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®

From discovery to technology explosion

- 1868: Discovery of DNA
- 1953: Watson and Crick propose double helix structure
- 1977: Sanger sequencing
- 1985: PCR
- 2000: Working draft human genome announced (Sanger method)

- 2005: 454 sequencer launch (pyrosequencing)
- 2006: Genome Analyzer launched (Solexa sequencing)
- 2007: SOLiD launched (ligation sequencing)
- 2009: Whole human genome no longer merits Nature/Science paper
- 2010: “third-gen” systems

$ human Genome

$3 billion

$2-3 million

$250k

$50k

$20k

<$1k

Frederick Sanger

1958 – Nobelova cena za určení struktury inzulínu

1975 - Dideoxy sekvenační metoda

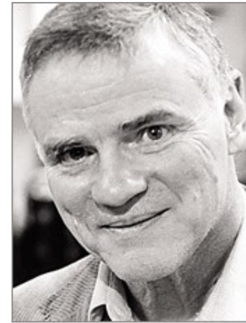
1977 – osekvenoval Φ-X174 (5,368 bp)

1980 – dostal druhou Nobelovu cenu za chemii

Později (polovina 80-tých let) osekvenoval bakteriofága λ pomocí shotgun metody (48,502 bp)

Sekvenování genomů

- **1986** **Leroy Hood: první automatický sekvenátor**



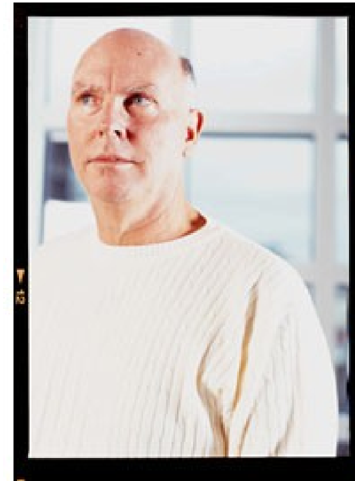
Leroy Hood

- **1986** **Human Genome Initiative**
- **1990** **započat projekt sekvenování lidského genomu (předpokládaná doba 15 let)**

Sekvenování genomů

- **1995** John Craig Venter sekvenoval první bakteriální genom

- **1996** první eukaryotický genom (kvasinka) sekvenován



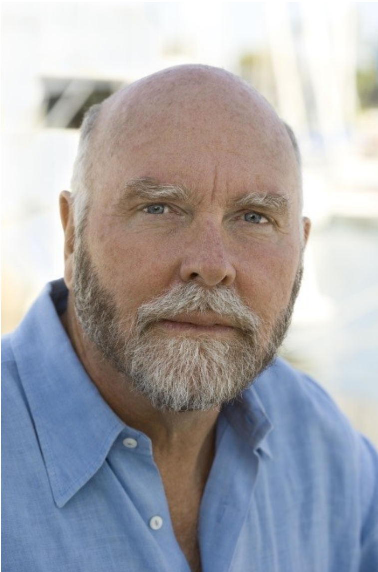
John Craig Venter

Craig Venter

Global Ocean Sampling Expedition

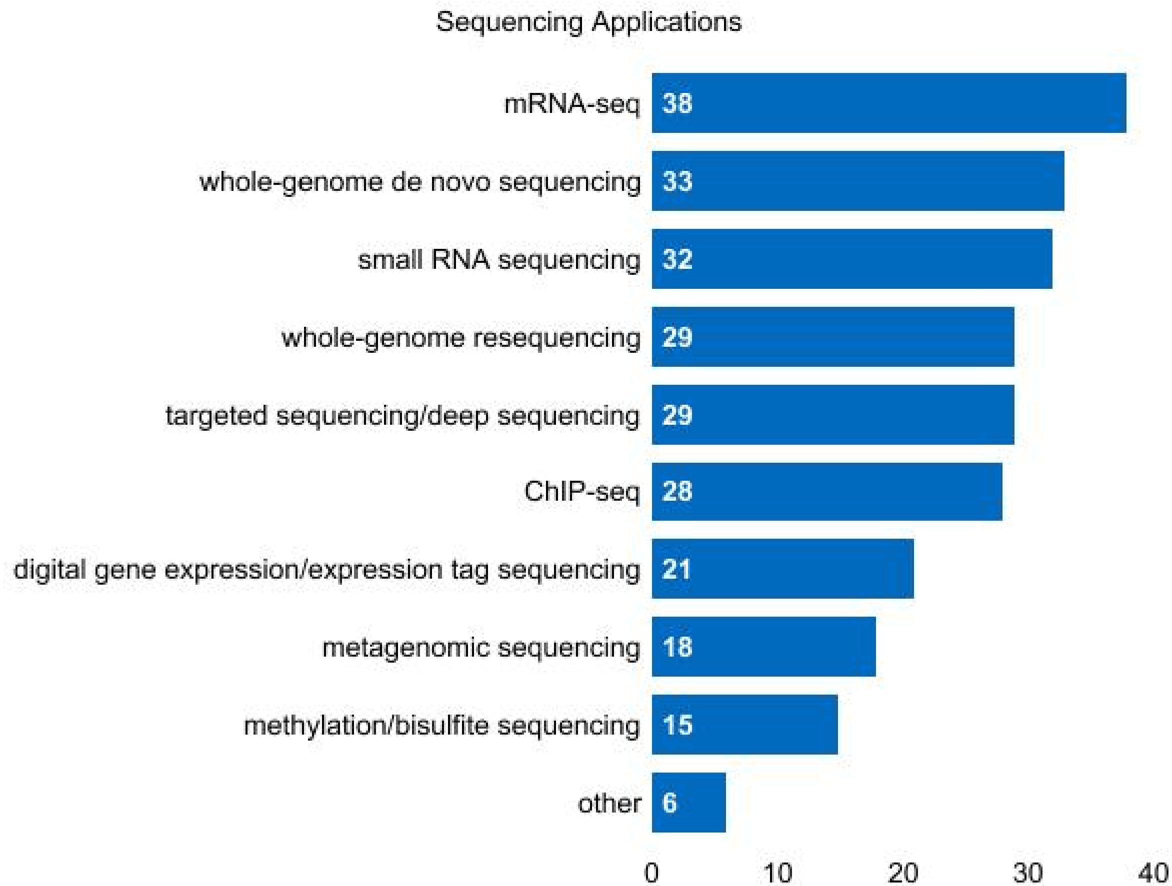
Synthetic genomics

Human Longevity Inc



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

Which applications are labs performing?

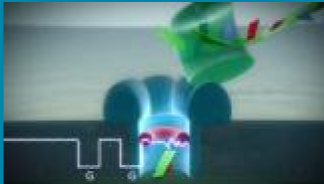


Oxford Nanopore

Application Specific

Adaptable protein nanopore:

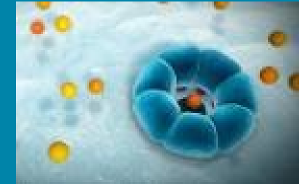
DNA Sequencing



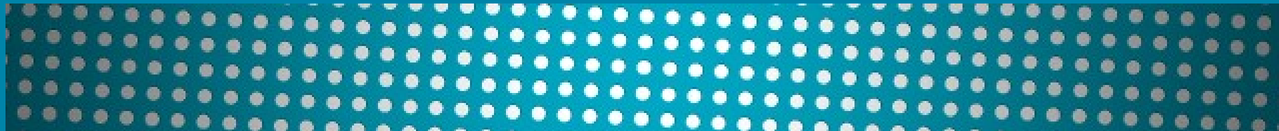
Proteins

Polymers

Small Molecules



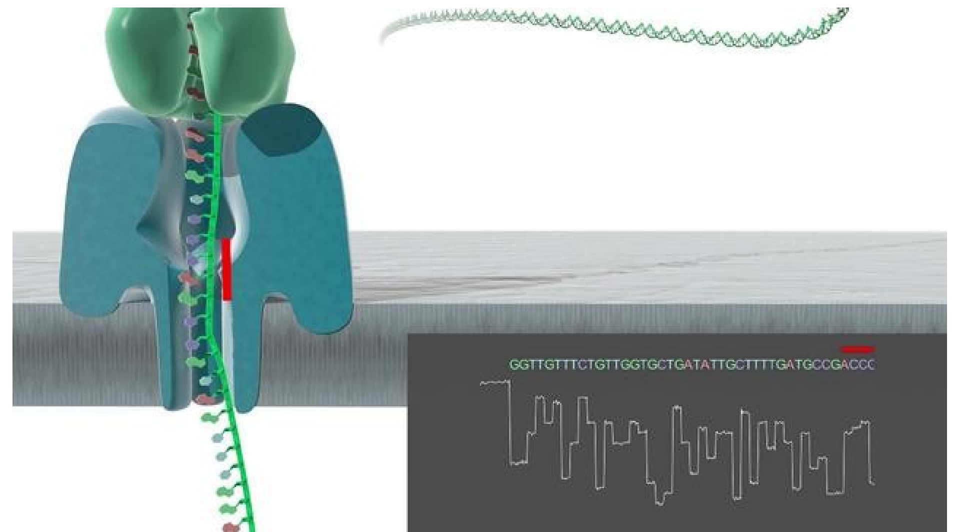
Generic Platform



Electronic read-out system

Nanopore

- Pros: Extremely long sequences, single molecule, portable (minION)
- Cons: Very high error rates (up to 38% reported)



DNA degradation

Mechanical damage during tissue homogenization.

Wrong pH and ionic strength of extraction buffer.

Incomplete removal / contamination with **nucleases**.

Phenol: too old, or inappropriately buffered (**pH 7.8 – 8.0**); incomplete removal.

Wrong pH of **DNA solvent** (acidic water).
Recommended: 1:10 TE for short-term storage, or 1xTE for long-term storage.

Vigorous pipetting (wide-bore pipet tips).

Vortexing of DNA in high concentrations.

Too many **freeze-thaw** cycles (*we tested 5, still Ok*).

Debatable: sequence-dependent

What are the main contaminants?

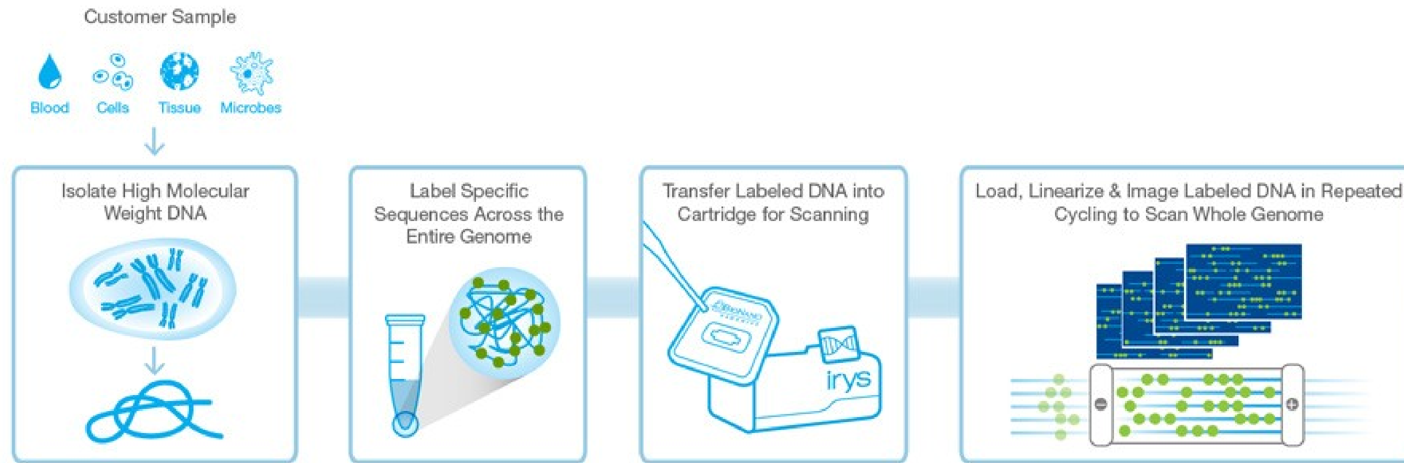
Polysaccharides
Lypopolysaccharides
Growth media residuals

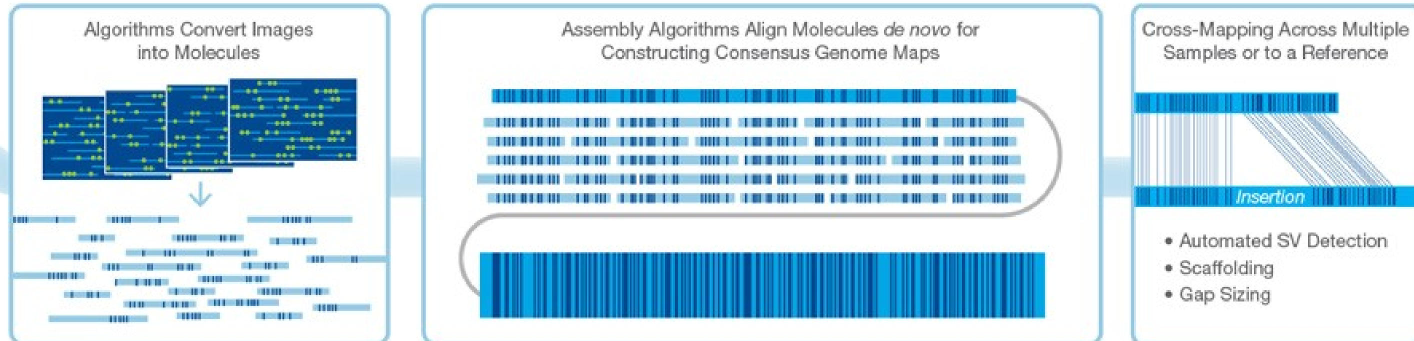
Chitin
Protein
Secondary metabolites
Pigments
Growth media residuals

Chitin
Fats
Proteins
Pigments

Polyphenols
Polysaccharides
Secondary metabolites
Pigments

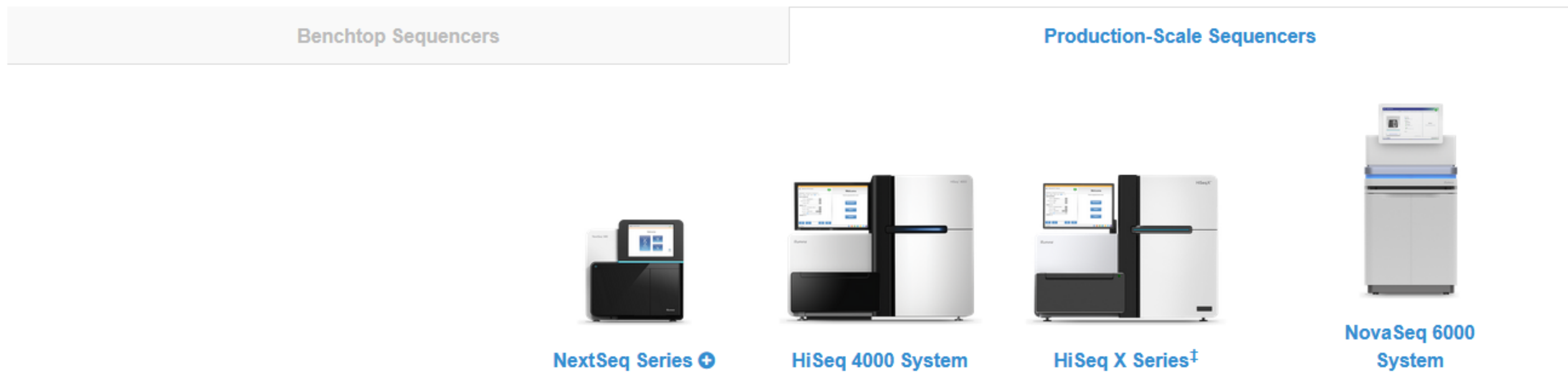
BioNano





Top sequencing companies

#1. Illumina
Revenues: $2.752 billion in 2017



Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●		●
Exome Sequencing	●	●		●
Targeted Gene Sequencing (amplicon, gene panel)	●	●		●
Whole-Transcriptome Sequencing	●	●		●
Gene Expression Profiling with mRNA-Seq	●	●		●
miRNA & Small RNA Analysis	●	●		●
DNA-Protein Interaction Analysis	●	●		●
Methylation Sequencing	●	●		●
Shotgun Metagenomics	●	●		●

#2. Thermo Fisher Scientific
Revenues: "Just under" **$418.36 million** in 2017
Ion AmpliSeq technology works for researchers using Illumina's NGS platforms, under the name AmpliSeq for Illumina. Thermo Fisher includes NGS within its life sciences solutions segment, which accounted for $5.73 billion of the company's total revenue of $20.918 billion.

#7. Pacific Biosciences of California (PacBio)
2017 revenues: $93.5 million aquired by Illumina 2019

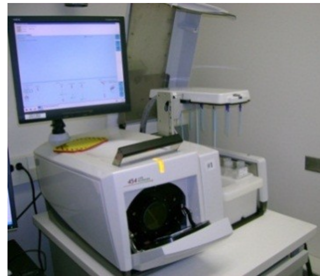
#10. Oxford Nanopore Technologies
2016 revenues: £4.5 million

	PacBio[1]		Oxford Nanopore[2]	
Instrument Specifications	RS II (P6-C4)	Sequel	MinION	PromethION
Average read length	10 – 15 kb	10 – 15 kb	Variable (up to 900 kb)[3,4]	*
Error rate	10 – 15 %	10 – 15 %	5 – 15 %[4,5]	*
Output	500 Mb – 1 Gb	5 Gb – 10 Gb	~5 Gb[4]	*
# of reads	~50k	~500k	Variable (up to 1M)[6,7]	*
Instrument price/Access fee[a]	$700k	$350k	$1000[8]	$135k bundle[9]
Run price	~$400	~$850	$500-$900[7]	*

#9. 10x Genomics
2017 revenues: $71 million
The company announced a new version of its Chromium *de novo* assembly solution, which includes a new version of the assembly software, Supernova 2.0. The company's offerings also include Linked-Reads, a sequencing technology designed to provide long-range information from short-read sequencing data. 10x Genomics—which completed a $55 million Series C financing in 2016—organizes genetic information based on what is known as "read clouds" to map the larger picture of the genome.

Sequencing without a limit?

- A rapid progress in next generation sequencing technologies promises to provide complete (reference) DNA sequences



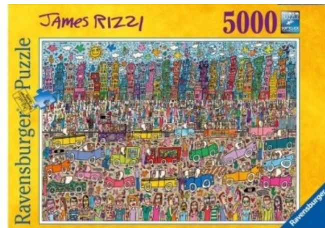
- **The bottleneck:**
  - NOT the sequencing capacity
  - BUT the ability to assemble many short reads with prevalence of repeated DNA (and polyploidy)

Genome sequencing

<u>Two strategies</u>

- **Whole genome shotgun (bottom-top)**
- **Clone-by-clone (top-bottom)**

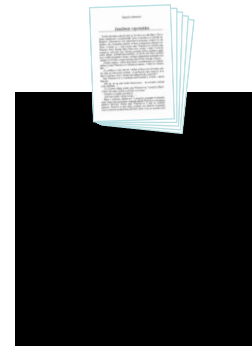
Whole genome shotgun

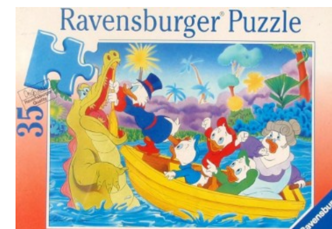


sequence assembly

Clone-by-clone

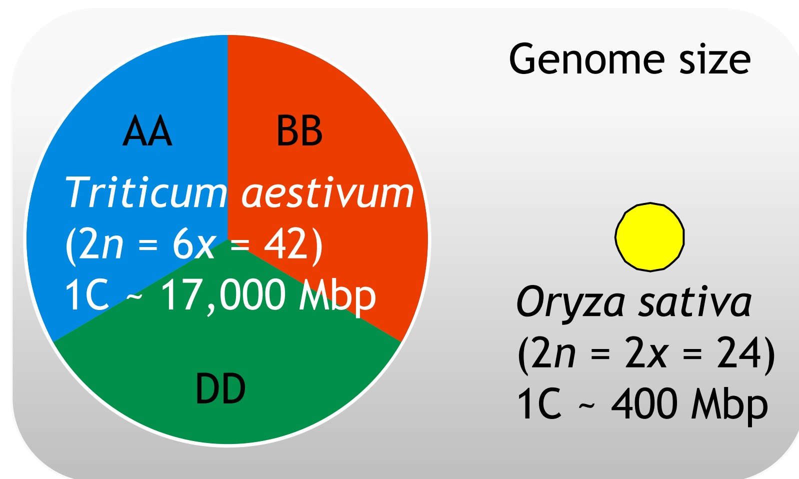
physical map construction

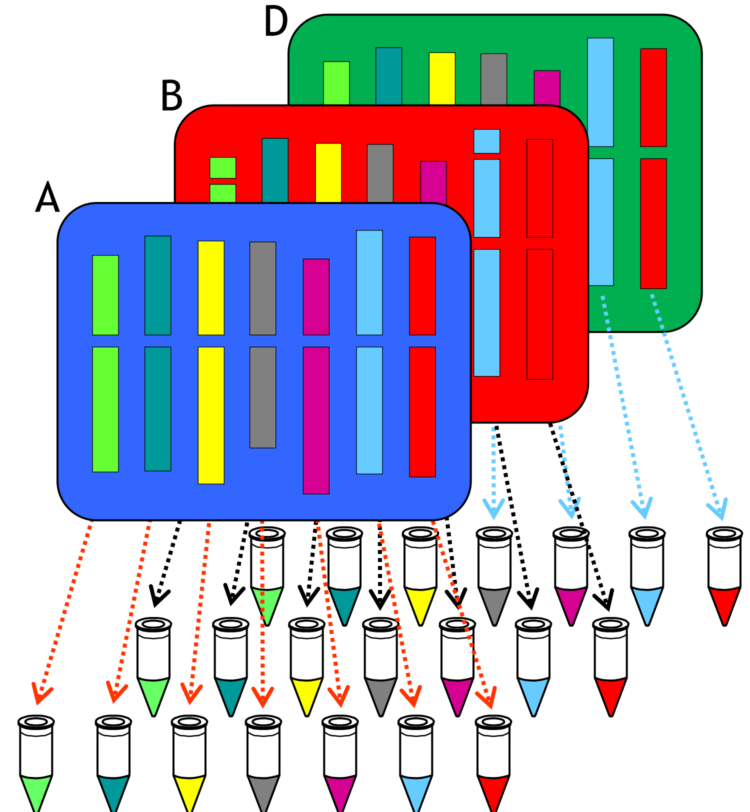




Chromosome genomics

- Aplication of genomics to flow-sorted chromosomes

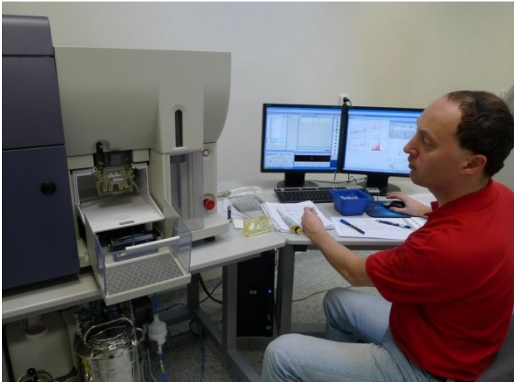




- Chromosomes: 605 - 995 Mbp (3.6 – 5.9% of the genome)
- Chromosome arms: 225 - 585 Mbp (1.3 – 3.4% of the genome)

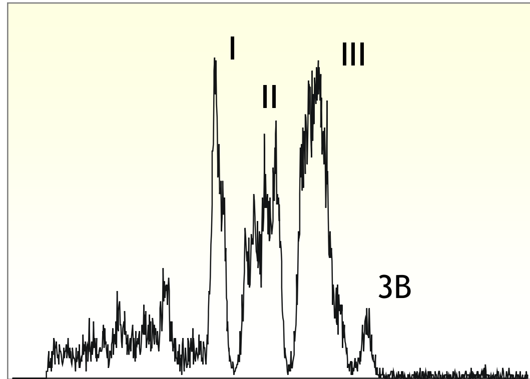
Doležel et al., Choromosome Research 15: 51, 2007

Chromosome sorting using flow cytometry



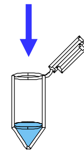
- Sample rate/ sec: ~1000 chromosomes
- Yield / day: 2 - 5x10^5 chromosomes

Creating chromosome-specific BAC libraries



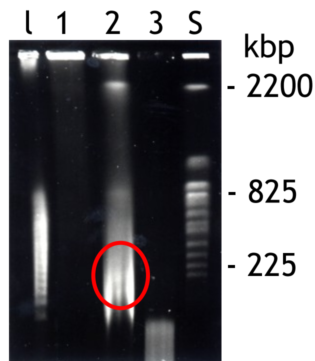
Chromosome sorting

5 x 10^6 flow-sorted chromosomes (~6 weeks of sorting)



Partial digestion

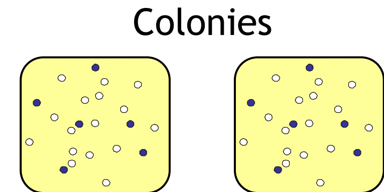
Size selection by PFGE

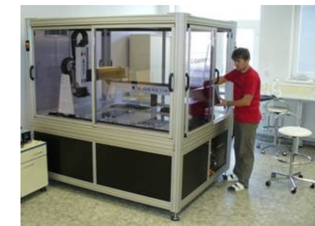


Ligation into a dephosphorylated BAC vector

Transformation of *Escherichia coli*

Colonies





Ordering into 384-well plates

Major challenges:

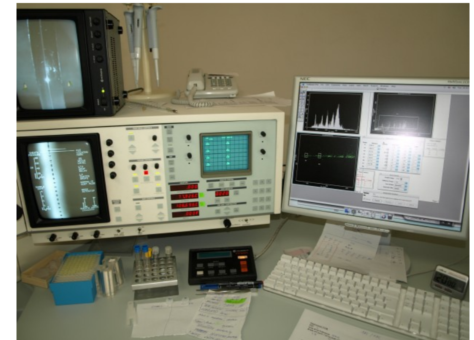
- Quantity of DNA (1 – 5 µg DNA)
- Quality of DNA (HMW)
- Cloning efficacy
- Insert size

