INTRODUCTION AU SÉQUENÇAGE À HAUT DÉBIT POUR LA GÉNOMIQUE

Claude THERMES

INSTITUT DE BILOGIE INTÉGRATIVE DE LA CELLULE – I2BC

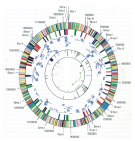
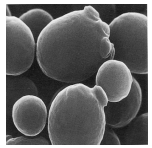
GIF-SUR-YVETTE

Diplôme Universitaire en Bioinformatique intégrative - DU-Bii - 08/03/2021

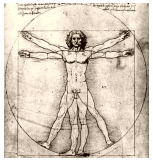
Premiers génomes entièrement séquencés

Médecine personnalisée >> milliers de génomes

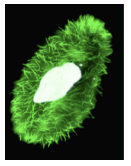
2 bactéries | Levure | Ver | Plante | Mouche | Humain | Souris | Poule | Chimpanzé | Paramécie | 2 humains



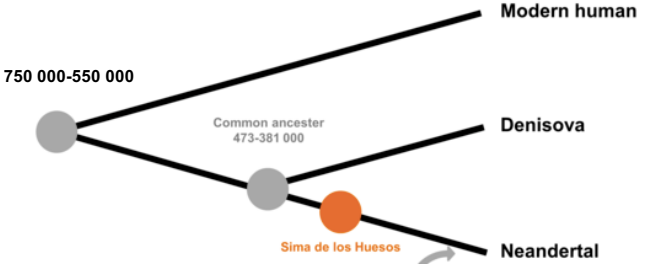







1000 génomes humains

1990 91 92 93 94 95 96 97 98 99 2000 01 02 03 04 05 06 07 08 09 12 ... 21



1rst generation : Sanger sequencing

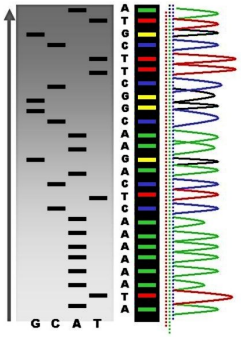
• Has been the major method up to 2005

Limitations

• Extremely high cost

• Long experimental set up times

• High DNA concentrations needed



2d generation

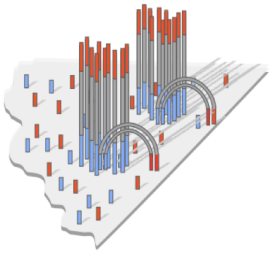
• Single DNA molecules replicated in clusters

