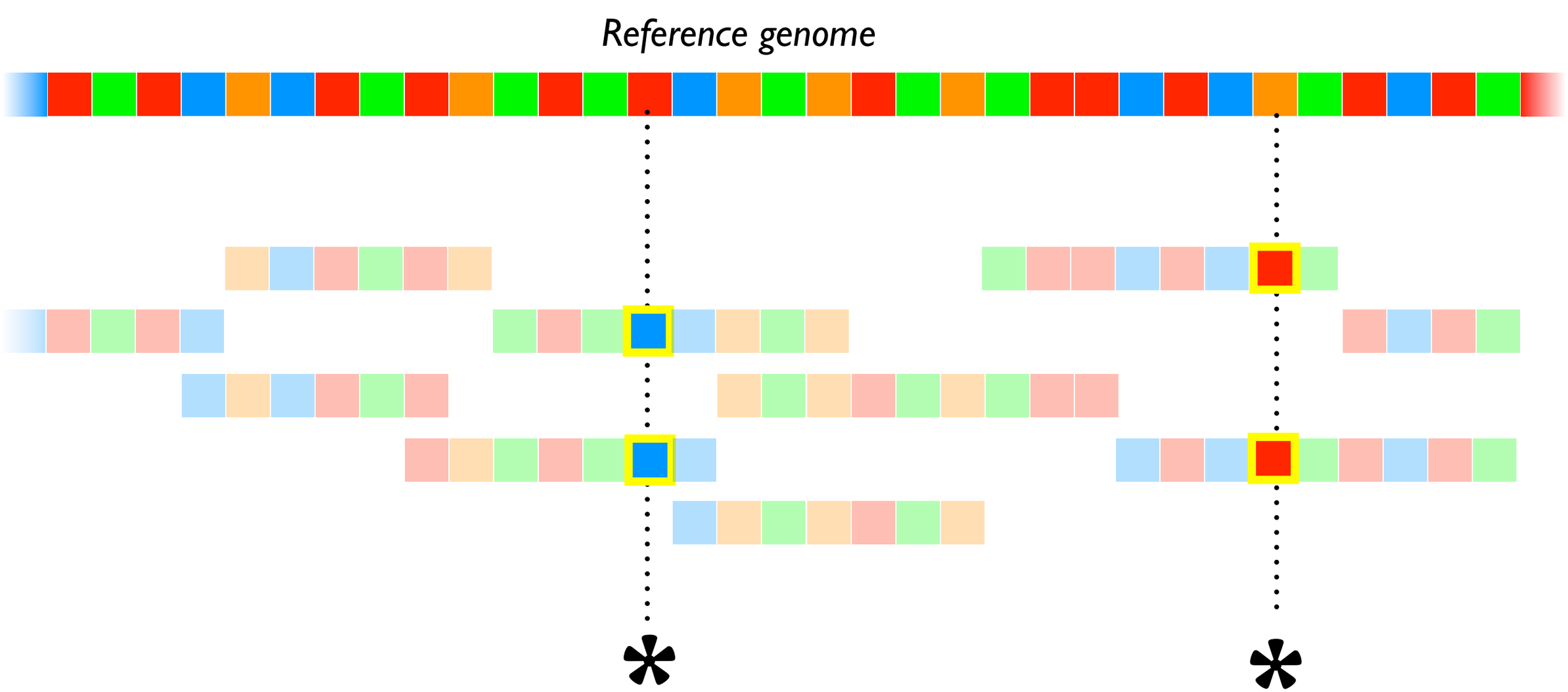
Genome Sequencing & Analysis Core Resource