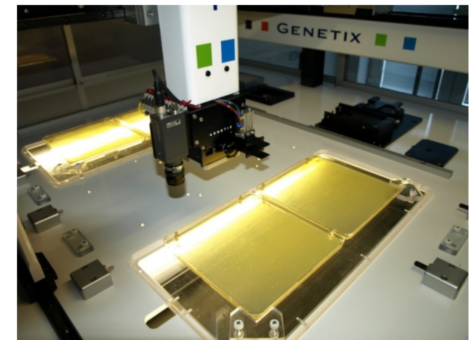
Šafář et al., Plant Journal 39: 986, 2004

Subgenomic BAC libraries

- Main advantages:
  - Chromosome specificity
  - Small number of clones (in wheat $\sim 5 \times 10^4$ vs. $>1 \times 10^6$)

- Subgenomic BAC libraries facilitate:
  - Targeted development of DNA markers (BAC end sequencing)
  - Positional gene cloning
  - Assembly of ready-to-sequence physical maps (BAC fingerprinting, WGP)



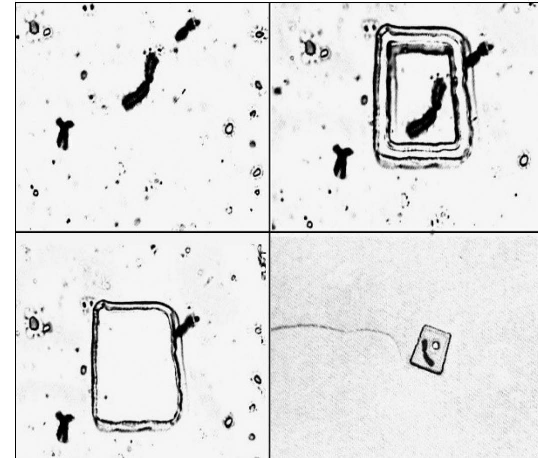


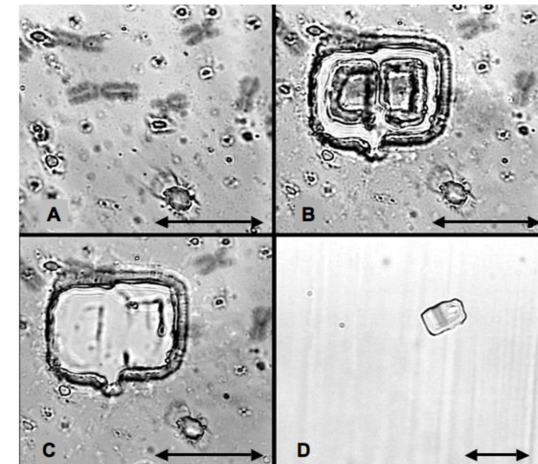
Šafář et al., Cytogent. Genome Research 129: 211, 2010

Laserová mikrodisekce

<u>Výhody:</u> vysoká čistota

<u>Nevýhody:</u> malý počet chromozomů, pracnost



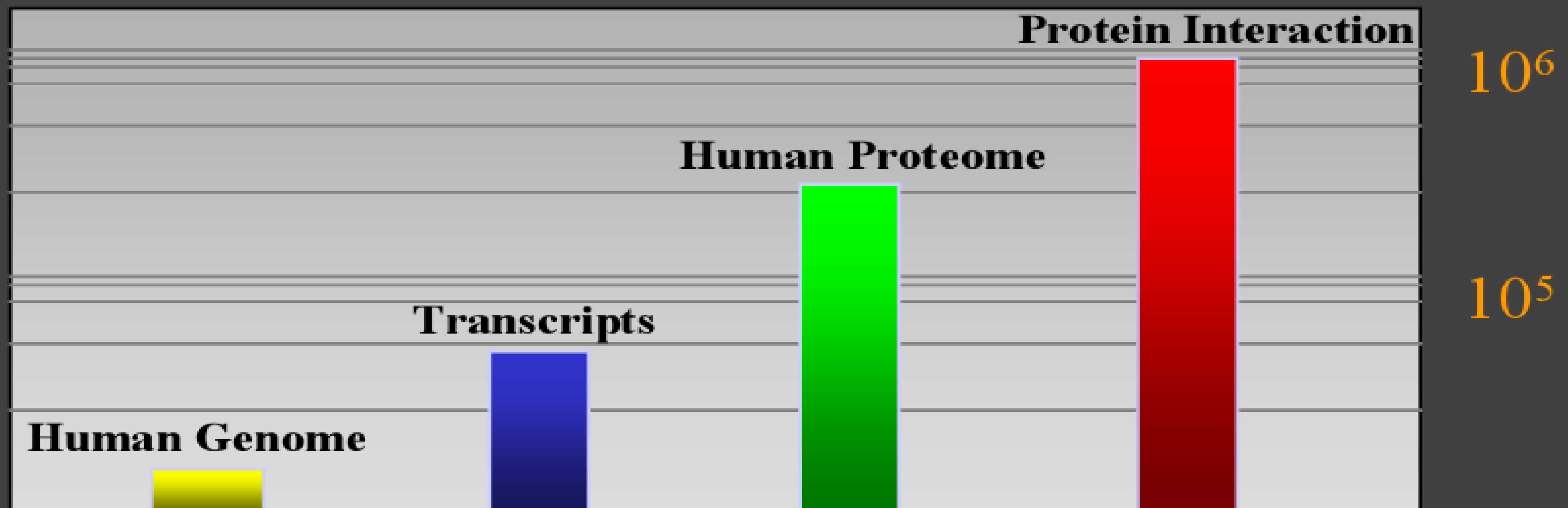


Genome: 30.000 genes

Transcriptome: 40-100.000 mRNAs

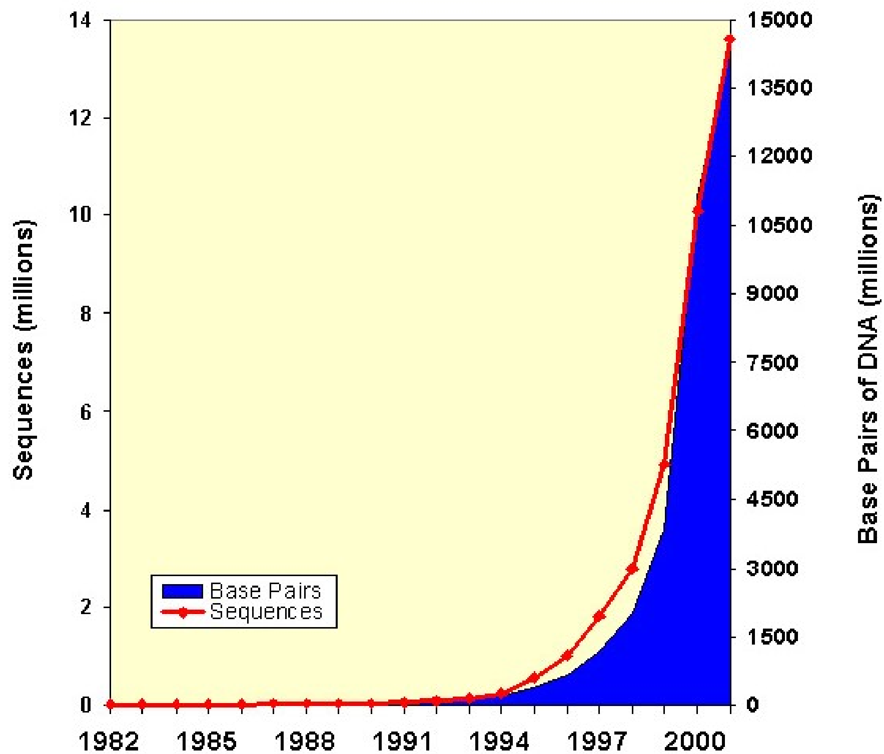
Proteome: 100-400.000 proteins

>1.000.000 interactions



Sekvenování genomů

GenBank vznikla v roce 1982 z Los Alamos Sequence Database



Walter Goad

Proč sekvenovat dál?

- Komparativní genomika
- Biomedicínský výzkum
- Osobní genom

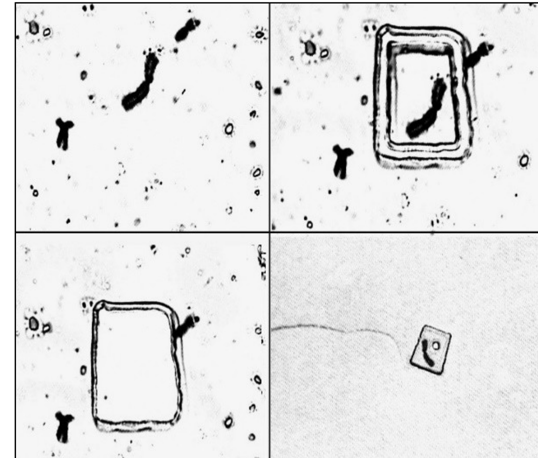
2010   Ideální lidský genom sekvenován

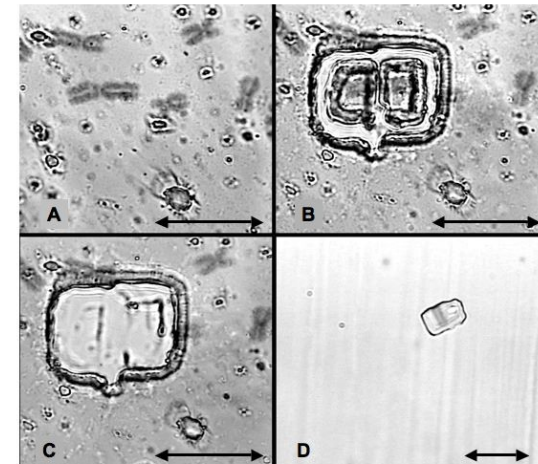
2010 Ideální lidský genom sekvenován

Laserová mikrodisekce

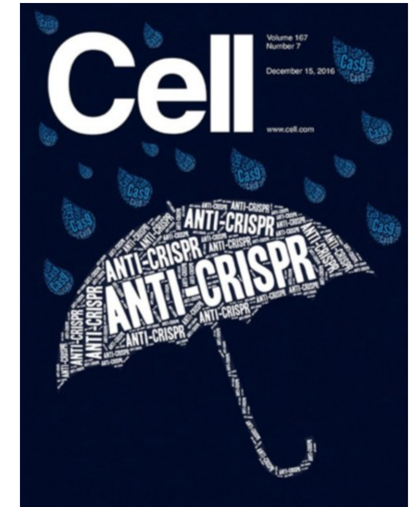
<u>Výhody:</u> vysoká čistota

<u>Nevýhody:</u> malý počet chromozomů, pracnost





Nová GMO revoluce - Molekulární nůžky CRISPR



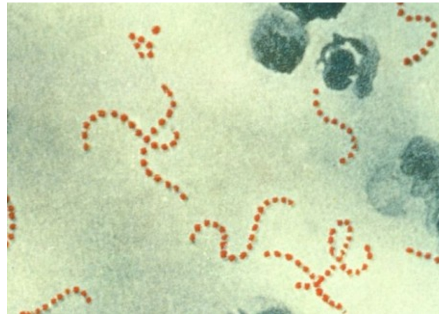
Co je to CRISPR/Cas systém?

- **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats
- Prokaryotický imunitní system, který brání buňku proti cizí DNA

Promotor 1 | Cas9 | Promotor 2 | guideRNA

Nukleáza

Naváděcí RNA



Streptococcus pyogenes
Cas9 donor

Cas9
gRNA
Target
PAM
dsDNA
Cleavage

Cílová DNA

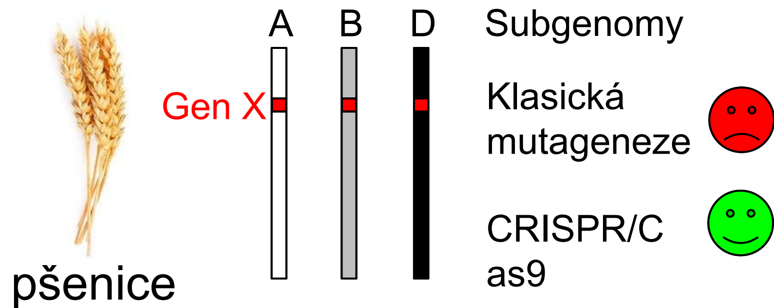
© Wikipedia

Mutace

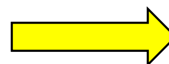
Proč CRISPR?

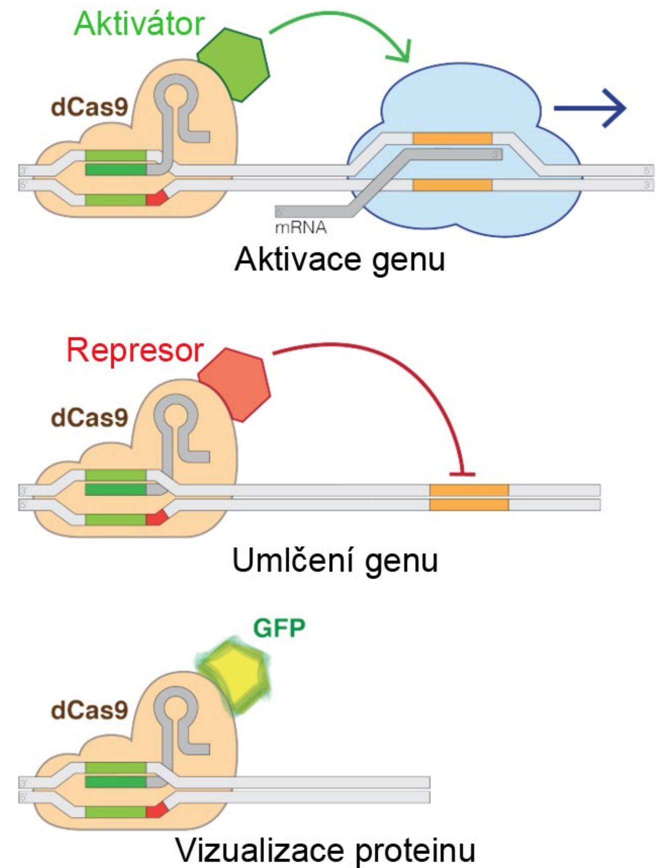
- Rychlost, přesnost, cena

- Vícenásobné mutace



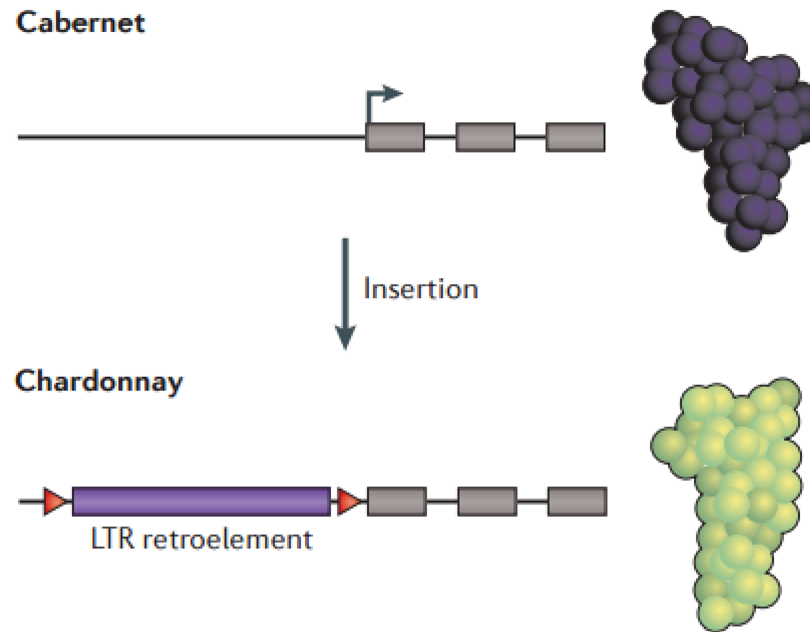
- Modifikace systému na „molekulárního poslíčka“





© Wikipedia, modifikováno

GMO versus evoluce



Bílé víno vzniklo před 7 tis. lety inzercí transpozonu do genu pro antocyan u původního červeného vína

V ČR klesá plocha osetá GMO plodinami

8 380 ha v roce 2008
6 480 ha v roce 2009
4500 ha v roce 2013

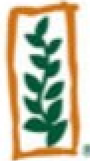
Kolik se vlastně GMO plodin pěstuje?

USA

- sója 94%
- bavlna 90%
- řepka 90%
- cukrová řepa 95%
- kukuřice 88%

GMOs: The Big Six

MONSANTO



syngenta

CHEMCHINA
syngenta

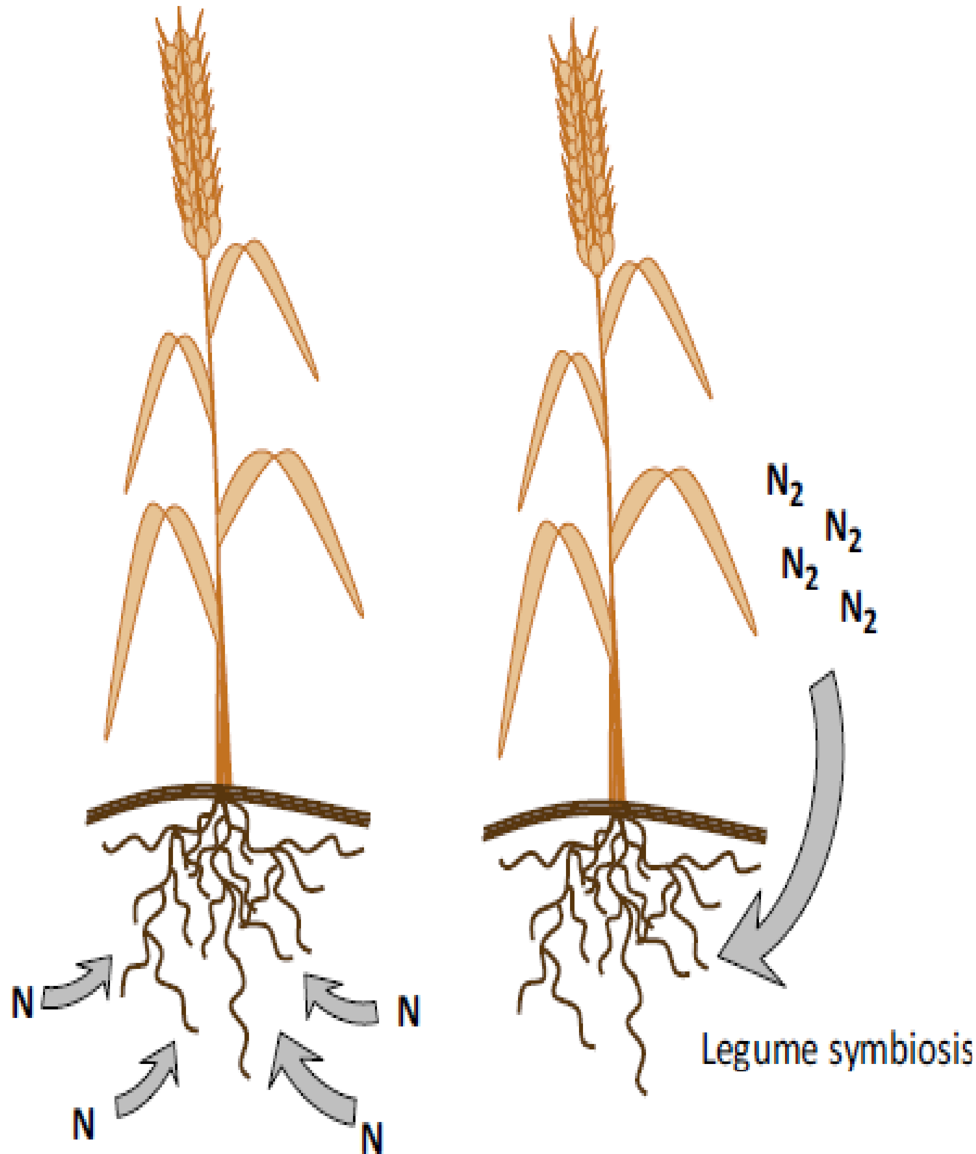
BASF

Vkládání cizorodých genů

Kukuřice odolná vůči zavíječi

Bacillus thuringiensis Bt delta endotoxin

"Syntetická" pšenice



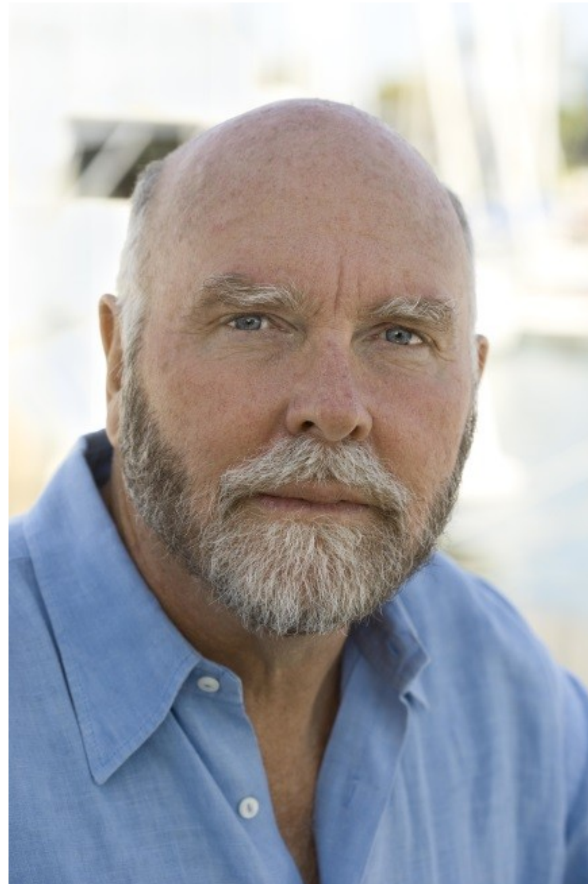
Symbióza mezi pšenicí a bakterií

1) odstranit geny rezistence vůči bakterii z genomu pšenice

2) do genomu vložit geny zodpovědné za symbiotické interakce

Craig Venter

Synthetic genomics



Synthia – umělý život (2016)

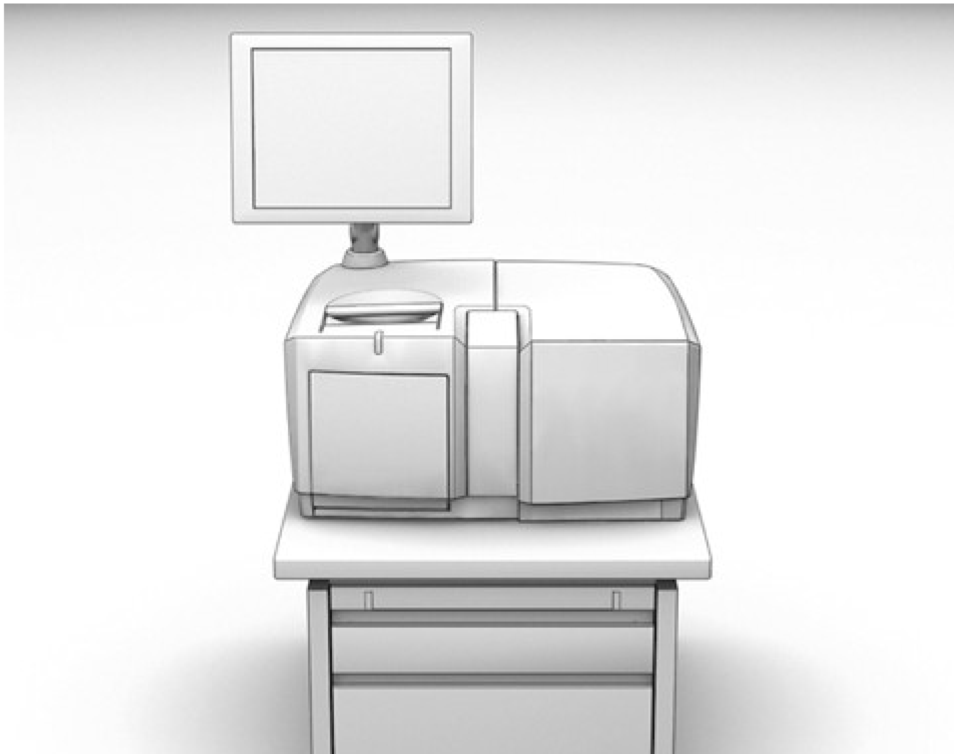
- Craig Venter: „první druh.... jehož rodičem je počítač... a je to také první druh, který má ve své DNA zapsán odkaz na své webové stránky“

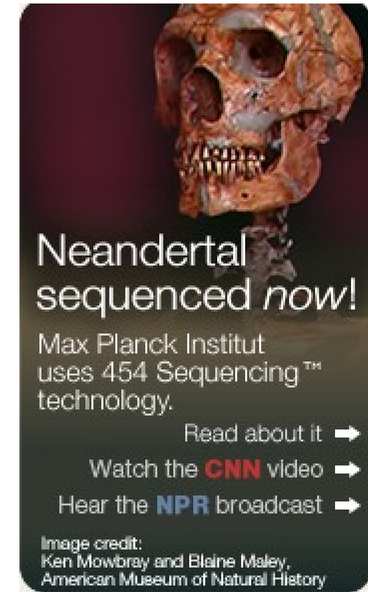
473 genů

  - *Richard Feynman: "What I cannot build, I cannot understand"*

Genome Sequencer 20 System
454 pyrosequencing (2005)