• Very high troughput

Limitations

• Maximum read length ≤ 300bp

Illumina



3rd generation

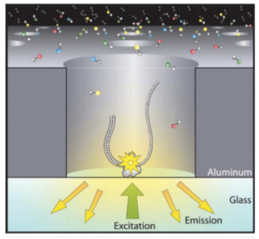
• Single molecules sequencing

• Very long reads

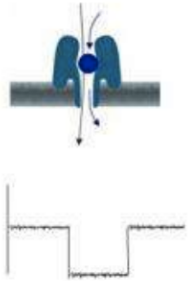
Limitations

• High error rates

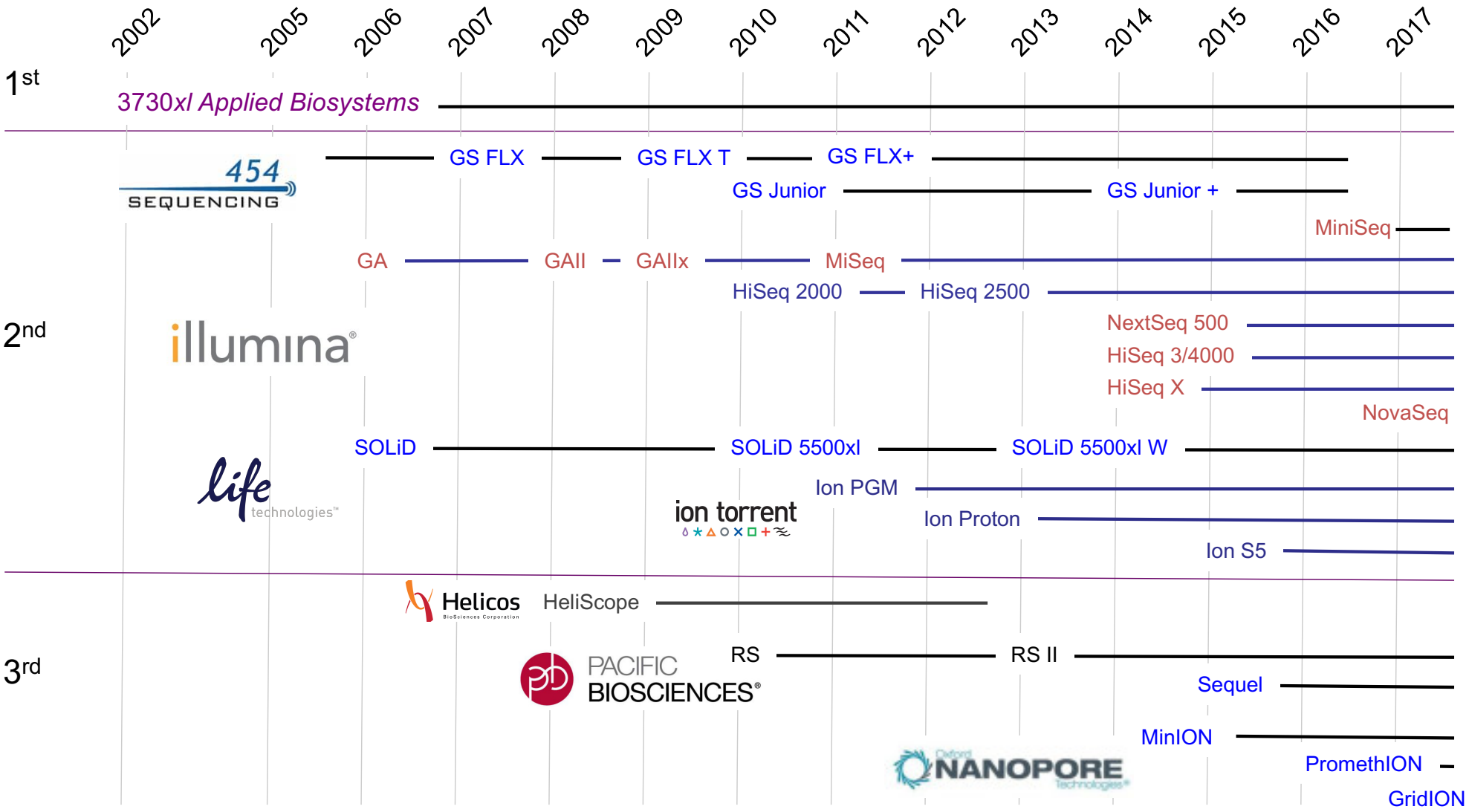
PacBio



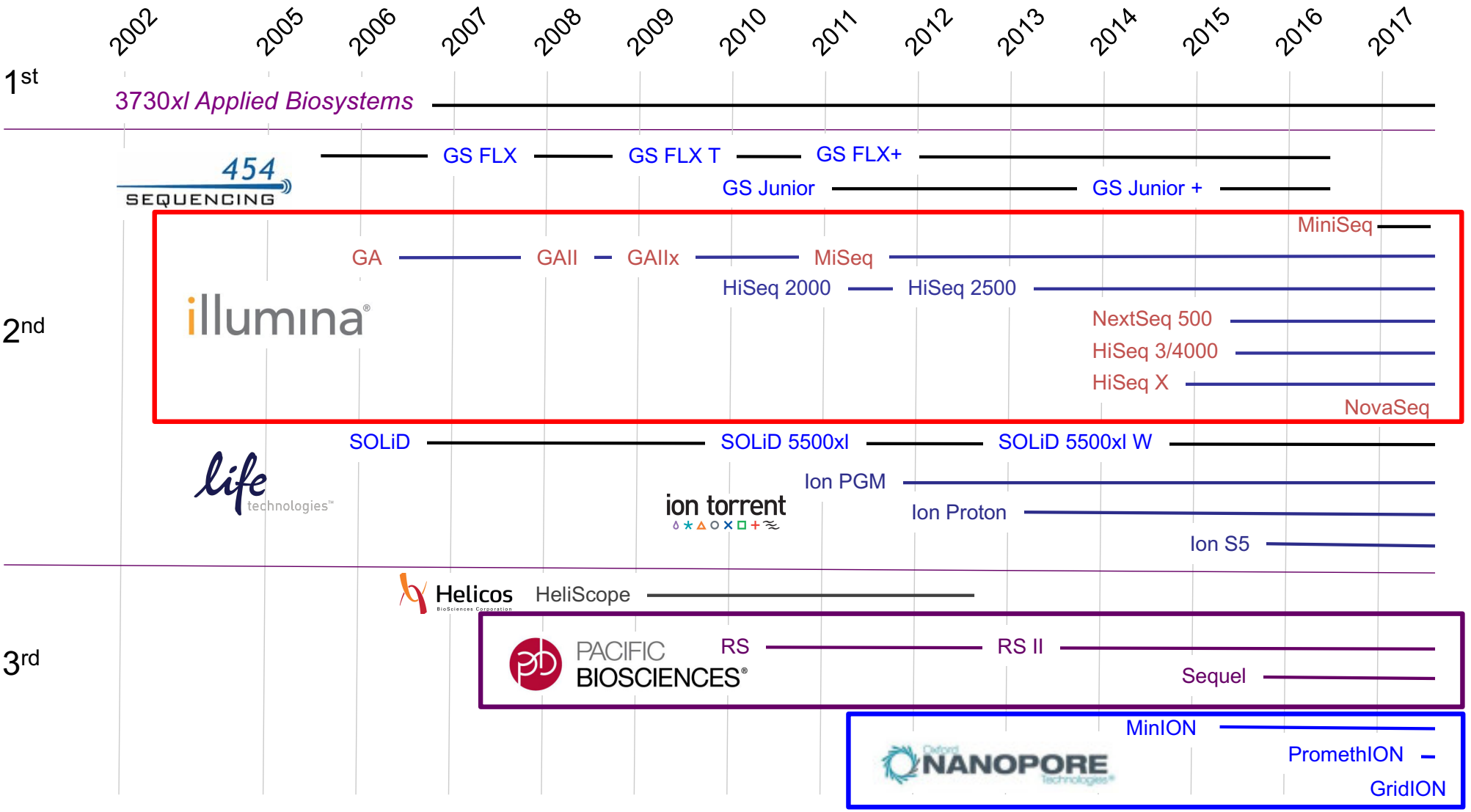
Oxford Nanopore



Sequencing technologies



Sequencing technologies



PART I

2d GENERATION SEQUENCING

Illumina : the winning technology

MiniSeq
25 million reads

MiSeq
25 millions reads, 2 x 300 bp

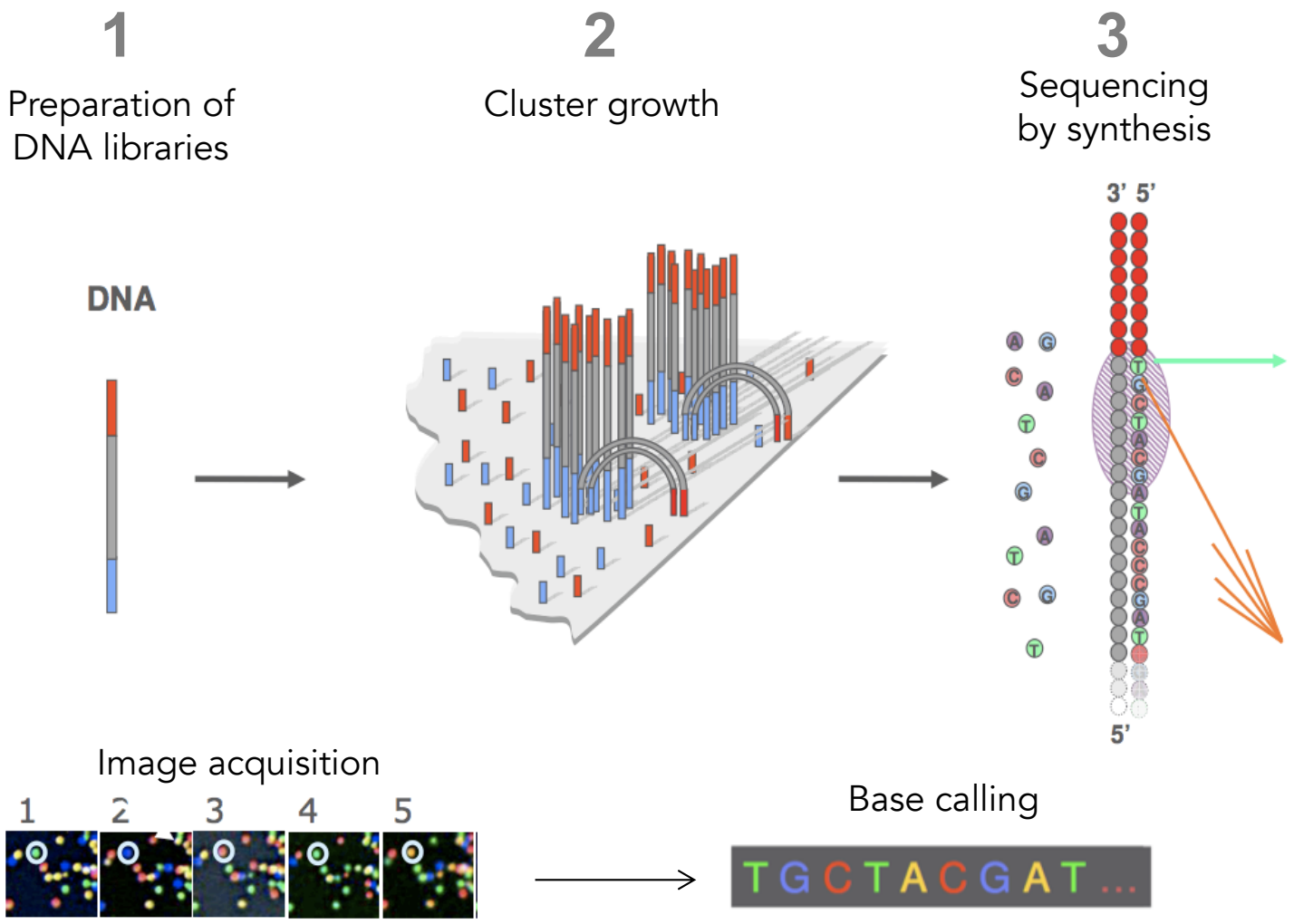
NextSeq
400 million reads

HisSeq 4000
5 billion reads

HisSeq X
6 billion reads

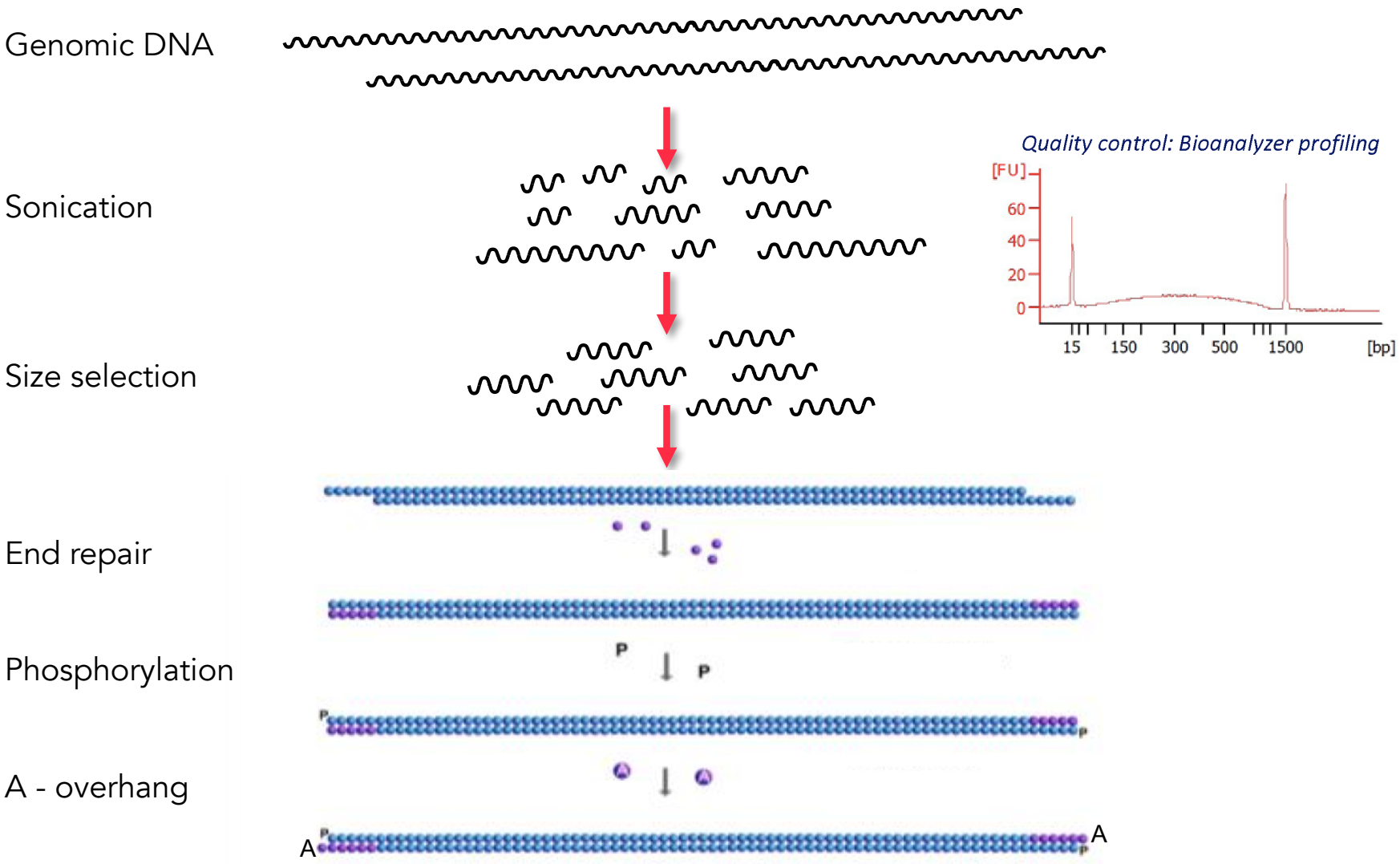
NovaSeq 6000
20 billion reads

General scheme of Illumina sequencing

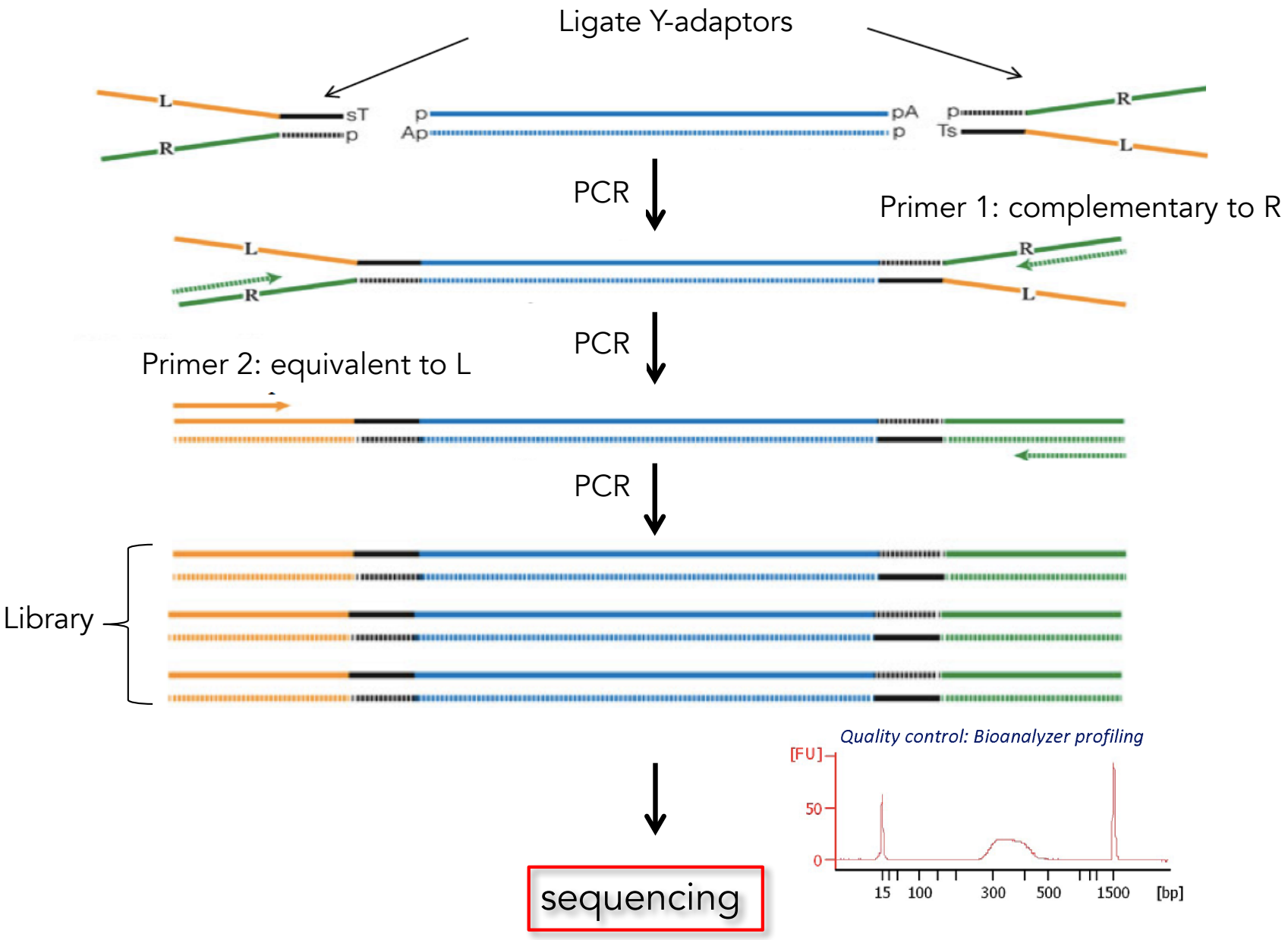


1 - Preparation of DNA-seq Libraries

Illumina TruSeq technology



1 - Preparation of DNA-seq Libraries



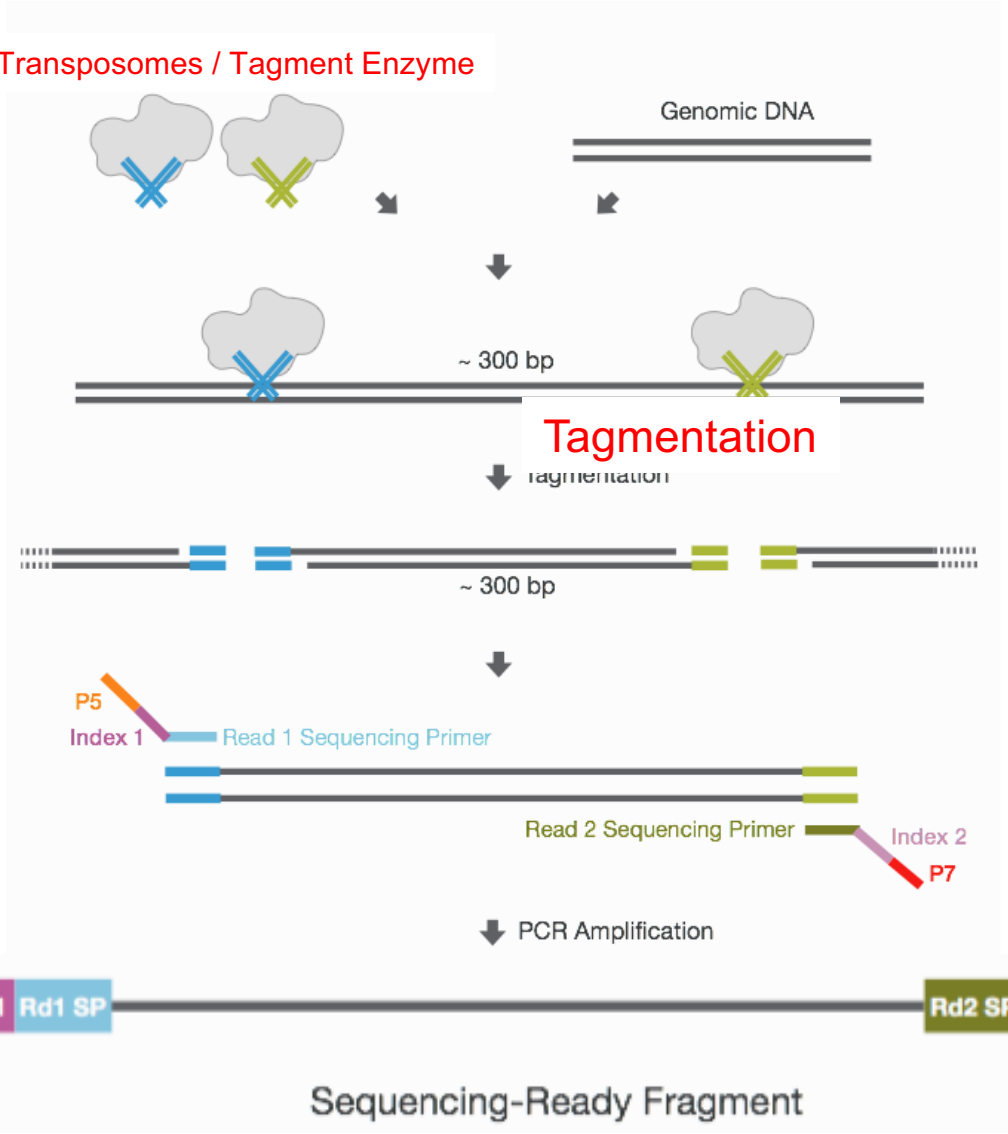
