Partie 6

Perspectives et outils d’avenirs

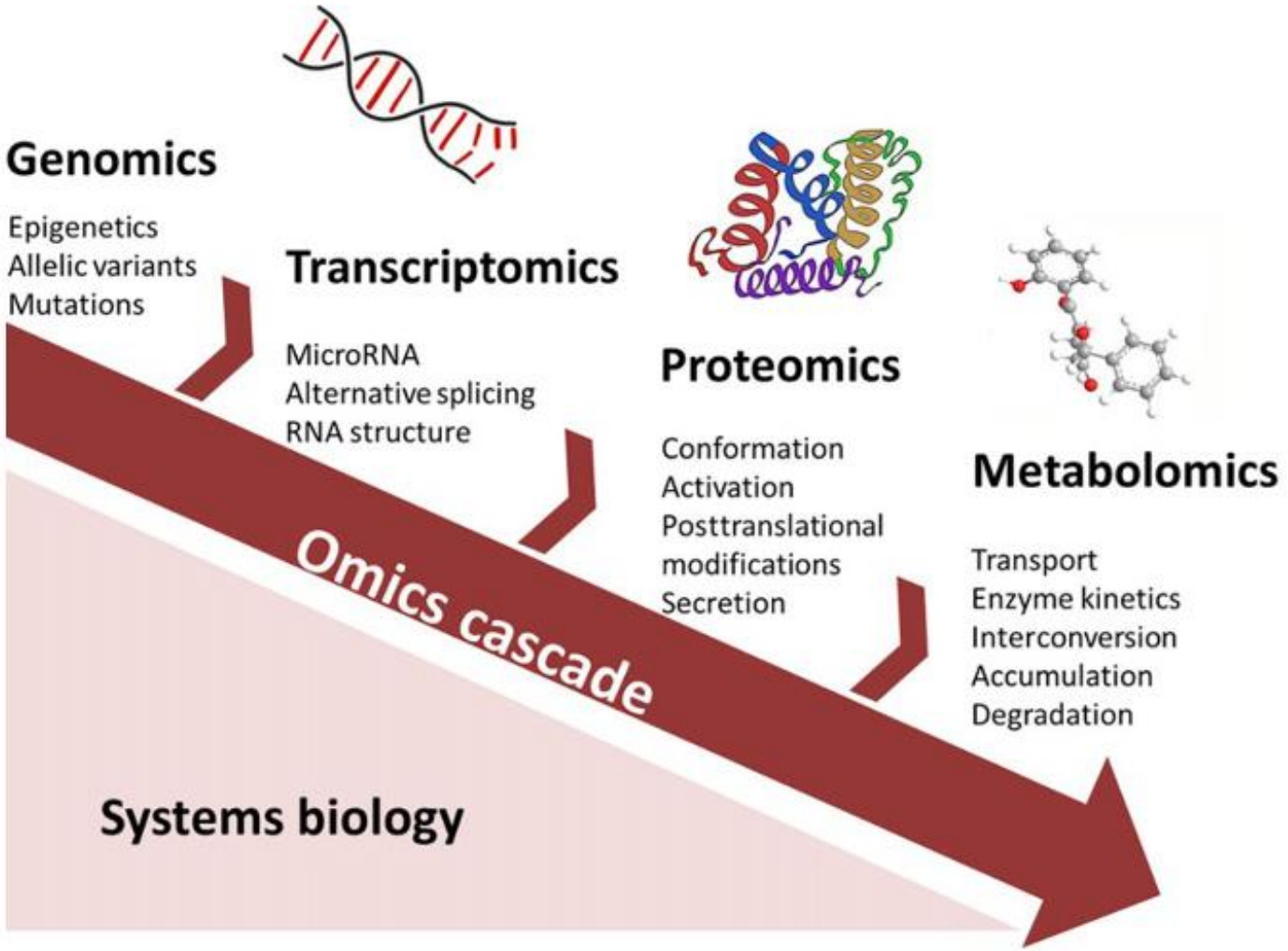
SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Groupe Hospitalier Pitié-Salpêtrière

Inserm
Institut national de la santé et de la recherche médicale

Cimi PARIS
Centre d'Immunologie et des Maladies Infectieuses
www.cimi-paris.upmc.fr

Omics sciences

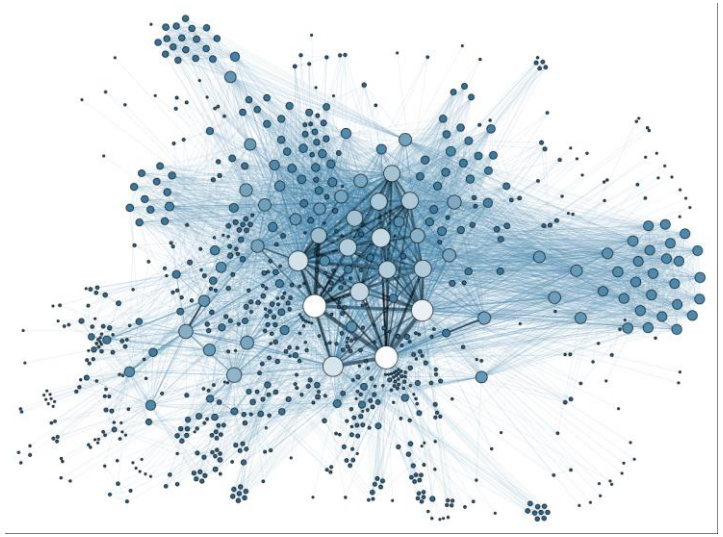


Carmen Bedia,
Chapter Two - Experimental Approaches in Omic Science
Editor(s): Joaquim Jaumot, Carmen Bedia, Romà Tauler,
Comprehensive Analytical Chemistry,
https://doi.org/10.1016/bs.coac.2018.07.002.

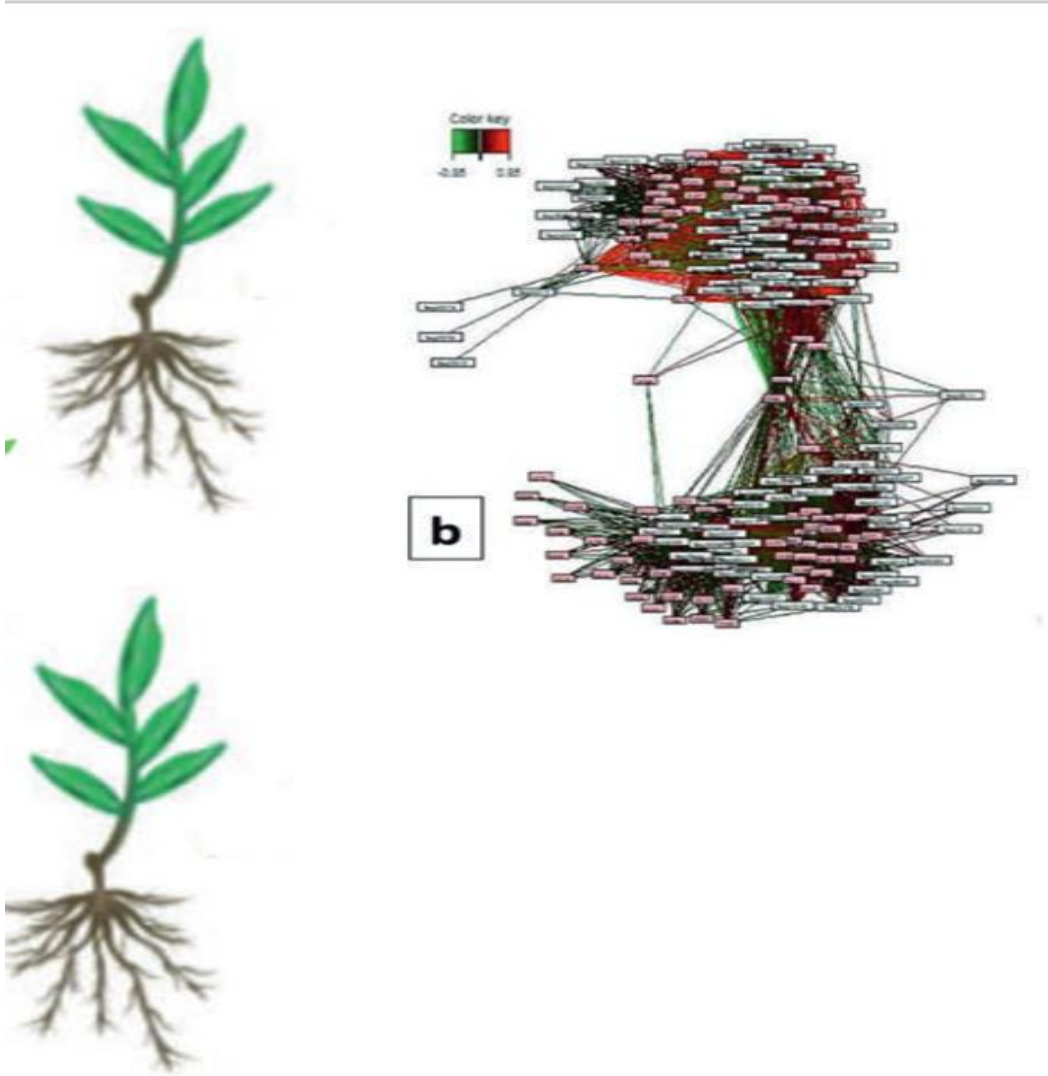
ome = totalité

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.

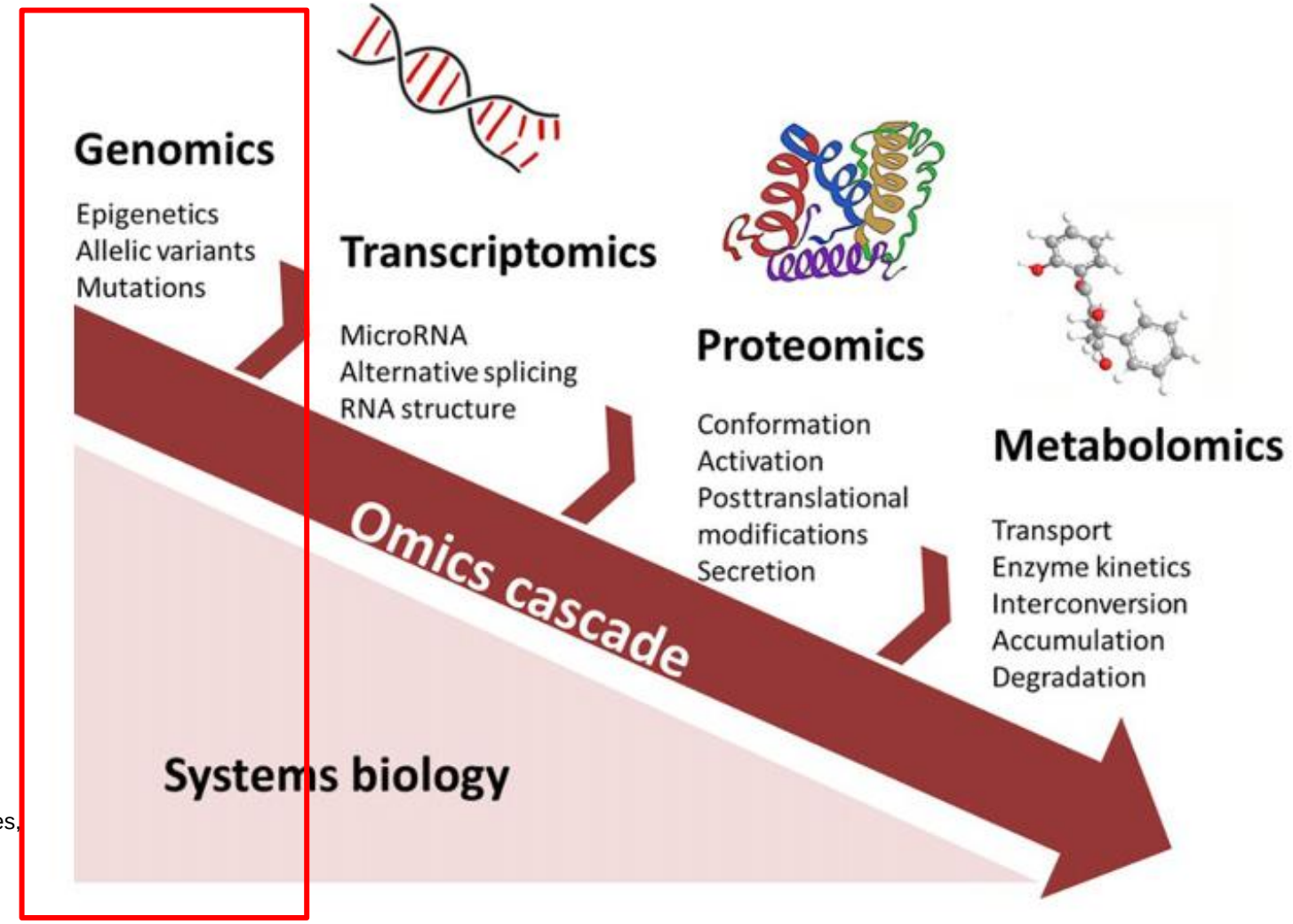


https://www.agropolis.fr/pdf/chapitres-dossier-thematique-systemes-complexes/systemes-complexes-comprehension-analyse.pdf