- http://www.454.com

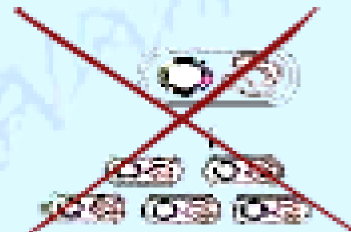




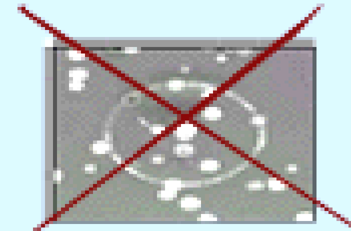
DNA library preparation

One sample preparation per genome

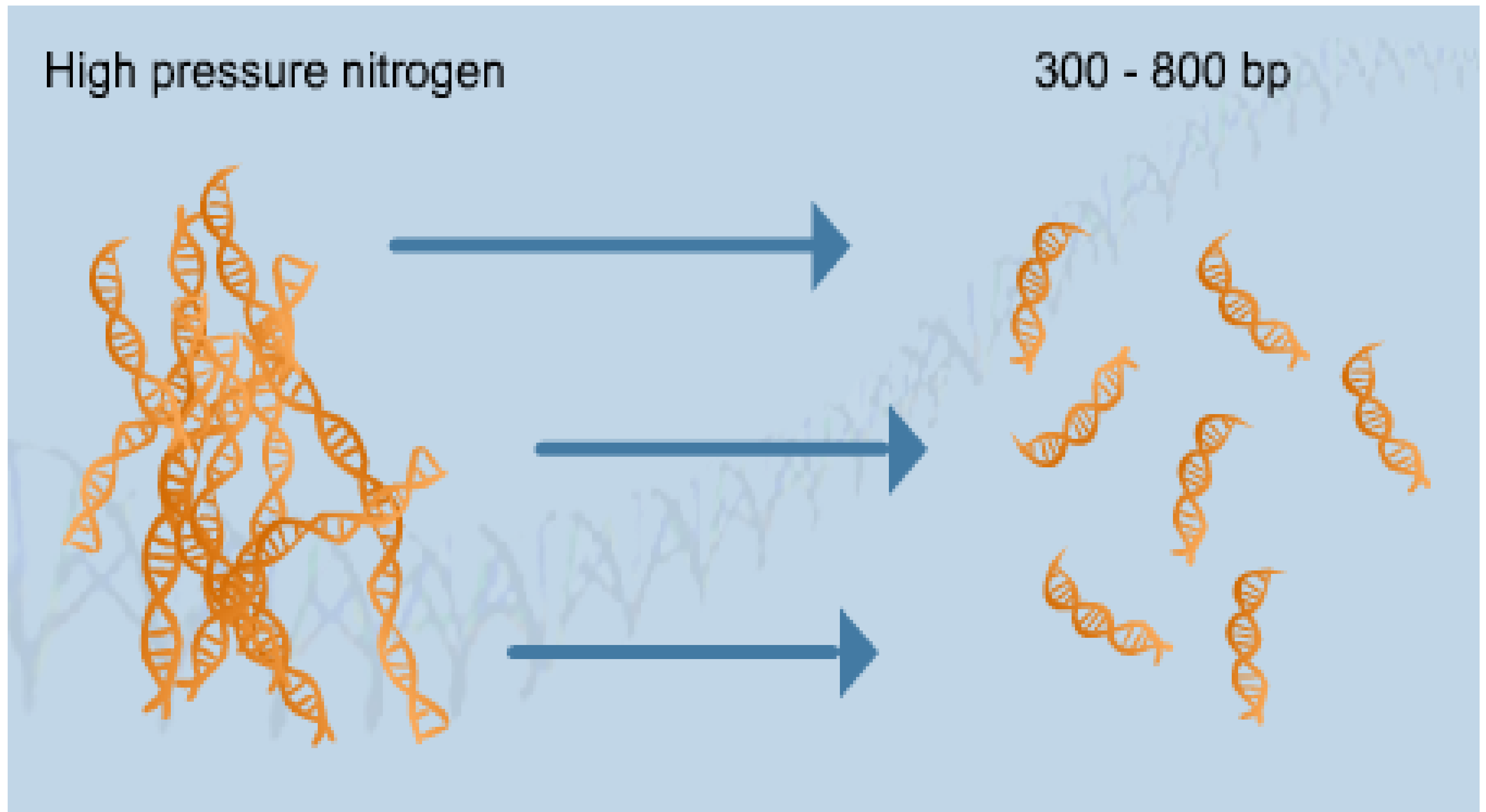
No Cloning



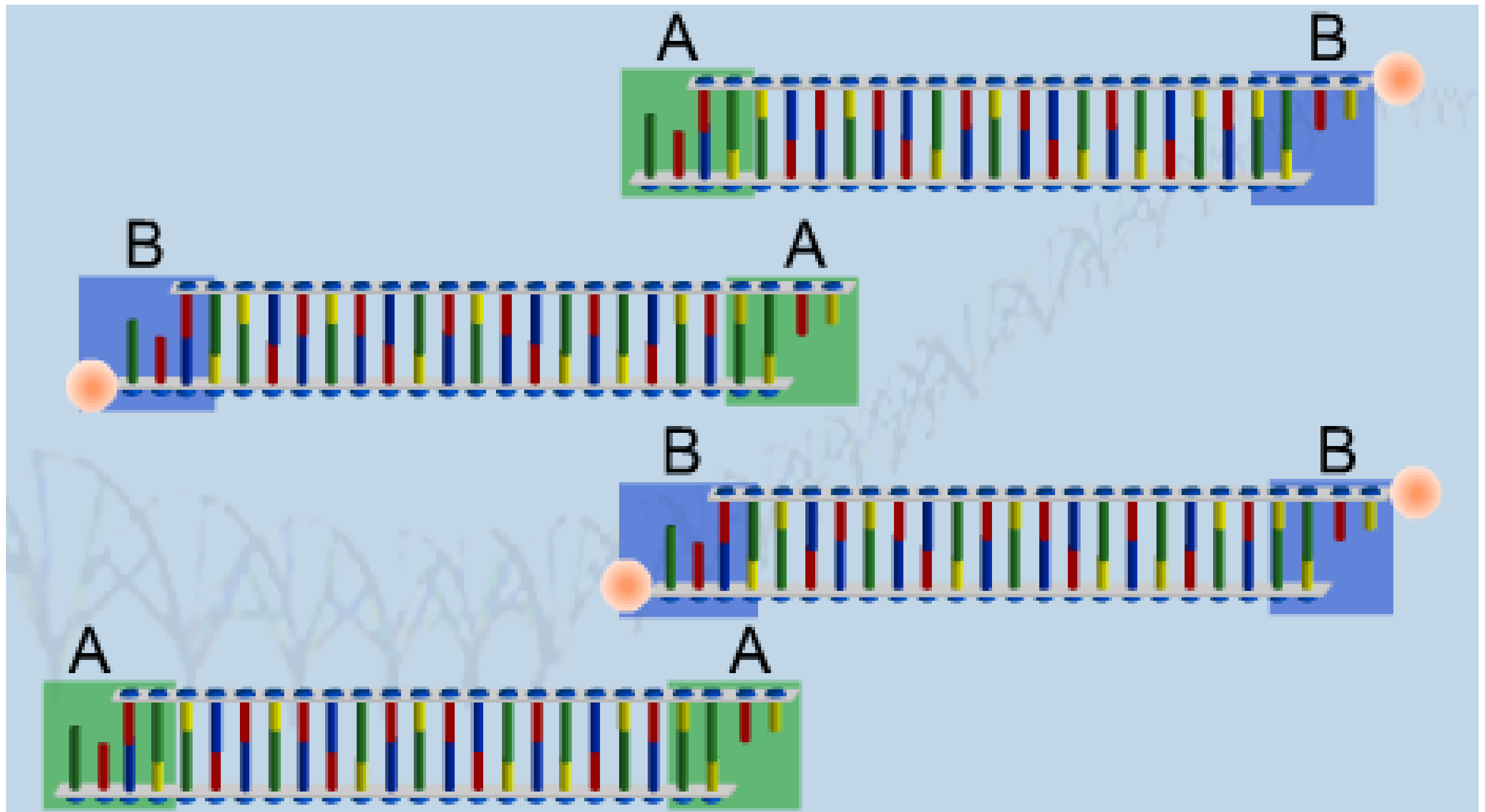
No Colony Picking



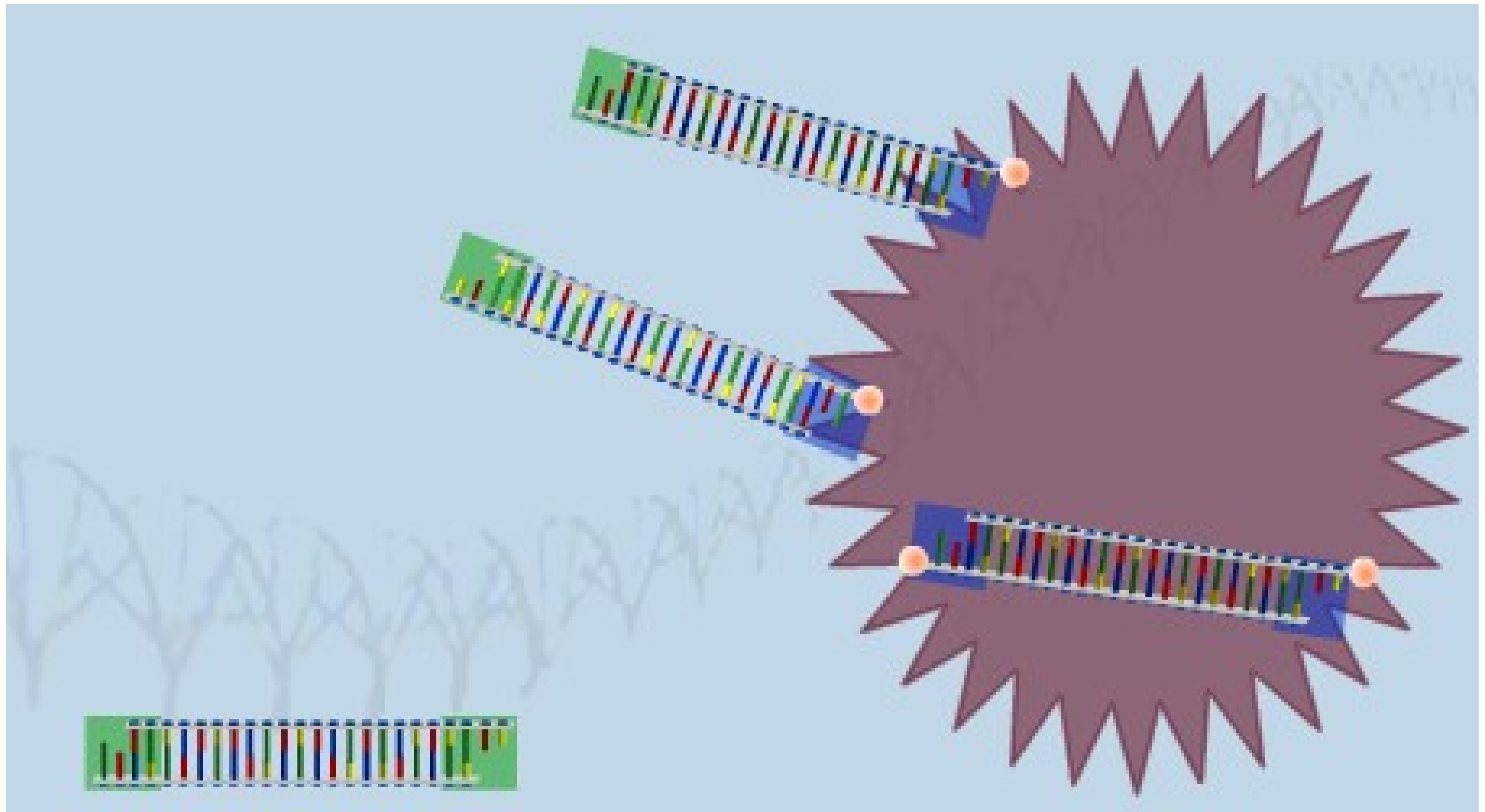
Fragmentace DNA



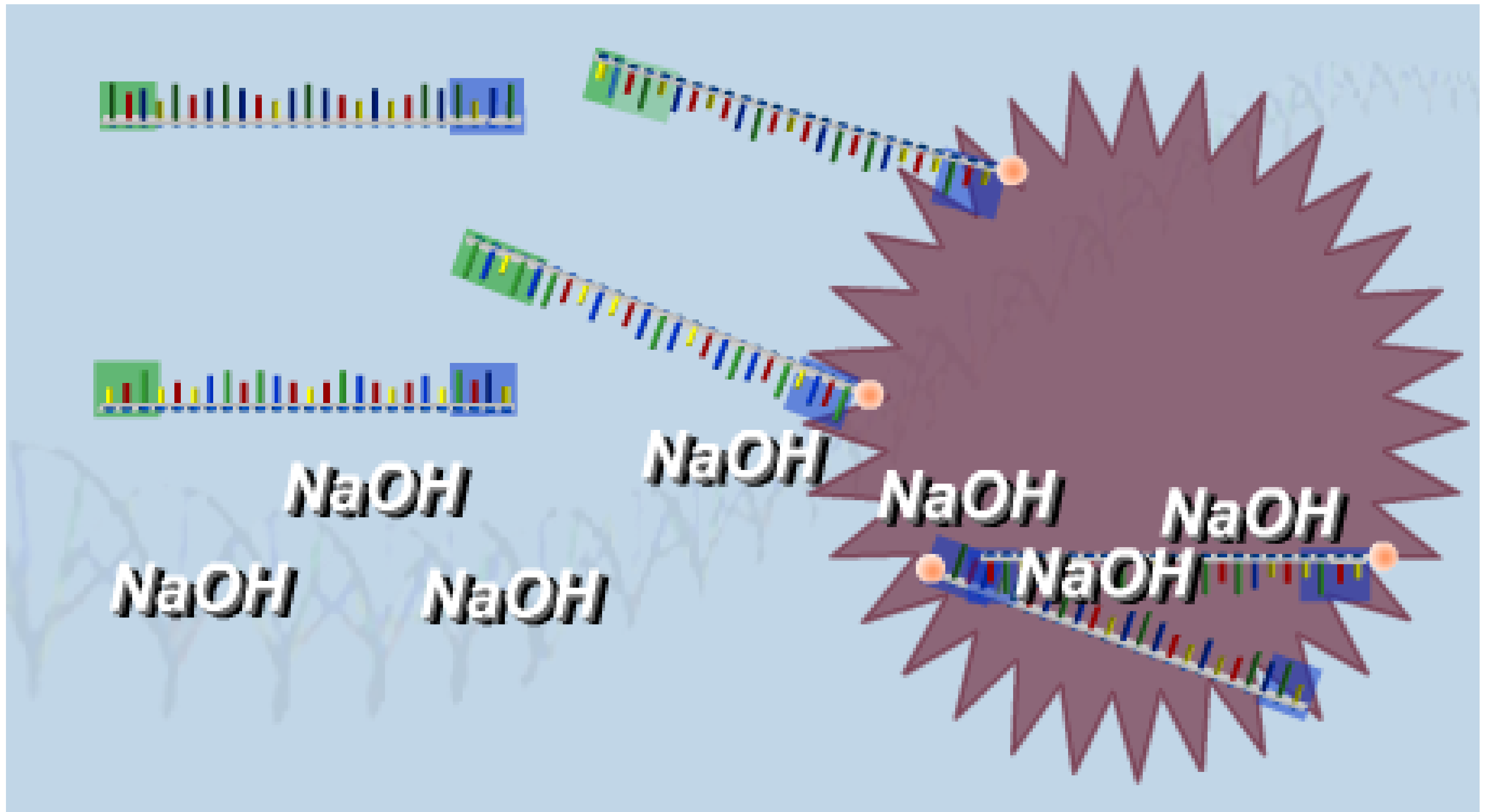
Ligace adaptoru

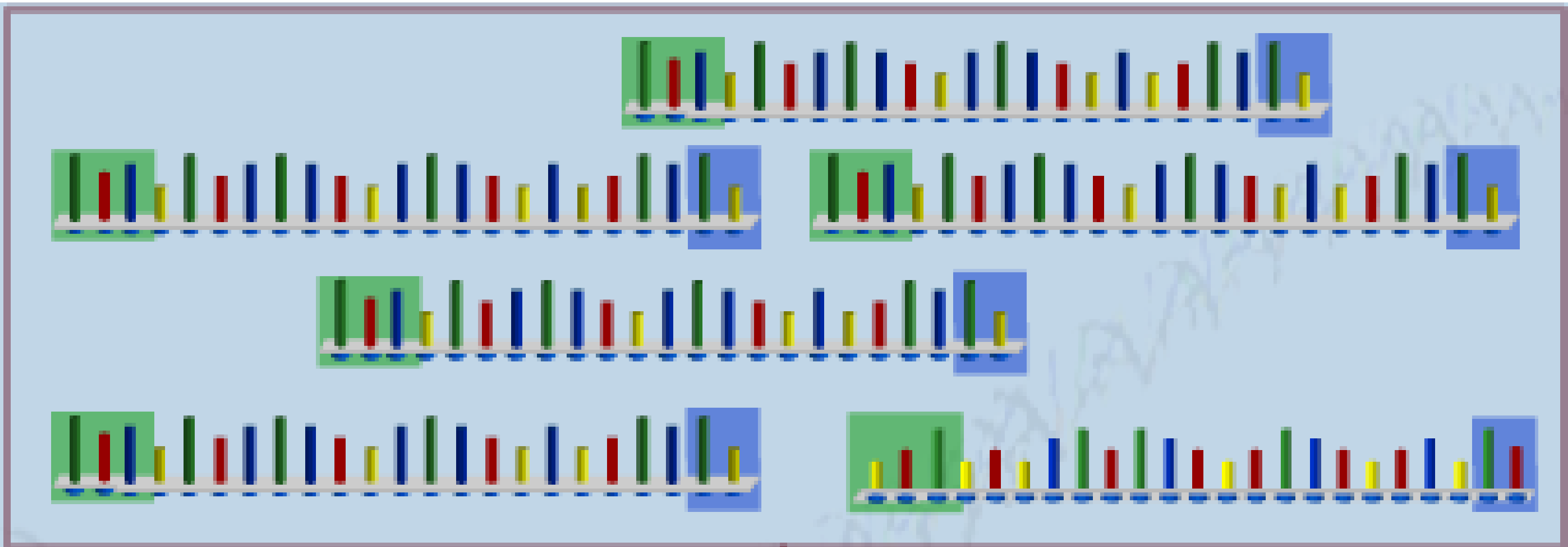


Vychytání DNA molekul

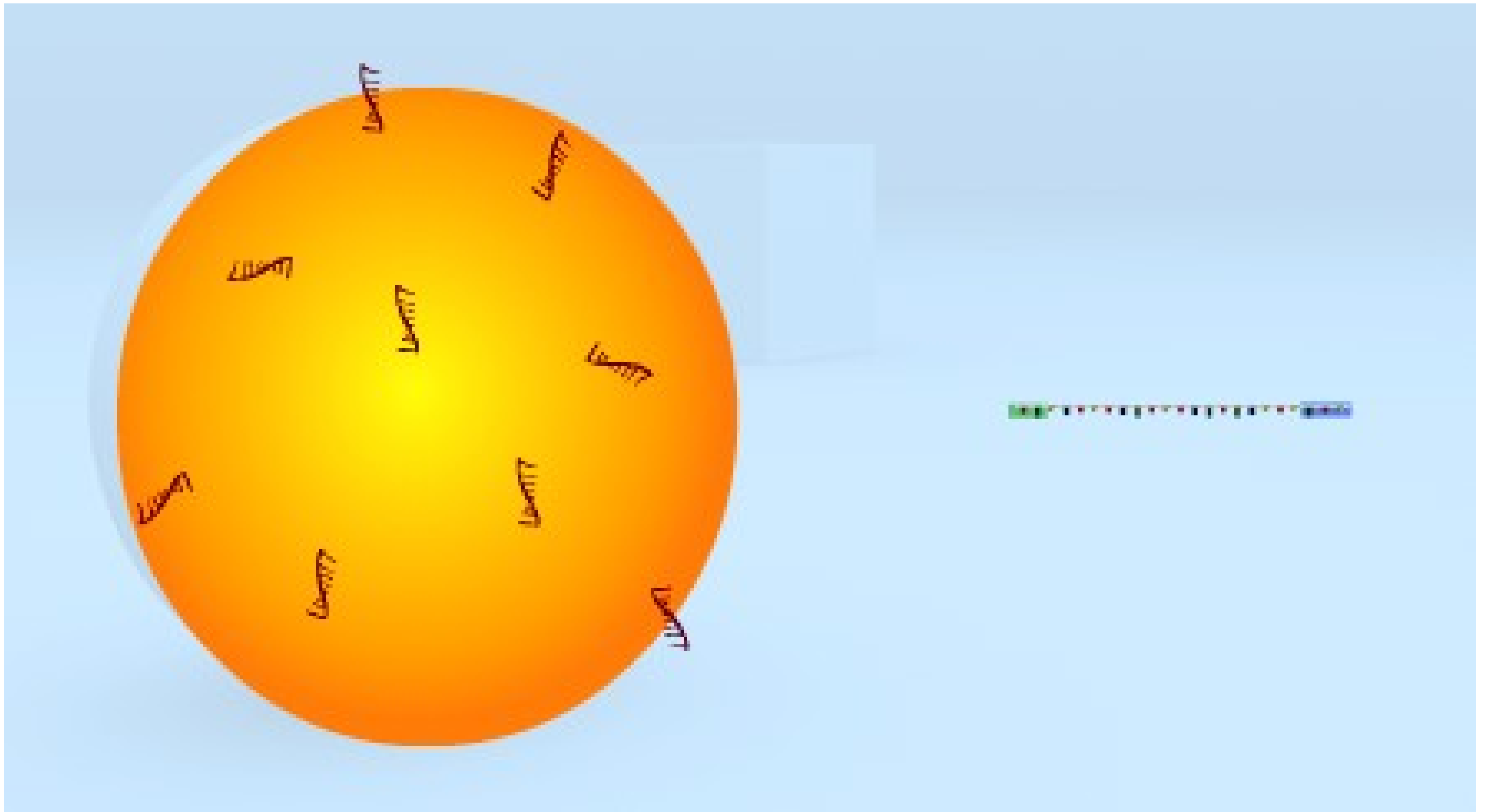


denaturace

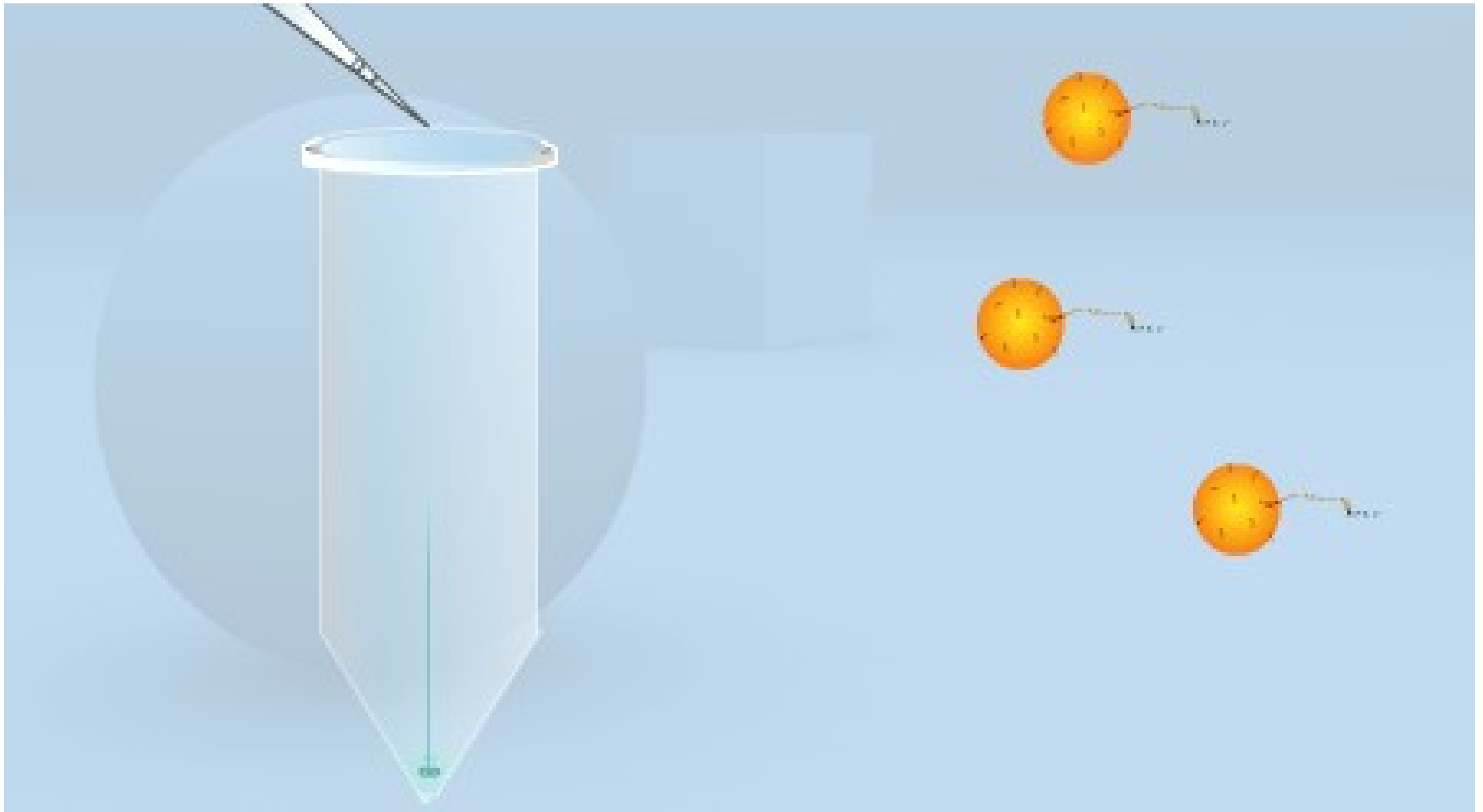



Single-stranded
template DNA
sstDNA

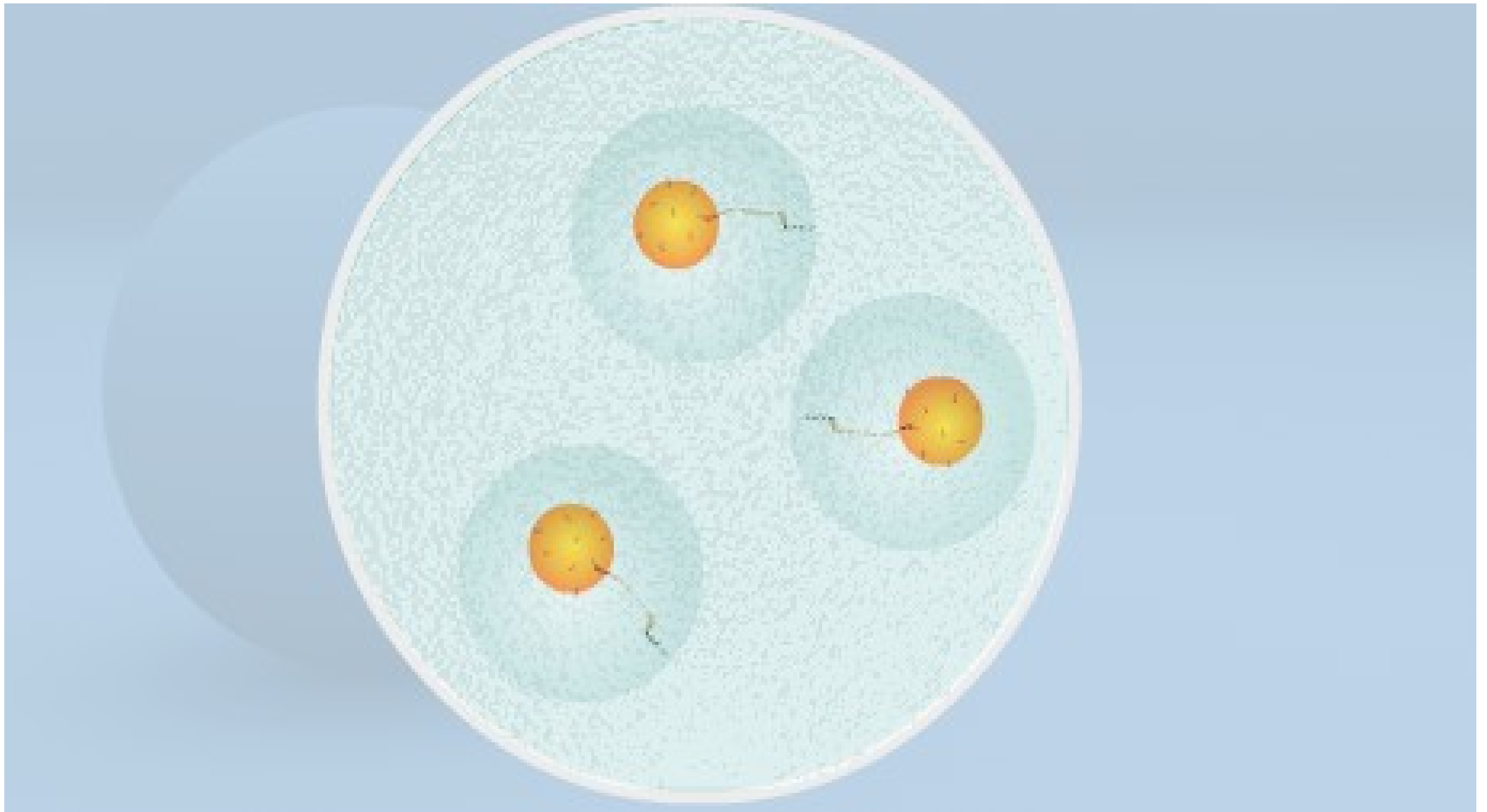
emPCR



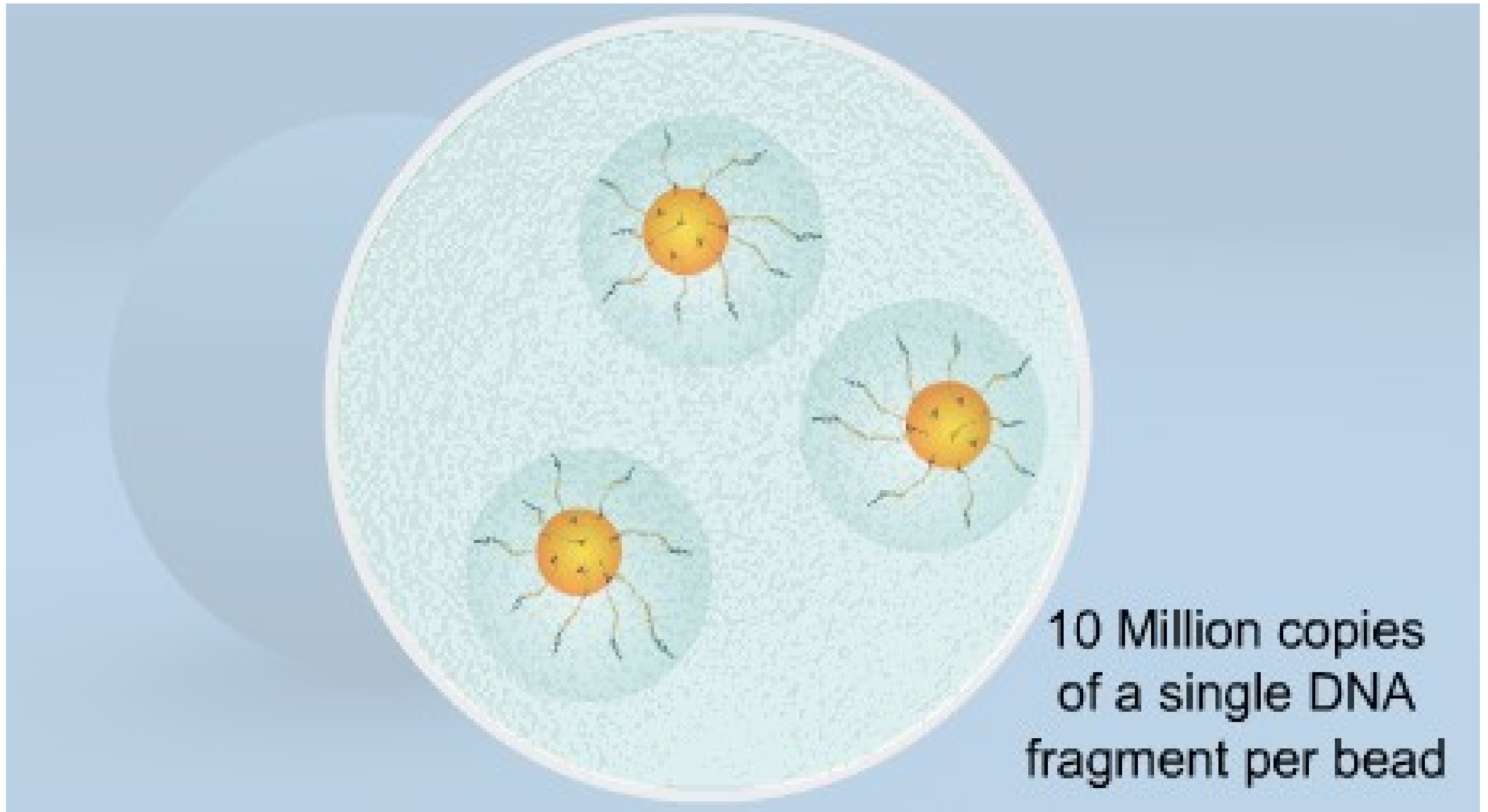
Vznik emulze (olej)