1 - Preparation of DNA-seq Libraries

Nextera "tagmentation"

Tagment Enzyme fragments DNA and attaches junction adapters (blue and green) to both ends of the tagmented molecule

Dual barcode approach

up to 96 indexed samples



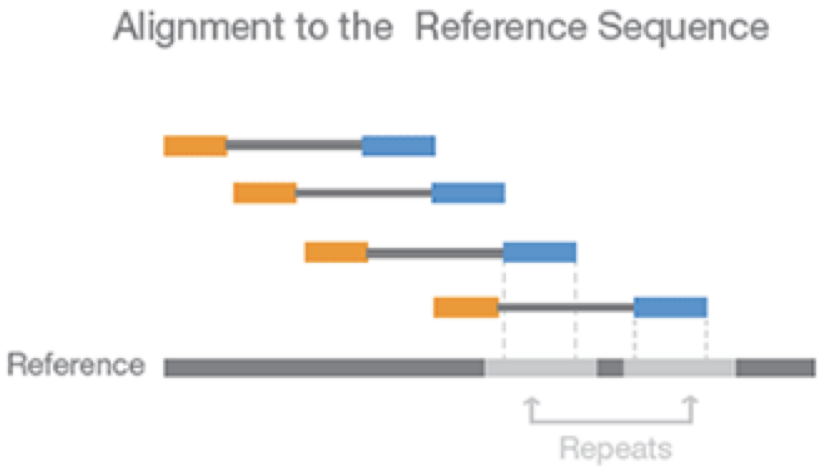
requires small quantities 1ng (bacteria) to 50 ng (human)

1 - Preparation of DNA-seq Libraries

SINGLE READ and PAIRED-END SEQUENCING

- **Single end**: Sequence one physical end of DNA fragment

- **Paired End:** Sequence both physical ends of DNA fragment
  - End distance: < 800nt



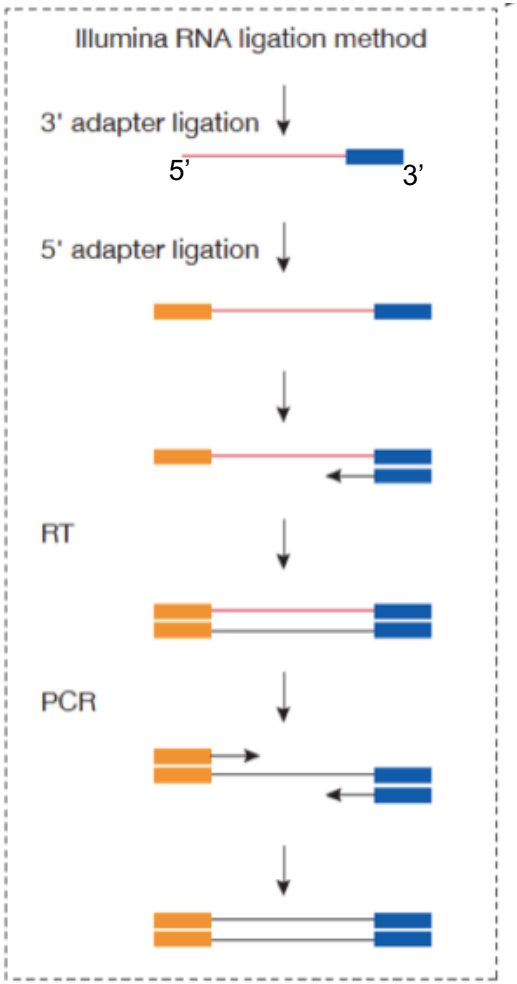
1 - Preparation of RNA-seq Libraries

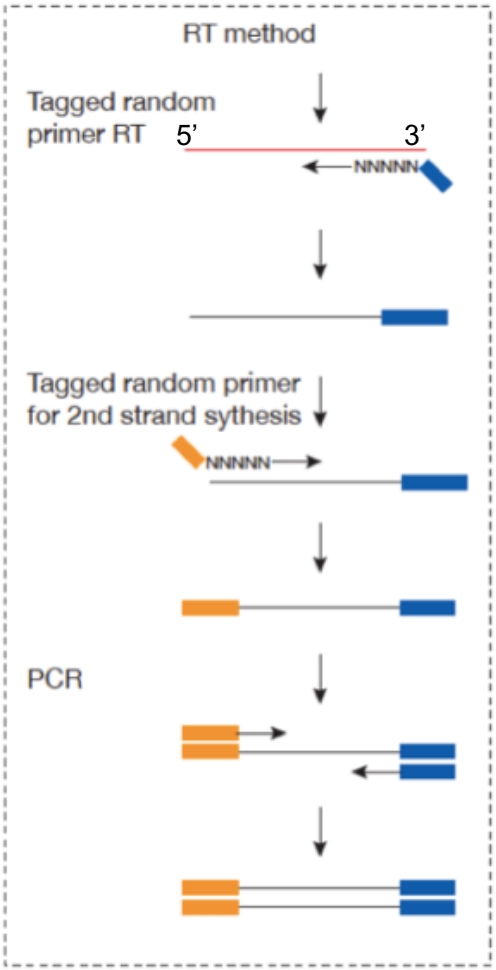
RNA after rRNA depletion

RNA fragmentation

both are directional

Kit smalRNA illumina

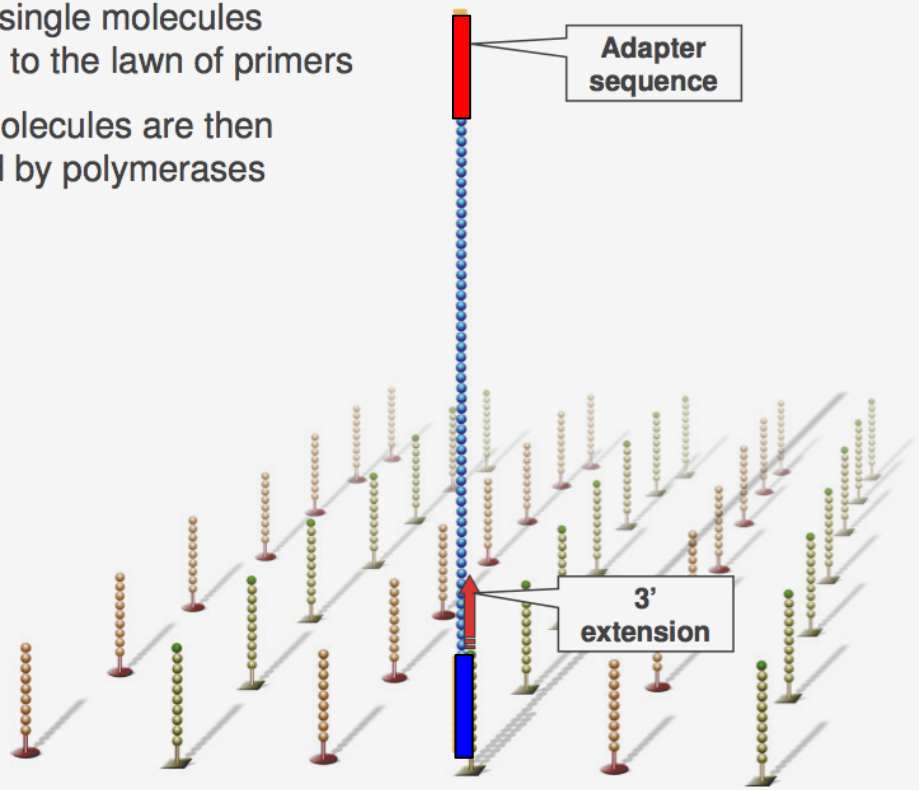




2 – Cluster growth

Cluster Generation:
Hybridize Fragment & Extend

- > 100 M single molecules hybridize to the lawn of primers
- Bound molecules are then extended by polymerases