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

Omics sciences



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

Génomes séquencés

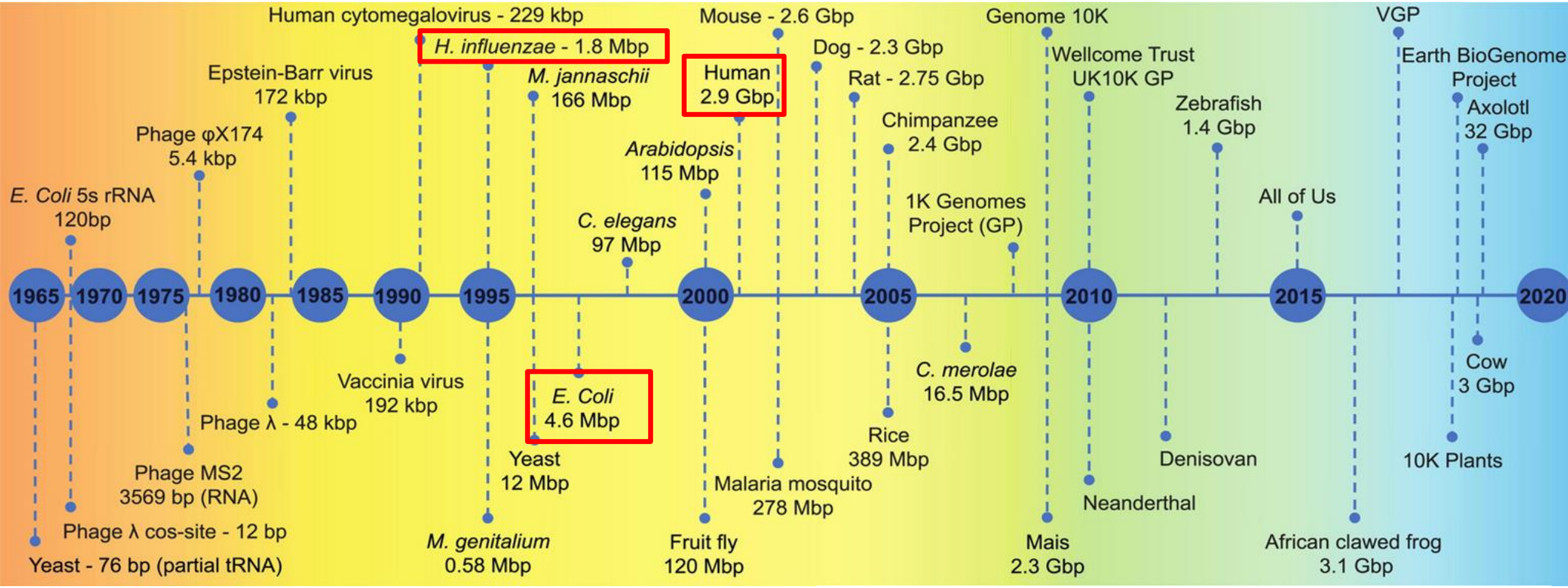


Fig. 1. Milestones in genome assembly. Timeline illustrating many of the major genome assembly achievements ranging from the beginning of the sequencing era to the large-scale genome projects currently ongoing. Each genome or genome project (GP) is placed under a color-coded background according to the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

Alice Maria Giani, Guido Roberto Gallo, Luca Gianfranceschi, Giulio Formenti,
Long walk to genomics: History and current approaches to genome sequencing and assembly,
Computational and Structural Biotechnology Journal, Volume 18, 2020,
https://doi.org/10.1016/j.csbj.2019.11.002.

Coût du séquencage

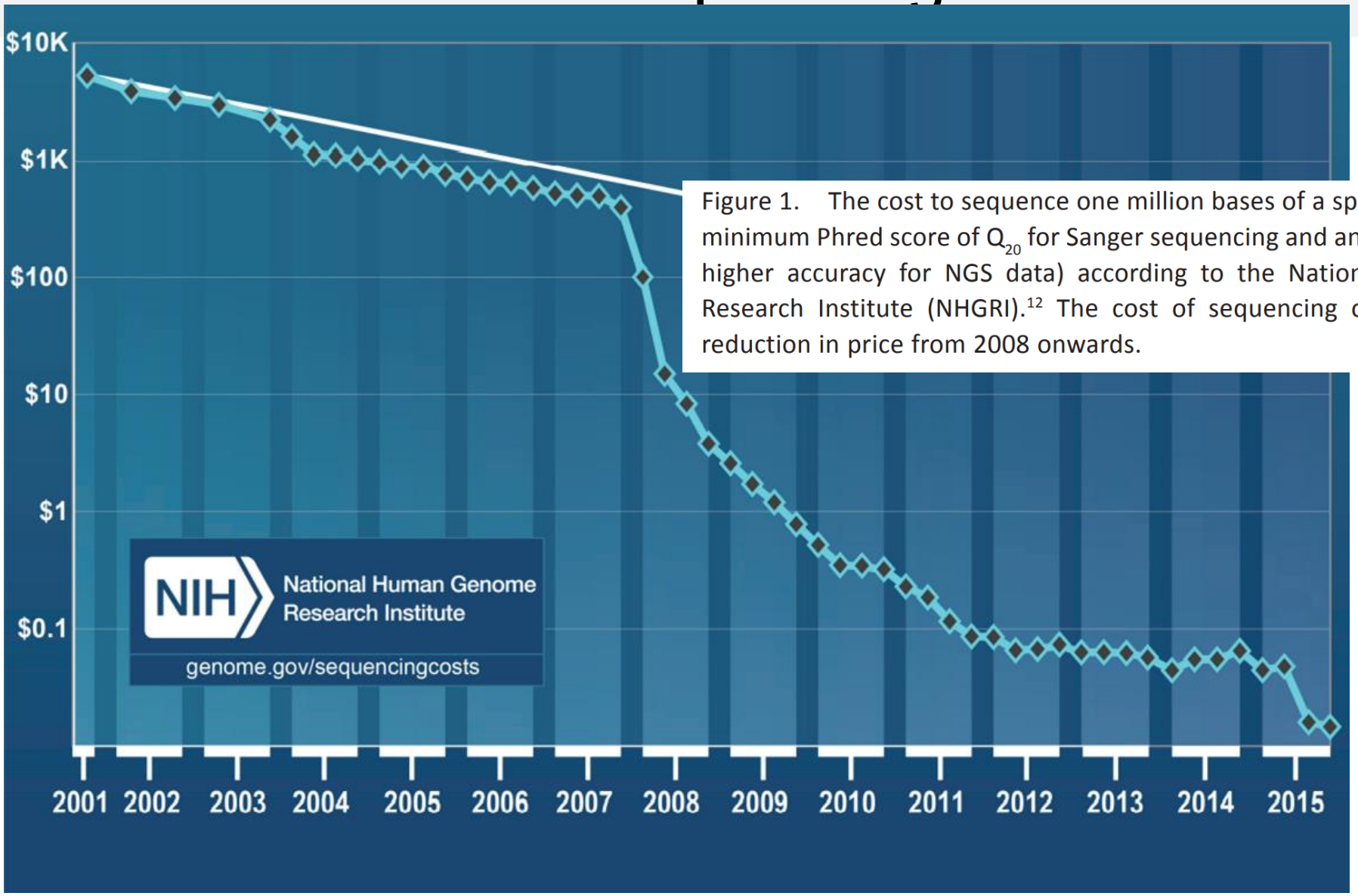


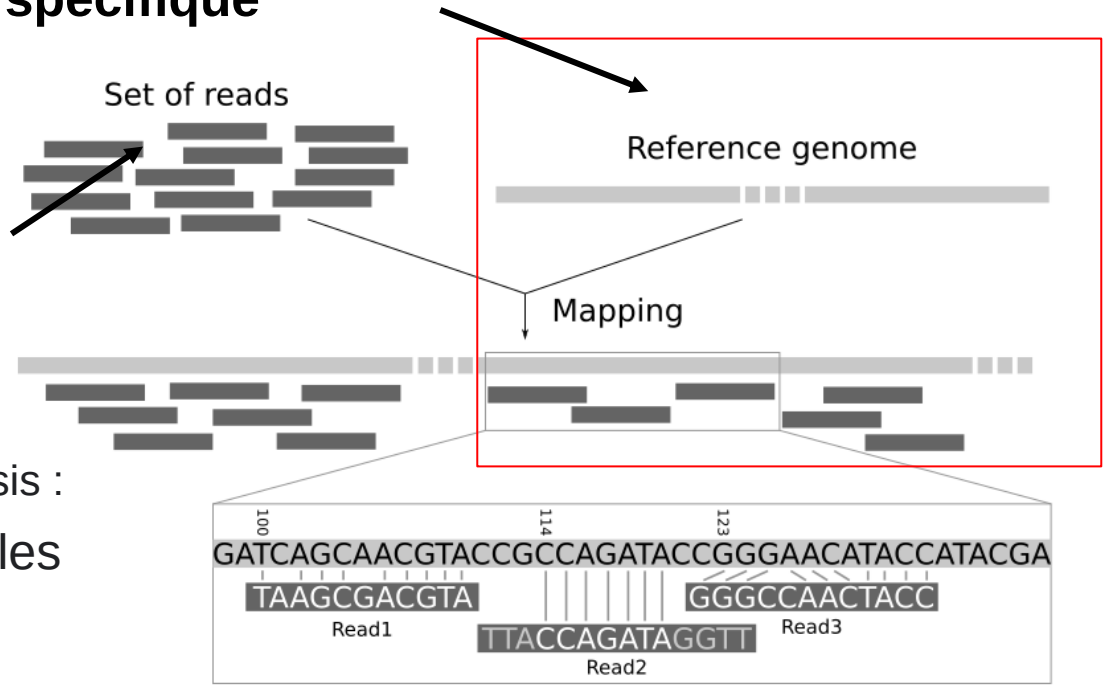
Figure 1. The cost to sequence one million bases of a specified quality (i.e. a minimum Phred score of Q_{20} for Sanger sequencing and an equivalent of Q_{20} or higher accuracy for NGS data) according to the National Human Genome Research Institute (NHGRI).[12] The cost of sequencing only made its rapid reduction in price from 2008 onwards.

A Practical Handbook of Next Generation Sequencing and Its Applications
https://doi.org/10.1142/10159 | September 2017

Quelques définitions....

- Il faut un **génome 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 génome 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)

https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

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 **DEL**etion d'une ou plusieurs bases
  - MNV (**M**ulti-**N**ucleotide **V**ariant) : plusieurs SNVs et/ou INDELS dans un bloc

SNV

DELetion

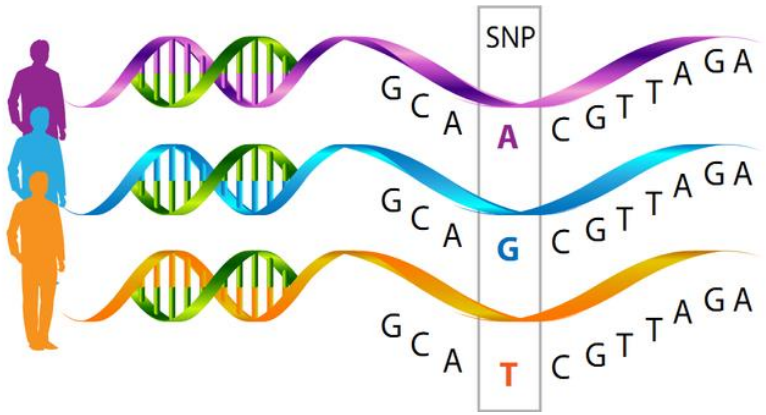
INsertion

Adapté F. morel

Quelques définitions....

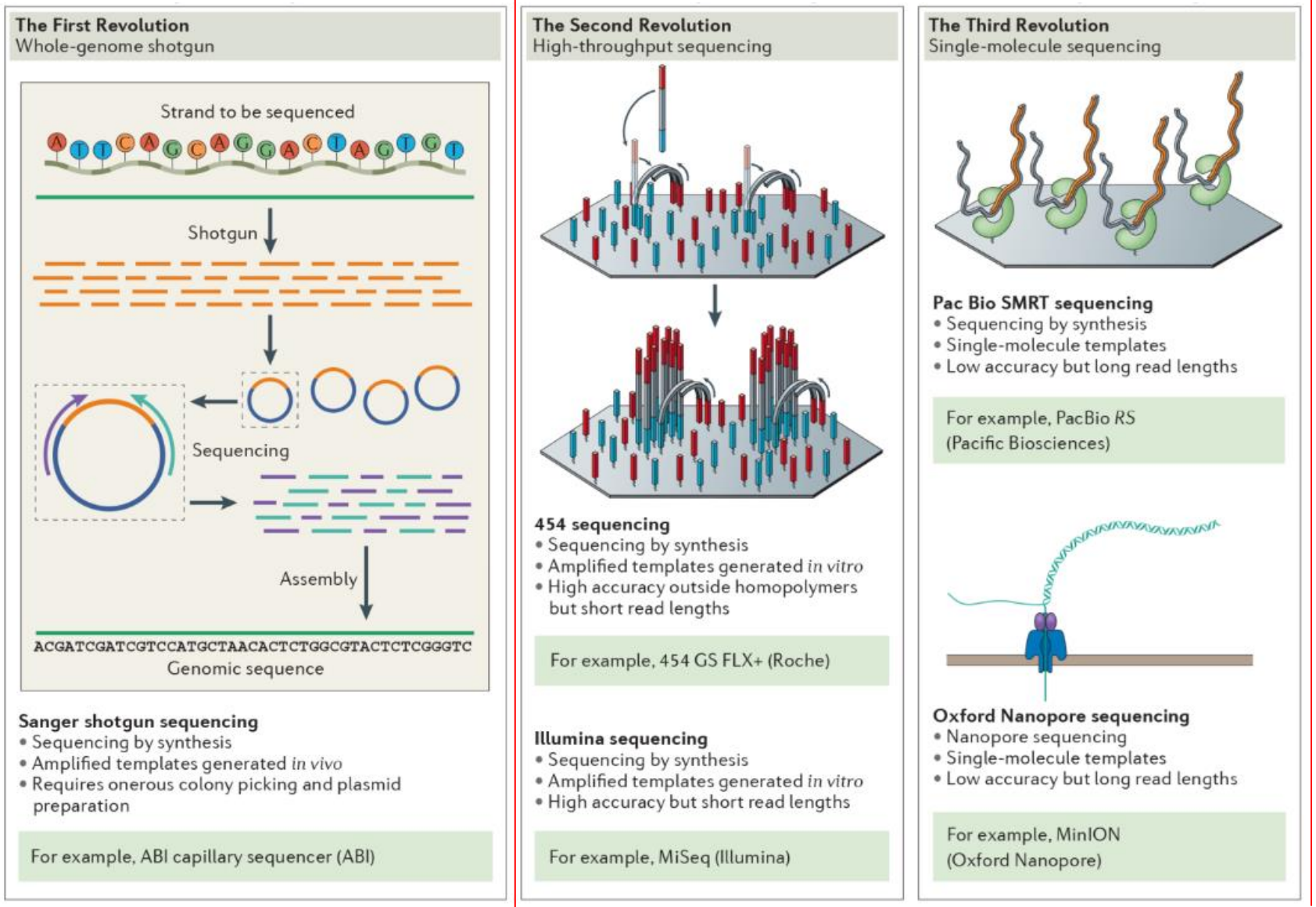
- Mutation : variation dans la séquence d'ADN (SNV, indel)
  - fréquence <1% dans une population
  - effet : gain de fonction ou perte de fonction
  - phénomène rare

