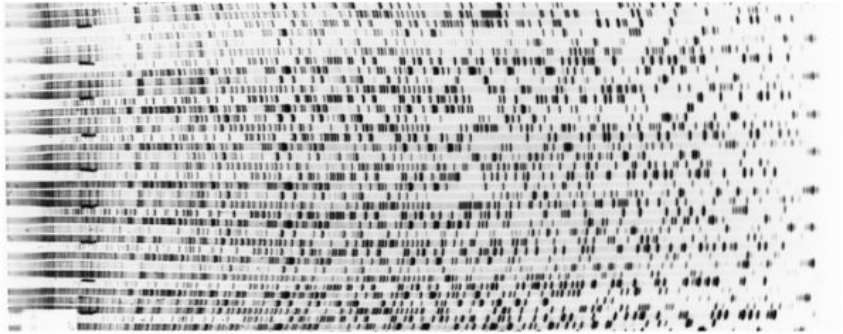
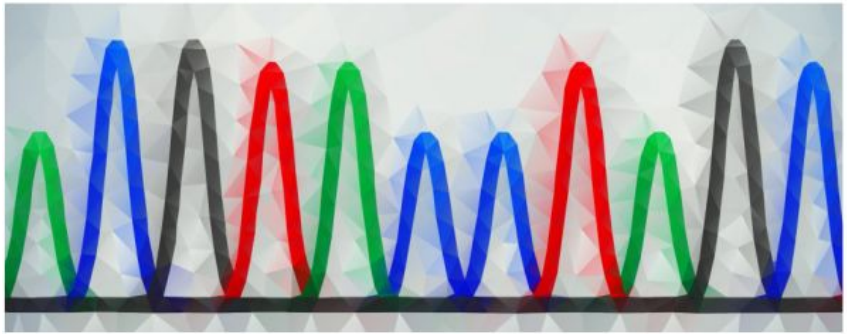
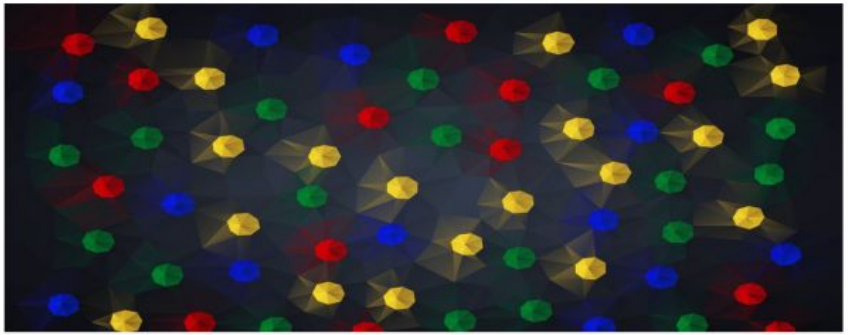
GENOME ANALYSIS & BIOINFORMATICS

