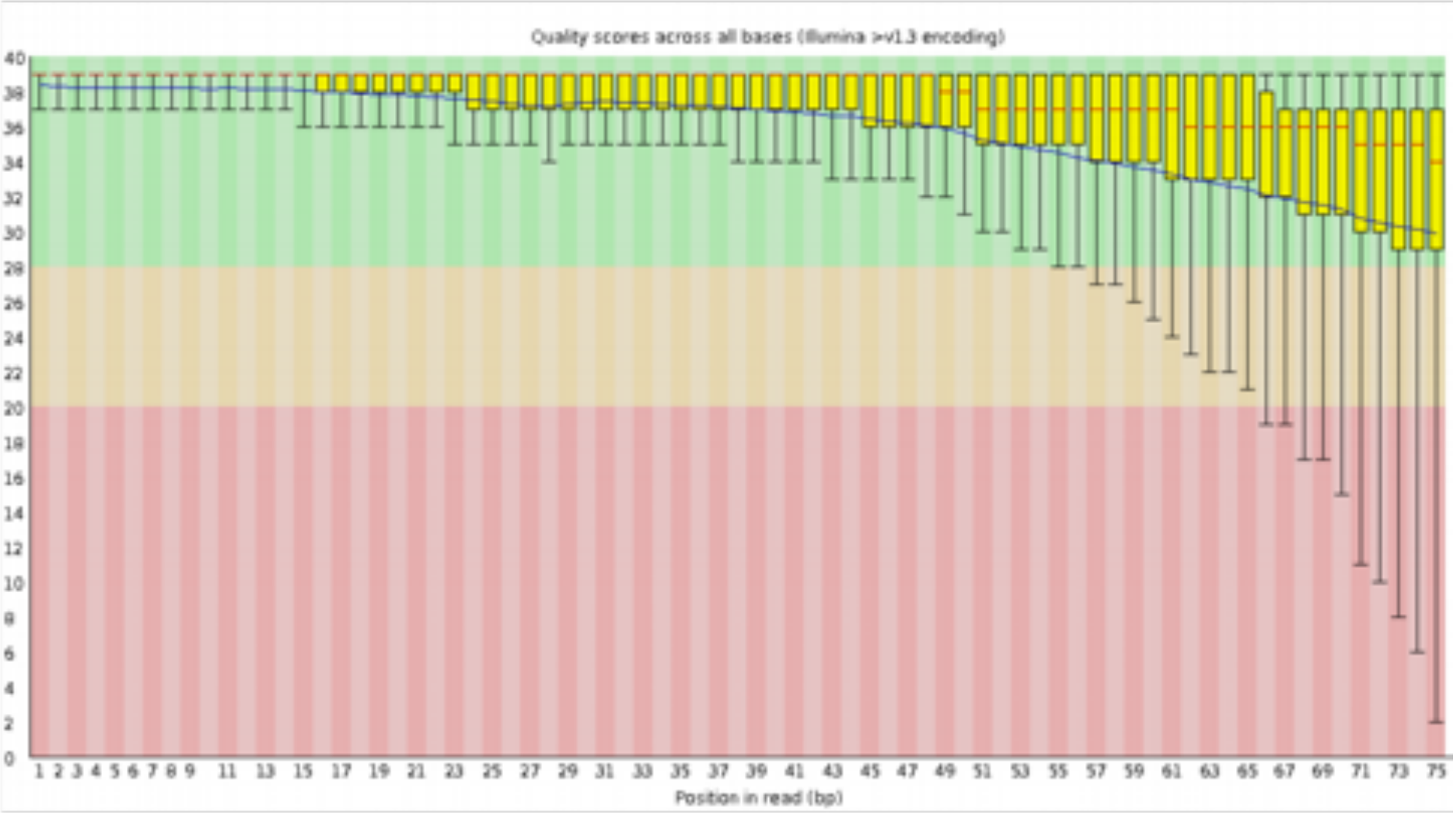
Sequencing



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)

