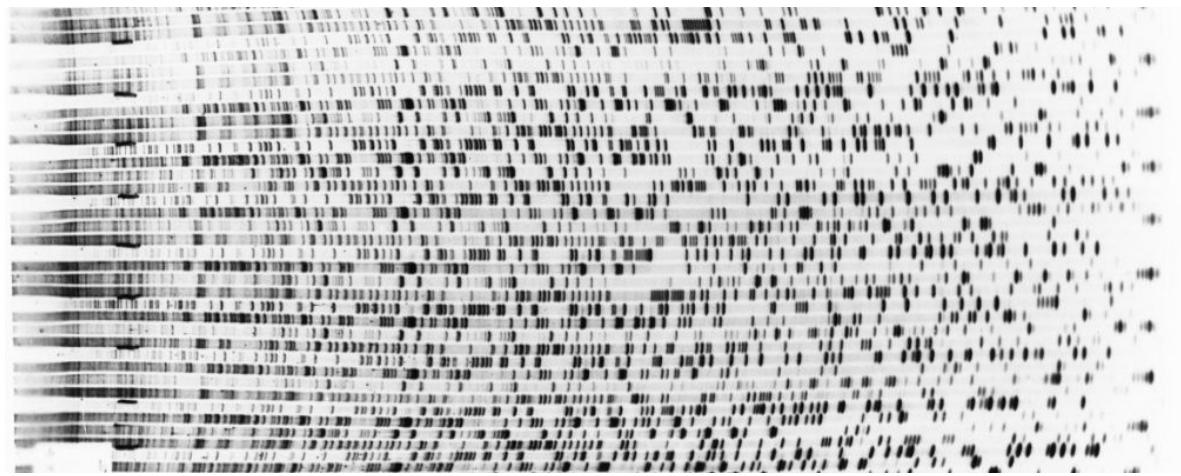
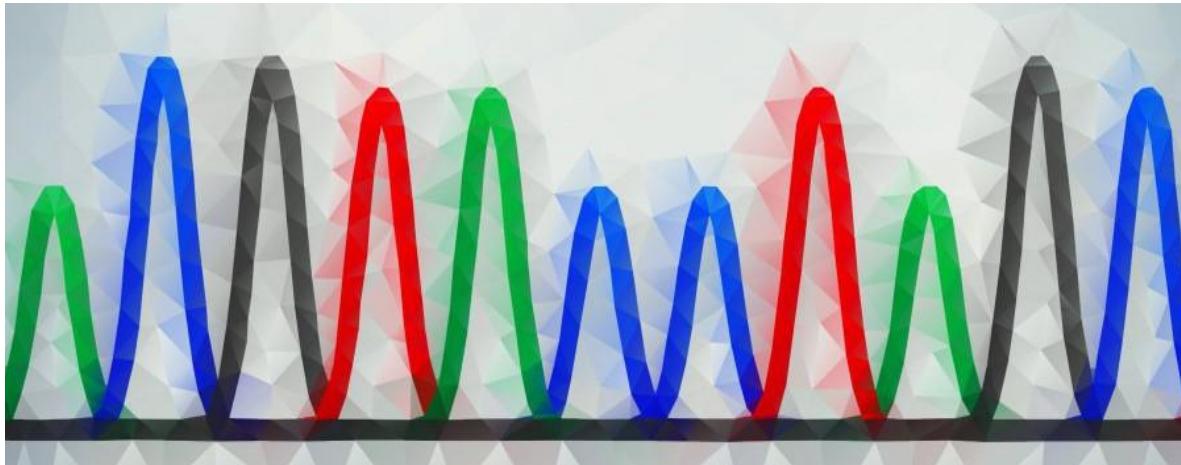


Next Generation Sequencing

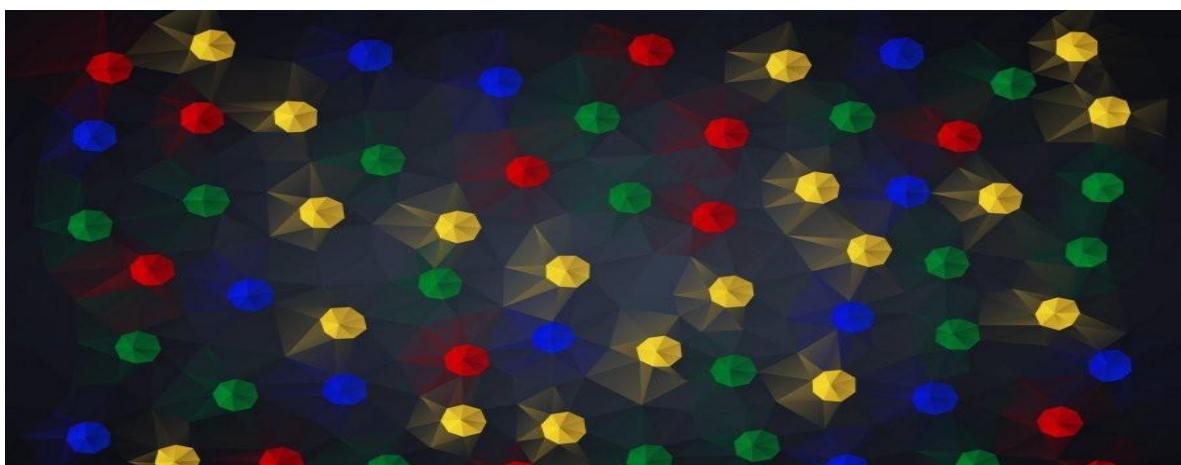
An Introduction



The future of SEQUENCING



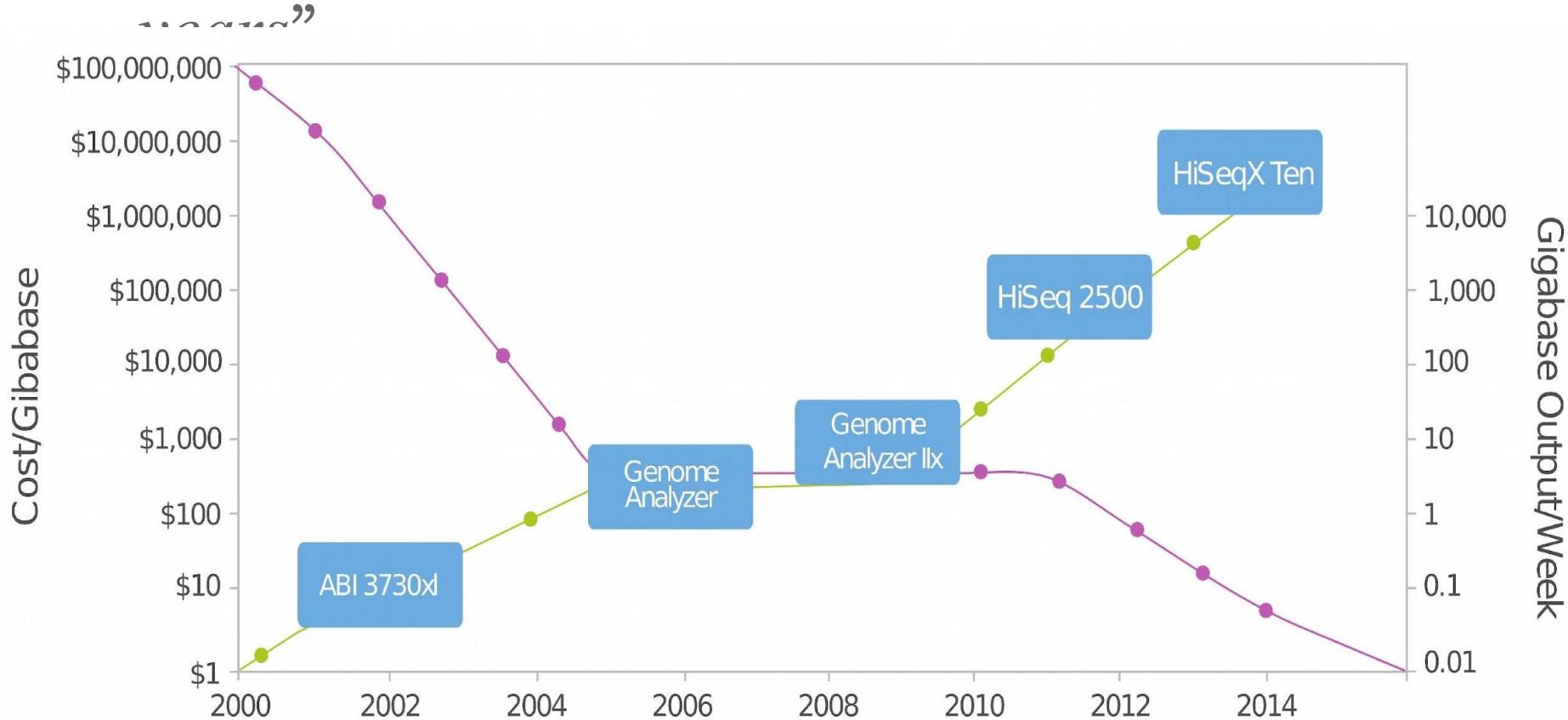
evermore
MASSIVE
evermore
PARALLEL
evermore
DATA