Week 1:
Introduction to Genomics & Sequencing Technologies





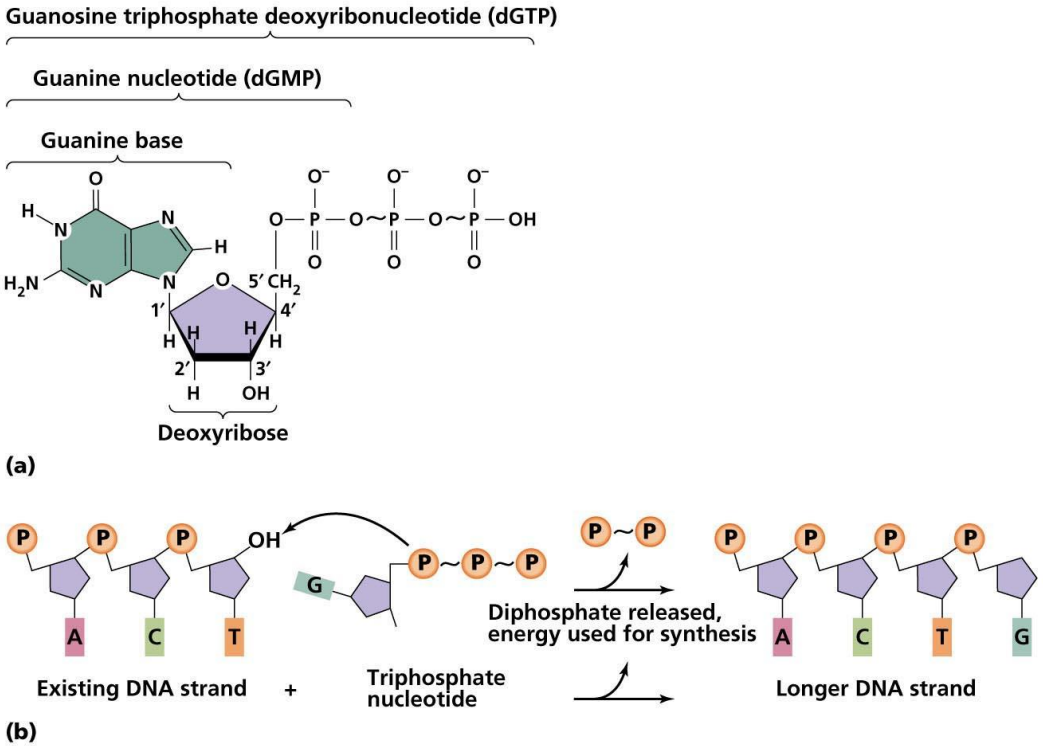


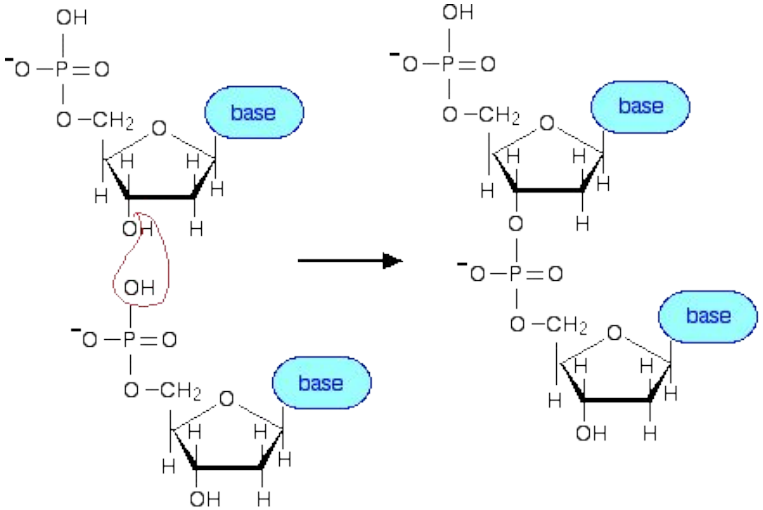
The future of SEQUENCING

ever more
MASSIVE

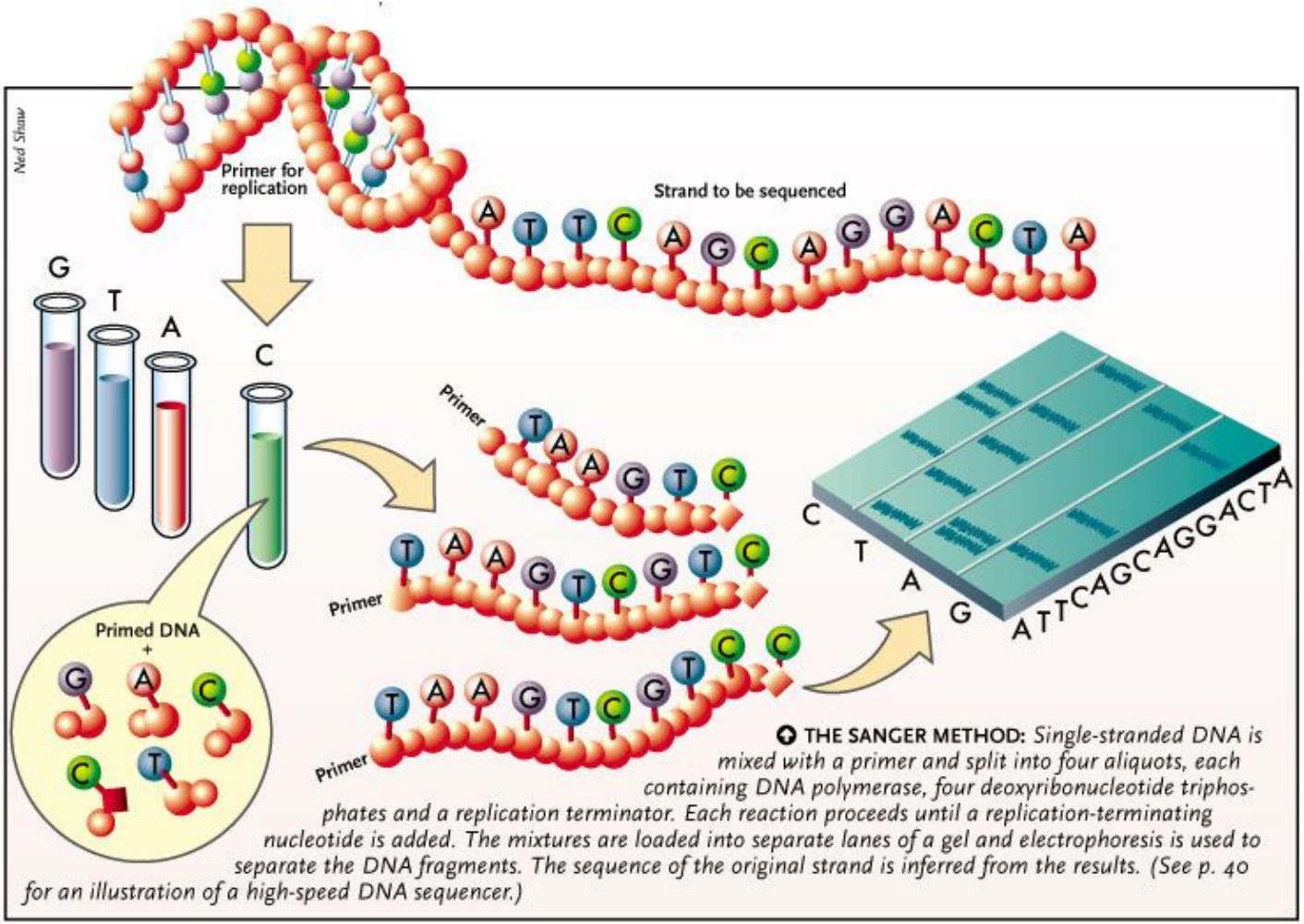
ever more
PARALLEL

ever more
DATA

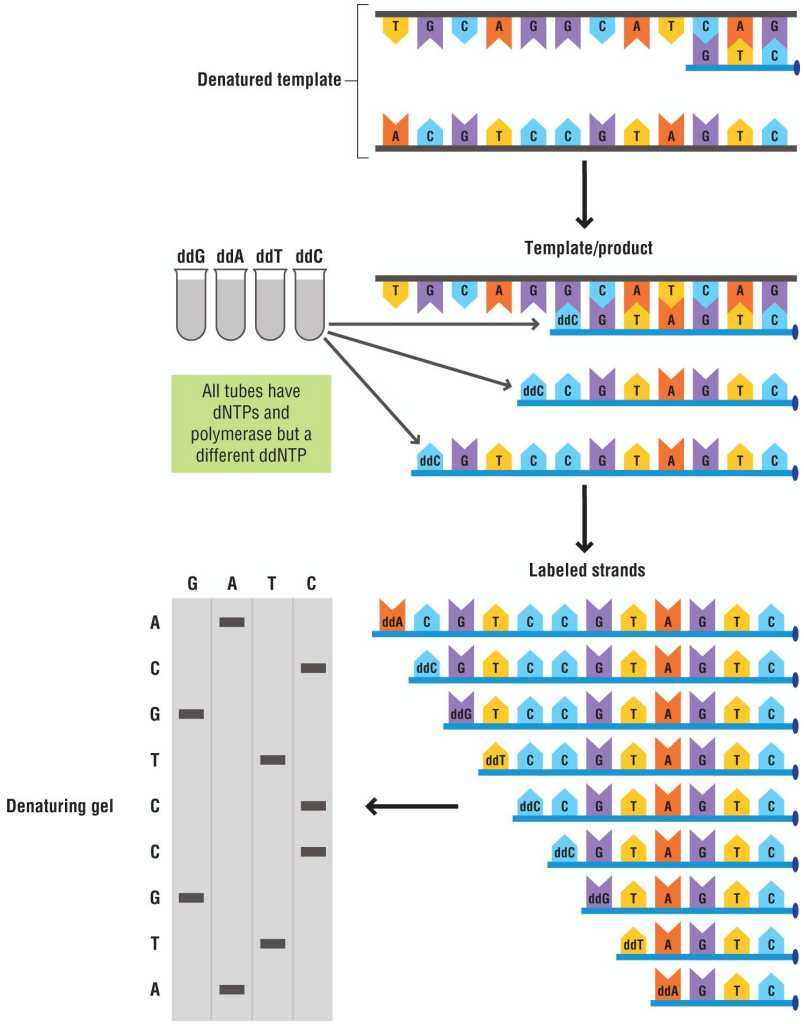


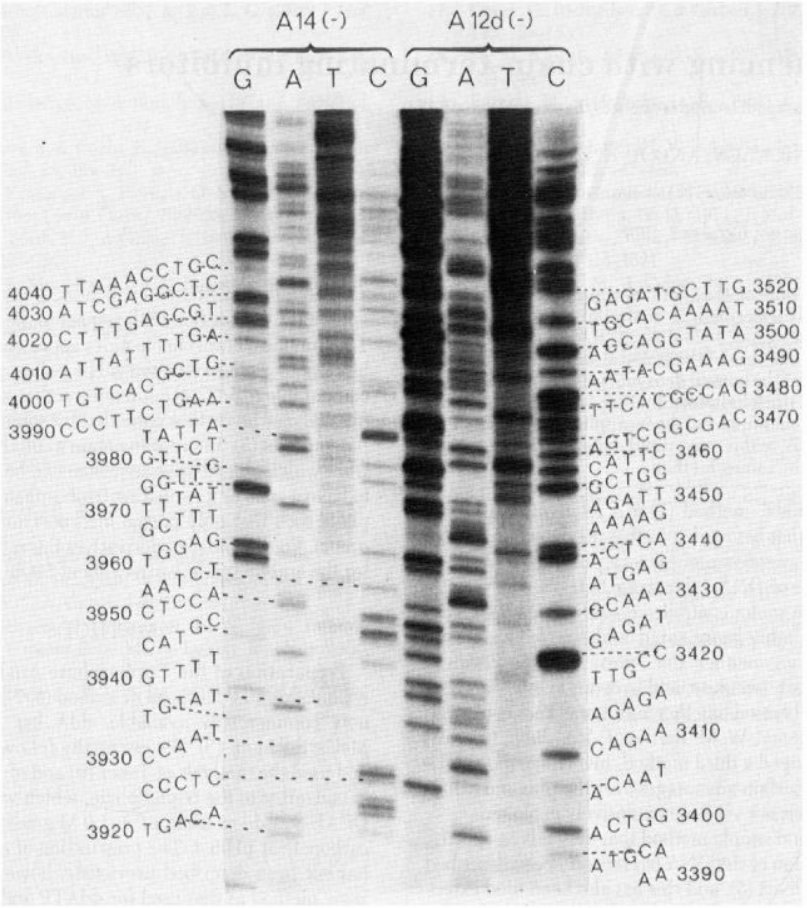


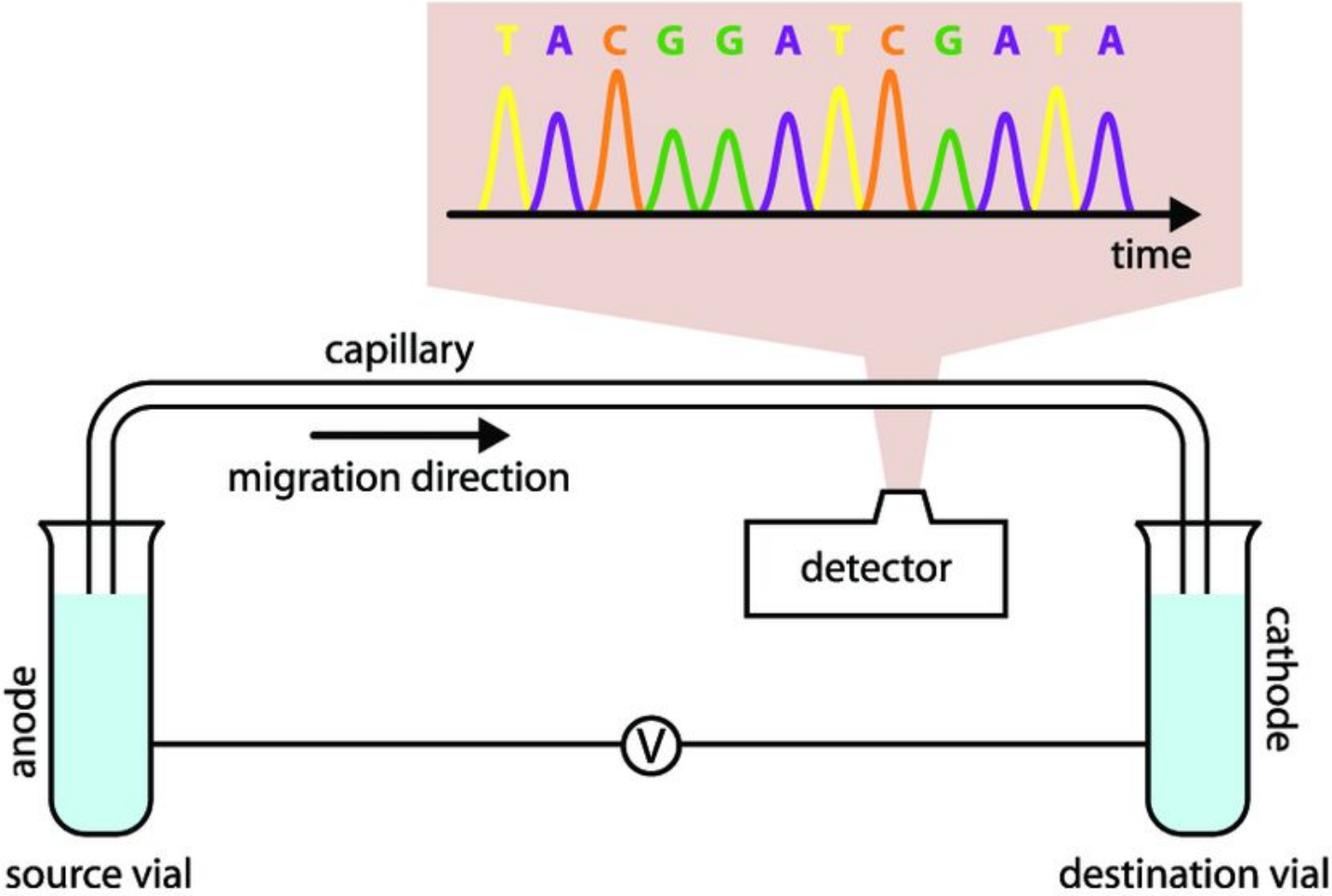
Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.