- **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*



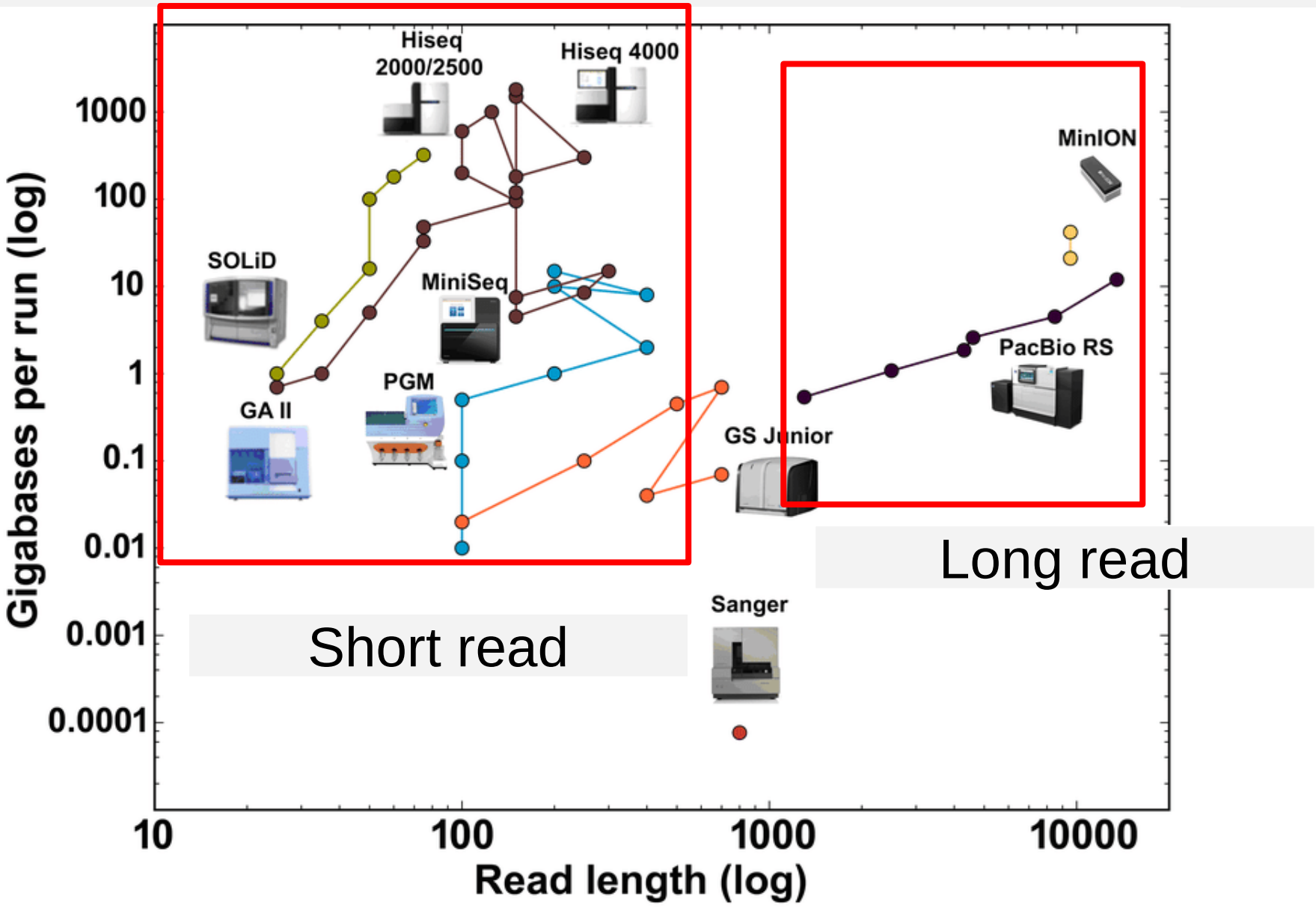
Les différentes technologies

NGS



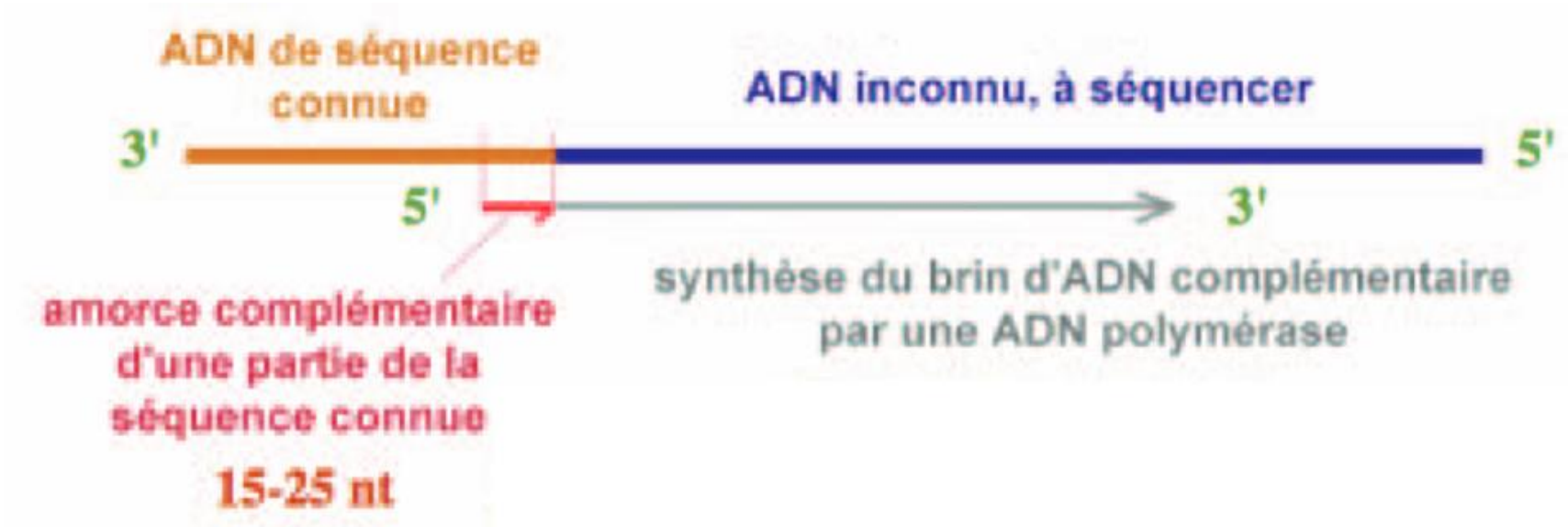
Nat Rev Microbiol. 2017 Dec;13(12):787-94

Les différentes technologies



Silva, Genivaldo. (2016). Who Is There and What are They Doing? An Agile and Computationally Efficient Framework for Genome Discovery and Annotation from Metagenomic Big Data.

Méthode sanger

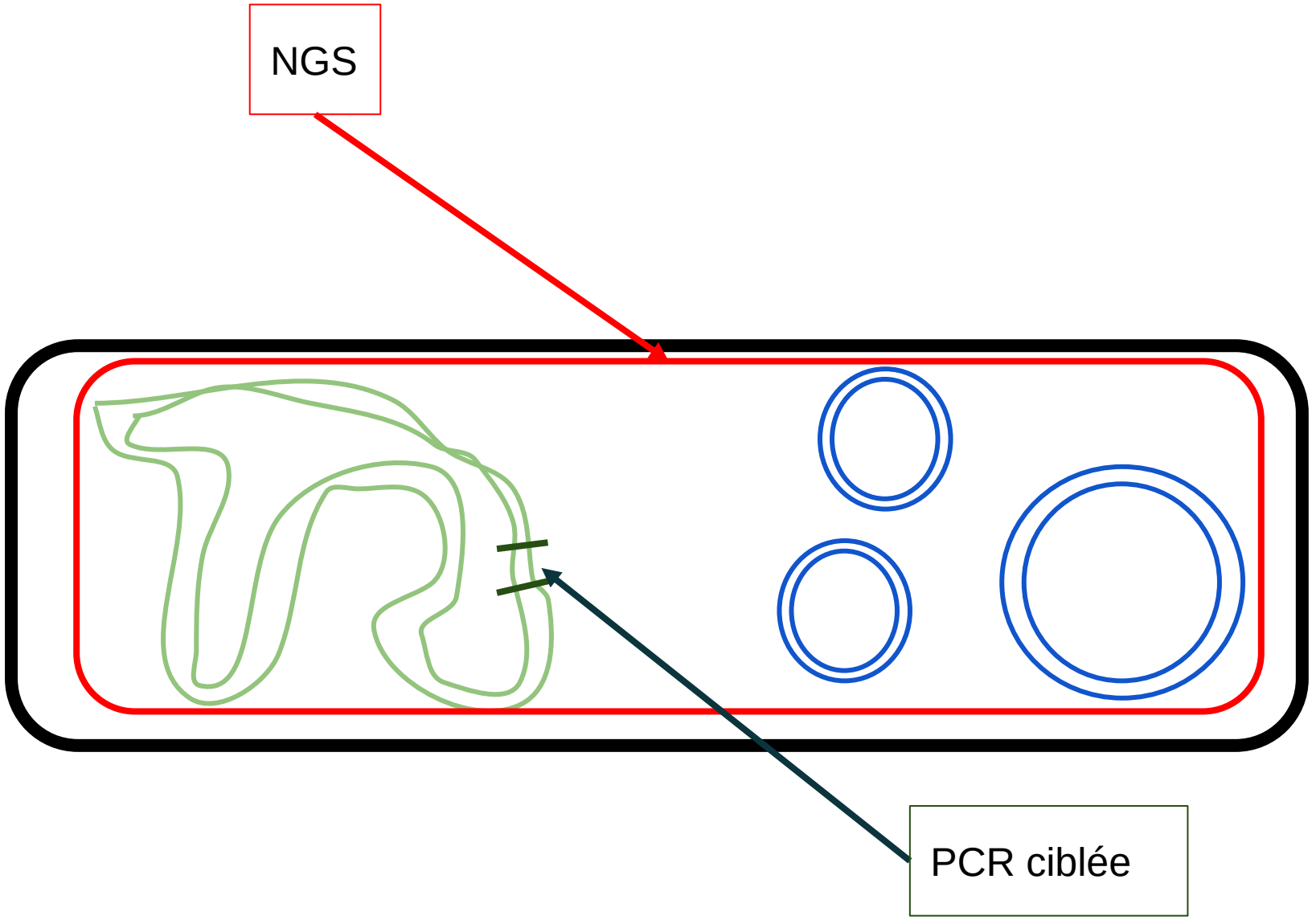


Préparer 4 solutions :

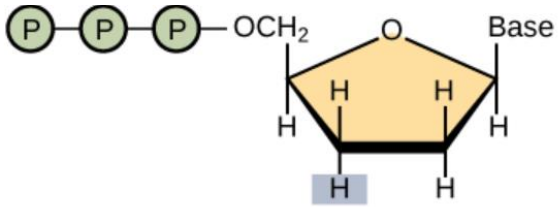
- le fragment qui doit être séquencé
- Amorce
- les 4 dNTP's (dCTP, dATP, dGTP, dTTP)
- l'ADN polymérase

https://facmed.univ-rennes1.fr/wkf//stock/RENNES20080328110058vdavidSEquencage.pdf

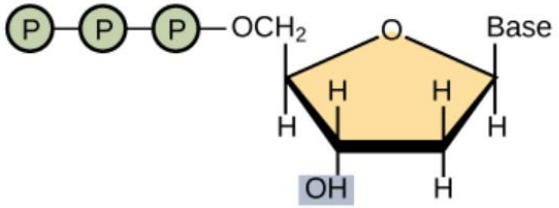
NGS *versus* PCR ciblée



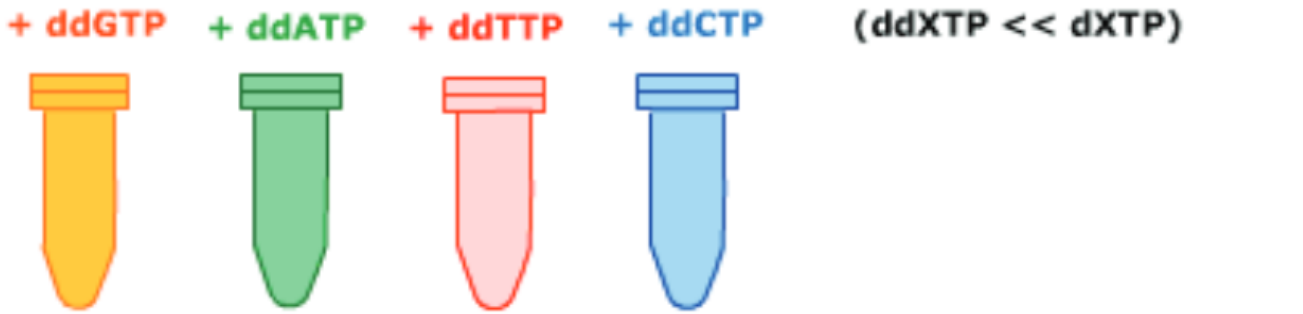
Méthode sanger



Dideoxynucleotide (ddNTP)