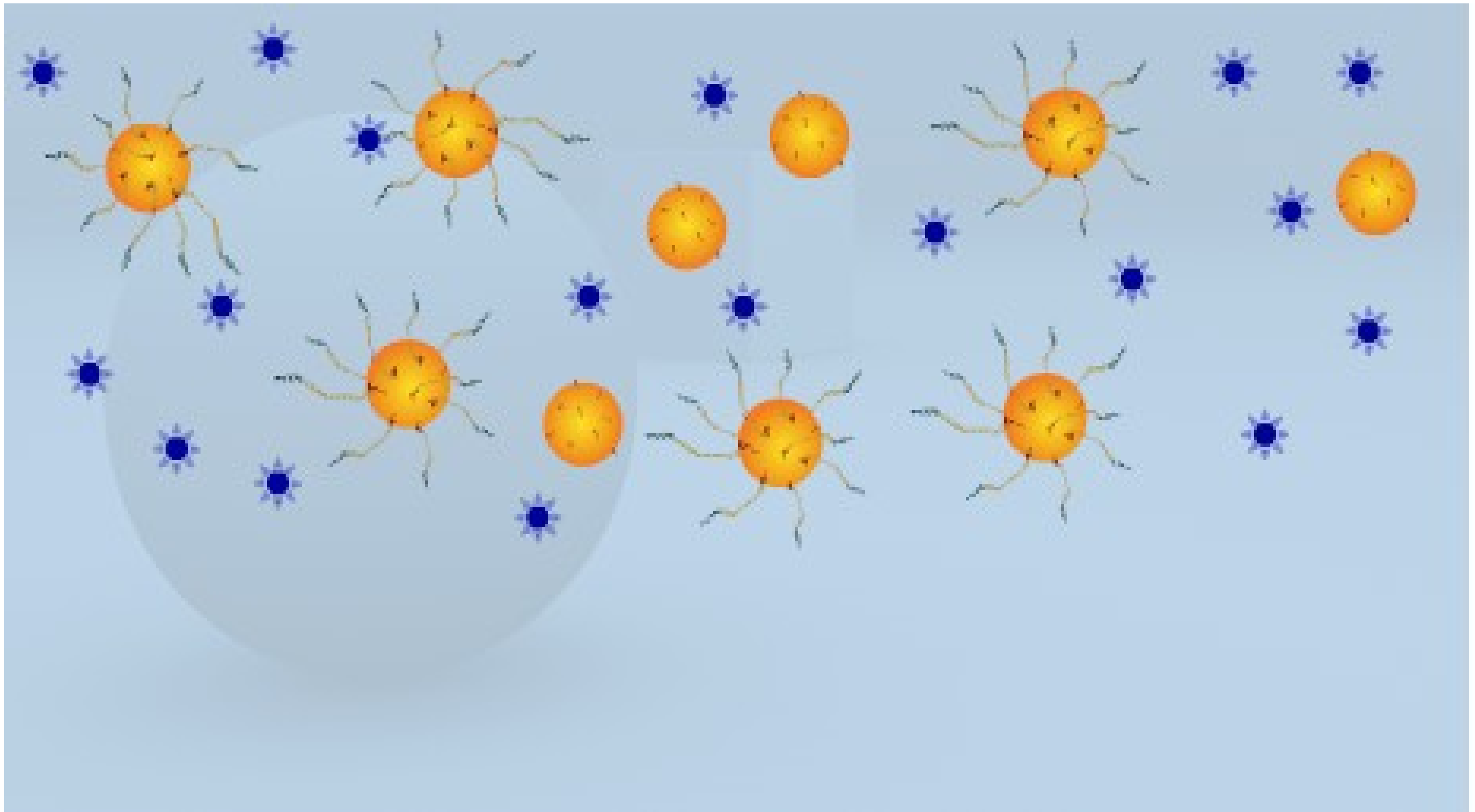
emPCR



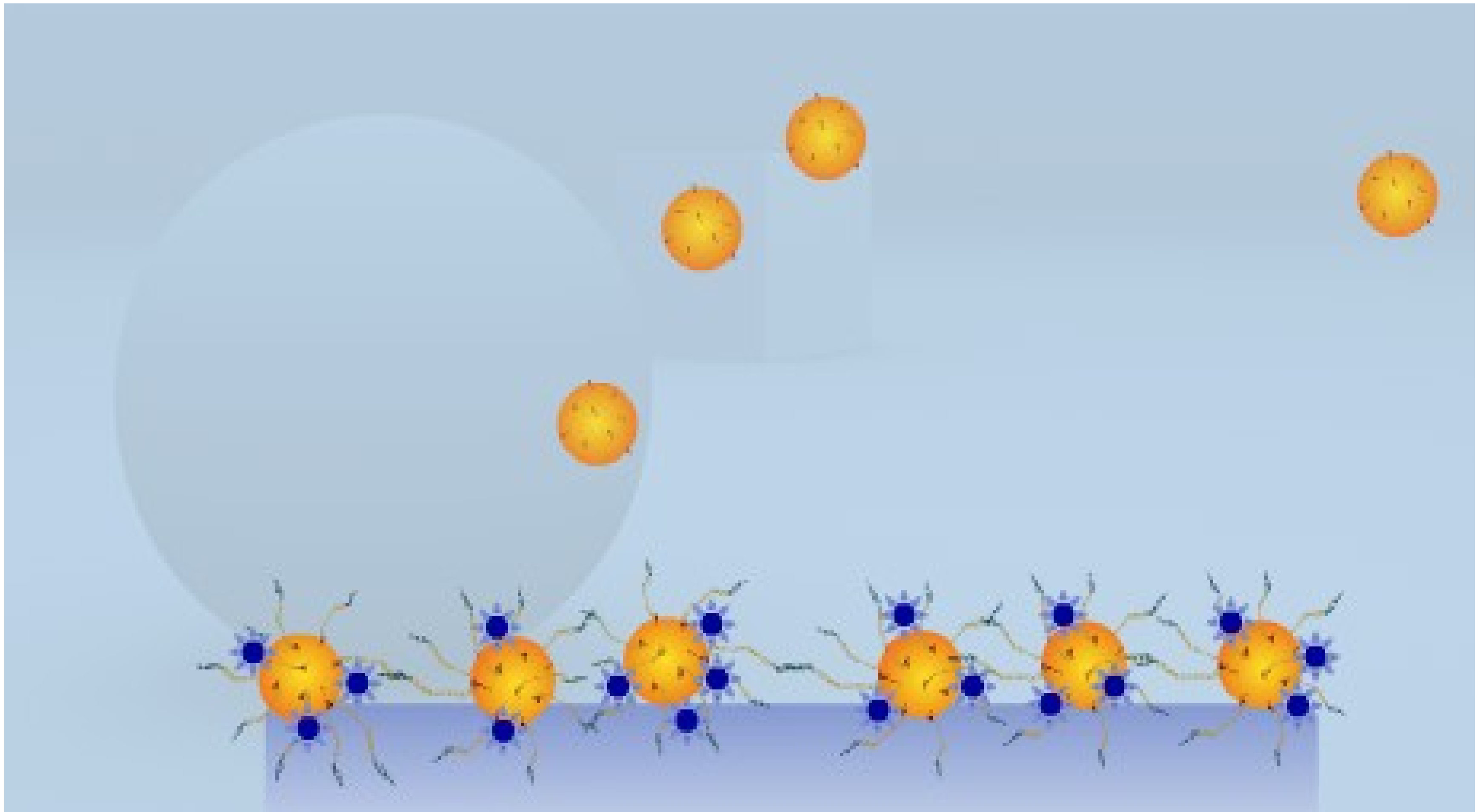
emPCR



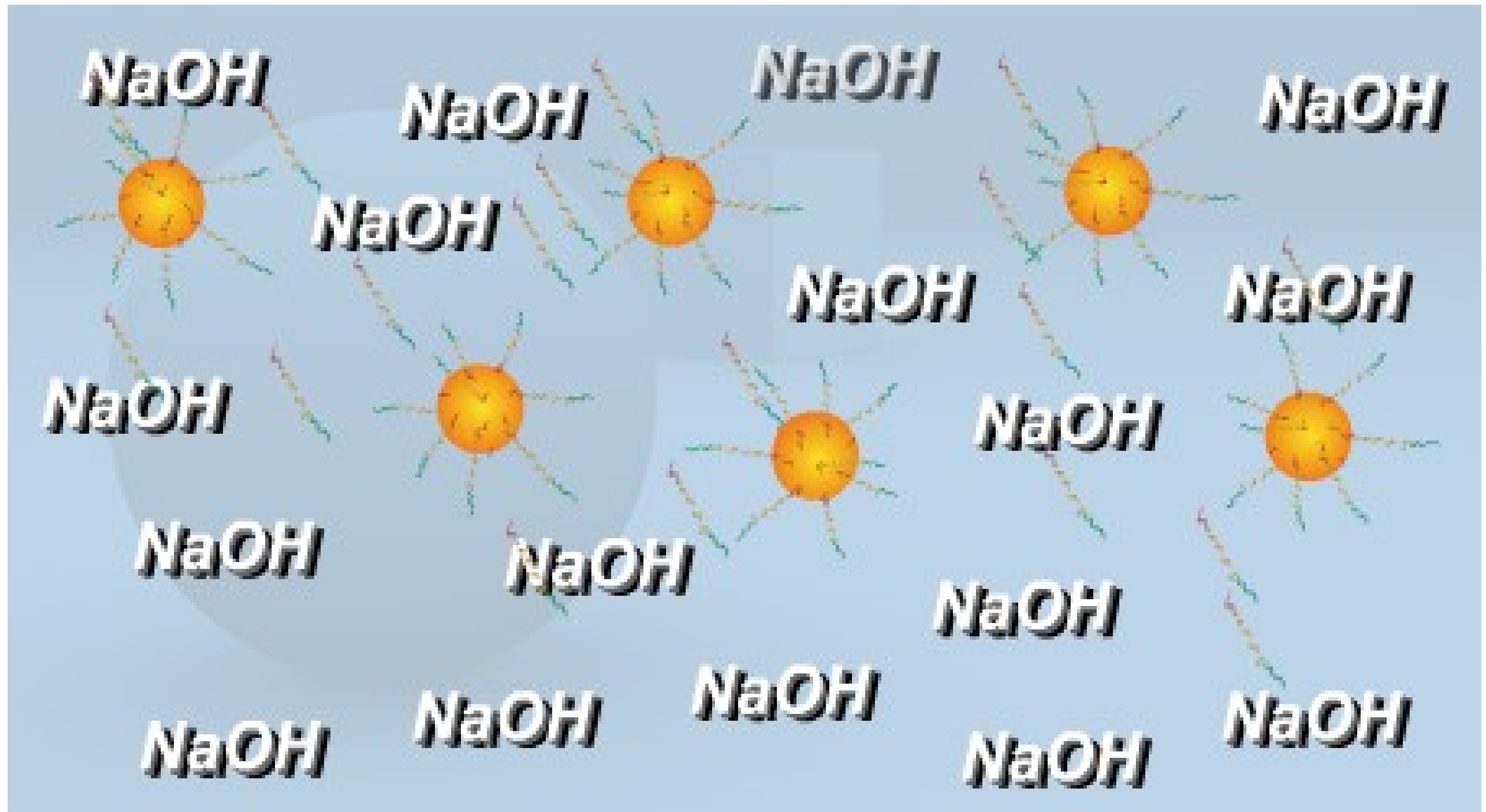
Vychytání kuliček



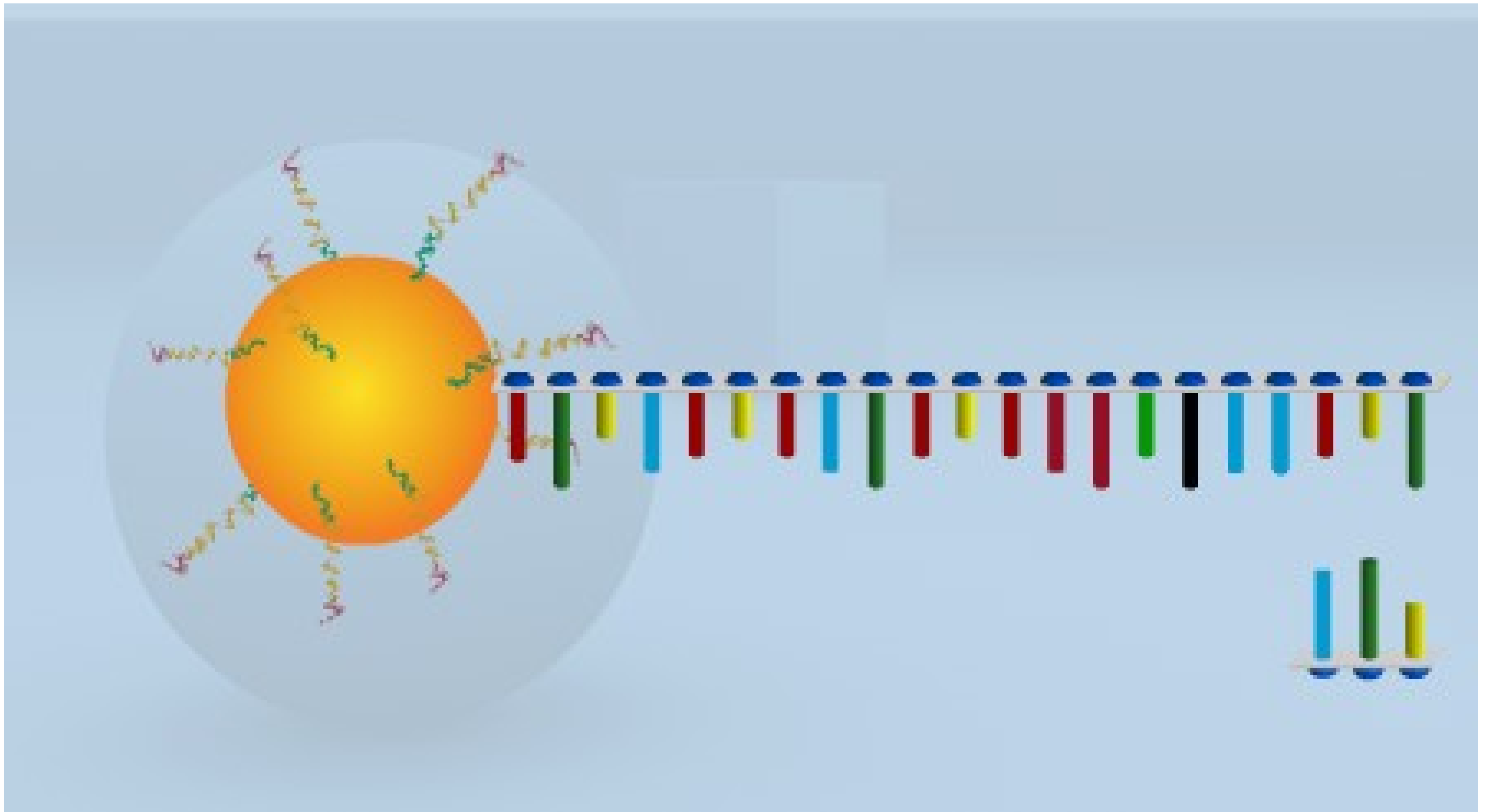
Vychytání kuliček



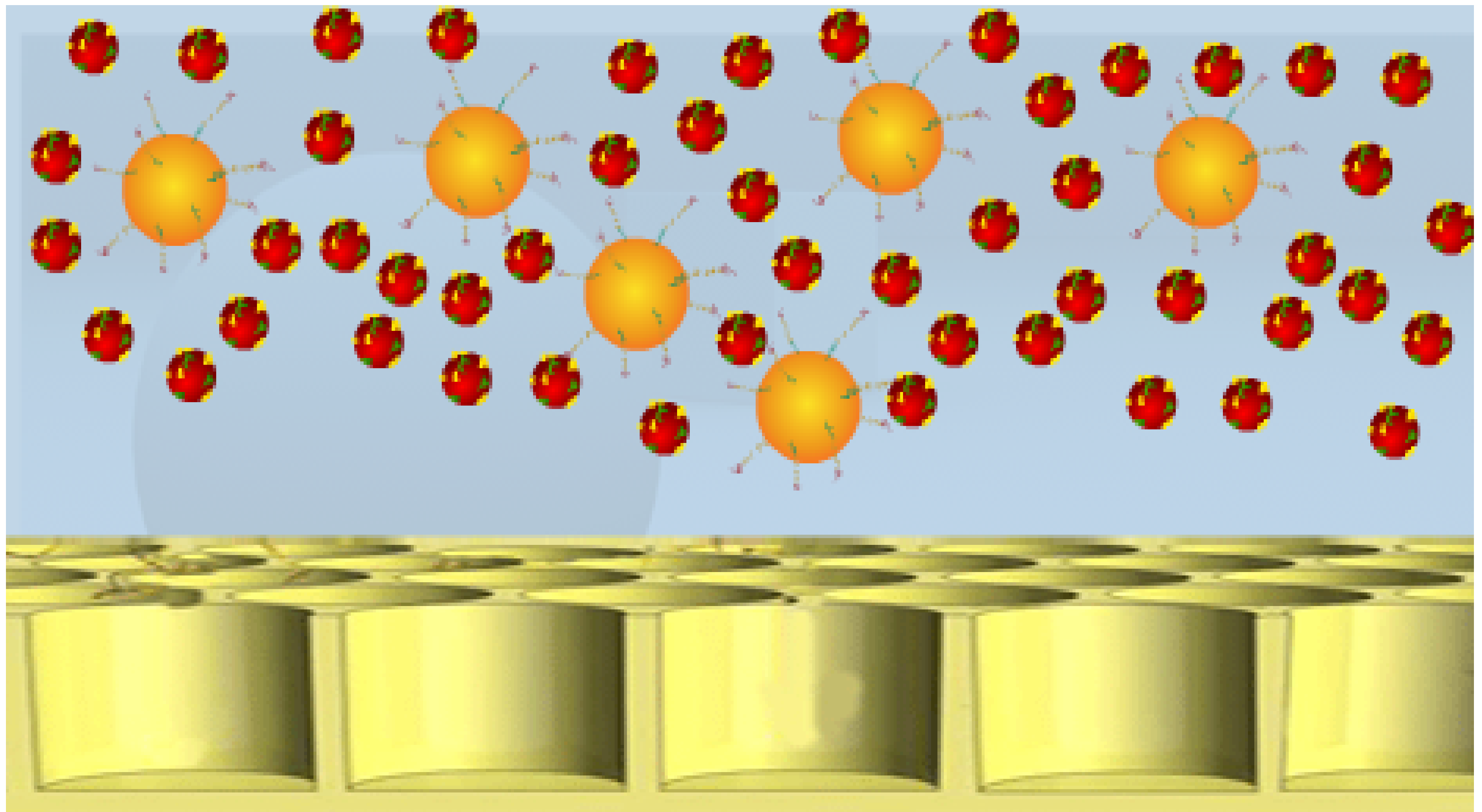
denaturace



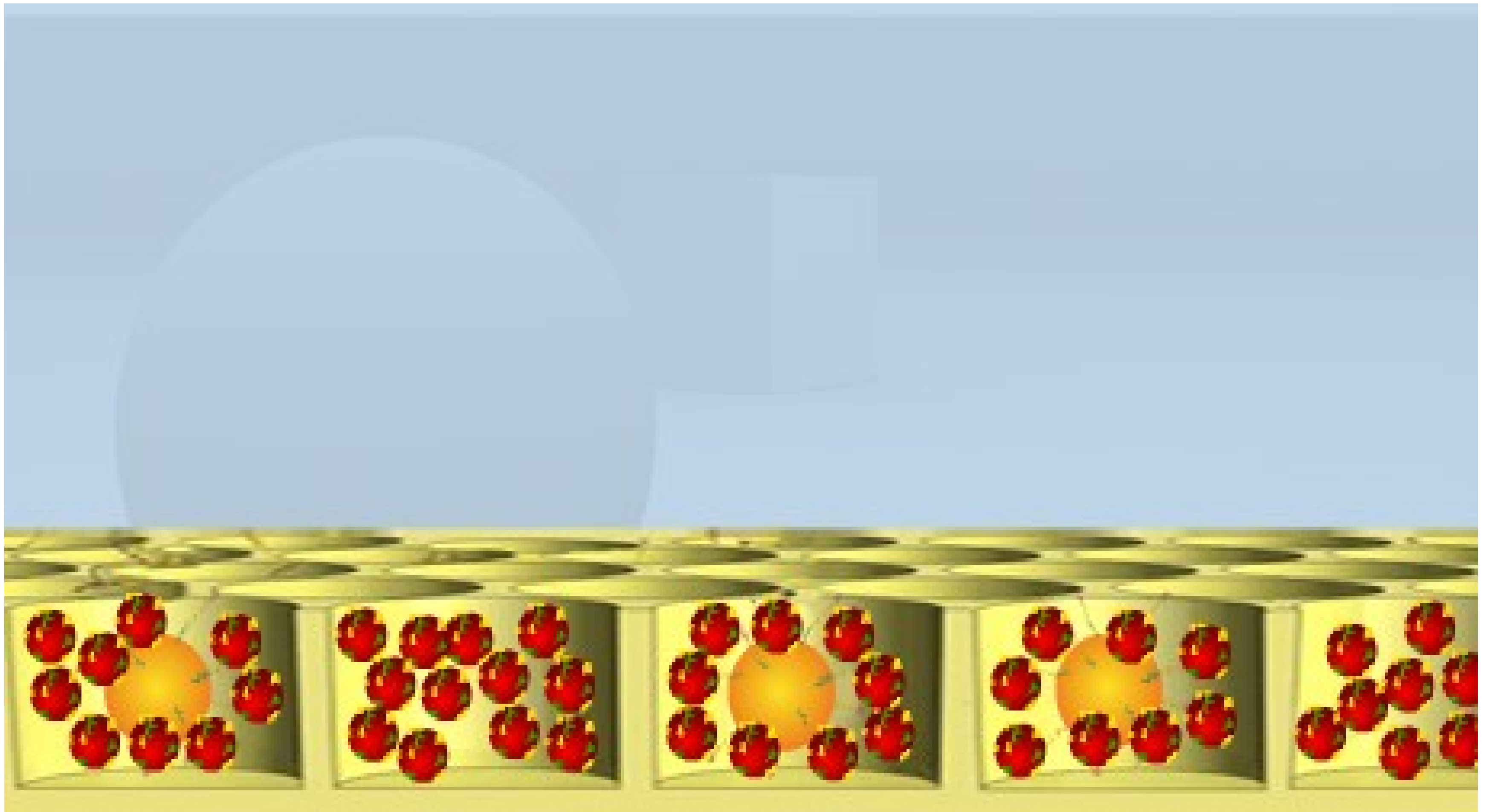
Sekvenační primer



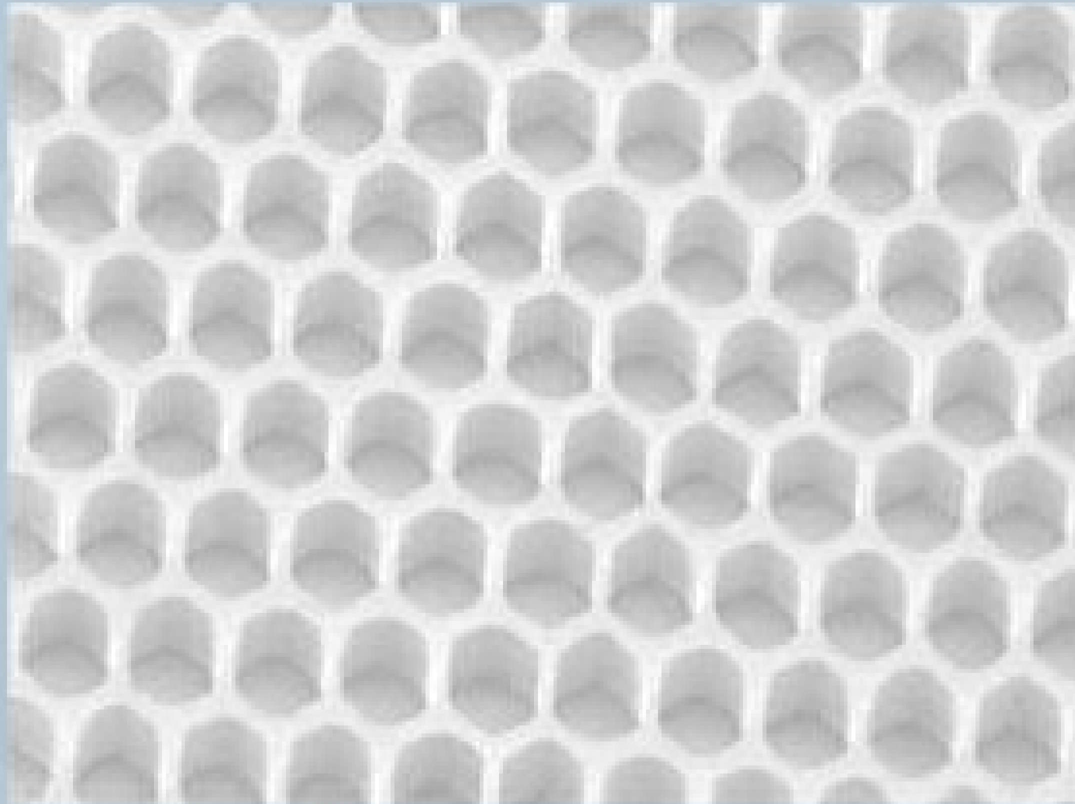
Disperze na sklíčko



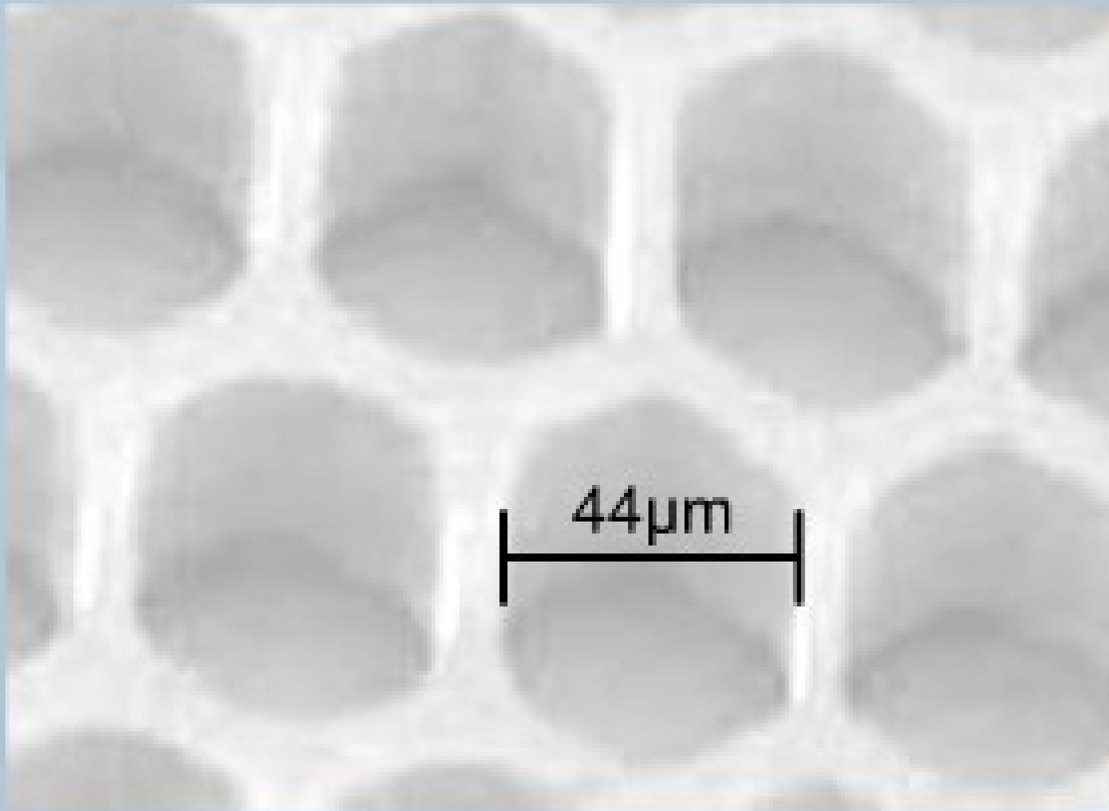
Disperze na sklíčko



Parametry mikroreaktorů

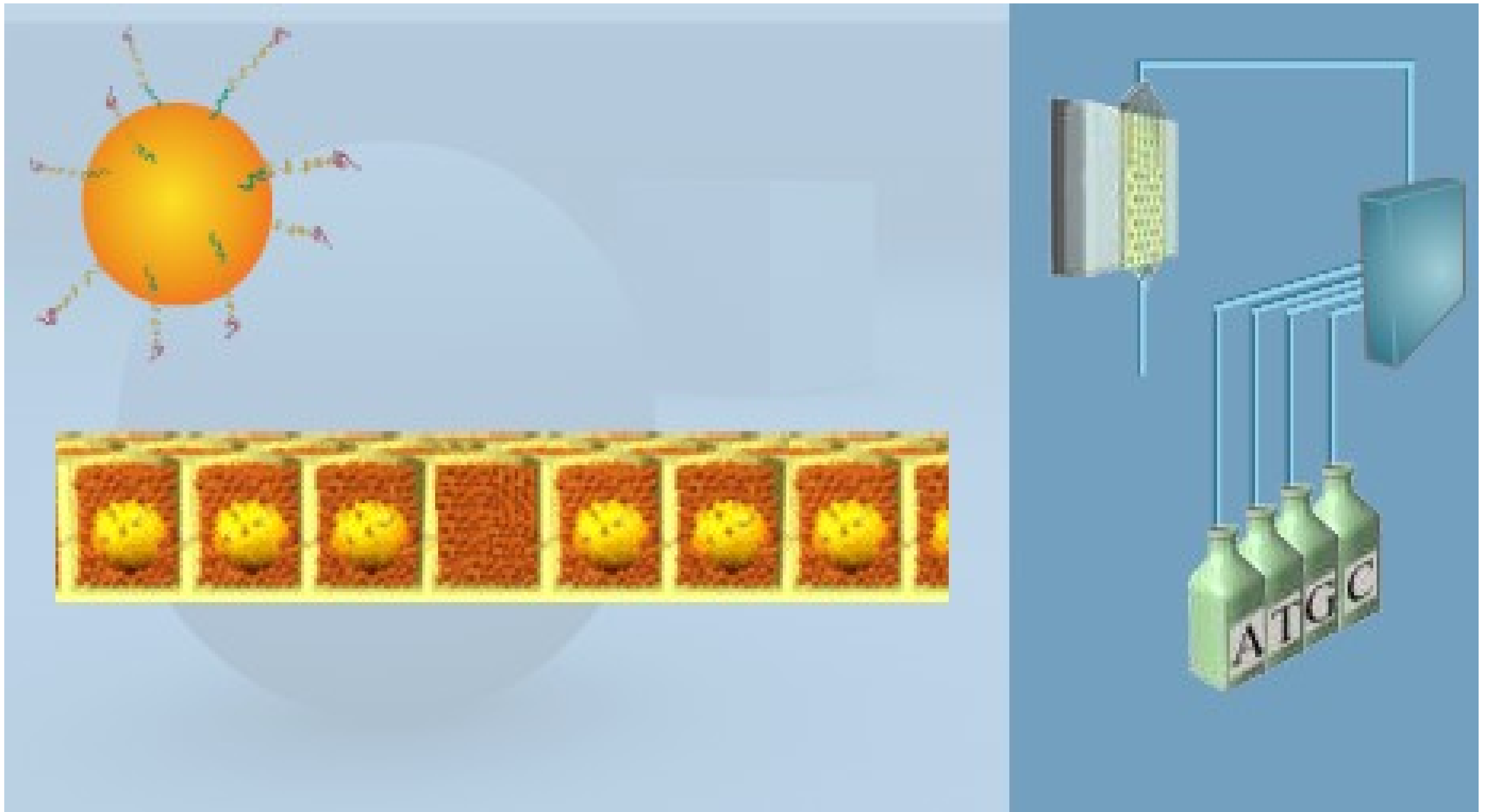


Parametry mikroreaktorů

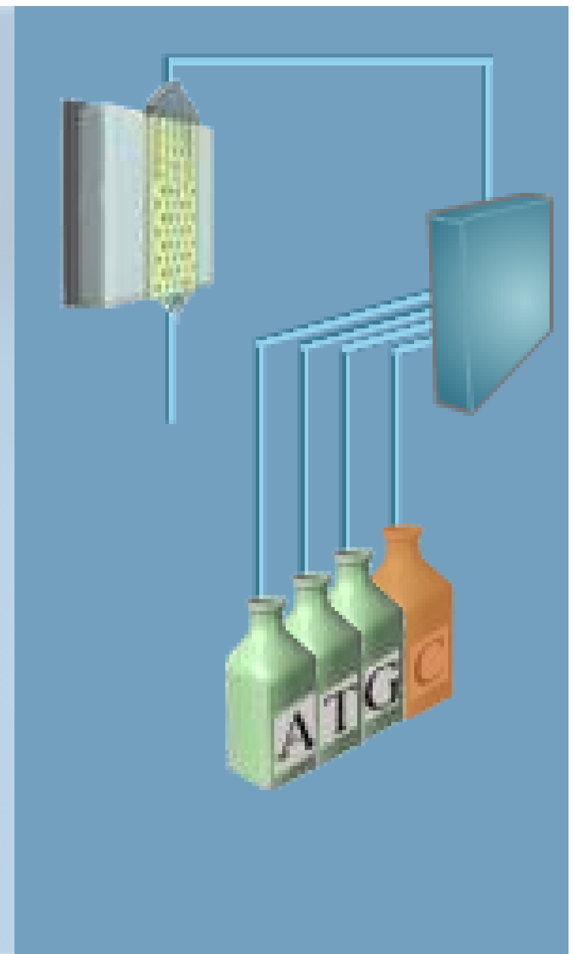


PicoTiterPlate device

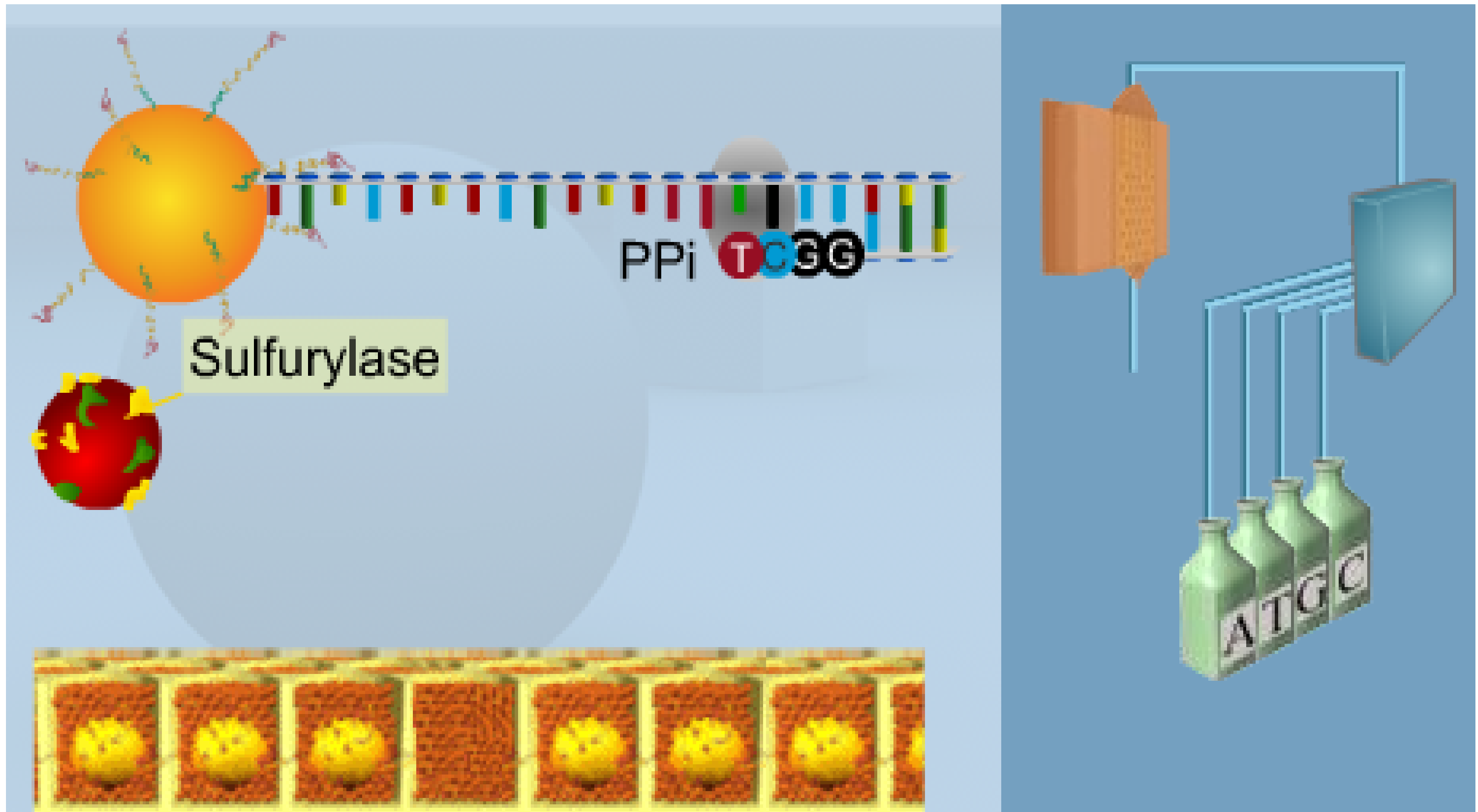
sekvenace



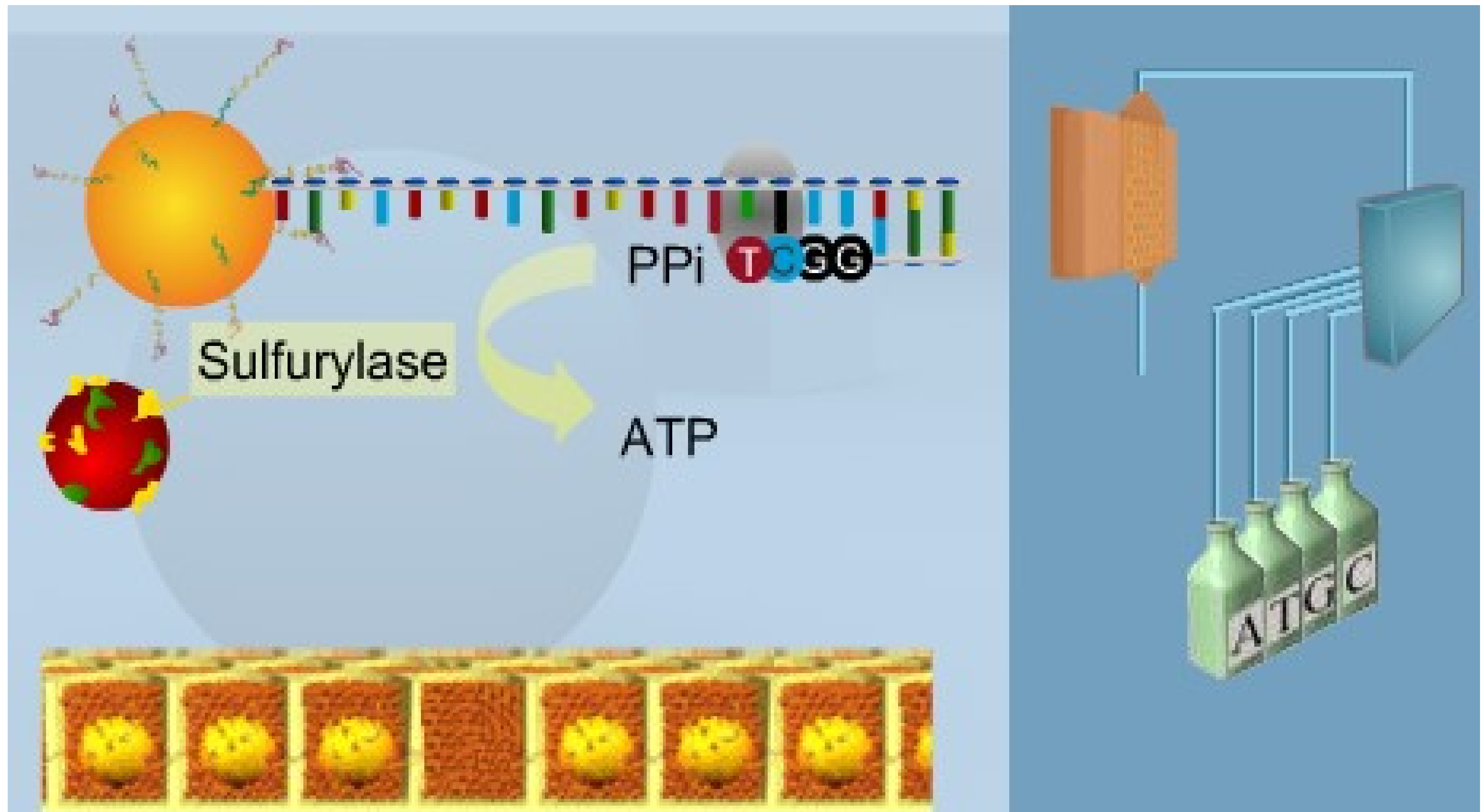
sekvenace



