Introduction NGS

Claude Thermes, Morgane Thomas-Chollier

École de bioinformatique AVIESAN-IFB-INSERM 2024

From the samples to the reads :what happens in a sequencing core facility ?

1 **Biologist brings the samples (DNA, RNA, cells)**

Platform

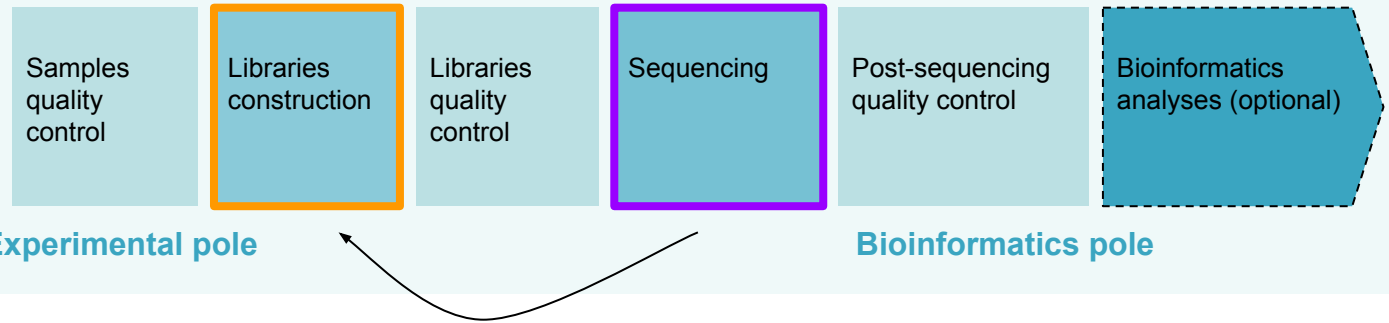
2

From the samples to the reads :what happens in a sequencing core facility ?

1 **Biologist brings the samples (DNA, RNA, cells)**

Platform

2



2 main activities

The protocol of library preparation is directly dependent on the sequencer (and on the sample type)

Different “generations” of sequencers

Sequencing

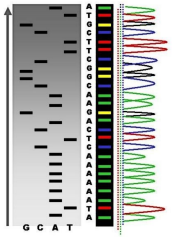
1rst generation : Sanger sequencing

- Has been the major methodology up to 2005

Limitations

- Extremely high cost
- Long experimental set up times
- High DNA concentrations needed

Sanger



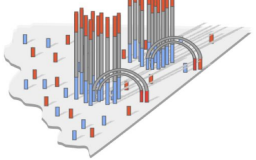
2d generation

- Very high troughput
- Low cost

Limitations

- Maximum read length ≤ 300bp

Illumina

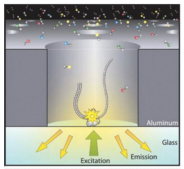


short reads

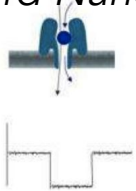
3rd generation

- Single molecules sequencing
- Very long reads

PacBio



Oxford Nanopore

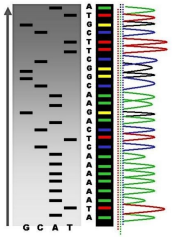


long reads

Different “generations” of sequencers

Sequencing

Sanger



1rst generation : Sanger sequencing

- Has been the major methodology up to 2005

Limitations

- Extremely high cost
- Long experimental set up times
- High DNA concentrations needed

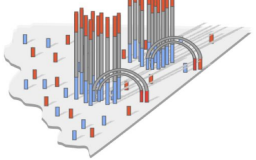
2d generation

- Very high troughput
- Low cost

Limitations

- Maximum read length ≤ 300bp

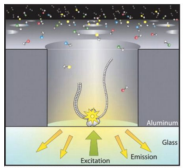
Illumina



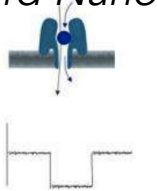
3rd generation

- Single molecules sequencing
- Very long reads

PacBio

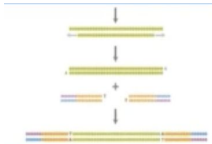


Oxford Nanopore

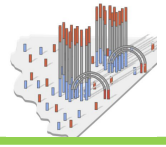


Illumina sequencing workflow

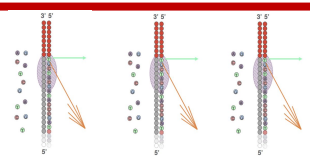
1 - Library preparation



2 - Cluster generation



3 - Sequencing



4 - Data analysis

```
cagaaactgcagattagcgtgtatatttatctgtttatgct
cagaaactgcagattagcgtgtatatttatctgtttgtgct
cagaaactgcagatttatgtgtatatttatctgtttatgct
cagaaactgcagatttttgtgtatatttatctgtttatgct
cagaaactggcggtgtatgtgtatatttatctgtttatgca
```

Libraries construction

Sequencing

Bioinformatics analyses (optional)

1 - Library preparation

Genomic DNA

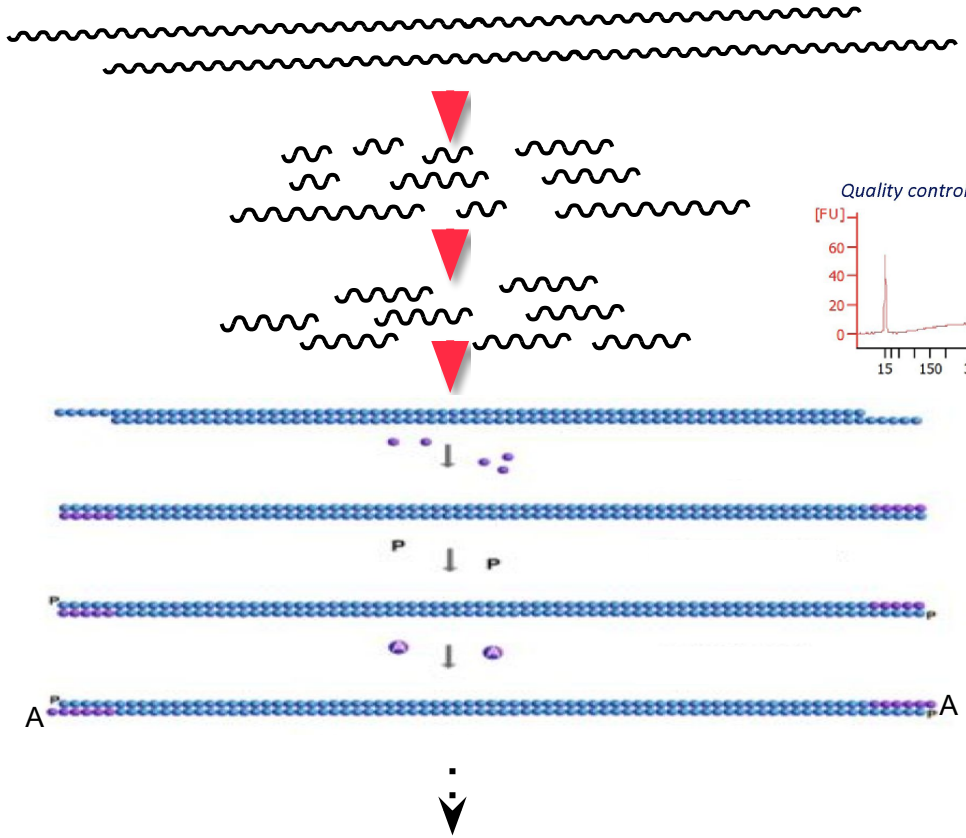
Sonication

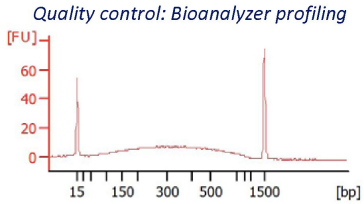
Size selection

End repair

Phosphorylation

A - overhang





1 - Library preparation

What is an adapter ?

