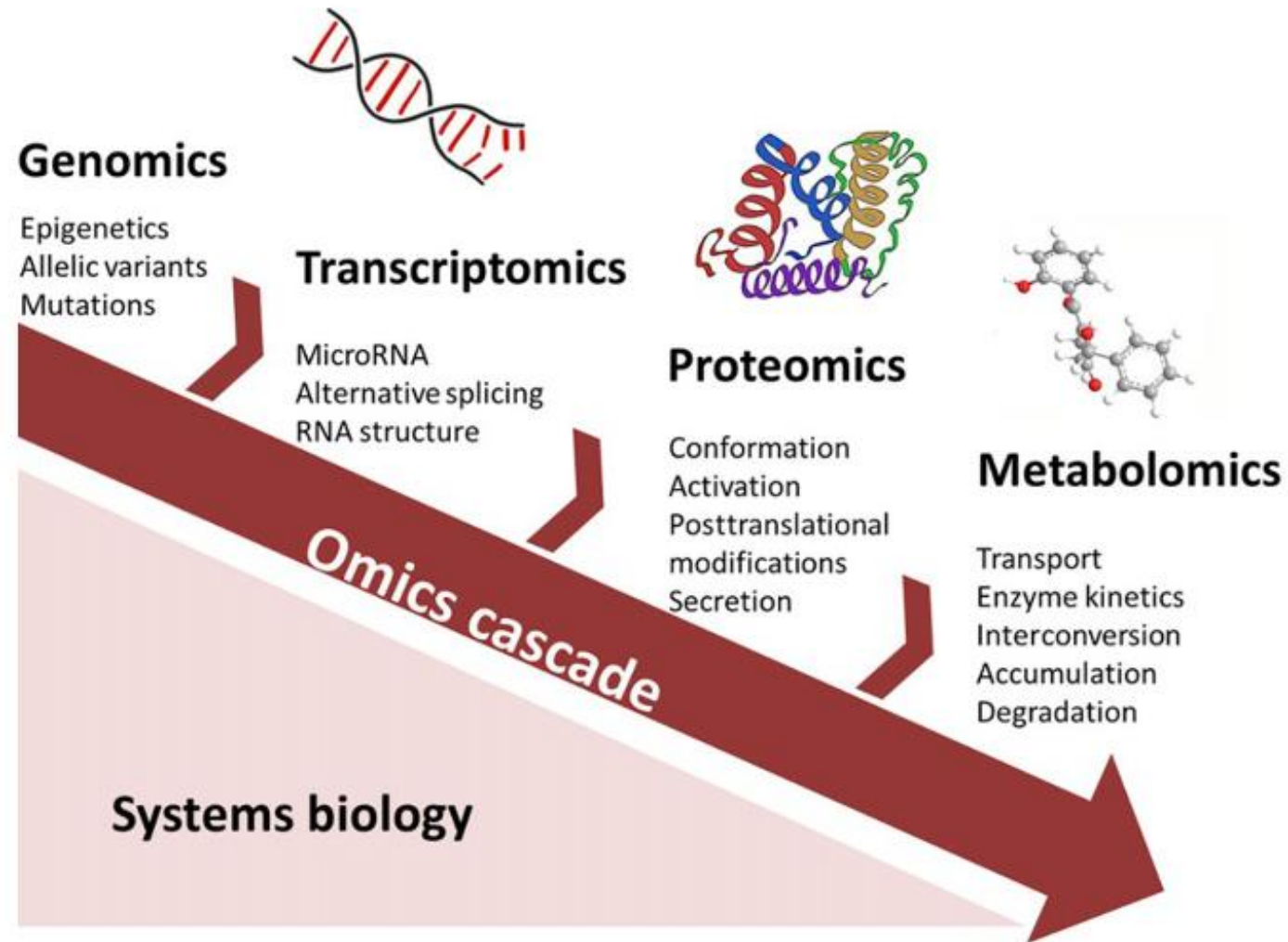


Partie 6

Perspectives et outils d'avenirs



Omics sciences

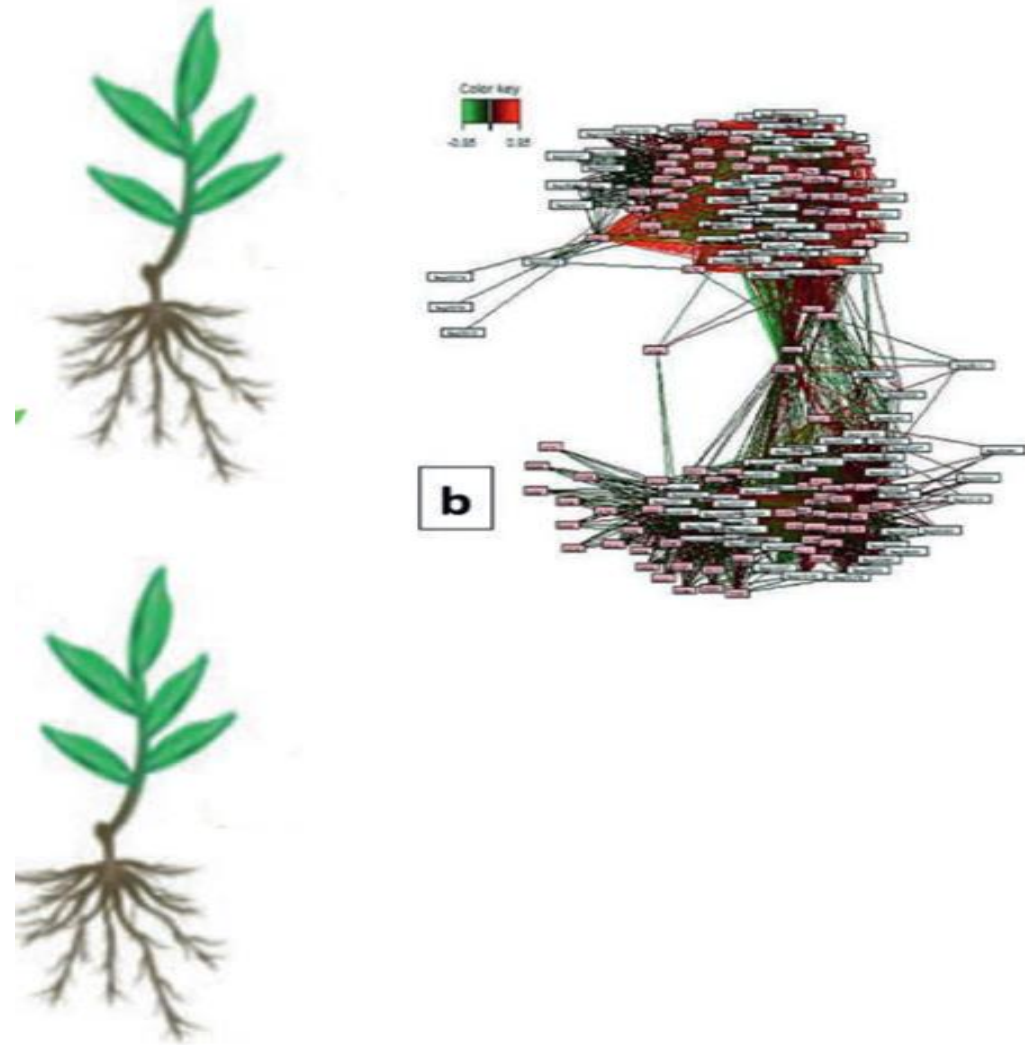
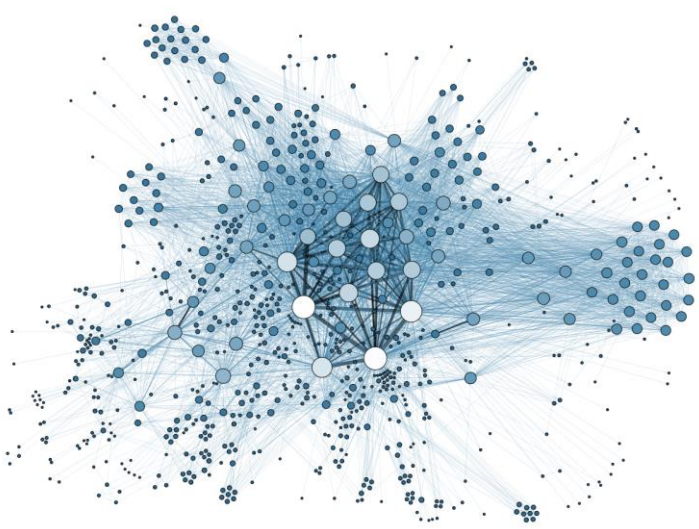


Carmen Bedia,
Chapter Two - Experimental Approaches in Omic Science
Editor(s): Joaquim Jaumot, Carmen Bedia, Romà Tauler,
Comprehensive Analytical Chemistry,
<https://doi.org/10.1016/bs.coac.2018.07.002>.

ome = totalité

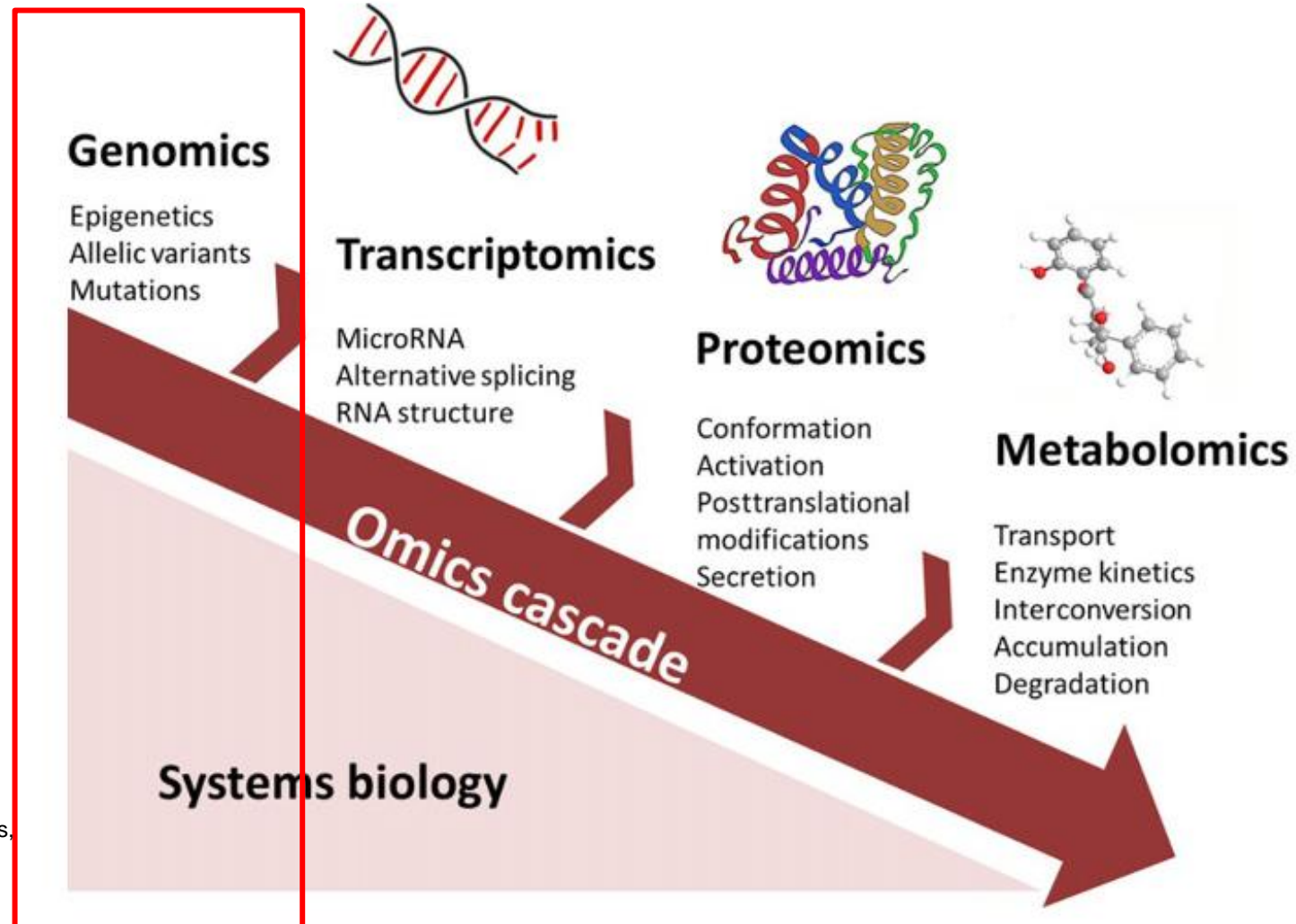
Un système complexe

Un système complexe est constitué de **nombreuses entités** dont les interactions produisent un **comportement global** qui ne peut être facilement expliqué à partir des seules propriétés individuelles des constituants.



