

# Introduction NGS

Claude Thermes, Morgane Thomas-Chollier

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



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



**Platform**

2

Samples  
quality  
control

Libraries  
construction

Libraries  
quality  
control

Sequencing

Post-sequencing  
quality control

Bioinformatics  
analyses (optional)

**Experimental pole**

**Bioinformatics pole**

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

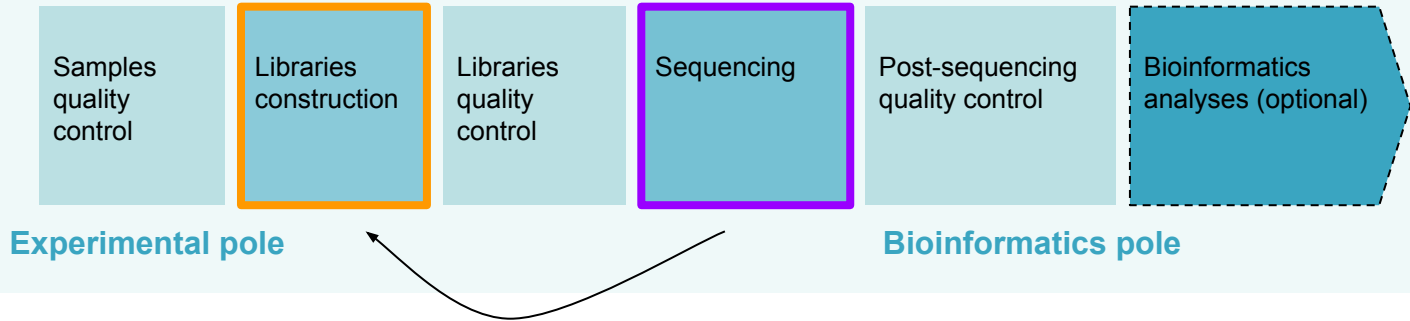


- 1 **Biologist brings the samples (DNA, RNA, cells)**



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

## 1st generation : Sanger sequencing

- Has been the major methodology up to 2005

### *Limitations*

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

## 2<sup>d</sup> generation

- Very high throughput
- Low cost

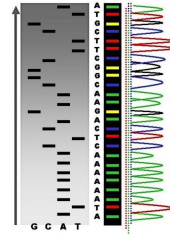
### *Limitations*

- Maximum read length  $\leq 300\text{bp}$

## 3<sup>rd</sup> generation

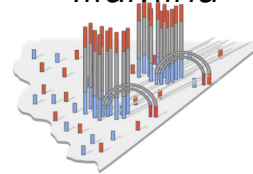
- Single molecules sequencing
- Very long reads

### *Sanger*



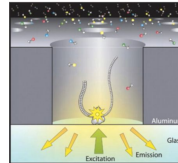
Sequencing

### *Illumina*

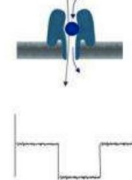


short reads

### *PacBio*



### *Oxford Nanopore*



long reads

# Different “generations” of sequencers

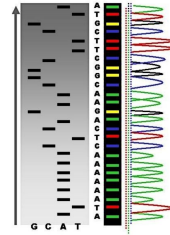
## 1st generation : Sanger sequencing

- Has been the major methodology up to 2005

### *Limitations*

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

## Sanger



Sequencing

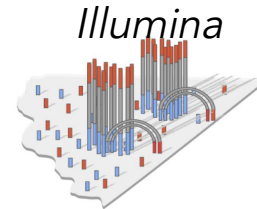
## 2<sup>d</sup> generation

- Very high throughput

- Low cost

### *Limitations*

- Maximum read length  $\leq$  300bp

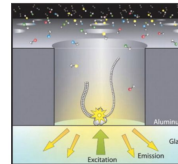


## 3<sup>rd</sup> generation

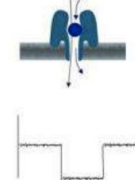
- Single molecules sequencing

- Very long reads

## PacBio



## Oxford Nanopore



# Illumina sequencing workflow

**1 - Library preparation**



The diagram illustrates the process of library preparation. It shows a DNA fragment being ligated with a sequencing adapter, resulting in a library of DNA fragments ready for sequencing.

Libraries construction

**2 - Cluster generation**



The diagram shows a flow cell with multiple lanes. DNA fragments are being amplified and clustered on a solid support, creating a dense array of clusters for sequencing.

Sequencing

**3 - Sequencing**



The diagram illustrates the sequencing by synthesis process. It shows three cycles of DNA synthesis where fluorescently labeled nucleotides are incorporated into the growing strand, and the signal is detected to determine the sequence.

**4 - Data analysis**

```
cagaaactgcagattagcgtgtatattatctggttatgct
cagaaactgcagattagcgtgtatattatctggttatgct
cagaaactgcagattatgtgtatattatctggttatgct
cagaaactgcagattttgtgtatattatctggttatgct
cagaaactgcggtgtatgtgtatattatctggttatgca
```

Bioinformatics analyses (optional)

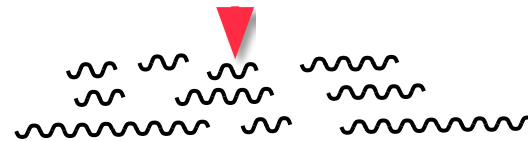
# 1 - Library preparation

---

Genomic DNA



Sonication



Size selection



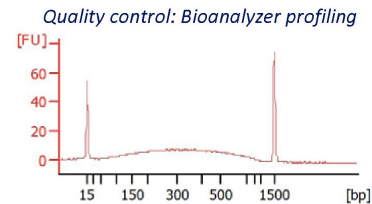
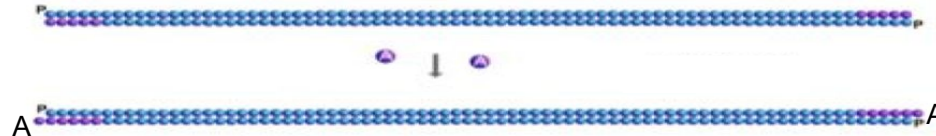
End repair