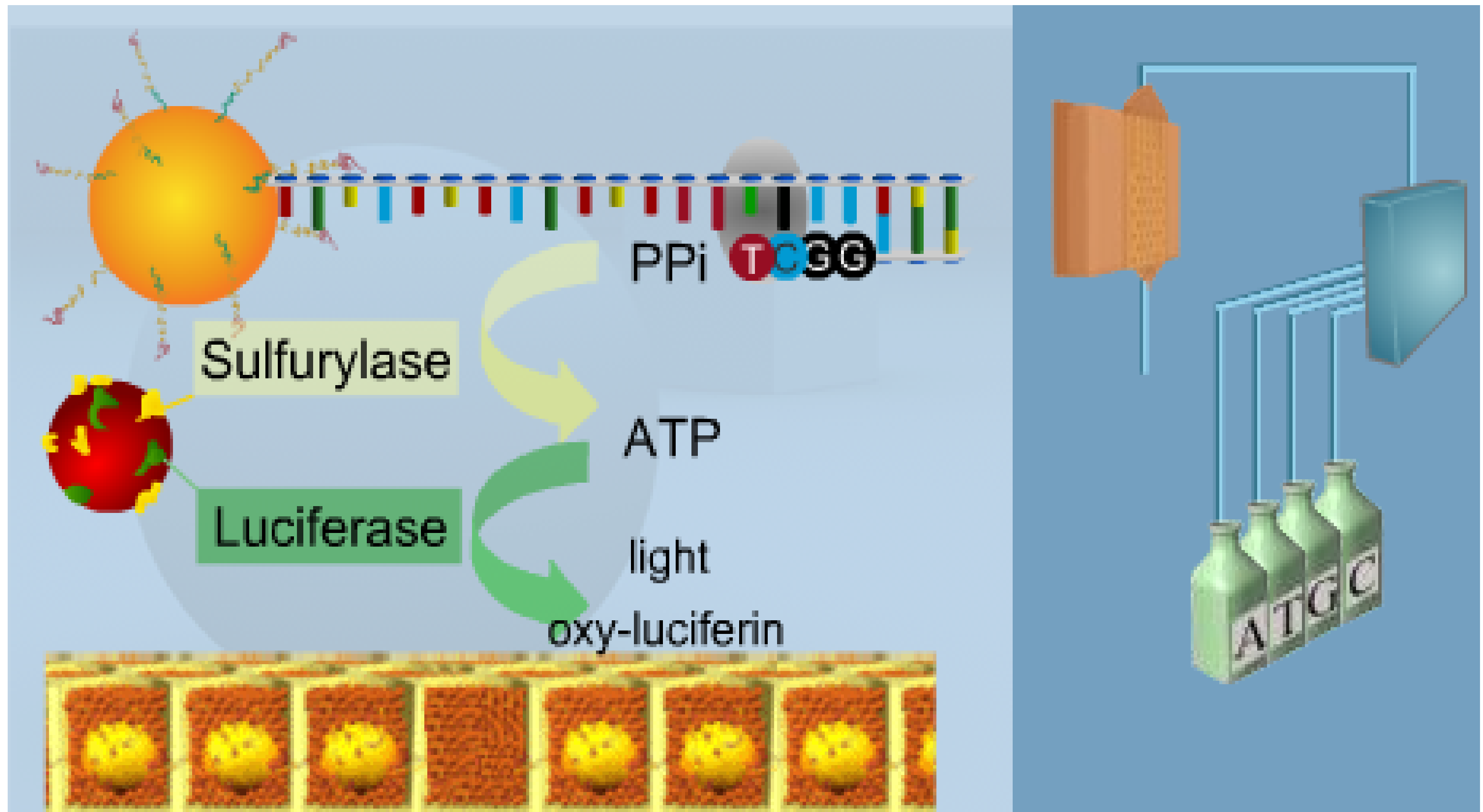
sekvenace



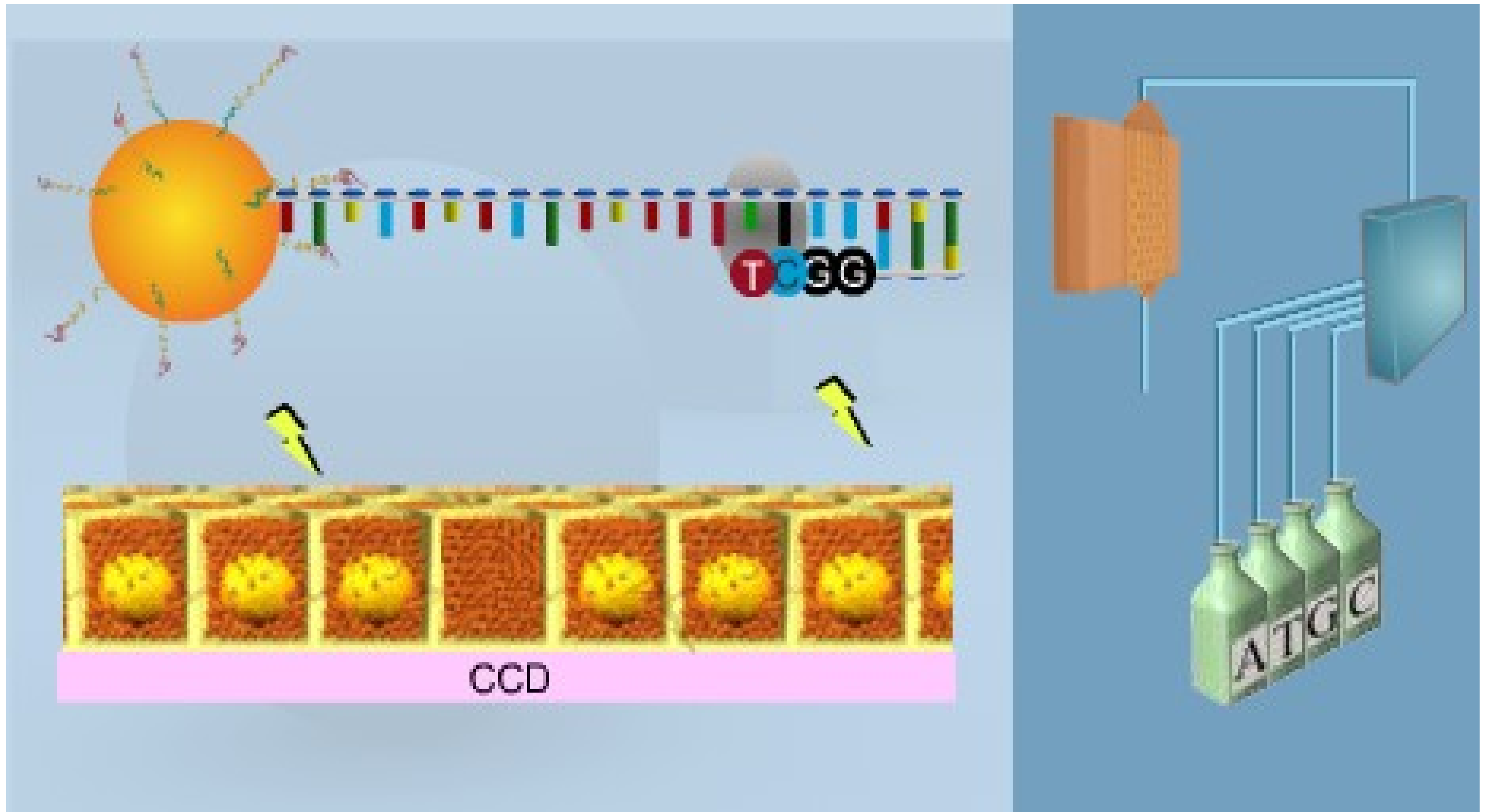
sekvenace



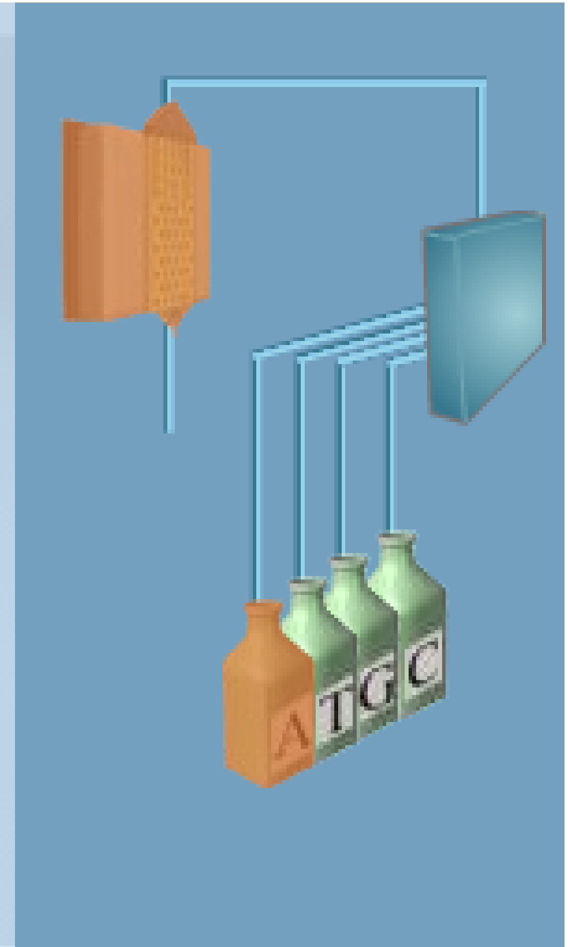
sekvenace



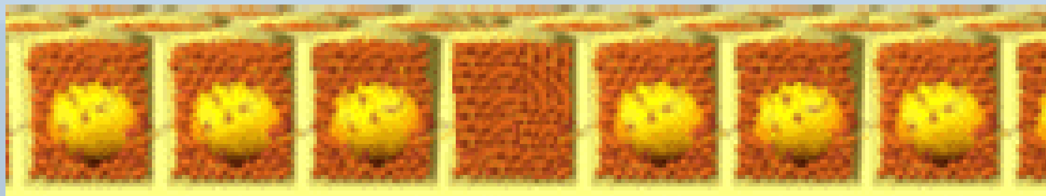
sekvenace



sekvenace



sekvenace



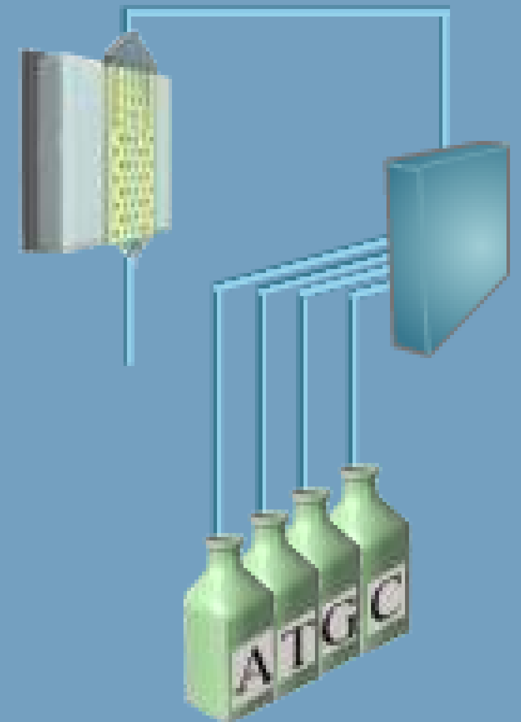
Massive parallelization of sequencing reactions

100 bases read length

X

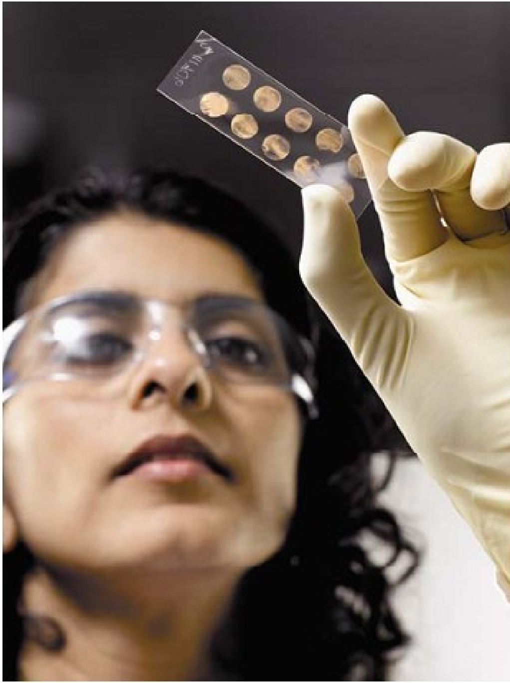
200 000

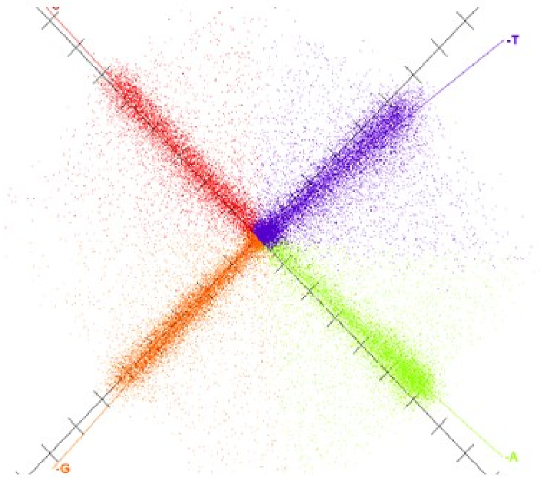
20 Million Bases



SOLID (Sequencing by Oligonucleotide Ligation and Detection)

2-base encoding sequencing (2007)



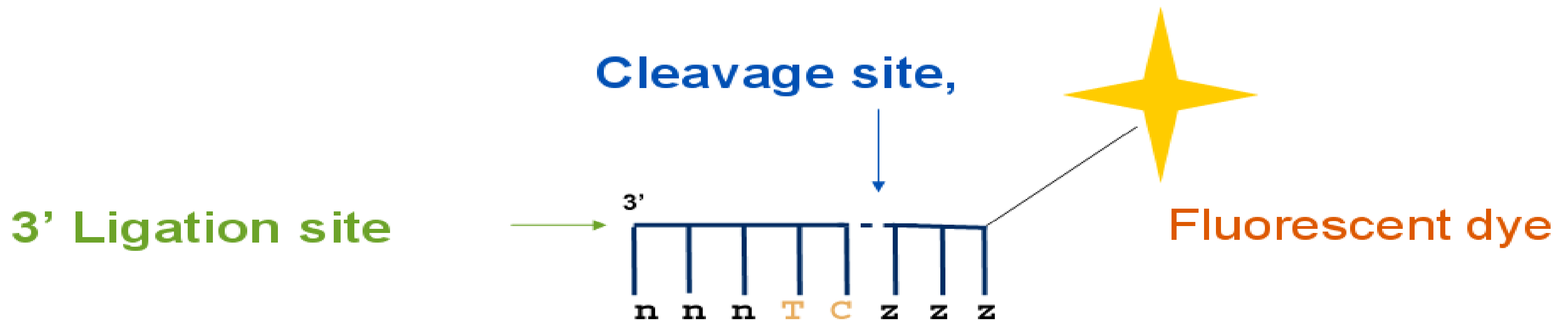


SOLiD™ System

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

Properties of the Probes