B. Intégration des données transcriptomiques (blanc) et métabolomiques (rose) chez l'eucalyptus (Favreau, in prep).

Omics sciences



Génomes séquencés

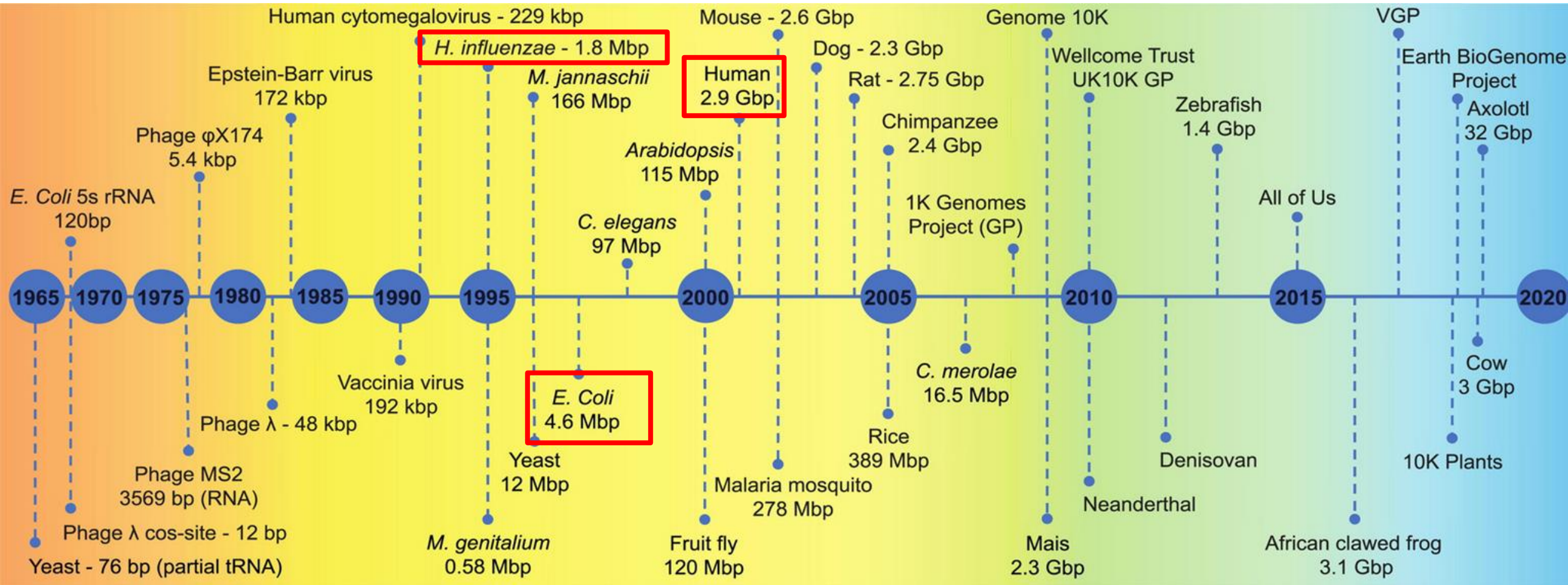
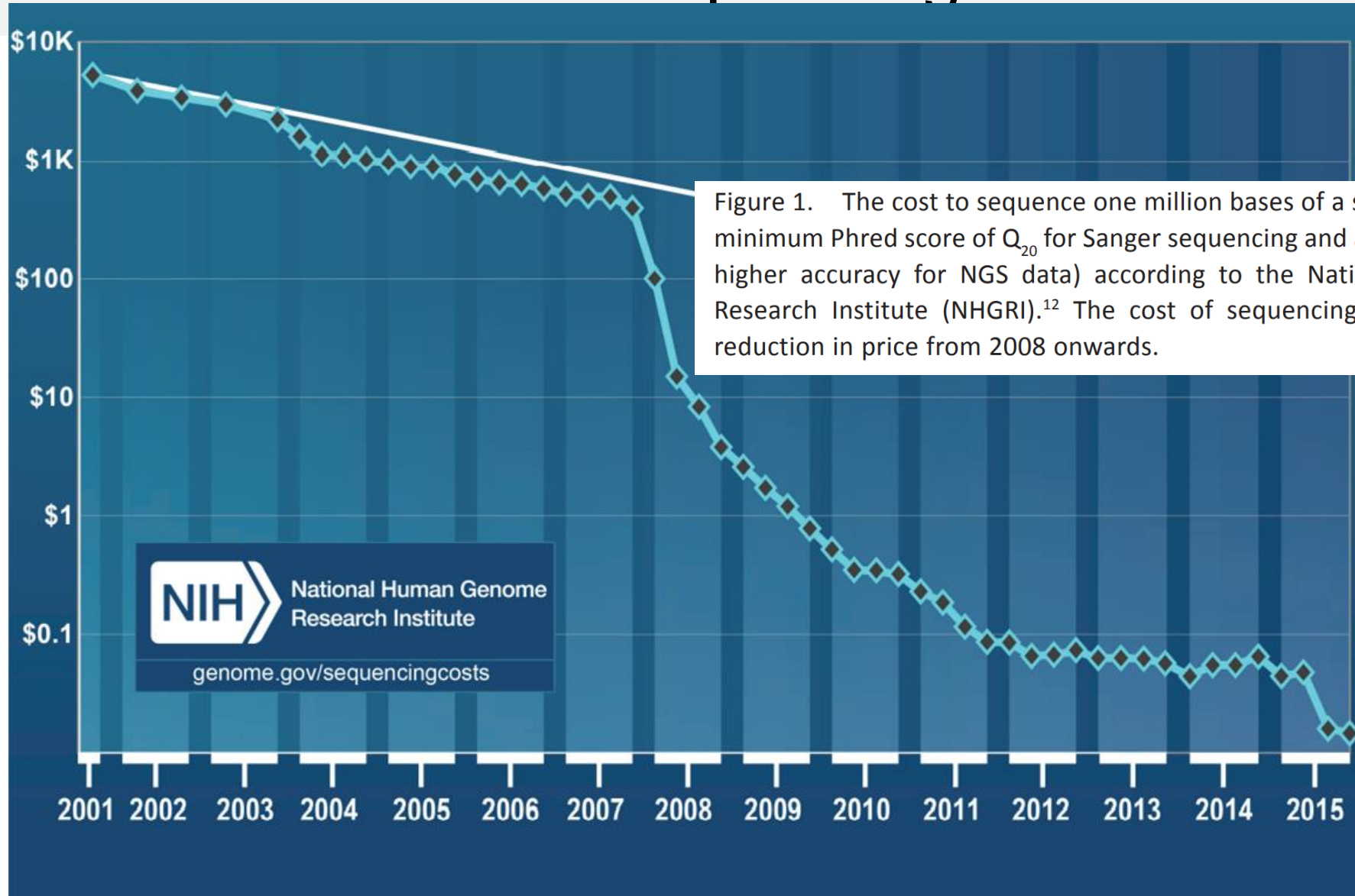


Fig. 1. Milestones in genome assembly. Timeline illustrating many of the major genome assembly achievements ranging from the beginning of the sequencing era to the large-scale genome projects currently ongoing. Each genome or genome project (GP) is placed under a color-coded background according to the sequencing approach adopted. Light red: early sequencing methods, Yellow: Sanger-based shotgun sequencing, Green: NGS, Light blue: TGS.

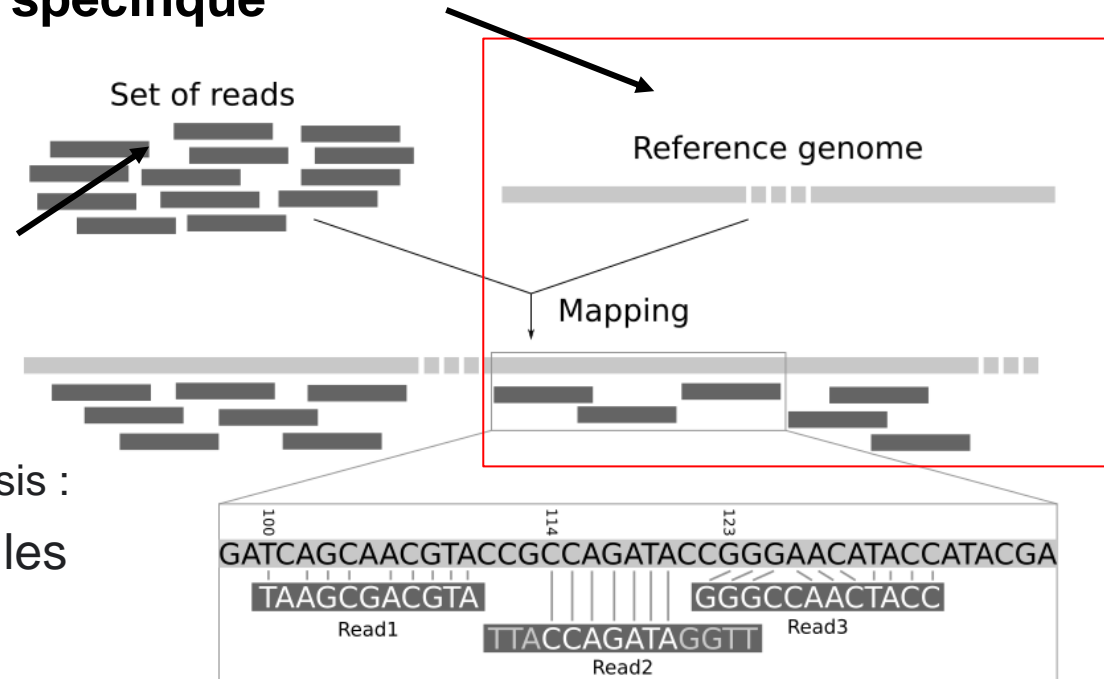
Coût du séquençage



Quelques définitions....

- Il faut un **génom**e de référence pour comparer nos séquences (reads) :
 - **base de données numérique de séquences d'acides nucléiques**, assemblée par des scientifiques comme exemple **représentatif de l'ensemble des gènes d'un organisme individuel idéalisé d'une espèce**
 - **ne représente pas** précisément l'ensemble des gènes d'un seul organisme individuel
 - **chaque population (ensemble d'individus partageant un ensemble de caractères communs)** isolée est caractérisée par un **génom**e spécifique

séquence déduite de paires de bases
correspondant à tout ou partie d'un seul
fragment d'ADN



Exemple, génome de référence de Mycobacterium tuberculosis :
H37Rv (souche de tuberculose la plus étudiée dans les
laboratoires de recherche)

Quelques définitions....

- **Variant** : Variation dans une séquence nucléotidique en comparaison avec une séquence de référence
 - **SNV** : **S**ingle **N**ucleotide **V**ariant
 - **INDEL** : **IN**sertion ou **DE**letion d'une ou plusieurs bases
 - **MNV** (**M**ulti-**N**ucleotide **V**ariant) : plusieurs SNVs et/ou INDELS dans un bloc



AACGGCCGTGAAC
AACGGCCAGTAAC

SNV



AACGGCCGTGAAC
AACGGCC-GTAAC

DEletion

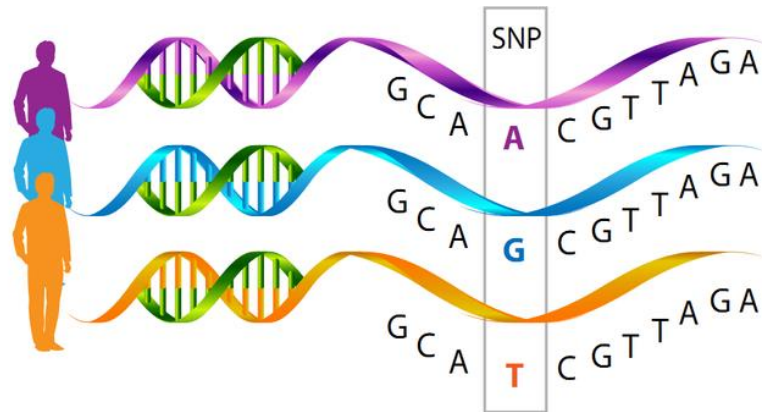


AACGGCCGTGAAC
AACGGCCAGCTGAAC

INsertion

Quelques définitions....

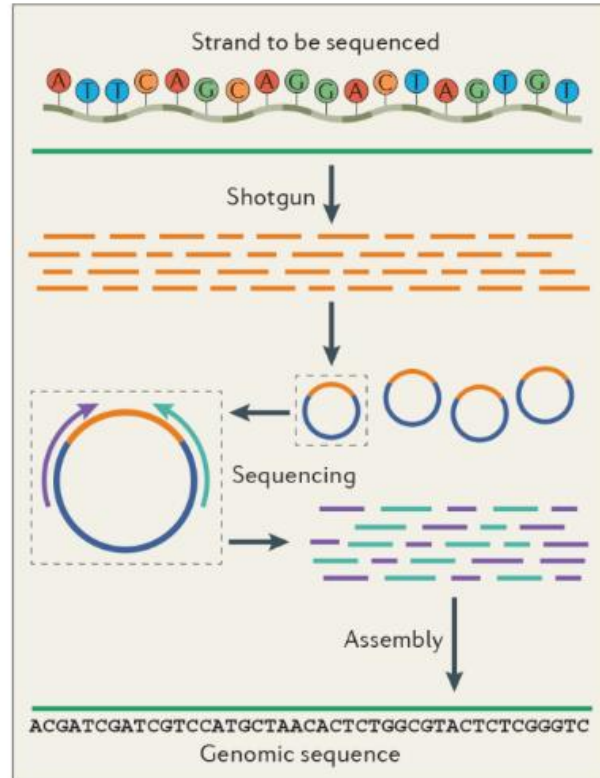
- **Mutation** : variation dans la séquence d'ADN (SNV, indel)
 - fréquence $< 1\%$ dans une population
 - effet : gain de fonction ou perte de fonction
 - phénomène rare
- **SNP (Single Nucleotide Polymorphism)** :
variant partagé dans la population ($> 1\%$)
Les SNP sont en fait des SNV dont la fréquence dans la population est élevée



Les différentes technologies

NGS

The First Revolution Whole-genome shotgun

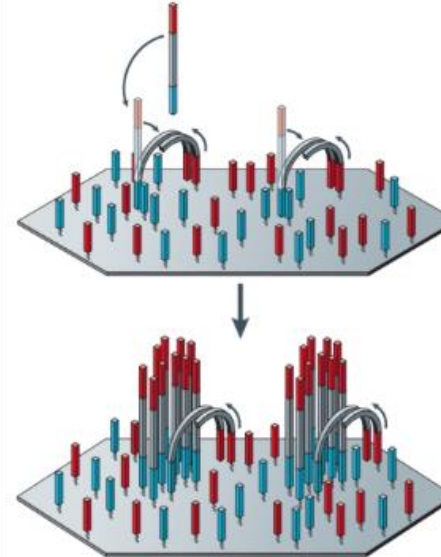


Sanger shotgun sequencing

- Sequencing by synthesis
- Amplified templates generated *in vitro*
- Requires onerous colony picking and plasmid preparation

For example, ABI capillary sequencer (ABI)

The Second Revolution High-throughput sequencing



454 sequencing

- Sequencing by synthesis
- Amplified templates generated *in vitro*
- High accuracy outside homopolymers but short read lengths