Adapters = DNA (~80nt), which attach to the **DNA fragments of interest** + **primers for amplification**. Adapters also **bind to the DNA linkers** on the flow cell's solid surface

Flow cell binding sequence: Platform-specific sequences for library binding to instrument

Sequencing primer sites: Binding sites for general sequencing primers

Sample indexes: Short sequences specific to a given sample library

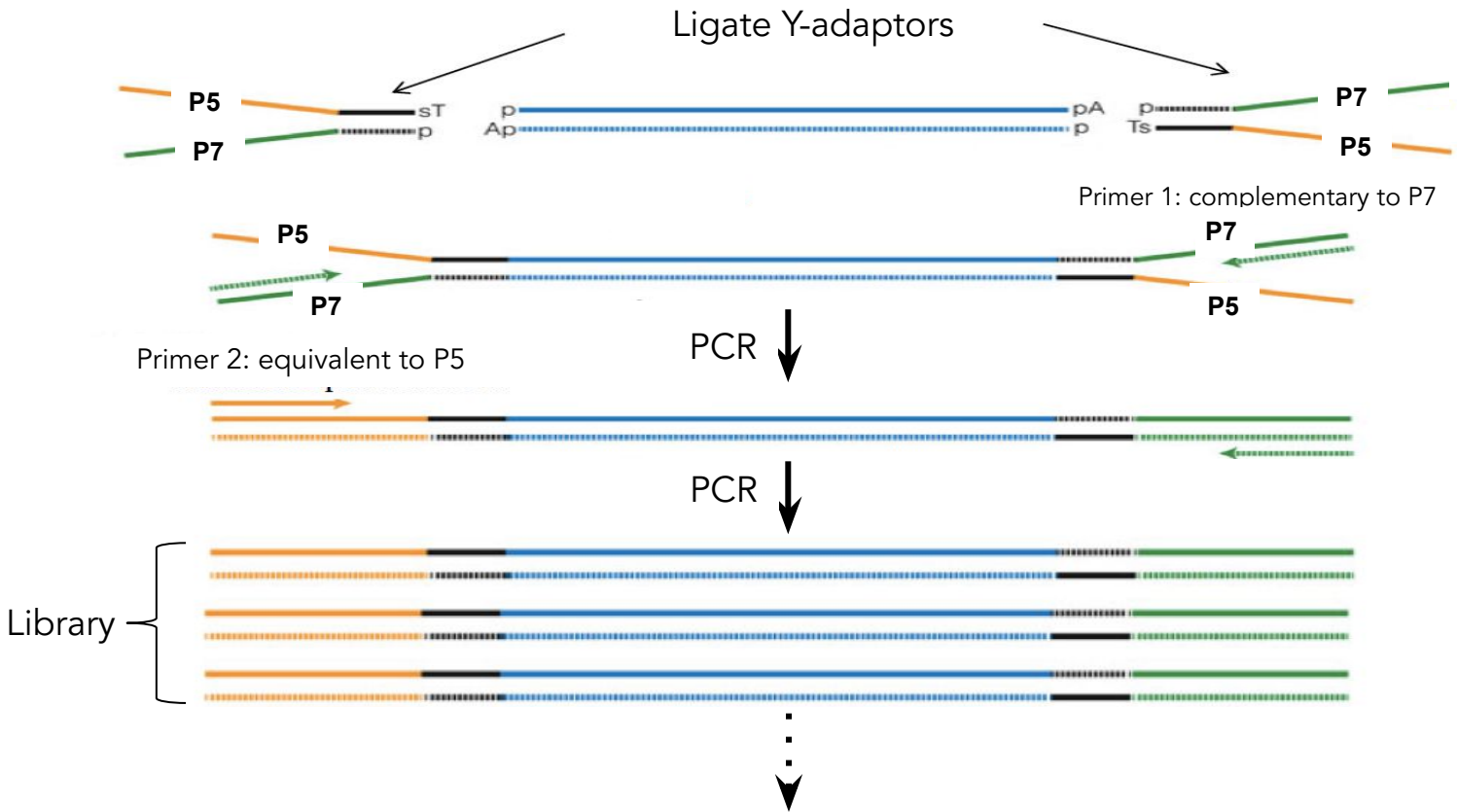
enables multiplexing of samples on a same flowcell

adapter : DNA not supposed to be sequenced, present for technical reasons (except for index)

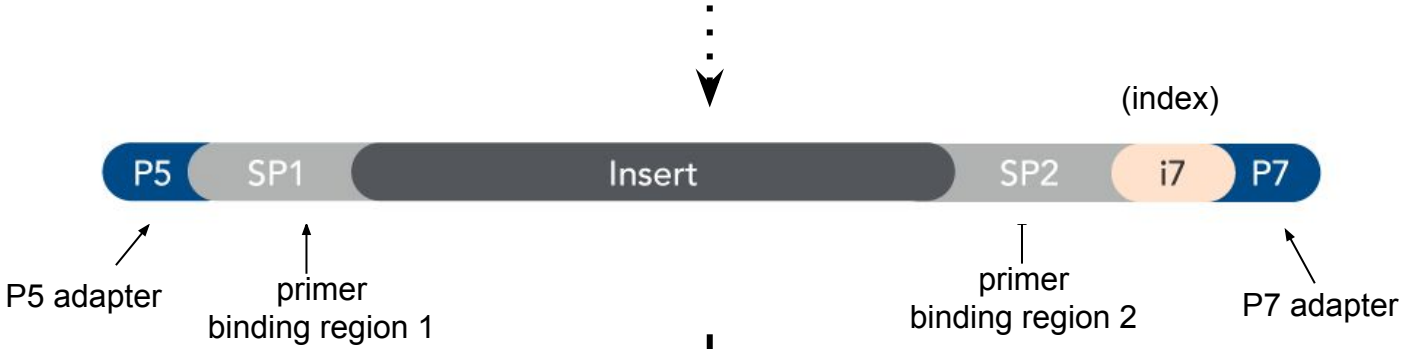
Insert: Target DNA or RNA fragment from a given sample library → this is the fragment we want to sequence

1 - Library preparation

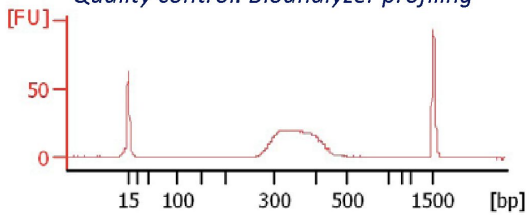
How are the adapters attached to the DNA of interest ?