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

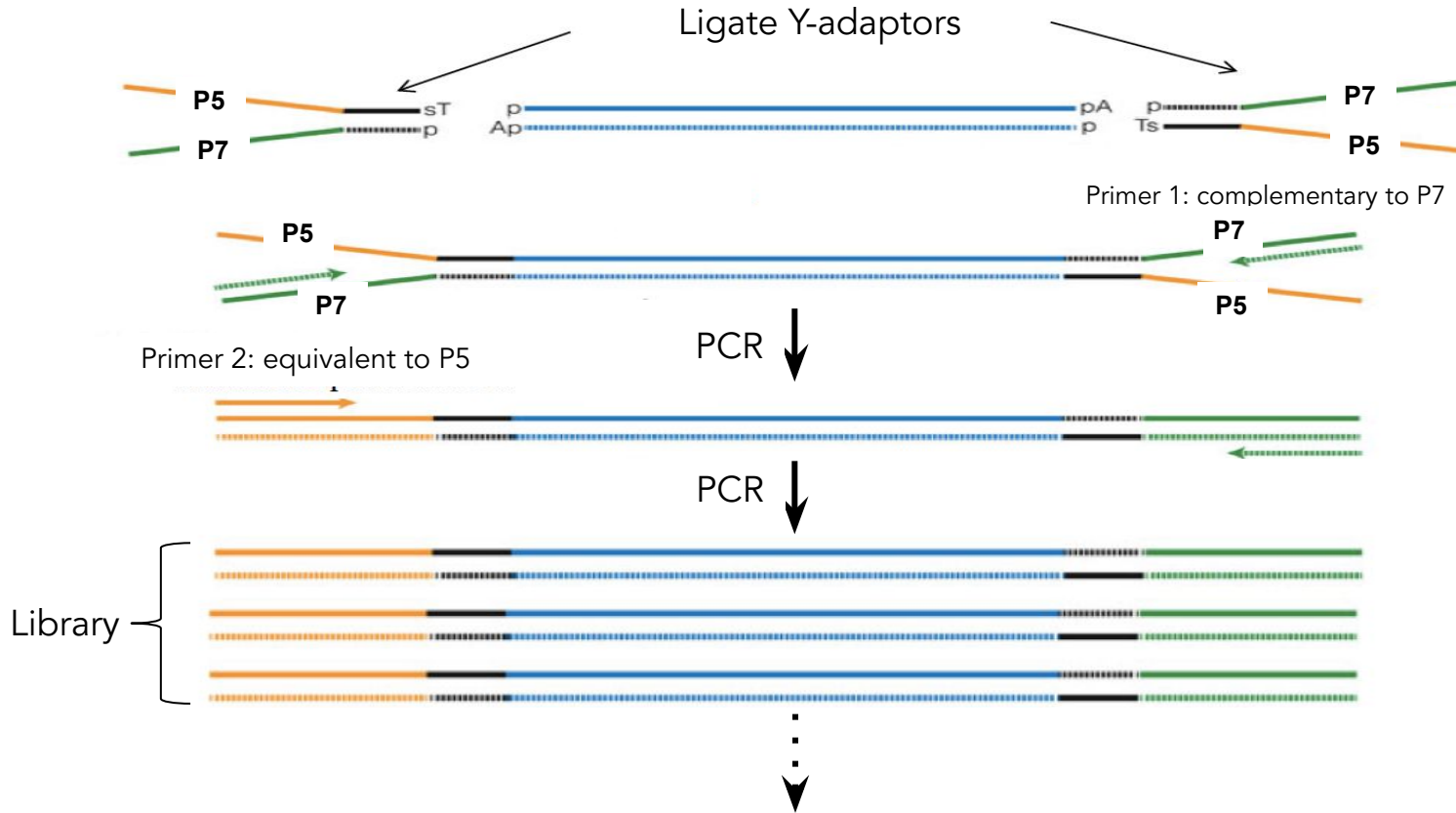
**Insert:** Target DNA or RNA fragment from a given sample library → this is the fragment we want to sequence

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



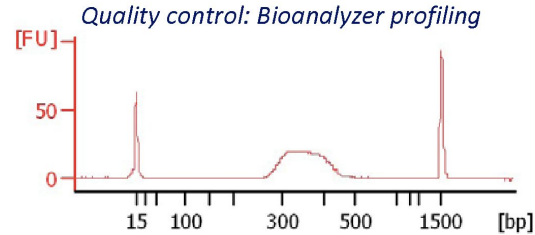
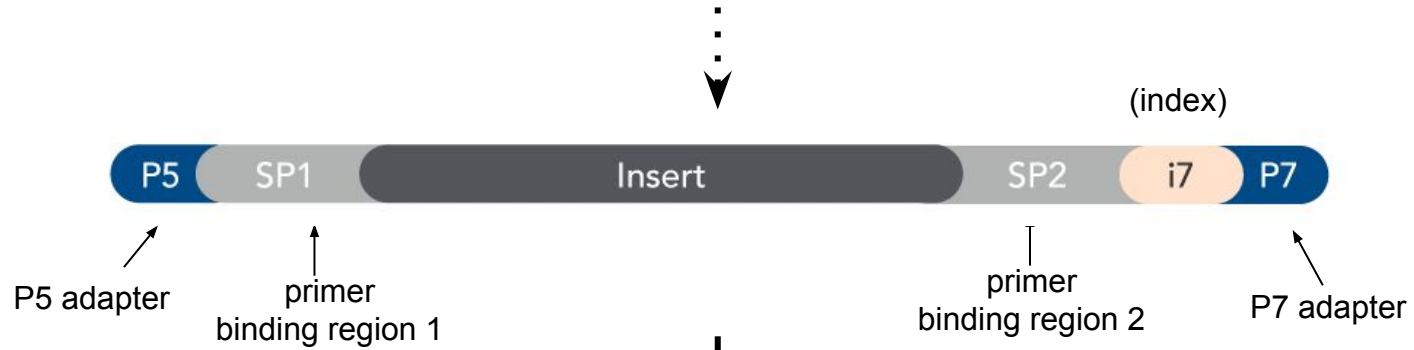
# 1 - Library preparation

How are the adapters attached to the DNA of interest ?