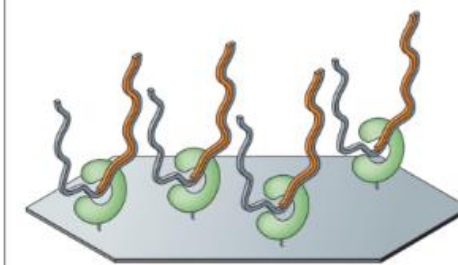
For example, 454 GS FLX+ (Roche)

Illumina sequencing

- Sequencing by synthesis
- Amplified templates generated *in vitro*
- High accuracy but short read lengths

For example, MiSeq (Illumina)

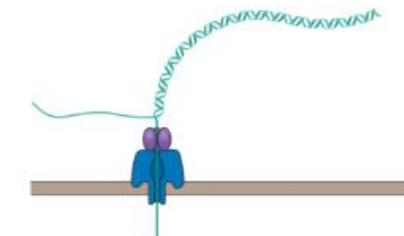
The Third Revolution Single-molecule sequencing



Pac Bio SMRT sequencing

- Sequencing by synthesis
- Single-molecule templates
- Low accuracy but long read lengths

For example, PacBio RS (Pacific Biosciences)

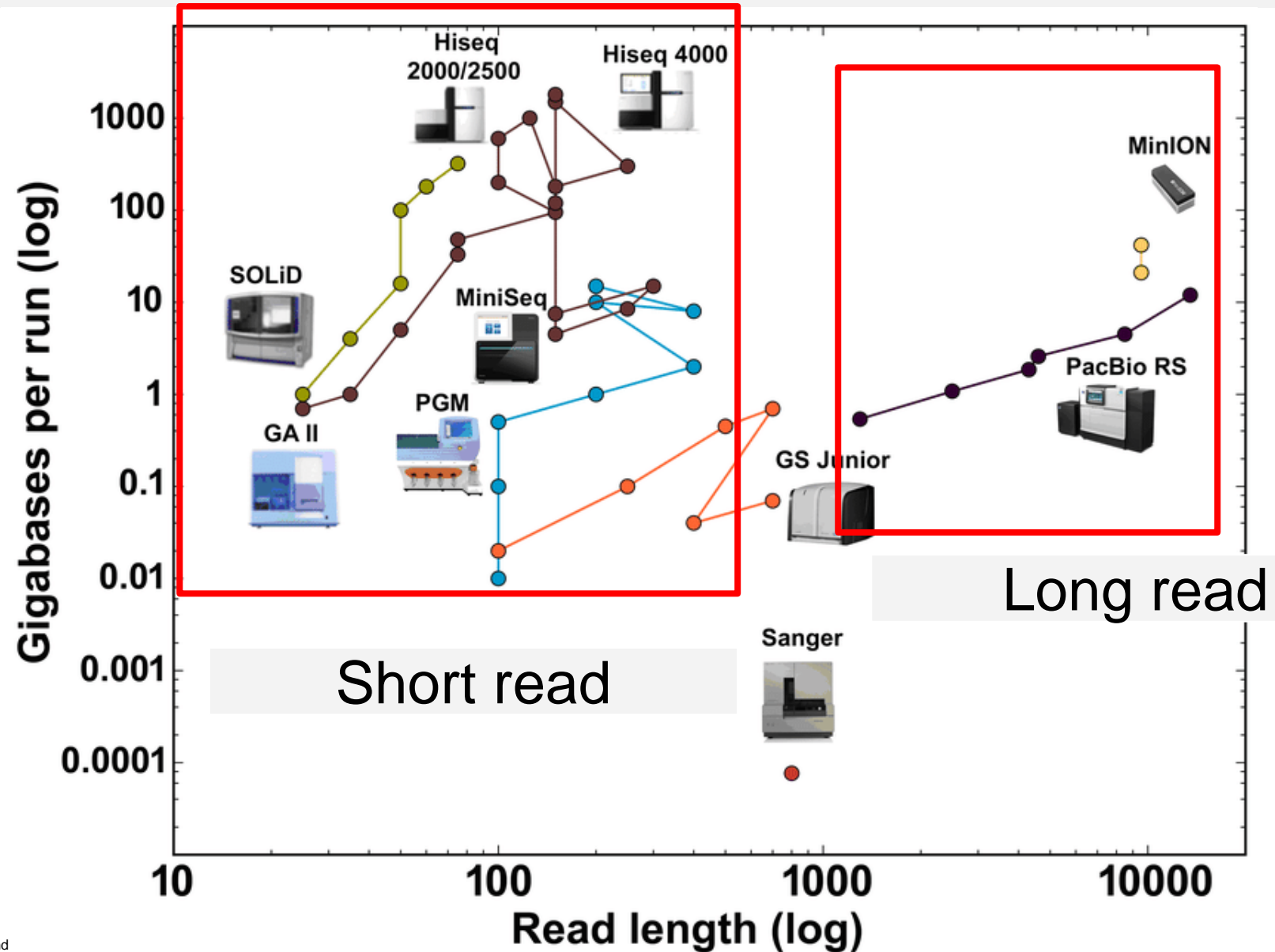


Oxford Nanopore sequencing

- Nanopore sequencing
- Single-molecule templates
- Low accuracy but long read lengths

For example, MinION (Oxford Nanopore)

Les différentes technologies



Les différentes technologies

Short read



MiniSeq
25 million reads



MiSeq
25 millions reads, 2 x 300 bp



NextSeq
400 million reads



HiSeq 4000
5 billion reads



HiSeq X
6 billion reads



NovaSeq 6000
20 billion reads

Taux d'erreur : <1%

Long read



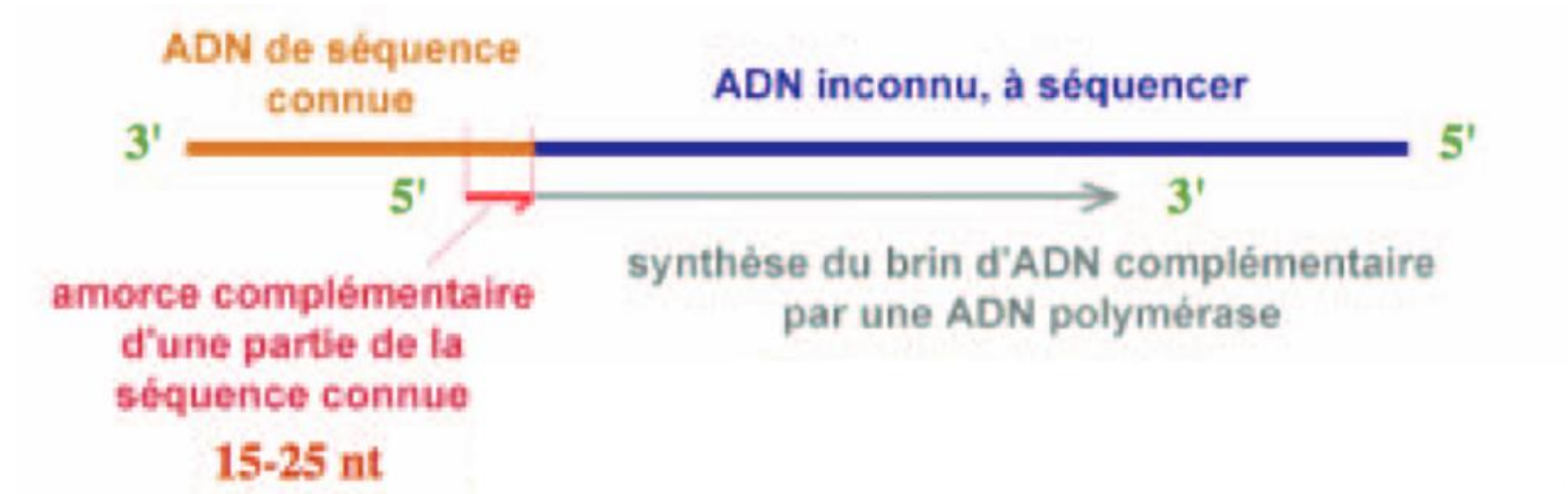
Sequel - Pacific Biosciences
Single molecules
Up to 80,000 bp long
Error rate \approx 10-15 % - CCS: <1%
Compensated by coverage



MinION - Oxford Nanopore
Single molecules
> 200 000 bp long
Error rate \approx 10-15 %
Compensated by coverage

Taux d'erreur : 5-10%

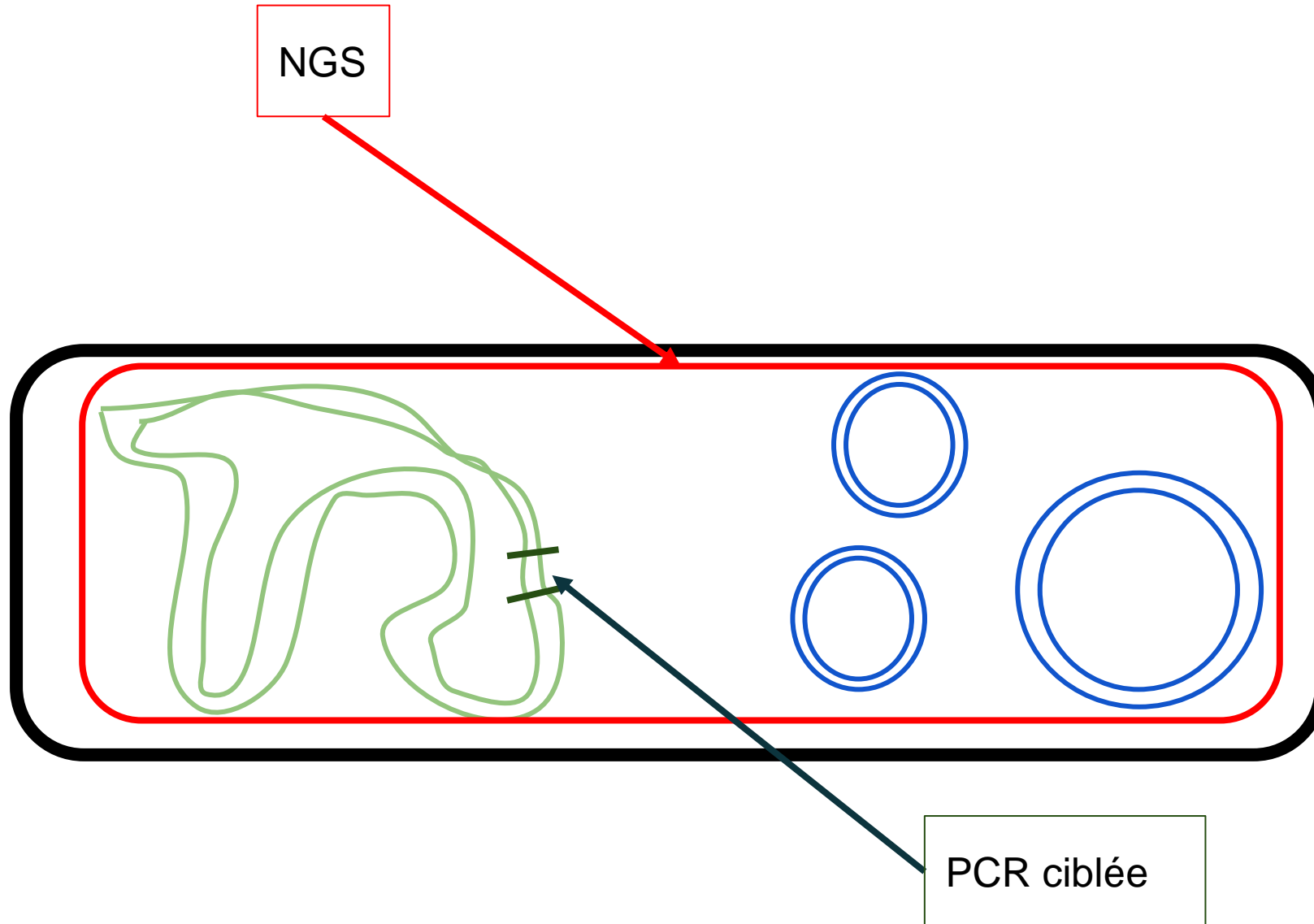
Méthode sanger



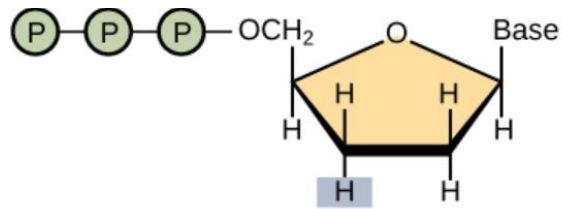
Préparer 4 solutions :

- le fragment qui doit être séquencé
- Amorce
- les 4 dNTP's (dCTP, dATP, dGTP, dTTP)
- l'ADN polymérase

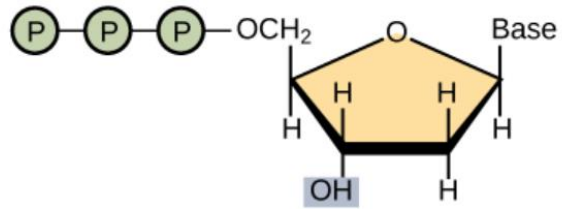
NGS *versus* PCR ciblée



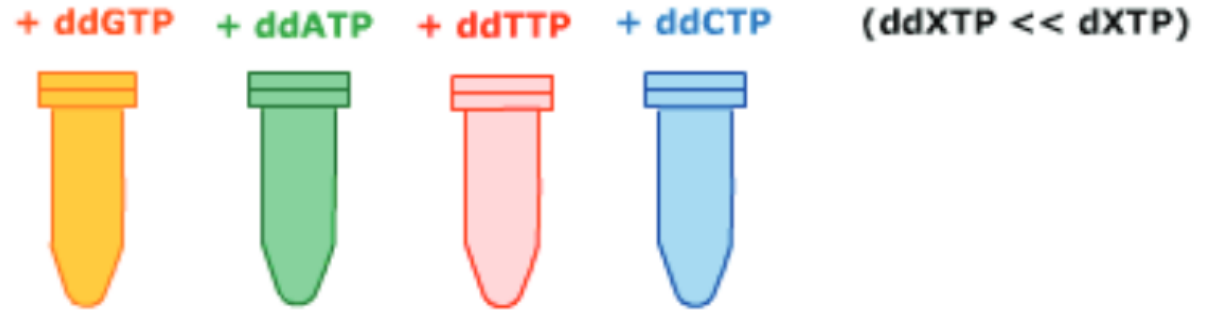
Méthode sanger



Dideoxynucleotide (ddNTP)