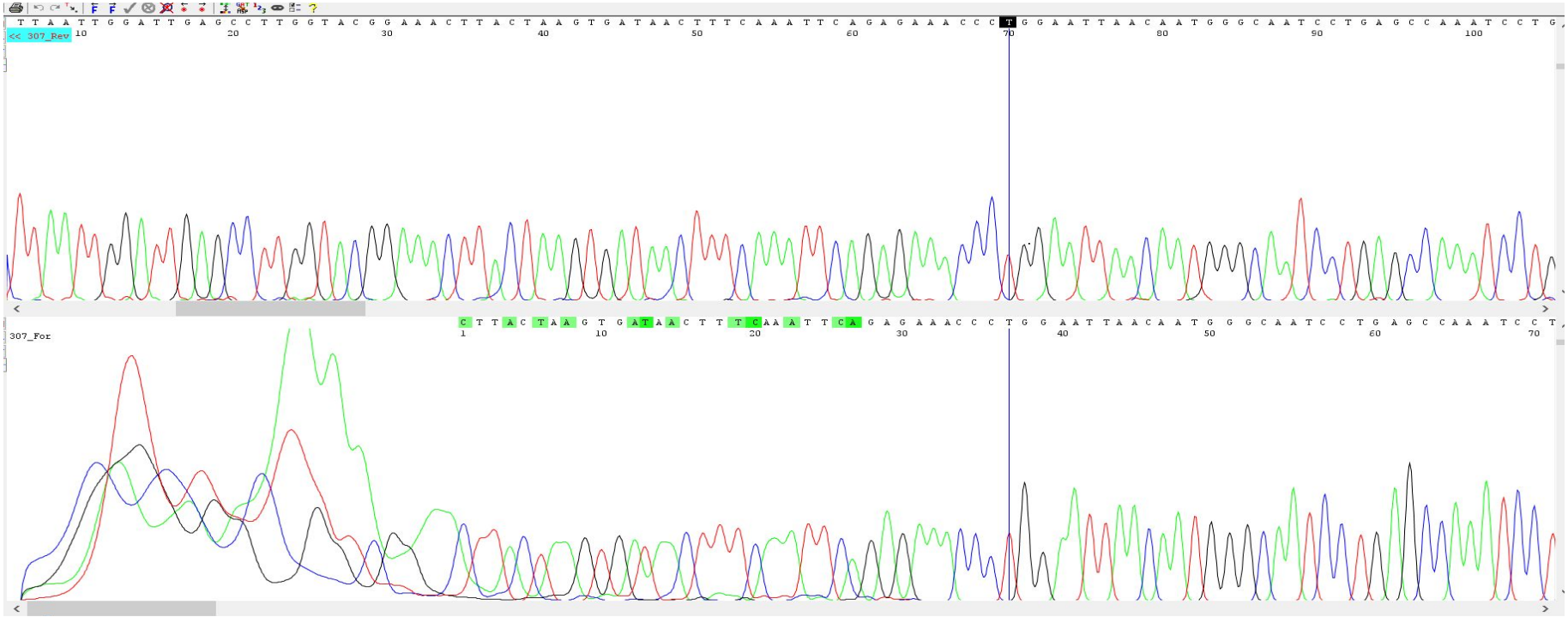


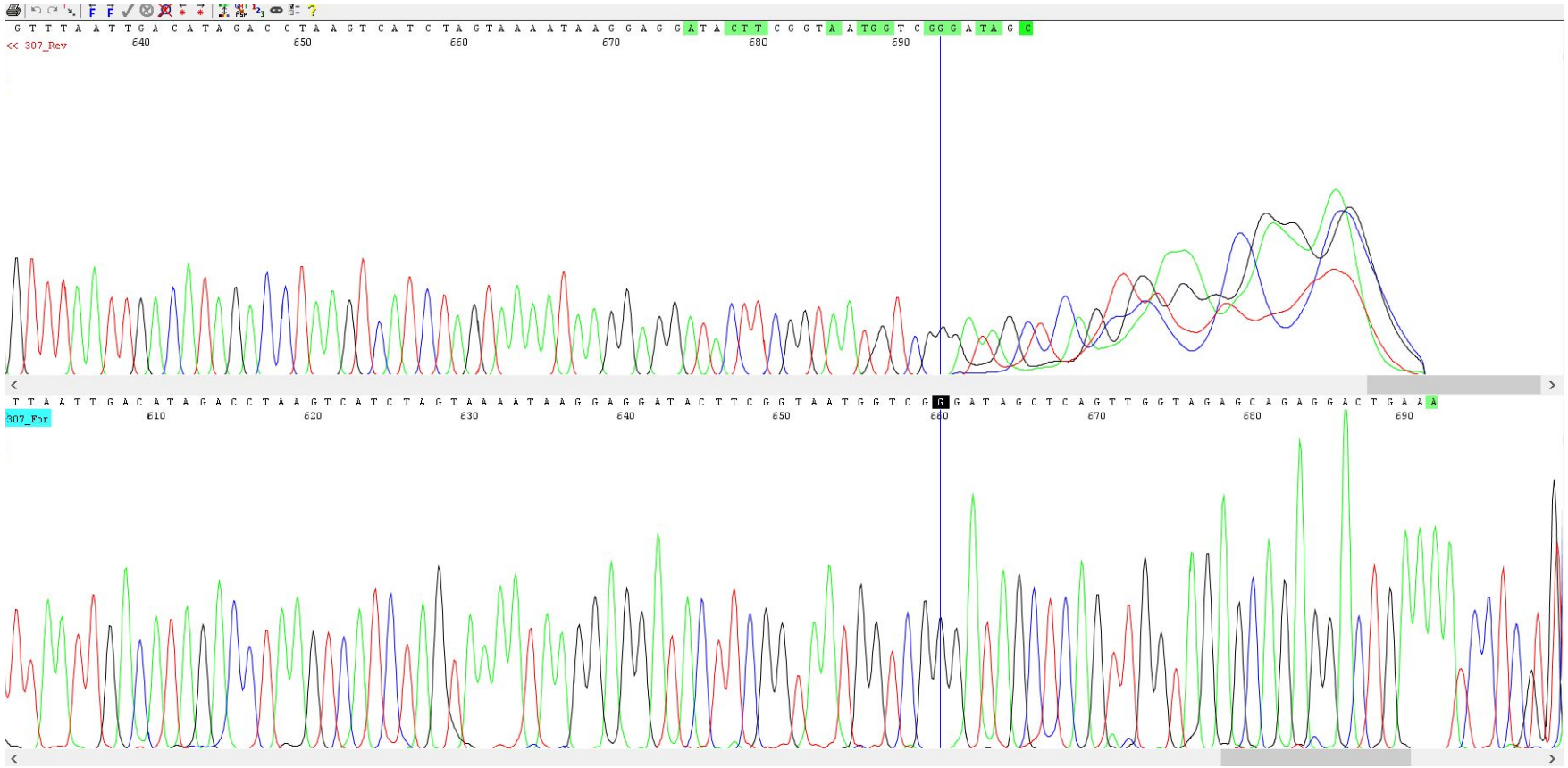
THE SANGER METHOD: *Single-stranded DNA is mixed with a primer and split into four aliquots, each containing DNA polymerase, four deoxyribonucleotide triphosphates and a replication terminator. Each reaction proceeds until a replication-terminating nucleotide is added. The mixtures are loaded into separate lanes of a gel and electrophoresis is used to separate the DNA fragments. The sequence of the original strand is inferred from the results. (See p. 40 for an illustration of a high-speed DNA sequencer.)*