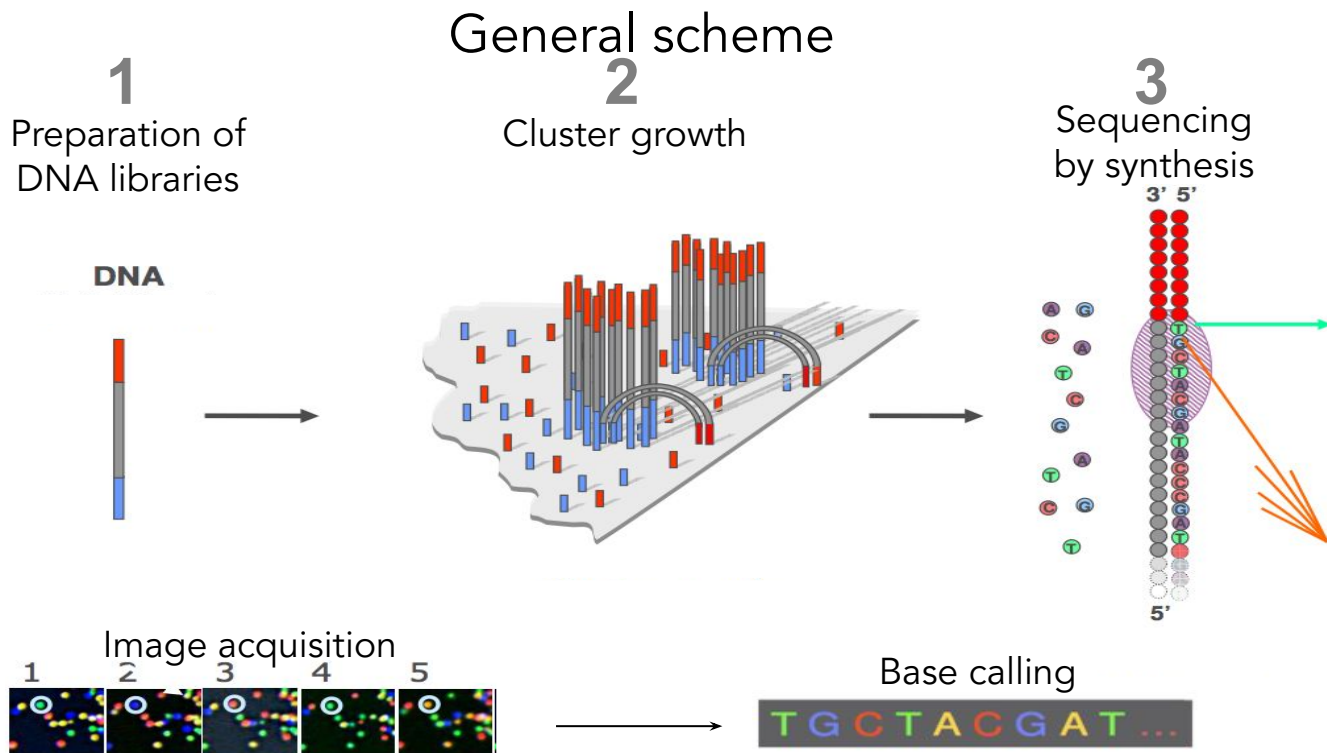
# 1 - Library preparation

---



cluster generation - sequencing

# Illumina sequencing



## 2 – Cluster generation

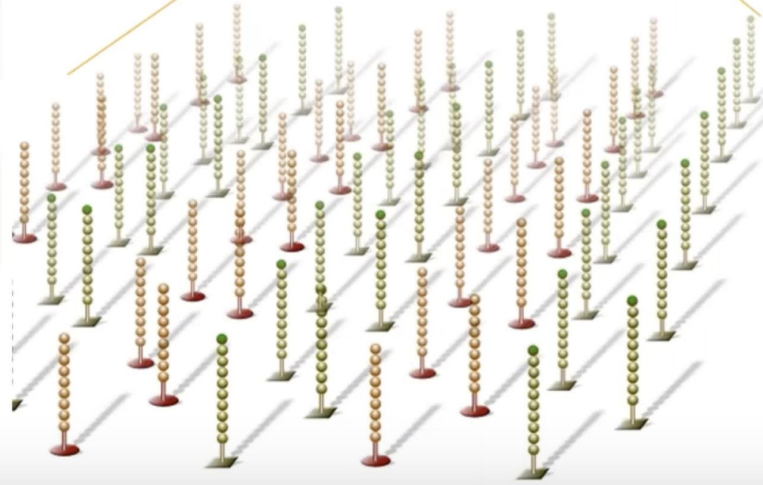
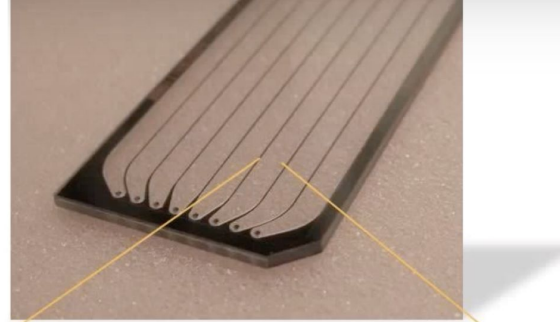
---

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



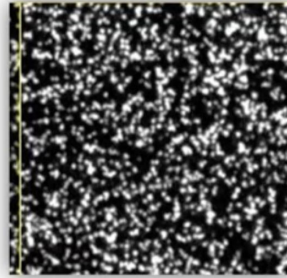
## 2 – Cluster generation

---

### What is a cluster ?

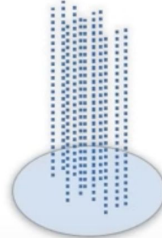
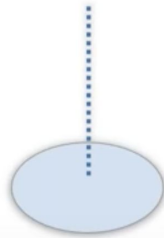
Clusters are a group of DNA strands positioned closely together

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

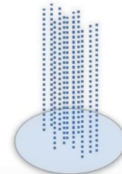
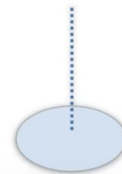
Single  
DNA  
Library