Deoxynucleotide (dNTP)



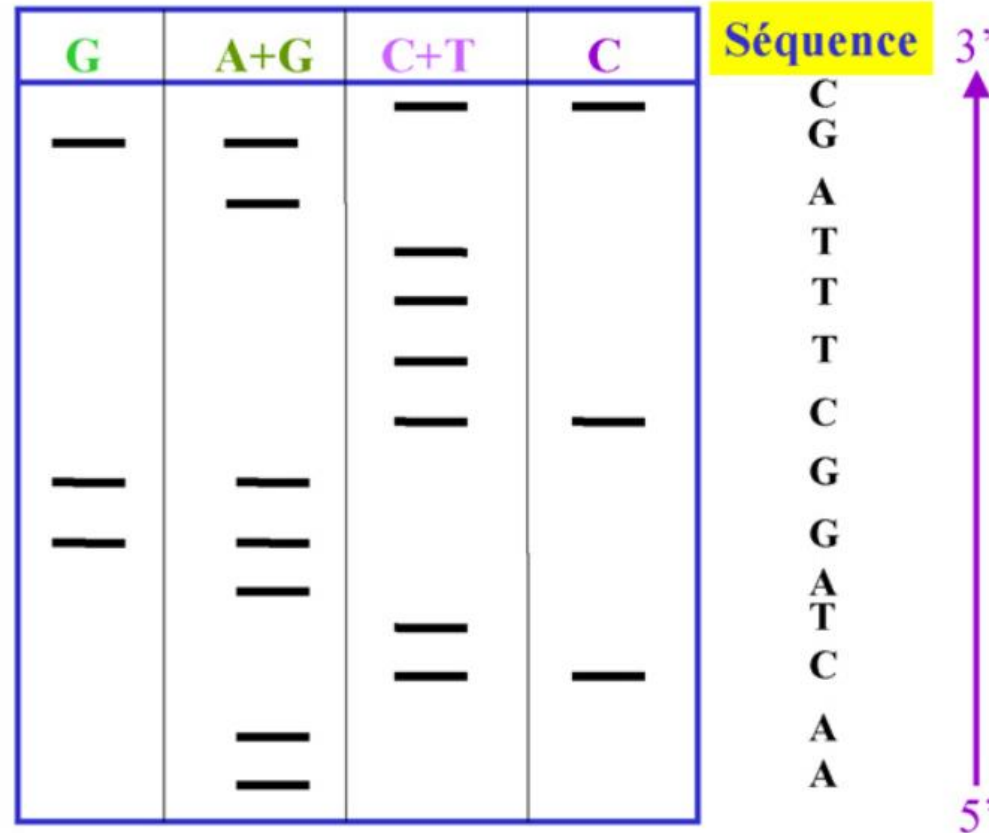
http://www.warpe.snv.jussieu.fr/cours/vt/images_10/sangerprinc.gif

4 solutions à préparer

- Dans chaque tube, on met de petites quantités d'un ddNTP fluorescent ou radioactif
- L'incorporation aléatoire d'un ddNTP stoppe la synthèse
- On obtient donc à la fin des réactions un ensemble de brins d'ADN de tailles variées, selon l'endroit où un ddNTP se sera inséré.

Méthode sanger

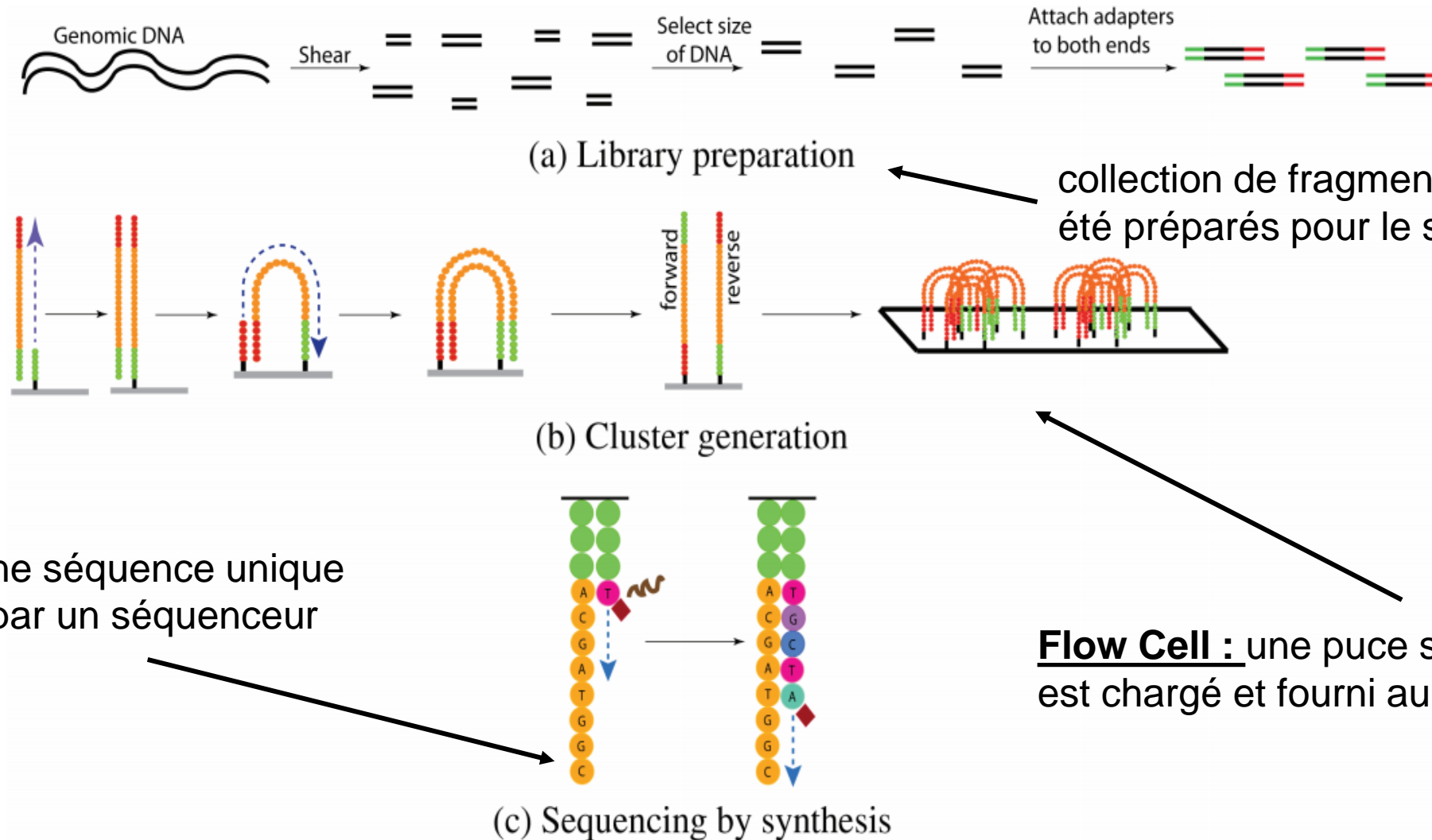
Autoradiogramme après traitement chimique des fragments



Séquence: 5' - AACTAGGCTTTAGC - 3'

Technique de Maxam-Gilber

Illumina sequencing

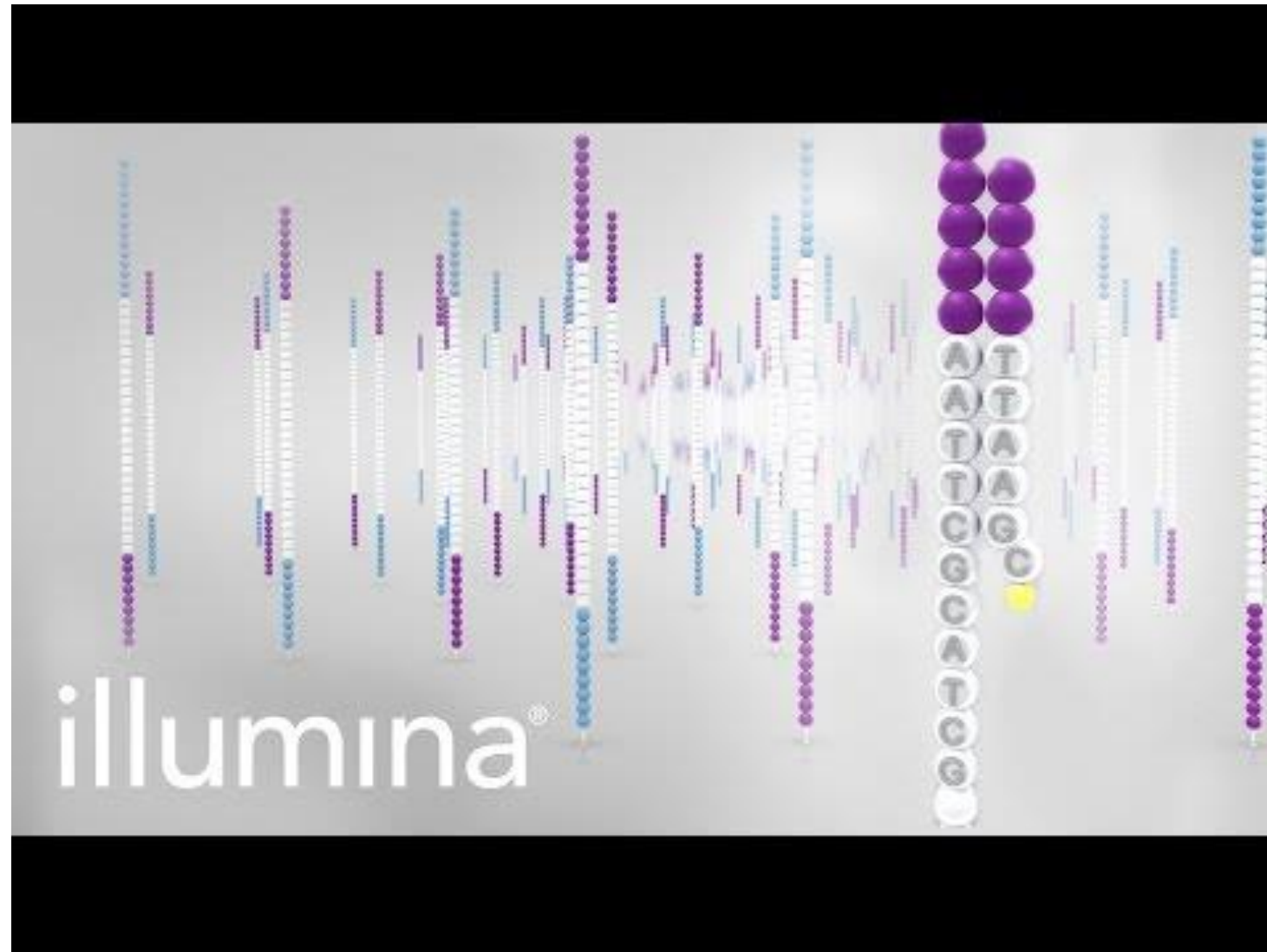


Read : une séquence unique produite par un séquenceur

Flow Cell : une puce sur laquelle l'ADN est chargé et fourni au séquenceur.