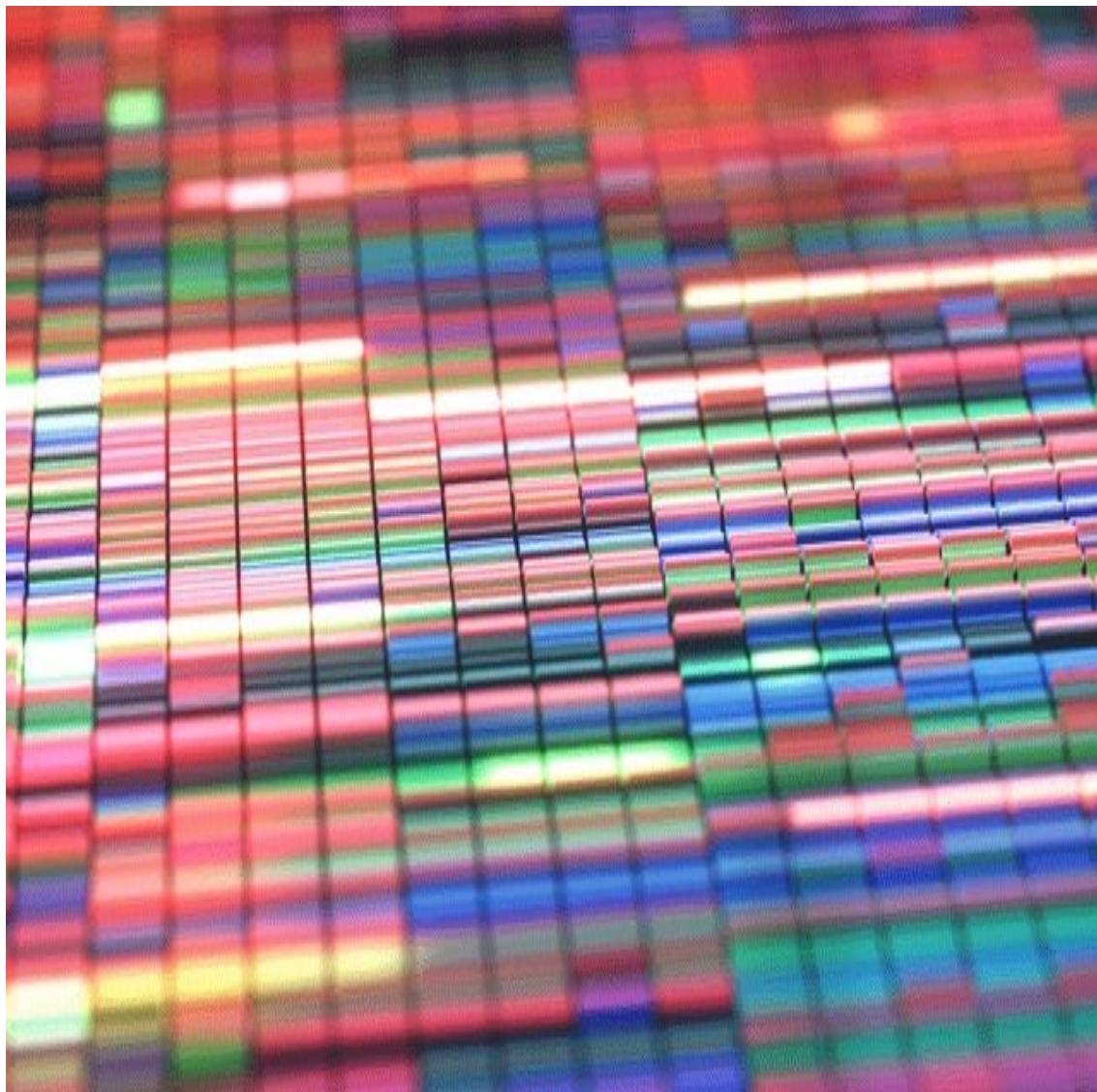


Next generation sequencing
has outpaced

MOORE'S LAW:

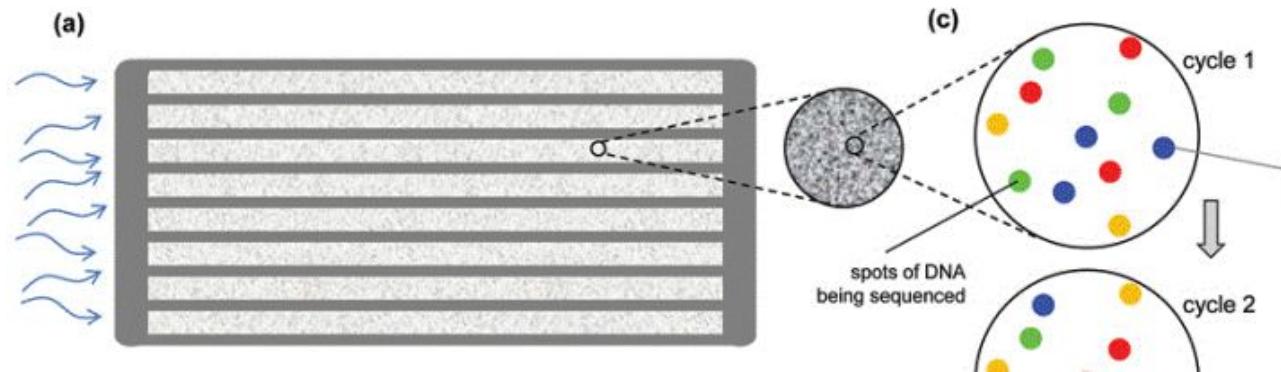
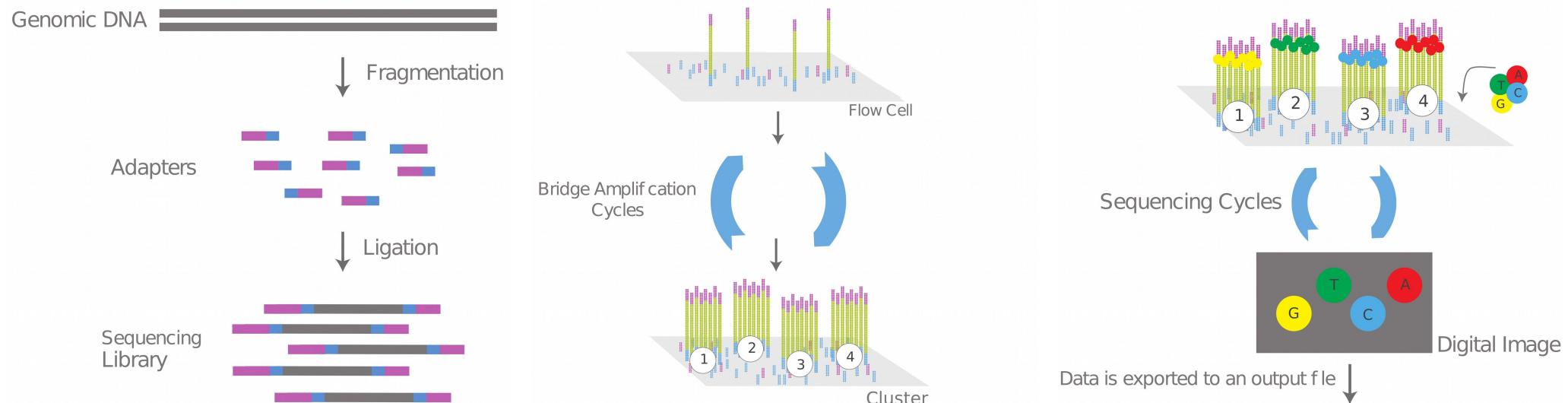
*“overall processing power of
computers will double every two
years”*



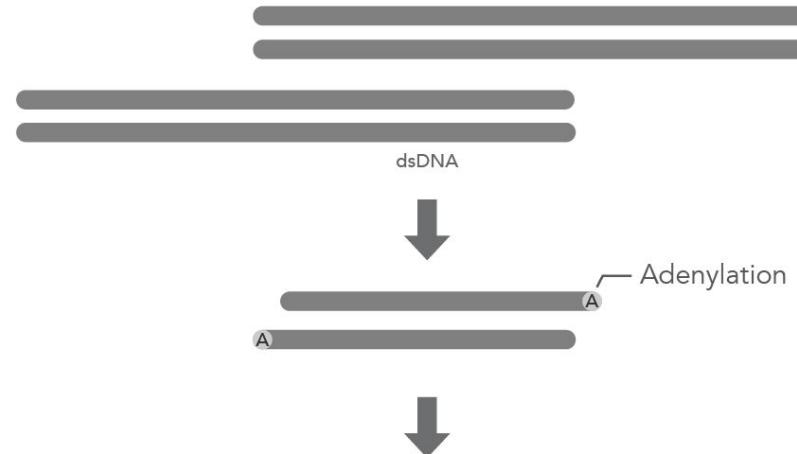


2nd Generation Sequencing

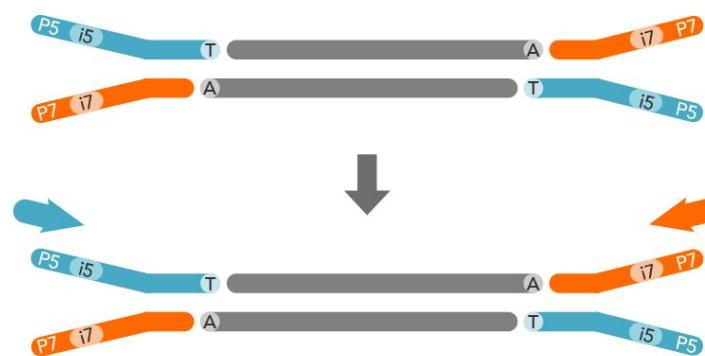
A quick look at 2nd Generation Sequencing



