Next-gen sequencing

Roman Hobza

CG920 Genomics

Lecture 7

Applied Biosystems
ABI 3730XL
1 Mb / day

Roche / 454
Genome Sequencer FLX
100 Mb / run

Illumina / Solexa
Genetic Analyzer
2000 Mb / run

Applied Biosystems
SOLiD
3000 Mb / run

PACIFIC BIOSCIENCES®

ion torrent

Oxford NANOPORE Technologies®

Illumina Inc. (ILMN) - NasdaqGS

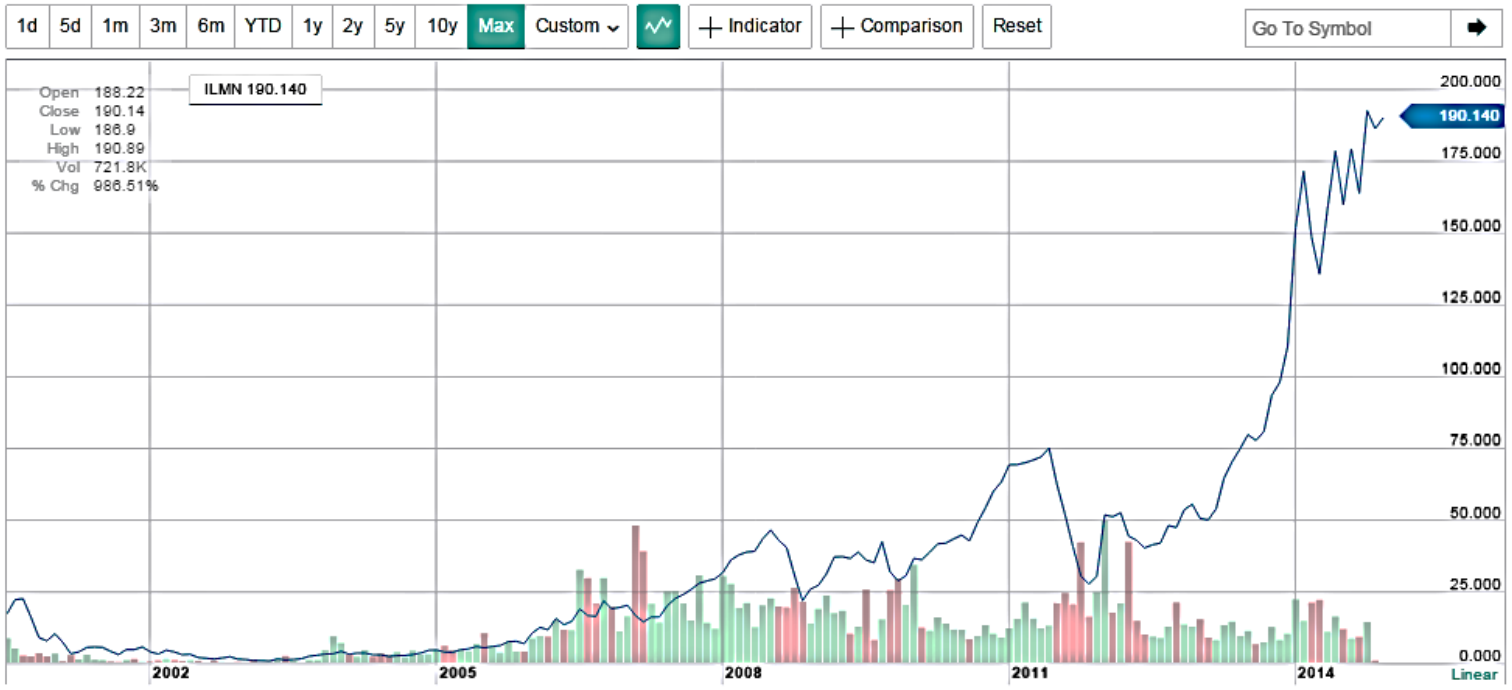
Add to Portfolio | Like 17

48.51 ↑0.07 (0.14%) 11:41AM EST - Nasdaq Real Time Price

Enter name(s) or symbol(s) | GET CHART | COMPARE | EVENTS | TECHNICAL INDICATORS | CHART SETTINGS | RESET

Pacific Biosciences of Californ

PACB


1d 5d 1m 3m 6m YTD 1y 2y 5y 10y Max Custom
Indicator
Comparison
Reset
Go To Symbol
Open 188.22
Close 190.14
Low 186.9
High 190.89
Vol 721.8K
% Chg 986.51%
ILMN 190.140
200.000
190.140
175.000
150.000
125.000
100.000
75.000
50.000
25.000
0.000
2002
2005
2008
2011
2014
Linear


1d 5d 1m 3m 6m YTD 1y 2y 5y 10y Max Custom
Indicator
Comparison
Reset
Go To Symbol
Open 1.88
Close 2
Low 1.77
High 2.21
Vol 913.2K
% Chg -88.1%
PACB 2.00
20.00
15.00
10.00
6.53
5.00
2.00
0.00
Jan 3, 2011
Jan 2, 2012
Jul 22, 2012
Jan 7, 2013
Jan 6, 2014
Linear

Comparison to other sequencing methods

	Ion Torrent [14][16]	454 Sequencing [17]	Illumina [18]	SOLiD [19]
Sequencing Chemistry	Ion semiconductor sequencing	Pyrosequencing	Polymerase-based sequence-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Emulsion PCR	Bridge amplification	Emulsion PCR
Mb per run	100	100	600,000	170,000
Time per run	1.5 hours	7 hours	9 days	9 days
Read length	200 bp	400 bp	2x150 bp	35x75 bp
Cost per run	$ 350 USD	$ 8,438 USD	$ 20,000 USD	$ 4,000 USD
Cost per Mb	$ 5.00 USD	$ 84.39 USD	$ 0.03 USD	$ 0.04 USD
Cost per instrument	$ 50,000 USD	$ 500,000 USD	$ 600,000 USD	$ 595,000 USD

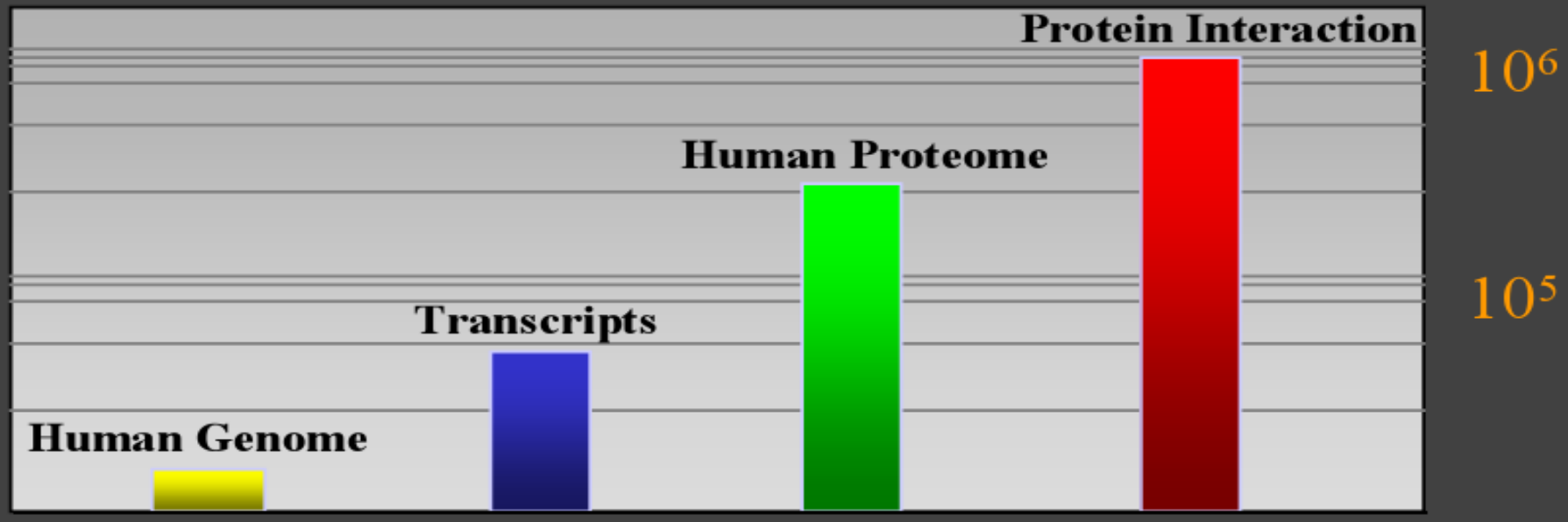
Genome: 30.000 genes

↓

Transcriptome: 40-100.000 mRNAs

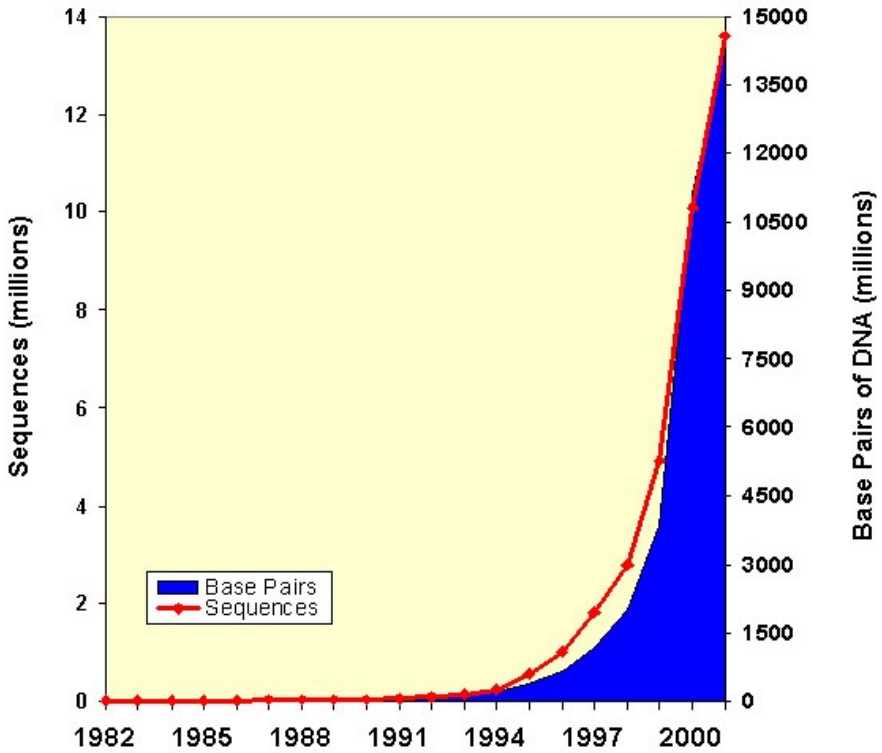
↓

Proteome: 100-400.000 proteins
>1.000.000 interactions



Sequencing of genomes

GenBank originated in 1982 from Los Alamos Sequence Database



Walter Goad

Why do we need sequencing?

- Comparative genomics
- Biomedicine research
- Personal genome

Frederick Sanger

1958 – Nobel prize – protein sequencing

1975 – dideoxy sequencing method

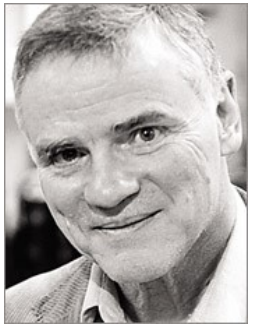
1977 – Φ-X174 (5,368 bp)

1980 – Nobel prize – DNA sequencing

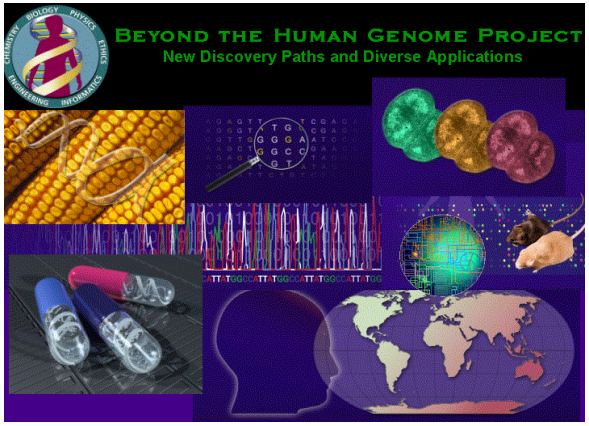
Phage λ - shotgun method (48,502 bp)

Genome sequencing

- **1986 Leroy Hood: automatic sequencer**
- **1986 Human Genome Initiative**
- **1990 HUGO**



Leroy Hood



Genome sequencing

- **1995** John Craig Venter – the first bacterial genome

- **1996** first eukaryotic genome (yeast)

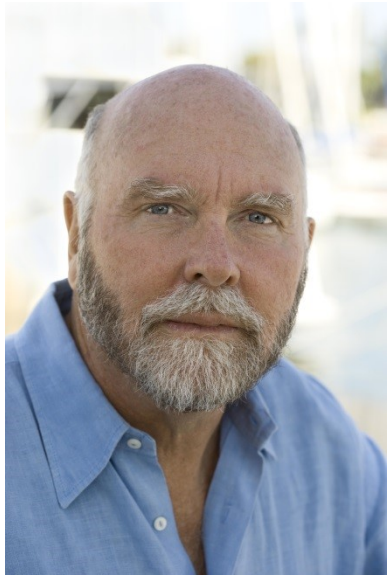
John Craig Venter

Craig Venter

Global Ocean Sampling Expedition

Synthetic genomics

Human Longevity Inc



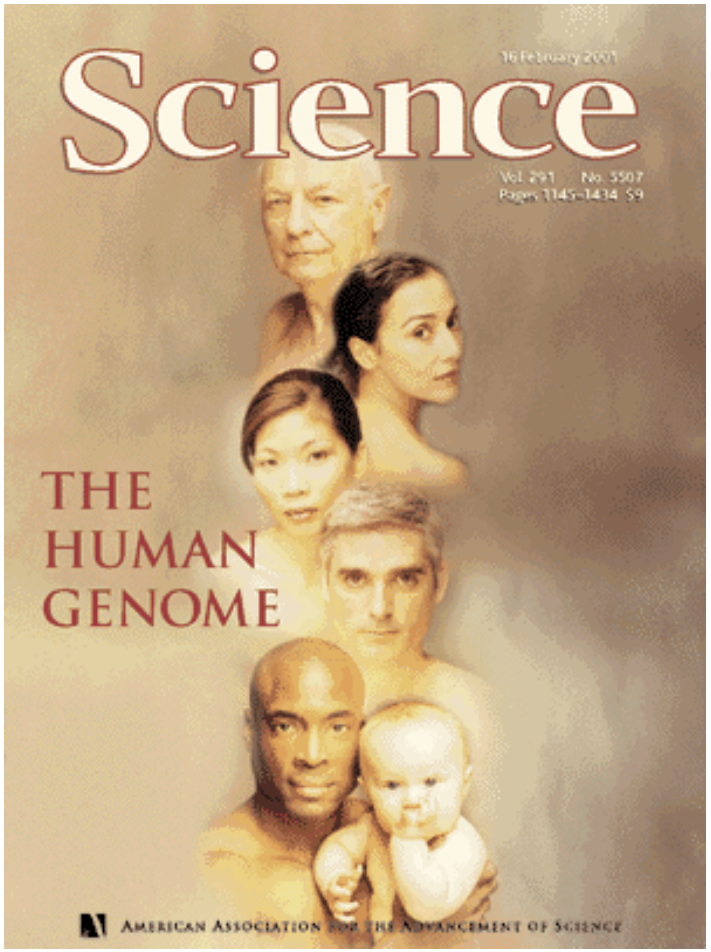
http://www.youtube.com/watch?v=J0rDFbrhjtI

Genome sequencing

- **1997** *E. coli sequence*
- **1998** *Caenorhabditis elegans* genome (the first multicellular genome)
- **1999** human chromosome 22

Genome sequencing

- **2000** *Drosophila melanogaster* genome
- **2001** Human Genome Sequencing: draft sequence

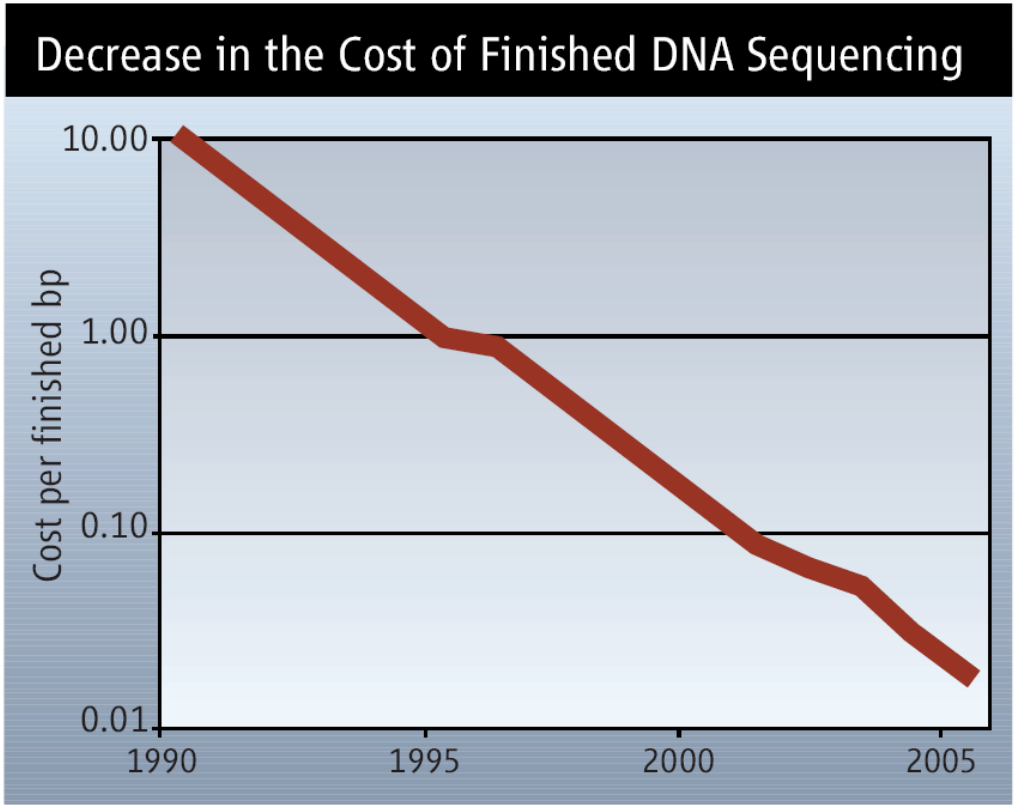


Genome sequencing

- **duben 2003** mouse draft genome
- **duben 2004** rat draft genome

2010 Perfect human genome

Race for sequences



Free fall. As with computer technology, the plunging cost of DNA sequencing has opened new applications in science and medicine.

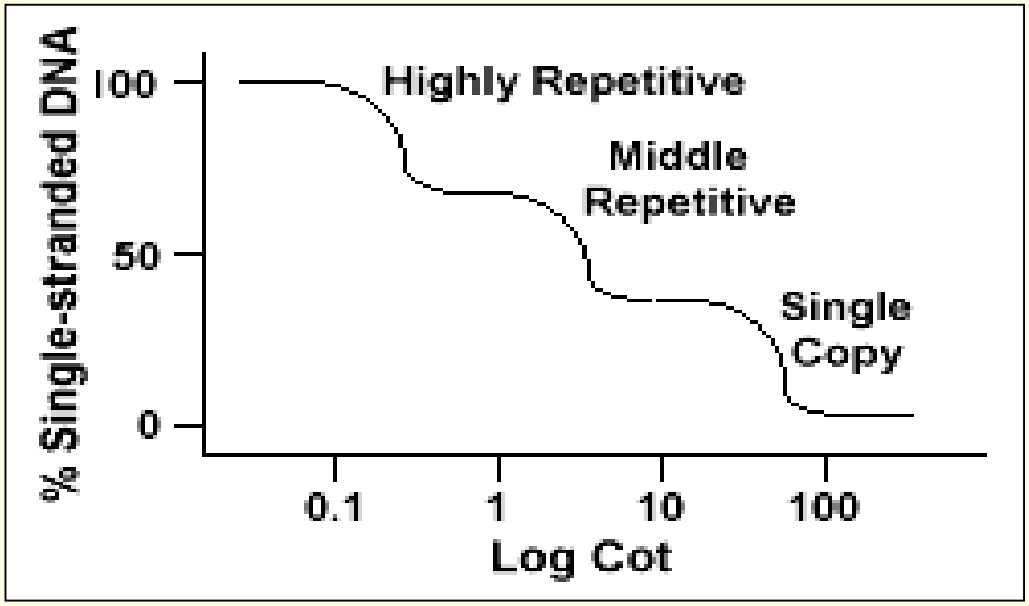
Human genome (first draft) – $300 million (2001)

Rhesus macaque – $22 million (2006)

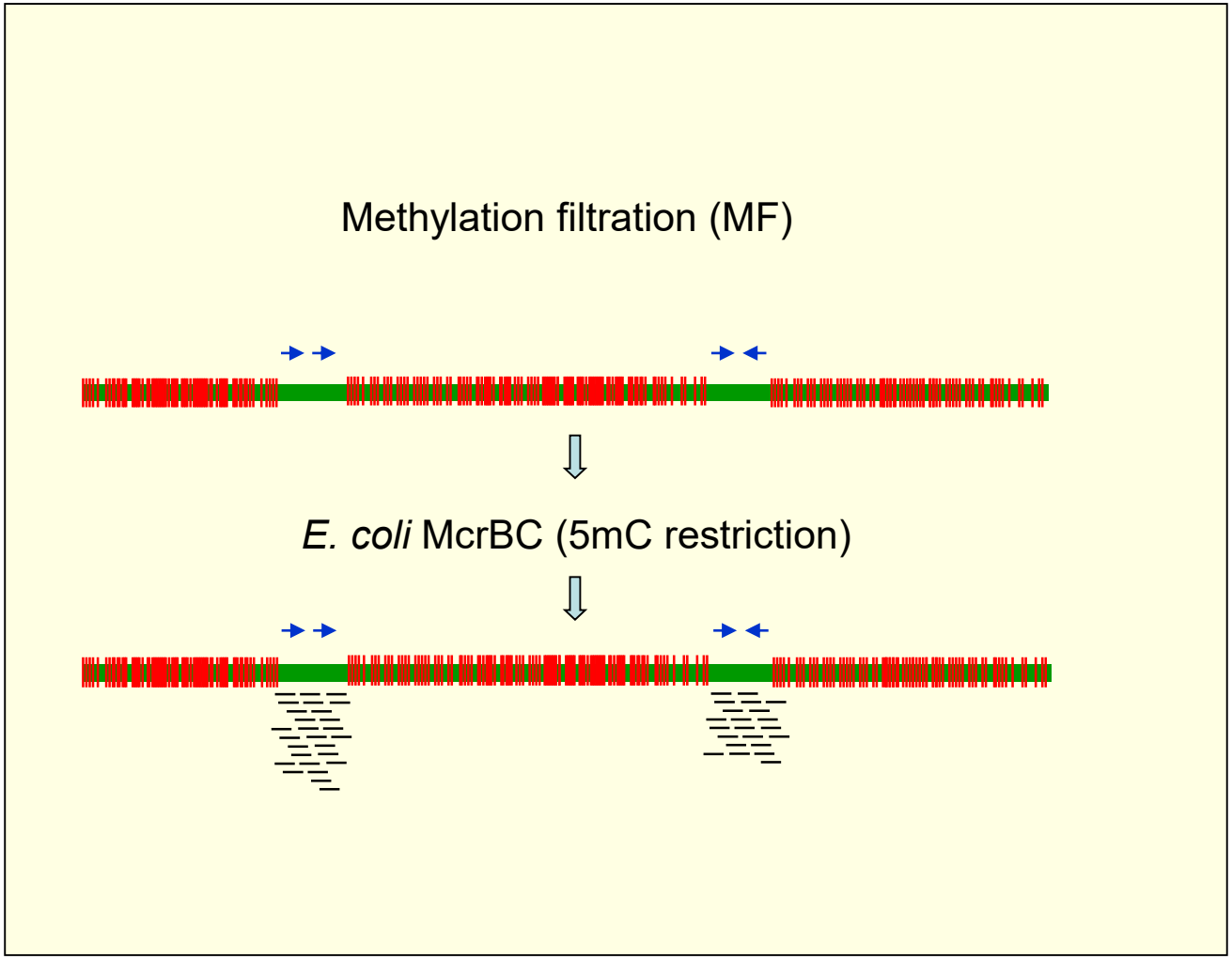
The Race for the $1000 Genome. ***Science*** 311: 1544 – 1546, **2006**