Deoxynucleotide (dNTP)



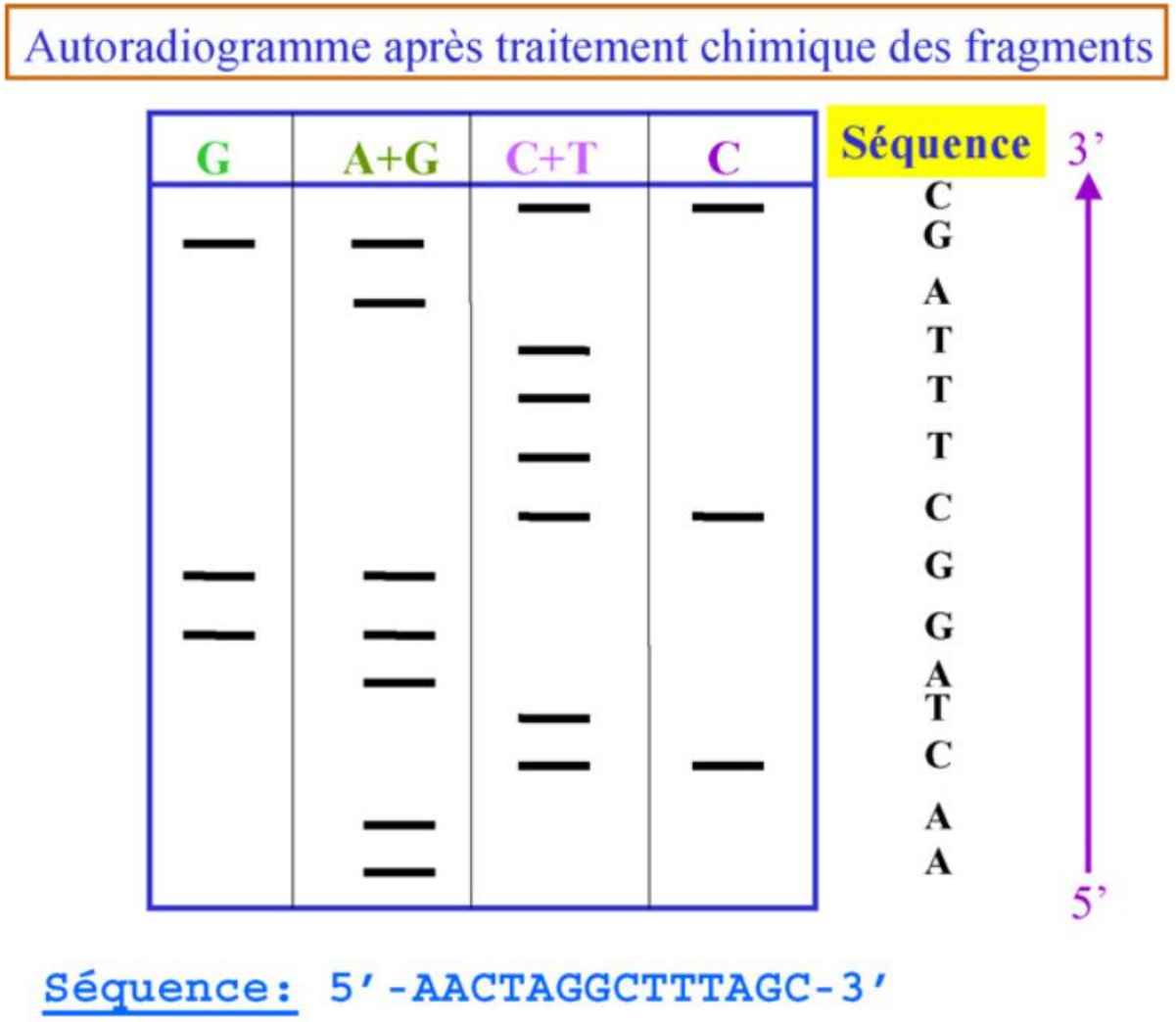
http://wwwarpe.snv.jussieu.fr/coursvt/images_10/sangerprinc.gif

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.

https://brainbrooder.com/lesson/476/17-3-1-strategies-used-in-sequencing-projects

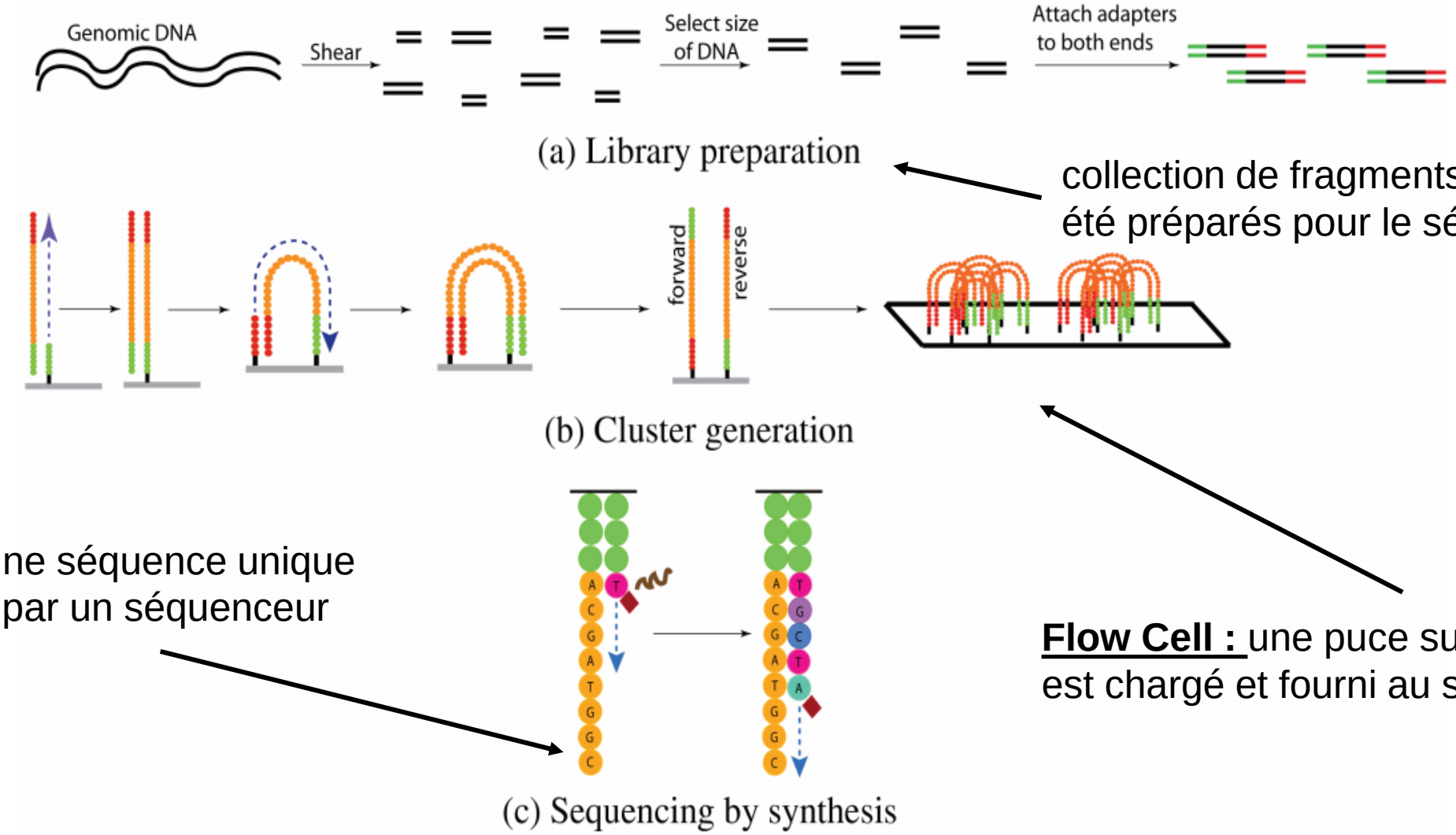
Méthode sanger



Technique de Maxam-Gilber

https://doi.org/10.1016/j.immbio.2008.07.016

Illumina sequencing



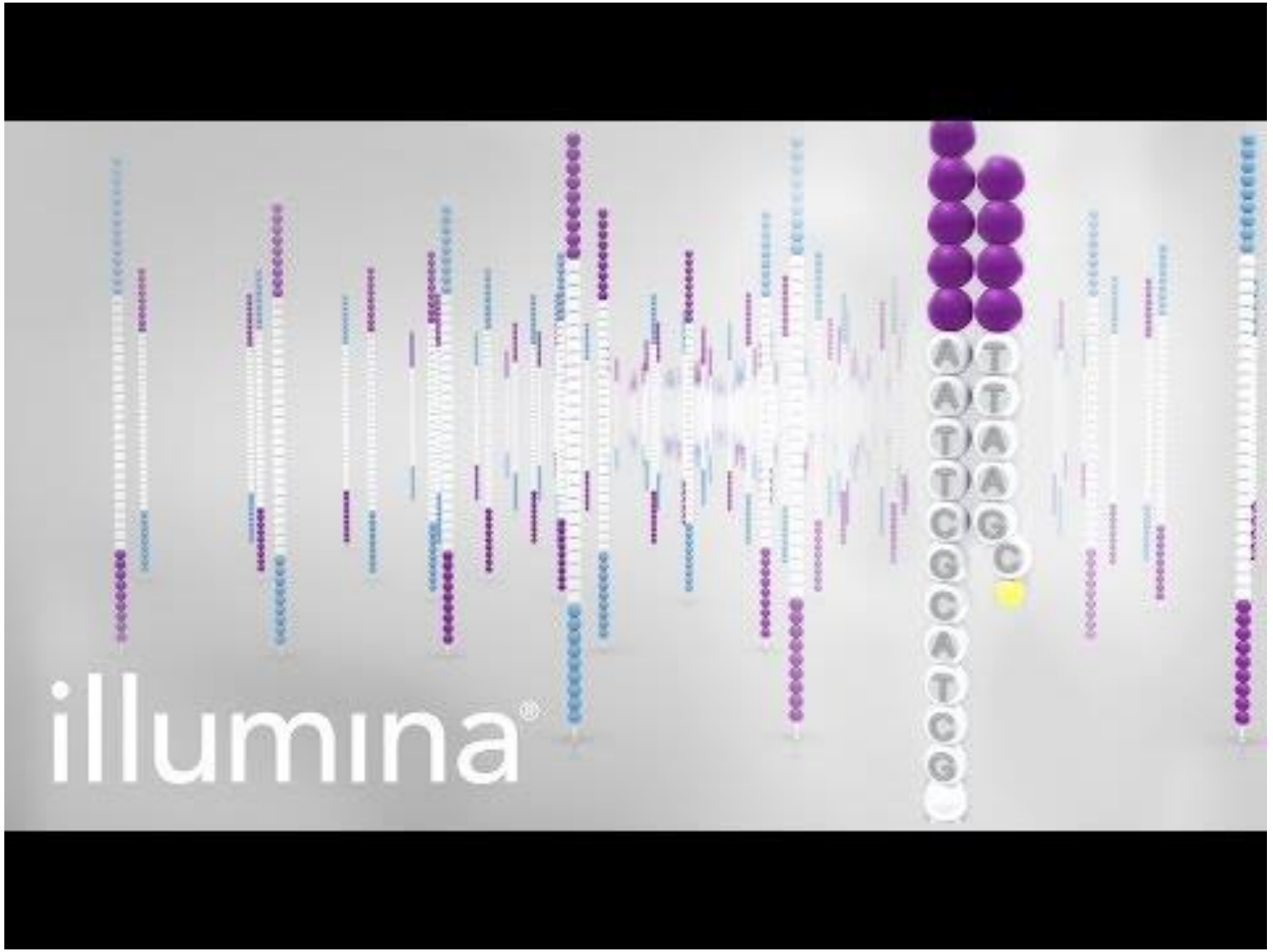
collection de fragments d'ADN qui ont été préparés pour le séquençage