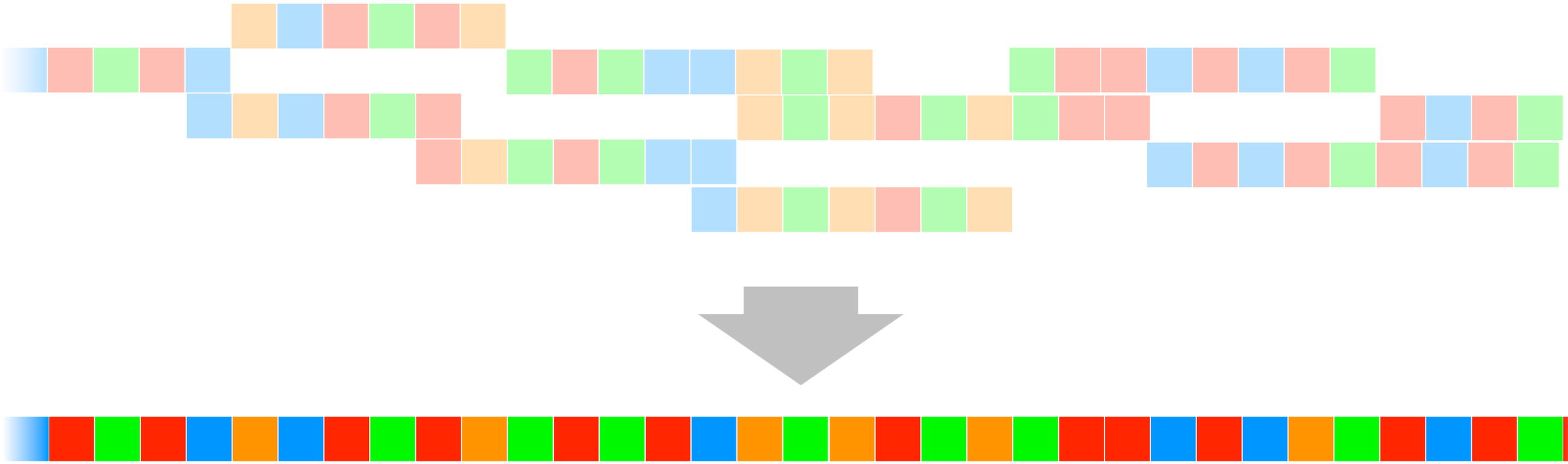
Olivier Fedrigo

DUKE INSTITUTE FOR GENOME SCIENCES & POLICY

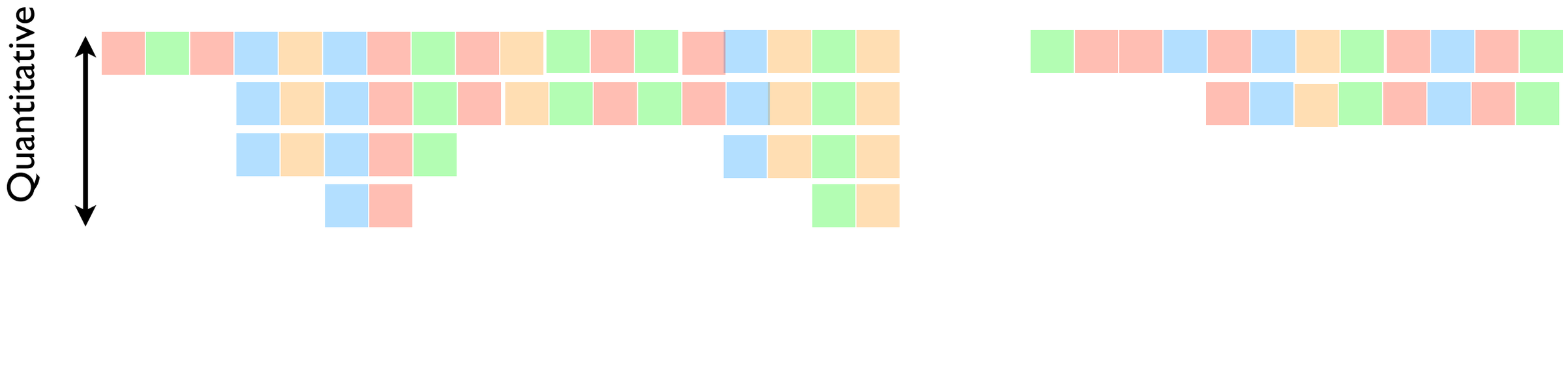


GENOME RESEQUENCING

Reference genome

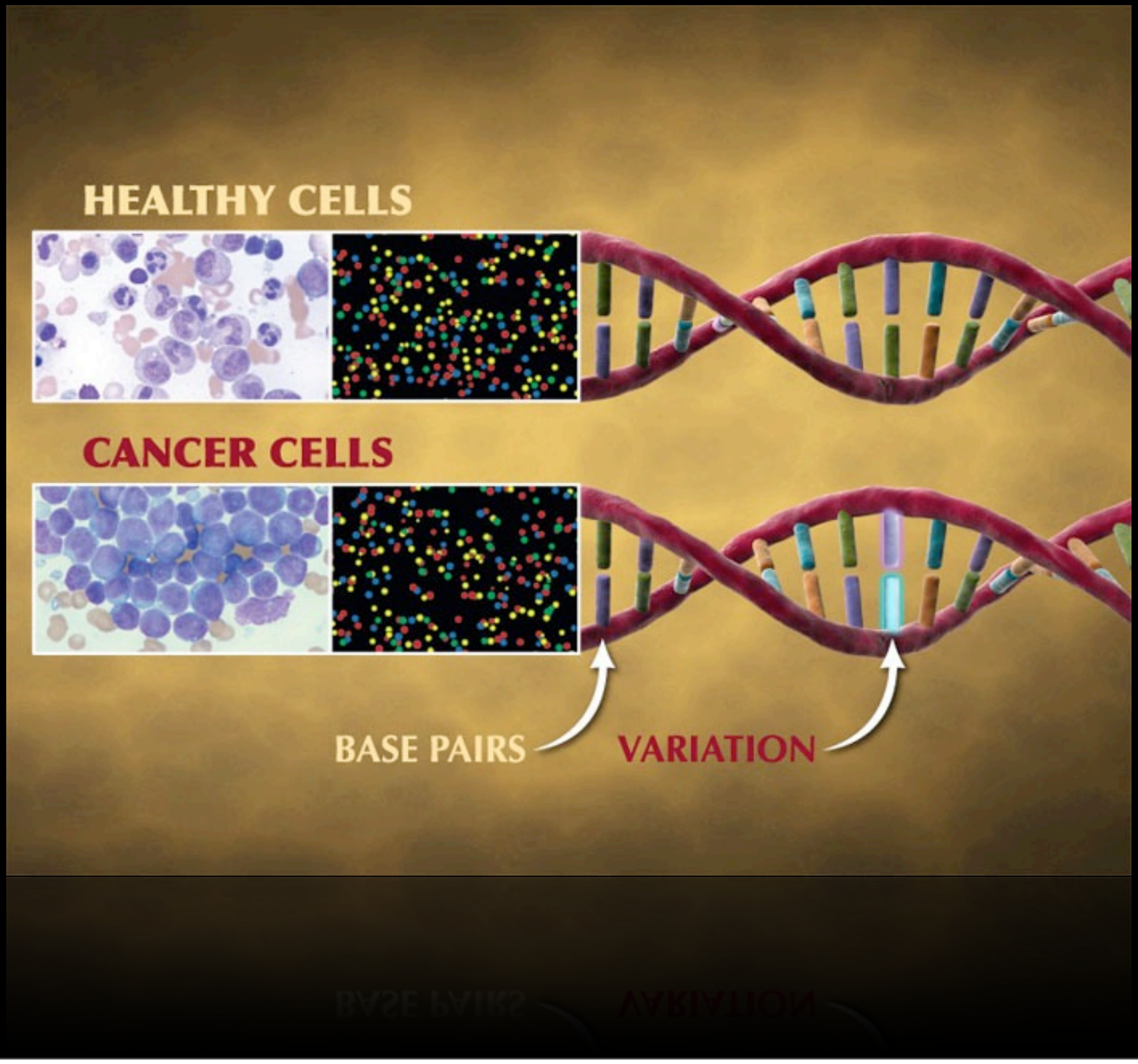


DE NOVO GENOME SEQUENCING

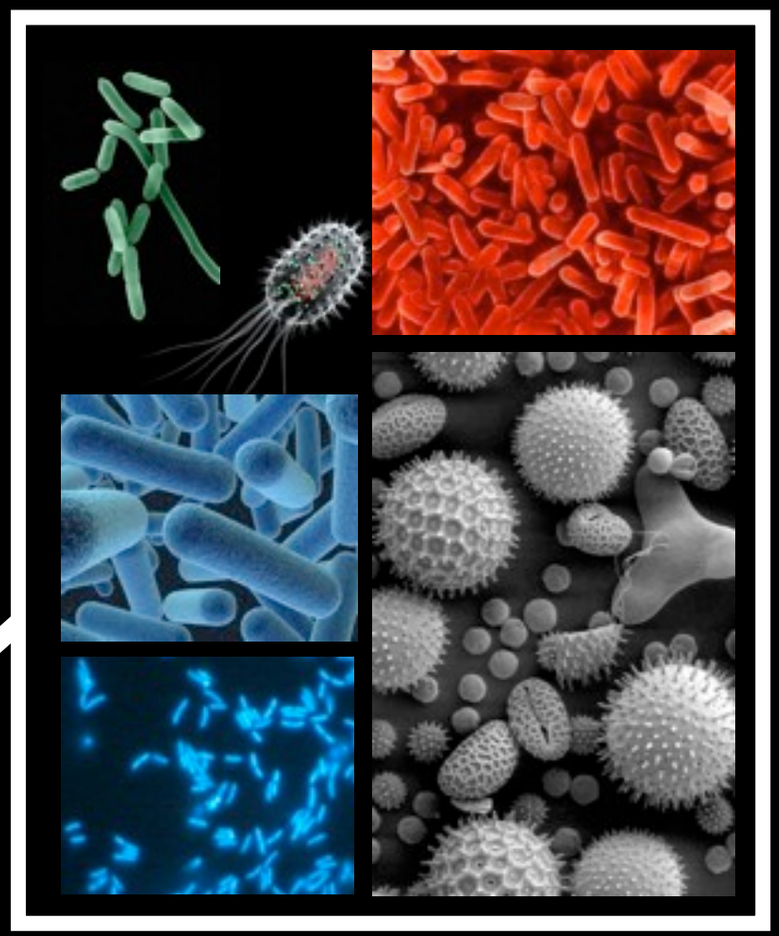


FUNCTIONAL GENOMICS

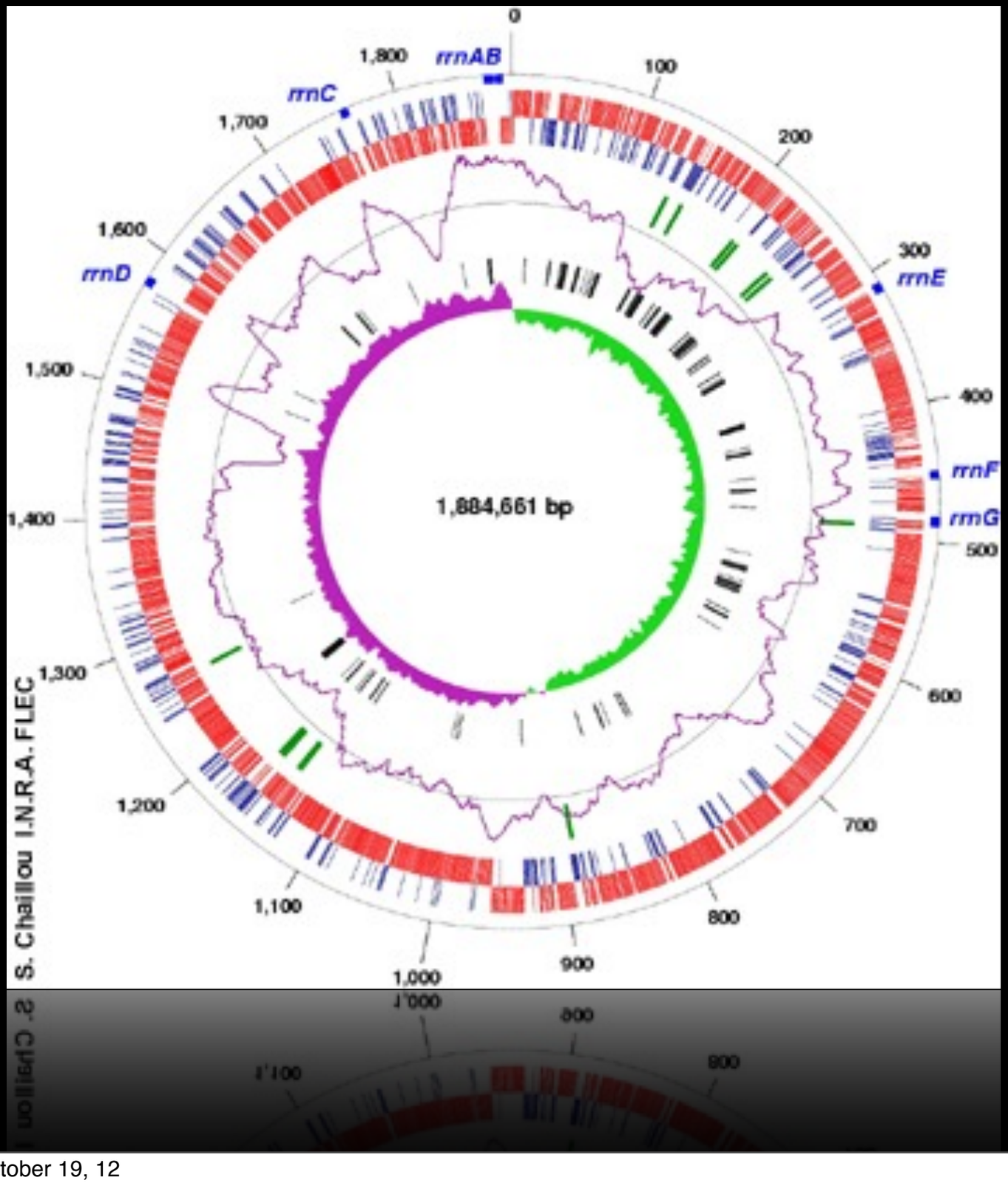
Medical Research



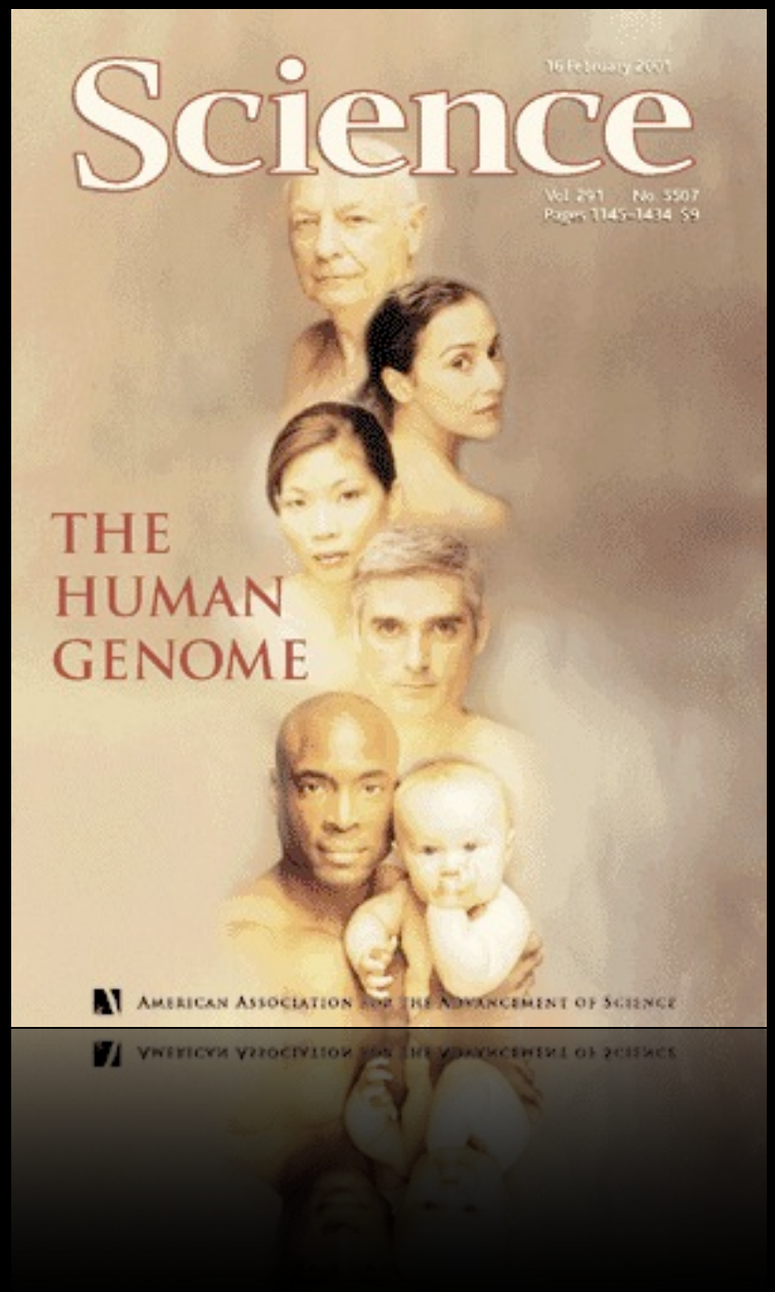
Metagenomics



Genome Sequencing



Lactobacillus sakei

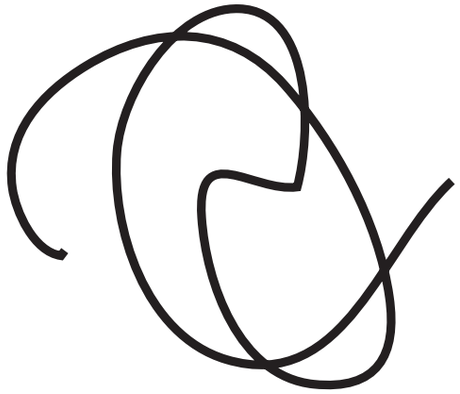


It took 13 years and 3billion$ to sequence the human genome (3 billion bases)

NEXT-GENERATION SEQUENCING

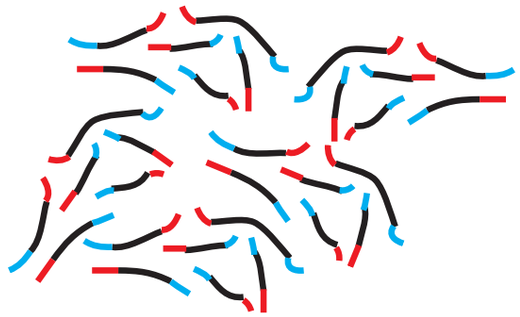
Second-Generation Sequencing

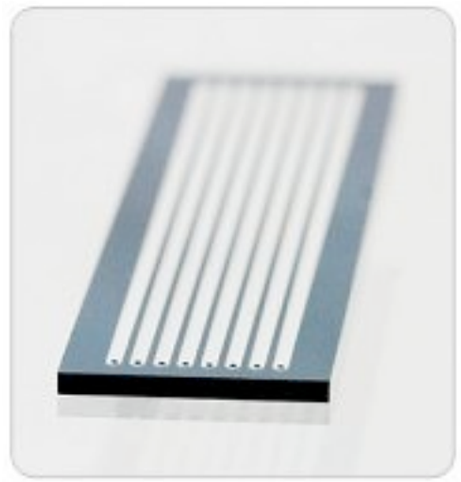
- Make library
- ~~Amplify signal~~ Third-Generation Sequencing
- Deposit sequences on a slide
- Imaging



