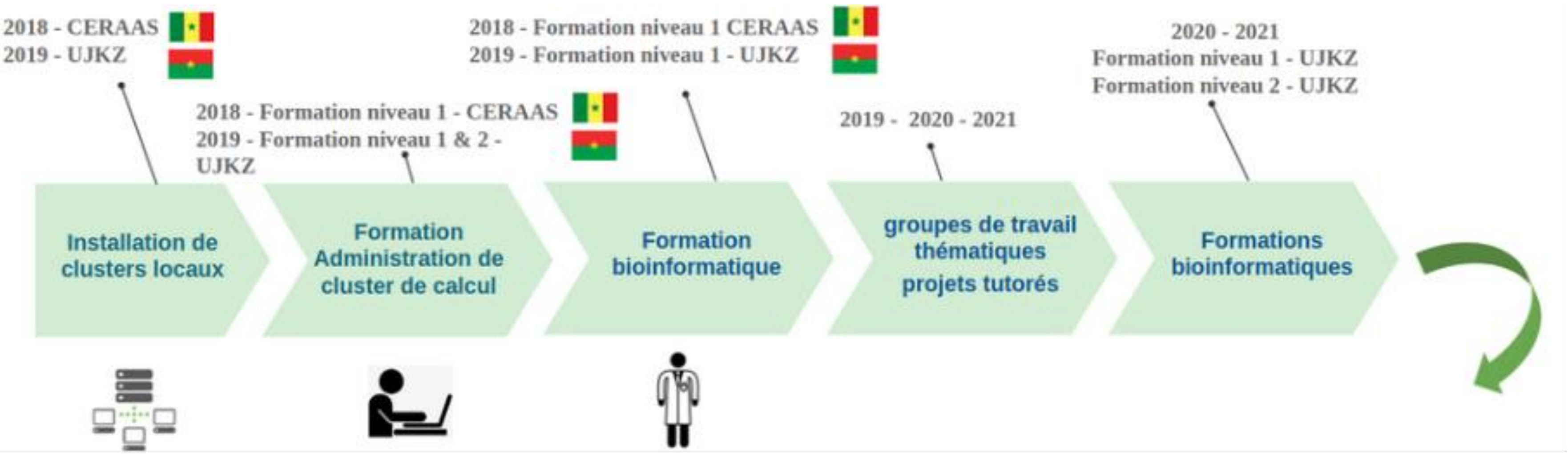
Projet Burkinabioinfo

Projet burkinabioinfo



2018 - CERAAS
2019 - UJKZ

2018 - Formation niveau 1 - CERAAS
2019 - Formation niveau 1 & 2 - UJKZ

2018 - Formation niveau 1 CERAAS
2019 - Formation niveau 1 - UJKZ

2019 - 2020 - 2021

2020 - 2021
Formation niveau 1 - UJKZ
Formation niveau 2 - UJKZ

Installation de clusters locaux

Formation Administration de cluster de calcul

Formation bioinformatique

groupes de travail thématiques
projets tutorés

Formations bioinformatiques

Administrateurs d'Afrique de l'Ouest (Burkina Faso, Côte d'Ivoire, Sénégal)

Chercheurs d'Afrique de l'Ouest (Burkina Faso, Bénin, Côte d'Ivoire, Mali, Sénégal)

A Moyen / long terme

- Clusters installés localement
- Réseau d'experts en administration de cluster
- Réseau de chercheurs experts en bioinformatique
- Intégration de modules de formation progressivement dans les cursus universitaires
- Mise en ligne des supports et cours, développement de cours accessibles en ligne

Programme & Logistique

13 apprenants

Objectifs : Initiation à la bioinformatique avec 2 cas pratique détection de variants structurants

- SNP : données short reads illumina
  Détections SNP > post Analyses
- Génomique comparatique : données long reads
  Détections de SNPs et de variants structuraux
  Assemblage et Comparaison de génomes assemblés

→ linux, Jupyter book, bash, python, protocoles bioinfos et bioanalyses

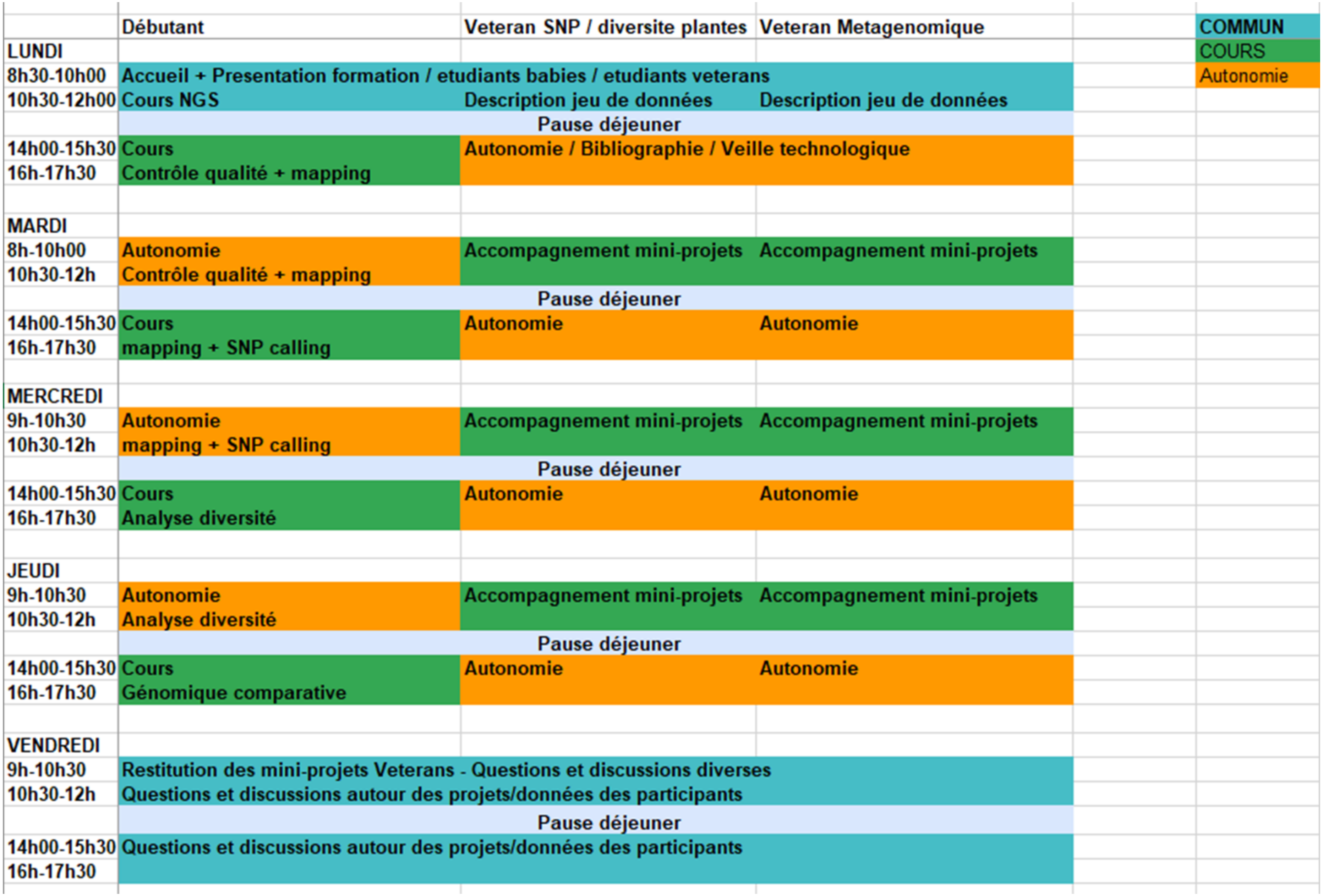
9 apprenants

Objectifs : 2 projets à mener

- Métagénomique à partir d'un échantillon prélevé dans un champs (virus, bactérie)
- Détection de SNPs à partir d'échantillons d'Ignames séquencés

→ Terminal Linux, Cluster UZB

Programme



	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*

tilapia du Nil, déterminisme du sexe, contrôle du sexe, population monosexe mâle, marqueurs chromosomiques

🇹🇬 *PALANGA Essowè*

Metagenomique, virus, interaction plante-parasite, phytopathologie

🇧🇫 *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

mateurs

Dereeper Alexis

Interaction plantes-pathogène, pangénomique des xanthomonas, diversité du riz, Kite surfeur

Tibiri B. Ezéchiel

Interaction plantes-pathogène,

Orjuela-Bouniol Julie

Assemblages de génomes, annotation, diversité, métagénomique
Développement des méthodes pour l'analyse de la diversité des plantes sans référence

Tranchant-Dubreuil
Christine

diversité des riz africains, mécanismes d'adaptation et de sélection, pangénomique, variants structuraux (SNP mais pas que)

Brunel Dominque
Centre Nationale de
Génotypage - INRAE

Programme

2 formations …
2 ambiances

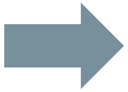
Mode “training”

- Session cours suivi par
- Session pratique en autonomie (individuel ou en groupe)
- Correction en groupe

Mode “projet”

- brainstroming en groupe, avec les formateurs
- projet en autonomie…
- debriefing collectif
- 2 projets en parallèle !

Des données différentes pour les 2 groupes avec des analyses différentes !!!



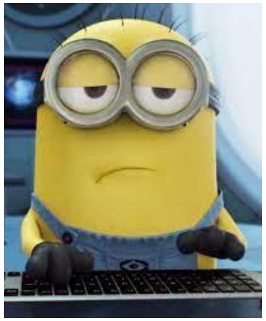
Apprendre à réaliser une analyse bioinformatique
Avoir un oeil critique lors de la (bio)analyse des données
Maîtriser linux, les outils bioinformatiques

auton

omie

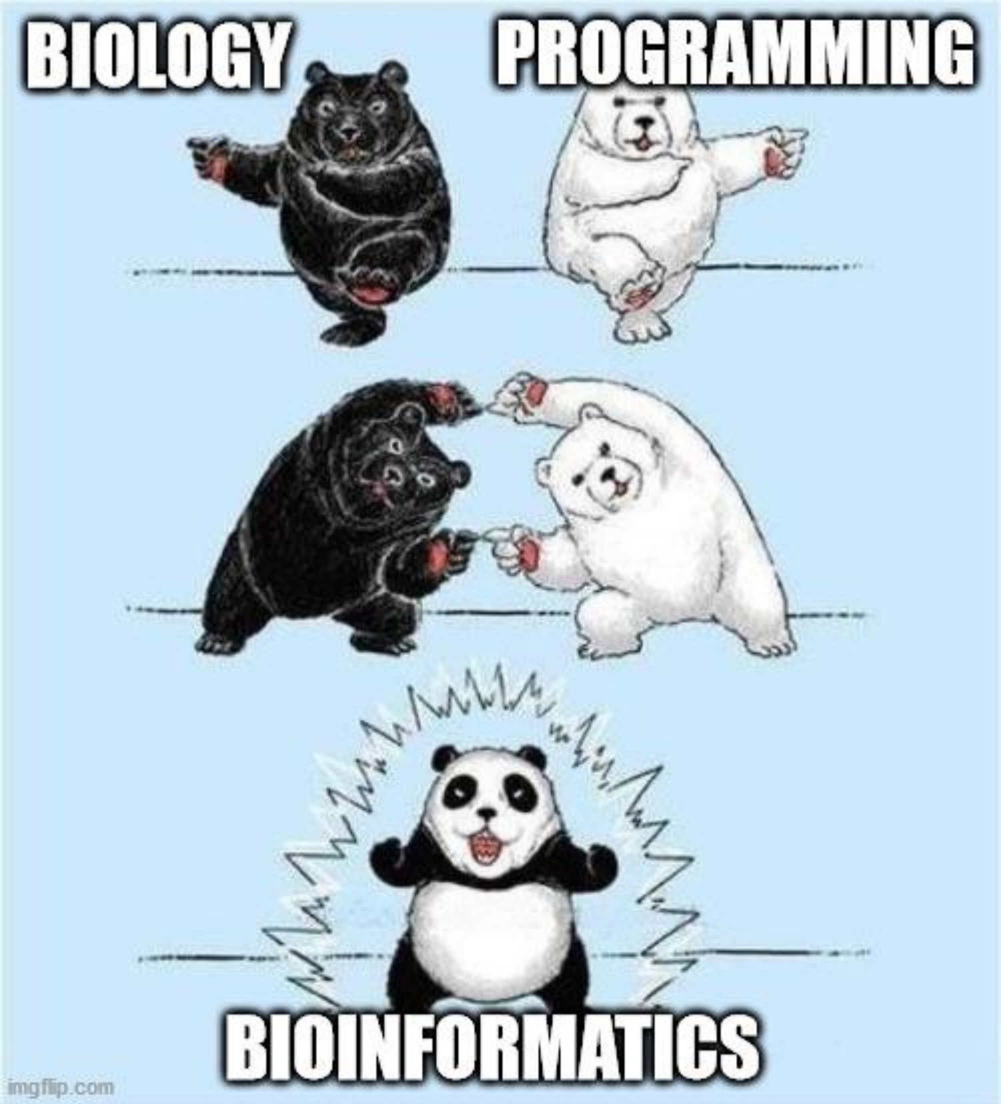
Mais aussi…

- Nos Adminsys burkinabé : Ousmane Barra, Seydou Konsimbo, Ndomassi Tando… Ezéchiel
- Le comité d’organisation : Ezéchiel, Fidéle Tiendrebeogo, Romaric Nanema, Isidore Boungoungou…
- Toutes nos tutelles, l’université JZK
- Le LMI Patho Bios : James Neya, Charlotte Tollenaere et Christophe Brugidou

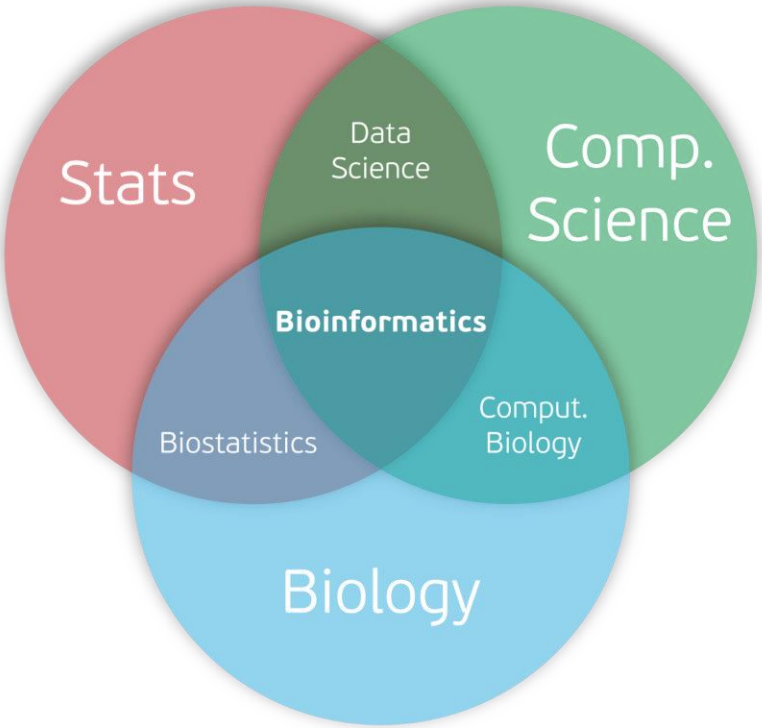


Introduction
Bioinformatics & Sequencing

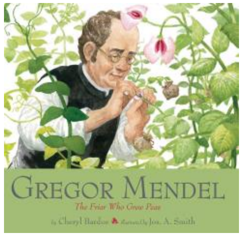
What is the bioinformatics ?

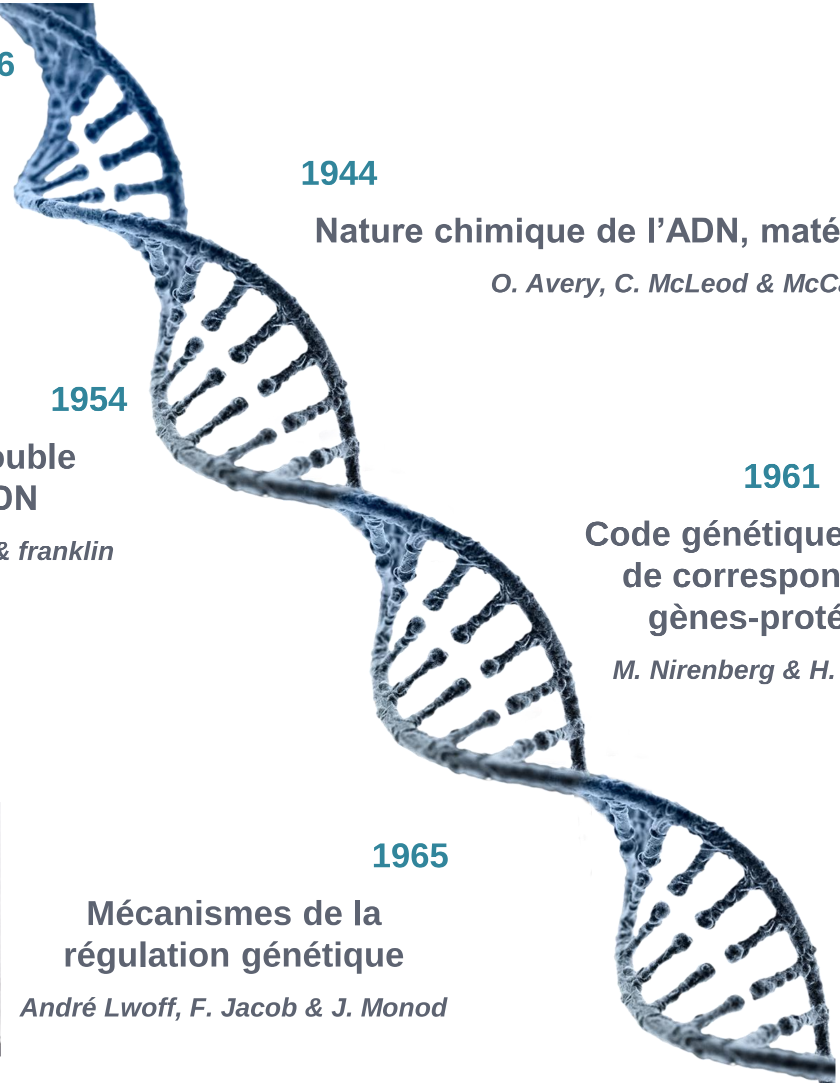


A interdisciplanary science



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

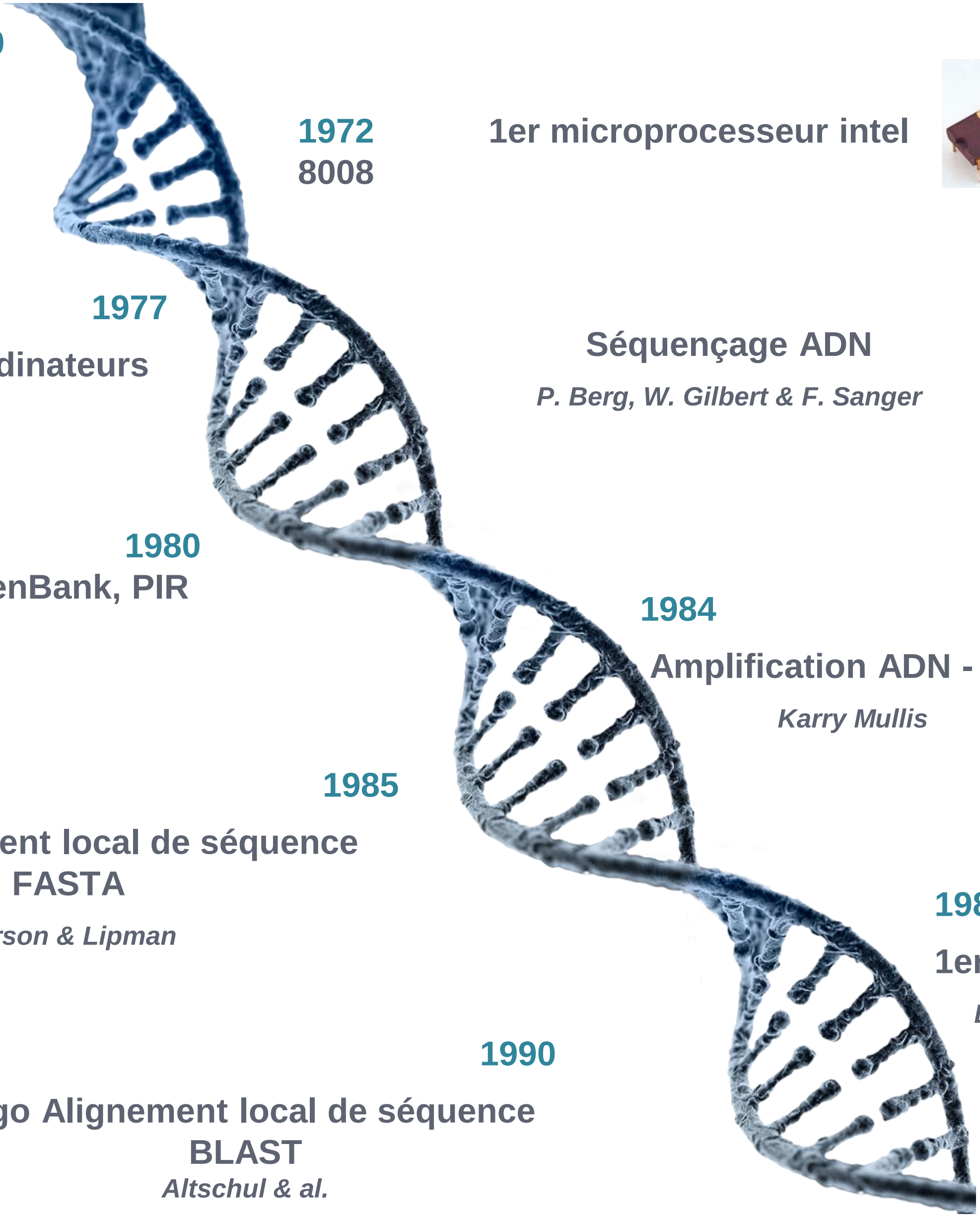
M. Nirenberg & H. Matthaei

1965

Mécanismes de la régulation génétique

André Lwoff, F. Jacob & J. Monod

De la génétique à la bioinformatique...



1970

Algo *Alignement global de séquence*

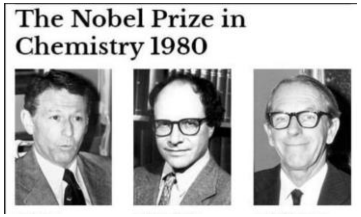
Needman, & Wunsh

1972
8008

1er microprocesseur intel

1977

Micro-ordinateurs



The Nobel Prize in Chemistry 1980

Séquençage ADN

P. Berg, W. Gilbert & F. Sanger

1980

Banque EMBL, GenBank, PIR

1984

Amplification ADN - PCR

Karry Mullis

1985

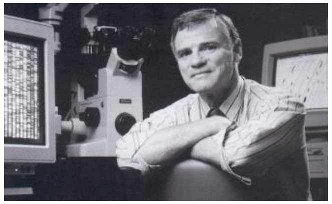
Algo Alignement local de séquence
FASTA

Person & Lipman

1987

1er séquenceur automaisé

L. Hood Société Applied Biosystems



1990

Algo Alignement local de séquence
BLAST
Altschul & al.

Your turn !

1 - How many base pairs (bp) are there in a human genome?

2 - How much did it cost to sequence the first human genome?

3 - How long did it take to sequence the first human genome?

Your turn !

1 - How many base pairs (bp) are there in a human genome?

3 billion

2 - How much did it cost to sequence the first human genome?

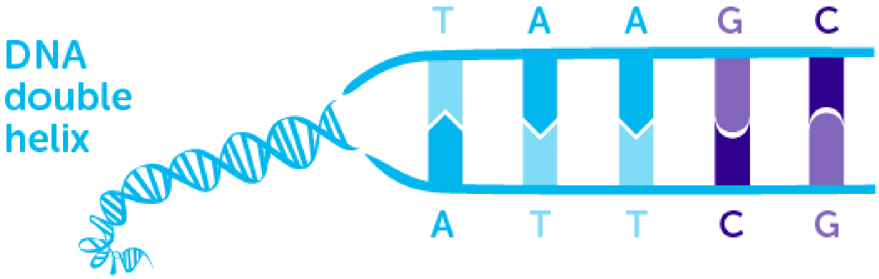
2.7 billions

3 - How long did it take to sequence the first human genome?

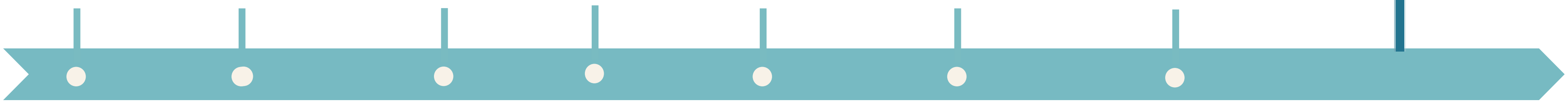
> 10 years

A little history of sequencing…

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

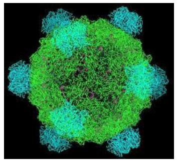


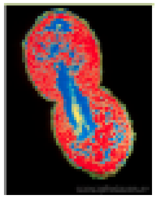
A little history of sequencing...

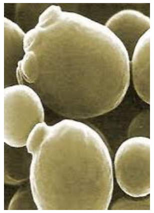


Next-generation sequencing

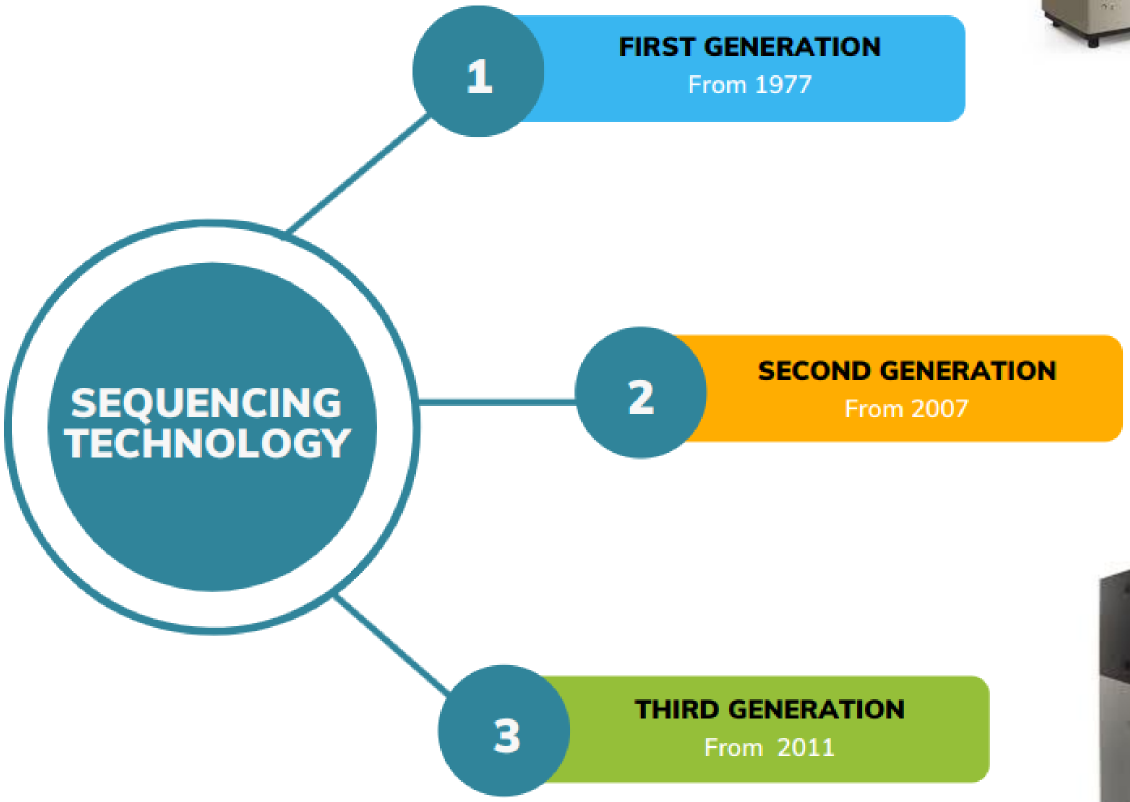
Virus phiX174	1977	5kb	11
H. influenza	1995	1,8 Mb	1,740
S. cerevisiae	1996	12 Mb	300
C. elegans	1998	100 Mb	22,000
A. thaliana	2000	125 Mb	~30,000
H. sapiens	2001	3,3 Gb	~21,000
O. sativa	2002	289 Mb	~35,000
Next-generation sequencing	2008		







Several sequencing technology



SEQUENCING TECHNOLOGY

1 FIRST GENERATION From 1977

2 SECOND GENERATION From 2007

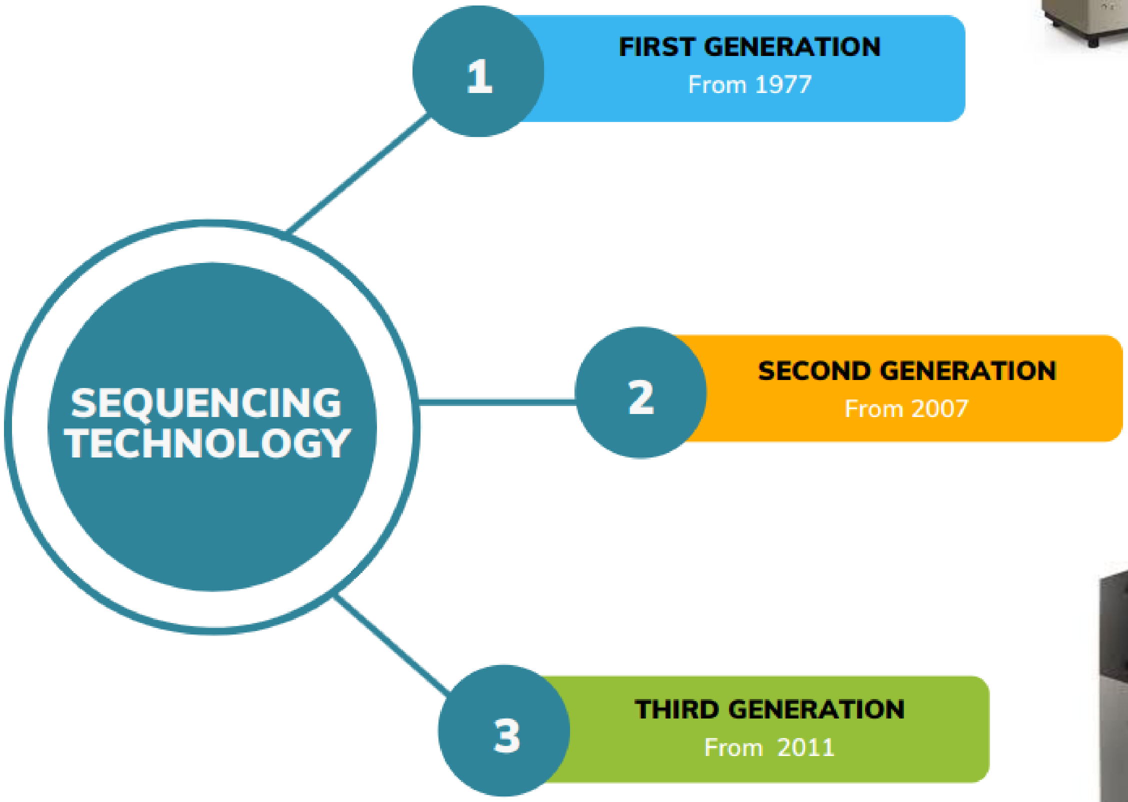
3 THIRD GENERATION From 2011

sanger

solexa, 454, illumina

PacBio, oxford
nanopore

Several sequencing technology



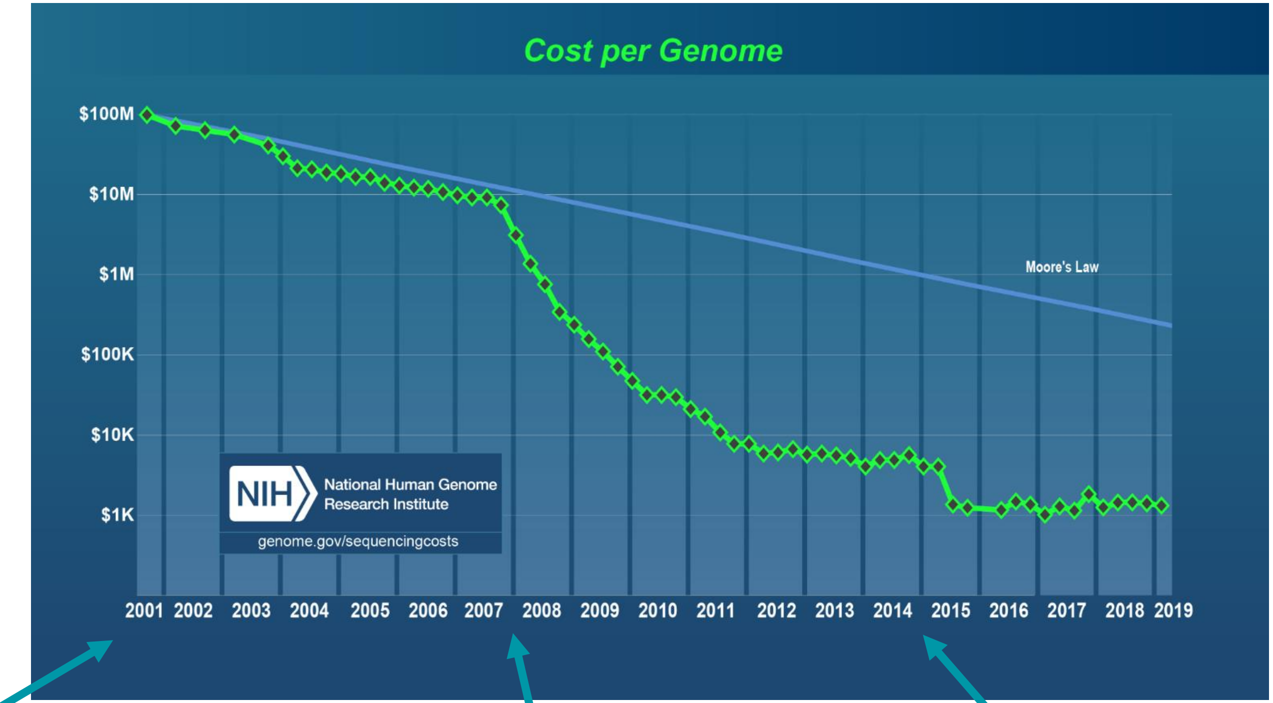
sanger

solexa, 454, illumina

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

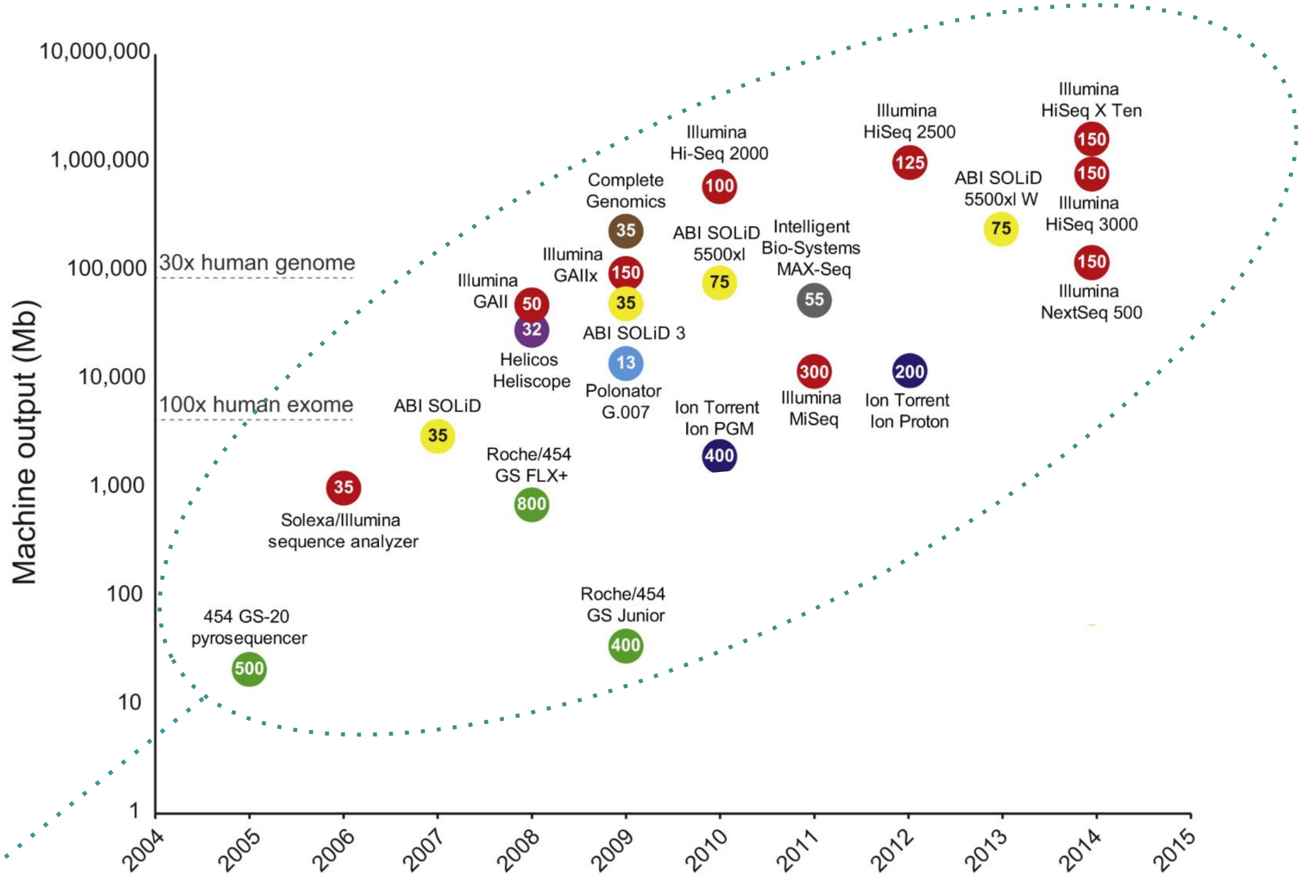
Illumina

3 **THIRD GENERATION**
From 2011

PacBio, ONT

From Sanger to 3rd sequencing technology

Une augmentation du débit de séquençage



2

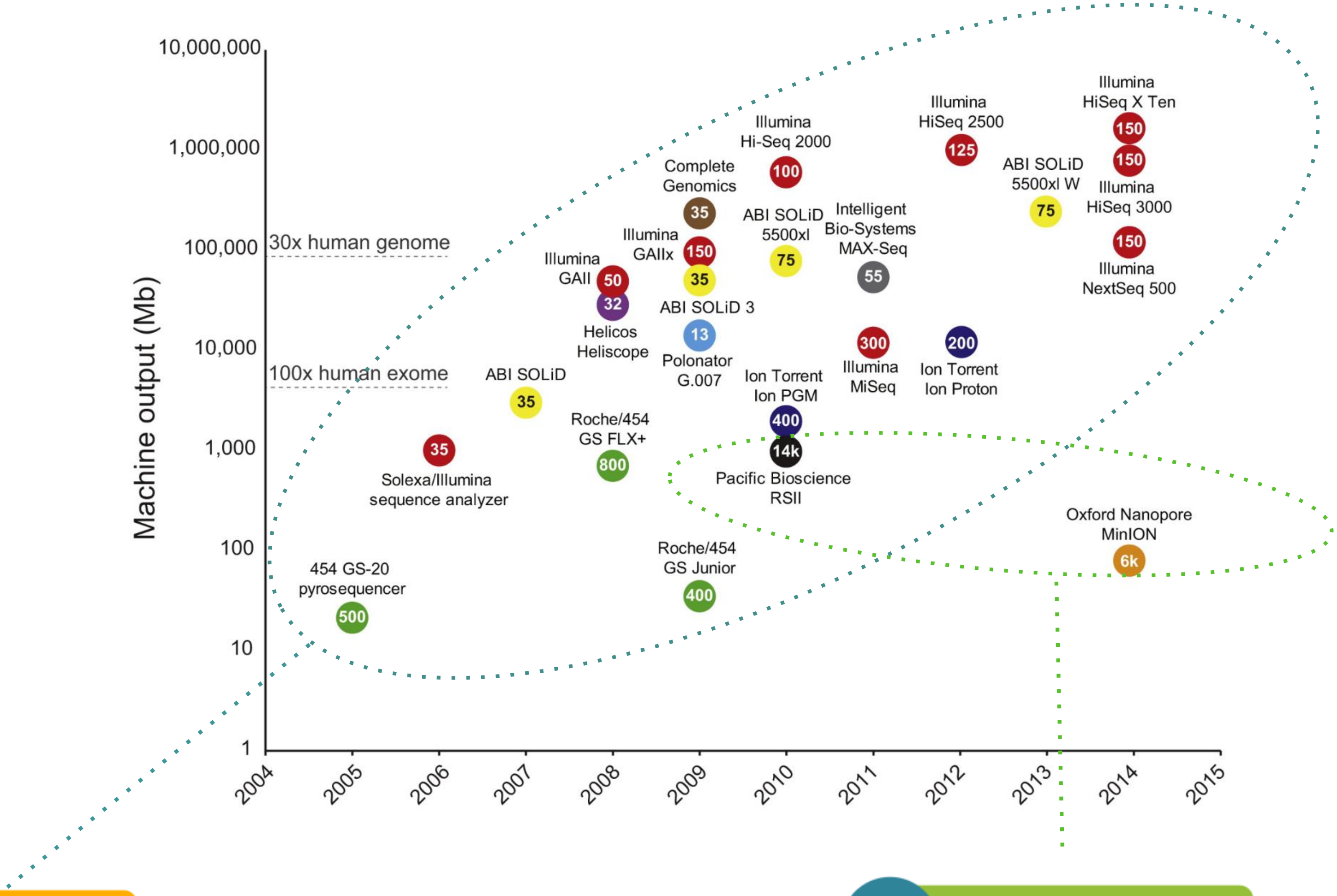
SECOND GENERATION

From 2007

Illumina

From Sanger to 3rd sequencing technology

Une augmentation du débit de séquençage



2 SECOND GENERATION From 2007

Illumina

3 THIRD GENERATION From 2011

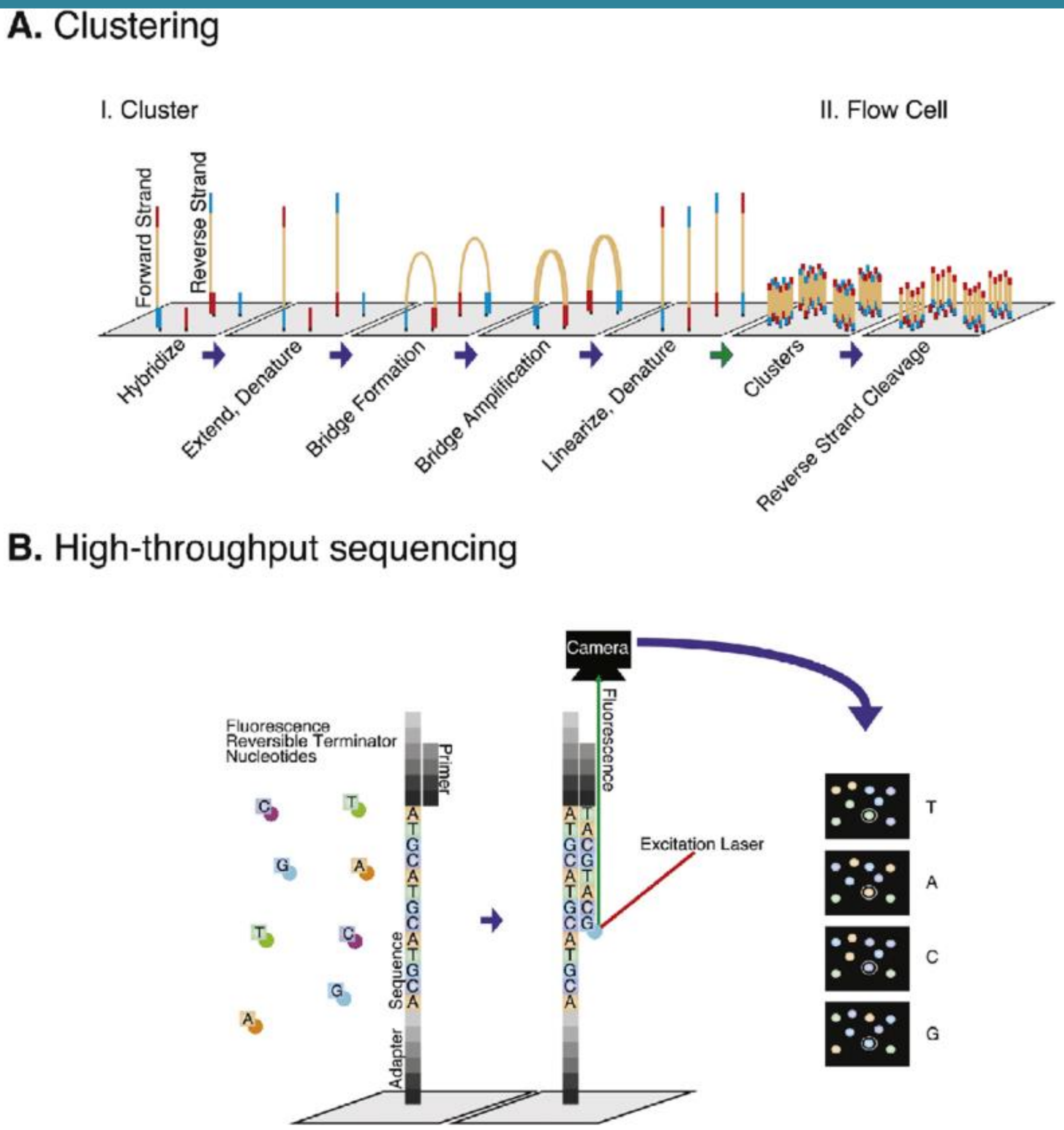
Short Reads ?
Long Reads ?