Complexity reduction

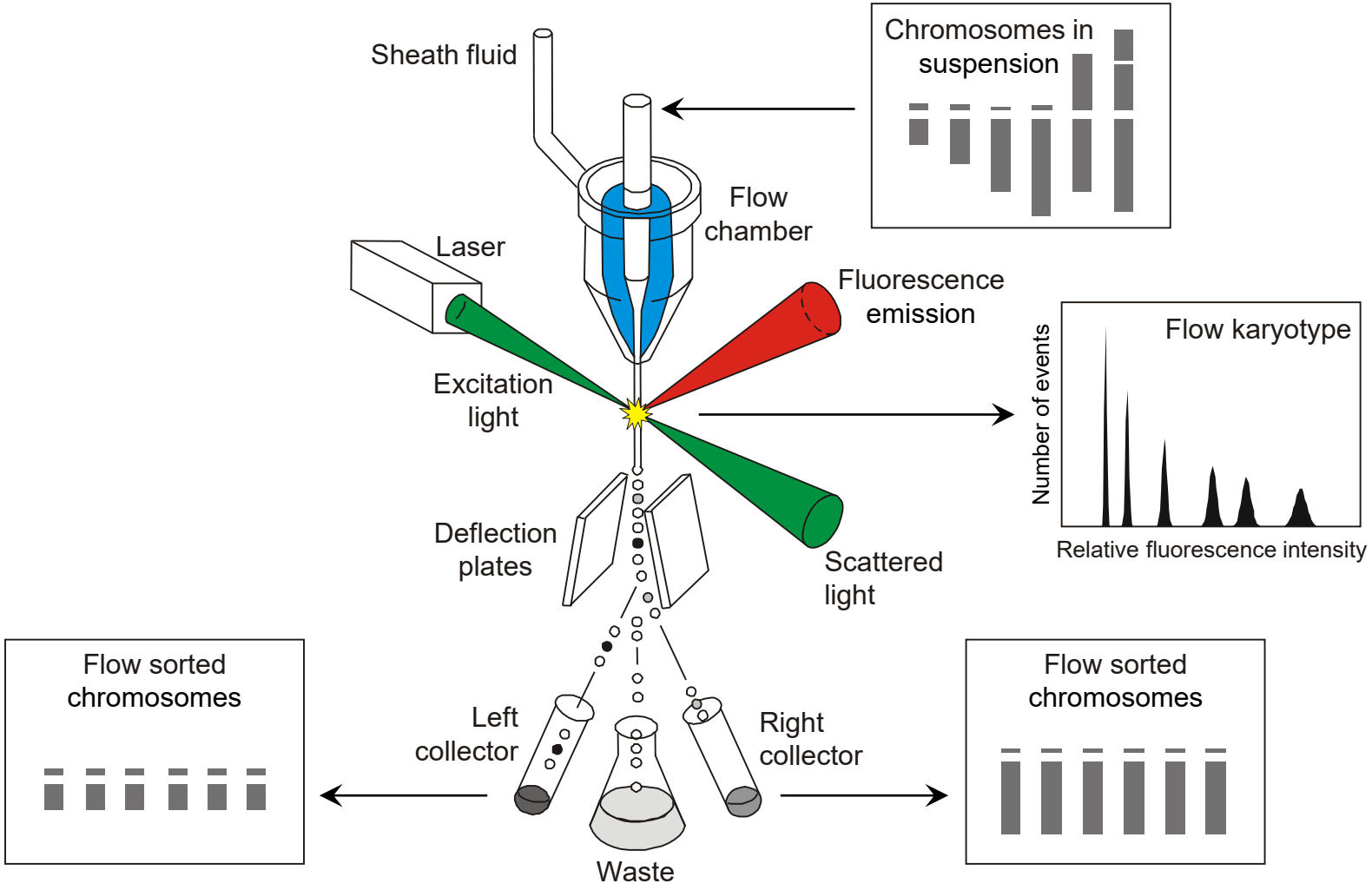
Hi-Cot selection



Complexity reduction



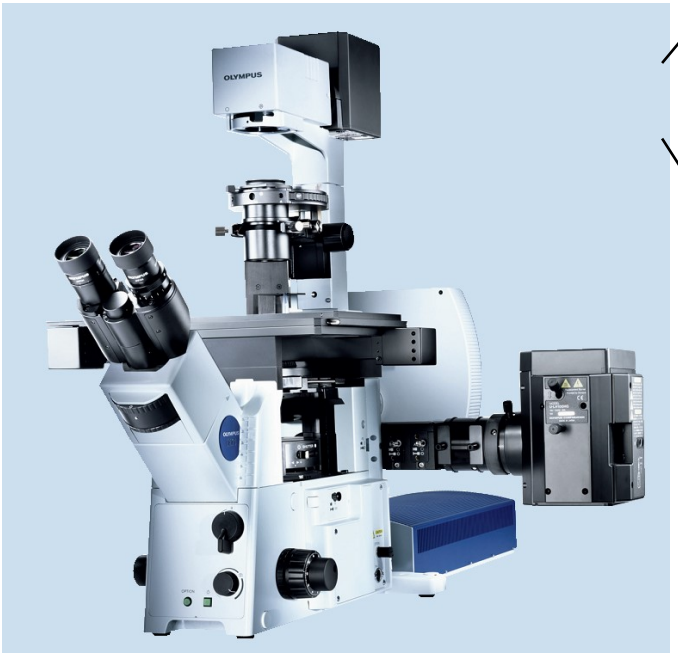
Complexity reduction – chromosome sorting

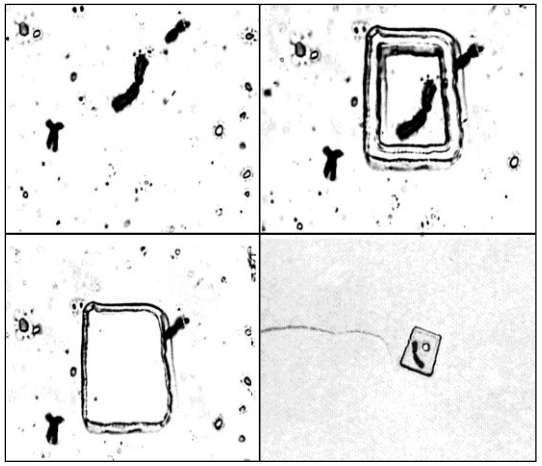


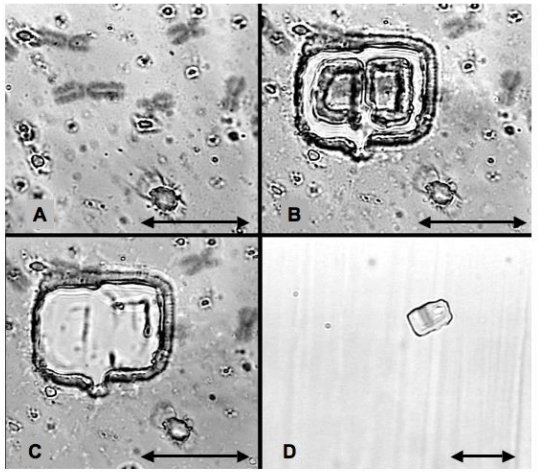
Laser microdissection

<u>Advantage:</u> purity

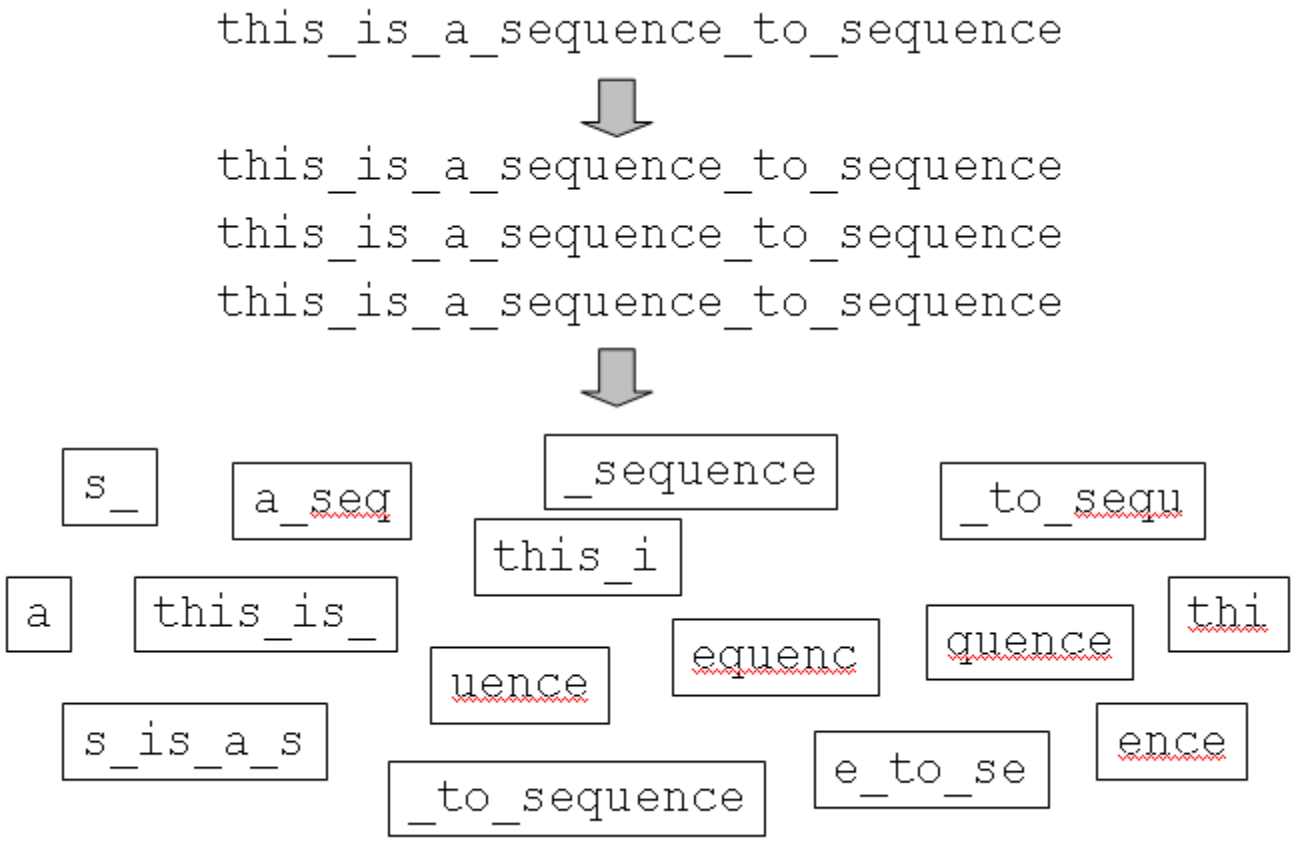
<u>Disadvantage:</u> small amount

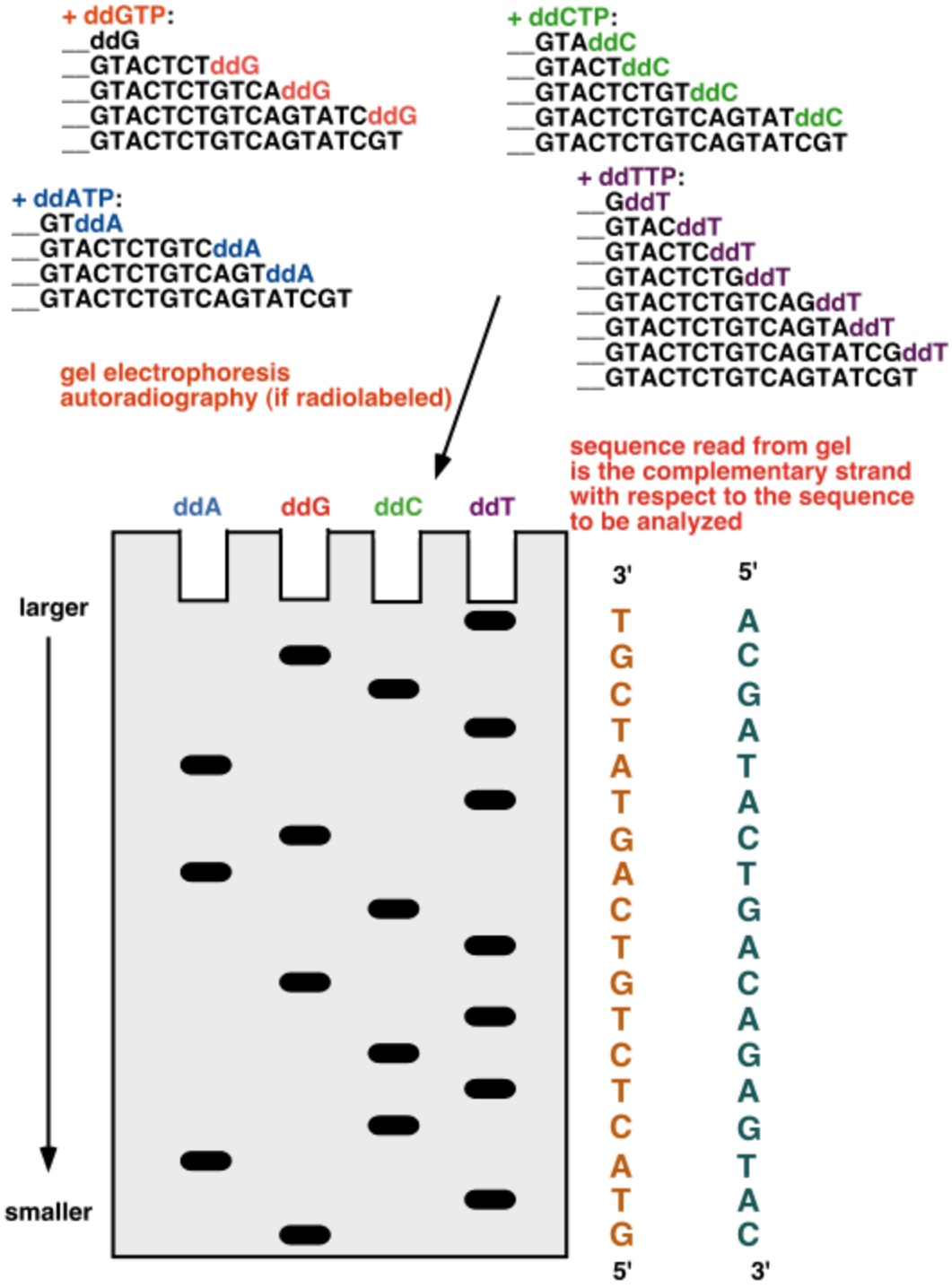






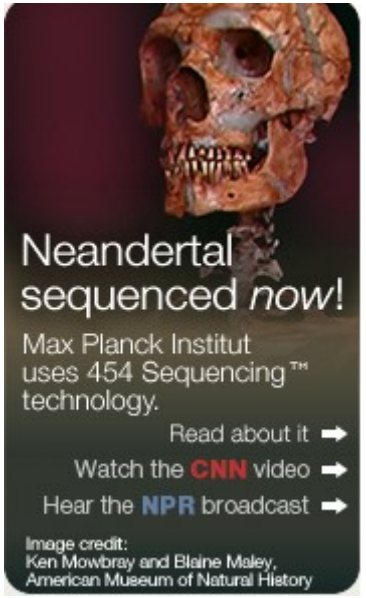
Methods


this_is_a_sequence_to_sequence
this_is_a_sequence_to_sequence
this_is_a_sequence_to_sequence
this_is_a_sequence_to_sequence
s_
a_seq
_sequence
_to_sequ
this_i
a
this_is_
thi
equenc
quence
uence
s_is_a_s
ence
e_to_se
_to_sequence


+ ddGTP:
__ddG
__GTACTCTddG
__GTACTCTGTCAddG
__GTACTCTGTCAGTATCddG
__GTACTCTGTCAGTATCGT
+ ddCTP:
__GTAddC
__GTACTddC
__GTACTCTGTddC
__GTACTCTGTCAGTATddC
__GTACTCTGTCAGTATCGT
+ ddATP:
__GTddA
__GTACTCTGTCddA
__GTACTCTGTCAGTddA
__GTACTCTGTCAGTATCGT
+ ddTTP:
__GddT
__GTACddT
__GTACTCddT
__GTACTCTGddT
__GTACTCTGTCAGddT
__GTACTCTGTCAGTAddT
__GTACTCTGTCAGTATCGddT
__GTACTCTGTCAGTATCGT
gel electrophoresis
autoradiography (if radiolabeled)
sequence read from gel
is the complementary strand
with respect to the sequence
to be analyzed
ddA
ddG
ddC
ddT
larger
smaller
3'
5'
T A
G C
C G
T A
A T
T A
G C
A T
C G
T A
G C
T A
C G
T A
A T
G C
5'
3'

Genome Sequencer 20 System
454 pyrosequencing (2005)