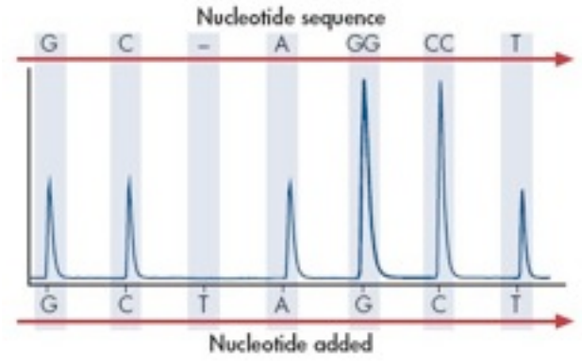
Shearing

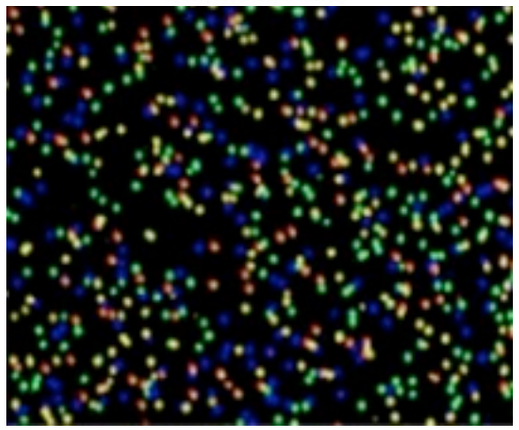
Adapters





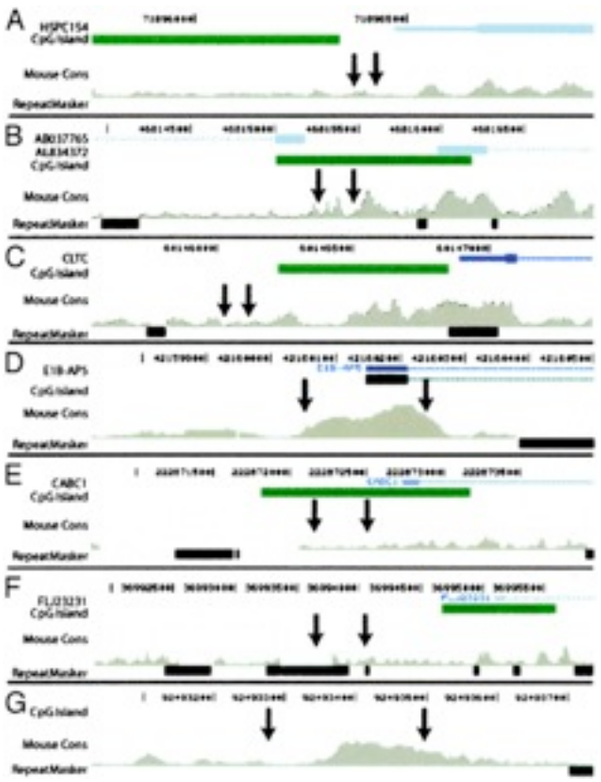
Amplification
+
Slide deposit



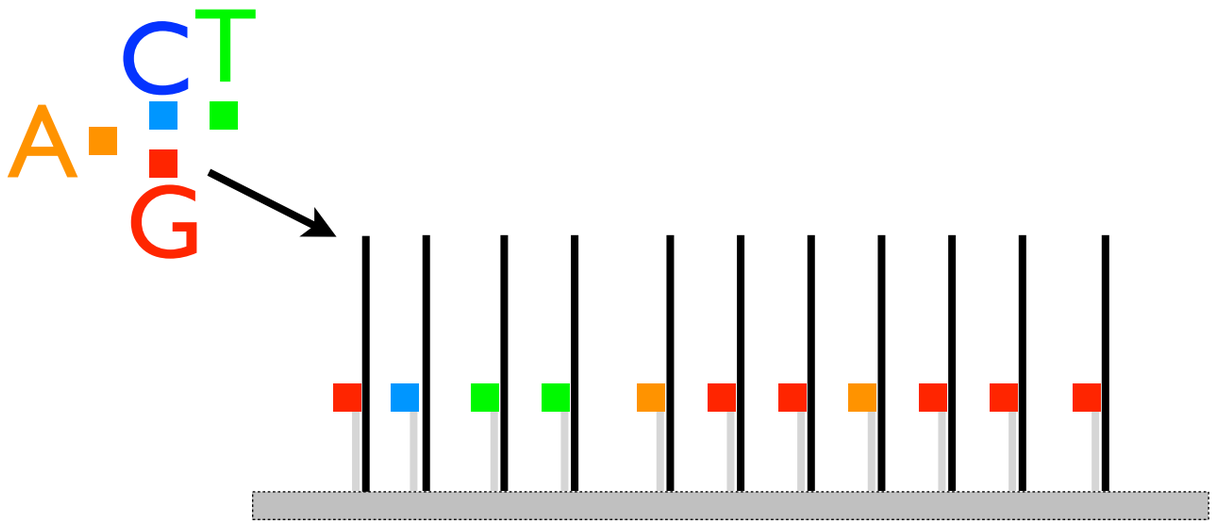


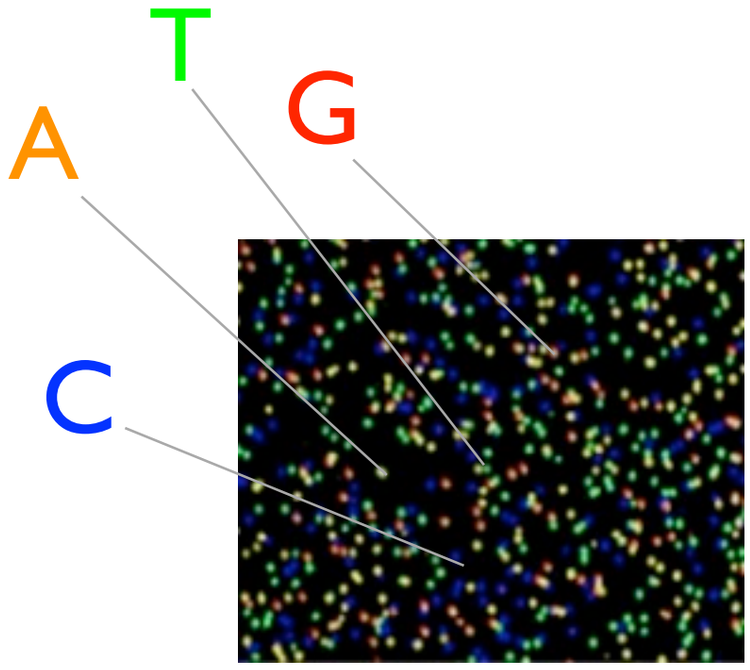
Sequencing

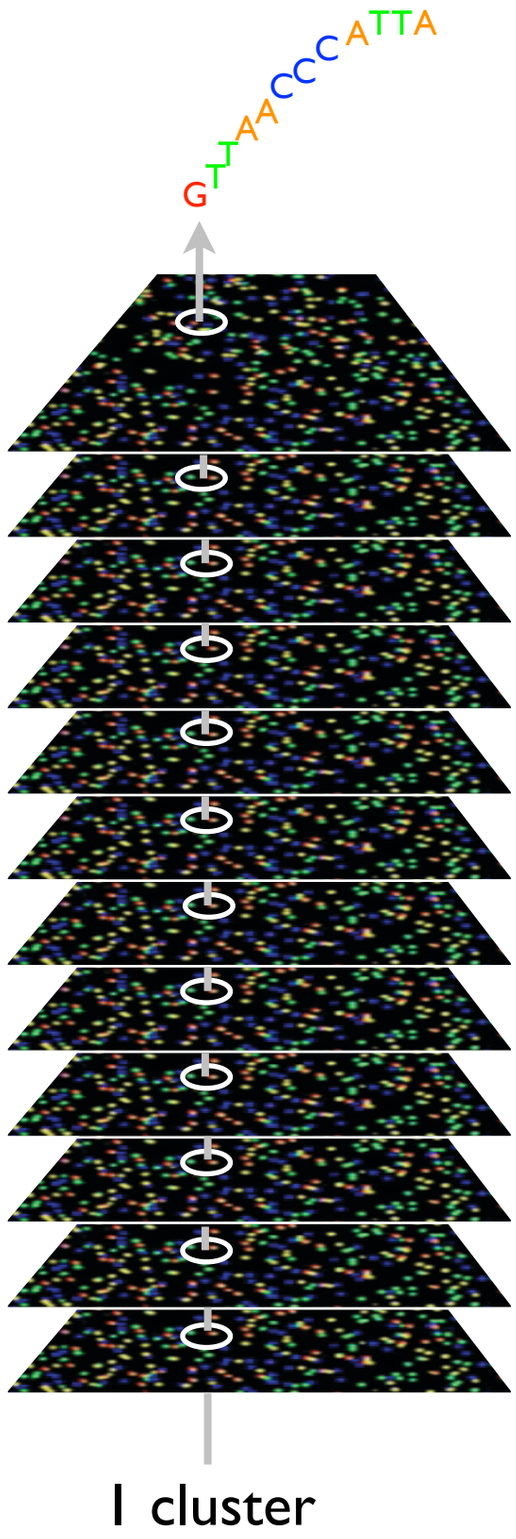
ACGTGTGT
ATTGTGTC
ACGTGTGG
TTGTGTGC
TGTGGTTT
GTGTGGGG
ACGTGTGT
ATTGTGTC
ACGTGTGG
TTGTGTGC
TGTGGTTT
GTGTGGGG



De novo assembly or
Mapping to reference


C
T
A
G


T
G
A
C


G
T
A
A
C
C
C
A
T
T
A
1 cluster

1977 — **Capillary-based Sanger sequencing:** *Applied Biosystems, etc.*

~1200 bp X 96/384 samples

2000 — **Pyrosequencing:** *Biotage*

up to 50 bp X 96/384 samples

2004 — **Massively parallel pyrosequencing:** *454-->Roche*

~800 bp X 1,200,000 reads per run

2005 — **Synthesis-based sequencing:** *Solexa-->Illumina*

up to 100 bp X 6 billion reads per run (2 flowcells)

2007 — **Ligation-based sequencing:** *Agencourt-->SOLiD (Applied Bios.)*

up to 75 bp X 1.4 billion reads per run

2011 — **Sequencing with pH:** *Ion Torrent*

up to 300bp X 5 million reads per run

2008 — **Single molecule sequencing:** *Helicos*

on the market; <50 bp X 100 million reads per run

2011 — **Single molecule sequencing:** *PacBio RS System*

~3kb, ~70,000 reads per smrtcell

"long reads" >250bp

Roche GS FLX Titanium (454)