Short Reads - Illumina technology

2

SECOND GENERATION

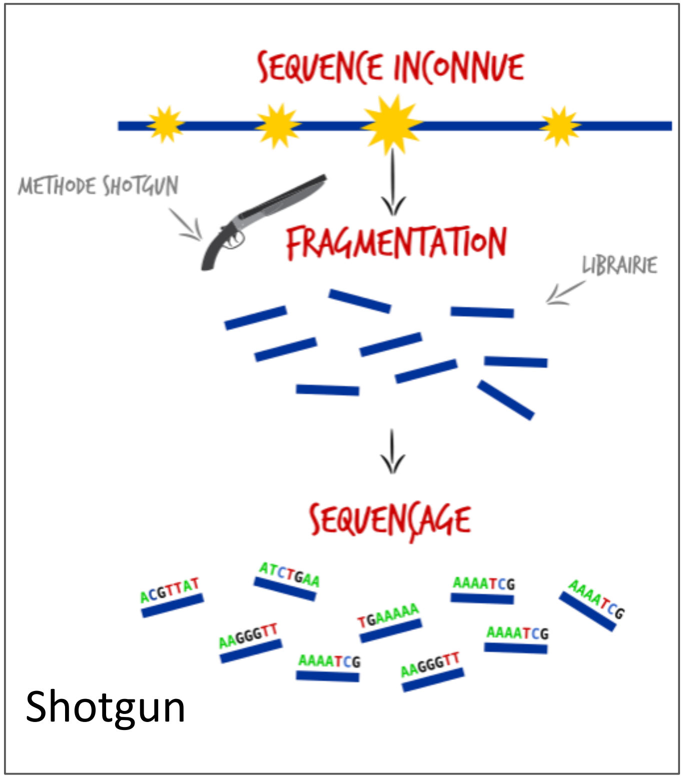
From 2007



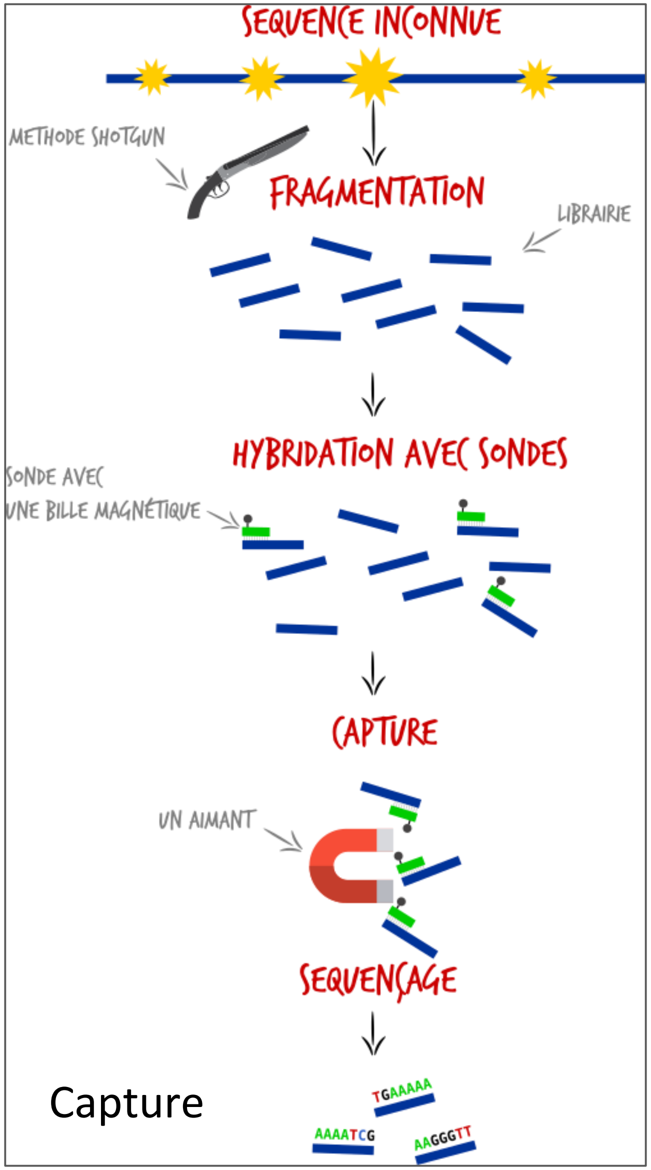
Vijender Chaitankar et al 2016

Short Reads - Illumina technology

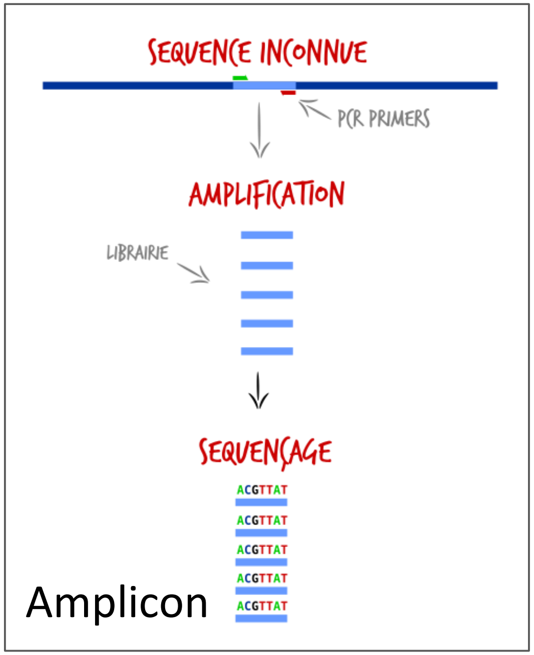
2 **SECOND GENERATION**
From 2007



Shotgun



Capture



Amplicon

https://dridk.me/ngs.html

Short Reads - Illumina - PROS & CONS

2 **SECOND GENERATION**
From 2007

- ✓ **Output volume** 20 billions of 150b reads, 6T

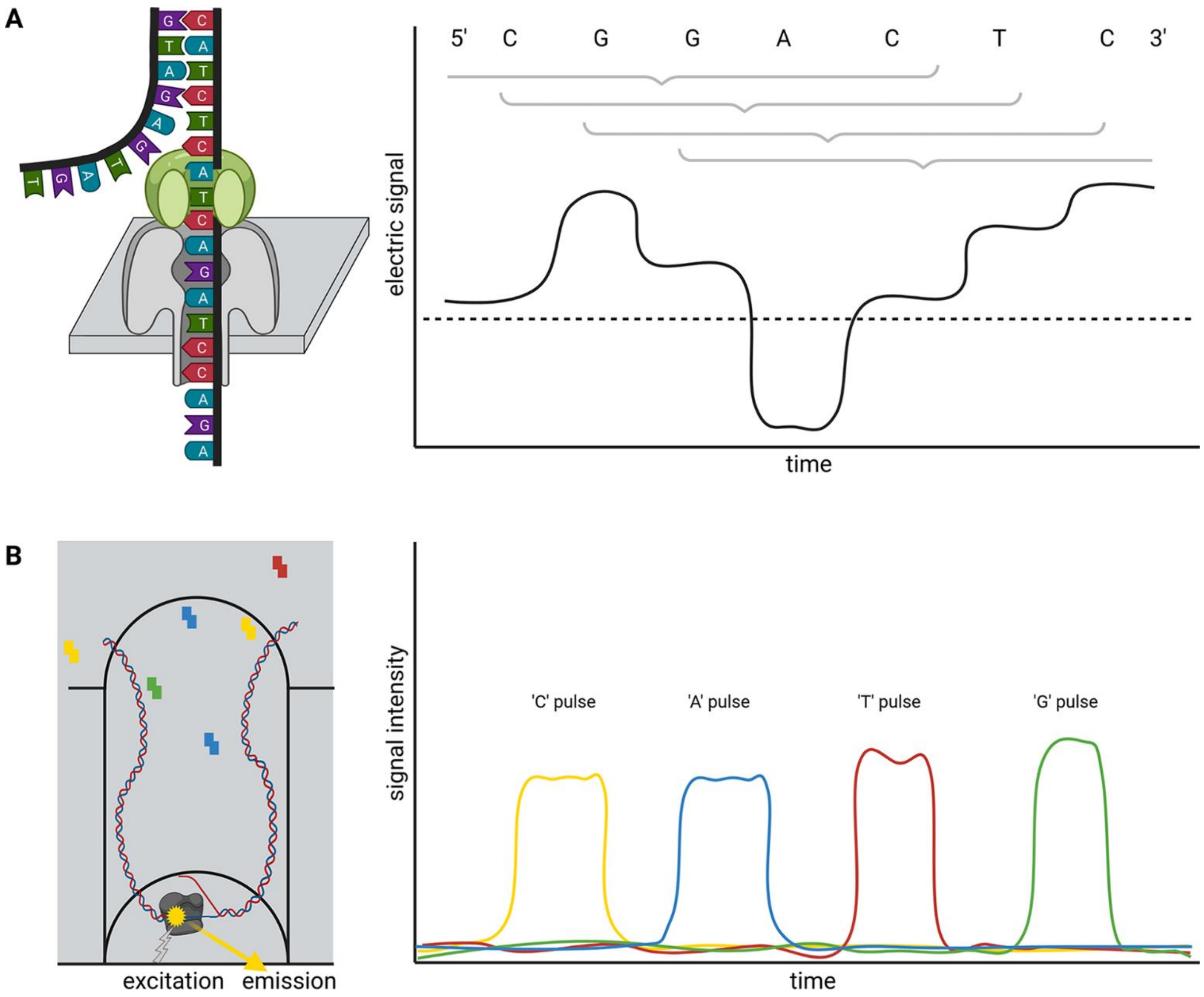
  NovaSeq6000
- ✓ **Accuracy** ~99 %
- ✓ **Run is cheap**
- ✓ **MySeq is cheap** ~60 000 USD per machine

- **Size** 150 + 150, *NovaSeq*

  but 400 pb, *MySeq*

Long Reads

3 **THIRD GENERATION**
From 2011



ONT is based on the translocation of a DNA or RNA strand through a nanopore located in an artificial membrane. Multiple nucleotides located in the nanopore determine the flow of ions through this nanopore in a specific way by physically blocking the space. This change in ion flux is recorded as an electric signal and further converted into sequence information.

Single-Molecule Real Time (**SMRT**) sequencing detects fluorescent light emitted from nucleotides upon incorporation into a DNA strand. The DNA polymerase is located at the bottom of a well and synthesises a new DNA strand. The integration into the new DNA strand keeps the nucleotide for a sufficiently long time in the well to allow detection.

Boas Pucker et al. 2022

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

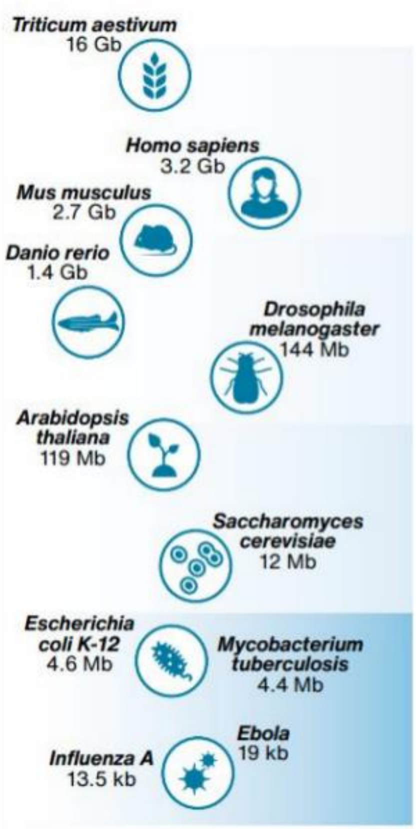
Plant genome project workflow from DNA extraction over ONT sequencing to data submission

	task	consumed time	hands-on time	equipment	estimated costs of consumables	estimated costs of lab equipment
A	plant incubation in darkness	2-3d	1h			
B	non-destructive sampling	-	1h			
C	DNA extraction	1d	8h	waterbath, centrifuge	$50	$1000 $8000
D	quality control	1h	1h	NanoDrop, Qubit	$20	
E	short fragment depletion	2h	1h	centrifuge	$50	
F	quality control	1h	1h	NanoDrop, Qubit	$20	$5000 $5000
G	library preparation & sequencing	1-5d	4-16h	centrifuge, magnetic rack, sequencer	$3000	$250 $1000
H	basecalling	1d	1h	computer with GPU		$3000
I	assembly	1-15d	1h			
J	polishing	1-5d	1h	compute cluster / cloud		
K	annotation	1-5d	1h			
L	data submission	2h	2h	fast internet connection		

Boas Pucker et al. 2022

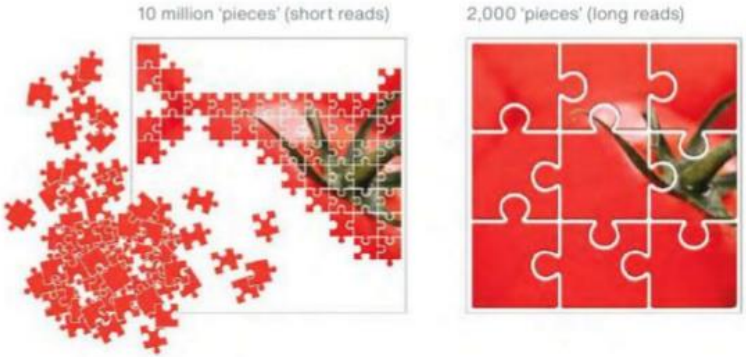
Long Reads

3 **THIRD GENERATION**
From 2011



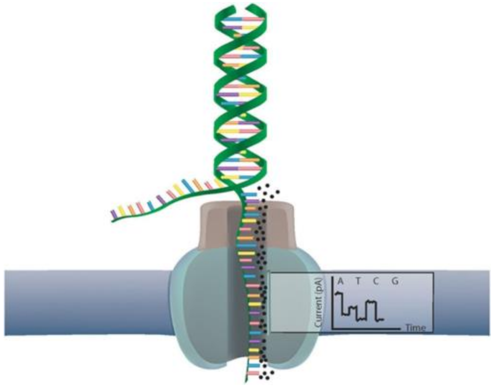
Microbial genomes | Human genomes | Animal genomes | Plant genomes

- 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


10 million 'pieces' (short reads)
2,000 'pieces' (long reads)

Long Reads - Oxford nanopore technology

3 **THIRD GENERATION**
From 2011



From Circulation Research

- No Amplification
- NO SYNTHESIS
- Very Long Length

Long Reads - Oxford nanopore - PROS & CONS

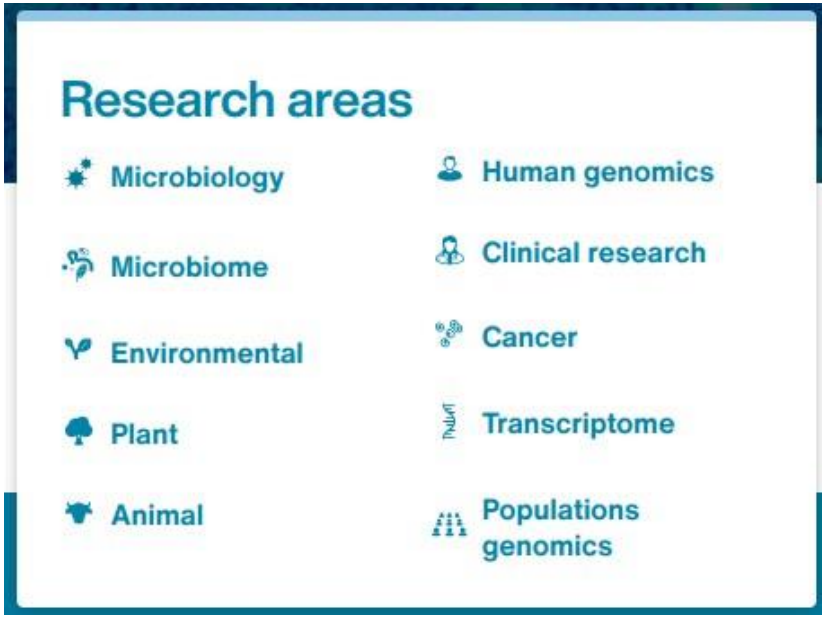
3 **THIRD GENERATION**
From 2011

- ✓ No Amplification, NO SYNTHESIS, Very Long Length
- ✓ 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 minimu
- ✓ **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



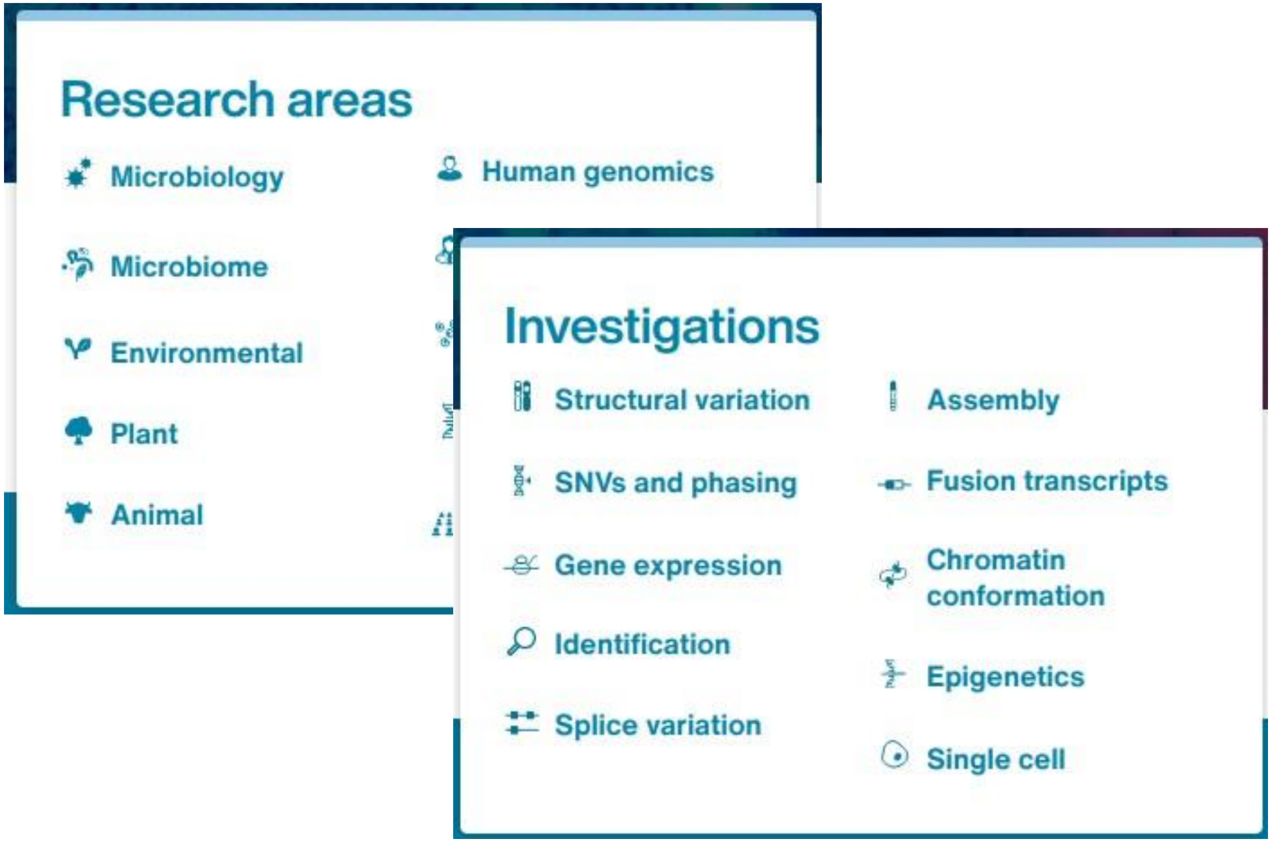
Research areas

- Microbiology
- Microbiome
- Environmental
- Plant
- Animal
- Human genomics
- Clinical research
- Cancer
- Transcriptome
- Populations genomics

From Nanopore website

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

3 THIRD GENERATION
From 2011



Research areas

- Microbiology
- Microbiome
- Environmental
- Plant
- Animal
- Human genomics

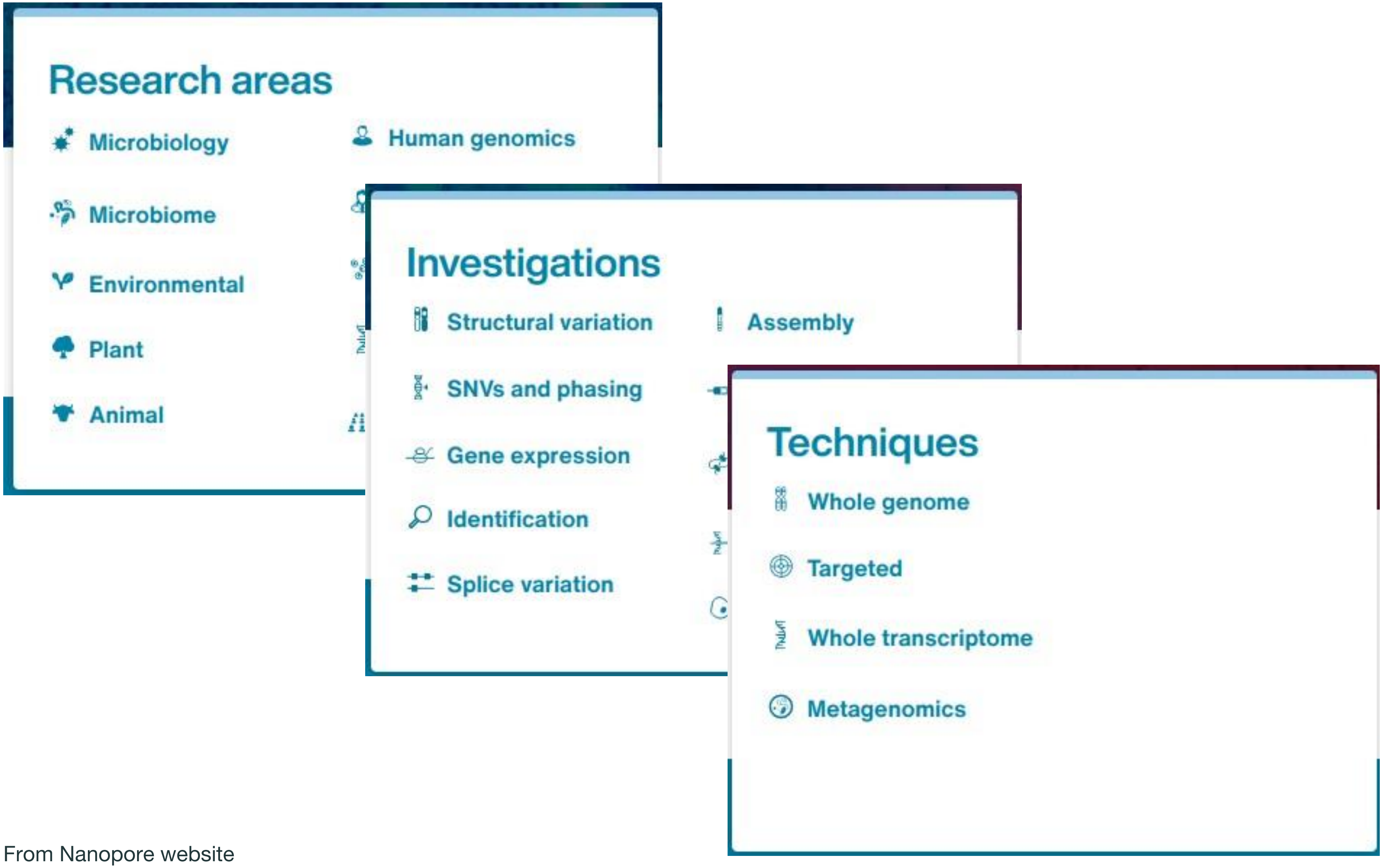
Investigations

- Structural variation
- SNVs and phasing
- Gene expression
- Identification
- Splice variation
- Assembly
- Fusion transcripts
- Chromatin conformation
- Epigenetics
- Single cell

From Nanopore website

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

3 **THIRD GENERATION**
From 2011



Research areas

- Microbiology
- Microbiome
- Environmental
- Plant
- Animal
- Human genomics

Investigations

- Structural variation
- SNVs and phasing
- Gene expression
- Identification
- Splice variation
- Assembly

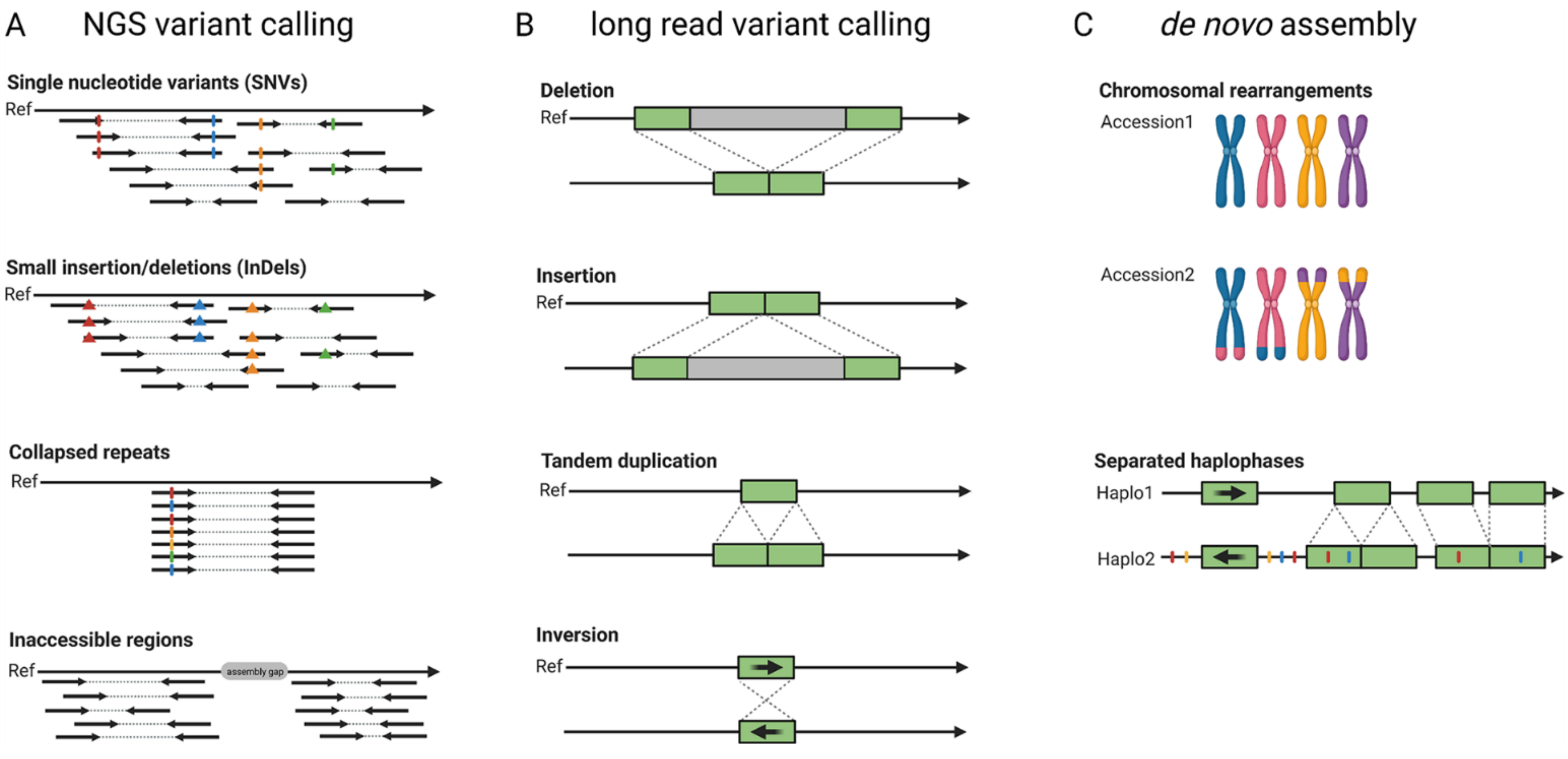
Techniques

- Whole genome
- Targeted
- Whole transcriptome
- Metagenomics

From Nanopore website

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

3 **THIRD GENERATION**
From 2011



Boas Pucker et al. 2022

Comparison

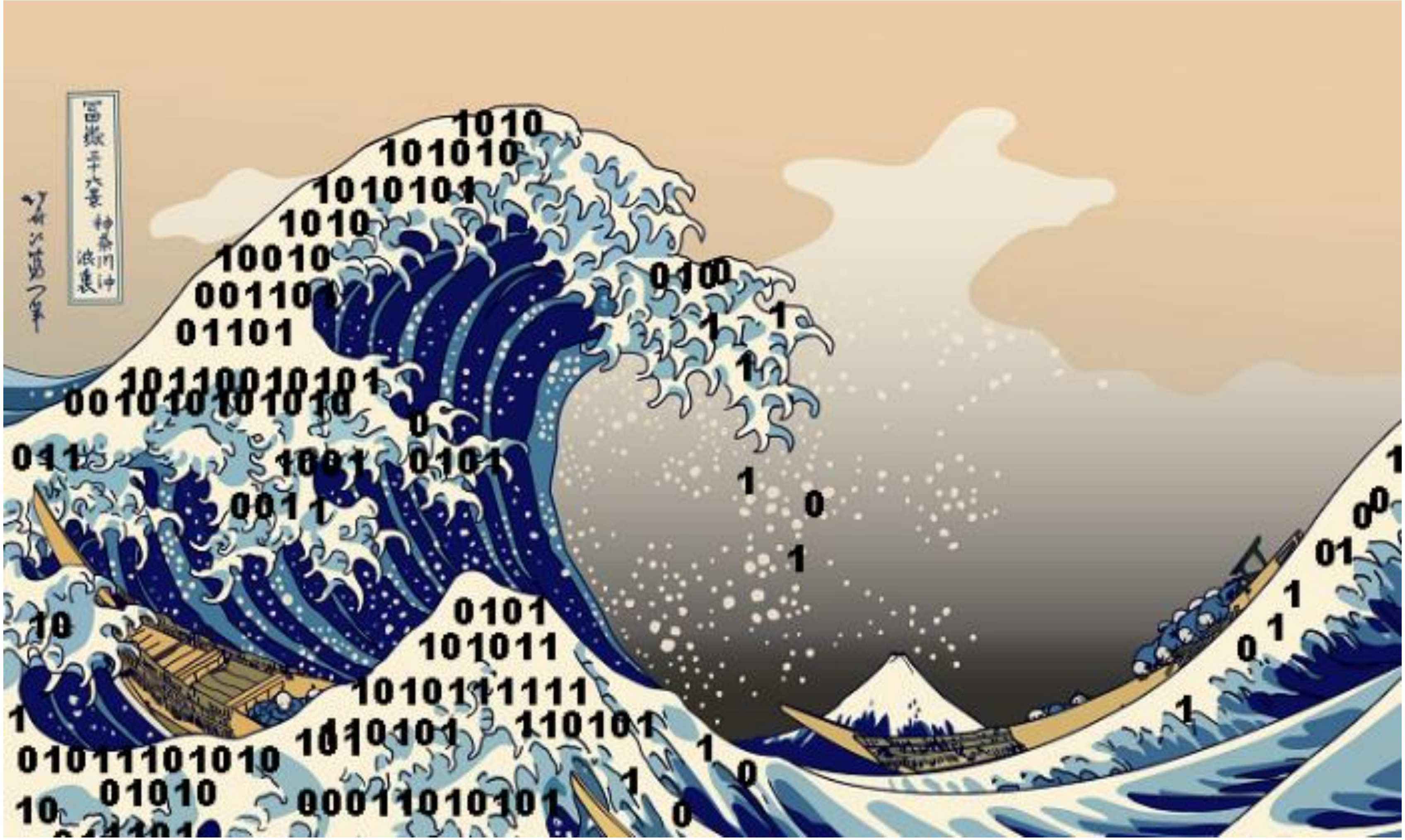
Evolution of sequencing technologies

Sequencers	1st generation : Sanger 3730xl (~ 2000)	2nd generation : Illumina HiSeq 2000 (~ 2006)	3rd generation : Pacific Biosciences RS II (~ 2012)
Method	Terminaison with dideoxynucleotides	Sequencing by synthesis with a polymerase	Real-time single molecule
Read length	400-900 nt	100 nt	10 000 nt
Error rate	0,001%	0,1%	15%
Amount of data produced at once	10^5 nt	10^{12} nt	10^{10} nt
Cost for 10^9 nt	2 000 000 €	80 €	1 000 €

Adapted from http://data-science-sequencing.github.io/lectures/lecture1/

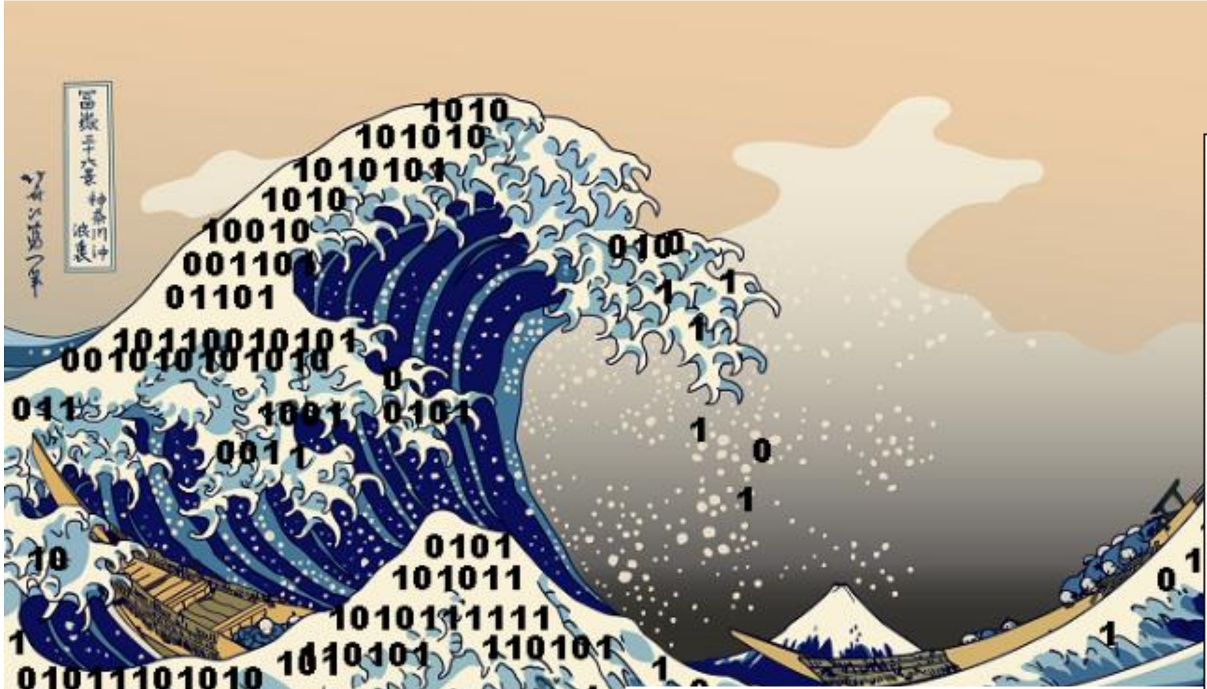
From Camille Rustenholz (Univ. Strasbourg, inrae) - Methods for plant genome assembly

From data rarity to data deluge

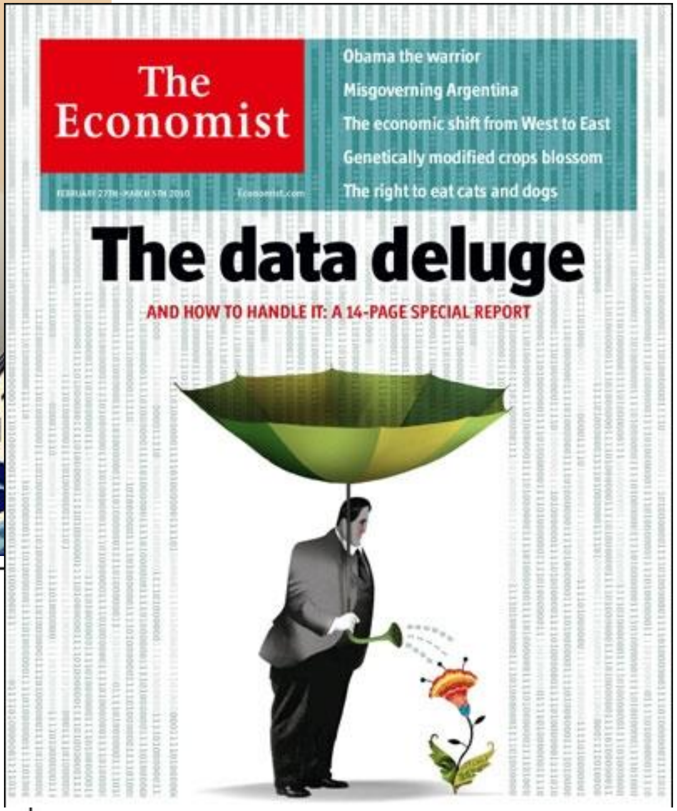


The Great Wave off Kanagawa, Hokusa @amitechsolutions.com

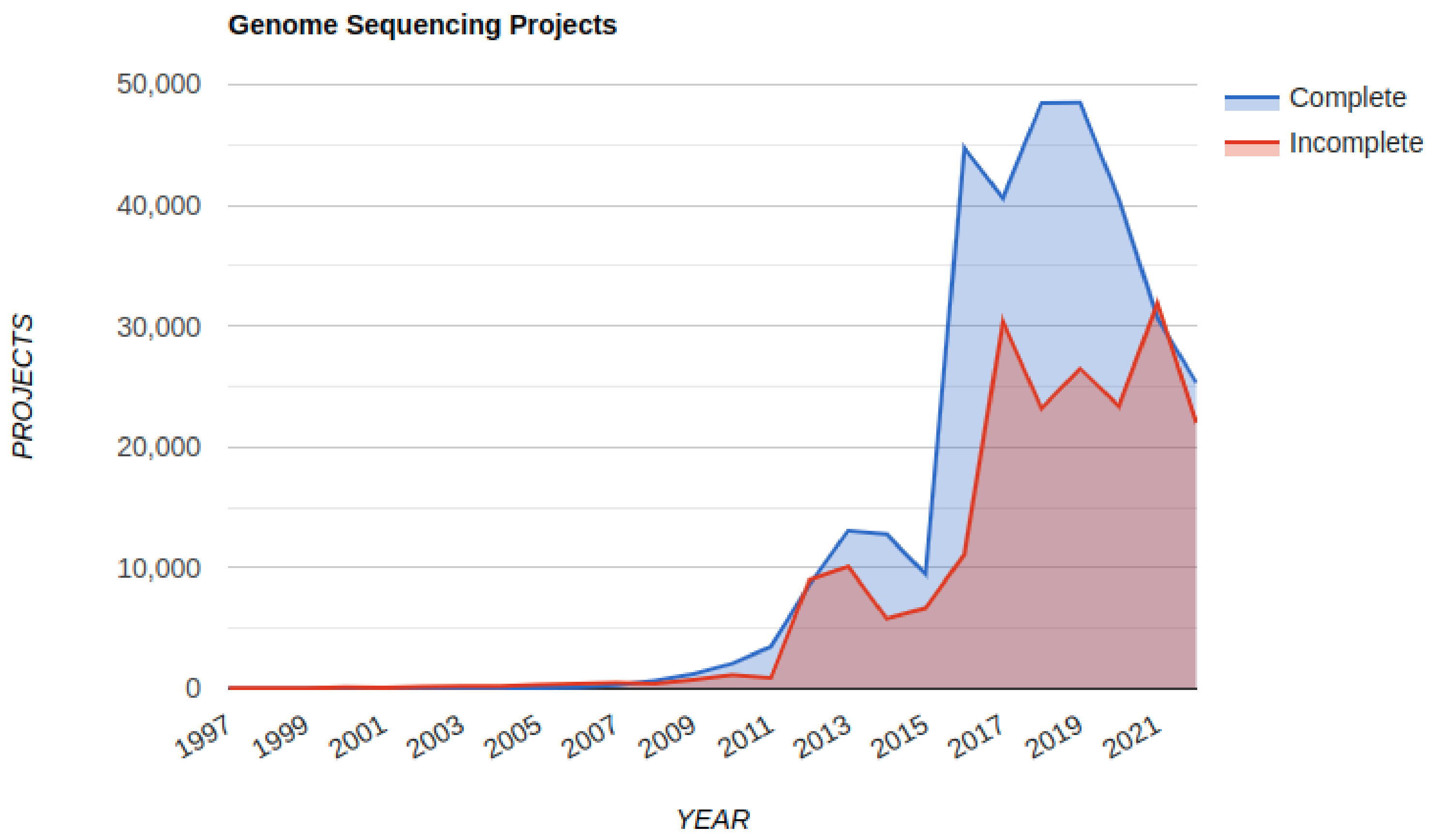
From data rarity to data deluge



© D.Fletcher for CloudTweaks.com


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

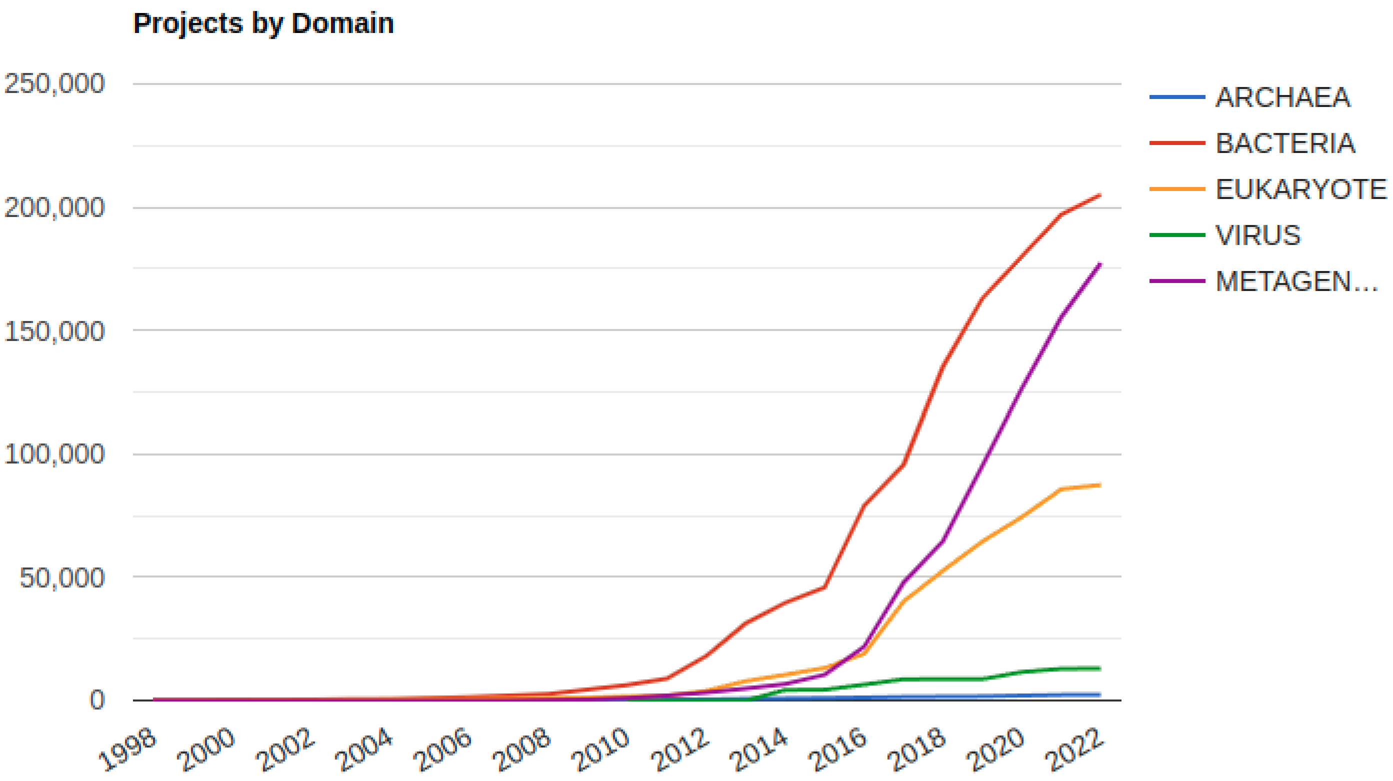
Genome Totals by year and status



GOLD
GENOMES ONLINE DATABASE

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

Project Totals by year and domain group



Projects by Domain

Phylogenetic distribution of Bacterial Genome Projects

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

Numerous resources available

Biological Databases

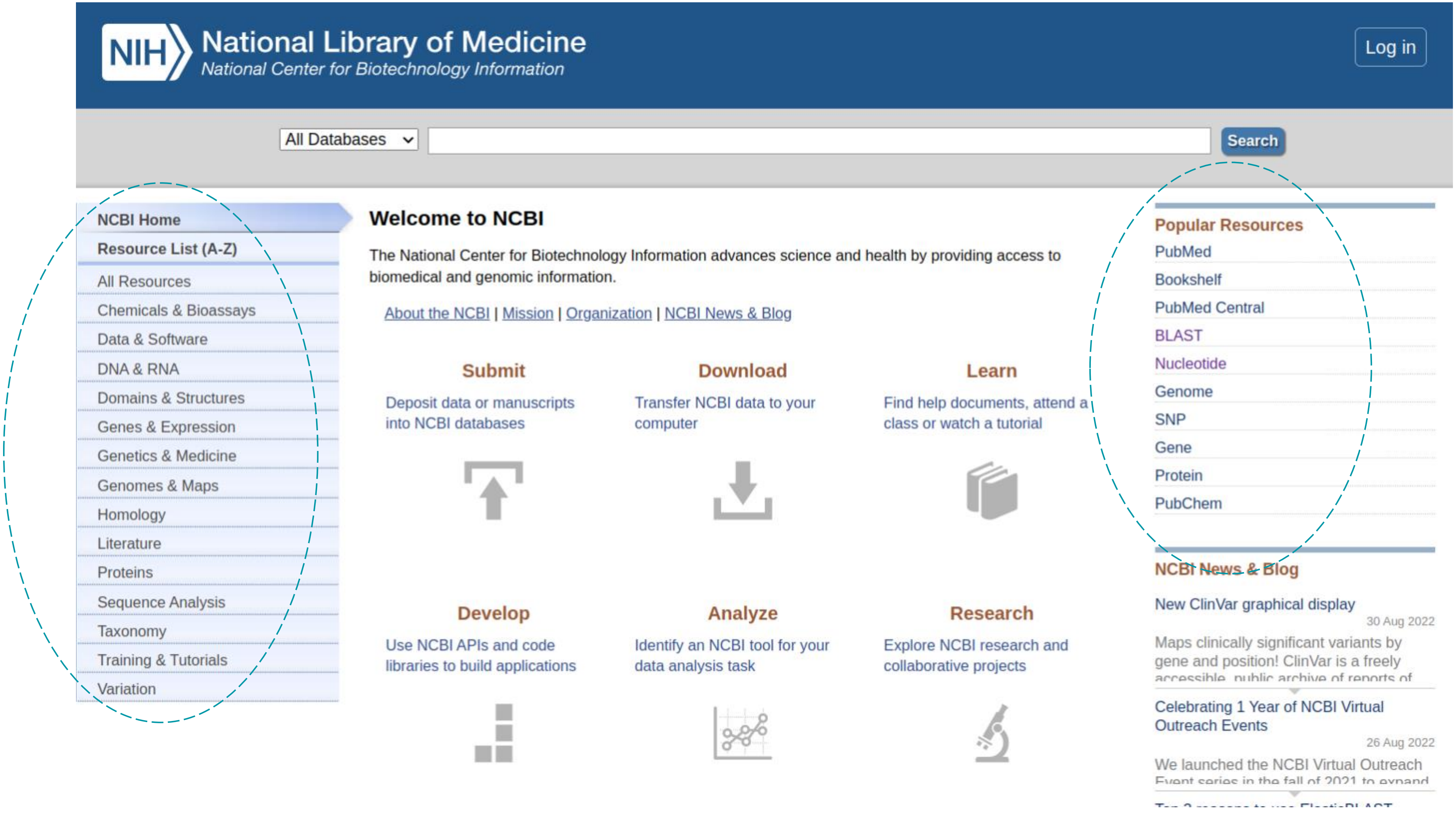
- ✓ **Sequence**
  - ○ Nucleic : **Genbank** EMBL-EBI DDBJ DNA Data Bank of Japan
  - ○ Proteic : SIB swissprot EBI TrEMBL

