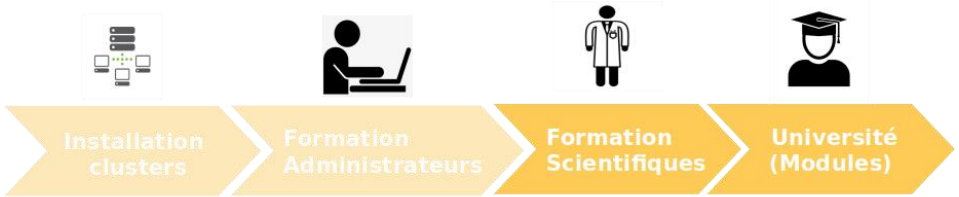
# Projet Burkinabioinfo

## Overview - Actions de formation




2018

2019

48h, 24 personnes

2 x 32h, 25 personnes

Programmes formation et organisation

# Programme & Logistique

| | Débutant | Veteran SNP / diversite plantes | Veteran Metagenomique | | COMMUN |
|---|---|---|---|---|---|
| LUNDI | | | | | COURS |
| 8h30-10h00 | Accueil + Presentation formation / etudiants babies / etudiants veterans | | | | Autonomie |
| 10h30-12h00 | Cours NGS | Description jeu de données | Description jeu de données | | |
| | Pause déjeuner | | | | |
| 14h00-15h30 | Cours | Autonomie / Bibliographie / Veille technologique | | | |
| 16h-17h30 | Contrôle qualité + mapping | | | | |
| | | | | | |
| MARDI | | | | | |
| 8h-10h00 | Autonomie | Accompagnement mini-projets | Accompagnement mini-projets | | |
| 10h30-12h | Contrôle qualité + mapping | | | | |
| | Pause déjeuner | | | | |
| 14h00-15h30 | Cours | Autonomie | Autonomie | | |
| 16h-17h30 | mapping + SNP calling | | | | |
| | | | | | |
| MERCREDI | | | | | |
| 9h-10h30 | Autonomie | Accompagnement mini-projets | Accompagnement mini-projets | | |
| 10h30-12h | mapping + SNP calling | | | | |
| | Pause déjeuner | | | | |
| 14h00-15h30 | Cours | Autonomie | Autonomie | | |
| 16h-17h30 | Analyse diversité | | | | |
| | | | | | |
| JEUDI | | | | | |
| 9h-10h30 | Autonomie | Accompagnement mini-projets | Accompagnement mini-projets | | |
| 10h30-12h | Analyse diversité | | | | |
| | Pause déjeuner | | | | |
| 14h00-15h30 | Cours | Autonomie | Autonomie | | |
| 16h-17h30 | Génomique comparative | | | | |
| | | | | | |
| VENDREDI | | | | | |
| 9h-10h30 | Restitution des mini-projets Veterans - Questions et discussions diverses | | | | |
| 10h30-12h | Questions et discussions autour des projets/données des participants | | | | |
| | Pause déjeuner | | | | |
| 14h00-15h30 | Questions et discussions autour des projets/données des participants | | | | |
| 16h-17h30 | | | | | |

*KIENDREBEOGO Touwendpoulimdé Isabelle*

**Génétique,Bm, Cancer du sein, Mutation, BRCA**

*AHONON Awovi Selom*

**S. rotundifolius, S. stenocarpa, plantes mineures, diversité**

*GBEKLEY Efui Holaly*

*TUINA Sévérin*

**Flux de gènes, Dynamique de la diversité, Sorghum bicolor,**

*DOSSIM Sika*

*BA Aminata Hamidou*

**Diversité S. rotundifolius: Profil morphologique/ génétique des morphotypes cultivés ( BF, Ghana)**

*TONDE Ignace*

**Solenostemon rotundifolius, interactions génotype x environnement , profil génitique,**

*SIRIMA Constant*

**RNA Seq, Plasmodium falciparum ACT-sensible/ ACT-résistant**

*BADOUM Emilie Salimata*

*ZOURE Abdou Azaque*

**Microbiome instestianal du moustique (Illumina) , Gène BRCA (Sanger)**

*ADAMOU IBRAHIM Maman Laouali*

**Analyse de la distribution génétiques et des régions liées au sexe des palmiers du Sahel**

## 13 Apprenants

*SANOU Estèle Pélagie*

*PALANGA Essowè*

**Metagenomique, virus, interaction plante-parasite, phytopathologie**

- tilapia du Nil, déterminisme du sexe, contrôle du sexe,

OUEDRAOGO Jacques

DANOU-KODJO Kodjovi Atassé

**Métagénomique-Variabilité génétique-Phytopathologie**

SAGNON Adama

**phosphate, solubilisation, bactéries, champignons**

SORY Siedou

**Diversité génétique/biochimique des cultivars d'ignames cultivées au Burkina Faso.**

NAME Pakyendou Estel

**Epidémiologie; Virus; ADN; CRESS; Séquençage**

ZONGO Saïdou

**Oxford Nanopore Technologie, Séquençage, Geminivirus, longs reads, NGS**

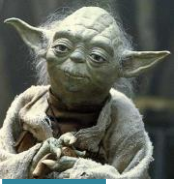
LALLOGO P. Doriane Tatiana

**SARS-CoV 2, facteurs génétiques, clairance, l'hôte humain, formes sévères.**

SAWADOGO Seydou

**Surveillance participative, maladies virales, racines et tubercules, séquençage**

# 8 Apprenants



Dominique

# Mais aussi…

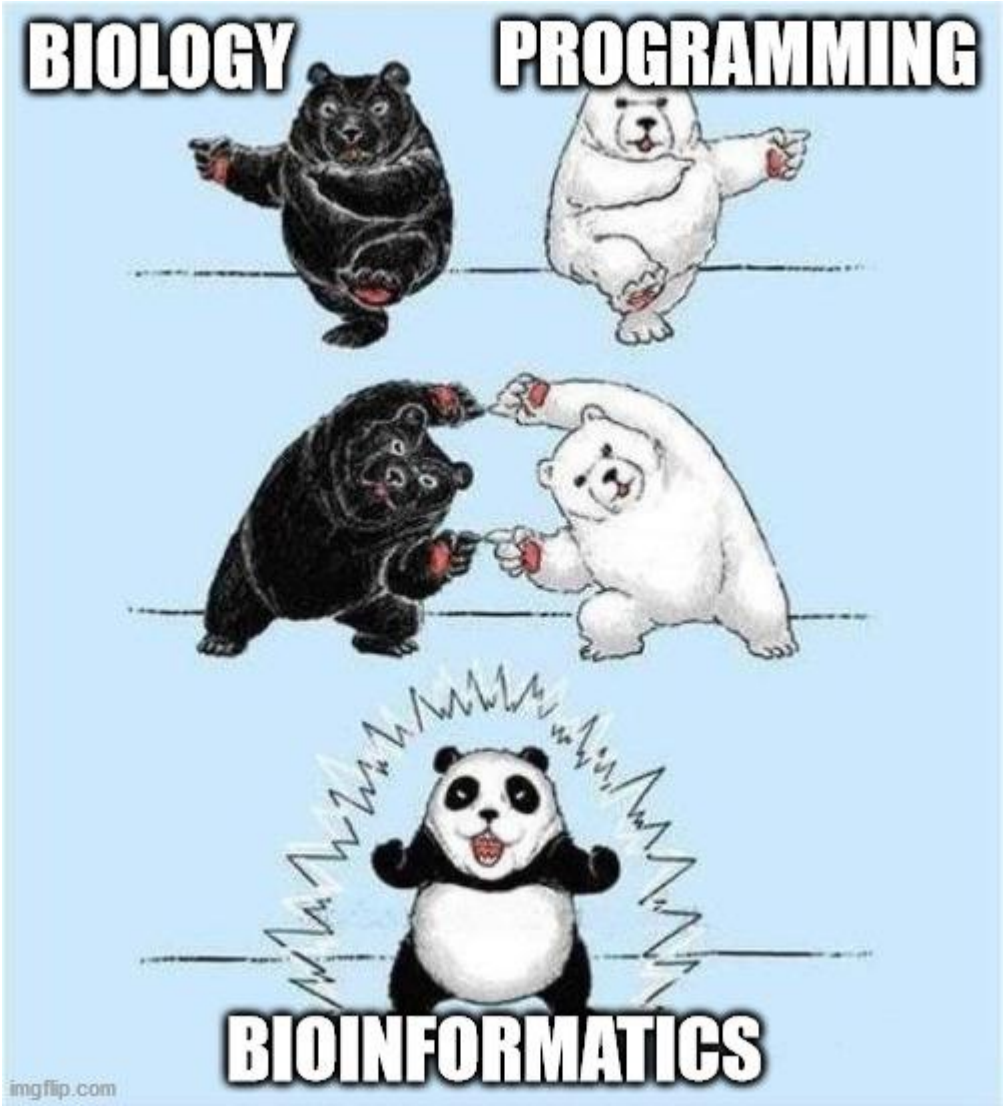
Admins : Seydou, Ousmanne Ndomassi

Fidele, romaric, Isidore
Le comité d organisation … ?
et bien d autres encore…. Neyra?

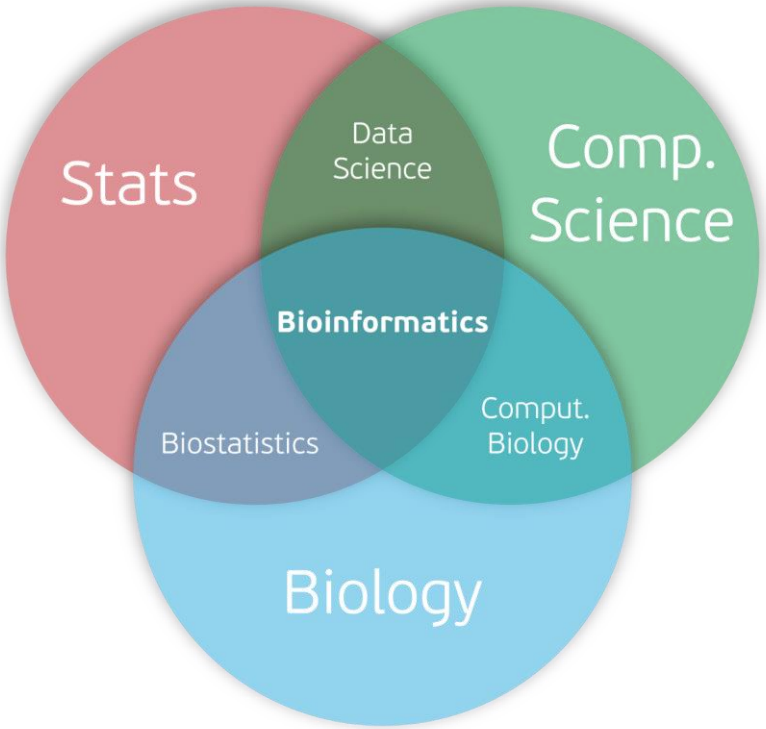


# Introduction Bioinformatics & Sequencing

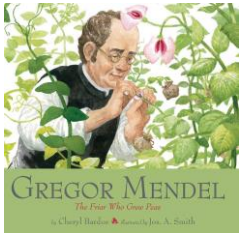
# What is the bioinformatics ?




## A interdisciplanary science



# De la génétique à la bioinformatique...




**1866**

**Lois de l'hérédité**

**1944**

**Nature chimique de l'ADN, matériel héréditaire**

***O. Avery, C. McLeod & McCarthy***


**1954**

**Structure en double hélice de l'ADN**

***J. Watson & F. Cricks & franklin***

**1961**

**Code génétique et règle de correspondance gènes-protéines**

***M. Nirenberg & H. Matthaei***

**1965**

**Mécanismes de la régulation génétique**

***André Lwoff, F. Jacob & J. Monod***

# De la génétique à la bioinformatique...

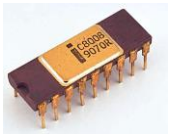
**1970**

**Algo *Alignement global de séquence***

***Needman, & Wunsh***

**1972**
**8008**

**1er microprocesseur intel**

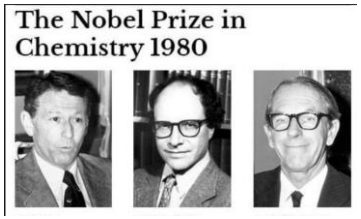


**1977**

**Micro-ordinateurs**

**Séquençage ADN**

***P. Berg, W. Gilbert & F. Sanger***




**1980**

**Banque EMBL, GenBank, PIR**

**1984**

**Amplification ADN - PCR**

***Karry Mullis***


**1985**

**Algo Alignement local de séquence FASTA**

***Person & Lipman***

**1987**

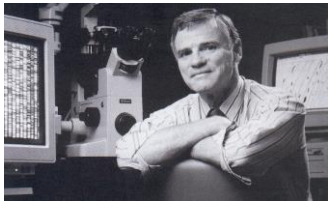
**1er séquenceur automaisé**

***L. Hood Société Applied Biosystems***

**1990**

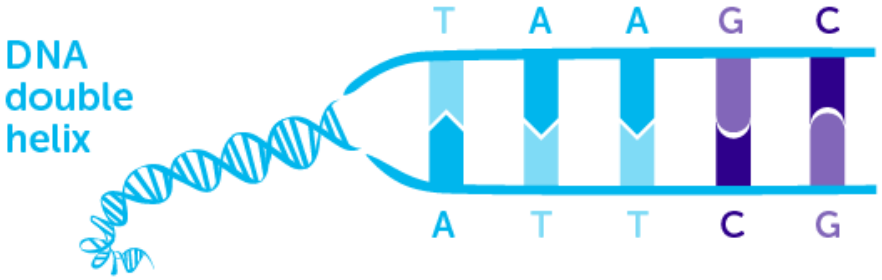
**Algo Alignement local de séquence BLAST**

***Altschul & al.***



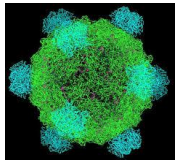
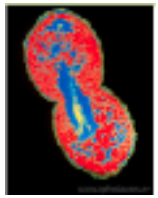
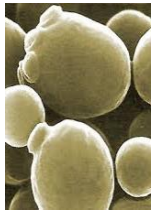
# A little history of sequencing...

*DNA sequencing : determining the order of the four bases or nucleotides that make up a given molecule of DNA



# A little history of sequencing...

Next-generation sequencing

| Virus phiX174 | H. influenza | S. cerevisiae | C. elegans | A. thaliana | H. sapiens | O. sativa |
|---|---|---|---|---|---|---|
| 1977 | 1995 | 1996 | 1998 | 2000 | 2001 | 2002 |
|  |  |  |   |  |   |  |
| 5kb<br>11 | 1,8 Mb<br>1,740 | 12 Mb<br>300 | 100 Mb<br>22,000 | 125 Mb<br>~30,000 | 3,3 Gb<br>~21,000 | 289 Mb<br>~35,000 |

# Several sequencing technology

SEQUENCING TECHNOLOGY

1 **FIRST GENERATION**
From 1977



sanger

2 **SECOND GENERATION**
From 2007



solexa, 454, illumina

3 **THIRD GENERATION**
From 2011


PacBio, oxford nanopore

# Several sequencing technology

SEQUENCING TECHNOLOGY

1 **FIRST GENERATION** From 1977

sanger

2 **SECOND GENERATION** From 2007

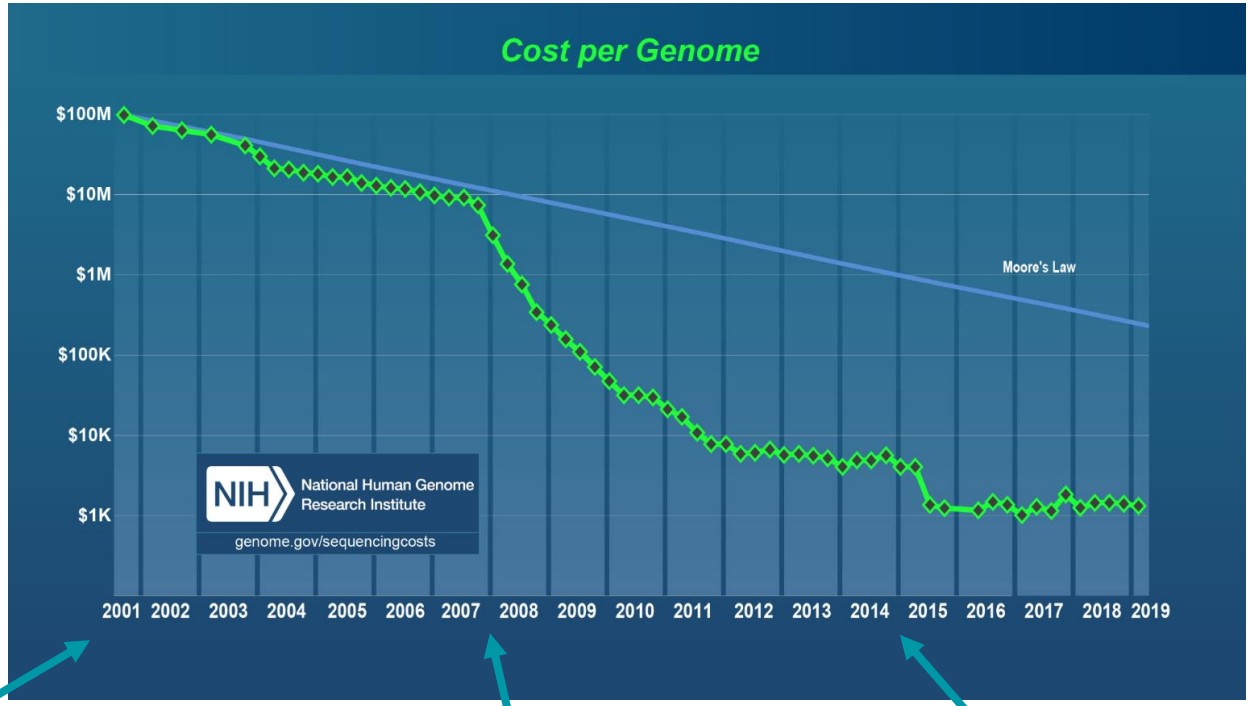
solexa, 454, illumina

3 **THIRD GENERATION** From 2011

PacBio, oxford nanopore

Sequencing output, price, reads size, sequencing quality

# From Sanger to 3rd sequencing technology




**1 FIRST GENERATION**
From 1977

sanger

**2 SECOND GENERATION**
From 2007

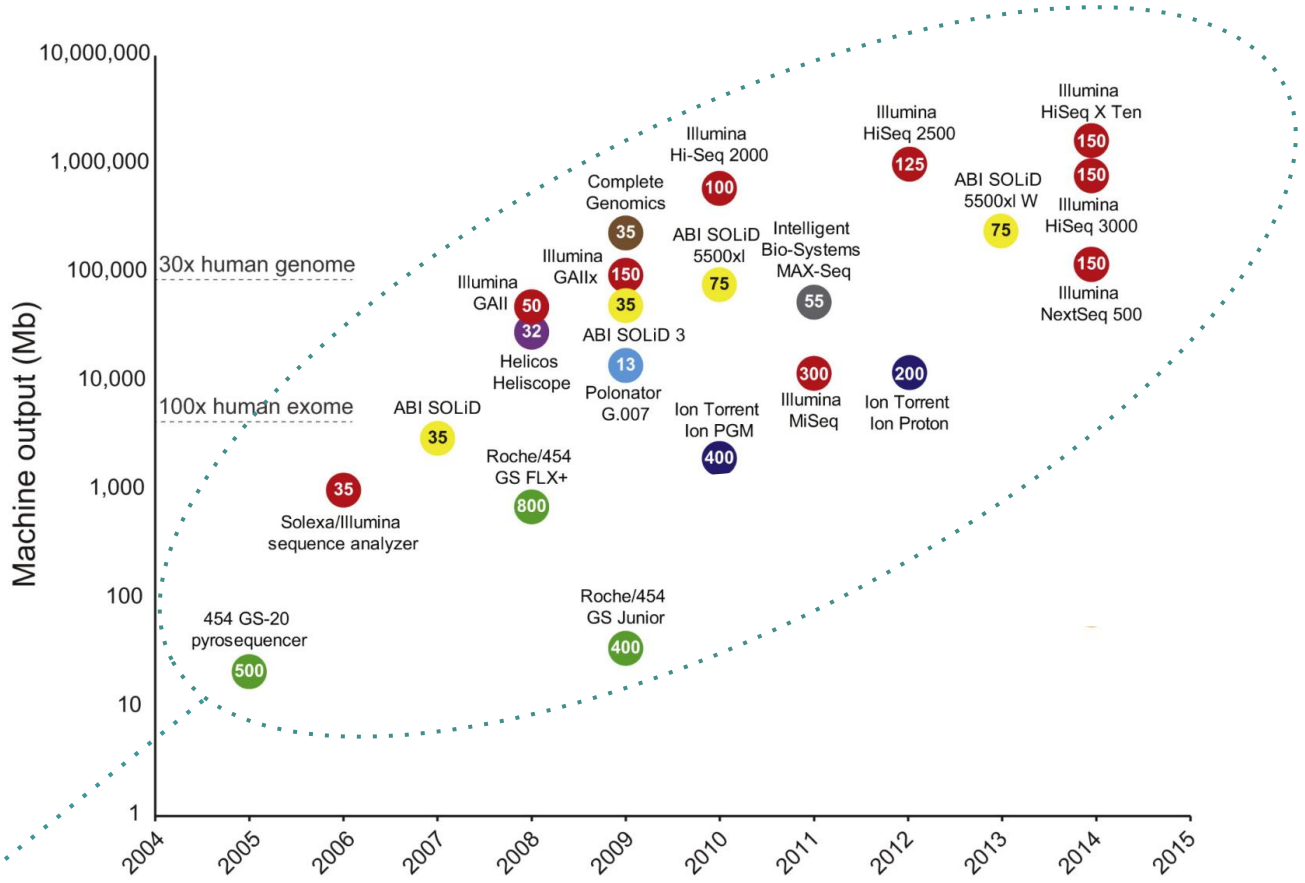
Illumina

**3 THIRD GENERATION**
From 2011

PacBio, ONT

# From Sanger to 3rd sequencing technology

## Une augmentation du débit de séquençage



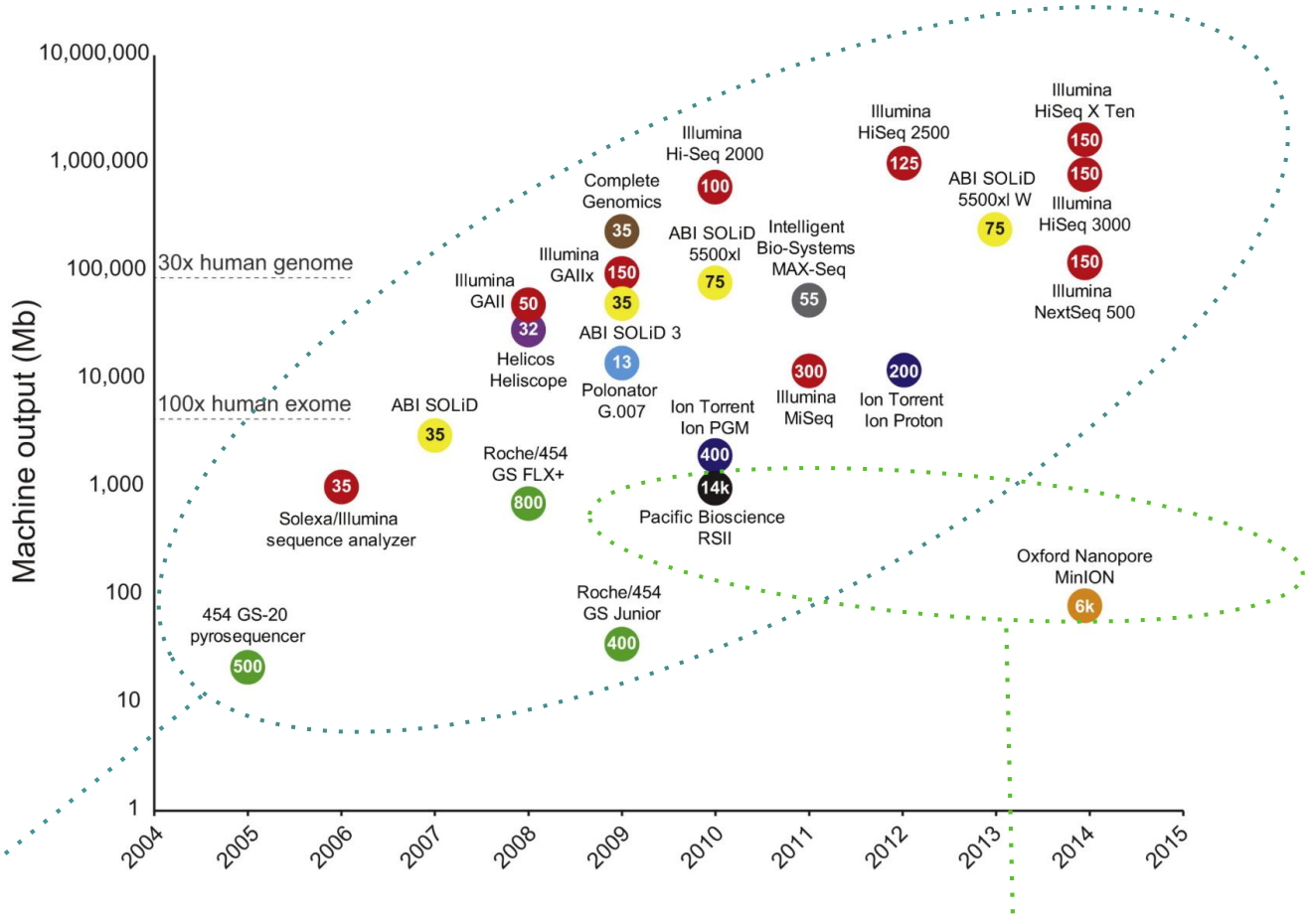

**2** **SECOND GENERATION**
From 2007

**Illumina**

# From Sanger to 3rd sequencing technology

## Une augmentation du débit de séquençage

REF




**2** SECOND GENERATION
From 2007

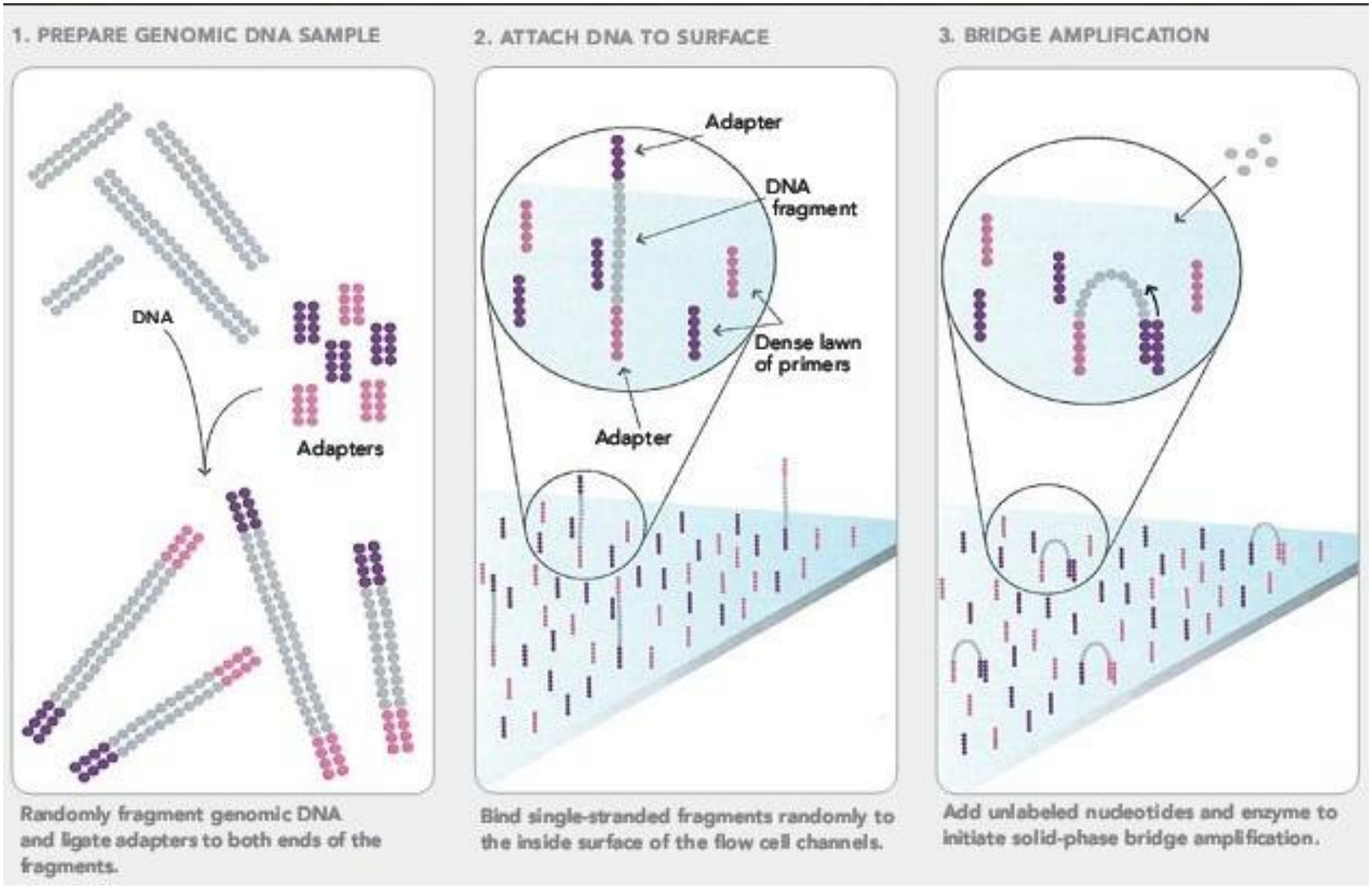
Illumina

**3** THIRD GENERATION
From 2011

# Short Reads ?
# Long Reads ?

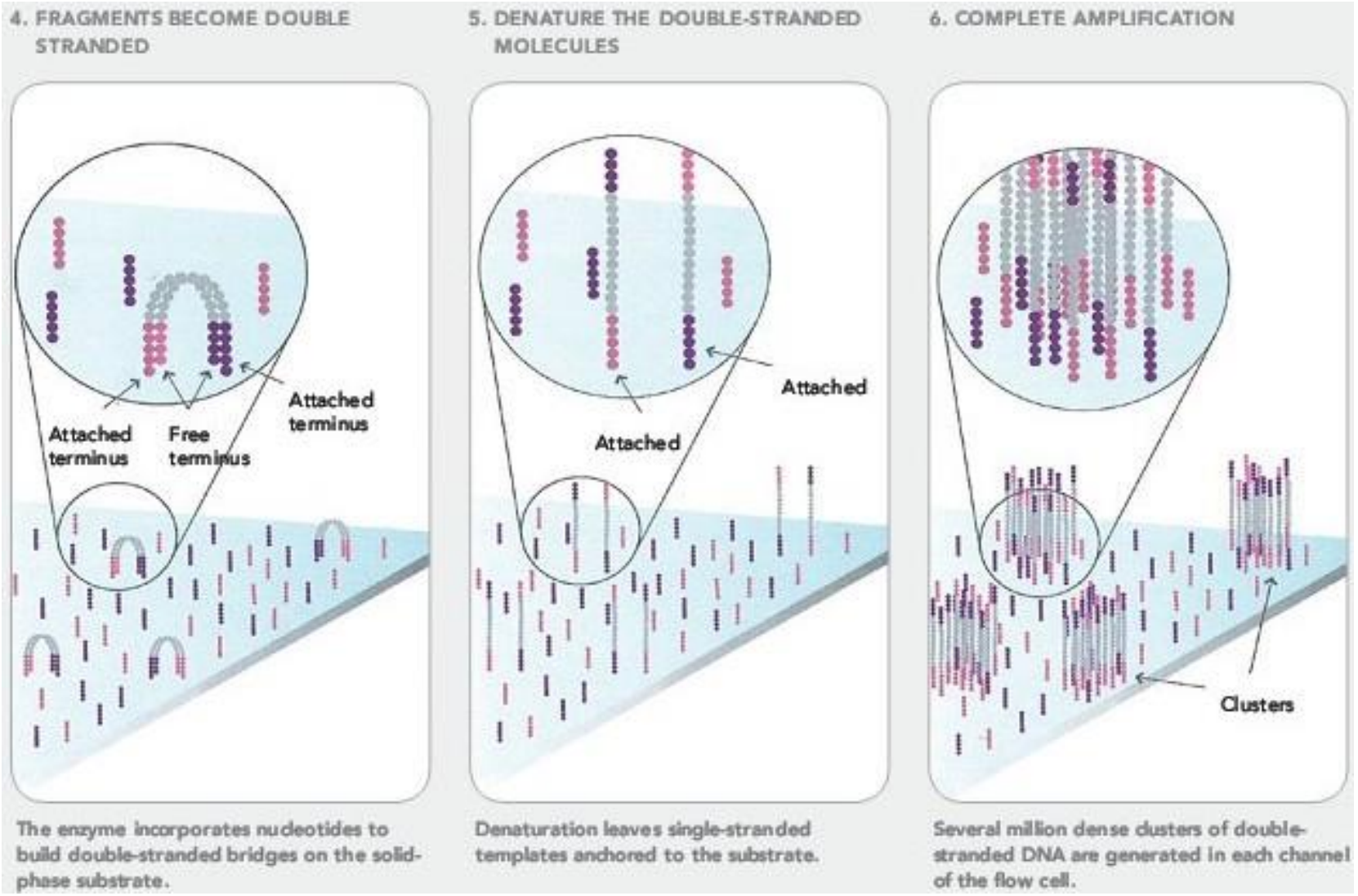
# Short Reads - Illumina technology

**2** **SECOND GENERATION**
From 2007



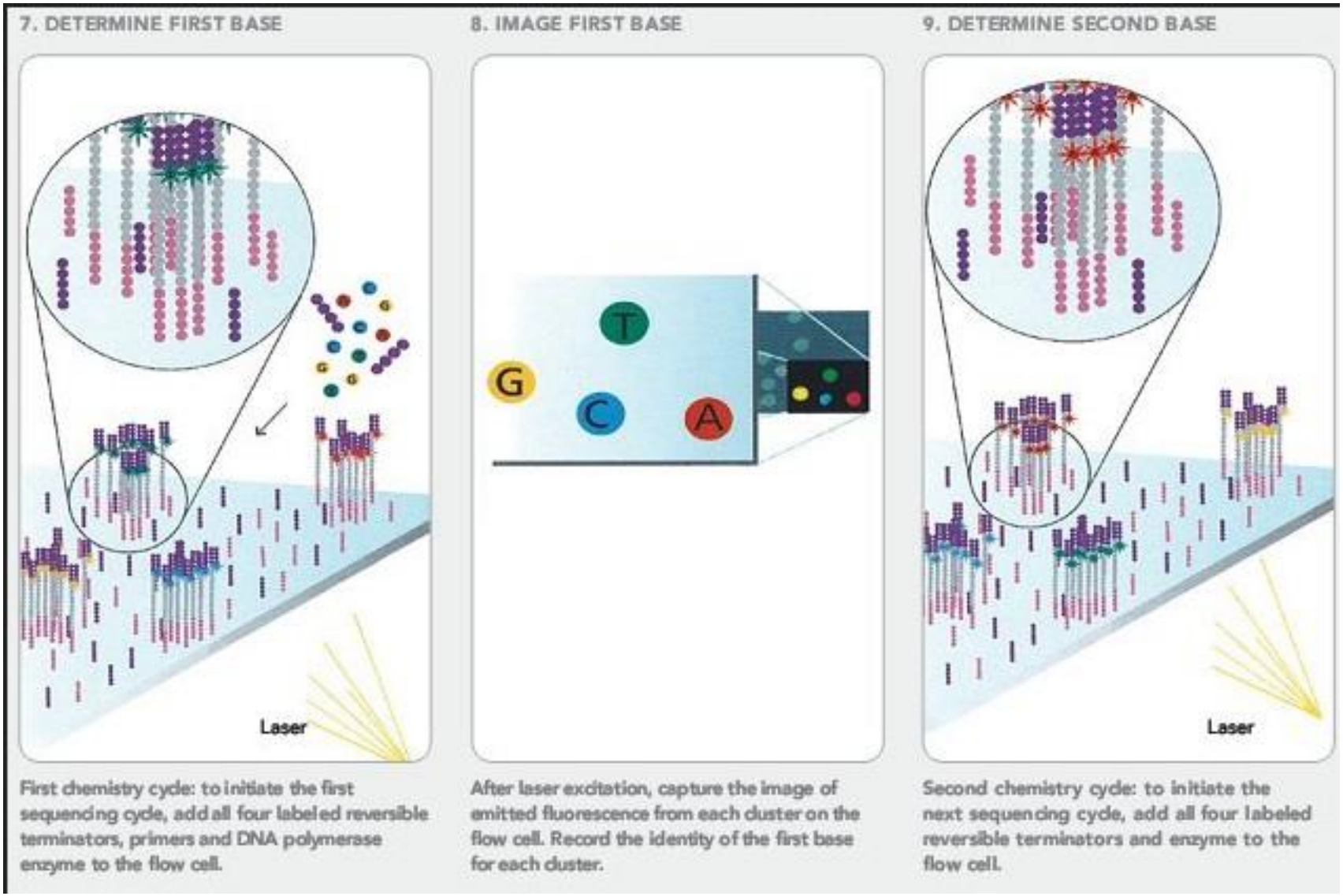
# Short Reads - Illumina technology

**2** **SECOND GENERATION**
From 2007



# Short Reads - Illumina technology

**2** **SECOND GENERATION**
From 2007



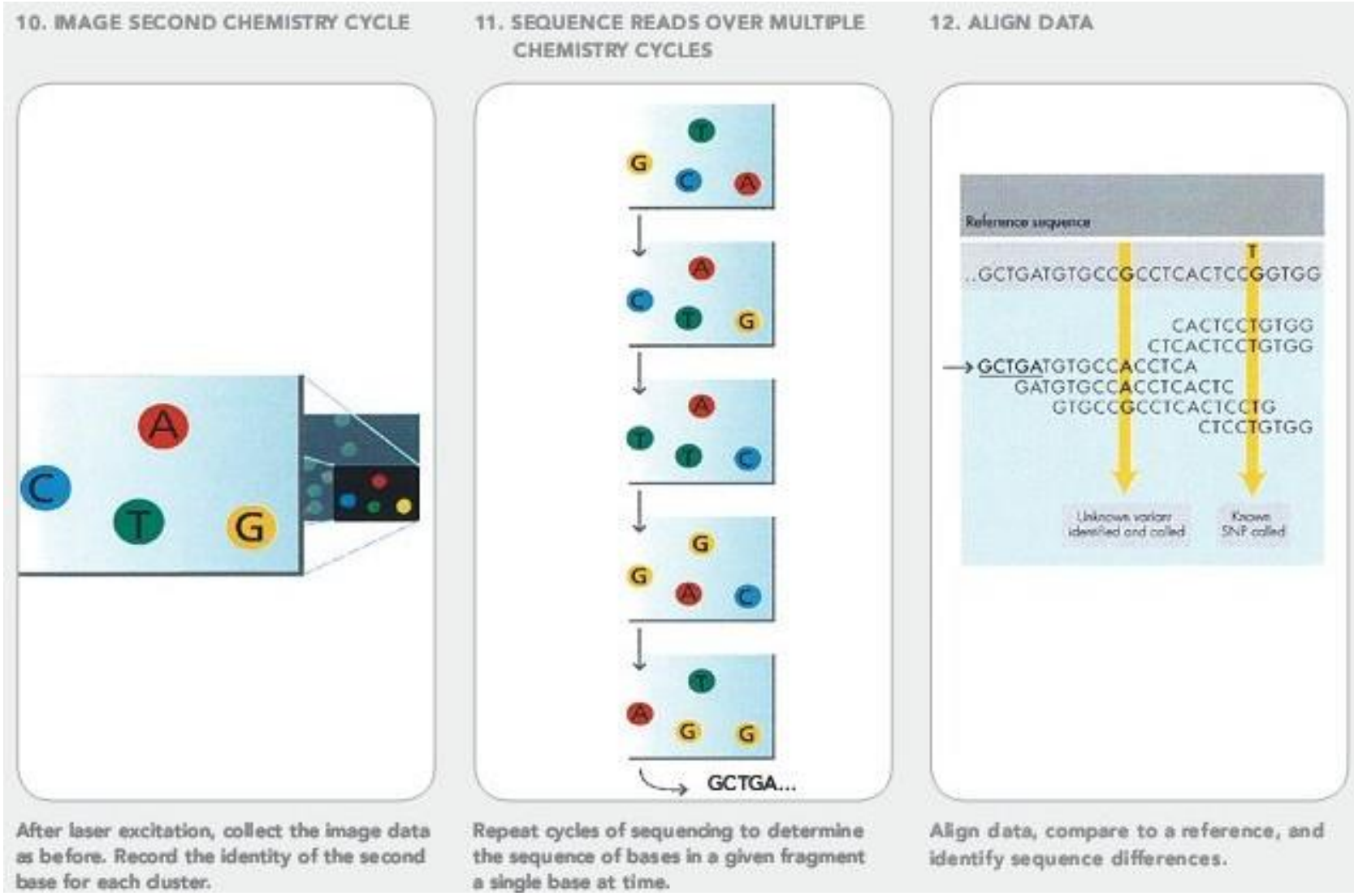

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

# Short Reads - Illumina technology

**2** **SECOND GENERATION**
From 2007



# Short Reads - Illumina - PROS & CONS

2 **SECOND GENERATION**
From 2007


- ✓ **Output volume** 20 billions of 150b reads, 6T
  *NovaSeq6000*
- ✓ **Accuracy** 99.99 % - but questionable
- ✓ **Run is cheap**
- ✓ **MySeq is cheap** ~60 000 USD per machine