PACBIO RS

Ion Torrent PGM

"short reads" ≤250bp

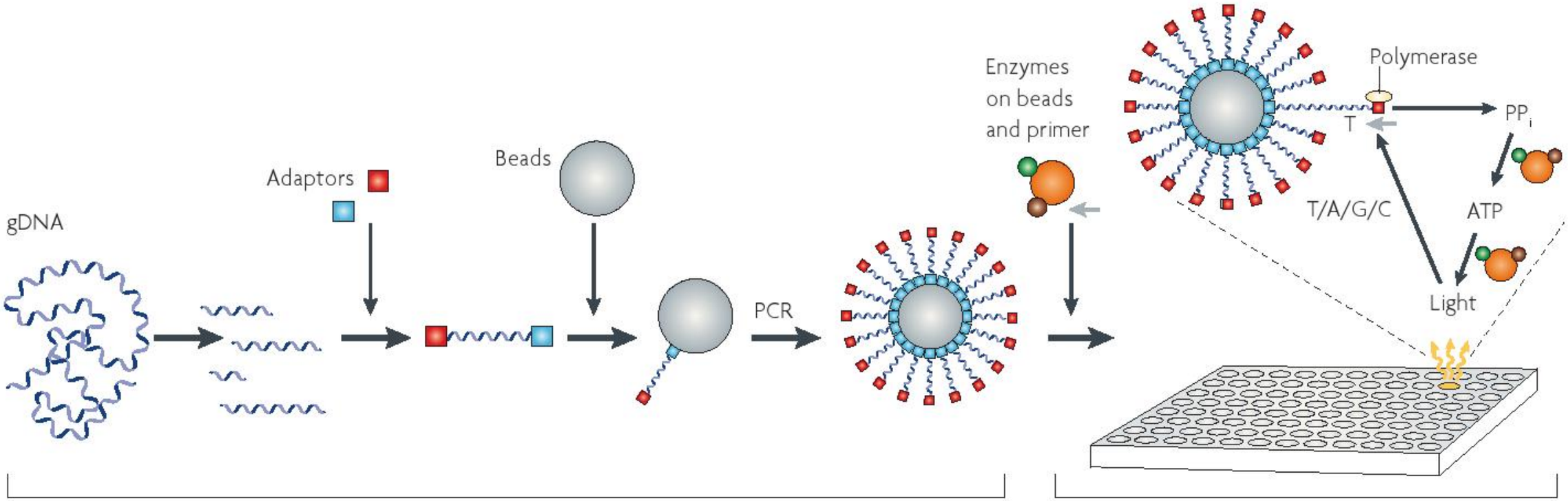
ABI SOLiD 5500xl

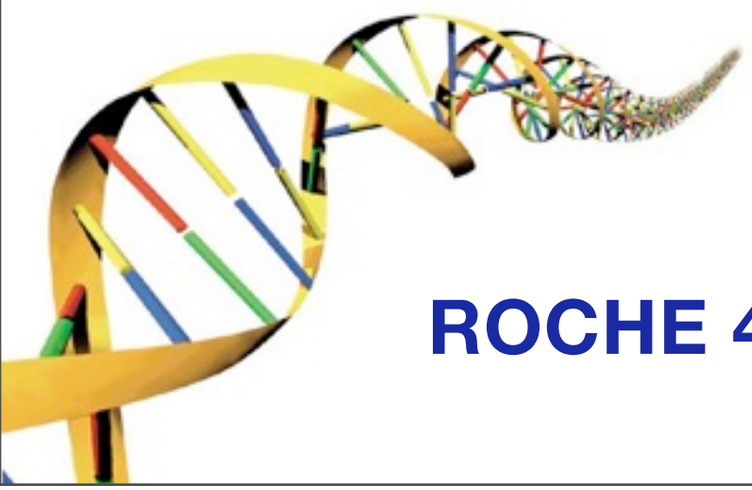
Illumina HiSeq 2000

Illumina MiSeq

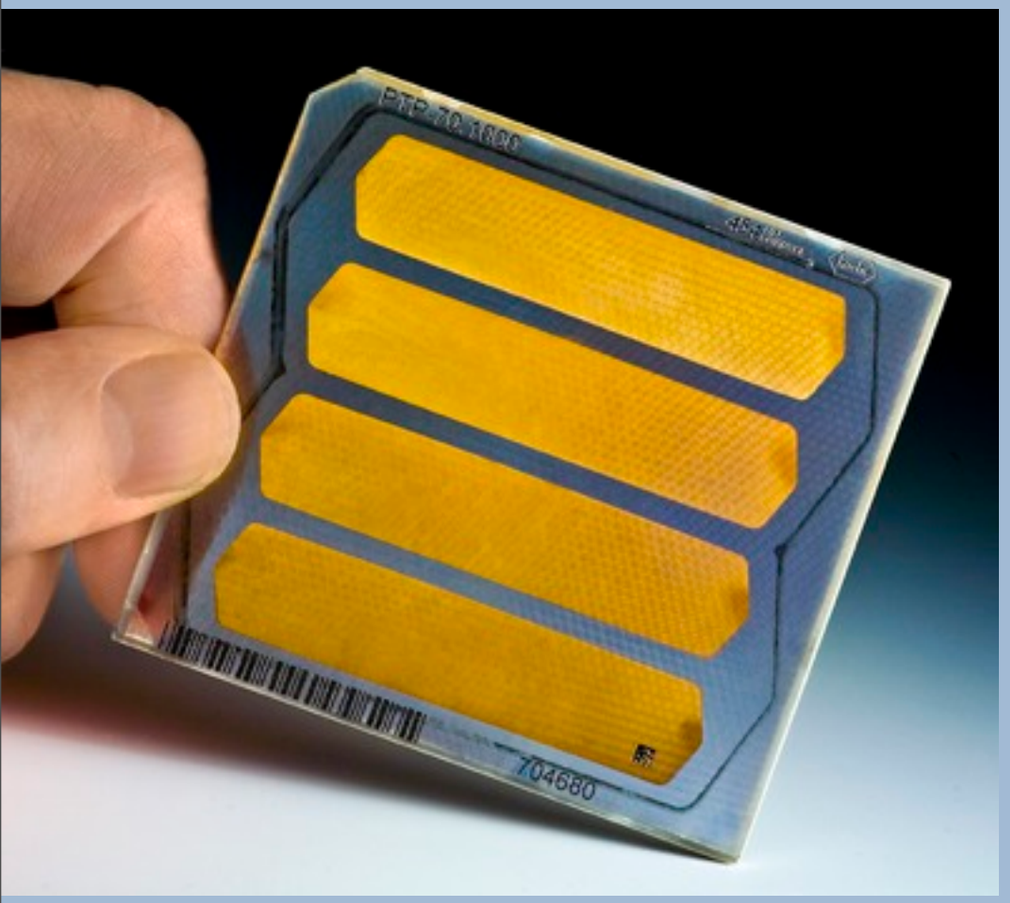
ROCHE 454


454 Life Sciences
454 LIFE SCIENCES
Genome Sequencer FLX
Roche

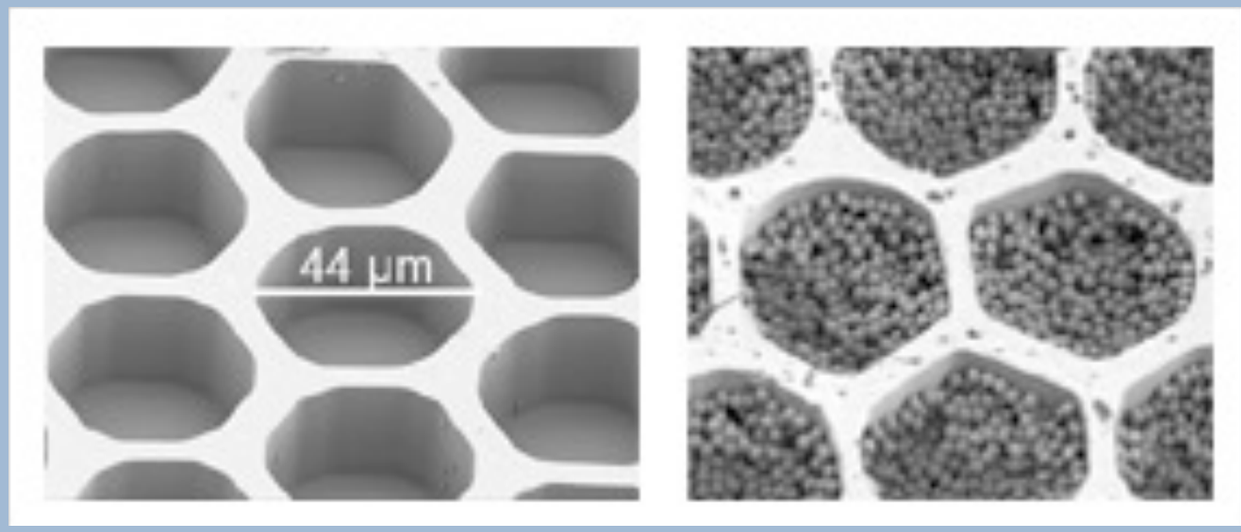

gDNA
Adaptors
Beads
PCR
Enzymes
on beads
and primer
Polymerase
T
PPi
T/A/G/C
ATP
Light
Sample preparation
Pyrosequencing

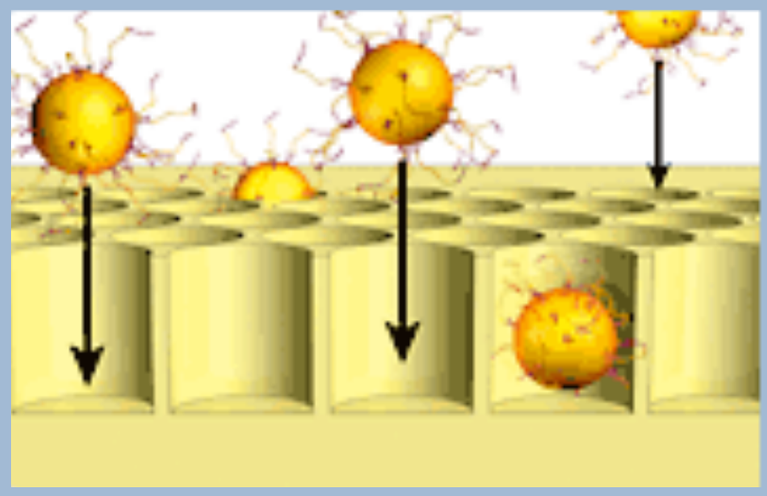


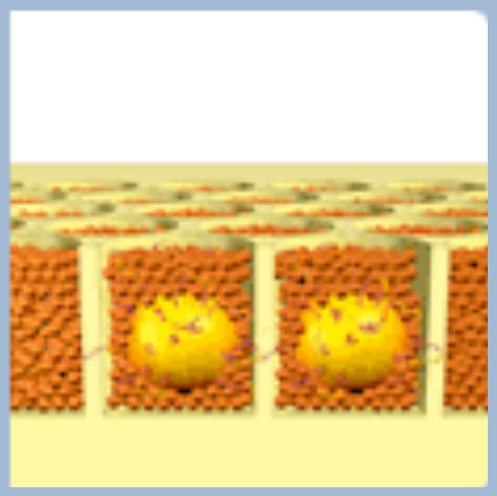
ROCHE 454

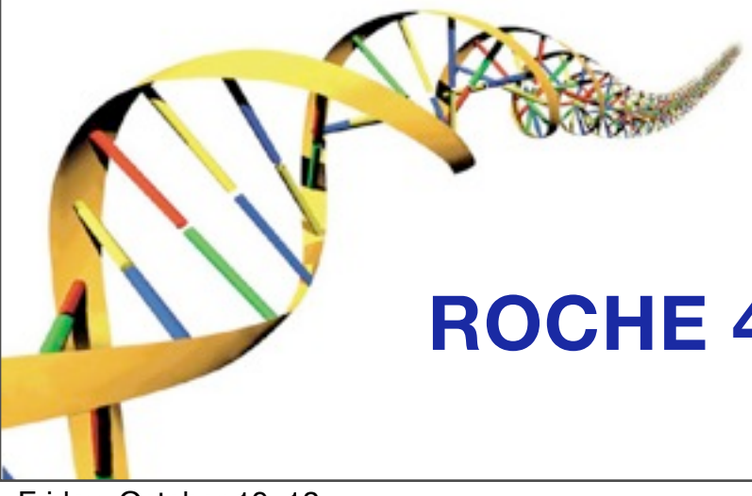


PicoTiterPlate (PTP)





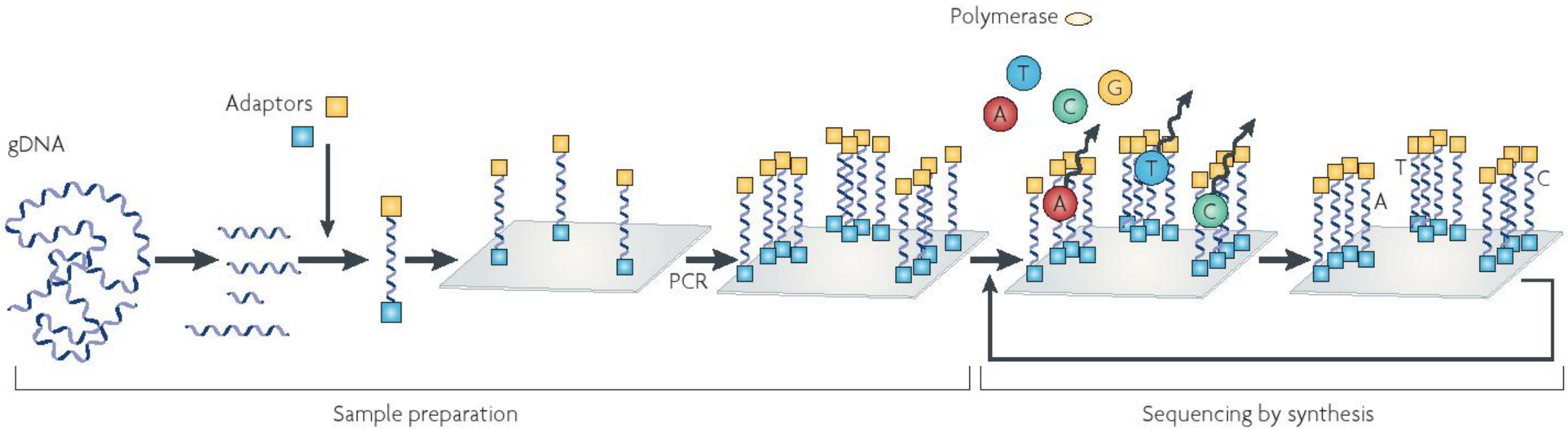




ROCHE 454

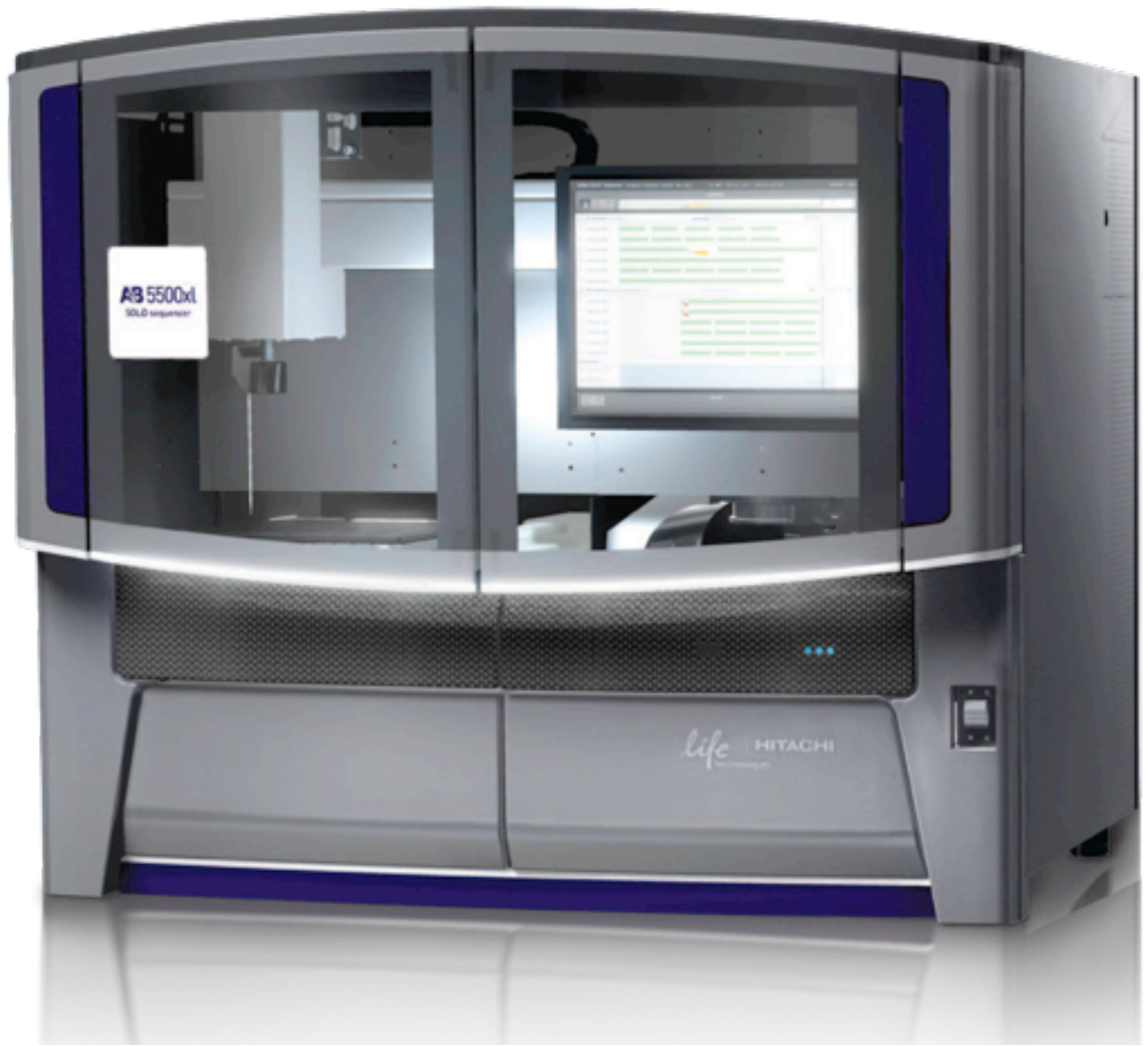
See video: http://www.youtube.com/watch?v=bFNjxKHP8Jc