1 - Library preparation



Quality control: Bioanalyzer profiling



cluster generation - sequencing

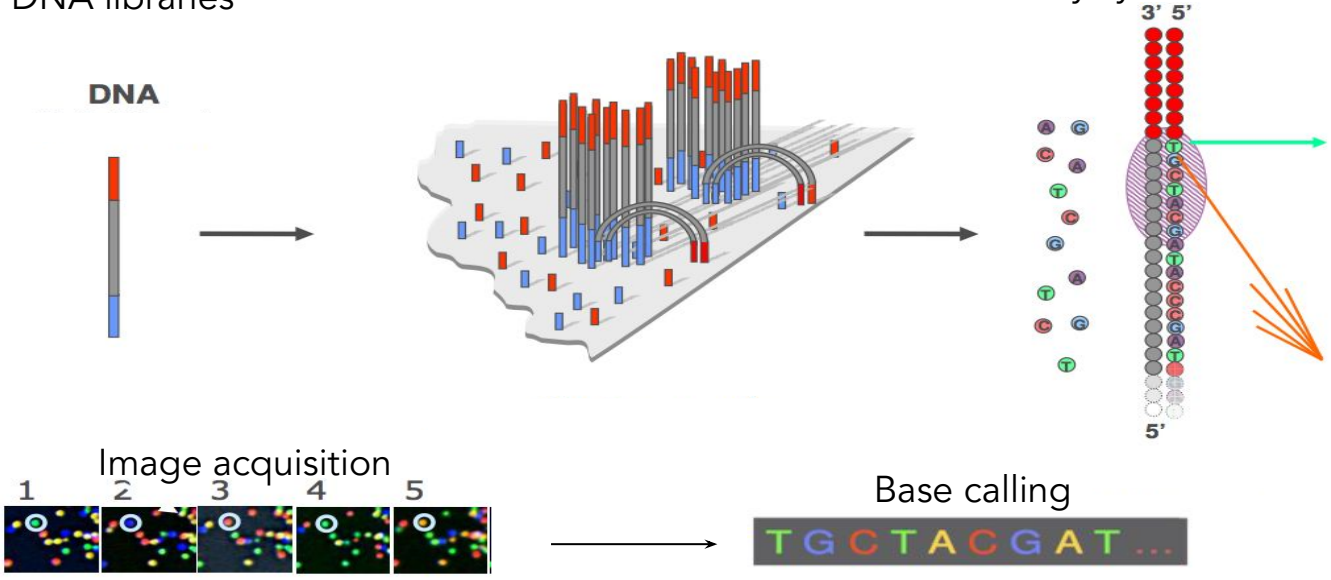
Illumina sequencing

General scheme

1
Preparation of DNA libraries

2
Cluster growth

3
Sequencing by synthesis



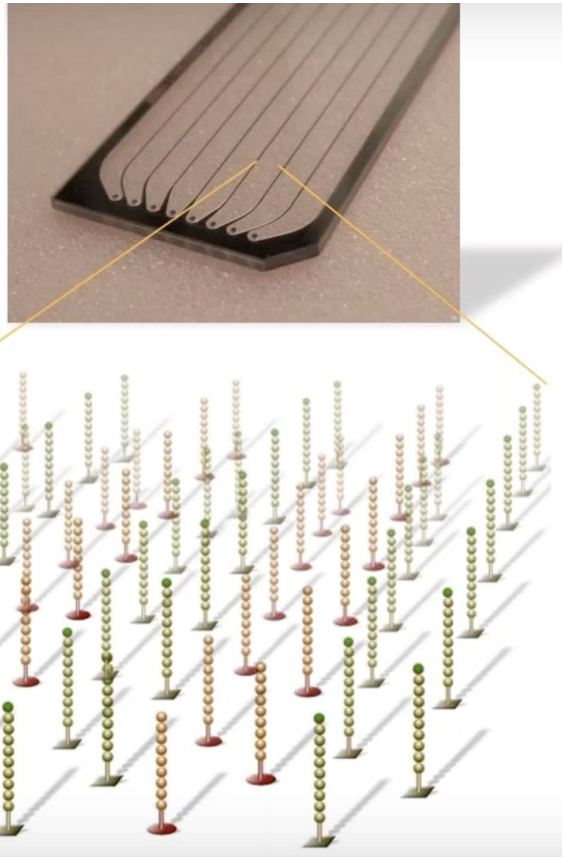
2 – Cluster generation

What is a flow cell ?

Cluster generation occurs on a flow cell

A flow cell is a thick glass slide with channels or lanes

Each lane is coated with a lawn of oligos complementary to library adapters

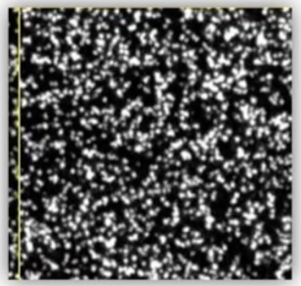


2 – Cluster generation

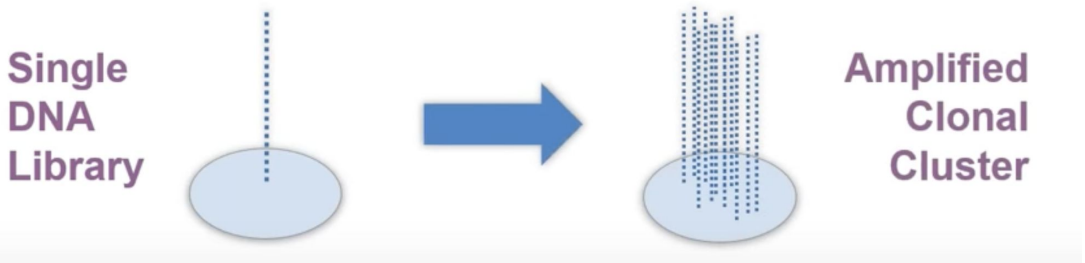
What is a cluster ?

Clusters are a group of DNA strands positioned closely together

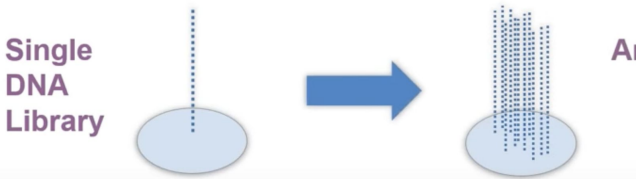
Each cluster represents thousands of copies of the same DNA strand in a 1–2 micron spot



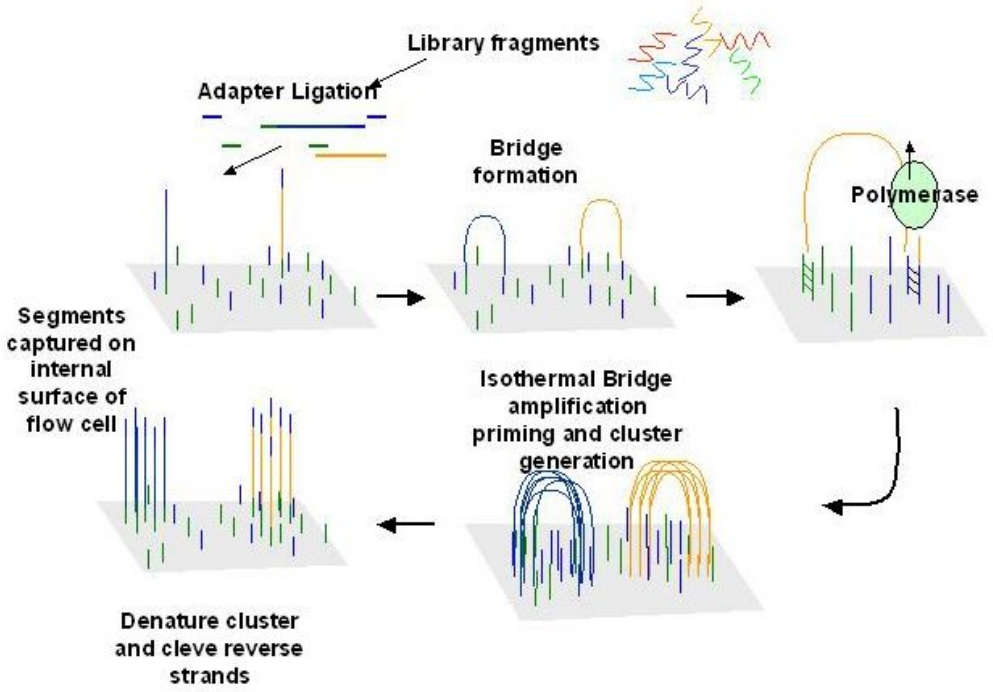
An image of fluorescently labelled clusters on a flow cell



2 – Cluster generation



How are cluster generated ?