Read : une séquence unique produite par un séquenceur

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

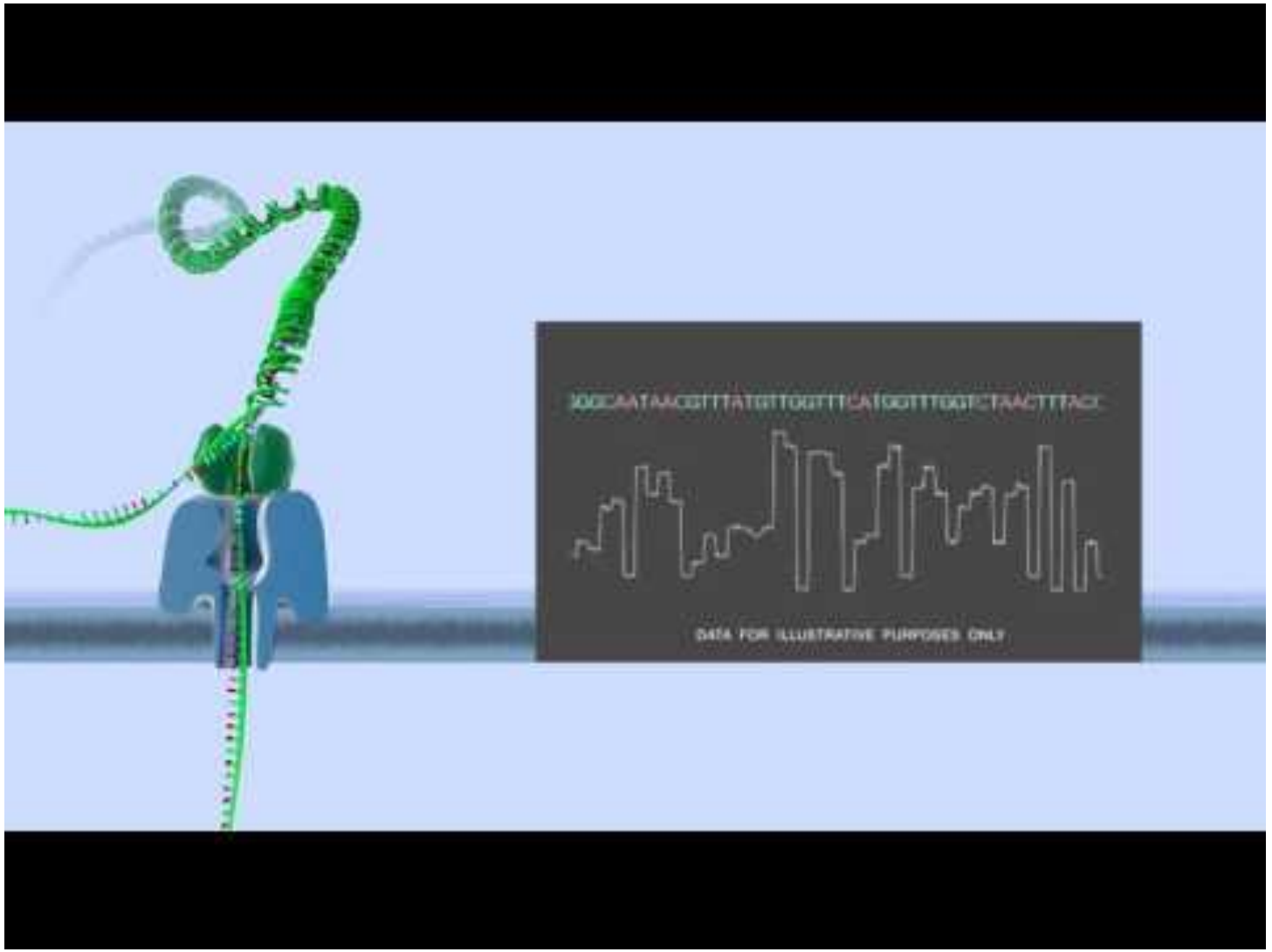
A Practical Handbook of Next Generation Sequencing and Its Applications
https://doi.org/10.1142/10159 | September 2017

Illumina sequencing

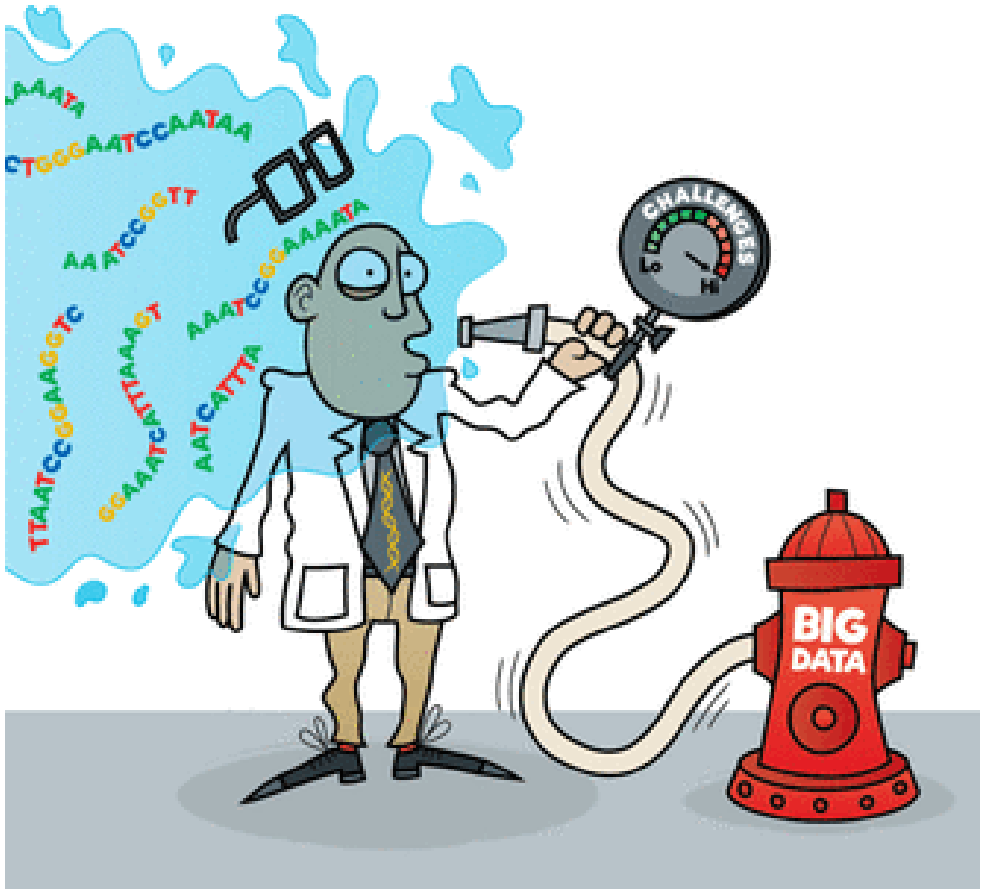


Minion sequencing

- **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



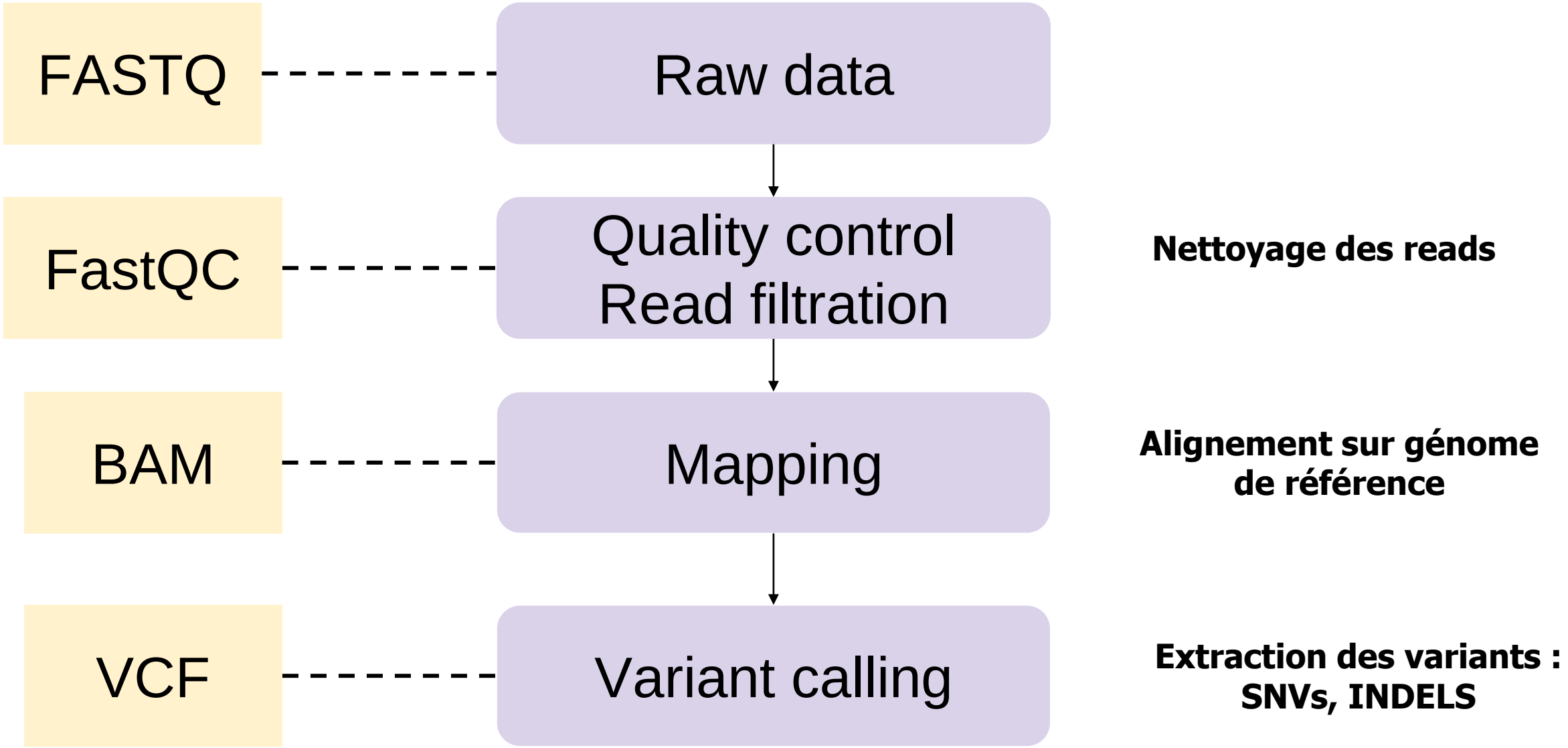

```
#!/bin/bash
```

The challenges of big data

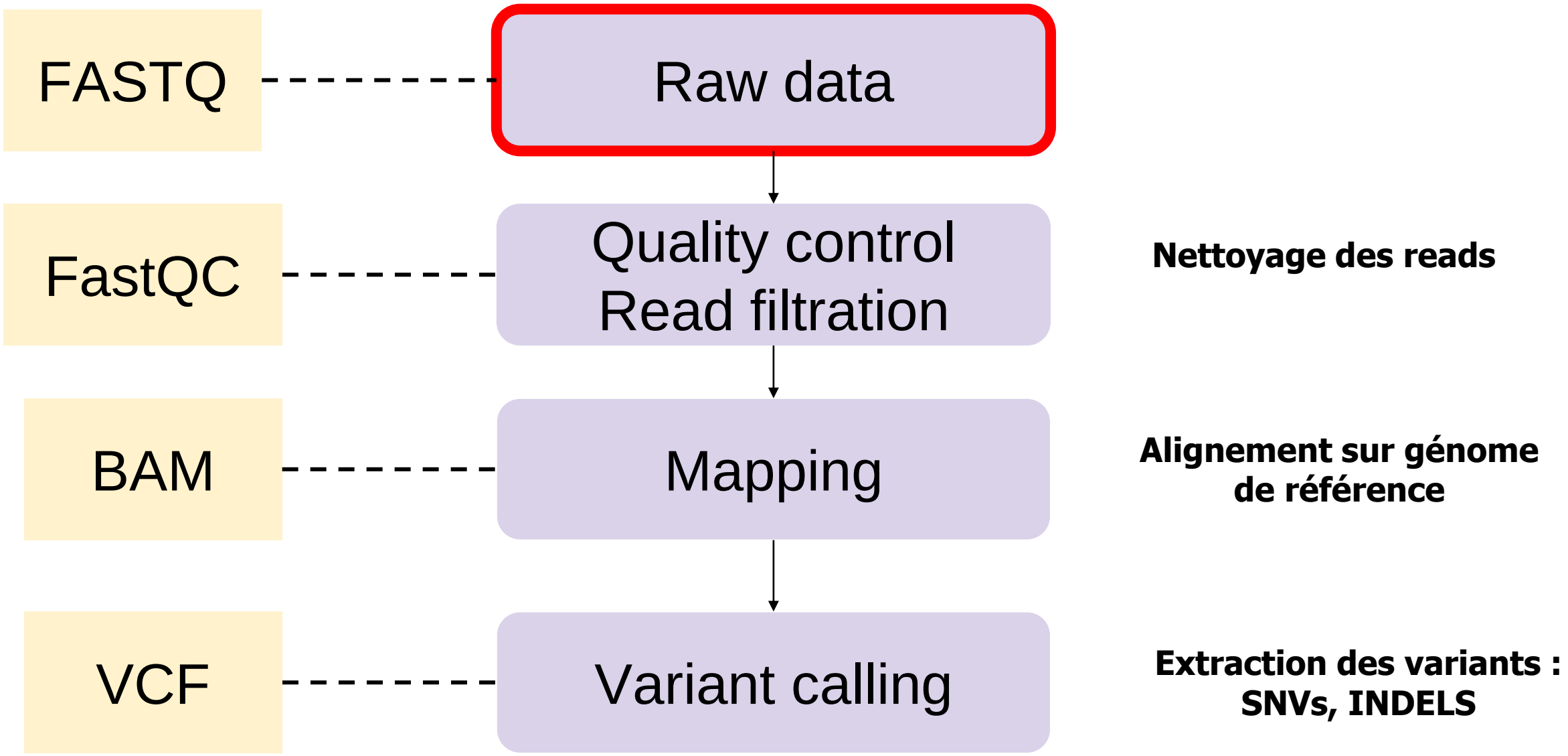
Elaine R. Mardis

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

Workflow d’analyse



Workflow d'analyse



FASTQ synthax

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  3IIIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
5  @ERR000589.42 EAS139_45:5:1:2:1293/1
6  AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
7  +
8  IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

Abuín Mosquera, José & Pichel, Juan & Pena, Tomas & Amigo, Jorge. (2016). SparkBWA: Speeding up the alignment of high-Throughput DNA sequencing data. PLOS ONE. 11. e0155461. 10.1371/journal.pone.0155461.