2 – Cluster growth

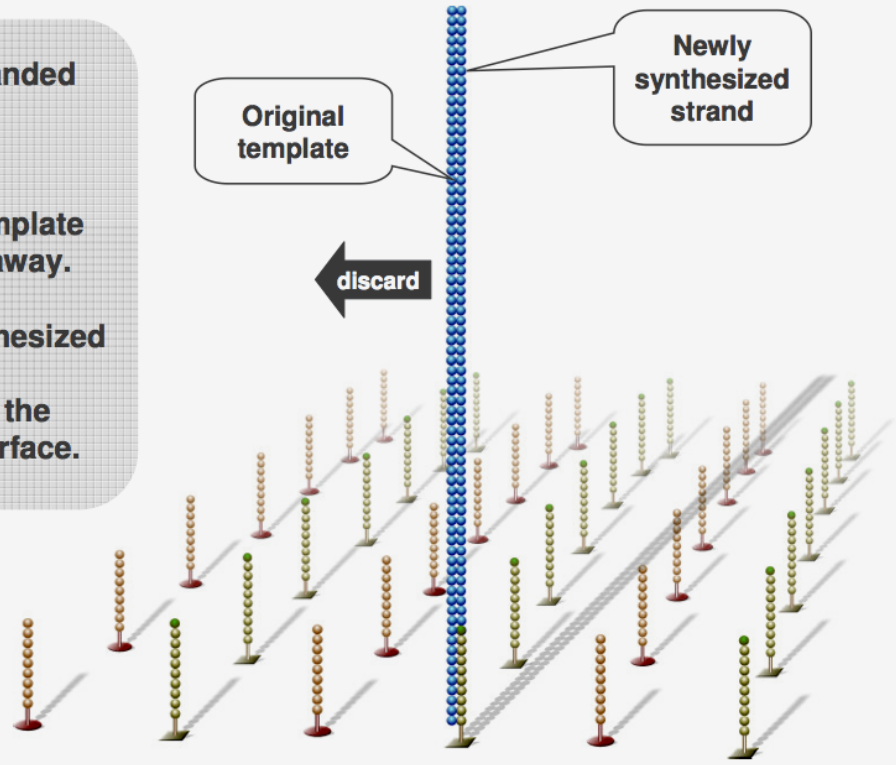
Cluster generation:

Denature double-stranded DNA

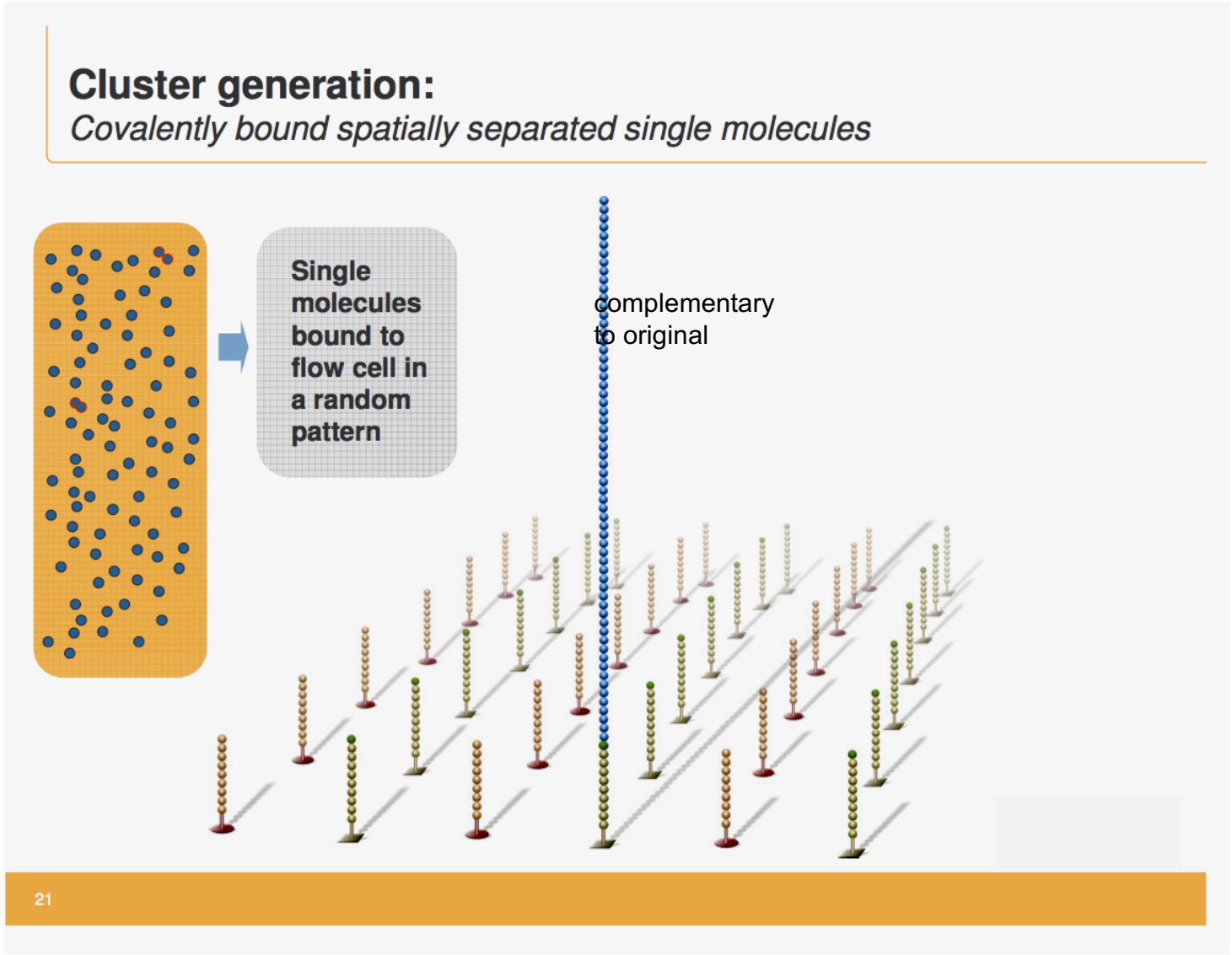
Double-stranded molecule is denatured.

Original template is washed away.

Newly synthesized covalently attached to the flow cell surface.



2 – Cluster growth

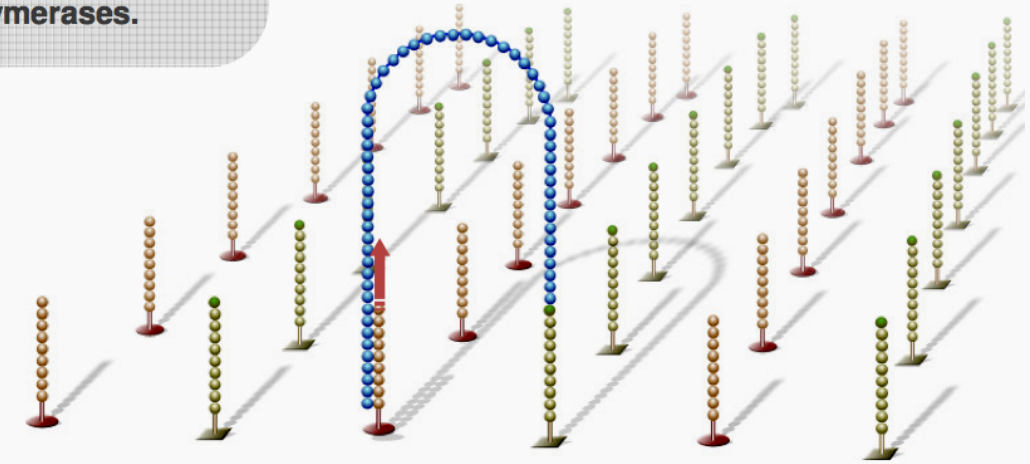


2 – Cluster growth

Cluster generation:
Bridge amplification

Single-strand flips over to hybridize to adjacent primers to form a bridge.

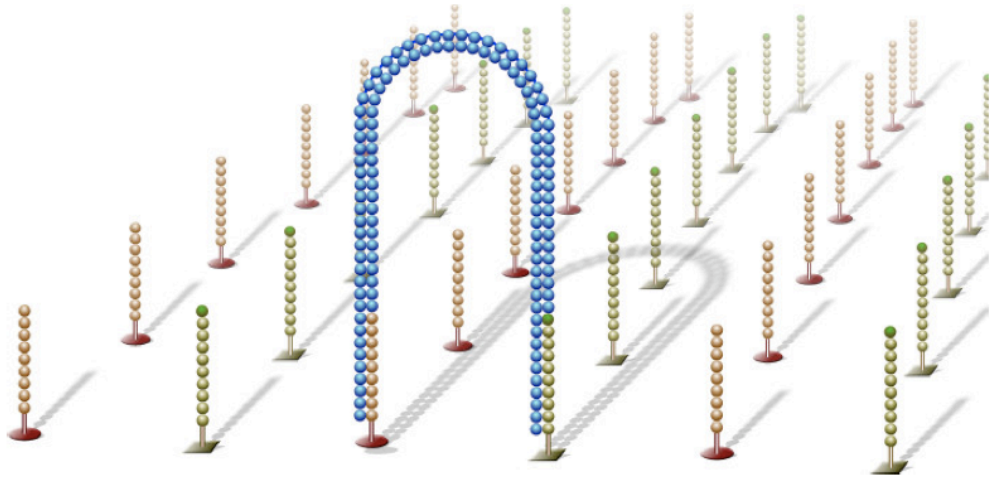
Hybridized primer is extended by polymerases.



2 – Cluster growth

Cluster generation:
Bridge amplification

→ **double-stranded bridge is formed.**



2 – Cluster growth

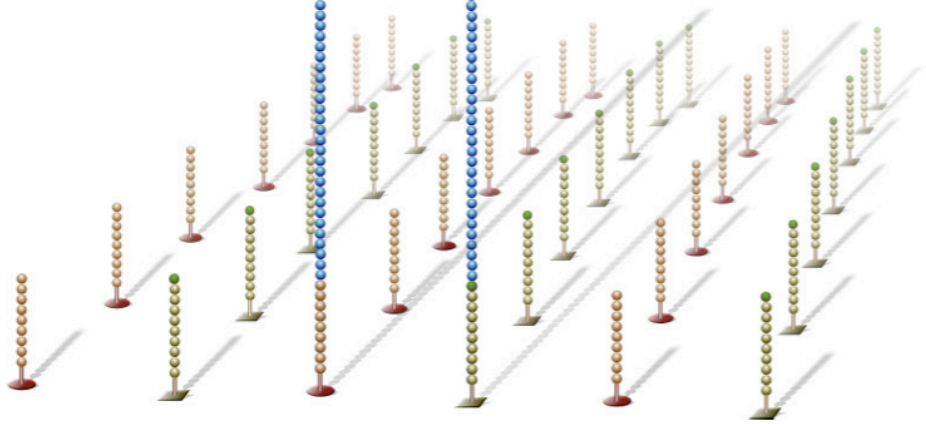
Cluster generation:
Bridge amplification

identical to original

complementary to original

Double-stranded bridge is denatured.

Result: Two copies of covalently bound single-stranded templates.

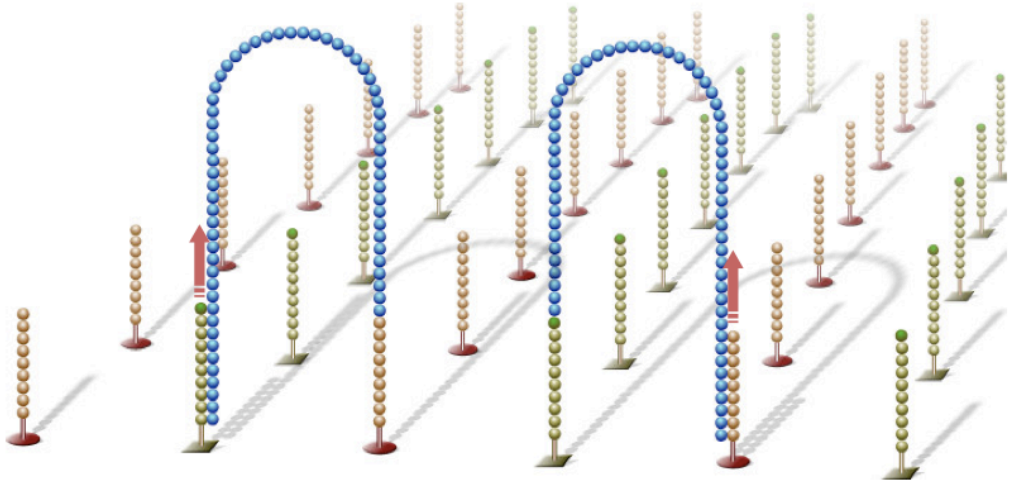


2 – Cluster growth

Cluster generation:
Bridge amplification

Single-strands flip over to hybridize to adjacent primers to form bridges.