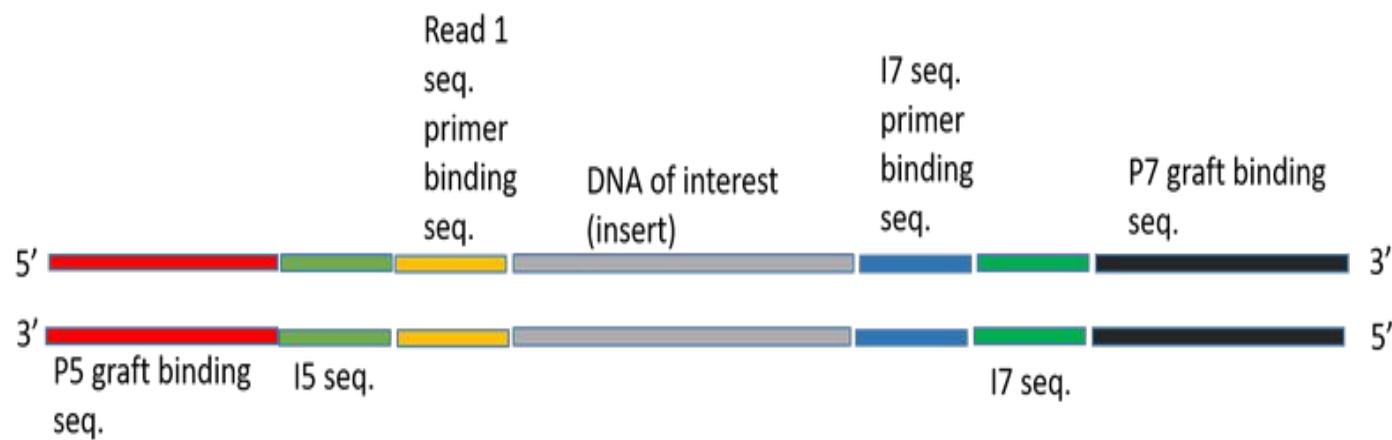
Fragmentation

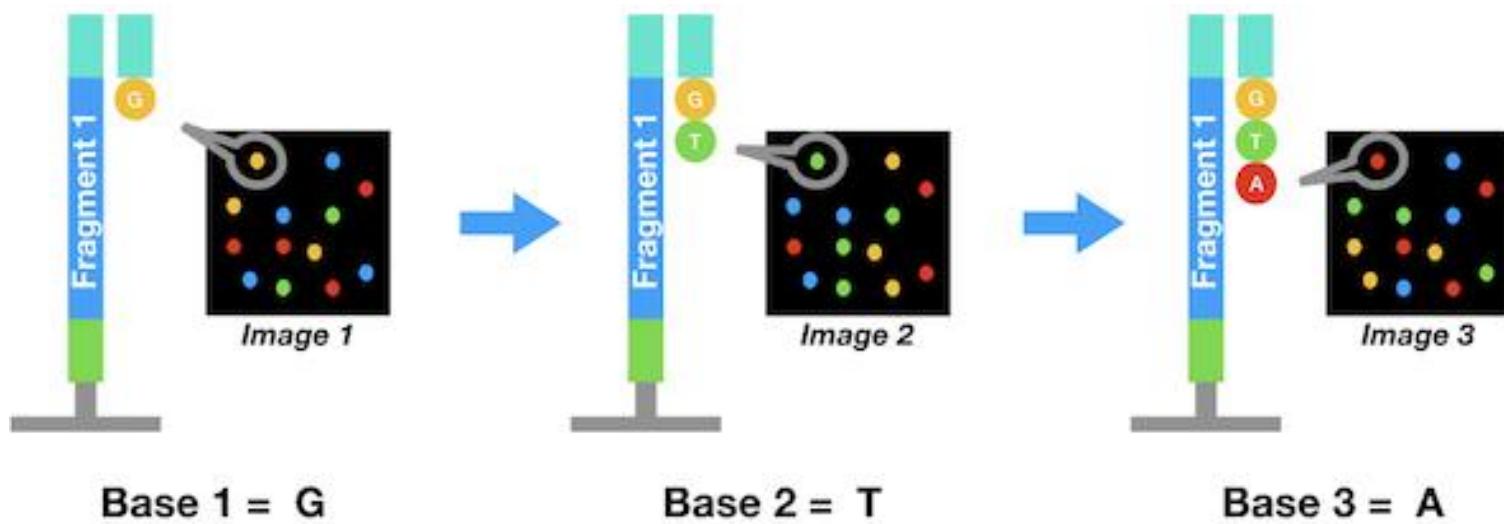
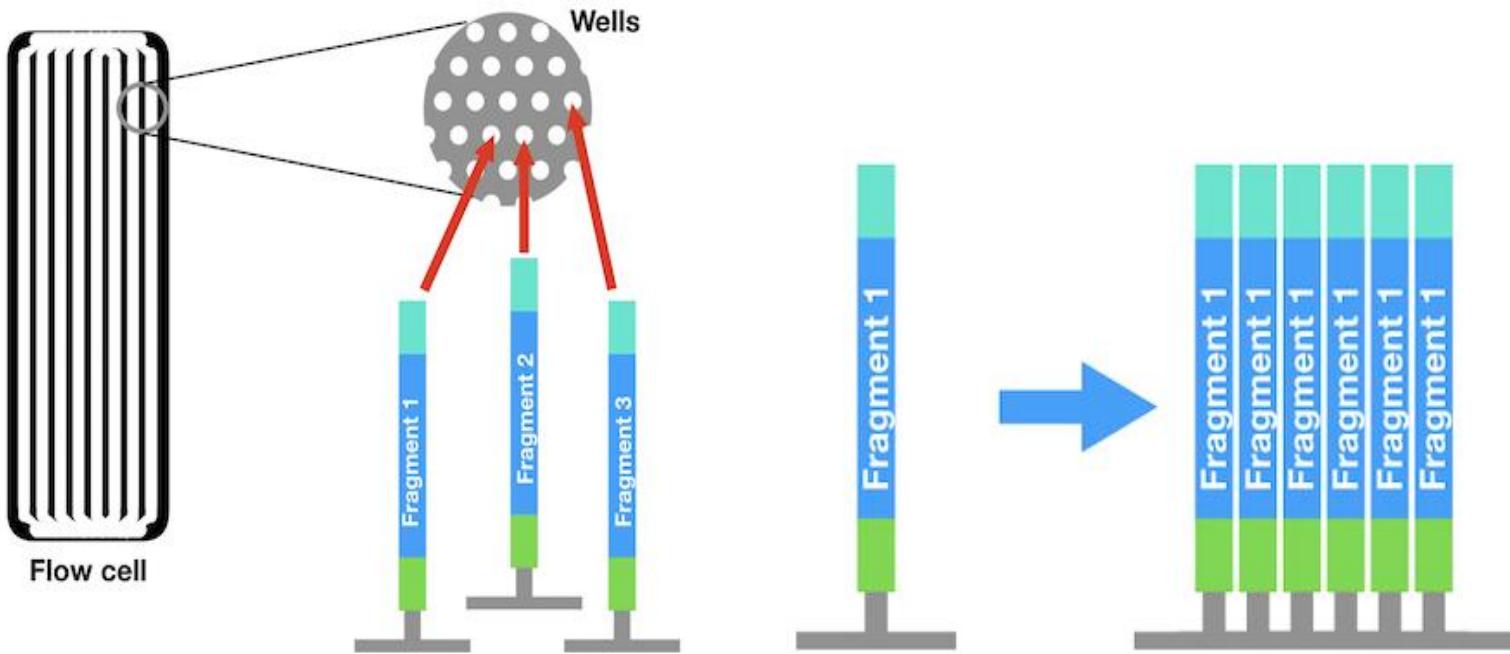