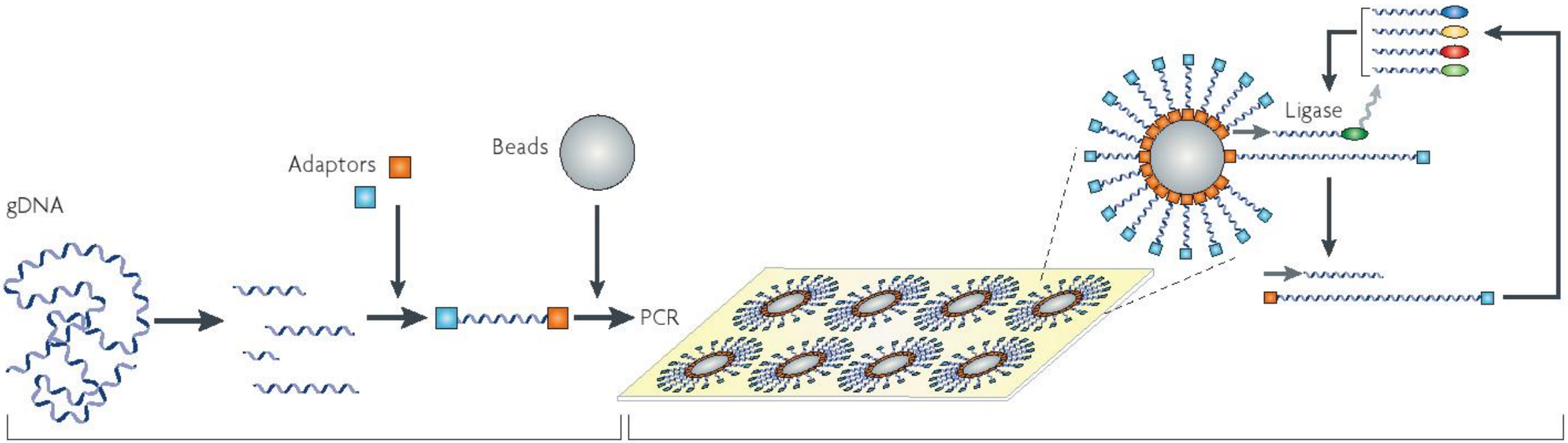
Illumina HiSeq 2000 and MiSeq

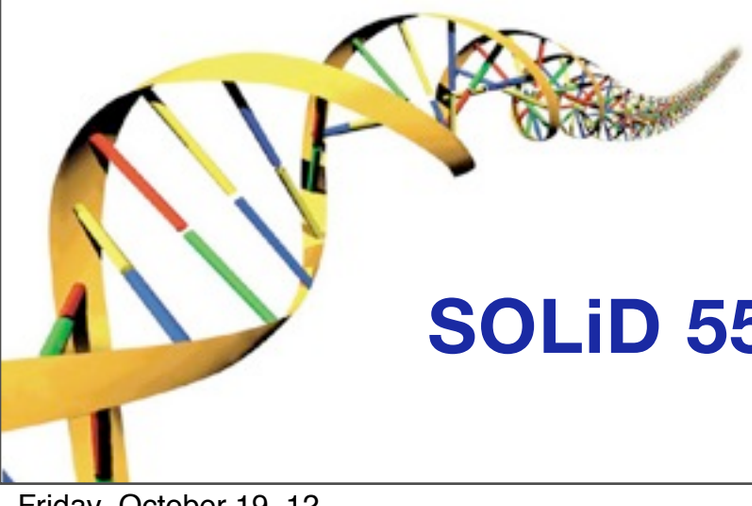


Illumina HiSeq and GAIIx

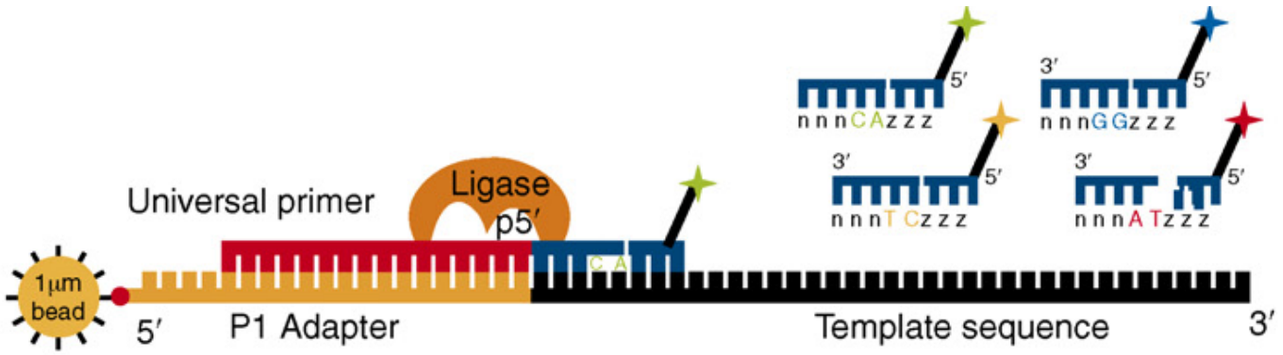
SOLiD 5500xl

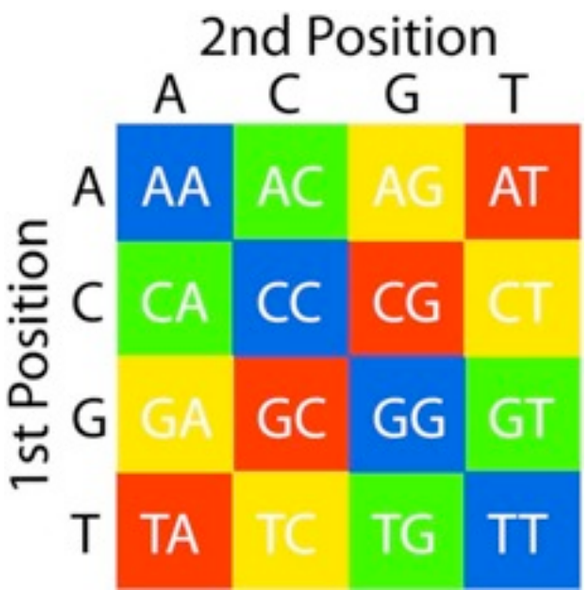


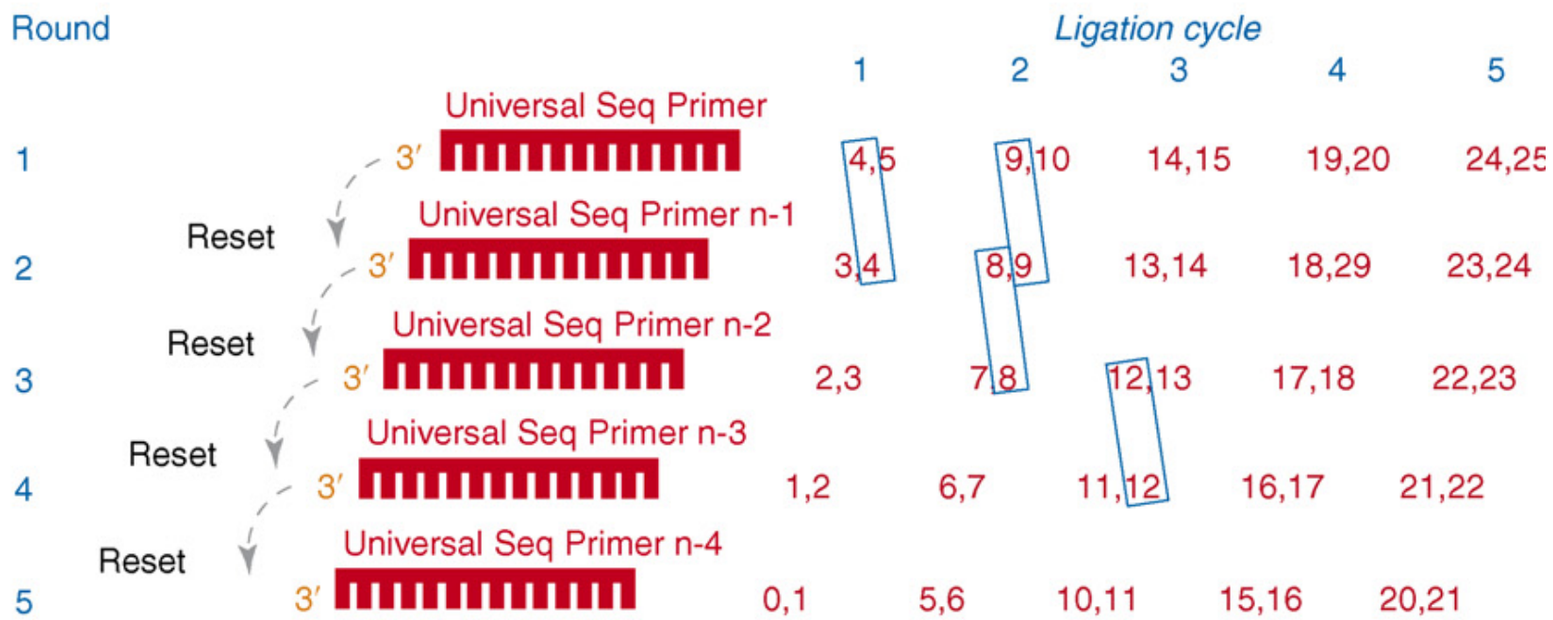


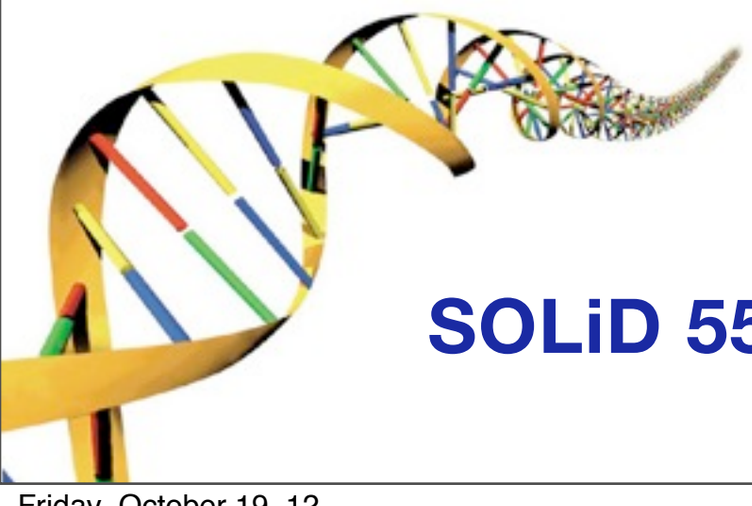


SOLiD 5500xl

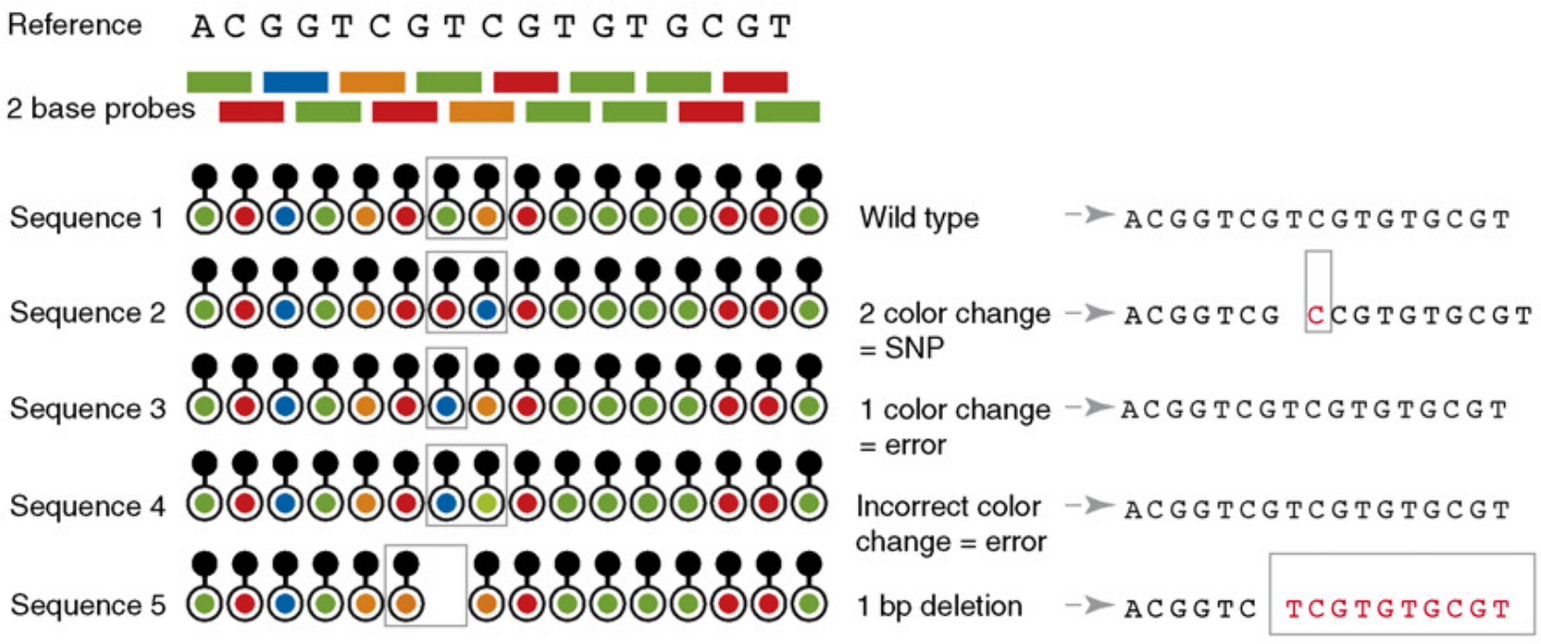








SOLiD 5500xl



PacBio RS System

Sequencing chemistry

Step 1: fluorescent phospholinked labeled nucleotides enter the ZMW (zero-mode waveguide)

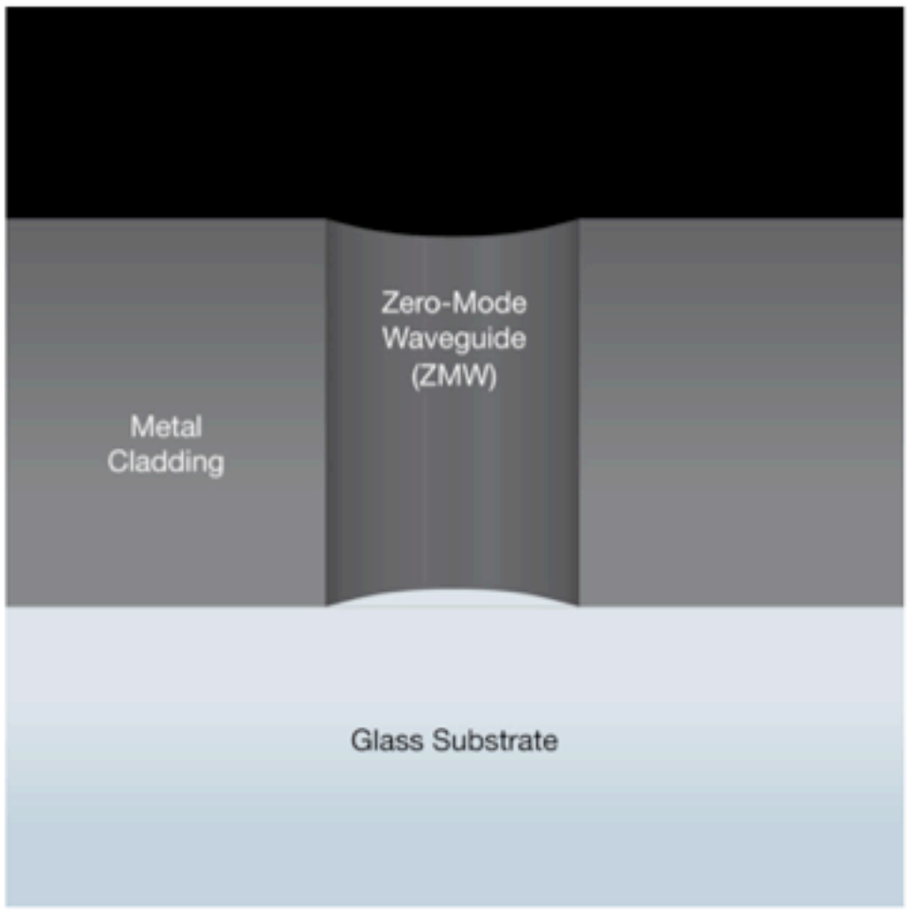
Step 2: the incorporated base is held in the detection volume for 10s of mS, releasing light