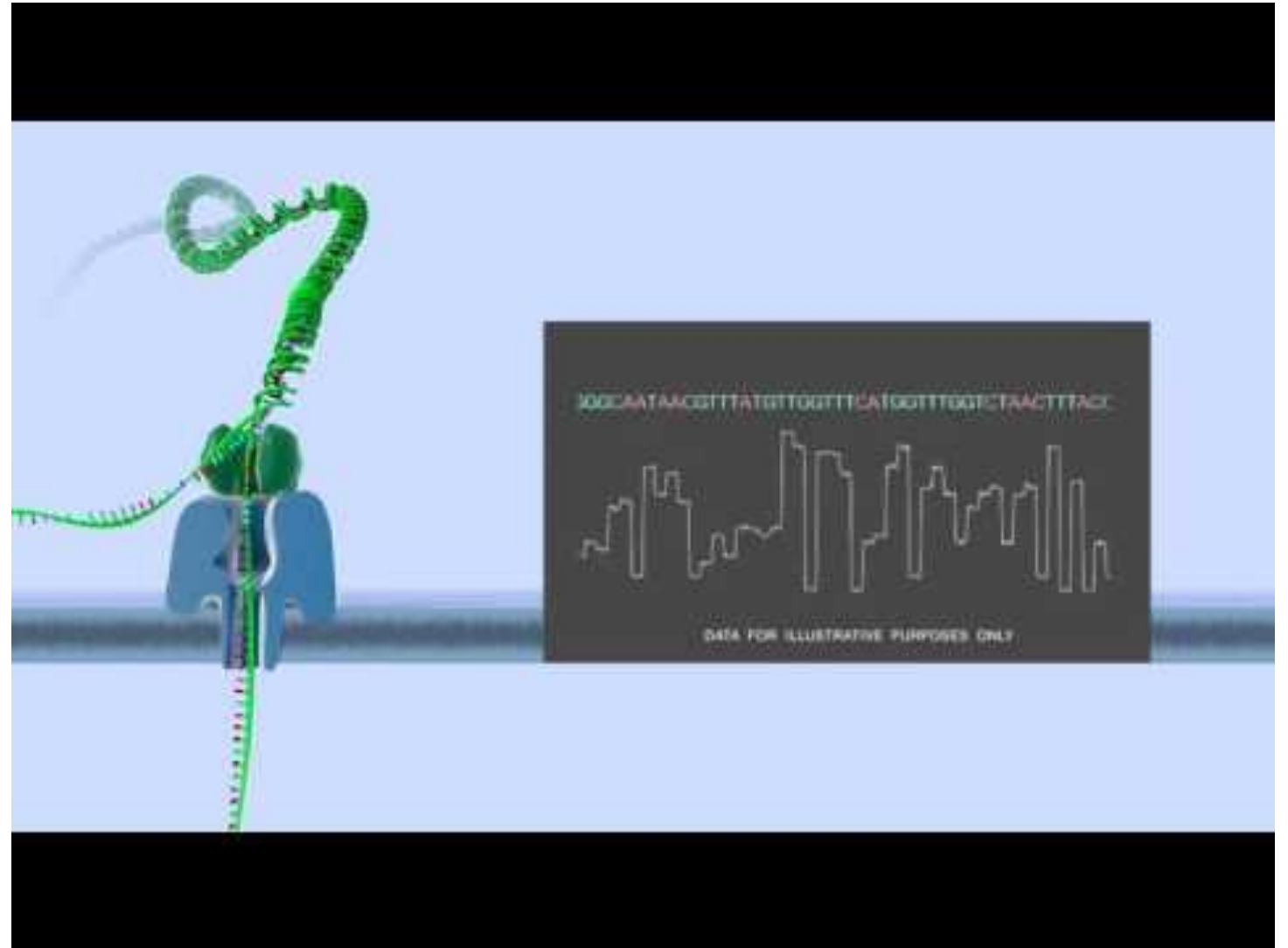
(c) Sequencing by synthesis

Illumina sequencing

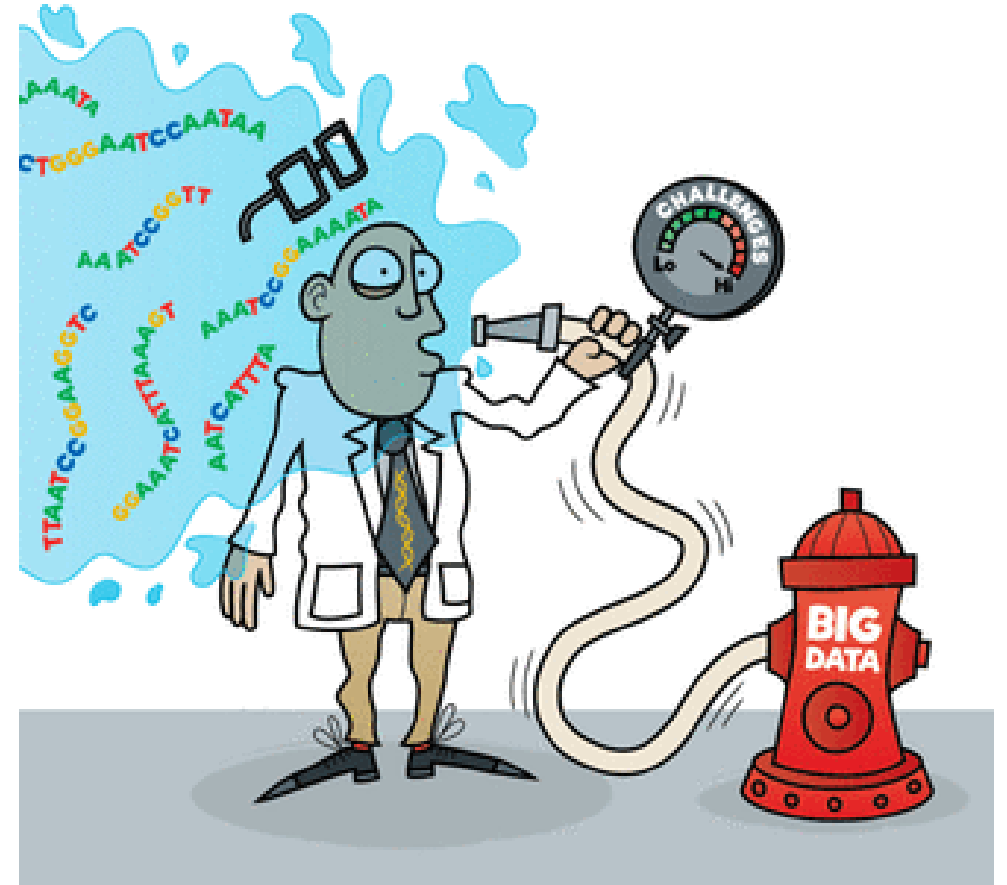


Minion sequencing

- **Nanopore** protéique couplé à une exonucléase
- Application d'un champ électrique
- **Exonucléase** détache successivement les bases
- **Variation d'impédance** selon A, C, G ou T



Workflow d'analyse



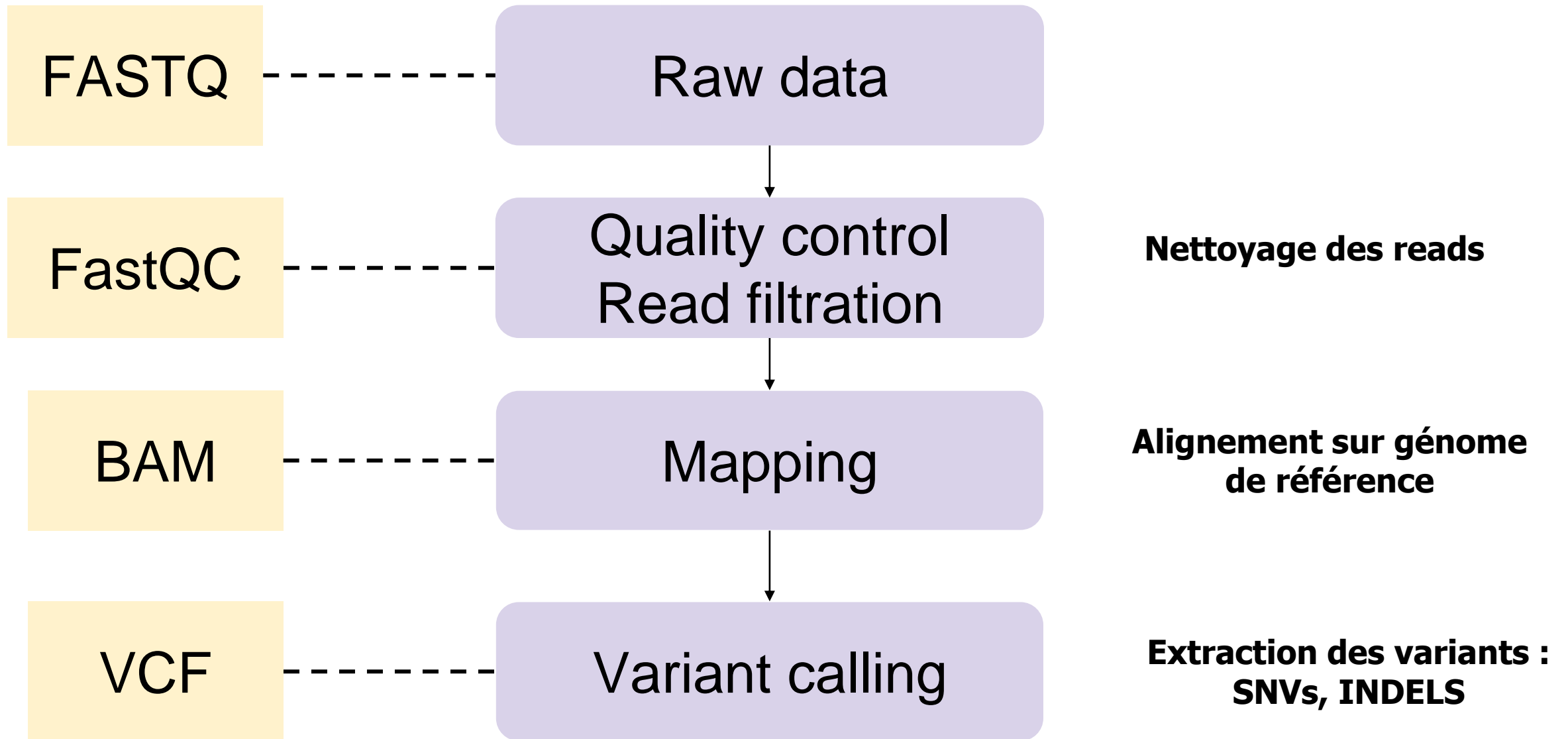
#!/bin/bash

The challenges of big data

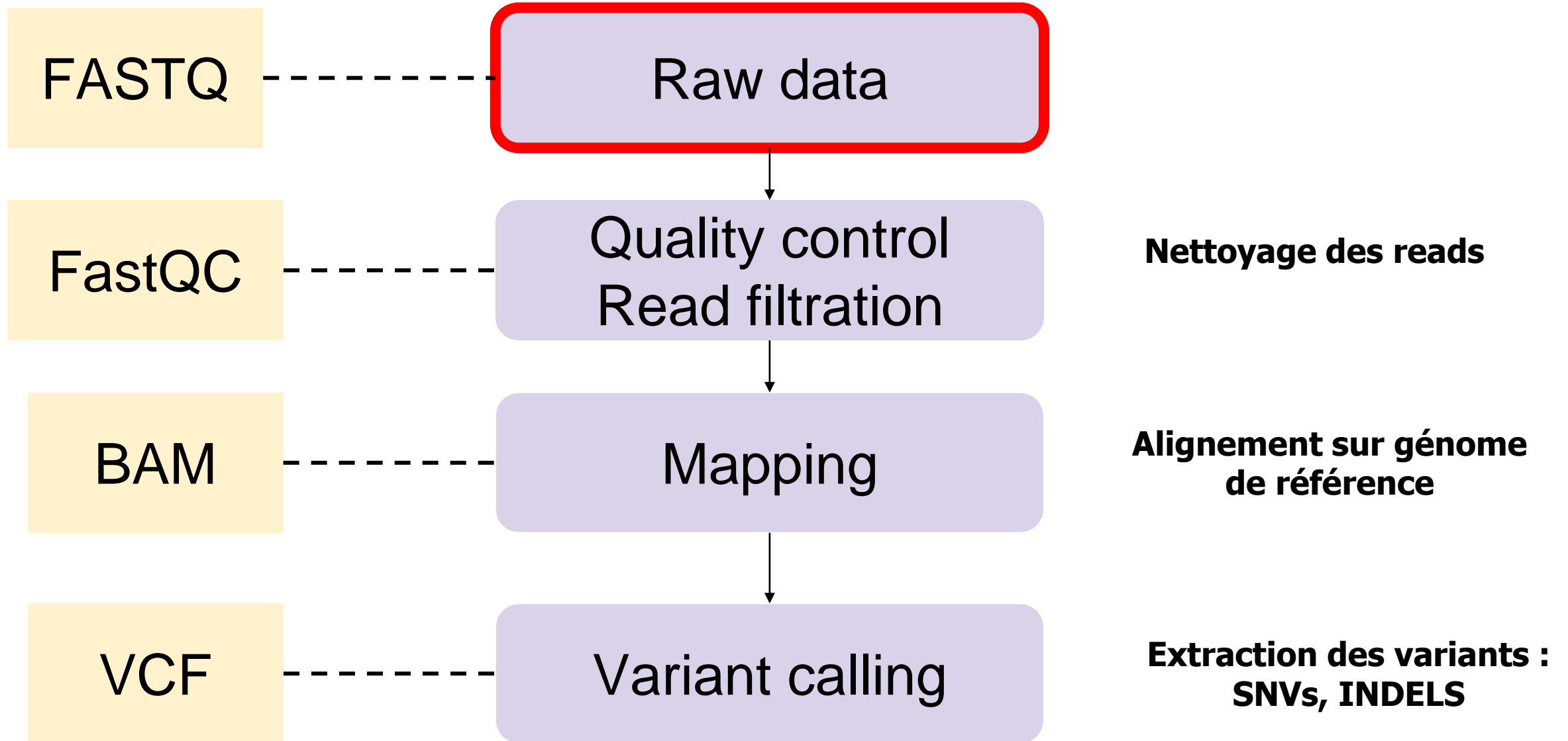
Elaine R. Mardis

Disease Models & Mechanisms 2016 9: 483-485; doi: 10.1242/dmm.025585

Workflow d'analyse



Workflow d'analyse



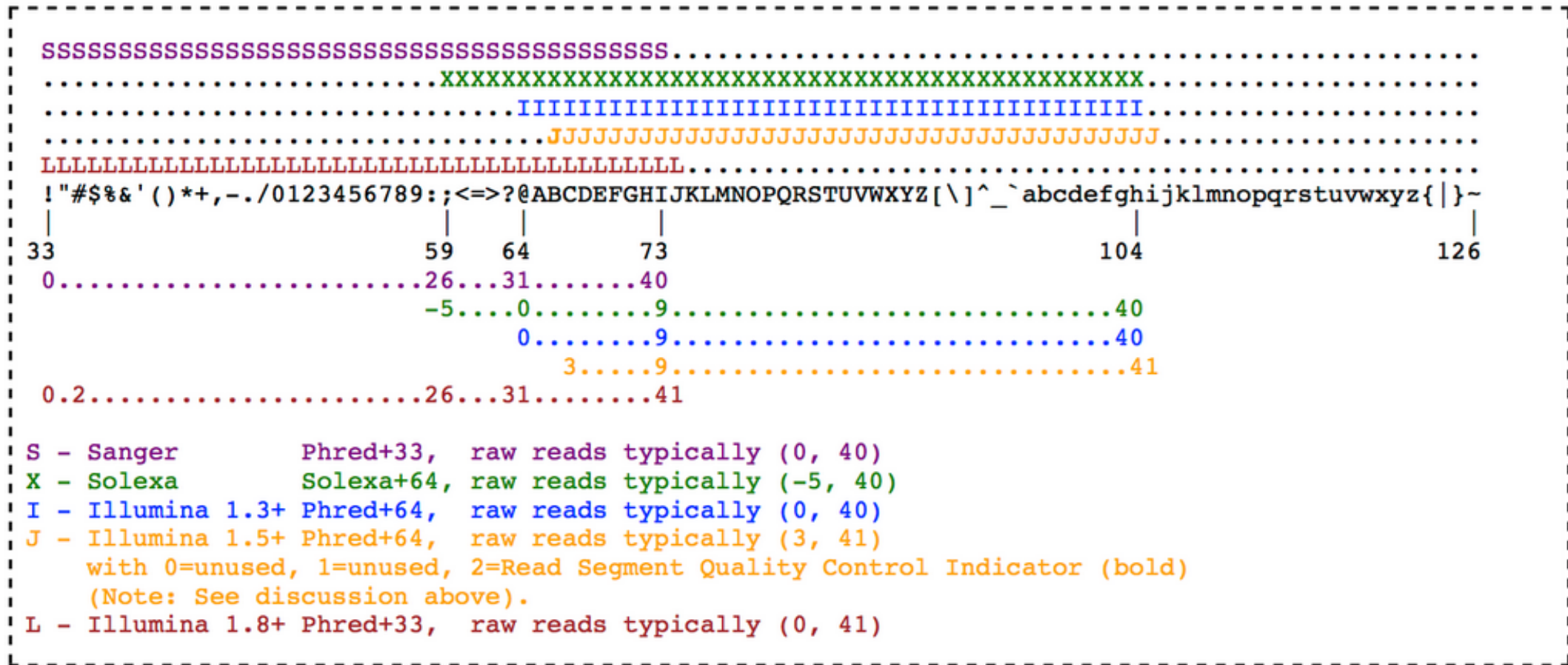
FASTQ syntax

Le format FASTQ (FASTA + mesure de qualité) est composé de 4 sections :