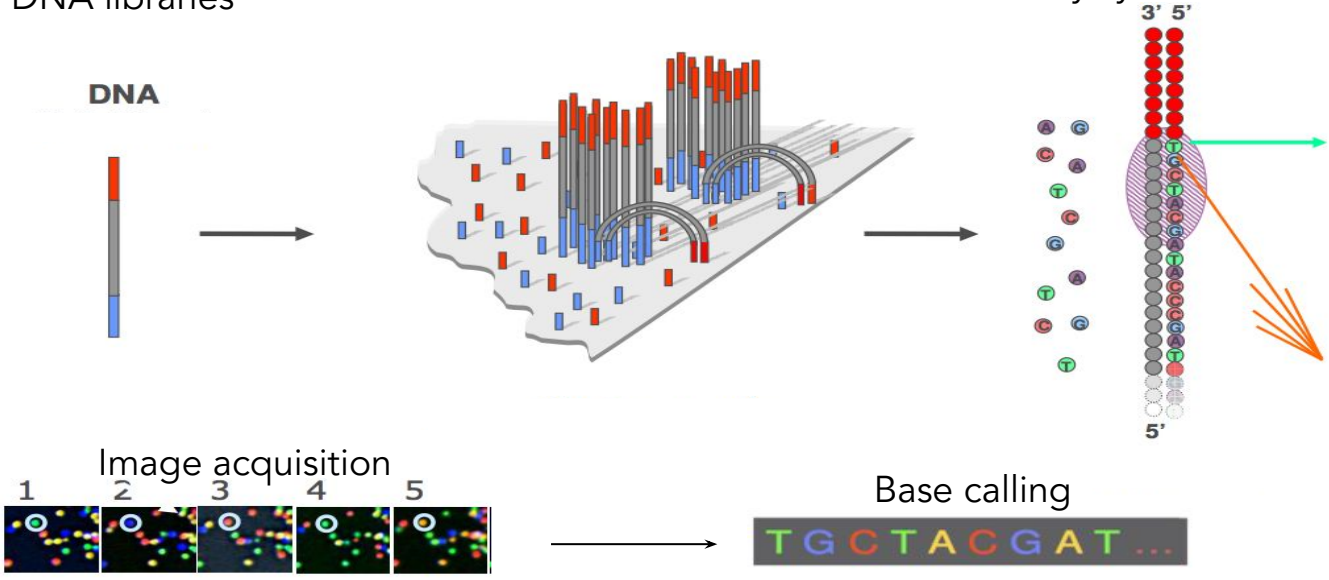
Illumina sequencing

General scheme

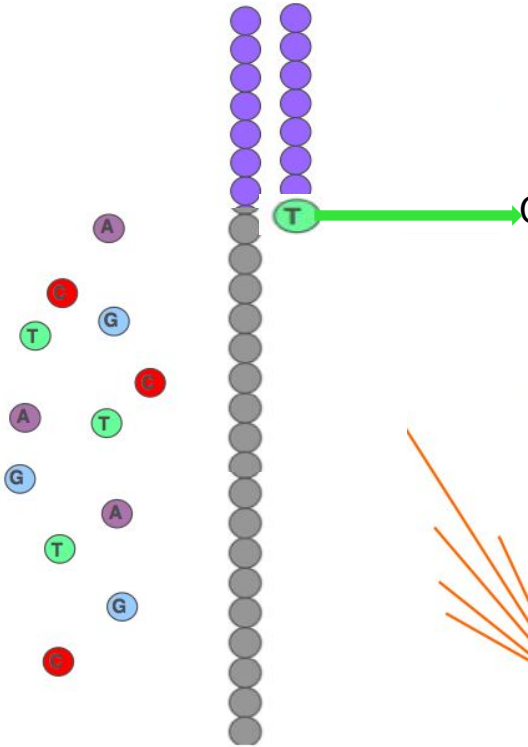
1
Preparation of DNA libraries

2
Cluster growth

3
Sequencing by synthesis



3 - Sequencing By Synthesis (SBS)

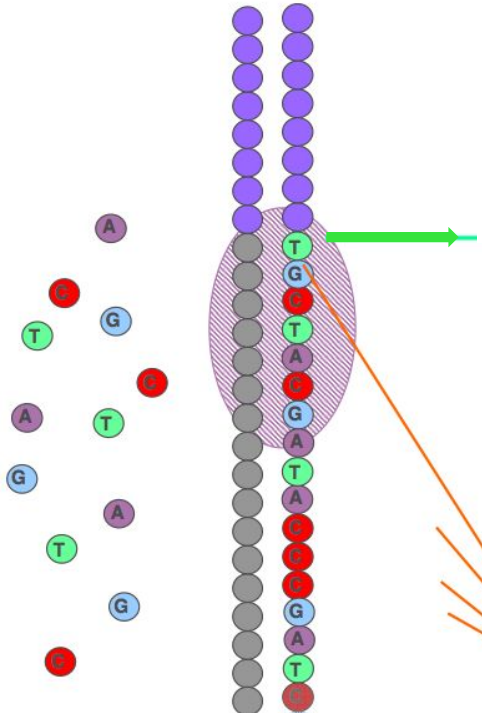


fluorescent-labeled terminator bound to each dNTP

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

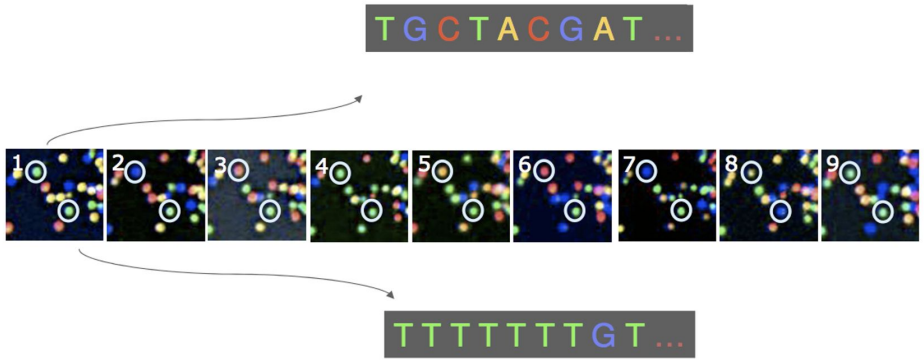
Cycle 2 : add sequencing reagents and repeat

4 terminator-bound dNTPs present during each cycle

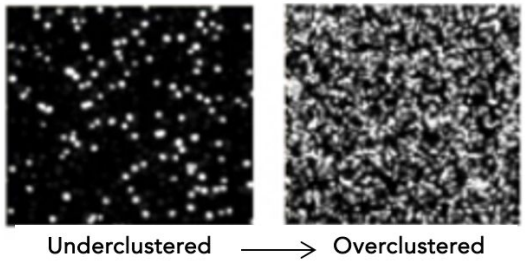
natural competition minimizes incorporation bias

LASER

3 – Sequencing



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



3 – Sequencing

What is a read ?

Read = extremity of the insert that is sequenced

DNA or cDNA insert

3 – Sequencing

What is a read ?

Read = extremity of the insert that is sequenced

DNA or cDNA insert

and what is a read for a bioinformatician ?

ATTTCGCATTTACGCTTTTA

Read = one sequence

3 – Sequencing

What is a read ?

Read = extremity of the insert that is sequenced

DNA or cDNA insert

and what is a read for a bioinformatician ?

ATTTCGCATTTACGCTTTTA

Read = one sequence

all reads = one file

3 – Sequencing

SINGLE READ and PAIRED-END SEQUENCING

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

- **Paired end**: Sequence both physical ends of DNA insert (generally fragment < 800nt)

3 – Sequencing

SINGLE READ and PAIRED-END SEQUENCING

- **Single end**: one file with all the **reads**

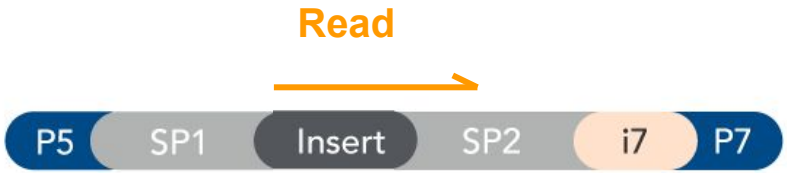
- **Paired end**: 2 files : one with **all reads1** and one with **all reads2**

3 – Sequencing

Possibility to find the adapter sequence in the read sequence ?

read length < insert length

Yes : If the sequencing length (e.g. 150 nt) is longer than the length of the small DNA inserts present in the library

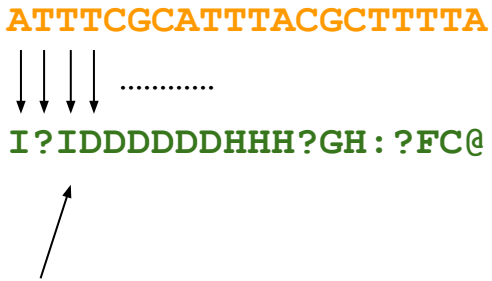


read length > insert length
the adapter is thus sequenced

3 – Sequencing

What is the quality of the reads ?