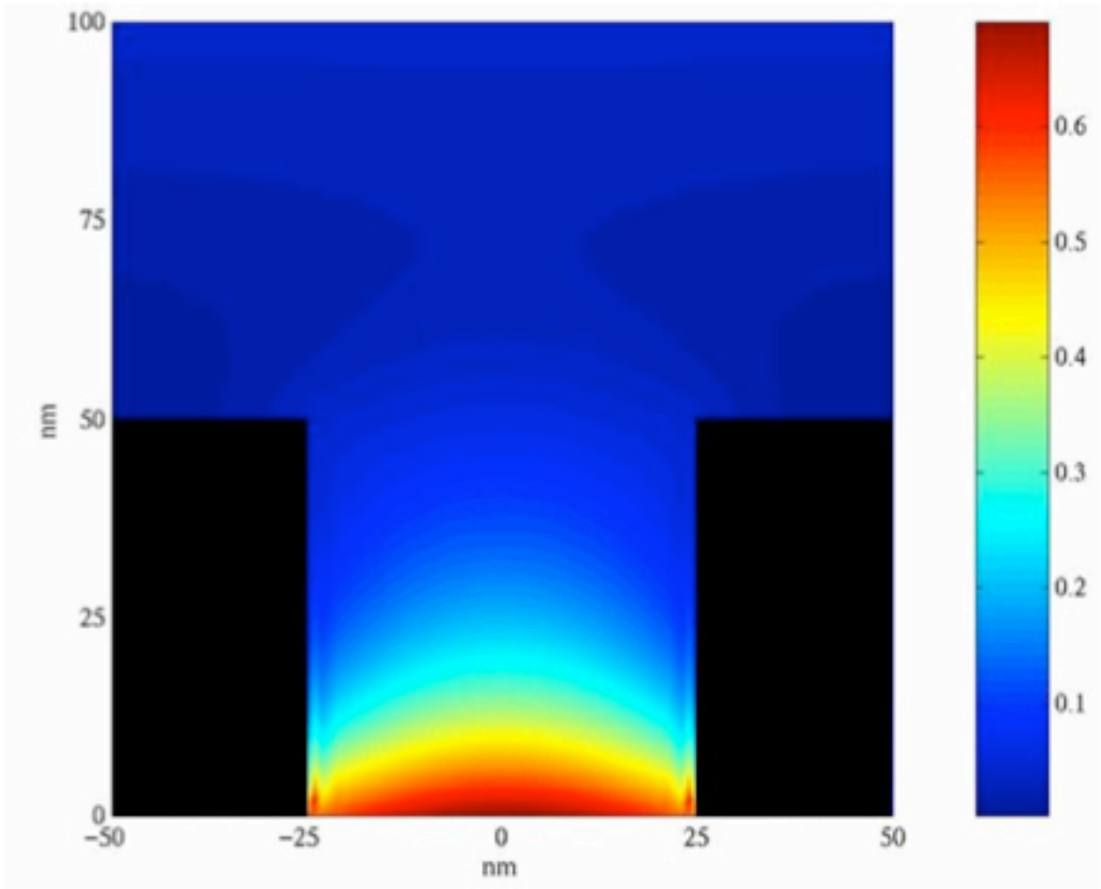
Step 3: the phosphate chain is cleaved, releasing the dye

Steps 4-5: the process repeats

Detection system

nanophotonic visualization: fluorescence present only in lower 20-30 nm





individual ZMW

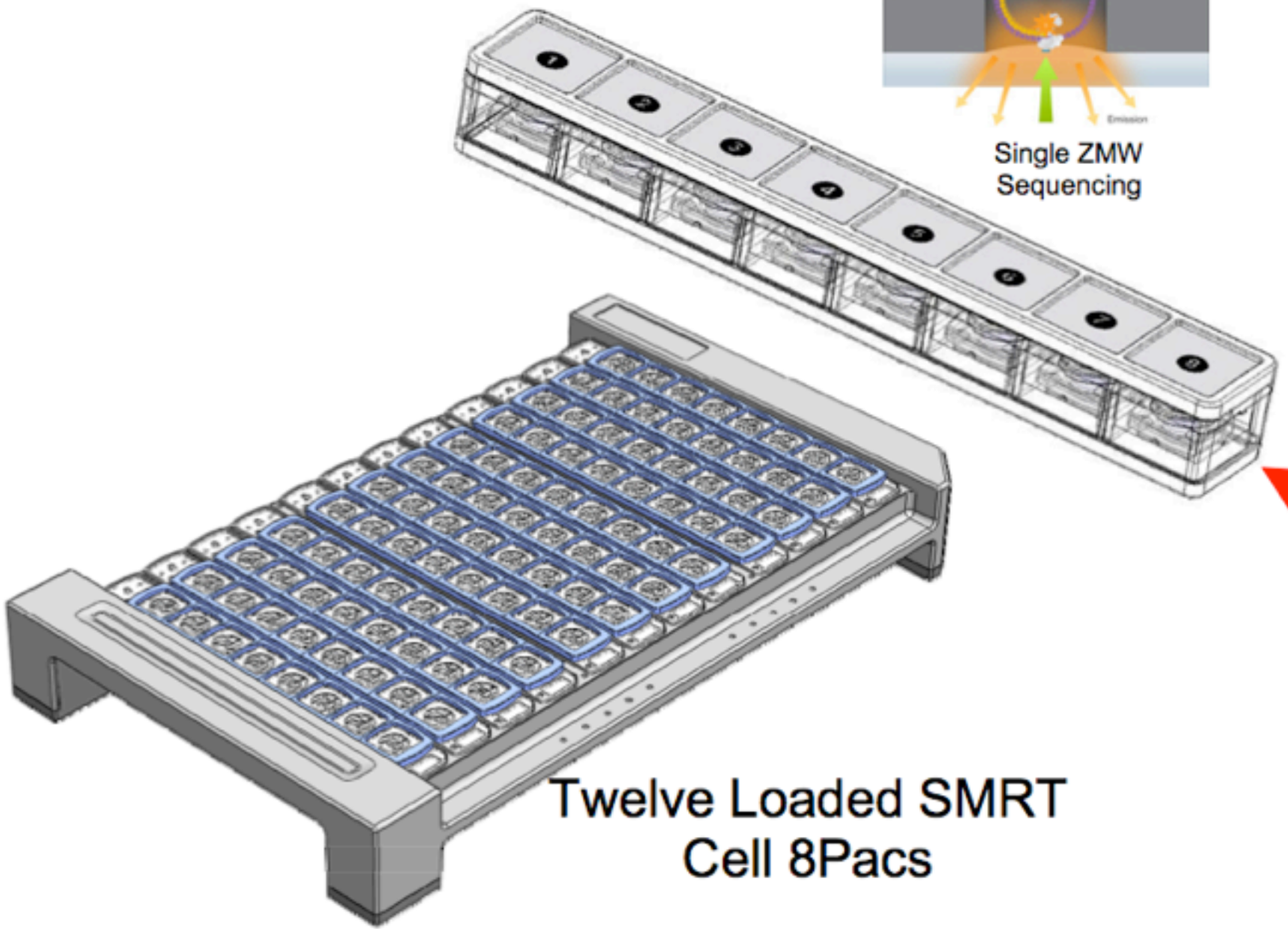
zero-mode waveguide

detection volume

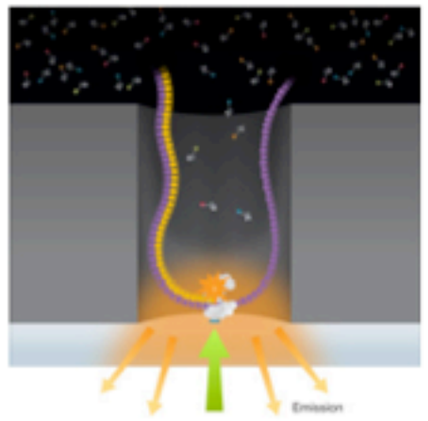
20 zeptoliters (10^{-21} liters)