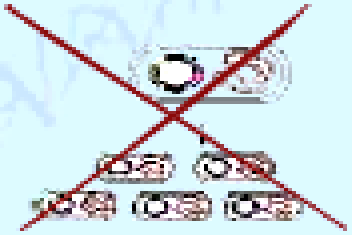
- http://www.454.com



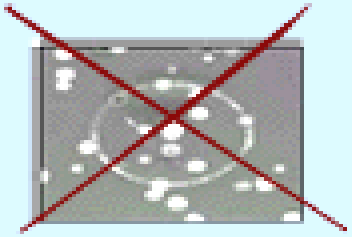
DNA library preparation

One sample preparation per genome

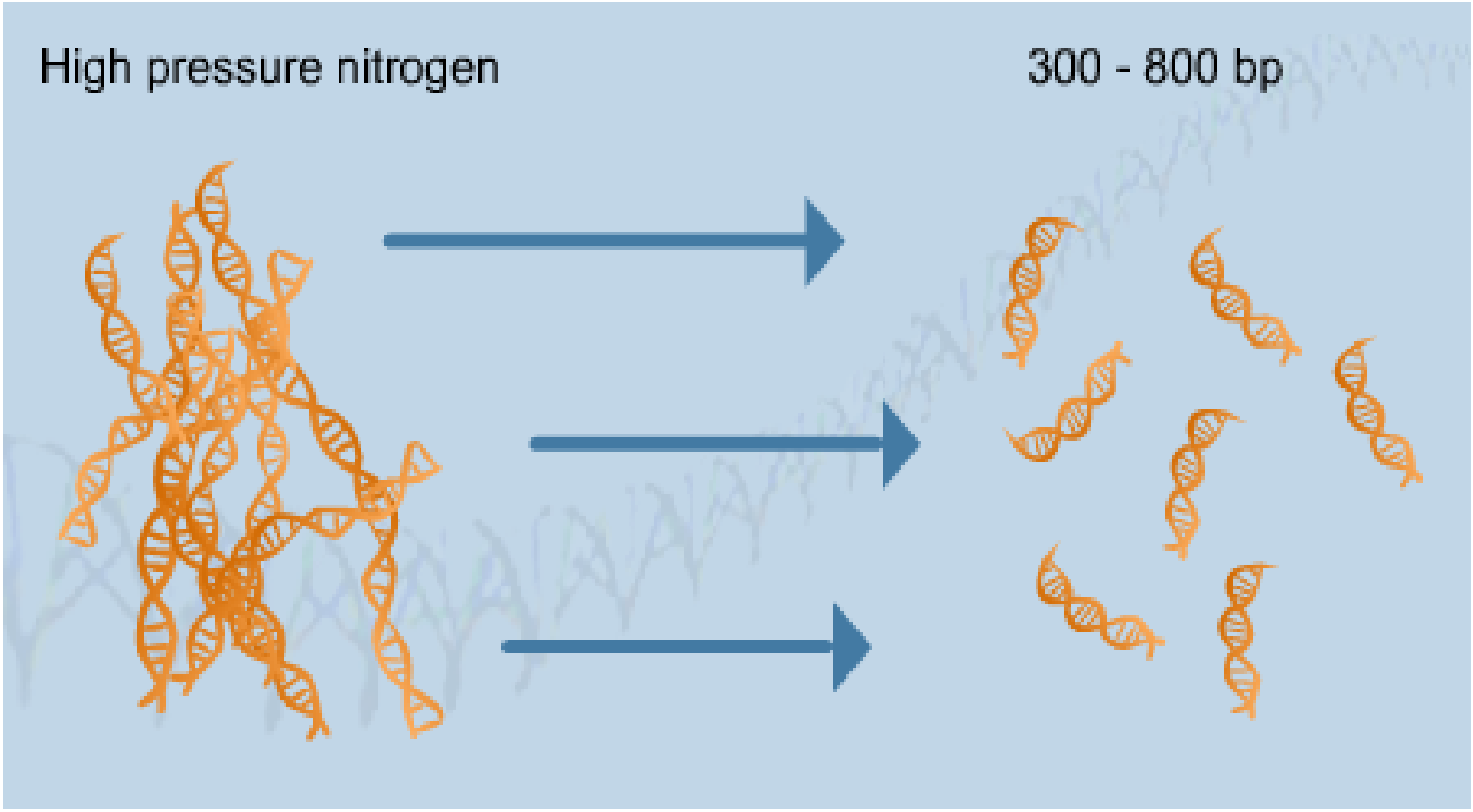
No Cloning



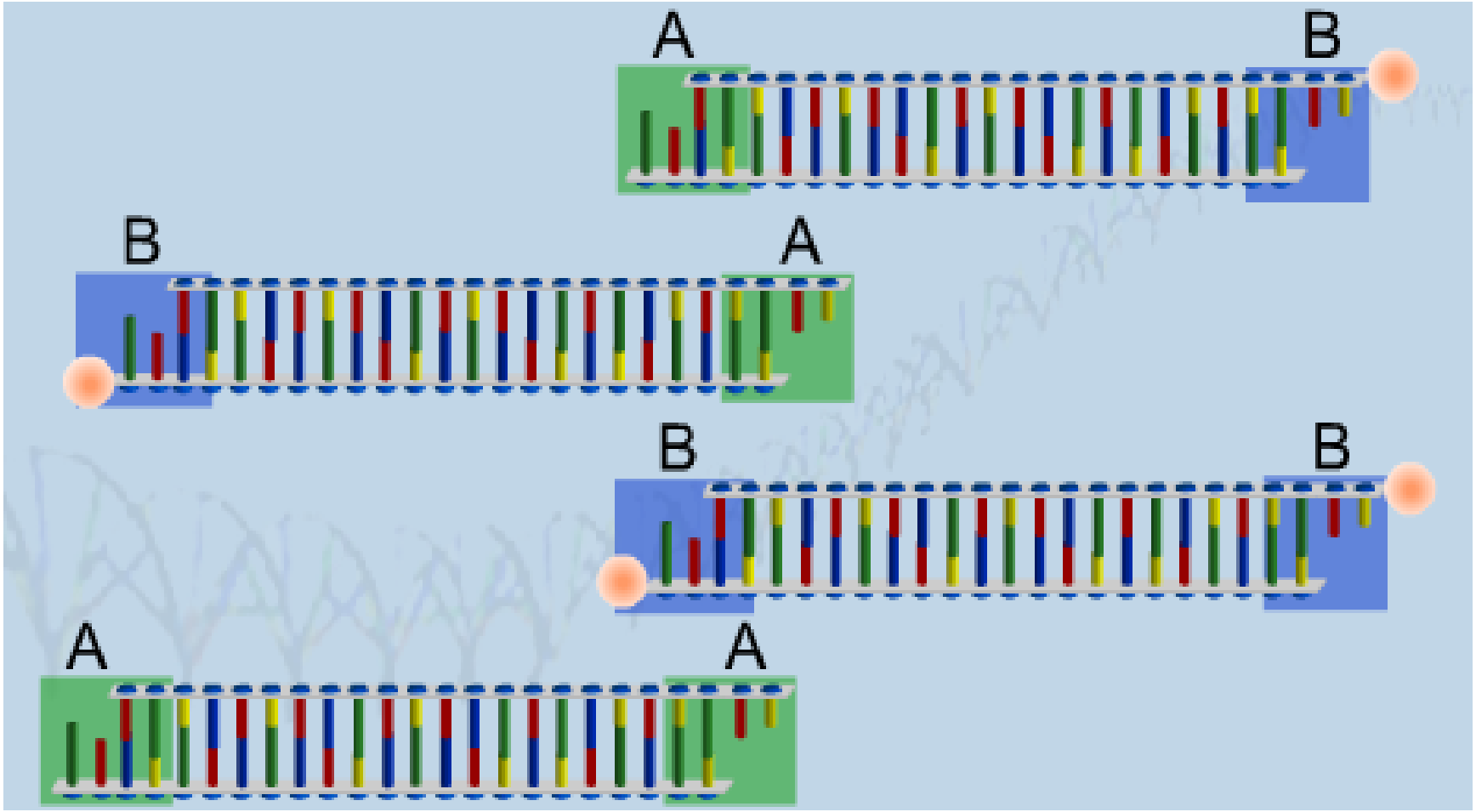
No Colony Picking



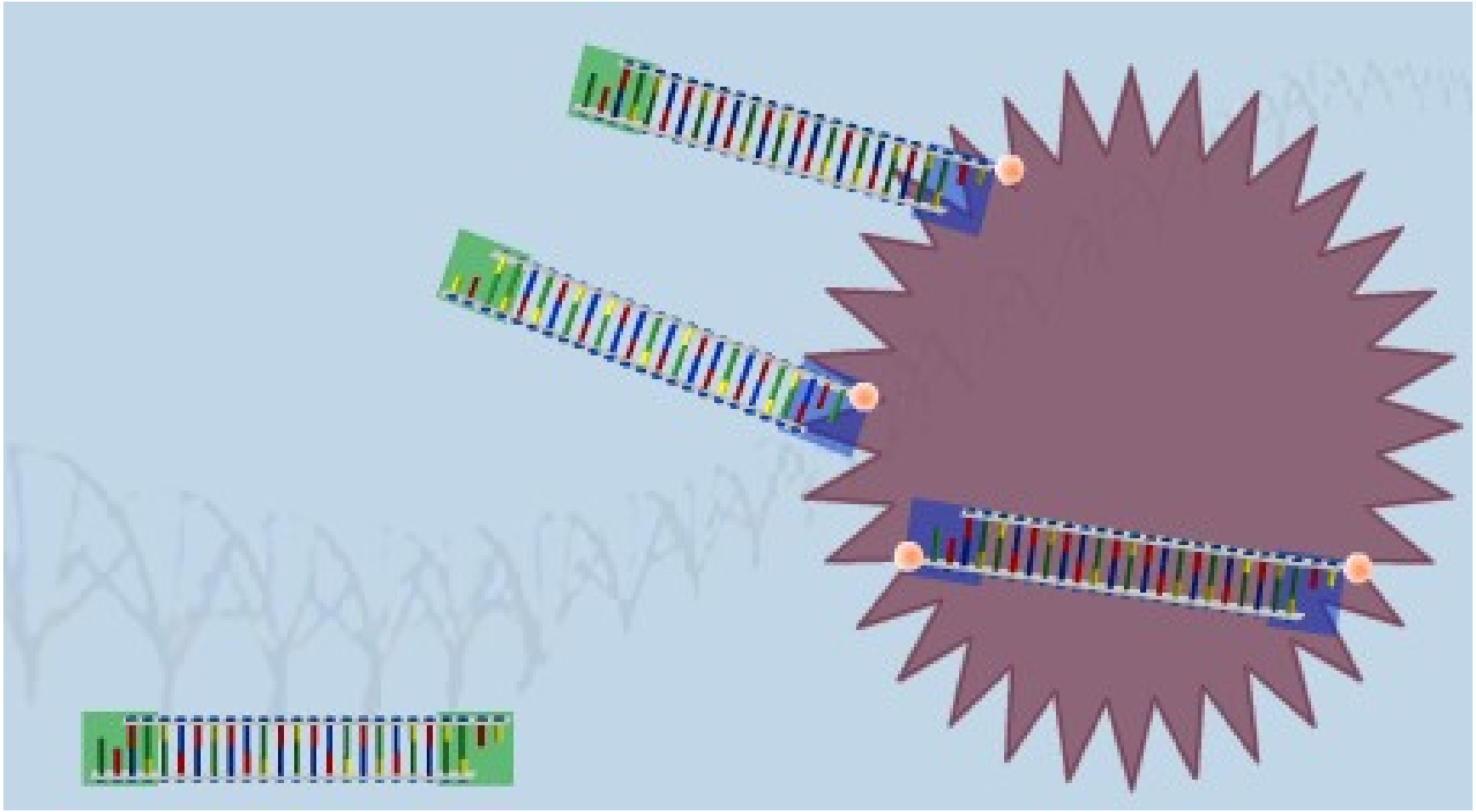
DNA fragmentation



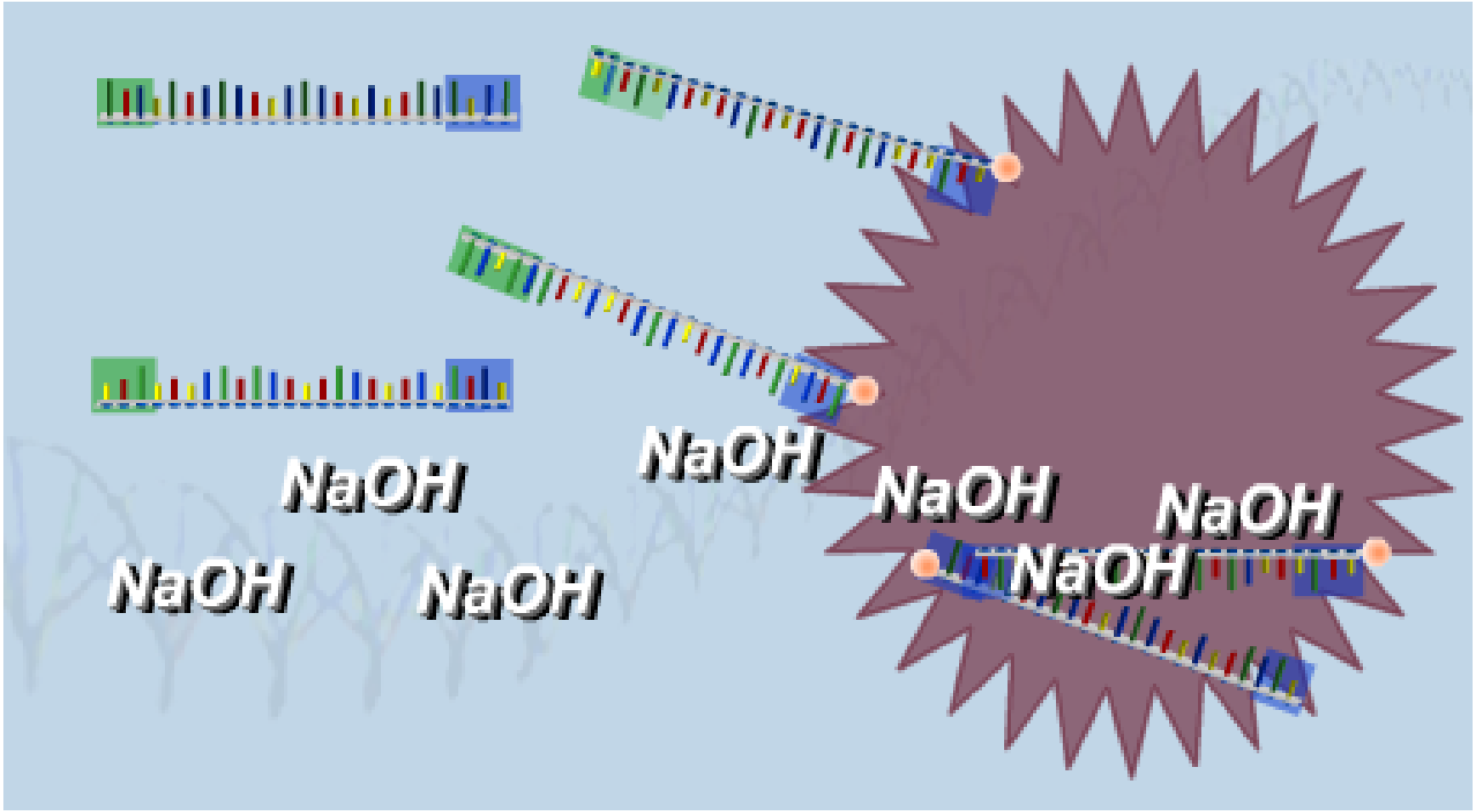
adaptor ligation

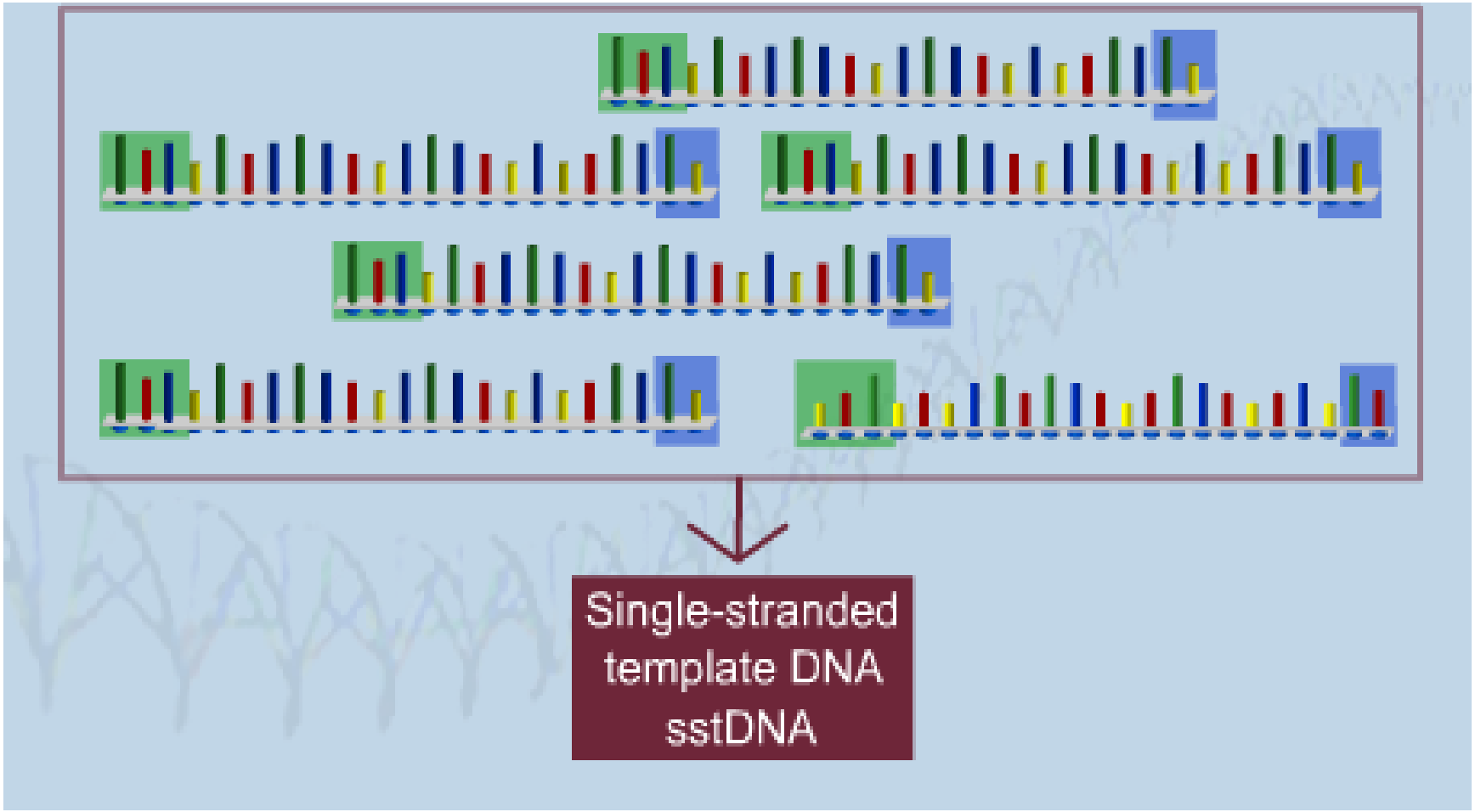


DNA capture

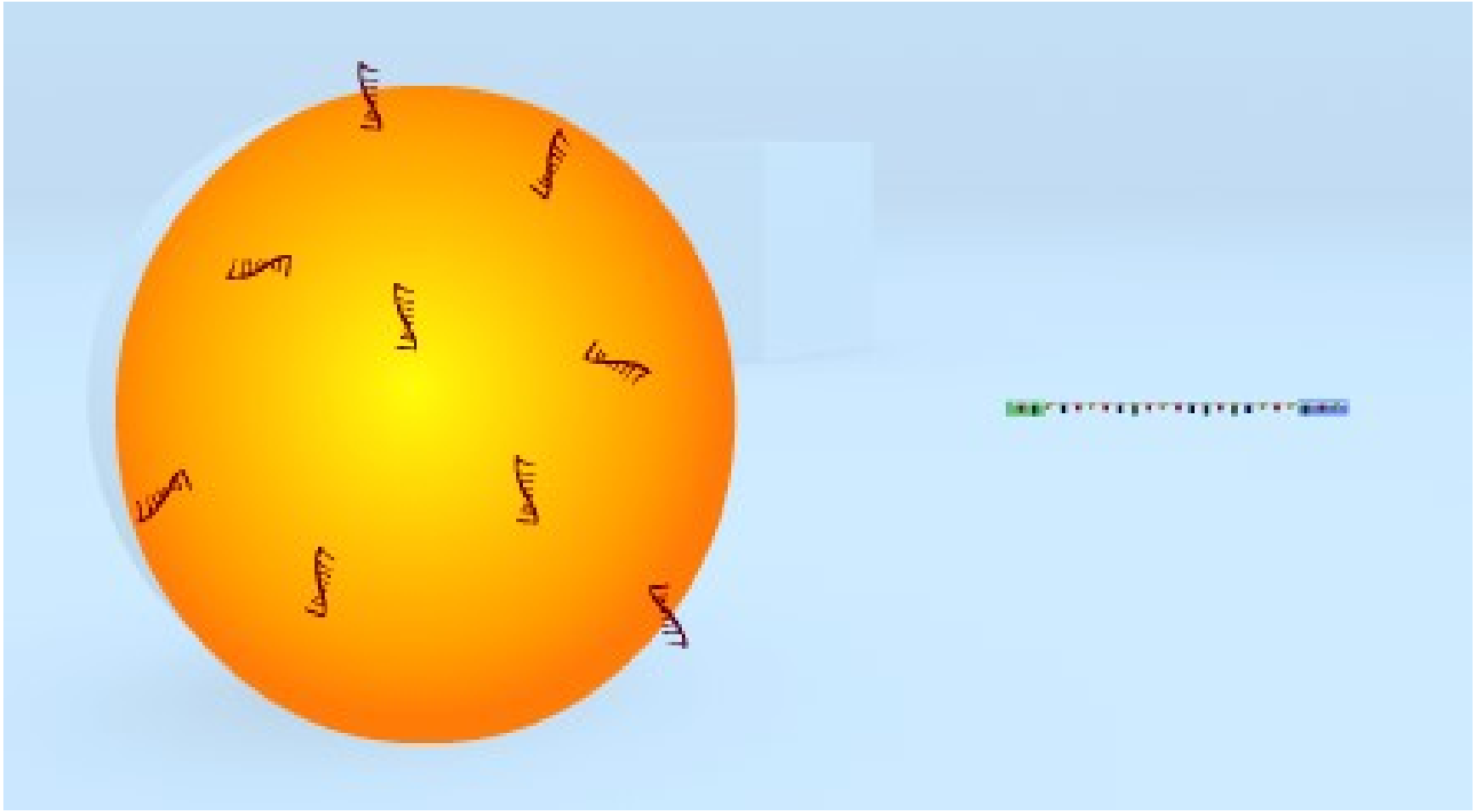


denaturation

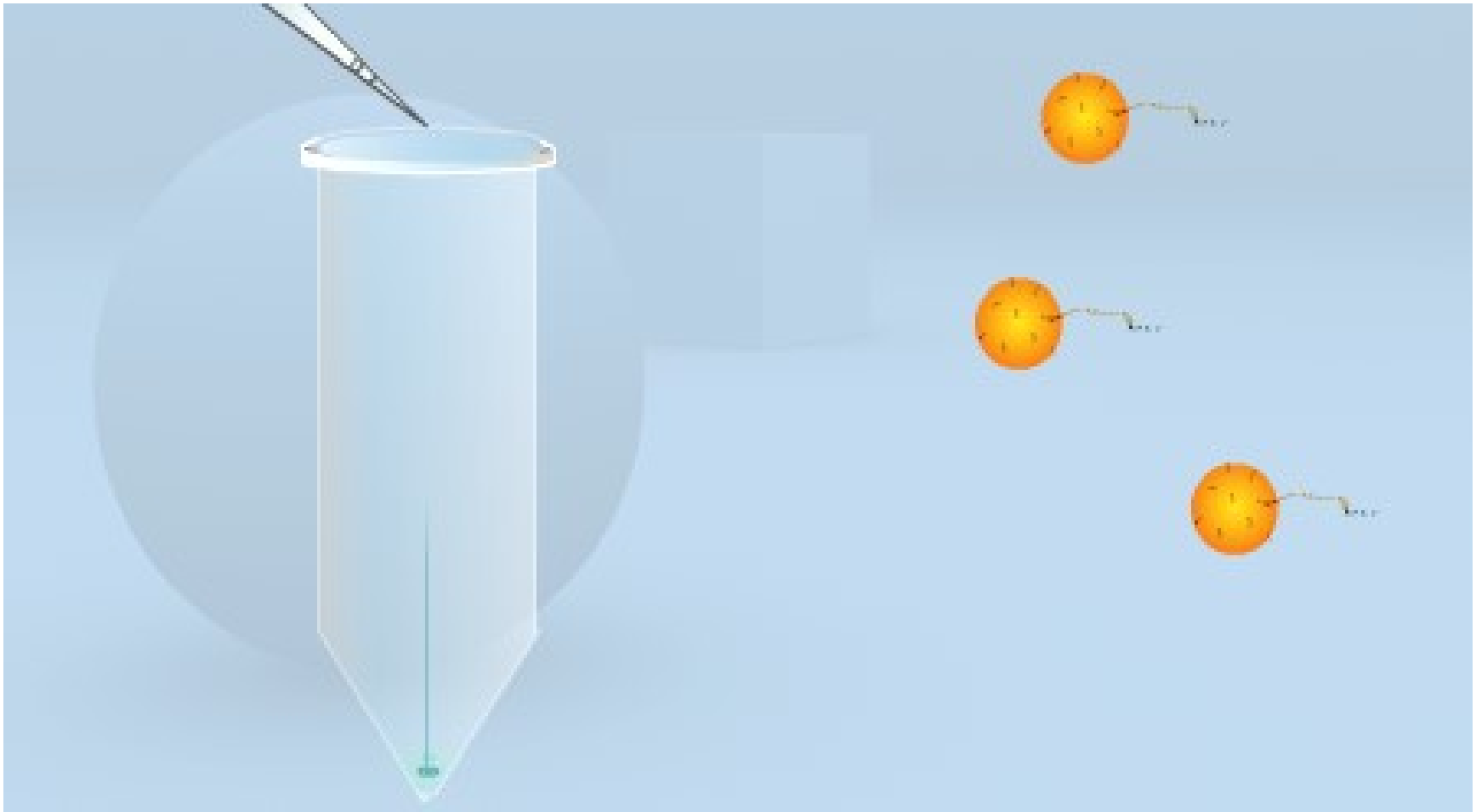



Single-stranded
template DNA
sstDNA