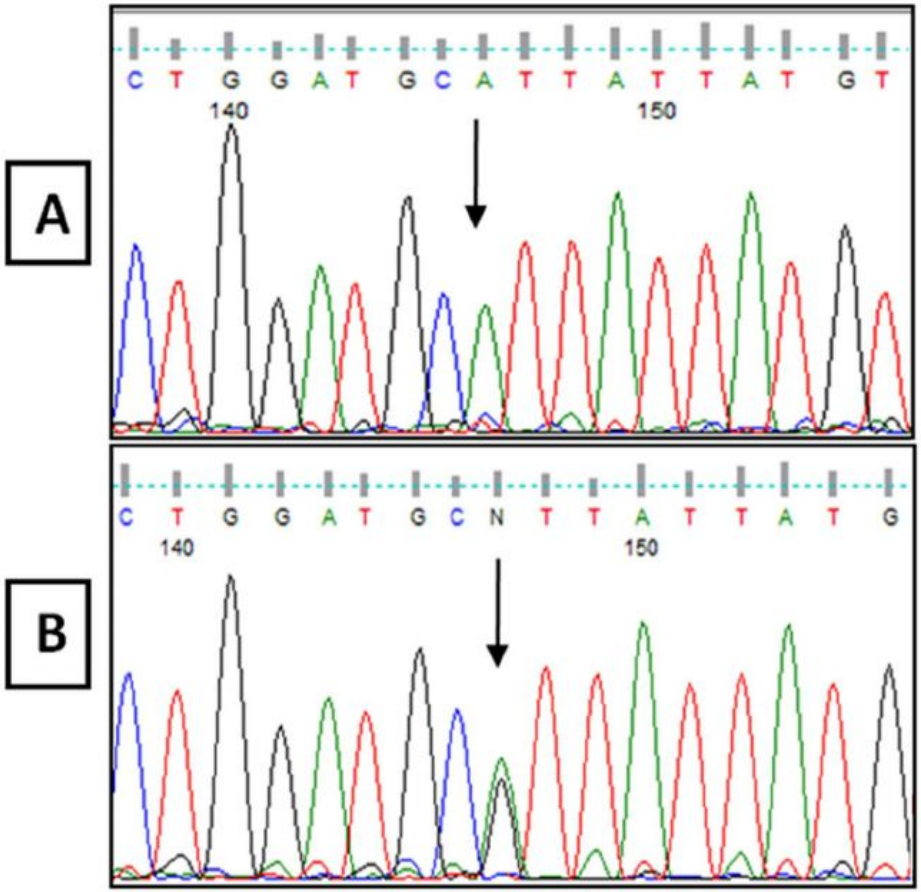

Denatured template
Template/product
ddG ddA ddT ddC
All tubes have dNTPs and polymerase but a different ddNTP
Labeled strands
Denaturing gel
G A T C

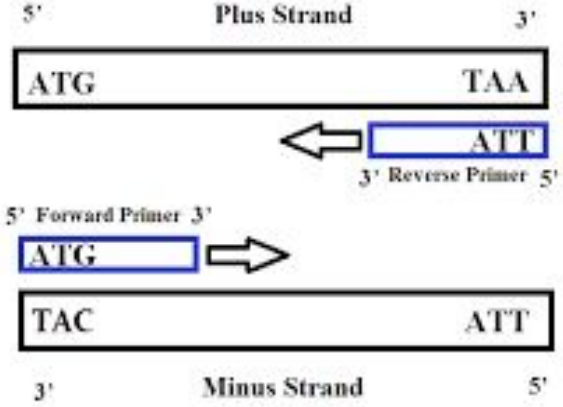

A14 (-)
A12d (-)
G A T C G A T C

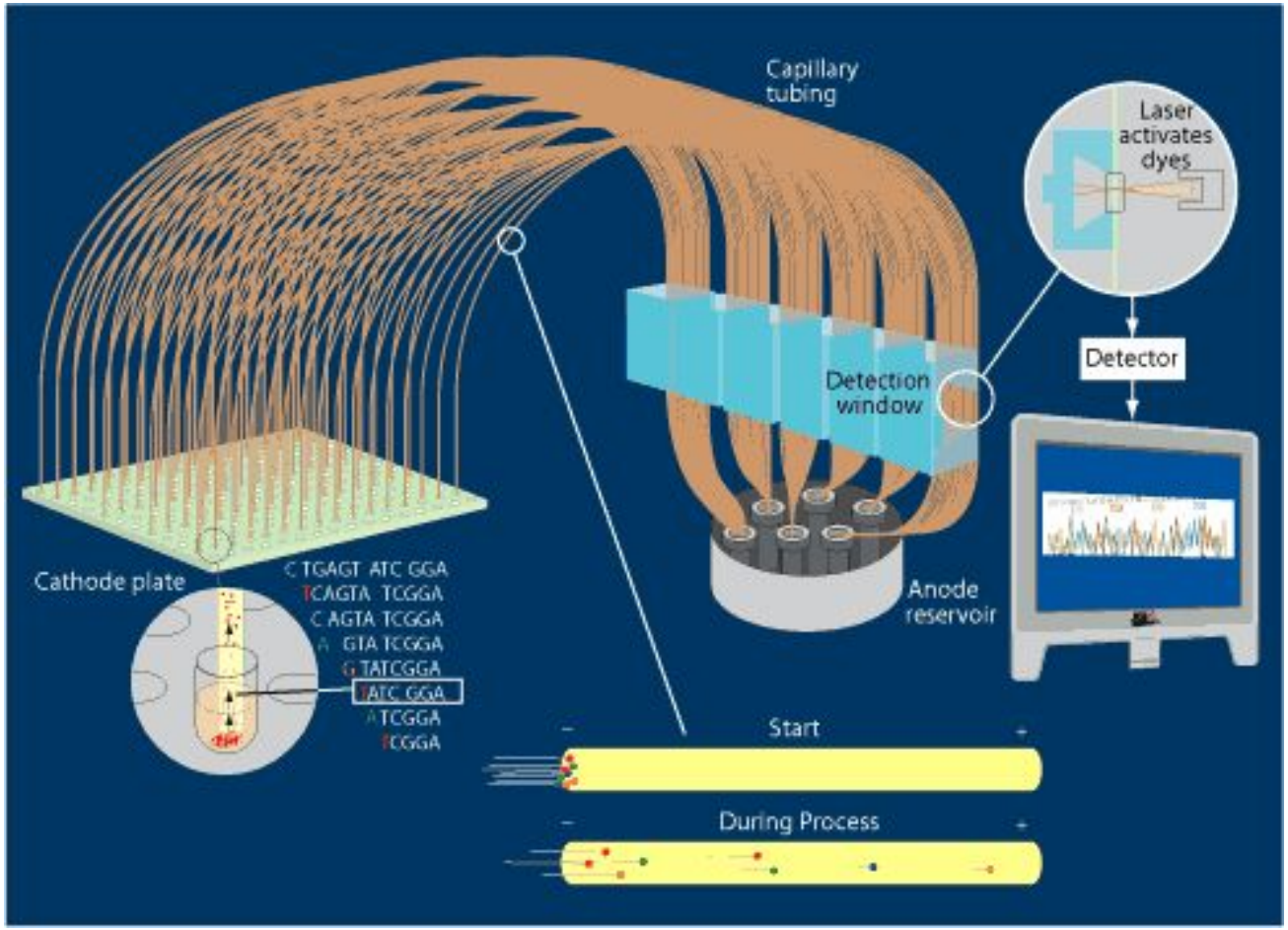

T A C G G A T C G A T A
time
capillary
migration direction
detector
anode
cathode
V
source vial
destination vial

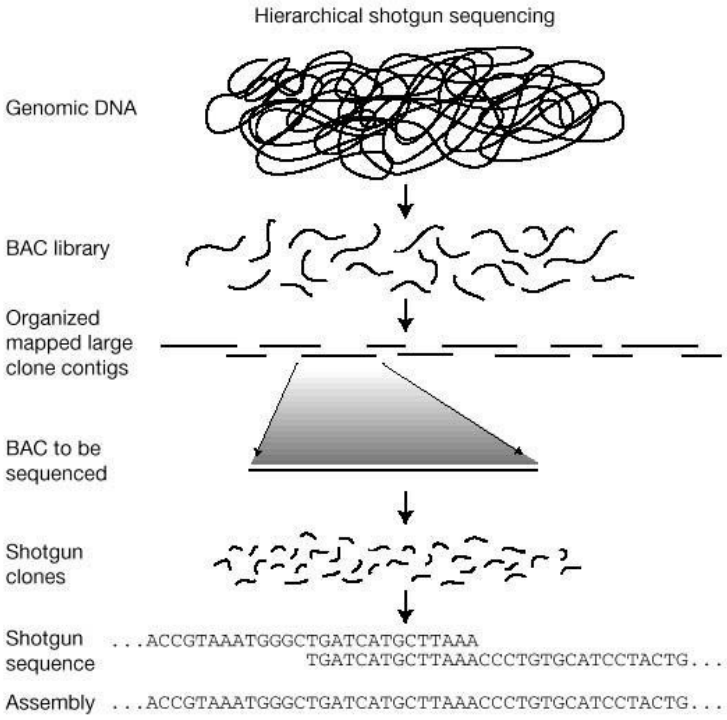

T T A A T T G G A T T G A G C C T T G G T A C G G A A A C T T A C T A A G T G A T A A C T T T C A A A T T C A G A G A A A C C C T G G A A T T A A C A A T G G G C A A T C C T G A G C C A A A T C C T G
<< 307_Rev
10 20 30 40 50 60 70 80 90 100
C T T A C T A A G T G A T A A C T T T C A A A T T C A G A G A A A C C C T G G A A T T A A C A A T G G G C A A T C C T G A G C C A A A T C C T
307_For
1 10 20 30 40 50 60 70