- **Size** 150 + 150, *NovaSeq*
  but 400 pb, *MySeq*

3 **THIRD GENERATION**
From 2011

# Two technologies

## Oxford Nanopore


MinION

GridION

PromethION

## Pacific BioScience

RSII

Sequel

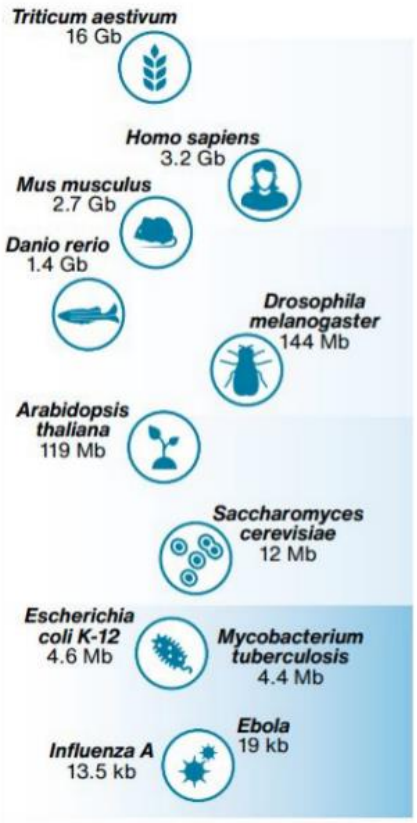
from Elixir GAAS 2018

# Long Reads - Oxford nanopore

**3** **THIRD GENERATION**
From 2011

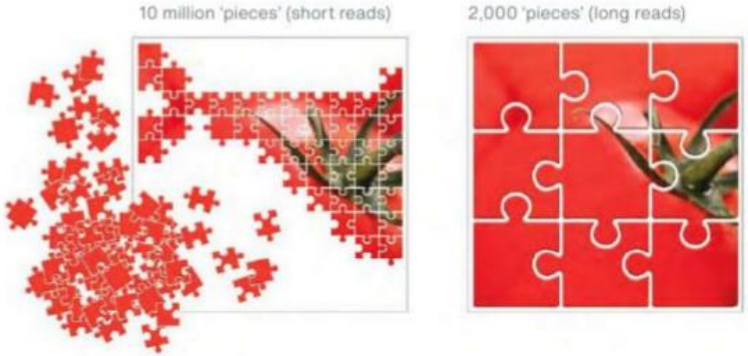

# Long Reads - Oxford nanopore

**3** **THIRD GENERATION**
From 2011



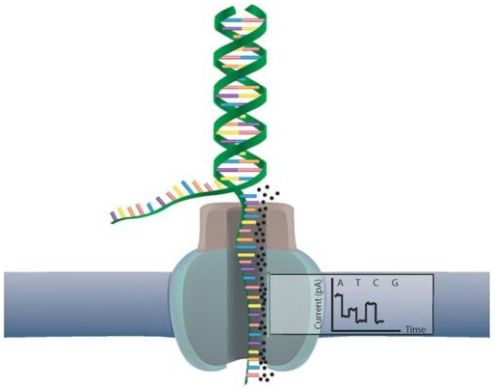

Microbial genomes | Human genomes | Animal genomes | Plant genomes

- Simplify de novo assembly and correct existing genomes
- They bridge repetitions and build less fragmented genomes. SV, repeats, phasing
- They come from technologies which do not amplify the DNA fragments and therefore have less coverage bias.
- They are affordable.
- Detecting base modifications : they provide methylation information
- Analysing long-read transcriptomes



# Long Reads - Oxford nanopore technology

**3** **THIRD GENERATION**
From 2011

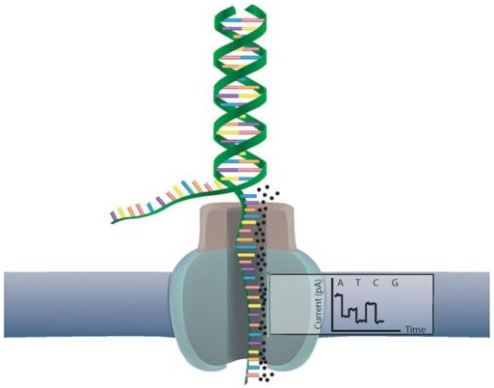



From Circulation Research

- No Amplification
- NO SYNTHESIS
- Very Long Length

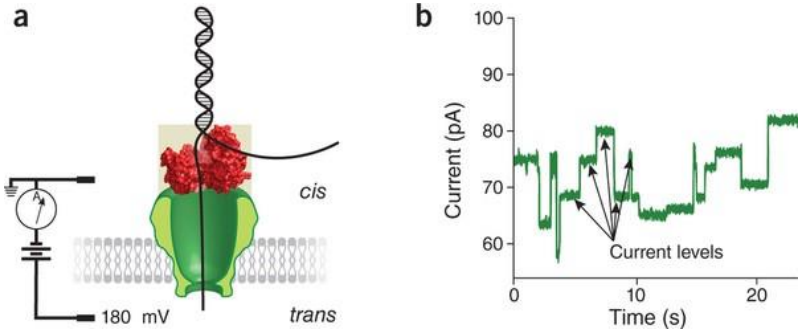
# Long Reads - Oxford nanopore technology

**3** **THIRD GENERATION**
From 2011




From Circulation Research

- No Amplification
- NO SYNTHESIS
- Very Long Length




- Magnetic fields variation measure
- *Minion*: USB key - sized
- Raw signal in *Fast5 format*, basecalled in *Fastq format*

# Long Reads - Oxford nanopore - PROS & CONS

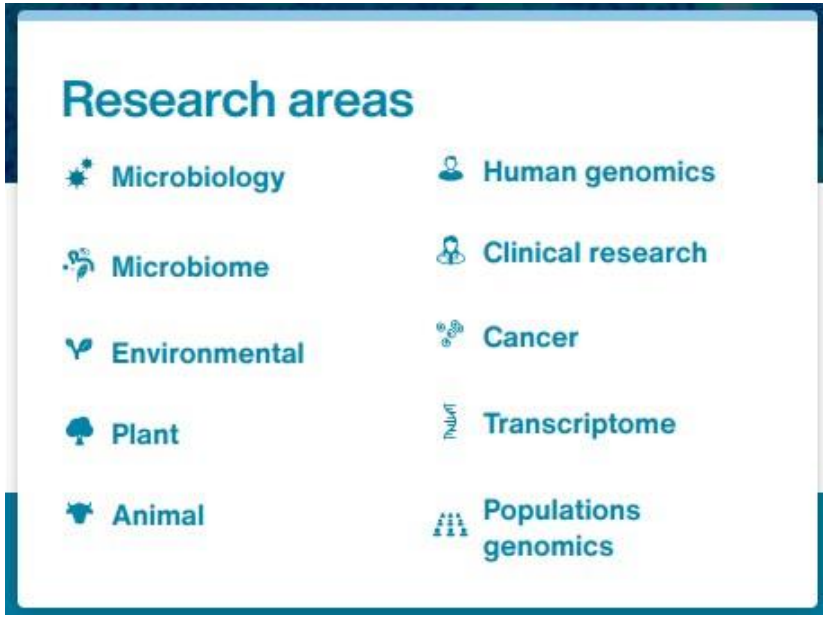
**3** **THIRD GENERATION**
From 2011


- ✓ Single strand direct sequencing
- ✓ Bases Modification detection in real-time
- ✓ Native RNA!
- ✓ **Read length** ~ 10-50kb more than 2Mb rep
- ✓ **Run cheap** 1,000 USD for 30Gb by now minimum
- ✓ **Machine cheap** 1,000 USD for Minion
- ✓ **Fast** 15mn library, 48-72h run


- Error Rate 3-8%, can be corrected, 1-2% in tests
- Quality of DNA/RNA limits the sequencing

# Long Reads - Oxford nanopore - What you can do with it ?

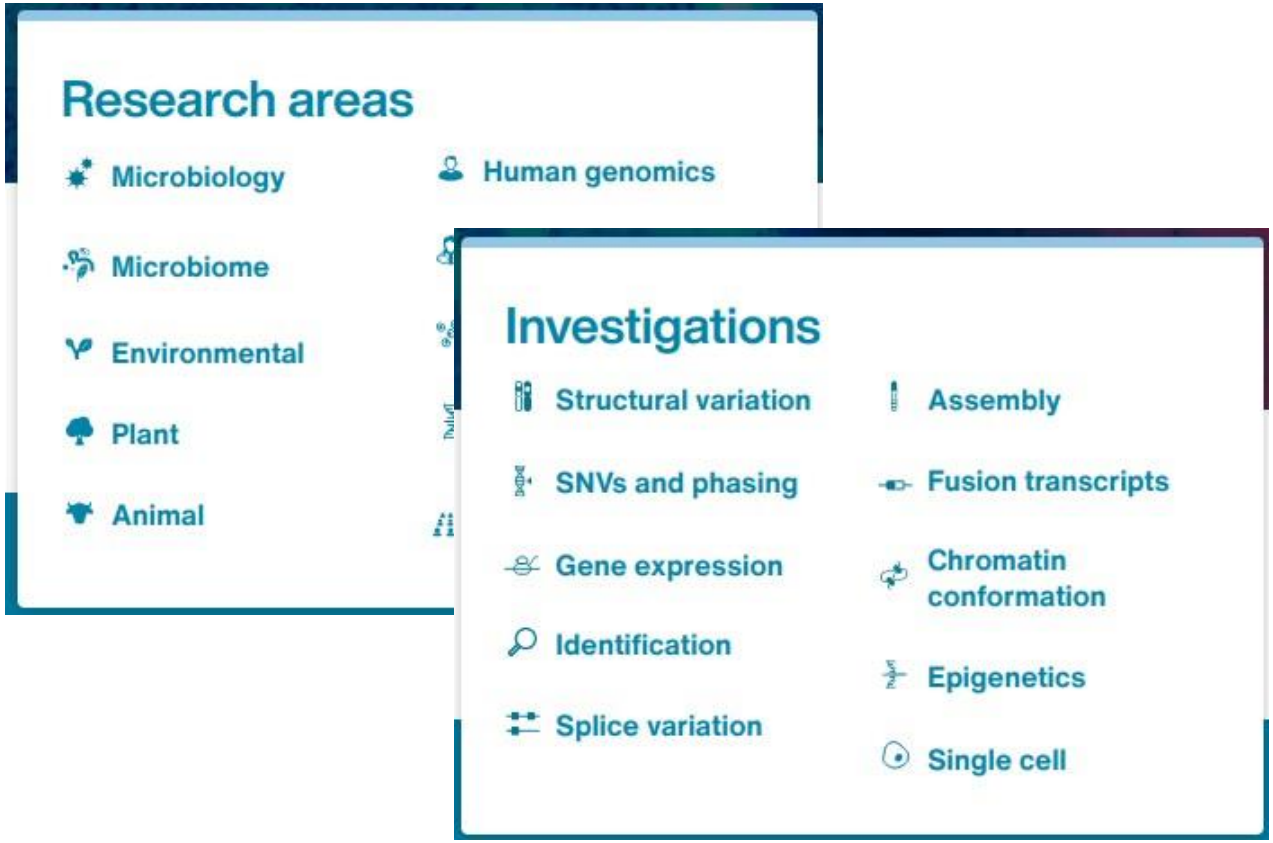
3 **THIRD GENERATION**
From 2011




From Nanopore website

# Long Reads - Oxford nanopore - What you can do with it ?

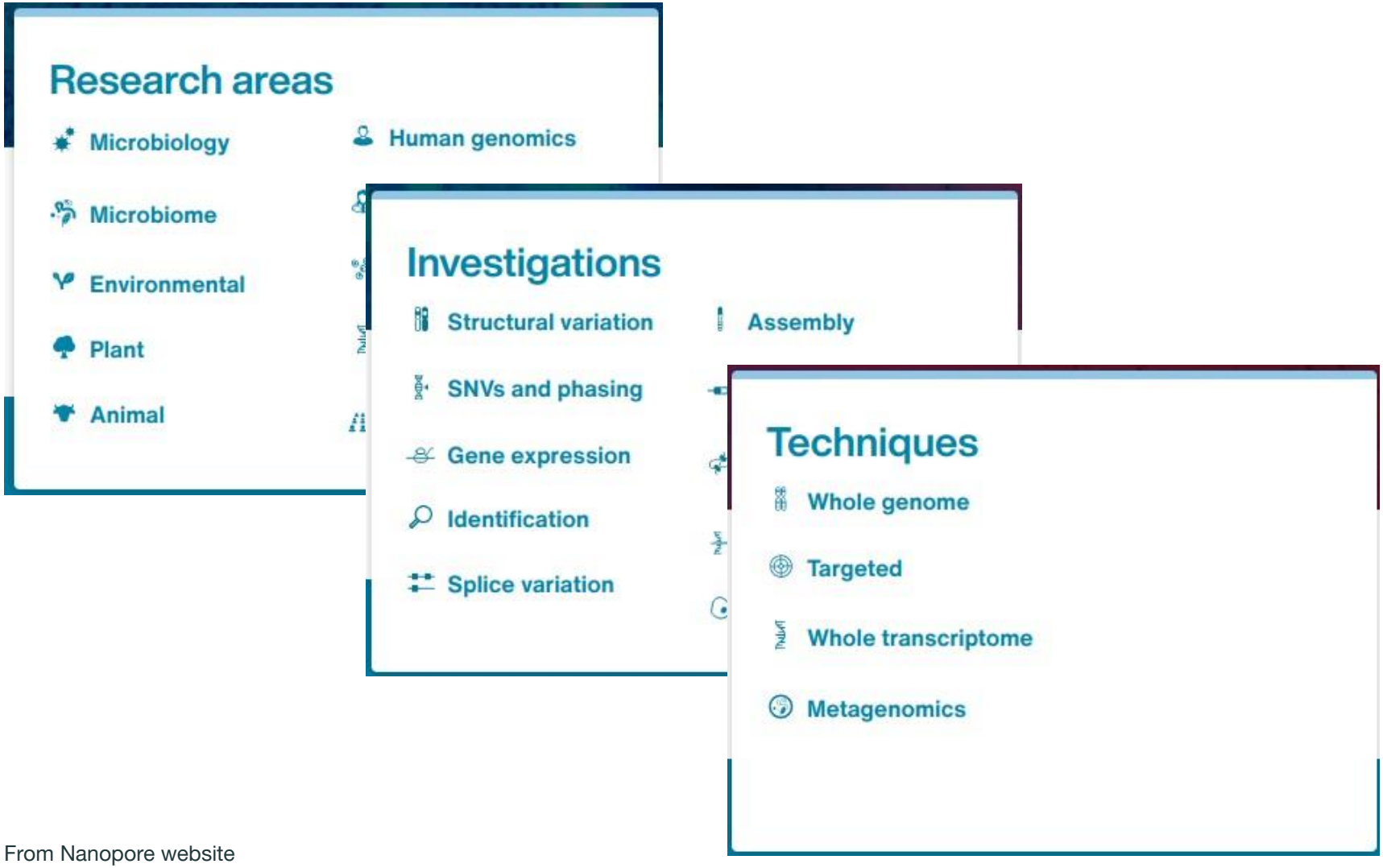
3 **THIRD GENERATION**
From 2011




From Nanopore website

# Long Reads - Oxford nanopore - What you can do with it ?

3 **THIRD GENERATION**
From 2011




From Nanopore website

# Comparison

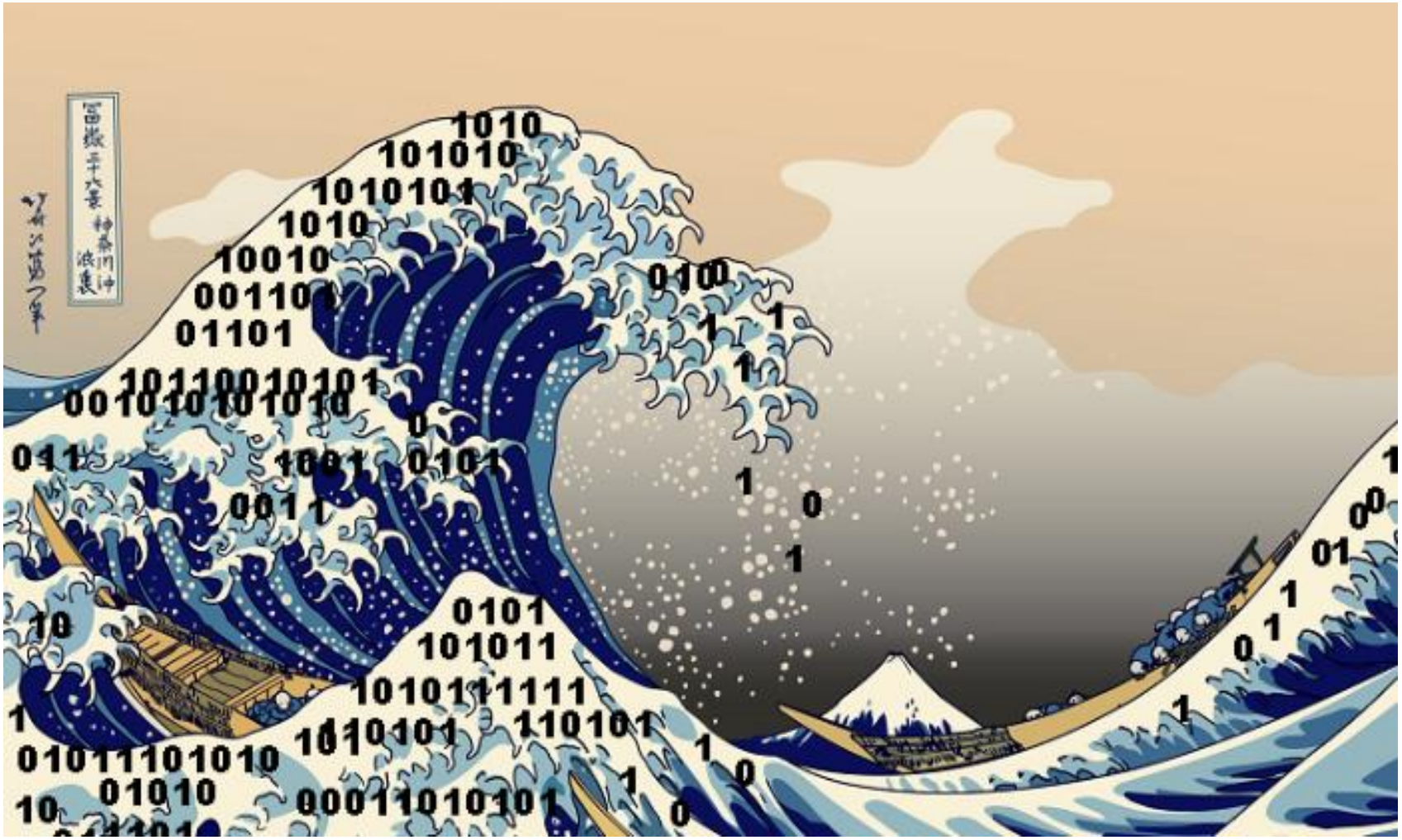
NGS platforms

| Platform | Template preparation | Chemistry | Max read length (bases) | Run times (days) | Max Gb per Run |
|---|---|---|---|---|---|
| **Illumina MiSeq** | Clonal Bridge Amplification | Reversible Dye Terminator | 2x300 | 0.17-2.7 | 15 |
| **Illumina HiSeq** | Clonal Bridge Amplification | Reversible Dye Terminator | 2x150 | 0.3-11[10] | 1000[11] |
| **Illumina Genome Analyzer IIX** | Clonal Bridge Amplification | Reversible Dye Terminator[12][13] | 2x150 | 2-14 | 95 |
| **Life Technologies SOLiD4** | Clonal-emPCR | Oligonucleotide 8-mer Chained Ligation[14] | 20-45 | 4-7 | 35-50 |
| **Life Technologies Ion Proton**[15] | Clonal-emPCR | Native dNTPs, proton detection | 200 | 0.5 | 100 |
| **Complete Genomics** | Gridded DNA-nanoballs | Oligonucleotide 9-mer Unchained Ligation[16][17][18] | 7x10 | 11 | 3000 |
| **Helicos Biosciences Heliscope** | Single Molecule | Reversible Dye Terminator | 35‡ | 8 | 25 |
| **Pacific Biosciences SMRT** | Single Molecule | Phospholinked Fluorescent | 10,000 (N50); 30,000+ (max)[19] | 0.08 | 0.5[20] |

| Platform | Read length (bp) | Isolates per run (max) | Run time | Instrument cost | Cost/ Mb |
|---|---|---|---|---|---|

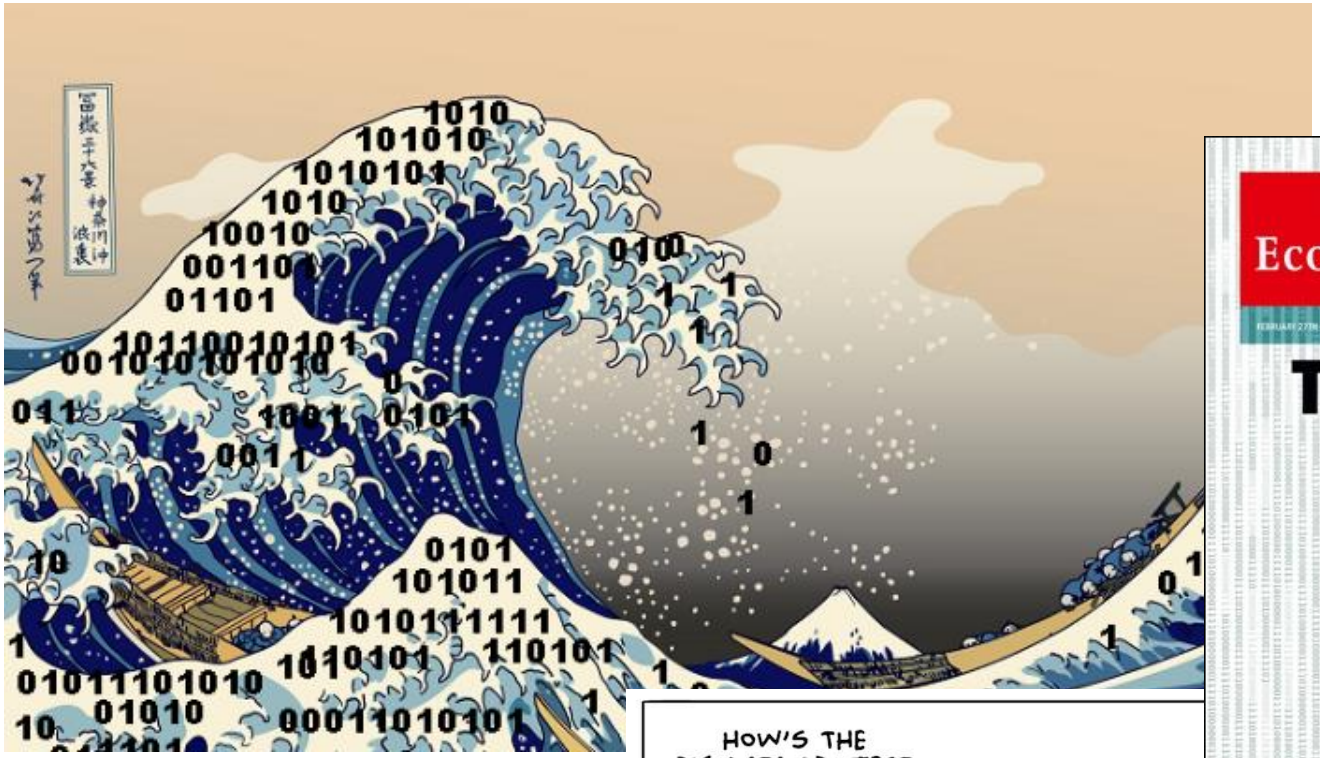
From wikipedia website

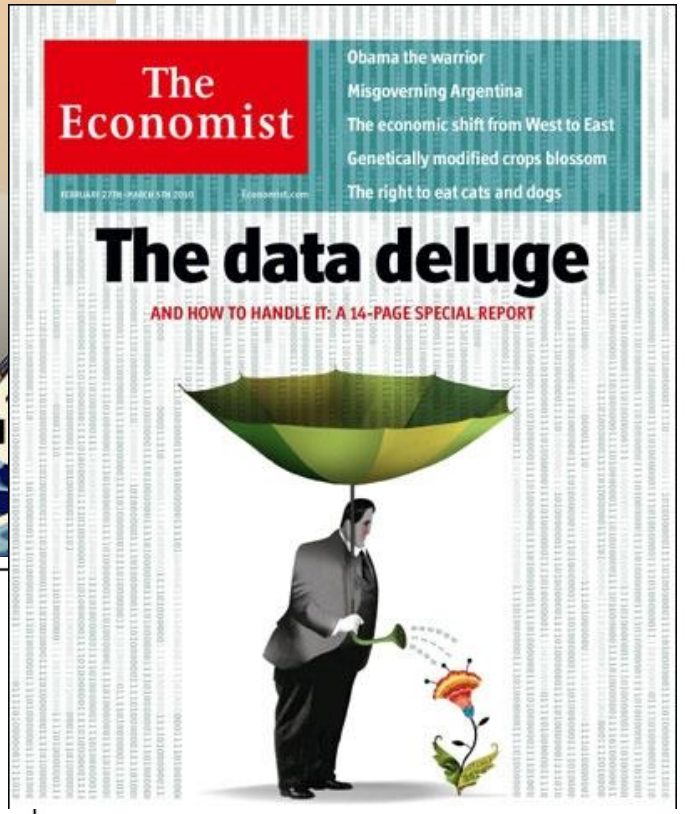
# From data rarity to data deluge



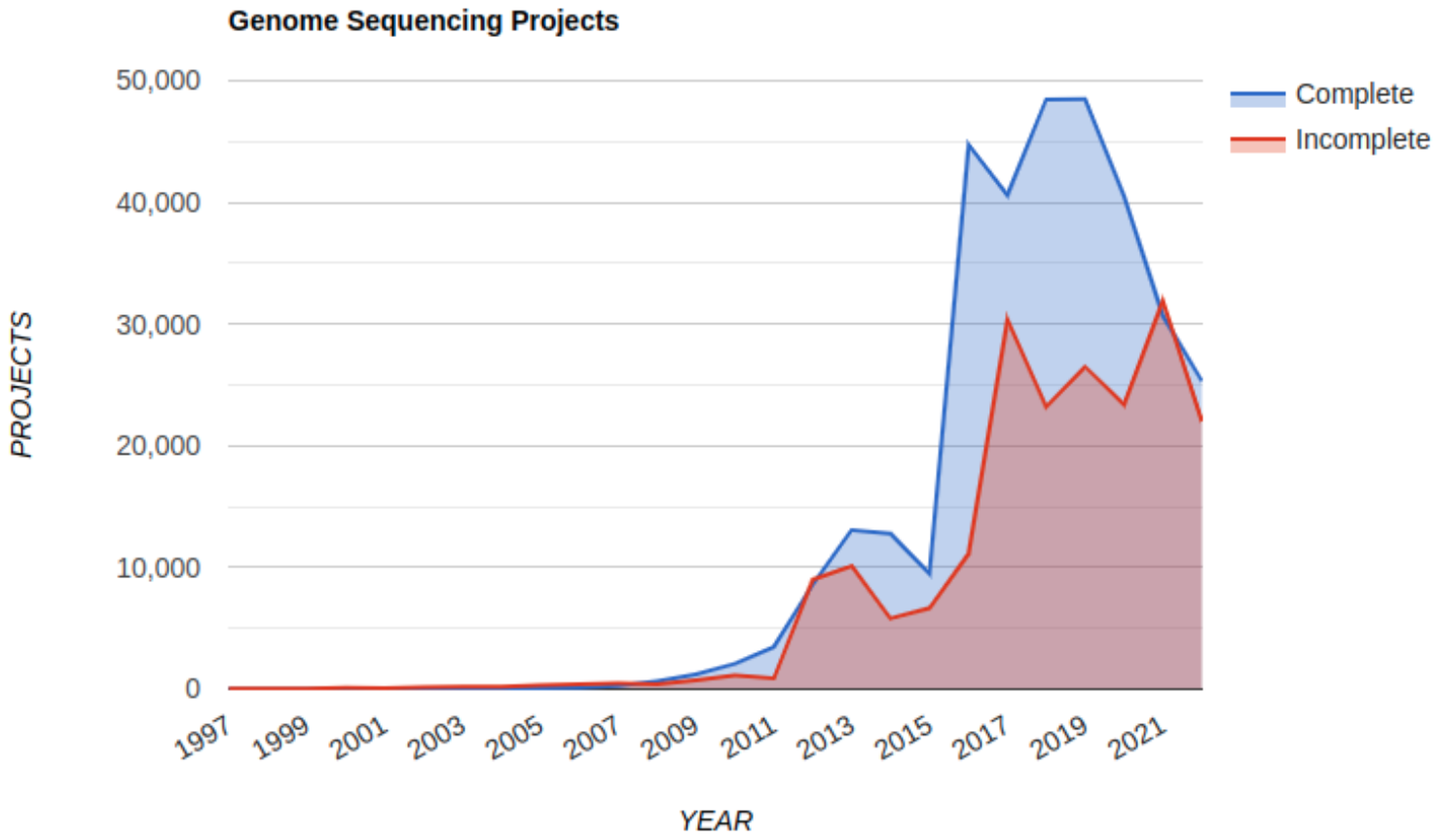
The Great Wave off Kanagawa, Hokusa @amitechsolutions.com