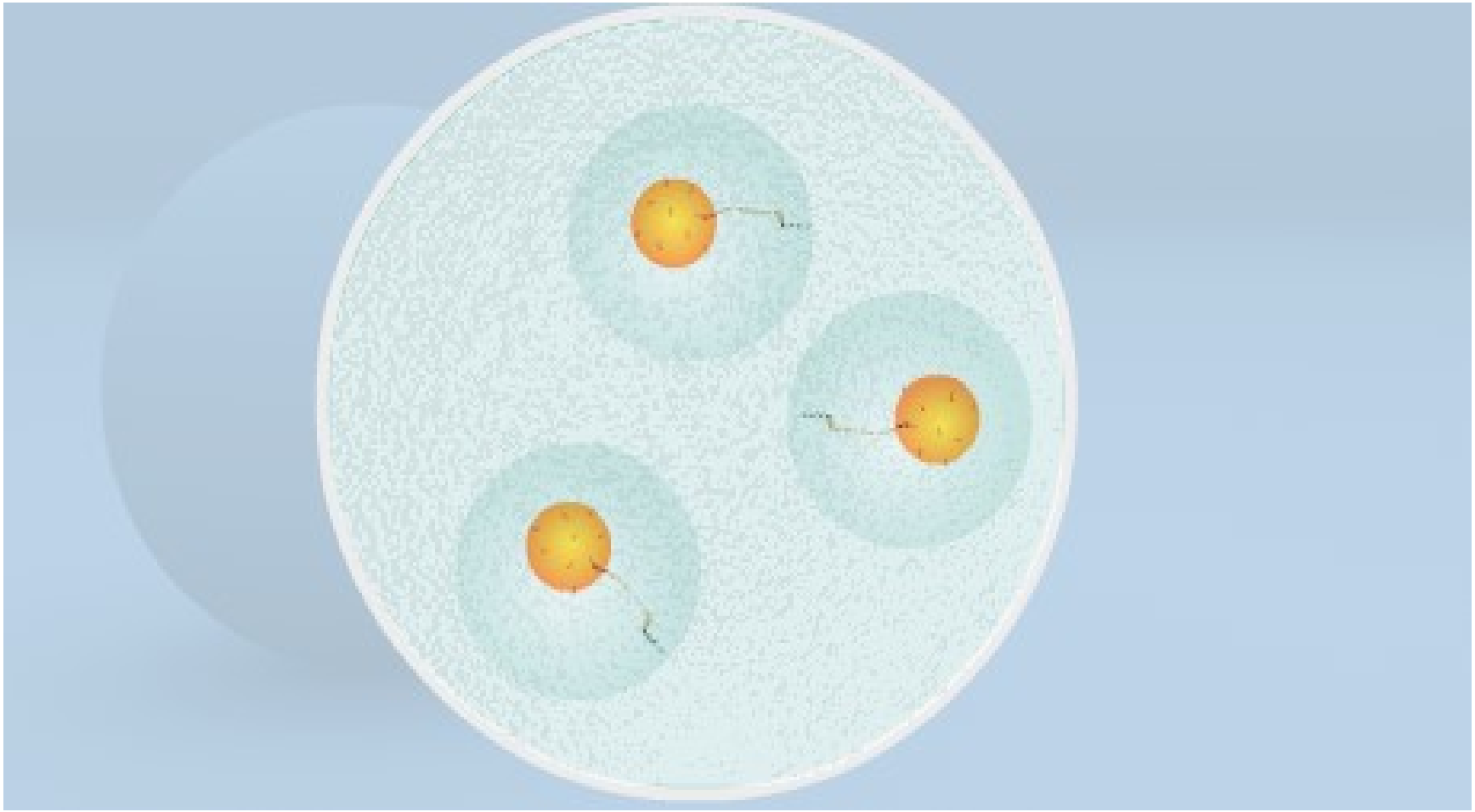
emPCR



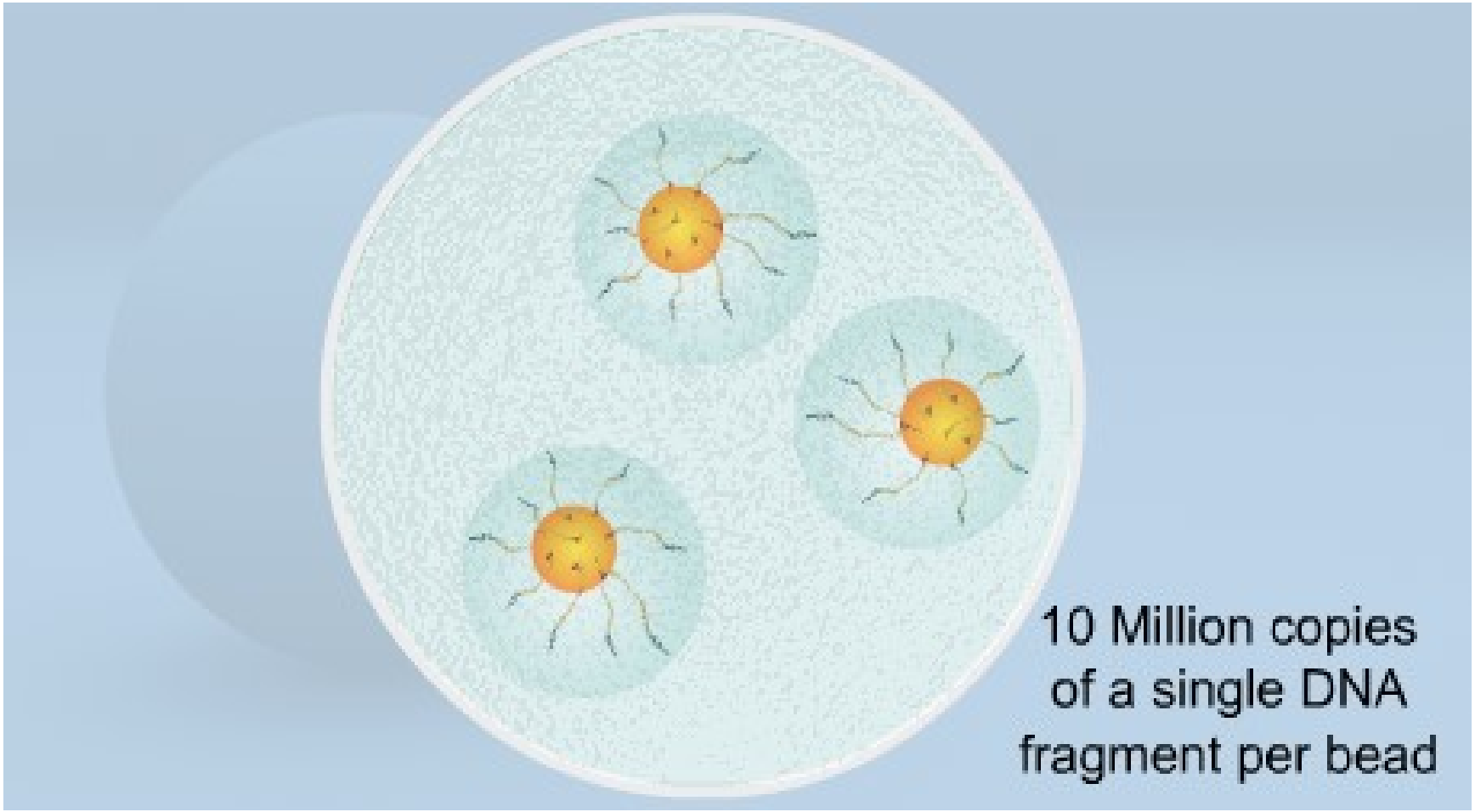
emulsion



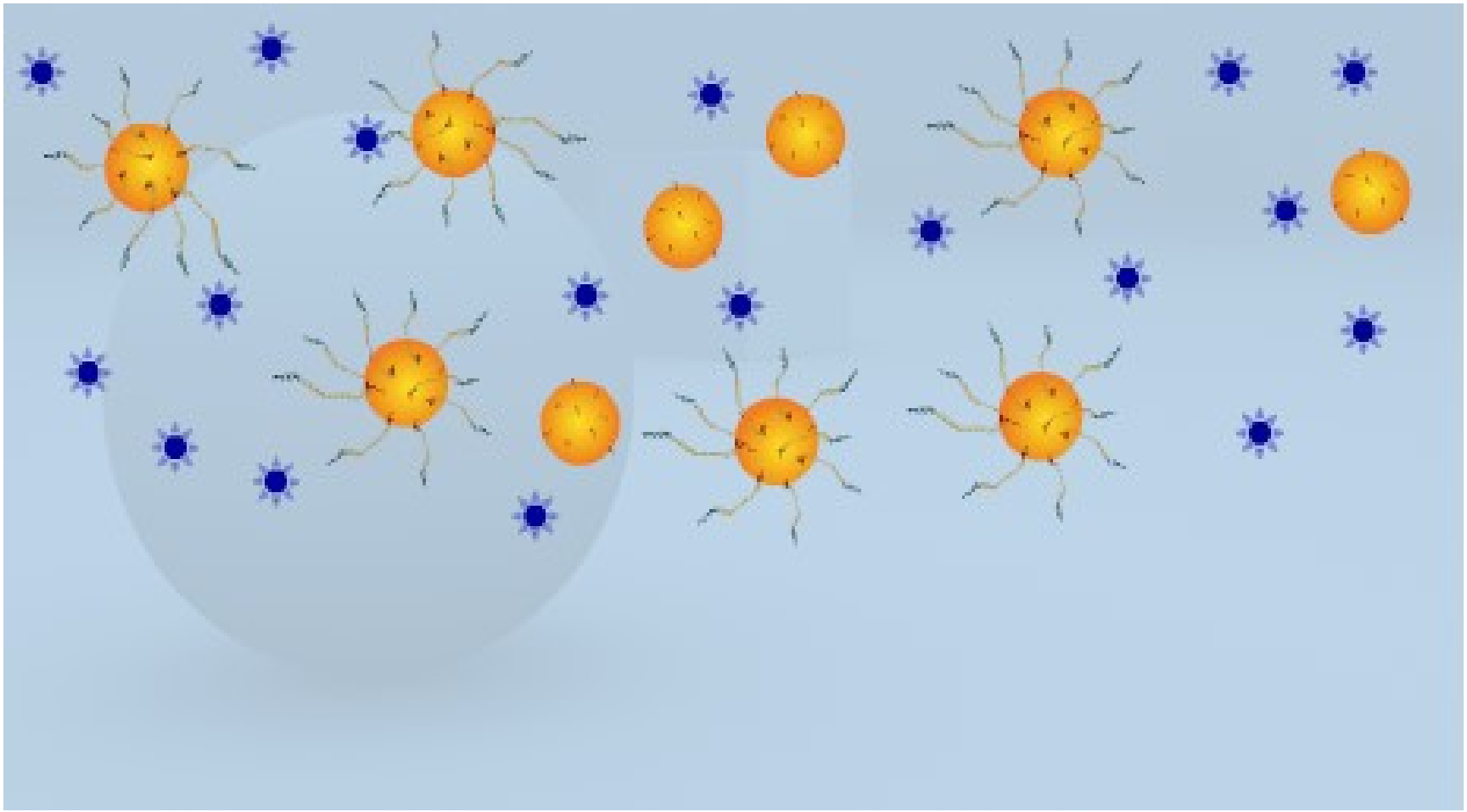
emPCR



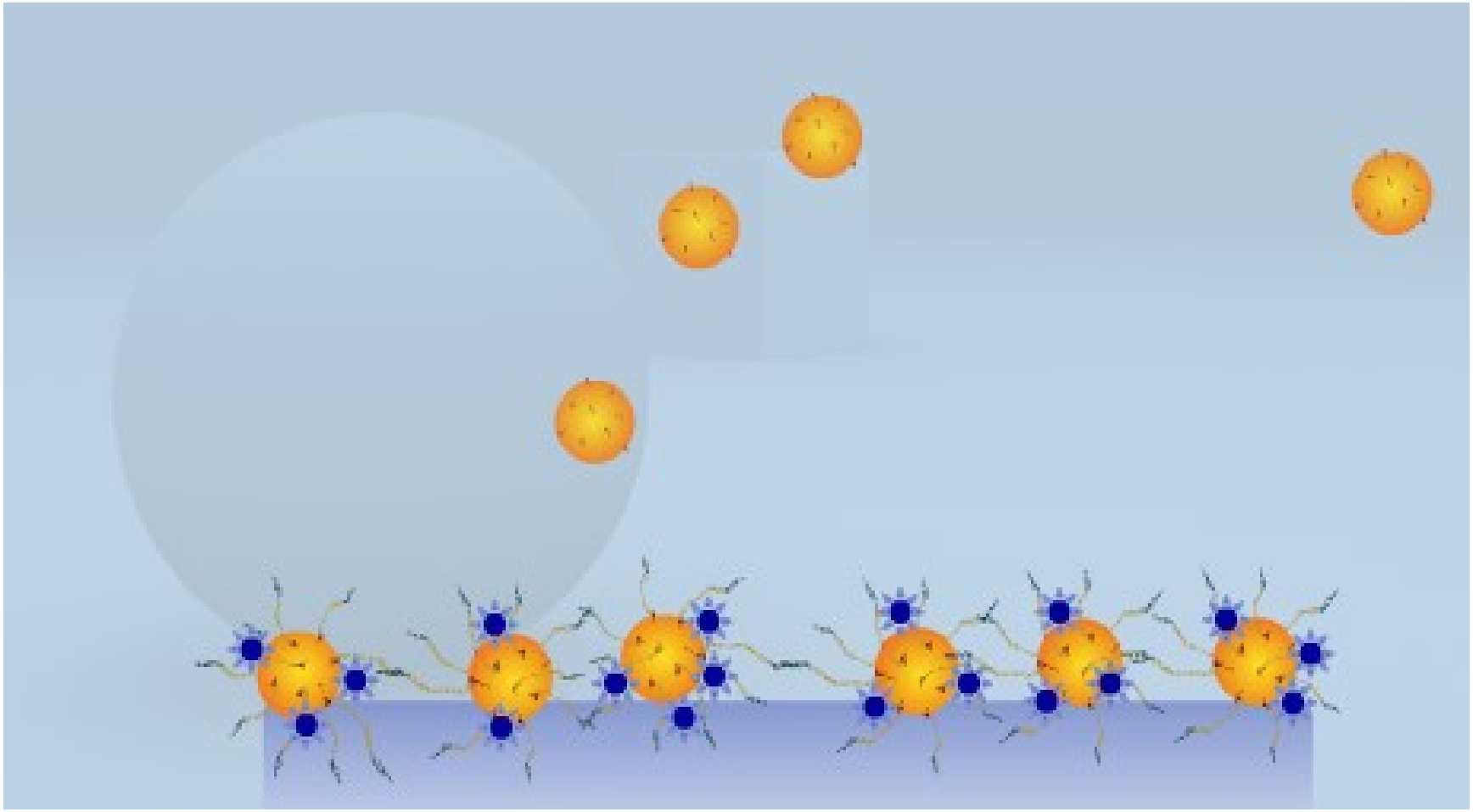
emPCR



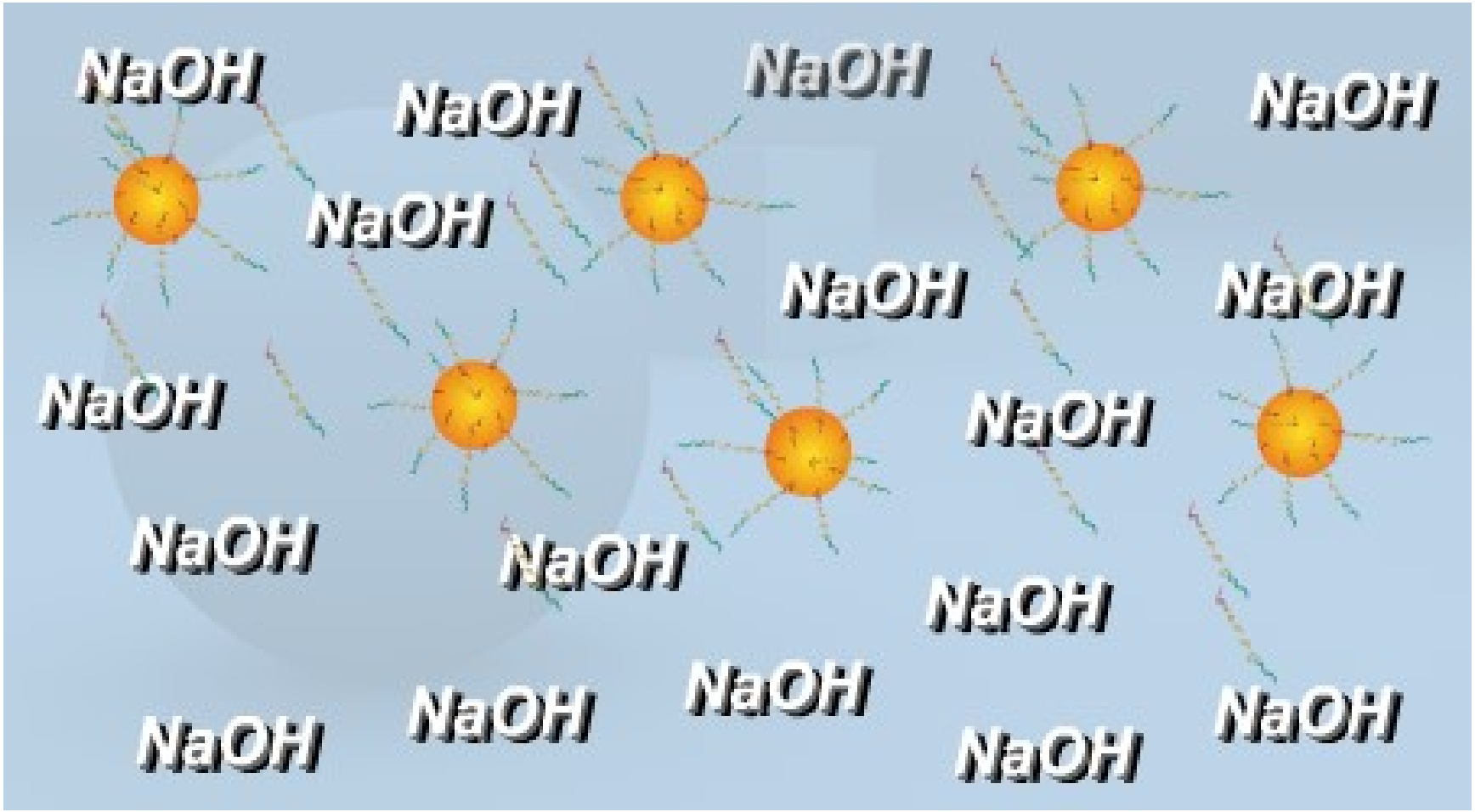
Bead capture



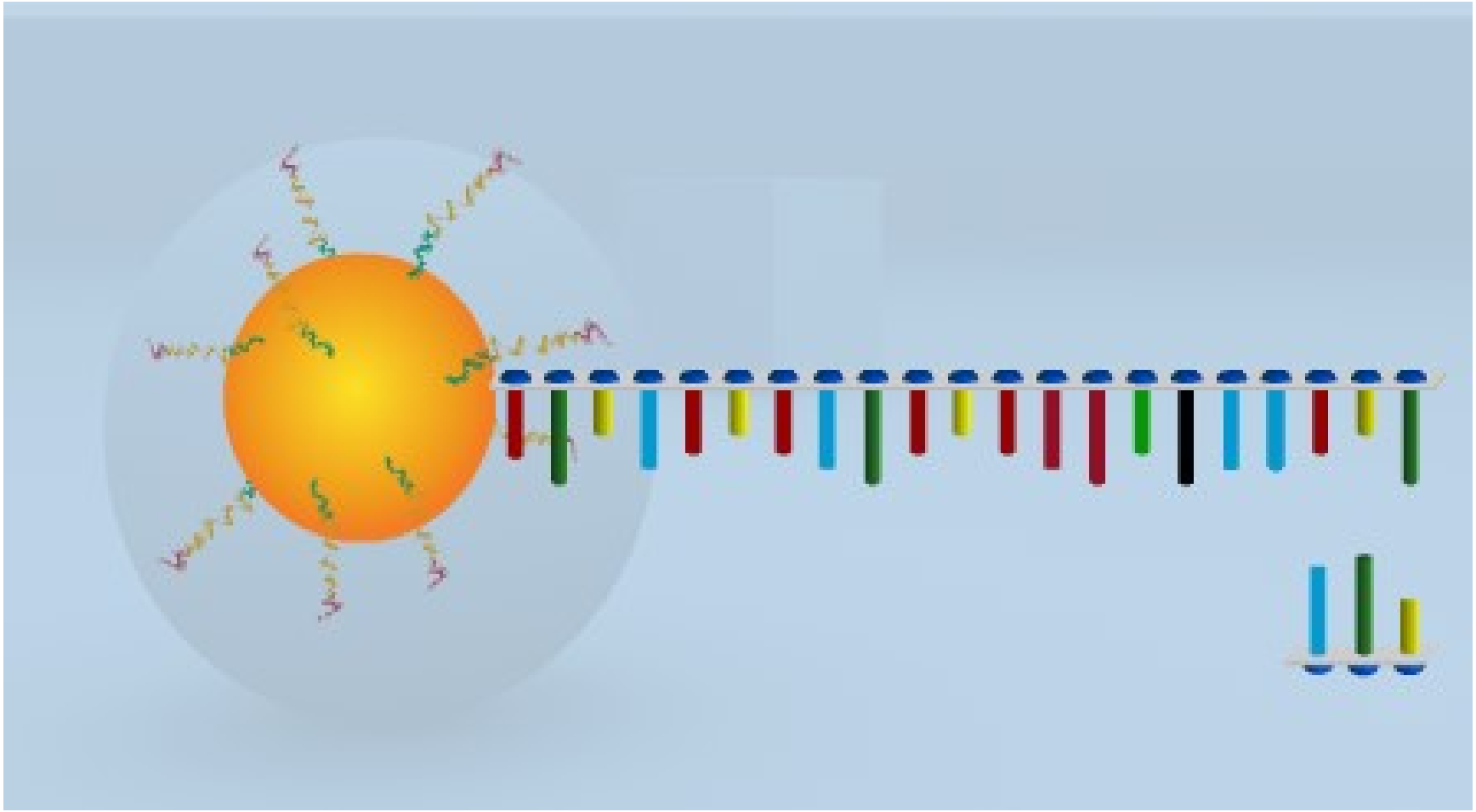
Bead capture



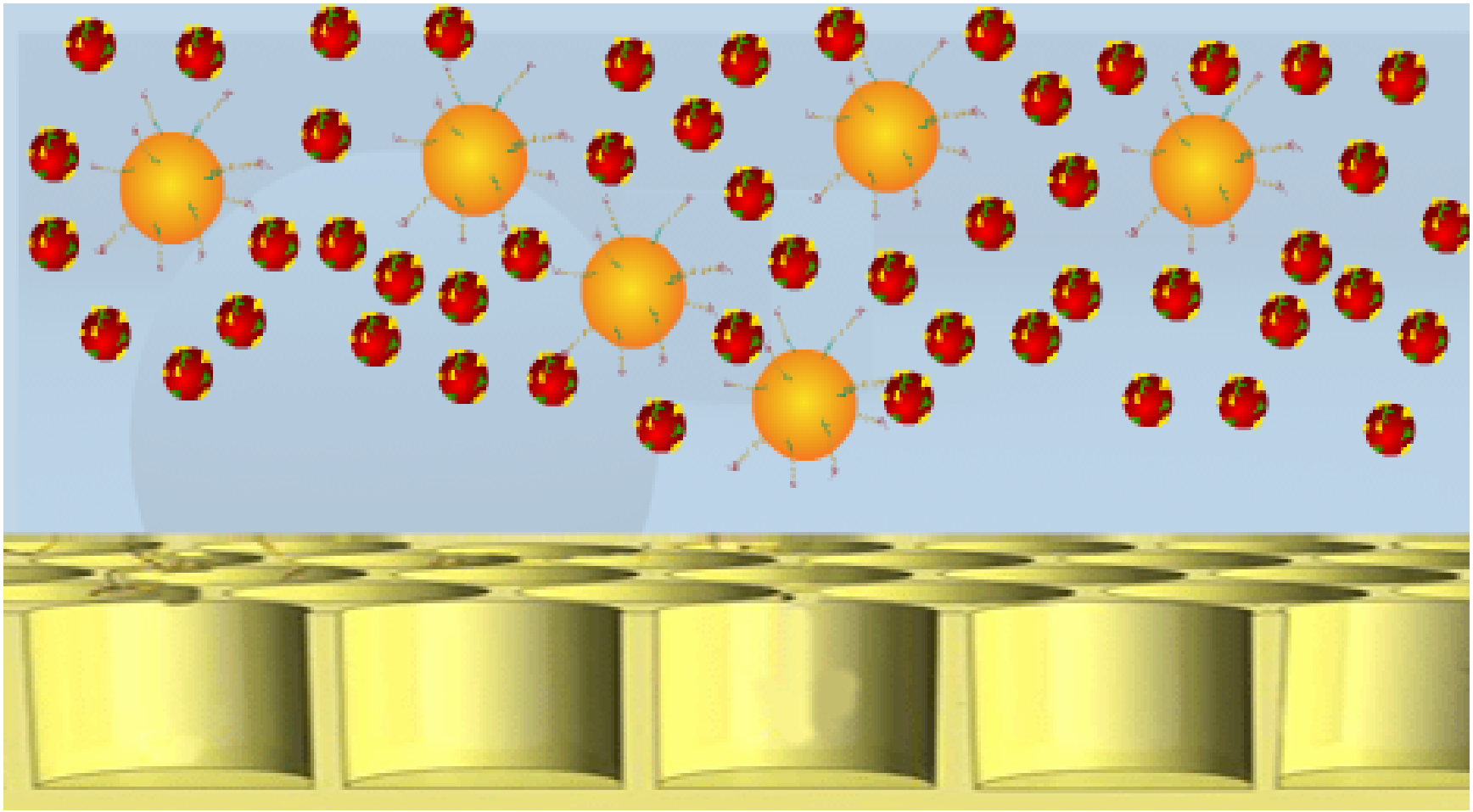
denaturation



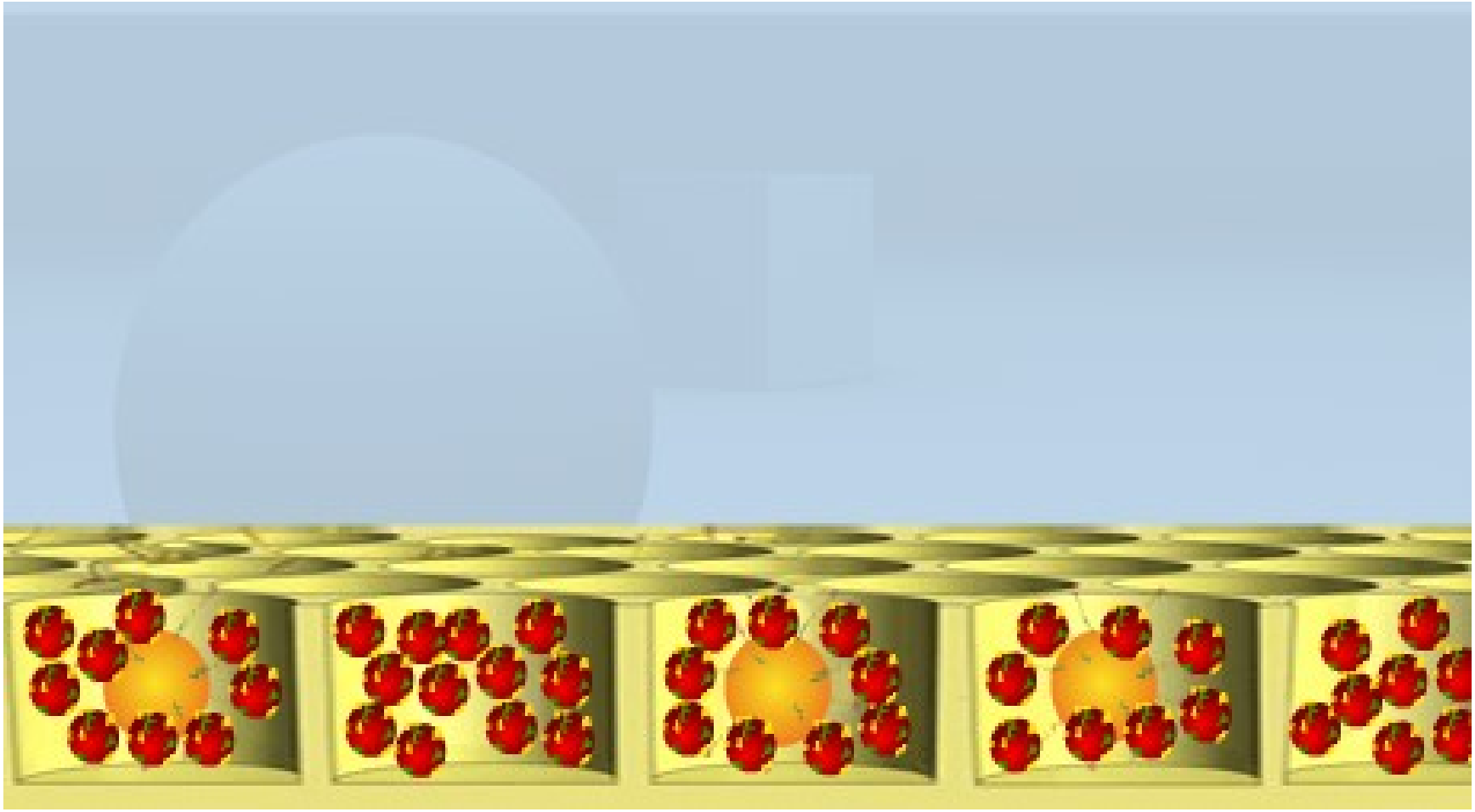
Sequencing primer



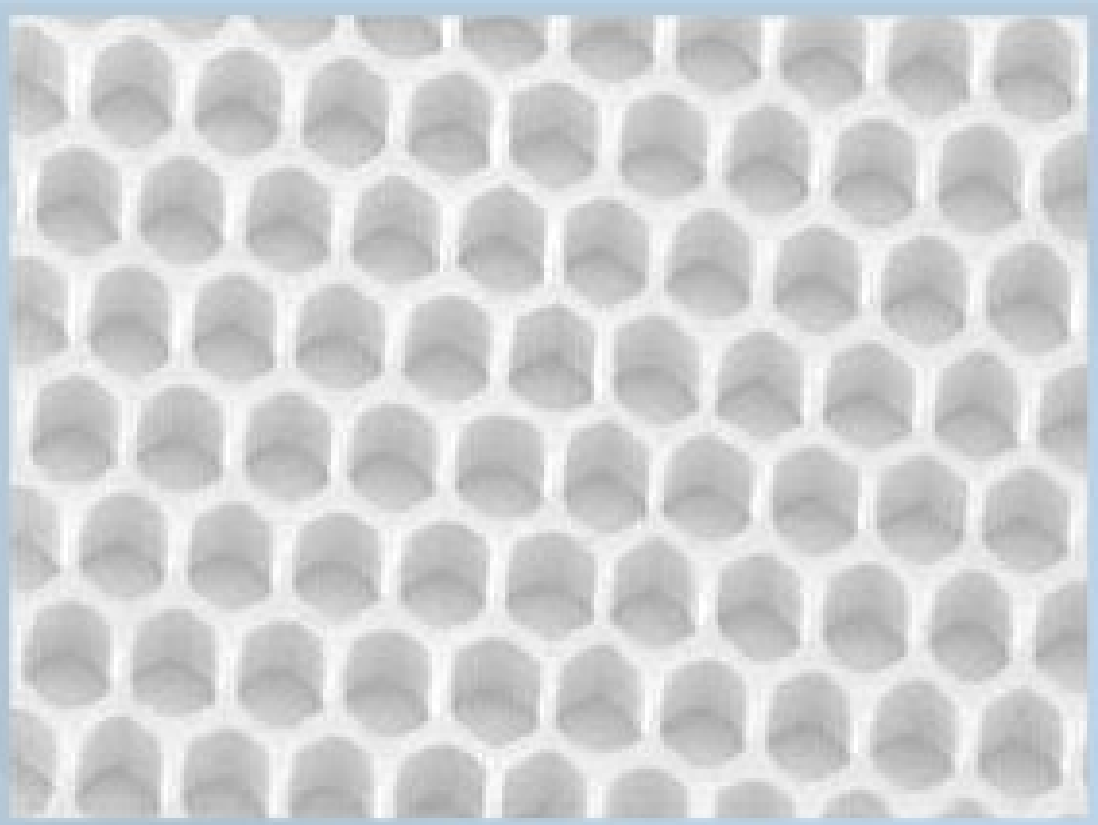
Dispersion



Dispersion



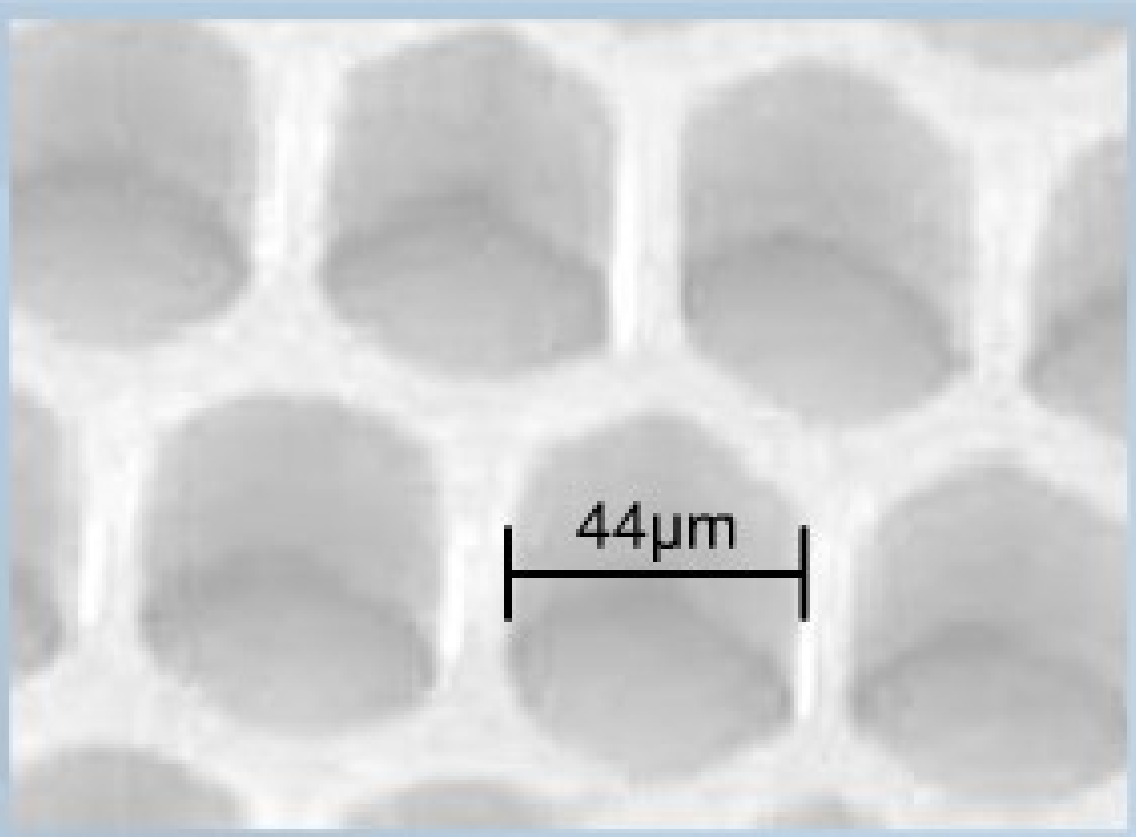
Microwells



1.6 million wells

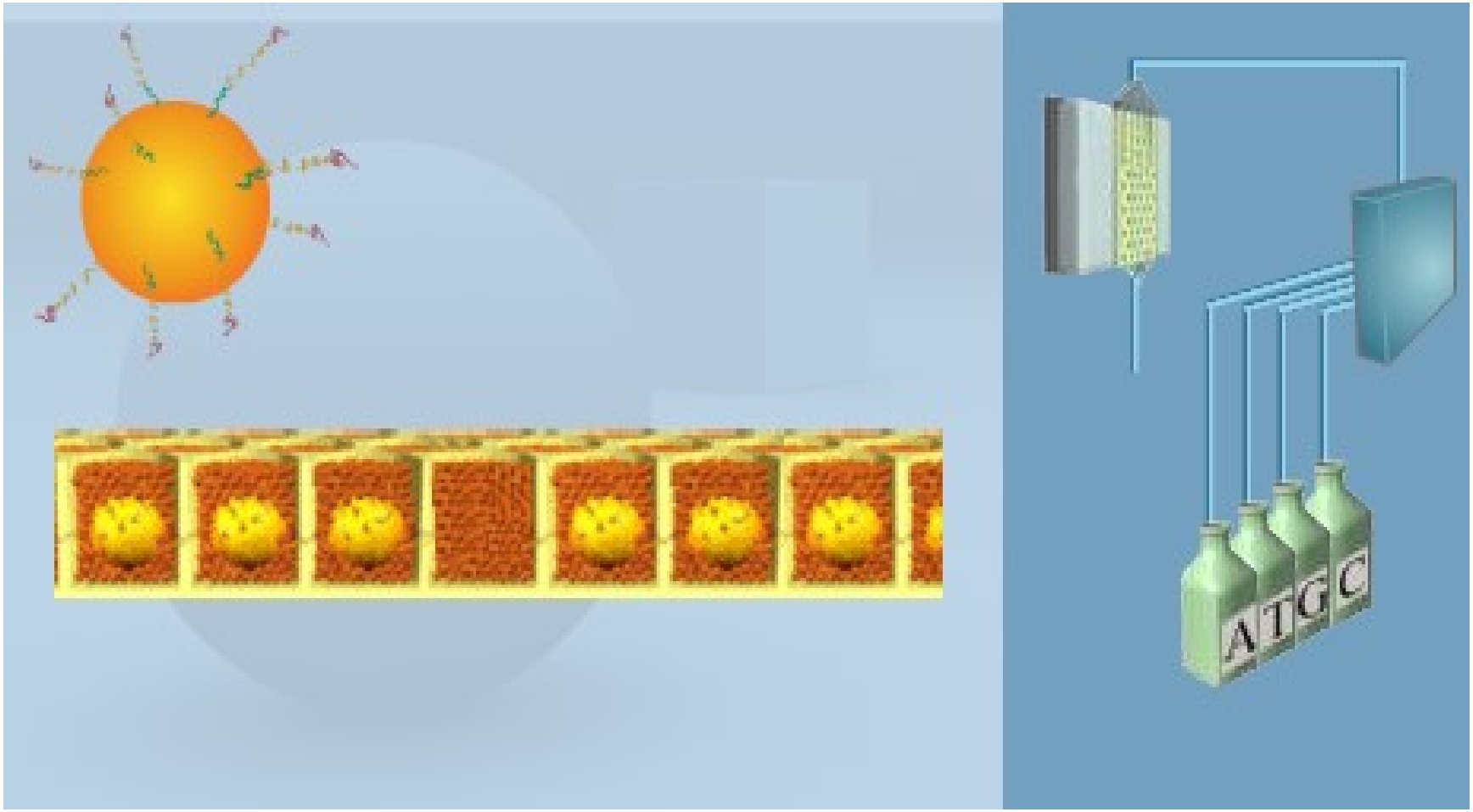
PicoTiterPlate device