Amplified  
Clonal  
Cluster

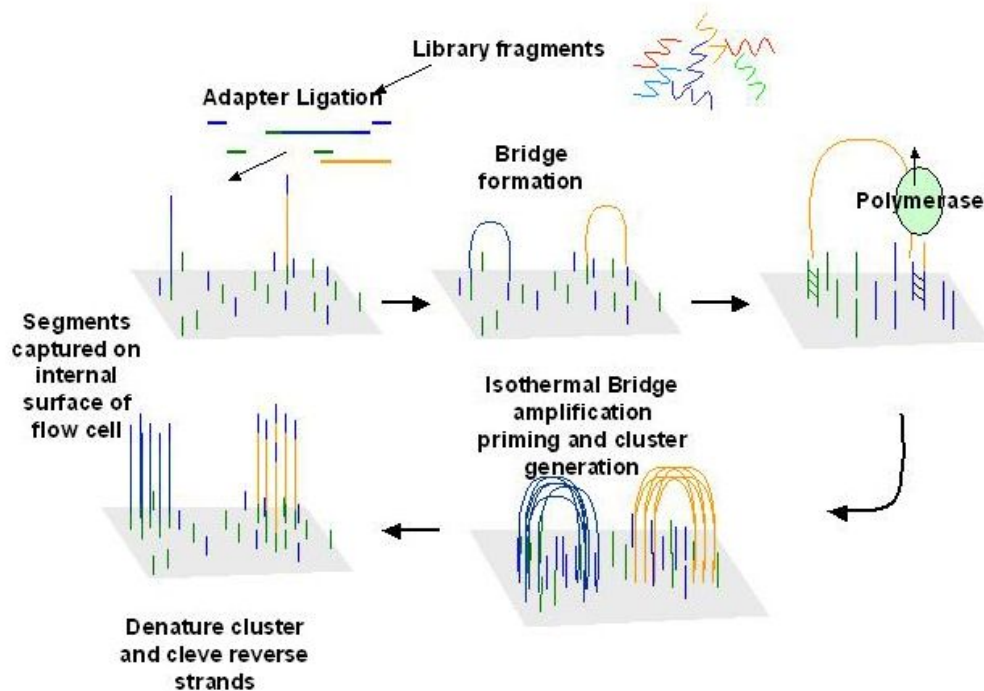
## 2 – Cluster generation

Single  
DNA  
Library

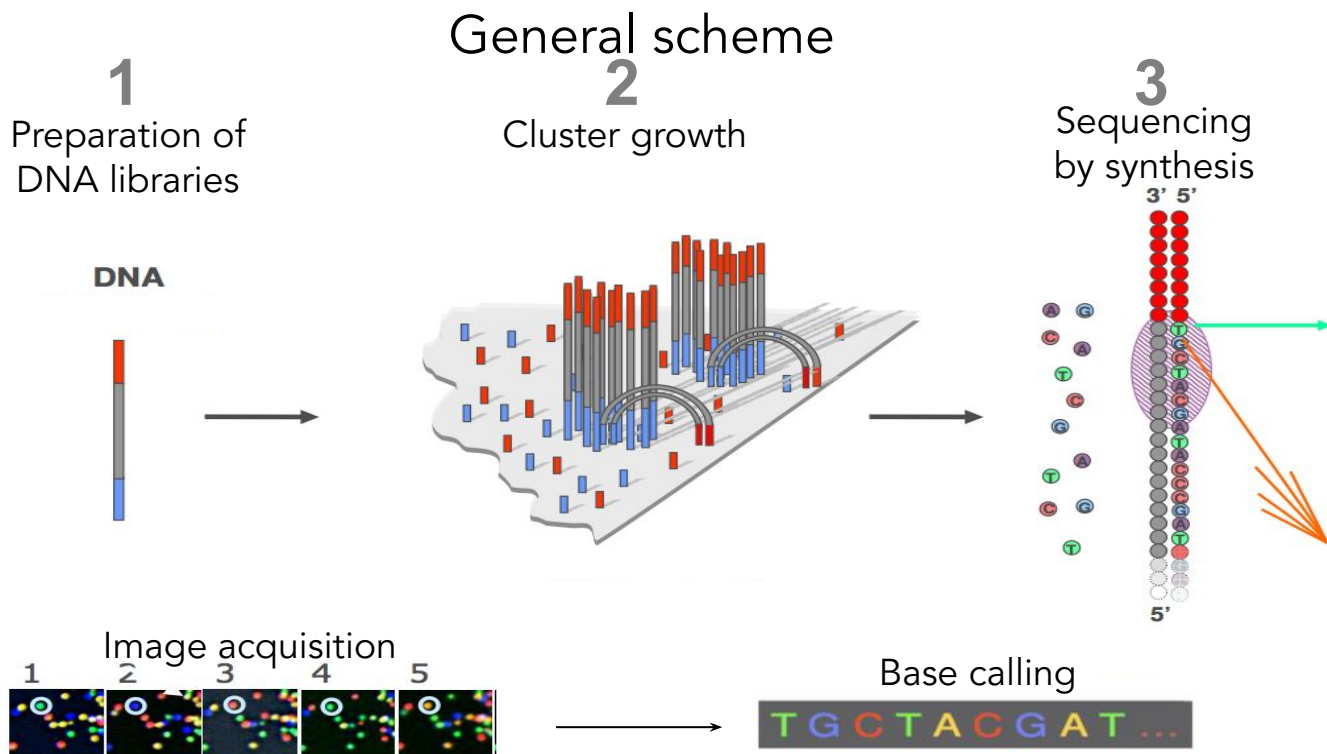


Amplified  
Clonal  
Cluster

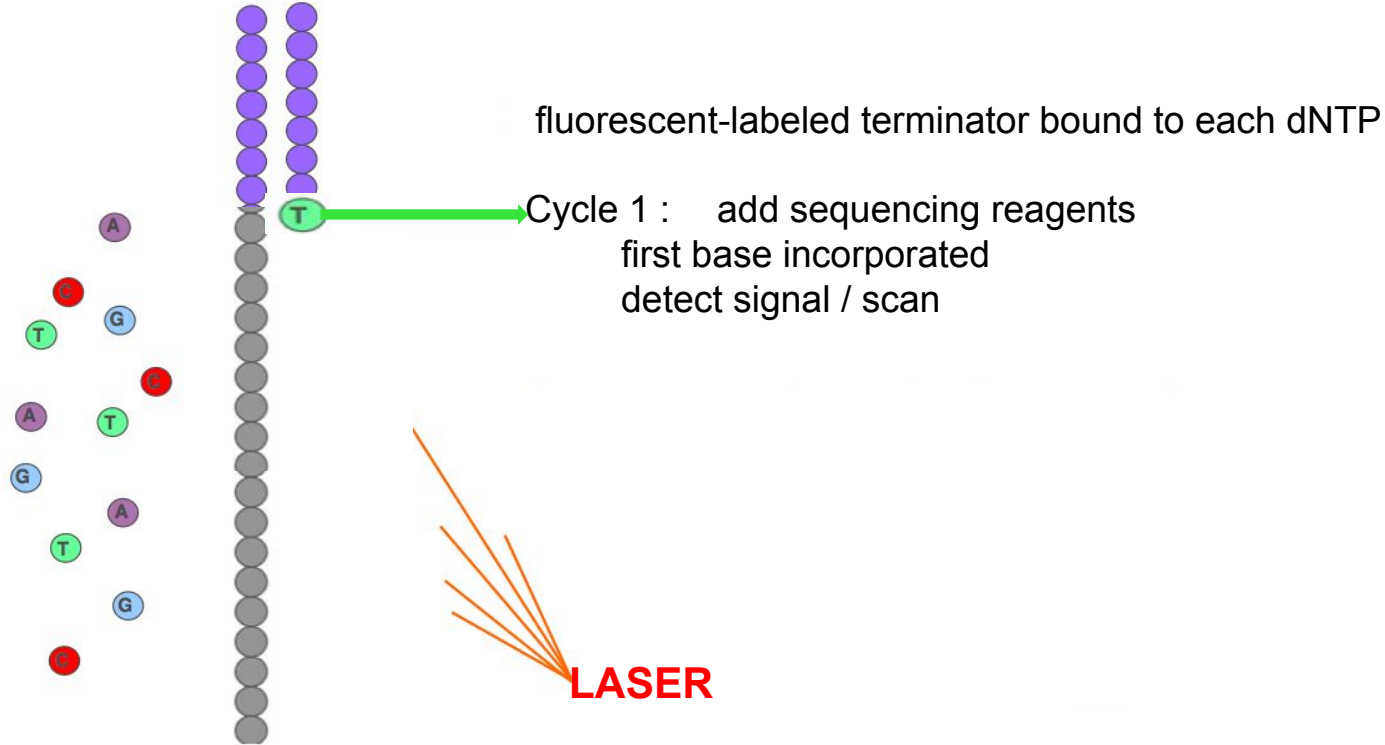
How are cluster generated ?