Spatial separation among dye, ligation & cleavage sites

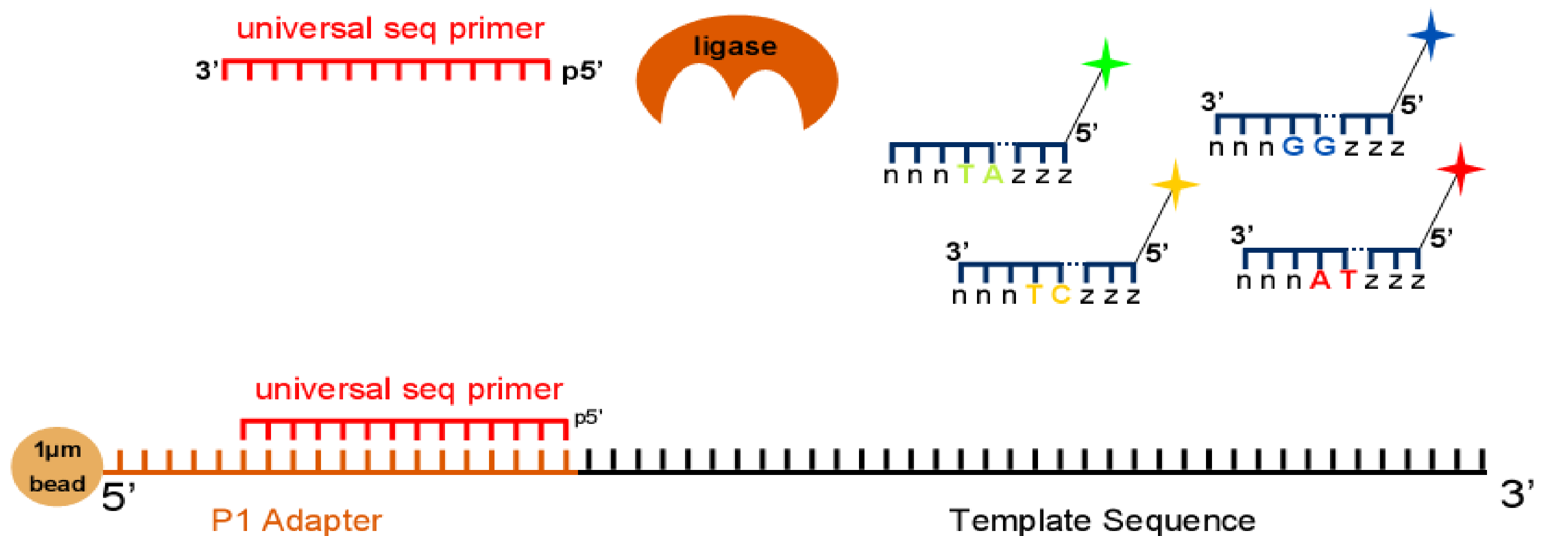


1,024 Octamer Probes (4^5)

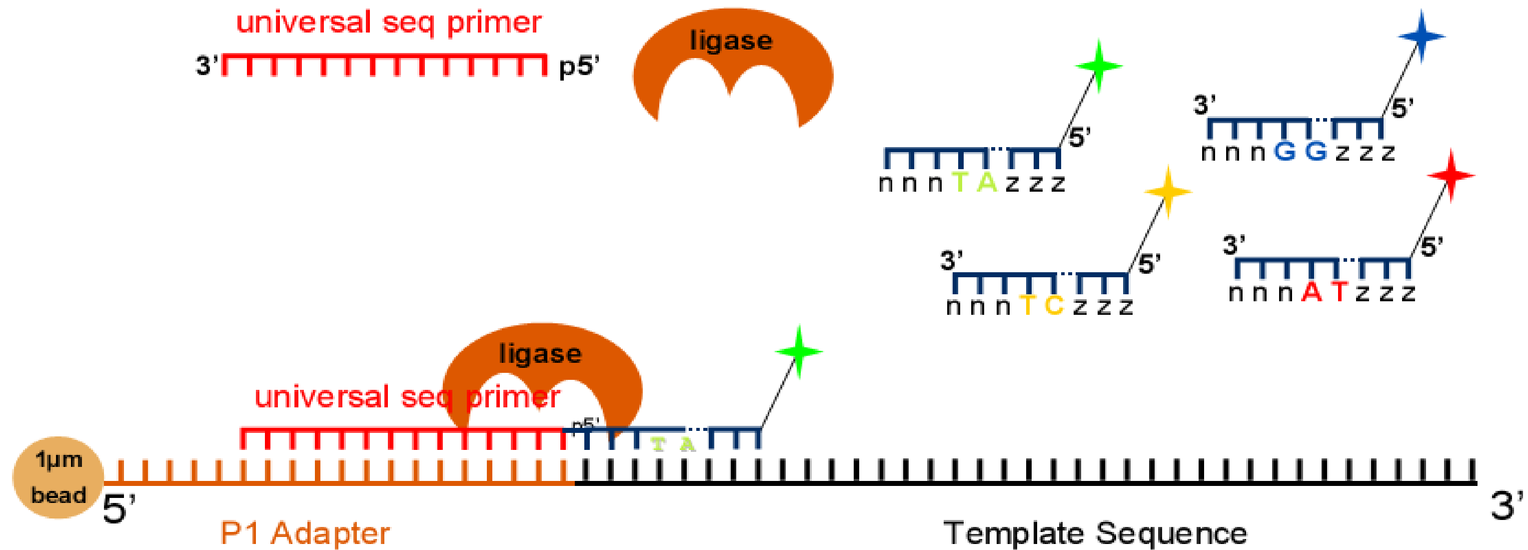
4 Dyes, 4 dinucleotides, 256 probes per dye

N= degenerate bases Z= Universal bases

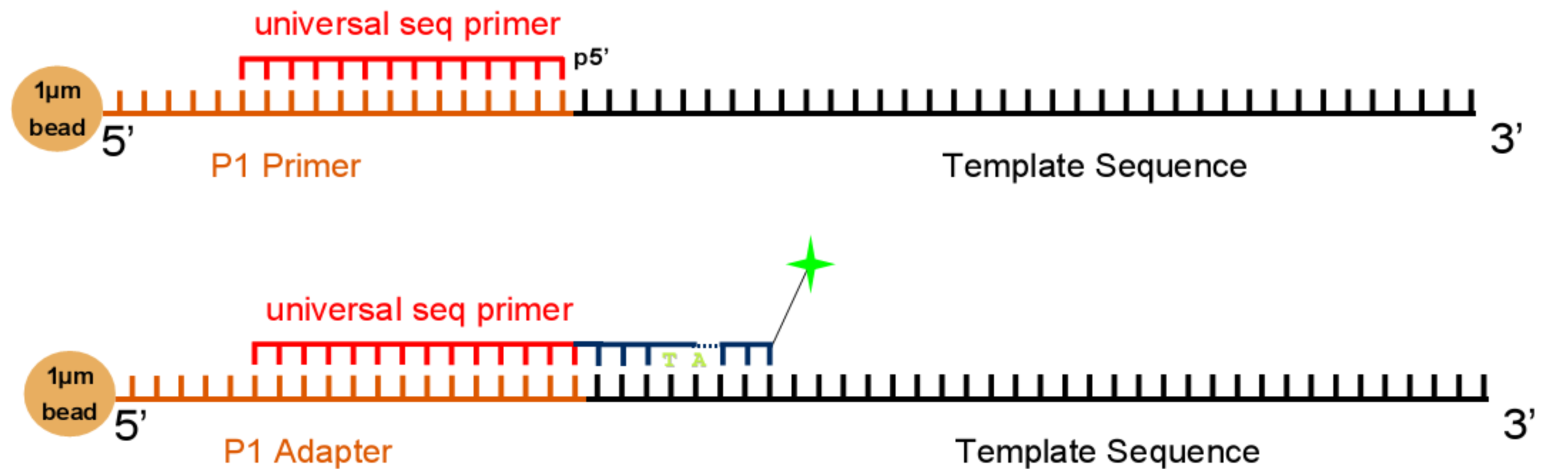
SOLiD Chemistry System 4-color ligation
Ligation reaction



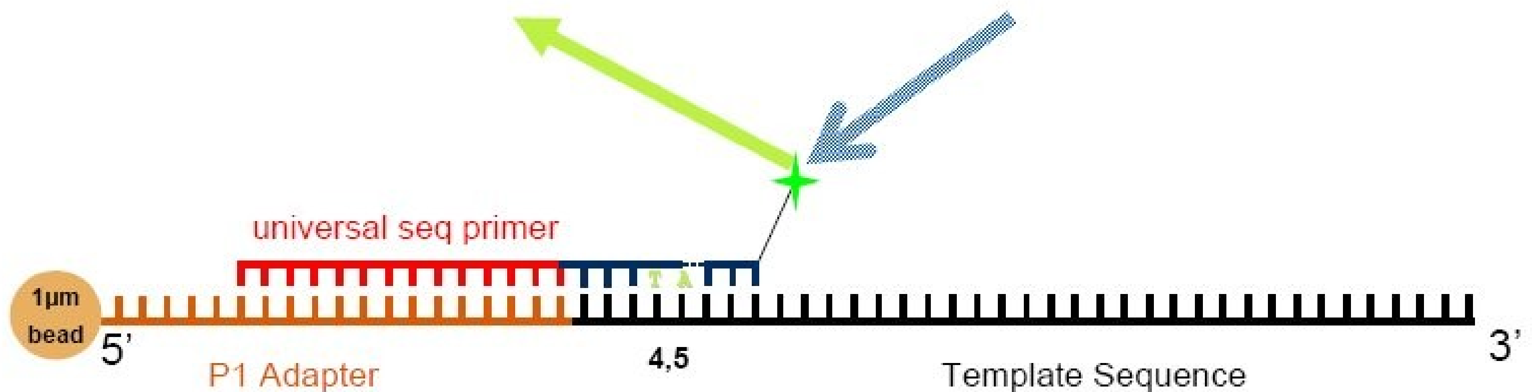
SOLiD Chemistry System 4-color ligation
Ligation reaction



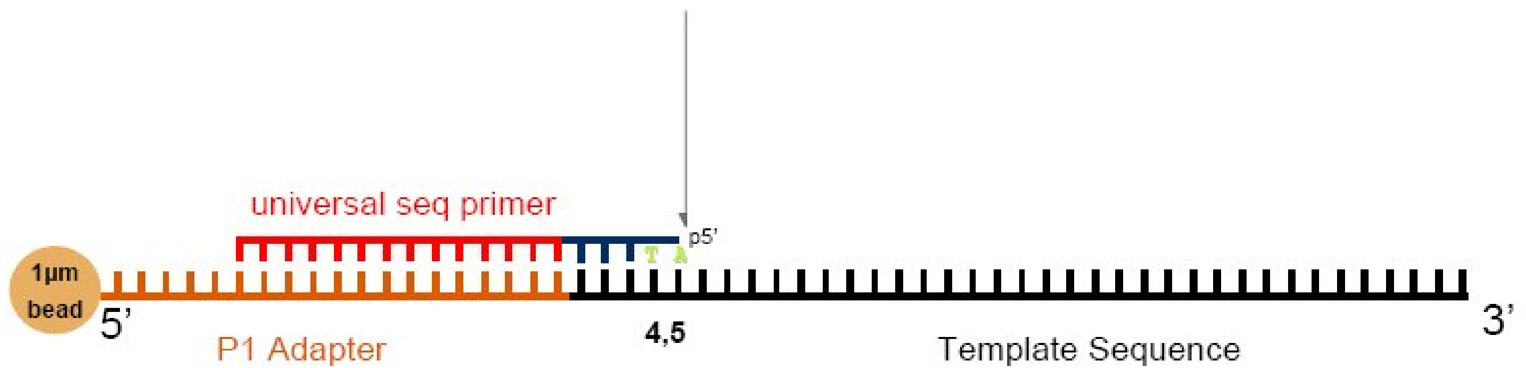
SOLiD Chemistry System 4-color ligation
De-Phosphorylation



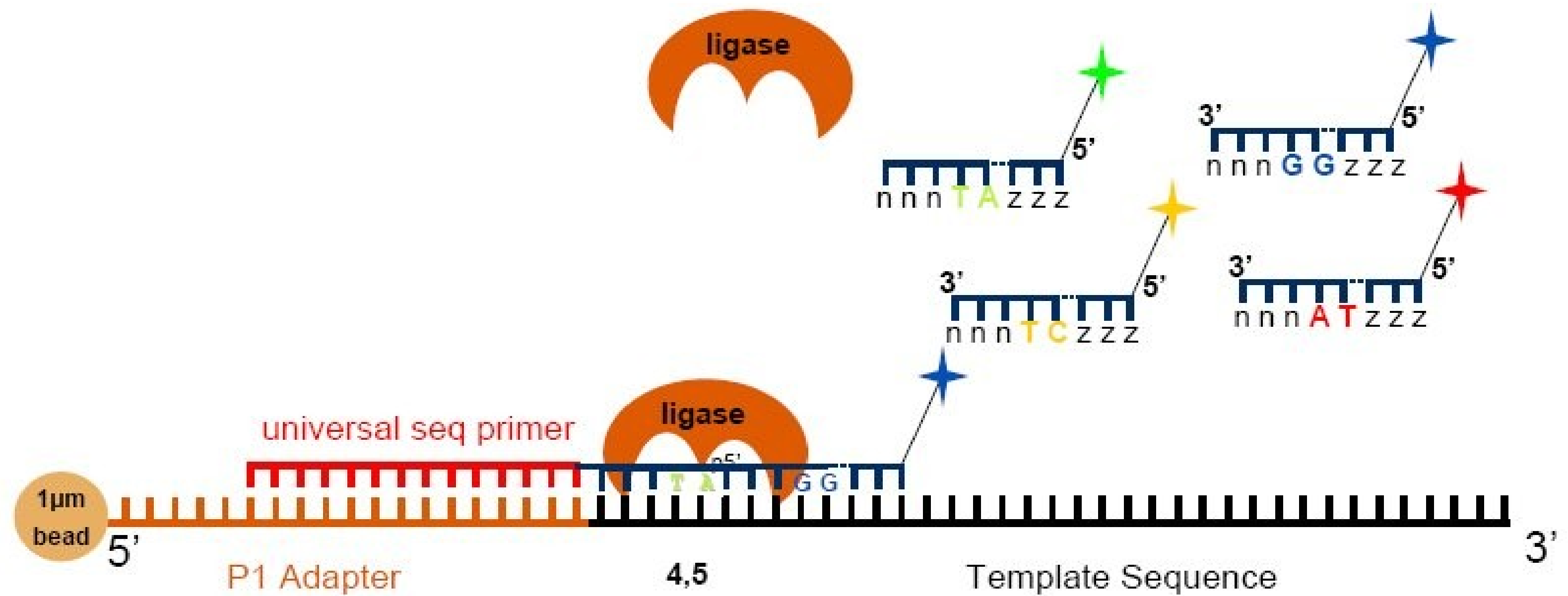
SOLiD Chemistry System 4-color ligation Visualization



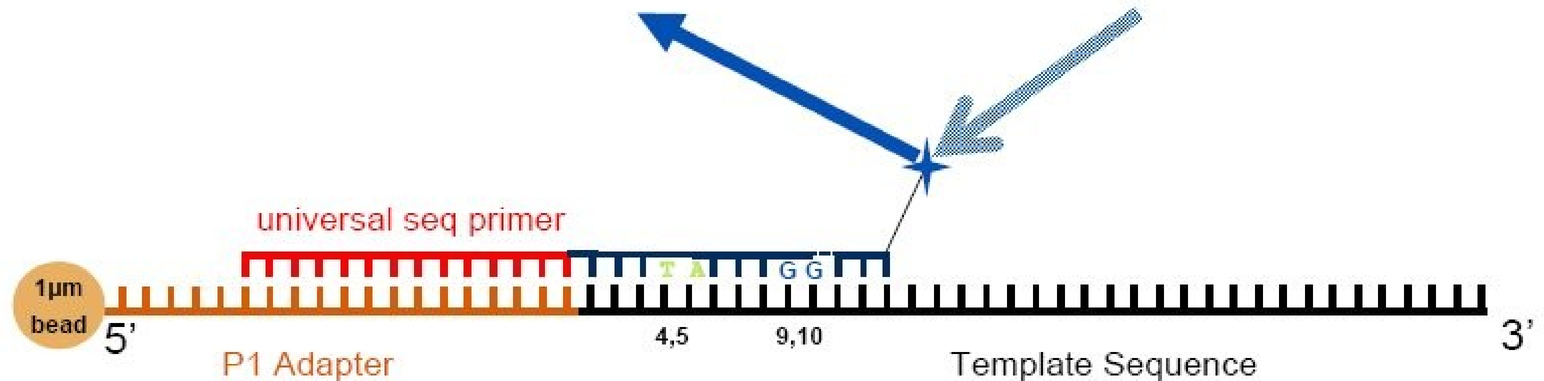
SOLiD Chemistry System 4-color ligation
Cleavage