    PIR, Pfam, Prosite

- ✓ **Structure** PDB SCOP CATH

- ✓ **Specialized** by organism, by sequence type

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



National Library of Medicine
National Center for Biotechnology Information

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

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

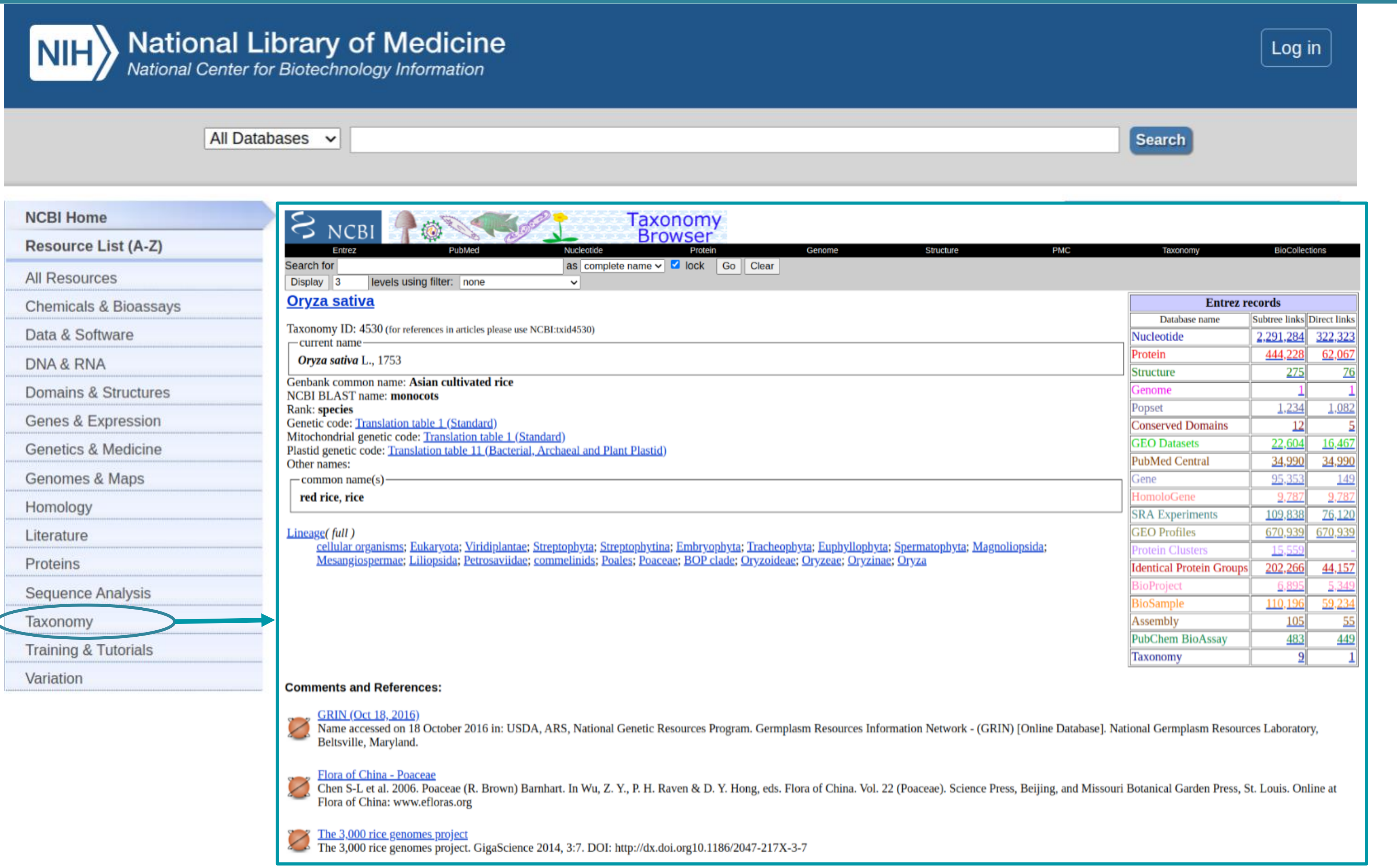
NCBI News & Blog

New ClinVar graphical display
30 Aug 2022
Maps clinically significant variants by gene and position! ClinVar is a freely accessible, public archive of reports of

Celebrating 1 Year of NCBI Virtual Outreach Events
26 Aug 2022
We launched the NCBI Virtual Outreach Event series in the fall of 2021 to expand

Top 3 reasons to use ElasticBLAST

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



NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
All Databases
Search

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

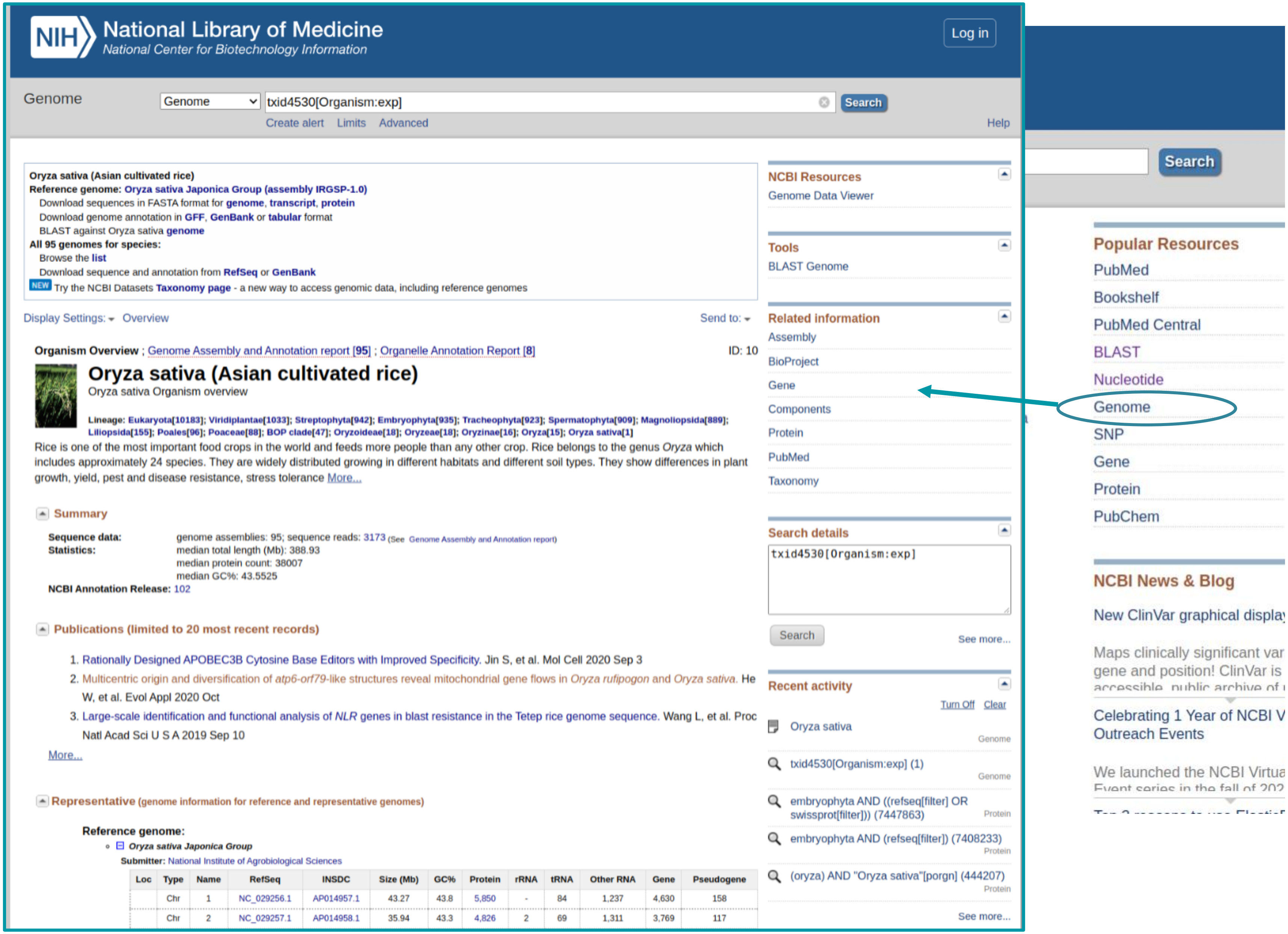
Comments and References:

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/


NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
Genome
Genome
Search
Create alert
Limits
Advanced
Help

txid4530[Organism:exp]

Oryza sativa (Asian cultivated rice)
Reference genome: **Oryza sativa Japonica Group (assembly IRGSP-1.0)**
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Oryza sativa **genome**
All 95 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**
NEW Try the NCBI Datasets **Taxonomy page** - a new way to access genomic data, including reference genomes

Display Settings: Overview   Send to:

Organism Overview ; Genome Assembly and Annotation report [95] ; Organelle Annotation Report [8]   ID: 10

Oryza sativa (Asian cultivated rice)

Oryza sativa Organism overview

Lineage: **Eukaryota[10183]; Viridiplantae[1033]; Streptophyta[942]; Embryophyta[935]; Tracheophyta[923]; Spermatophyta[909]; Magnoliopsida[889]; Liliopsida[155]; Poales[96]; Poaceae[88]; BOP clade[47]; Oryzoideae[18]; Oryzeae[18]; Oryzinae[16]; Oryza[15]; Oryza sativa[1]**

Rice is one of the most important food crops in the world and feeds more people than any other crop. Rice belongs to the genus *Oryza* which includes approximately 24 species. They are widely distributed growing in different habitats and different soil types. They show differences in plant growth, yield, pest and disease resistance, stress tolerance More...

Summary

Sequence data: genome assemblies: 95; sequence reads: 3173 (See Genome Assembly and Annotation report)
Statistics: median total length (Mb): 388.93
median protein count: 38007
median GC%: 43.5525
NCBI Annotation Release: 102

Publications (limited to 20 most recent records)

1. Rationally Designed APOBEC3B Cytosine Base Editors with Improved Specificity. Jin S, et al. Mol Cell 2020 Sep 3
2. Multicentric origin and diversification of *atp6-orf79*-like structures reveal mitochondrial gene flows in *Oryza rufipogon* and *Oryza sativa*. He W, et al. Evol Appl 2020 Oct
3. Large-scale identification and functional analysis of *NLR* genes in blast resistance in the Tetep rice genome sequence. Wang L, et al. Proc Natl Acad Sci U S A 2019 Sep 10

More...

Representative (genome information for reference and representative genomes)

Reference genome:
- ⊟ ***Oryza sativa Japonica Group***

Submitter: National Institute of Agrobiological Sciences

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Chr	1	NC_029256.1	AP014957.1	43.27	43.8	5,850	-	84	1,237	4,630	158
	Chr	2	NC_029257.1	AP014958.1	35.94	43.3	4,826	2	69	1,311	3,769	117

NCBI Resources
Genome Data Viewer

Tools
BLAST Genome

Related information
Assembly
BioProject
Gene
Components
Protein
PubMed
Taxonomy

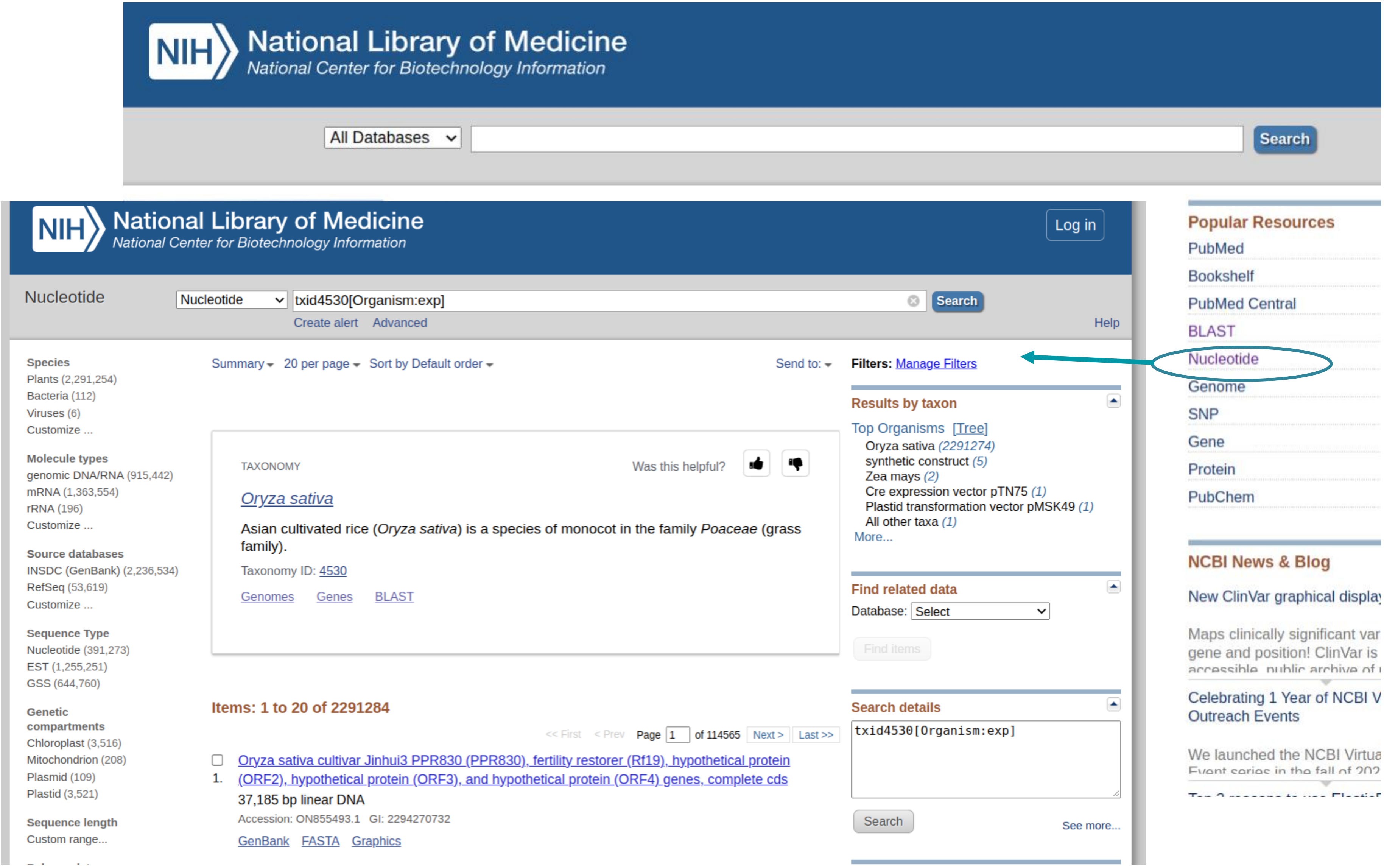
Search details
txid4530[Organism:exp]
Search   See more...

Recent activity
Turn Off Clear
Oryza sativa — Genome
txid4530[Organism:exp] (1) — Genome
embryophyta AND ((refseq[filter] OR swissprot[filter])) (7447863) — Protein
embryophyta AND (refseq[filter]) (7408233) — Protein
(oryza) AND "Oryza sativa"[porgn] (444207) — Protein
See more...

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog
New ClinVar graphical display
Maps clinically significant var gene and position! ClinVar is accessible public archive of
Celebrating 1 Year of NCBI V Outreach Events
We launched the NCBI Virtua Event series in the fall of 202

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



Nucleotide: txid4530[Organism:exp]

Species
Plants (2,291,254)
Bacteria (112)
Viruses (6)
Customize ...

Molecule types
genomic DNA/RNA (915,442)
mRNA (1,363,554)
rRNA (196)
Customize ...

Source databases
INSDC (GenBank) (2,236,534)
RefSeq (53,619)
Customize ...

Sequence Type
Nucleotide (391,273)
EST (1,255,251)
GSS (644,760)

Genetic compartments
Chloroplast (3,516)
Mitochondrion (208)
Plasmid (109)
Plastid (3,521)

Sequence length
Custom range...

TAXONOMY — Was this helpful?

Oryza sativa

Asian cultivated rice (*Oryza sativa*) is a species of monocot in the family *Poaceae* (grass family).

Taxonomy ID: 4530

Items: 1 to 20 of 2291284

Page 1 of 114565

1. Oryza sativa cultivar Jinhui3 PPR830 (PPR830), fertility restorer (Rf19), hypothetical protein (ORF2), hypothetical protein (ORF3), and hypothetical protein (ORF4) genes, complete cds
37,185 bp linear DNA
Accession: ON855493.1 GI: 2294270732

Results by taxon

Top Organisms [Tree]
Oryza sativa (2291274)
synthetic construct (5)
Zea mays (2)
Cre expression vector pTN75 (1)
Plastid transformation vector pMSK49 (1)
All other taxa (1)
More...

Find related data
Database: Select

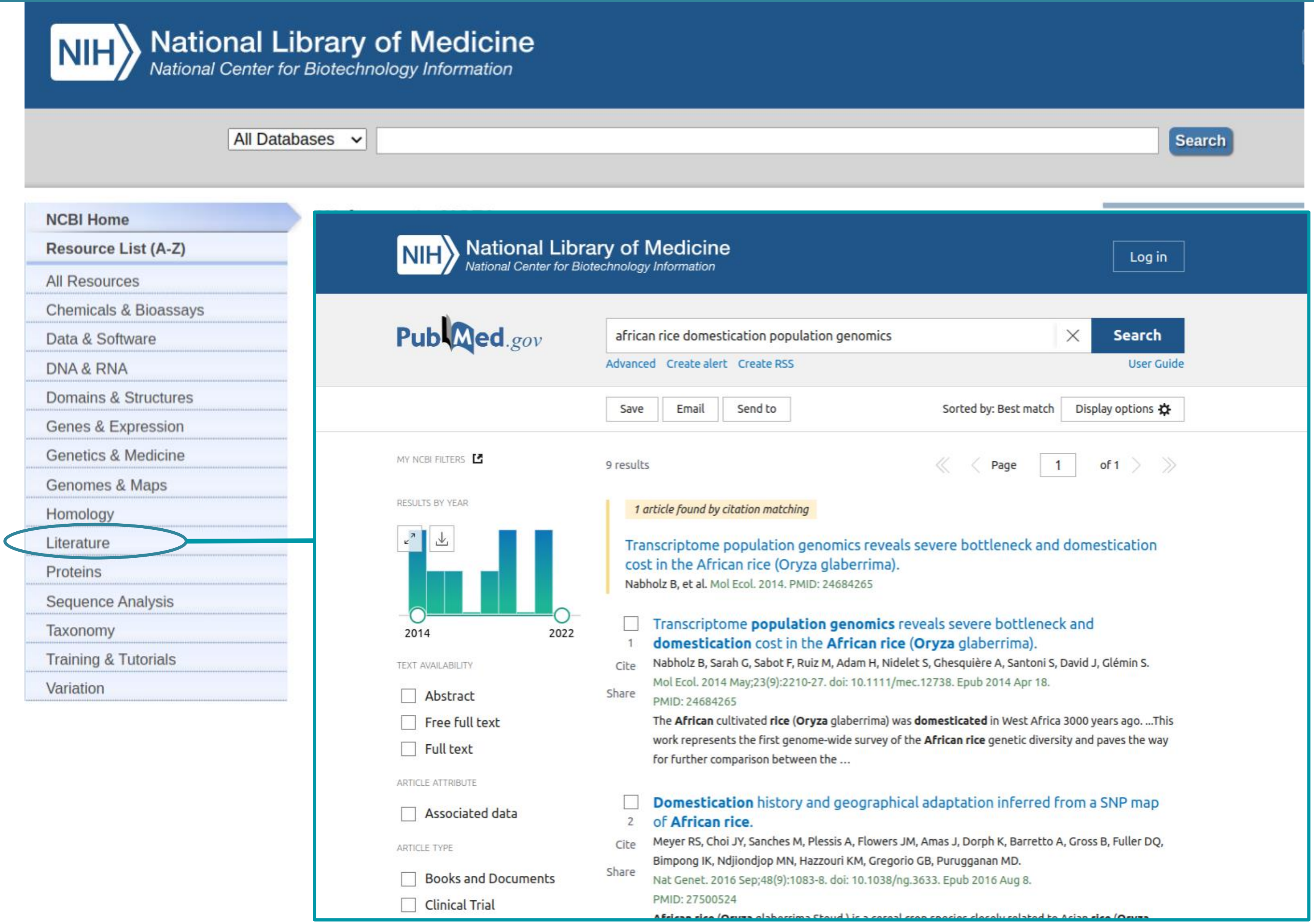
Search details
txid4530[Organism:exp]

New ClinVar graphical display

Maps clinically significant var gene and position! ClinVar is accessible, public archive of

Celebrating 1 Year of NCBI V Outreach Events

We launched the NCBI Virtua Event series in the fall of 202 https://www.ncbi.nlm.nih.gov/



NIH National Library of Medicine
National Center for Biotechnology Information

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

NIH National Library of Medicine
National Center for Biotechnology Information
Log in

PubMed.gov

african rice domestication population genomics ✕ Search

Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Best match Display options

MY NCBI FILTERS

9 results Page 1 of 1

RESULTS BY YEAR

1 article found by citation matching

Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima).
Nabholz B, et al. Mol Ecol. 2014. PMID: 24684265

TEXT AVAILABILITY
□ Abstract
□ Free full text
□ Full text

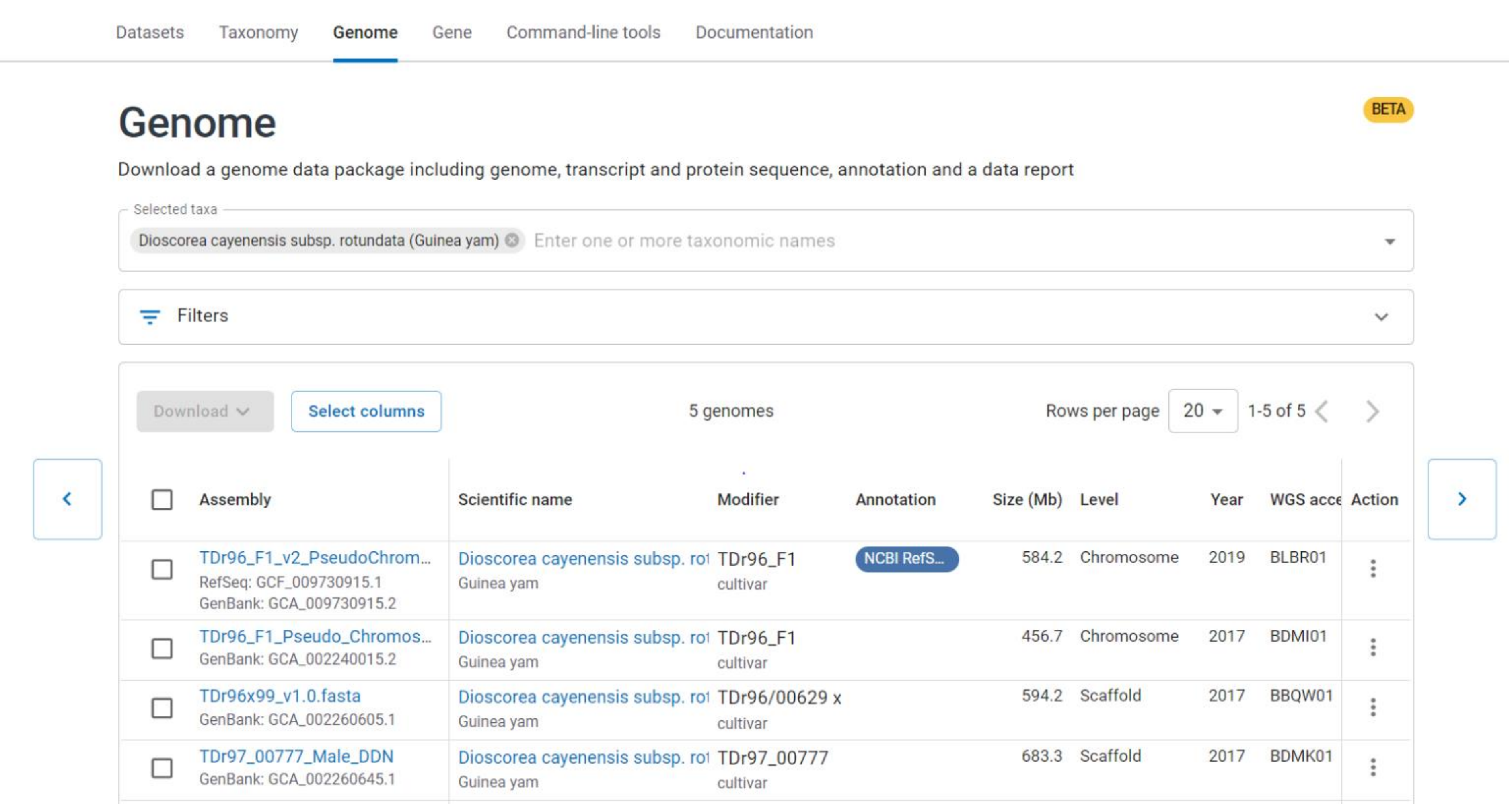
ARTICLE ATTRIBUTE
□ Associated data

ARTICLE TYPE
□ Books and Documents
□ Clinical Trial

□ 1 Cite Share
Transcriptome **population genomics** reveals severe bottleneck and **domestication** cost in the **African rice** (**Oryza** glaberrima).
Nabholz B, Sarah G, Sabot F, Ruiz M, Adam H, Nidelet S, Ghesquière A, Santoni S, David J, Glémin S.
Mol Ecol. 2014 May;23(9):2210-27. doi: 10.1111/mec.12738. Epub 2014 Apr 18.
PMID: 24684265
The **African** cultivated **rice** (**Oryza** glaberrima) was **domesticated** in West Africa 3000 years ago. ...This work represents the first genome-wide survey of the **African rice** genetic diversity and paves the way for further comparison between the ...

□ 2 Cite Share
Domestication history and geographical adaptation inferred from a SNP map of **African rice**.
Meyer RS, Choi JY, Sanches M, Plessis A, Flowers JM, Amas J, Dorph K, Barretto A, Gross B, Fuller DQ, Bimpong IK, Ndjiondjop MN, Hazzouri KM, Gregorio GB, Purugganan MD.
Nat Genet. 2016 Sep;48(9):1083-8. doi: 10.1038/ng.3633. Epub 2016 Aug 8.
PMID: 27500524
African rice (**Oryza** glaberrima Steud.) is a cereal crop species closely related to Asian **rice** (**Oryza**

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



Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

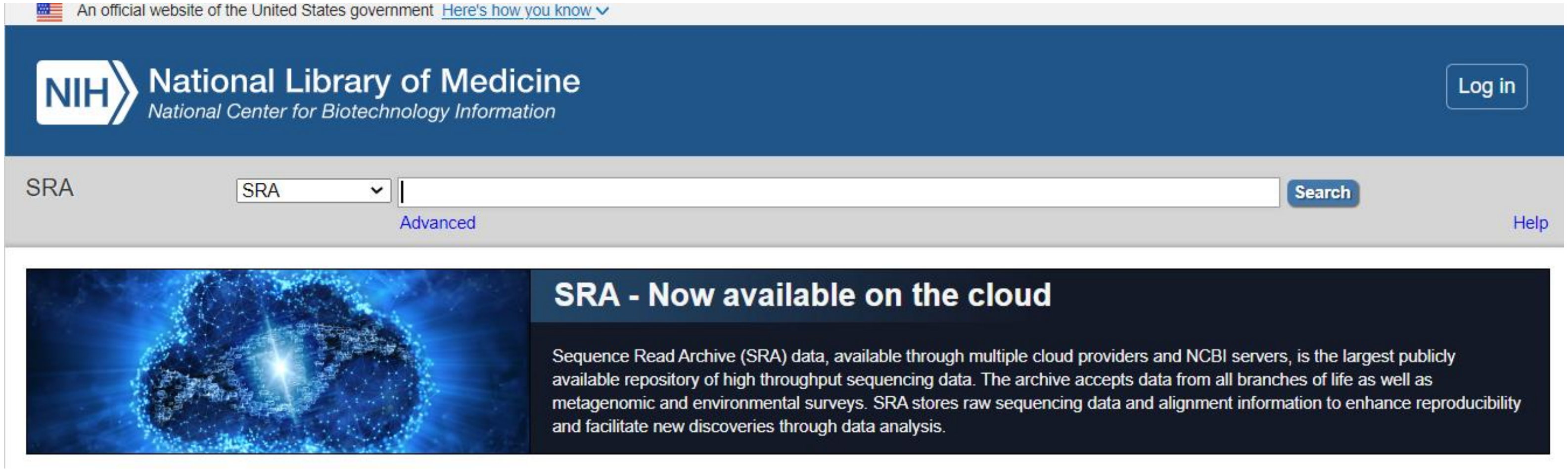
Selected taxa: Dioscorea cayenensis subsp. rotundata (Guinea yam) ⊗ Enter one or more taxonomic names

5 genomes

Rows per page 20 1-5 of 5

□	Assembly	Scientific name	Modifier	Annotation	Size (Mb)	Level	Year	WGS acce	Action
□	TDr96_F1_v2_PseudoChrom... RefSeq: GCF_009730915.1 GenBank: GCA_009730915.2	Dioscorea cayenensis subsp. rot Guinea yam	TDr96_F1 cultivar	NCBI RefS...	584.2	Chromosome	2019	BLBR01	⋮
□	TDr96_F1_Pseudo_Chromos... GenBank: GCA_002240015.2	Dioscorea cayenensis subsp. rot Guinea yam	TDr96_F1 cultivar		456.7	Chromosome	2017	BDMI01	⋮
□	TDr96x99_v1.0.fasta GenBank: GCA_002260605.1	Dioscorea cayenensis subsp. rot Guinea yam	TDr96/00629 x cultivar		594.2	Scaffold	2017	BBQW01	⋮
□	TDr97_00777_Male_DDN GenBank: GCA_002260645.1	Dioscorea cayenensis subsp. rot Guinea yam	TDr97_00777 cultivar		683.3	Scaffold	2017	BDMK01	⋮

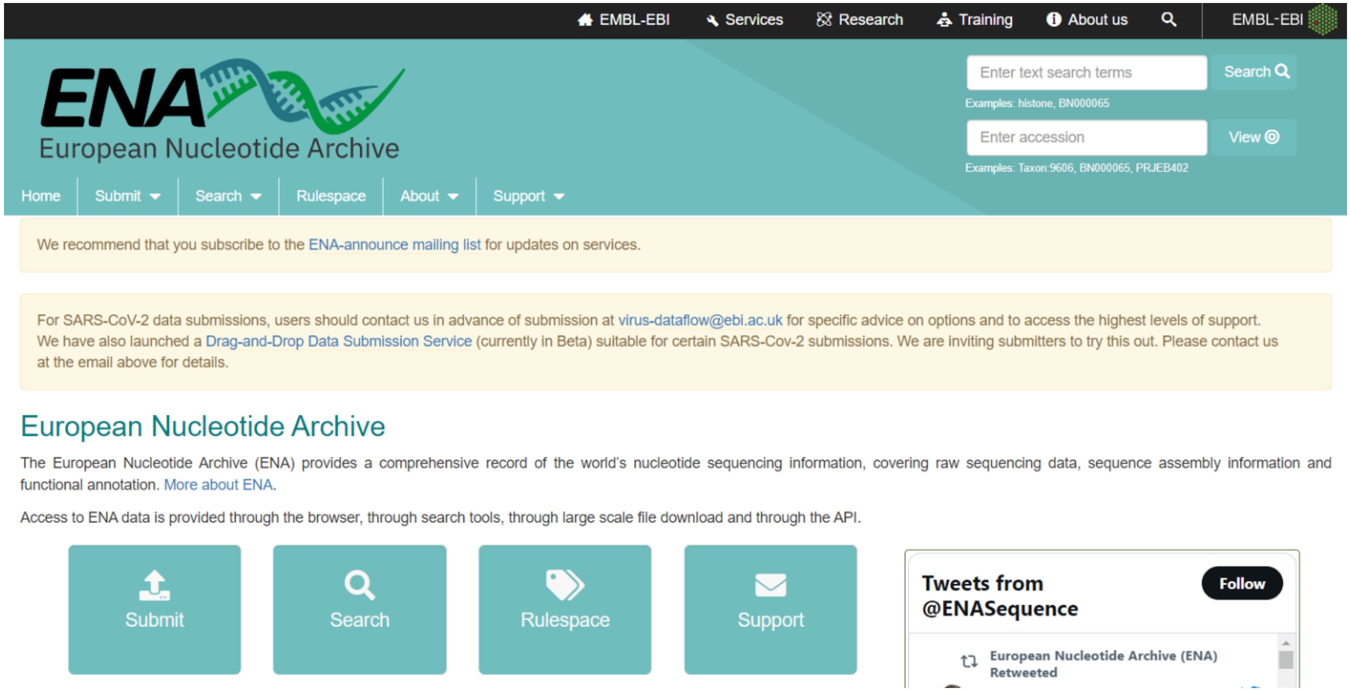
SRA (Sequence Reads Archive) / ENA (European Nucleotide Archive)



An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
SRA
SRA
Search
Advanced
Help

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.



We recommend that you subscribe to the ENA-announce mailing list for updates on services.

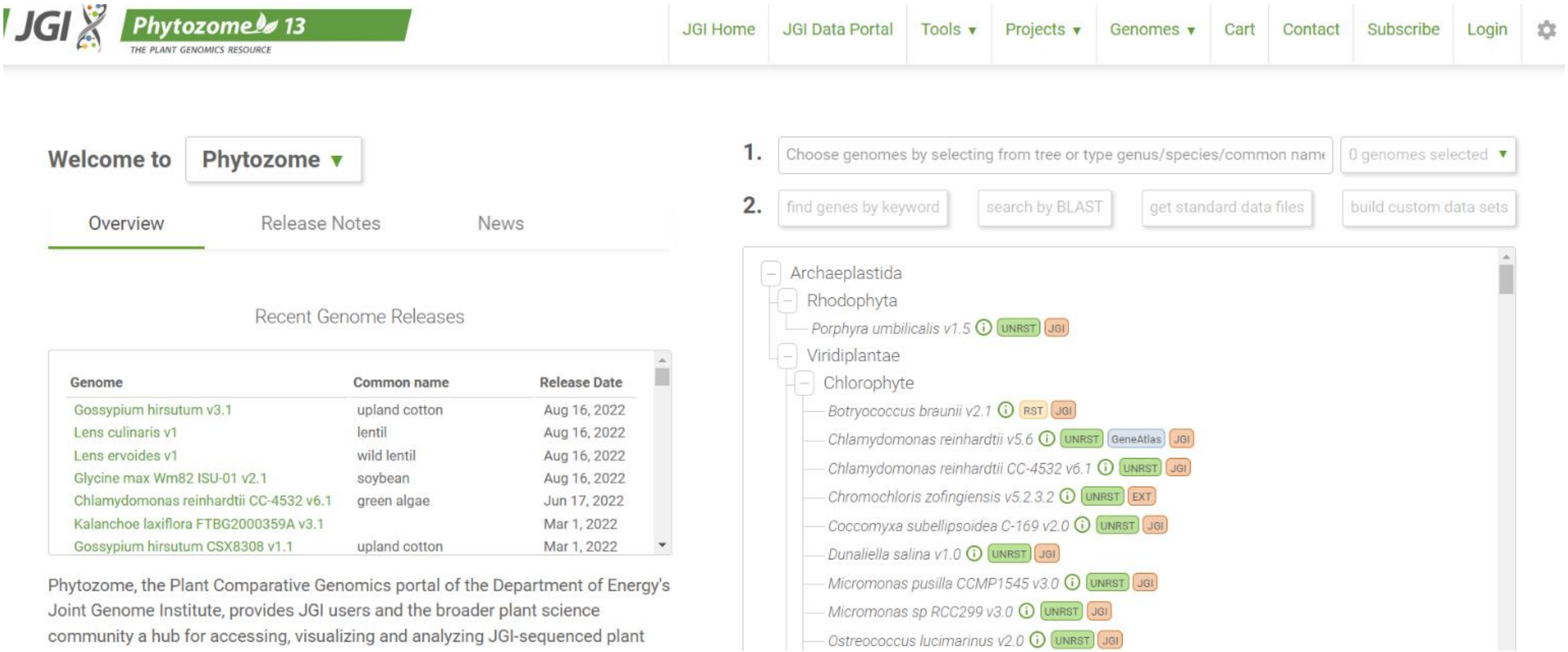
For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a Drag-and-Drop Data Submission Service (currently in Beta) suitable for certain SARS-Cov-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

Phytozome



Welcome to Phytozome ▾

Overview | Release Notes | News

Recent Genome Releases

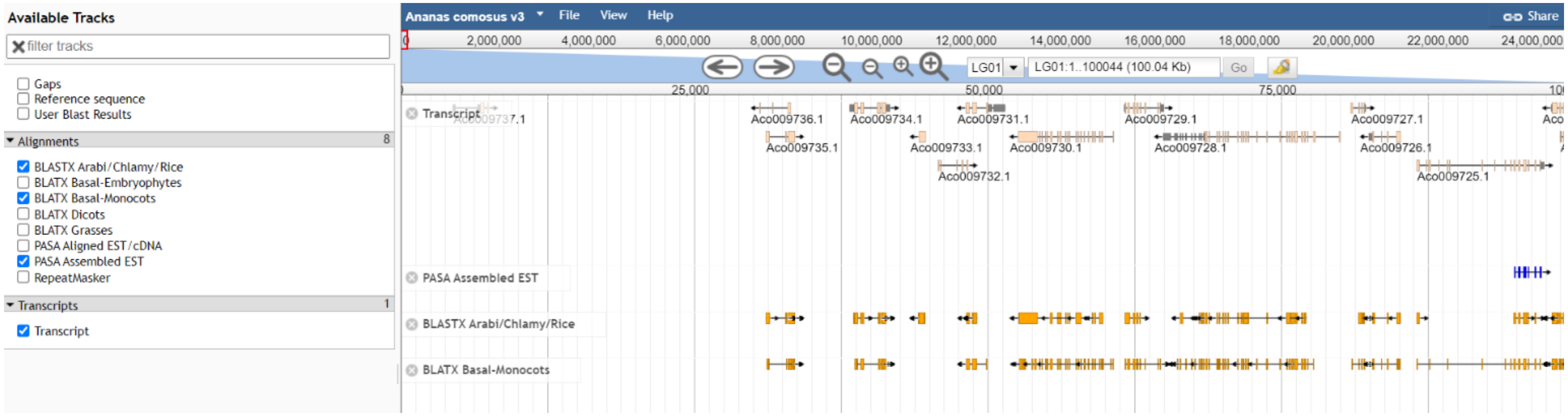
Genome	Common name	Release Date
Gossypium hirsutum v3.1	upland cotton	Aug 16, 2022
Lens culinaris v1	lentil	Aug 16, 2022
Lens ervoides v1	wild lentil	Aug 16, 2022
Glycine max Wm82 ISU-01 v2.1	soybean	Aug 16, 2022
Chlamydomonas reinhardtii CC-4532 v6.1	green algae	Jun 17, 2022
Kalanchoe laxiflora FTBG2000359A v3.1		Mar 1, 2022
Gossypium hirsutum CSX8308 v1.1	upland cotton	Mar 1, 2022

Phytozome, the Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute, provides JGI users and the broader plant science community a hub for accessing, visualizing and analyzing JGI-sequenced plant

1. Choose genomes by selecting from tree or type genus/species/common name | 0 genomes selected ▾

2. find genes by keyword | search by BLAST | get standard data files | build custom data sets

- Archaeplastida
  - Rhodophyta
    - *Porphyra umbilicalis v1.5* UNRST JGI
  - Viridiplantae
    - Chlorophyte
      - *Botryococcus braunii v2.1* RST JGI
      - *Chlamydomonas reinhardtii v5.6* UNRST GeneAtlas JGI
      - *Chlamydomonas reinhardtii CC-4532 v6.1* UNRST JGI
      - *Chromochloris zofingiensis v5.2.3.2* UNRST EXT
      - *Coccomyxa subellipsoidea C-169 v2.0* UNRST JGI
      - *Dunaliella salina v1.0* UNRST JGI
      - *Micromonas pusilla CCMP1545 v3.0* UNRST JGI
      - *Micromonas sp RCC299 v3.0* UNRST JGI
      - *Ostreococcus lucimarinus v2.0* UNRST JGI



Available Tracks

filter tracks

☐ Gaps
☐ Reference sequence
☐ User Blast Results

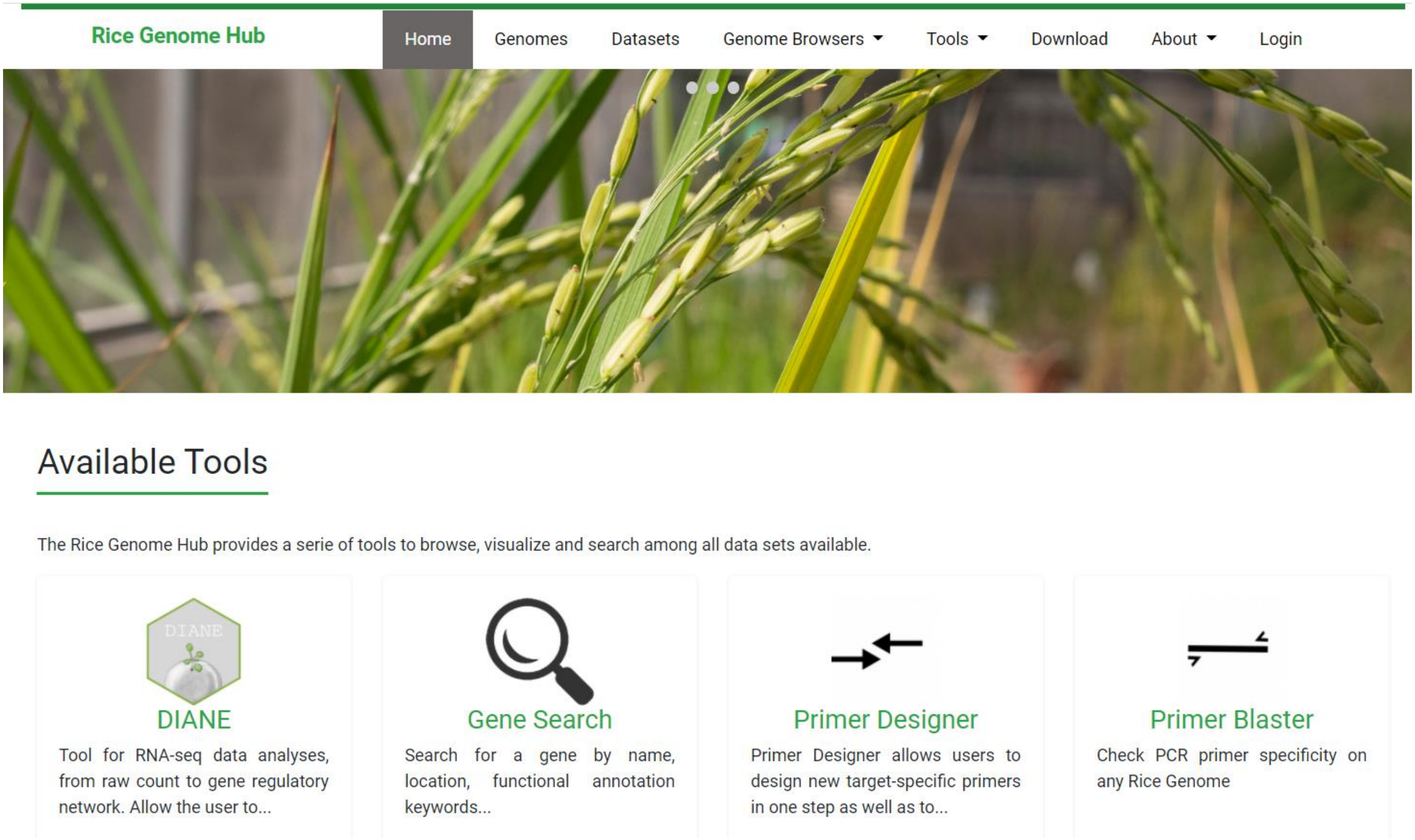
▾ Alignments 8

☑ BLASTX Arabi/Chlamy/Rice
☐ BLATX Basal-Embryophytes
☑ BLATX Basal-Monocots
☐ BLATX Dicots
☐ BLATX Grasses
☐ PASA Aligned EST/cDNA
☑ PASA Assembled EST
☐ RepeatMasker

▾ Transcripts 1

☑ Transcript

Rice Genome Hub



Available Tools

The Rice Genome Hub provides a serie of tools to browse, visualize and search among all data sets available.