1. il utilise le **symbole @**. Il est suivi d'un ID et d'autres textes optionnels
2. La deuxième section contient la séquence mesurée (typiquement sur une seule ligne)
3. La troisième section est marquée **par le signe +**
4. La dernière ligne code les **valeurs de qualité**

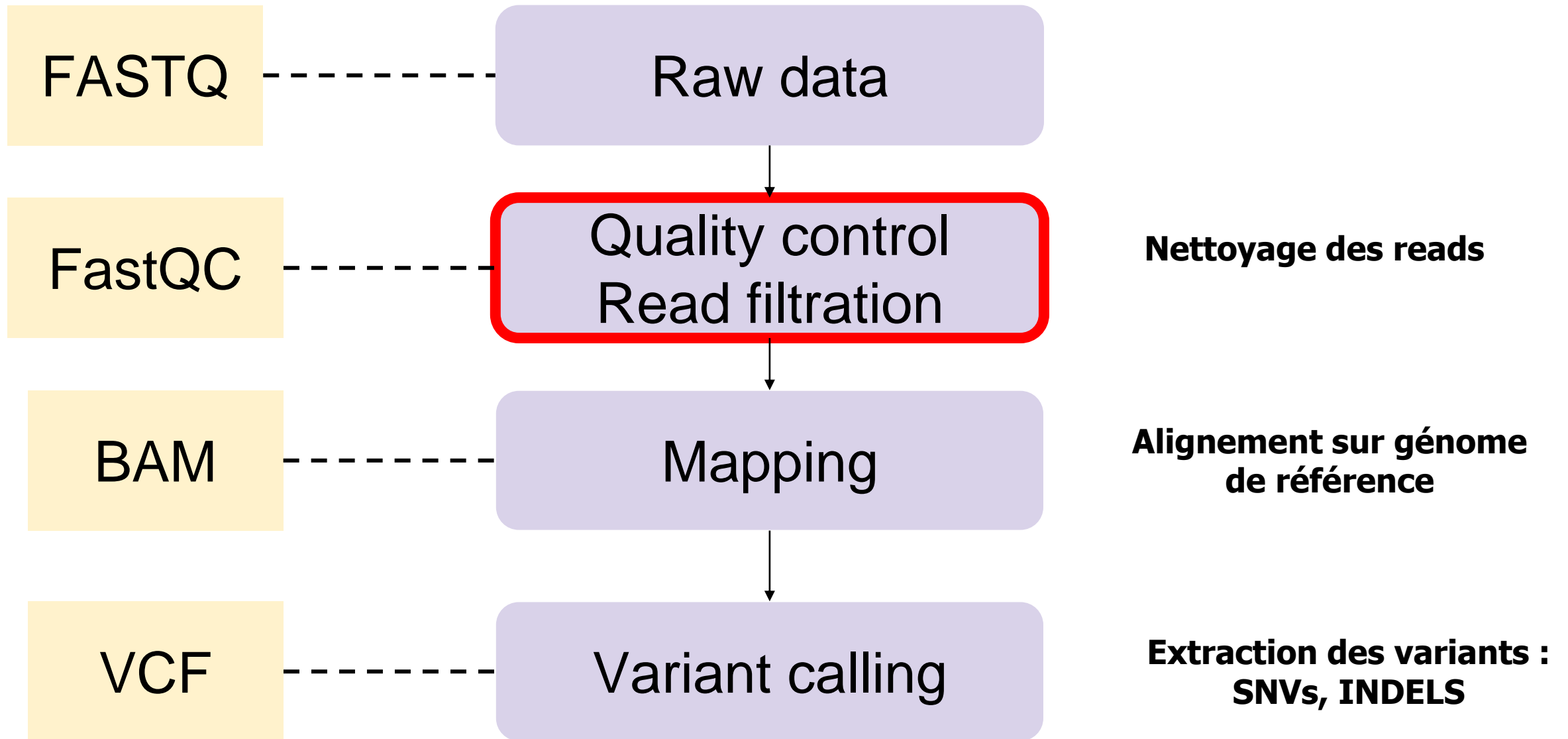
```
1 @ERR000589.41 EAS139_45:5:1:2:111/1
2 CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
3 +
4 3IIIIIIIIIIII>1IIIF9BG08E00I%IG+&?(4)%00646.C1#&(
5 @ERR000589.42 EAS139_45:5:1:2:1293/1
6 AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTTAAAAGAAAT
7 +
8 IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

FASTQ syntax



- Quality scores are typically represented using a Phred scoring scheme, where a read quality value = $-10 * \log_{10}(\text{error probability})$
- For example,
 - Quality = 10 => error rate = 10% => base call has 90% confidence
 - Quality = 20 => error rate = 1% => base call has 99% confidence
 - Quality = 30 => error rate = 0.1% => base call has 99.9% confidence
- See [Phred quality score](#) for more details.

Workflow d'analyse



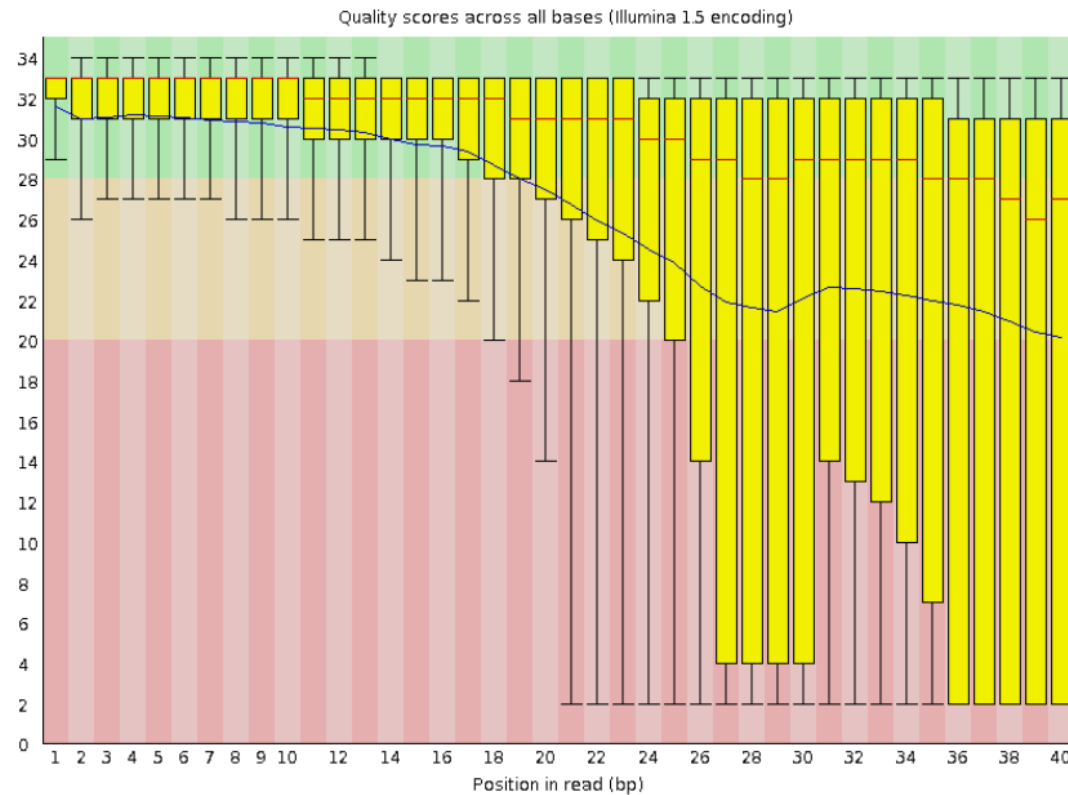
Quality control/ FastQC report

FastQC Report

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ⚠ Per base sequence content
- ⚠ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ⚠ Sequence Duplication Levels
- ⚠ Overrepresented sequences
- ✓ Adapter Content

✗ Per base sequence quality



$$Q = -10 \log_{10} P \quad \longrightarrow \quad P = 10^{-\frac{Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality control/ Trimming