End repair and A-tailing



PCR amplification

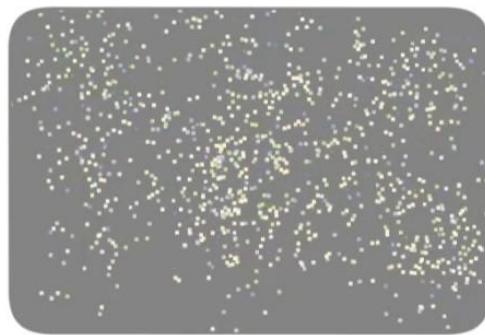




2nd Generation Sequencing

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

Sequencing



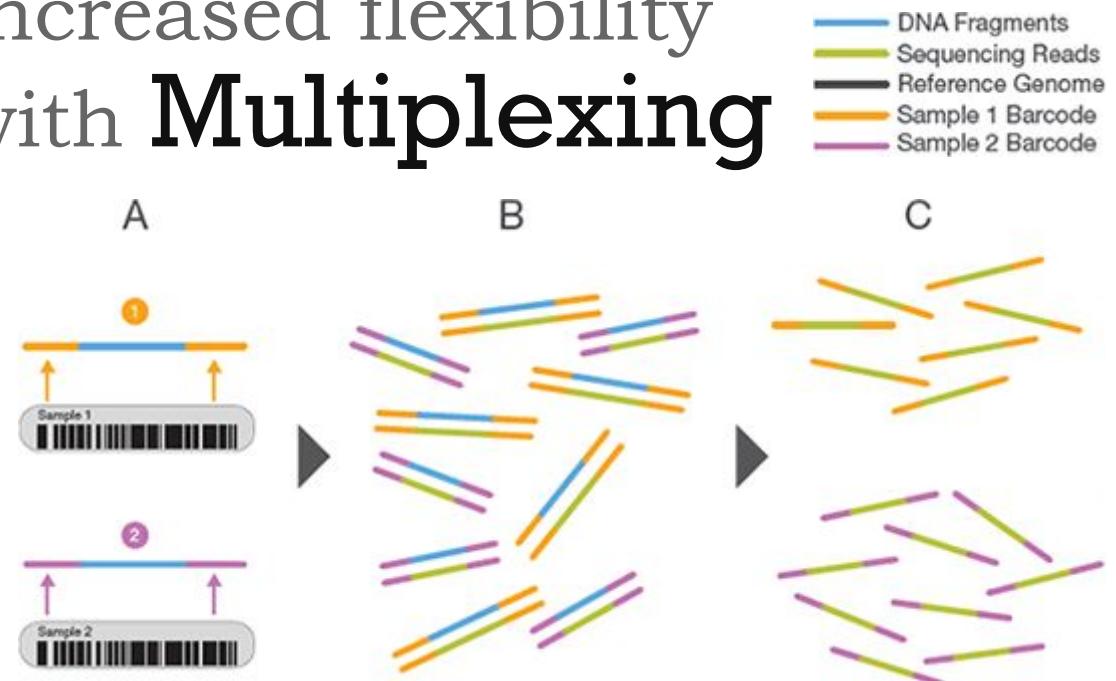
Flow cell

```
GAAACAAAAGCAATTGACA/  
CTTACGCCGTACTACCTCA/  
AGTAAGAAACAAAAGCAATT  
ACGCCGTACTACCTCAGCA  
CCTCAGCAGTAGTAAGAAA/  
GAAACAAAAGCAATTGACA/  
CTTACGCCGTACTACCTCA/  
AGTAAGAAACAAAAGCAATT  
ACGCCGTACTACCTCAGCA  
CCTCAGCAGTAGTAAGAAA/  
GAAACAAAAGCAATTGACA/  
CTTACGCCGTACTACCTCA/  
AGTAAGAAACAAAAGCAATT  
ACGCCGTACTACCTCAGCA  
CCTCAGCAGTAGTAAGAAA/  
GAAACAAAAGCAATTGACA/  
CTTACGCCGTACTACCTCA/  
AGTAAGAAACAAAAGCAATT  
ACGCCGTACTACCTCAGCA
```

Increased coverage with Paired-end sequencing



Increased flexibility with Multiplexing



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

```
GTTGAGGCTTGCCTTTGGTACGCTGGACTTTGT  
GTACTCGTCGCTGCCTTGAGGCTTGCCTTTGGT  
ATGGTACGCTGGACTTTGTAGGATAACCCCTCGCTT  
TTGCCTTATGGTACGCTGGACTTTGTAGGATAACC  
CTTGCCTTATGGTACGCTGGACTTTGTAGGATAAC  
TTGCCTTATGGTACGCTGGACTTTGTAGGATAAC  
GCGTTATGGTACGCTGGACTTTGTAGGATAACCC  
GAGGCTTGCCTTATGGTACGCTGGACTTTGTAGG  
GCGTTGAGGCTTGCCTTATGGTACGCTGGATT  
CGTTATGGTACGCTGGACTTTGTAGGATAACCC  
ATGGTACGCTGGACTTTGTAGGATAACCCCTCGCTT  
GTTTATGGTACGCTGGACTTTGTAGGATAACCC  
TCTCGTGCCTCGCTGCCTTGAGGCTTGCCTTA  
TGCTCGTGCCTGCCTTGAGGCTTGCCTTATGGTA  
GCTCGTGCCTGCCTTGAGGCTTGCCTTATGGTAC  
TATGGTACGCTGGACTTTGTAGGATAACCC  
TCGTGCTCGCTGCCTTGAGGCTTGCCTTGGTAC  
CGTCGCTGCCTTGAGGCTTGCCTTATGGTACGCT  
GTTGAGGCTTGCCTTATGGTACGCTGGCTTT  
TTGCCTTATGGTACGCTGGACTTTGTAGGATAAC
```

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

The
FASTQ FORMAT
for efficient storage & information

The diagram illustrates the structure of a FASTQ sequence. It consists of four lines of text: 1) The Sequence ID, which starts with '@' and includes a sample identifier and lane information. 2) The Sequence itself, represented by a string of nucleotide bases (A, T, C, G). 3) A plus sign ('+') indicating the start of quality score data. 4) The Quality scores, shown as a series of ASCII characters where each character's position corresponds to a base in the sequence. Red arrows point from red labels on the left to these elements: a dot and arrow points to the Sequence ID, another dot and arrow points to the Sequence, and a third dot and arrow points to the Quality scores.

• Sequence ID
@HWUSI-EAS100R:6:73:941:1973#0/1

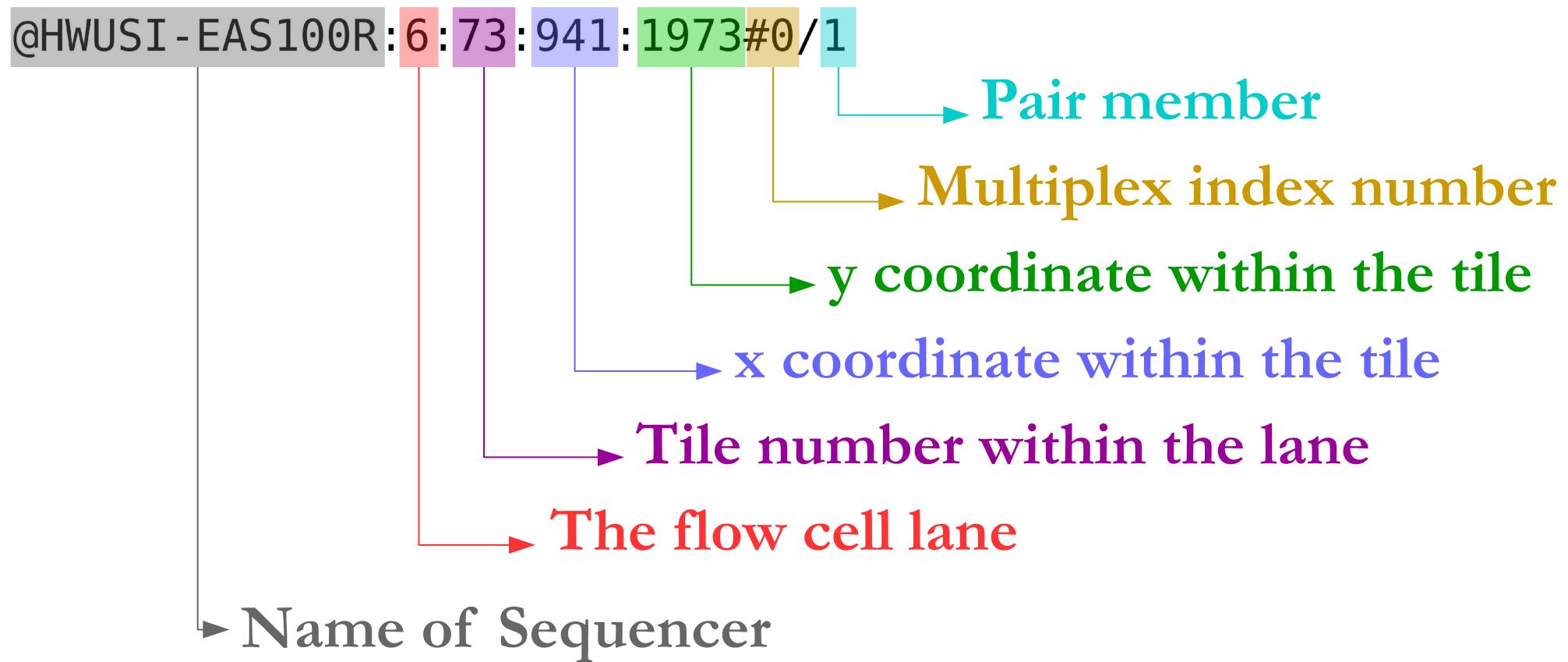
Sequence
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT
+