# From data rarity to data deluge





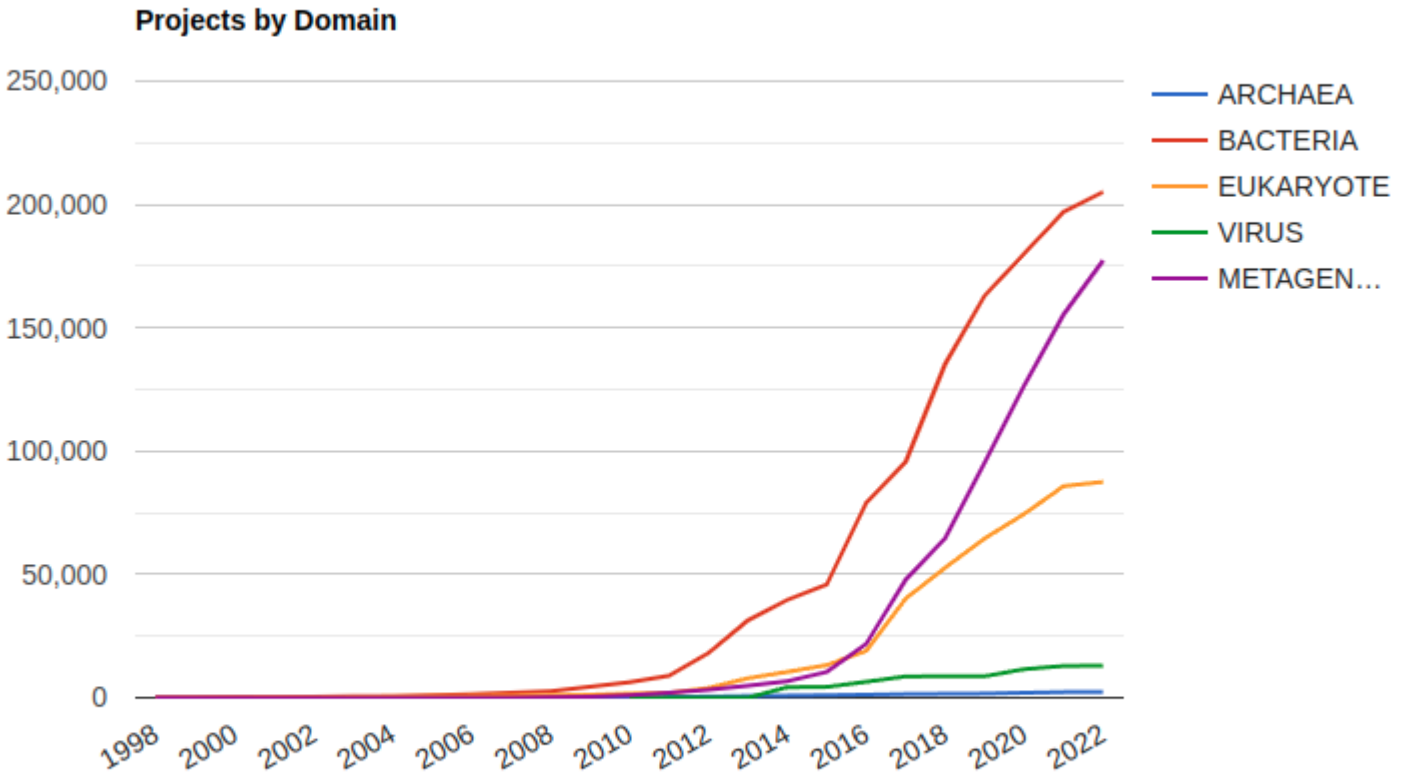

# Genome Totals by year and status




GOLD
GENOMES ONLINE DATABASE

from https://gold.jgi.doe.gov/statistics

# Project Totals by year and domain group




**Phylogenetic distribution of Bacterial Genome Projects**

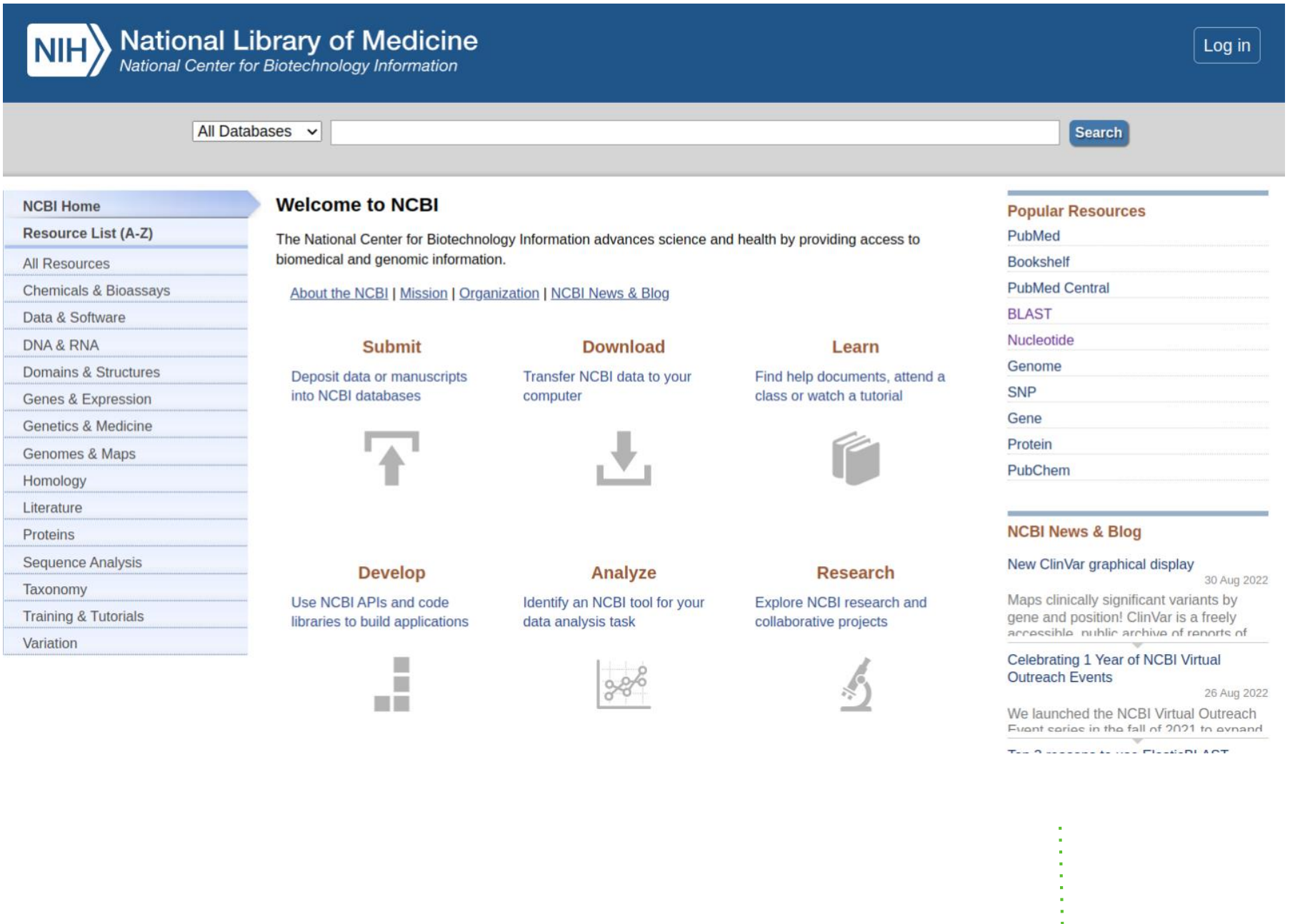
GOLD
GENOMES ONLINE DATABASE

from https://gold.jgi.doe.gov/statistics

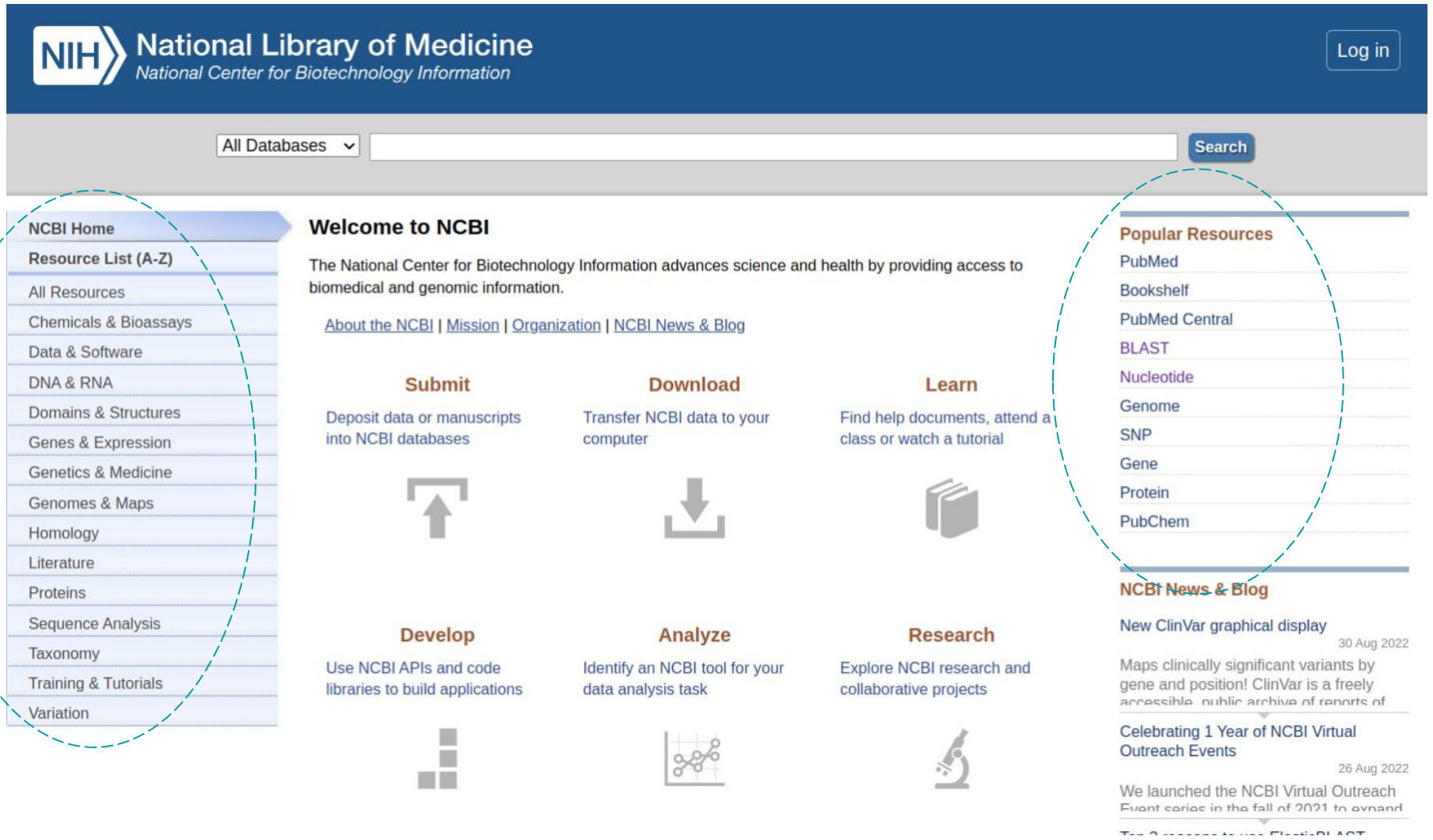
## Biological Databases

- ✓ **Sequence**
  - Nucleic : Genbank EMBL-EBI DDBJ DNA Data Bank of Japan
  - Proteic : swissprot TrEMBL
    PIR, Pfam, Prosite
- ✓ **Structure** PDB SCOP CATH
- ✓ **Specialized** by organism, by sequence type

https://www.ncbi.nlm.nih.gov/



https://www.ncbi.nlm.nih.gov/



https://www.ncbi.nlm.nih.gov/

National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases | Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | BioCollections

Search for [ ] as complete name lock Go Clear

Display 3 levels using filter: none

## Oryza sativa

Taxonomy ID: 4530 (for references in articles please use NCBI:txid4530)

current name

***Oryza sativa*** L., 1753

Genbank common name: **Asian cultivated rice**
NCBI BLAST name: **monocots**
Rank: **species**
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 1 (Standard)
Plastid genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:

common name(s)

**red rice, rice**

Lineage( *full* )
cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; Liliopsida; Petrosaviidae; commelinids; Poales; Poaceae; BOP clade; Oryzoideae; Oryzeae; Oryzinae; Oryza

**Entrez records**

| Database name | Subtree links | Direct links |
|---|---|---|
| Nucleotide | 2,291,284 | 322,323 |
| Protein | 444,228 | 62,067 |
| Structure | 275 | 76 |
| Genome | 1 | 1 |
| Popset | 1,234 | 1,082 |
| Conserved Domains | 12 | 5 |
| GEO Datasets | 22,604 | 16,467 |
| PubMed Central | 34,990 | 34,990 |
| Gene | 95,353 | 149 |
| HomoloGene | 9,787 | 9,787 |
| SRA Experiments | 109,838 | 76,120 |
| GEO Profiles | 670,939 | 670,939 |
| Protein Clusters | 15,559 | - |
| Identical Protein Groups | 202,266 | 44,157 |
| BioProject | 6,895 | 5,349 |
| BioSample | 110,196 | 59,234 |
| Assembly | 105 | 55 |
| PubChem BioAssay | 483 | 449 |
| Taxonomy | 9 | 1 |

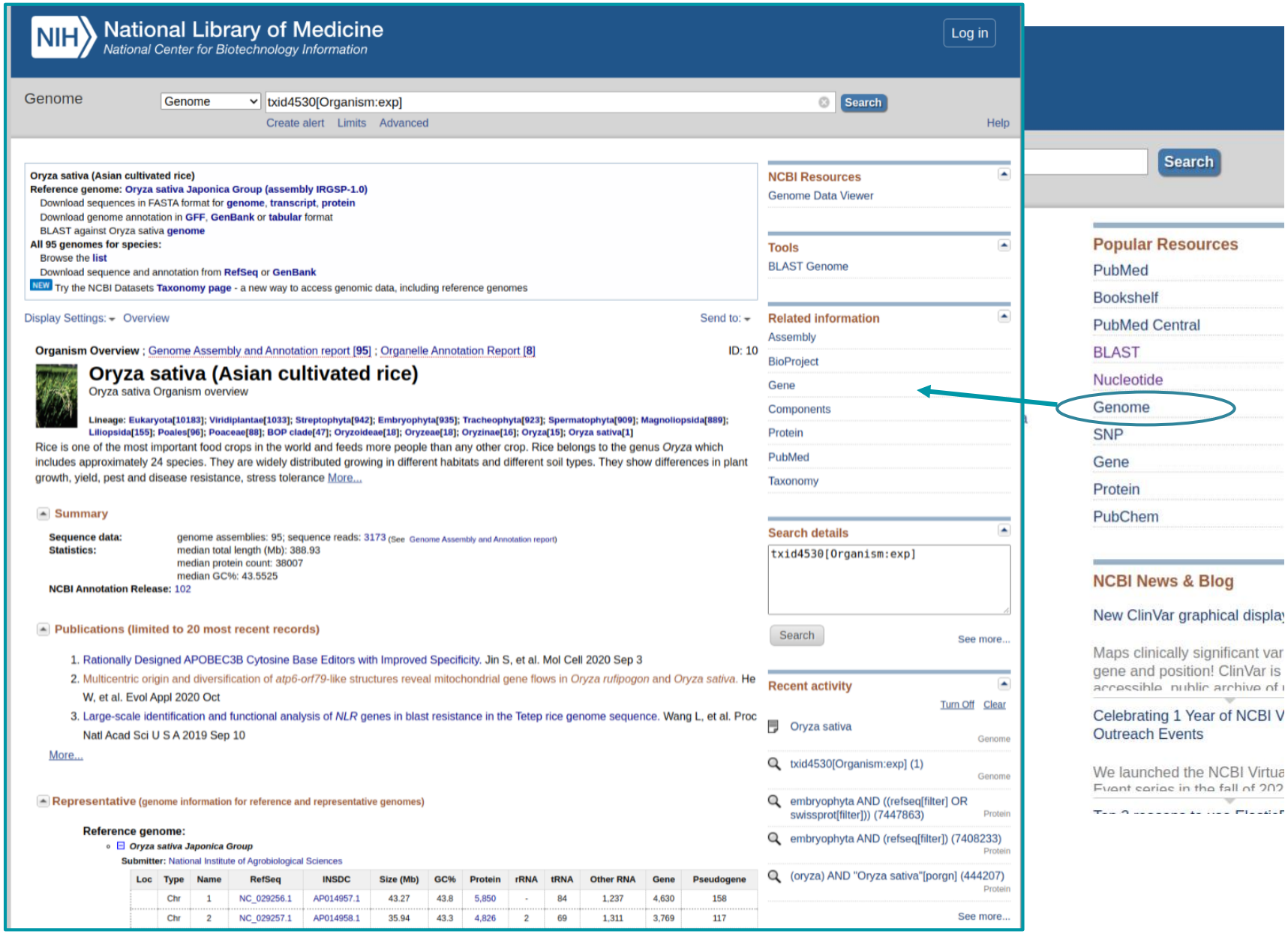
**Comments and References:**

GRIN (Oct 18, 2016)
Name accessed on 18 October 2016 in: USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network - (GRIN) [Online Database]. National Germplasm Resources Laboratory, Beltsville, Maryland.

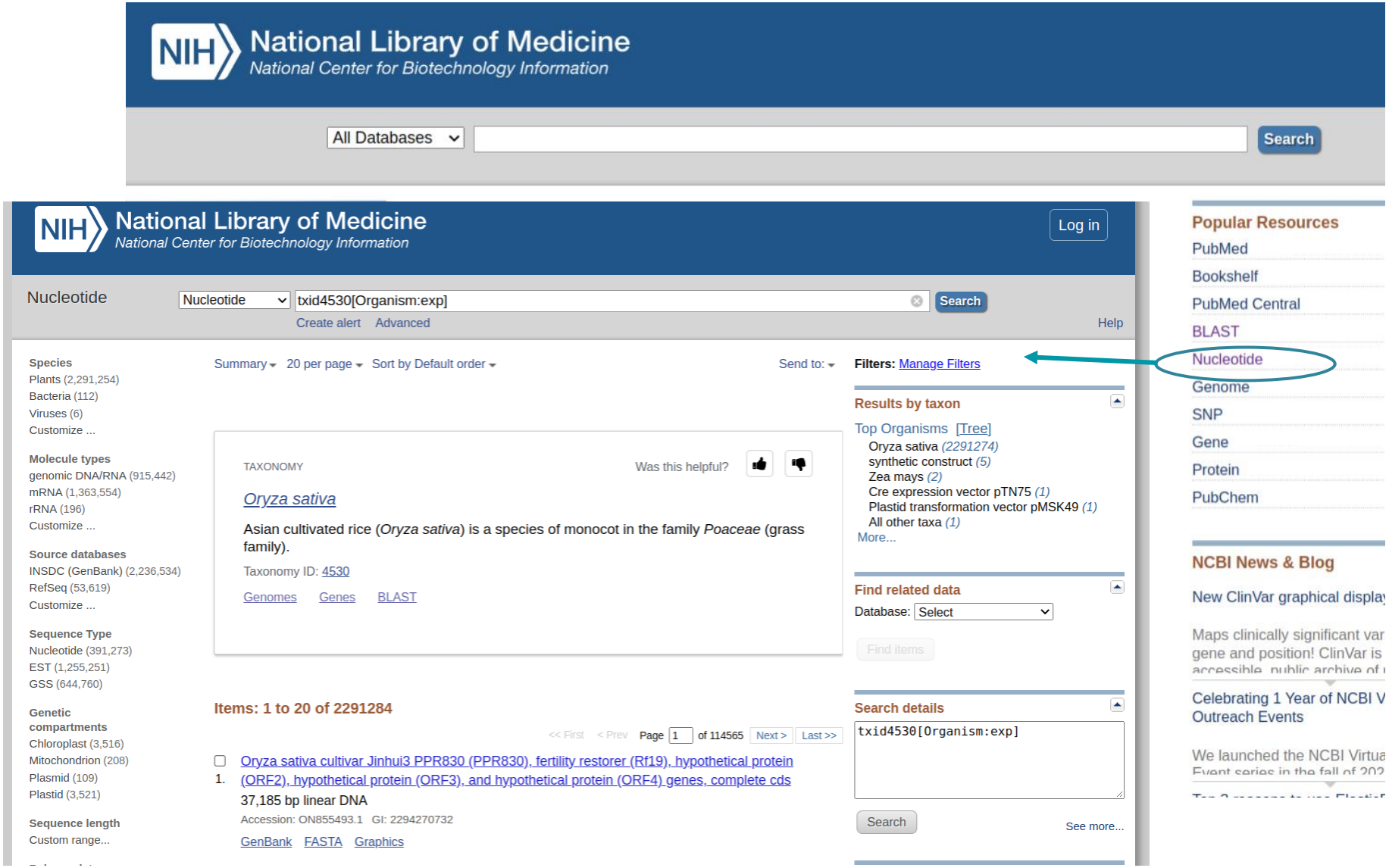
Flora of China - Poaceae
Chen S-L et al. 2006. Poaceae (R. Brown) Barnhart. In Wu, Z. Y., P. H. Raven & D. Y. Hong, eds. Flora of China. Vol. 22 (Poaceae). Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis. Online at Flora of China: www.efloras.org

The 3,000 rice genomes project
The 3,000 rice genomes project. GigaScience 2014, 3:7. DOI: http://dx.doi.org10.1186/2047-217X-3-7

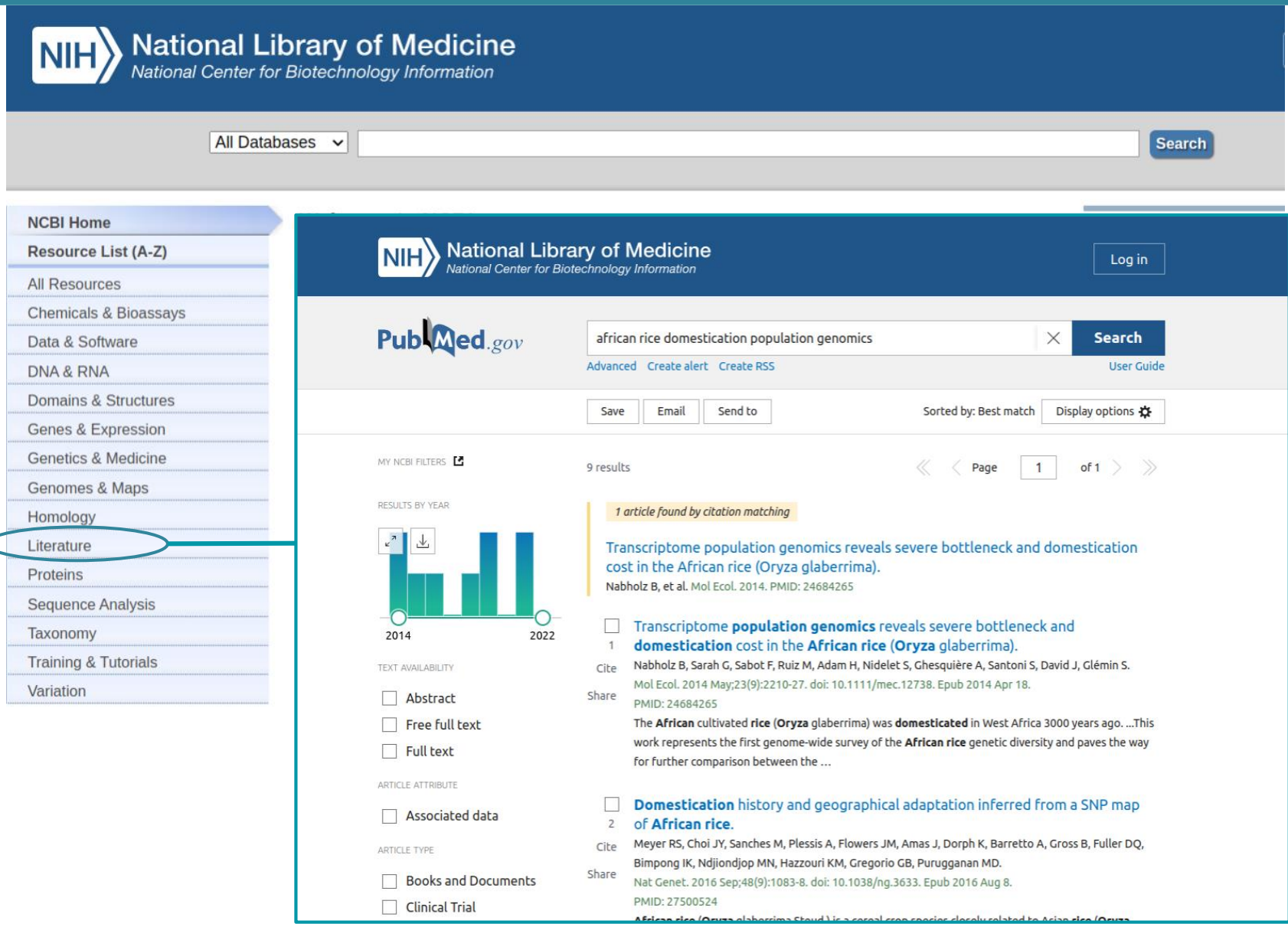
https://www.ncbi.nlm.nih.gov/



https://www.ncbi.nlm.nih.gov/



https://www.ncbi.nlm.nih.gov/



https://www.ncbi.nlm.nih.gov/genomes

# SRA / ENA (Eureopean Nucleotide Archive)

# Phytozome: Ressources génomiques de plantes

# Sequencing project

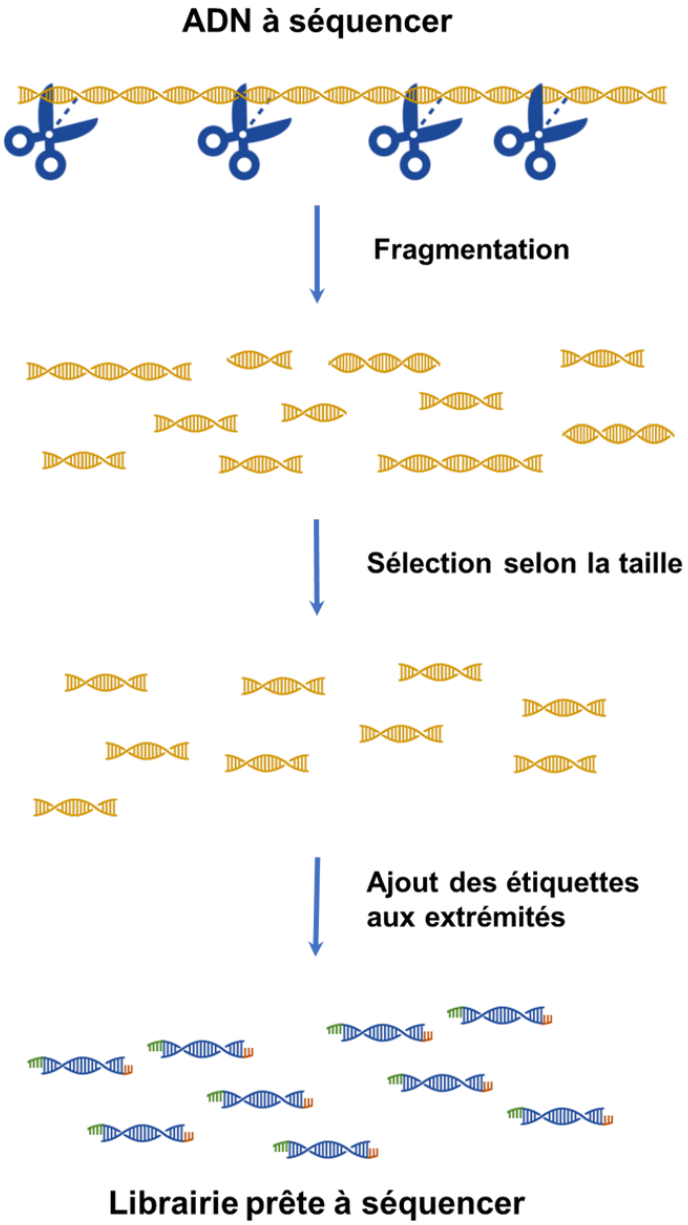
# OVERVIEW OF DNA SEQUENCING PROJECT

## Design expérimental

- Question scientifique
- quelle stratégie ? Quel échantillonnage ?
- Quelle stratégie bioinfo ?
- Quel méthodo de séquençage ? Quelle couverture de séquençage ?
- Quel volume de données brut? Sur quel cluster les analyses bioinformatiques vont-elles être tournées ?
- Qui va analyser mes données ?
- Où est ce que je vais stocker mes données?

# OVERVIEW OF DNA SEQUENCING PROJECT

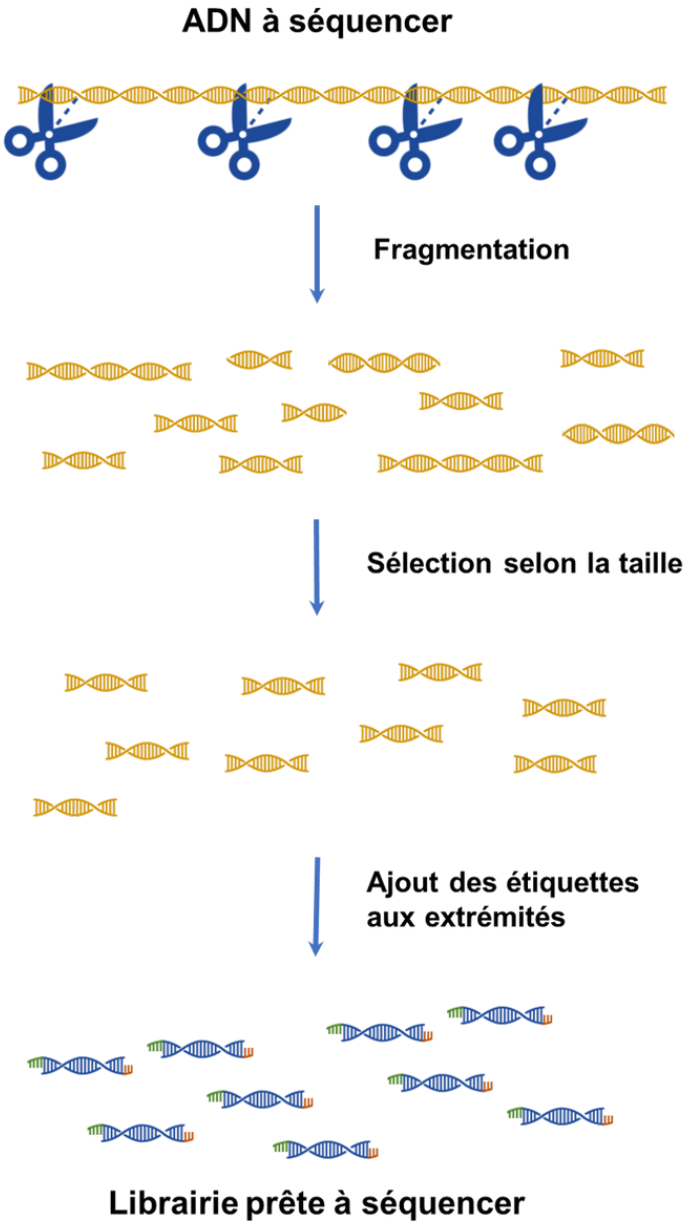
Design expérimental → Préparation banque



# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque

- Adaptateurs
- Contamination



# OVERVIEW OF DNA SEQUENCING PROJECT

# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

- Qualité de séquençage
- Profondeur de séquençage

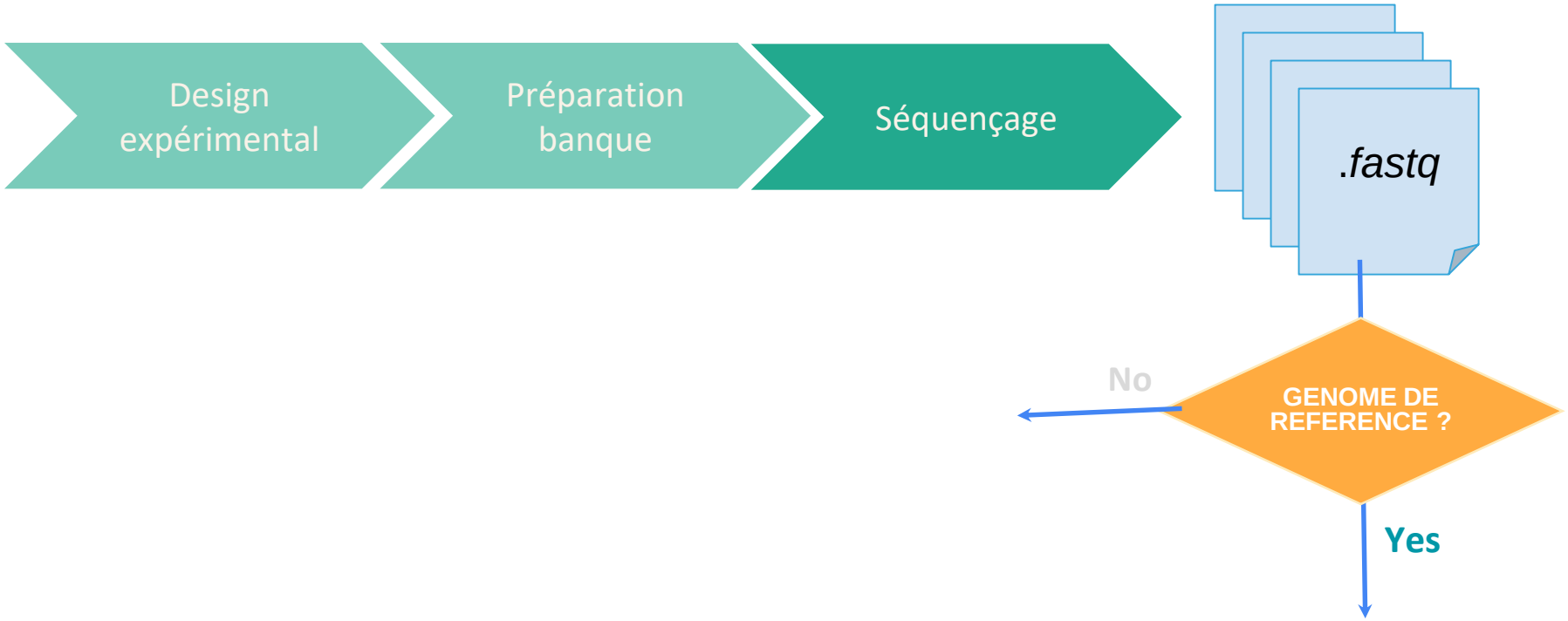
# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq


Genomic DNA is fragmented (not Nanopore) and sequenced -> millions of small sequences (reads) from random parts of the genome

Depending on sequence technology, reads can be from 100 bp up to 100kb in length

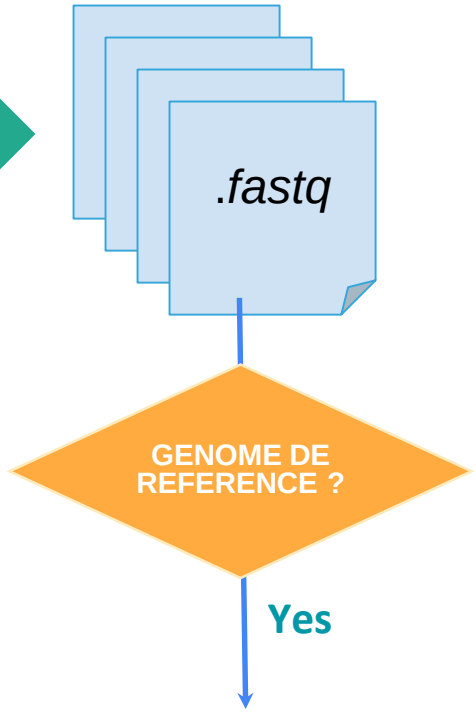
# OVERVIEW OF DNA SEQUENCING PROJECT




Adapted from Ross Whetten…

# OVERVIEW OF DNA SEQUENCING PROJECT

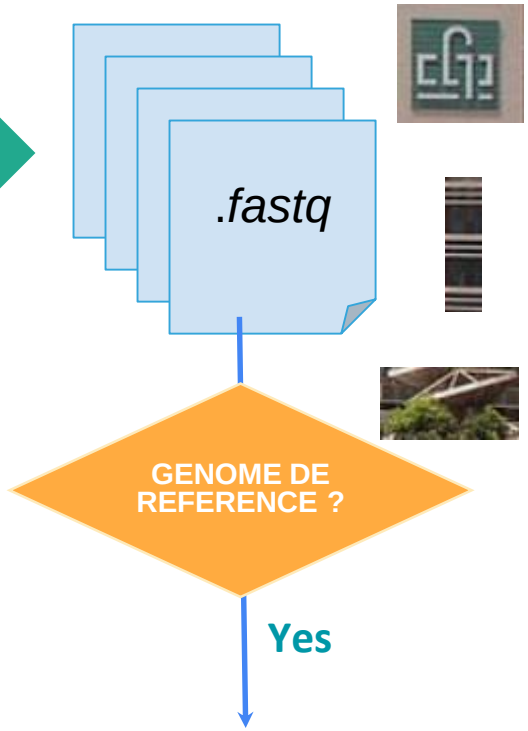
Design expérimental → Préparation banque → Séquençage




Adapted from Ross Whetten…

# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage




Adapted from Ross Whetten…

# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

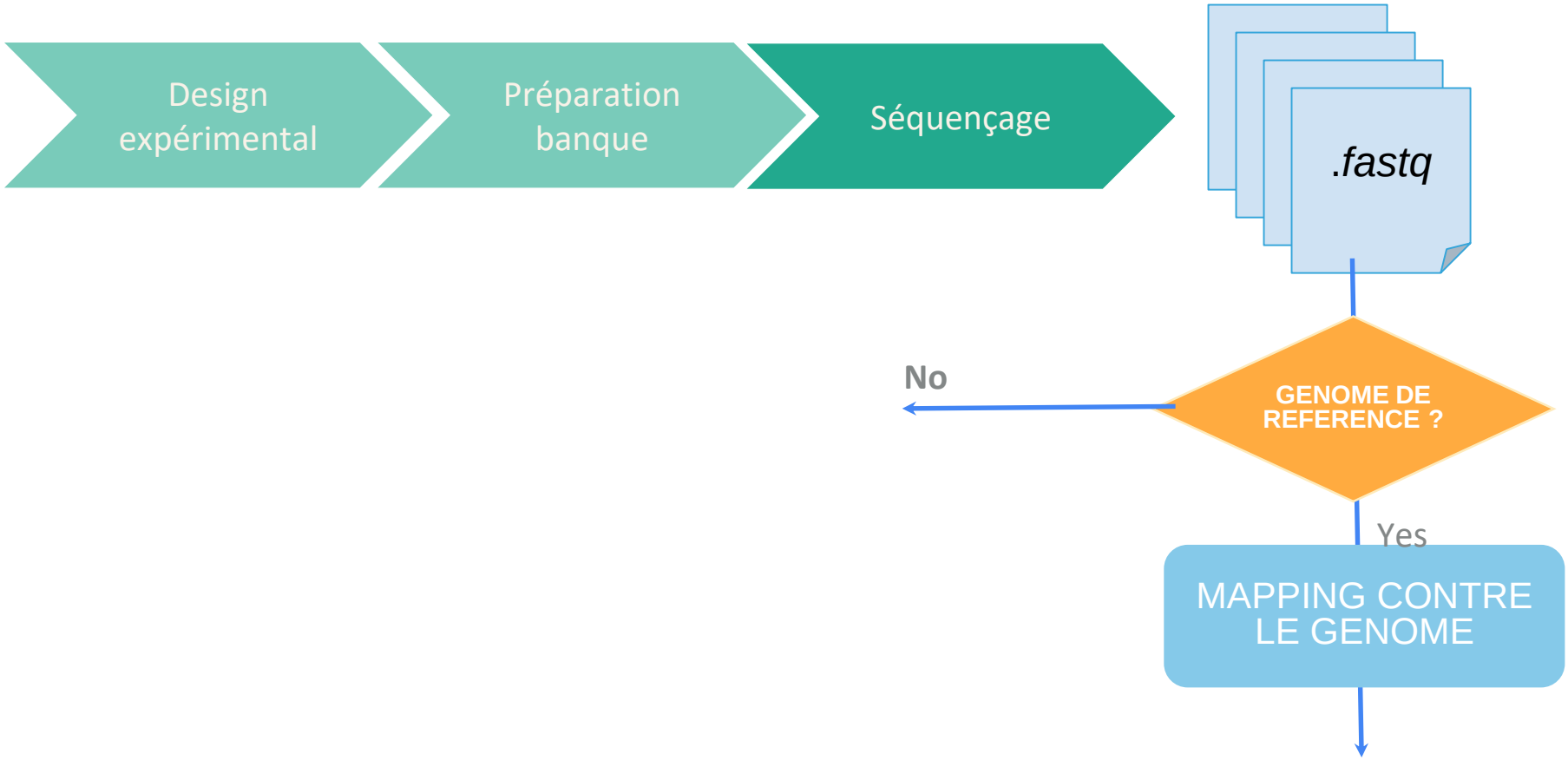
GENOME DE REFERENCE ?