DIANE

Tool for RNA-seq data analyses, from raw count to gene regulatory network. Allow the user to...

Gene Search

Search for a gene by name, location, functional annotation keywords...

Primer Designer

Primer Designer allows users to design new target-specific primers in one step as well as to...

Primer Blaster

Check PCR primer specificity on any Rice Genome

Sequencing project

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental

- Question scientifique => quelle stratégie ? Quel échantillonnage ?

  Quelle stratégie bioinfo ?

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental

- Question scientifique => quelle stratégie ? Quel échantillonnage ?
  Quelle stratégie bioinfo ?
- Quel méthodo de séquençage ? Quelle couverture de séquençage ?

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 ?

OVERVIEW OF DNA SEQUENCING PROJECT

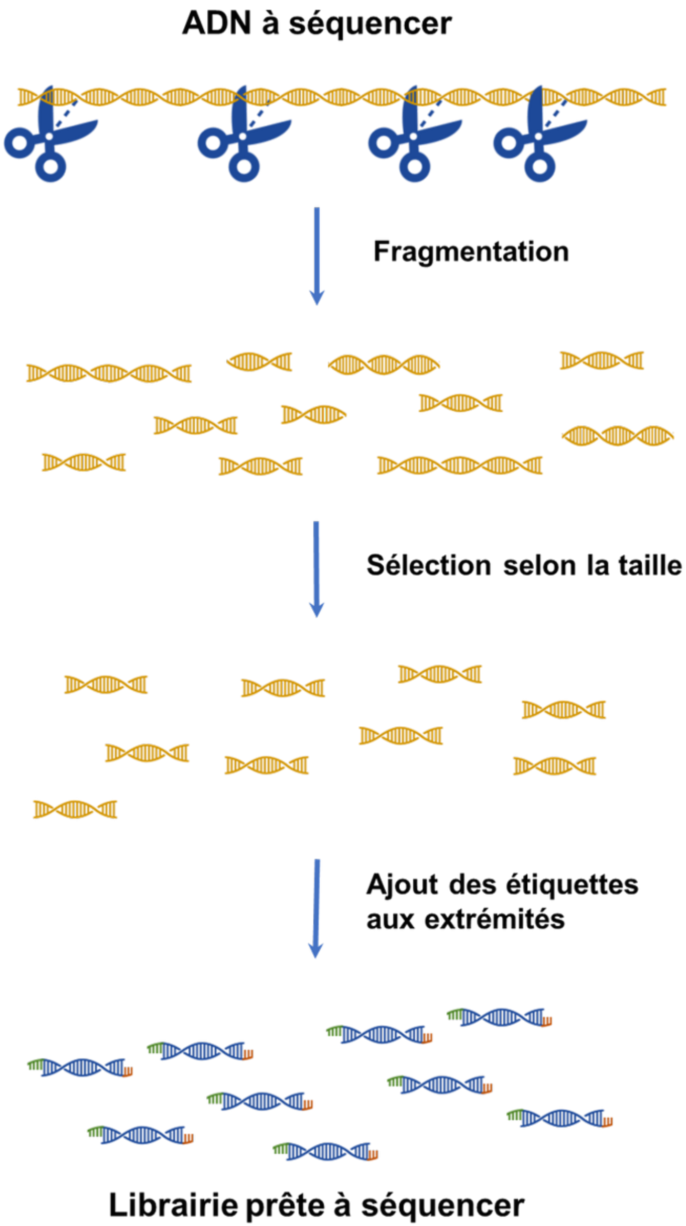
Design expérimental

- 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


ADN à séquencer

Fragmentation

Sélection selon la taille

Ajout des étiquettes aux extrémités

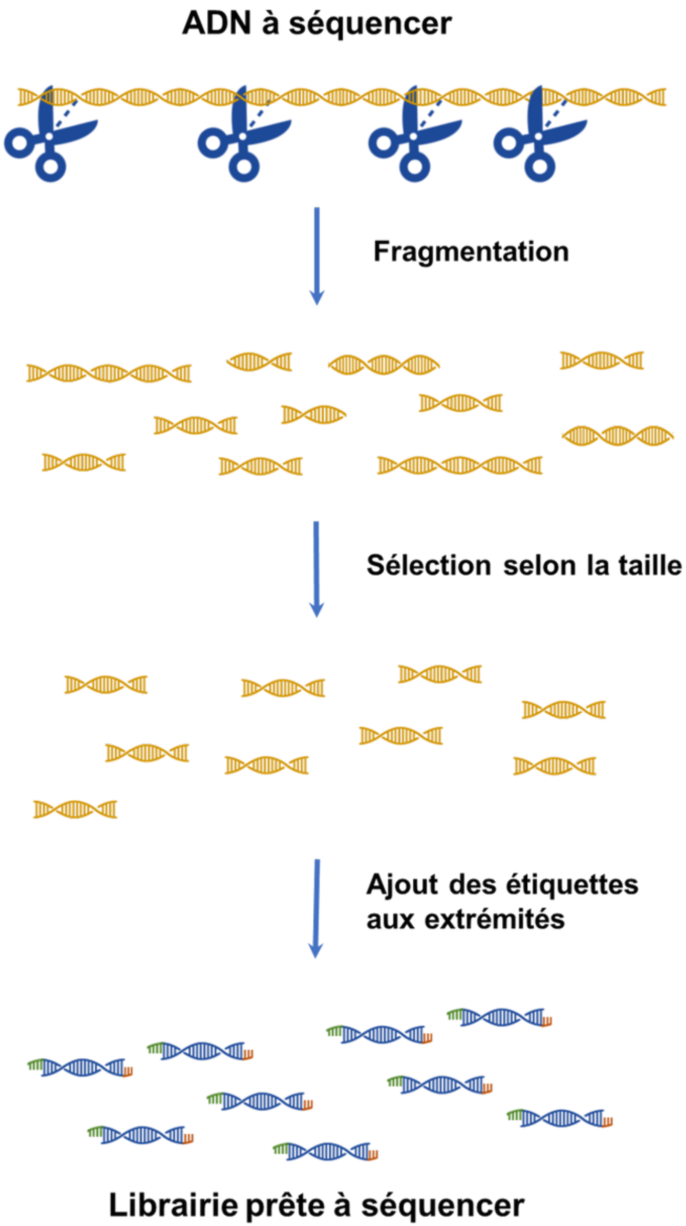
Librairie prête à séquencer

OVERVIEW OF DNA SEQUENCING PROJECT

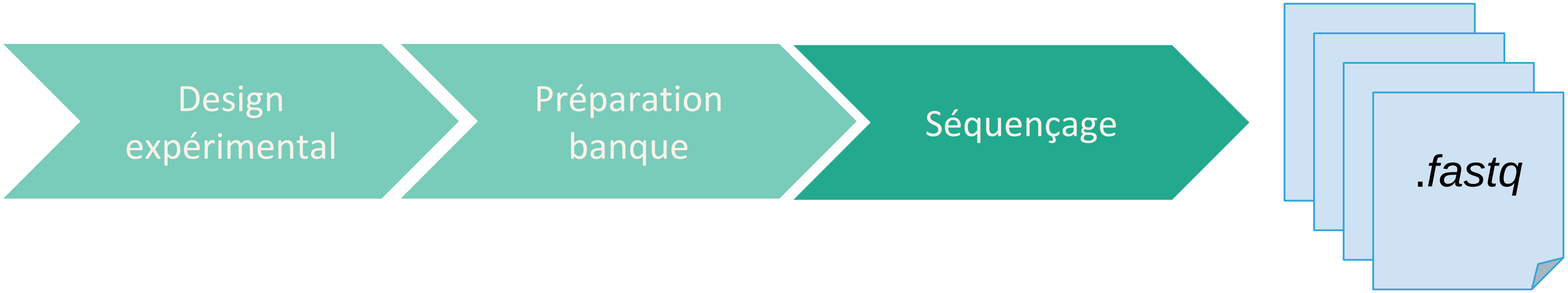
Design expérimental

Préparation banque

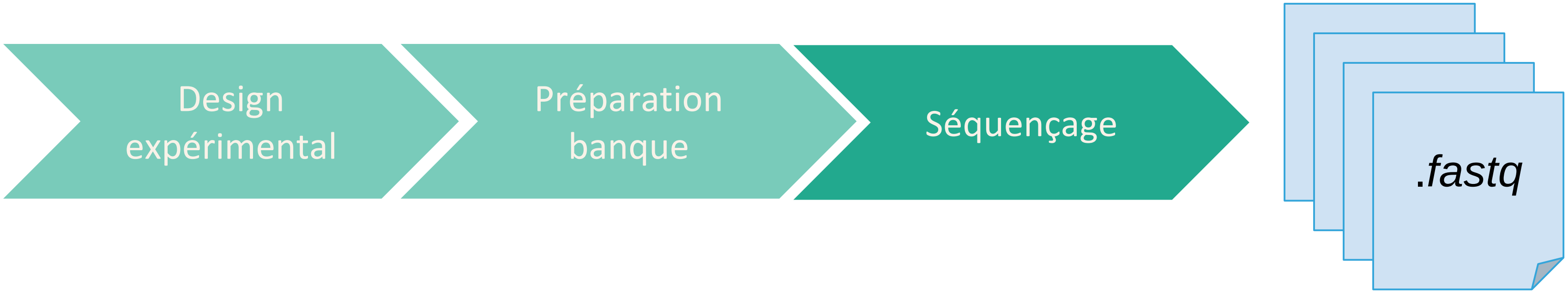
- Adaptateurs
- Contamination


ADN à séquencer
Fragmentation
Sélection selon la taille
Ajout des étiquettes aux extrémités
Librairie prête à séquencer

OVERVIEW OF DNA SEQUENCING PROJECT

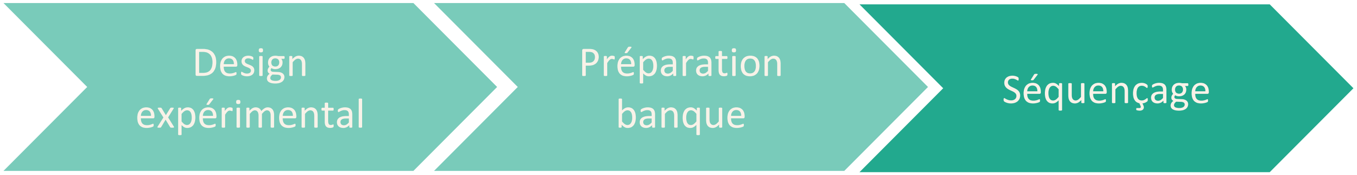


OVERVIEW OF DNA SEQUENCING PROJECT



- Qualité de séquençage
- Profondeur de séquençage

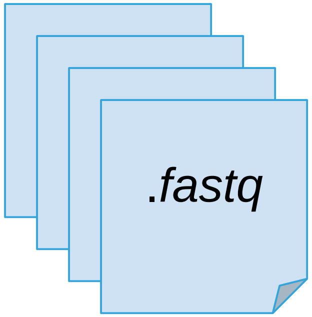
OVERVIEW OF DNA SEQUENCING PROJECT



Design expérimental

Préparation banque

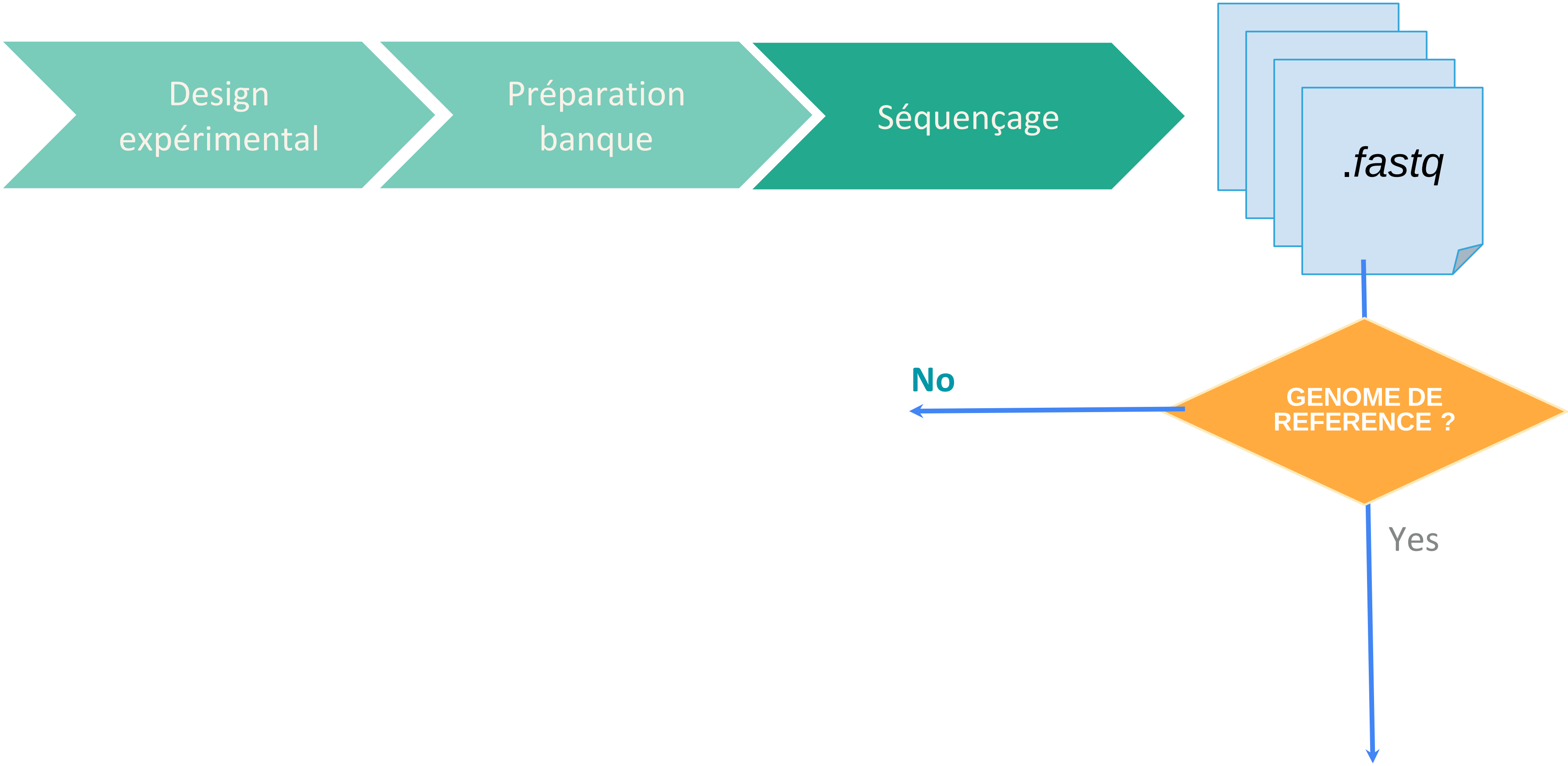
Séquençage



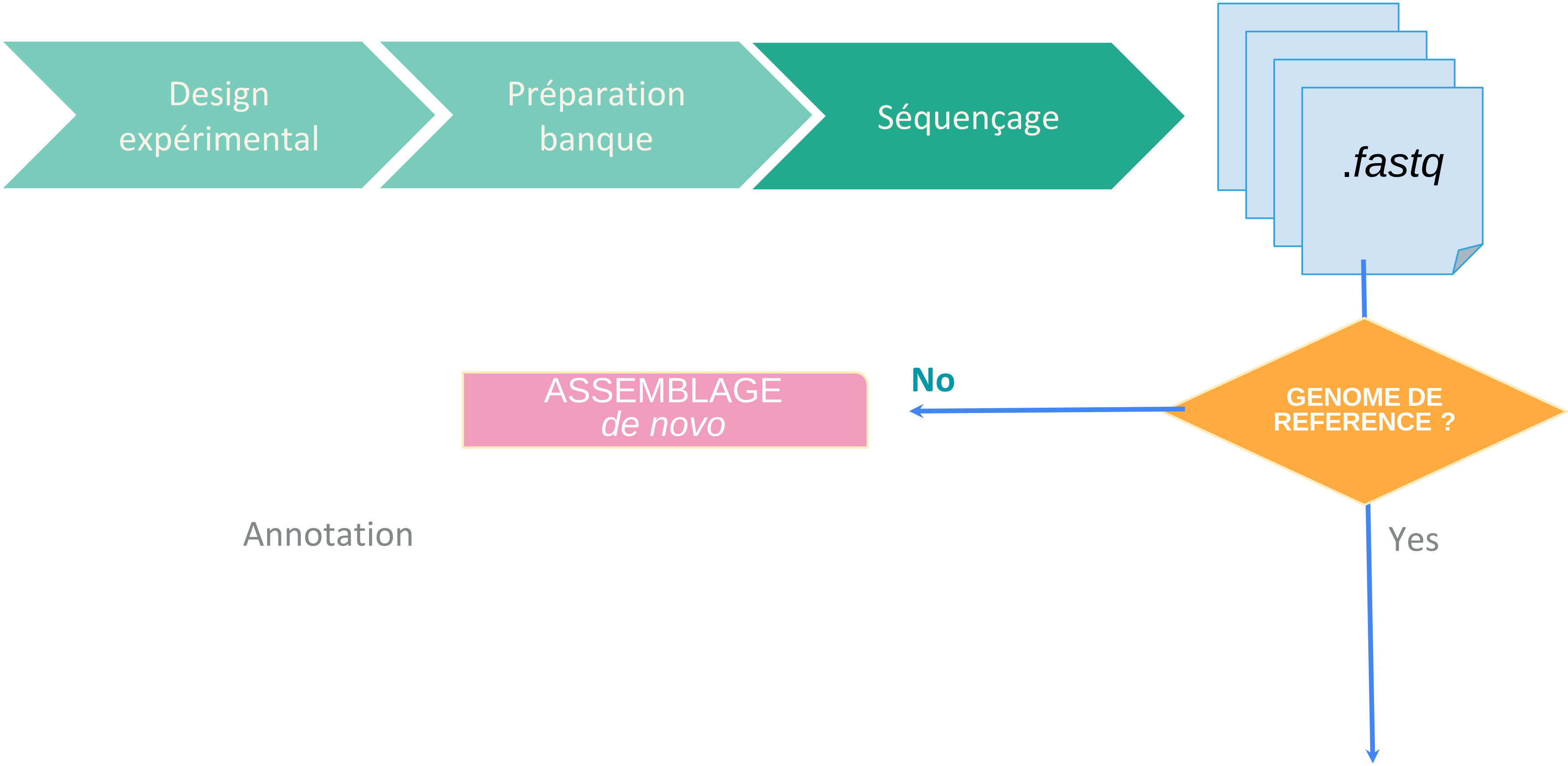
.fastq

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



OVERVIEW OF DNA SEQUENCING PROJECT



Design expérimental

Préparation banque

Séquençage

.fastq

GENOME DE REFERENCE ?

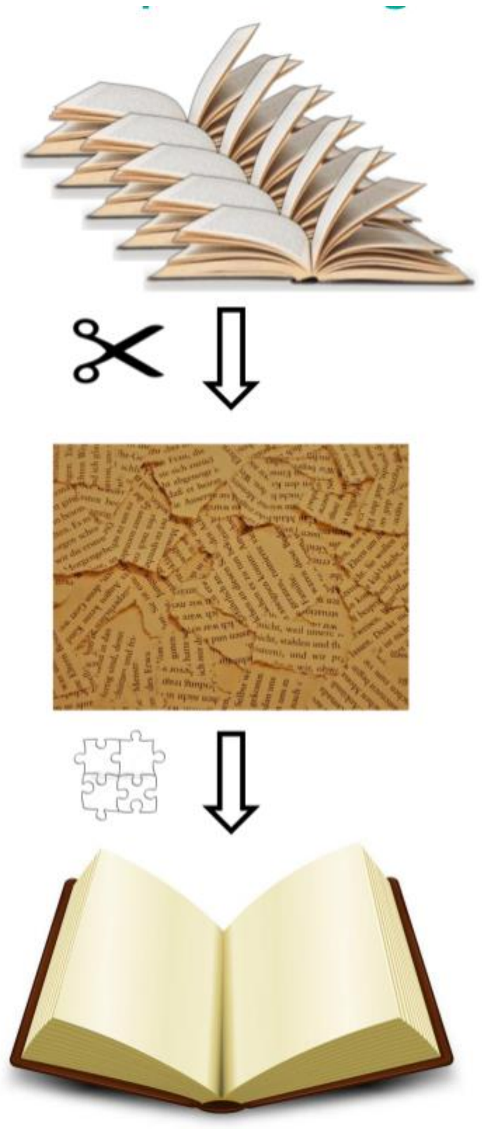
No

ASSEMBLAGE *de novo*

Yes

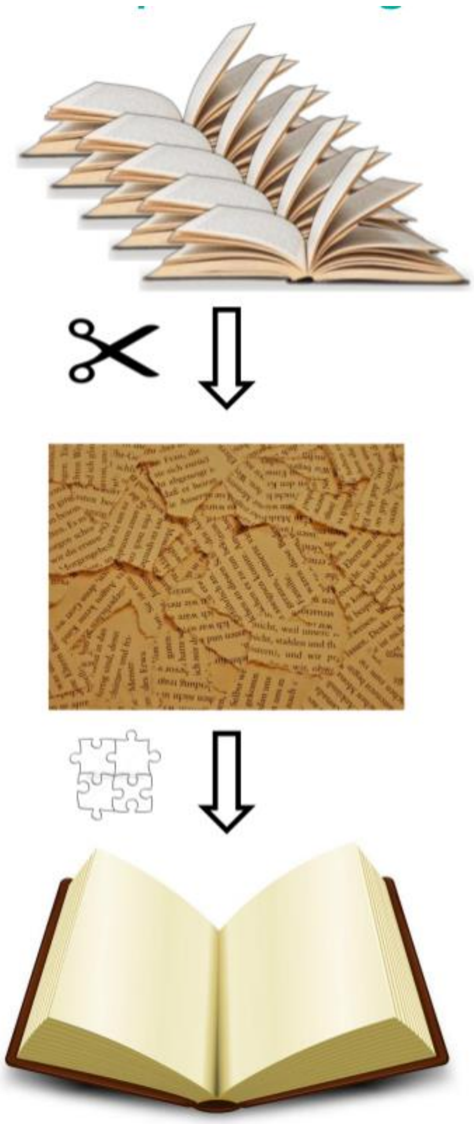
Annotation

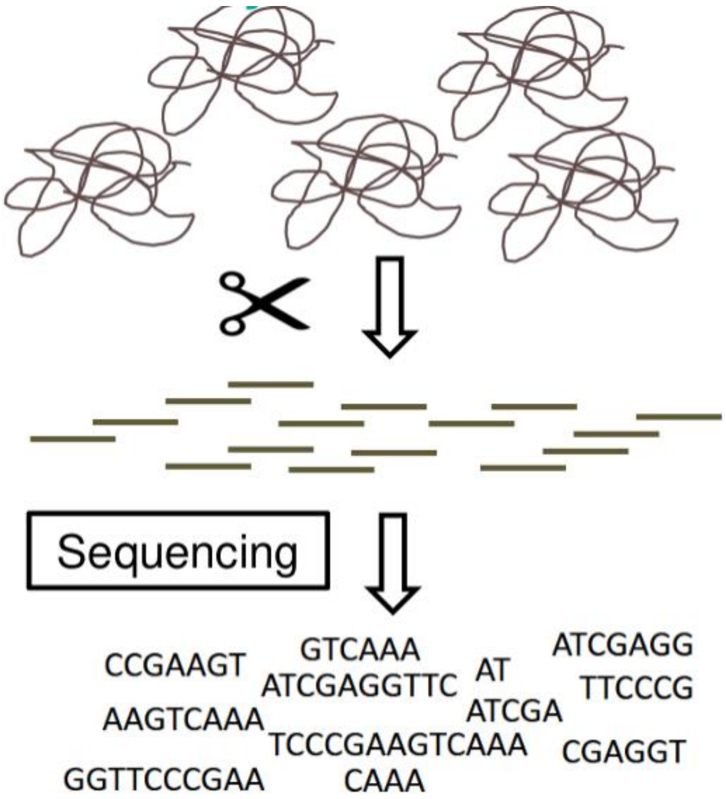
Sequencing and genome assembly



From Camille Rustenholz (Univ. Strasbourg, inrae) - Methods for plant genome assembly

Sequencing and genome assembly



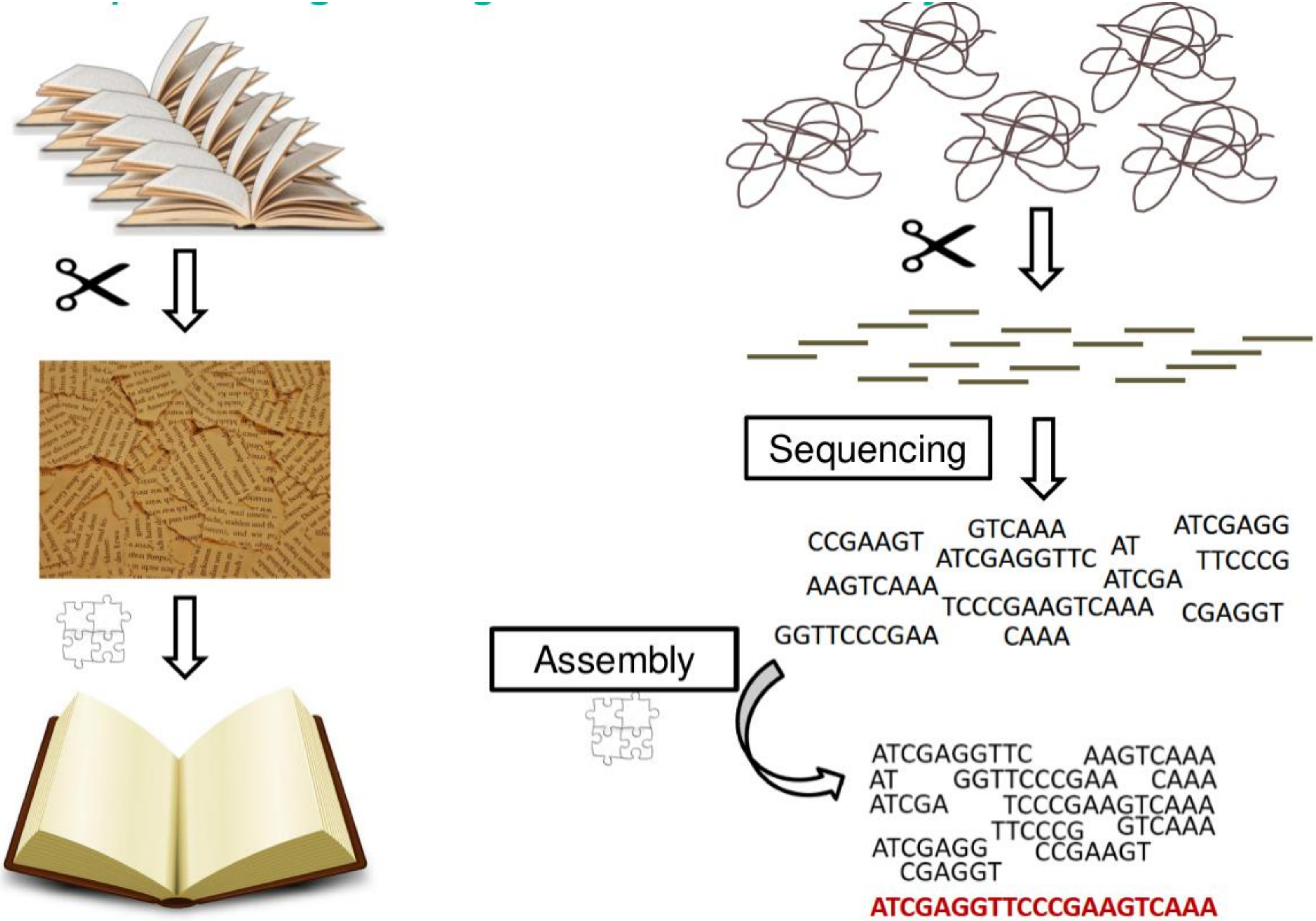


Sequencing

CCGAAGT
GTCAAA
ATCGAGGTTC
AT
ATCGAGG
TTCCCG
ATCGA
AAGTCAAA
TCCCGAAGTCAAA
CGAGGT
GGTTCCCGAA
CAAA

From Camille Rustenholz (Univ. Strasbourg, inrae) - Methods for plant genome assembly

Sequencing and genome assembly



Sequencing

CCGAAGT GTCAAA ATCGAGGTTC AT ATCGAGG TTCCCG
AAGTCAAA ATCGA TCCCGAAGTCAAA CGAGGT
GGTTCCCGAA CAAA

Assembly

ATCGAGGTTC AAGTCAAA
AT GGTTCCCGAA CAAA
ATCGA TCCCGAAGTCAAA
TTCCCG GTCAAA
ATCGAGG CCGAAGT
CGAGGT

ATCGAGGTTCCCGAAGTCAAA

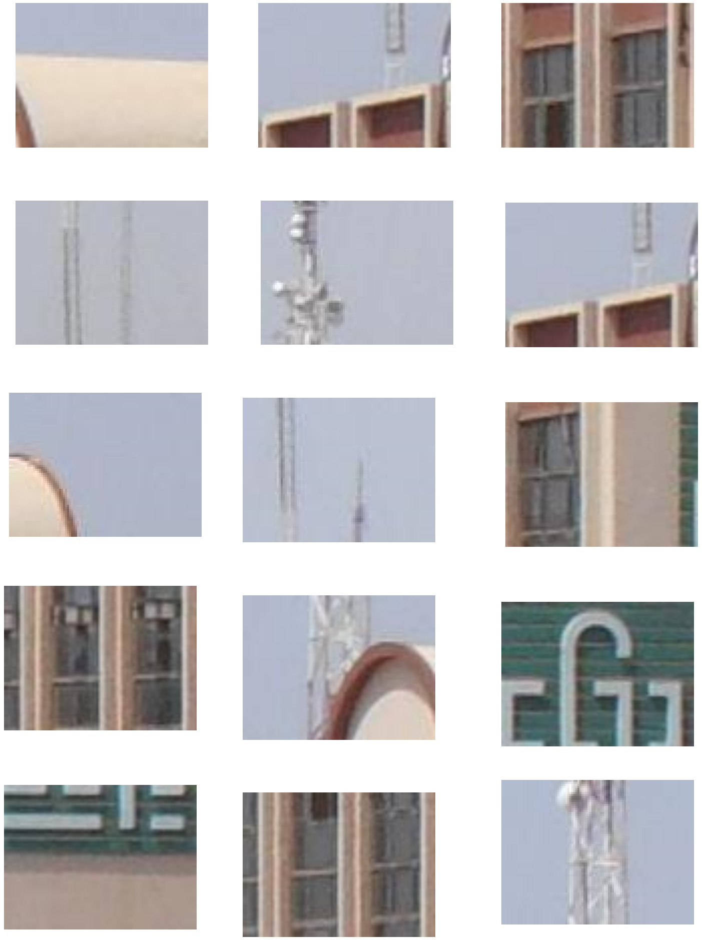
From Camille Rustenholz (Univ. Strasbourg, inrae) - Methods for plant genome assembly

Sequencing and genome assembly

Puzzle 400 pièces “petite taille”

Sequencing and genome assembly

Puzzle 400 pièces “petite taille”



+ 100 pièces “ciel” + …

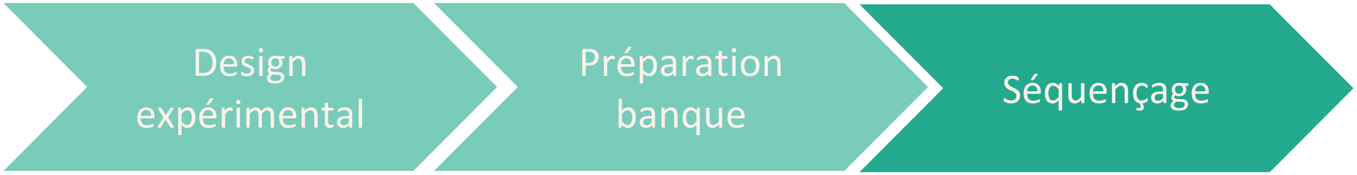
Puzzle 100 pièces - “grande taille”

Sequencing and genome assembly

Puzzle 100 pièces “grande taille”

+ ~ 20 pièces “ciel”

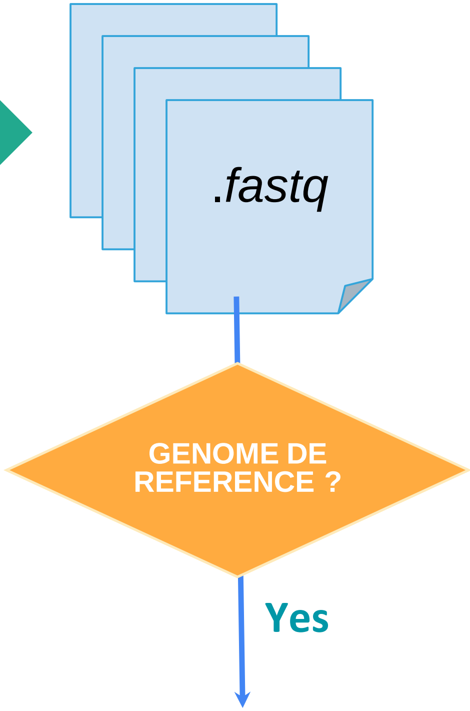
OVERVIEW OF DNA SEQUENCING PROJECT



Design expérimental

Préparation banque

Séquençage



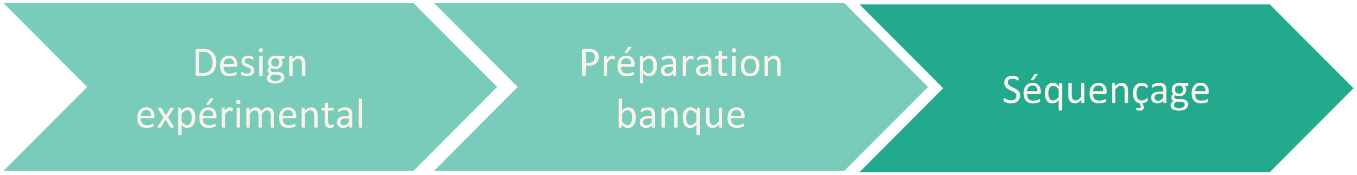
.fastq

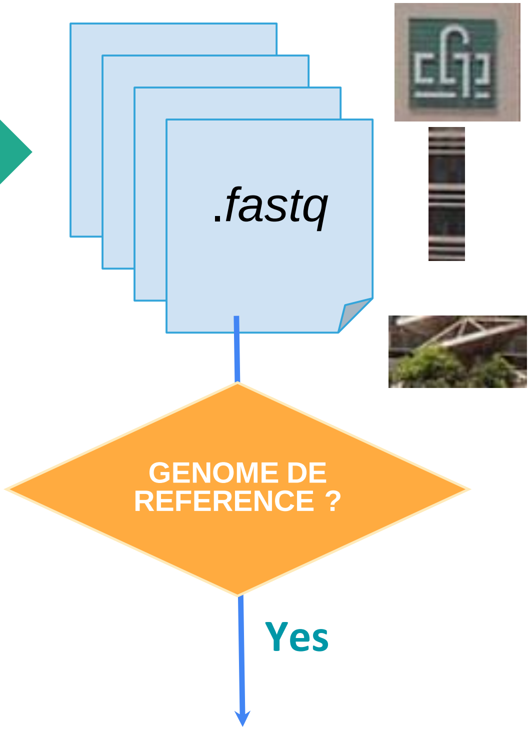
GENOME DE REFERENCE ?

Yes

Adapted from Ross Whetten...

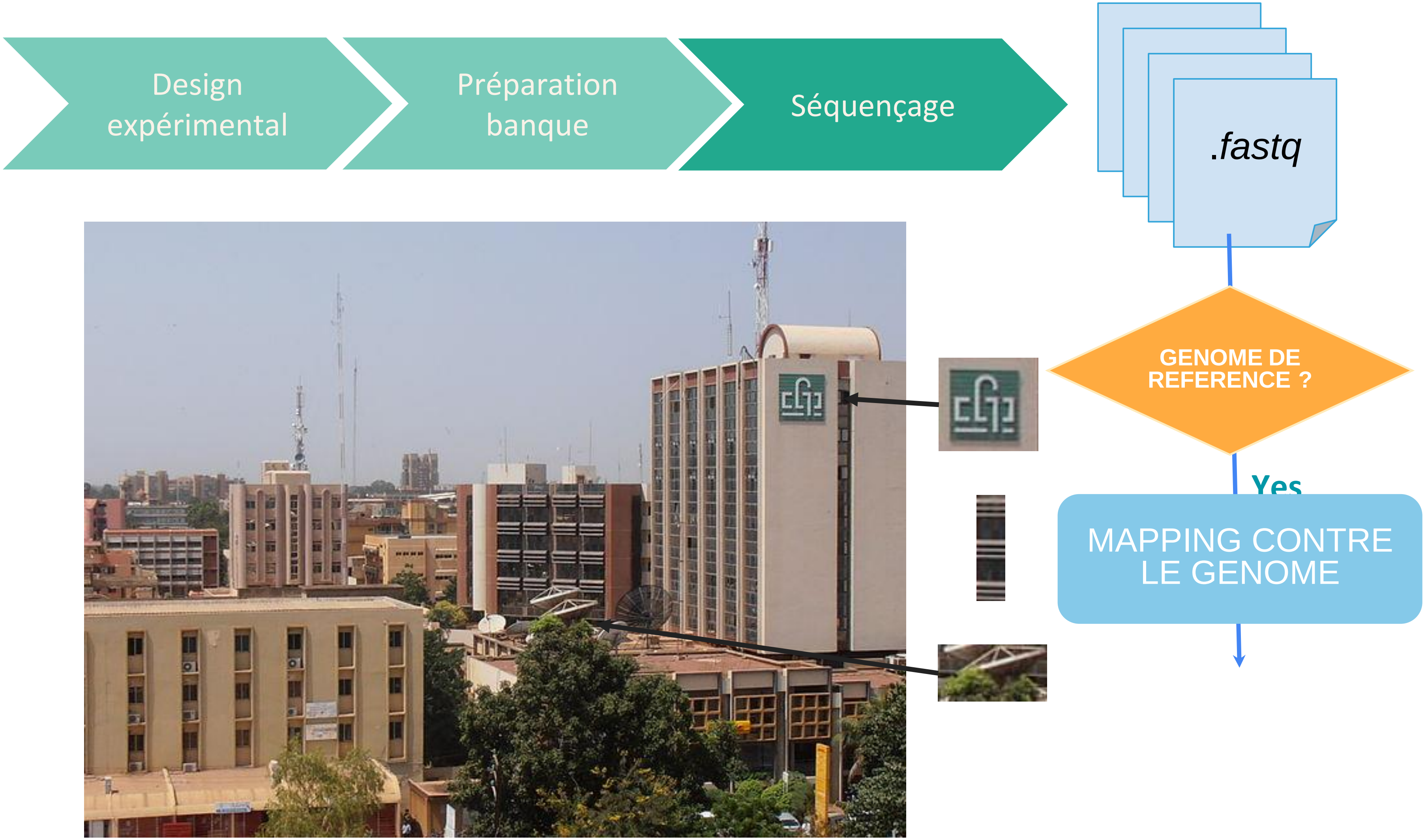
OVERVIEW OF DNA SEQUENCING PROJECT





Adapted from Ross Whetten…

OVERVIEW OF DNA SEQUENCING PROJECT



Design expérimental

Préparation banque

Séquençage

.fastq

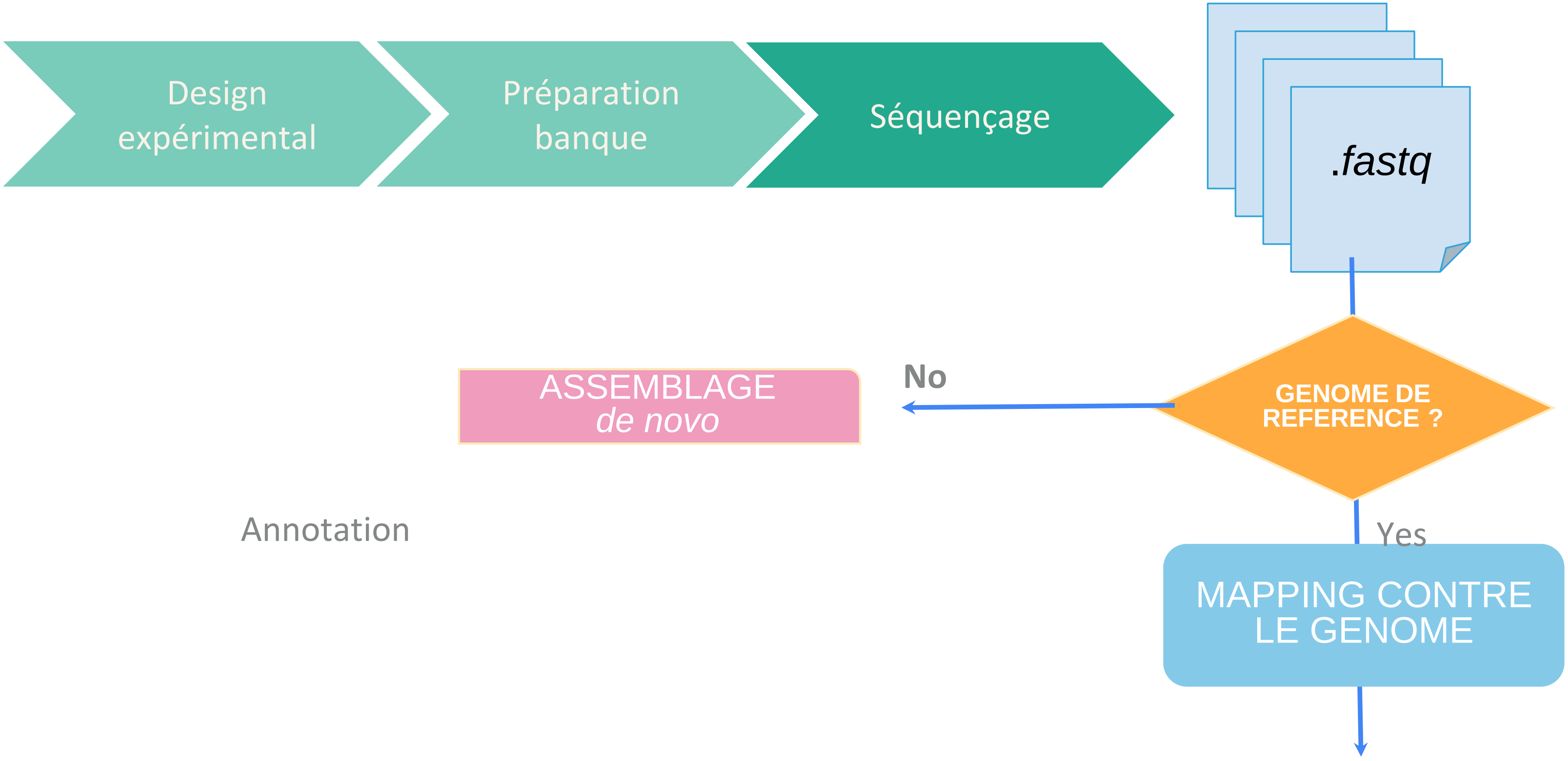
GENOME DE REFERENCE ?