• Quality scores
! ' ' * ((((***)+) %%%++) (%%%) . 1*** - +* ' ')) **55CCF>>>>>CCCCCCCC65

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
GGCACTTCGATGCAGGCCATGACGGCCGAGCTCTGGAAGCGCAGGTCGGTCTTGAAGTCCTGAGCGATCTCCCTGACCAACCGCTGGAAGGGCAGCTTA
+

• Barcode #2

• RAD cut-site

The

FASTQ FORMAT

Phred Quality Scores

logarithmically related to the
base-calling error probabilities

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

or

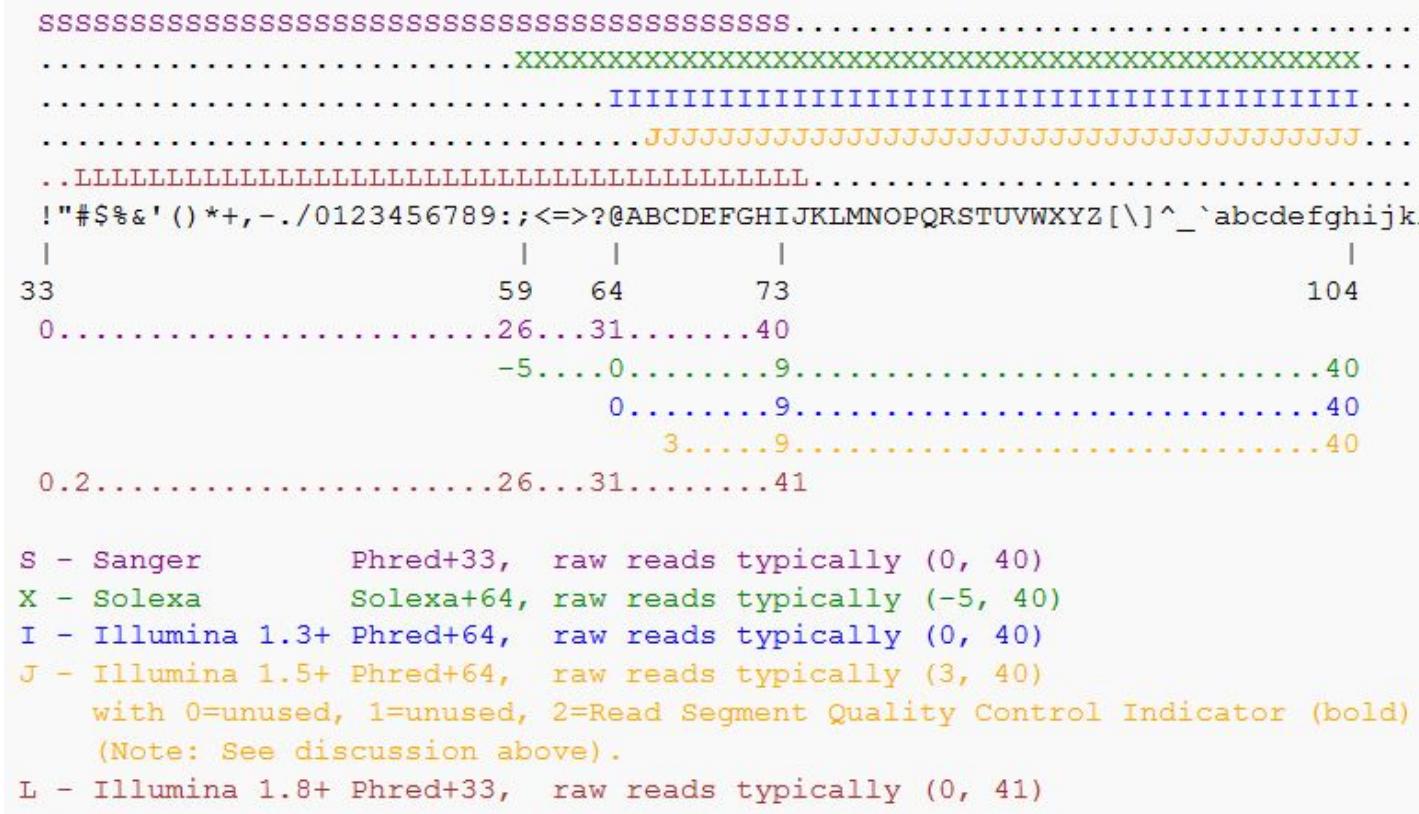
$$P = 10^{\frac{-Q}{10}}$$

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

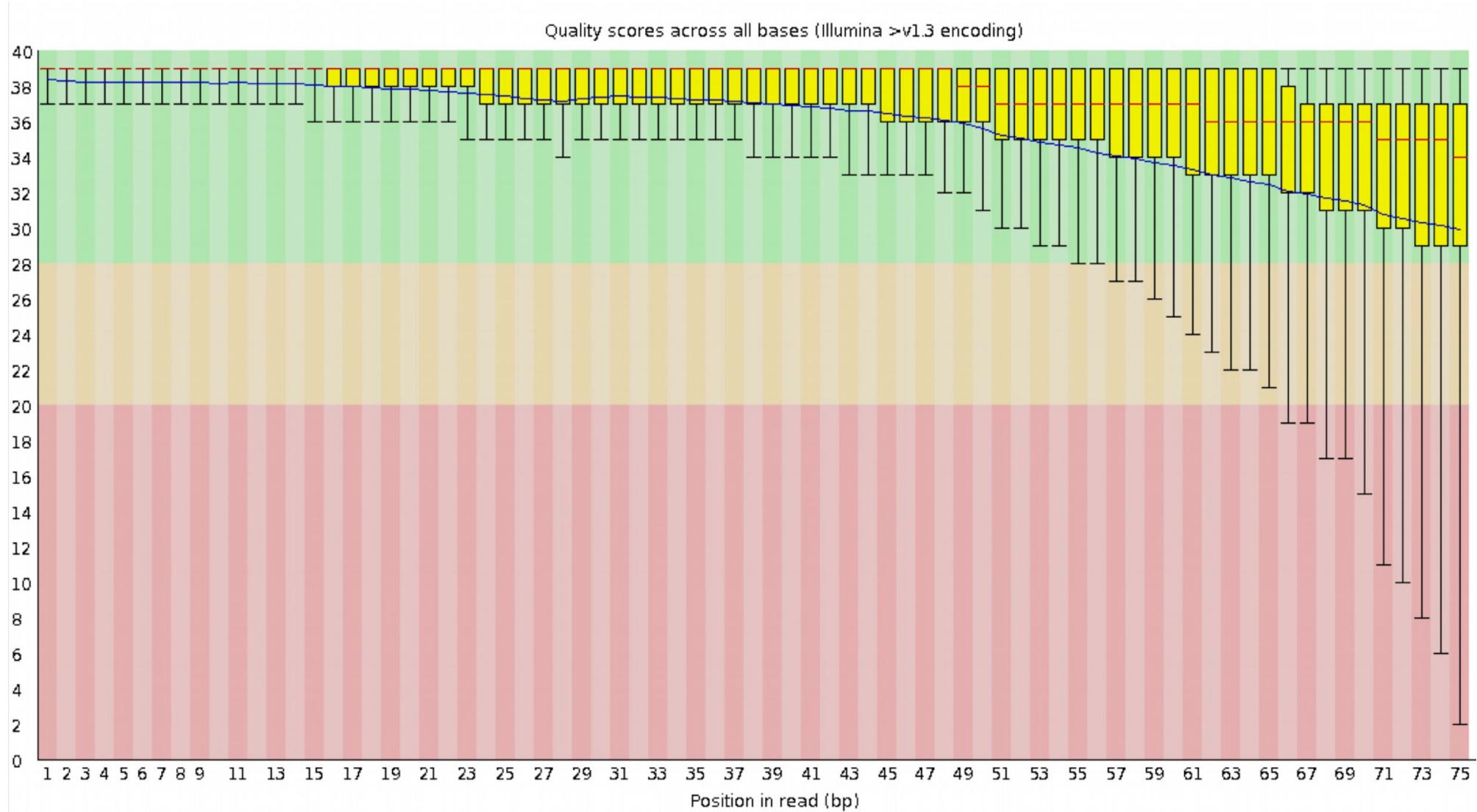
+

! ' ' * (((***+)) % % % + +) (% % % %) . 1 * * * - + * ' ')) * * 5 C C F > > > > C C C C C C C 6 5



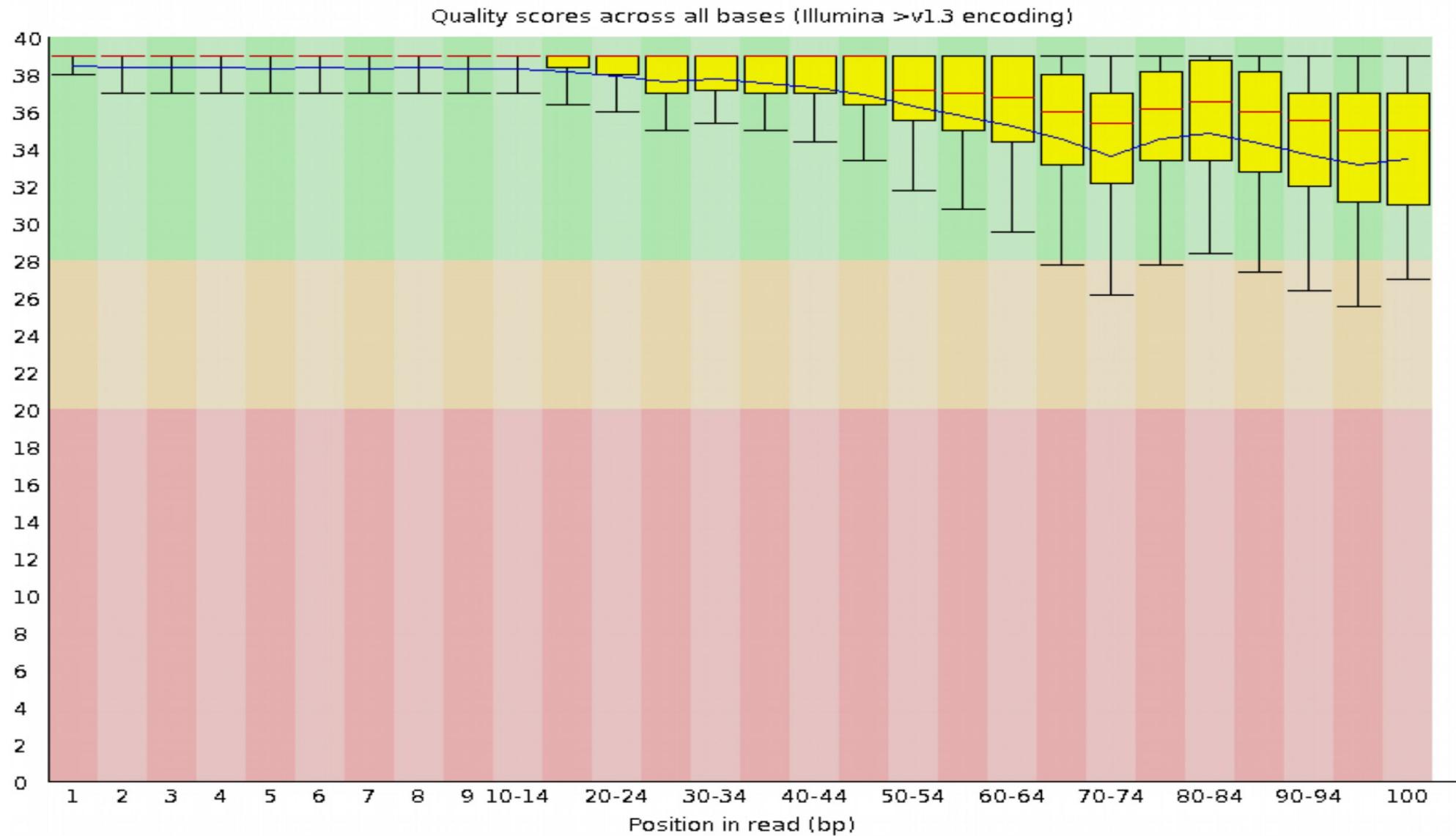
We can use quality scores to