The sequencer outputs both
- the nucleotides of the reads
- a quality value indicating how "sure" the sequencer is that the nucleotide is the right one



Each symbol corresponds to a quality value from bad to excellent
cf. FASTQ format in the next course

Vocabulary/concepts important to remember

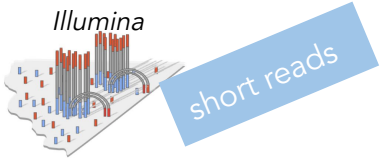
- library (banque)

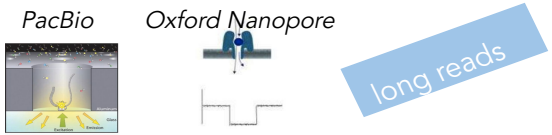
- adapter (adaptateur)

- read (lecture)

- single-end versus paired-end sequencing

- short-read (2nd génération) versus long-read (3rd génération séquencing)





- adapter in reads ?

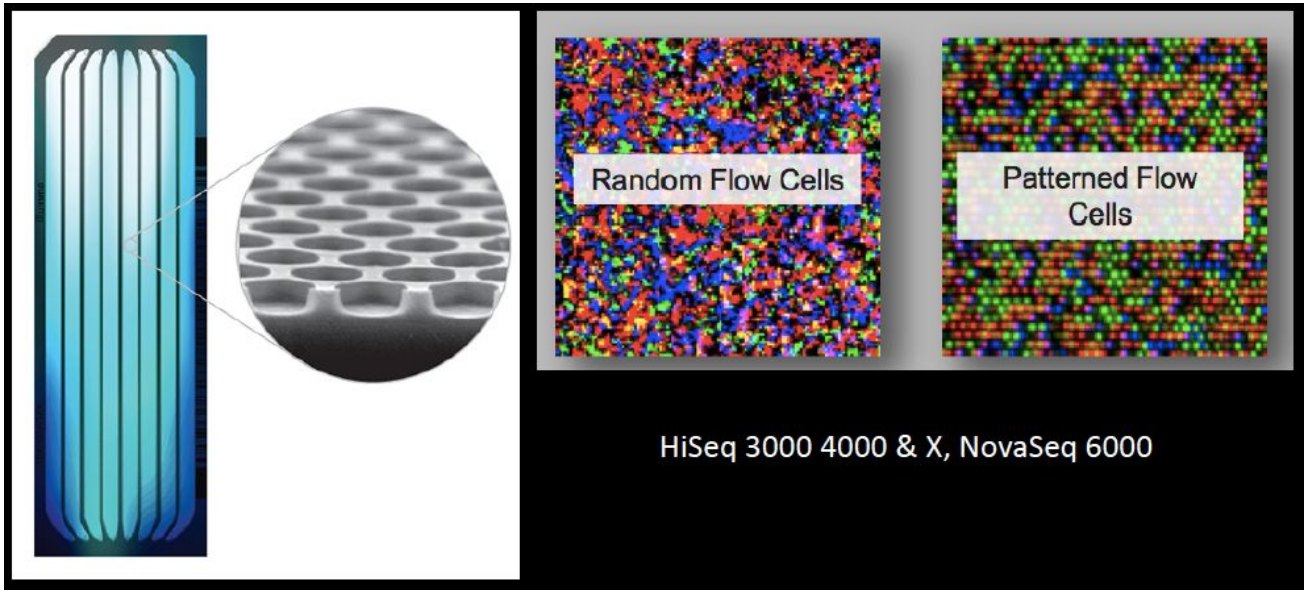
- read quality

```
ATTTCGCATTTACGCTTTTA
I?IDDDDDDHHH?GH:?FC@
```