# Illumina sequencing

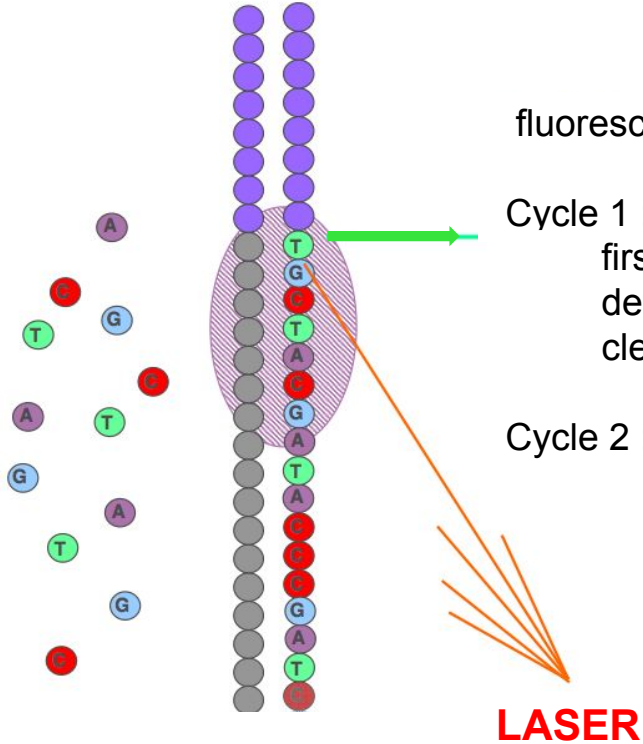


### 3 - Sequencing By Synthesis (SBS)





### 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

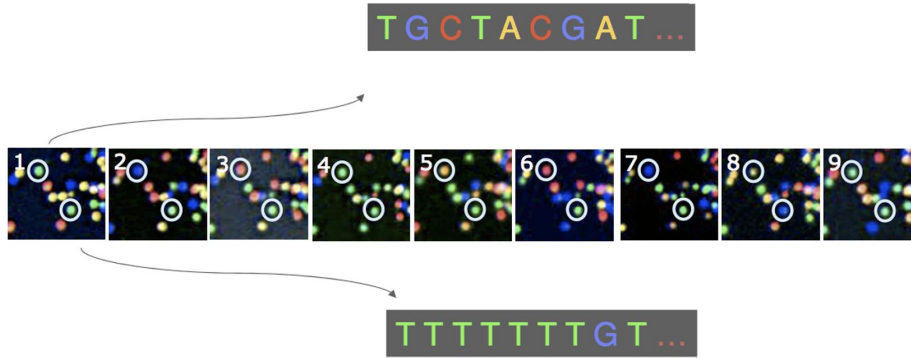
Cycle 2 : add sequencing reagents and repeat

4 terminator-bound dNTPs present during each cycle

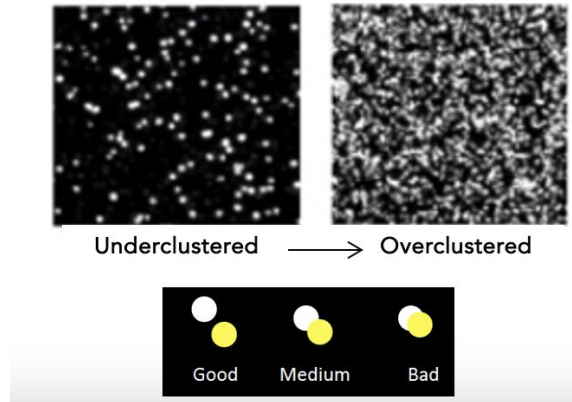
natural competition minimizes incorporation bias

# 3 – Sequencing

---



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



## 3 – Sequencing

---

P5

SP1

Insert

SP2

i7

P7

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 ?

**ATTTCGCATTACGCTTTTA**

**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 ?

**all reads** = one file

**ATTTCGCATTTACGCTTTTA**

**Read** = one sequence



ATTTCGCATTTACGCTTTTA  
CCTCGCATTTACGCTCCTAT  
CGCATTTACGCTCCTATCTC



### 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  $< 800\text{nt}$ )



### 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 ?



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





### 3 – Sequencing

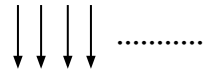
---

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

**ATTTTCGCATTTACGCTTTTA**



**I ? I D D D D D D H H H ? G H : ? F C @**



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)



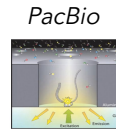
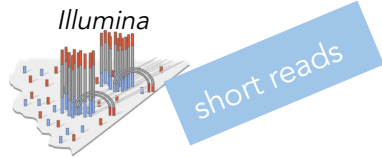
- read (lecture)



- single-end versus paired-end sequencing



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



- adapter in reads ?