Yes

MAPPING CONTRE LE GENOME

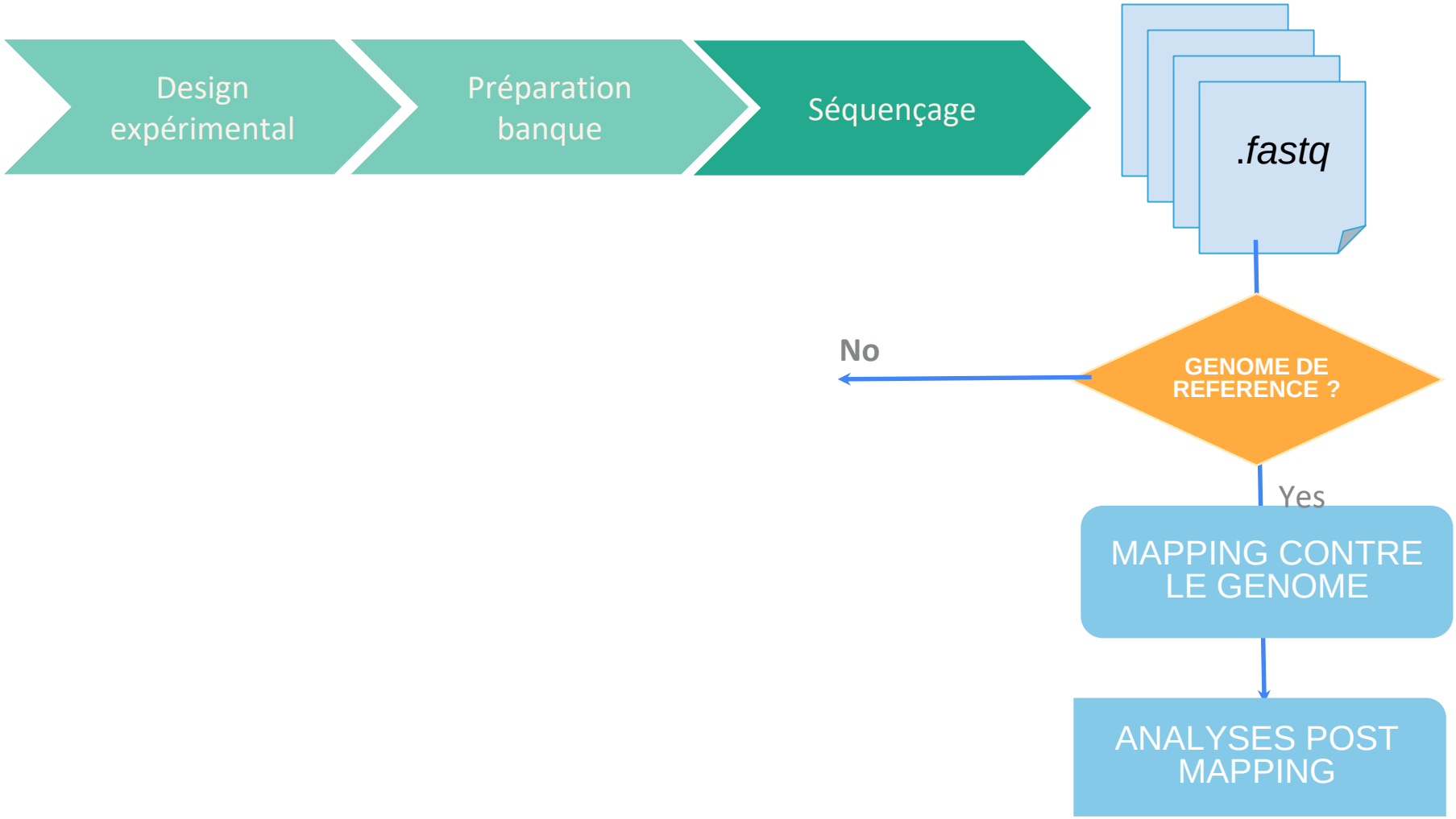
Adapted from Ross Whetten…

# OVERVIEW OF DNA SEQUENCING PROJECT




Adapted from Ross Whetten…

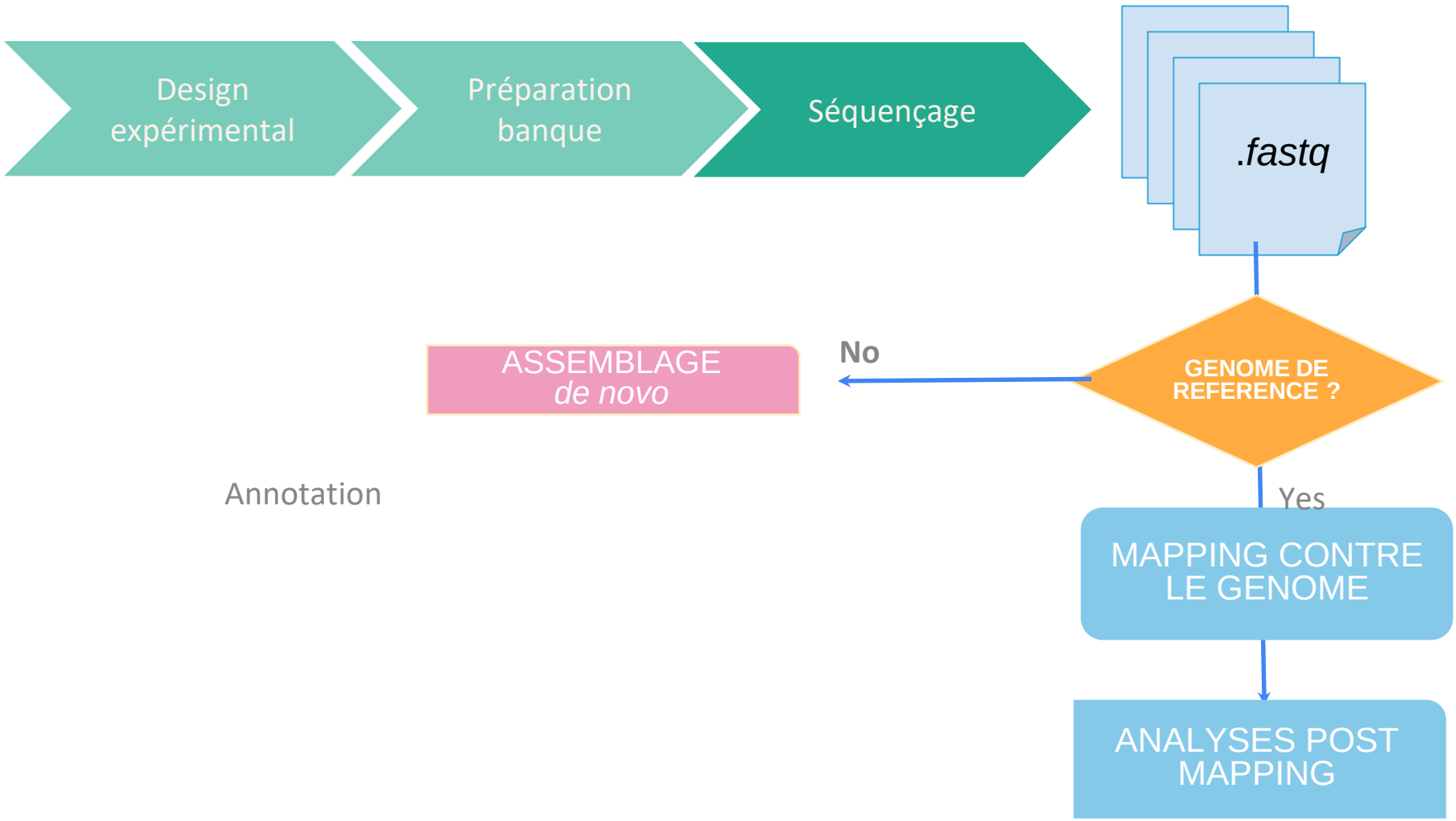
# OVERVIEW OF DNA SEQUENCING PROJECT




Adapted from Ross Whetten…

SNP, GWAS? expression différentielle

# OVERVIEW OF DNA SEQUENCING PROJECT




Adapted from Ross Whetten…

# What metagenomics is ?

Metagenomics ( Environmental Genomics or Community Genomics) is the study of genomes recovered from environmental samples without the need for culturing them

Metagenomics processes data using bioinformatics tools

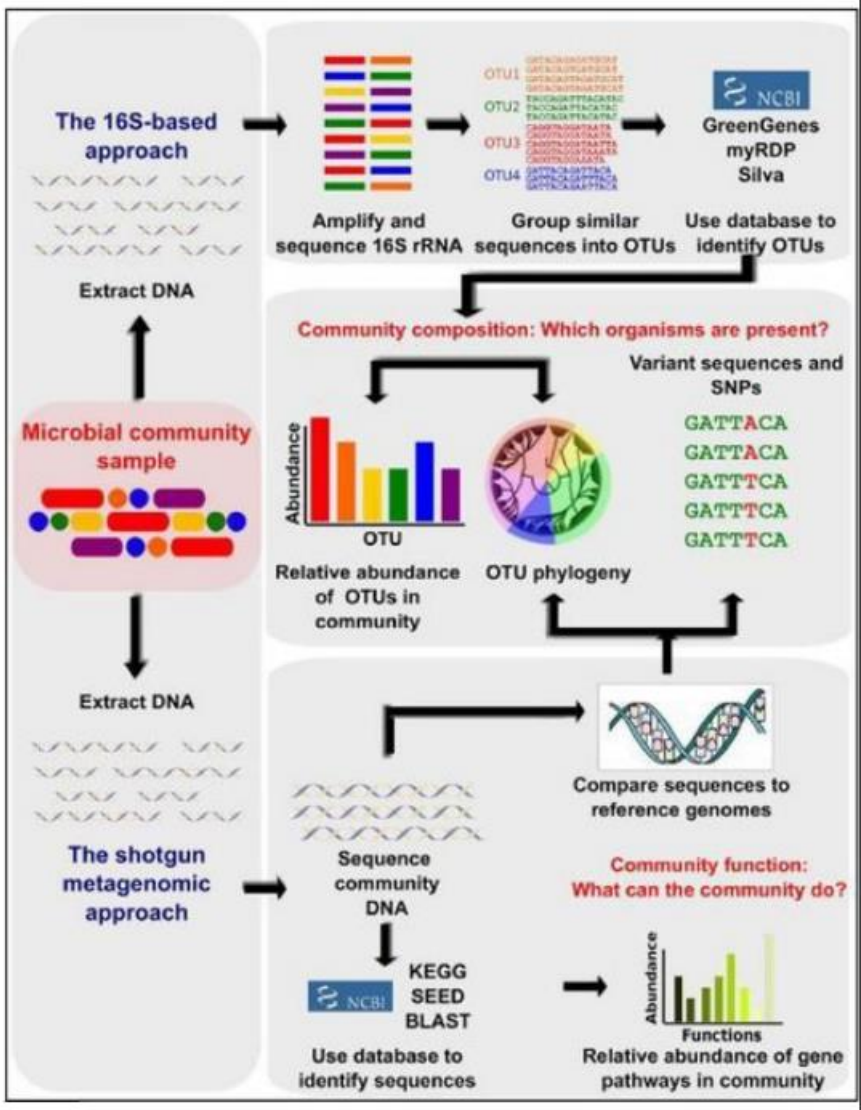
=> Organisms can be studied directly in their environments bypassing the need to isolate each species

=> There are significant advantages for viral metagenomics, because of difficulties cultivating the appropriate host

# Two main strategies in metagenomics

We can distinguish targeted metagenomics or shot-gun metagenomics :

- 16S rRNA metabarcoding is used to characterize the bacterial communities of an environment
- Whole-genome sequencing when the goal is to identify gene functions and pathways, or reconstruct microbial genomes.




Morgan XC, Huttenhower (2012) Plos Comput Biol

# Markers genes vs Shotgun metagenomics

| Marker Gene Profiling | Shotgun Metagenomics Profiling |
|---|---|
| Less expensive (~$100 per sample) | Still very expensive (~$1000 per sample) |
| Computational needs can be met by desktop / small server computers | Usually requires huge computational resources (cluster of computers) |
| Provides mainly taxonomic profiling | Provides both taxonomic and functional profiling |
| For 16S, majority of genes can be assigned at least to phylum level | Many more unassigned gene fragments (“wasted” data) |
| Relatively free of host DNA contamination | Prone to host DNA contamination |

# Strategies in Diversity Characterisation

| Technique | Advantages and challenges | Main applications |
|---|---|---|
| Metataxonomics using amplicon sequencing of the 16S or 18S rRNA gene or ITS | + Fast and cost-effective identification of a wide variety of bacteria and eukaryotes<br>– Does not capture gene content other than the targeted genes<br>– Amplification bias<br>– Viruses cannot be captured | * Profiling of what is present<br>* Microbial ecology<br>* rRNA-based phylogeny |
| Metagenomics using random shotgun sequencing of DNA or RNA | + No amplification bias<br>+ Detects bacteria, archaea, viruses and eukaryotes<br>+ Enables *de novo* assembly of genomes<br>– Requires high read count<br>– Many reads may be from host<br>– Requires reference genomes for classification | * Profiling of what is present across all domains<br>* Functional genome analyses<br>* Phylogeny<br>* Detection of pathogens |
| Meta-transcriptomics using sequencing of mRNA | + Identifies active genes and pathways<br>– mRNA is unstable<br>– Multiple purification and amplification steps can lead to more noise | * Transcriptional profiling of what is active |

Breitwieser et al. 2016

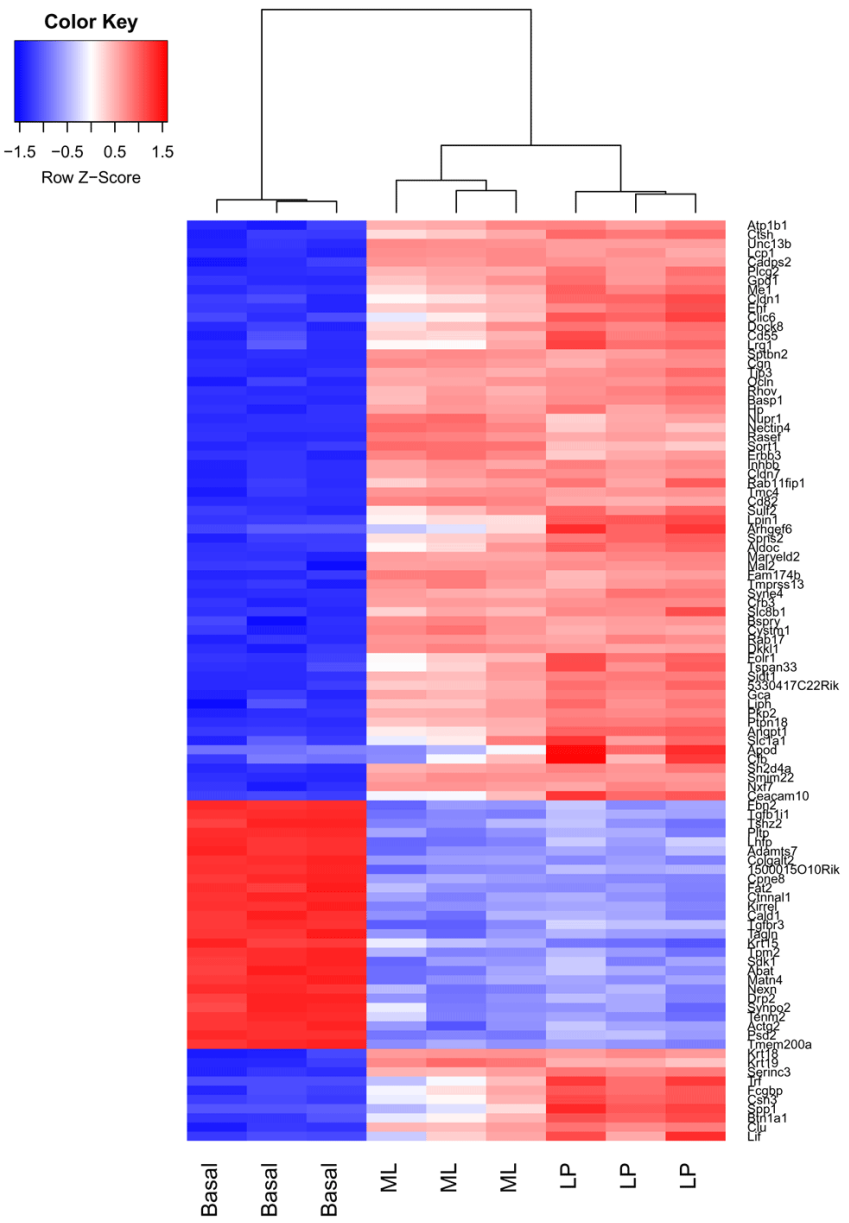
## Projets métagénomiques

4900 projets sur NCBI (avril 2018)

- Sable de plage
- Moustique
- Corail
- Glace
- Air de la ville de Singapour
- Surface de la cuvette des toilettes
- Fromages
- ...

# Pourquoi faire du RNAseq ?

- **L'analyse d'expression différentielle** (différence d'expression dans des conditions précises) au niveau transcriptomique.
- Etude de **l'épissage alternatif** (isoformes) et recherche de nouveaux transcrits.
- **Recherche d'allèles spécifiques** et quantification de leur expression.
- **Construction d'un transcriptome** de novo pour les organismes non modèles.



# Application: Transcriptomics / RNASeq

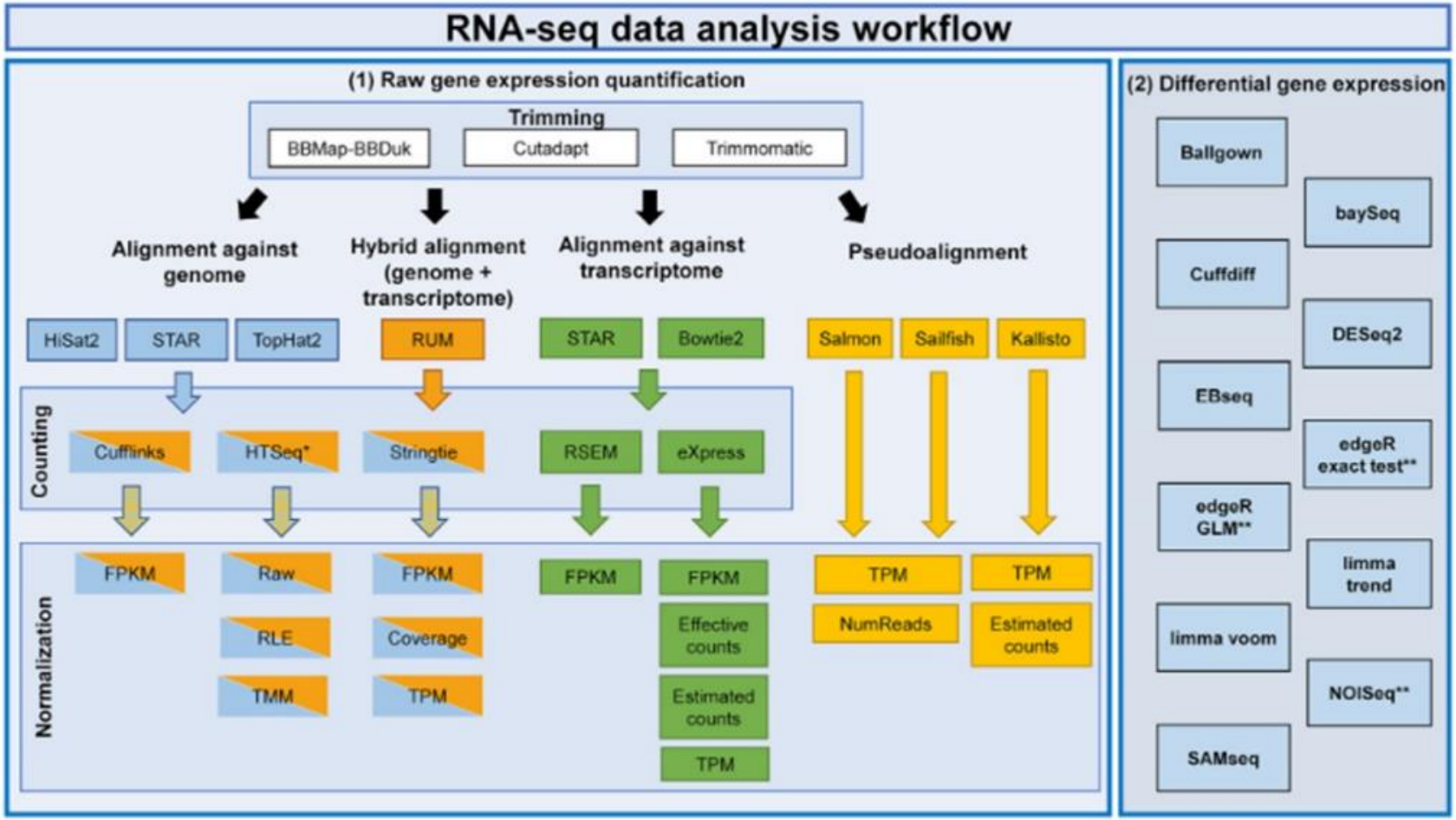



**Figure 1.** RNA-seq analysis workflow. Left panel (1) represents the raw gene expression quantification workflow. Every box contains the algorithms and methods used for the RNA-seq analysis at trimming, alignment, counting, normalization and pseudoalignment levels. The right panel (2) represents the algorithms used for the differential gene expression quantification. **HTSeq* was performed in two modes: union and intersection-strict. ***EdgeR exact* test, *edgeR GLM* and *NOISeq* have internally three normalization techniques that were evaluated separately.

# Application: Transcriptomics / RNASeq

⇒ Comparaison entre conditions expérimentales différentes

Ex:

- Comparaison plante infectée/saine
- Comparaison d'expression à différentes altitudes
- Comparaison ombre/soleil

⇒ Comparaison dans le temps (time series): cinétique

Ex:

- Cinétique d'infection de pathogènes
- Étude du rythme circadien sur l'expression de gènes

=> logiciels dédiés pour ce type de problèmatique

# Studying Genetic Diversity using Single Nucleotidic Polymorphism (SNP)

# Population Variation

Rice pathoogene ?

# Theory of evolution - C. Darwin 1859

add images

**Understanding how individuals of a same species vary**

✓ **Variations** between individuals

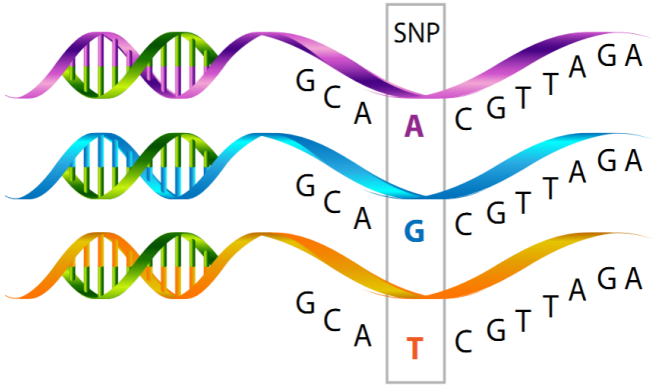
✓ **Natural selection** in a population

+ Each individual = unique combination of traits

+ Inherited varaitions that confer an advantage (increasing an organism's chance of survival) will be pass to offspring

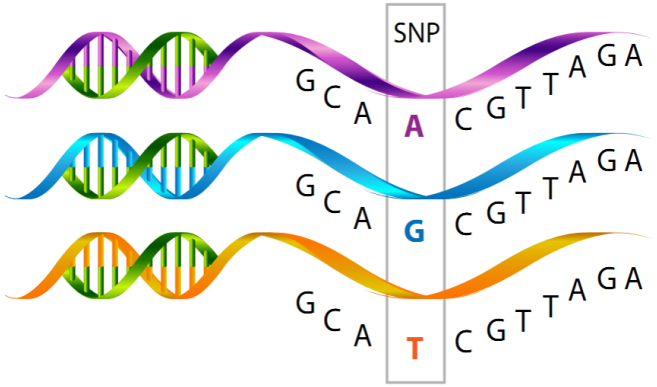
# Mutations & Variations as main source of genetic diversity

## Single Nucleotide Polymorphism

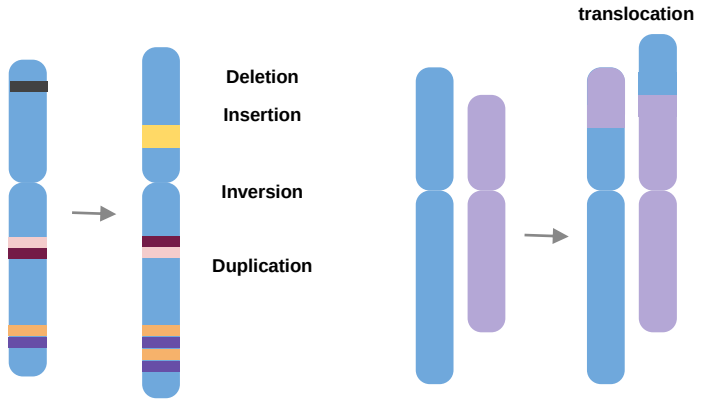


# Mutations & Variations as main source of genetic diversity

## Single Nucleotide Polymorphism

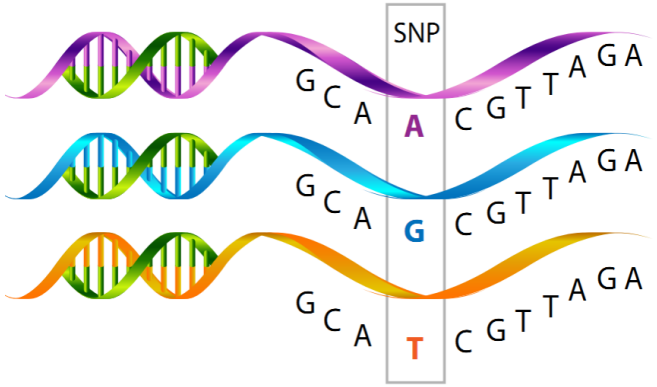



## Structural Variations

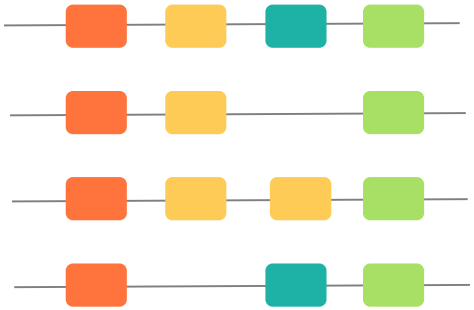


# Mutations & Variations as main source of genetic diversity

## Single Nucleotide Polymorphism




## Structural Variations

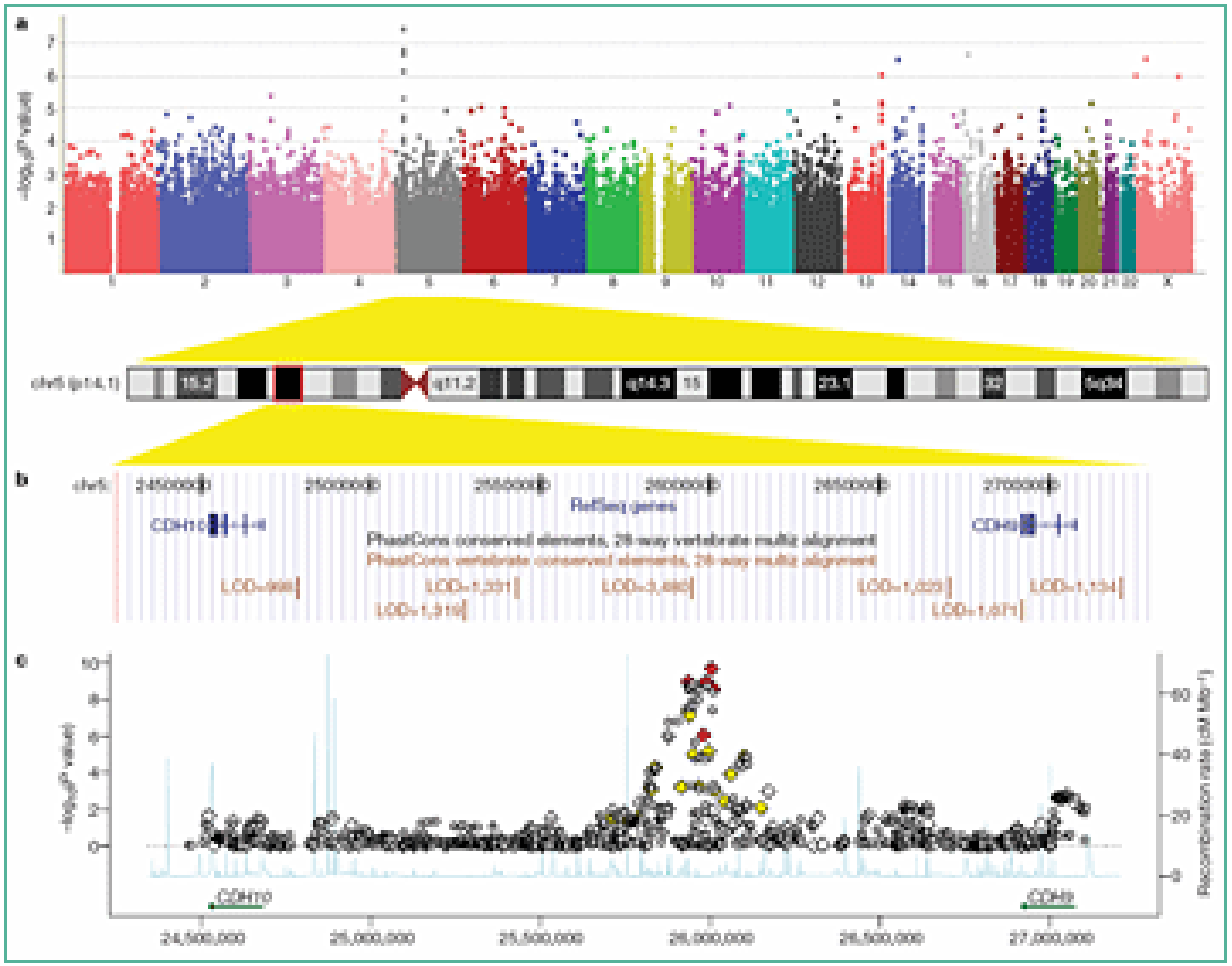


**Presence Absence Variation (PAV)**

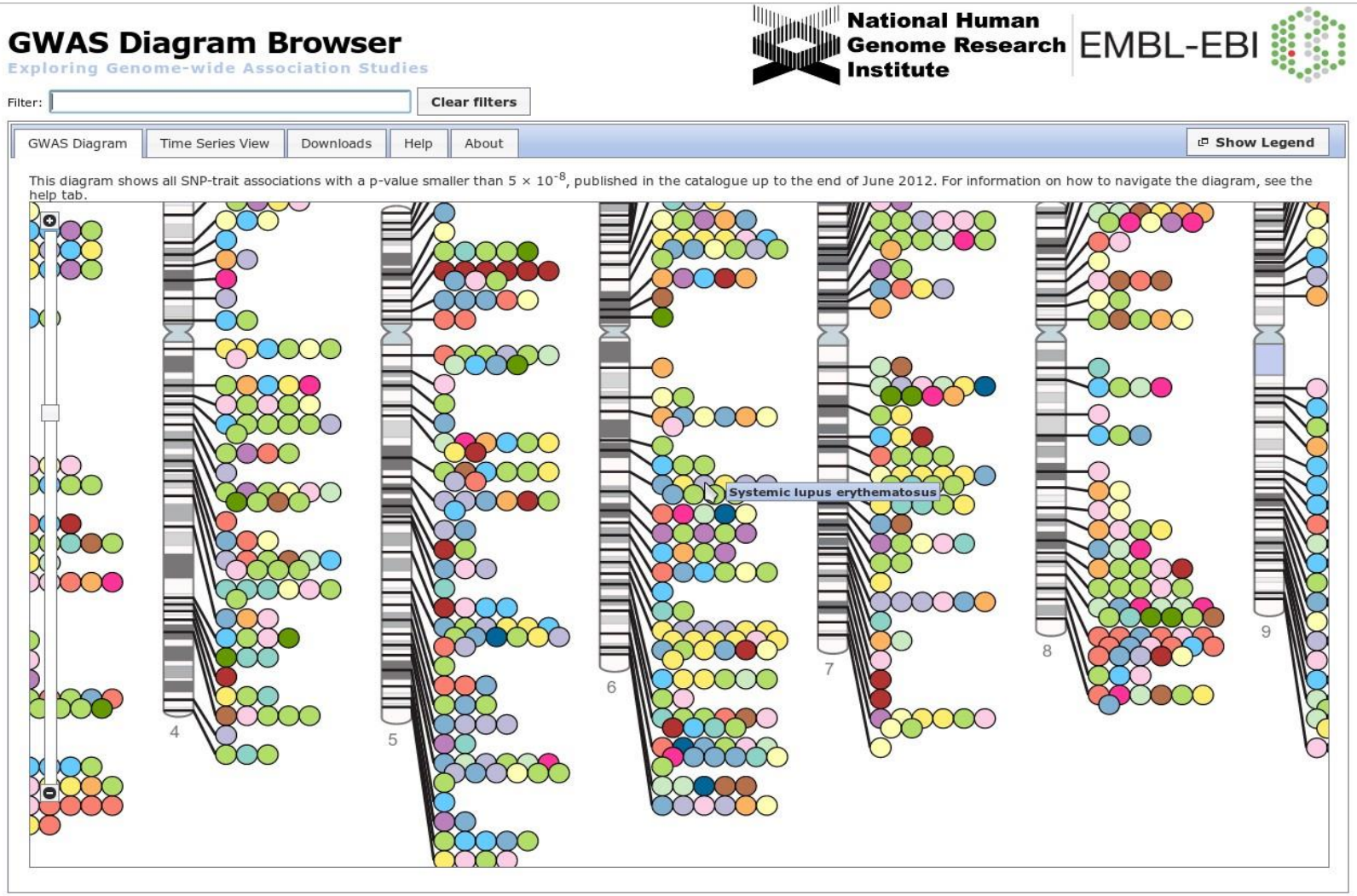
Deletion, duplication, copy number variation, mobile element insertion

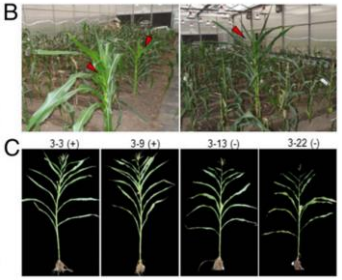
# GWAS (Genome Wide Association Studies)

Pour chaque marqueur, statistiques d'association entre le génotype et le phenotype
=> Manhattan plot



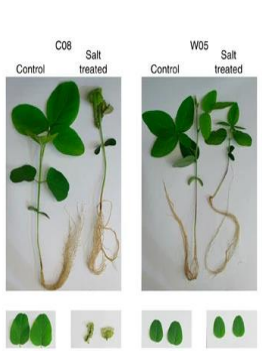
# GWAS (Genome Wide Association Studies)