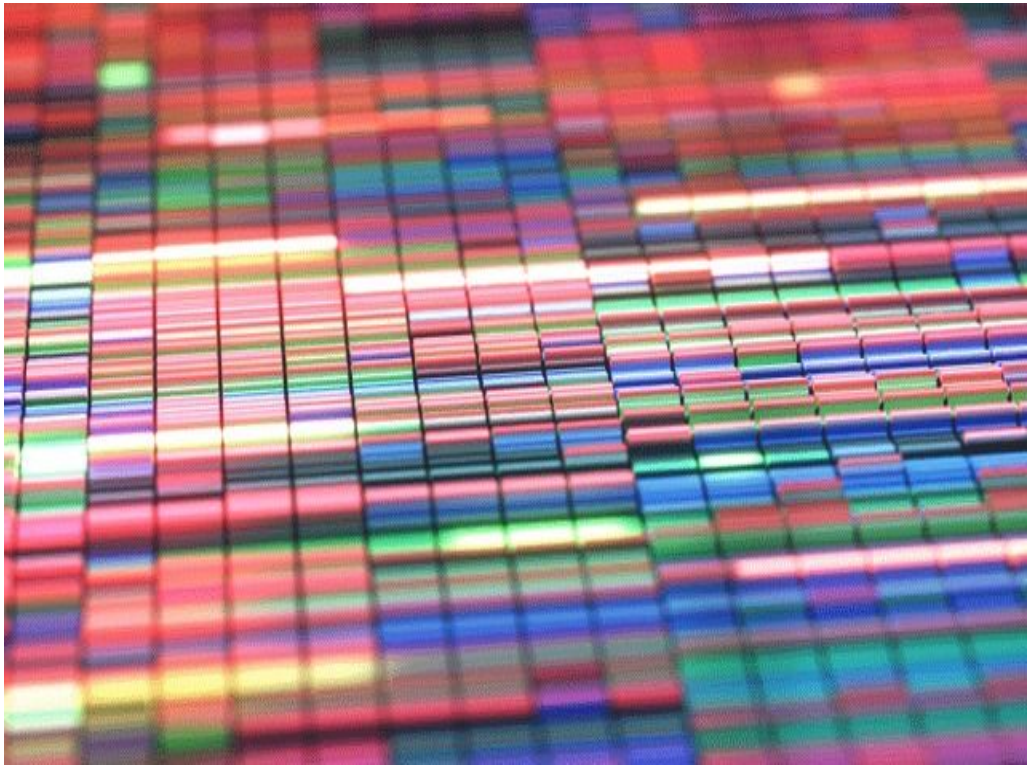

G T T T A A T T G A C A T A G A C C T A A G T C A T C T A G T A A A A T A A G G A G G A T A C T T C G G T A A T G G T C G G G A T A G C
<< 307_Rev
640
650
660
670
680
690
T T A A T T G A C A T A G A C C T A A G T C A T C T A G T A A A A T A A G G A G G A T A C T T C G G T A A T G G T C G G G A T A G C T C A G T T G G T A G A G C A G A G G A C T G A A A
307_For
610
620
630
640
650
660
670
680
690


A
C T G G A T G C A T T A T T A T G T
140
150
B
C T G G A T G C N T T A T T A T G
140
150

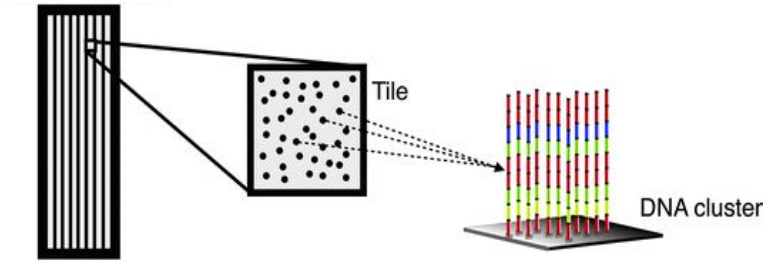

5'
Plus Strand
3'
ATG
TAA
ATT
3' Reverse Primer 5'
5' Forward Primer 3'
ATG
TAC
ATT
3'
Minus Strand
5'

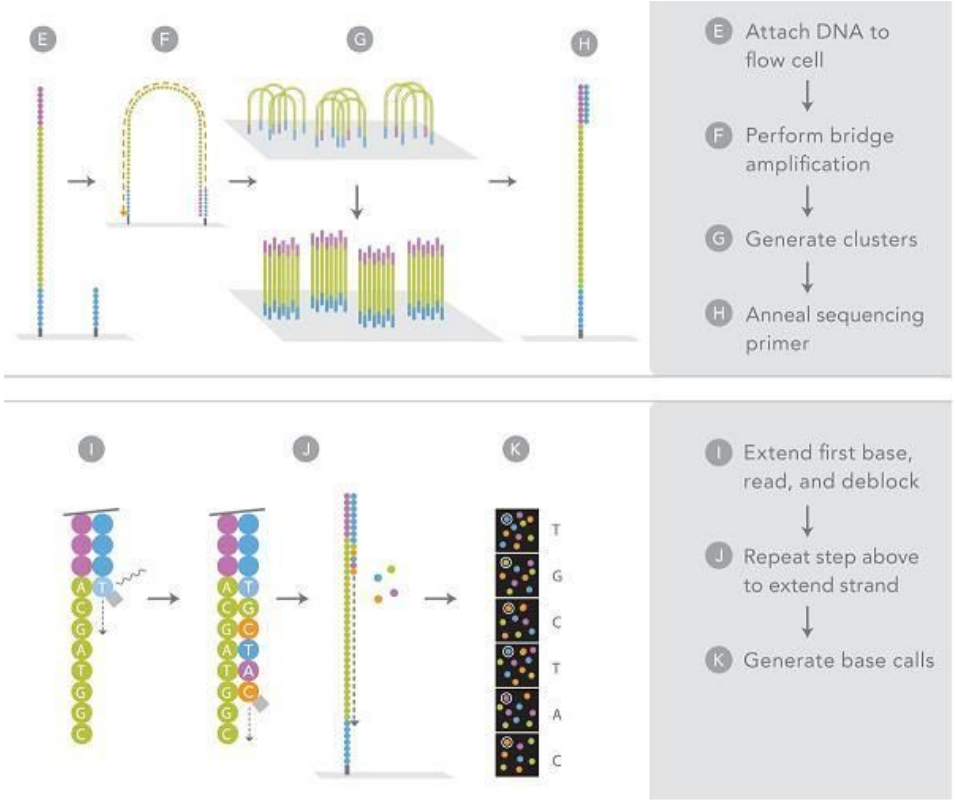

Capillary tubing
Laser activates dyes
Detector
Detection window
Anode reservoir
Cathode plate
C TGAGT ATC GGA
TCAGTA TCGGA
C AGTA TCGGA
A GTA TCGGA
G TATCGGA
TATC GGA
A TCGGA
TCGGA
Start
During Process

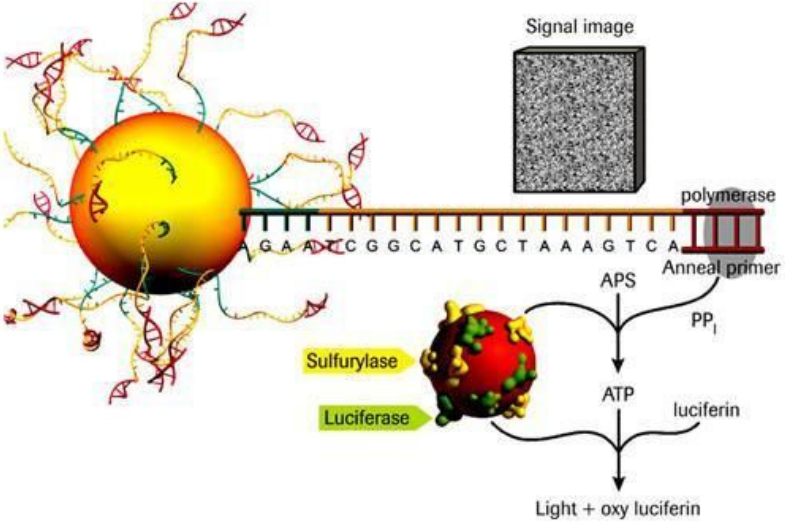

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...