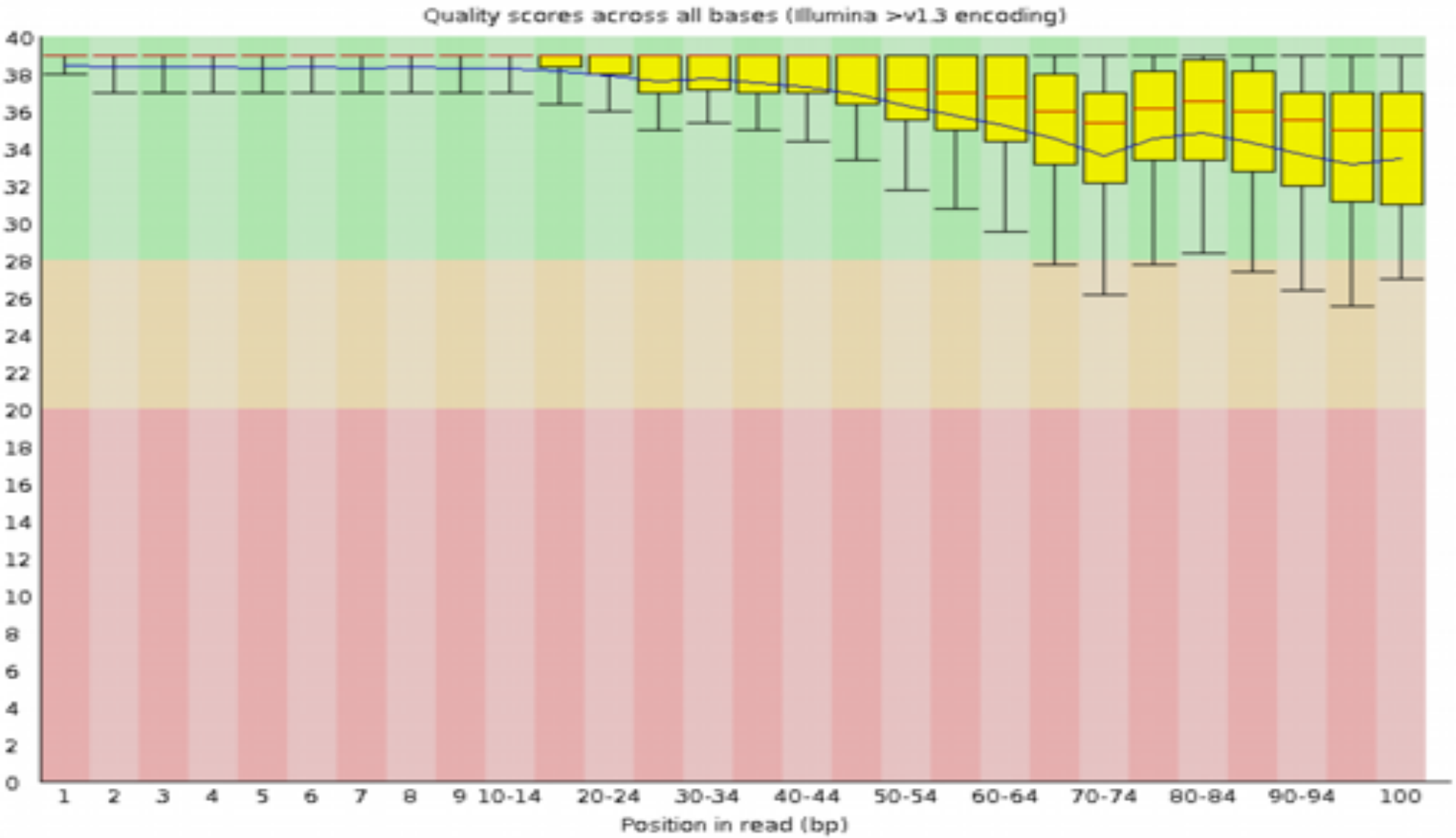
We can use quality scores to

Remove bad reads



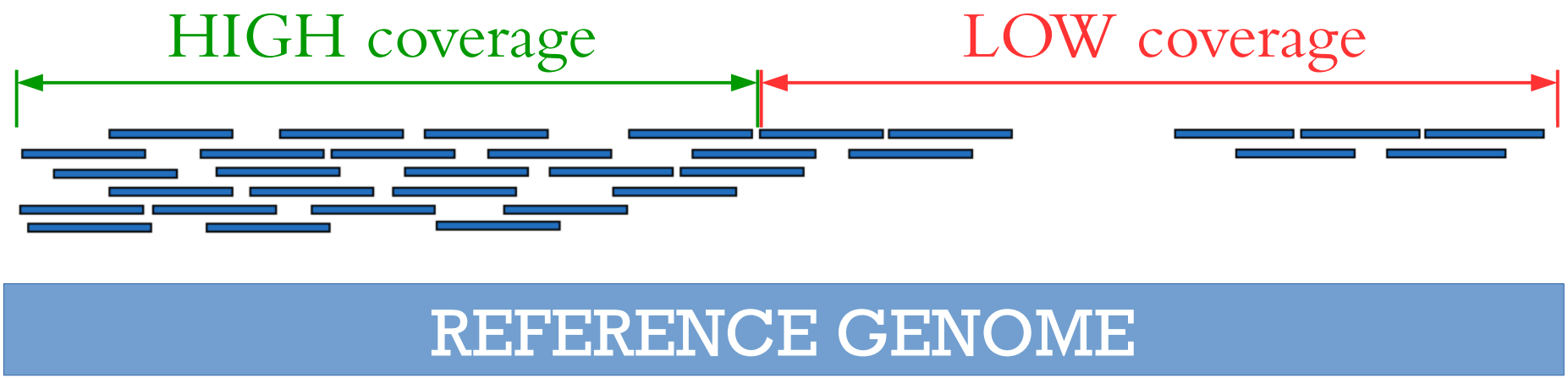
We can use quality scores to

Remove bad reads



Matching reads to a reference

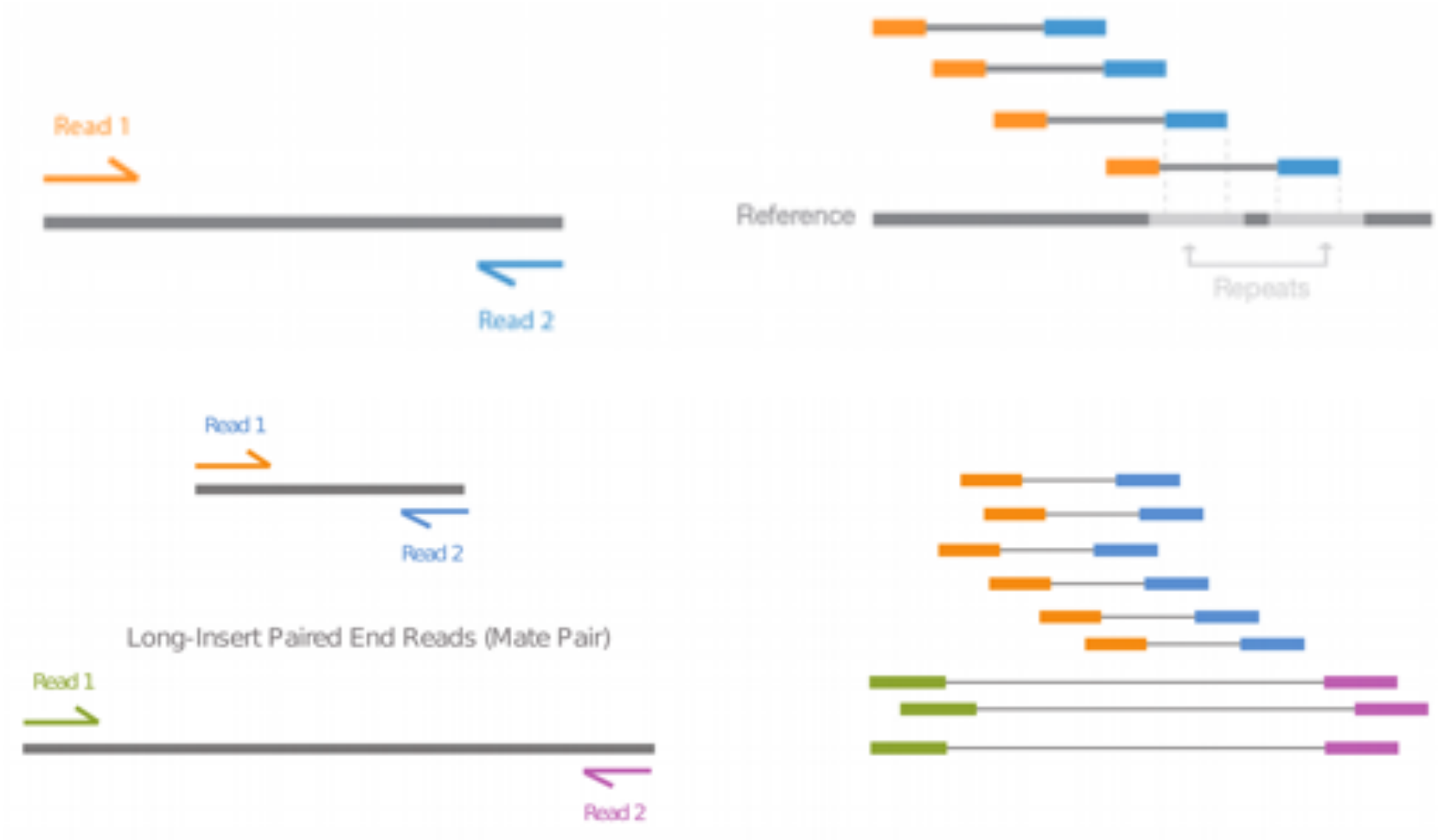
MAPPING



BWA BOWTIE SOAP NOVOALIGN

Mapping is more effective with

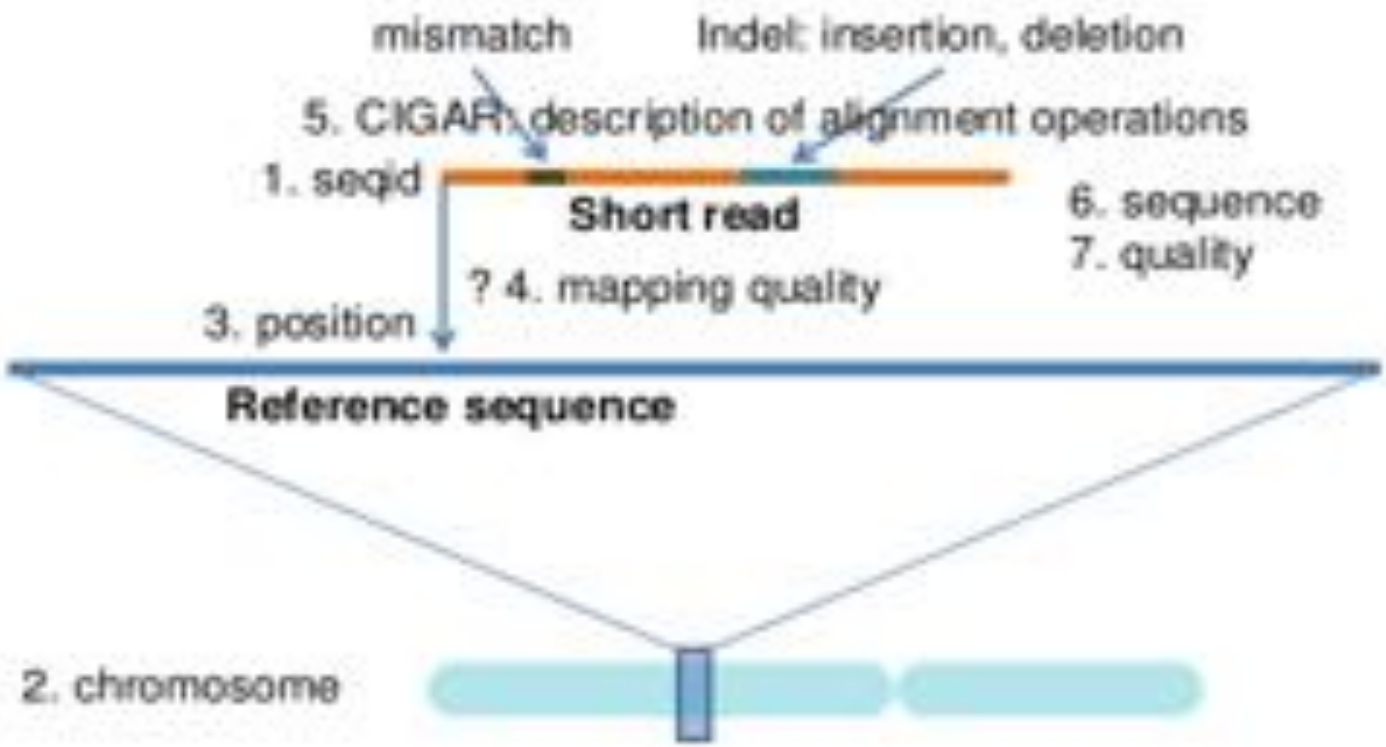
PAIRED-END DATA



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 ALIGNMENTS

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

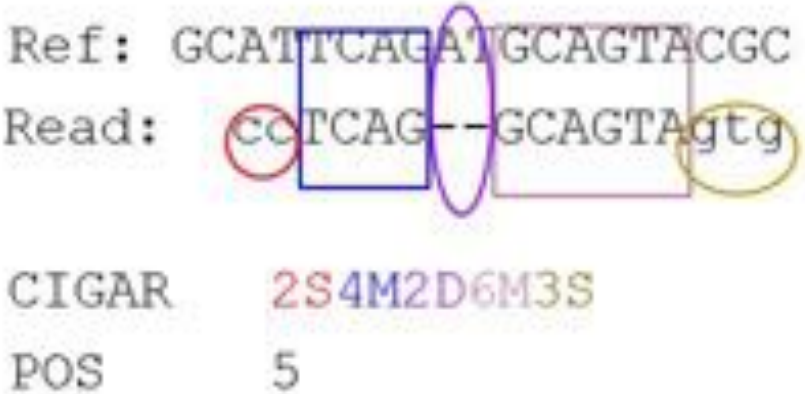
SAM ALIGNMENTS

Information about individual read alignments

Name	Flag	Ref	Pos	Mqual	Cigar	Rnext	Pnext	Len	Seq	Squal