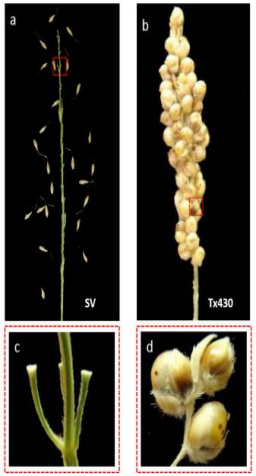

FromYang et al., 2013


From Li et al. 2012

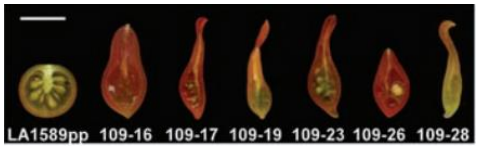



From Qi et al. 2014

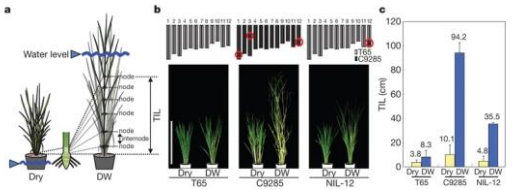
FromYang et al., 2014



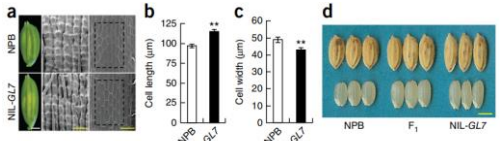

From Lin et al. 2012




From Xiao et al. 2008



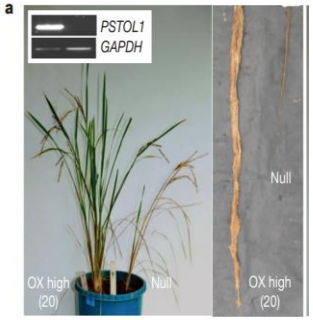

From Hattori et al. 2009



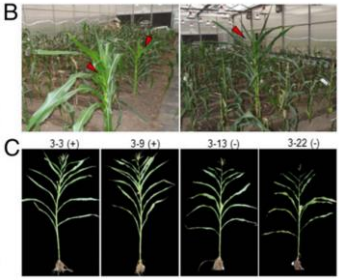

From Wang et al. 2015


From Bai et al. 2017


From Xu et al. 2006

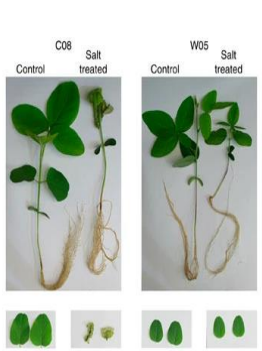



From Gamuyao et al. 2012



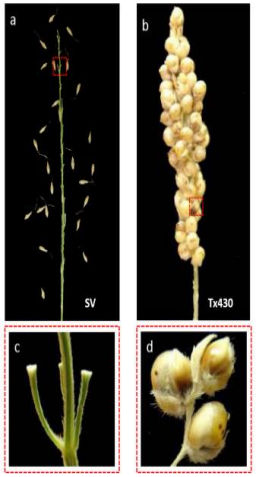

FromYang et al., 2013


From Li et al. 2012




From Qi et al. 2014

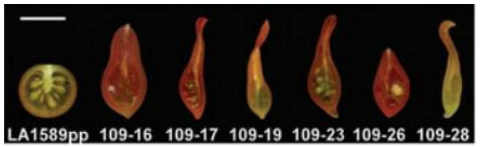
FromYang et al., 2014



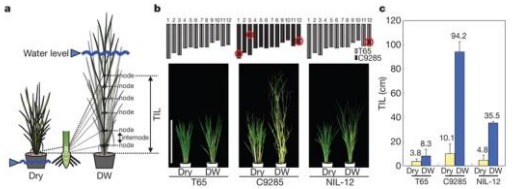

From Lin et al. 2012

**Is One Reference genome enough**

**to capture all genetic diversity ?**

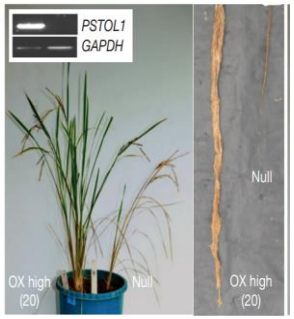



From Xiao et al. 2008

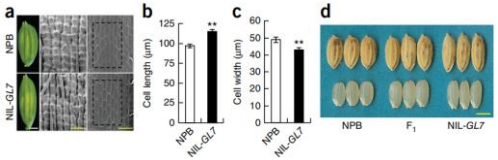



From Hattori et al. 2009


From Xu et al. 2006



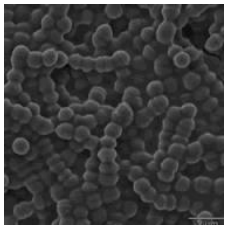

From Gamuyao et al. 2012




From Wang et al. 2015


From Bai et al. 2017

# Gene number variations within a species



*Streptococcus agalactiae*

## Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[a], Jonathan Crabtree[a], Amanda L. Jones[g], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a],

*Tetellin et al., 2005*

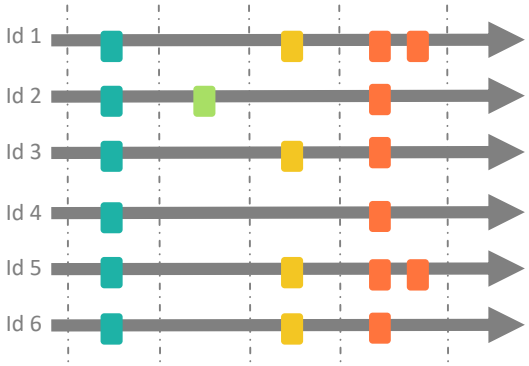
- 8 strains sequenced
- SNP variations

Large number of genes not shared between isolates

20% genome variability and 80 % shared by all isolates

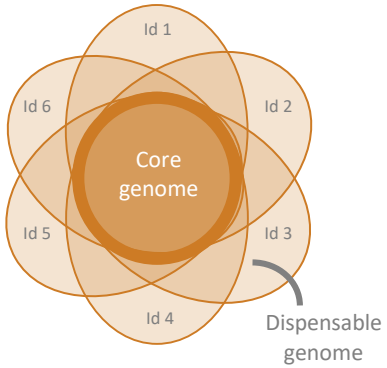
***Pangenome concept***

# Pangenome concept

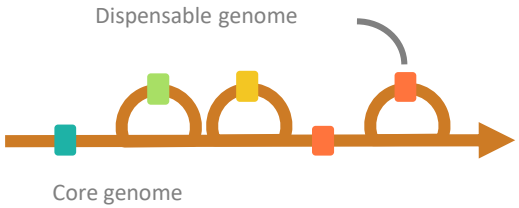



## Pangenome

Collection of genes or sequences found in all individuals of a population (intra or inter species)



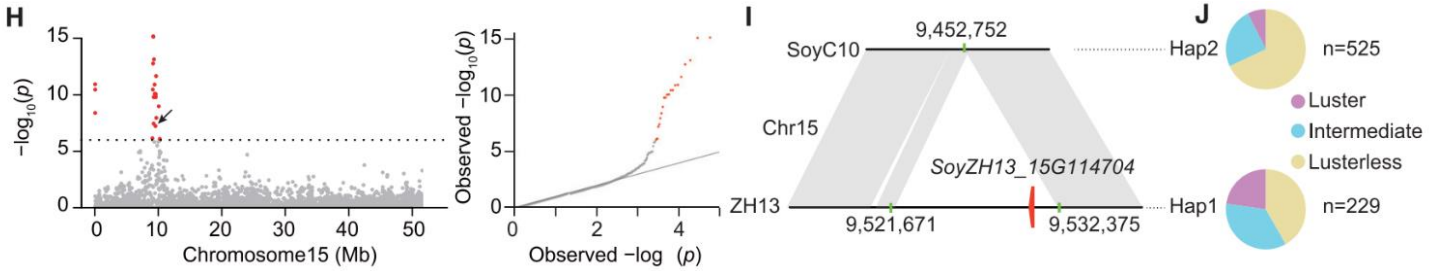

- **Core genome** : present in all individuals
- **Dispensable genome** : absent from one or several individuals (also called variable, accessory,...)



# What's else ?

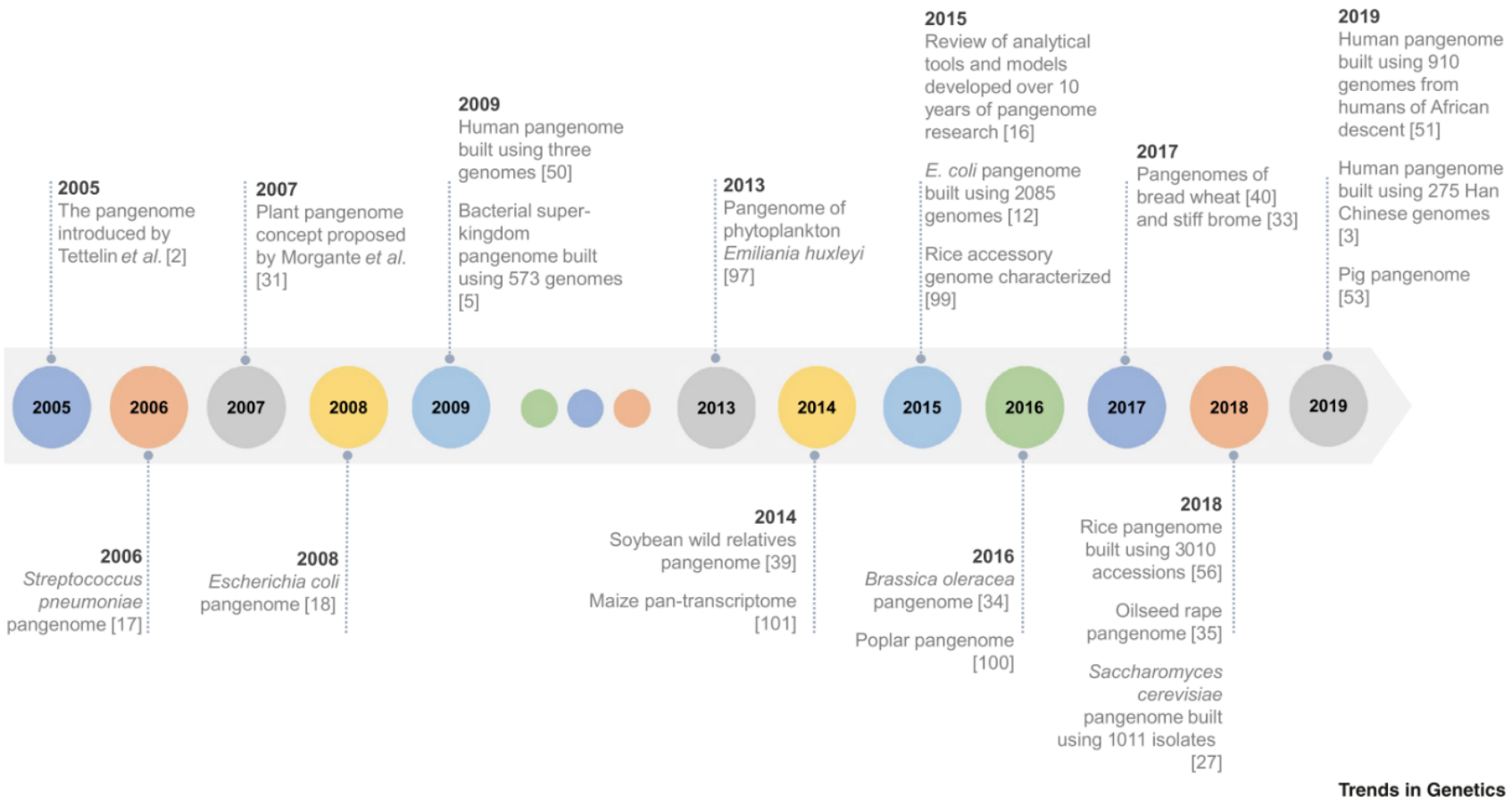
- 12,150 genes absent from the reference (18 cultivars)

- 27,175 genes absent from the reference (26 cultivars)



# From the first pangenome analyse by Tettelin & al.

Over 20 eukaryotic pangenomes constructed (12 Mb to 17 Gb)




Golicz et al., 2020

# How to detect SV ?


2 formations => 2 ambiances

Mode training mais bcp
de pratique par soi meme
ou collectivement

Mode projet

Des données différentes pour les 2 groupes et des analyses légèrement différentes !!!

C est le but de la formation

**=> Détecter des variants (SNP, variants structuraux) à partir de données de séquençage short et long reads.**

2 formations => 2 ambiances

Mode training mais bcp de pratique par soi meme ou collectivement

Mode projet

Des données différentes pour les 2 groupes et des analyses légèrement différentes !!!

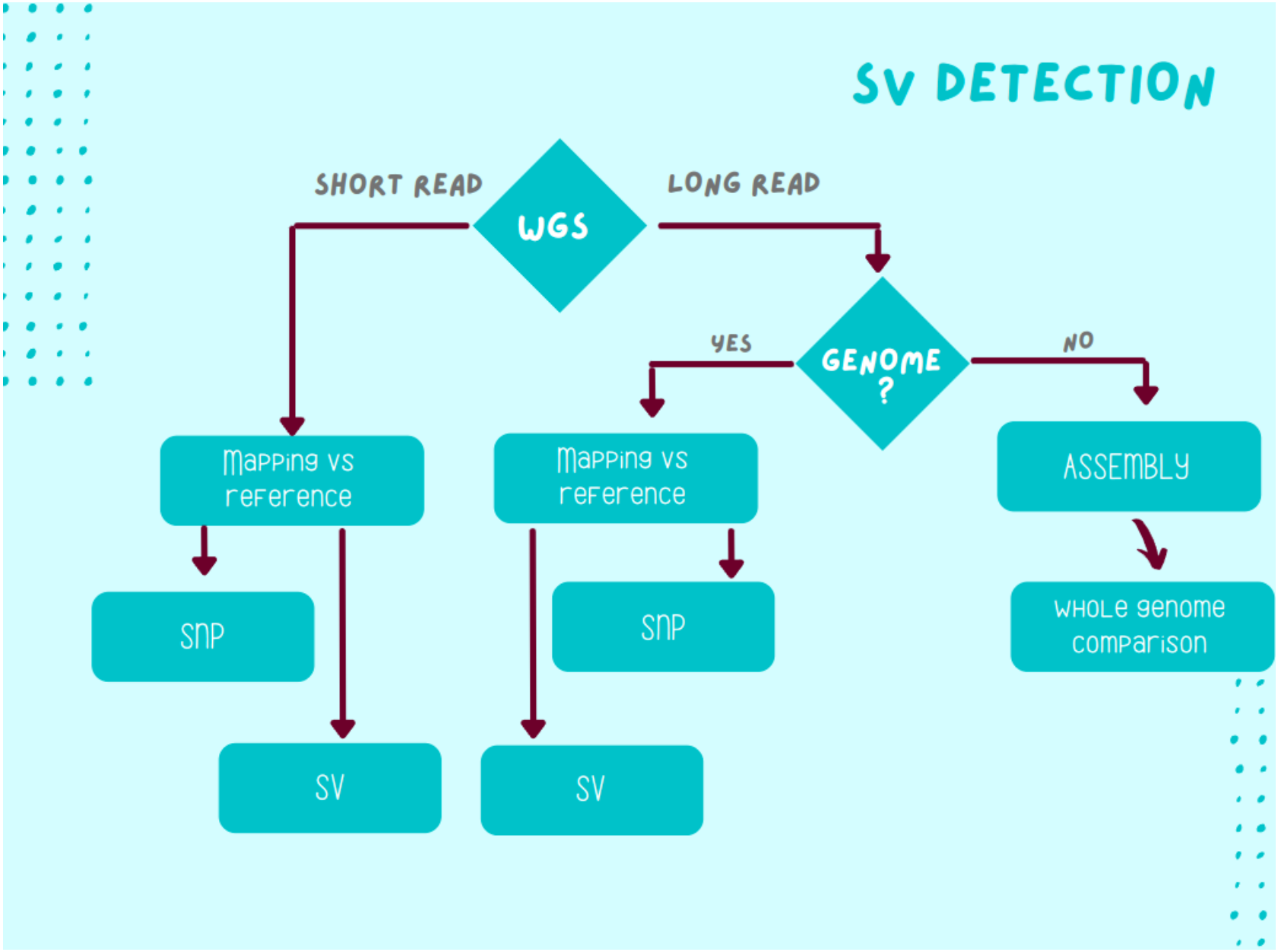
# How to detect SV ?

C est le but de la formation

=> Protocole

=> Je dois réaliser mon analyse après avoir suivi une formation, comment je fais

# Un même plan de bataille... ou pas !!!



**Training** 

**#data**

# What data will we use for our training ?


From
Wikimedia

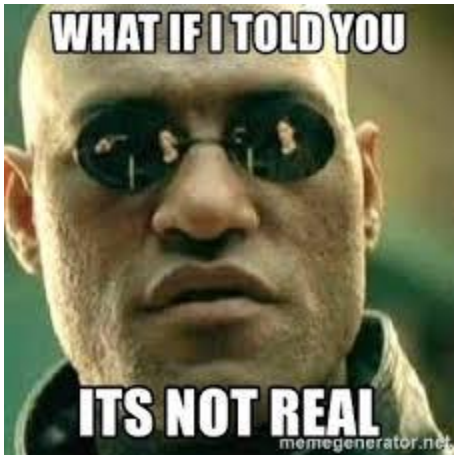
Diploid Asian Rice, *Oryza sativa*

*Chara*

# What data will we use for our training ?

Diploid Asian Rice, *Oryza*

From
Wikimedia



# What data will we use for our training ?

Diploid Asian Rice, *Oryza*

From Wikimedia

add images

1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones
3. Introduce
    - SNP (1-10%),
    - indel (10b-10kb),
    - duplications...
4. Generate short & long reads for each clone...
5. ~~Torturing students with these data~~

illumina®

Oxford NANOPORE™ Technologies

# Projet SNP

MISSION IMPOSSIBLE

NOM DE CODE : “PROJET SNP”

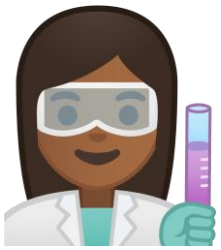
Votre mission si vous l’acceptez…

# #LIEU : Burkina Faso

# #MISSION :

Le Docteur kezako, chercheuse non spécialiste en bioinformatique a réalisé une longue prospection ***en Afrique.***

Elle a notamment ramené des échantillons d'ignames (elle pense que c'est de l'igname) qui présentent une diversité phénotypique particulièrement intéressante dans le contexte climatique actuel.



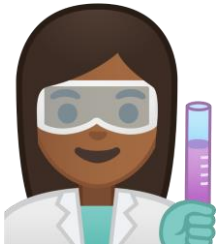
- Avons nous collecté une nouvelle espèce d'igname ?
- Ou avons nous collecté des ignames domestiqués ? sauvages ?

# #MISSION :

Malgrè son emploi du temps très chargée, **elle a séquencé 10 individus**

Elle met à votre disposition ces données de séquençage ainsi 5 collègues qui pourront vous assister mais leur temps est précieux car ils ont une autre mission à mener en parallèle…

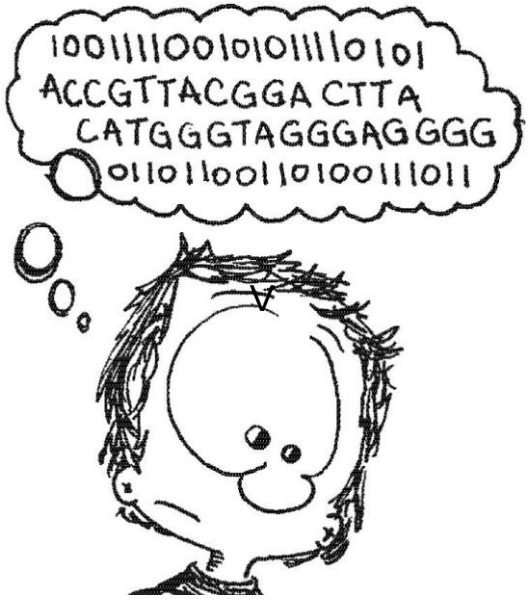
Dominique



Je compte sur vous !!!!

# #DATA :

Décrire où ils trouvent les données





# A vous de jouer !

# Metagenomic

# Bioinformatics resources

# Comment utiliser Linux?

- 2 façons d'utiliser linux :

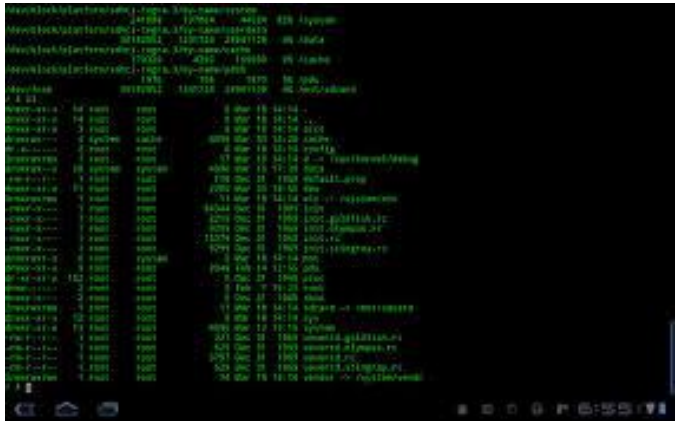
  en ***mode graphique***

# Comment utiliser Linux?

- 2 façons d'utiliser linux :

  en ***mode graphique***

  en ***mode console***

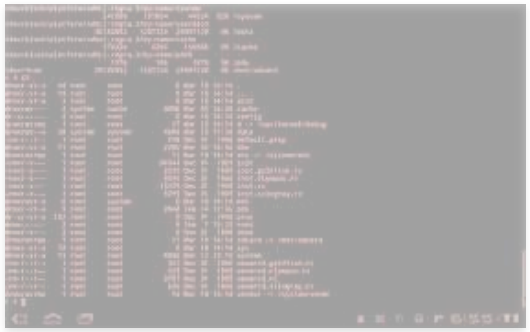


# Pourquoi utiliser Linux ?

- Nombreux programmes rapides & puissants
- Facile de lier des commandes/programmes entre eux (workflow)
- Nombreux outils bioinformatique disponibles
- Pas besoin de ressources matérielles importantes
- 90% des serveurs fonctionnent sous Linux

# Pourquoi utiliser Linux ?

- Nombreux programmes rapides & puissants
- Facile de lier des commandes/programmes entre eux (workflow)
- Nombreux outils bioinformatique disponibles
- Pas besoin de ressources matérielles importantes
- 90% des serveurs fonctionnent sous Linux



**Pas d'interfaces graphiques**

**Convivialité de la ligne de commande ?**

# Pourquoi utiliser Linux ?

Nécessité de la pratique et de l'expérience

⇔ **Investissement non négligeable pour de bons résultats rapidement**

# Let's discover Jupyter !

***Working environment***

# What is jupyter book ?

- One of the most popular tool among data scientists to perform data analysis
- Provides a complete environment in which numerous programming languages can be use through a simple web browser

ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala

# Why use jupyter book ?

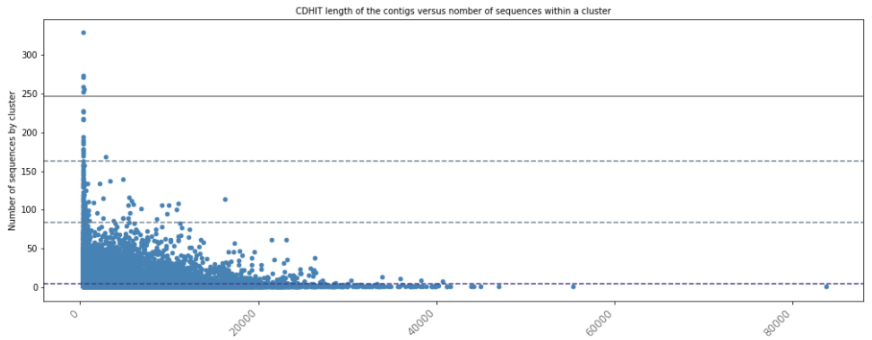
An unique interface/file where text,code and output codes can be mixed :

- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook

*All ctgs*

Entrée [30]:

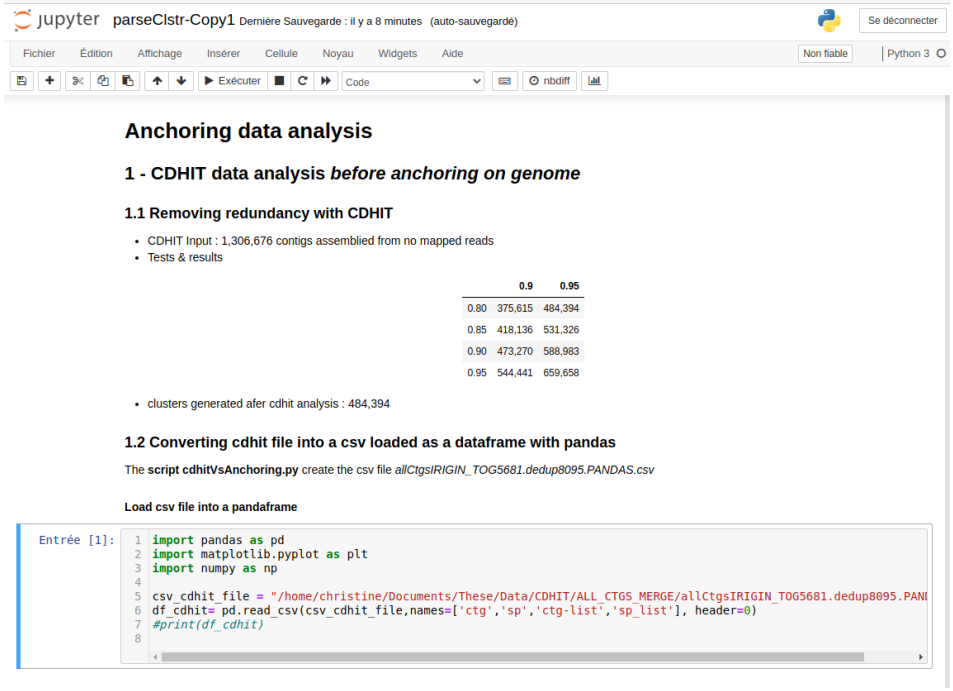
```
plt.figure(figsize=(17, 6))
ax = plt.gca()
df_cdhit[(df_cdhit.sp == 'Ob')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
df_cdhit[(df_cdhit.sp == 'Og')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
plt.axhline(y=5, color='darkslateblue', linestyle='--')
plt.axhline(y=84, color='slategrey', linestyle='--')
plt.axhline(y=163, color='slategrey', linestyle='--')
plt.axhline(y=247, color='grey', linestyle='-')

plt.title("CDHIT length of the contigs versus nomber of sequences within a cluster", fontsize=10)
plt.xlabel('Cluster length')
plt.ylabel('Number of sequences by cluster')
plt.xticks(
    rotation=45,
    horizontalalignment='right',
    fontweight='light',
    fontsize='12',
)
plt.show()
```



# Why use jupyter book ?

An unique interface/file where text,code and output codes can be mixed :

- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook
- explanations, formulas, charts can be added



# Lab notebook for science data ?


- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

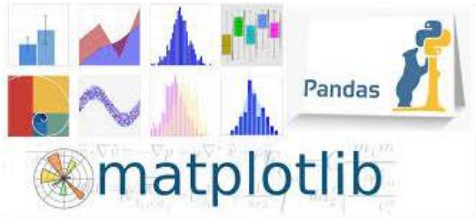
# How to become a super datascientist ?

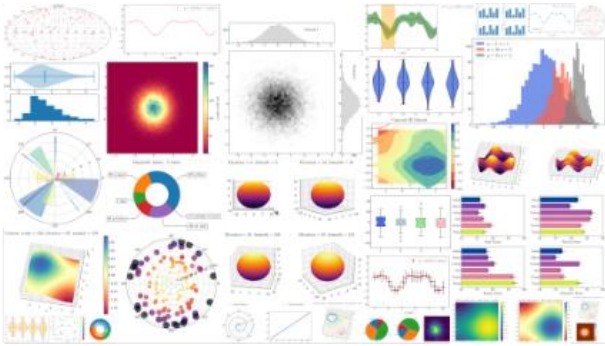
- facilement importer des fichiers tabulés dans des dataframes, similaires aux dataframes sous R.

  (et exporter)

- manipuler ces tableaux de données / DataFrames

- facilement tracer des graphes à partir de ces DataFrames grâce à matplotlib

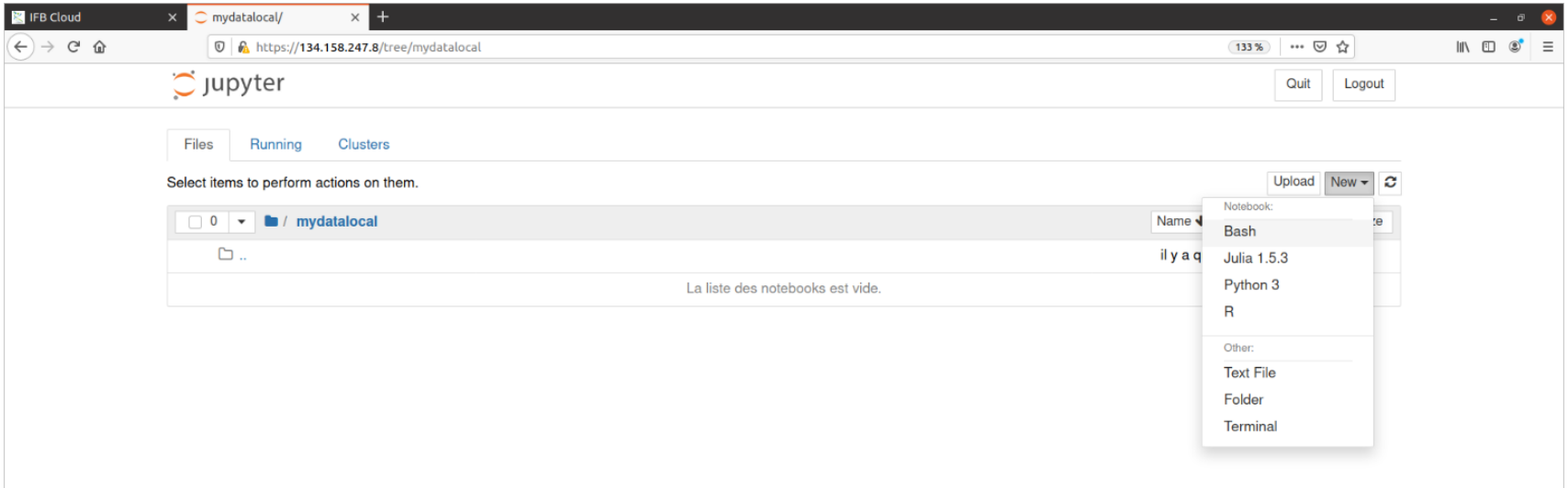
# How will you use Jupyter Notebook ?

- Launch our analyses through a jupyter book within a virtual machine launched via the IFB cloud “BIOSPHERE”

# How will you use Jupyter Notebook ?

- Launch our analyses through a jupyter book within a virtual machine launched via the IFB cloud “BIOSPHERE”

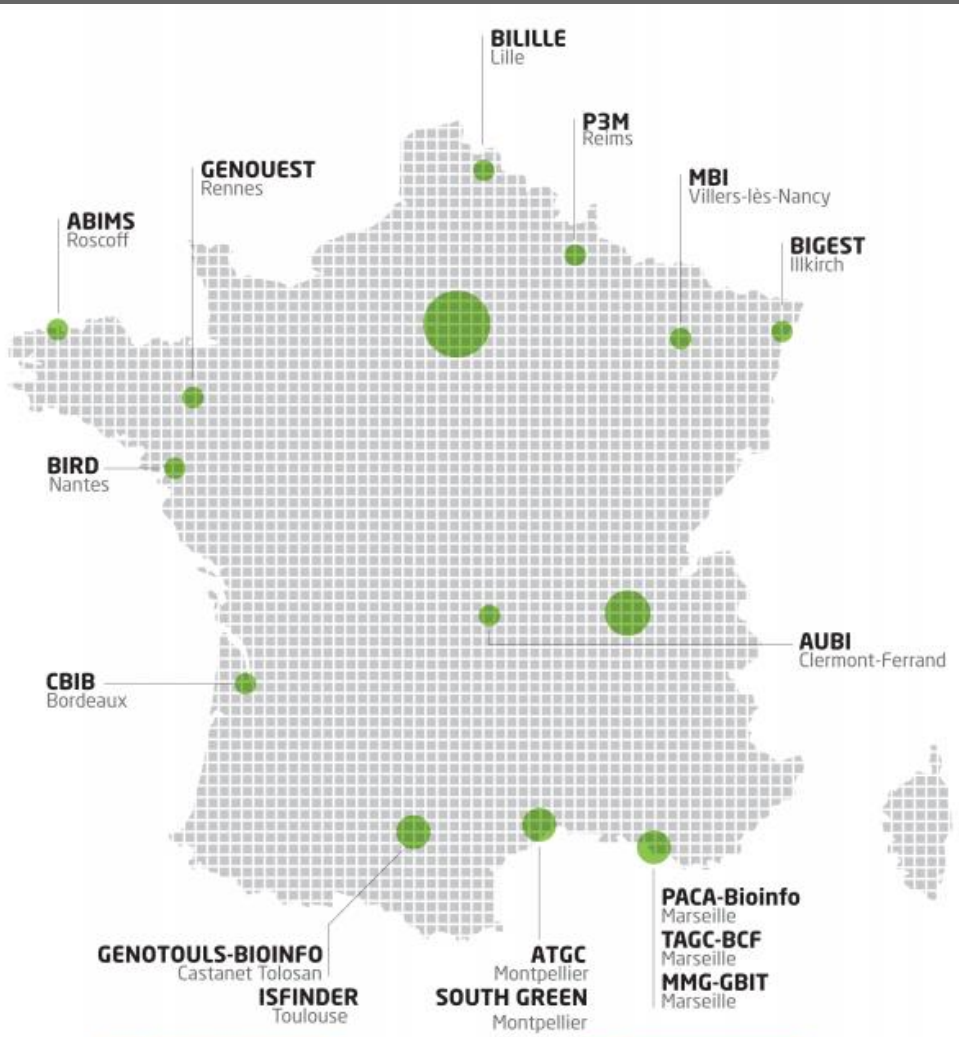
- Through this virtual machine, we will create jupyter books and execute all our analysis



# IFB ?

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

22 plateformes-membres
7 plateformes contributrices
8 équipes associées
>400 experts (~200 FTE)



# Biosphere, IFB CLOUD FOR LIFE SCIENCES

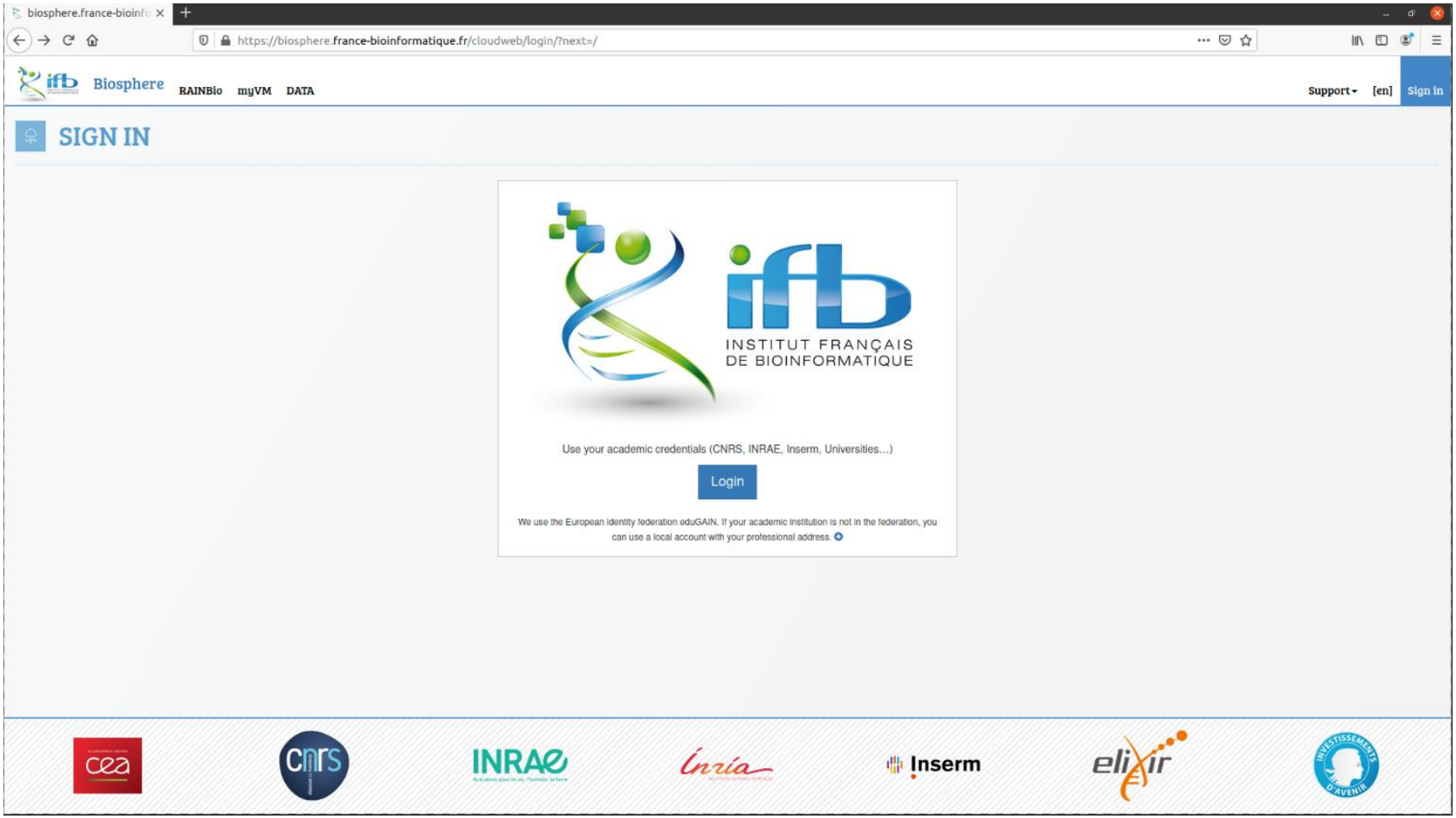
- A federation of clouds, which relies on interconnected IFB's infrastructures, providing distributed services to analyze life science data

- .Access to a large set of virtual machines (computing ressources, bioinformatics tool)

- Used for scientific production in the life sciences, developments, and also to support events like cloud and scientific training sessions, hackathons or workshops.