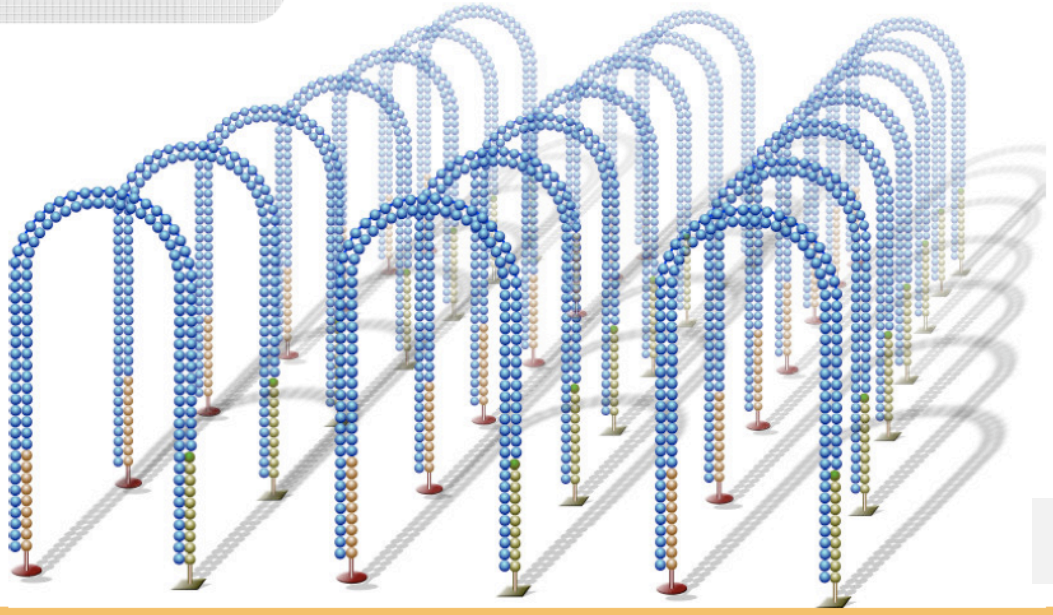
Hybridized primer is extended by polymerase.



2 – Cluster growth

Cluster generation:
Bridge amplification

Bridge amplification cycle repeated till multiple bridges are formed

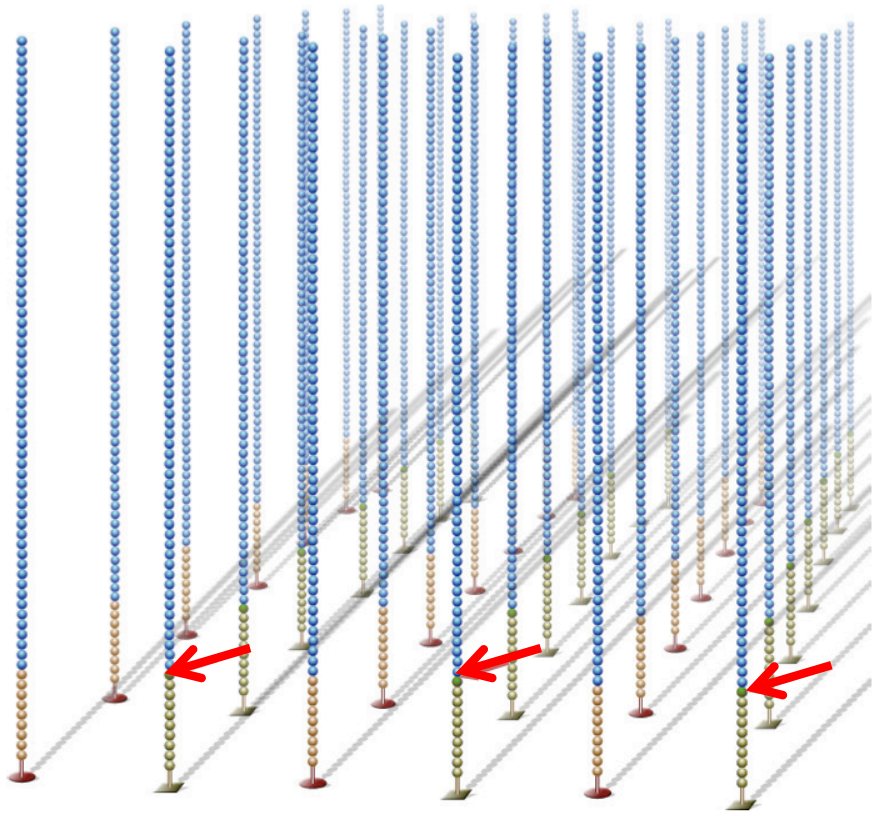


2 – Cluster growth

Cluster generation

dsDNA bridges denatured.

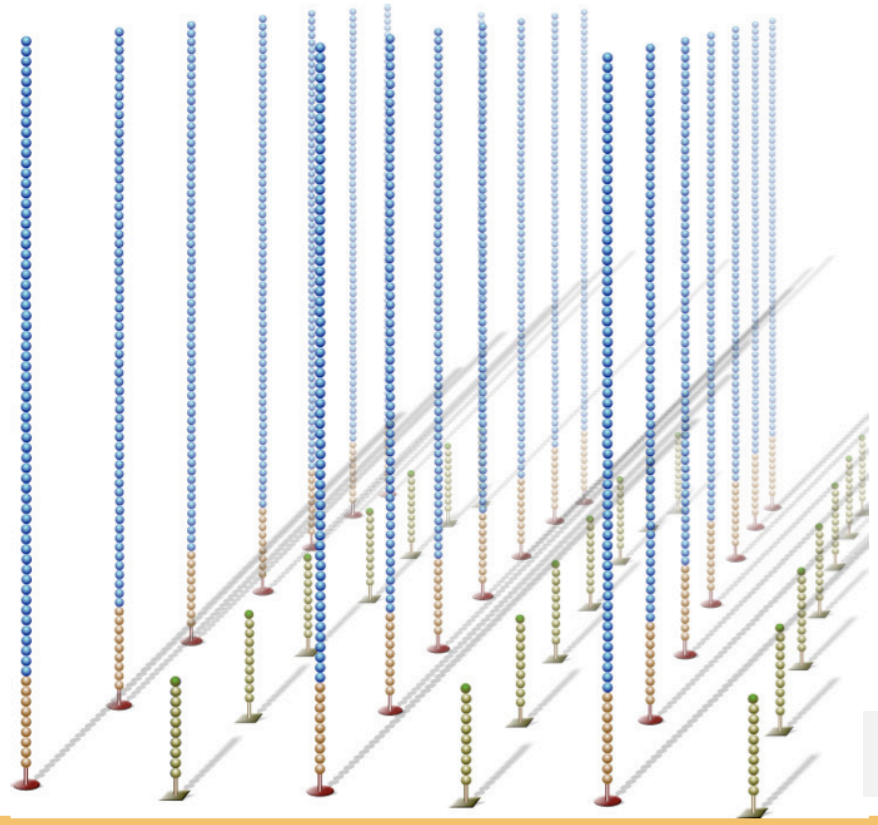
Reverse strands cleaved and washed away.



2 – Cluster growth

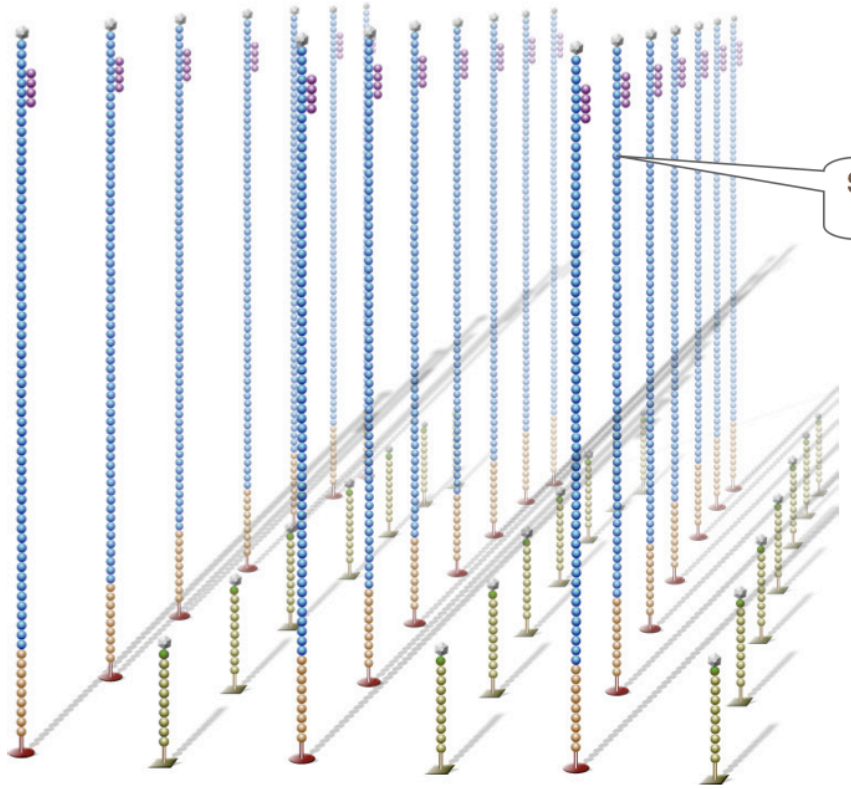
Cluster generation

… leaving a cluster with forward strands only.

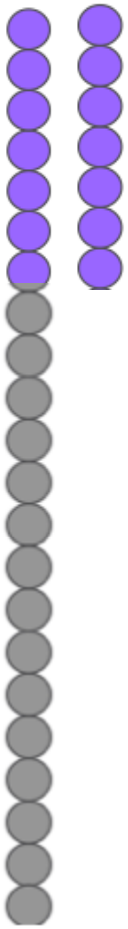


3 – Sequencing

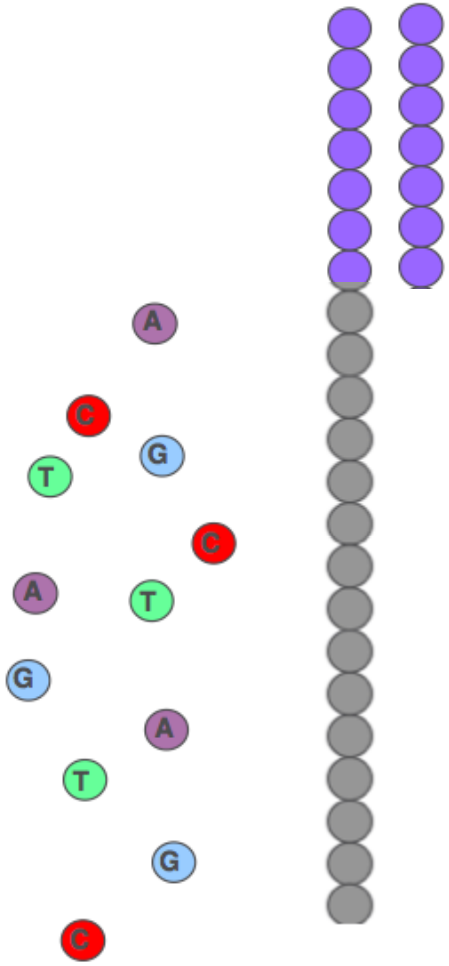
Sequencing primer is hybridized to adapter sequence.