2nd Generation Sequencing

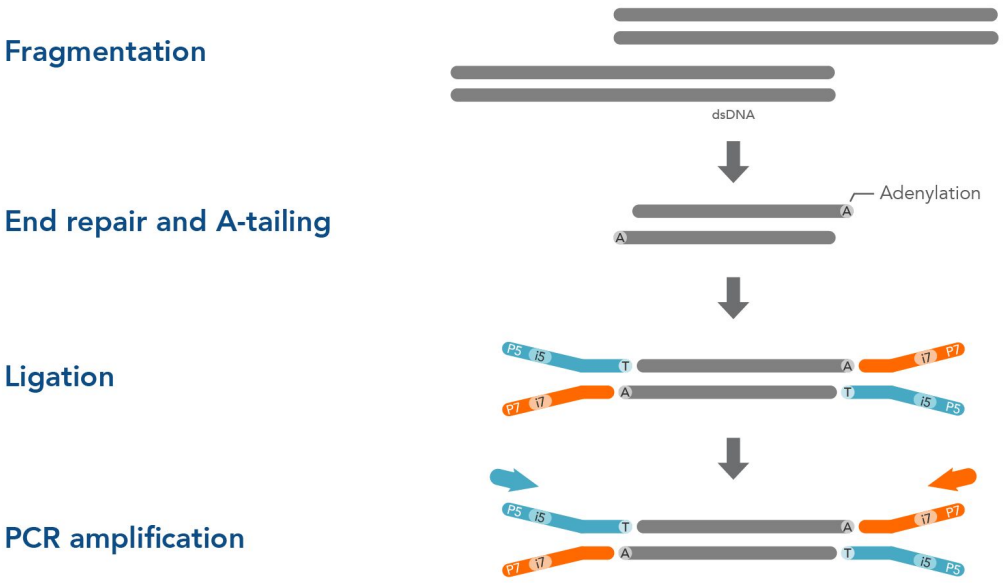


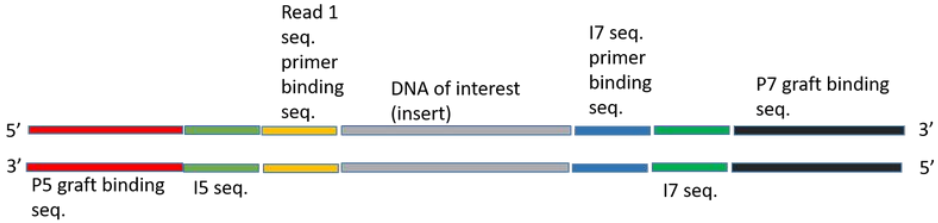

E
F
G
H
E Attach DNA to flow cell
F Perform bridge amplification
G Generate clusters
H Anneal sequencing primer
I
J
K
T
G
C
T
A
C
I Extend first base, read, and deblock
J Repeat step above to extend strand
K Generate base calls


Signal image
polymerase
G A A T C G G C A T G C T A A A G T C A
Anneal primer
APS
PPi
Sulfurylase
Luciferase
ATP
luciferin
Light + oxy luciferin

Library preparation and NGS sequencing steps:

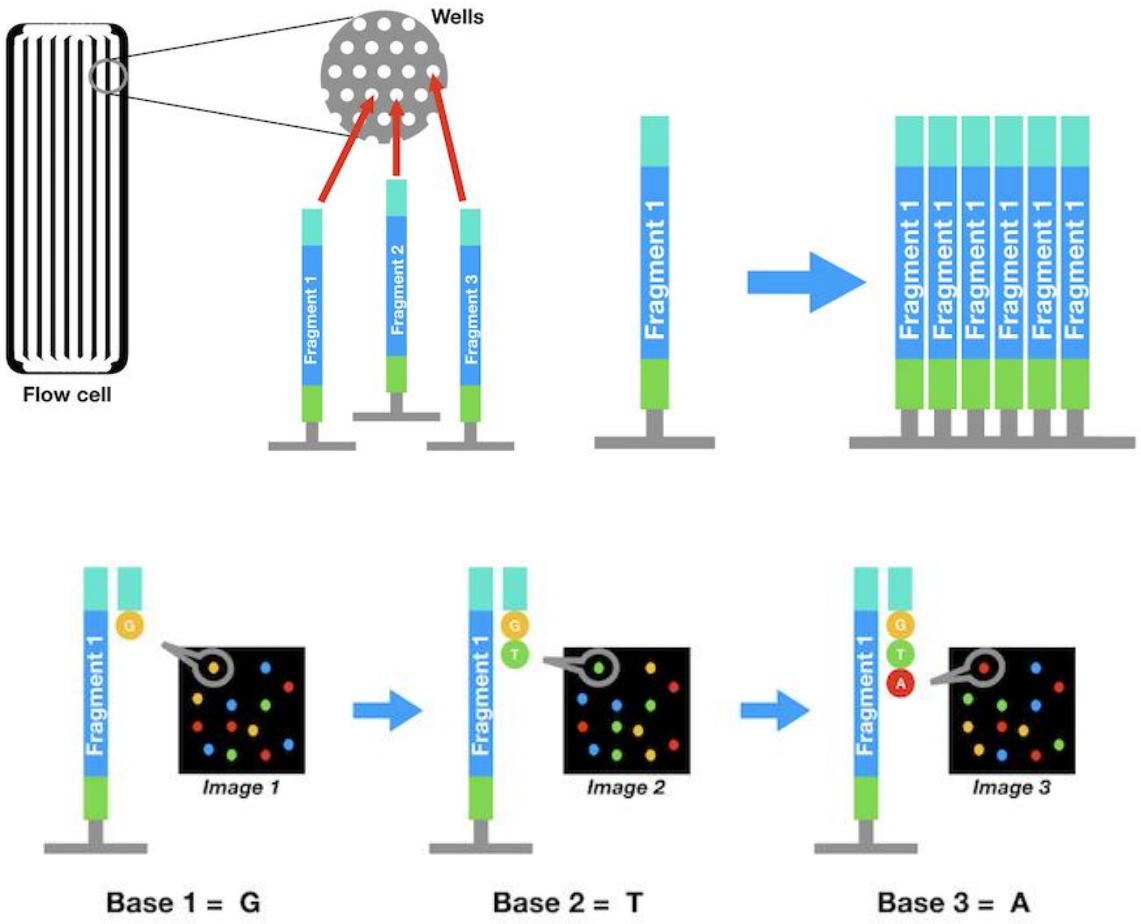
1. Shatter genomic DNA
2. Ligate adapters to both ends of the fragments.
3. PCR amplify the fragments with adapters



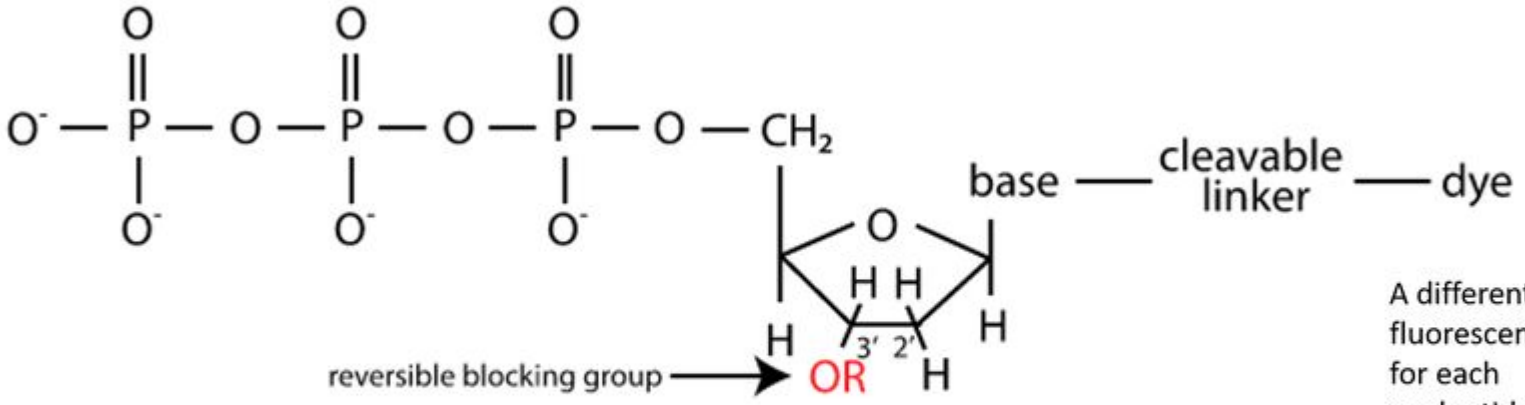


Library preparation and NGS sequencing steps:

4. Spread DNA molecules across flowcells (for Illumina technology). Goal is to get exactly one DNA molecule per flowcell lawn of primers. This depends purely on probability, based on the concentration of DNA.
5. Use bridge PCR to amplify the single molecule on each lawn so that you can get a strong enough signal to detect. Usually this requires several hundred or low thousands of molecules.
6. Sequence by synthesis of complementary strand via reversible terminator chemistry (Illumina)



3'-O-blocked reversible terminator



A different fluorescent dye for each nucleotide (4 in total)

This can be manipulated chemically to regenerate an –OH allowing synthesis to continue one step at a time.

Read 1 and Read 2 will "light up" on the **same spot** on the Illumina chip, which tells the system they are related as **paired ends**.