- read quality

ATTTCGCATTACGCTTTTA  
I?IDDDDDH?GH:?FC@

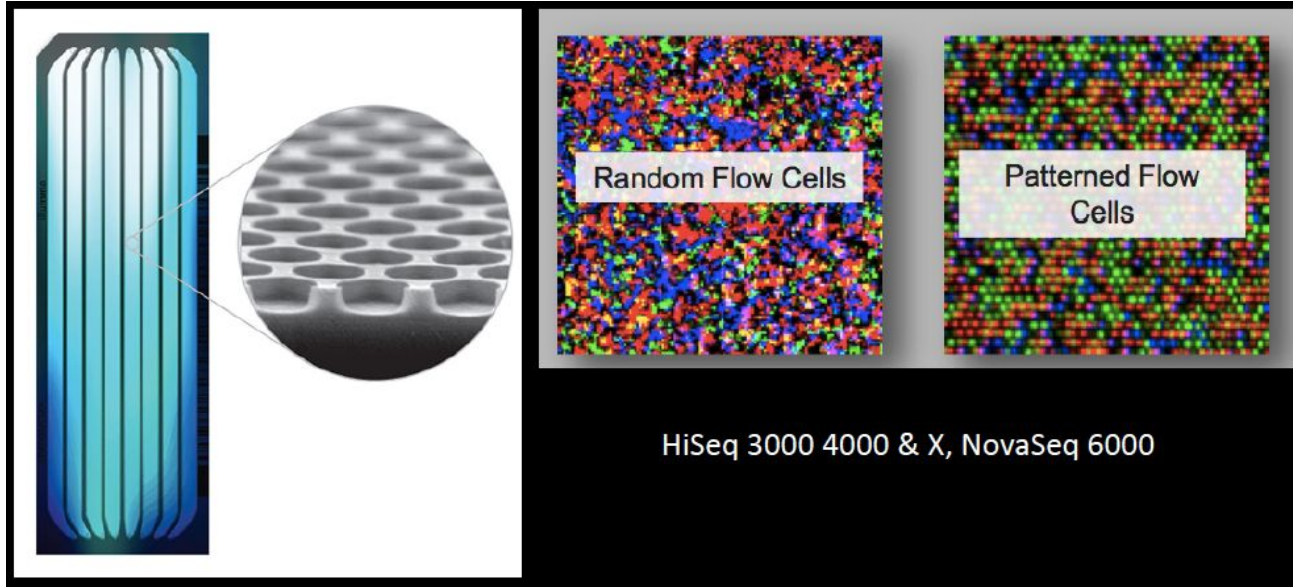
Supplementary

### 3 – Sequencing

---

#### Patterned flow cells

- Improves regularity of densities and qualities
- Reduces analysis time



## 3 – Sequencing

---

“Dephasing” due to partial blockage of DNA synthesis

Cycle 1 reading with strong signals



Cluster generation

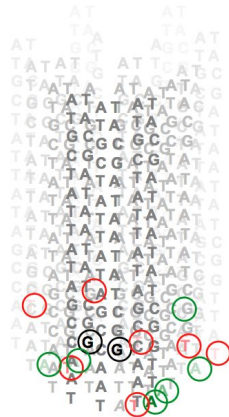
Cycle 1 read as: **T** **T**

### 3 – Sequencing

---

“Dephasing” due to partial blockage of DNA synthesis

Later Cycles with More Errors

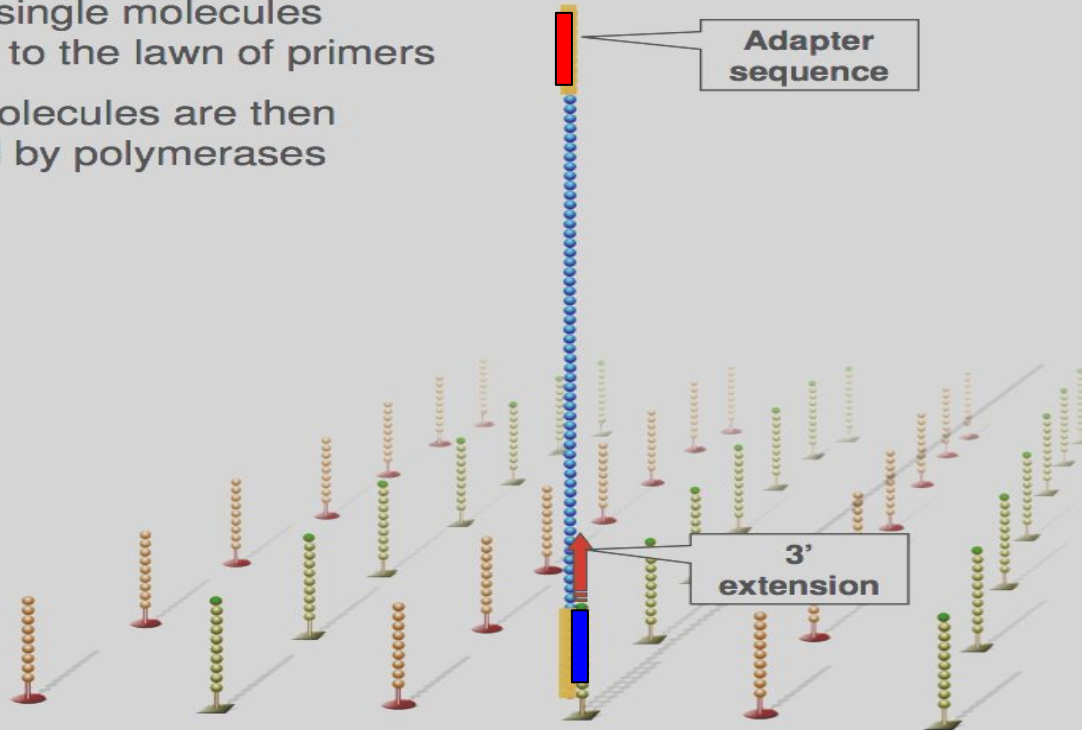


Cluster generation

Cycle 1	read as:	T	T
Cycle 2	read as:	A	A
Cycle 3	read as:	C	C
Cycle 4	read as:	G	G
Cycle 5	read as:	A	A
Cycle 6	read as:	T	T
Cycle 7	read as:	A	A
Cycle 8	read as:	A	A
Cycle 9	read as:	T	T
Cycle 10	read as:	A	A
Cycle 11	read as:	T	?
Cycle 12	read as:	C	?
Cycle 13	read as:	G	?
Cycle 14	read as:	G	?
Cycle 15	read as:	T	?
Cycle 16	read as:	T	?