Supplementary

Patterned flow cells

- Improves regularity of densities and qualities
- Reduces analysis time

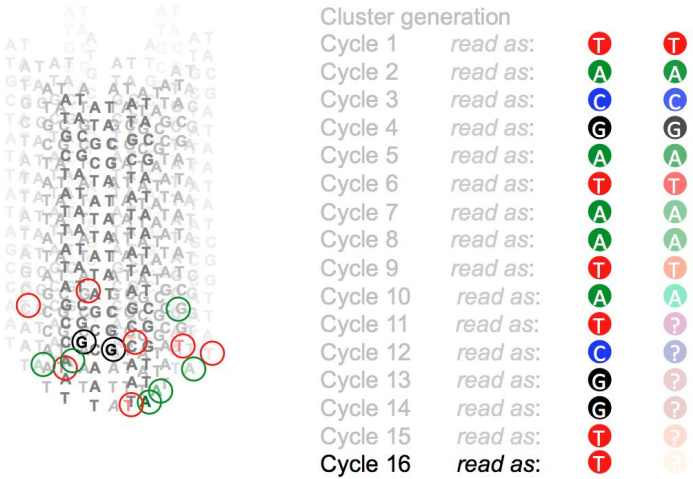


3 – Sequencing

“Dephasing” due to partial blockage of DNA synthesis

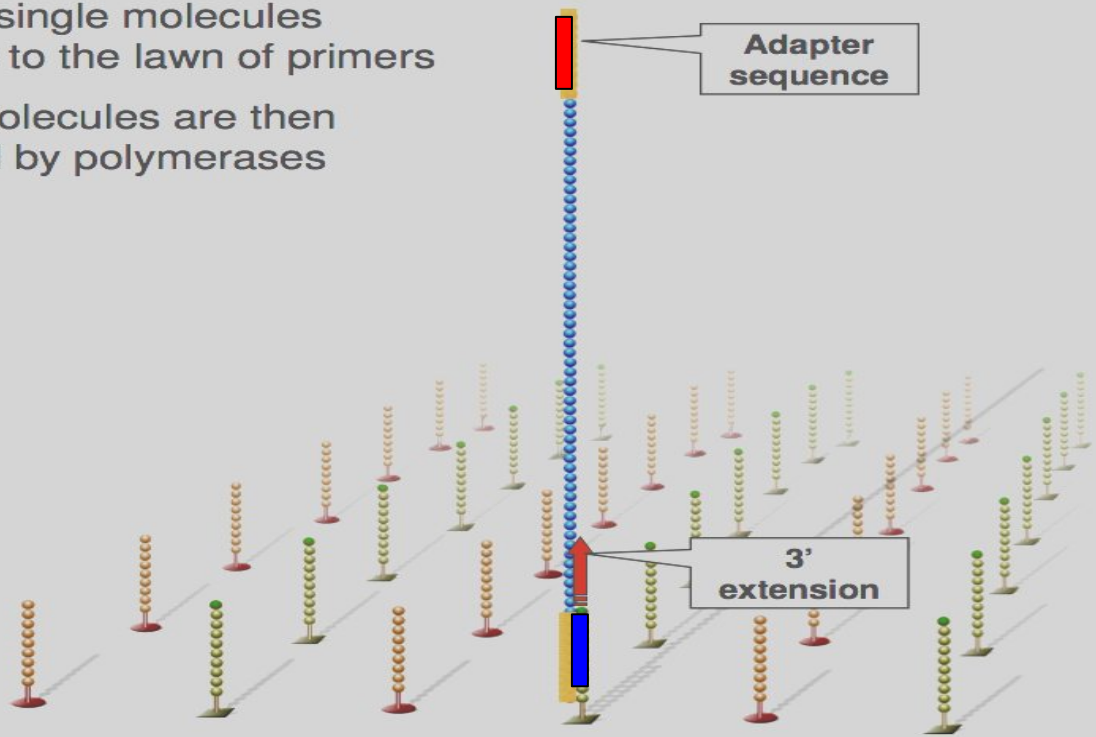
3 – Sequencing

“Dephasing” due to partial blockage of DNA synthesis



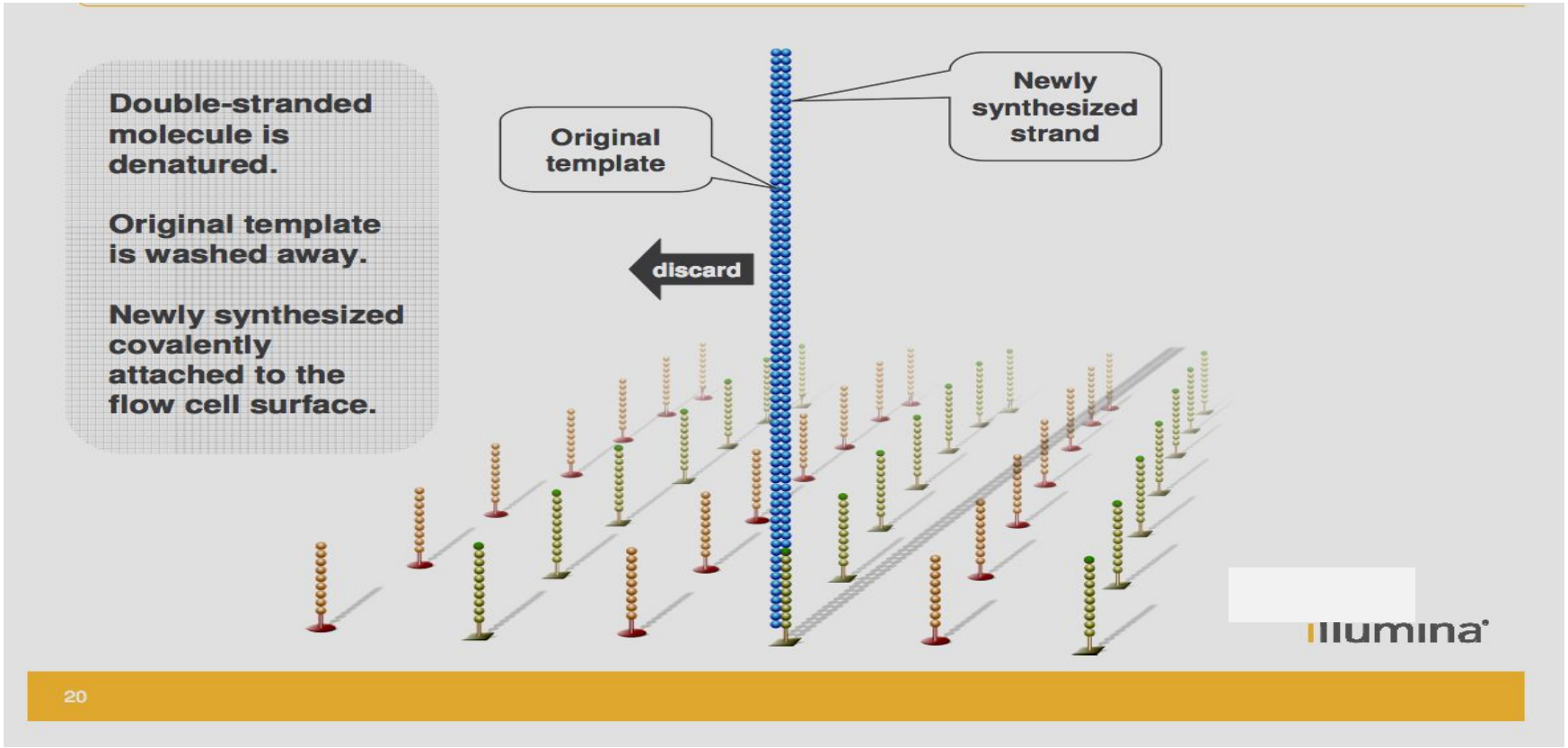