FASTQ synthax

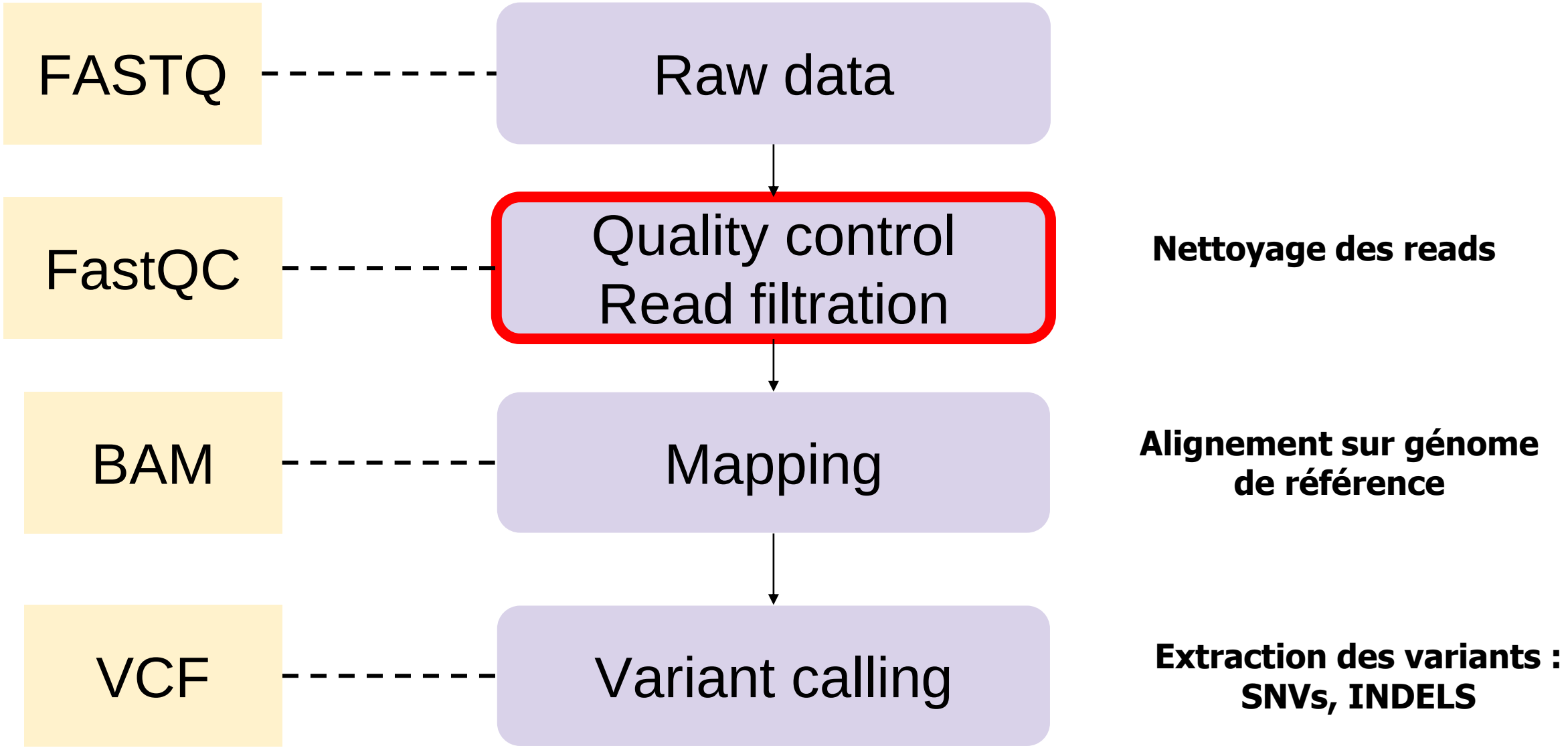
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ....................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................41
  0.2......................26...31.......41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

- Quality scores are typically represented using a Phred scoring scheme, where a read quality value = −10 * log10 (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.

http://barcwiki.wi.mit.edu/wiki/SOPs/qc_shortReads

Workflow d'analyse



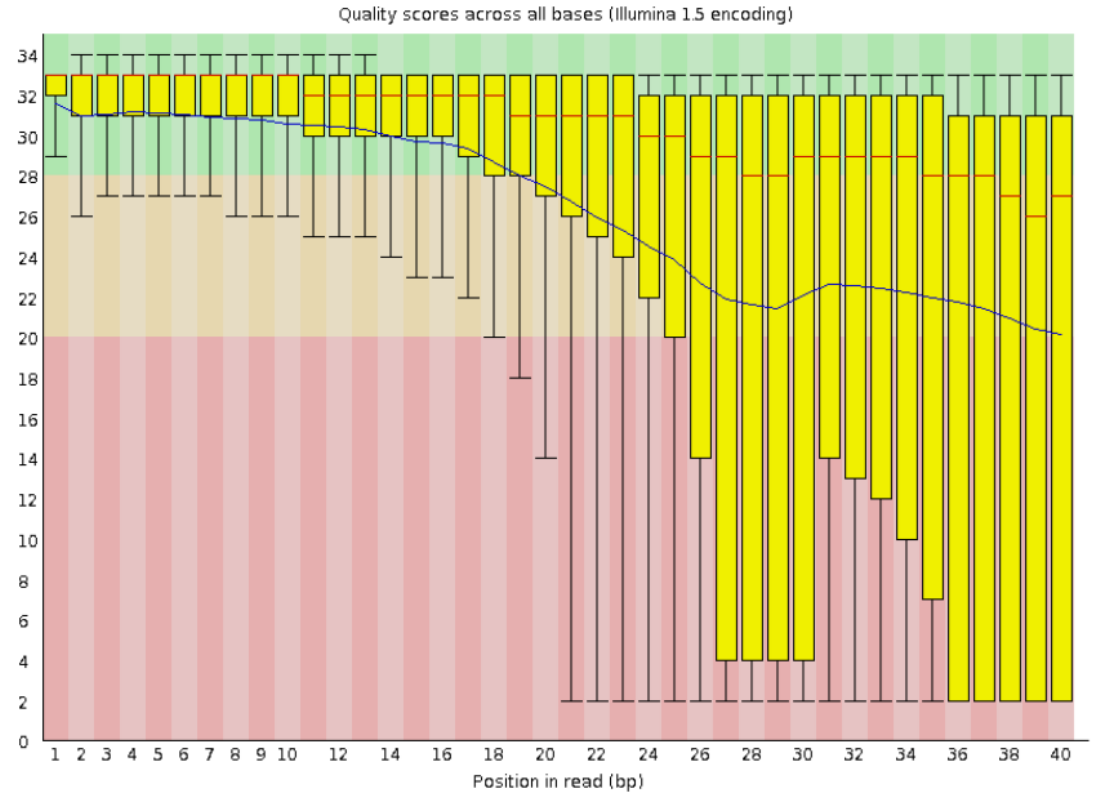
Quality control/ FastQC report

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



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

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%

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Quality control/ Trimming

Error correction depends on application…

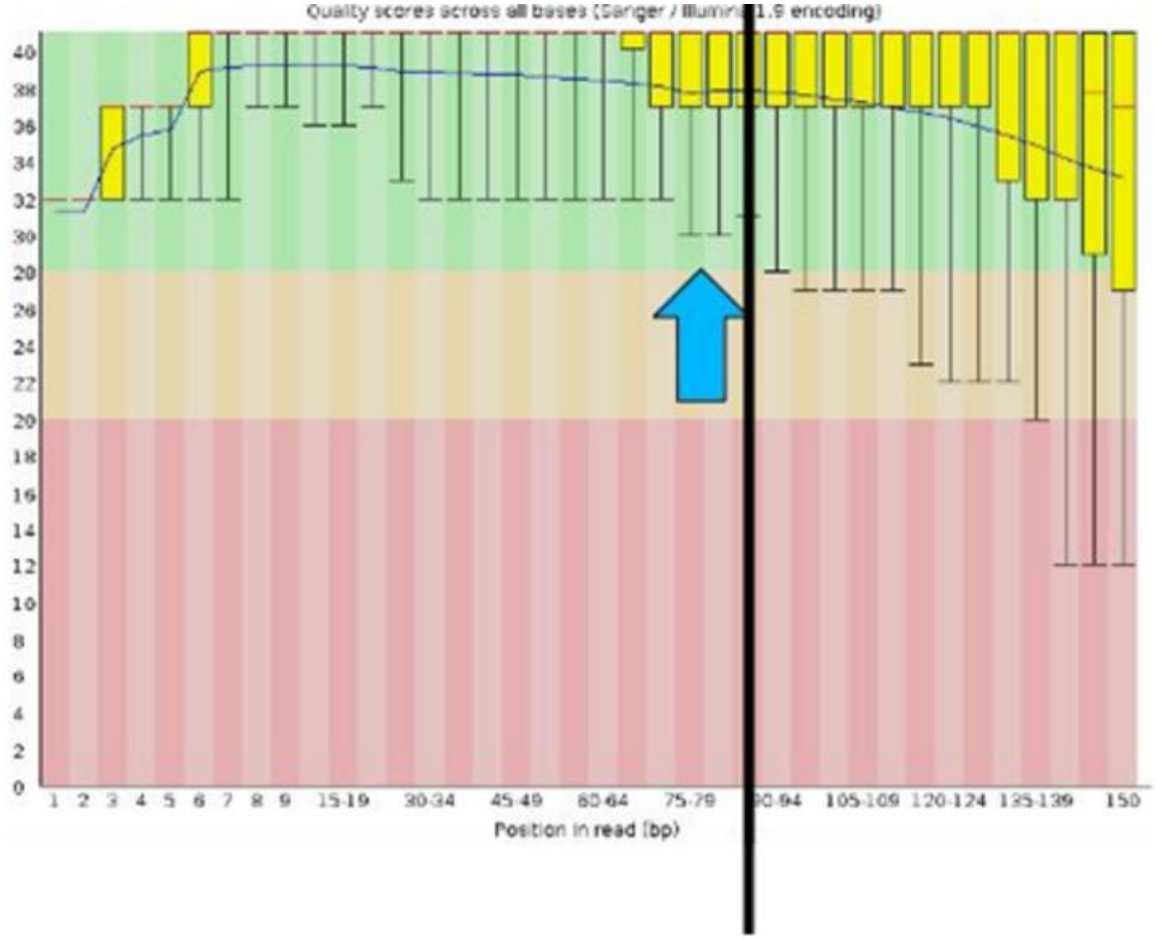
Ex. 1 SNP calling:

Phred 30: 0.1% P error

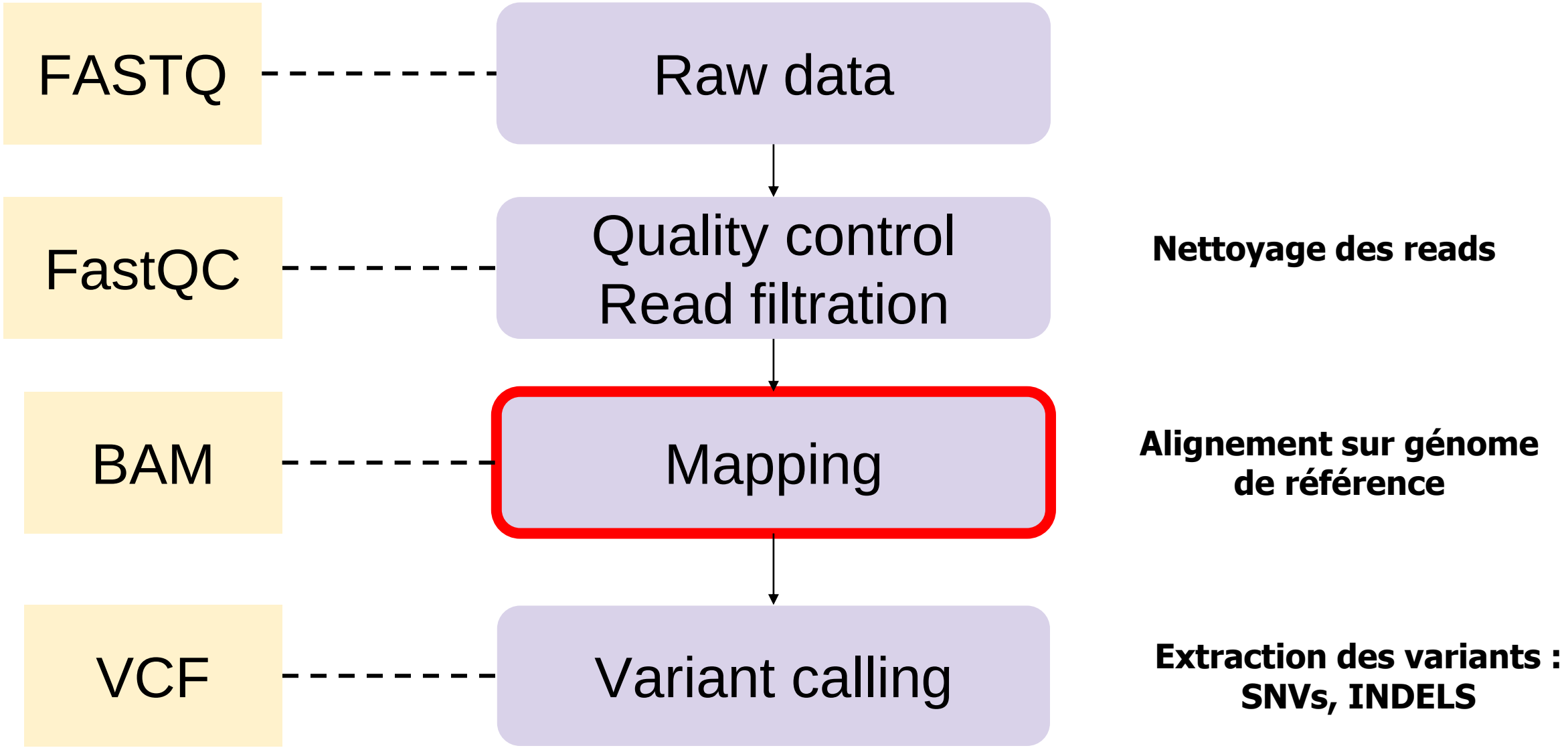
Quality-based trimming

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

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

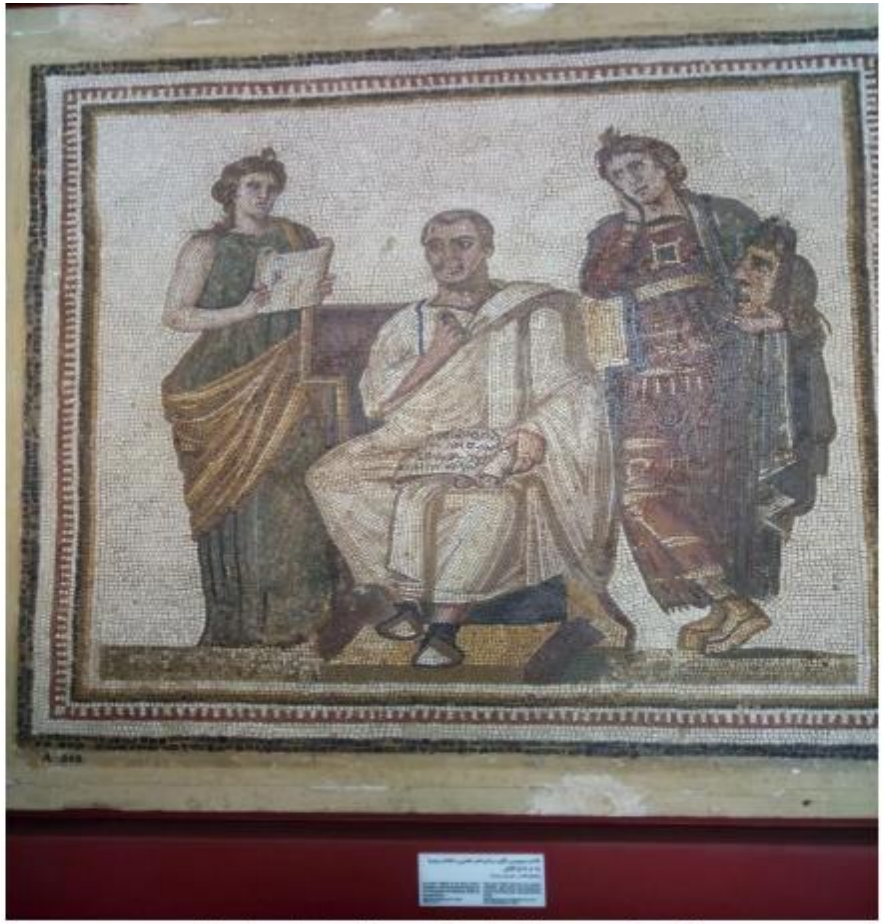


Workflow d'analyse



Mapping

Reads

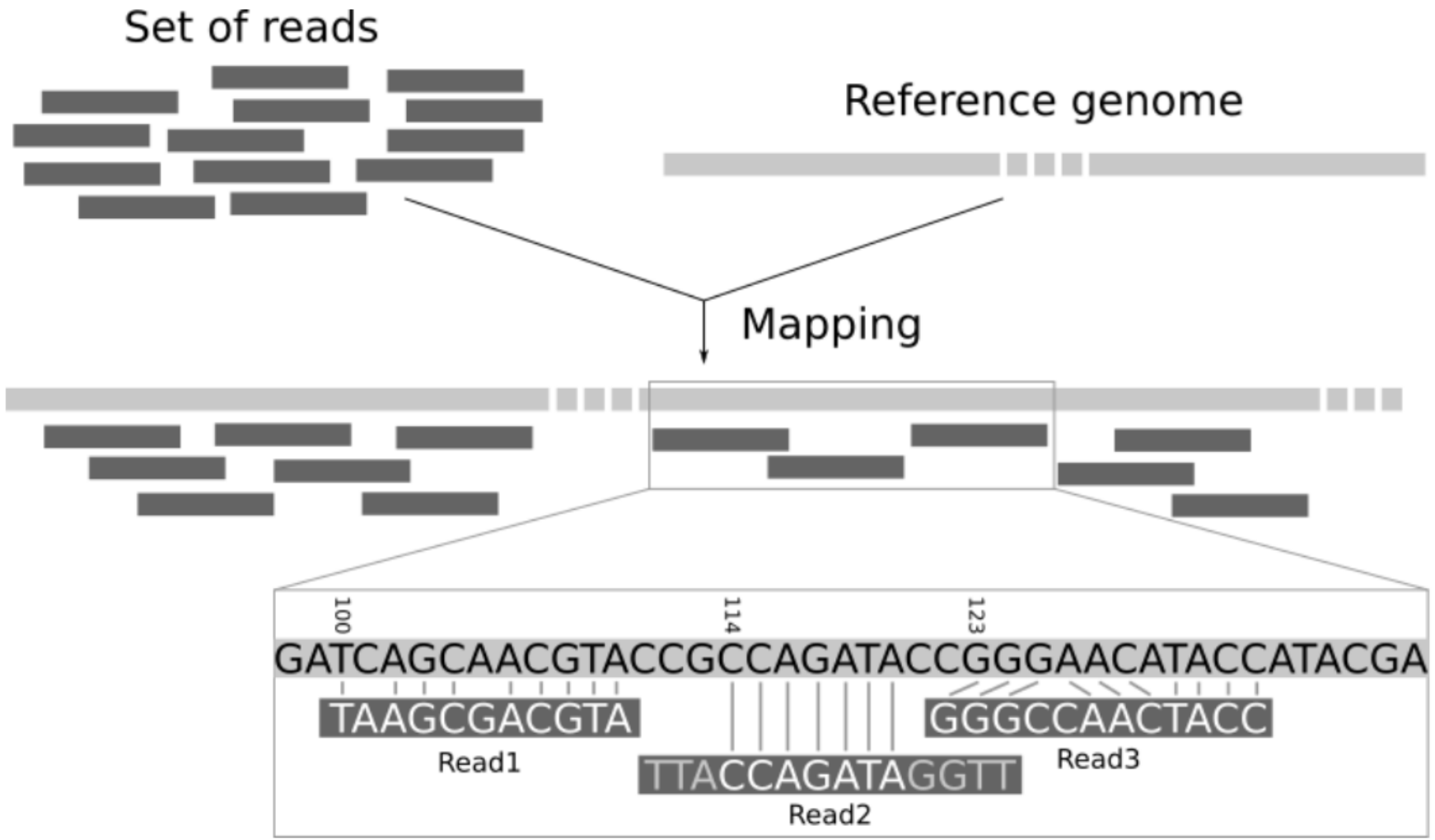


[Virgil and the Muses, Bardo Museum, Tunis]

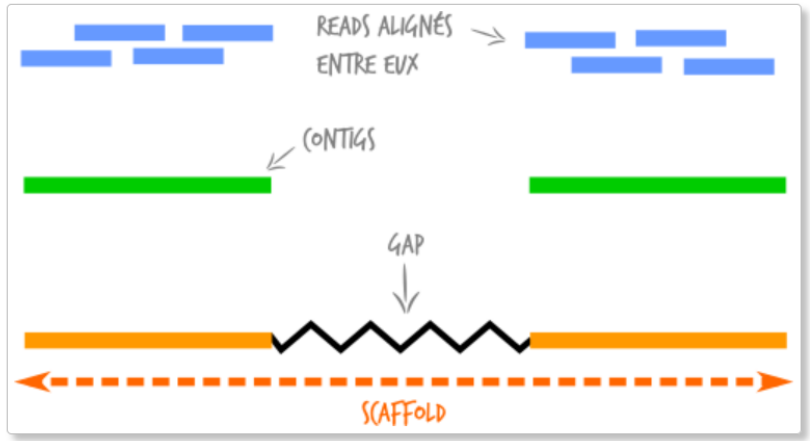
Génome assemblé

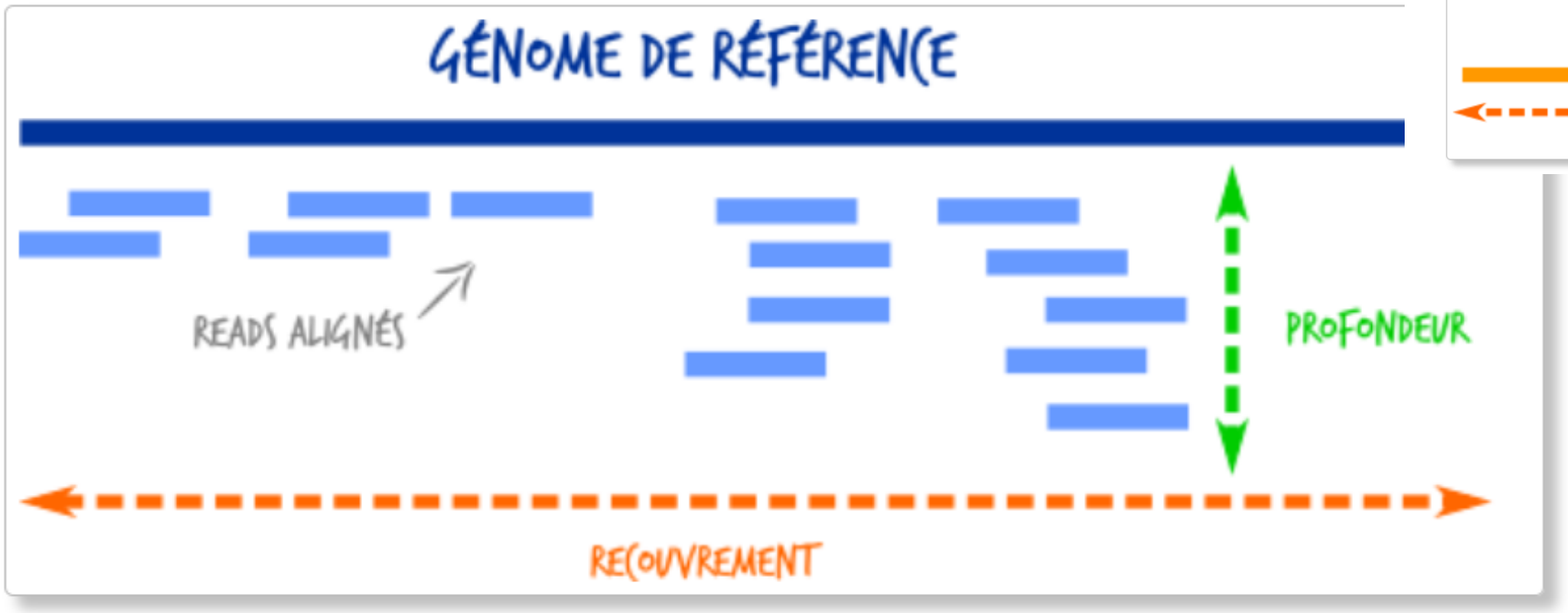
Fichiers SAM (Sequence Alignment/Map)/ Fichier BAM

Mapping


Set of reads
Reference genome
Mapping
100
114
123
GATCAGCAACGTACCGCCAGATACCGGGAACATACCATACGA
TAAGCGACGTA
Read1
TTACCAGATAGGTT
Read2
GGGCCAACTACC
Read3