Parameters of microreactors

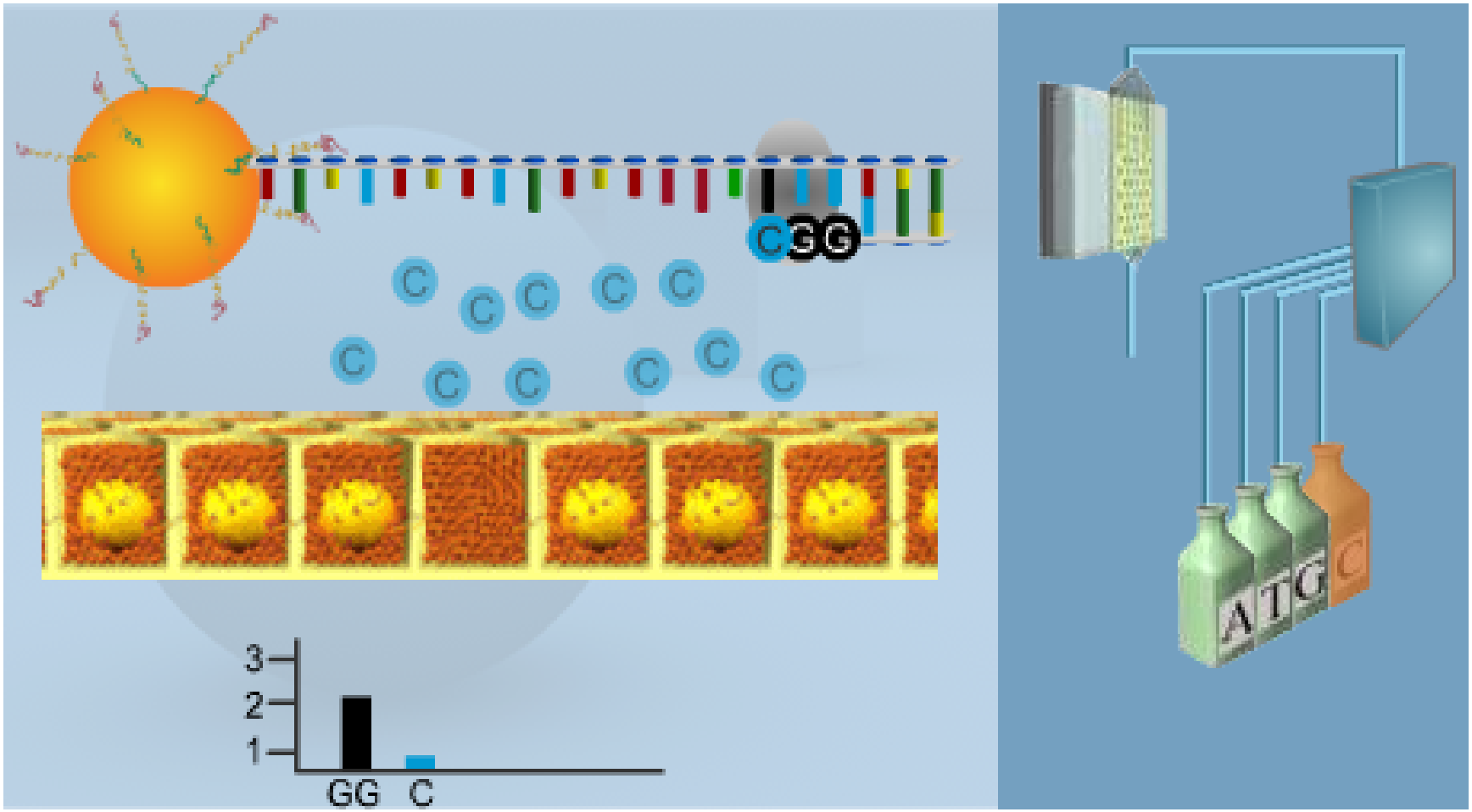


PicoTiterPlate device

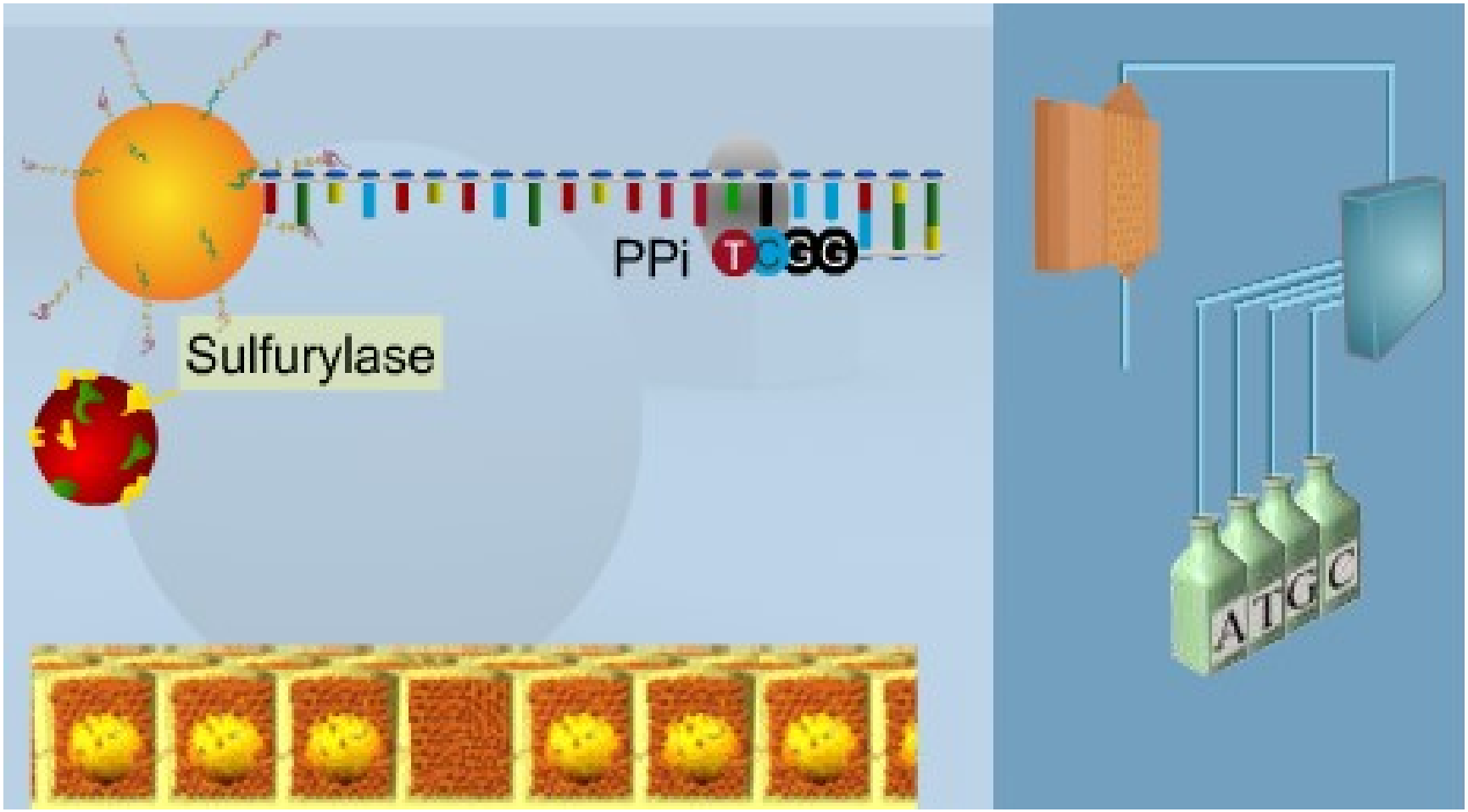
Sequencing



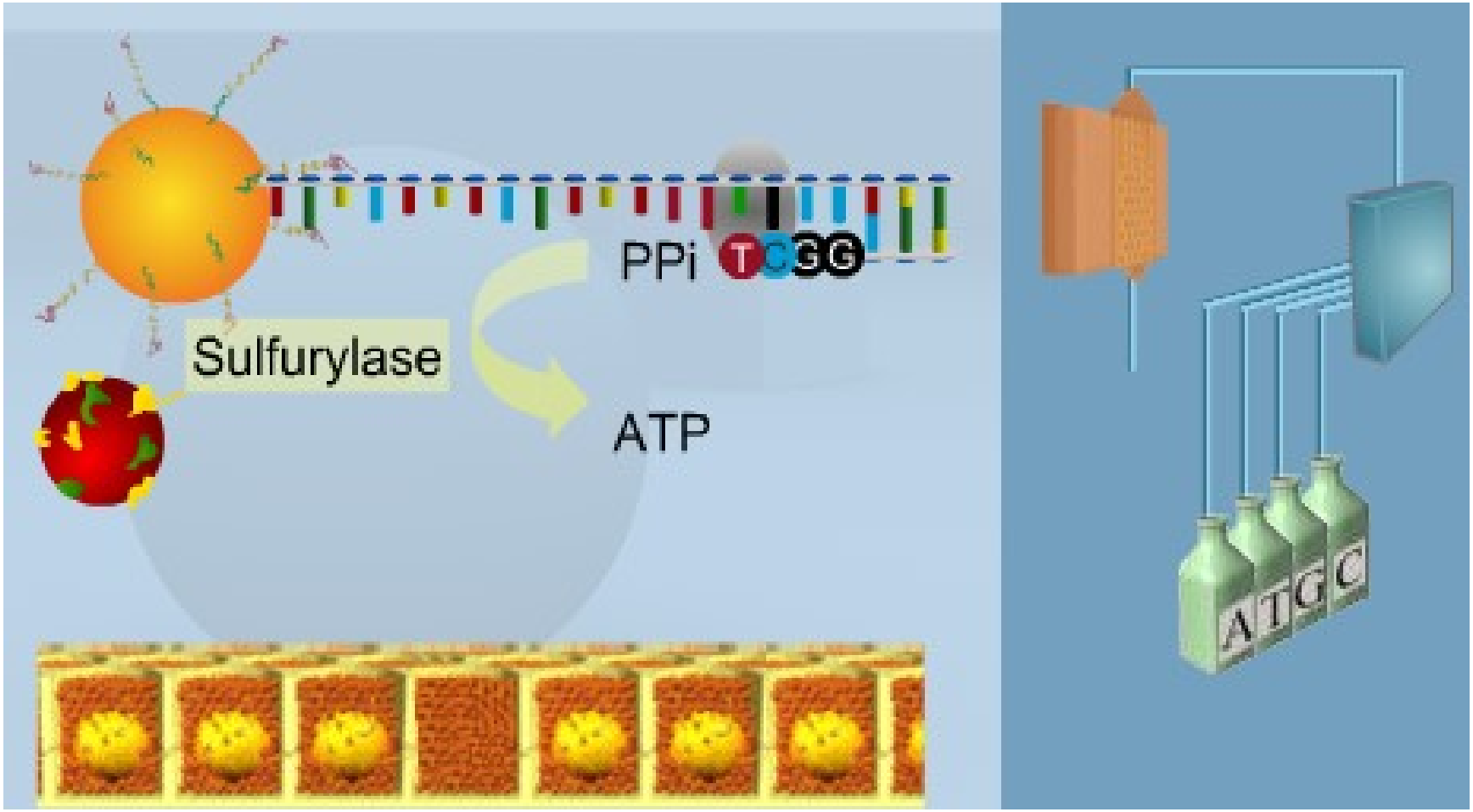
Sequencing



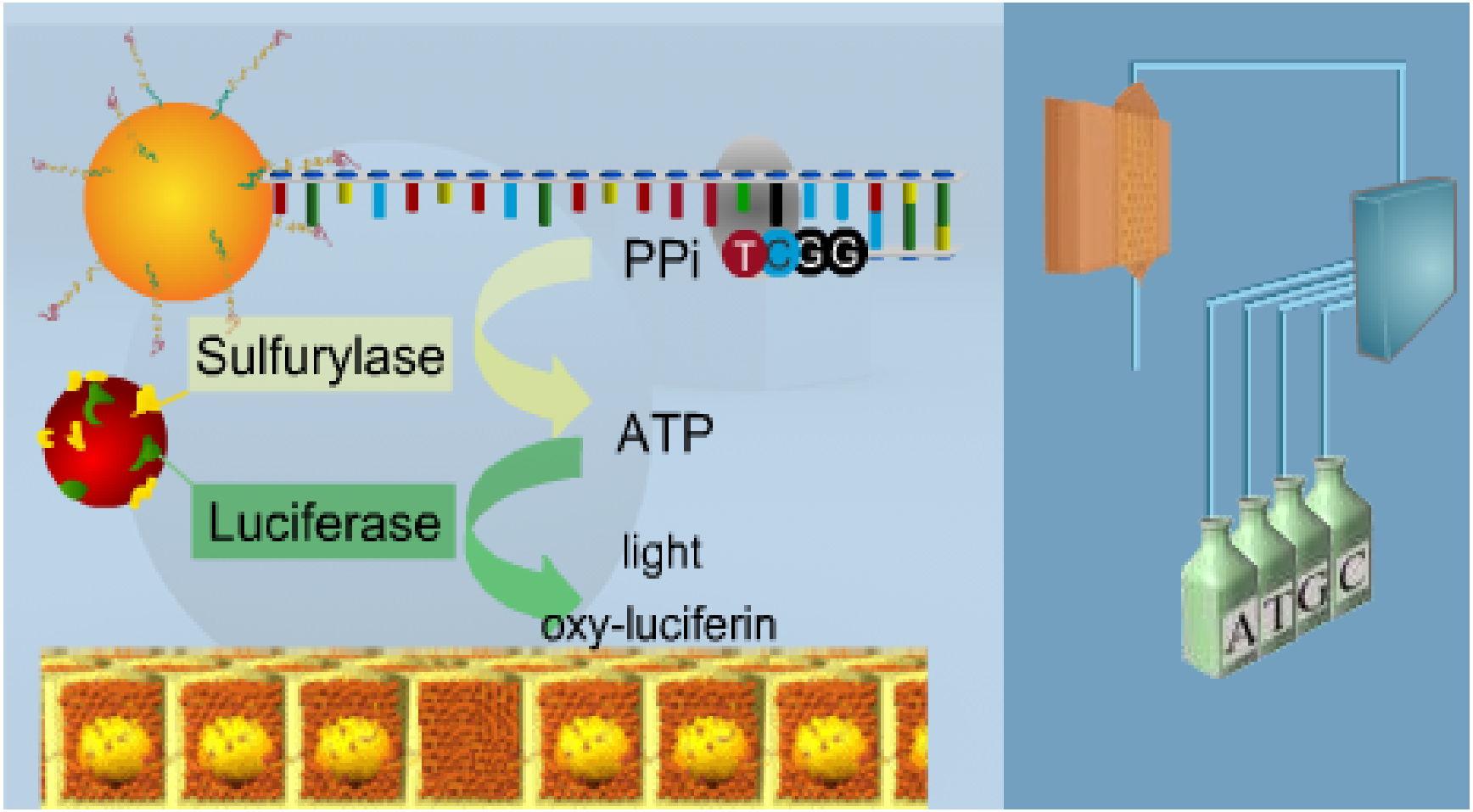
Sequencing



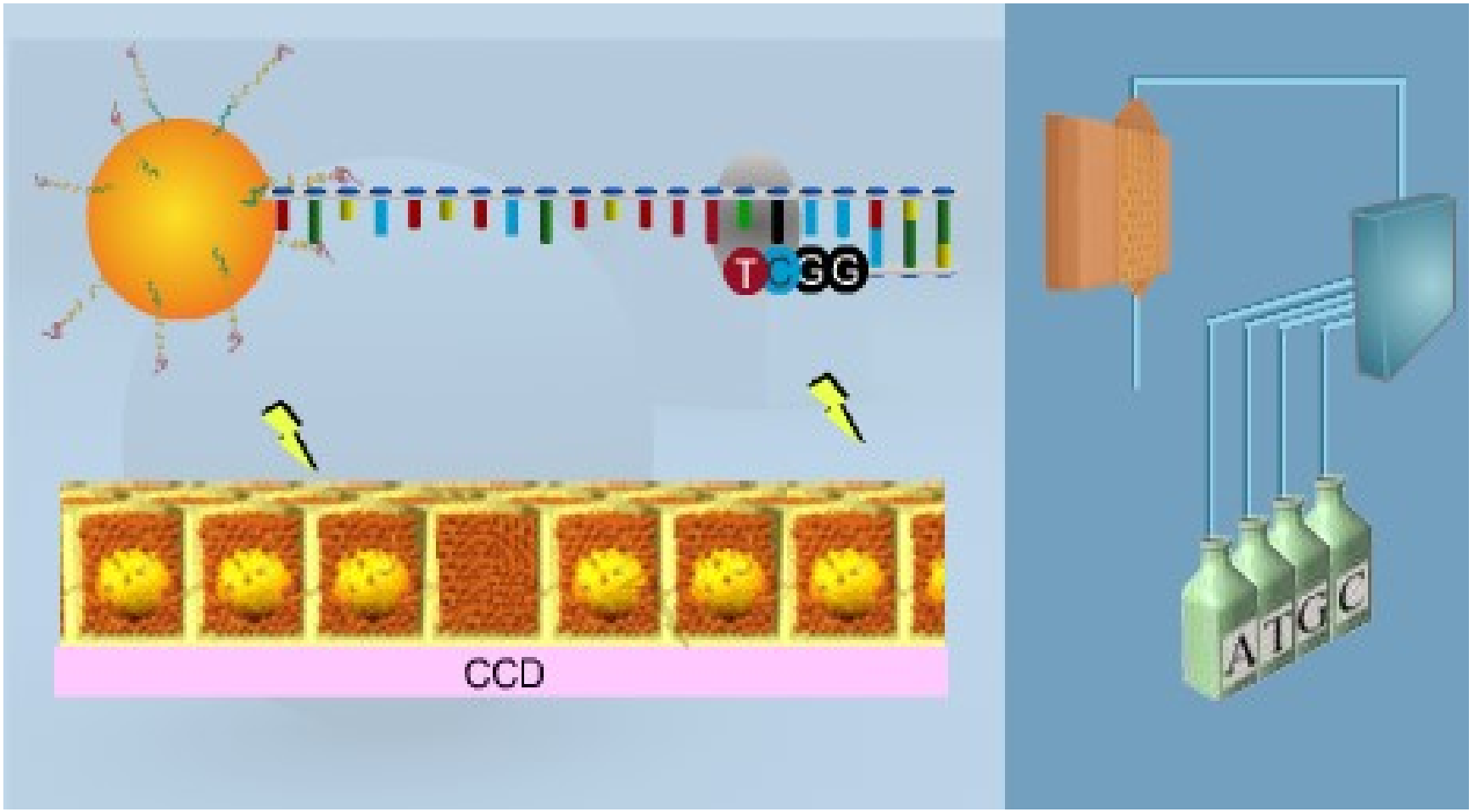
Sequencing



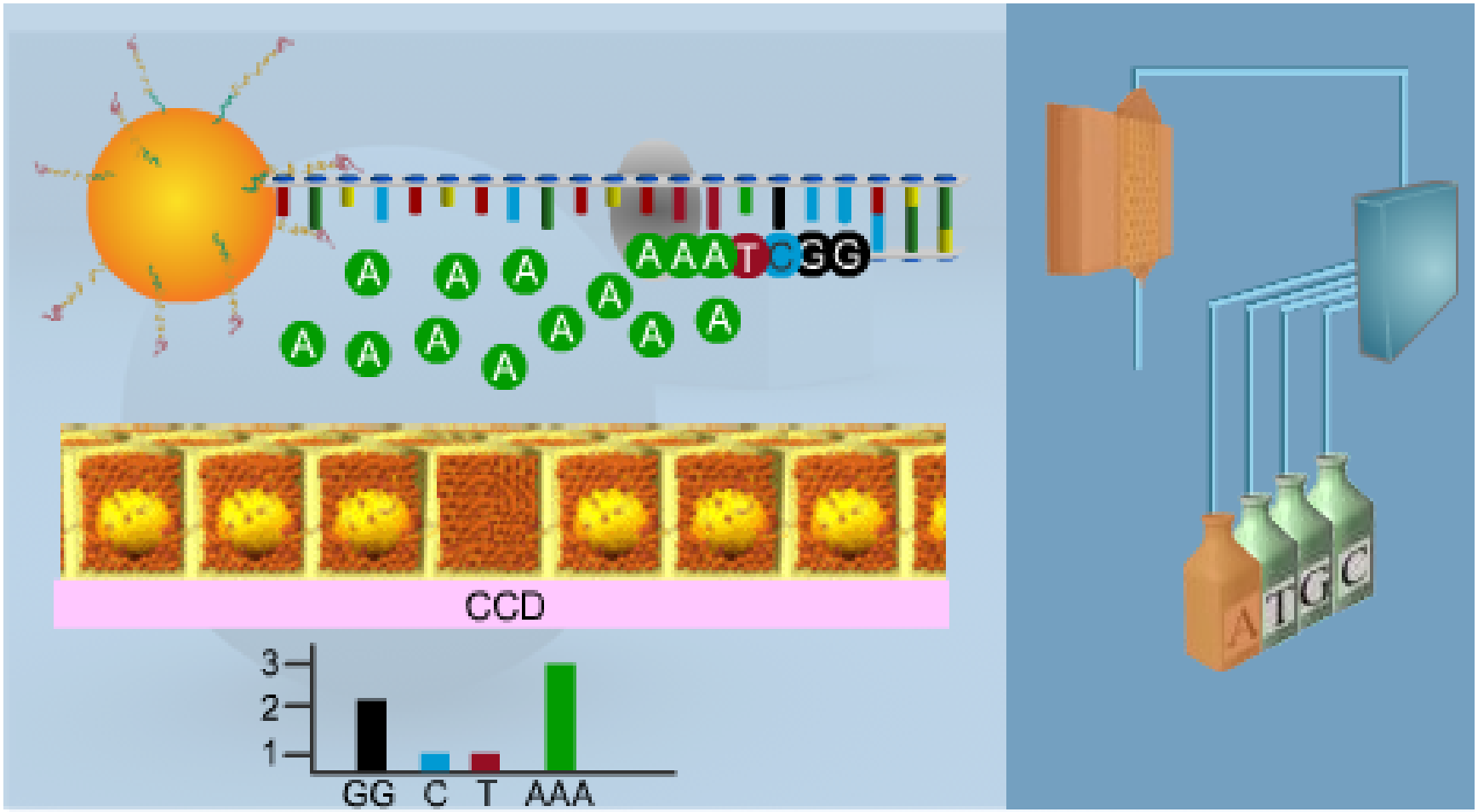
Sequencing


PPi
T C G G
Sulfurylase
ATP
Luciferase
light
oxy-luciferin
A T G C

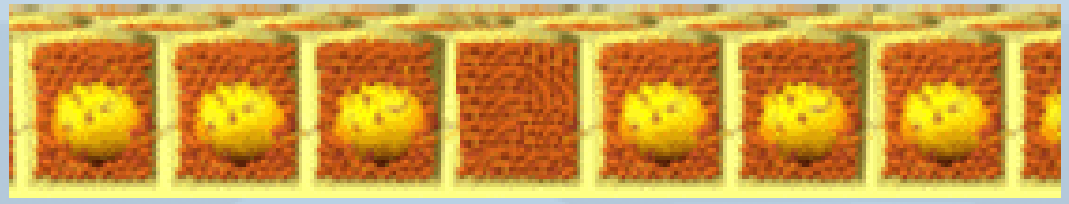
Sequencing



Sequencing



Sequencing



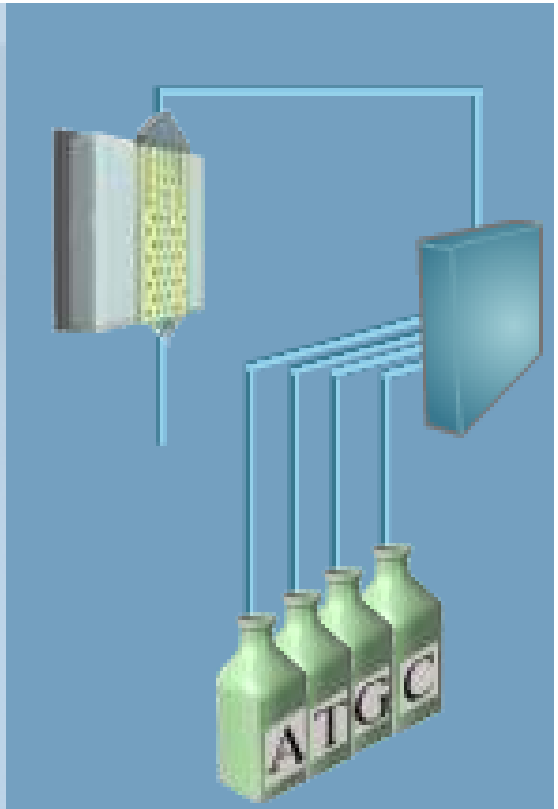
Massive parallelization of sequencing reactions