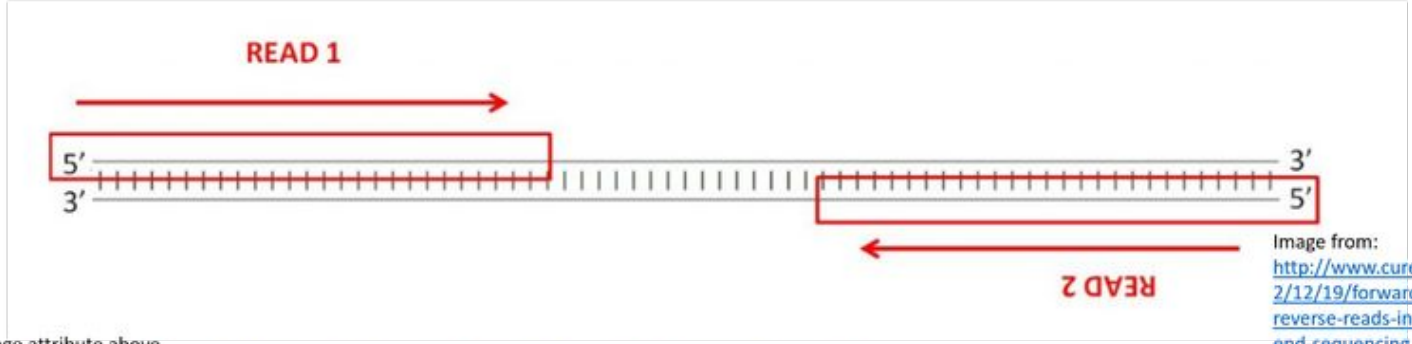
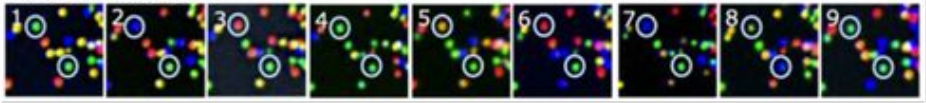


Image from: http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/

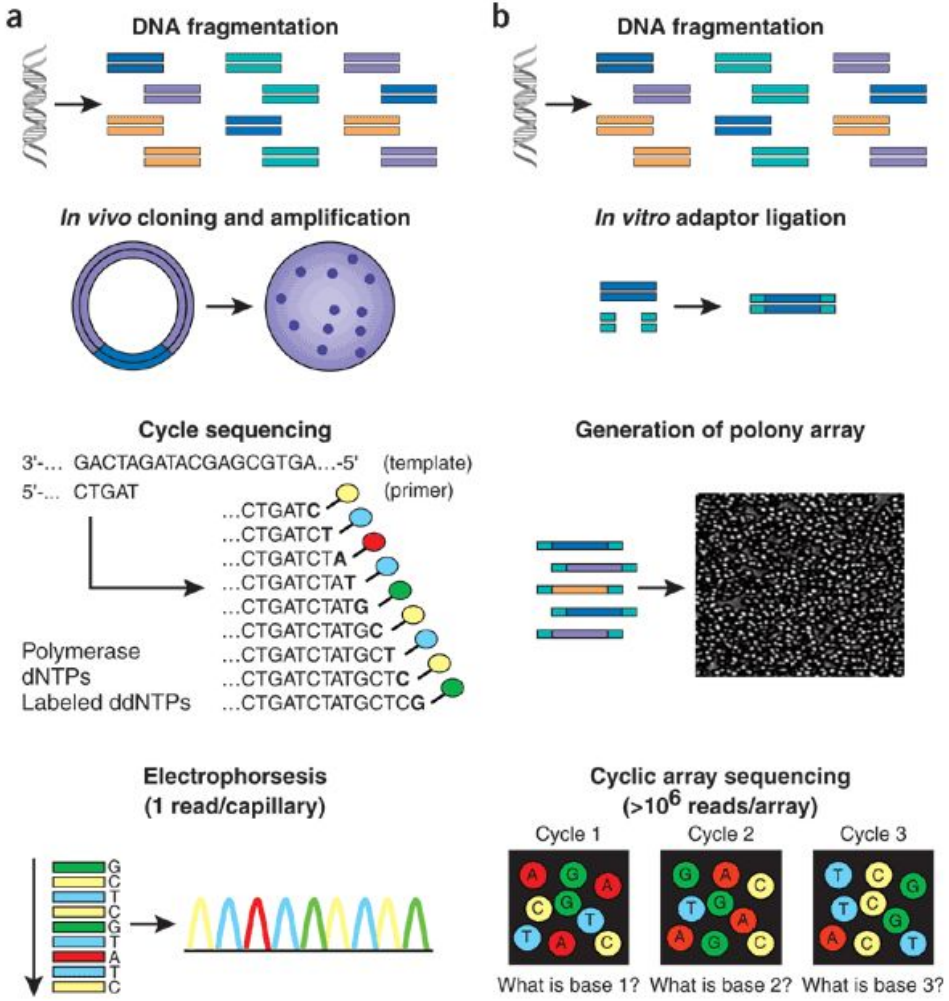
Image attribute above.



Each coloured dot is a spot on an Illumina chip that has a clonal cluster. The dots will light up to show what nucleotide is being added. This will occur in essentially four stages. The first is for Read 1 (flash flash flash and so on), then for the I7 index (flash flash flash etc.), then for the I5 index (flash flash flash etc), and finally for Read 2 (flash flash flash etc.).

https://kscbioinformatics.wordpress.com/2017/02/13/illumina-sequencing-for-dummies-samples-are-sequenced/

https://www.youtube.com/watch?v=HMyCqWhwB8E


a
DNA fragmentation
In vivo cloning and amplification
Cycle sequencing
3'-... GACTAGATACGAGCGTGA...-5' (template)
5'-... CTGAT (primer)
...CTGATC
...CTGATCT
...CTGATCTA
...CTGATCTAT
...CTGATCTATG
...CTGATCTATGC
...CTGATCTATGCT
...CTGATCTATGCTC
...CTGATCTATGCTCG
Polymerase
dNTPs
Labeled ddNTPs
Electrophorsesis
(1 read/capillary)
G
C
T
C
G
T
A
T
C
b
DNA fragmentation
In vitro adaptor ligation
Generation of polony array
Cyclic array sequencing
(>10^6 reads/array)
Cycle 1
Cycle 2
Cycle 3
What is base 1?
What is base 2?
What is base 3?