Yes

MAPPING CONTRE LE GENOME

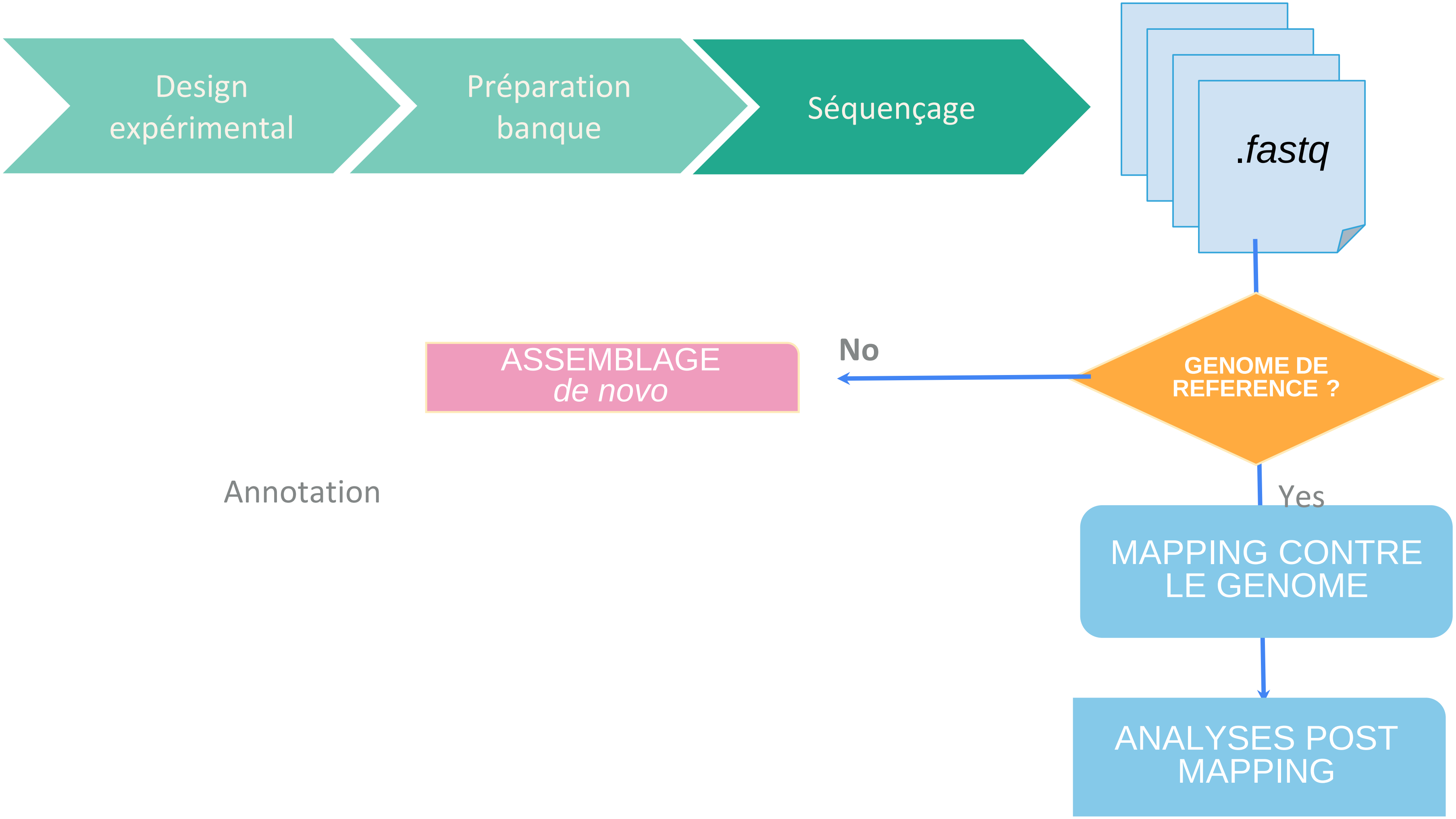
Adapted from Ross Whetten…

OVERVIEW OF DNA SEQUENCING PROJECT



Adapted from Ross Whetten…

OVERVIEW OF DNA SEQUENCING PROJECT



Annotation

Adapted from Ross Whetten…

SNP, GWAS? expression différentielle

What metagenomics is ?

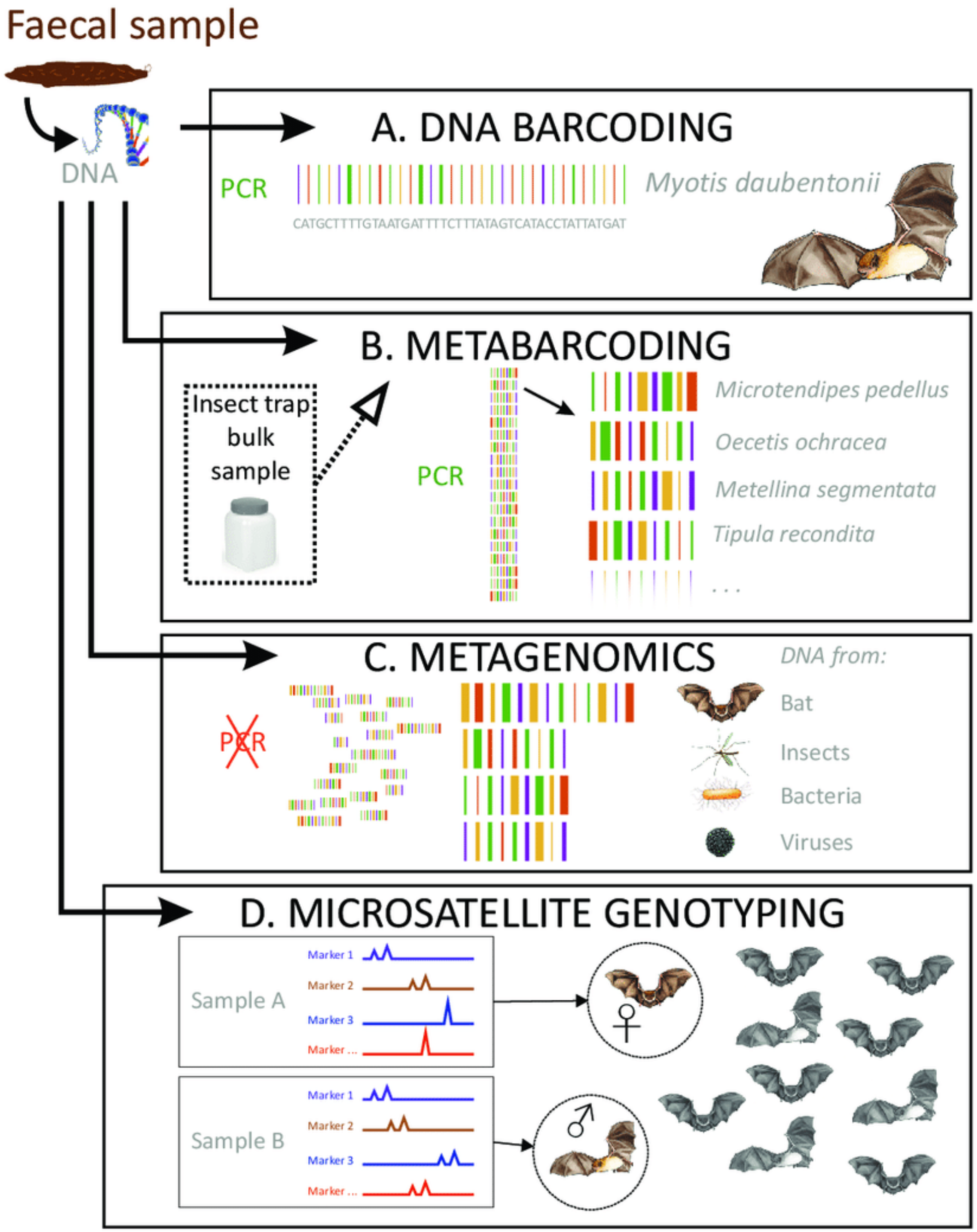
Metagenomics (Environmental Genomics or Community Genomics) is the study of genomes recovered from environmental samples without the need for culturing them

Metagenomics processes data using bioinformatics tools

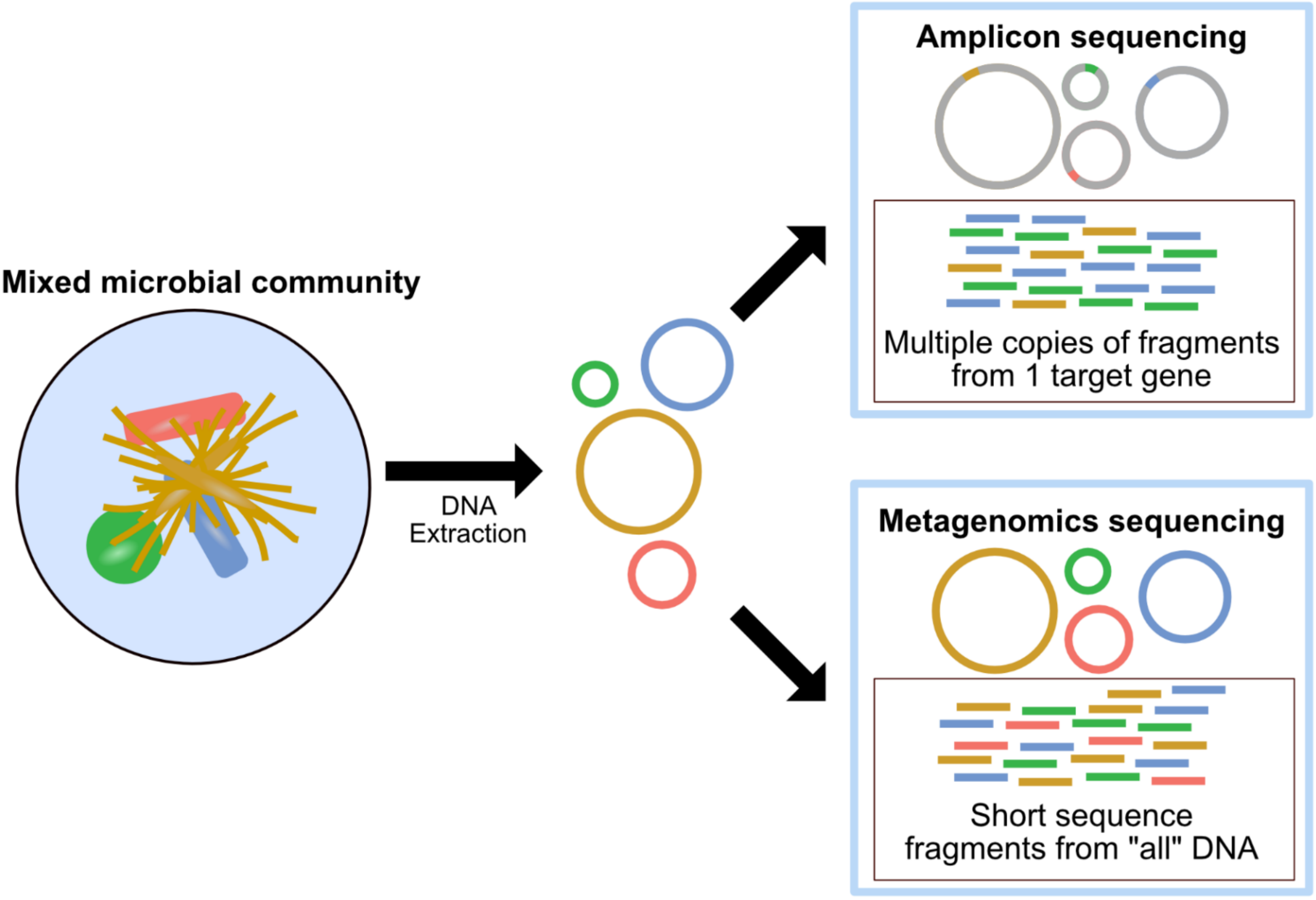
=> Organisms can be studied directly in their environments bypassing the need to isolate each species

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

Application: Metagenomics



Application: Metagenomics



Mixed microbial community

DNA Extraction

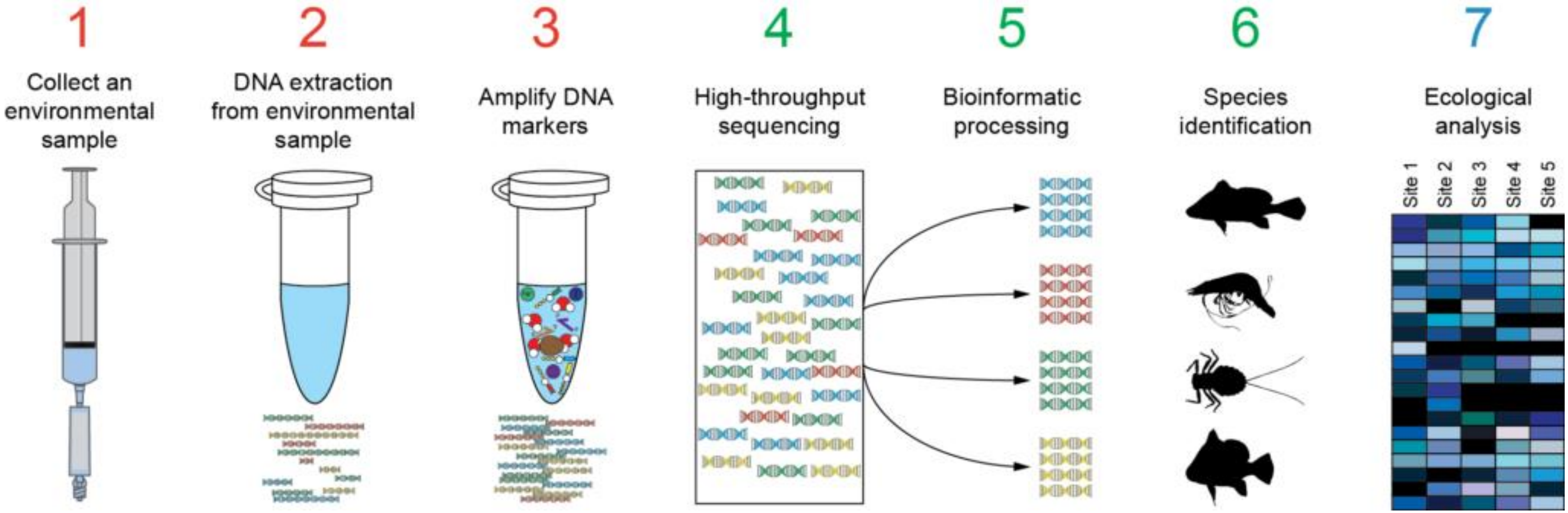
Amplicon sequencing

Multiple copies of fragments from 1 target gene

Metagenomics sequencing

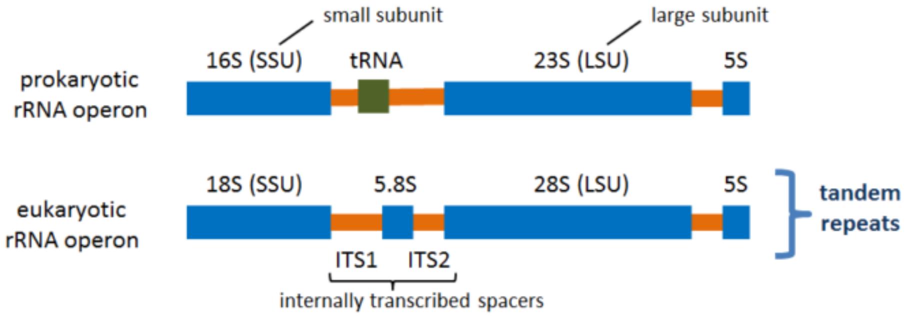
Short sequence fragments from "all" DNA

Application: Metabarcoding



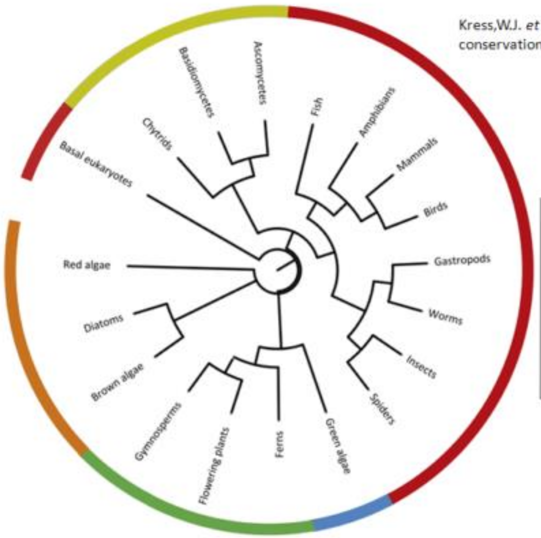
http://www.sixthresearcher.com/amplicon-sequencing-and-high-throughput-genotyping-metagenomics/

Application: Metabarcoding



Type	LSU	SSU
prokaryotic	5S - 120 bp 23S - 2906 bp	16S - 1542 bp
eukaryotic	5S - 121 bp 5.8S - 156 bp 28S - 5070 bp	18S - 1869 bp

Which barcode to choose?



Kress,W.J. *et al.* (2014) DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.*, **30**, 25–35.

Tree of life

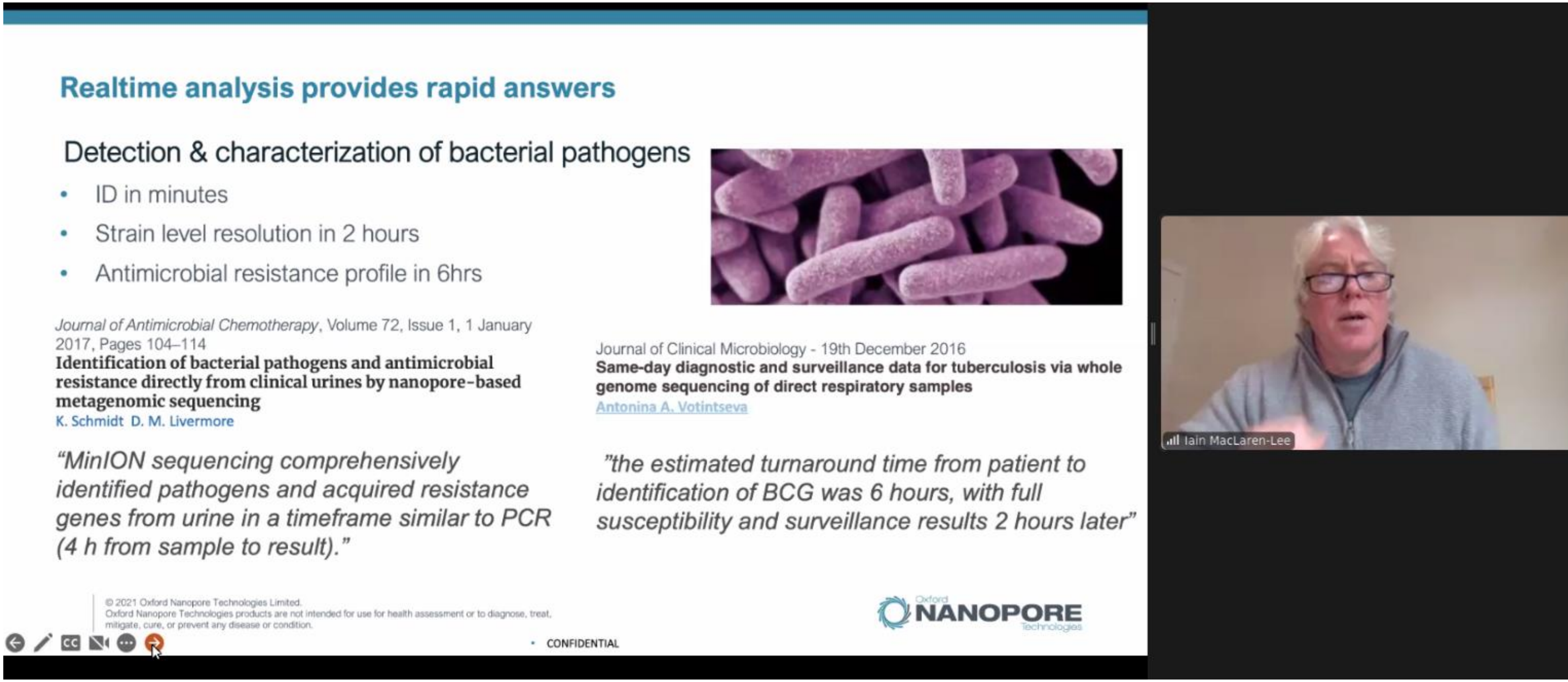
Key:	Color	Clade	Primary barcode(s)	Secondary barcode(s)
	🟥	Animals	CO1	CO1, 16S
	🟨	Fungi	ITS	LSU D1/D2
	🟦	Green algae	*tufA*	LSU D2/D3
	🟩	Land plants	*rbcL/matK*	*psbA-trnH*/ITS
	🟧	Algae	CO1-5P	LSU D2/D3
		Bacteria/ Archae	16S	RIF

CO1: cytochrome c oxidase subunit 1
ITS: internally transcribed spacer
LSU: large subunit rRNA
D1/D2/D3: divergent domains
RIF: DnaA replication initiation factor

http://www.barcodeoflife.org/

http://www.sixthresearcher.com/amplicon-sequencing-and-high-throughput-genotyping-metagenomics/

Application: Metagenomic



Realtime analysis provides rapid answers

Detection & characterization of bacterial pathogens

- ID in minutes
- Strain level resolution in 2 hours
- Antimicrobial resistance profile in 6hrs

Journal of Antimicrobial Chemotherapy, Volume 72, Issue 1, 1 January 2017, Pages 104–114
Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing
K. Schmidt D. M. Livermore

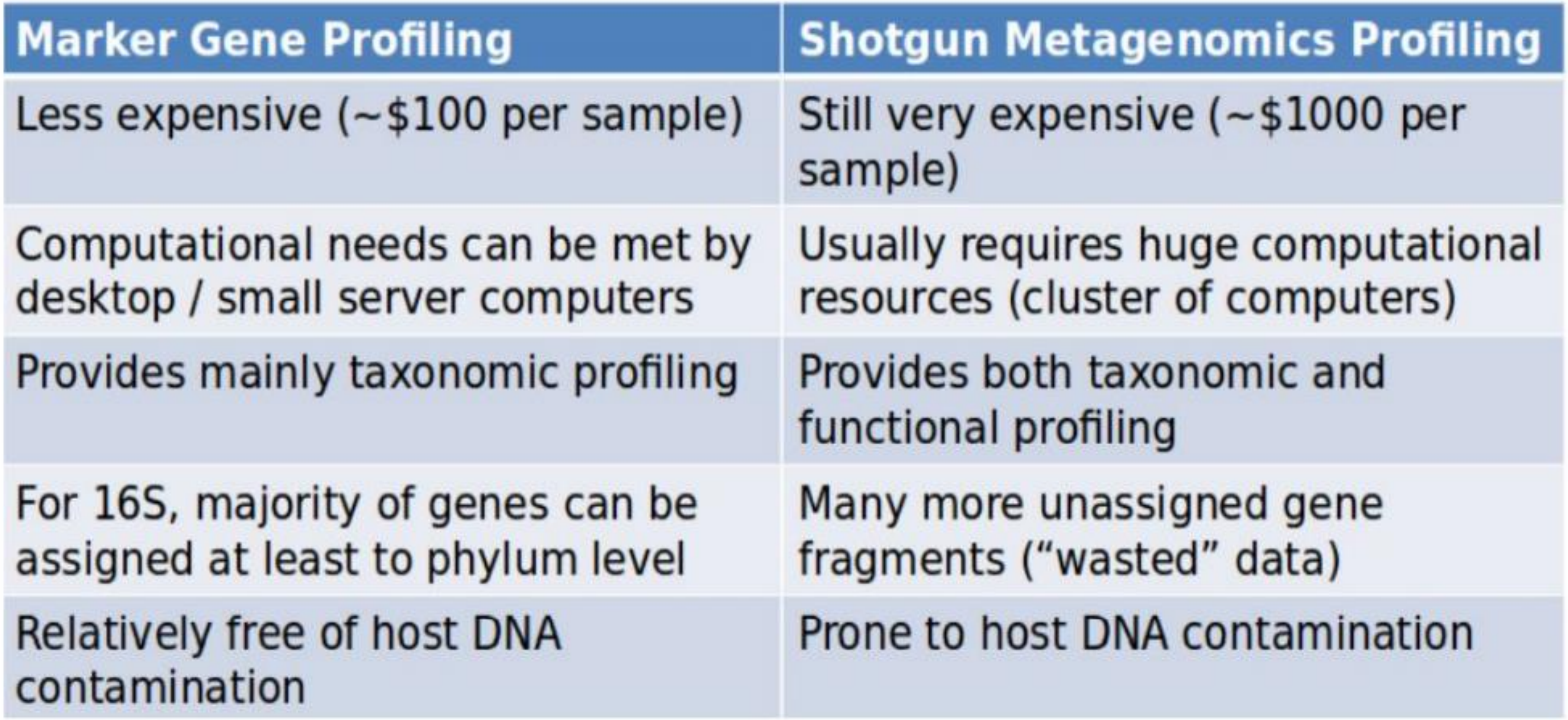
"MinION sequencing comprehensively identified pathogens and acquired resistance genes from urine in a timeframe similar to PCR (4 h from sample to result)."

Journal of Clinical Microbiology - 19th December 2016
Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples
Antonina A. Votintseva

"the estimated turnaround time from patient to identification of BCG was 6 hours, with full susceptibility and surveillance results 2 hours later"

CONFIDENTIAL

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

Application: Transcriptomics / RNASeq

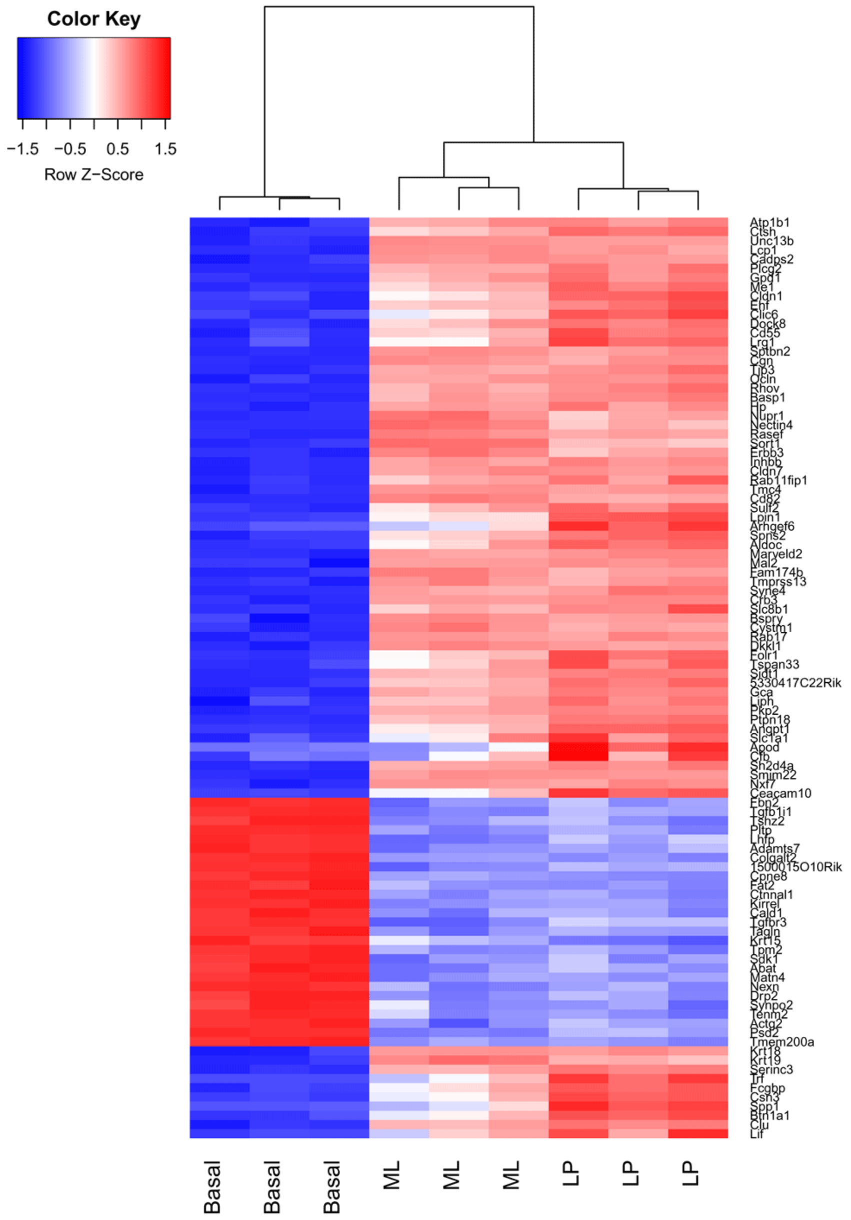
Pourquoi faire du RNAseq ?

- **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.

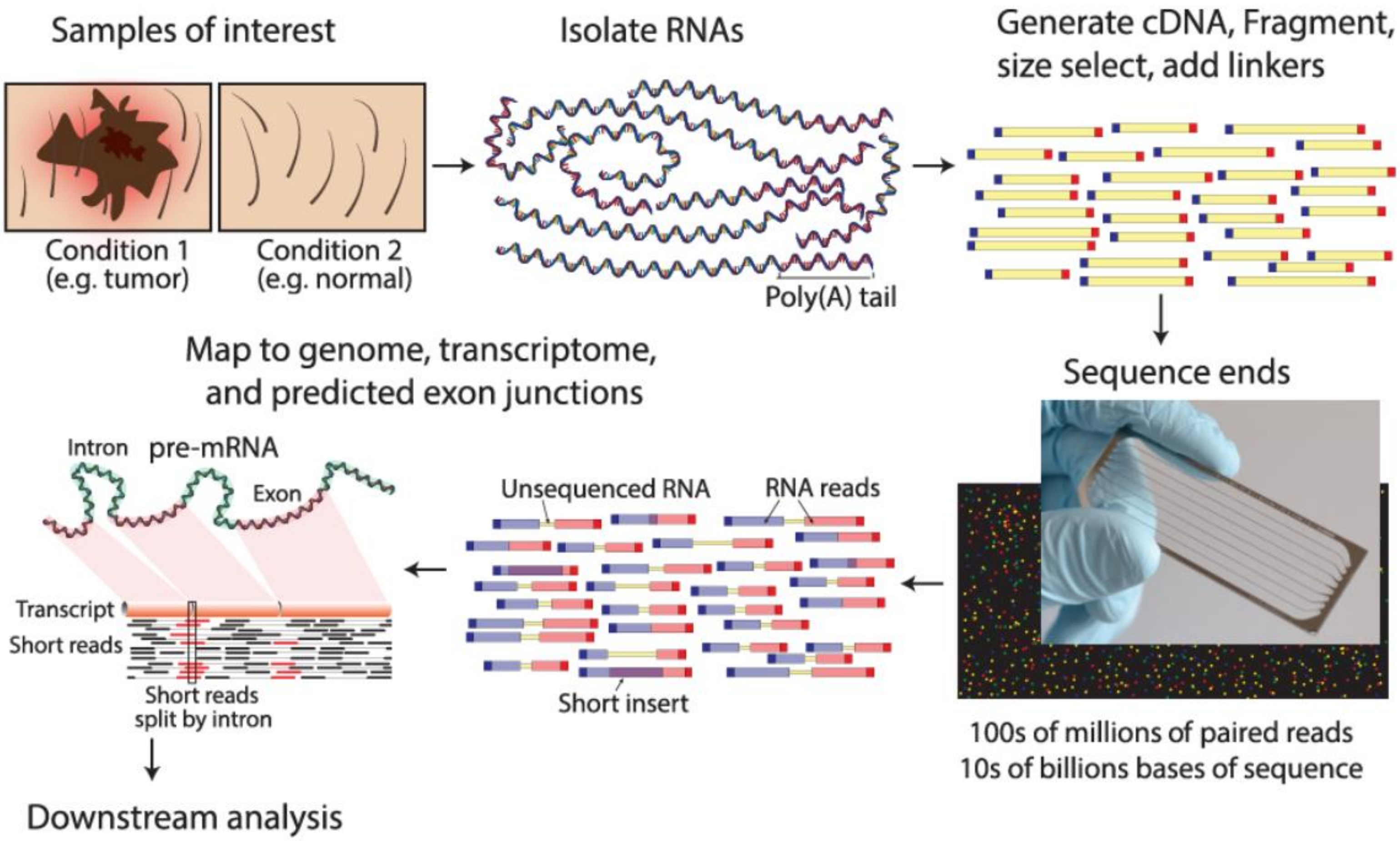
- **Recherche d'allèles spécifiques** et quantification de leur expression.

- **Construction d'un transcriptome** de novo pour les organismes non modèles.

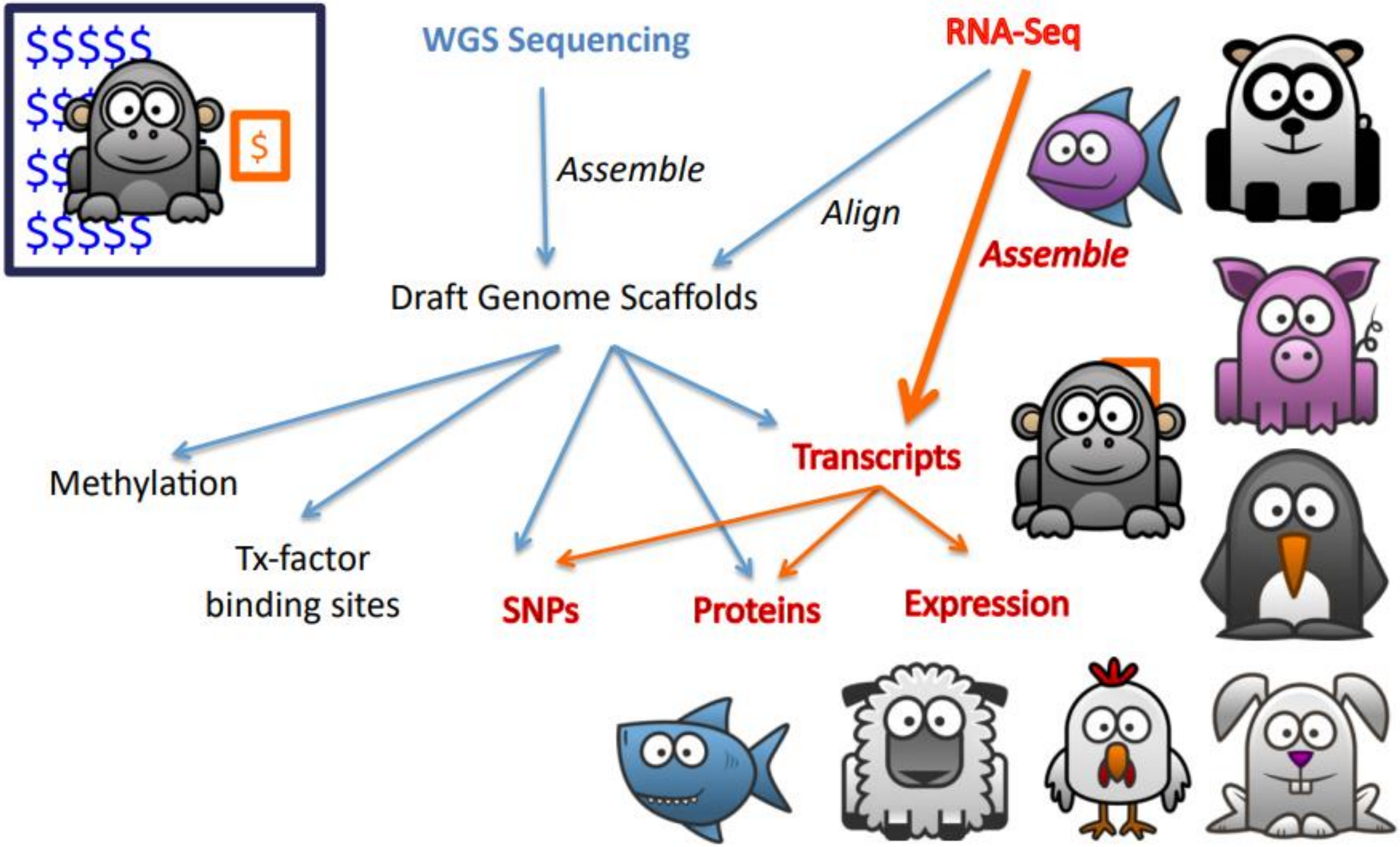


Application: Transcriptomics / RNASeq

RNA sequencing

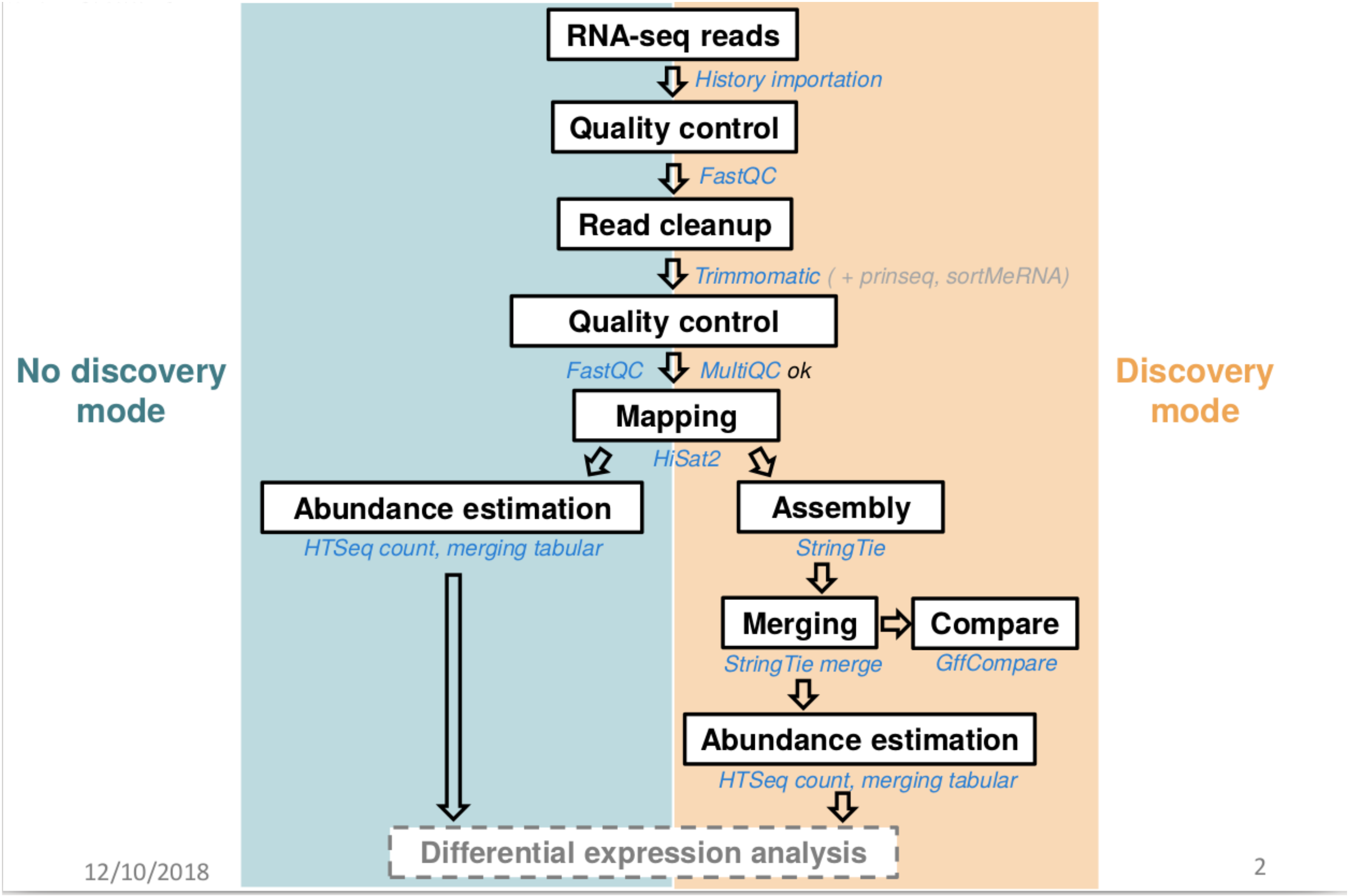


Application: Transcriptomics / RNASeq



From SG-ABiMS RNAseq Training Sept 2019

Application: Transcriptomics / RNASeq



No discovery mode

Discovery mode

RNA-seq reads

History importation

Quality control

FastQC

Read cleanup

Trimmomatic (+ prinseq, sortMeRNA)

Quality control

FastQC MultiQC ok

Mapping

HiSat2

Abundance estimation

HTSeq count, merging tabular

Assembly

StringTie

Merging

StringTie merge

Compare

GffCompare

Abundance estimation

HTSeq count, merging tabular

Differential expression analysis

12/10/2018

2

From Abims RNAseq Training oct 2018

Application: Transcriptomics / RNASeq

⇒ Comparaison entre conditions expérimentales différentes

Ex:

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

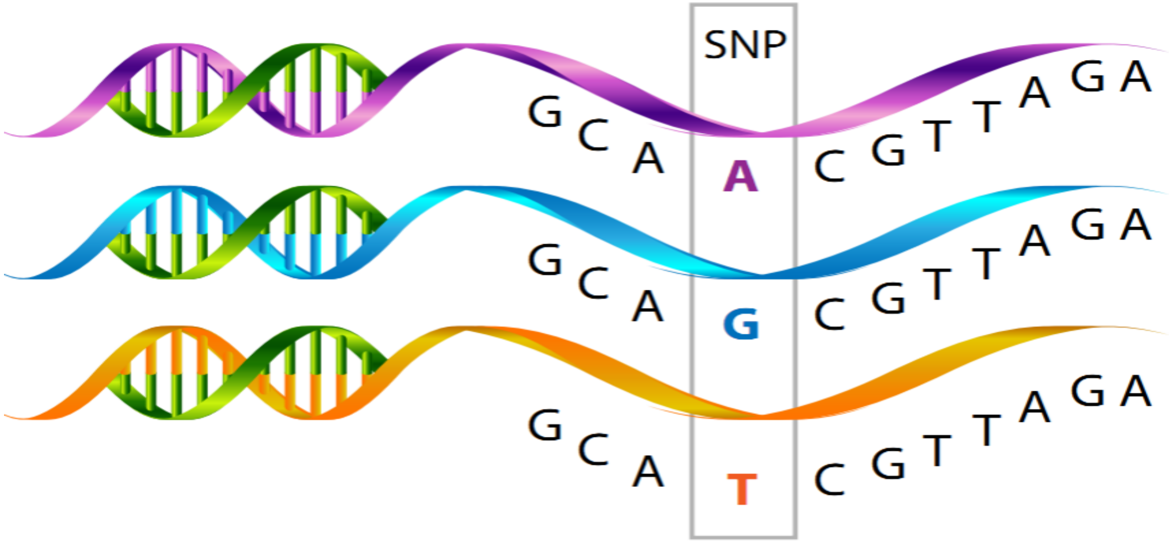
⇒ Comparaison dans le temps (time series): cinétique

Ex:

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

=> logiciels dédiés pour ce type de problèmatique

Single Nucleotide Polymorphism



Application : Global genotyping & Population genomics

Origin of domestication and evolutionary history of African crop?