100 bases read length

X

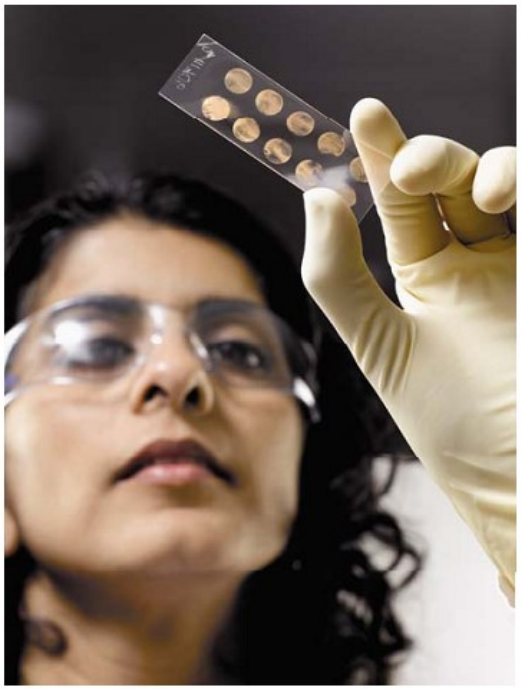
200 000

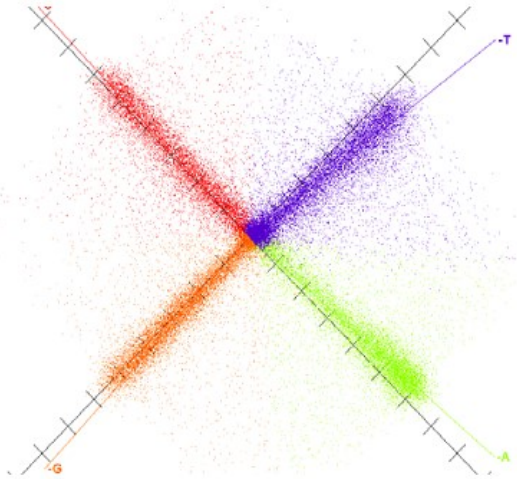
20 Million Bases



SOLID (Sequencing by Oligonucleotide Ligation and Detection)

2-base encoding sequencing (2007)





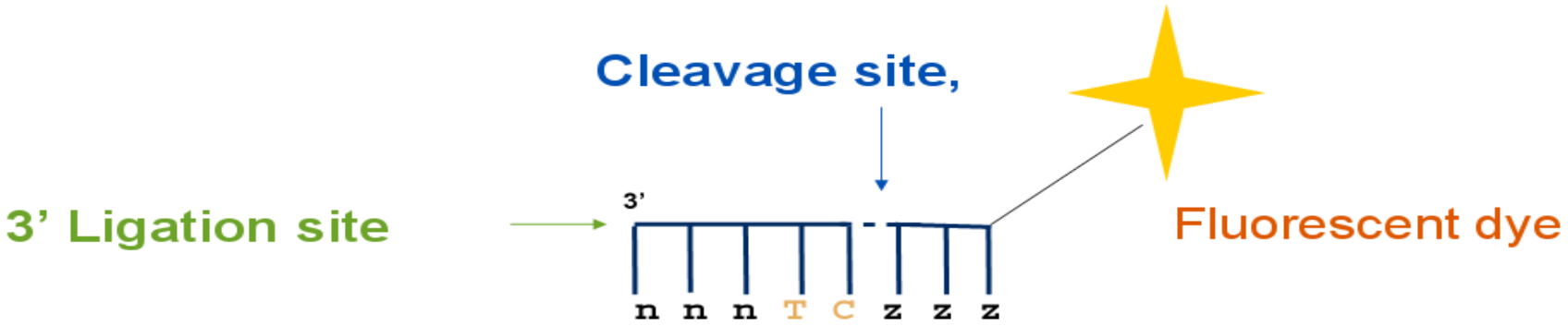
Applied Biosystems

SOLiD™ System

Sequencing by **O**ligonucleotide **Li**gation and **D**etection

Properties of the Probes

Spatial separation among dye, ligation & cleavage sites

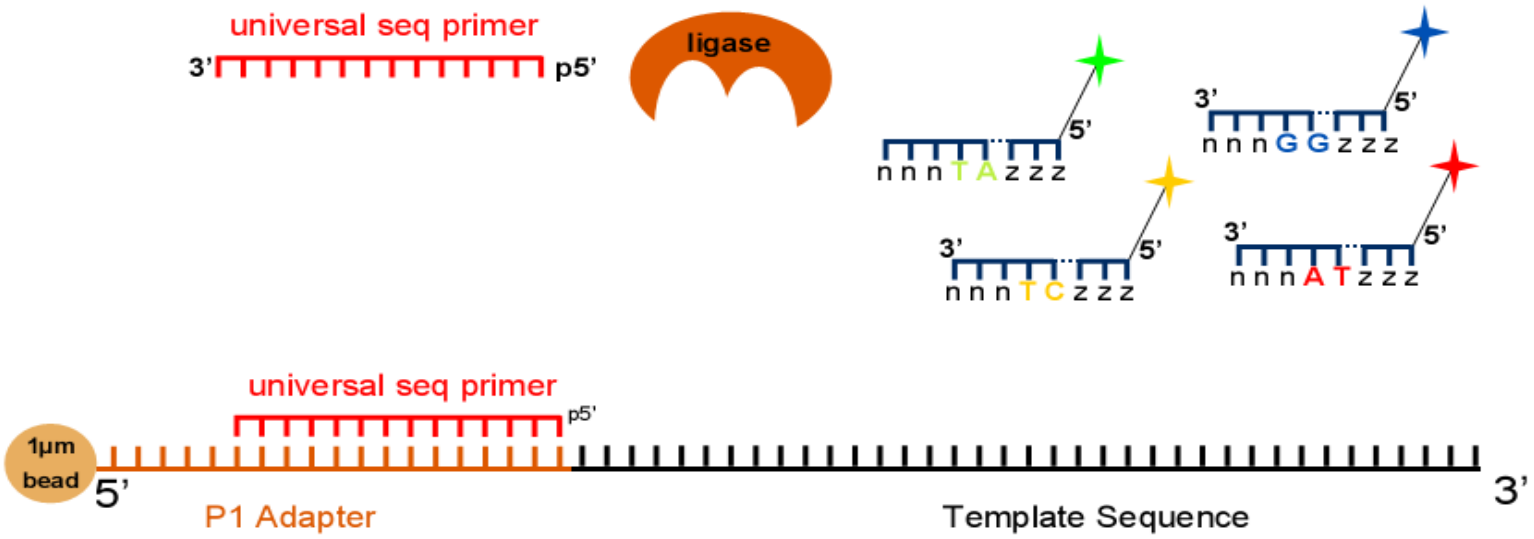


1,024 Octamer Probes (4^5)

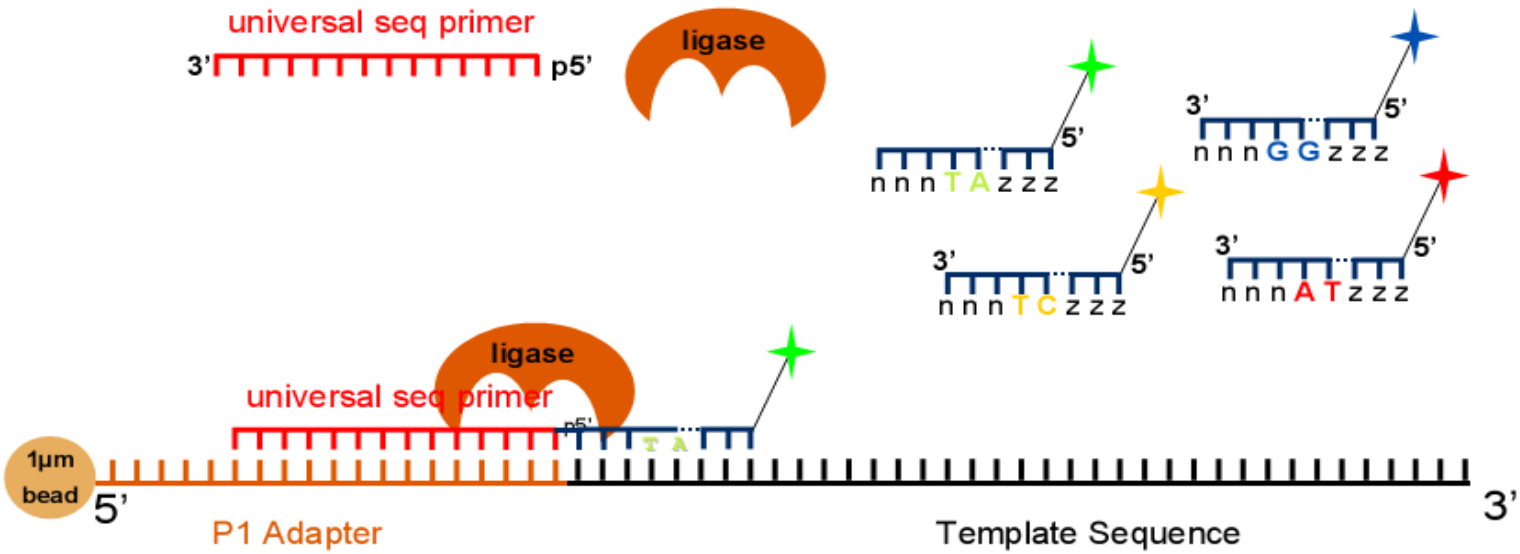
4 Dyes, 4 dinucleotides, 256 probes per dye

N= degenerate bases Z= Universal bases

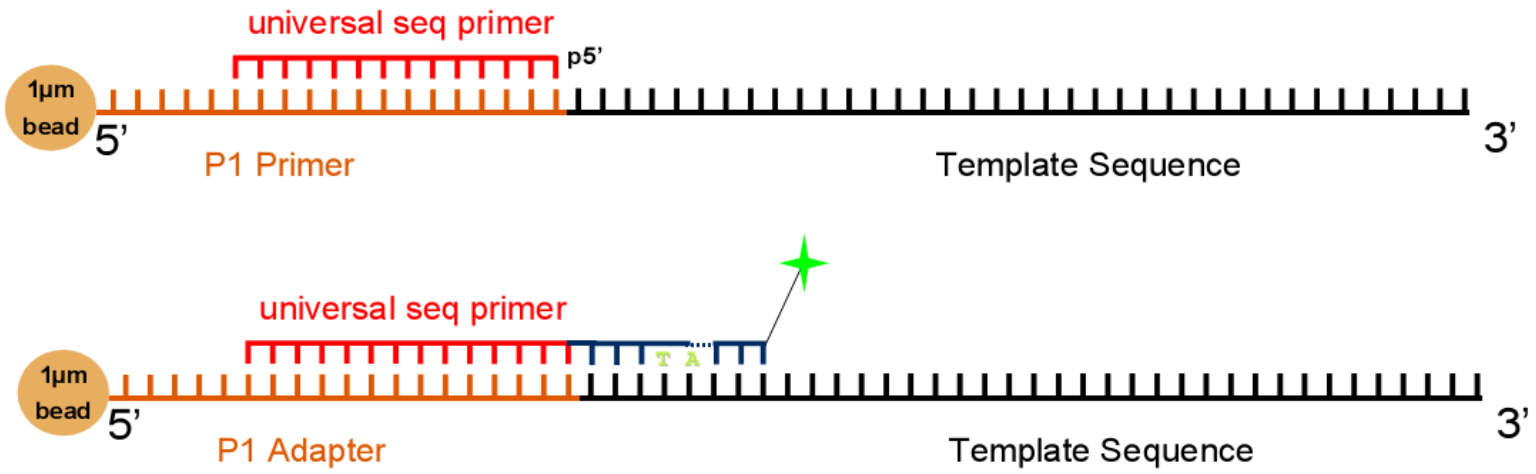
SOLiD Chemistry System 4-color ligation
Ligation reaction



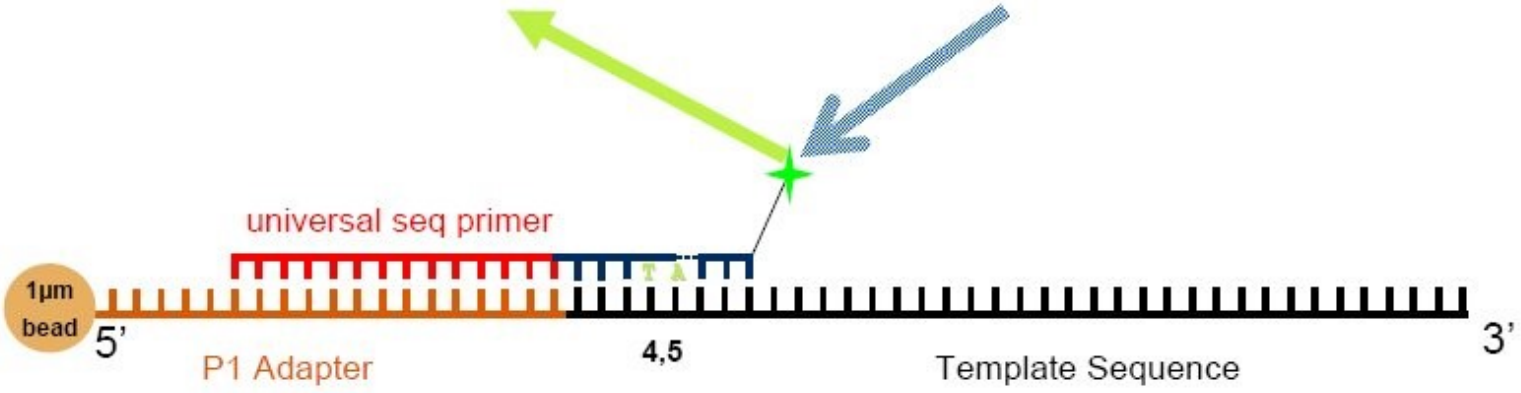
SOLiD Chemistry System 4-color ligation
Ligation reaction



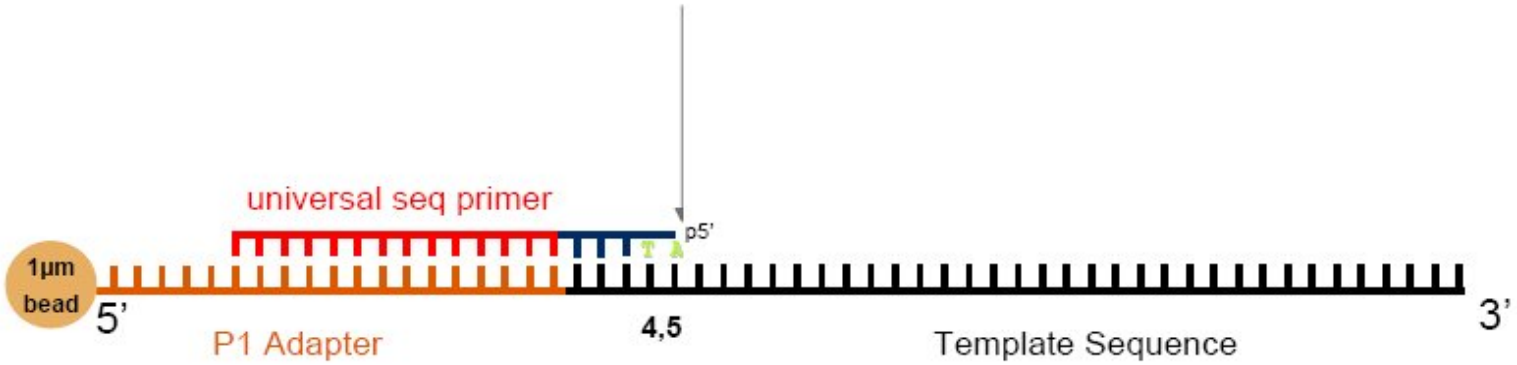
SOLiD Chemistry System 4-color ligation De-Phosphorylation



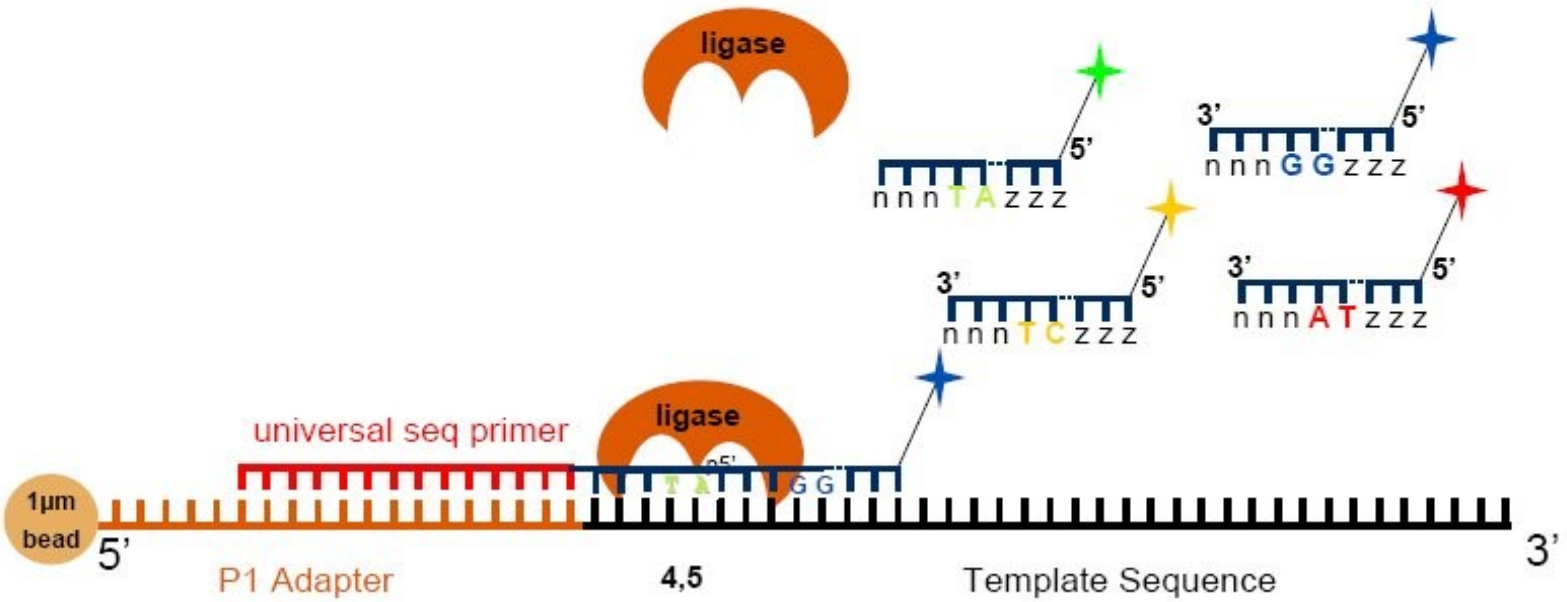
SOLiD Chemistry System 4-color ligation Visualization



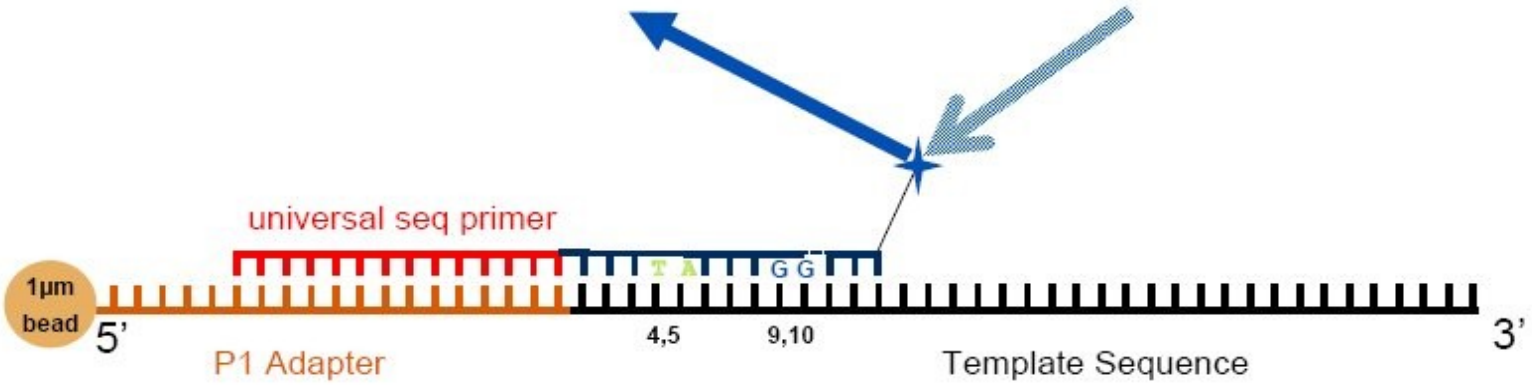
SOLiD Chemistry System 4-color ligation
Cleavage



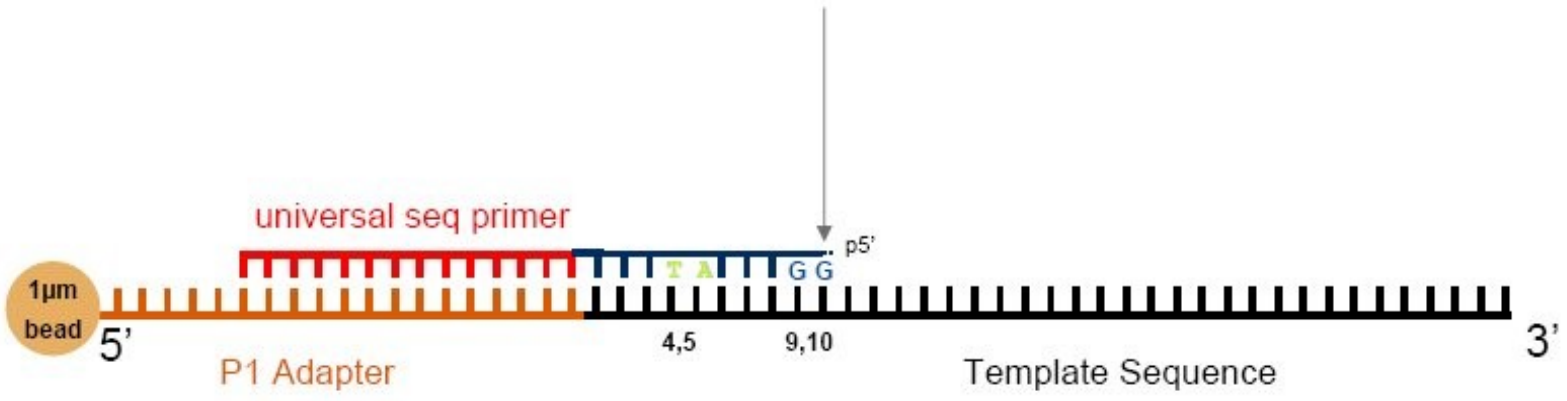
SOLiD Chemistry System 4-color ligation
Ligation (2nd cycle)



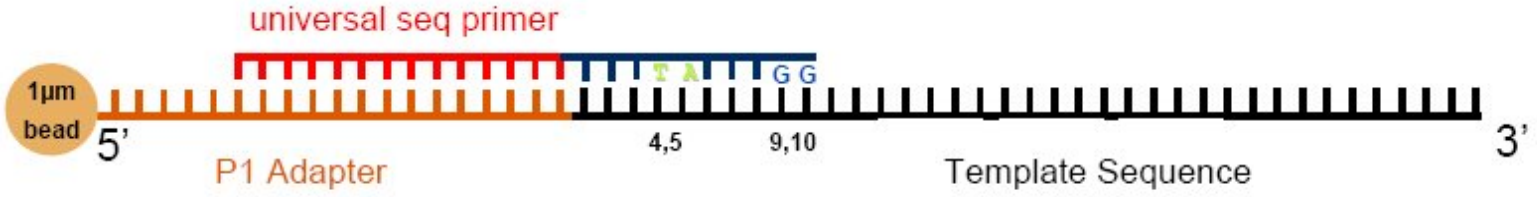
SOLiD Chemistry System 4-color ligation Visualization (2nd cycle)



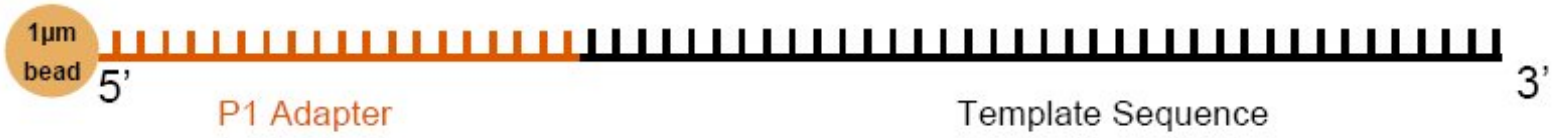
SOLiD Chemistry System 4-color ligation Cleavage (2nd cycle)