3 - Sequencing By Synthesis (SBS)

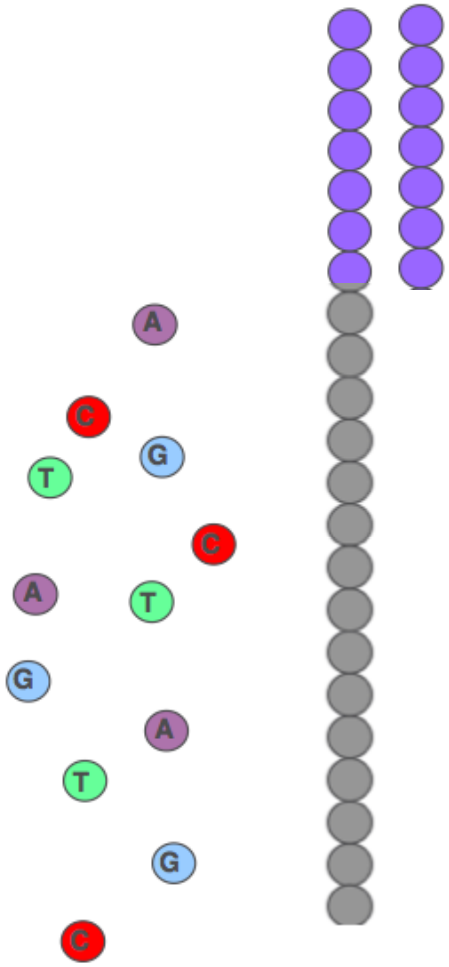


3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

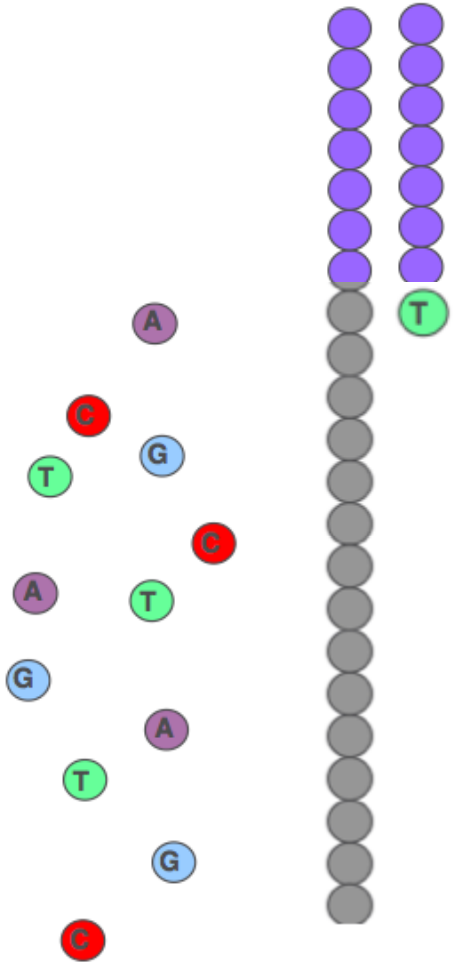
3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents

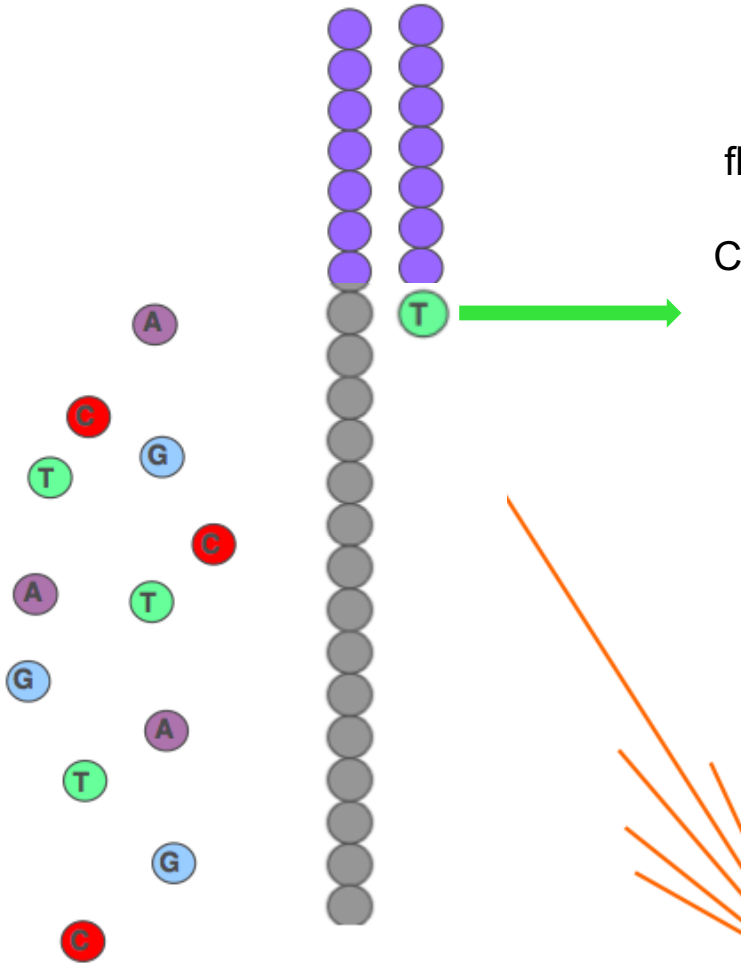
3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated

3 - Sequencing By Synthesis (SBS)

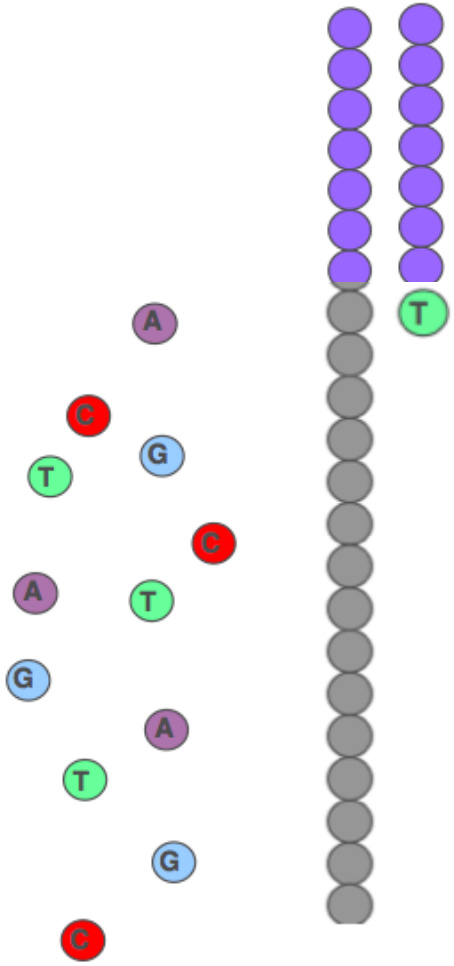


fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan

LASER

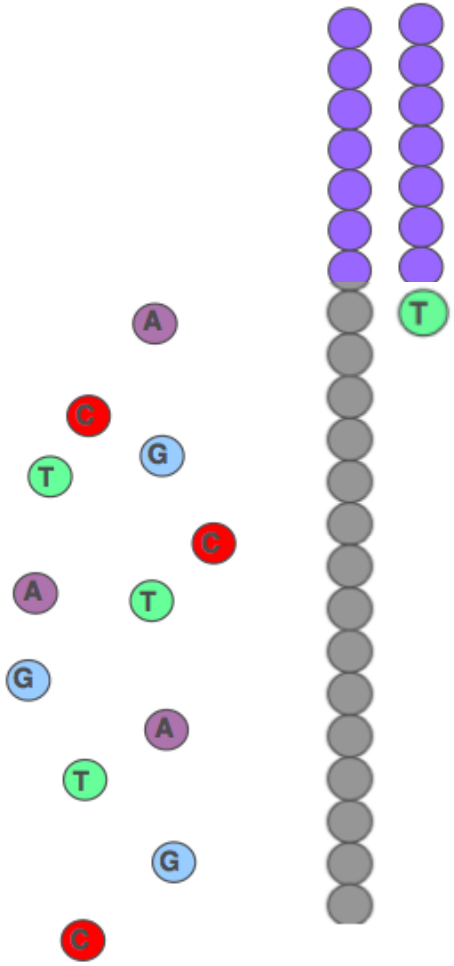
3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

3 - Sequencing By Synthesis (SBS)

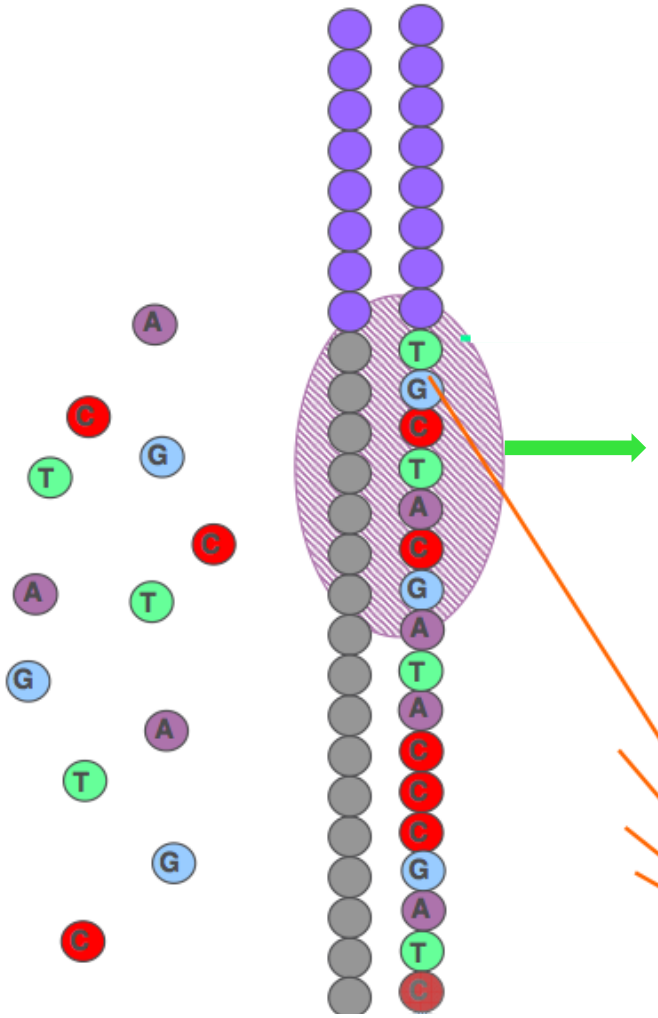


fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

Cycle 2 : add sequencing reagents and repeat

3 - Sequencing By Synthesis (SBS)