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

Sequence ID

Sequence

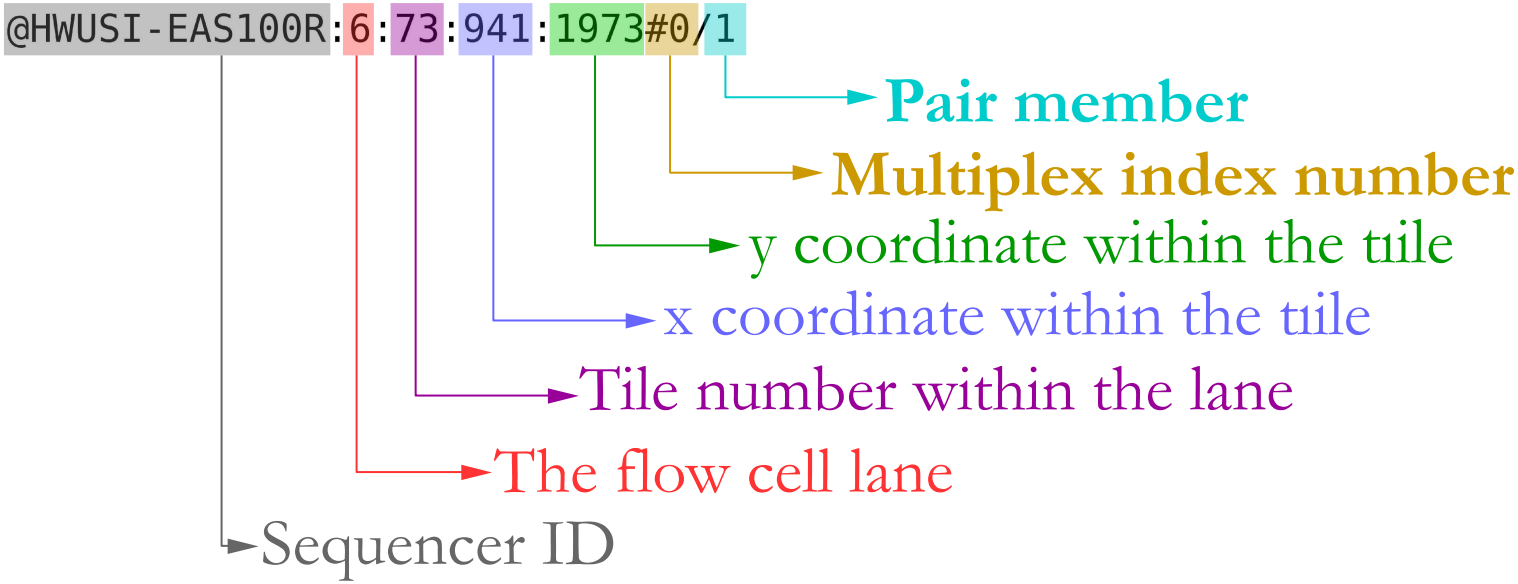
```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Quality scores

The

FASTQ FORMAT

for efficient storage & information



The FASTQ FORMAT

for efficient storage & information

Barcode #1

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

Barcode #2

RAD cut-site

The FASTQ FORMAT

Phred Quality Scores

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%

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

or

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

logarithmically related to the base-calling error probabilities

The

FASTQ FORMAT

Phred Quality Scores

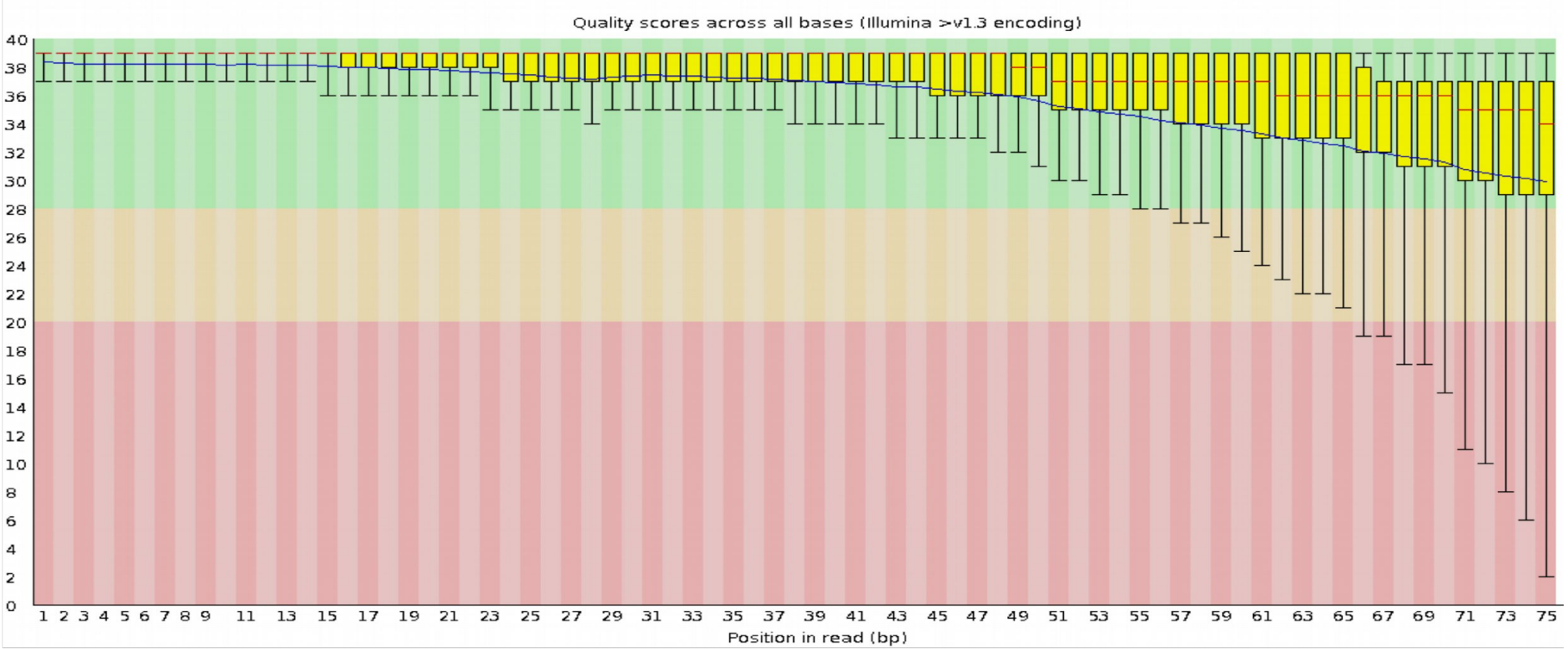
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.................................
 ........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX....
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII....
 ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ....
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijkl
 |                         |    |        |                              |
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)
```

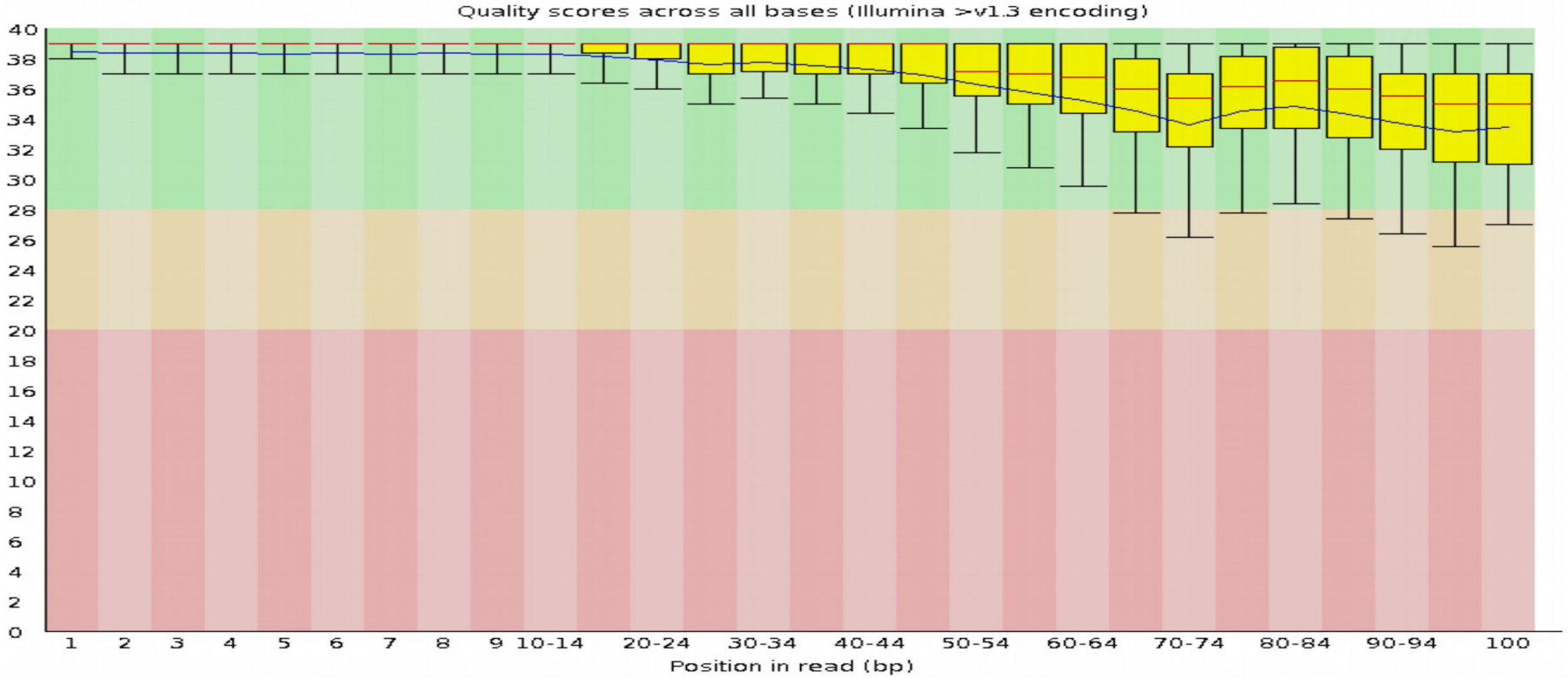
We can use quality scores to

Remove bad reads



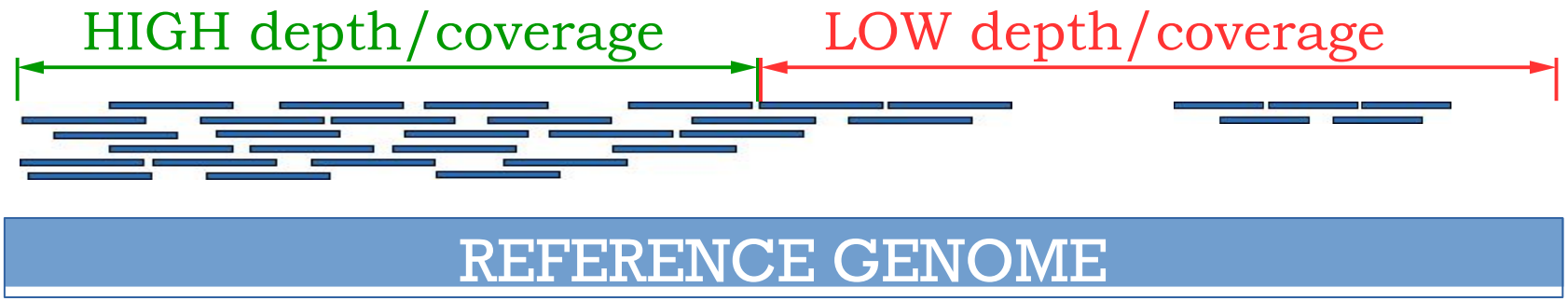
We can use quality scores to

Remove bad reads



Matching reads to a reference

MAPPING



BWA BOWTIE SOAP NOVOALIGN

Remove

CLONES

that can artificially bias depth

- PCR duplicates are identical copies of the same DNA molecule
- PCR duplicates arise because we *intentionally* create multiple copies of each original genomic DNA molecule so that we can have enough of them
- PCR duplicates occur when two copies of the same original molecule get onto different primer lawns in a flowcell.

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 COLUMNS

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 COLUMNS

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

```
Ref:  GCATTCAGATGCAGTACGC
Read:   ccTCAG--GCAGTAgtg

CIGAR   2S4M2D6M3S
POS     5
```

```
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>...

sample_01.bam (611MB)

sample_01.sorted.bam

sample_01.sorted.bam.bai

downstream analysis

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/