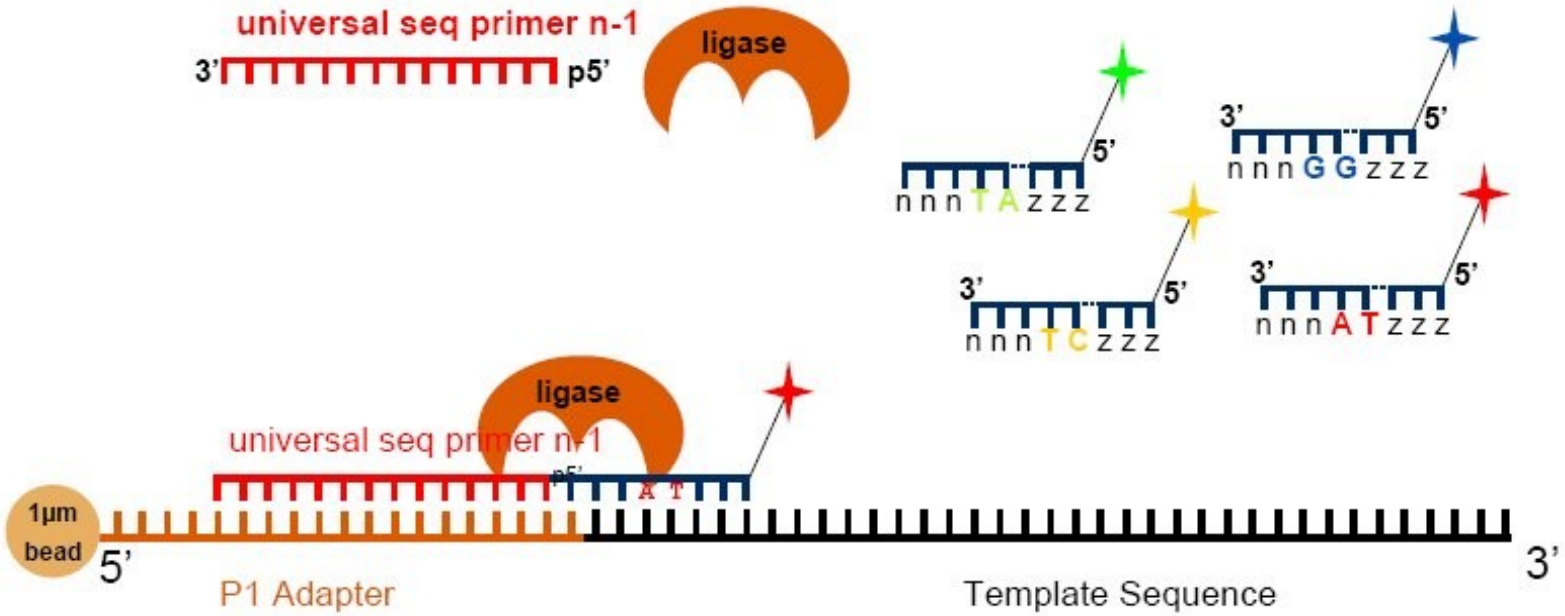
SOLiD Chemistry System 4-color ligation interrogates every 5th base



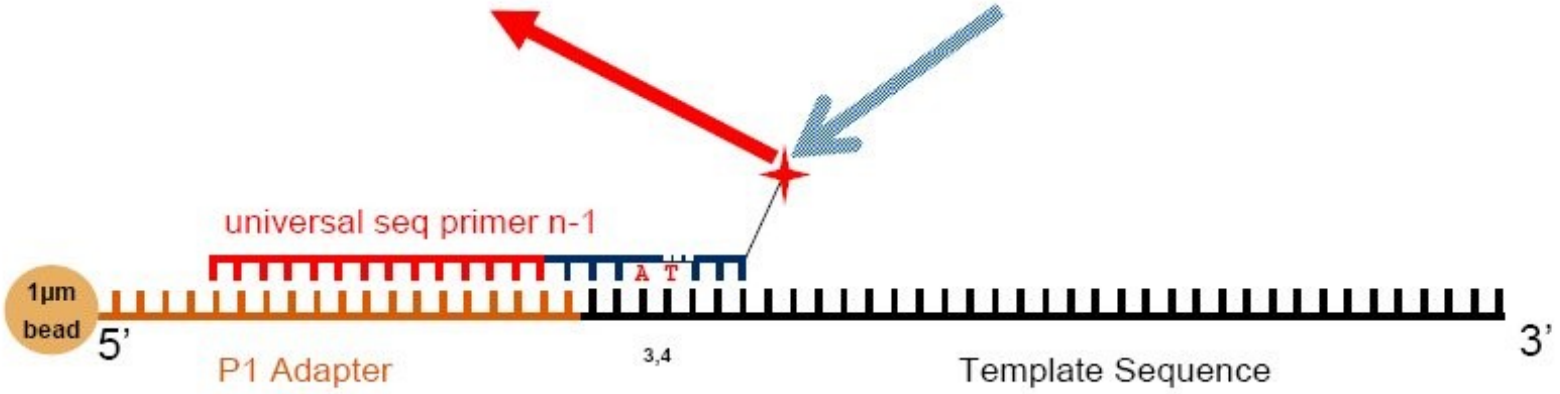
SOLiD Chemistry System 4-color ligation Reset



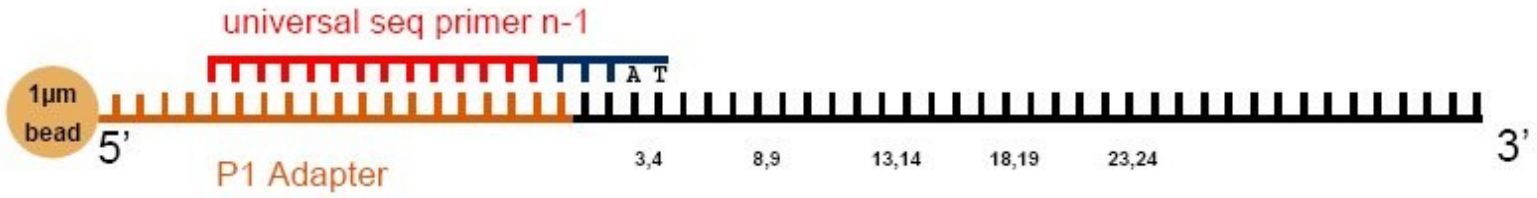
SOLiD Chemistry System 4-color ligation (1st cycle after reset)



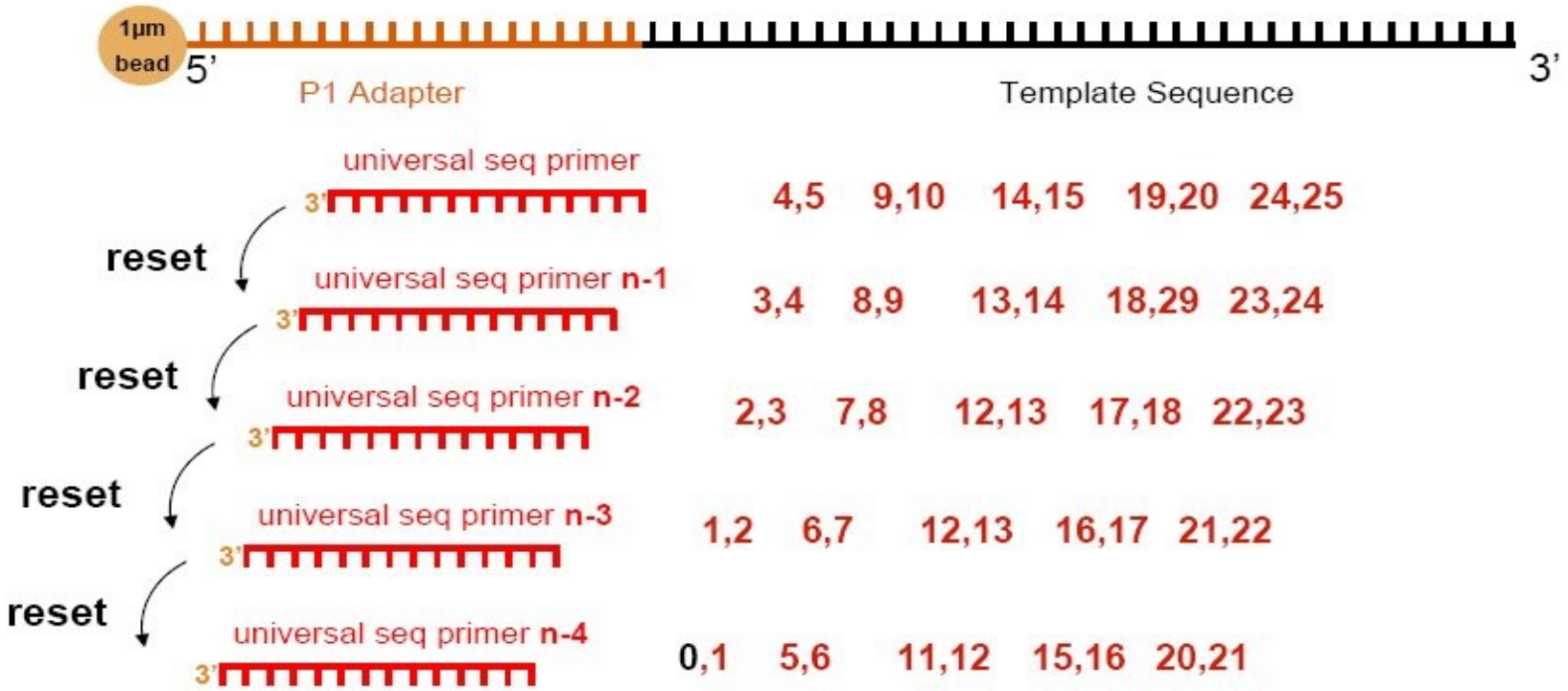
SOLiD Chemistry System 4-color ligation (1st cycle after reset)



SOLiD Chemistry System 4-color ligation (2nd Round)

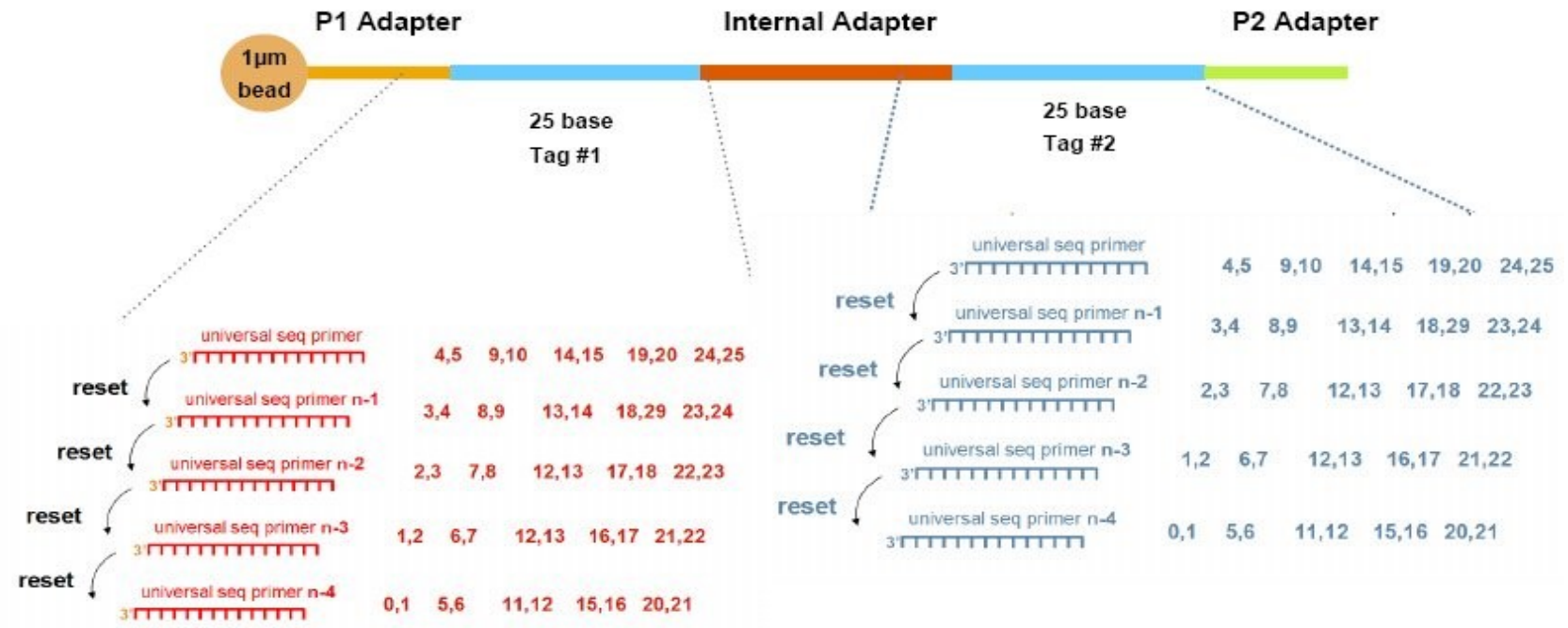


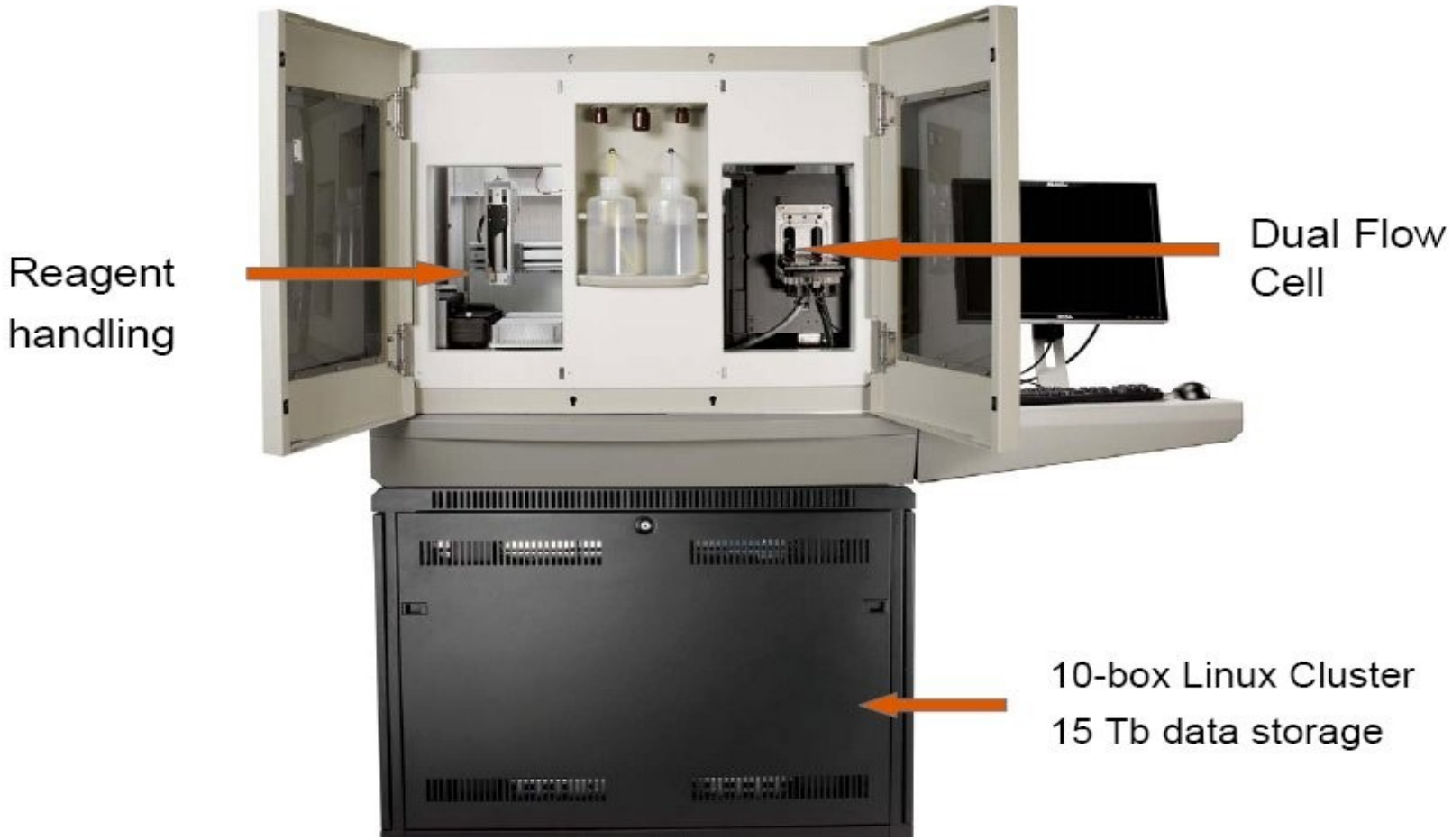
Sequential rounds of sequencing
Multiple cycles per round



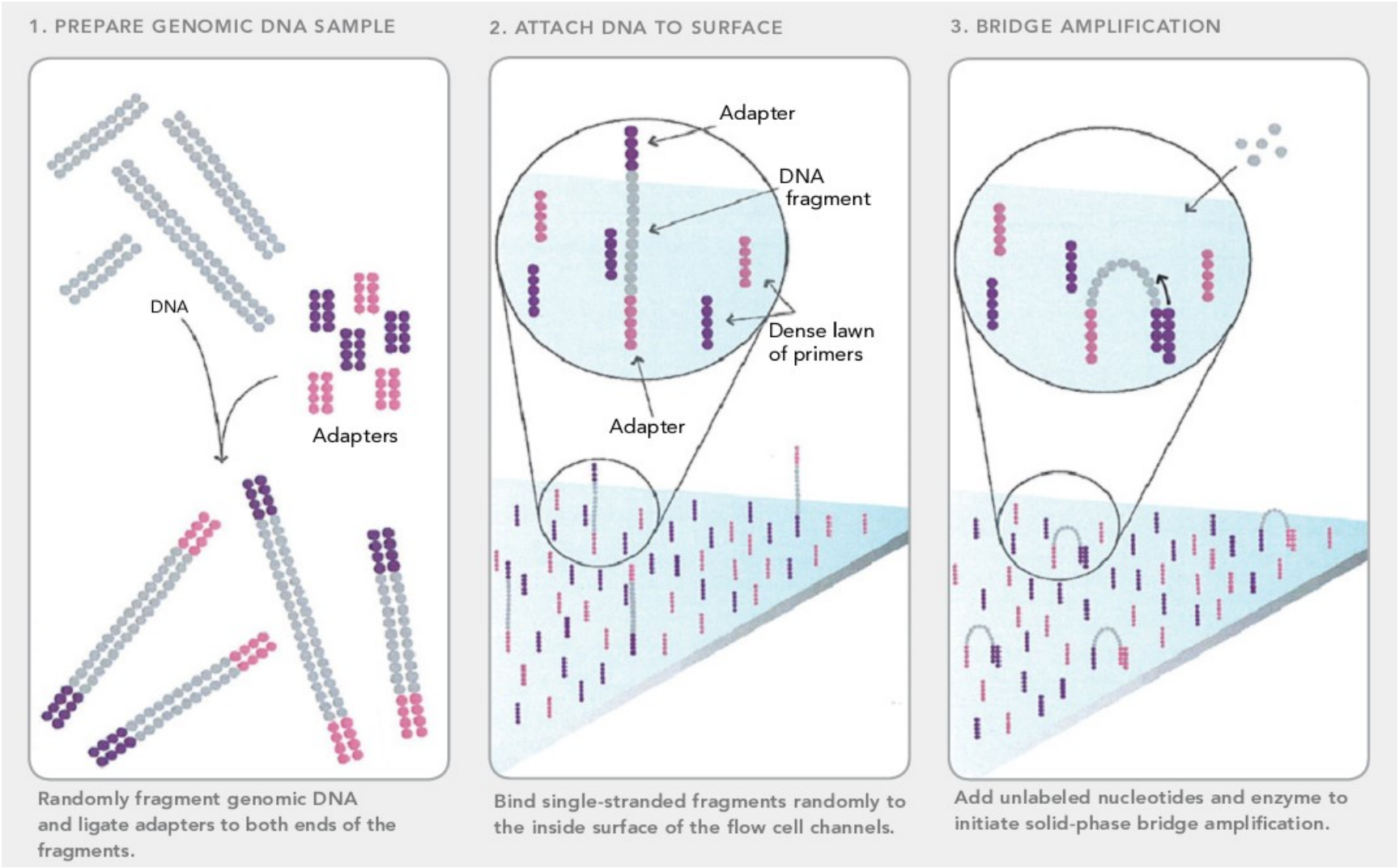
Paired End two sequences generated

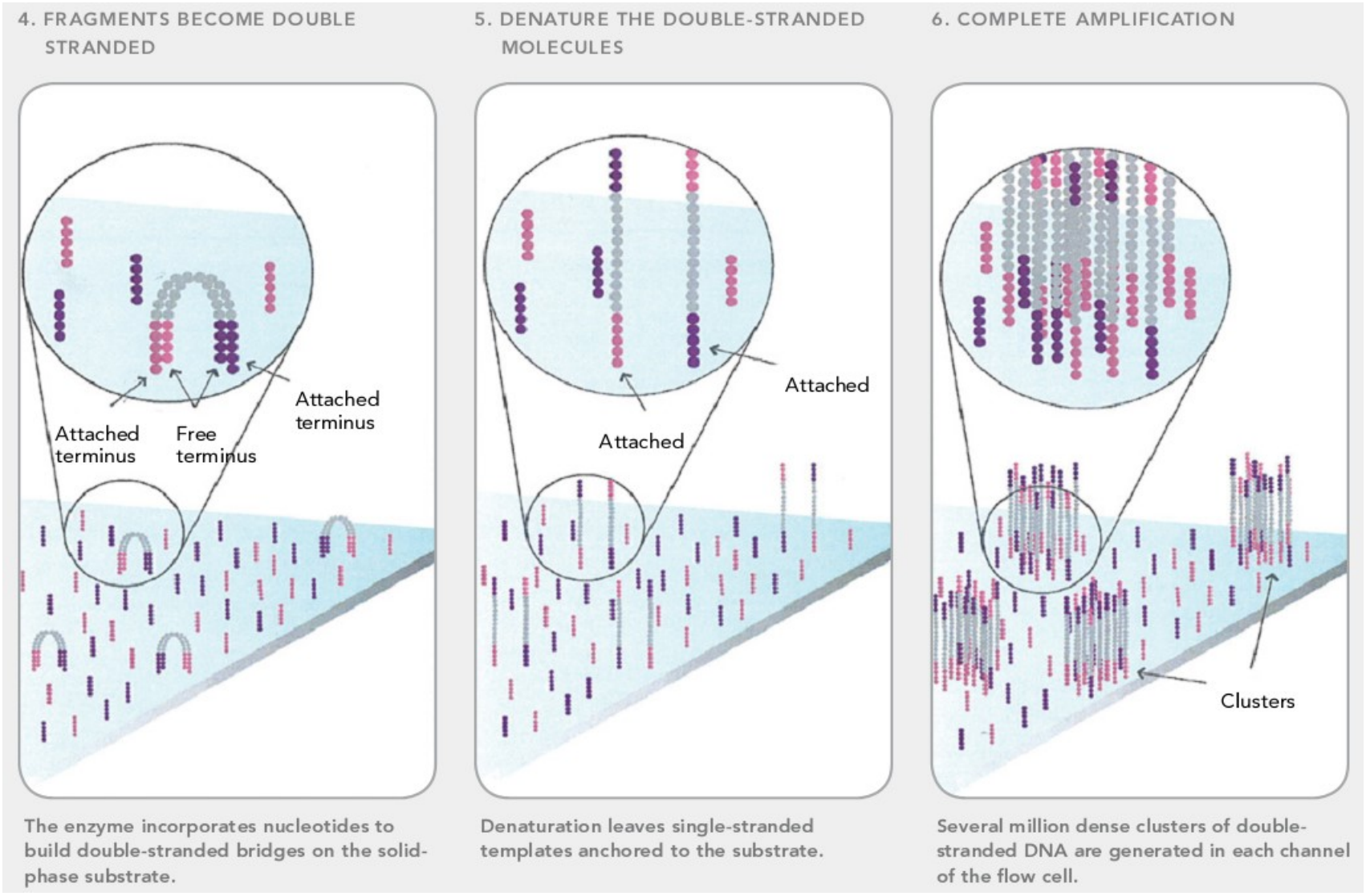
Sequential rounds of sequencing
Multiple cycles per round



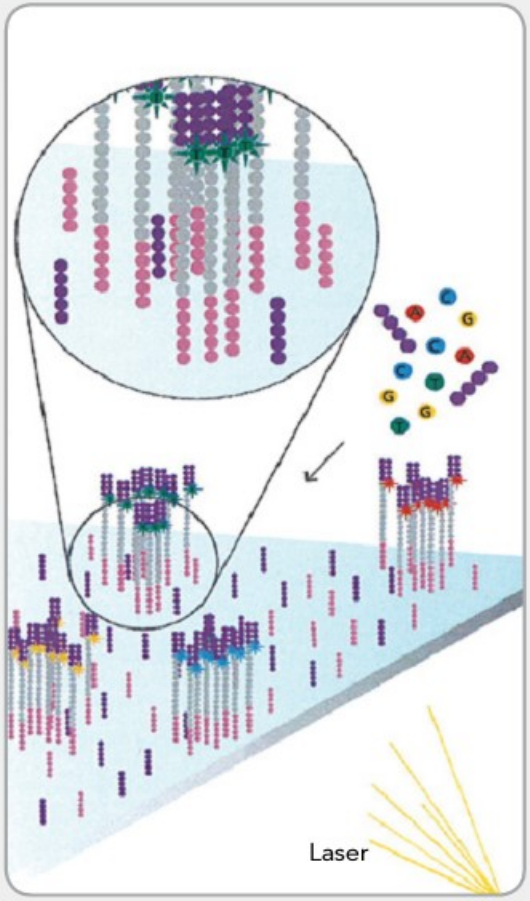

Reagent handling
Dual Flow Cell
10-box Linux Cluster
15 Tb data storage

Solexa (2007)


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.