SMRT Cell Arrangement

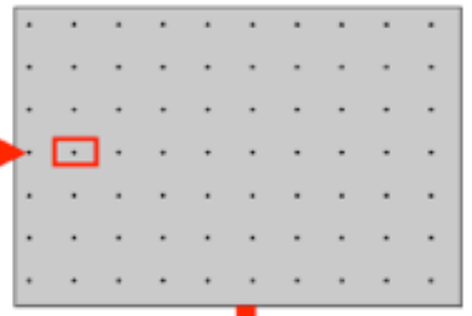
SMRT Cell 8Pac



Twelve Loaded SMRT Cell 8Pacs

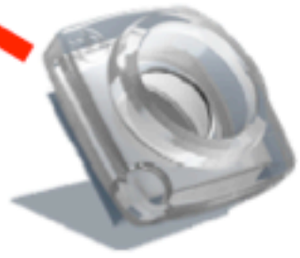


Single ZMW Sequencing



Array of ZMWs

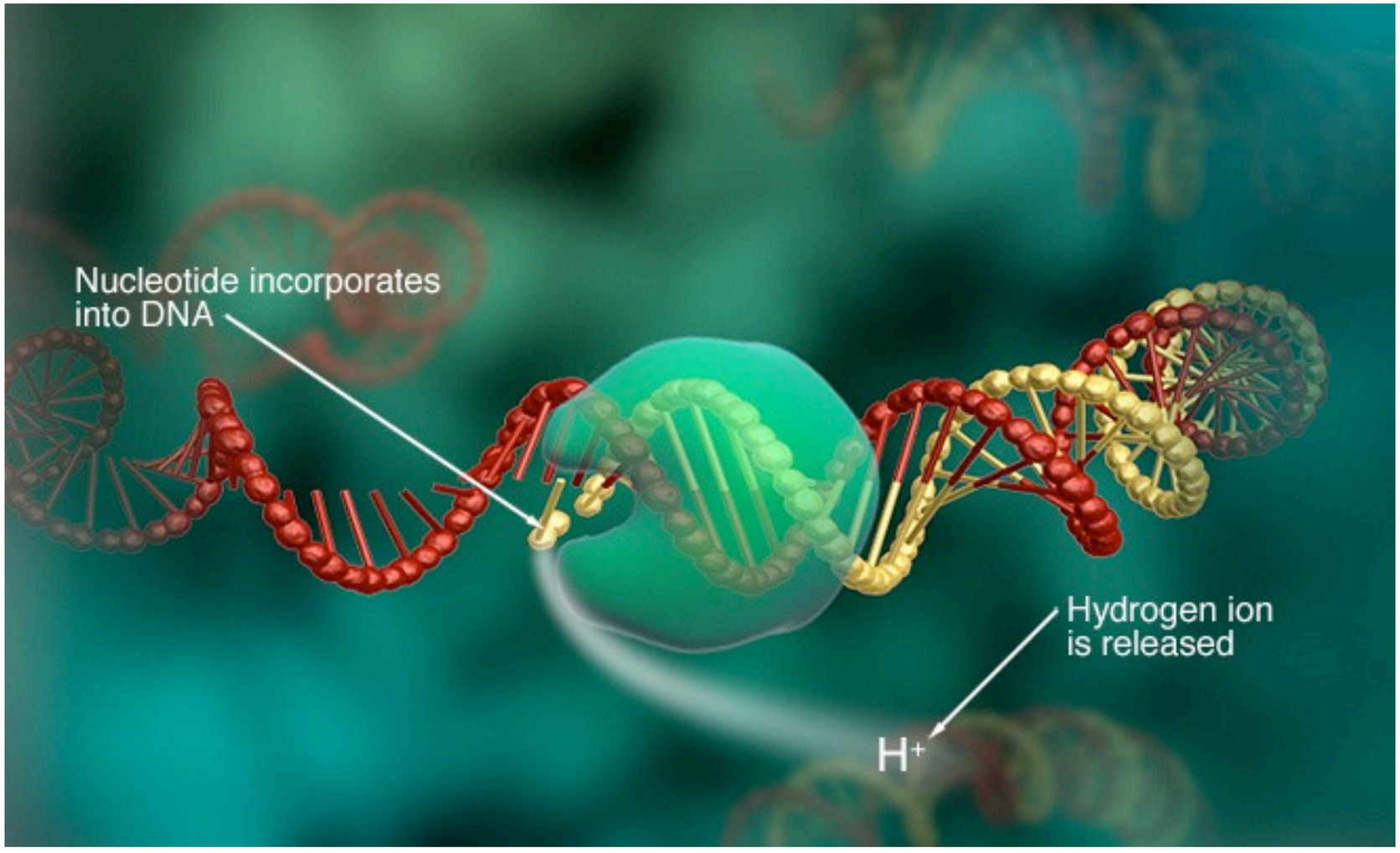
SMRT Cell Pattern

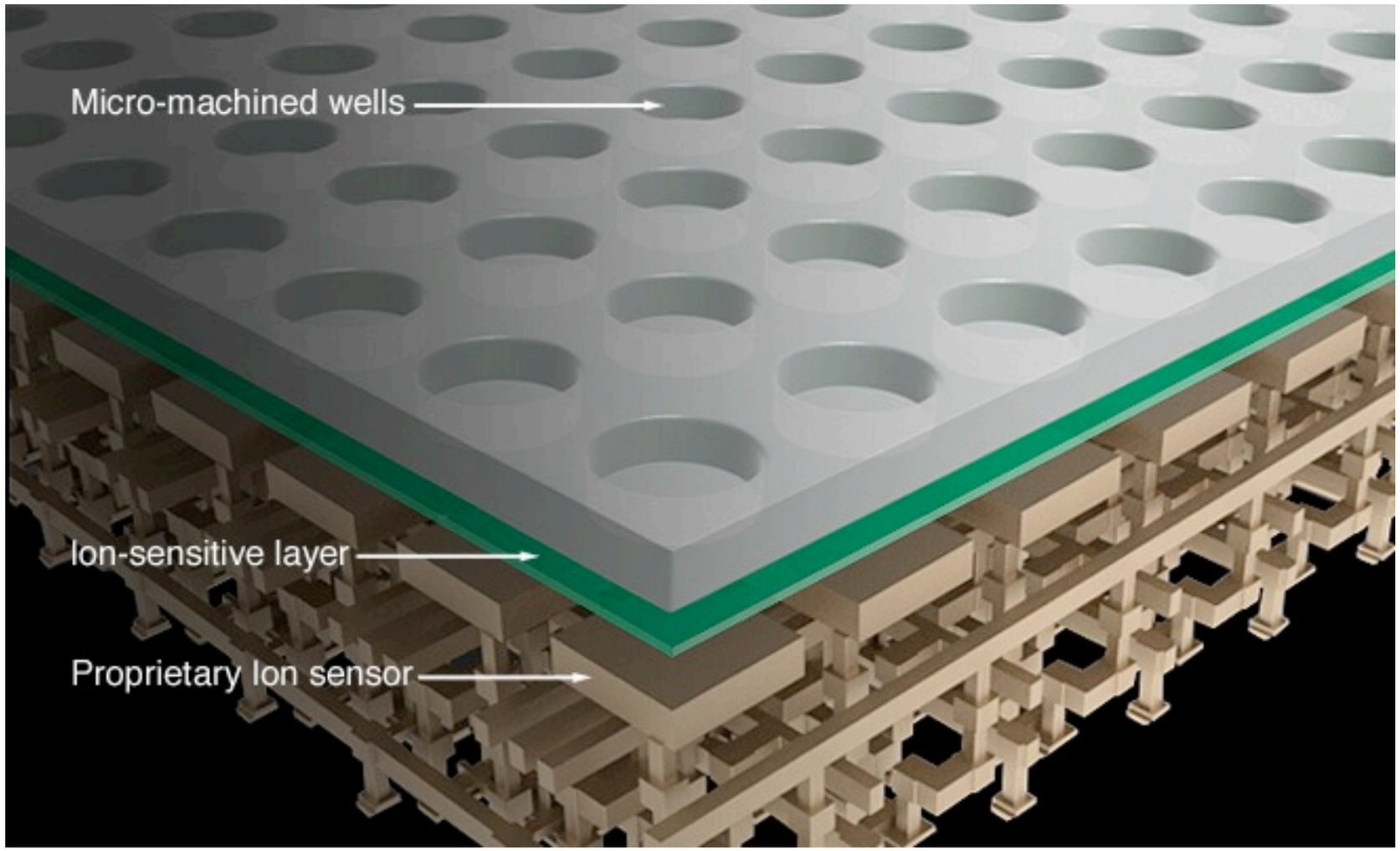


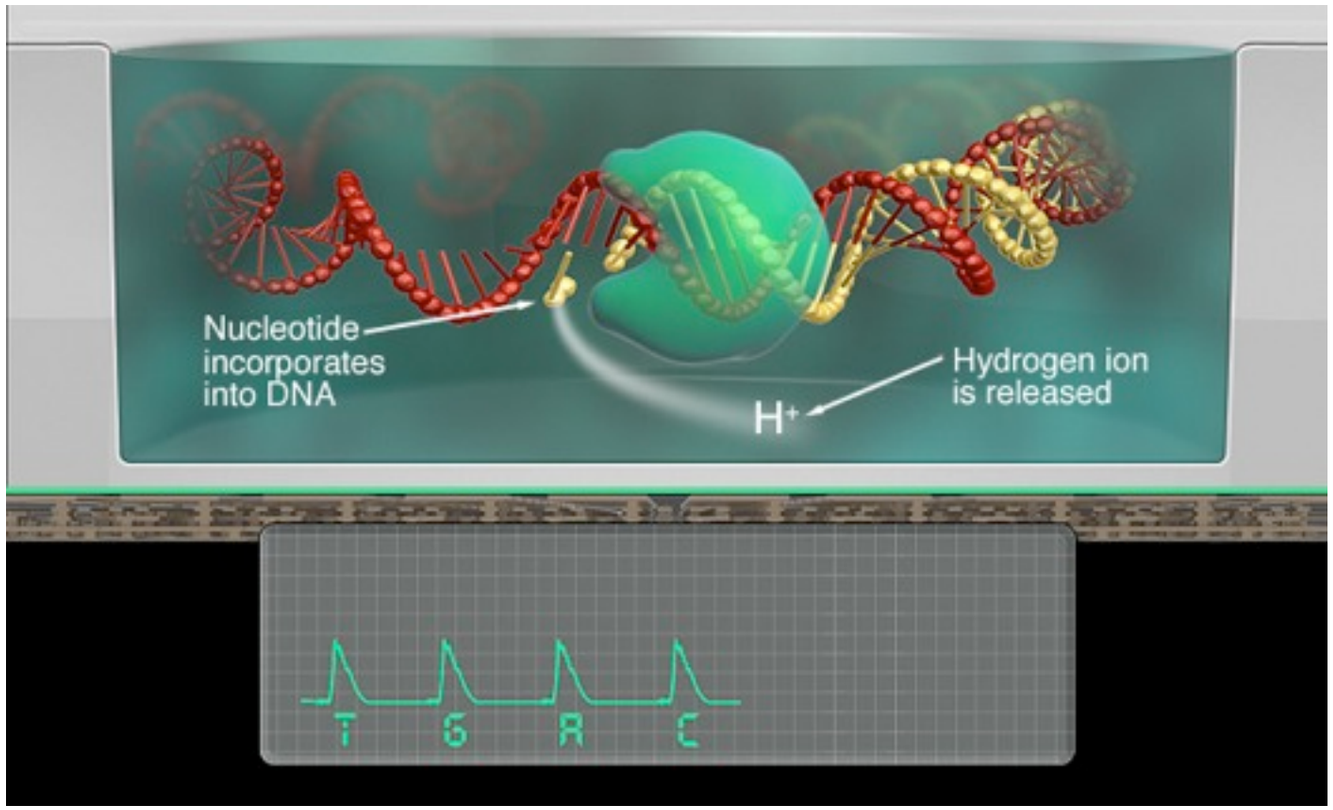
SMRT Cell

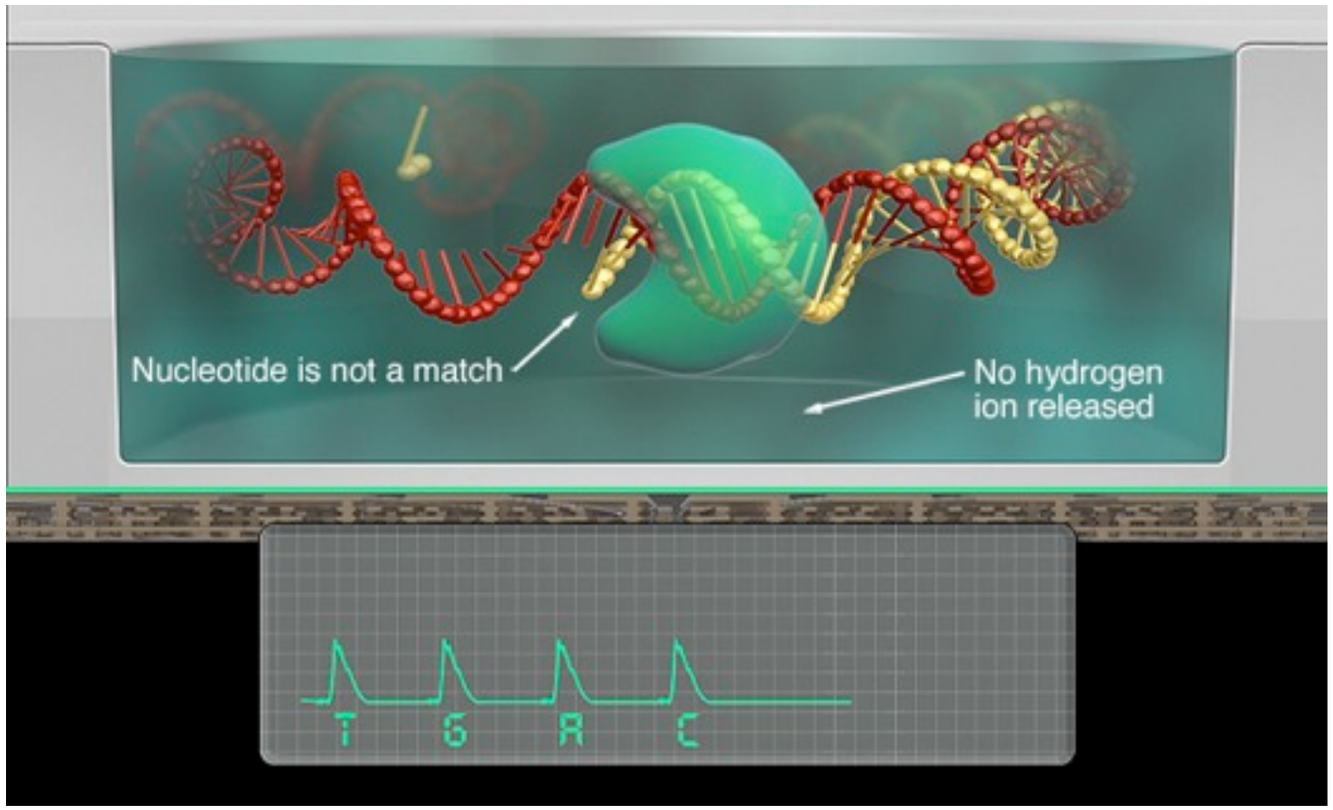
2x75,000 ZMWs

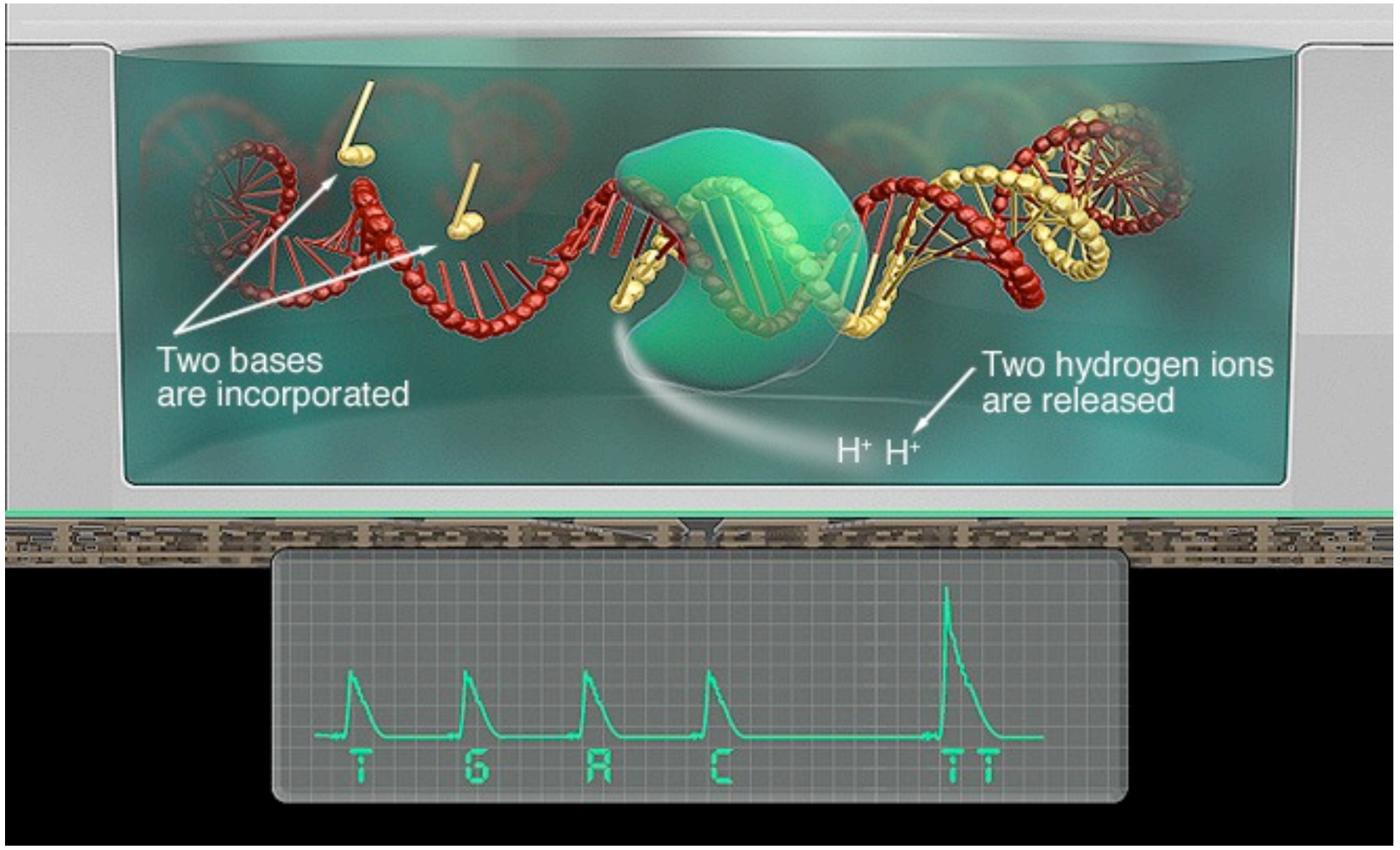
Ion Torrent PGM


Nucleotide incorporates
into DNA
Hydrogen ion
is released
H+


Micro-machined wells
Ion-sensitive layer
Proprietary Ion sensor


Nucleotide incorporates into DNA
H+
Hydrogen ion is released
T G A C


Nucleotide is not a match
No hydrogen ion released
T G A C


Two bases
are incorporated
Two hydrogen ions
are released
H+ H+
T G A C TT

FASTQ file

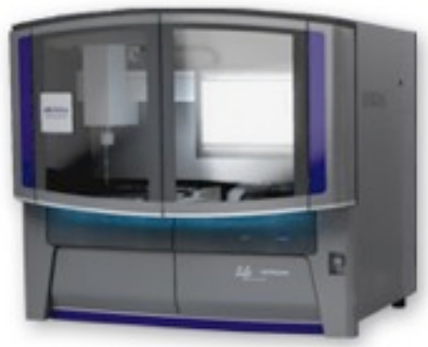
Read name
Read seq
Read name
Read qual

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_```^a``a`^a_^__]a_]\]`a_____`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
@HWI-EAS121:4:100:1783:1837#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATATCGTATGCCGTCTTCTGCTTTAATAAAAAAAAA
+HWI-EAS121:4:100:1783:1837#0/1
aaaaaab`aaaaaa\aaabaaaZ`b`baaaaTYXZ\Q\YZ[^_]MOOQPMHDPRFTTNHH[GMJDRODDDHNNWTUVXPG
@HWI-EAS121:4:100:1783:1127#0/1
TGCTTCTACCGGAGGGAGTACAATGTCTTCCACTGTGATCATCAACTGAATGATCCCCTTCCCAACTGAAATCCTCCTTT
```
```

PACBIO RS

Roche GS FLX+ (454)

Ion Torrent PGM



ABI SOLiD5500xl

Illumina MiSeq

Illumina HiSeq

long reads

short reads

Emulsion PCR

Clusters

Synthesis

Pyrosequencing

H+

Ligation