2 – Cluster generation

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

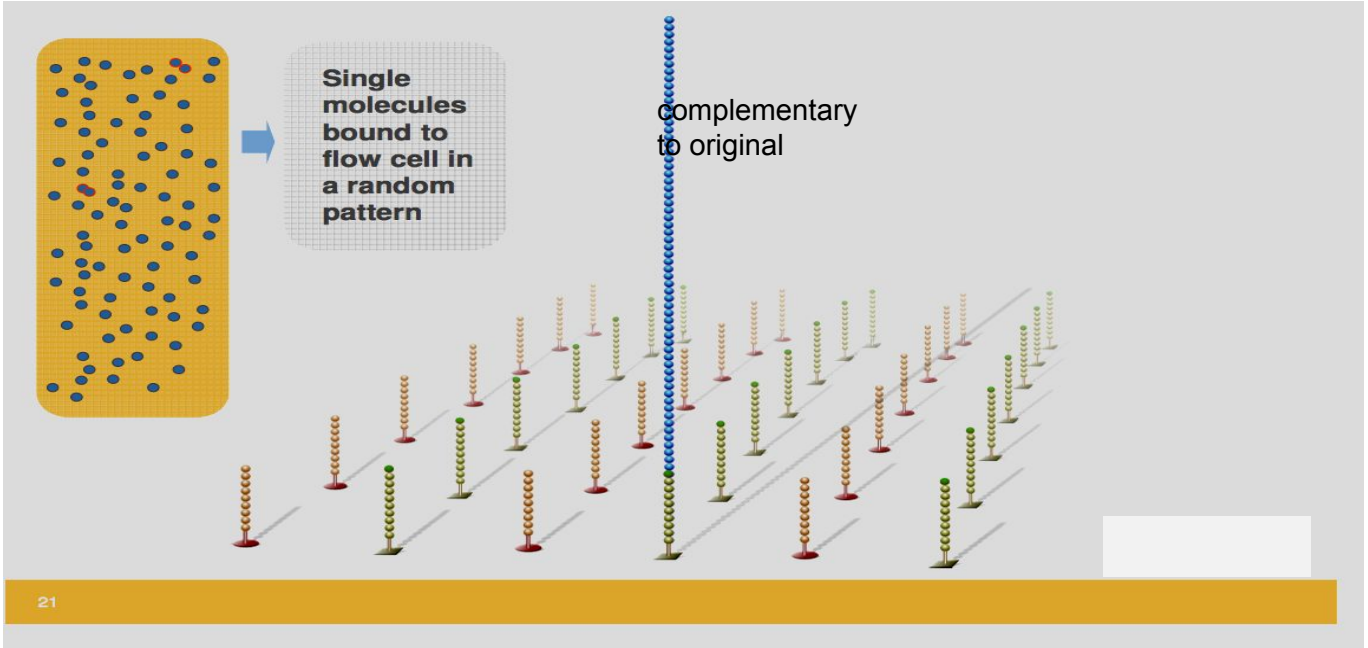


illumina

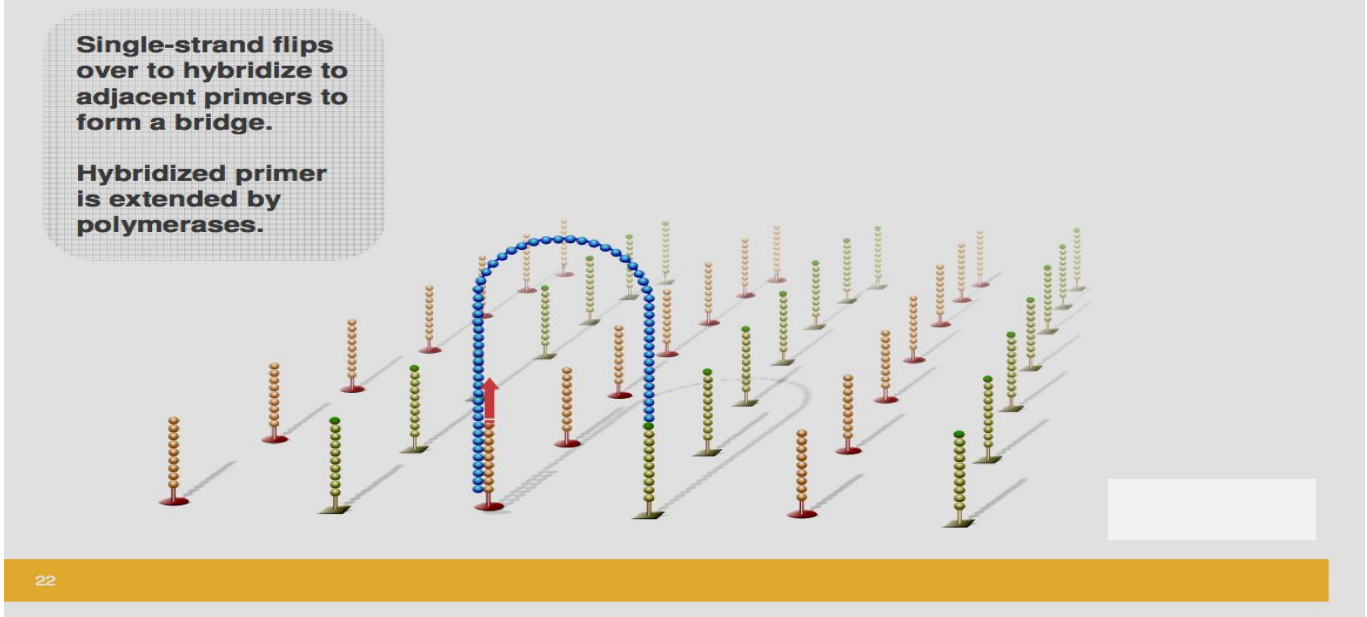
2 – Cluster generation



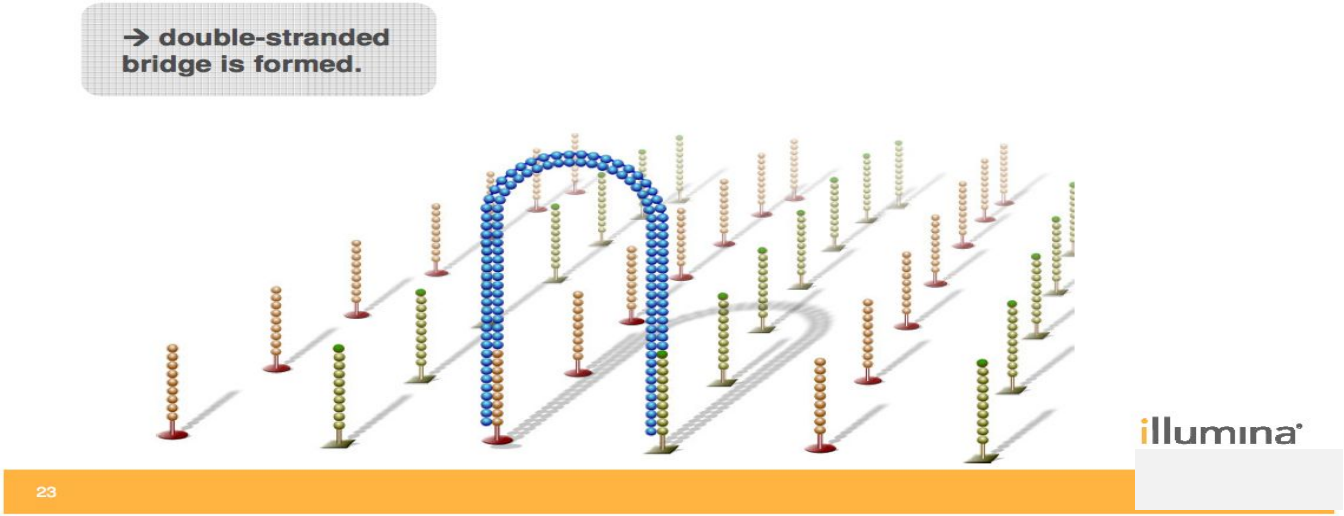
2 – Cluster generation



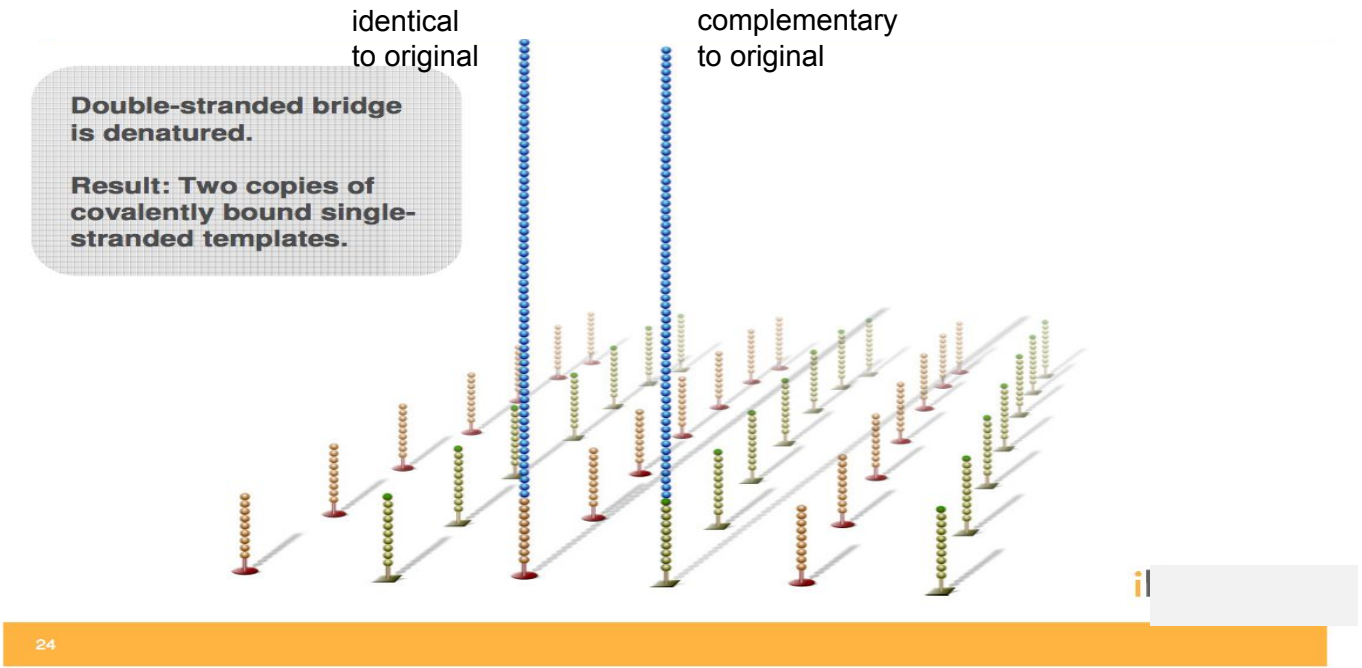
2 – Cluster generation



2 – Cluster generation



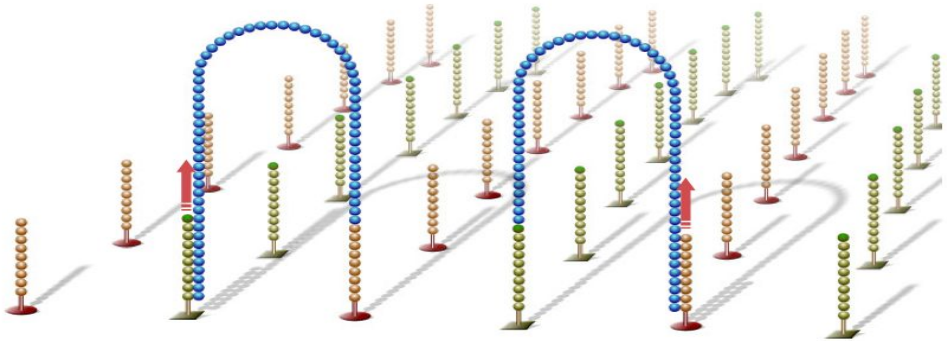