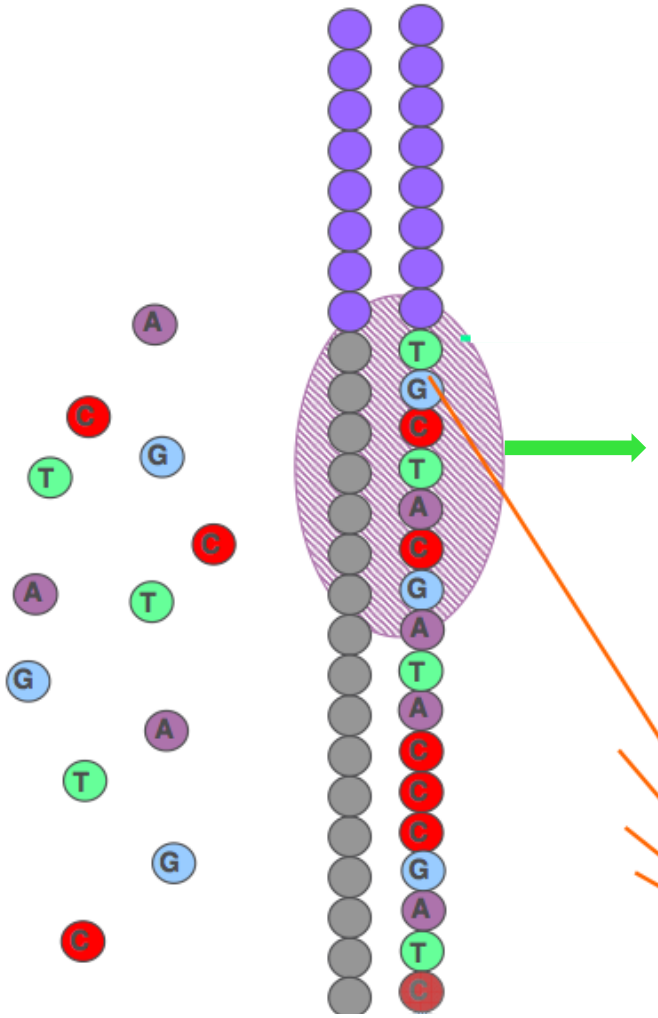
fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

Cycle 2 : add sequencing reagents and repeat

LASER

3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

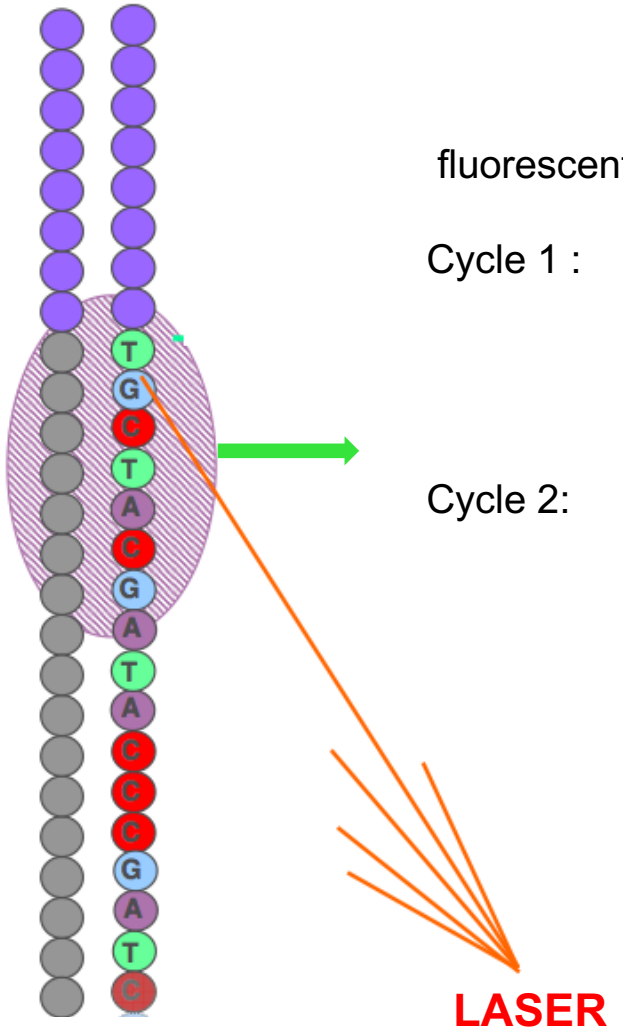
Cycle 2 : add sequencing reagents and repeat

4 terminator-bound dNTPs present during each cycle

natural competition minimizes incorporation bias

LASER

3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

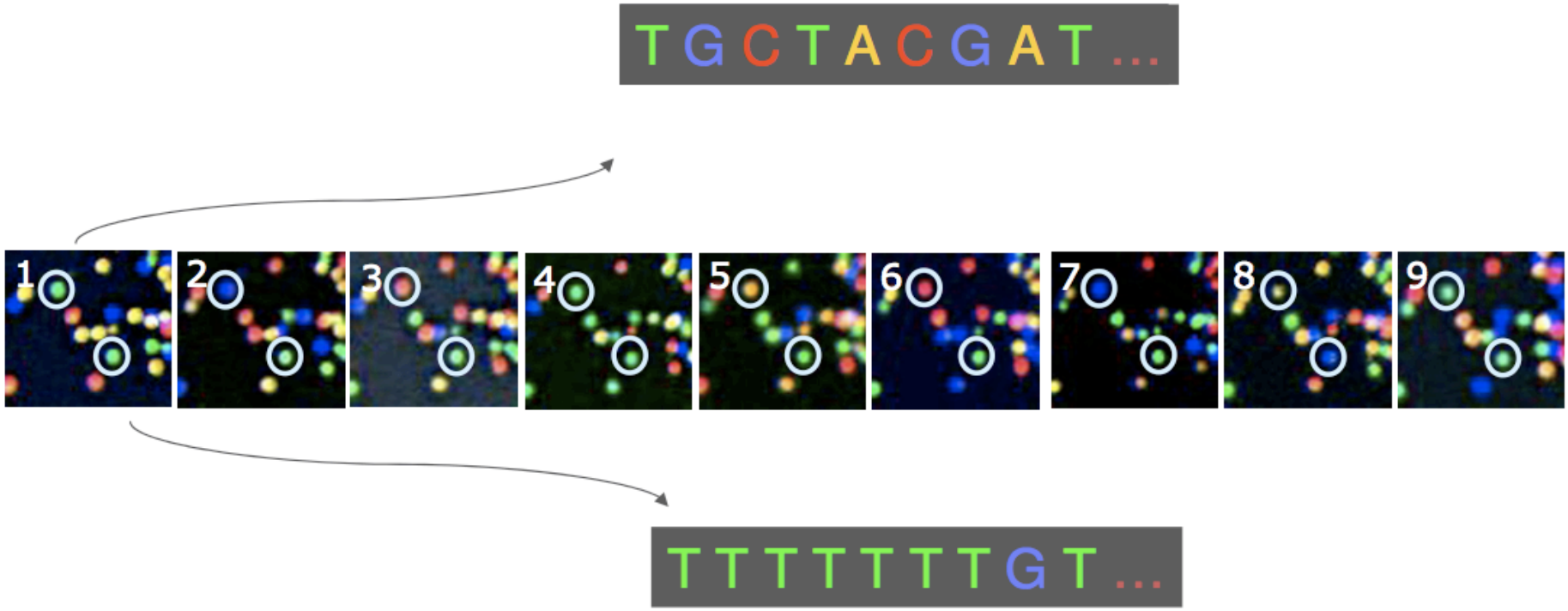
Cycle 2: add sequencing reagents and repeat

LASER

$(C_{eff})^{RL} = 0.5$

Cycle Efficiency (%)	Read-length (bases)
90.0	7
95.0	14
98.0	35
99.0	69
99.5	149
99.9	693

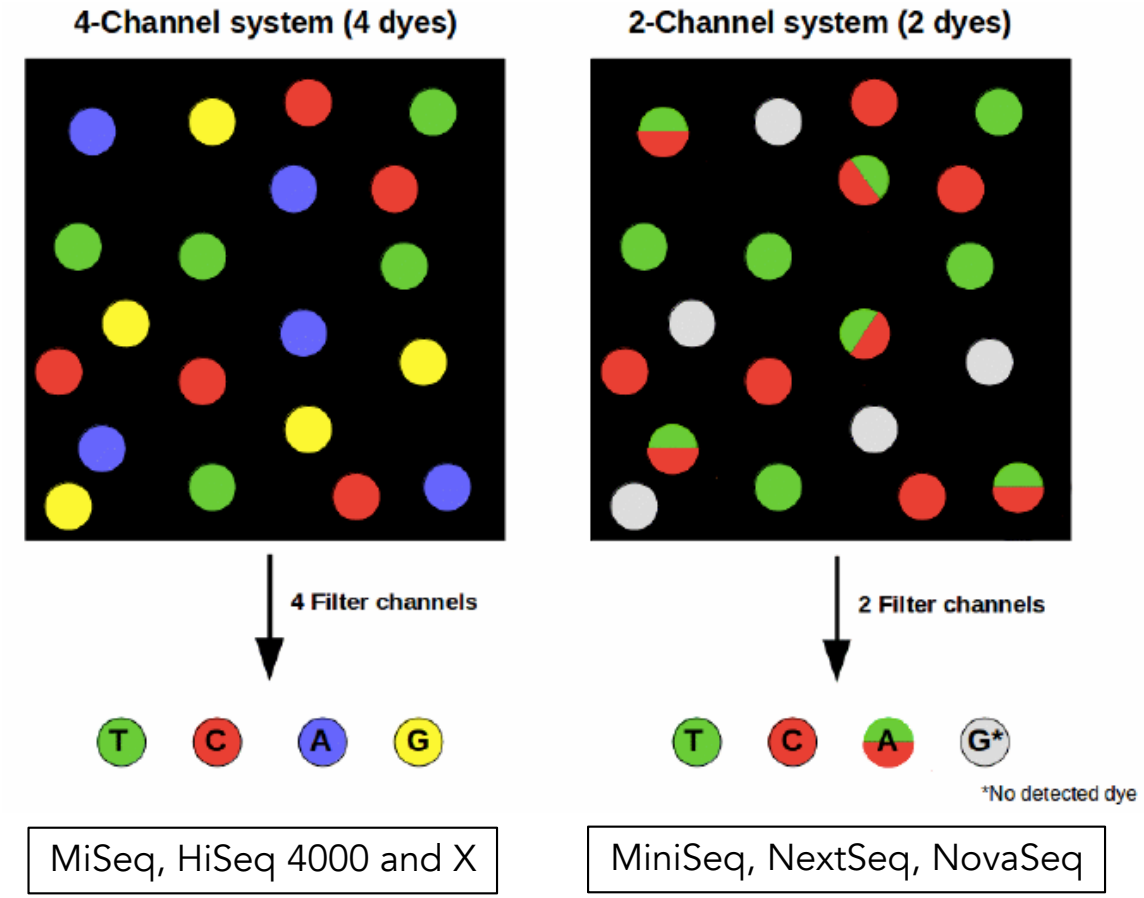
Base calling from raw data



The identity of each base of a cluster is read off from sequential images.