## 2 – Cluster generation

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

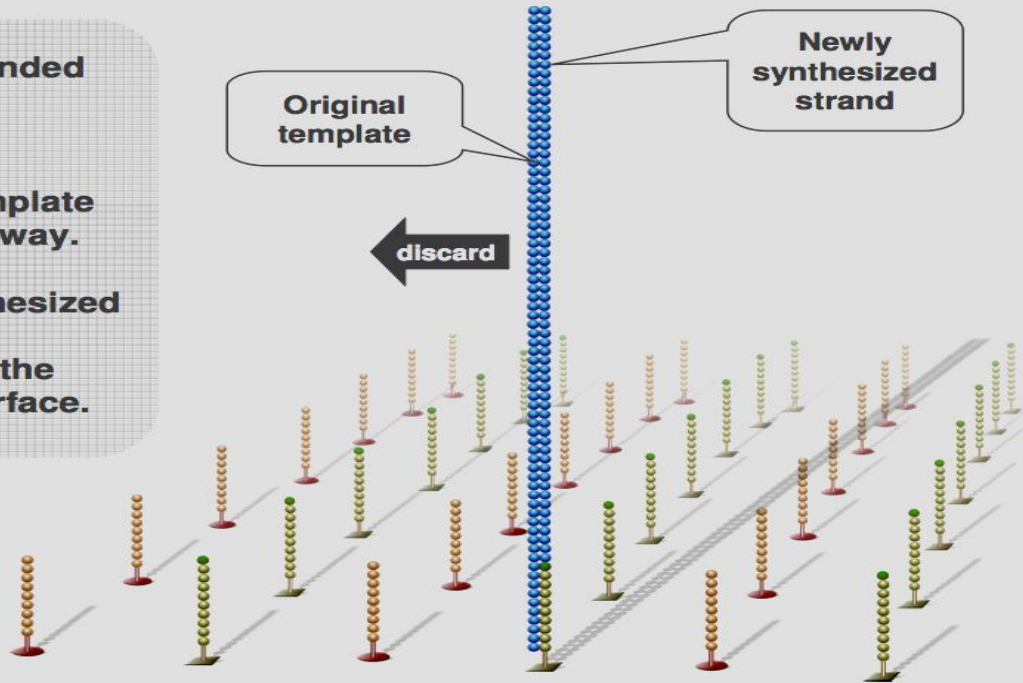


## 2 – Cluster generation

**Double-stranded molecule is denatured.**

**Original template is washed away.**

**Newly synthesized covalently attached to the flow cell surface.**

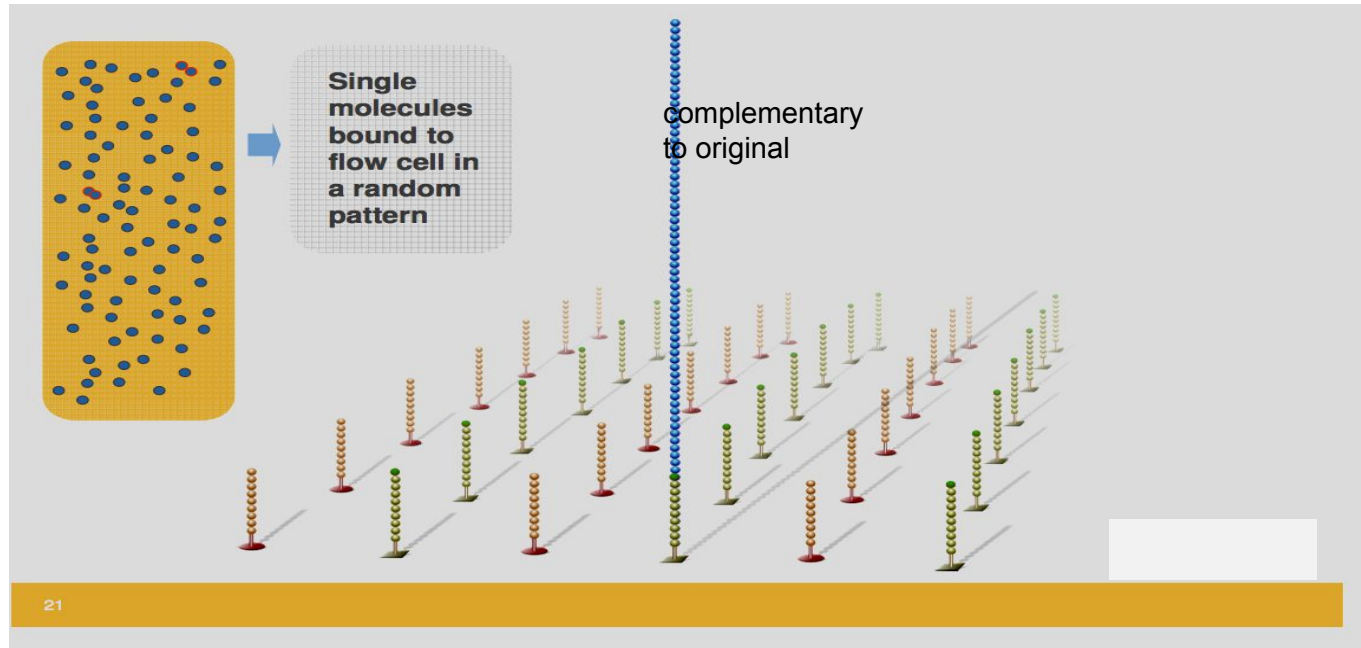


illumina



## 2 – Cluster generation

---

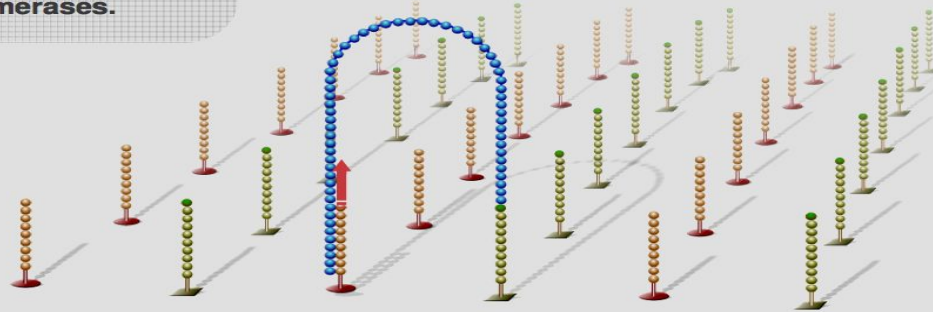


## 2 – Cluster generation

---

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

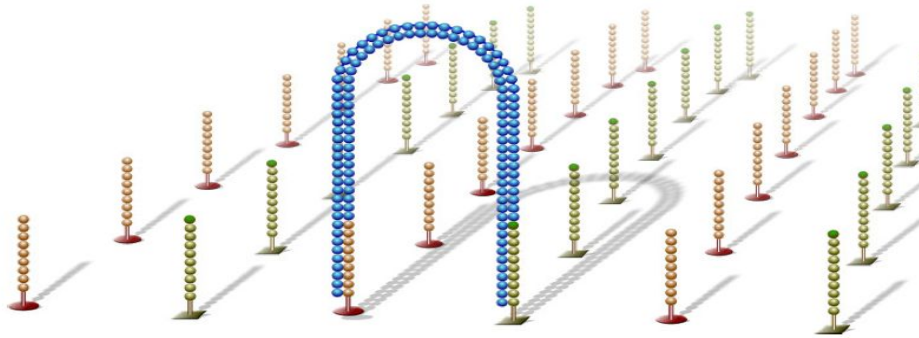
**Hybridized primer is extended by polymerases.**



## 2 – Cluster generation

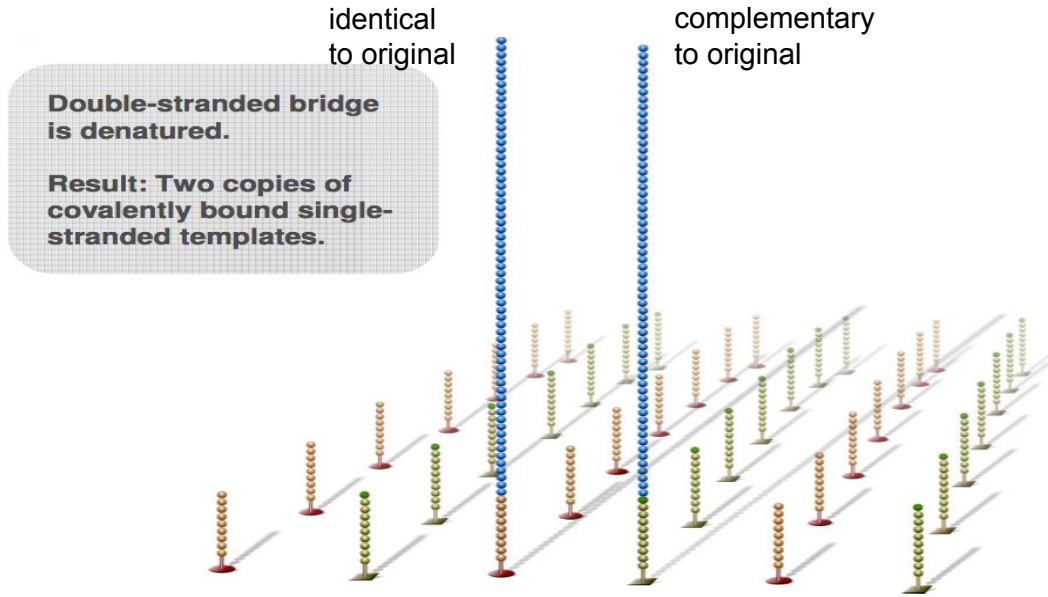
---

→ double-stranded bridge is formed.