Mapping/ Profondeur et couverture



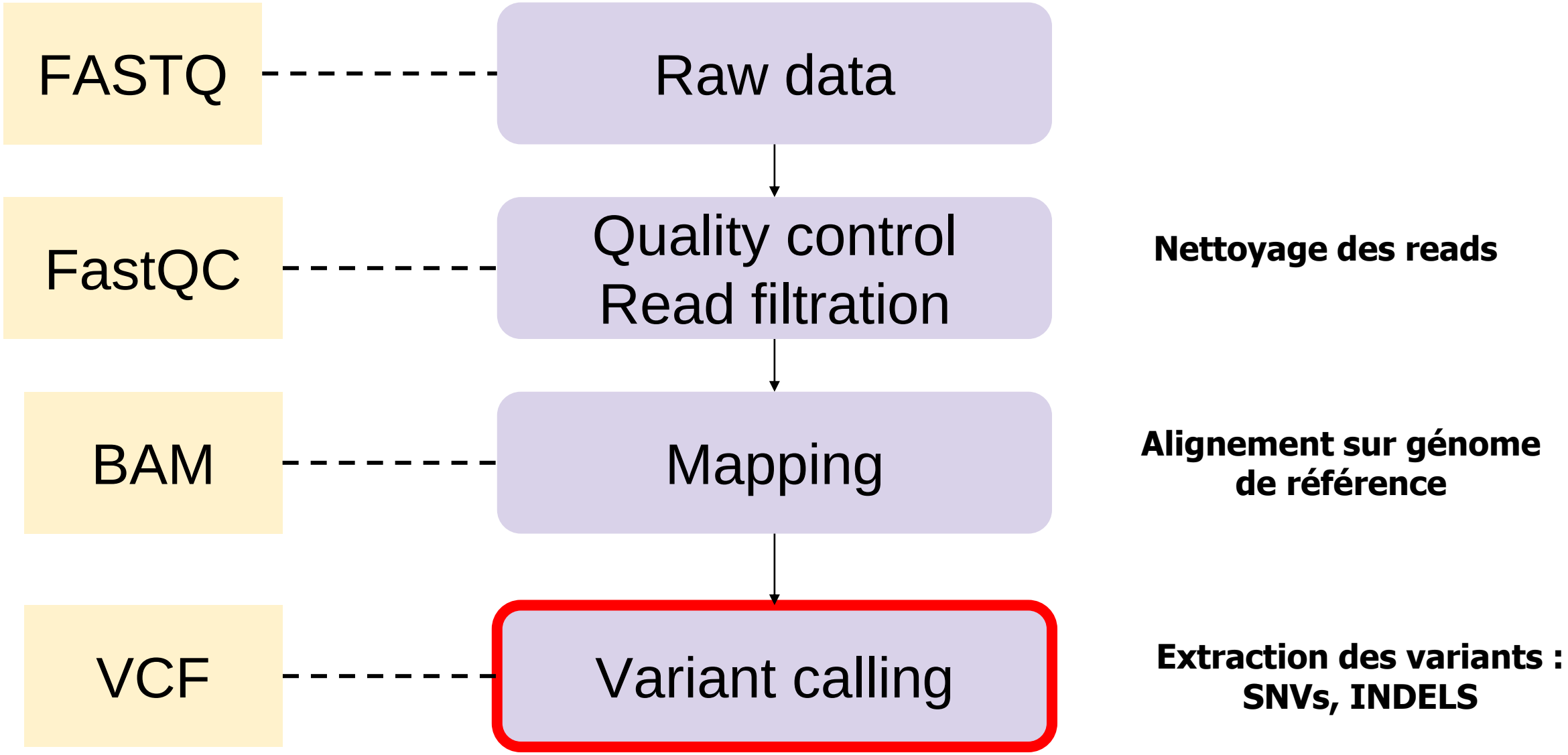


nombre de lecture de chaque base, exprimée en X

zone couverte par au moins une lecture, exprimée en %

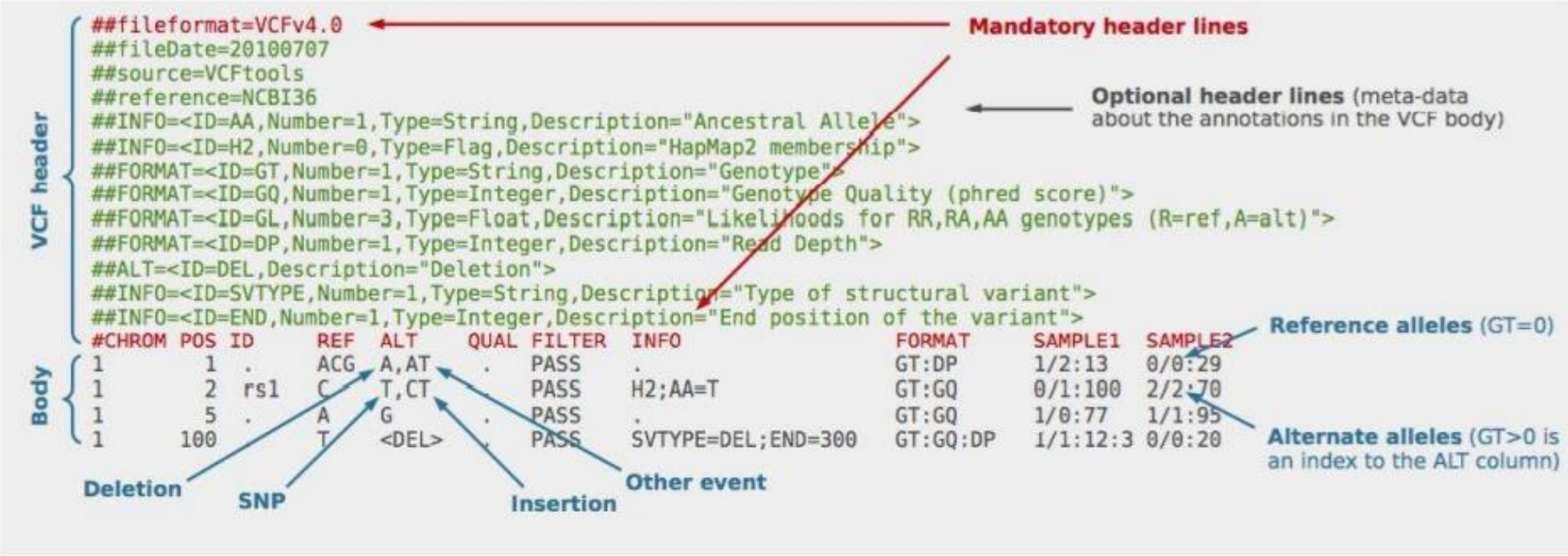
Le séquençage de nouvelle génération // Sacha Schutz // bioinformatique génétique médecine (dridk.me)

Workflow d'analyse



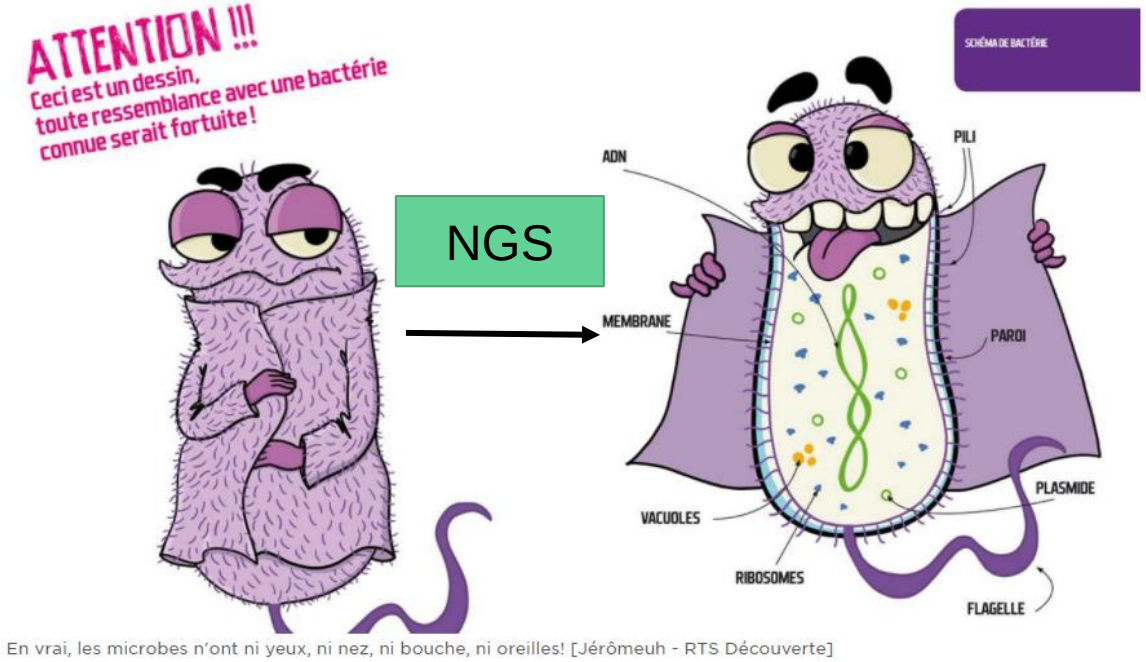
Questions

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



résultats dans fichier « VCF » = Variant Calling File

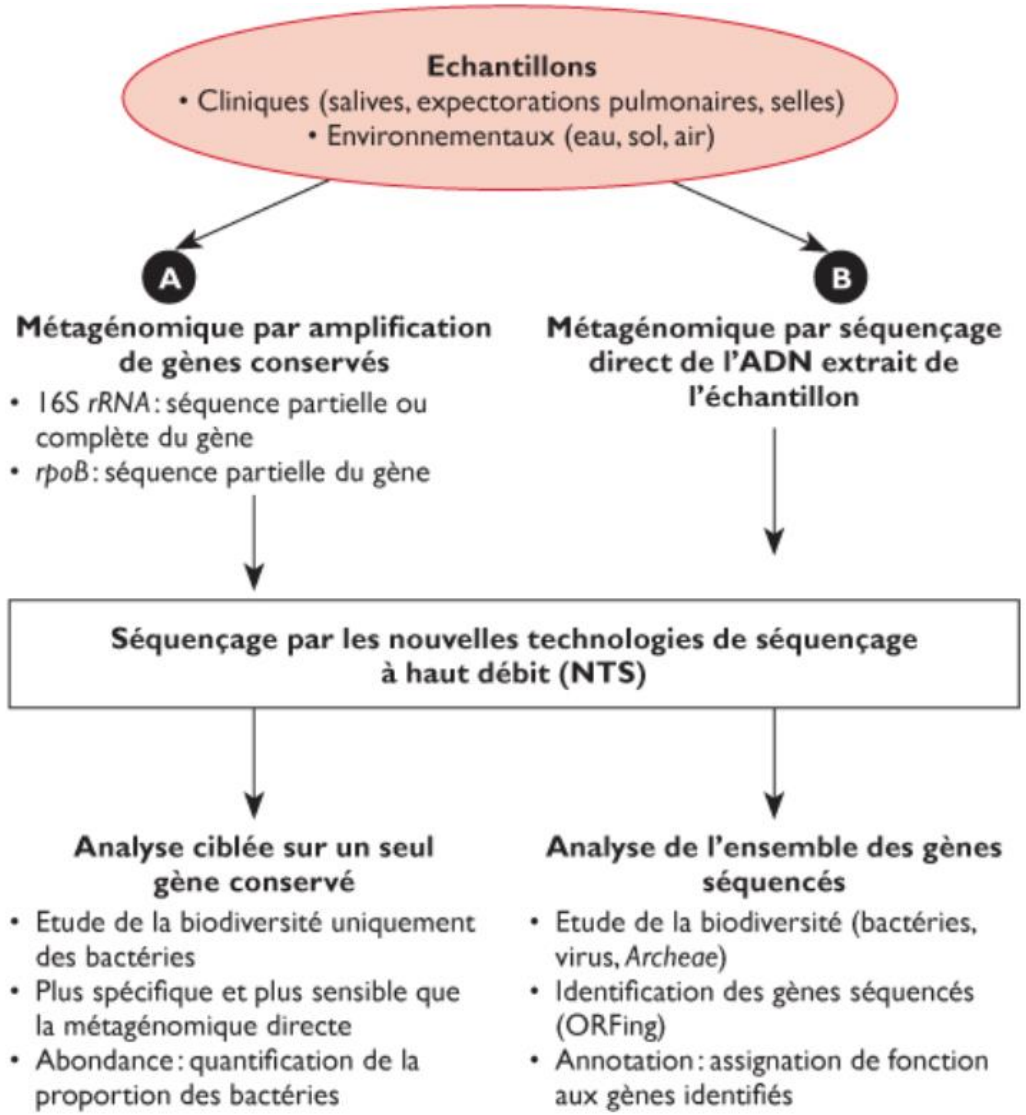
Application/ Métagénomique



NGS

En vrai, les microbes n'ont ni yeux, ni nez, ni bouche, ni oreilles! [Jérômeuh - RTS Découverte]

Certaines bactéries sont non cultivables

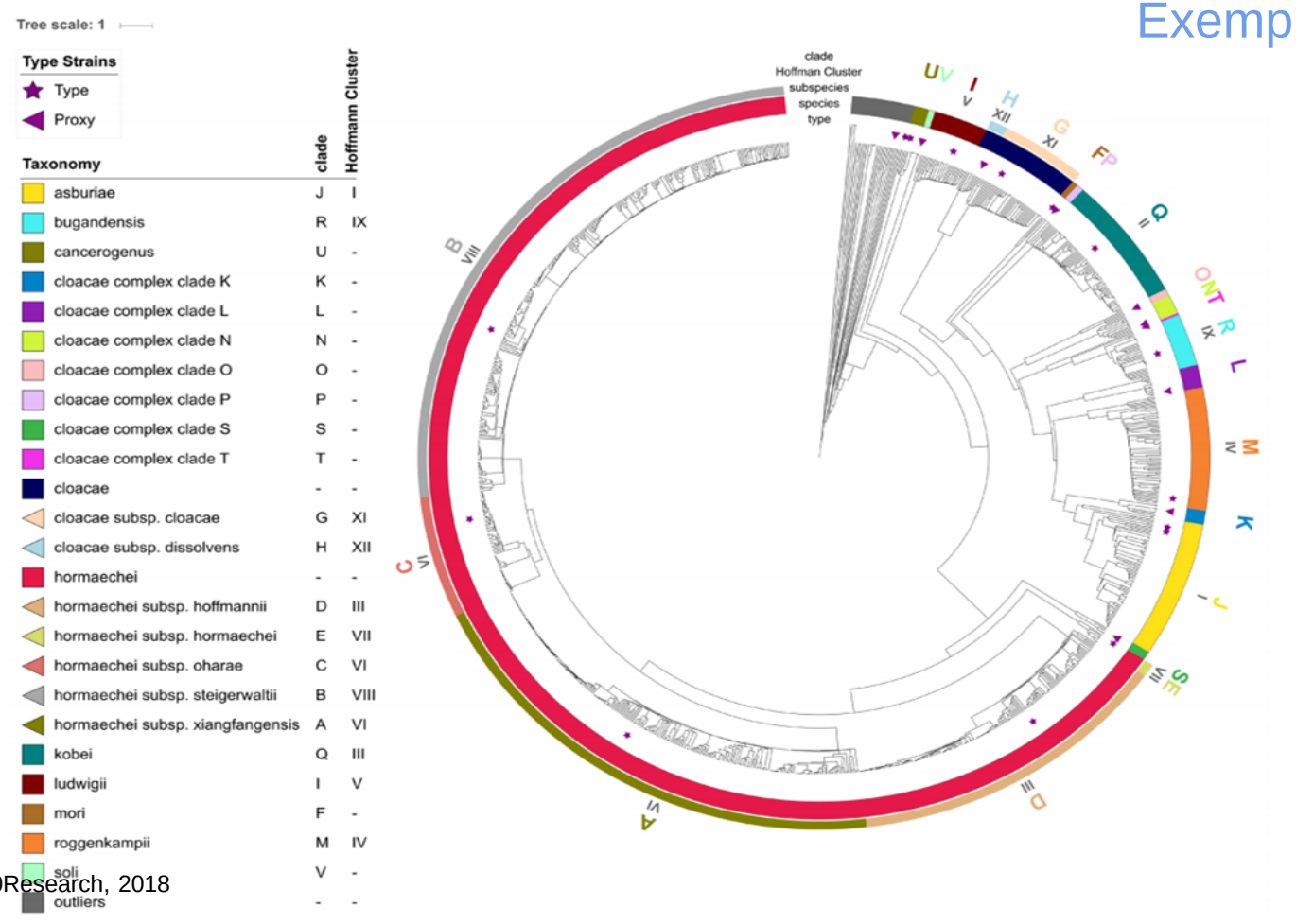


Rev Med Suisse 2014; volume 10. 2155-2161

https://www.rts.ch/decouverte/sante-et-medecine/corps-humain/microbes/4658406-les-microbes-expliques-aux-enfants.html

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 (n)	Mutation (n)	Total # isolates	Sensitivity % (CI)	specificity % (CI)	PPV % (CI)	NPV % (CI)
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)	

Connie Lam, Elena Martinez, Taryn Crighton, Catriona Furlong, Ellen Donnan, Ben J. Marais, Vitali Sintchenko,Value of routine whole genome sequencing for Mycobacterium tuberculosis drug resistance detection,International Journal of Infectious Diseases, 2021, https://doi.org/10.1016/j.ijid.2021.0

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 !