Remove bad reads



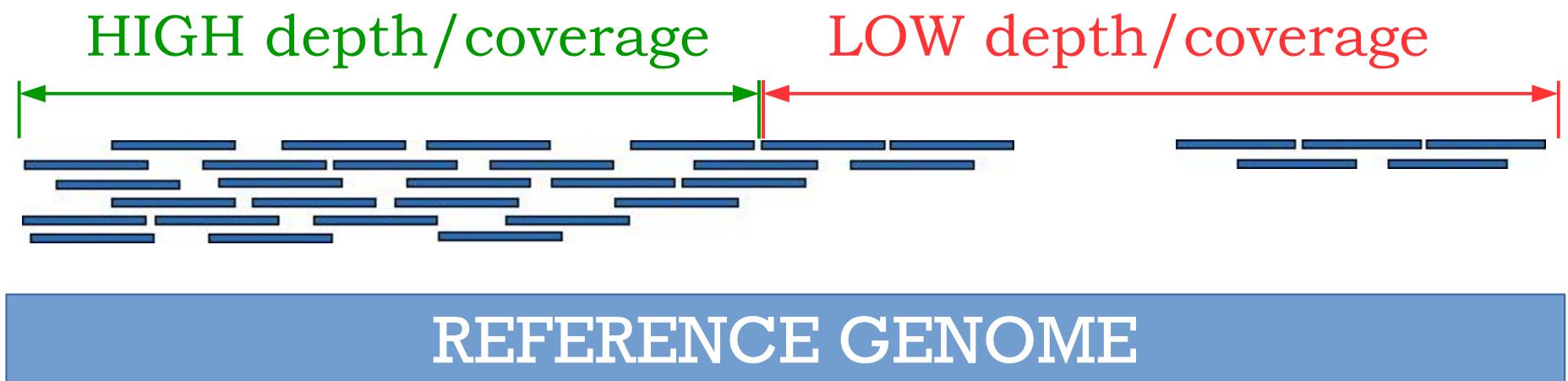
We can use quality scores to

Remove bad reads



Matching reads to a reference

MAPPING



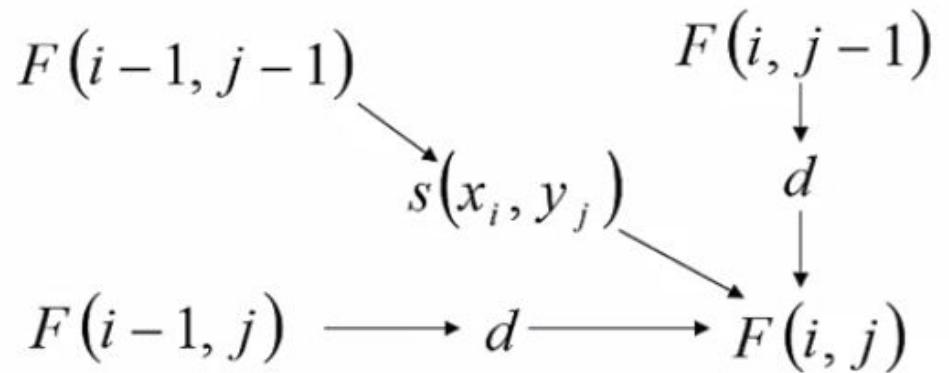
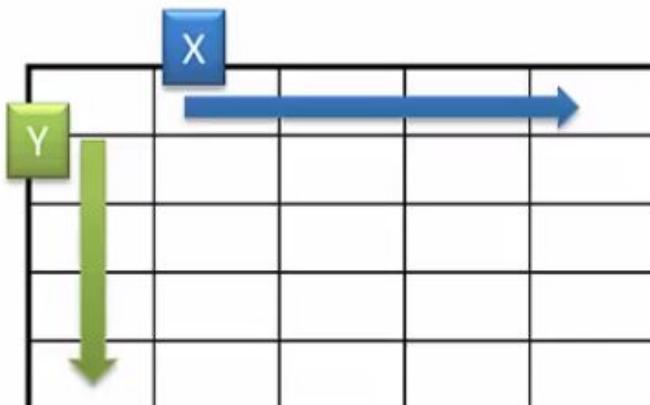
BWA BOWTIE SOAP NOVOALIGN

Dynamic Programming

Needleman–Wunsch algorithm

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

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



<i>A</i>	-35	-27	-19	-9	-3	1
<i>G</i>	-30	-22	-14	-4	2	2
<i>T</i>	-25	-17	-9	-1	-1	7
<i>C</i>	-20	-12	-4	0	4	3
<i>G</i>	-15	-7	1	5	4	-1
<i>C</i>	-10	-2	6	1	-4	-9
<i>A</i>	-5	3	-2	-7	-12	-17
-	0	-5	-10	-15	-20	-25
-	<i>A</i>	<i>C</i>	<i>T</i>	<i>G</i>	<i>T</i>	

ACGCTGA

A - - CTGT

ACGCTGA

AC - - TGT

How can we solve the read alignment problem?