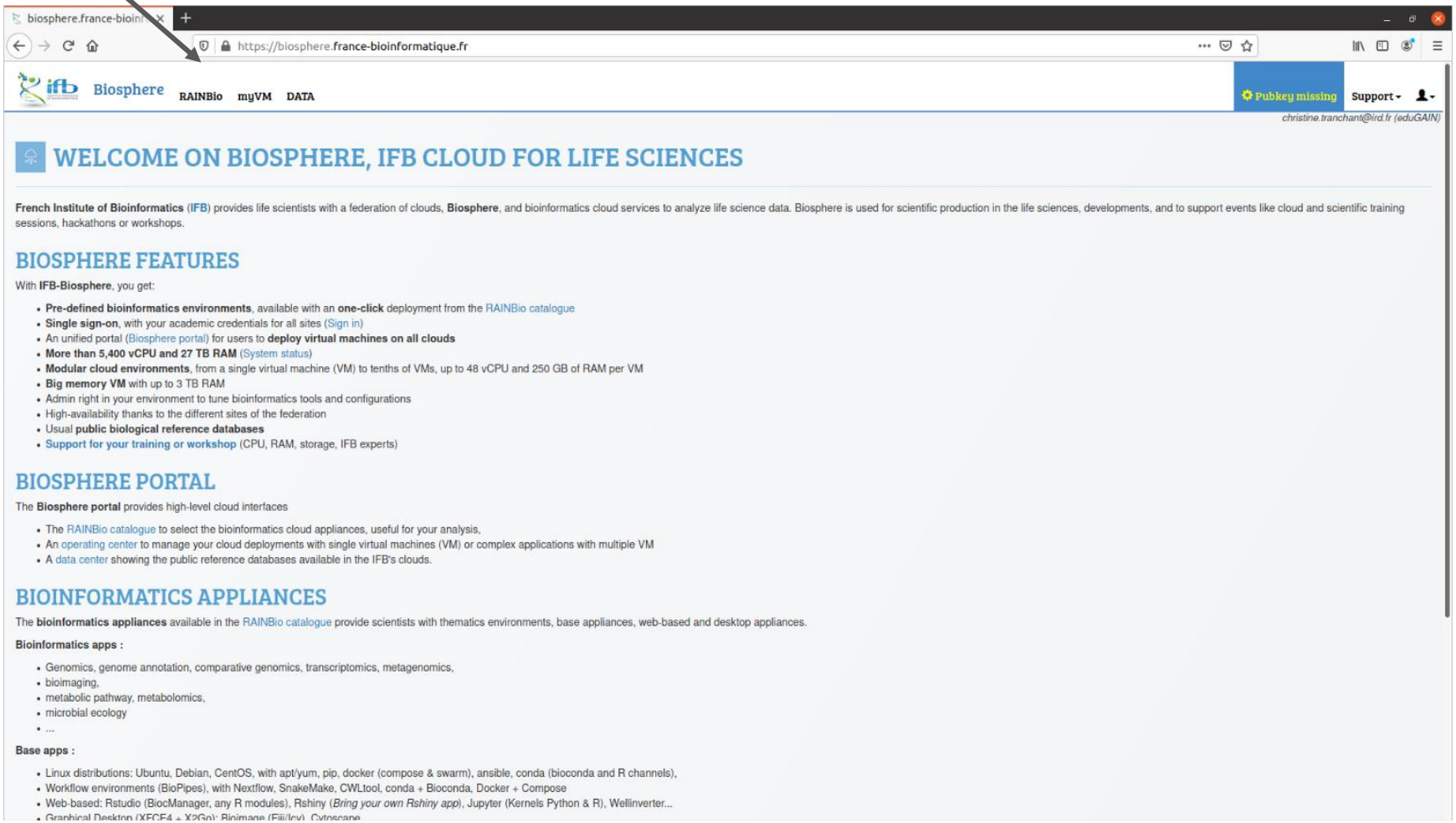
# Let's start with biosphere

- Open the biosphere website : https://biosphere.france-bioinformatique.fr/cloud/ and sign in



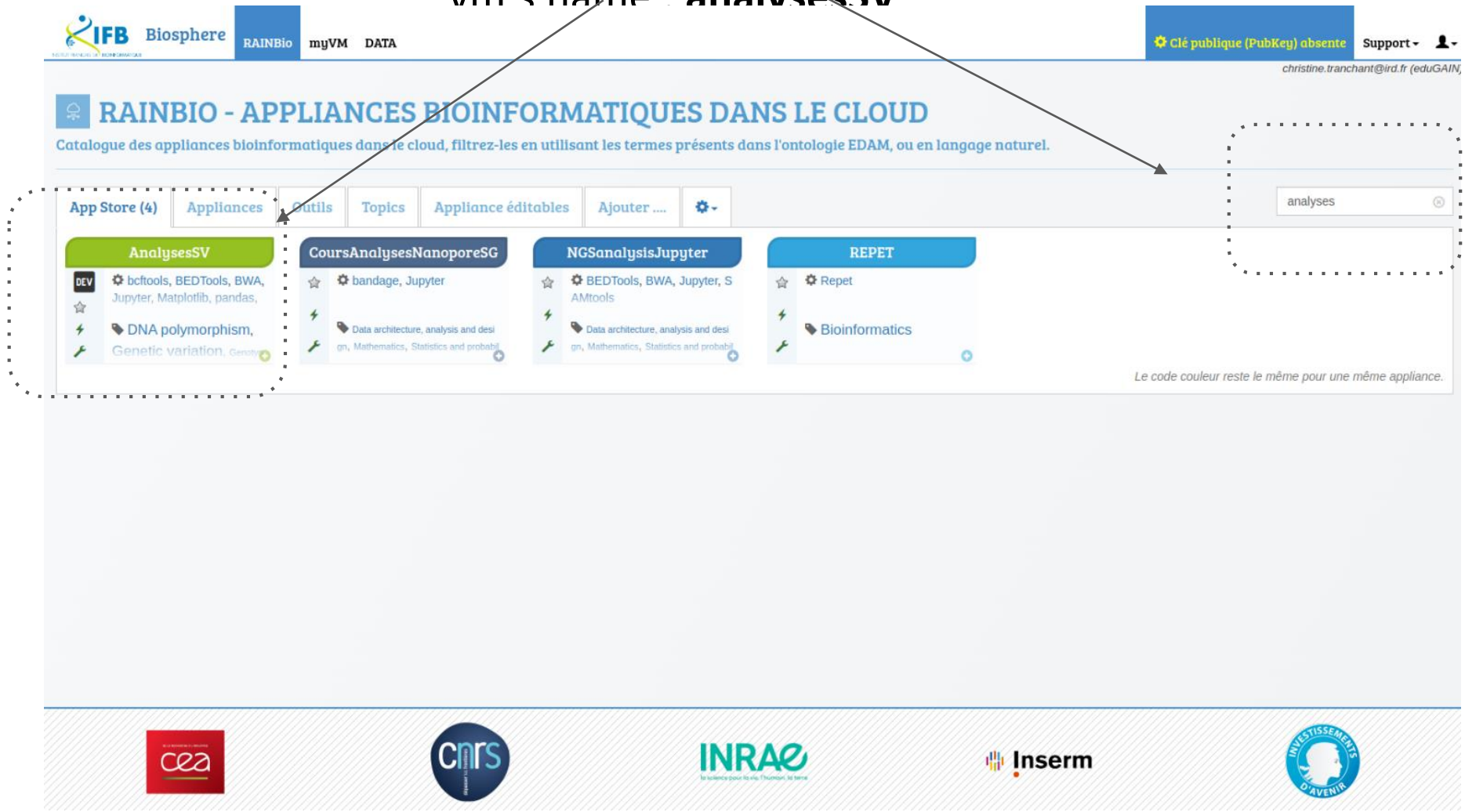
# Connected / here we are

RAINBIO catalog to access our Virtual Machine (VM)

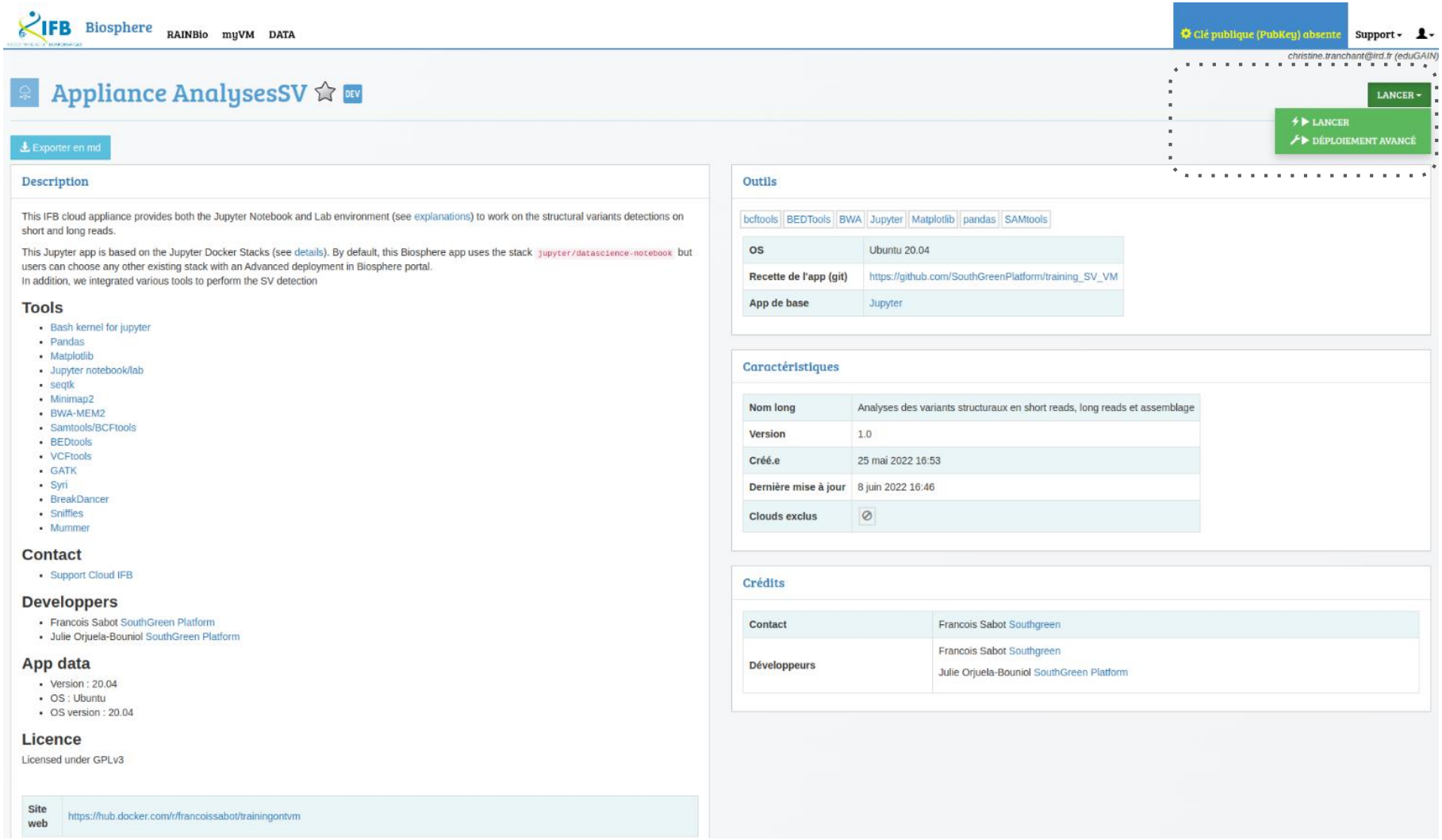


# Searching for the vm we will use

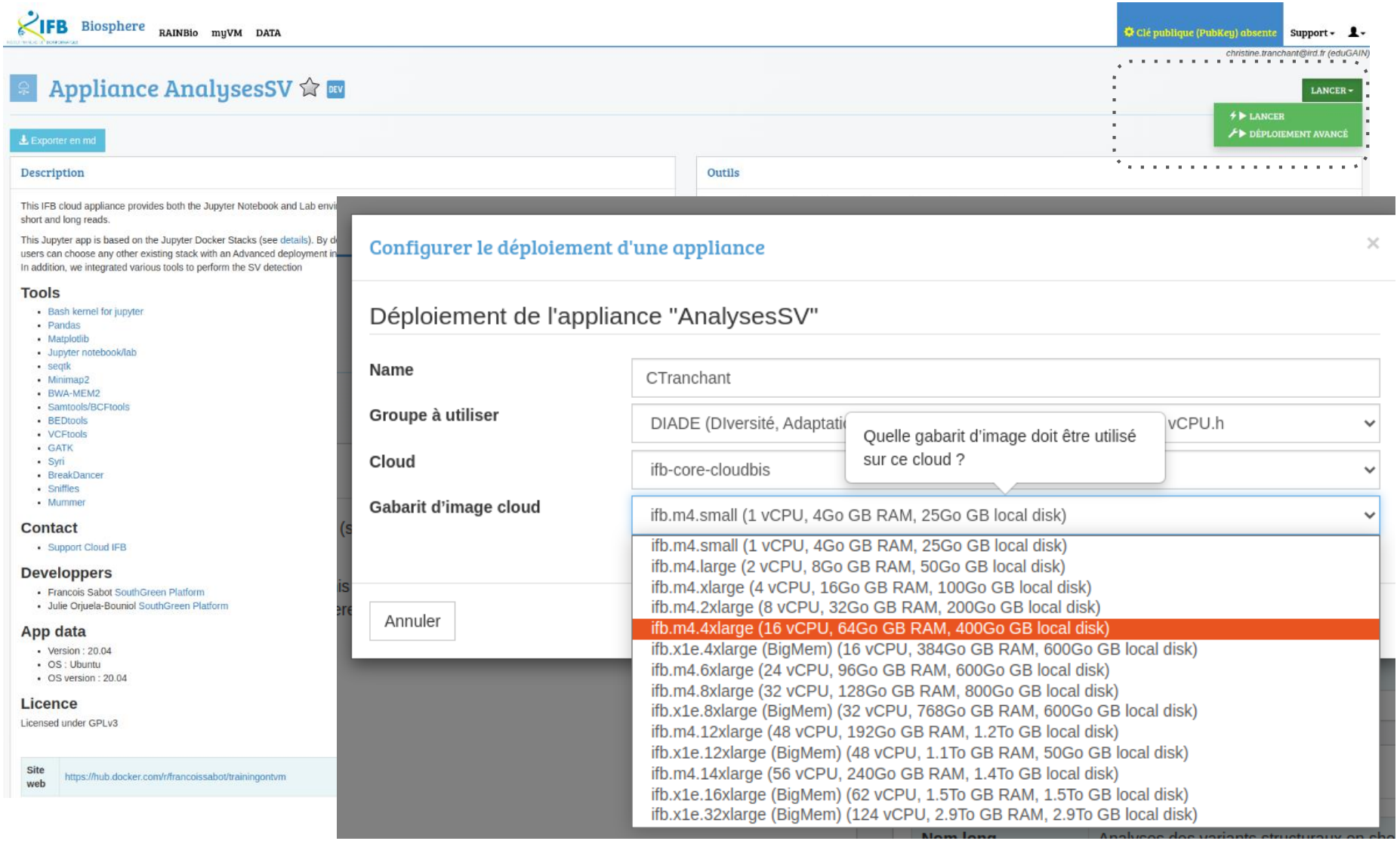
vm's name : **analysesSV**



# Let's run your vm through the cloud

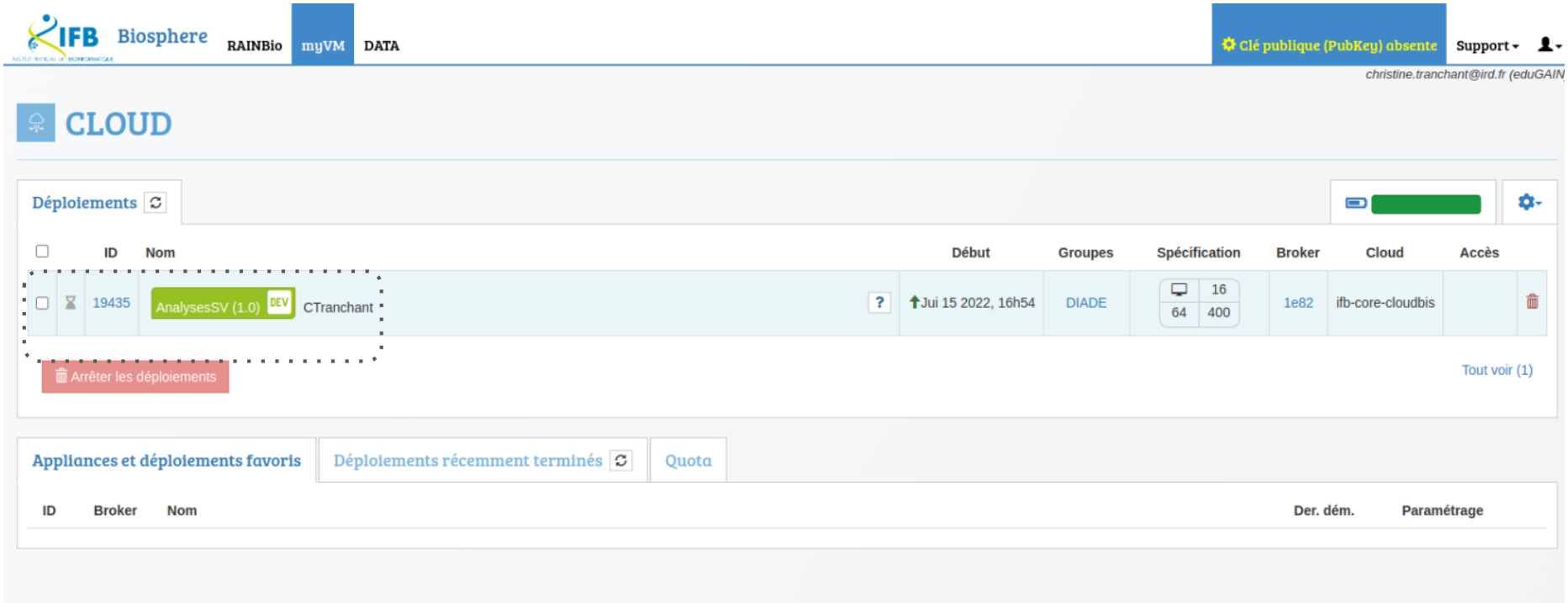


# Let’s run your vm through the cloud



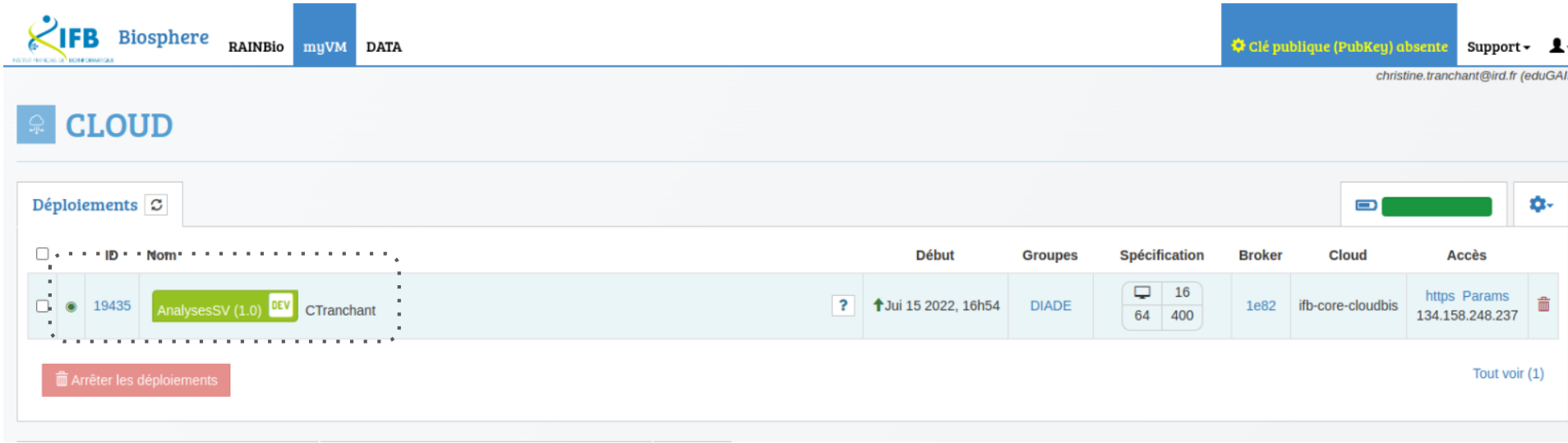
# Let’s run your vm through the cloud

Loading…



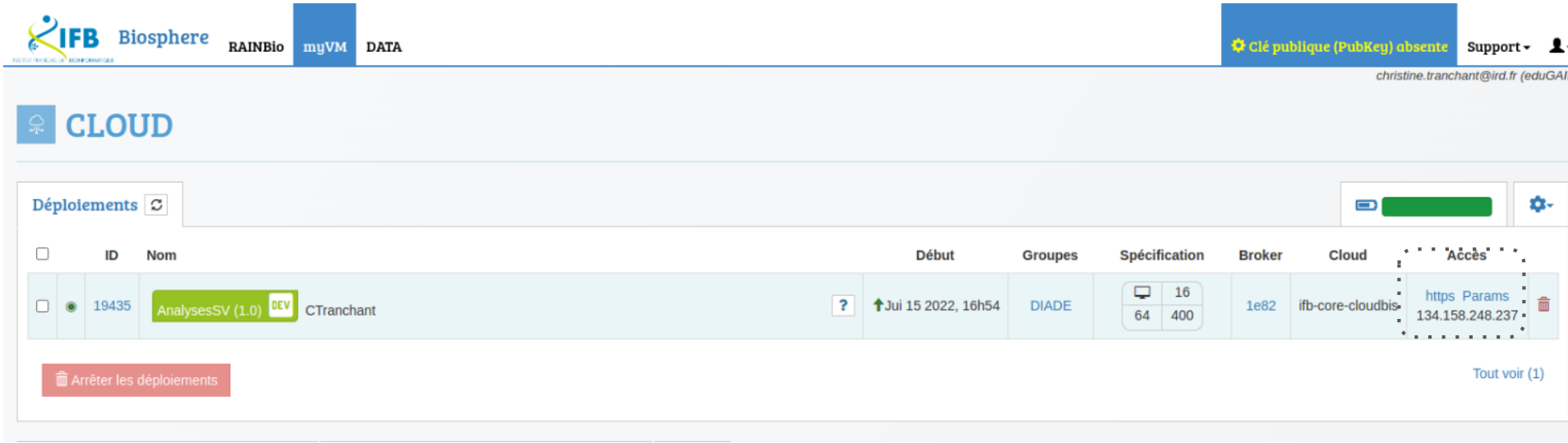
# Let’s run your vm through the cloud

ready !



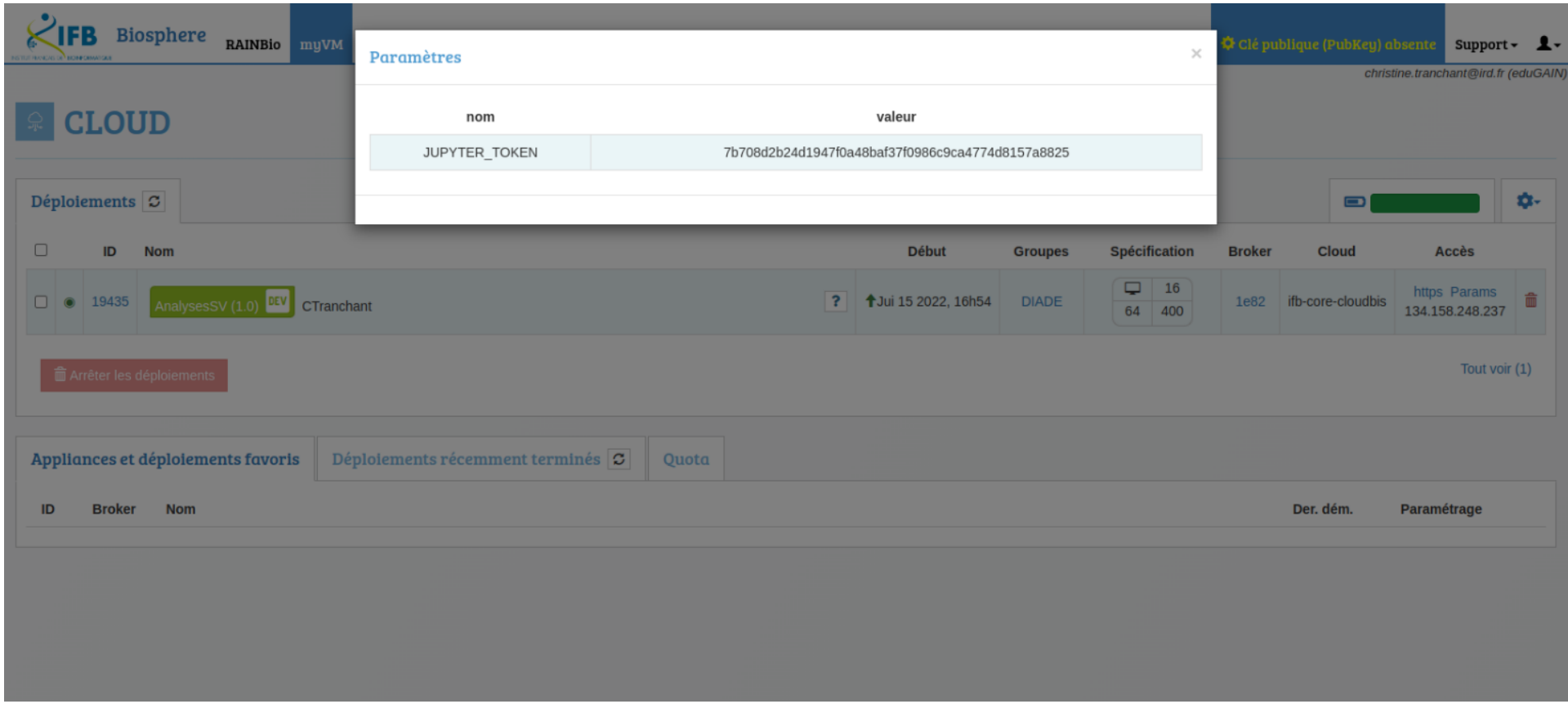
# Let’s run your vm through the cloud

get the url… link “https”



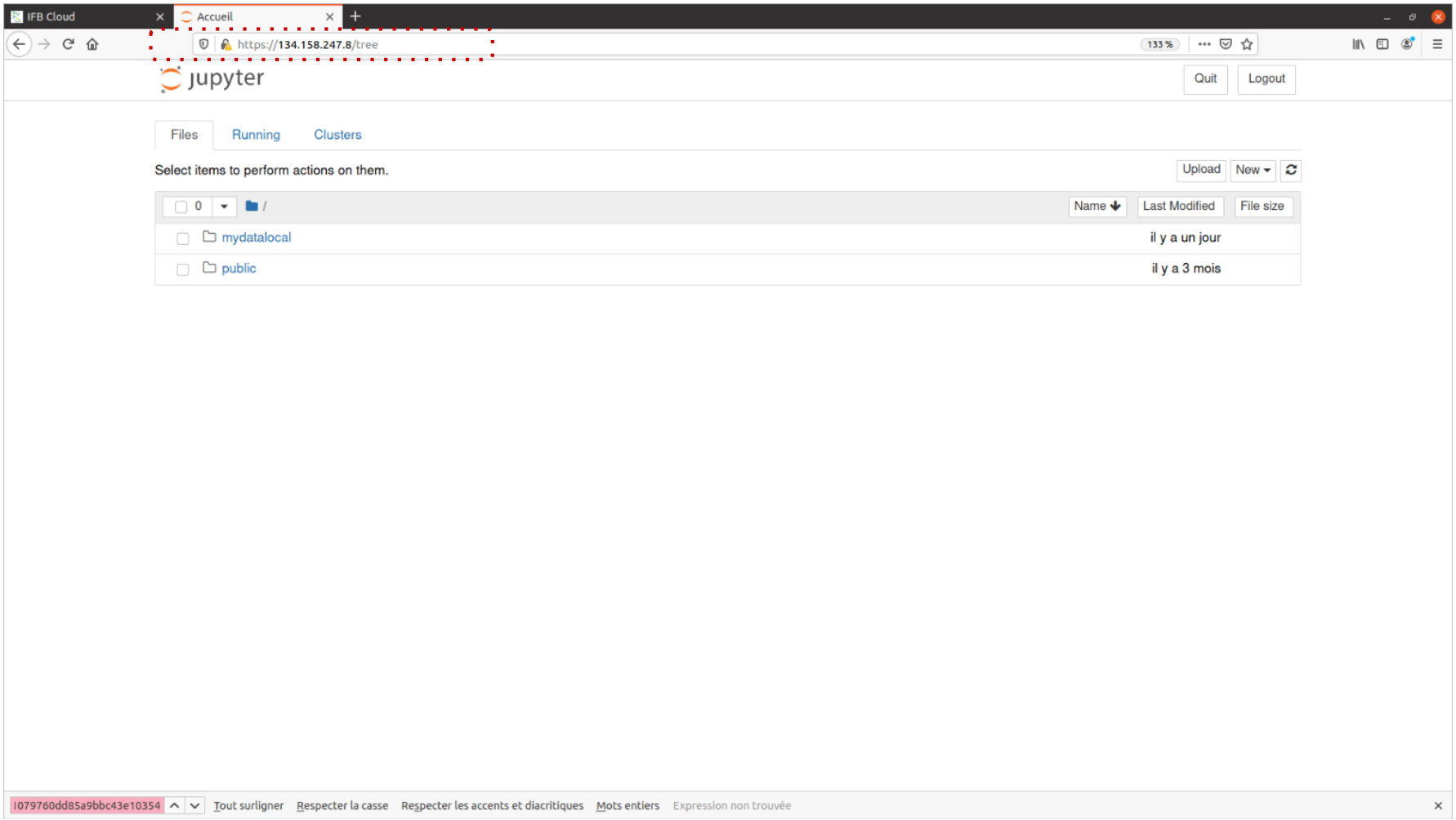
# Let’s run our vm through the cloud

Get the token identifiant… link “Params”



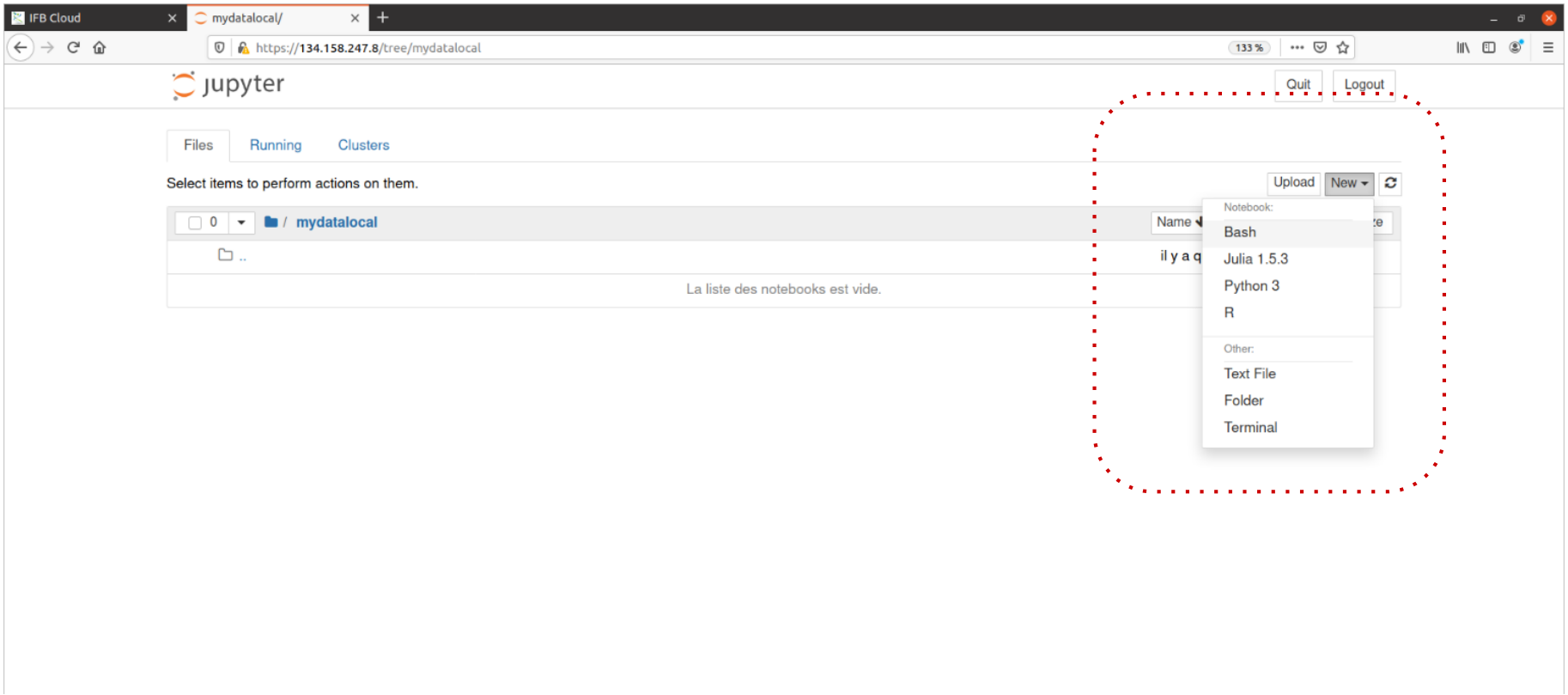
# Let’s run our vm through the cloud

Open your vm (https link) to access to your own jupyter lab



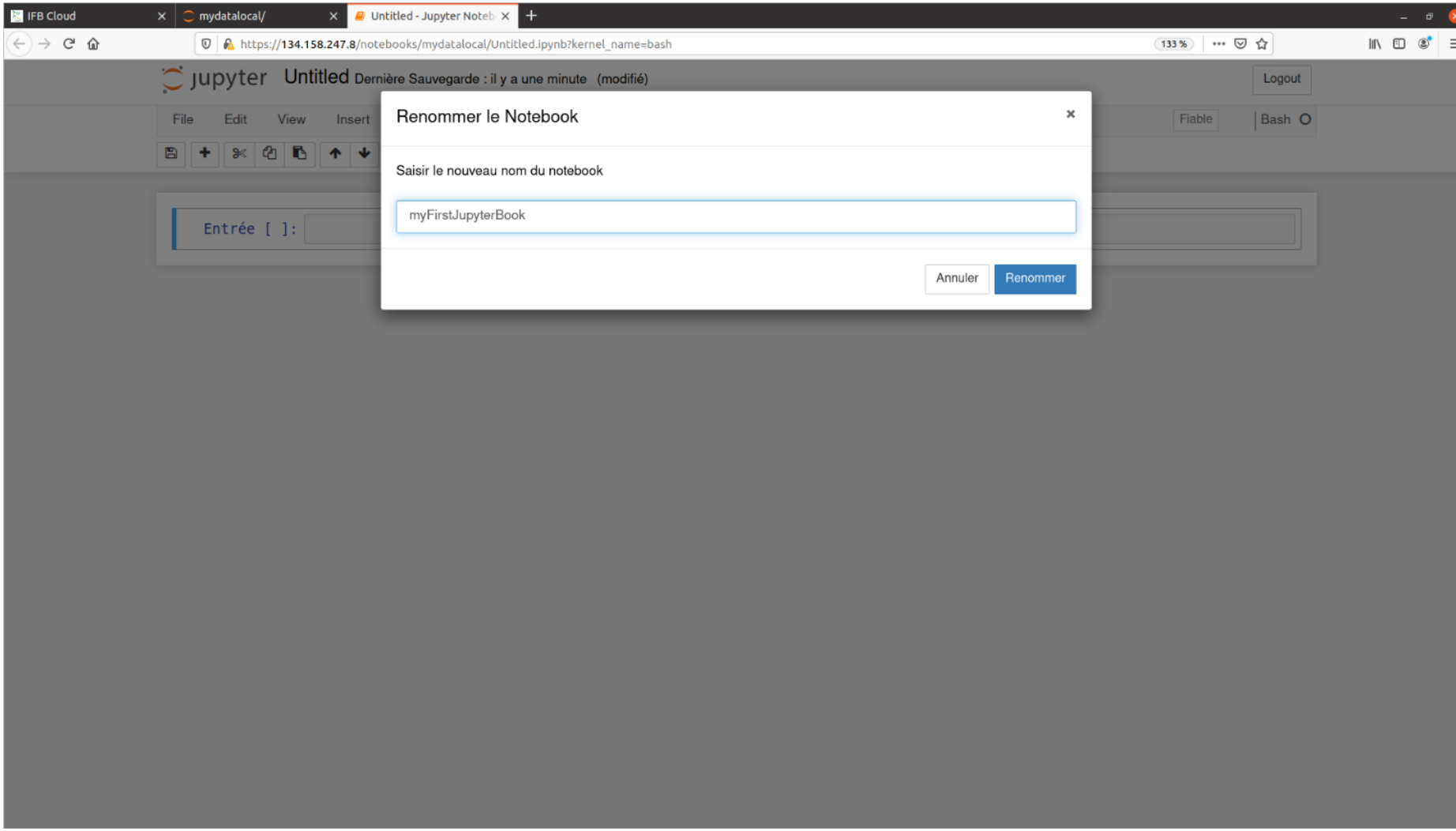
# Create your first jupyter book

Go into the directory “work” and create a new jupyter book
-> kernel : bash



# Rename your first jupyter book

myFirstJupyterBook

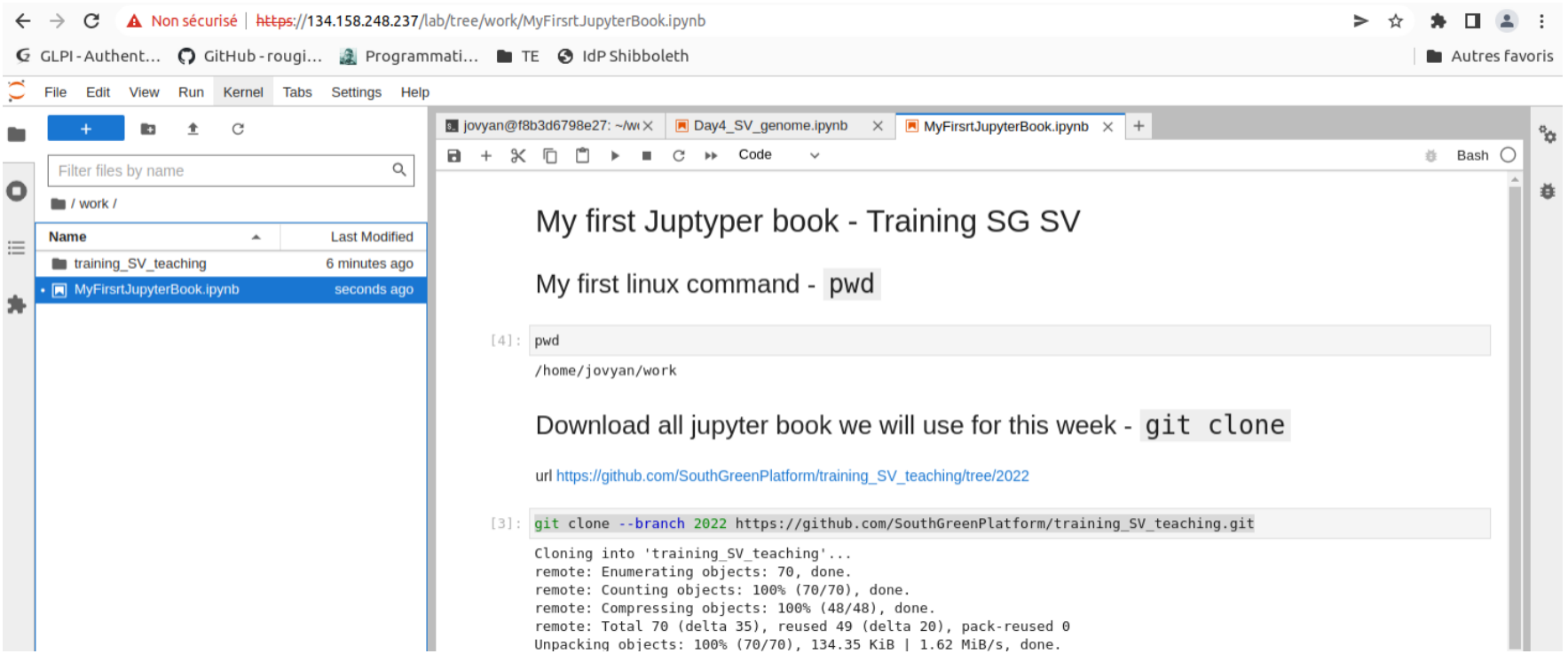


# Run your first bask command - *git clone*

- All jupyterbook used for practice are here :
  https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022