illumina

## 2 – Cluster generation

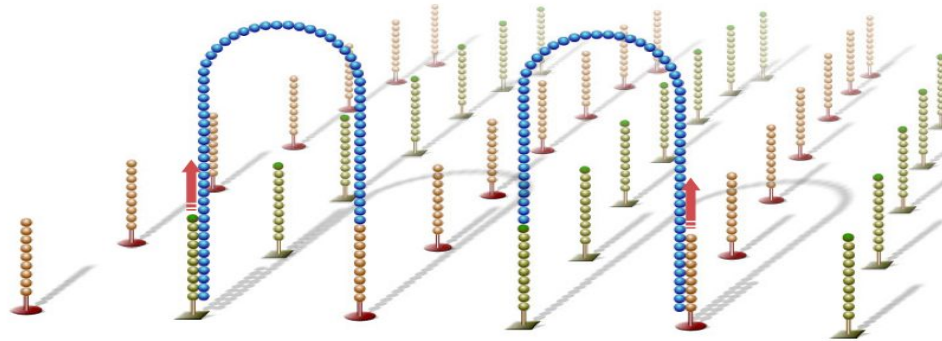


## 2 – Cluster generation

---

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

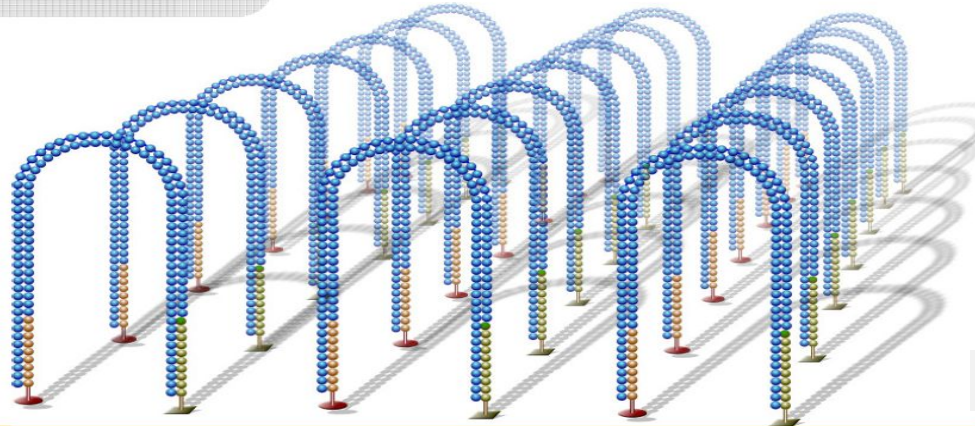
**Hybridized primer is extended by polymerase.**



## 2 – Cluster generation

---

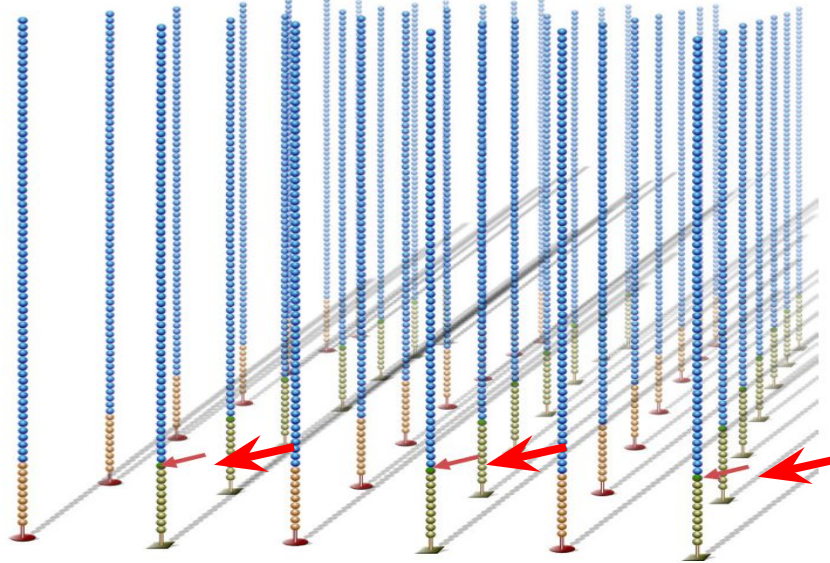
**Bridge amplification  
cycle repeated till  
multiple bridges  
are formed**



## 2 – Cluster generation

dsDNA  
bridges  
denatured.

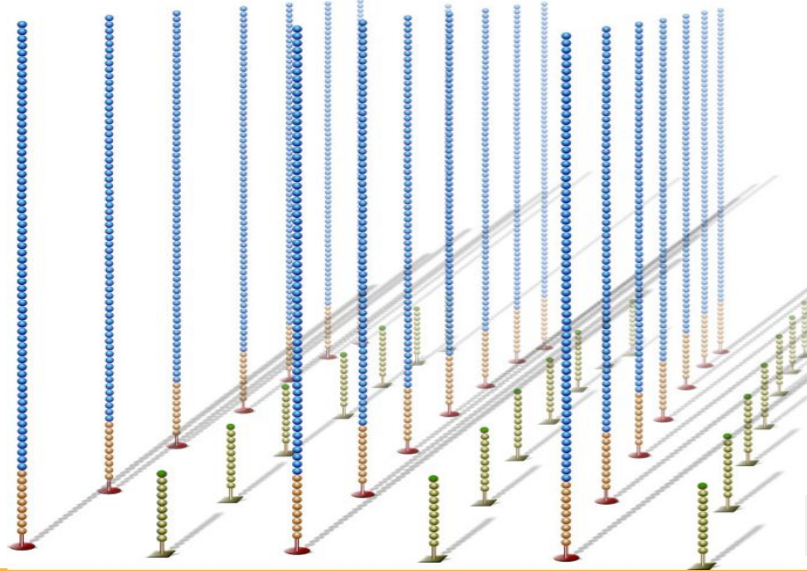
Reverse  
strands  
cleaved  
and  
washed  
away.



Cleavage of  
a chemically  
modified  
nucleotide

## 2 – Cluster generation

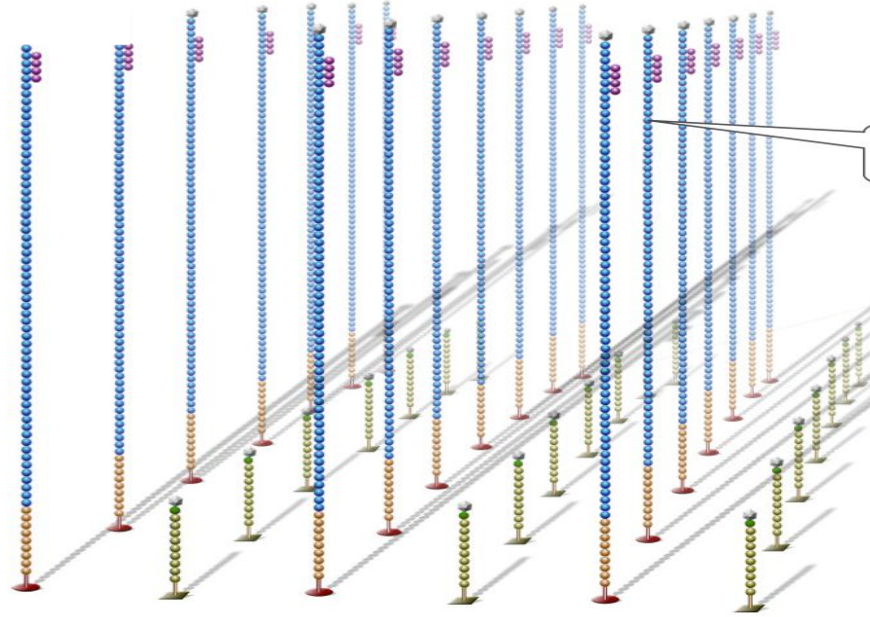
... leaving  
a cluster  
with forward  
strands only.





# 3 – Sequencing

Sequencing primer is hybridized to adapter sequence.



Sequencing primer