Read

CTCAAACCTCTGACCTTGGTATCCACCCGCCTAGGCCTTC x million

Reference

GATCACAGGTCTATCACCTATTAAACACTCACGGGAGCTCTCATGCATTGGTATTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTC
GCACTATCTGTCTTGATTCTGCCTCATCTTATTATTCACGTTCAATATT
ACAGGGAAACATACTTAACCTAAAGTGTGTTAATTAAATTATGCTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTCCACACAGACATCATAACAAAAATTCCACCA
AACCCCCCTCCCCGCTCTGCCACAGCA
ACAAAGAACCTAACACCAGCCTAACCA
TTTAACAGTCACCCCCAACTAACAA
CTCATCAATACAACCCCCGCCAT
CCCCGAACCAACCAACCCAAAC
GCAATACACTGACCCGCTAACAC
CTAGCCTTCTATTAGCTTCTAG
TCACCCCTCTAAATCACACGAT
AAAACGCTTAGCCTAGCCACAC
ACGAAAGTTAACTAAGCTATACT
GGTCACACGATTAACCCAAGTCAT
TCCCCAATAAAGCTAAACTCACCTG
TACGAAAGTGGCTTTAACATATCTGAAC
TACCCCACATGCTTAGCCCTAACCTCAACAC
CACTAGAGCCACAGCTTAAACTCAAAGGACCTGGCGGTGCTTCAT
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCCCTTGTCA
CCGCATCTTCAGCAAACCCGTATGAAGGCTACAAAGTAAGCGCAAGTAC
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTCA
AAAACATACGATAGCCTTATGAAACTTAAGGGTCAAGGGTGGGATTAGCAGTA
AGTAGAGTGTCTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCGTACCC
AAGTATACTTCAAAGGACATTTAACTAAACCCCTACGCATTATATAGAGGAGACAA
CGTAACCTCAAACCTCTGCCTTGGTATCCACCCGCCTGGCTACCTGCATAATGAAG
AAGCACCCAACCTAACCTTAGGAGATTCAACTTAACCTGACCGCTCTGAGCTAAACCTA
GCCCAACCCACTCCACCTTACCGAGACAACCTTAGCCAACCCATTACCCAAATAA
AGTATAGGCAGATAAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATATAAGCAAGGACTAACCCCTATACCTCTGCATAATGAA
TTAACTAGAAATAACTTGTCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
ACCTAAGAACAGCTAAAGAGCACACCCGCTATGTAGCTAAAGATAGTGGAGATTATA
GGTAGAGGCAGAACCTACCGAGCCTGGTATAGCTGGTGTCCAAGATAGAATCTTAG

x million

—Index—

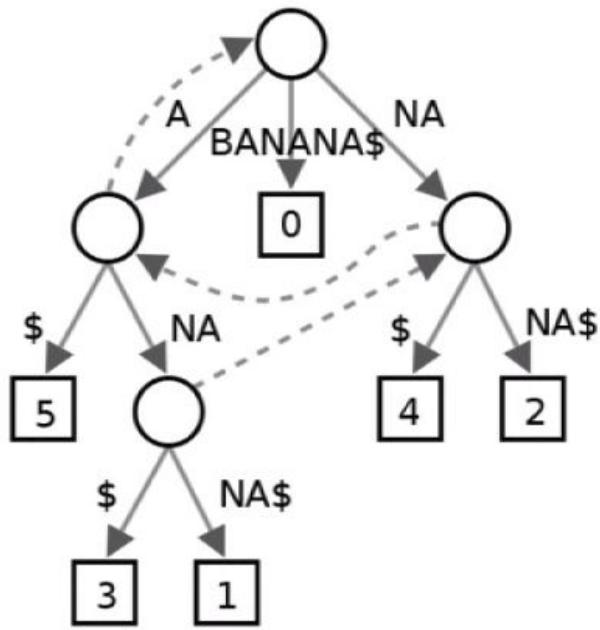
—A—

about the author 128, 132, 412
account info 295
active table of contents 34, 120-124, 238-239,
 285-286, 354, 366, 370
ACX 465-467
Adobe 506
advertising 434, 439-449
age 312
aggregator 17-18, 322
alignment 68, 101-103, 105-106, 229-230, 261-262, 353-
 354, 380, 389
Alt codes 39
Amazon Associates 415
Amazon Follow 430, 437, 480
Amazon Giveaway 436-439
Amazon Marketing Services (AMS) 439-449
Android 167-169, 171, 371-375
apostrophe 40, 42-44
app 141-142
Apple 169, 342, 372, 506

automatic renewal 327-329, 341, 343
Automatically Update 73-75, 94, 144
AZK 371

—B—

back matter 124-129
background 47, 93, 181, 184, 192-193, 246, 252-253, 355,
 370, 385, 390
bank information 295
Barnes & Noble 506
biography 128, 132, 410
black 47, 93, 184, 192, 252-253, 355, 370, 385, 390
Blackberry 372-373
blank line 27-28, 110, 112-114, 276-277, 284-285, 385
blank page 354, 385-386
block indent 50, 52, 67, 82, 106-107, 234-235
blog 411, 429, 479
Blogger 429
bloggers 327, 430
blurb 300-306, 364, 406, 411-412, 417, 477
blurry 162-164, 172, 175, 193, 246, 387, 389
body text 66, 68, 79-82, 92-94, 115, 233-235



Suffix tree
 $\geq 45 \text{ GB}$

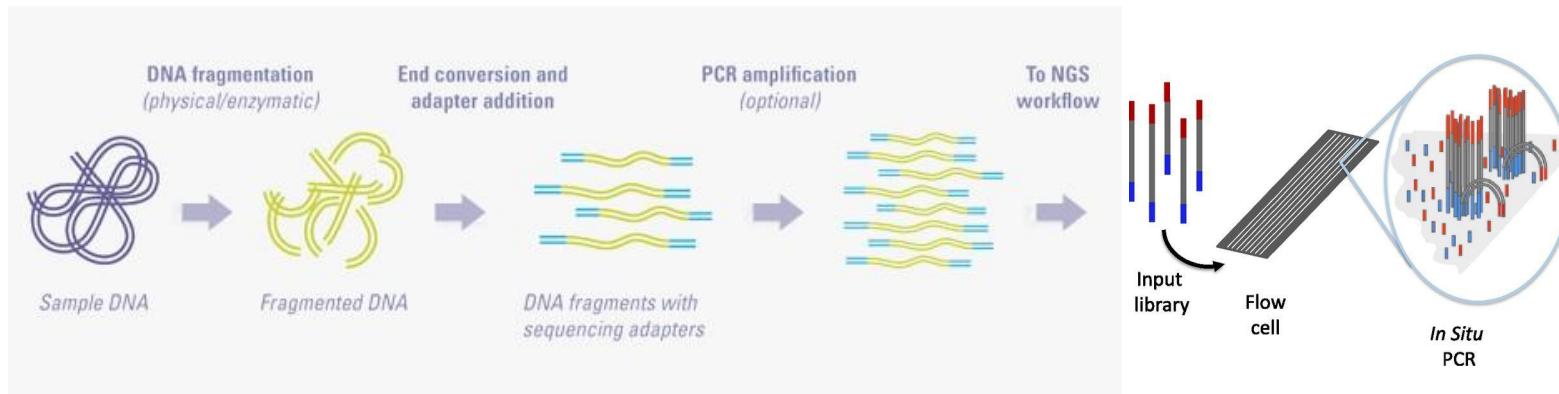
6	\$
5	A\$
3	ANAS\$
1	ANANAS\$
0	BANANA\$
4	NA\$
2	NANAS\$

Suffix array
 $\geq 12 \text{ GB}$

\$ BANANA
A \$ BANAN
ANA \$ BAN
ANANA \$ B
BANANA \$
NA \$ BANA
NANA \$ BA

FM Index
 $\sim 1 \text{ GB}$

Remove **CLONES** that can artificially bias depth



1. Shatter 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

Remove
CLONES
that can artificially bias depth

TCTCGTGCTCGTCGCTGCAGGCTTGCCTTA
TCGTGCTCGTCGCTGCAGGCTTGCCTTTTG
GTACTCGTCGCTGCAGGCTTGCCTTTTGTTGGT
TGCTCGTCGCTGCAGGCTTGCCTTAATGGTA
GCTCGTCGCTGCAGGCTTGCCTTAATGGTAC
CGTCGCTGCAGGCTTGCCTTAATGGTACGCT
GCGTTGAGGCTTGCCTTAATGGTACGCTGGATT
GTTGAGGCTTGCCTTTGTTGGTACGCTGGACTTGT

Possible PCR clones



CTCTCGTGCTCGTCGCTGCGTTGAGGCTTATGGTACGCTGGACTTGTAGGATAACCCTCGCTTC

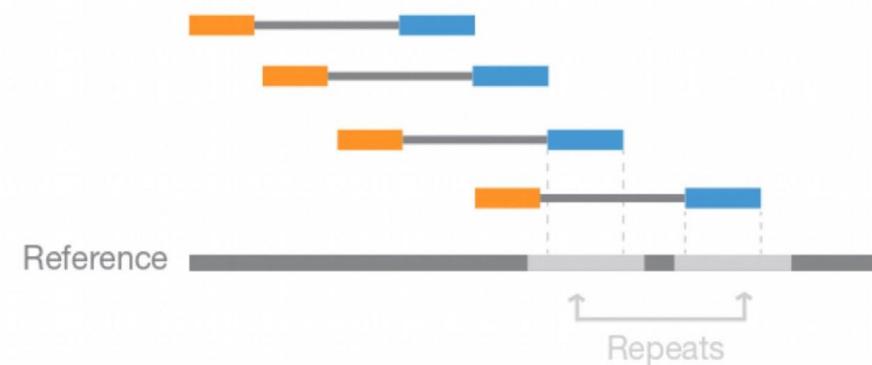
Remove
CLONES
that can artificially bias depth