2 – Cluster generation



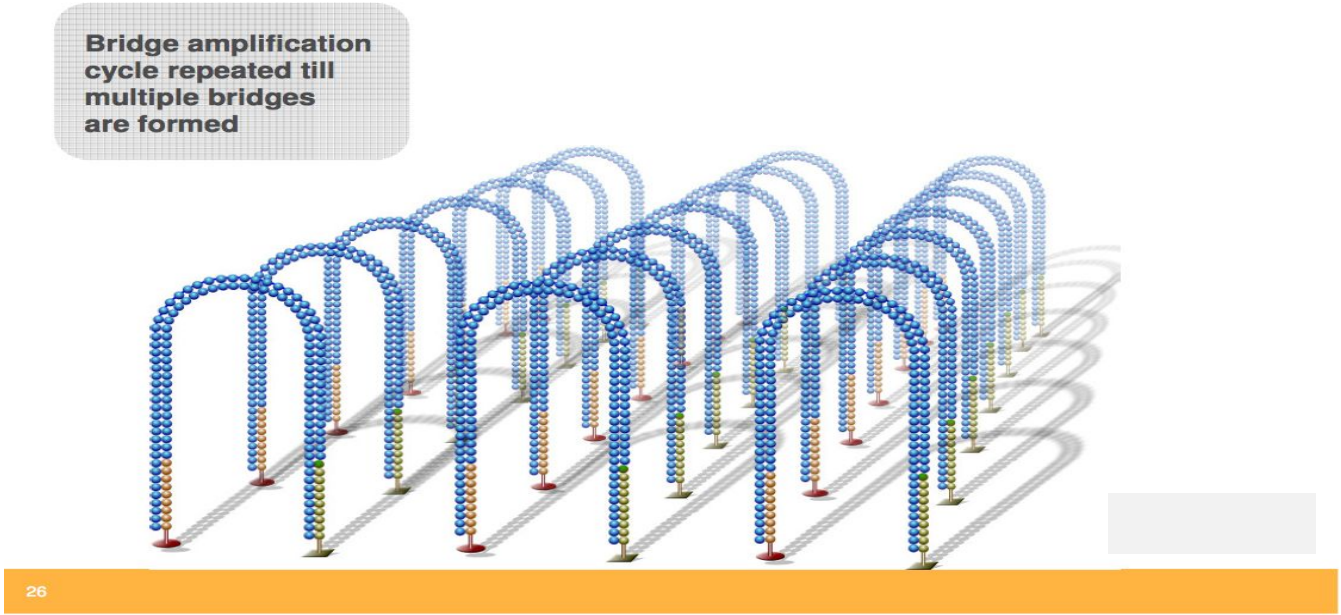
2 – Cluster generation

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

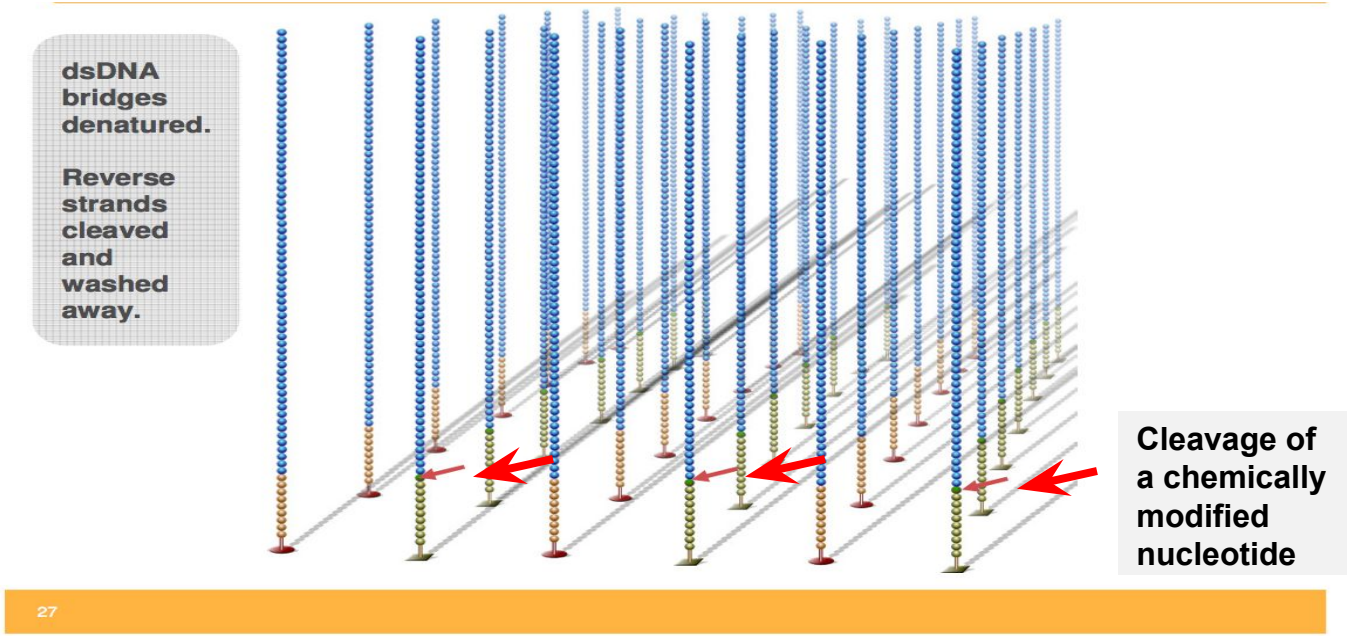
Hybridized primer is extended by polymerase.



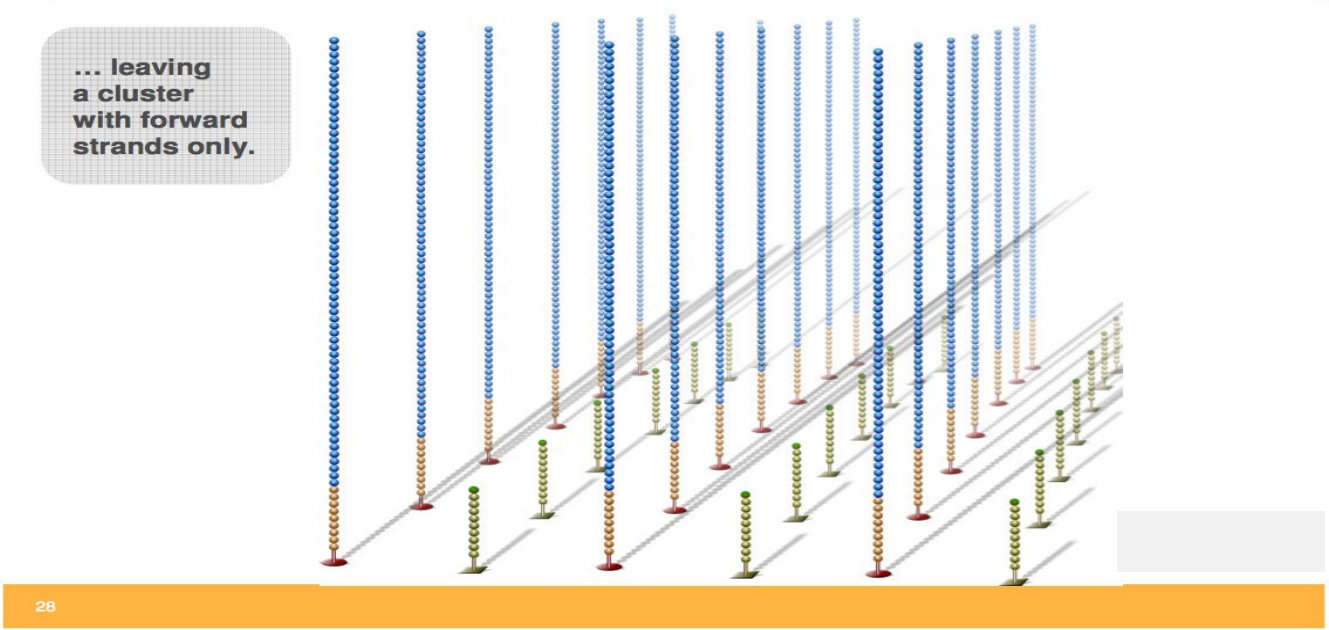
2 – Cluster generation



2 – Cluster generation



2 – Cluster generation



3 – Sequencing