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>...

SAM CIGAR STRING

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



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**	**113**		

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

SAM FLAG : 113 000001110001

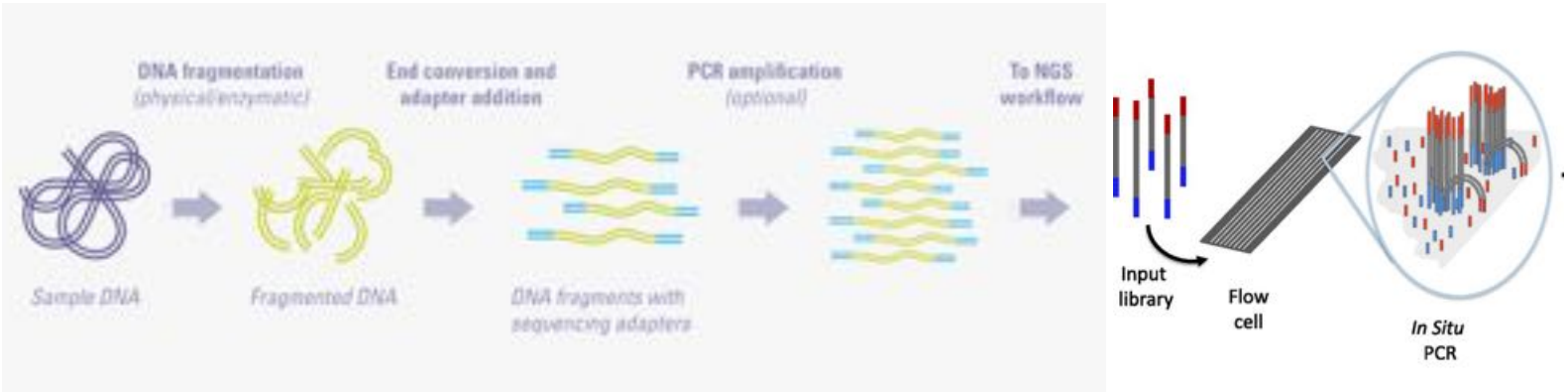
#	Binary	Decimal	Hexadecimal	Description
1	1	1	0x1	Read paired
2	10	2	0x2	Read mapped in proper pair
3	100	4	0x4	Read unmapped
4	1000	8	0x8	Mate unmapped
5	10000	16	0x10	Read reverse strand
6	100000	32	0x20	Mate reverse strand
7	1000000	64	0x40	First in pair
8	10000000	128	0x80	Second in pair
9	100000000	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:	**000001110001**	**113**		

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

Remove

CLONES

that can artificially bias coverage



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

Remove

CLONES

that can artificially bias coverage

```
 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 coverage

Possible PCR clones

```
 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 coverage

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

The

BAM FORMAT

Compressed, binary, indexed version of SAM

sample_01.sam (2.5 GB)

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>...

sample_01.bam (611 MB)

sample_01.sorted.bam

sample_01.sorted.bam.bai

downstream analysis

NGS FLOW CHART	FILE FORMAT	PROGRAMS
Raw sequence reads	Fastq	
↓		
De-multiplex & remove low quality reads	Fastq	Custom scripts Fastqc/Fastx-toolkit
↓		
Map reads to reference genome	SAM/BAM	BWA/Bowtie Soap/Novoalign
↓		
Filter unpaired, unmapped & duplicate reads	SAM/BAM	SAMtools/Picard

DOWNSTREAM ANALYSIS

USEFUL LINKS

SAMtools: http://www.htslib.org

Picard tools: 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-toolkit: http://hannonlab.cshl.edu/fastx_toolkit/

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