TCTCGTGC~~T~~CGCTGCGTTGAGGCTTGC~~G~~TTA
TCGTGCTCG~~T~~CGCTGCGTTGAGGCTTGC~~G~~TTT~~T~~TG
GTA~~T~~CTCG~~T~~CGCTGCGTTGAGGCTTGC~~G~~TTT~~T~~GGT
TGCTCG~~T~~CGCTGCGTTGAGGCTTGC~~G~~TTA~~T~~GGTA
GCTCG~~T~~CGCTGCGTTGAGGCTTGC~~G~~TTA~~T~~GGTAC
CGTCGCTGCGTTGAGGCTTGC~~G~~TTA~~T~~GGTACGCT
GC~~G~~TTGAGGCTTGC~~G~~TTA~~T~~GGTACGCTGGATT~~T~~T
GTTGAGGCTTGC~~G~~TTT~~T~~GGTACGCTGGACTTGT
T~~A~~TGGTACGCTGGACTTGTAGGATA~~C~~CC~~C~~TCGCTT
A~~T~~GGTACGCTGGACTTGTAGGATA~~C~~CC~~C~~TCGCTT
A~~T~~GGTACGCTGGACTTGTAGGATA~~C~~CC~~C~~TCGCTT

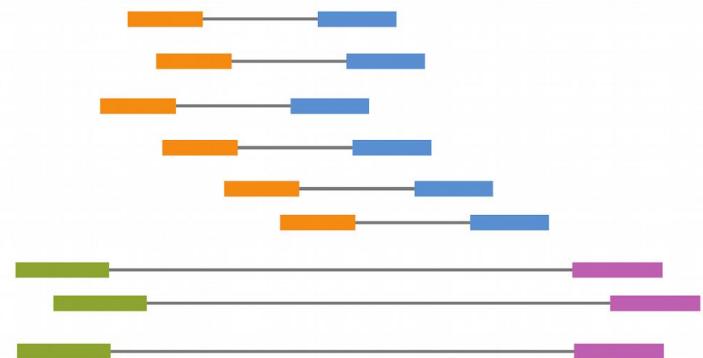
CTCTCGTGC~~T~~CGCTGCGTTGAGGCTTGC~~G~~TTA~~T~~GGTACGCTGGACTTGTAGGATA~~C~~CC~~C~~TCGCTTTC

Mapping is more effective with

PAIRED-END DATA

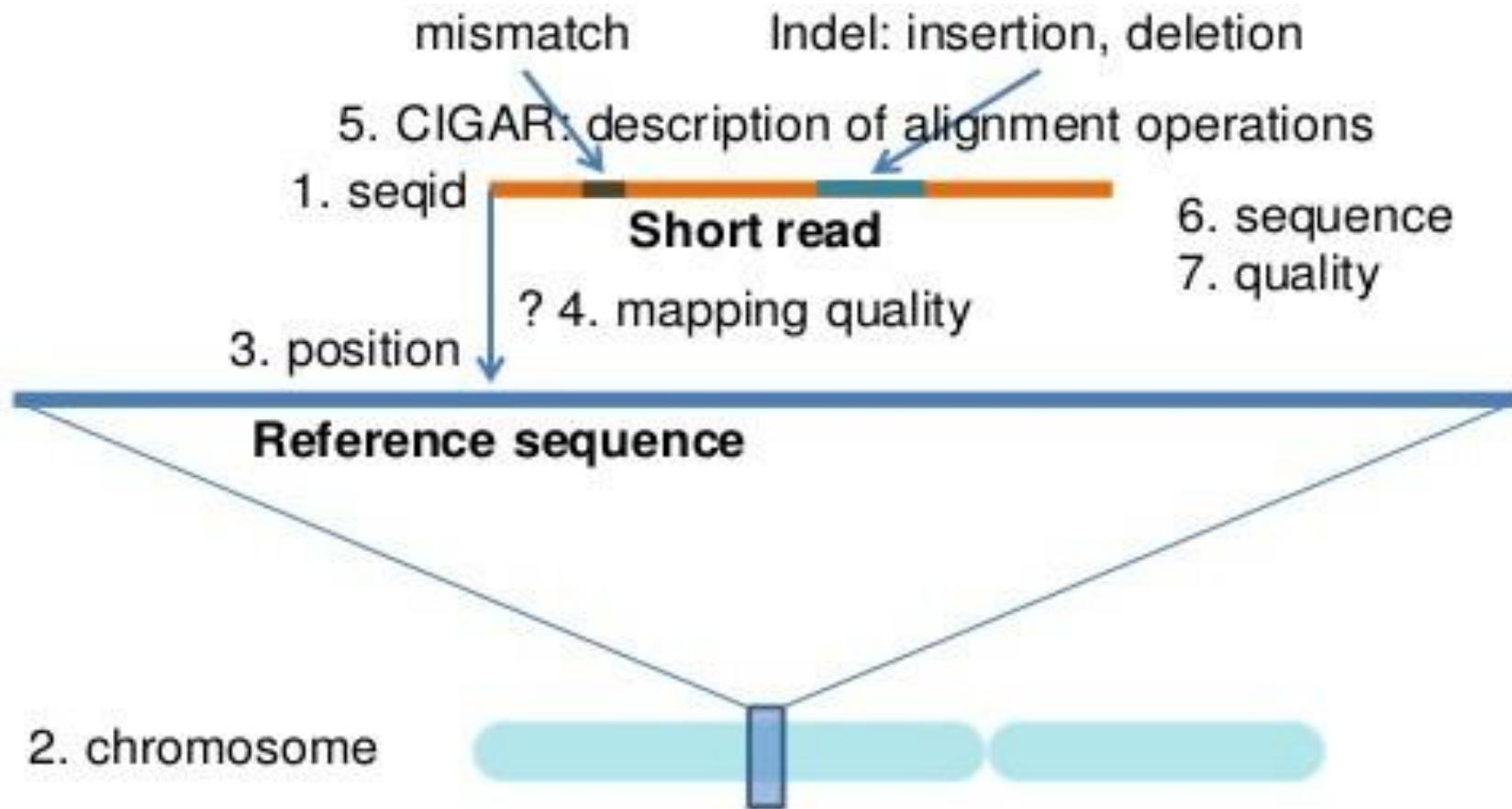


Long-Insert Paired End Reads (Mate Pair)



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

	Flag	Pos	Cigar	Pnext	Seq		
Name							
Ref						Squal	
1:497:R:-272+13M17D24M	113	1	497	37	37M	CGGGTCT...	0;=====...
19:20389:F:275+18M2D19M	99	1	176	0	37M	TATGACT...	>>>>>...
19:20389:F:275+18M2D19M	147	1	179	0	18M2D19M	GTAGTAC...	;44999;...
9:21597+10M2I25M:R:-209	83	1	216	0	8M2I27M	CACCACA...	<;9<<5>...

Diagram illustrating the mapping of SAM header fields to their corresponding values in the alignment records:

- Flag:** Indicated by a red double-headed arrow pointing to the first column.
- Pos:** Indicated by a blue double-headed arrow pointing to the second column.
- Cigar:** Indicated by a yellow double-headed arrow pointing to the third column.
- Pnext:** Indicated by an orange double-headed arrow pointing to the fourth column.
- Seq:** Indicated by a red double-headed arrow pointing to the fifth column.
- Name:** Indicated by a grey box below the first column.
- Ref:** Indicated by a purple box below the second column.
- Mqual:** Indicated by a green box below the third column.
- Rnext:** Indicated by a cyan box below the fourth column.
- Len:** Indicated by a blue box below the fifth column.
- Squal:** Indicated by a brown box below the sixth column.

SAM CIGAR STRING

M: match/mismatch

I: insertion

D: deletion

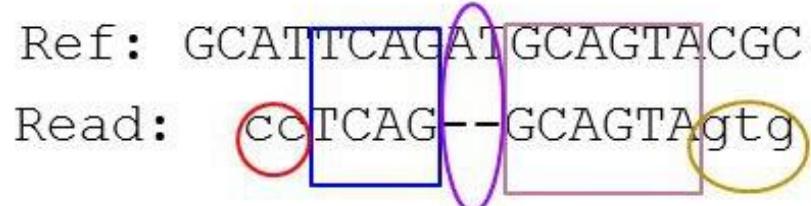
P: padding

N: skip

S: soft-clip

H: hard-clip

Ref: GCATTTCAGATGCAGTACGC

Read: 

CIGAR 

POS 5

REF: CACGATCA**GACCGATAACGTCCGA

READ1: CGATCAGAGACCGATA

READ2: ATCA*AGACCGATAAC

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  
19:20389:F:275+18M2D19M  
19:20389:F:275+18M2D19M  
9:21597+10M2I25M:R:-209
```

113	1	497	37	37M	15	100	0	CGGGTCT...	0;=====...
99	1	176	0	37M	=	179	314	TATGACT...	>>>>>...
147	1	179	0	18M2D19M	=	176	-314	GTAGTAC...	;44999;...
83	1	216	0	8M2I27M	=	214	-244	CACCACA...	<;9<<5>...



sample_01.bam (611MB)

sample_01.sorted.bam

sample_01.sorted.bam.bai

downstream analysis

NGS FLOW CHART

Raw sequence reads



Demultiplex and remove
low quality reads



Map reads to reference
genome



Filter unpaired,
unmapped and duplicate
reads

FILE FORMAT

Fastq

Fastq

SAM/BAM

SAM/BAM

PROGRAMS

Customscripts
Fastqc/Fastx-toolkit

BWA/Bowtie
Soap/Novoalign

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>

L

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

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