Synthesis

~10 million reads
~3kb
low accuracy

~1 million reads
~800bp

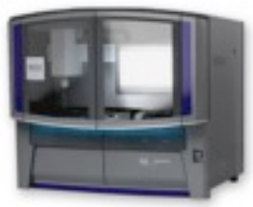
~7 million reads
~300bp

medium-high accuracy

~3 billion reads
up to 75bp/read
highest accuracy

~6 billion reads
up to 100bp/read
medium-high accuracy

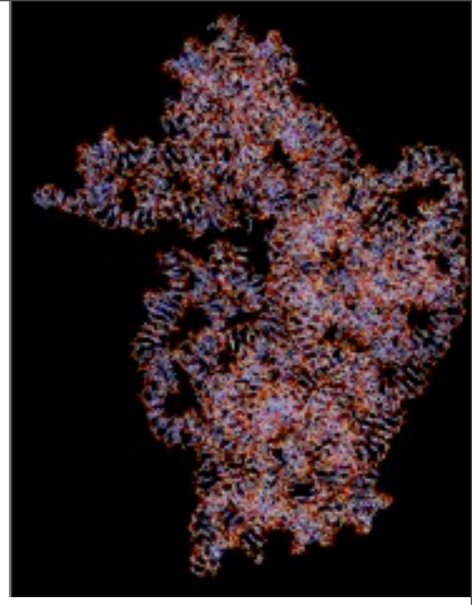
~15 million reads
up to 250bp/read
medium-high accuracy

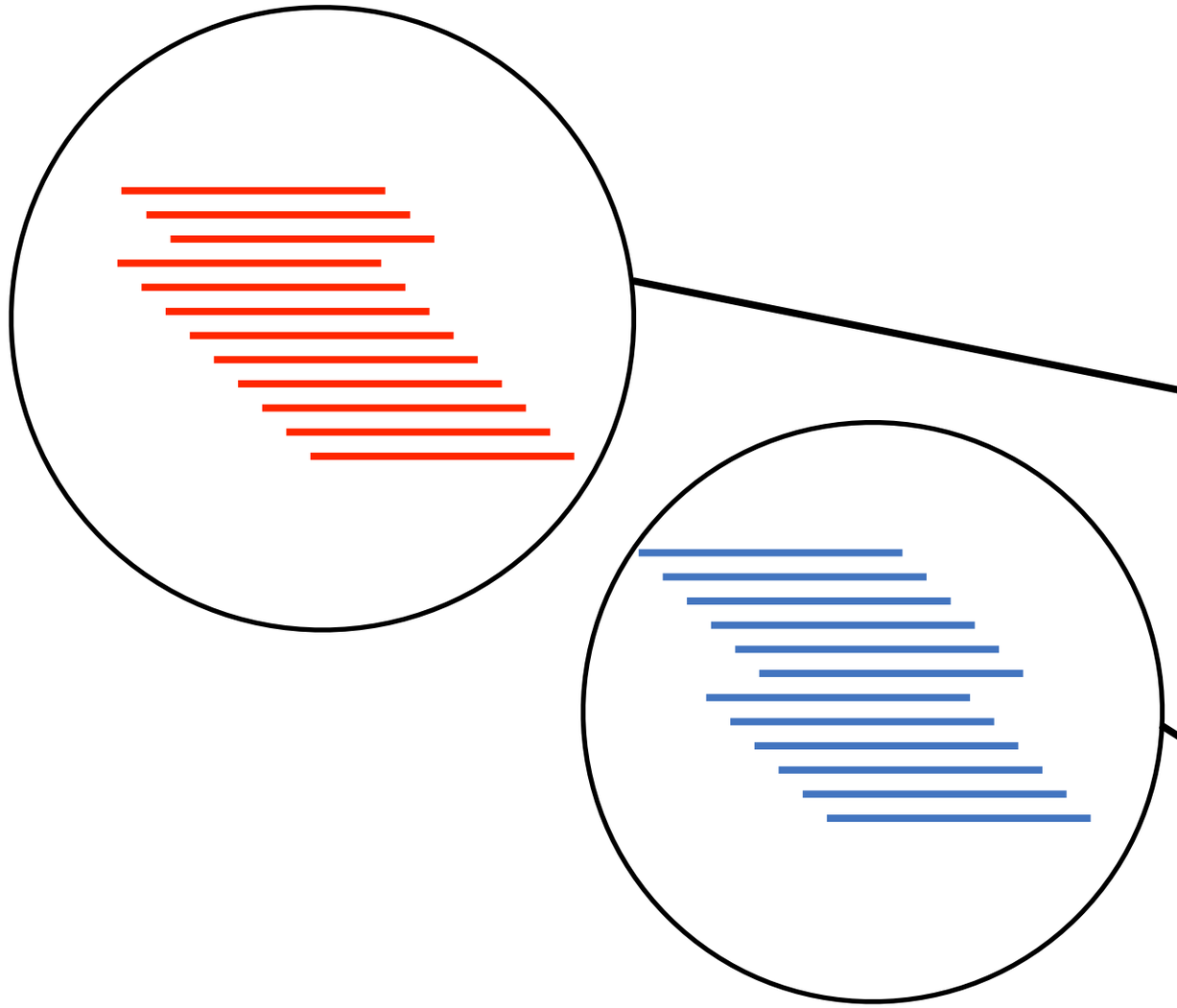
	Pros	Cons
 Roche GS FLX + (454)	•long reads •good for repeats •relatively fast	•throughput •homopolymers •cost
 ABI SOLiD5500xl	•throughput •accuracy	•short reads •bad with repeats
 Illumina HiSeq	•highest throughput •longer reads than SOLiD	•short reads •bad with repeats •issues with low diversity
 Illumina MiSeq	•cheap and fast	•issues with low diversity •bad with long repeats •throughput
 PACBIO RS	•cheap and fast •the longest reads	•the lowest throughput •lowest accuracy
 Ion Torrent PGM	•cheap and fast	•throughput •homopolymers

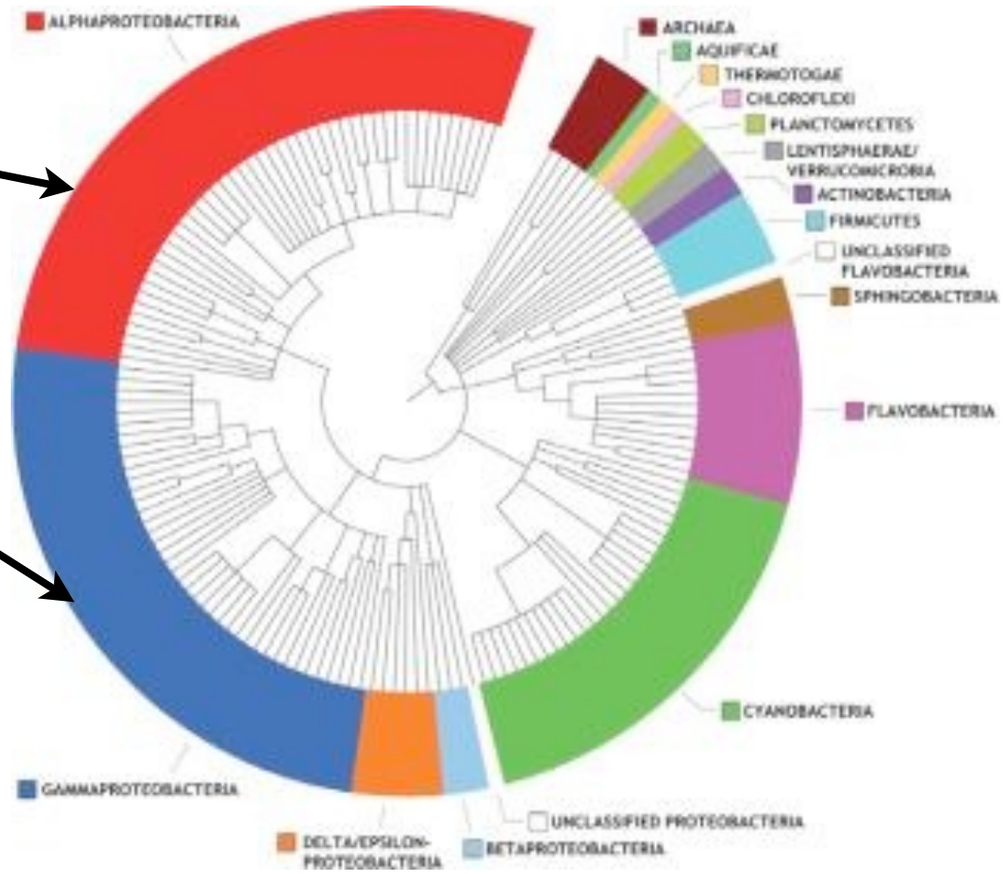
Parameters for applications

- read length: better assembly
- accuracy: better SNP calling
- throughput: better coverage
- cost

Metagenomics: using a genomic marker (e.g. 16S rRNA) (Amplicon)

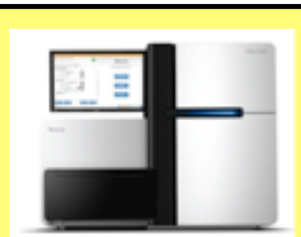




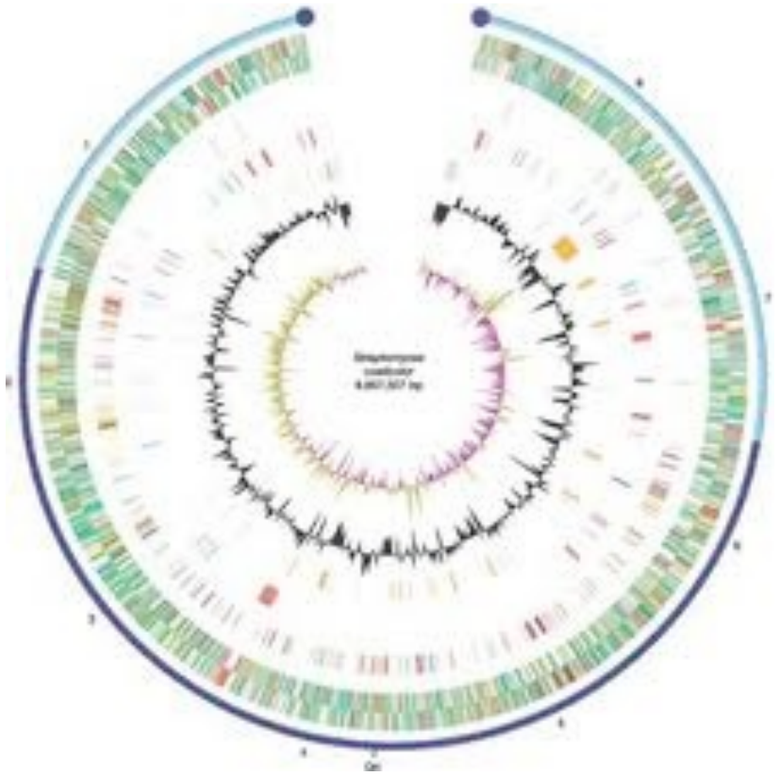


Long amplicon (more specific)

Short amplicon (less specific)



De novo bacterial genome sequencing



Easier to assemble

More difficult but possible

SNP calling (mapping)

Bacterial genome re-sequencing --SNP calling

Human genome re-sequencing --SNP calling

requires >~30x

Less accuracy


Good accuracy