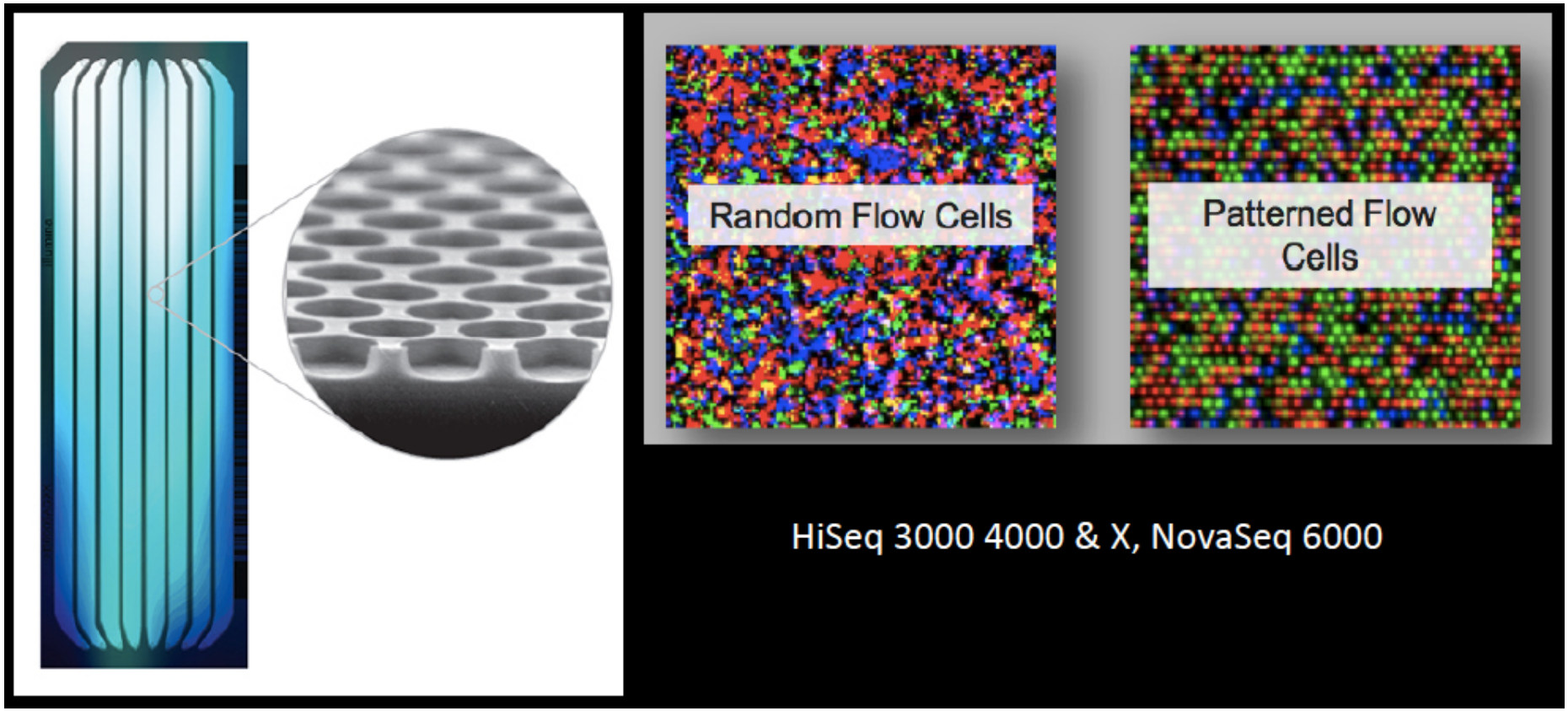
4 channel vs 2 channel detection

2 channel detection can cause sequencing errors due to phasing issues and can cause polyG tracts at the end of fragments



Patterned flow cells

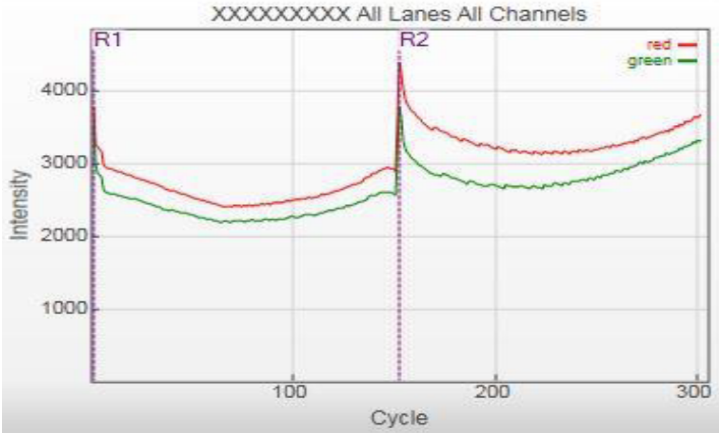
- Improves regularity of densities and qualities
- Reduces analysis time



Sequencing qualities

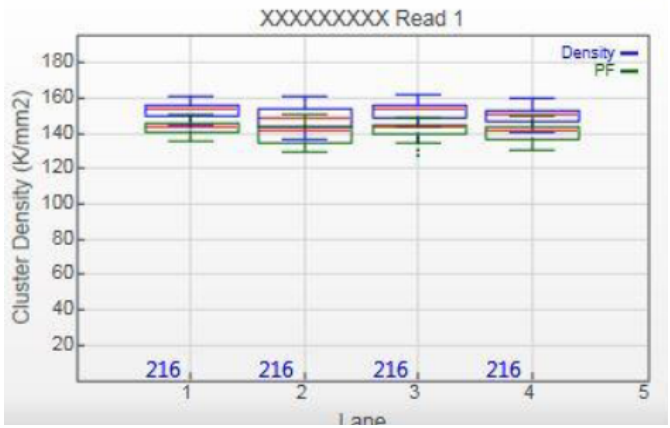
Intensities

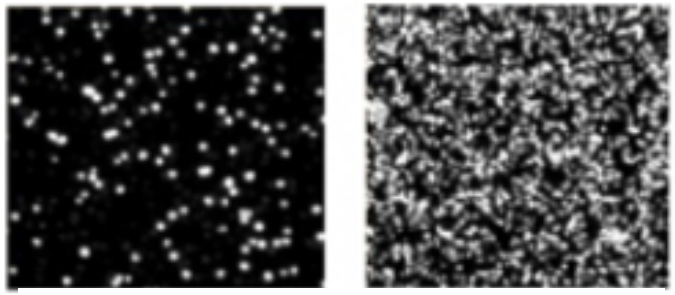
Intensity values of the four incorporated bases are measured to determine if a clustering failure has occurred



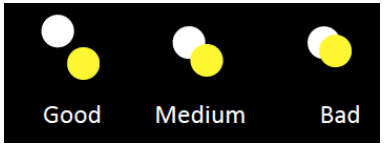
Cluster densities

Density of clusters for each tile in thousands/mm^2. Overloading of the library may cause merging of clusters, lowering of the % PF and reduction of the quality score (Q30)





Underclustered ⟶ Overclustered



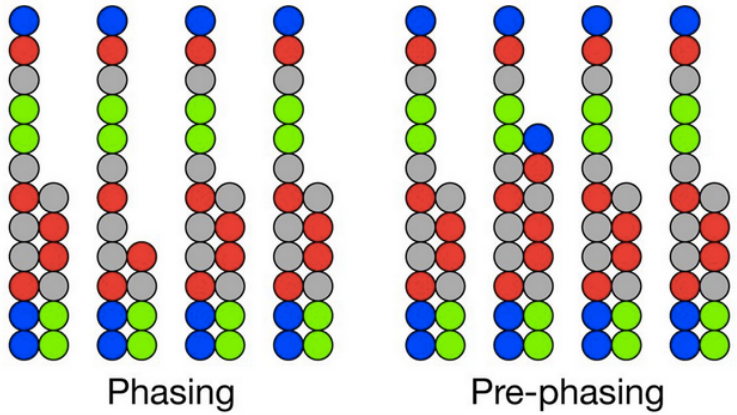
Percentage of reads passing filter (% PF)

Reflects the purity of the signal from each cluster. Ratio of the brightest intensity divided by the sum of the brightest and second brightest intensities for each cycle. Optimal values for MiSeq and HiSeq 2500 from 80-95%.

Sequencing qualities

Phasing/Prephasing

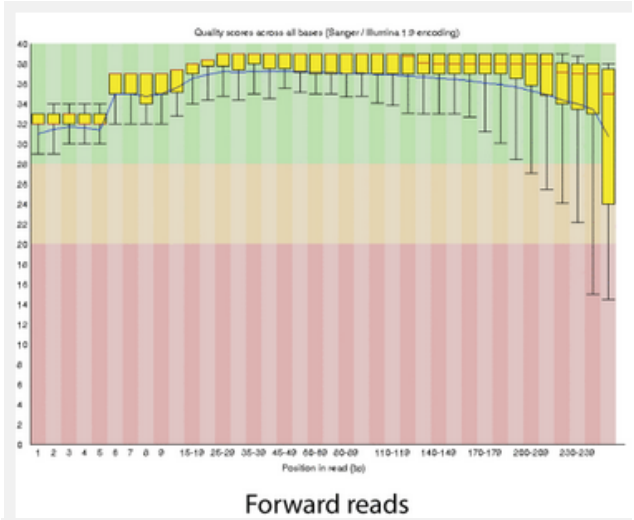
Percentages of bases that fell behind or jumped ahead the current cycle within a read. It increases with cycle number, hampering correct base identification for long reads. Optimal values are below 0.5 or 0.2% depending on platform. Issues may arise when read length is above specifications.

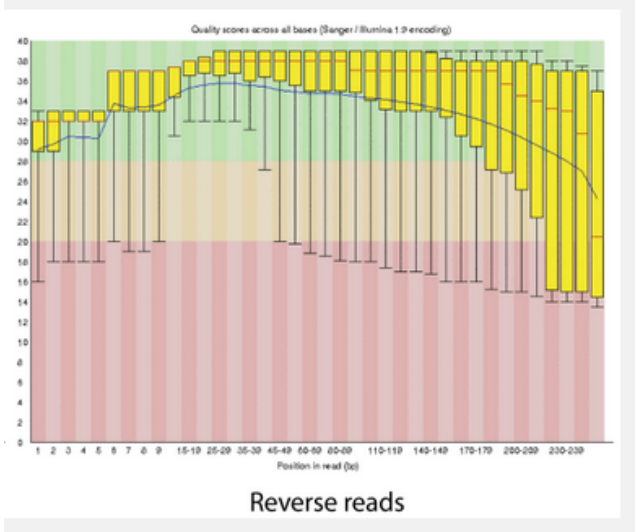


Q-score

Base calling error probabilities P : $Q = -10 \log_{10} P$

% Q-score > 30 : percentage of bases that have a a probability of incorrect base calling of 1/1000. Low Q scores can increase false-positive variant calls, which can result in inaccurate conclusions.

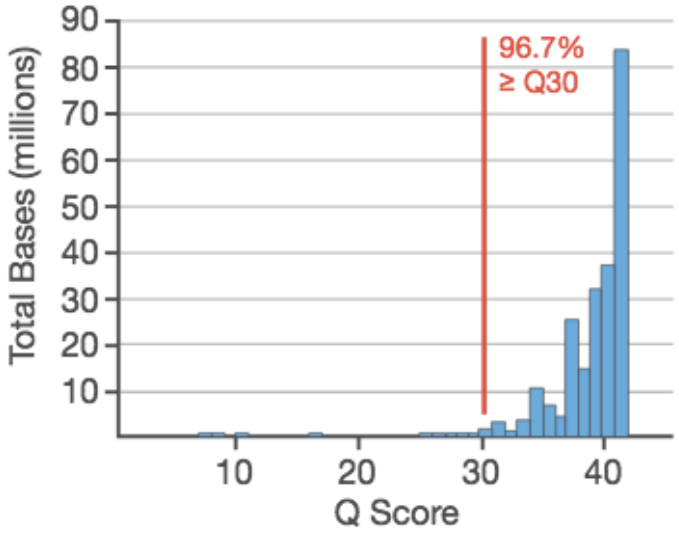




Sequencing qualities

Percentage of alignment

% align: percentage of the reads aligned to PhiX control genome. It is important to estimate the error rate. Indication on the success of the clustering process.



Error rate

Calculated for each read based on the aligned % of PhiX control sample. Important to determine if the sequencing has proceeded as expected.