Where, when, how, (why) ?

African rice

Pearl millet

Yam

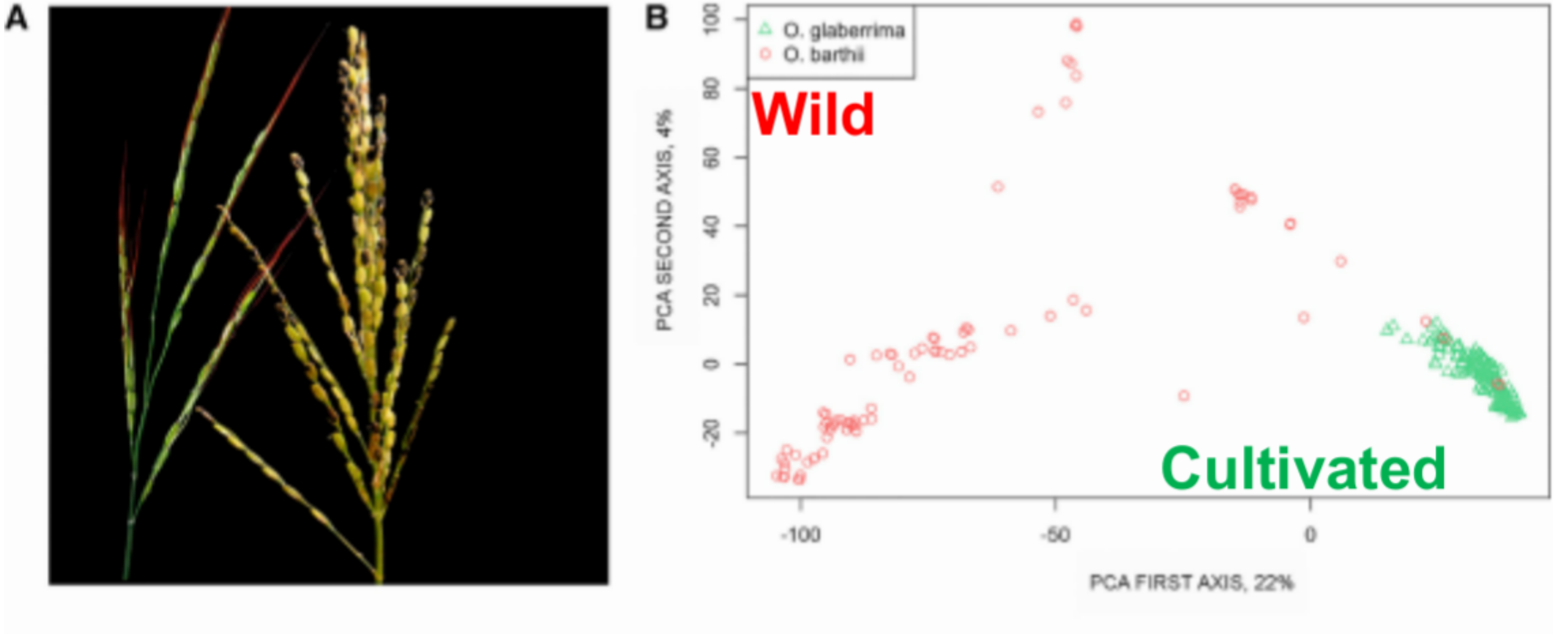
Fonio

Sorghum

From Y. Vigouroux

Application : African rice domestication

246 fully resequenced genomes
3 051 681 SNPs



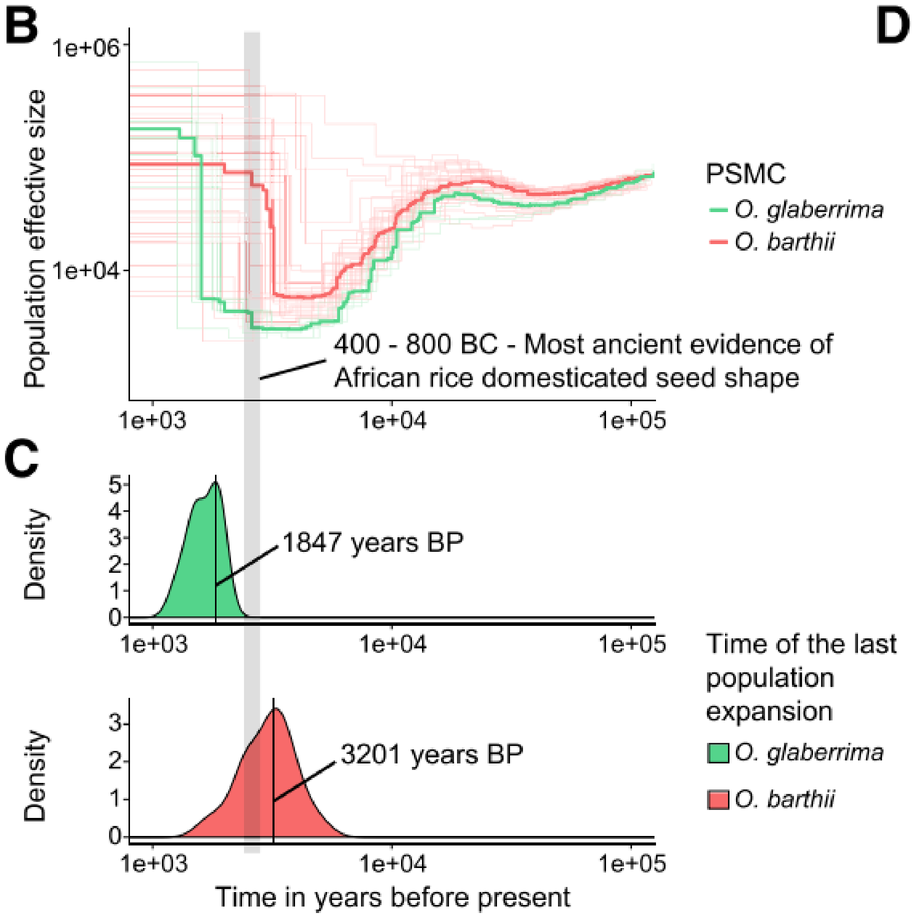
Cubry P, Tranchant-Dubreuil C, Thuillet AC, Monat C, *et al.* Current Biol 2018

From Y. Vigouroux

Application : Genomics help untangled past events

WHEN ?

Pairwise Sequentially Markovian Coalescent (PSMC)

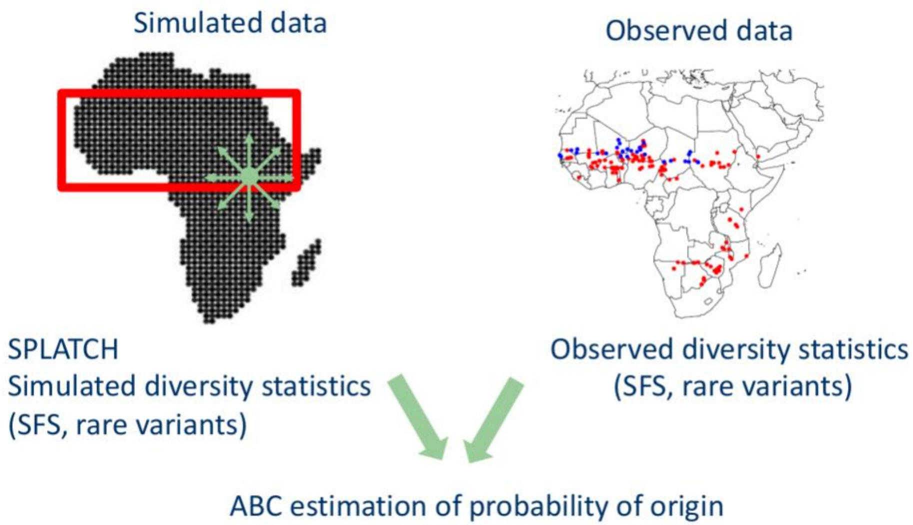


The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes

From Cubry et al, 2018

Application : Aproximate Bayesian spatial approach

WHERE ?



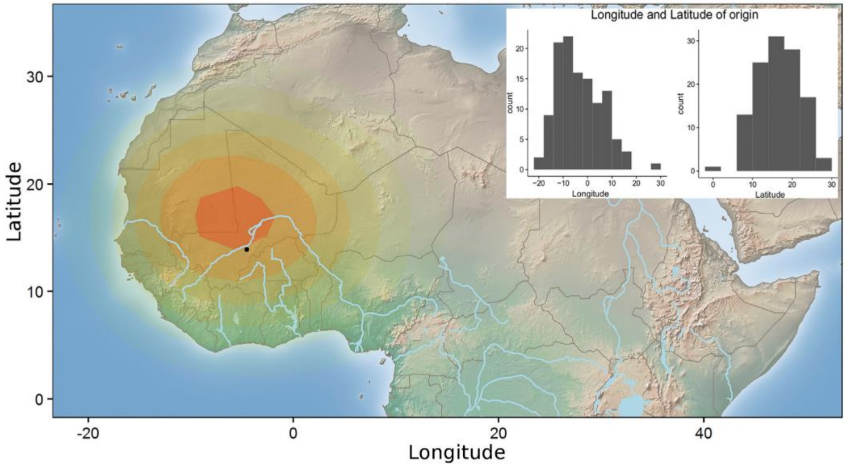
Simulated data

Observed data

SPLATCH
Simulated diversity statistics
(SFS, rare variants)

Observed diversity statistics
(SFS, rare variants)

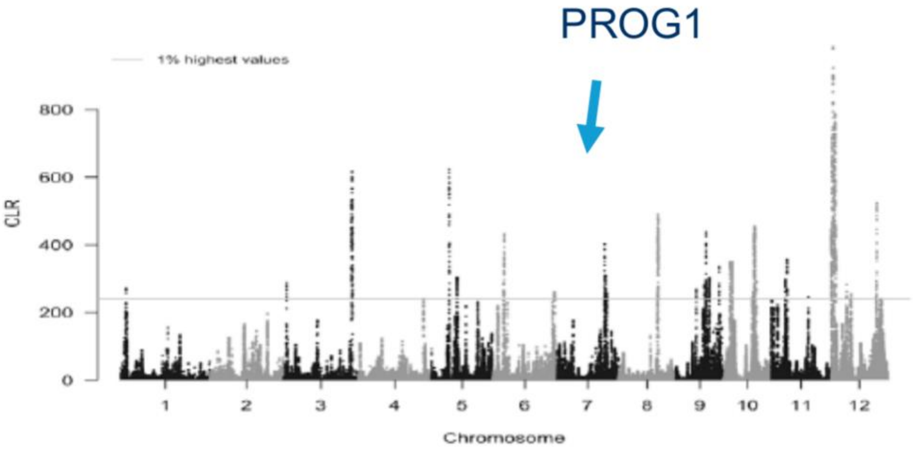
ABC estimation of probability of origin

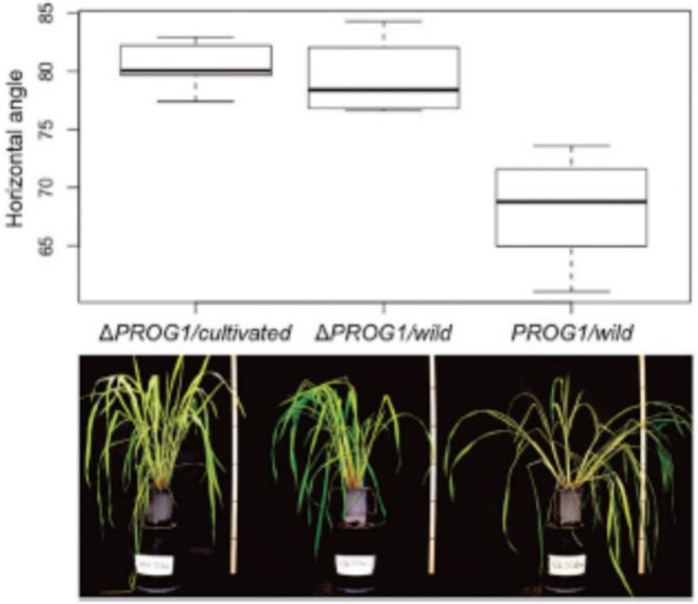


From Cubry et al, 2018

Application : Detection of selection

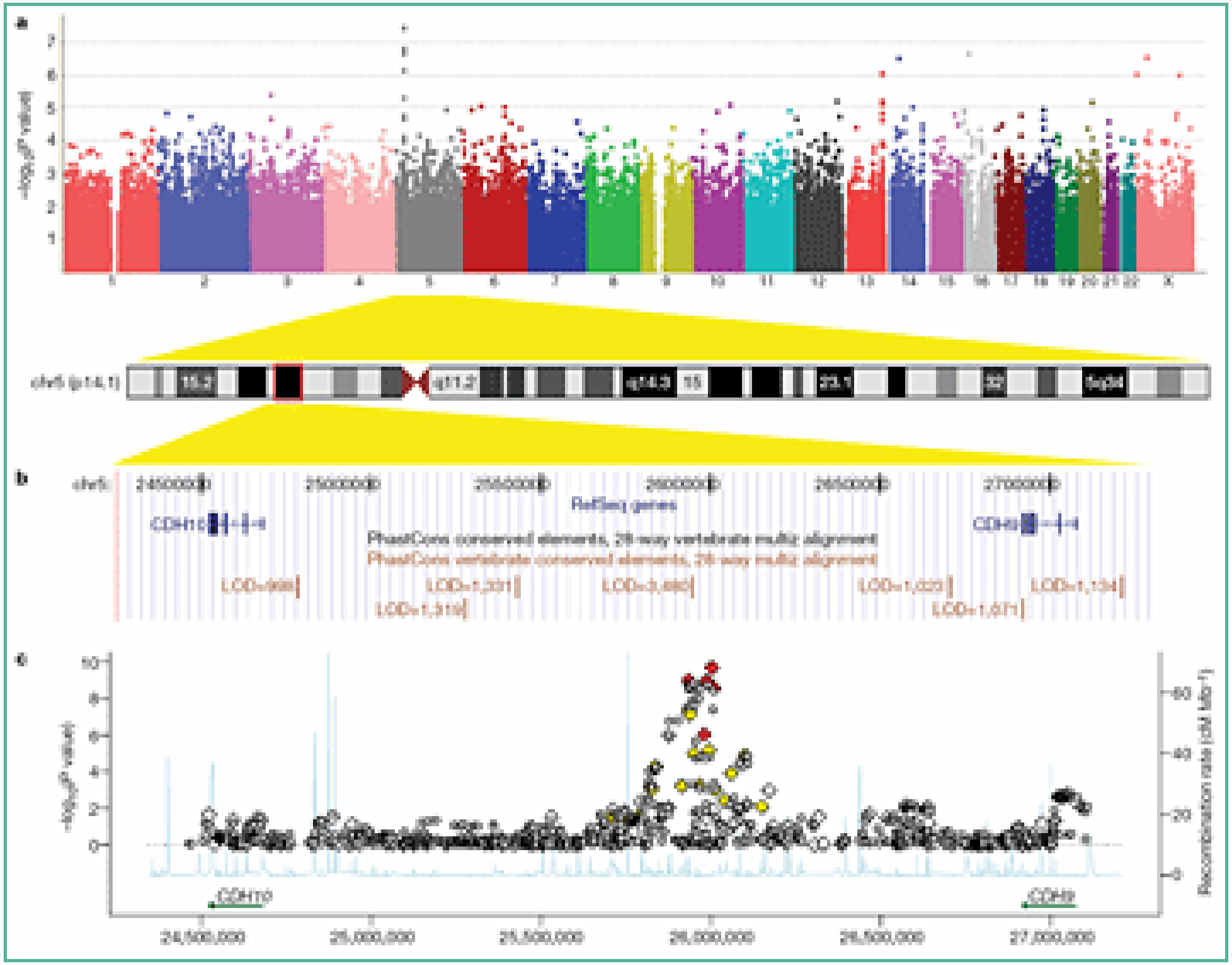
Prostrate growth 1



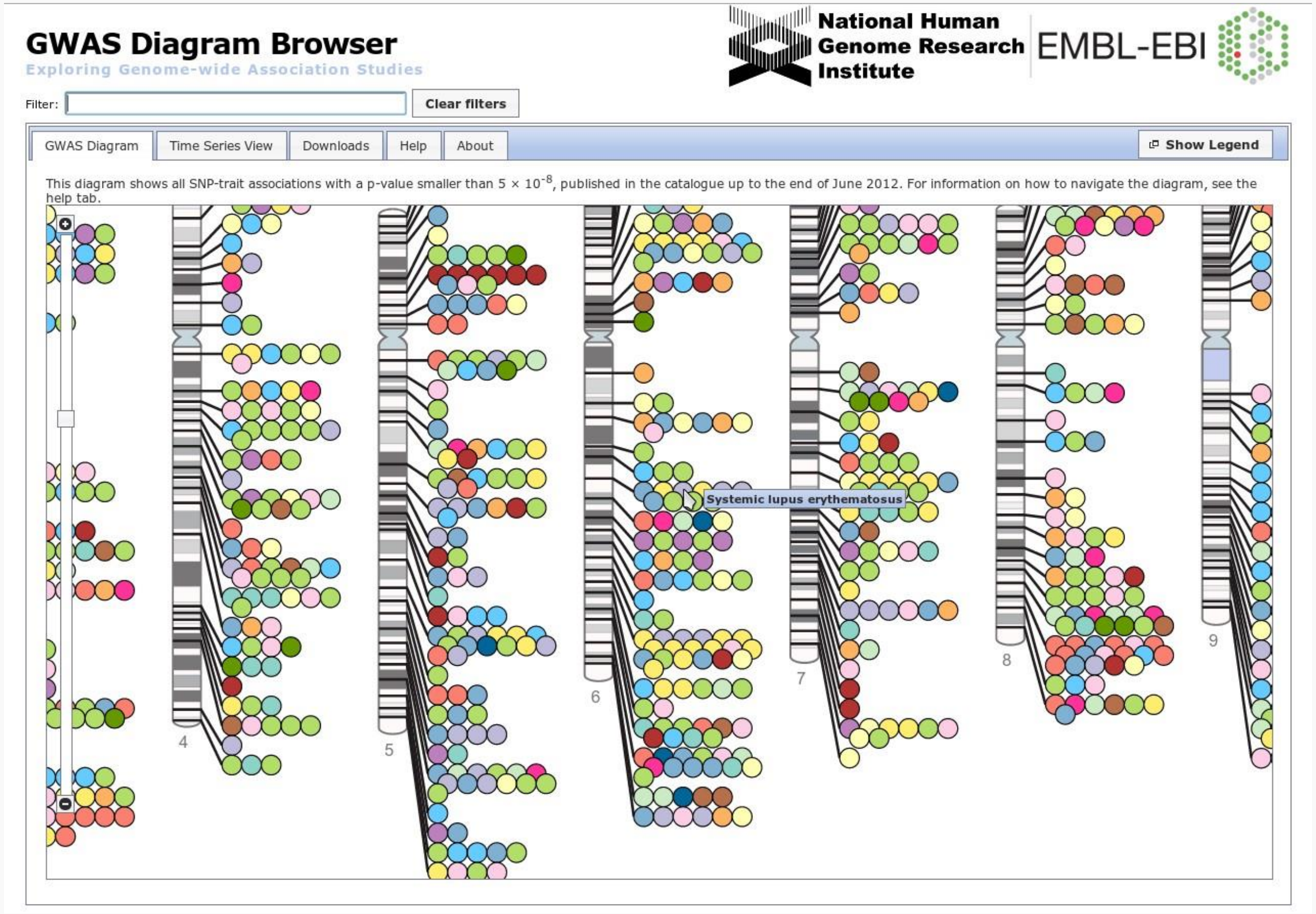


Prog1 deletion

Example in GWAs & Population Genomics



Example in GWAs & Population Genomics



GWAS Diagram Browser

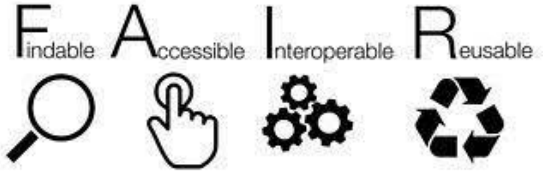
Exploring Genome-wide Association Studies

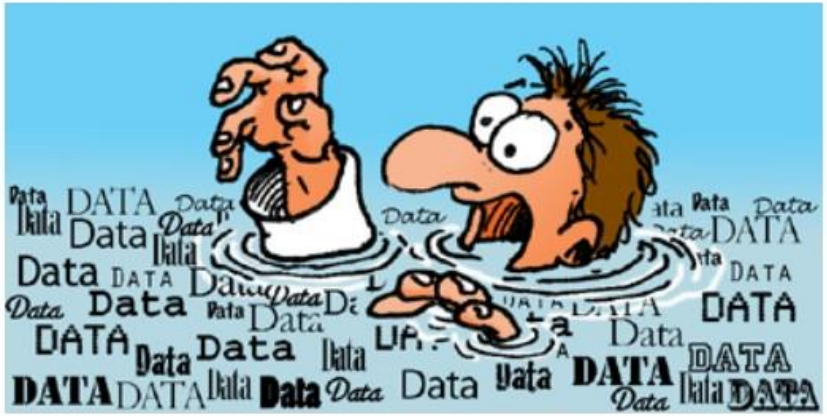
Filter:

This diagram shows all SNP-trait associations with a p-value smaller than 5 × 10^{-8}, published in the catalogue up to the end of June 2012. For information on how to navigate the diagram, see the help tab.

Be Careful to data drowning!

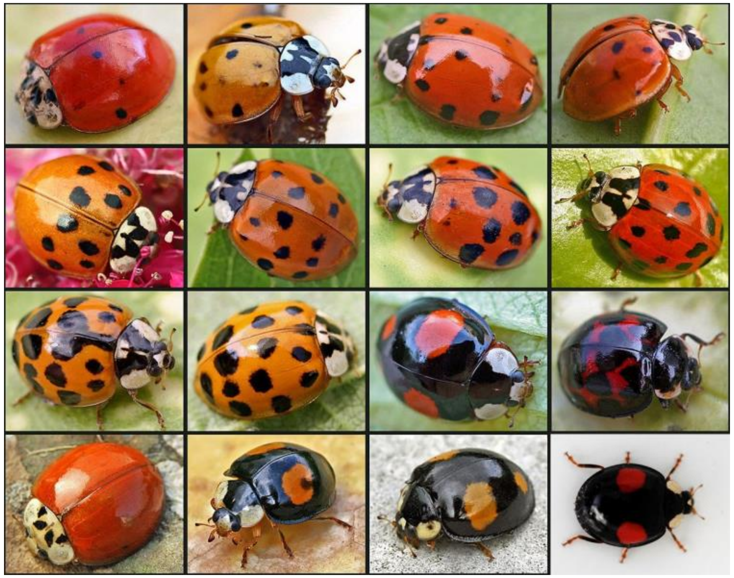
Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com



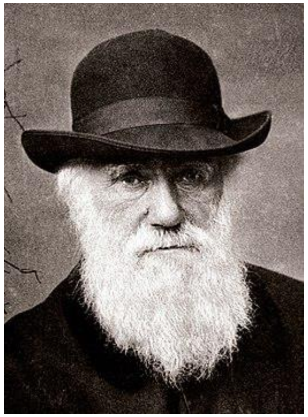


Studying Genetic Diversity using Single Nucleotidic Polymorphism (SNP)

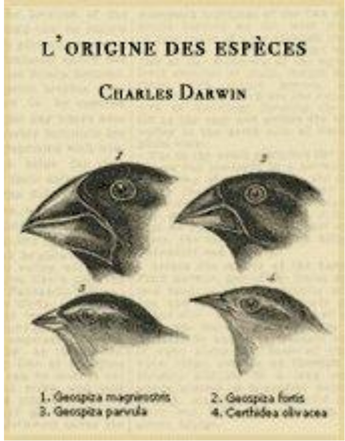
Population Variation



Theory of evolution - C. Darwin 1859



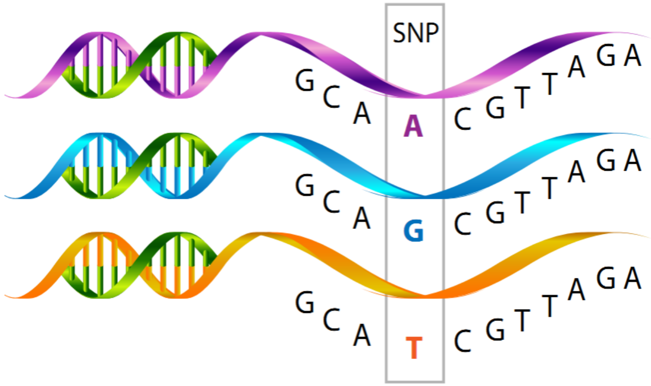
Understanding how individuals of a same species vary



- ✓ **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

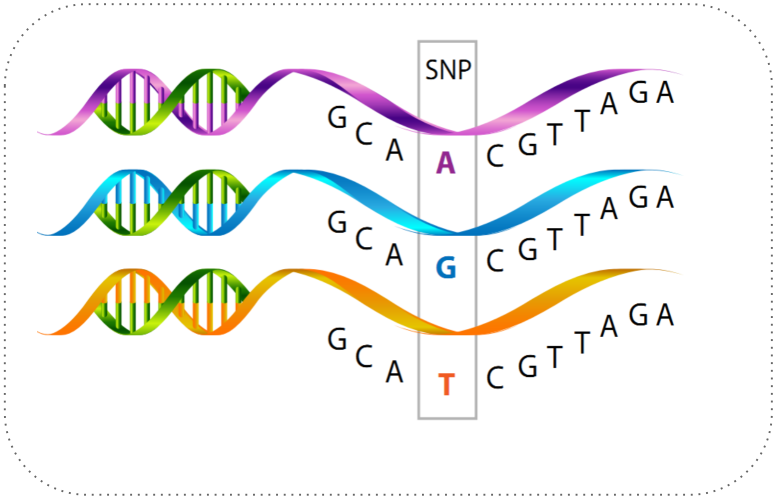
Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism

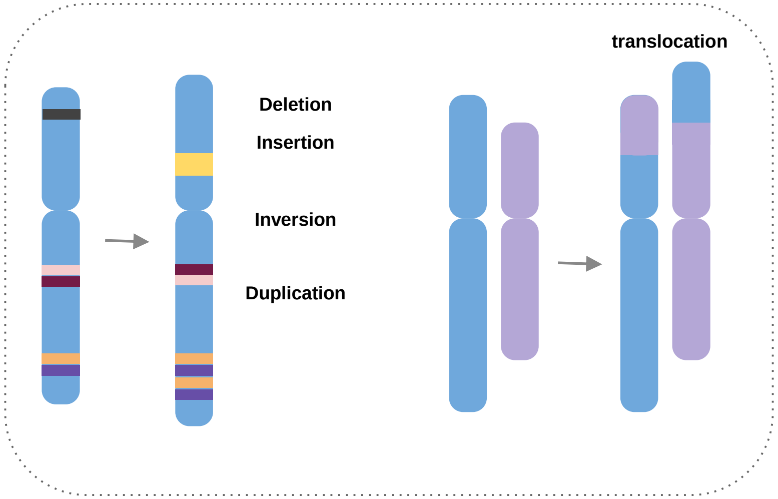


Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism

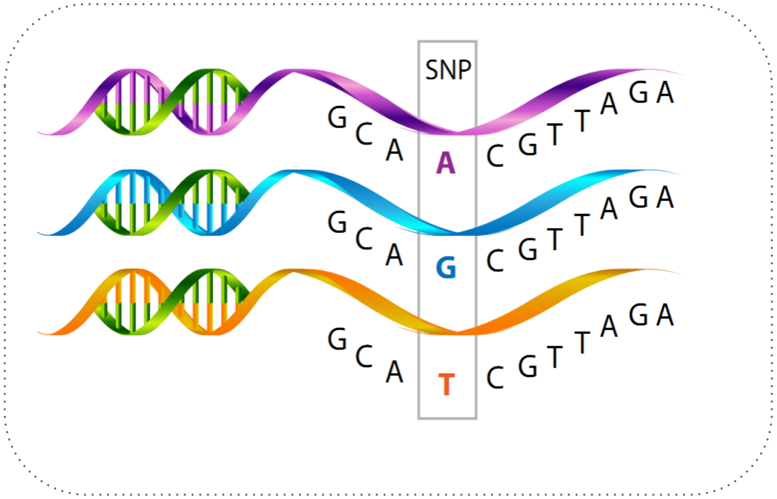


Structural Variations

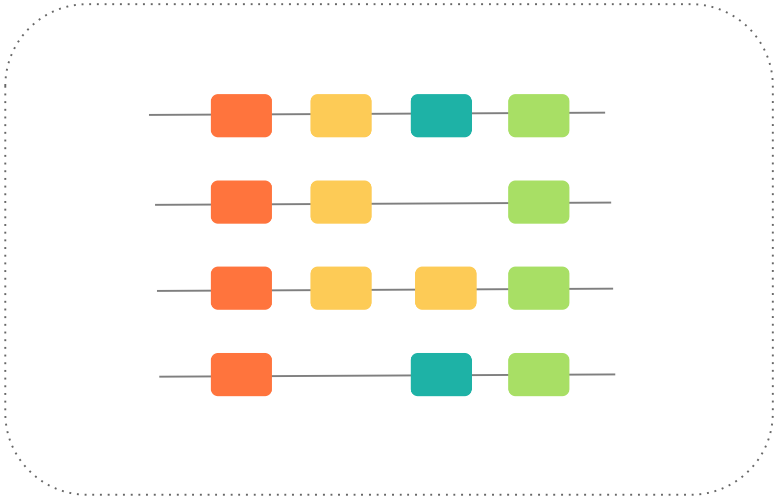


Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism

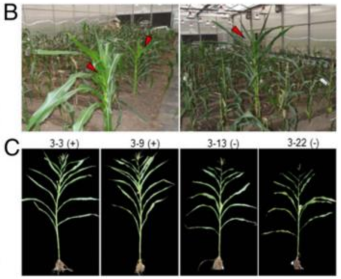


Structural Variations



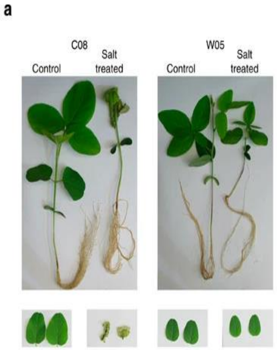
Presence Absence Variation (PAV)

Deletion, duplication, copy number variation, mobile element insertion



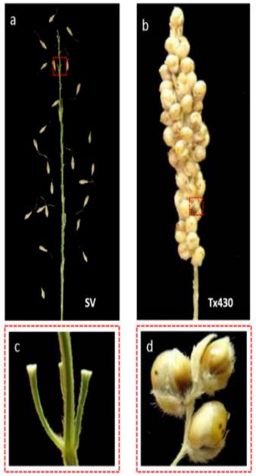
FromYang et al., 2013

From Li et al. 2012

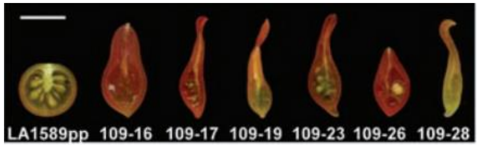


From Qi et al. 2014

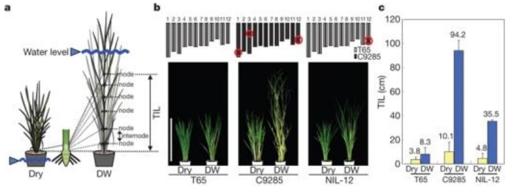
FromYang et al., 2014



From Lin et al. 2012

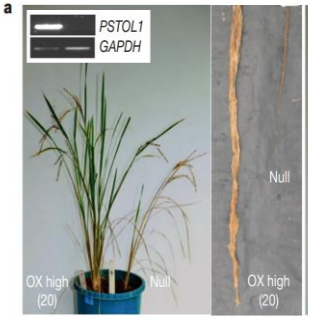


From Xiao et al. 2008

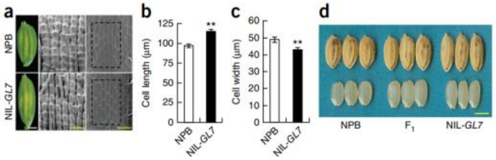


From Hattori et al. 2009

From Xu et al. 2006

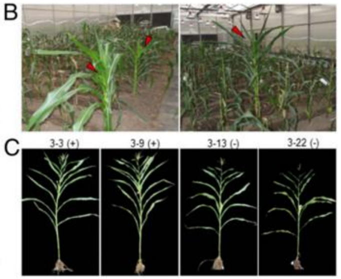


From Gamuyao et al. 2012



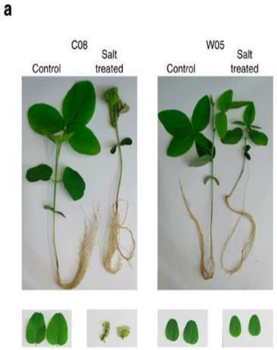
From Wang et al. 2015

From Bai et al. 2017



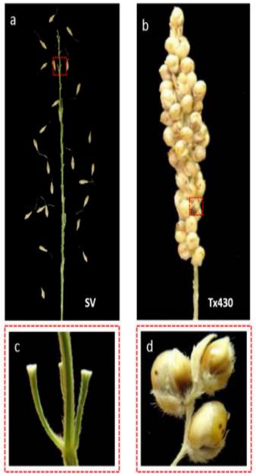
FromYang et al., 2013

From Li et al. 2012



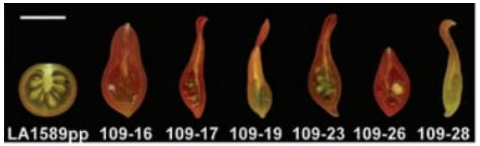
From Qi et al. 2014

FromYang et al., 2014

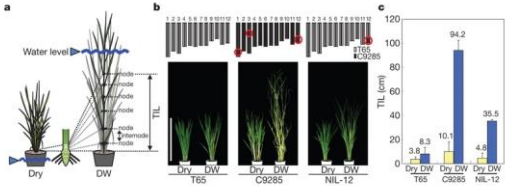


From Lin et al. 2012

Is One Reference genome enough to capture all genetic diversity ?

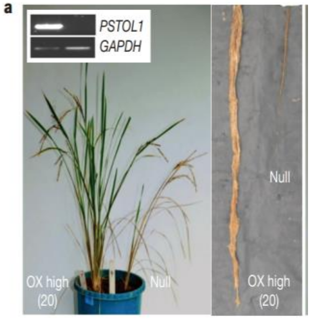


From Xiao et al. 2008

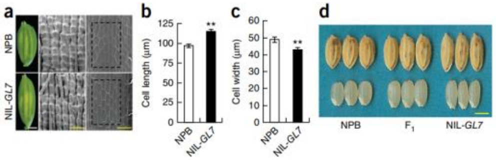


From Hattori et al. 2009

From Xu et al. 2006



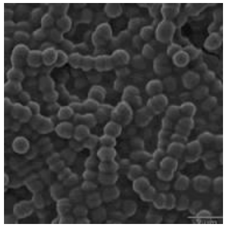
From Gamuyao et al. 2012



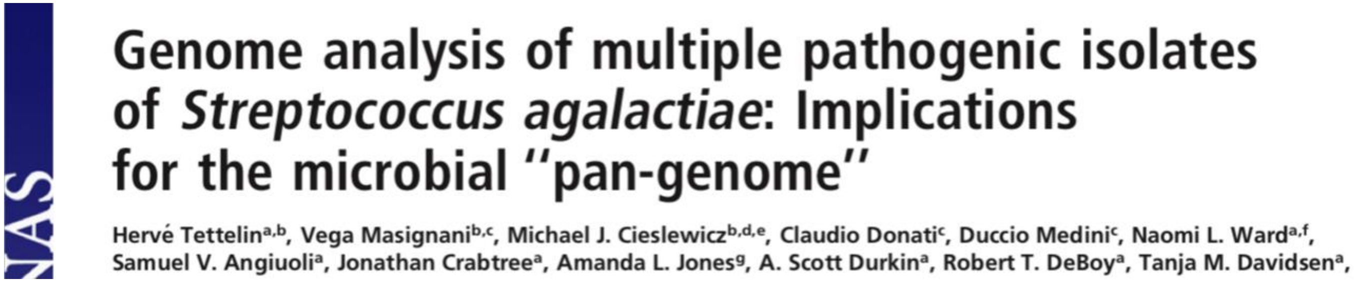
From Wang et al. 2015

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

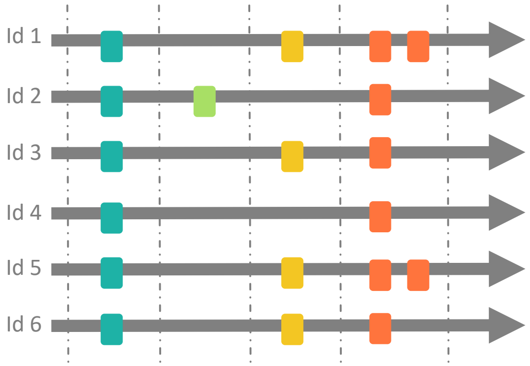
- 8 strains sequenced
- SNP variations

Large number of genes not shared between isolates

20% genome variability and 80 % shared by all isolates

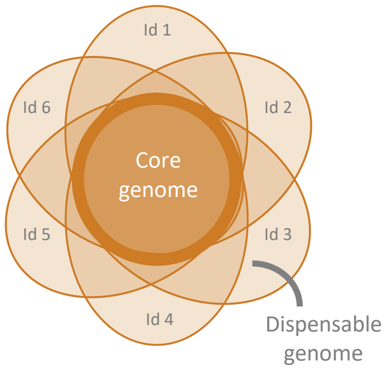
Pangenome concept

Pangenome concept

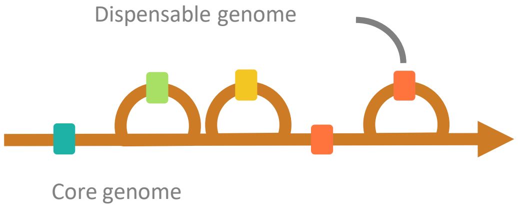


Pangenome

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



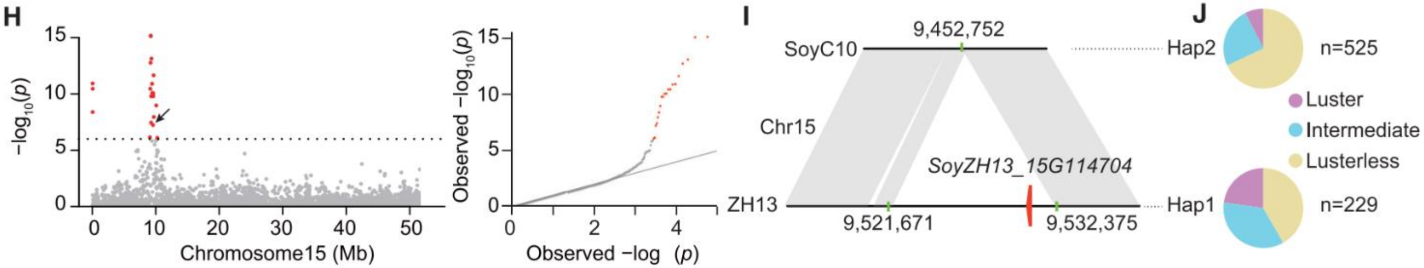
- **Core genome** : present in all individuals
- **Dispensable genome** : absent from one or several individuals (also called variable, accessory,...)



What's else ?

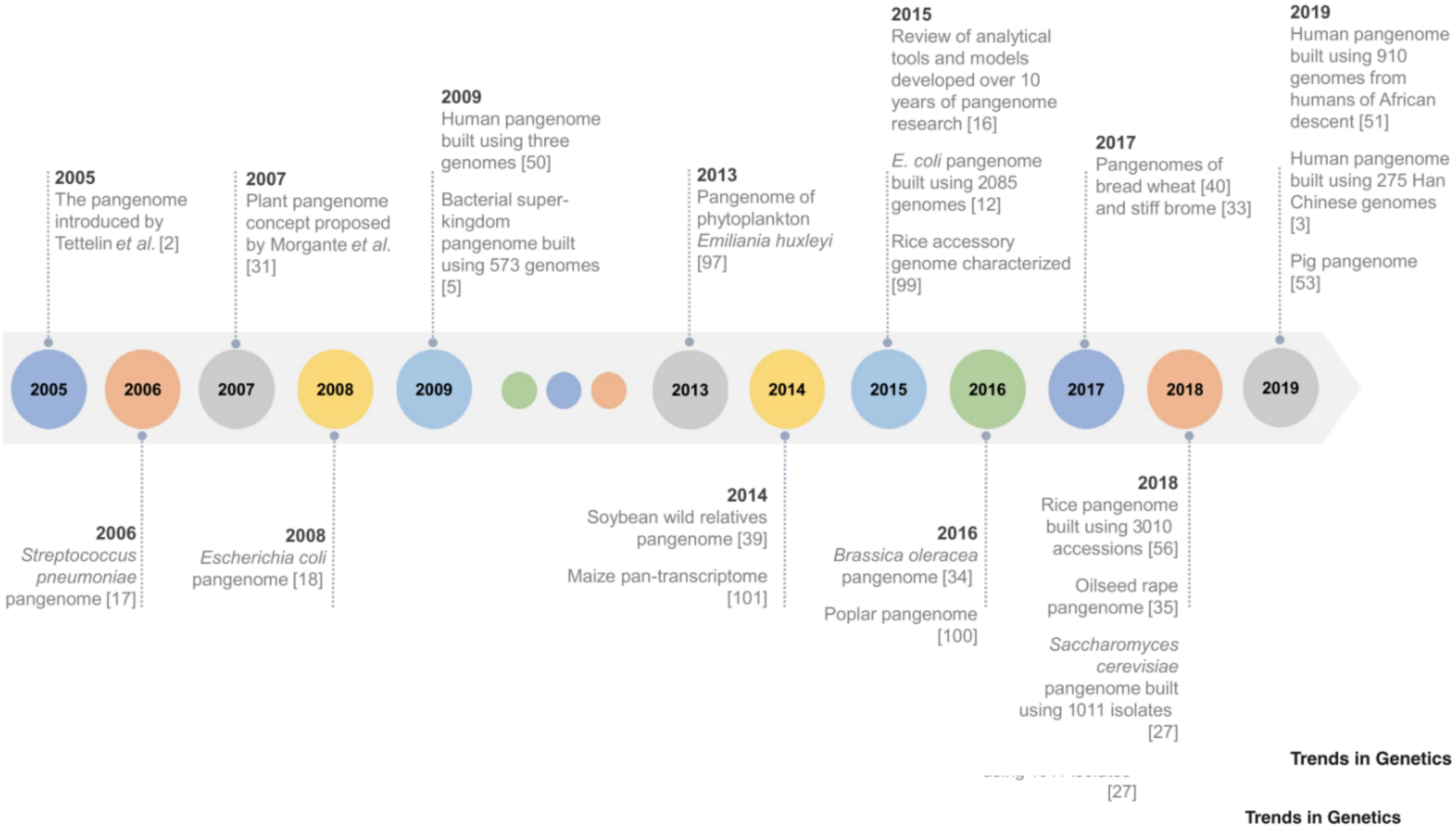
- 12,150 genes absent from the reference (18 cultivars)

- 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)



2005
The pangenome introduced by Tettelin *et al.* [2]

2006
Streptococcus pneumoniae pangenome [17]

2007
Plant pangenome concept proposed by Morgante *et al.* [31]

2008
Escherichia coli pangenome [18]

2009
Human pangenome built using three genomes [50]

Bacterial super-kingdom pangenome built using 573 genomes [5]

2013
Pangenome of phytoplankton *Emiliania huxleyi* [97]

2014
Soybean wild relatives pangenome [39]

Maize pan-transcriptome [101]

2015
Review of analytical tools and models developed over 10 years of pangenome research [16]

E. coli pangenome built using 2085 genomes [12]

Rice accessory genome characterized [99]

2016
Brassica oleracea pangenome [34]

Poplar pangenome [100]

2017
Pangenomes of bread wheat [40] and stiff brome [33]

2018
Rice pangenome built using 3010 accessions [56]

Oilseed rape pangenome [35]

Saccharomyces cerevisiae pangenome built using 1011 isolates [27]

2019
Human pangenome built using 910 genomes from humans of African descent [51]

Human pangenome built using 275 Han Chinese genomes [3]

Pig pangenome [53]

Trends in Genetics

using 1011 isolates
[27]

Trends in Genetics

Golicz et al., 2020

How to detect SV ?