- Download all the jupyter books with the command *git clone*

git clone --branch 2022_burkina
https://github.com/SouthGreenPlatform/training_SV_teaching.git



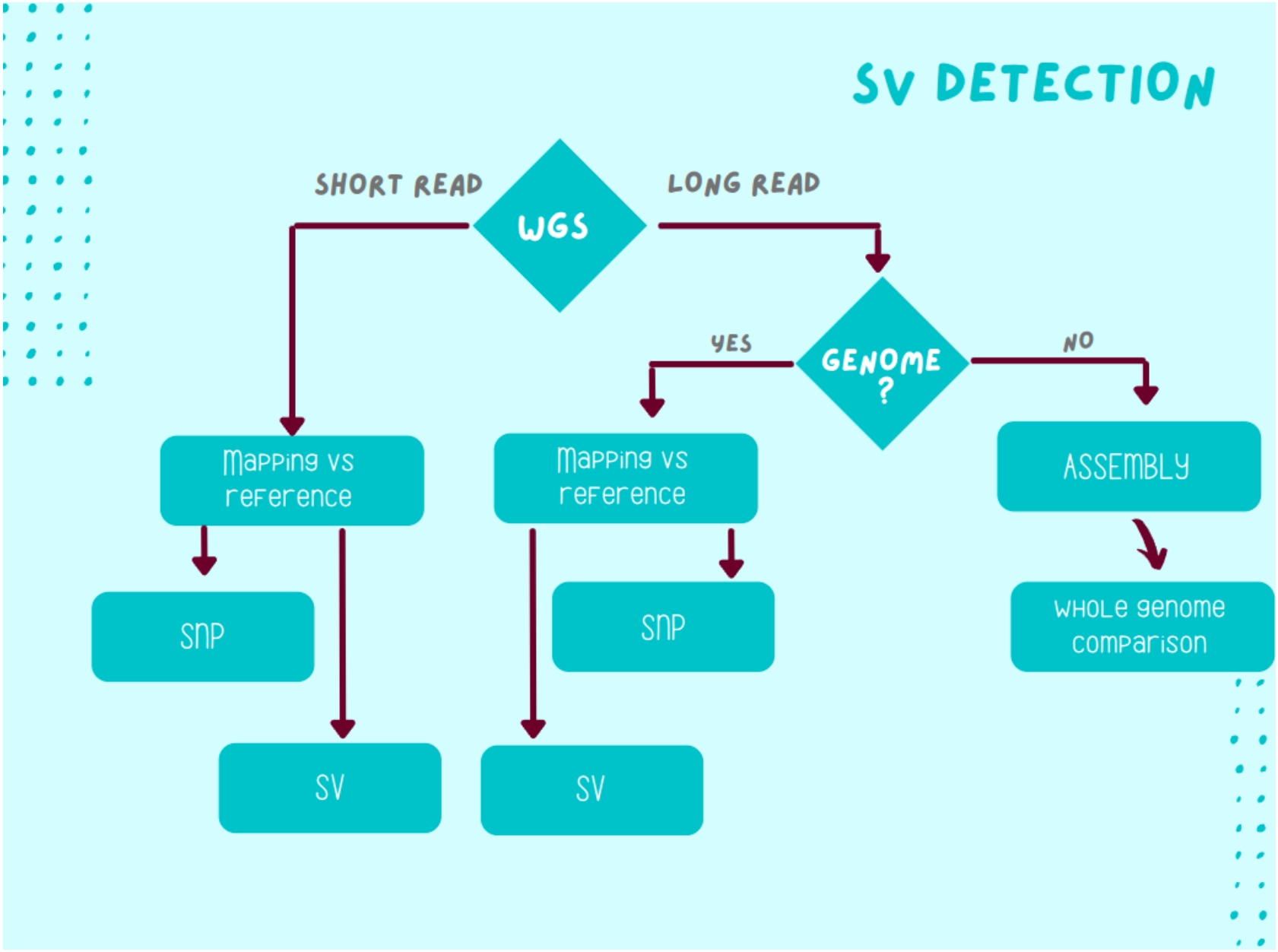
# Détection de variants à partir de données de séquençage short & long reads

**Alexis Dereeper** - UMR PHIM

**Julie Orjuela** - UMR DIADE

**Christine Tranchant-Dubreuil** - UMR DIADE

# Un même plan de bataille... ou pas !!!



# Objectifs

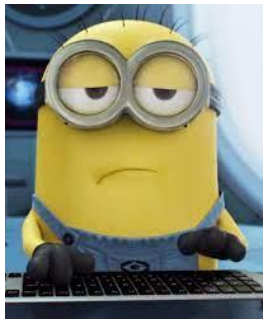
**Détecter des variants (SNP, variants structuraux) à partir de données de séquençage short et long reads.**

## Applications :

- Mapper des reads contre un génome ***bwa***
- Détecter des SNPs à partir du mapping de reads - ***bcftools***
- Analyser les données SNPs brutes (ex: stats, filtres) - ***vcftools, bcftools***
- Exemples d'études possibles à partir de SNPs - ***SNIPlay***

**Avec jupyter book** : lancer les commandes + analyser les résultats

=> Avoir un plan de bataille opérationnel



# RAW SEQUENCING DATA

# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

- Statistics
- Sequencing quality ?

  Adaptators ? Contaminants ?

# fastq format

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTTGTAGTCATTCCGTGTTGGTTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTTGGT
+
@@@DDDD>FFFAFBEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTGCAGTAACGACCTATAAATTTTAAAGTAGC
+
@CFFFFFGGHHHHIJJJJIEE<HHHIJJIGBHGGEEIIJJEIEIJIHHJFIIJJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTTCGGGTATGGCTCTAGTAGCTTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIEEHIII<CGHIJIJIIJ:?FC9DGAFGHII?DGBFIIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFGHHHHJIIIIIJJJIIGJCHHEIII>GHIG@GHIDHGJIIFIFHIJJIJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGTAAGAGTTAAAAAACAATTAATGAGCAACTGAGTTC
+
@@CFFFFFHFFHFGIJJIHHIIJJIIIHIJJJECGHIJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTTGTCGGGGGTTGGGAAATTGTCACTTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```

**1 sequence/read = 4 lines**

- read id, starting by **@**
- read sequence
- Comment line starting by ***+***

(usually contains read id).

- read Quality for each base

# PHRED SCORE

- Séquenceur assigne à chaque base séquencée un score lié à la probabilité que la base appelée soit fausse

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

*Ewing 1998*

- Ce score (PHRED score) varie entre 0 et 50

| 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 10000 | 99,99% |
| 50 | 1 in 100000 | 99.999 % |

# How to code quality score for each base with one letter ?

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTTGTAGTCATTCCGTGTTGGTTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTTGGT
+
@@@DDDD>FFFAFBEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTGCAGTAACGACCTATAAATTTTAAAGTAGC
+
@CFFFFFGGHHHHIJJJJIEE<HHHIJJIGBHGGEEIIJJEIEIJIHHJFIIJJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTTCGGGTATGGCTCTAGTAGCTTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIEEHIII<CGHIJIJIIJ:?FC9DGAFGHII?DGBFIIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFGHHHHJIIIIIJJJIIGJCHHEIII>GHIG@GHIDHGJIIFIFHIJJIJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGTGTGTAAGAGTTAAAAAACAATTAATGAGCAACTGAGTTC
+
@@CFFFFFHFFHFGIJJIHHIIJJIIIHIJJJECGHIJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTTGTCGGGGGTTGGGAAATTGTCACTTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```

## 1 sequence/read = 4 lines

- read id, starting by **@**
- read sequence
- Comment line starting by ***+***

(usually contains read id).

- read Quality for each base

# How to code quality score for each base with one letter ?

*Code ASCII*

## ASCII Table

<linuxhint/>

| Code | Char | Code | Char | Code | Char | Code | Char |
|---|---|---|---|---|---|---|---|
| 0 | NUL (null) | 32 | SPACE | 64 | @ | 96 | ` |
| 1 | SOH (start of heading) | 33 | ! | 65 | A | 97 | a |
| 2 | STX (start of text) | 34 | " | 66 | B | 98 | b |
| 3 | ETX (end of text) | 35 | # | 67 | C | 99 | c |
| 4 | EOT (end of transmission) | 36 | $ | 68 | D | 100 | d |
| 5 | ENQ (enquiry) | 37 | % | 69 | E | 101 | e |
| 6 | ACK (acknowledge) | 38 | & | 70 | F | 102 | f |
| 7 | BEL (bell) | 39 | ' | 71 | G | 103 | g |
| 8 | BS (backspace) | 40 | ( | 72 | H | 104 | h |
| 9 | TAB (horizontal tab) | 41 | ) | 73 | I | 105 | i |
| 10 | LF (NL line feed, new line) | 42 | * | 74 | J | 106 | j |
| 11 | VT (vertical tab) | 43 | + | 75 | K | 107 | k |
| 12 | FF (NP form feed, new page) | 44 | , | 76 | L | 108 | l |
| 13 | CR (carriage return) | 45 | - | 77 | M | 109 | m |
| 14 | SO (shift out) | 46 | . | 78 | N | 110 | n |
| 15 | SI (shift in) | 47 | / | 79 | O | 111 | o |
| 16 | DLE (data link escape) | 48 | 0 | 80 | P | 112 | p |
| 17 | DC1 (device control 1) | 49 | 1 | 81 | Q | 113 | q |
| 18 | DC2 (device control 2) | 50 | 2 | 82 | R | 114 | r |
| 19 | DC3 (device control 3) | 51 | 3 | 83 | S | 115 | s |
| 20 | DC4 (device control 4) | 52 | 4 | 84 | T | 116 | t |
| 21 | NAK (negative acknowledge) | 53 | 5 | 85 | U | 117 | u |
| 22 | SYN (synchronous idle) | 54 | 6 | 86 | V | 118 | v |
| 23 | ETB (end of trans. block) | 55 | 7 | 87 | W | 119 | w |
| 24 | CAN (cancel) | 56 | 8 | 88 | X | 120 | x |
| 25 | EM (end of medium) | 57 | 9 | 89 | Y | 121 | y |
| 26 | SUB (substitute) | 58 | : | 90 | Z | 122 | z |
| 27 | ESC (escape) | 59 | ; | 91 | [ | 123 | { |
| 28 | FS (file separator) | 60 | < | 92 | \ | 124 | \| |
| 29 | GS (group separator) | 61 | = | 93 | ] | 125 | } |
| 30 | RS (record separator) | 62 | > | 94 | ^ | 126 | ~ |
| 31 | US (unit separator) | 63 | ? | 95 | _ | 127 | DEL |

# How to code quality score for each base with one letter ?

## Code ASCII

| Code Char |
|---|
| 64 @ |
| 65 A |
| 66 B |
| 67 C |
| 68 D |
| 69 E |
| 70 F |
| 71 G |
| 72 H |
| 73 I |
| 74 J |
| 75 K |
| 76 L |
| 77 M |
| 78 N |
| 79 O |
| 80 P |
| 81 Q |
| 82 R |
| 83 S |
| 84 T |
| 85 U |
| 86 V |
| 87 W |
| 88 X |
| 89 Y |
| 90 Z |
| 91 [ |
| 92 \ |
| 93 ] |
| 94 ^ |
| 95 _ |

| Phred Quality Score |
|---|
| 0 .. 50 |

| Code Char |
|---|
| 96 ` |
| 97 a |
| 98 b |
| 99 c |
| 100 d |
| 101 e |
| 102 f |
| 103 g |
| 104 h |
| 105 i |
| 106 j |
| 107 k |
| 108 l |
| 109 m |
| 110 n |
| 111 o |
| 112 p |
| 113 q |
| 114 r |
| 115 s |
| 116 t |
| 117 u |
| 118 v |
| 119 w |
| 120 x |
| 121 y |
| 122 z |
| 123 { |
| 124 \| |
| 125 } |
| 126 ~ |
| 127 DEL |

# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

- Statistics
- Sequencing quality ? Adaptators ?

  Contaminants ?

# OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

- Statistics
- Sequencing quality ? Adaptators ?

  Contaminants ?

→ Basic statistics and quality control checks using **fastqc**

**fastqc** to get some basic statistics and to do some quality control checks

```
# fastqc command
fastqc /path2fastq/AX8798.fastq  -o path2fastqcDIR

fastqc /path2fastq/*  -o path2fastqcDIR
```

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

**[command line] manuel :**
https://manpages.ubuntu.com/manpages/trusty/man1/fastqc.1.html#:~:text=DESCRIPTION,of%20problem%20in%20your%20data

# FastQC : Basic Statistics

## Basic Statistics

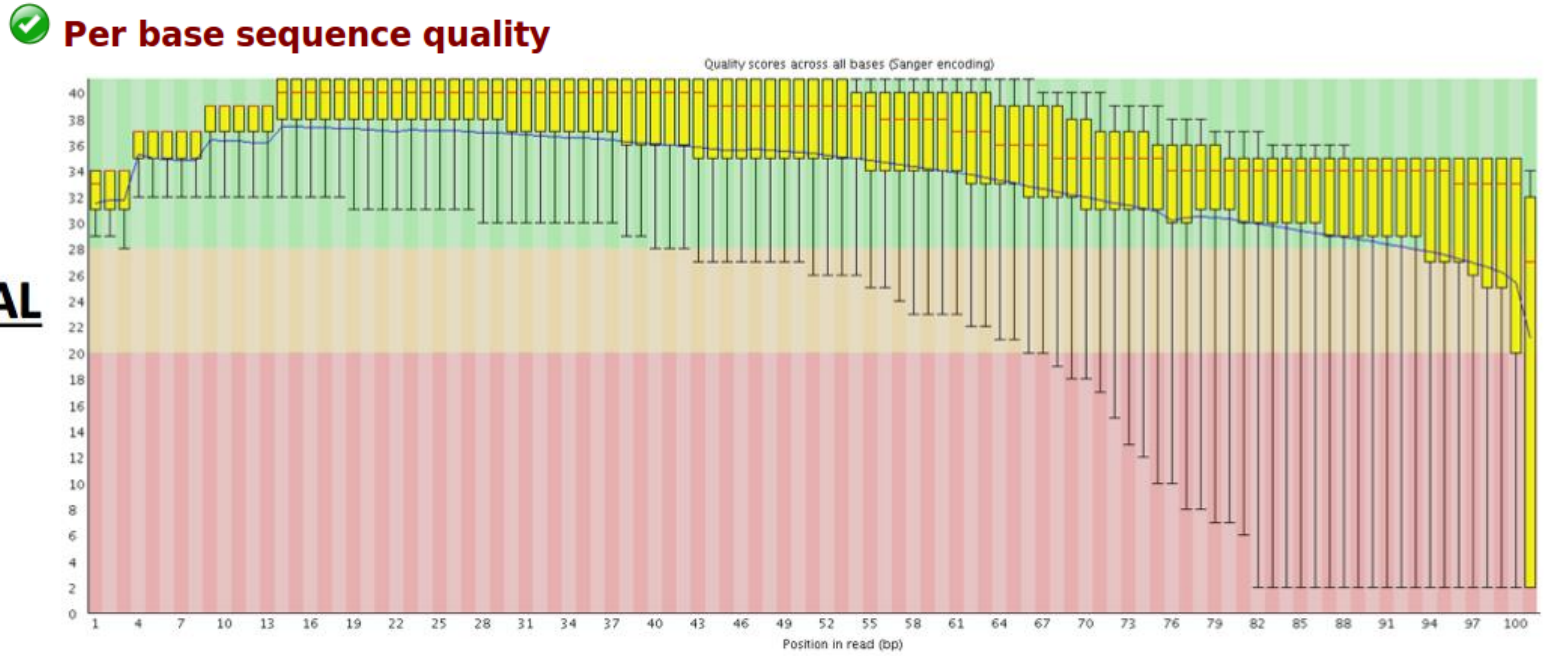
| Measure | Value |
|---|---|
| Filename | ATR_AOSE_15.read1.fastq |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 680123611 |
| Filtered Sequences | 0 |
| Sequence length | 30-101 |
| %GC | 47 |

# FastQC : Per base sequence quality

This plot shows the base quality score distribution for all reads in a lane, with each read position considered independently.

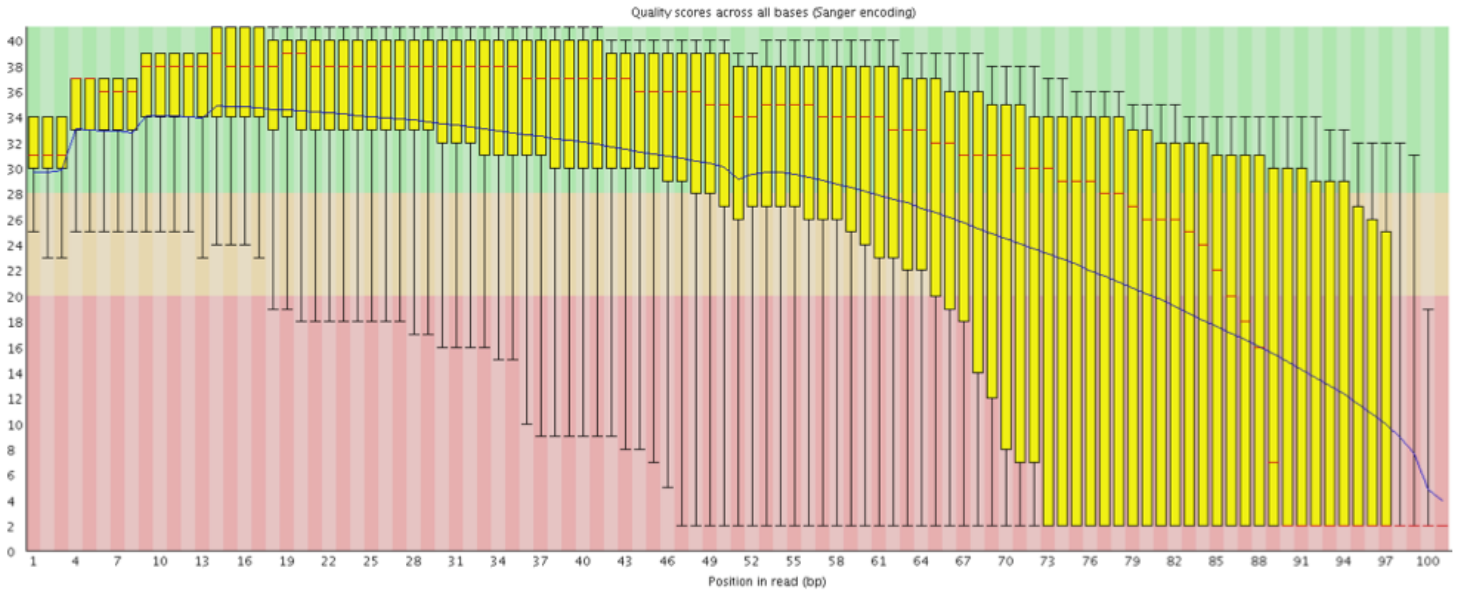
- x-axis = position in read (bp)
- y-axis = Phred-like base quality score [pink=0-20, tan=20-30, green=30-40]
- red bar = median score, blue line = mean score
- yellow box = 25th to 75th percentile, black whiskers = 10th to 90th percentile

**GOOD/NORMAL LANE**

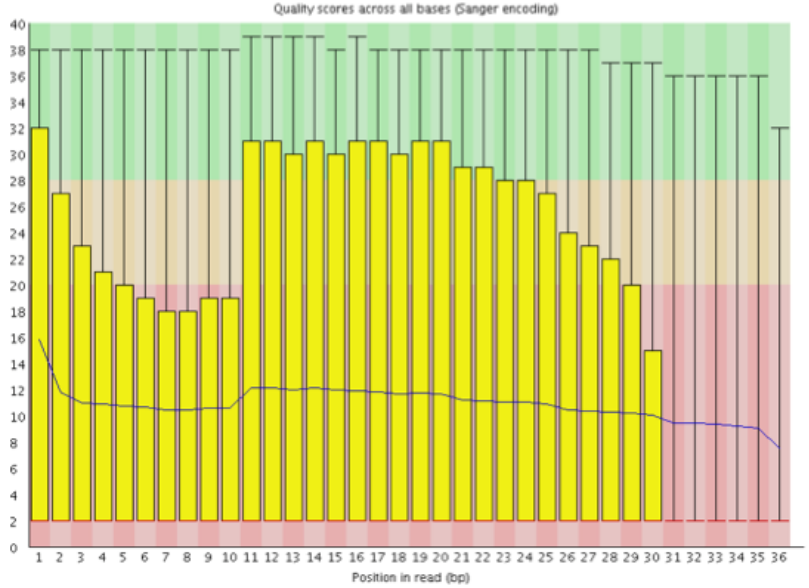


# FastQC : Per base sequence quality

**SALVAGEABLE LANE**

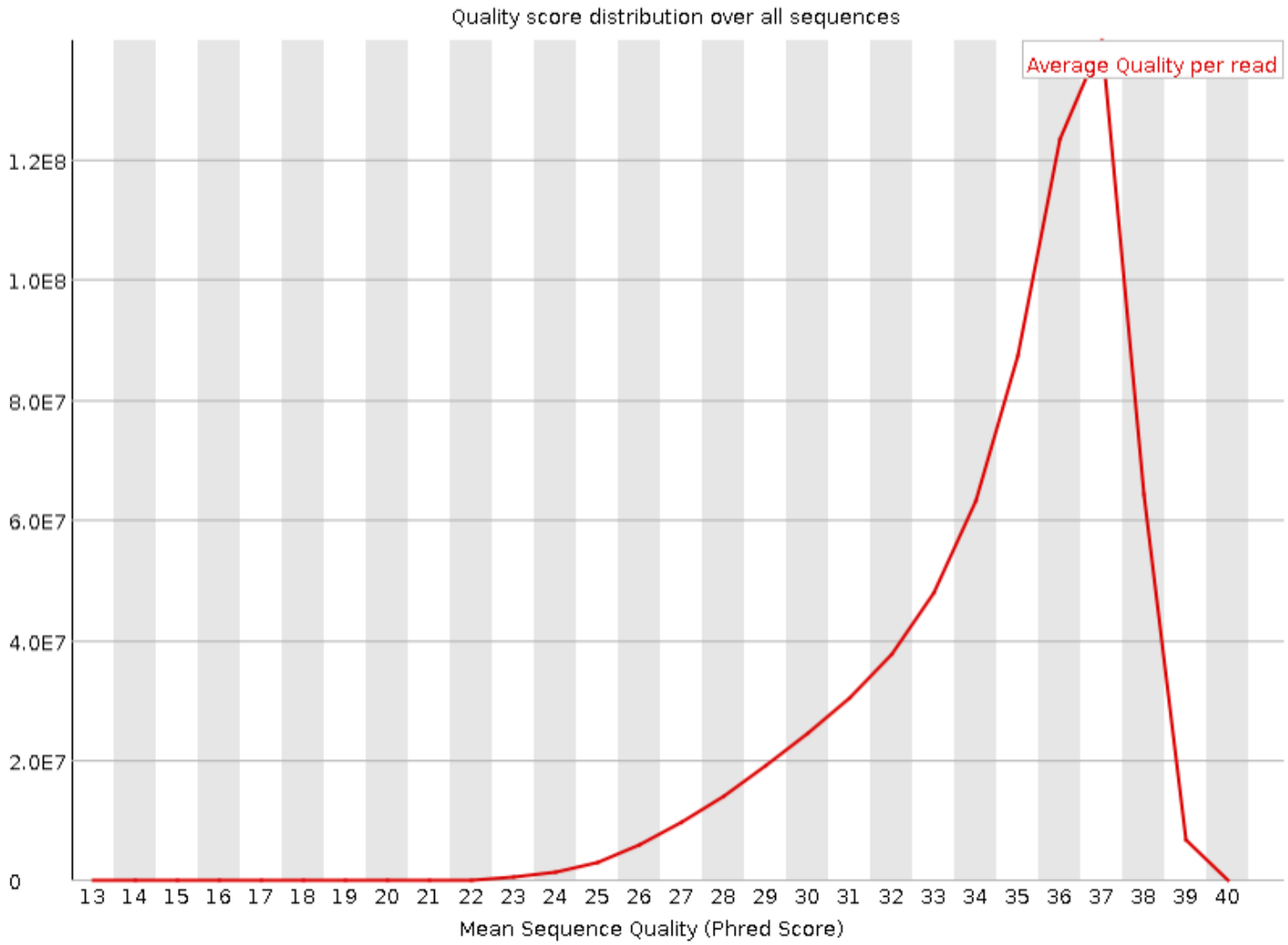



**FAILED LANE**



# FastQC: Per sequence quality scores

## Per sequence quality scores

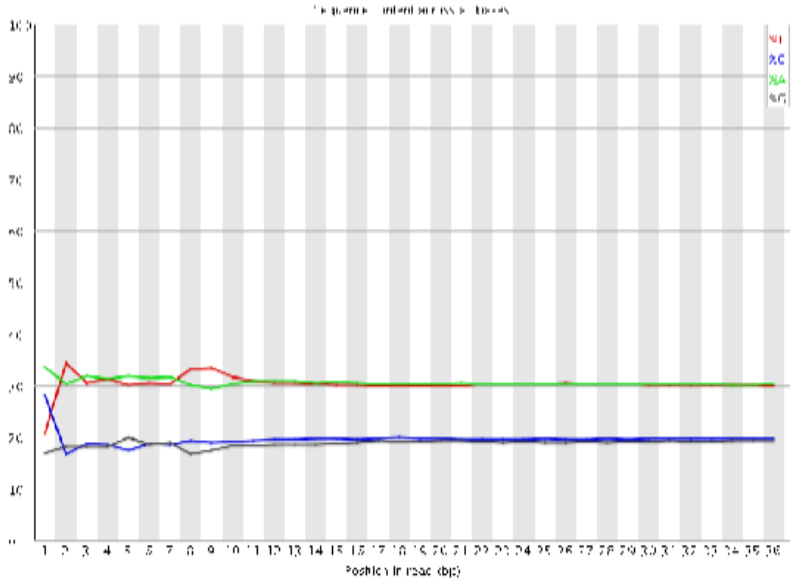


# FastQC: Per base sequence content

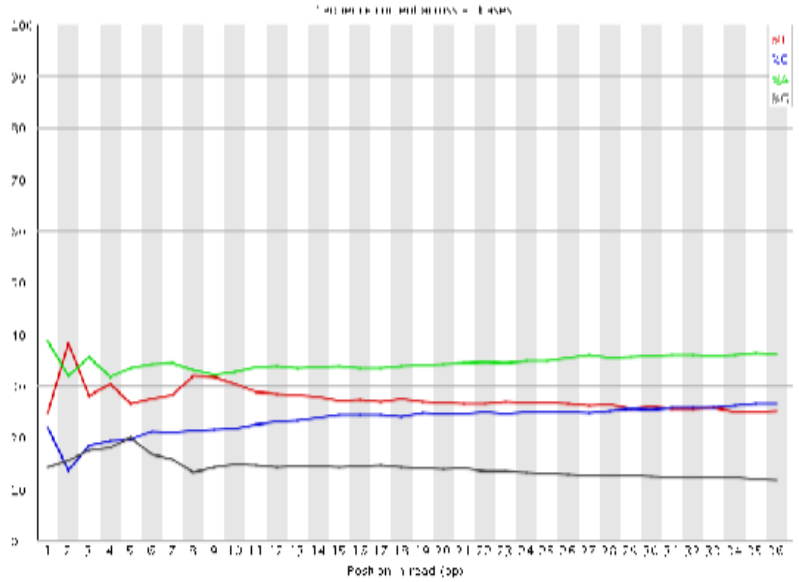
This plot shows the nucleotide distribution per read position for all reads in a lane.

- x-axis = position in read (bp)
- y-axis = % of all reads in the lane
- colors refer to individual nucleotides: **A**, **C**, **G**, **T**

**GOOD LANE**



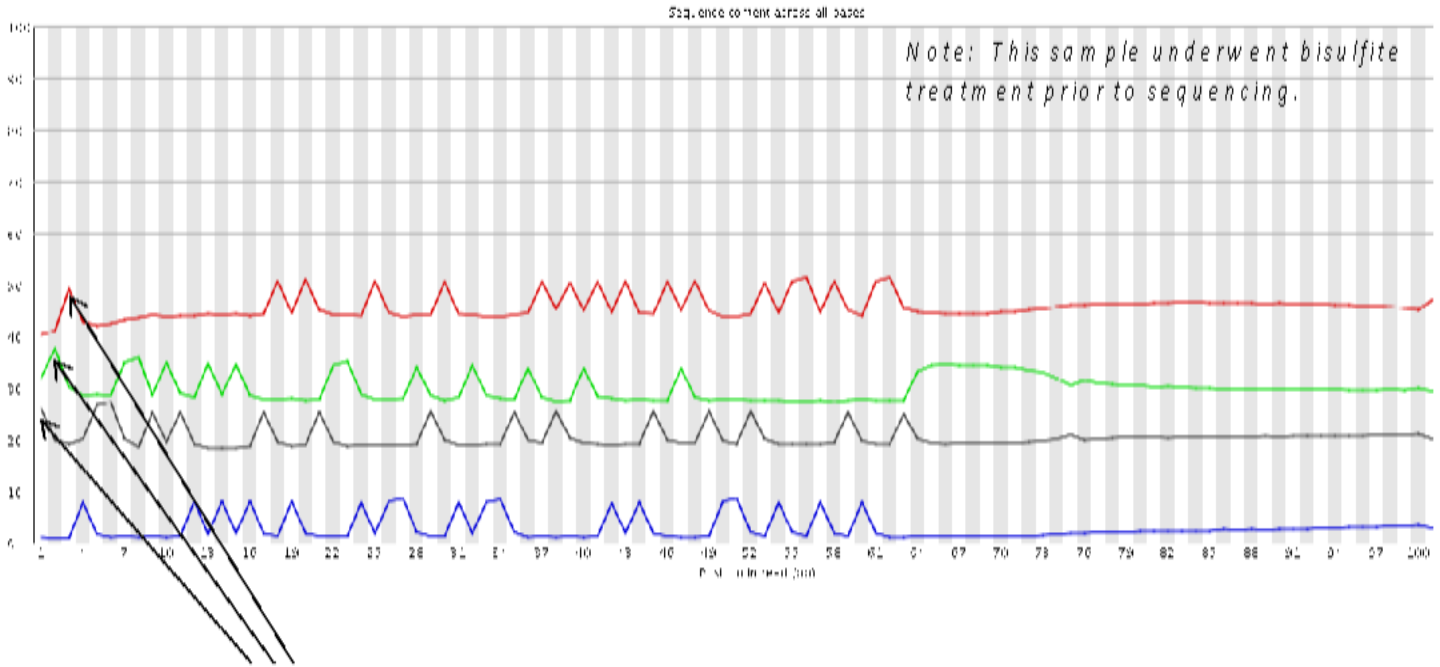
**BAD LANE**



**Can this be fixed?** No.

# FastQC: Per base sequence content

This lane has a different problem – one sequence motif is highly over-represented.




primer/adapter sequence: **GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG**

In this lane, ~10% of reads have the adapter sequence & the rest are normal.

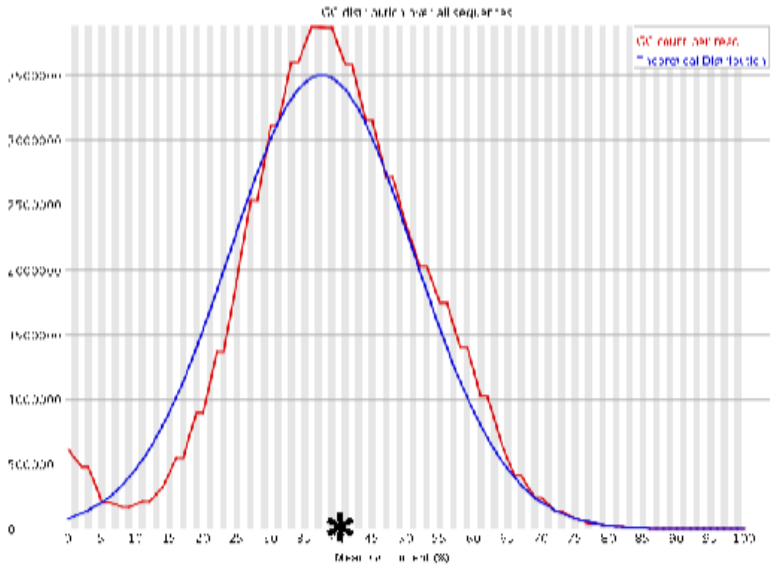
**<u>Can this be fixed?</u> Yes. Simply remove the reads w/ adapter contamination, and everything that's left should be fine. (Talk to a bioinformatics analyst for help.)**

# FastQC: Per sequence GC content

This plot shows the distribution of GC content per read for all reads in a lane.