4. FRAGMENTS BECOME DOUBLE STRANDED
5. DENATURE THE DOUBLE-STRANDED MOLECULES
6. COMPLETE AMPLIFICATION
Attached terminus
Attached terminus
Free terminus
Attached
Attached
Clusters
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
Denaturation leaves single-stranded templates anchored to the substrate.
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE

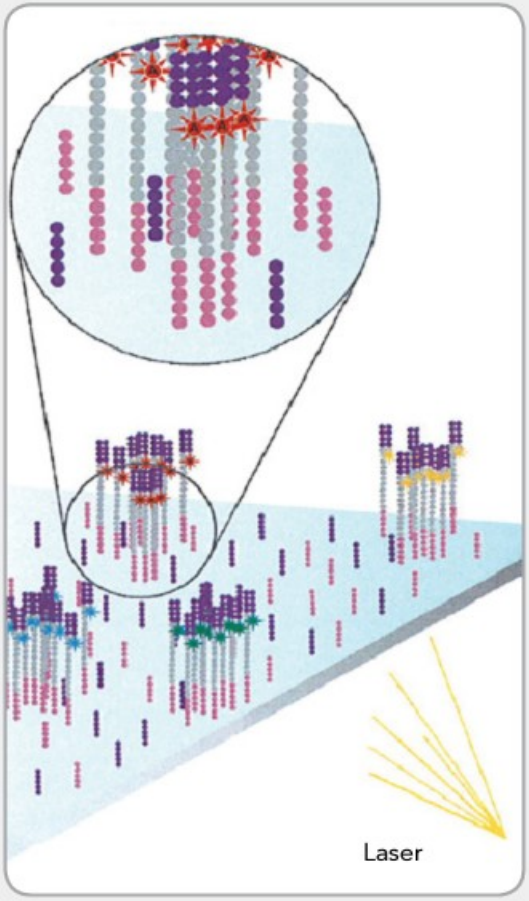


First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

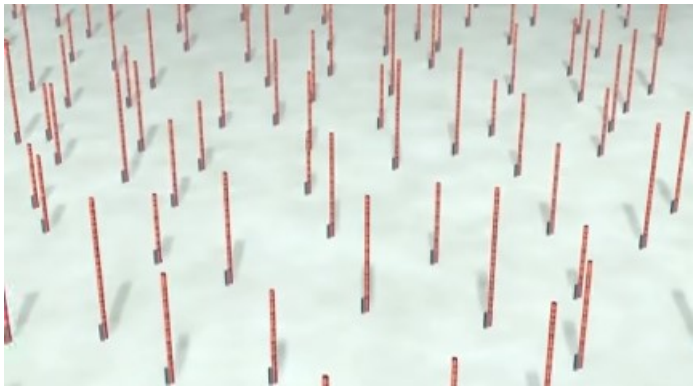


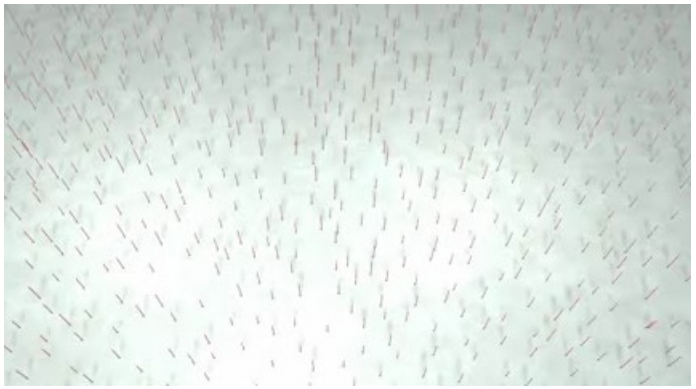
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

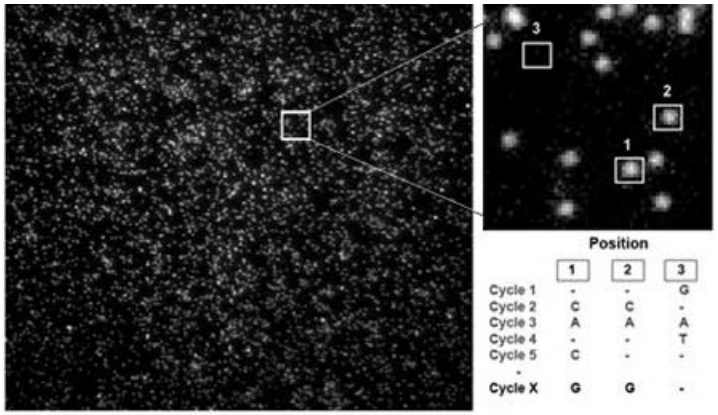
HELICOS (2008)

Helicos BioSciences Corporation

True Single Molecule Sequencing (tSMS)

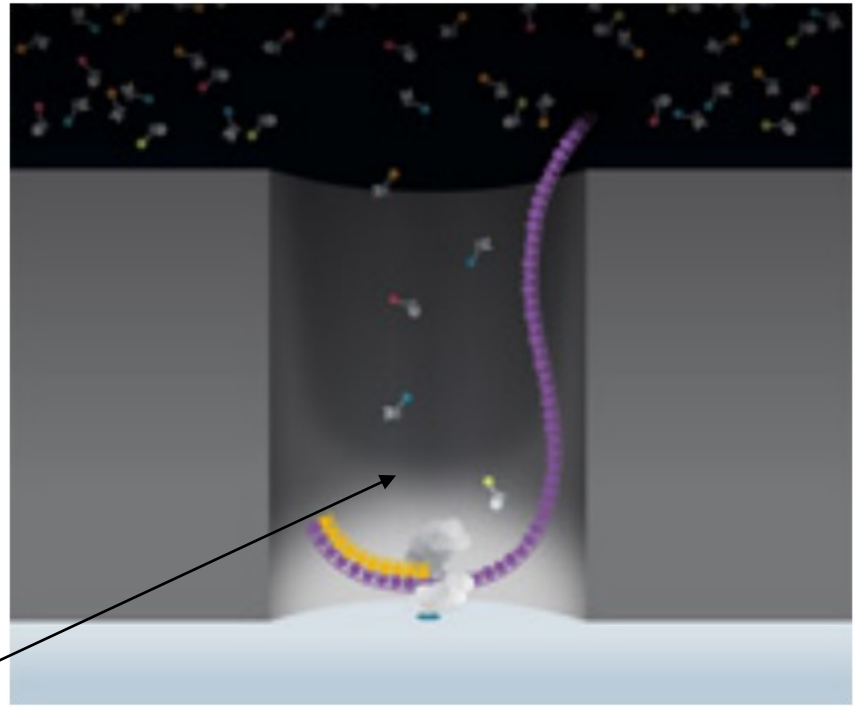






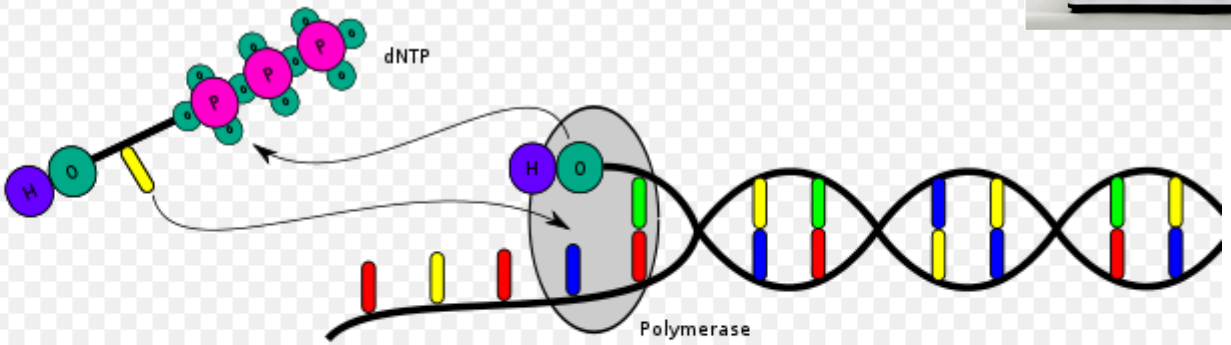
Single Molecule Real-Time (SMRT)

Pacific Biosciences

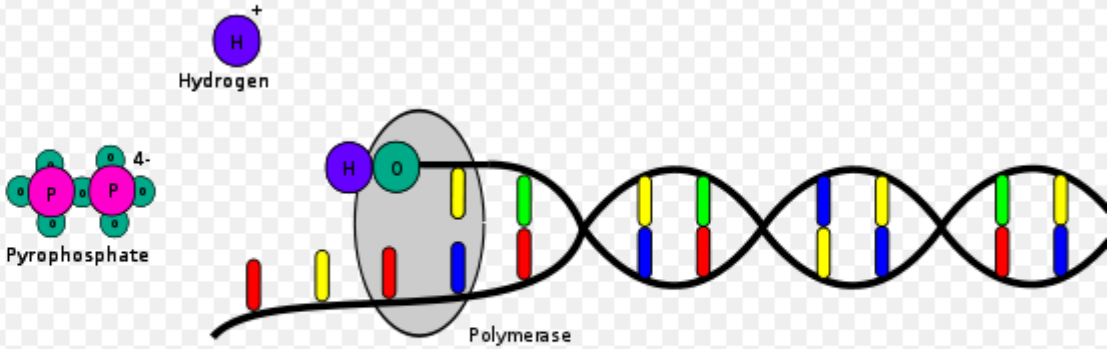


20 zeptoliters

Ion Torrent

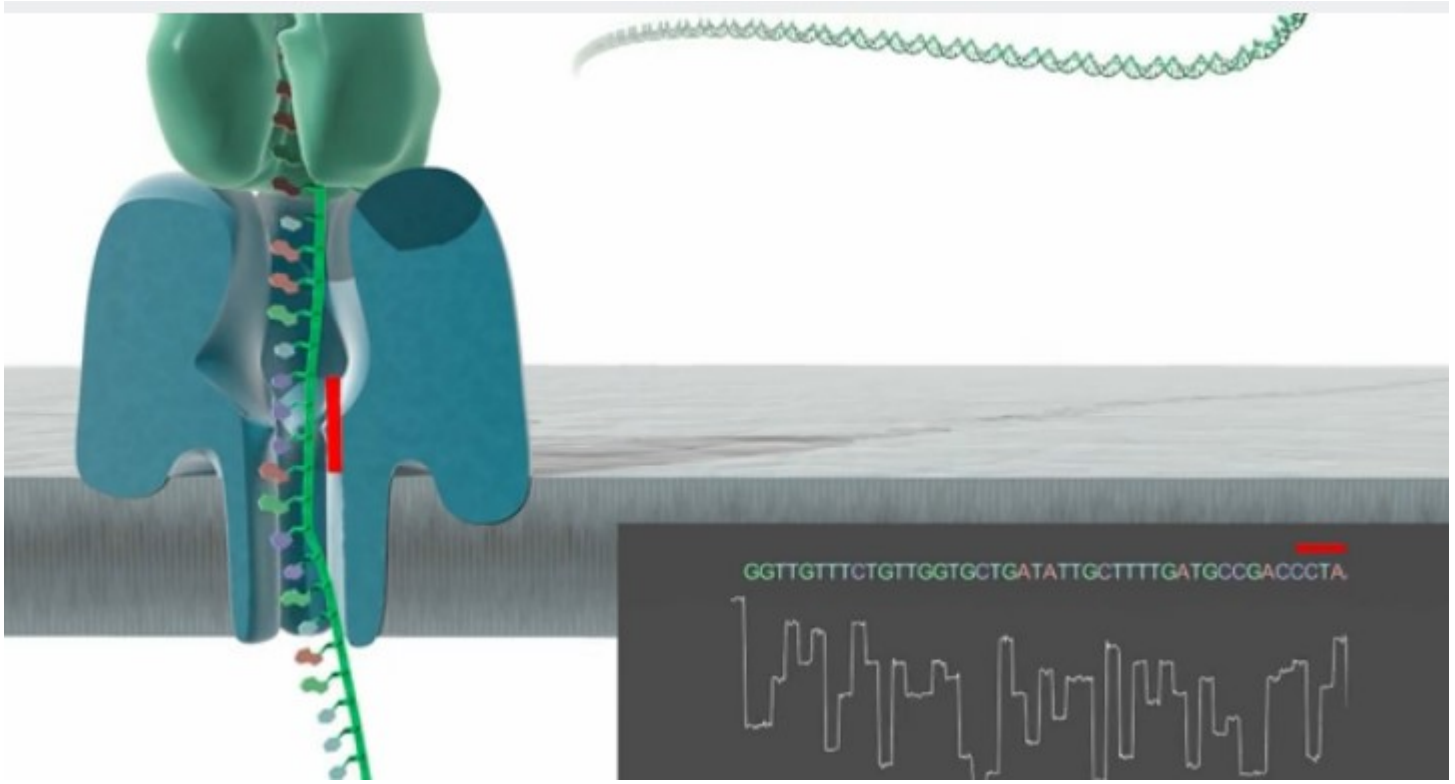


Polymerase integrates a nucleotide.



Hydrogen and pyrophosphate are released.

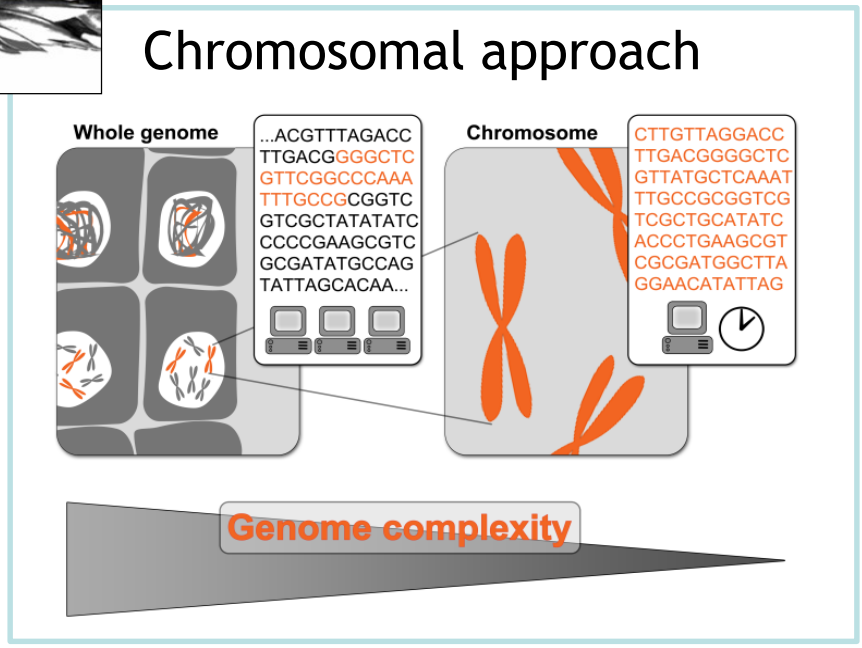
Oxford nanopore



Challenges in Genome Sequencing

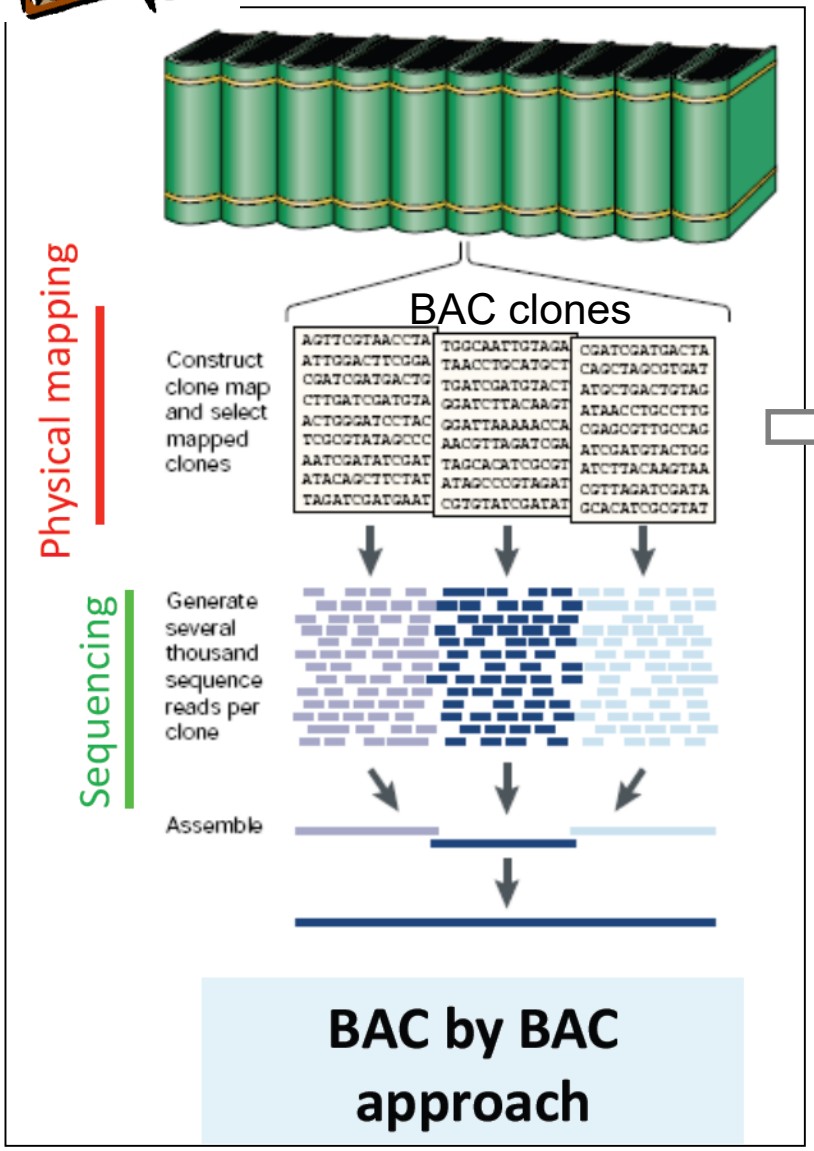
De novo genome assemblies using only short read data of NGS technologies are generally incomplete and highly fragmented due to

- Large duplications - chromosomal approach, BAC-by-BAC sequencing
- High proportion of repetitive DNA - challenge!

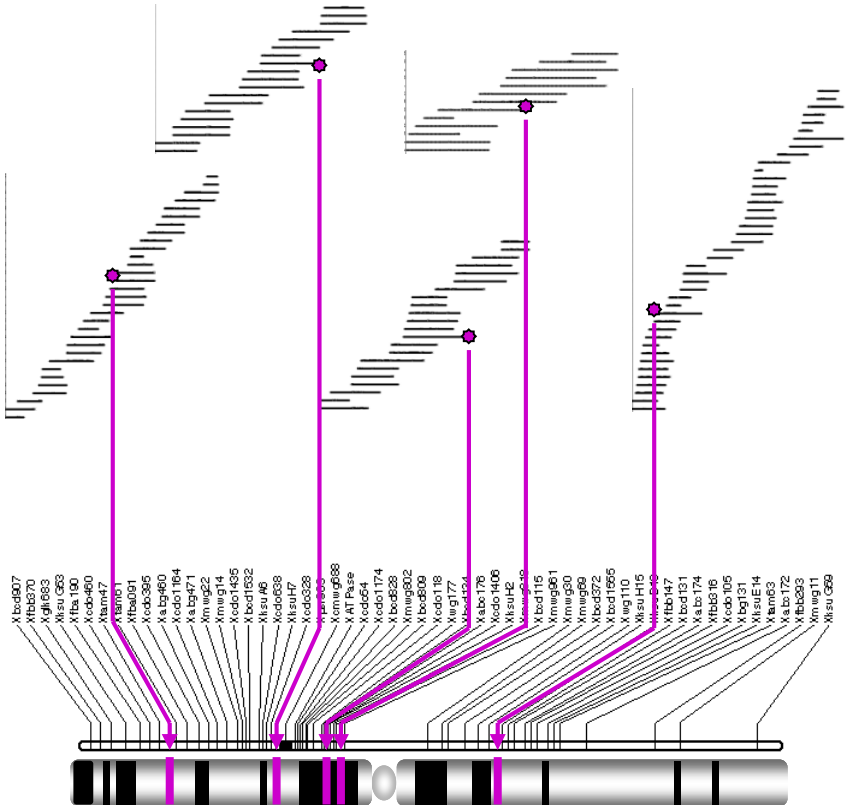


- Large genome size (~17 Gb)
- Polyploidy (3 subgenomes)

BAC-BY-BAC SEQUENCING

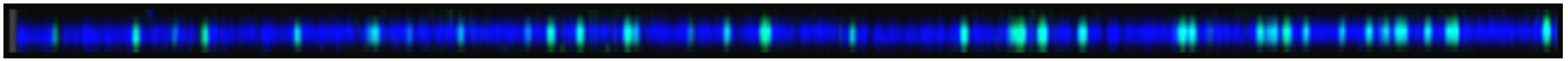


- **Physical map** is composed of contigs of overlapping BAC clones
- BAC contigs are landed on the chromosome through markers comprised in the contigs



SOLUTIONS FOR THE REPEATS

- **Long mate-pair reads** > 10 kb
- **Long read technologies** – PacBio, Oxford Nanopore
- **Optical mapping**
  - Single-molecule mapping of genomic DNA **hundreds of kilobases to several megabases in size**
  - Creates **sequence-motif maps**, which provide long-range template for ordering genomic sequences
  - **Visualisation of reality** "Seeing is Believing"

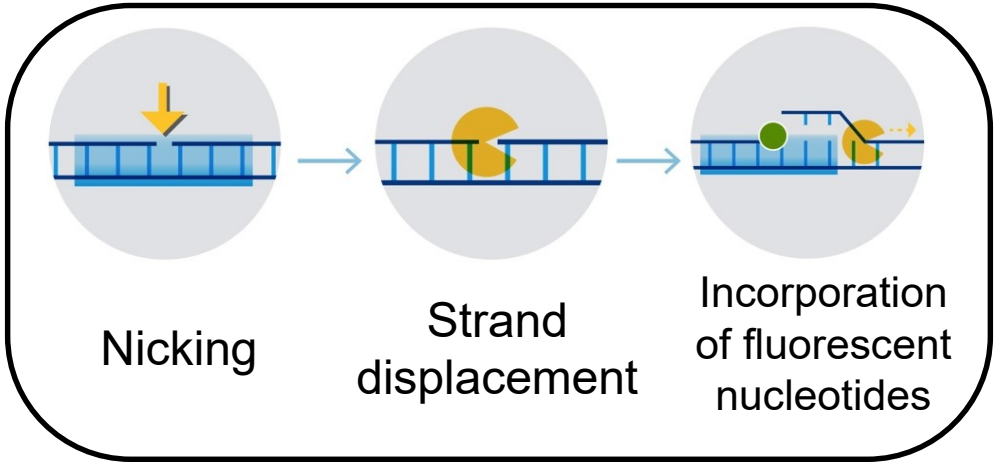


Optical Mapping

Three enzymatic approaches

- **restriction enzymes:**
  sequence-specifically cleave DNA immobilized on a surface

- **nicking enzymes:**
  fluorescent labelling of the nicking site in solution (BioNano Genomics - Irys)

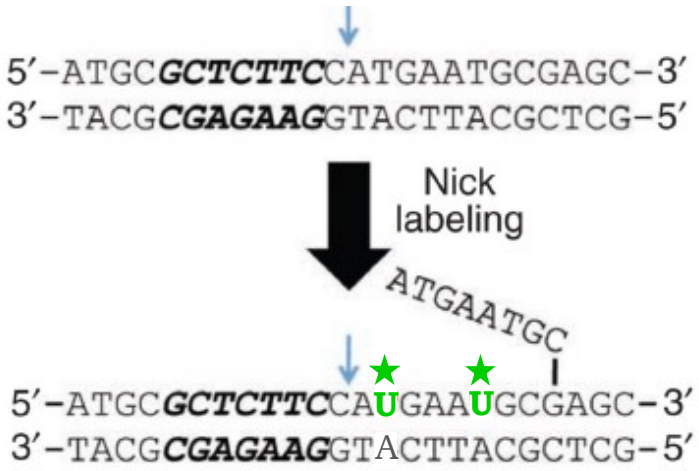


- **methyltransferase enzymes:**
  labelling with ultra-high density

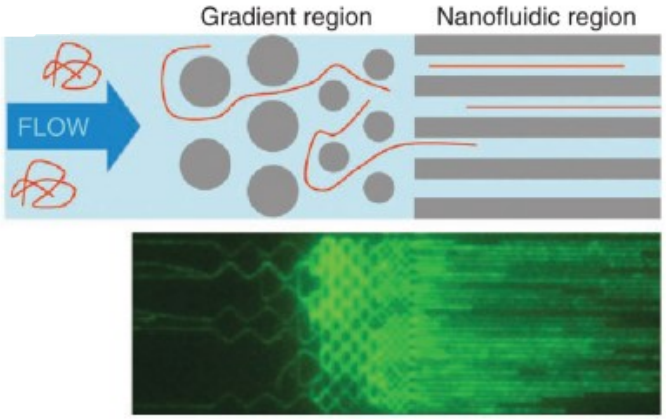
BioNano Genome mapping on Nanochannel Arrays

1 Sequence-specific labeling

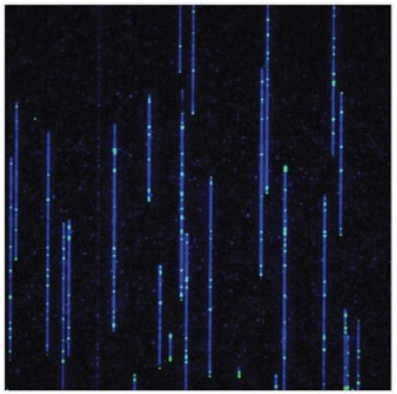
Nickase (Nt.BspQI)



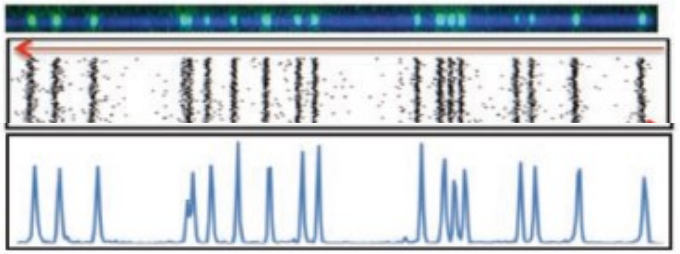
2 DNA linearization



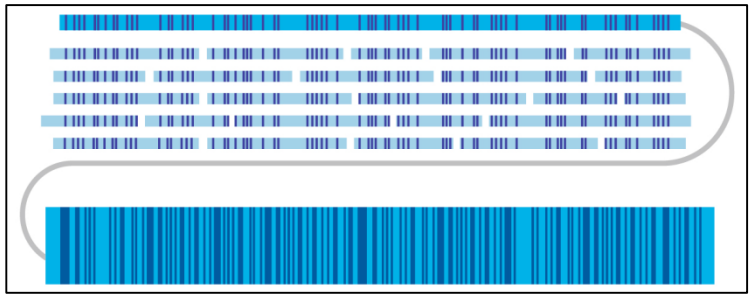
3 Fluorescence imaging



4 Map construction



5 Building consensus map



Lam et al., Nat. Biotechnol. 30(8) 2012