2 formations …

2 ambiances

Mode “training”

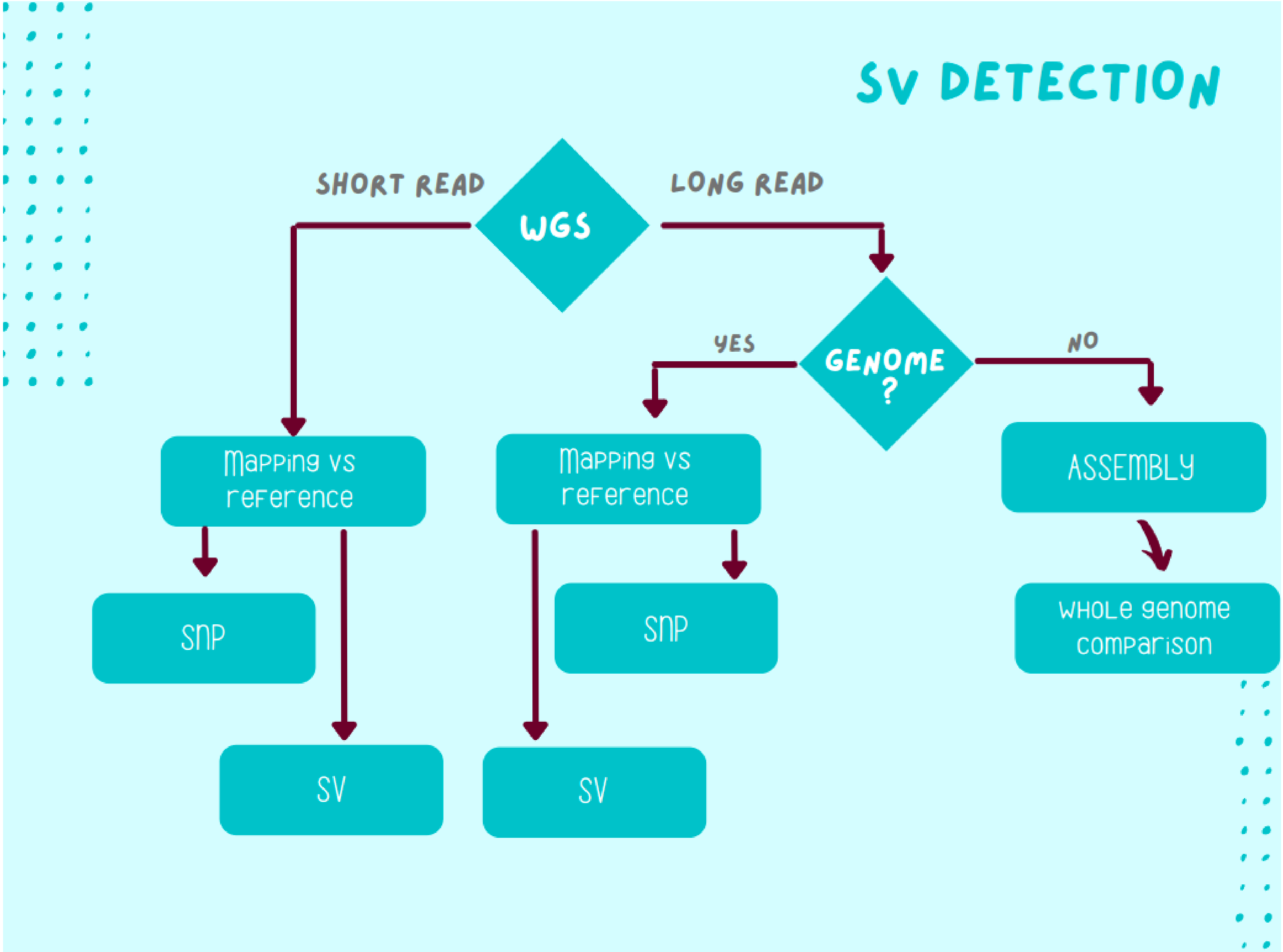
- Session cours suivi par
- Session pratique en autonomie (individuel ou en groupe)
- Correction

Mode “projet”

- brainstroming en groupe, avec les formateurs
- projet en autonomie…
- debriefing collectif
- 2 projets en parallèle !

Des données différentes pour les 2 groupes avec des analyses différentes !!!

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



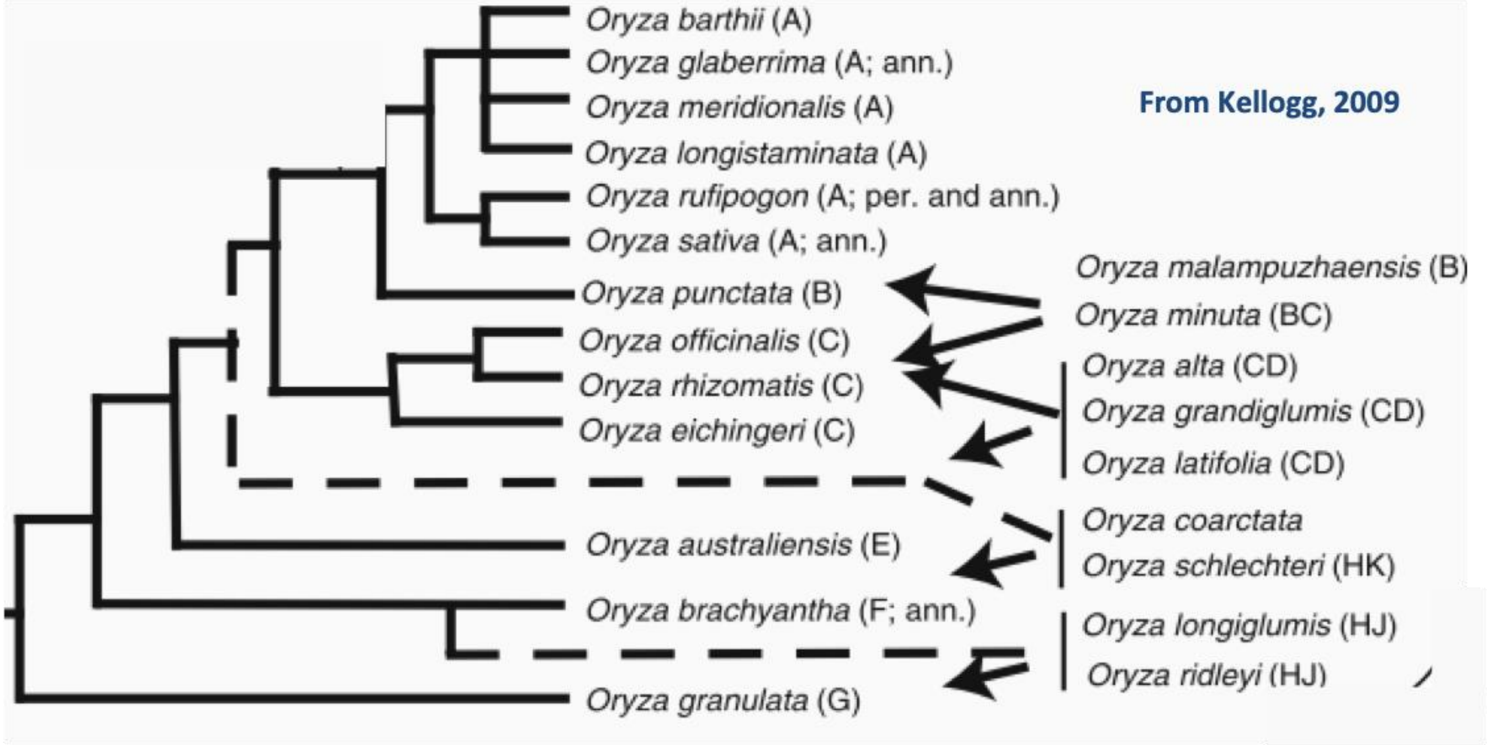
Détection de variants
à partir de données de séquençage short
& long reads

#data

What data will we use for our training ?

Genre Oryza

- 21 espèces sauvages



From Kellogg, 2009

From Wikimedia

What data will we use for our training ?

Genre Oryza

- 21 espèces sauvages
- 2 espèces domestiquées
    - *Oryza sativa*
    - *Oryza glaberrima*

From Wikimedia

What data will we use for our training ?

Oryza sativa

- Culture céréalière importante
- aliment de base de plus de la moitié de la population humaine
- Espèces diploïdes, 2n = 24 (genome AA)
- Céréale avec un génome petit
- Plante modèle
- Domestiquée ~10000 ans - O. rufipogon

From Wikimedia

What data will we use for our training ?

20 indidividus d’O. sativa ⇔ 20 clones avec une diversité intéressante

⇓

Séquençage short and long reads

illumina®

Oxford NANOPORE™ Technologies

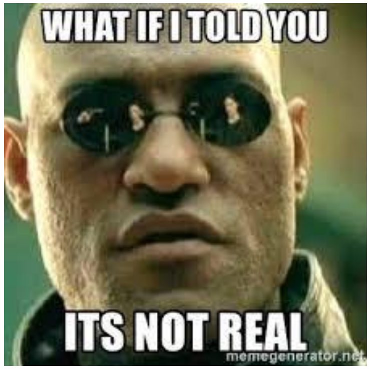
What data will we use for our training ?

20 indidividus d'O. sativa ⇔ 20 clones
avec une diversité intéressante

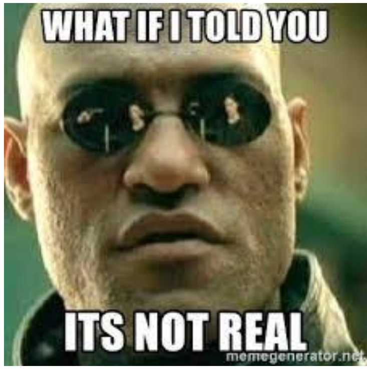
Séquençage short and long reads

illumina®

Oxford NANOPORE™ Technologies



What data will we use for our training ?



20 indidividus d'O. sativa ⇔ 20 clones
avec une diversité intéressante

1. Extract 1 Mb from the Chromosome 1
2. Create 20 exact clones
3. Introduce mutations with bioinformatics program
   a. SNP : from 1 to 10%
   b. indel : between 10bp and 10kb
   c. duplications
4. Getting 20 clones with different mutations that were sequenced in silico (short & long reads)

Projet SNP

MISSION ~~IM~~POSSIBLE

NOM DE CODE : “PROJET SNP”

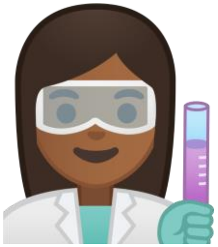
Votre mission si vous l’acceptez…

#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.



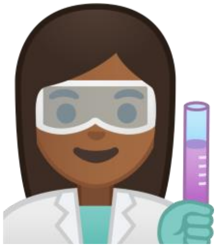
- Avons nous collecté une nouvelle espèce d'igname ?
- Ou avons nous collecté des ignames domestiqués ? sauvages ?

#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…

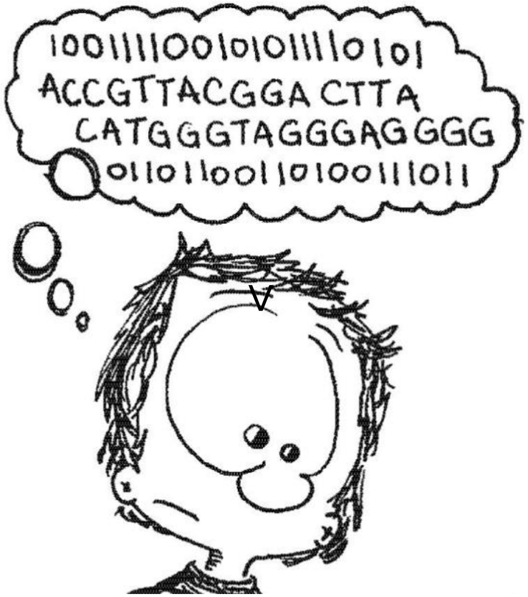
Dominique



Je compte sur vous !!!!

#DATA :

Décrire où seront les données à partir de mardi…



A vous de jouer !

Metagenomic

#MISSION :

Comment caractériser la diversité métagénomique de l'échantillon mystère ?

NOM DE CODE : "MÉTAGÉNOMIQUE"

#MISSION :

La productrice Mme. BOBODOU voudrais savoir pourquoi son champ d’ananas est peu productif

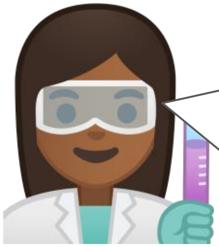
Elle a vu que les feuilles de la plante etaient plus jaunes que d’habitude… Elle s'inquiète!

Elle a collecté quelques feuilles et a fait séquencer son échantillon mystère par la technologie Oxford Nanopore à un collègue de l’UJKZ.

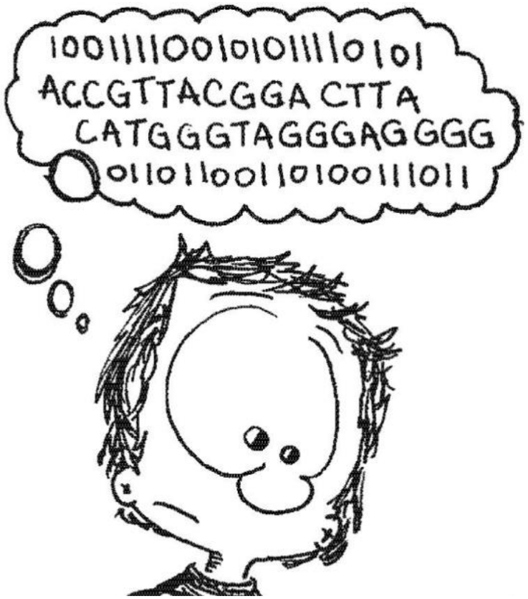
=> Aidez-lui à caractériser cet échantillon !

#MISSION :

- **Aidez-lui à caractériser cet échantillon !**



- Au laboratoire quelques marqueurs PCR sont négatifs pour les bactéries et les champignons pathogènes? aussi pour certains virus. Il s'agit d'une nouvelle espèce ?



A vous de jouer !

Bioinformatics resources

On va travailler sous Linux !

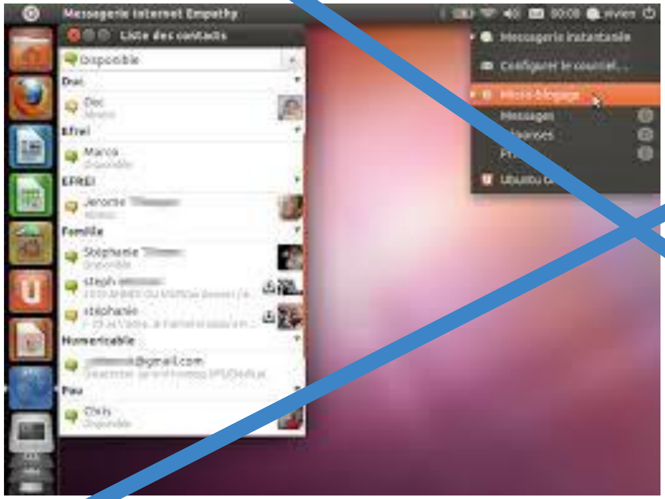
- 2 façons d’utiliser linux :

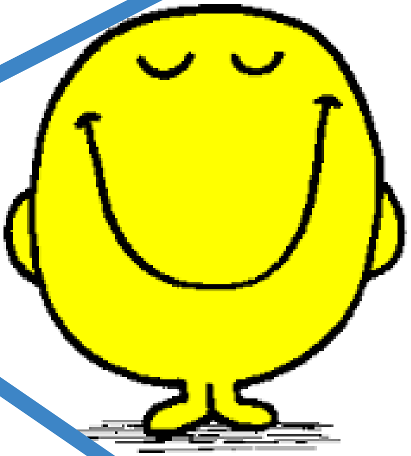
  en ***mode graphique***

On va travailler sous Linux !

- 2 façons d'utiliser linux :

en ***mode graphique***

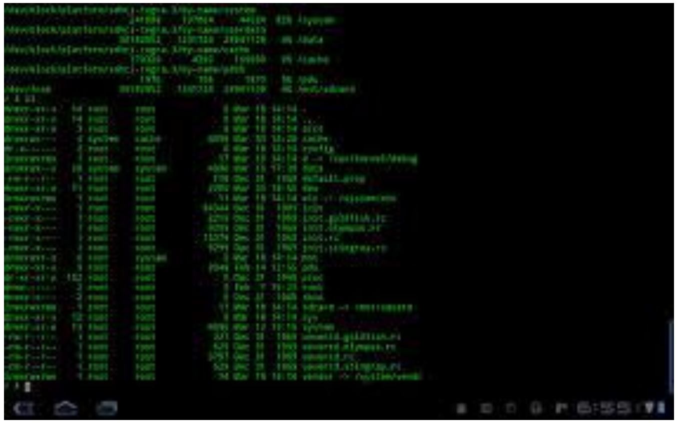




En mode terminal

- 2 façons d’utiliser linux :

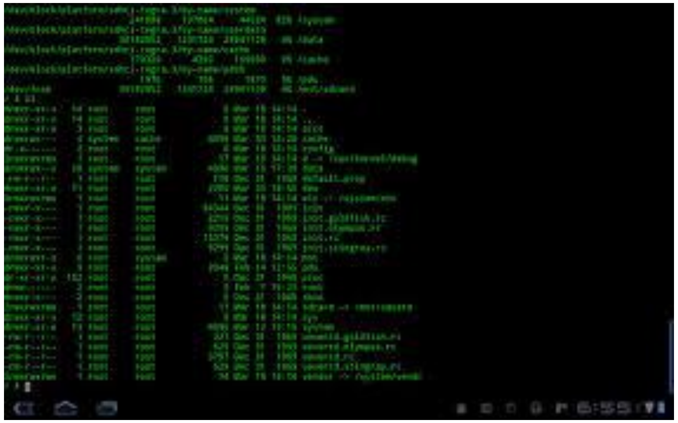
  en ***mode console***



En mode terminal

- 2 façons d’utiliser linux :

en ***mode console***



Sur le cluster de l’université !

En mode jupyter book

- Une troisième façon d’utiliser linux :

  en ***mode jupyter bool***

➡ Sur le cloud IFB!

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

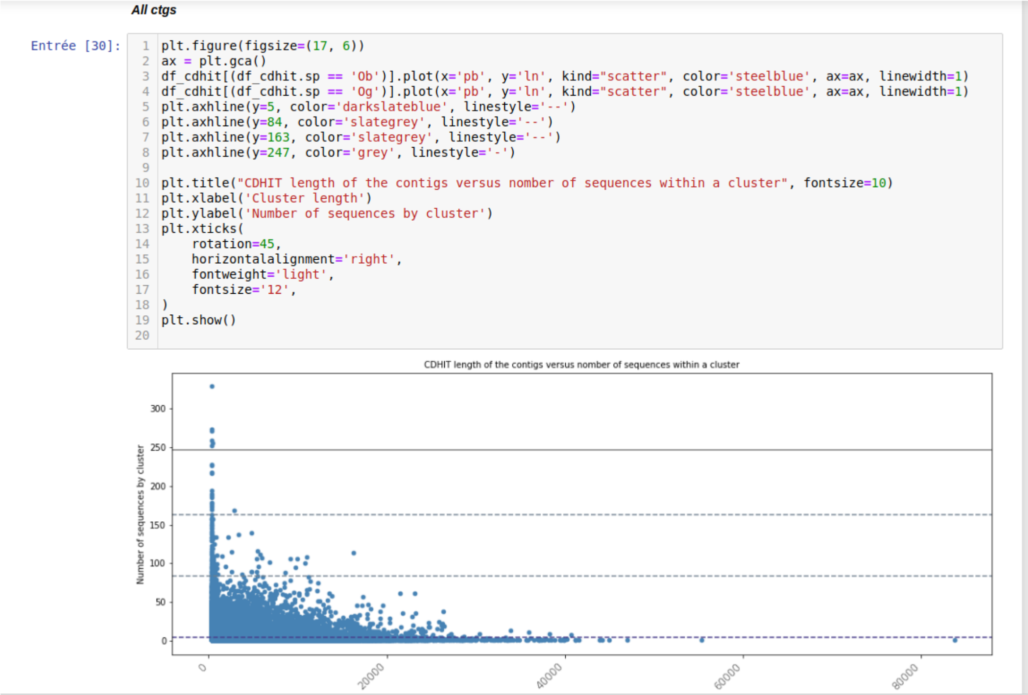
ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala

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]:

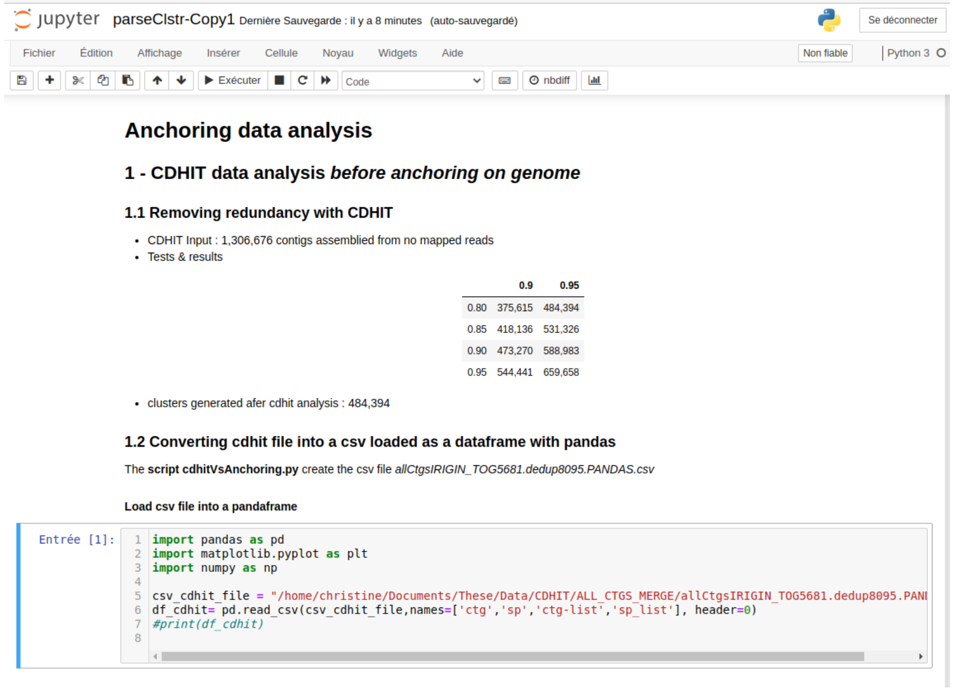
```
plt.figure(figsize=(17, 6))
ax = plt.gca()
df_cdhit[(df_cdhit.sp == 'Ob')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
df_cdhit[(df_cdhit.sp == 'Og')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
plt.axhline(y=5, color='darkslateblue',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

```

CDHIT length of the contigs versus nomber of sequences within a cluster
Number of sequences by cluster
300
250
200
150
100
50
0
0
20000
40000
60000
80000

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



Anchoring data analysis

1 - CDHIT data analysis *before anchoring on genome*

1.1 Removing redundancy with CDHIT

- CDHIT Input : 1,306,676 contigs assemblied from no mapped reads
- Tests & results

	0.9	0.95
0.80	375,615	484,394
0.85	418,136	531,326
0.90	473,270	588,983
0.95	544,441	659,658

- clusters generated afer cdhit analysis : 484,394

1.2 Converting cdhit file into a csv loaded as a dataframe with pandas

The **script cdhitVsAnchoring.py** create the csv file *allCtgsIRIGIN_TOG5681.dedup8095.PANDAS.csv*