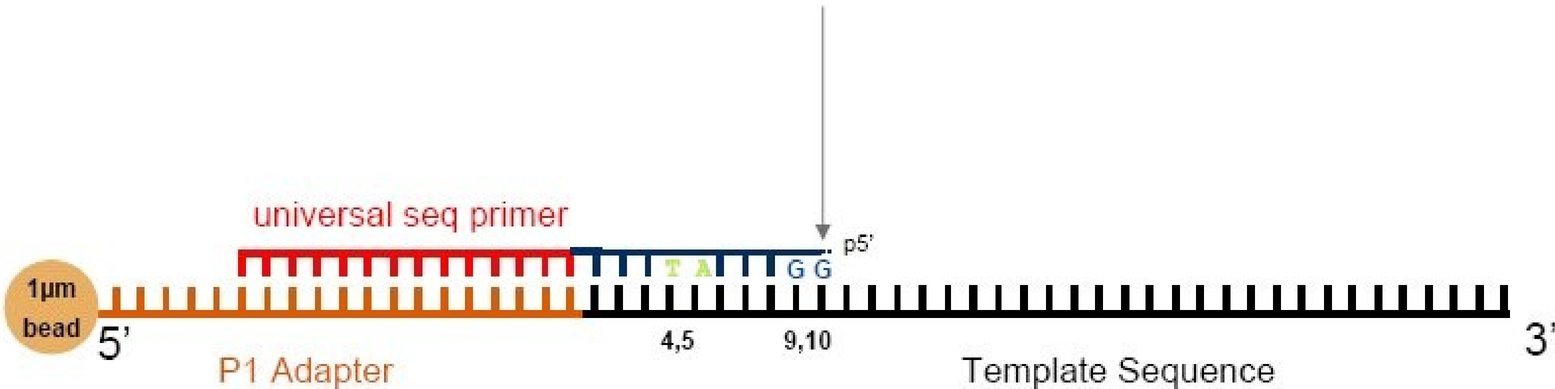
SOLiD Chemistry System 4-color ligation
Ligation (2nd cycle)



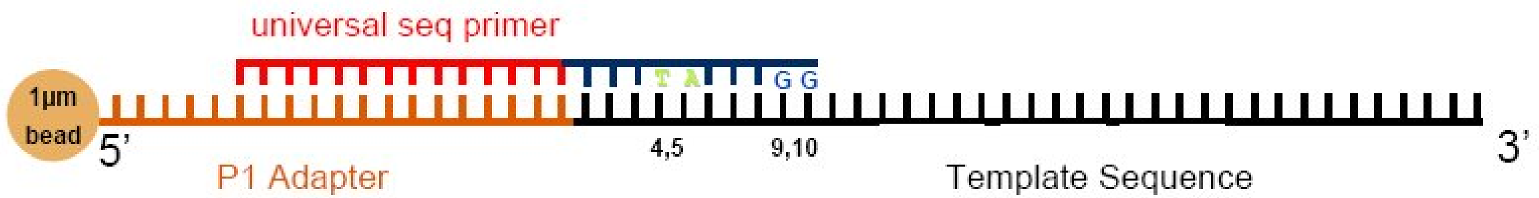
SOLiD Chemistry System 4-color ligation Visualization (2nd cycle)



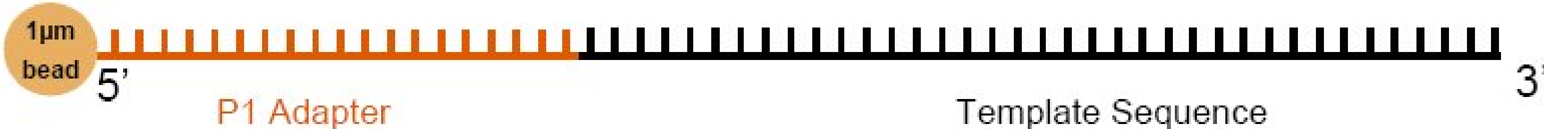
SOLiD Chemistry System 4-color ligation Cleavage (2nd cycle)



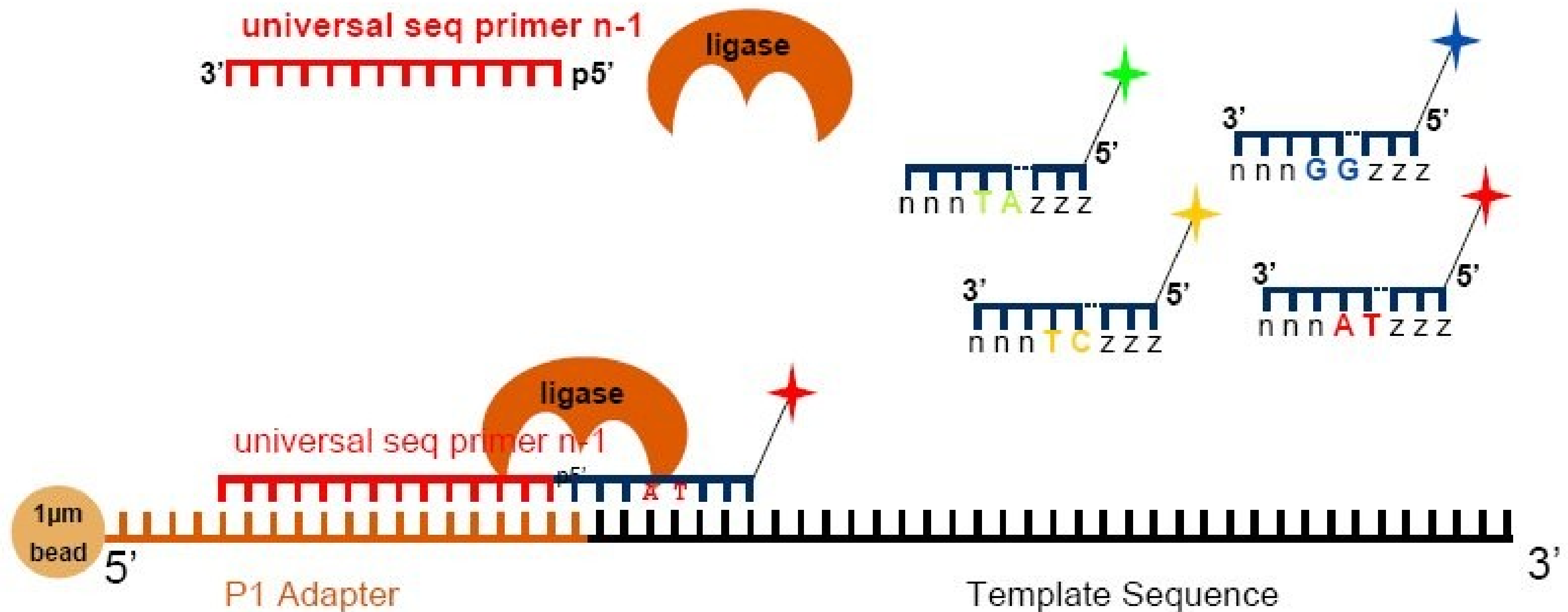
SOLiD Chemistry System 4-color ligation interrogates every 5th base



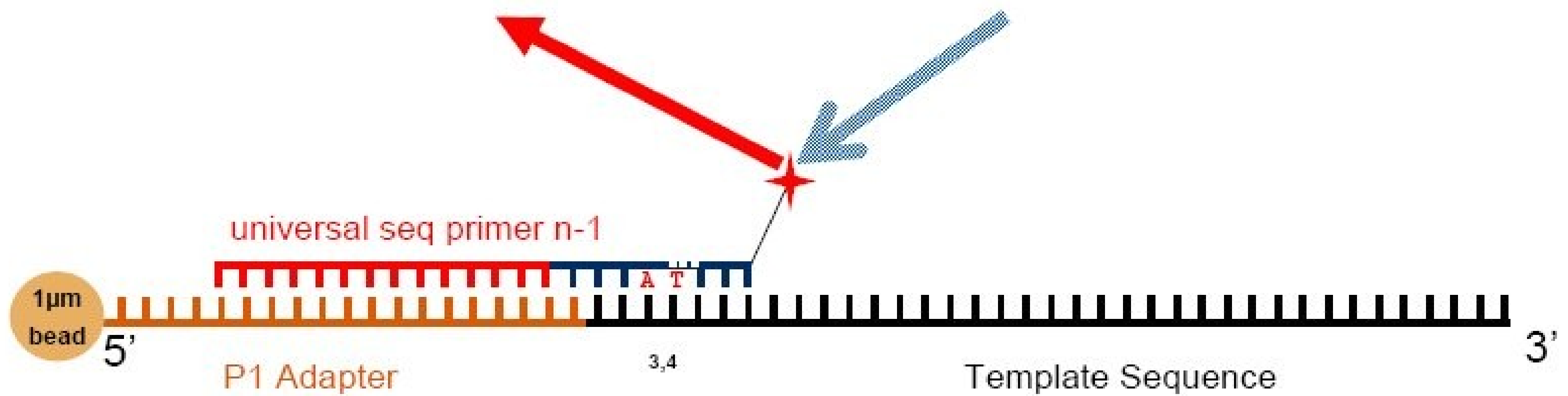
SOLiD Chemistry System 4-color ligation
Reset



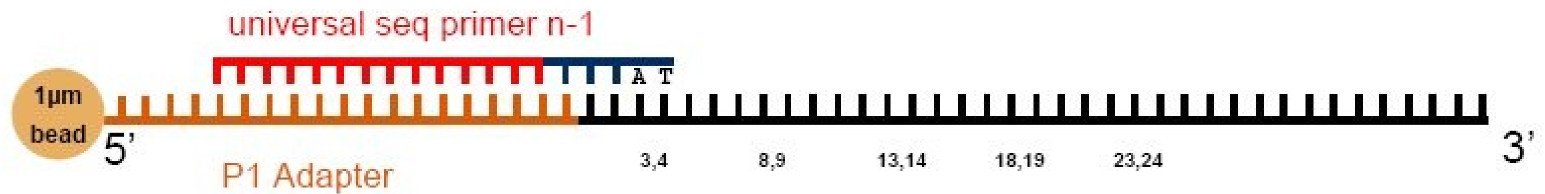
SOLiD Chemistry System 4-color ligation
(1st cycle after reset)



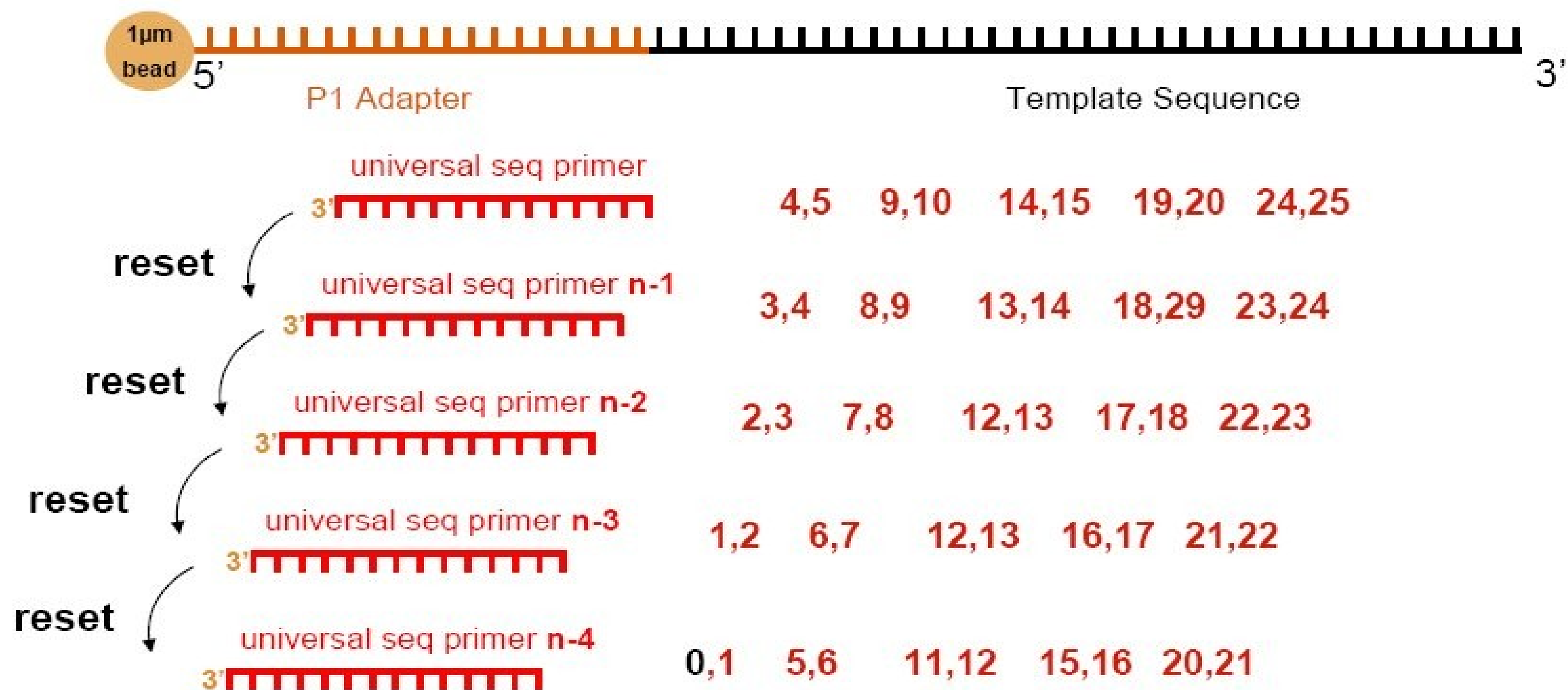
SOLiD Chemistry System 4-color ligation
(1st cycle after reset)



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

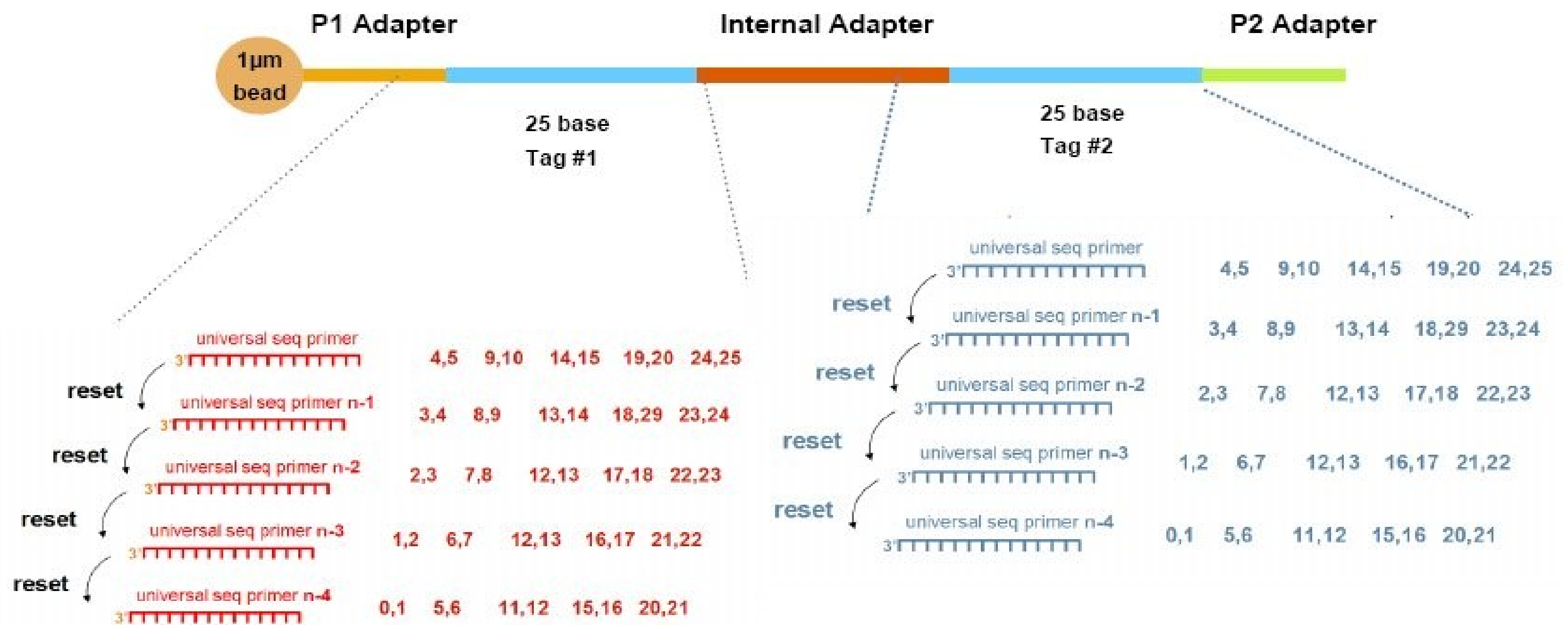


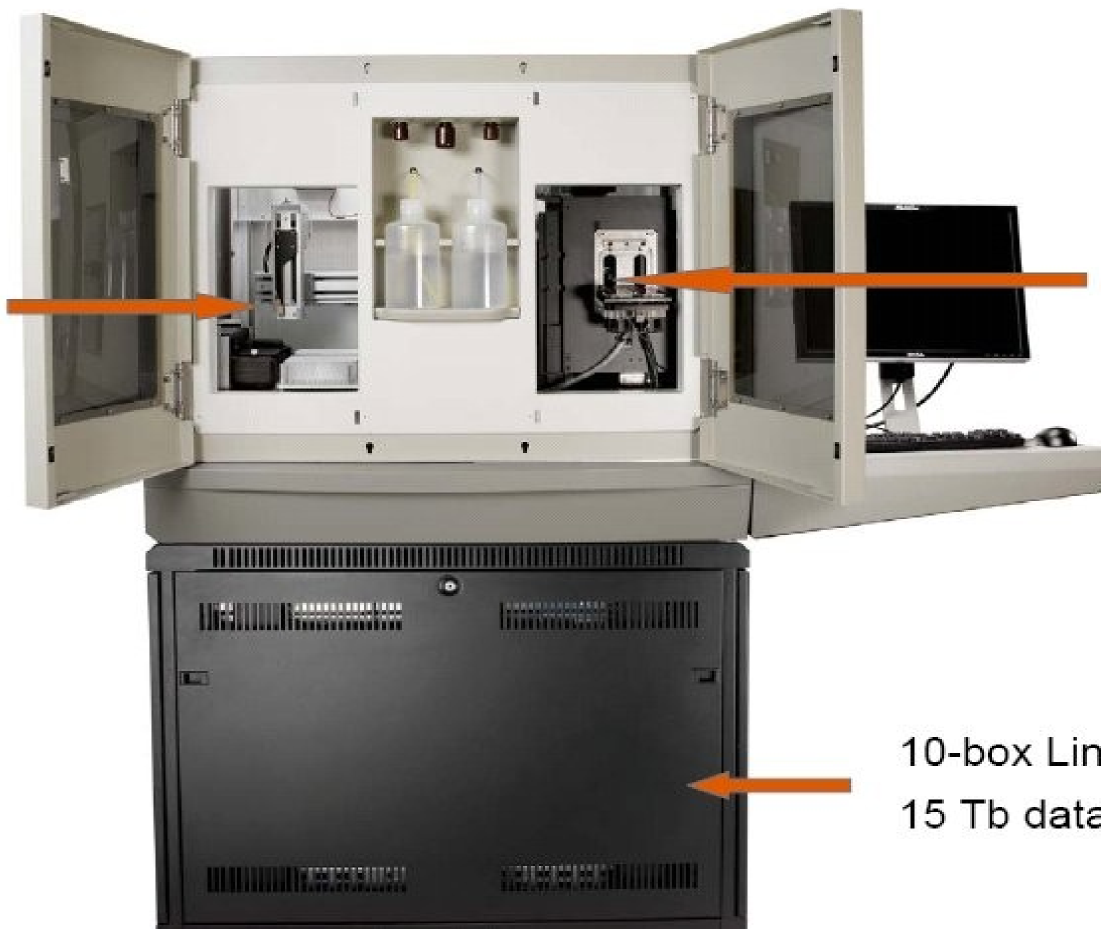
Sequential rounds of sequencing
Multiple cycles per round



Paired End two sequences generated

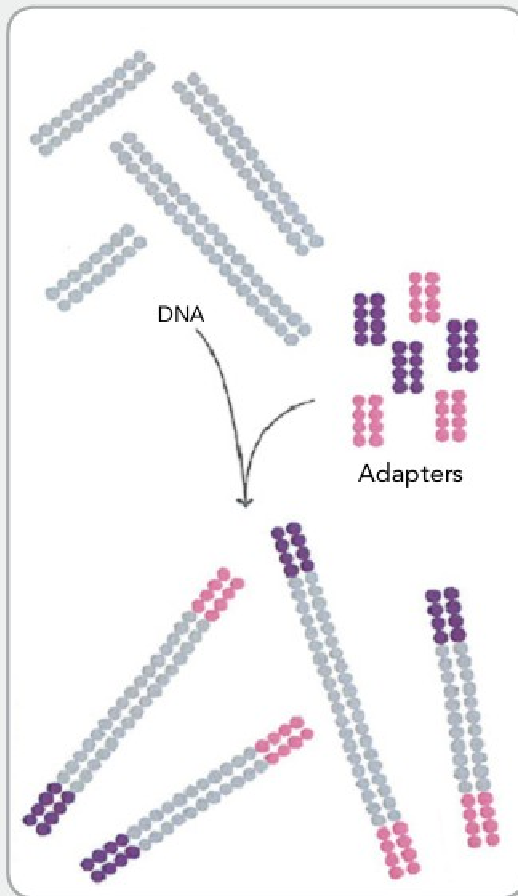
Sequential rounds of sequencing
Multiple cycles per round




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

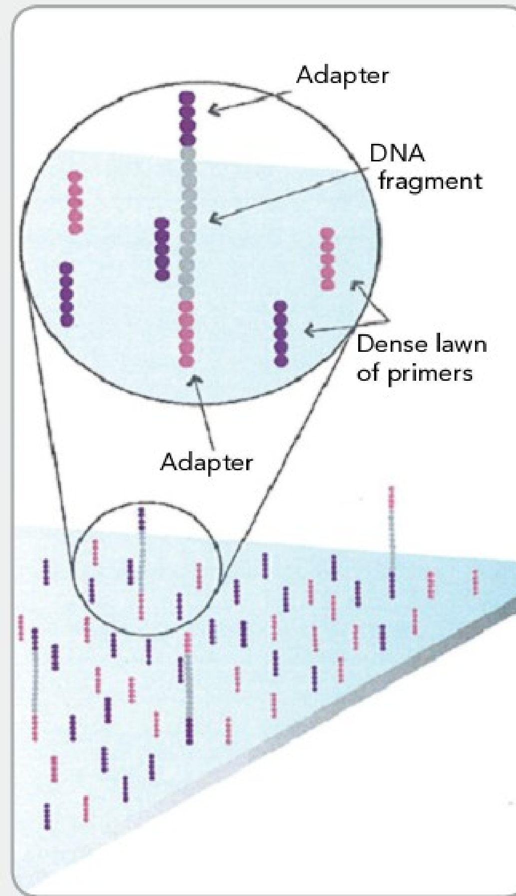
Solexa (2007)

1. PREPARE GENOMIC DNA SAMPLE



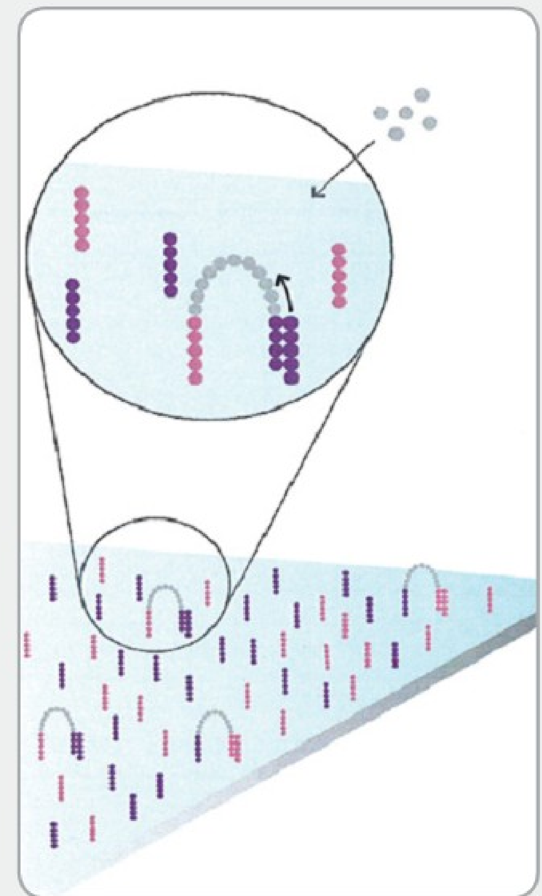
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



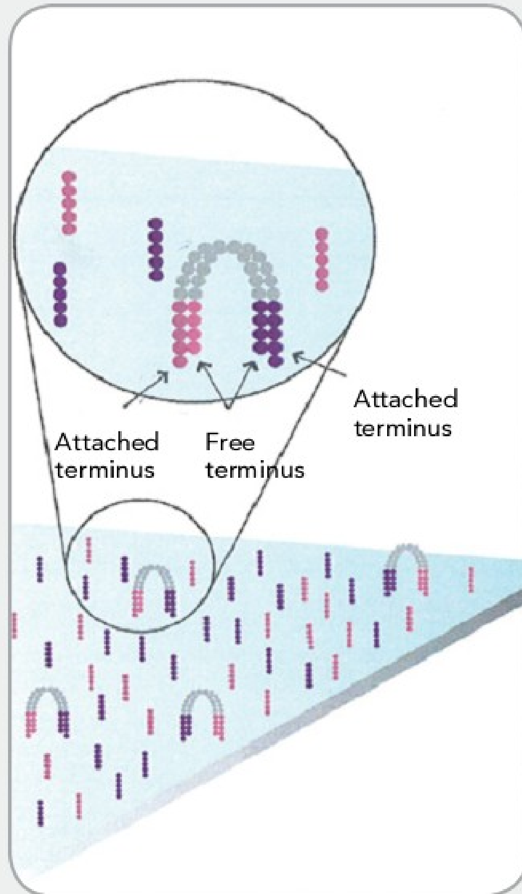
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION



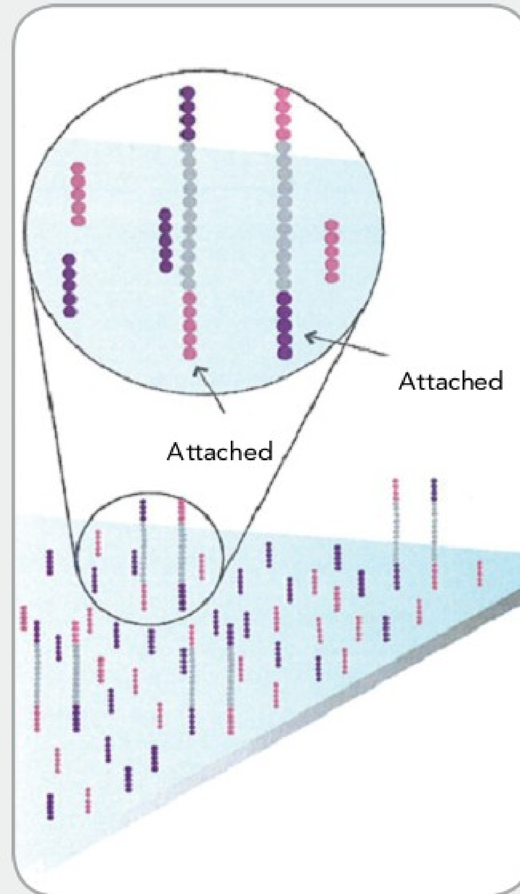
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED



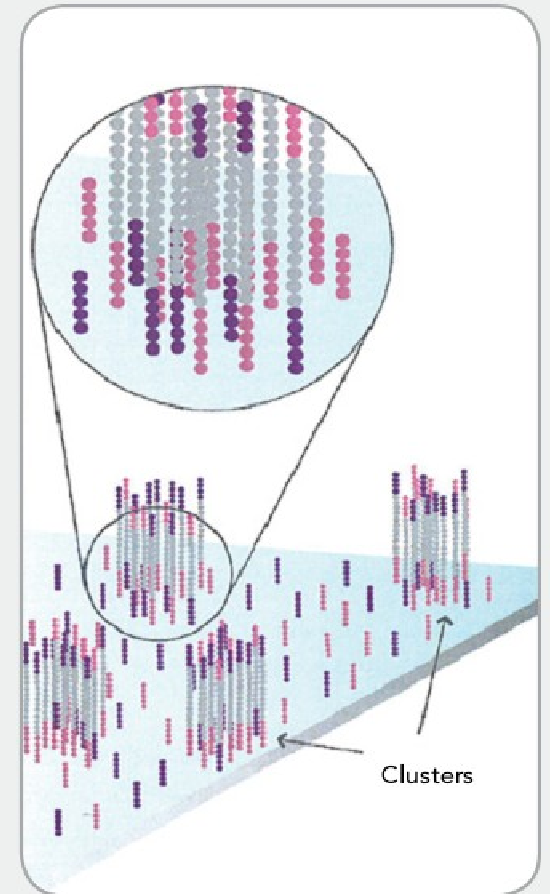
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



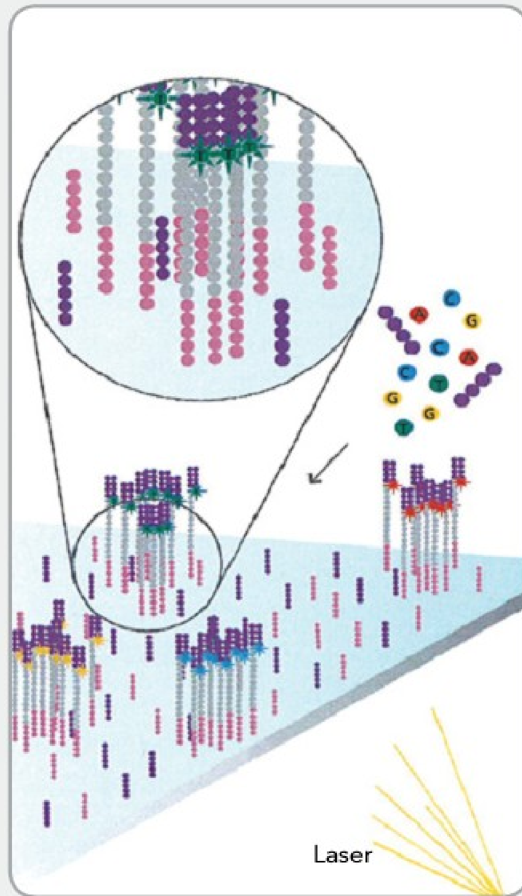
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



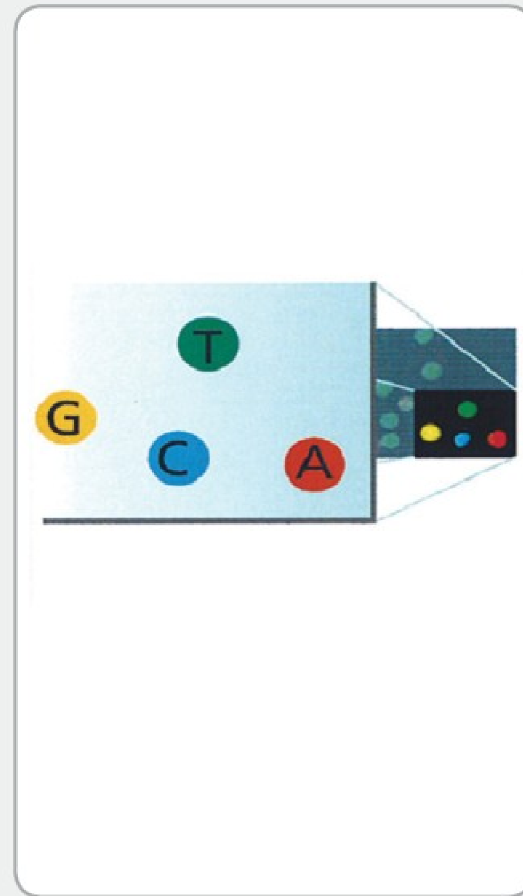
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE



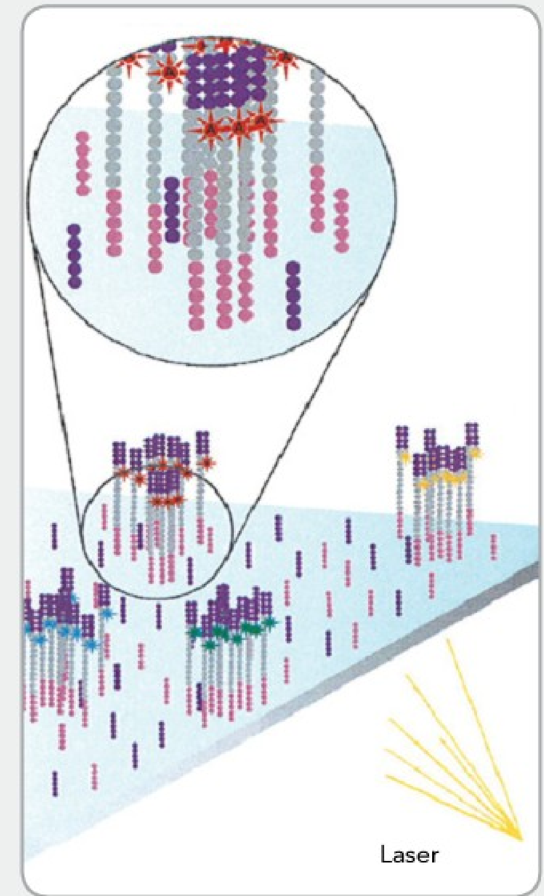
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

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.

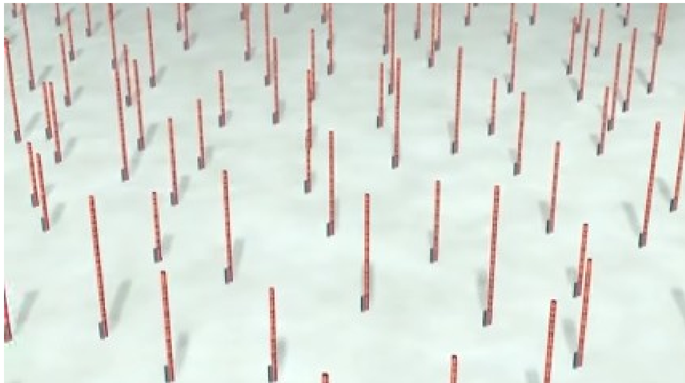
9. DETERMINE SECOND BASE

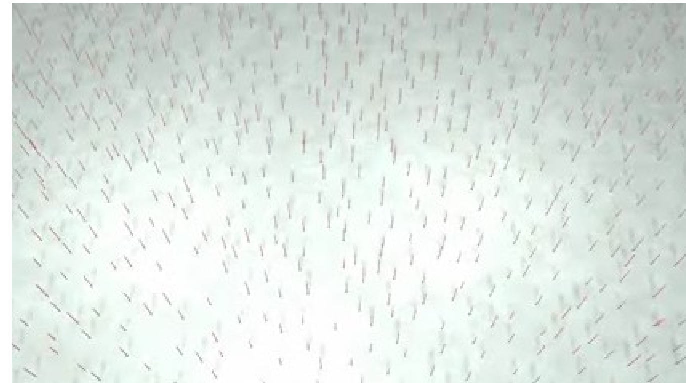


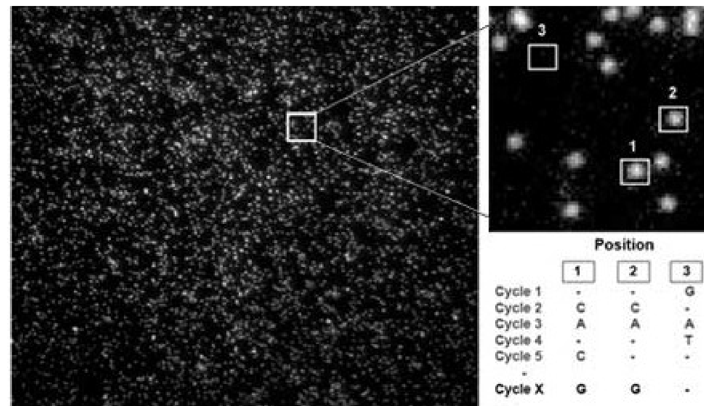
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

HELICOS (2008)

True Single Molecule Sequencing (tSMS)



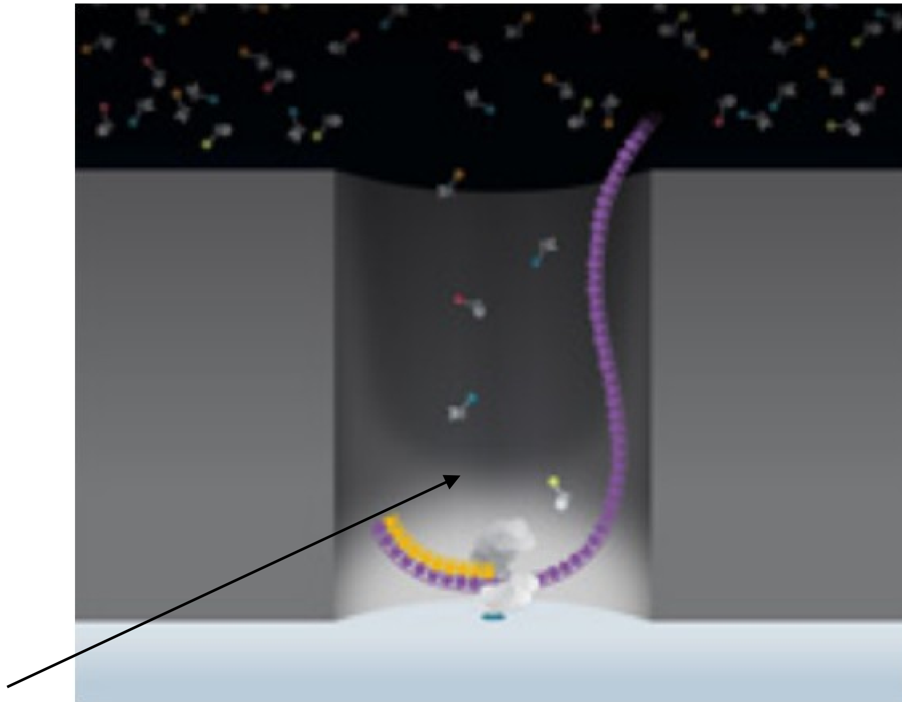




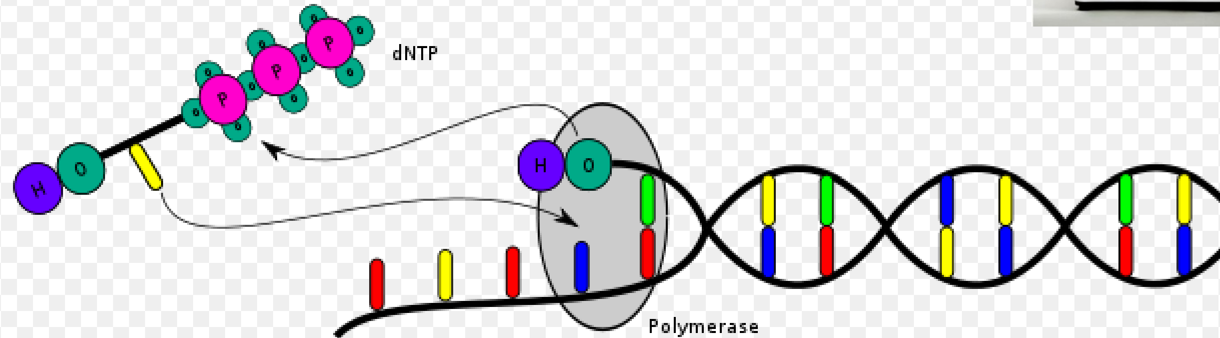
Single Molecule Real-Time (SMRT)

Pacific Biosciences

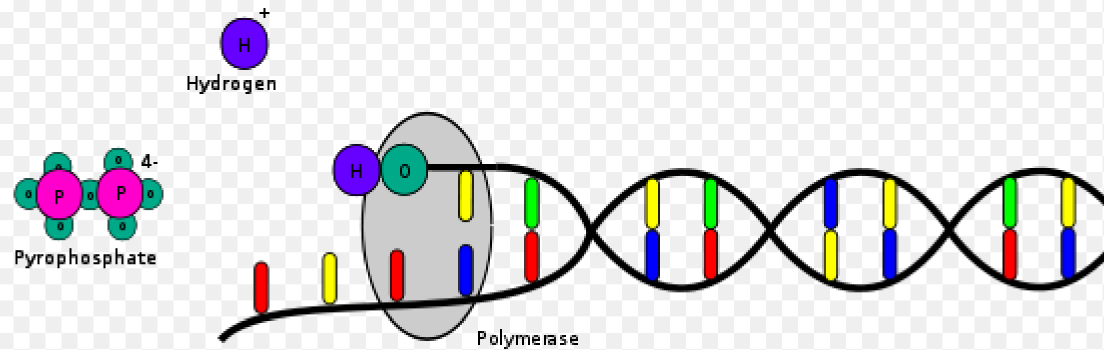
20 zeptolitrů



Ion Torrent



Polymerase integrates a nucleotide.



Hydrogen and pyrophosphate are released.

Oxford nanopore

GGTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCTA

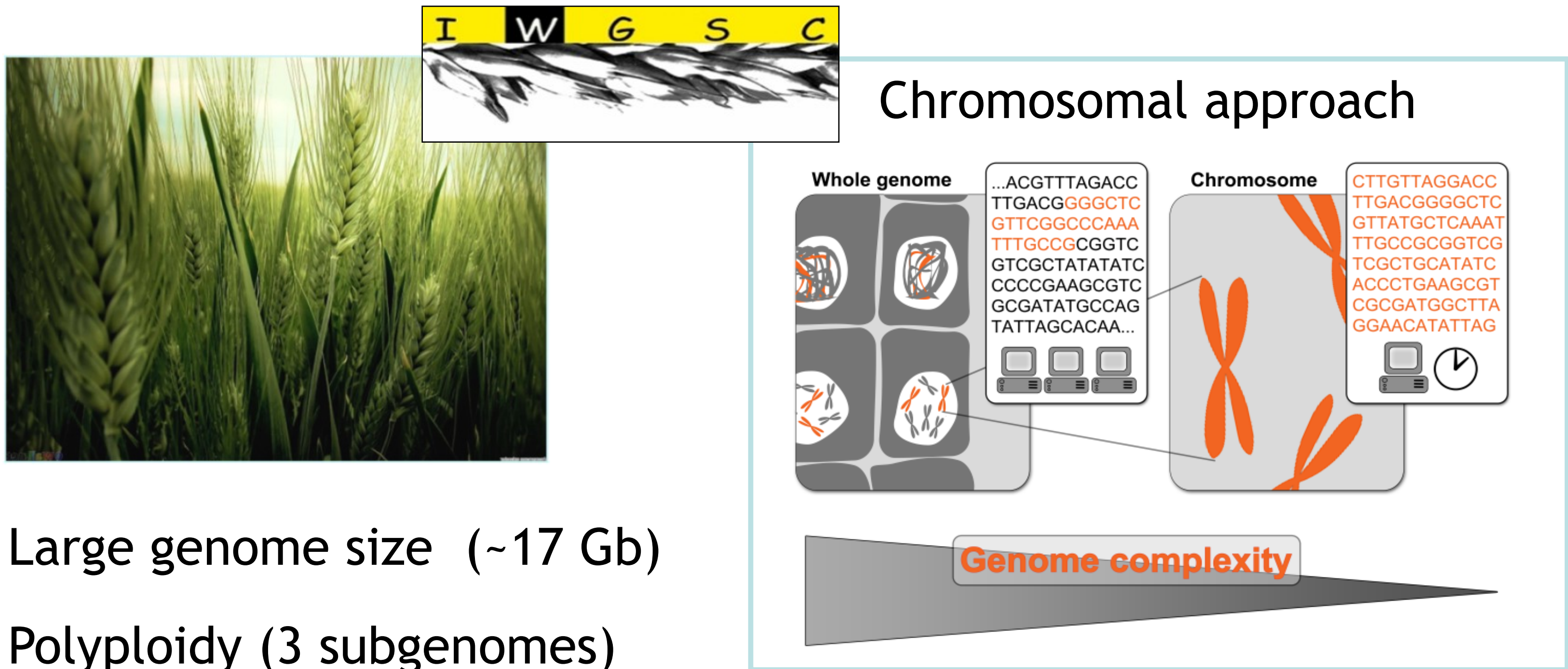
Další technologie

- Mikroelektroforéza
- Sekvenování na bázi microarray

Challenges in Genome Sequencing

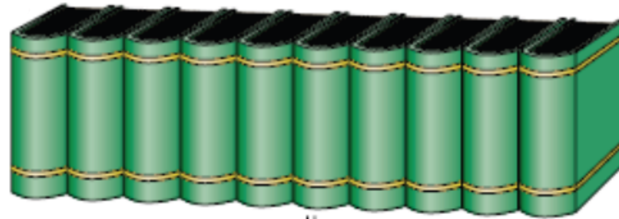
De novo genome assemblies using only short read data of NGS technologies are generally incomplete and highly fragmented due to

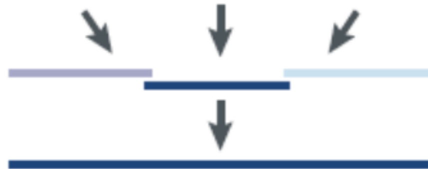
- Large duplications - chromosomal approach, BAC-by-BAC sequencing
- High proportion of repetitive DNA - challenge!



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

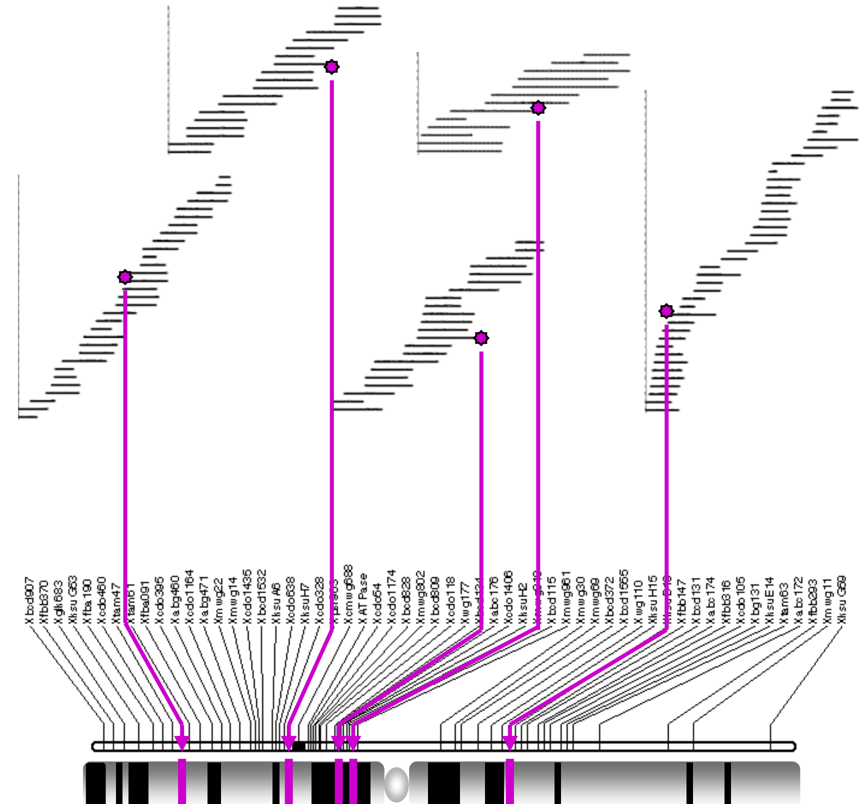
BAC-BY-BAC SEQUENCING





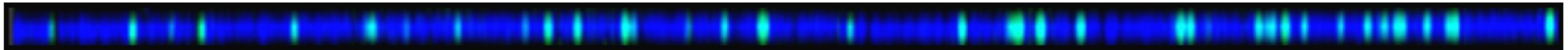
BAC by BAC approach

- **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

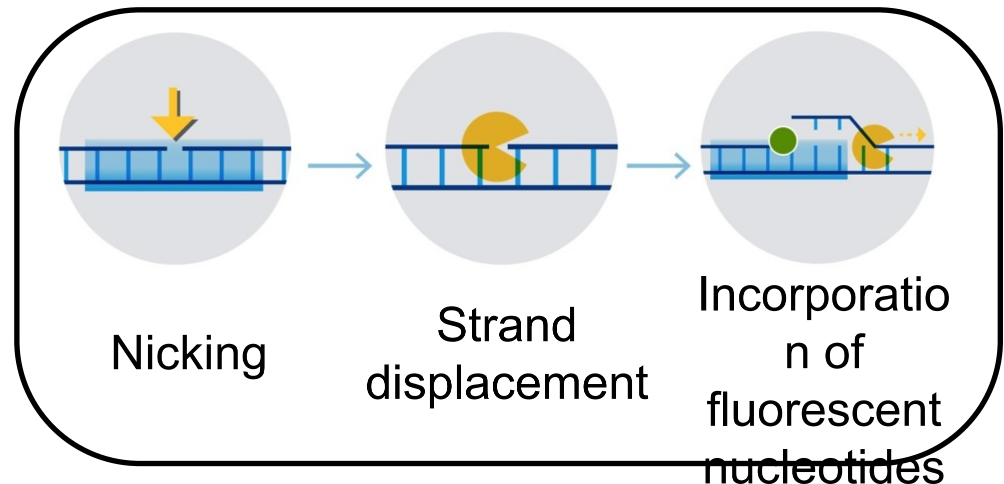
- **Long mate-pair reads** > 10 kb
- **Long read technologies** – PacBio, Oxford Nanopore
- **Optical mapping**
  - Single-molecule mapping of genomic DNA **hundreds of kilobases to several megabases in size**
  - Creates **sequence-motif maps**, which provide long-range template for ordering genomic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



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 Nanochanel Arrays

1 Sequence-specific labeling

Nickase (Nt.BspQI)

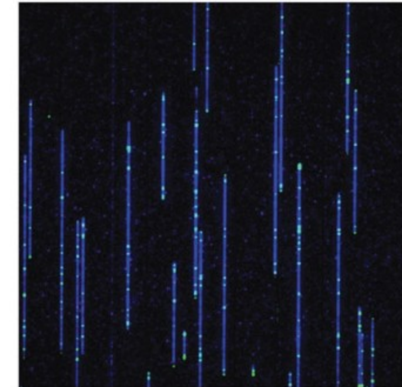
5′-ATGC**GCTCTTC**CATGAATGCGAGC-3′
3′-TACG**CGAGAAG**GTACTTACGCTCG-5′

Nick labeling

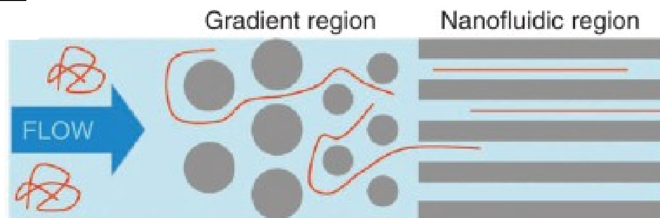
ATGAATGC

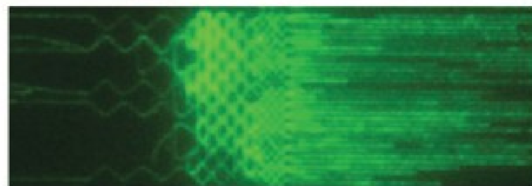
5′-ATGC**GCTCTTC**CAUGAAUGCGAGC-3′
3′-TACG**CGAGAAG**GTACTTACGCTCG-5′

3 Fluorescence imaging

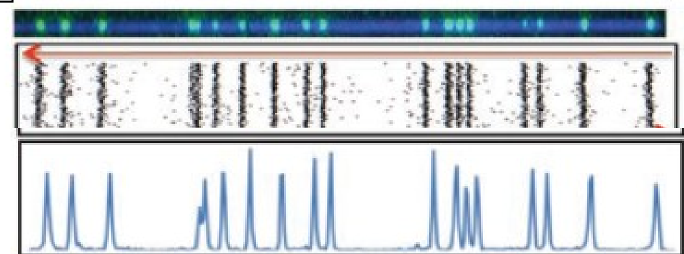


2 DNA linearization

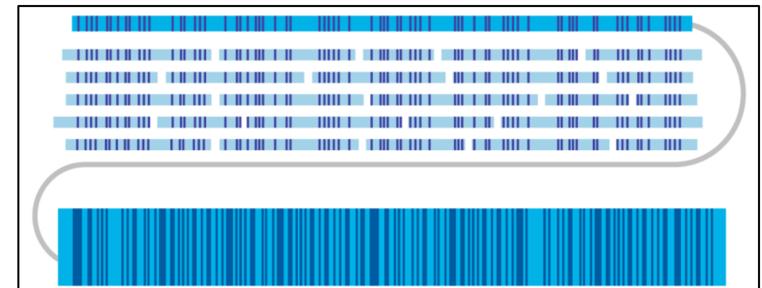




4 Map construction



5 Building consensus map



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