Error correction depends on application...

Ex. 1 SNP calling:

Phred 30: 0.1% P error

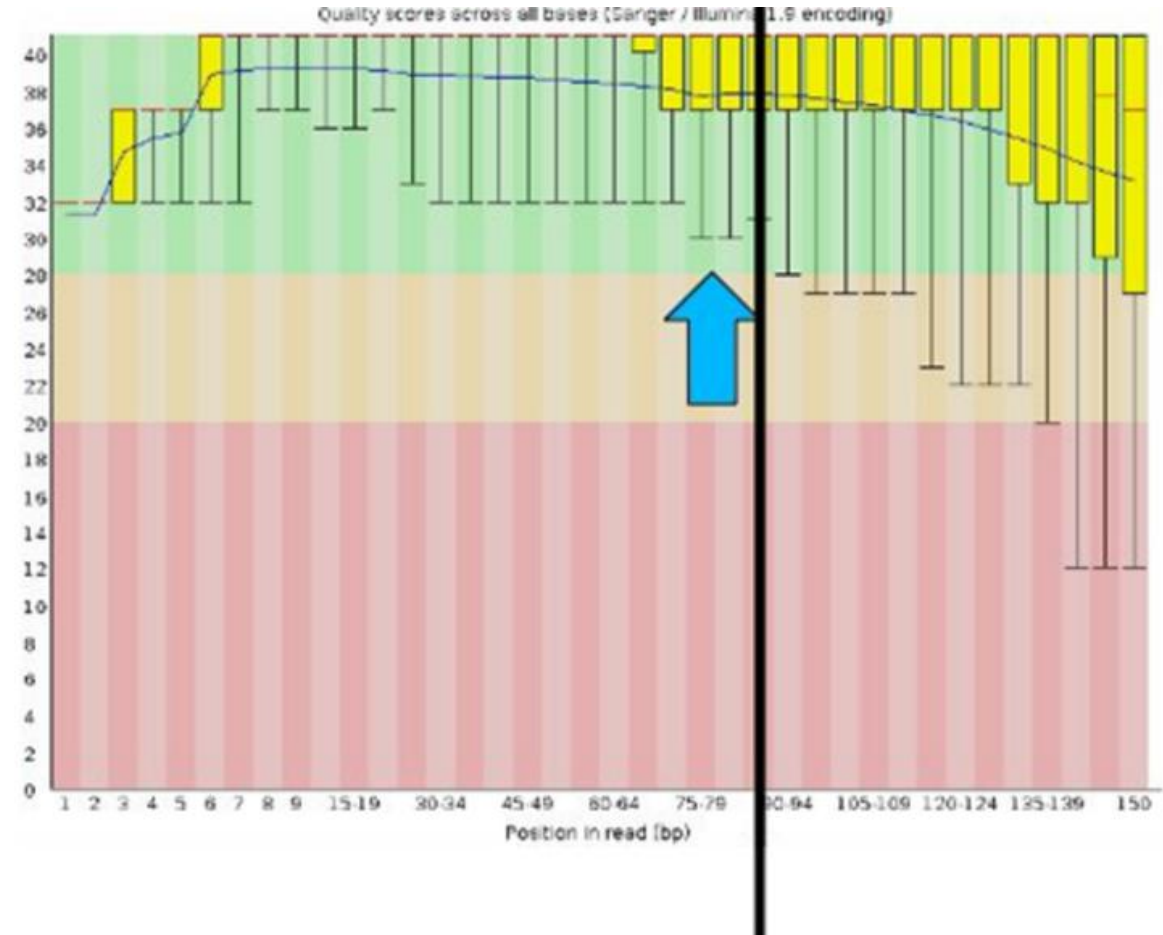
Quality-based trimming

Fastx

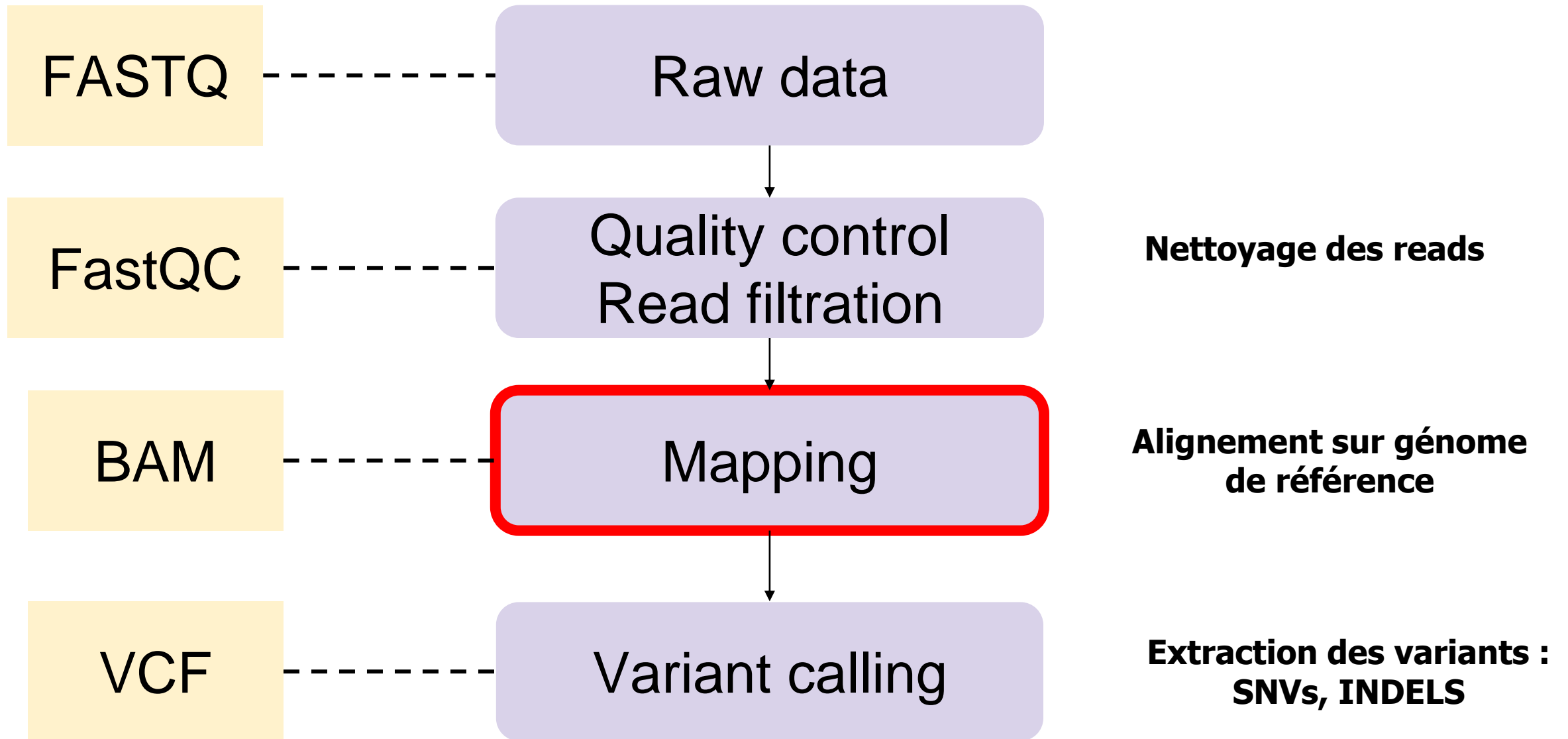
http://hannonlab.cshl.edu/fastx_toolkit/download.html

Trimmomatic

<http://www.usadellab.org/cms/?page=trimmomatic>



Workflow d'analyse



Mapping



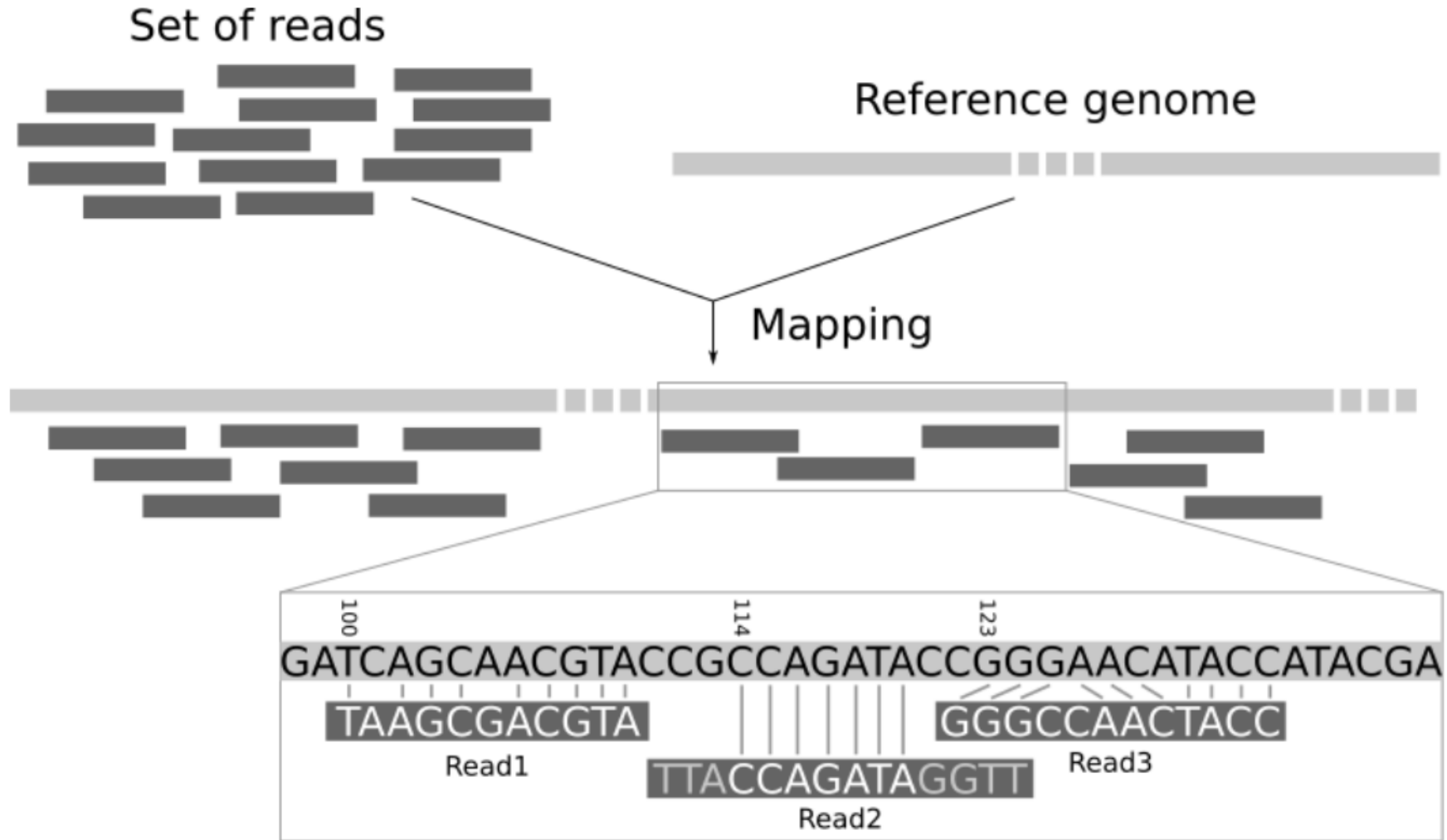
Reads



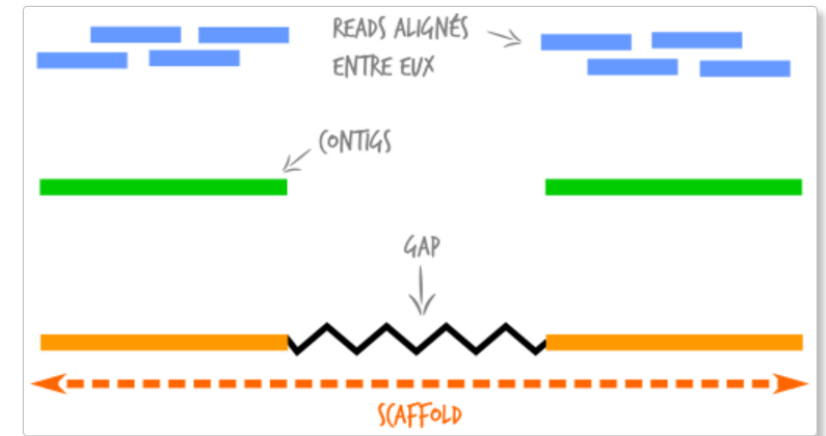
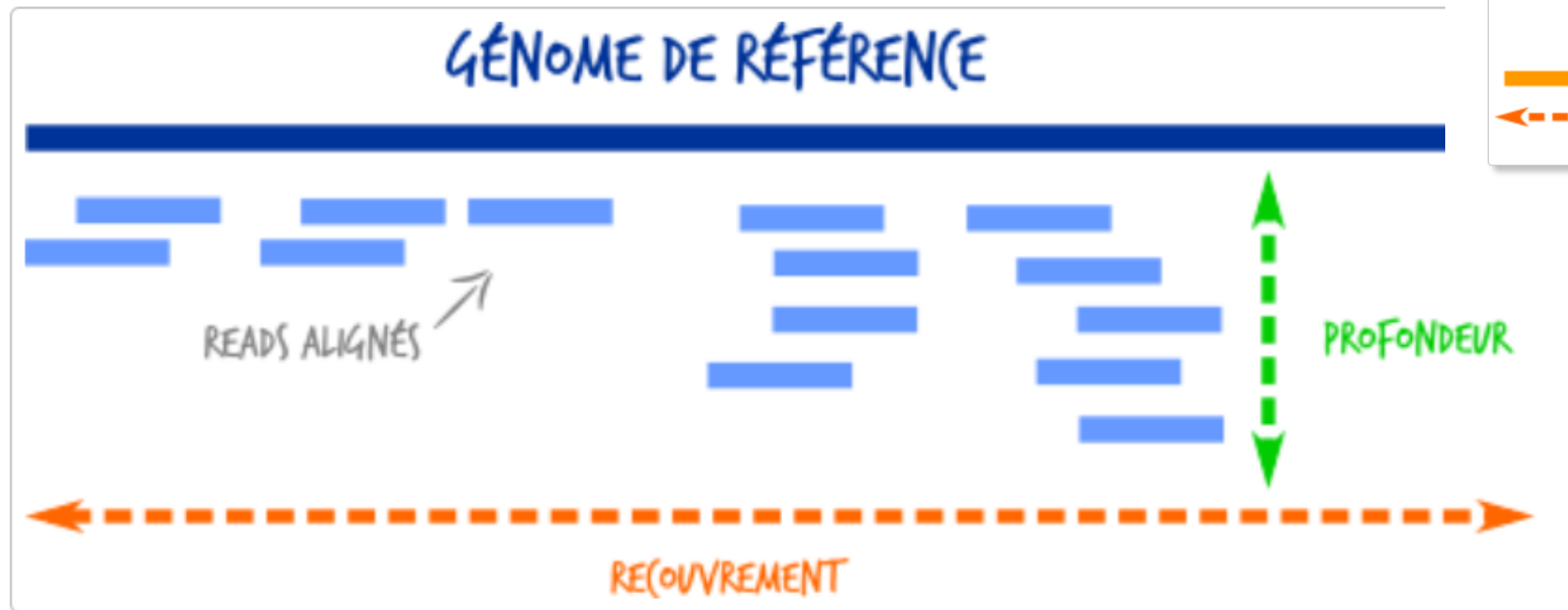
[Virgil and the Muses, Bardo Museum, Tunis]

Génome assemblé

Mapping



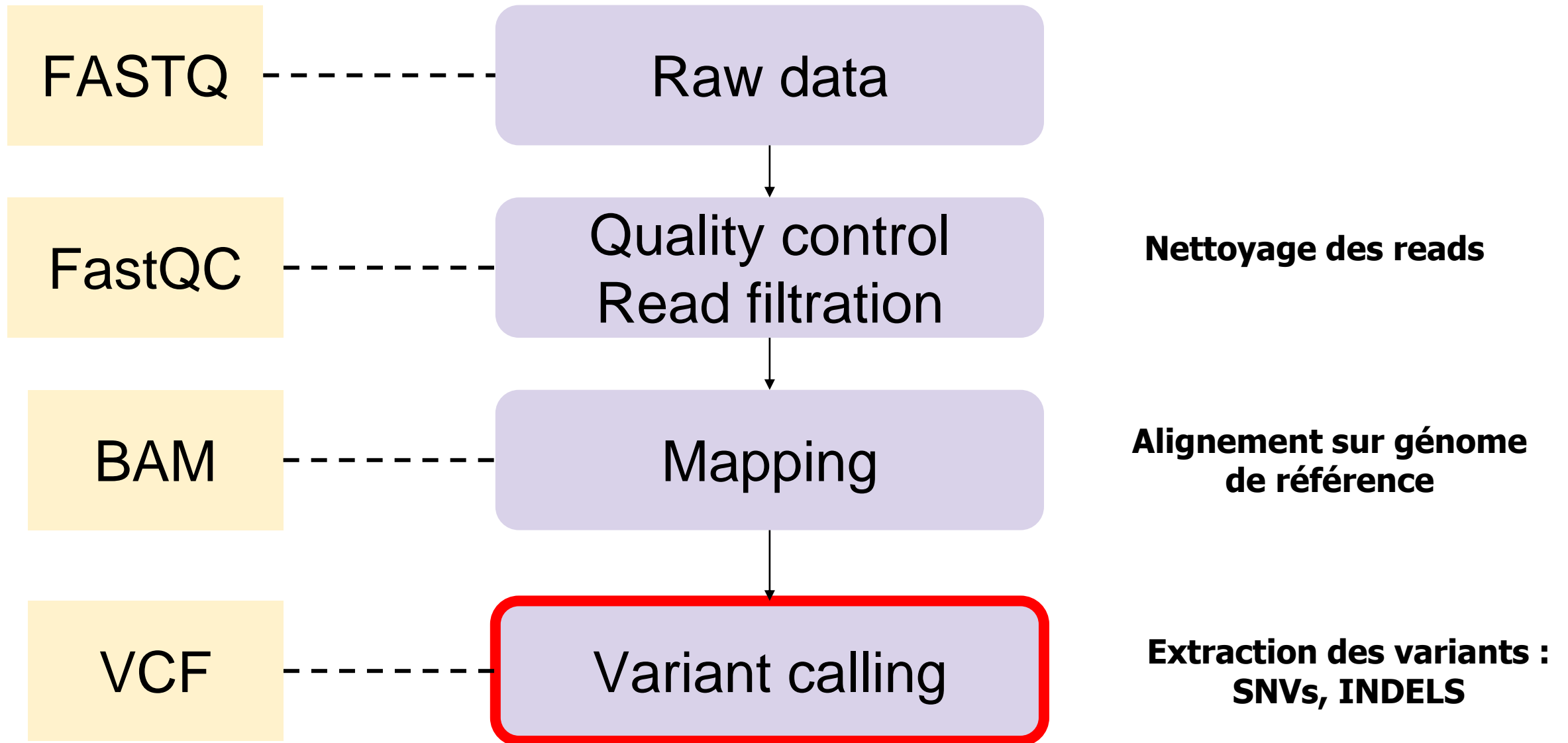
Mapping/ Profondeur et couverture



nombre de lecture de
chaque base, exprimée
en X

zone couverte par au moins une lecture, exprimée en %

Workflow d'analyse



Questions

Détection automatisée des variants (SNVs, Indels) à partir d'un fichier contenant des données de séquençage alignées (BAM)

The diagram illustrates the structure of a VCF file, divided into a **VCF header** and a **Body**.

VCF header: Contains mandatory and optional header lines.

- Mandatory header lines:** Lines starting with `##` that define the file format and reference genome.
- Optional header lines:** Lines starting with `##` that provide meta-data about the annotations in the VCF body.

Body: Contains variant records. Each record is a line of text with fields separated by tabs.

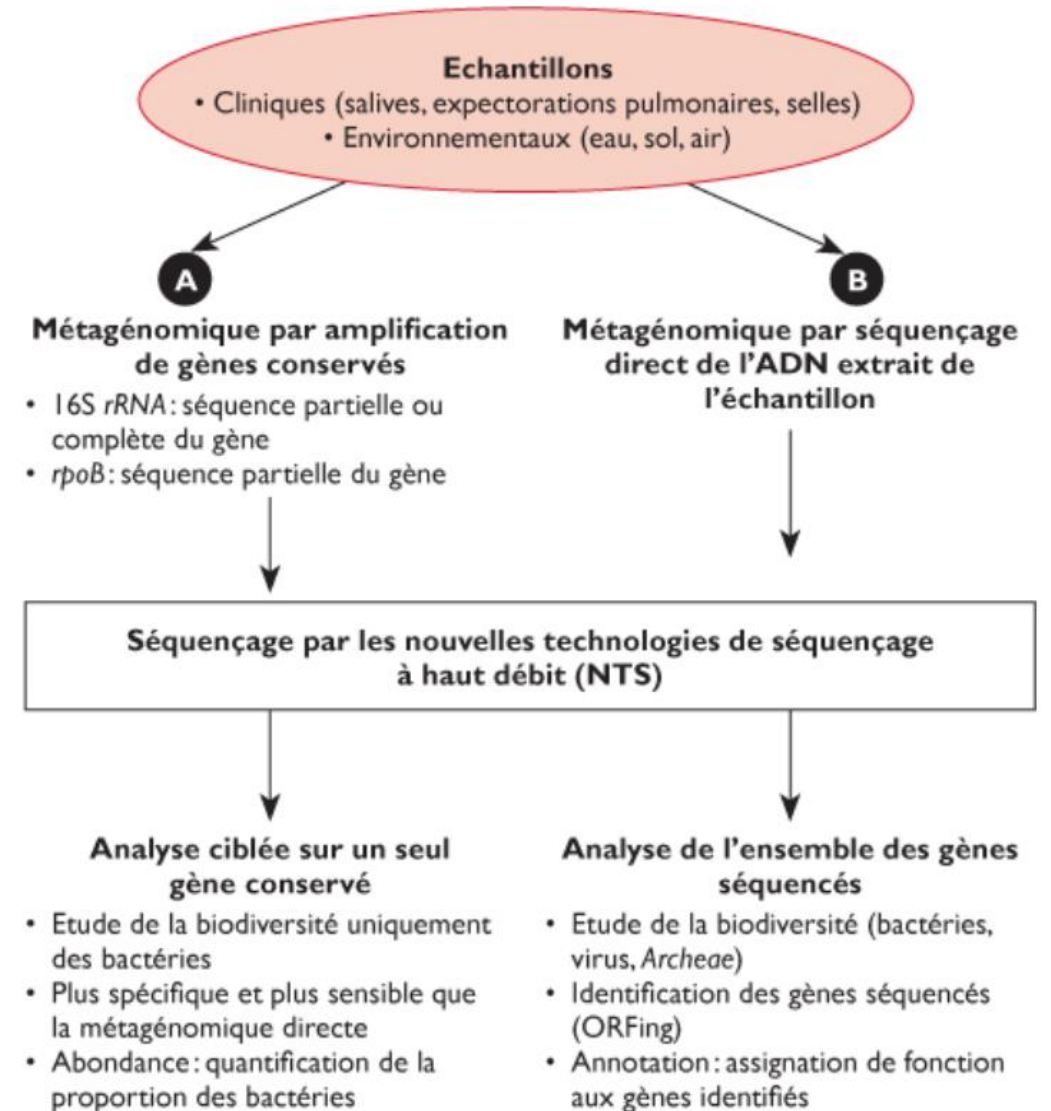
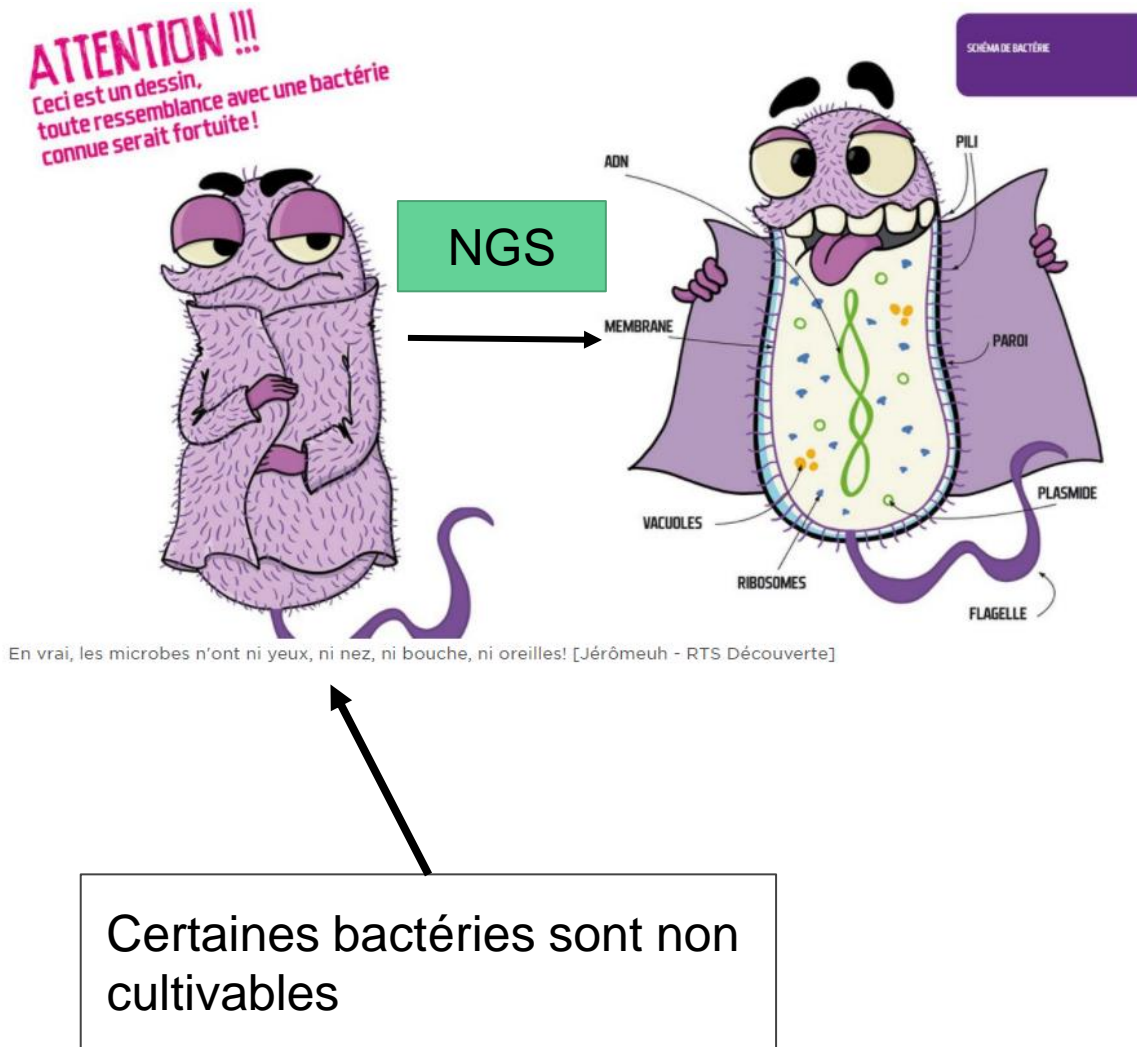
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0/1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1/0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Annotations:

- Reference alleles (GT=0):** Alleles present in the reference genome (e.g., A, C, G, T).
- Alternate alleles (GT>0 is an index to the ALT column):** Alleles found in the sample that differ from the reference.
- Deletion:** Indicated by `` in the ALT field.
- SNP:** Single Nucleotide Polymorphism.
- Insertion:** Indicated by a sequence not in the REF field.
- Other event:** Other types of structural variants.

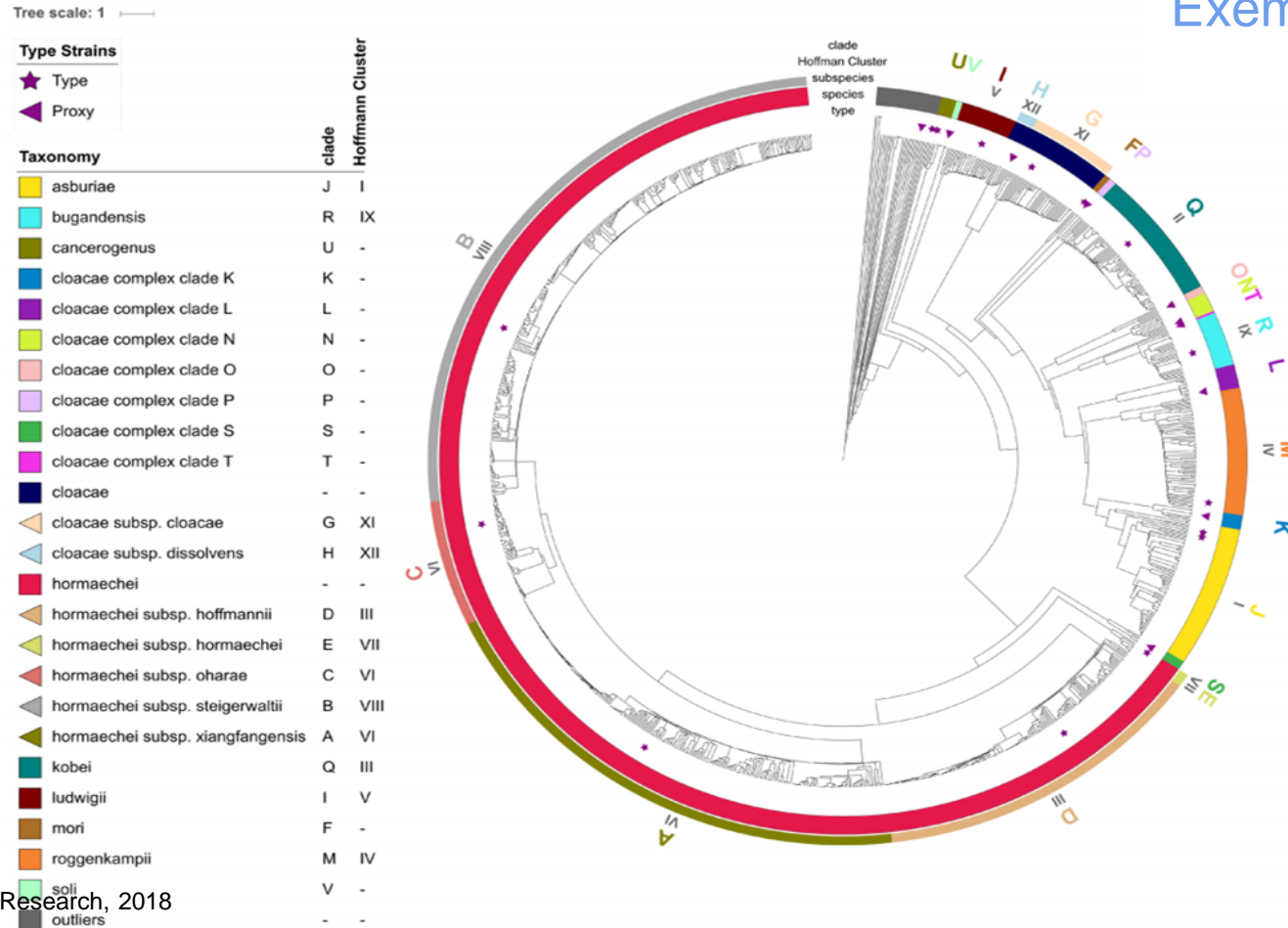
résultats dans fichier « VCF » = Variant Calling File

Application/ Métagénomique



Application/Phylogénie (ANI)

Exemple : complexe *Enterobacter cloacae*



Sutton et al., F1000Research, 2018

Diversité génomique
au moins 7 espèces d'intérêt clinique

Application/Resistome

Drugs	pDST	Wild type	Mutation	Total #	Sensitivity % (CI)	specificity % (CI)	PPV % (CI)	NPV % (CI)
		(n)	(n)	isolates				
RIF	Susceptible	1072	0	1072	100	100	100	100
	Resistant	0	31	31	(88.78-100)	(99.66-100)		
INH	Susceptible	966	12	978	93	98.8	90.7	99.2
	Resistant	8	117	125	(87.78-97.2)	(97.93-99.79)	(84.72-94.49)	(98.45-99.59)
PZA	Susceptible	1076	6	1082	85.71	98.8	75	99.72
	Resistant	3	18	21	(63.66-96.95)	(97.93-99.38)	(57.0-87.16)	(99.2-99.9)
EMB	Susceptible	1082	3	1085	100	99.72	85.71	100
	Resistant	0	18	18	(81.47-100)	(99.19-99.14)	(65.9-94.89)	

Conclusion NGS

- NGS = nombreux axes de recherches
- Méthode puissante (étude épidémiologique, résitome, phylogénie...)
- Nécessite l'utilisation de la biologie intégrative (système complexe)
- Attention à la reproductibilité des analyses bioinformatiques

Merci !