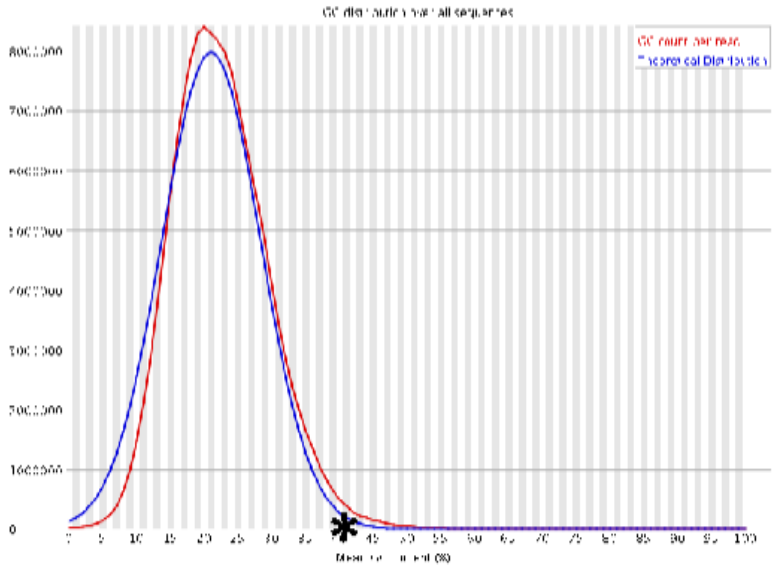
- x-axis = mean GC content (%)
- y-axis = # of reads
- red: observed read count, blue: theoretical distribution (given observed)

**GOOD LANE**



*mouse genome ≈ 40% GC*

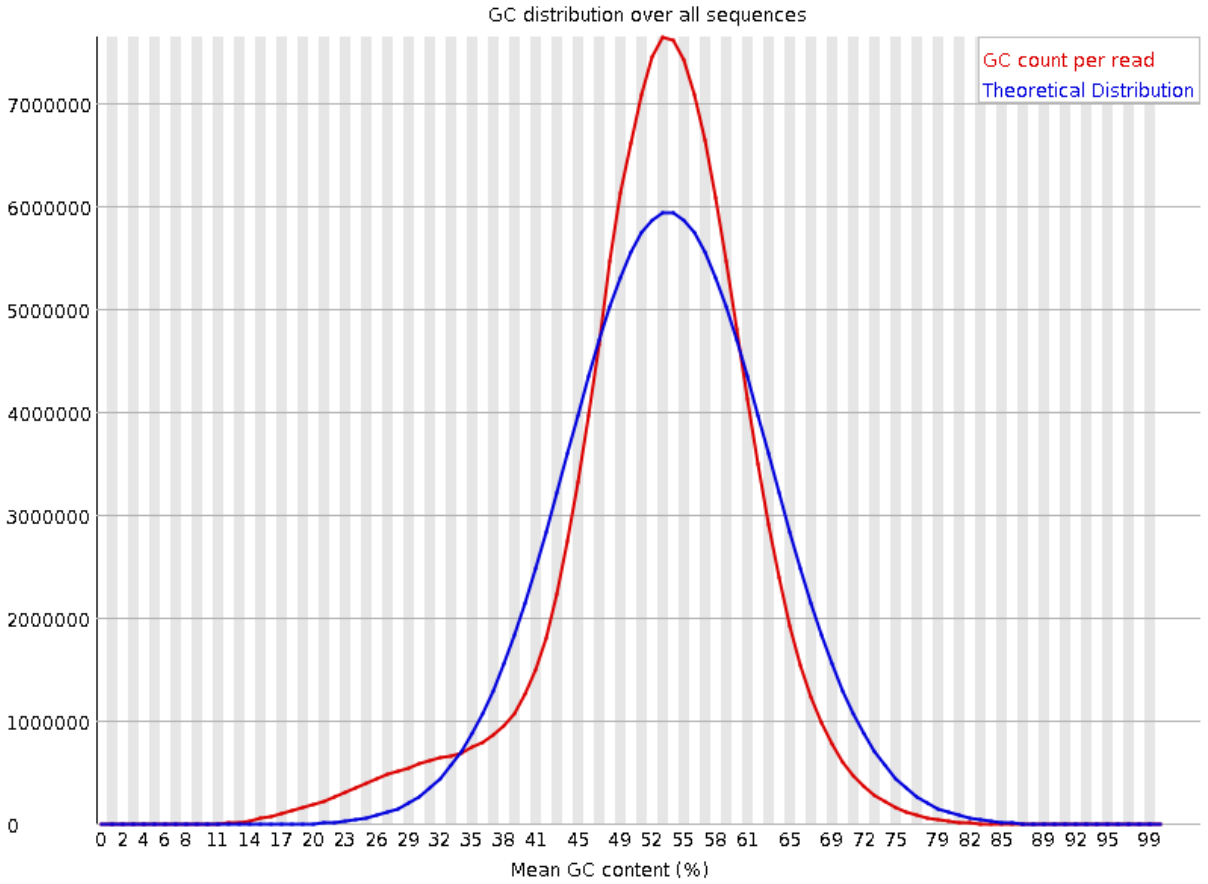
**BAD LANE**



**Can this be fixed?** No.

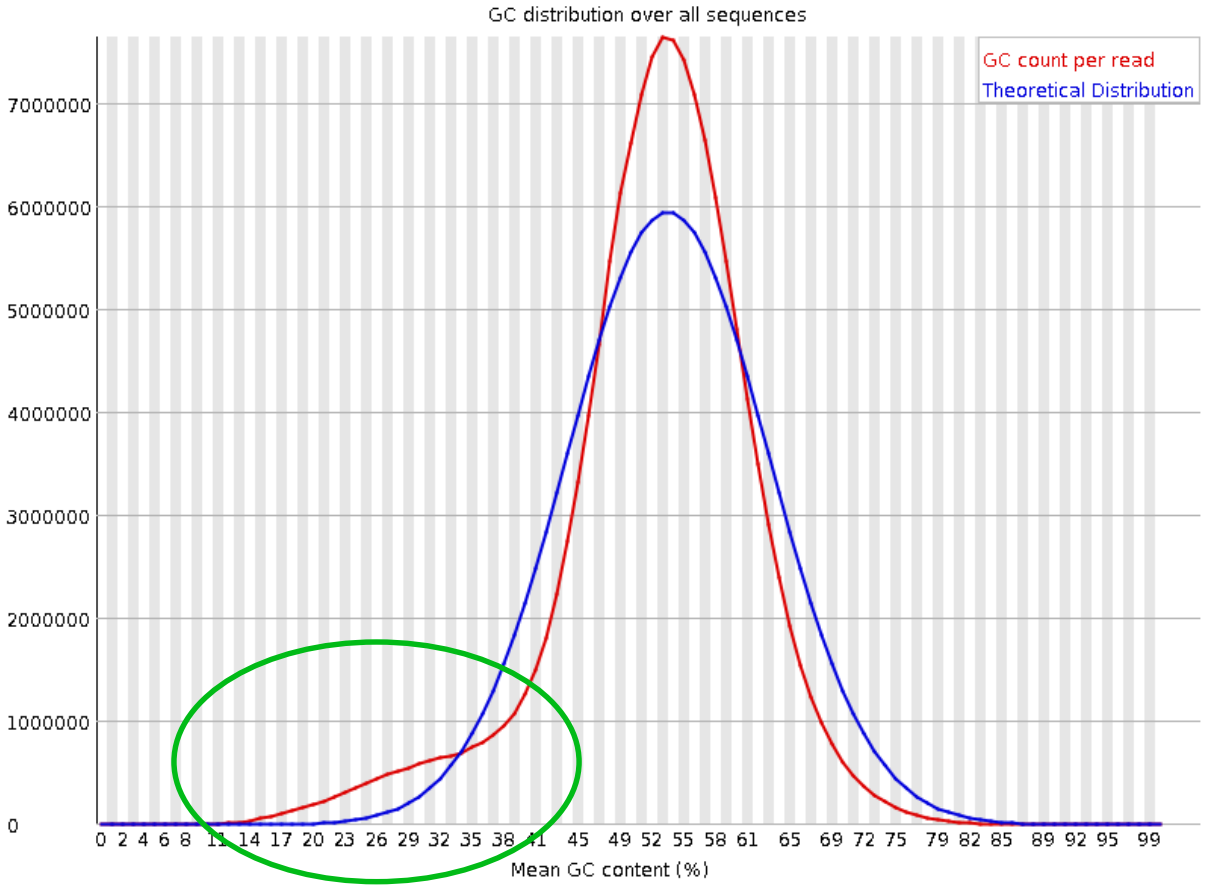
# FastQC: Per sequence GC content

- A contamination ?



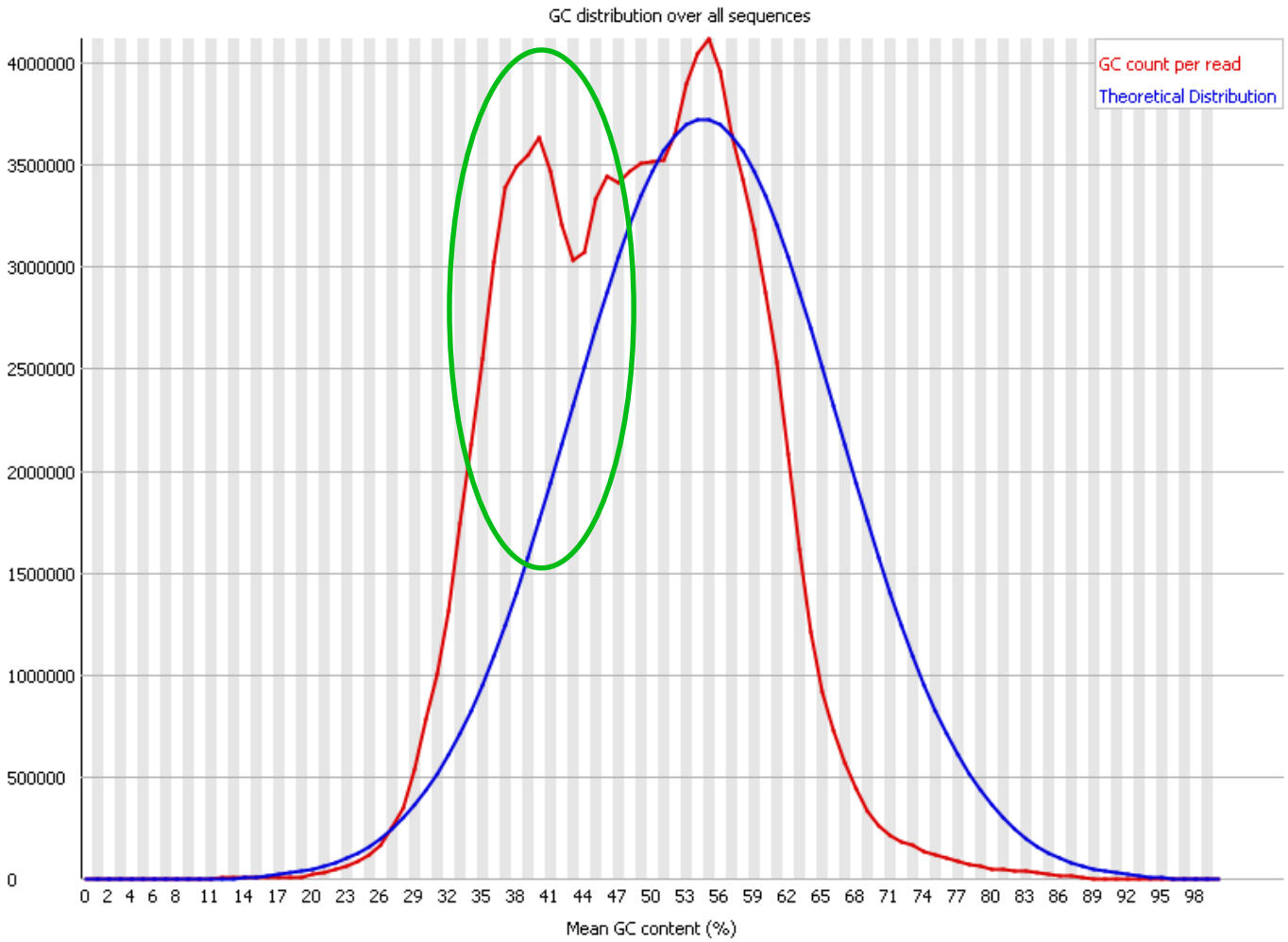
# FastQC: Per sequence GC content

- A contamination ?

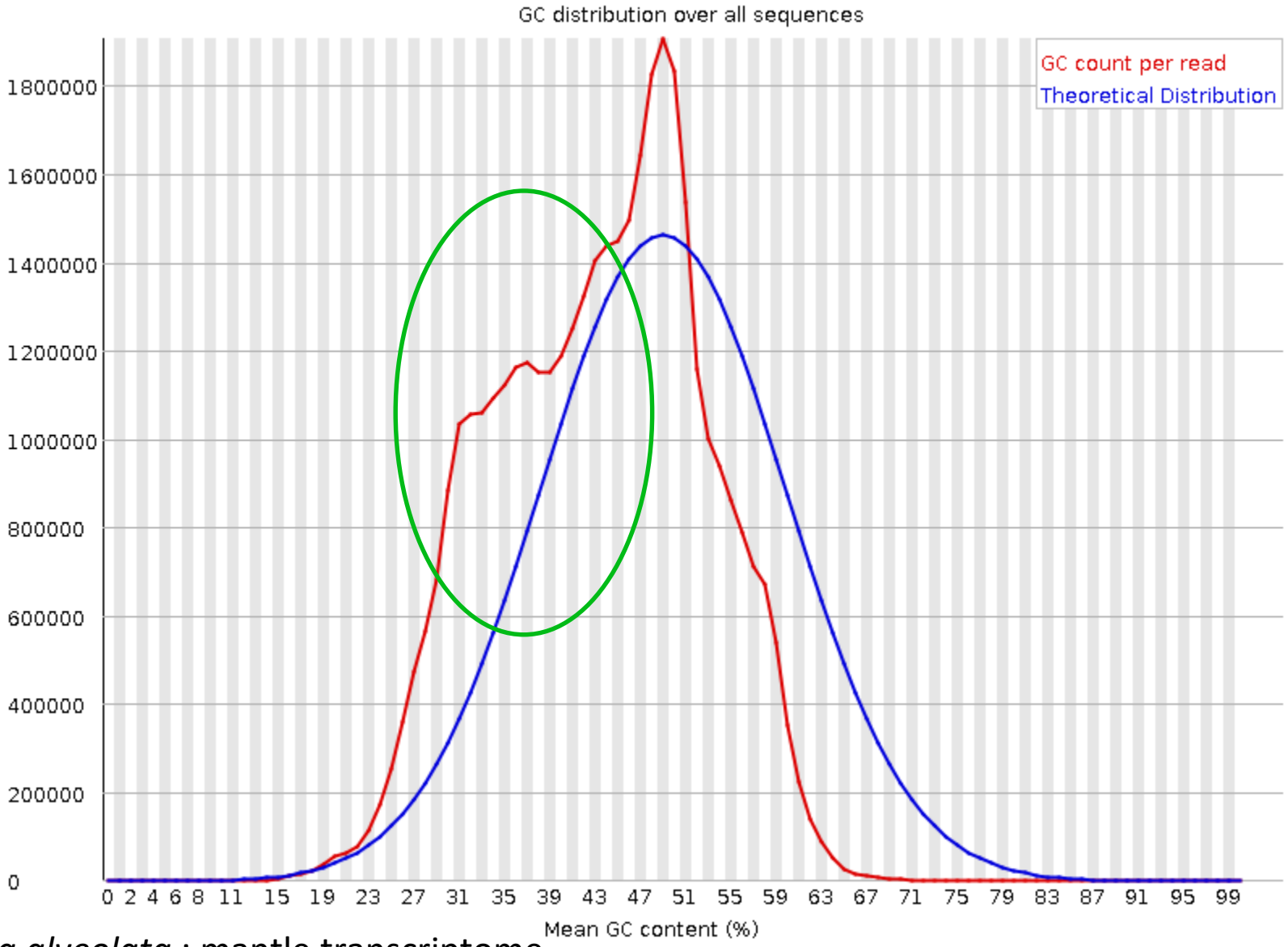



**Can this be fixed ?** Maybe…

# FastQC: Per sequence GC content



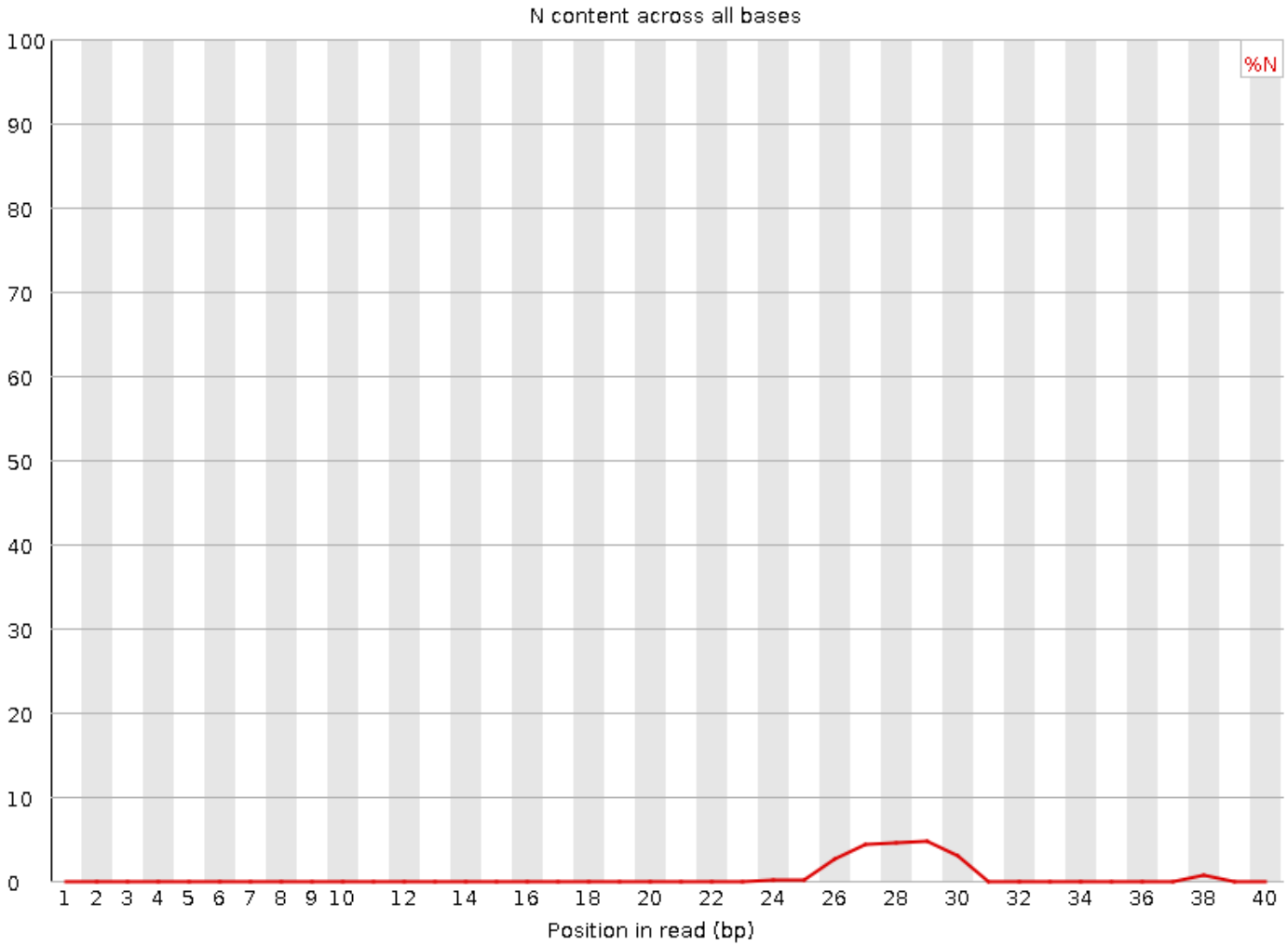
# Third-party contamination : detection




*Sabellaria alveolata* : mantle transcriptome

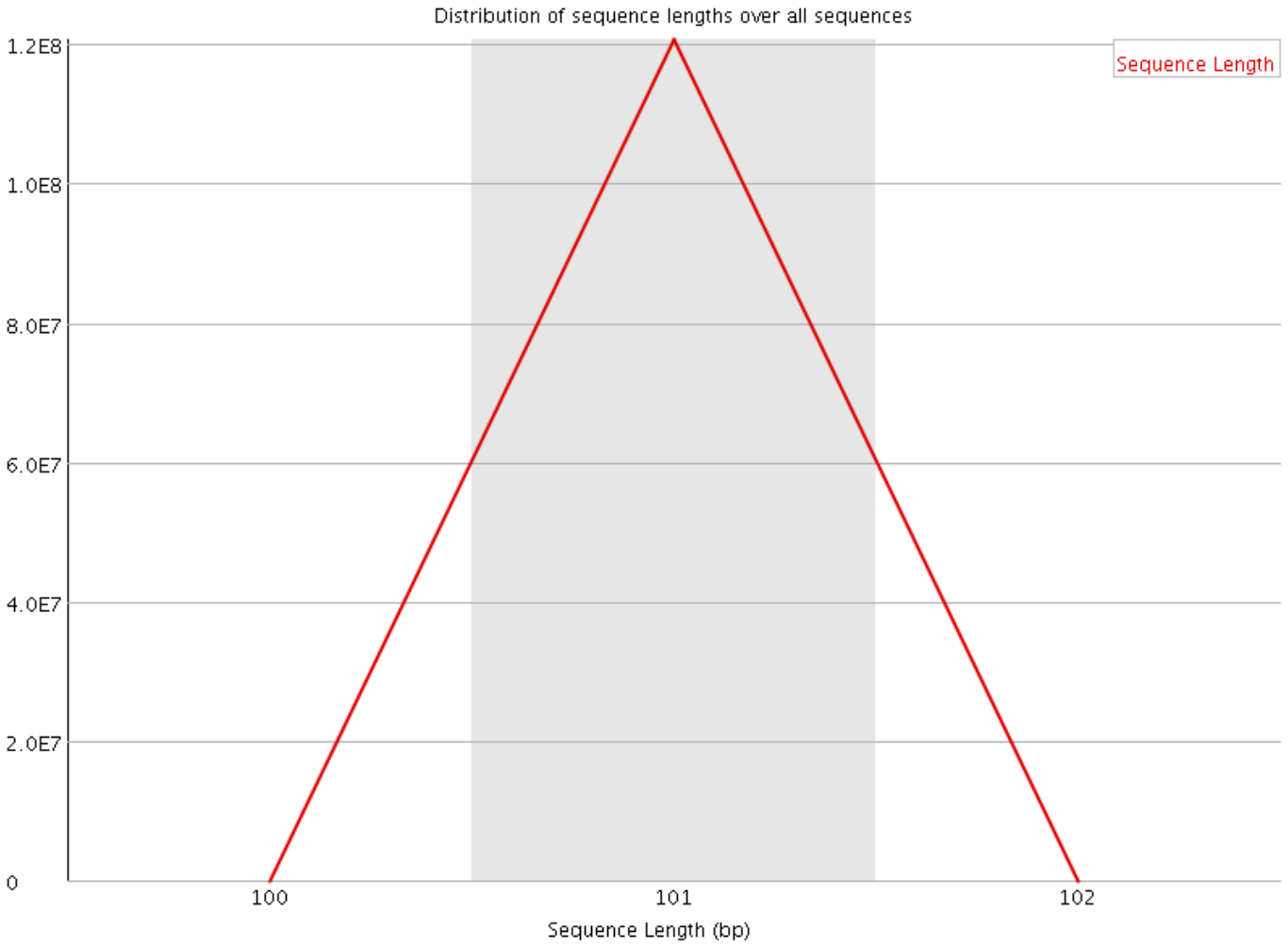
# FastQC: Per base N content

## Per base N content



# FastQC: Sequence Length Distribution

## Sequence Length Distribution

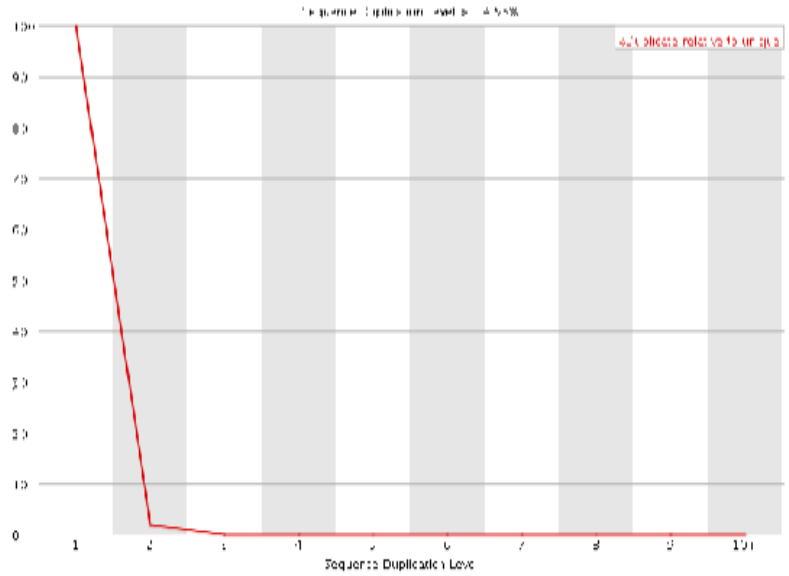


# FastQC: Sequence Duplication Levels

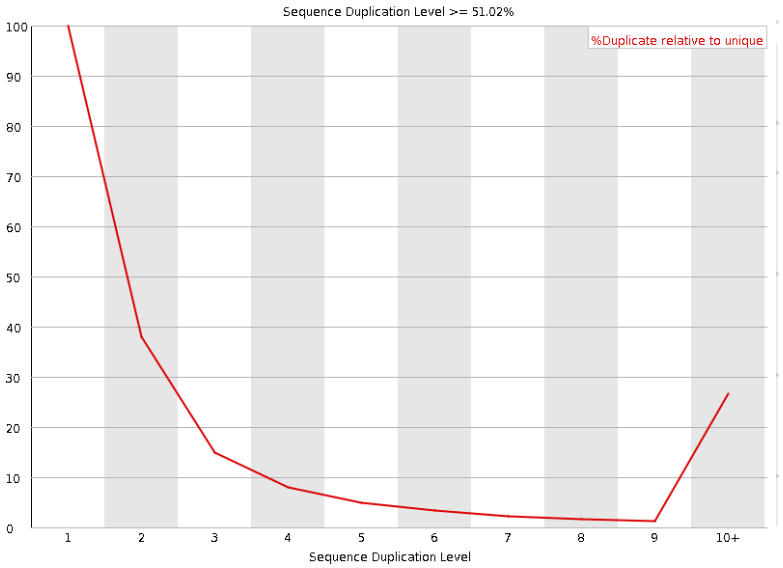
This plot shows the degree of duplication for a subset of reads in a lane.

- x-axis = sequence duplication level
- y-axis = % duplicates relative to unique reads

## GOOD LANE



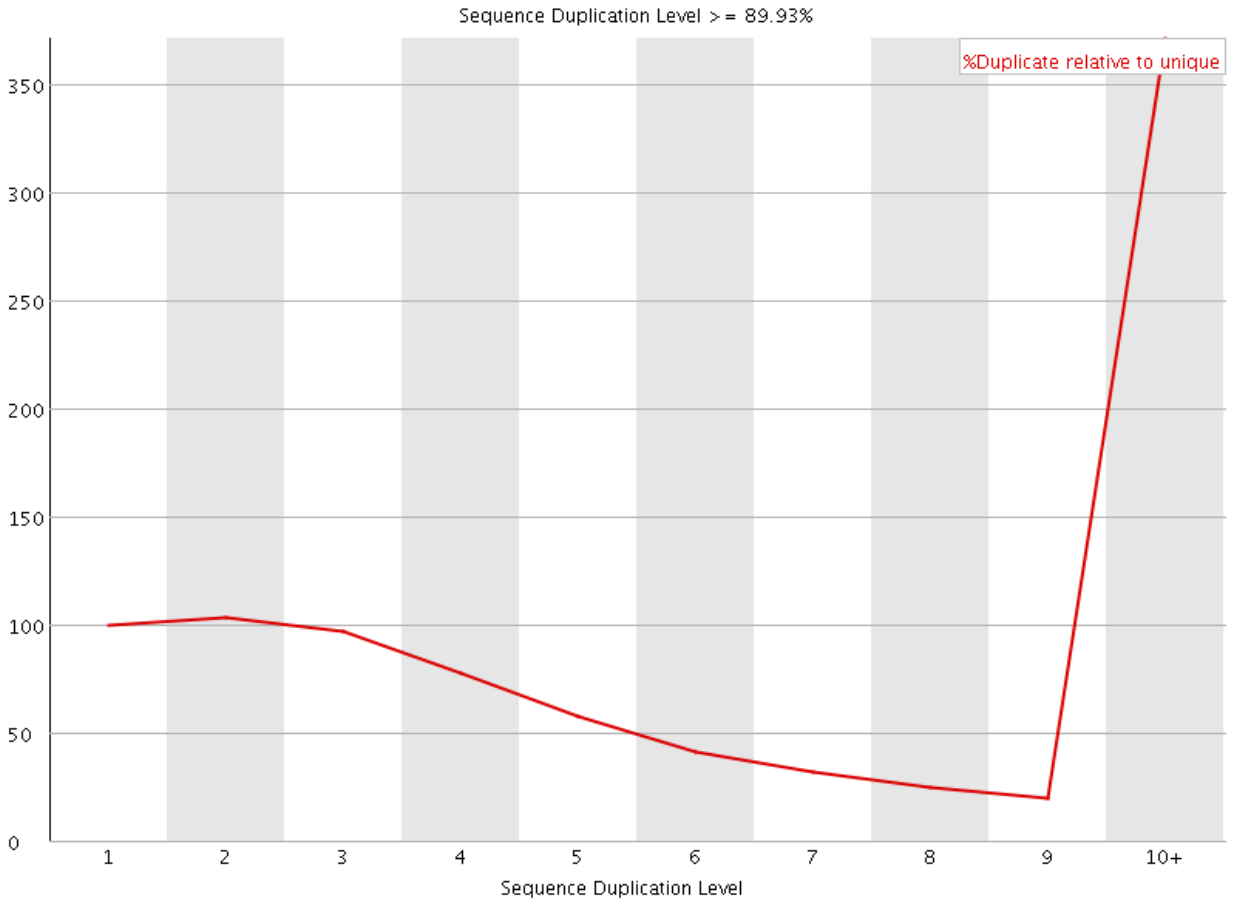
## BAD LANE




**Can this be fixed?** **Maybe.**

# FastQC: Sequence Duplication Levels

## Sequence Duplication Levels




**Can this be fixed? Hem…**

# FastQC: Overrepresented sequences

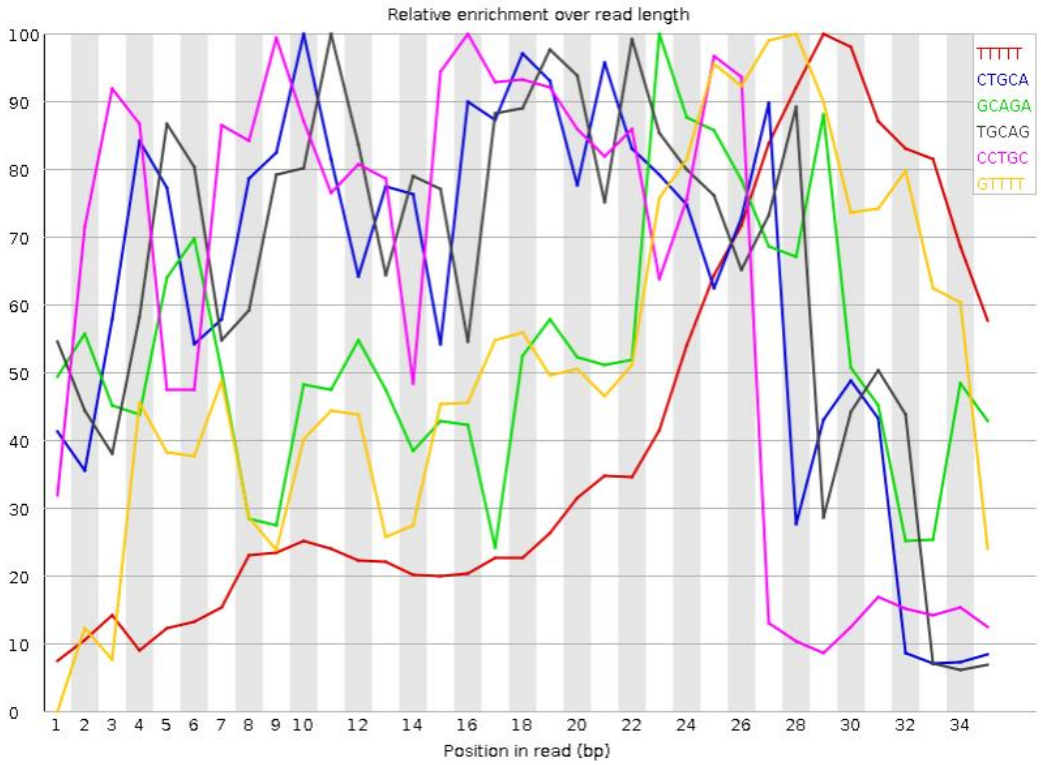
## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC | 2065 | 0.5224039181558763 | No Hit |
| GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG | 2047 | 0.5178502762542754 | No Hit |
| ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA | 2014 | 0.5095019327680071 | No Hit |
| CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT | 1913 | 0.4839509420979134 | No Hit |
| GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA | 1879 | 0.47534961850600066 | No Hit |
| AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT | 1846 | 0.4670012750197325 | No Hit |
| TCATGGAAGCGATAAAACTCTGCAGGTTGGATACGCCAAT | 665 | 0.16823177025358726 | No Hit |
| TCTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATAC | 627 | 0.15861852623909656 | No Hit |
| GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT | 624 | 0.1578595859221631 | Illumina Paired End PCR Primer 2 (100% over 40bp) |
| CCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACA | 613 | 0.15507680476007366 | No Hit |
| CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC | 599 | 0.15153508328105078 | Illumina Paired End PCR Primer 2 (96% over 25bp) |
| TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG | 585 | 0.1479933618020279 | No Hit |
| CGCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTT | 552 | 0.13964501831575965 | No Hit |
| CTCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGC | 532 | 0.1345854162028698 | No Hit |
| CTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATACG | 515 | 0.13028475440691342 | No Hit |
| CTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCGC | 505 | 0.12775495335046852 | No Hit |
| GCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTTG | 411 | 0.10397482341988626 | No Hit |

Adapter dimers
rRNA
Satellite sequences

# FastQC: Kmer Content

## Kmer Content

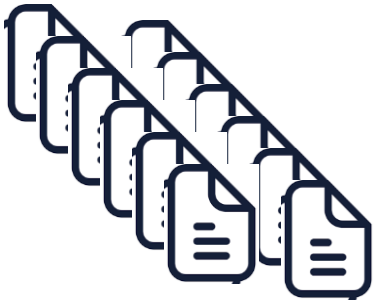



| Sequence | Count | Obs/Exp Overall | Obs/Exp Max | Max Obs/Exp Position |
|---|---|---|---|---|
| TTTTT | 192940 | 8.590186 | 21.06293 | 29 |
| CTGCA | 90975 | 7.7906475 | 12.251836 | 10 |
| GCAGA | 84910 | 7.163295 | 13.539302 | 23 |
| TGCAG | 92470 | 7.002405 | 10.671717 | 11 |
| CCTGC | 57235 | 5.4987235 | 8.729035 | 16 |
| GTTTT | 108205 | 5.324498 | 10.243909 | 28 |
| CAACC | 49005 | 5.2869425 | 9.85526 | 13 |
| ATCGC | 58320 | 4.9942355 | 8.029807 | 29 |
| CCAAC | 46220 | 4.9864807 | 9.408141 | 12 |
| AAAAA | 62285 | 4.7588468 | 8.0126295 | 5 |
| CAGAG | 56370 | 4.7555633 | 7.148592 | 20 |
| ACCTG | 55315 | 4.736902 | 7.919266 | 15 |
| CGCCA | 44035 | 4.7130895 | 8.830201 | 35 |
| GGGGG | 63675 | 4.67525 | 16.94222 | 27 |
| GCAGG | 55380 | 4.6350074 | 17.521912 | 19 |
| AAAAC | 51945 | 4.452569 | 8.159592 | 24 |
| TATCG | 64615 | 4.4271946 | 8.394971 | 34 |
| GCTGG | 58505 | 4.3952427 | 10.37436 | 18 |
| AACCT | 50775 | 4.382863 | 7.691214 | 14 |
| TTATC | 70080 | 4.3444843 | 7.810299 | 33 |
| TTTTA | 87340 | 4.332125 | 7.8541703 | 28 |
| TTTAT | 86645 | 4.297653 | 7.9511886 | 35 |
| CGCTT | 54695 | 4.2042785 | 6.9374876 | 31 |

**fastqc** to get some basic statistics and to do some quality control checks

```
# fastqc command
fastqc /path2fastq/AX8798.fastq  -o path2fastqcDIR

fastqc /path2fastq/*  -o path2fastqcDIR
```

**fastqc** generate one report by fastq file

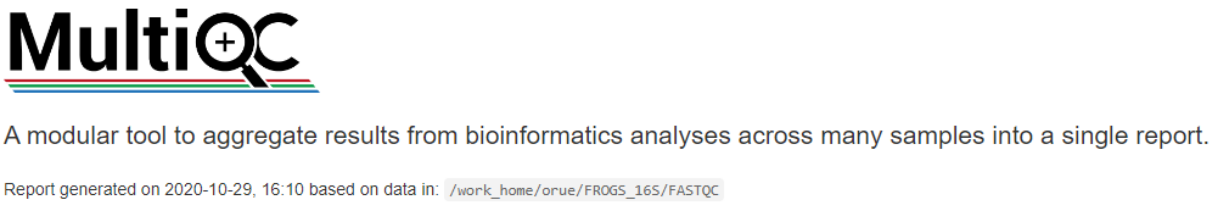


**With numerous fastq and fastqc report => use MultiQC**

**MultiQC** : a modular tool to summarise results from a bioinformatics analyse performed on many samples into a single report

```
# MultiQC command
multiqc path2fastqcDIR
```

https://multiqc.info/



## QUALITE DE SEQUENÇAGE & « NETTOYAGE »

### cutadapt, trimmomatics

- Détection et retrait des adaptateurs et primers
- Retrait des queue polyA/T
- Détection des séquences contaminantes, ARN ribosomal
- Masquage des bases avec  phred score bas par N
- Séquences courtes après retriat des adaptateurs