Load csv file into a pandaframe

Entrée [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

csv_cdhit_file = "/home/christine/Documents/These/Data/CDHIT/ALL_CTGS_MERGE/allCtgsIRIGIN_TOG5681.dedup8095.PAN
df_cdhit= pd.read_csv(csv_cdhit_file,names=['ctg','sp','ctg-list','sp_list'], header=0)
#print(df_cdhit)

```

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

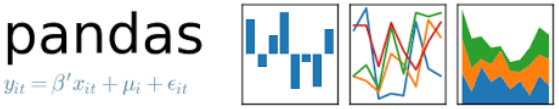
How to become a super datascientist ?

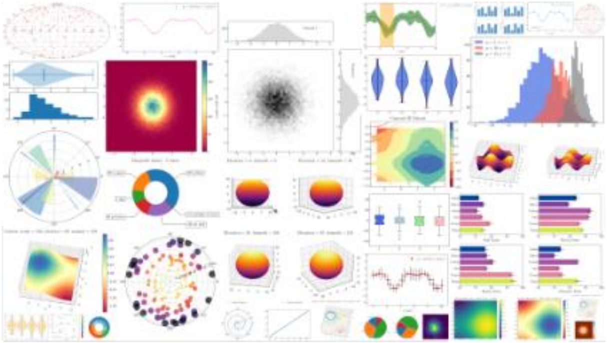
- facilement importer des fichiers tabulés dans des dataframes, similaires aux dataframes sous R.

  (et exporter)

- manipuler ces tableaux de données / DataFrames

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





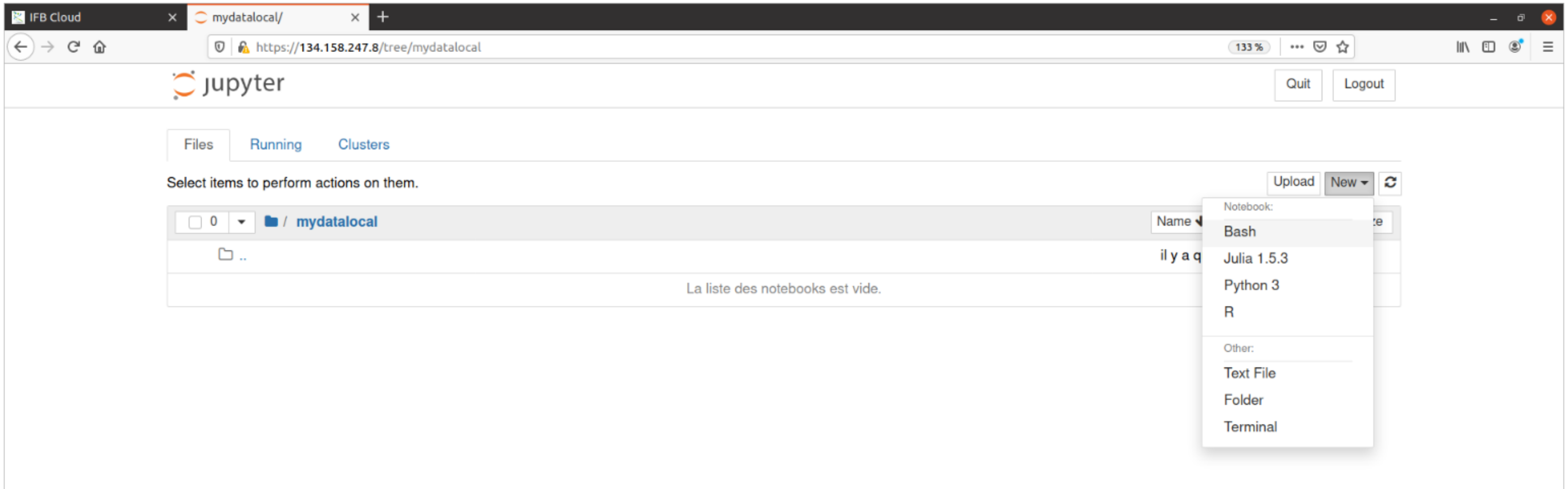
How will you use Jupyter Notebook ?

- 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



https://134.158.247.8/tree/mydatalocal

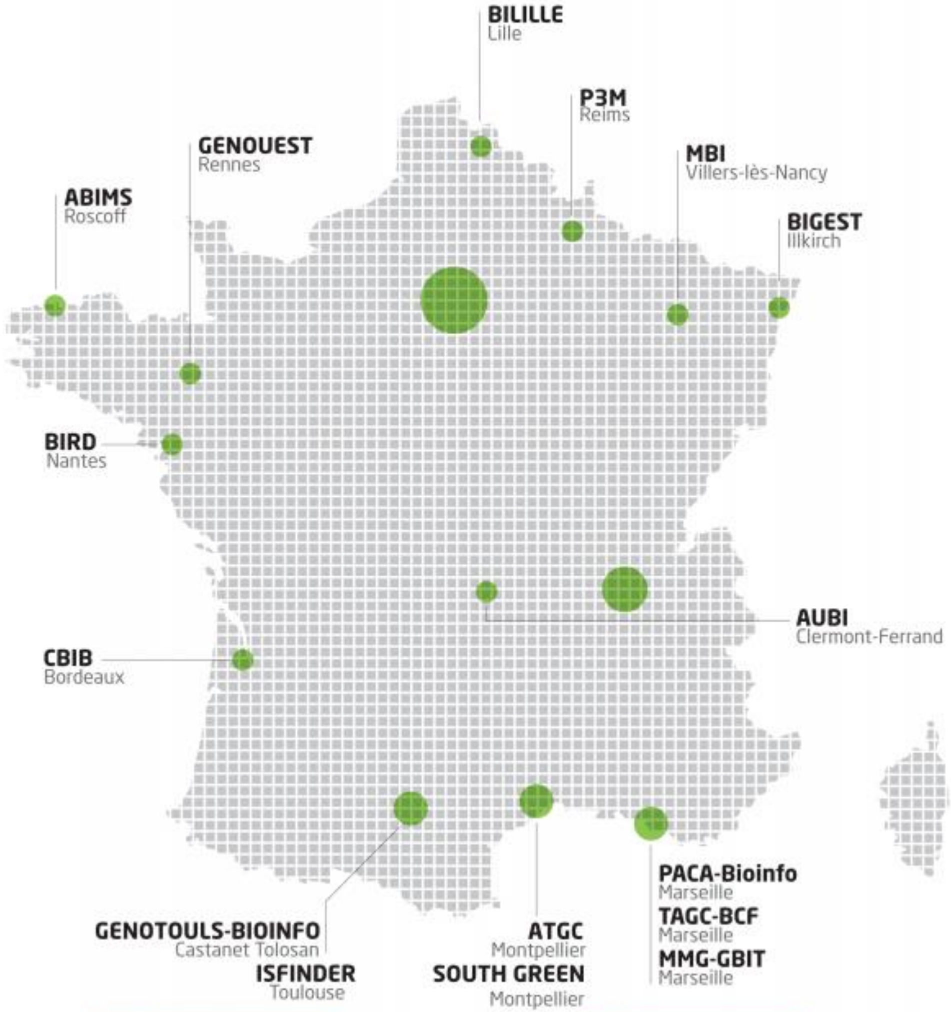
Select items to perform actions on them.

La liste des notebooks est vide.

IFB ?

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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



RÉGION PARISIENNE

EBIO
Orsay
INSTITUT CURIE
Paris
IGR
Villejuif
MICROSCOPE
Evry
MIGALE
Jouy-en-Josas

C3BI
Paris
RPBS
Paris
URGI
Versailles
ORPHANET
Paris
ICONICS
Paris
IFB CORE
Evry

RÉGION LYONNAISE

INCA-SLC
Lyon
PRABI-HCL
Lyon
PRABI-AMSB
Villeurbanne
PRABI-Lyon-Grenoble
Villeurbanne
PRABI-Lyon-Gerland
Lyon

Biosphere, IFB CLOUD FOR LIFE SCIENCES

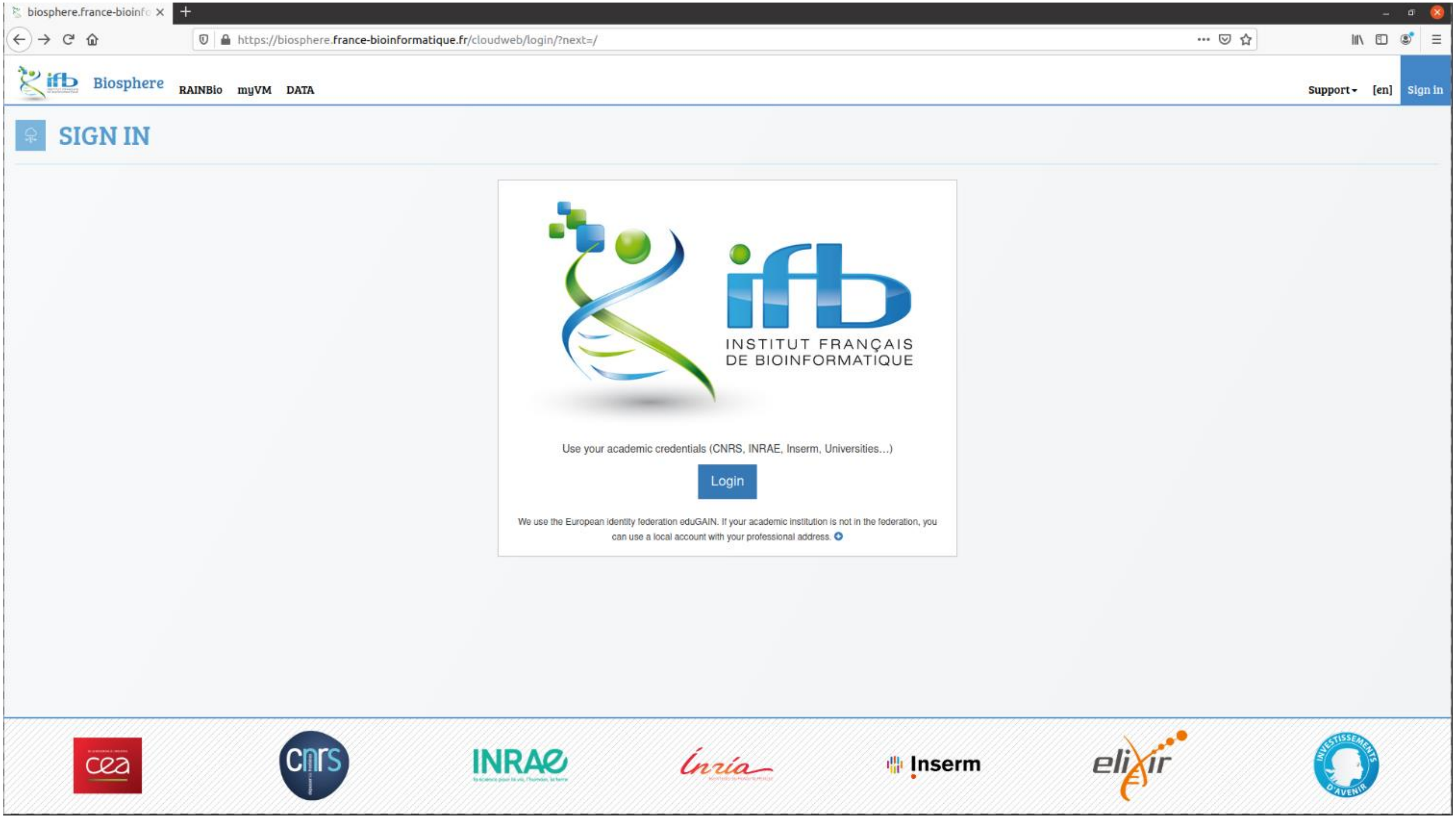
- A federation of clouds, which relies on interconnected IFB's infrastructures, providing distributed services to analyze life science data

- .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



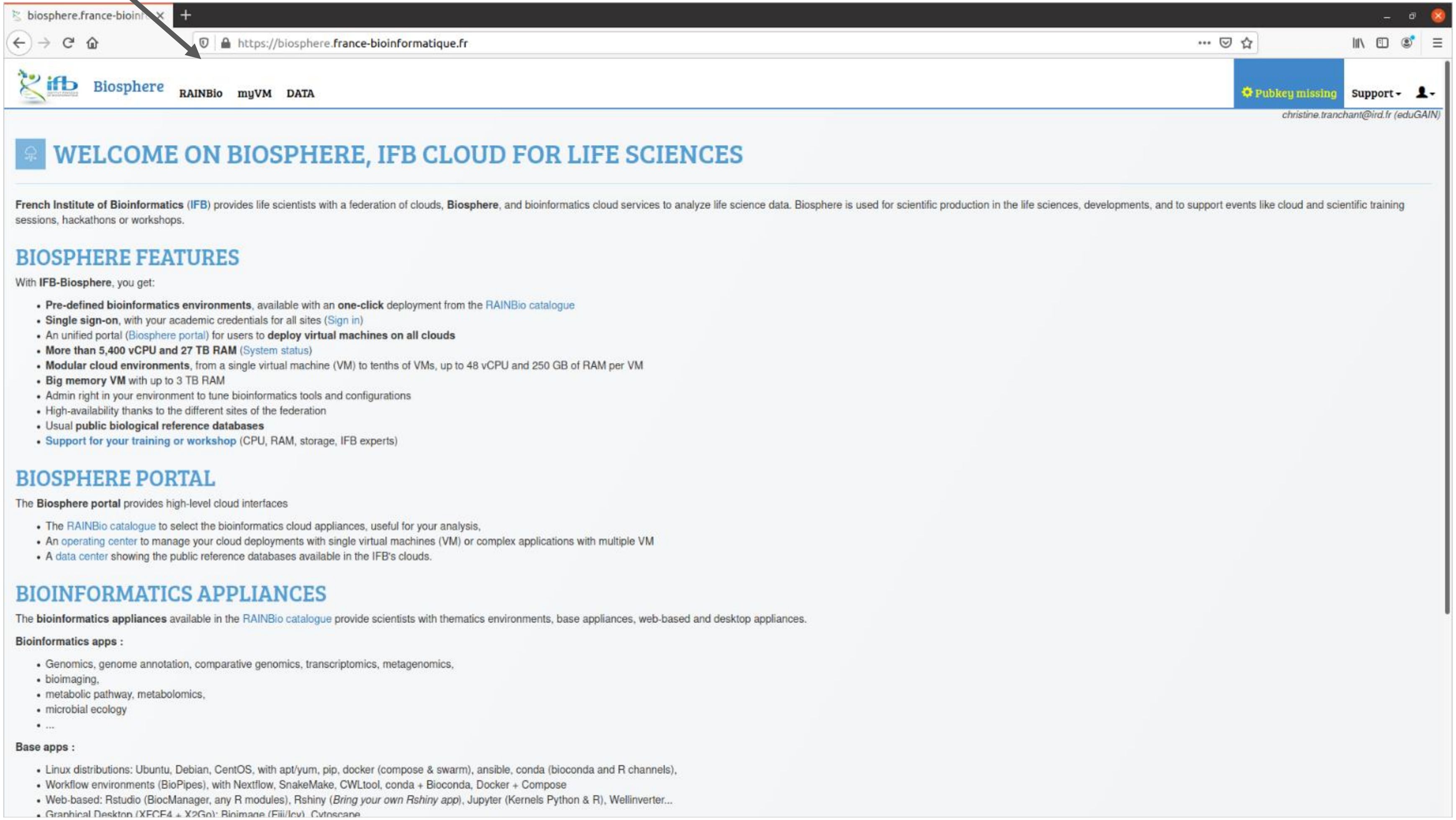
SIGN IN

Use your academic credentials (CNRS, INRAE, Inserm, Universities...)

We use the European identity federation eduGAIN. If your academic institution is not in the federation, you can use a local account with your professional address.

Connected / here we are

RAINBIO catalog to access our Virtual Machine (VM)


biosphere.france-bioinf
https://biosphere.france-bioinformatique.fr
ifb
Biosphere
RAINBio
myVM
DATA
Pubkey missing
Support
christine.tranchant@ird.fr (eduGAIN)

WELCOME ON BIOSPHERE, IFB CLOUD FOR LIFE SCIENCES

French Institute of Bioinformatics (IFB) provides life scientists with a federation of clouds, **Biosphere**, and bioinformatics cloud services to analyze life science data. Biosphere is used for scientific production in the life sciences, developments, and to support events like cloud and scientific training sessions, hackathons or workshops.

BIOSPHERE FEATURES

With **IFB-Biosphere**, you get:

- **Pre-defined bioinformatics environments**, available with an **one-click** deployment from the RAINBio catalogue
- **Single sign-on**, with your academic credentials for all sites (Sign in)
- An unified portal (Biosphere portal) for users to **deploy virtual machines on all clouds**
- **More than 5,400 vCPU and 27 TB RAM** (System status)
- **Modular cloud environments**, from a single virtual machine (VM) to tenths of VMs, up to 48 vCPU and 250 GB of RAM per VM
- **Big memory VM** with up to 3 TB RAM
- Admin right in your environment to tune bioinformatics tools and configurations
- High-availability thanks to the different sites of the federation
- Usual **public biological reference databases**
- **Support for your training or workshop** (CPU, RAM, storage, IFB experts)

BIOSPHERE PORTAL

The **Biosphere portal** provides high-level cloud interfaces

- The RAINBio catalogue to select the bioinformatics cloud appliances, useful for your analysis,
- An operating center to manage your cloud deployments with single virtual machines (VM) or complex applications with multiple VM
- A data center showing the public reference databases available in the IFB's clouds.

BIOINFORMATICS APPLIANCES

The **bioinformatics appliances** available in the RAINBio catalogue provide scientists with thematics environments, base appliances, web-based and desktop appliances.

Bioinformatics apps :

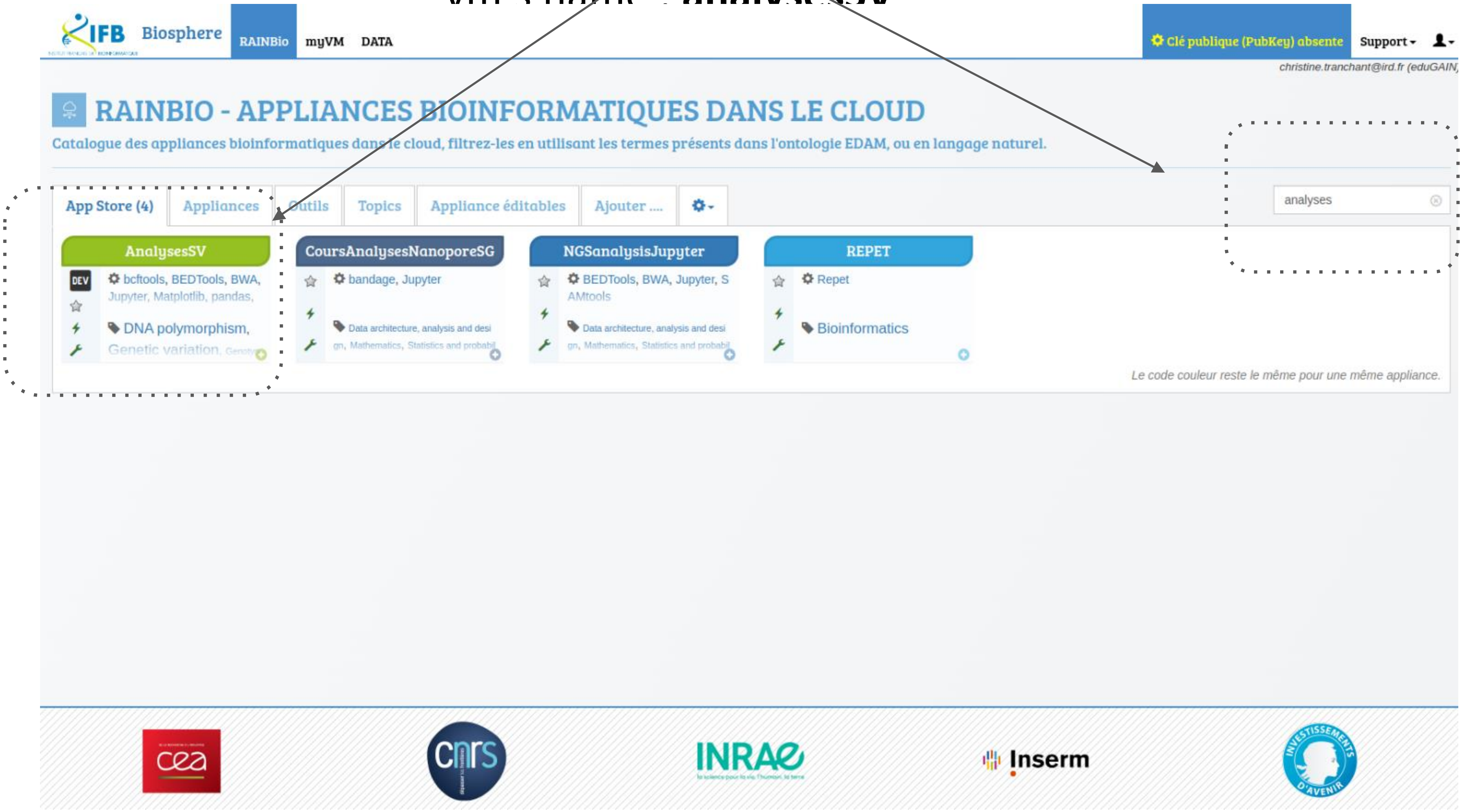
- Genomics, genome annotation, comparative genomics, transcriptomics, metagenomics,
- bioimaging,
- metabolic pathway, metabolomics,
- microbial ecology
- ...

Base apps :

- Linux distributions: Ubuntu, Debian, CentOS, with apt/yum, pip, docker (compose & swarm), ansible, conda (bioconda and R channels),
- Workflow environments (BioPipes), with Nextflow, SnakeMake, CWLtool, conda + Bioconda, Docker + Compose
- Web-based: Rstudio (BiocManager, any R modules), Rshiny (*Bring your own Rshiny app*), Jupyter (Kernels Python & R), Wellinverter...
- Graphical Desktop (XFCE4 + X2Go): Bioimage (Fiji/Icy), Cytoscape

Searching for the vm we will use

vm's name : **analysesSV**



IFB Biosphere
RAINBio
myVM
DATA
Clé publique (PubKey) absente
Support
christine.tranchant@ird.fr (eduGAIN)

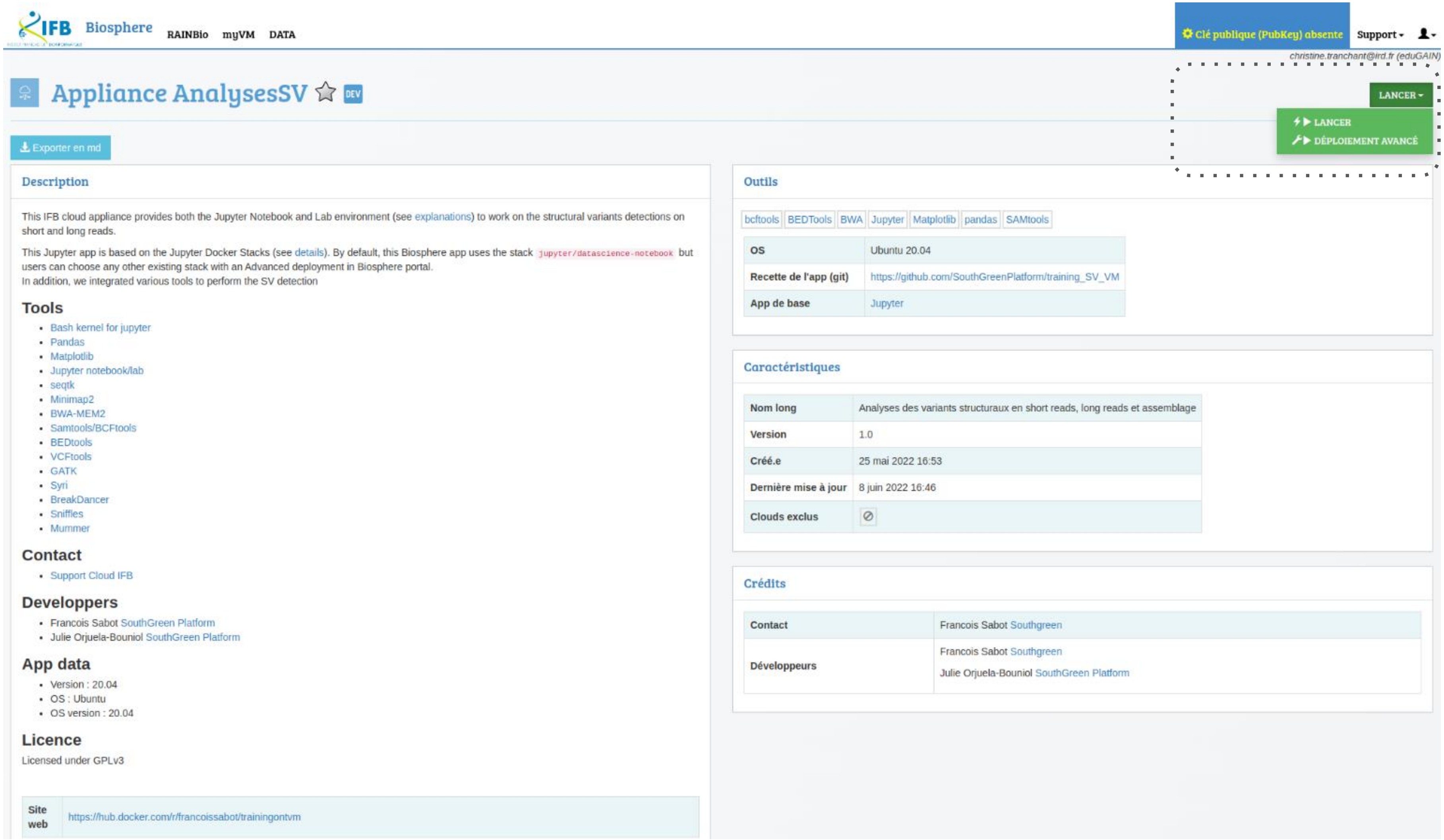
RAINBIO - APPLIANCES BIOINFORMATIQUES DANS LE CLOUD

Catalogue des appliances bioinformatiques dans le cloud, filtrez-les en utilisant les termes présents dans l'ontologie EDAM, ou en langage naturel.

AnalysesSV	CoursAnalysesNanoporeSG	NGSanalysisJupyter	REPET
DEV bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas, DNA polymorphism, Genetic variation, Genoty	bandage, Jupyter Data architecture, analysis and design, Mathematics, Statistics and probabil	BEDTools, BWA, Jupyter, SAMtools Data architecture, analysis and design, Mathematics, Statistics and probabil	Repet Bioinformatics

Le code couleur reste le même pour une même appliance.

Let’s run your vm through the cloud



Appliance AnalysesSV

Description

This IFB cloud appliance provides both the Jupyter Notebook and Lab environment (see explanations) to work on the structural variants detections on short and long reads.

This Jupyter app is based on the Jupyter Docker Stacks (see details). By default, this Biosphere app uses the stack `jupyter/datascience-notebook` but users can choose any other existing stack with an Advanced deployment in Biosphere portal.
In addition, we integrated various tools to perform the SV detection

Tools

- Bash kernel for jupyter
- Pandas
- Matplotlib
- Jupyter notebook/lab
- seqtk
- Minimap2
- BWA-MEM2
- Samtools/BCFtools
- BEDtools
- VCFtools
- GATK
- Syri
- BreakDancer
- Sniffles
- Mummer

Contact

- Support Cloud IFB

Developpers

- Francois Sabot SouthGreen Platform
- Julie Orjuela-Bouniol SouthGreen Platform

App data

- Version : 20.04
- OS : Ubuntu
- OS version : 20.04

Licence

Licensed under GPLv3

Site web	https://hub.docker.com/r/francoissabot/trainingontvm

Outils

bcftools BEDTools BWA Jupyter Matplotlib pandas SAMtools

OS	Ubuntu 20.04
Recette de l'app (git)	https://github.com/SouthGreenPlatform/training_SV_VM
App de base	Jupyter

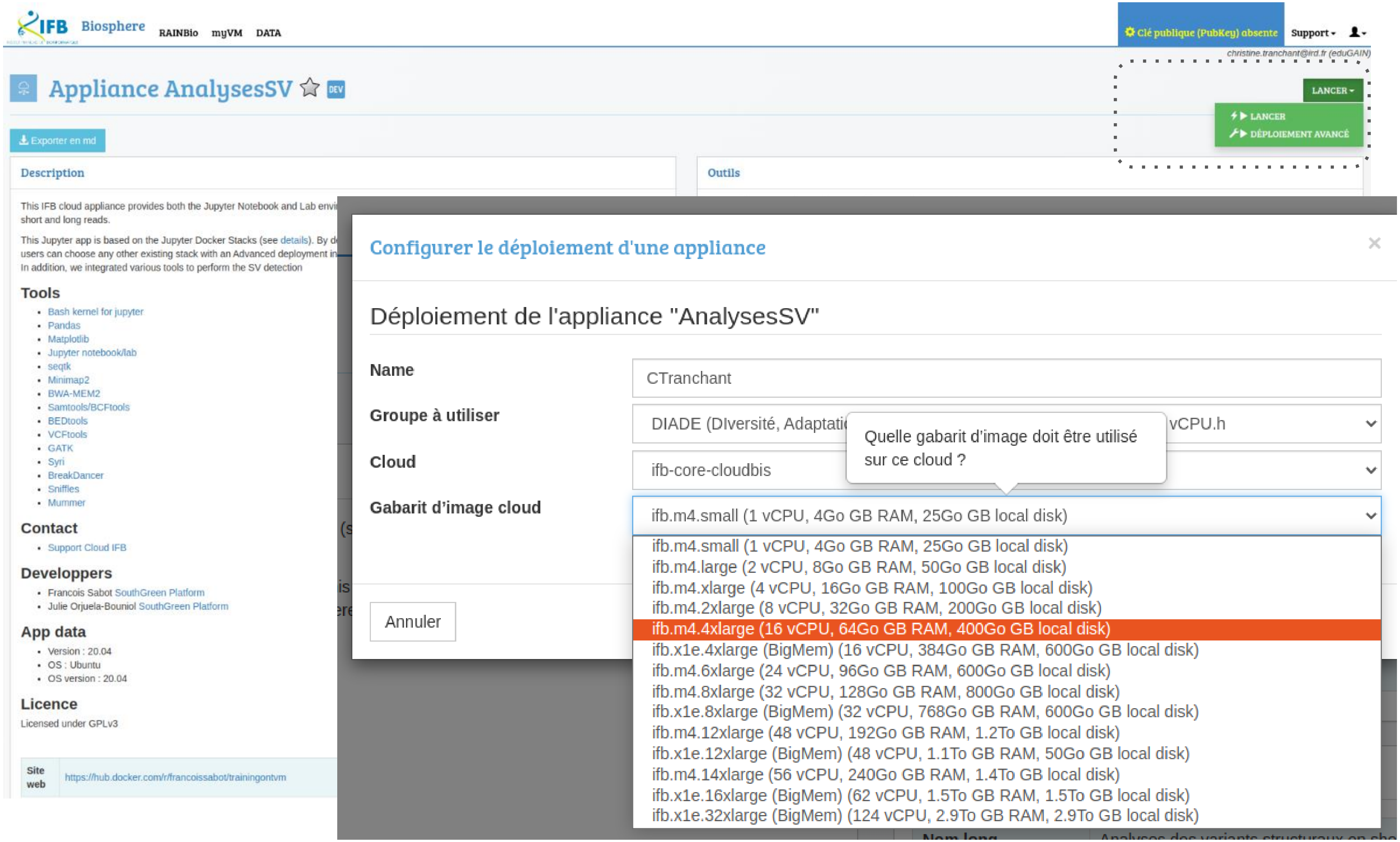
Caractéristiques

Nom long	Analyses des variants structuraux en short reads, long reads et assemblage
Version	1.0
Créé.e	25 mai 2022 16:53
Dernière mise à jour	8 juin 2022 16:46
Clouds exclus	⊘

Crédits

Contact	Francois Sabot Southgreen
Développeurs	Francois Sabot Southgreen Julie Orjuela-Bouniol SouthGreen Platform

Let’s run your vm through the cloud



Appliance AnalysesSV DEV

Description

This IFB cloud appliance provides both the Jupyter Notebook and Lab envi
short and long reads.

This Jupyter app is based on the Jupyter Docker Stacks (see details). By d
users can choose any other existing stack with an Advanced deployment in
In addition, we integrated various tools to perform the SV detection

Tools

- Bash kernel for jupyter
- Pandas
- Matplotlib
- Jupyter notebook/lab
- seqtk
- Minimap2
- BWA-MEM2
- Samtools/BCFtools
- BEDtools
- VCFtools
- GATK
- Syri
- BreakDancer
- Sniffles
- Mummer

Contact

- Support Cloud IFB

Developpers

- Francois Sabot SouthGreen Platform
- Julie Orjuela-Bouniol SouthGreen Platform

App data

- Version : 20.04
- OS : Ubuntu
- OS version : 20.04

Licence

Licensed under GPLv3

Site web	https://hub.docker.com/r/francoissabot/trainingontvm

Configurer le déploiement d'une appliance ×

Déploiement de l'appliance "AnalysesSV"

Name	CTranchant
Groupe à utiliser	DIADE (DIversité, Adaptati … vCPU.h
Cloud	ifb-core-cloudbis
Gabarit d'image cloud	ifb.m4.small (1 vCPU, 4Go GB RAM, 25Go GB local disk)

Quelle gabarit d'image doit être utilisé sur ce cloud ?

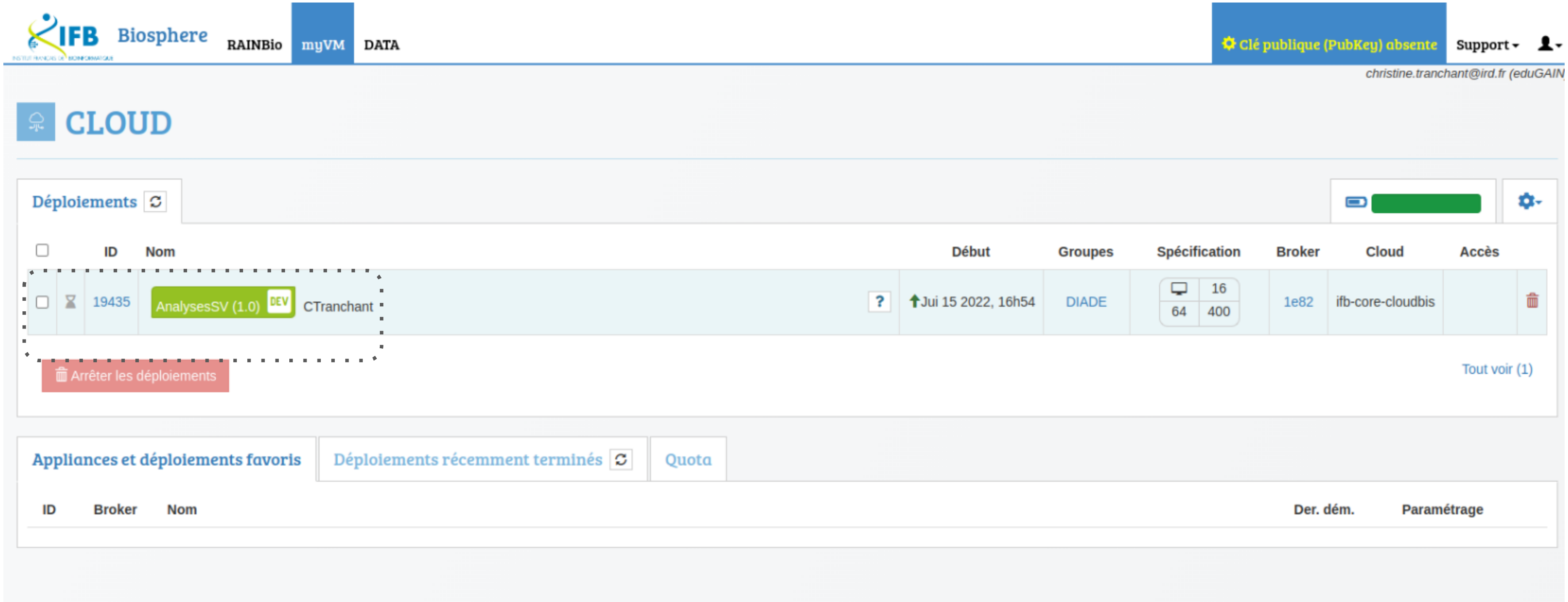
- ifb.m4.small (1 vCPU, 4Go GB RAM, 25Go GB local disk)
- ifb.m4.large (2 vCPU, 8Go GB RAM, 50Go GB local disk)
- ifb.m4.xlarge (4 vCPU, 16Go GB RAM, 100Go GB local disk)
- ifb.m4.2xlarge (8 vCPU, 32Go GB RAM, 200Go GB local disk)
- ifb.m4.4xlarge (16 vCPU, 64Go GB RAM, 400Go GB local disk)
- ifb.x1e.4xlarge (BigMem) (16 vCPU, 384Go GB RAM, 600Go GB local disk)
- ifb.m4.6xlarge (24 vCPU, 96Go GB RAM, 600Go GB local disk)
- ifb.m4.8xlarge (32 vCPU, 128Go GB RAM, 800Go GB local disk)
- ifb.x1e.8xlarge (BigMem) (32 vCPU, 768Go GB RAM, 600Go GB local disk)
- ifb.m4.12xlarge (48 vCPU, 192Go GB RAM, 1.2To GB local disk)
- ifb.x1e.12xlarge (BigMem) (48 vCPU, 1.1To GB RAM, 50Go GB local disk)
- ifb.m4.14xlarge (56 vCPU, 240Go GB RAM, 1.4To GB local disk)
- ifb.x1e.16xlarge (BigMem) (62 vCPU, 1.5To GB RAM, 1.5To GB local disk)
- ifb.x1e.32xlarge (BigMem) (124 vCPU, 2.9To GB RAM, 2.9To GB local disk)

Annuler

Nom long   Analyses des variants structuraux en sho

Let's run your vm through the cloud

Loading…



IFB Biosphere RAINBio myVM DATA

Clé publique (PubKey) absente Support

christine.tranchant@ird.fr (eduGAIN)

CLOUD

Déploiements

□		ID	Nom		Début	Groupes	Spécification	Broker	Cloud	Accès	
□	⌛	19435	AnalysesSV (1.0) DEV CTranchant	?	Jui 15 2022, 16h54	DIADE	16 64 400	1e82	ifb-core-cloudbis		

Arrêter les déploiements

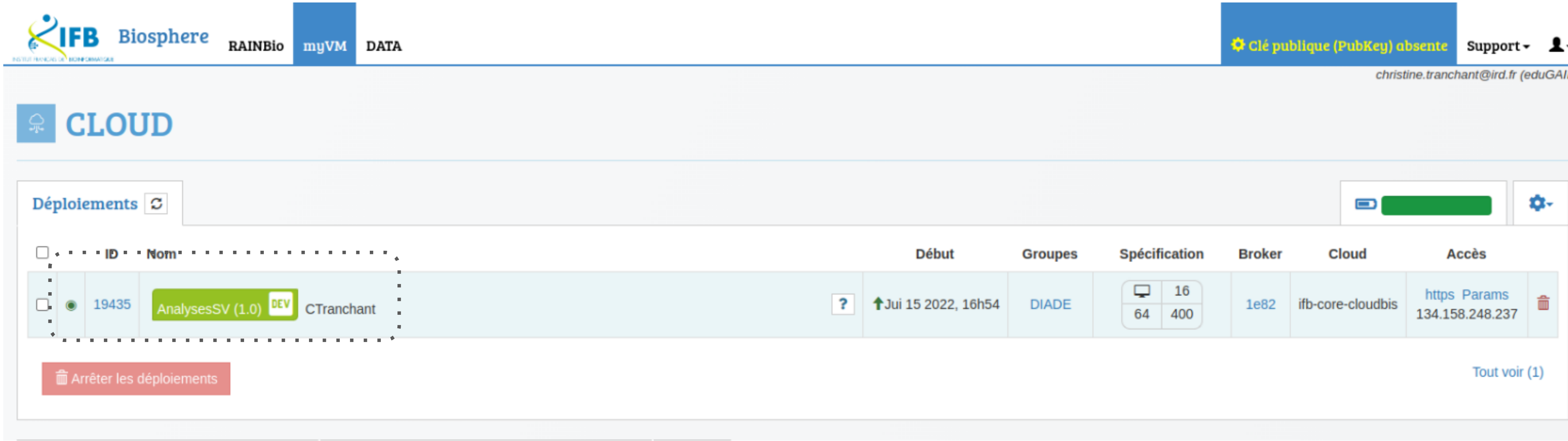
Tout voir (1)

Appliances et déploiements favoris | Déploiements récemment terminés | Quota

ID	Broker	Nom	Der. dém.	Paramétrage

Let's run your vm through the cloud

ready !



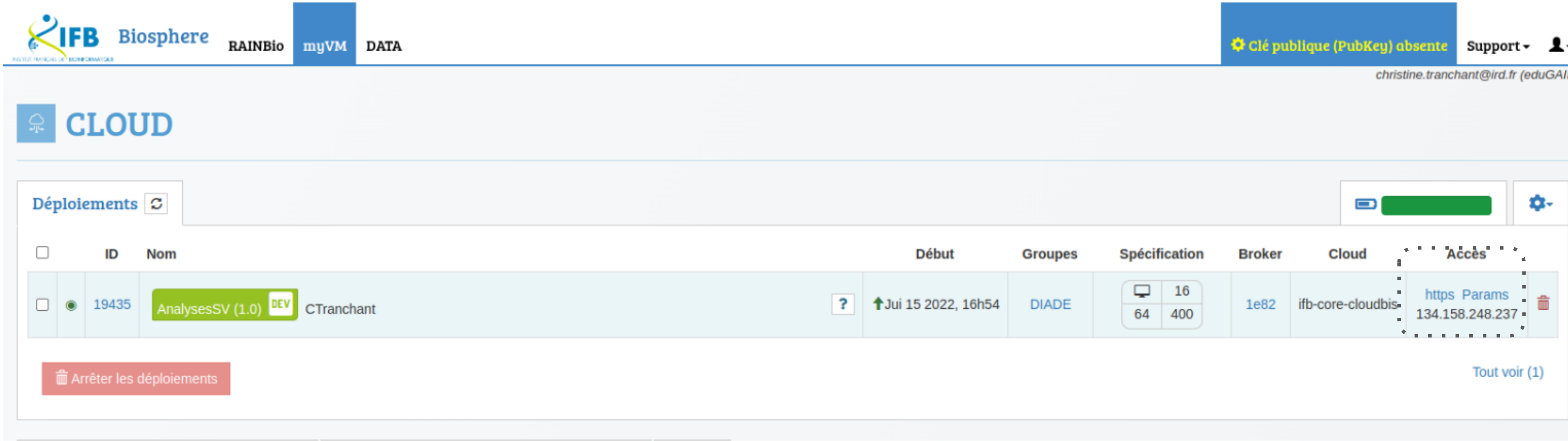
christine.tranchant@ird.fr (eduGAI

ID	Nom		Début	Groupes	Spécification	Broker	Cloud	Accès
19435	AnalysesSV (1.0) DEV CTranchant	?	Jui 15 2022, 16h54	DIADE	16 64 400	1e82	ifb-core-cloudbis	https Params 134.158.248.237

South Green bioinformatics platform

Let’s run your vm through the cloud

get the url… link “https”



IFB Biosphere | RAINBio | myVM | DATA | Clé publique (PubKey) absente | Support

christine.tranchant@ird.fr (eduGAI

CLOUD

Déploiements

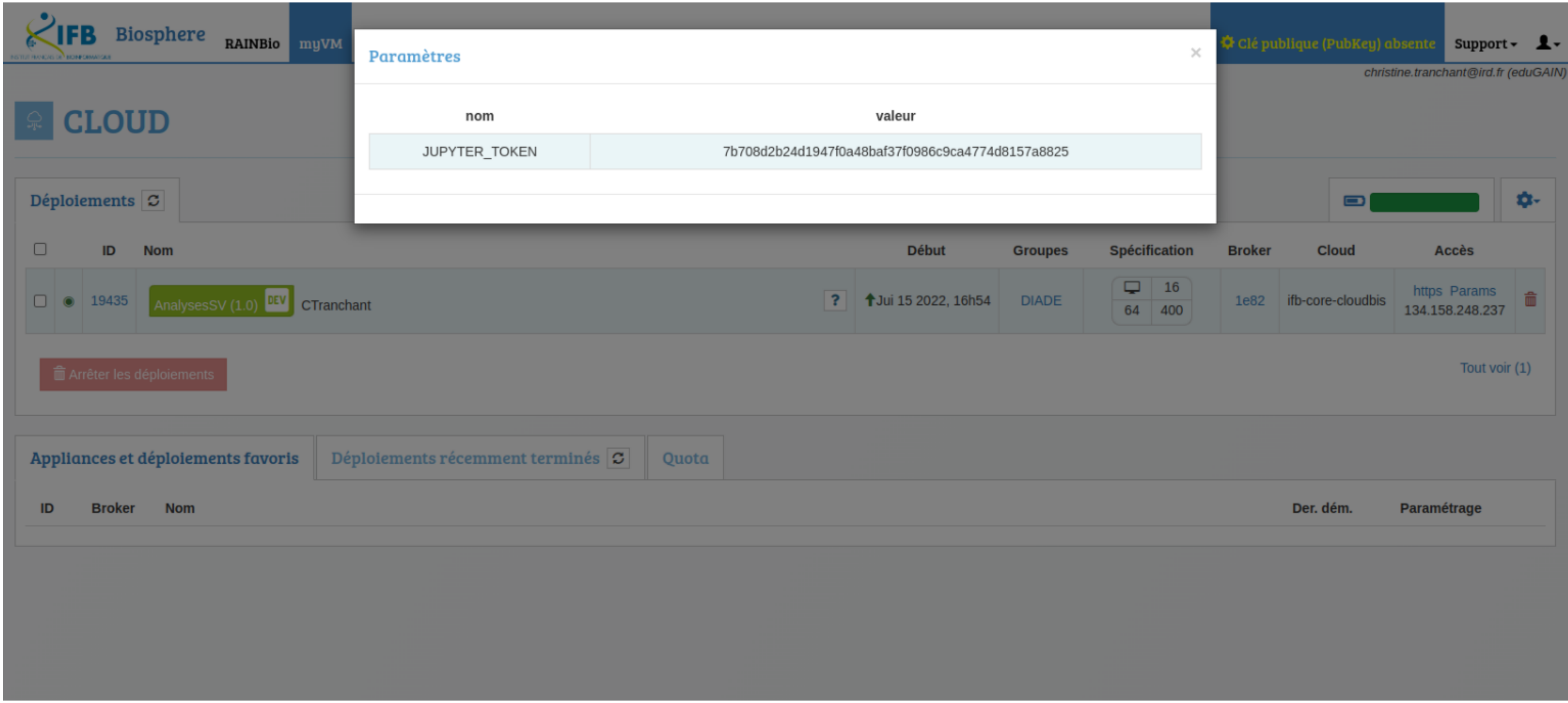
□		ID	Nom		Début	Groupes	Spécification	Broker	Cloud	Accès
□	●	19435	AnalysesSV (1.0) DEV CTranchant	?	Jui 15 2022, 16h54	DIADE	16 64 400	1e82	ifb-core-cloudbis	https Params 134.158.248.237

Arrêter les déploiements

Tout voir (1)

Let’s run our vm through the cloud

Get the token identifiant… link “Params”



IFB Biosphere
RAINBio
myVM
Clé publique (PubKey) absente
Support
christine.tranchant@ird.fr (eduGAIN)
CLOUD
Déploiements
Arrêter les déploiements
Tout voir (1)
Appliances et déploiements favoris
Déploiements récemment terminés
Quota

Paramètres

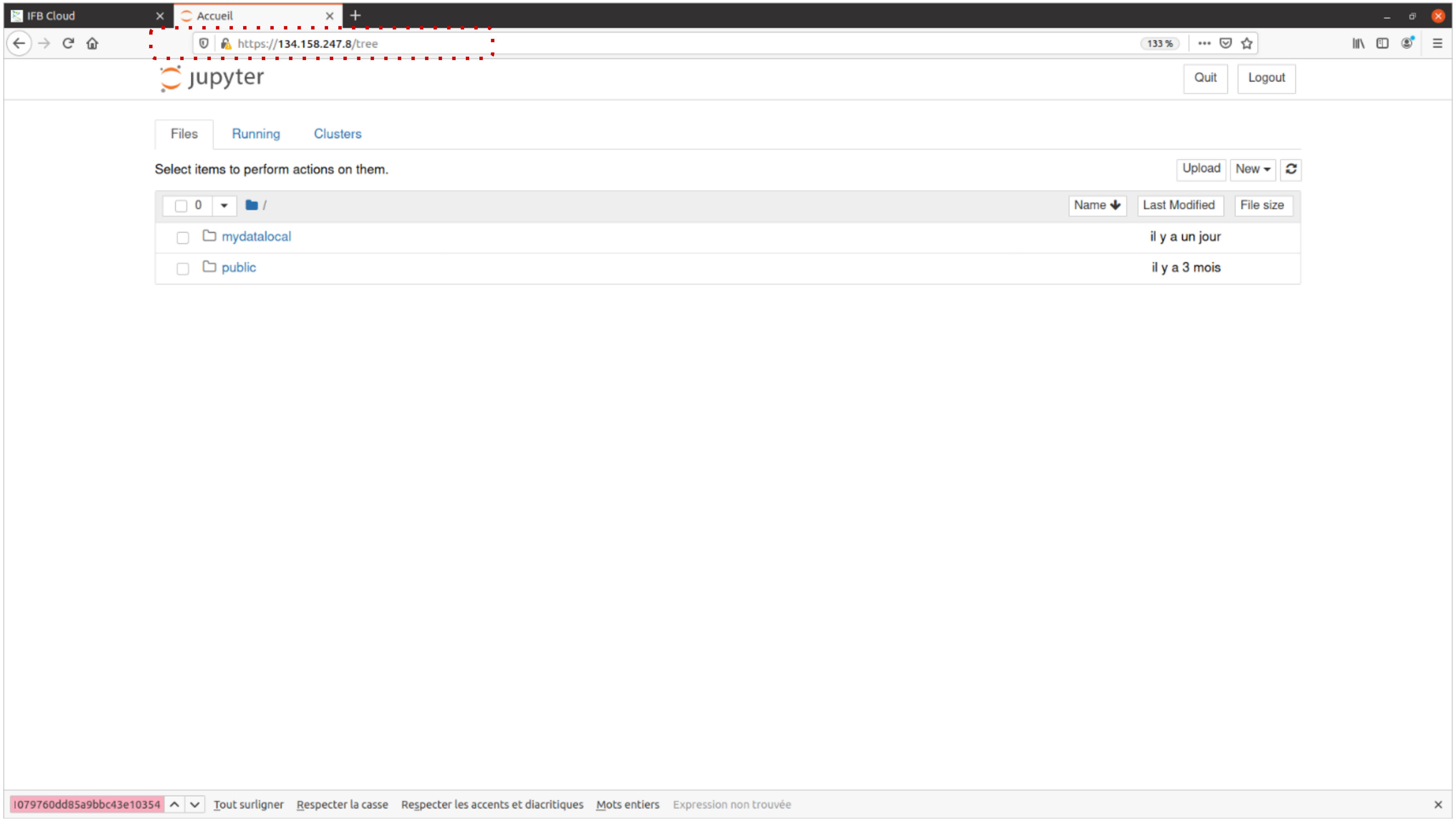
nom	valeur
JUPYTER_TOKEN	7b708d2b24d1947f0a48baf37f0986c9ca4774d8157a8825

☐		ID	Nom		Début	Groupes	Spécification	Broker	Cloud	Accès
☐	●	19435	AnalysesSV (1.0) DEV CTranchant	?	Jui 15 2022, 16h54	DIADE	16 64 400	1e82	ifb-core-cloudbis	https Params 134.158.248.237

ID	Broker	Nom	Der. dém.	Paramétrage

Let’s run our vm through the cloud

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



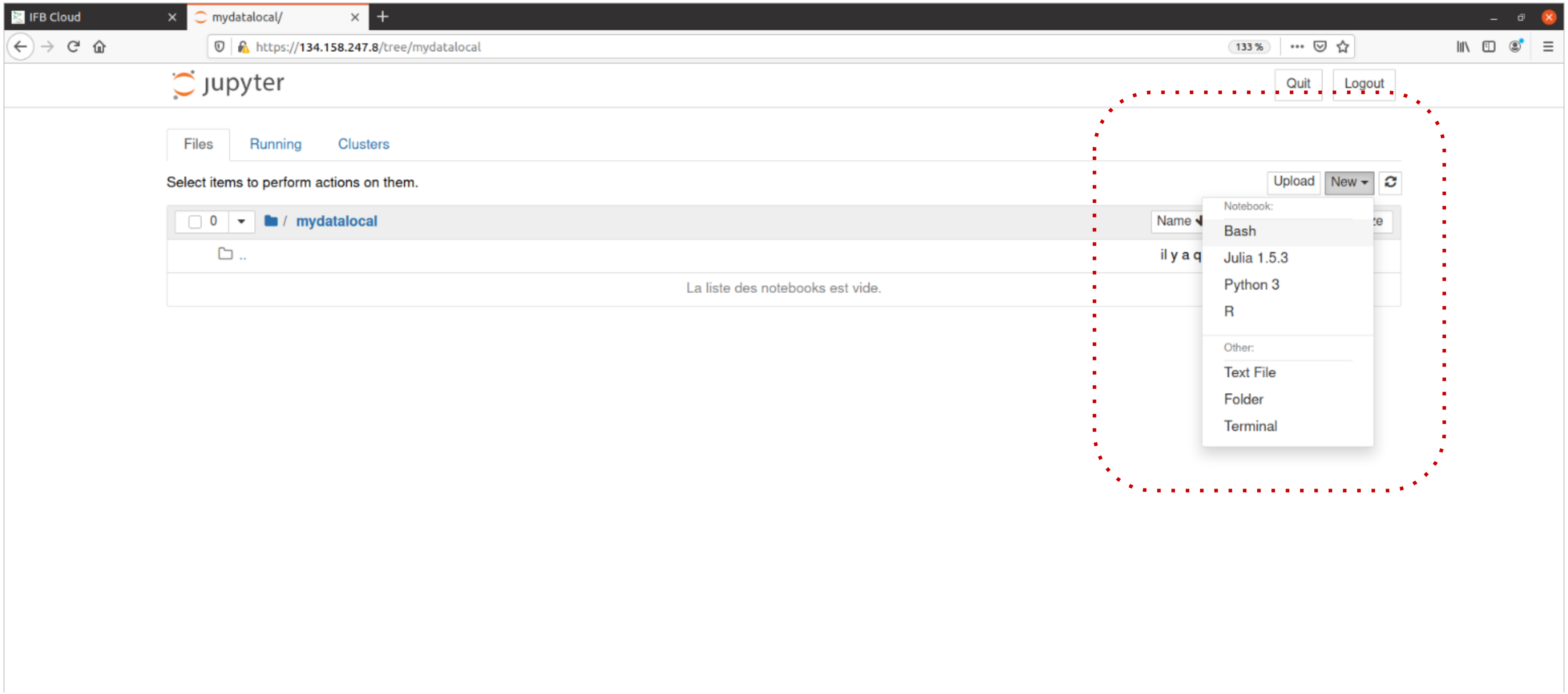
Select items to perform actions on them.

Name	Last Modified
mydatalocal	il y a un jour
public	il y a 3 mois

l079760dd85a9bbc43e10354

Create your first jupyter book

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

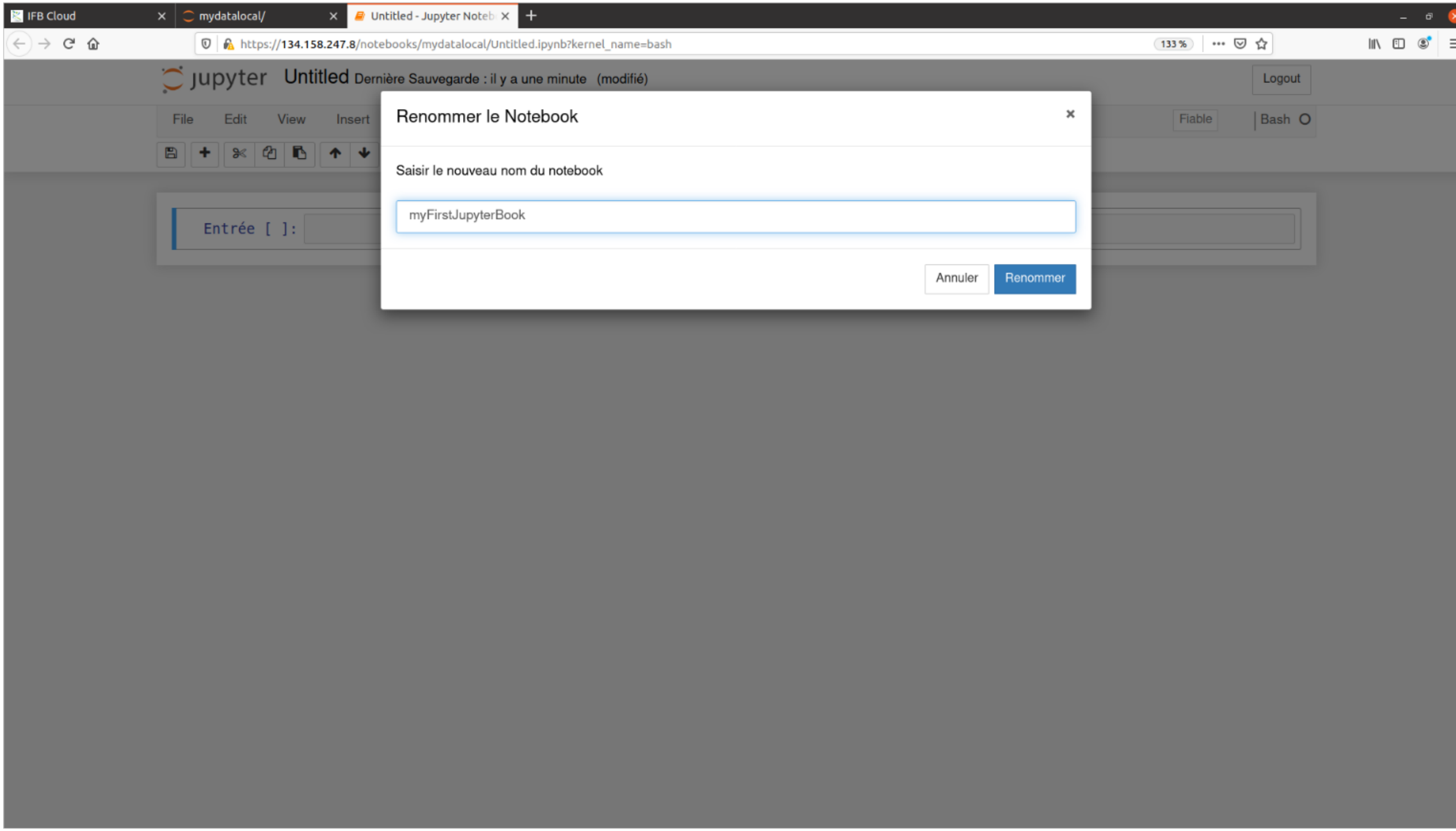


Select items to perform actions on them.

La liste des notebooks est vide.

Rename your first jupyter book

myFirstJupyterBook



IFB Cloud
mydatalocal/
Untitled - Jupyter Noteb
https://134.158.247.8/notebooks/mydatalocal/Untitled.ipynb?kernel_name=bash
133 %
jupyter
Untitled Dernière Sauvegarde : il y a une minute (modifié)
Logout
File
Edit
View
Insert
Fiable
Bash
Entrée [ ]:

Renommer le Notebook

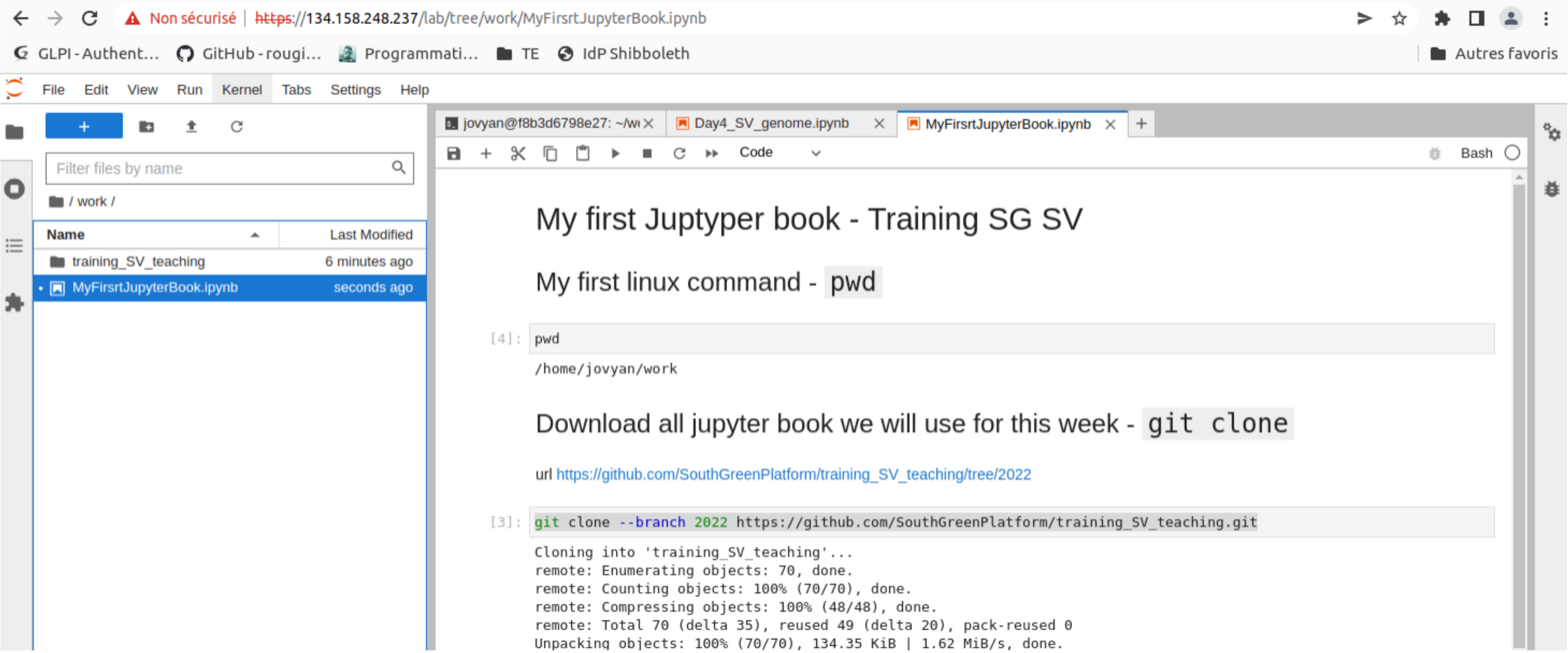
Saisir le nouveau nom du notebook

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



/ work /

Name	Last Modified
training_SV_teaching	6 minutes ago
MyFirsrtJupyterBook.ipynb	seconds ago

My first Juptyper book - Training SG SV

My first linux command - `pwd`

```
[4]: pwd
/home/jovyan/work
```

Download all jupyter book we will use for this week - `git clone`

url https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022

```
[3]: git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git
Cloning into 'training_SV_teaching'...
remote: Enumerating objects: 70, done.
remote: Counting objects: 100% (70/70), done.
remote: Compressing objects: 100% (48/48), done.
remote: Total 70 (delta 35), reused 49 (delta 20), pack-reused 0
Unpacking objects: 100% (70/70), 134.35 KiB | 1.62 MiB/s, done.
```

Nécessité de la pratique et de l’expérience

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

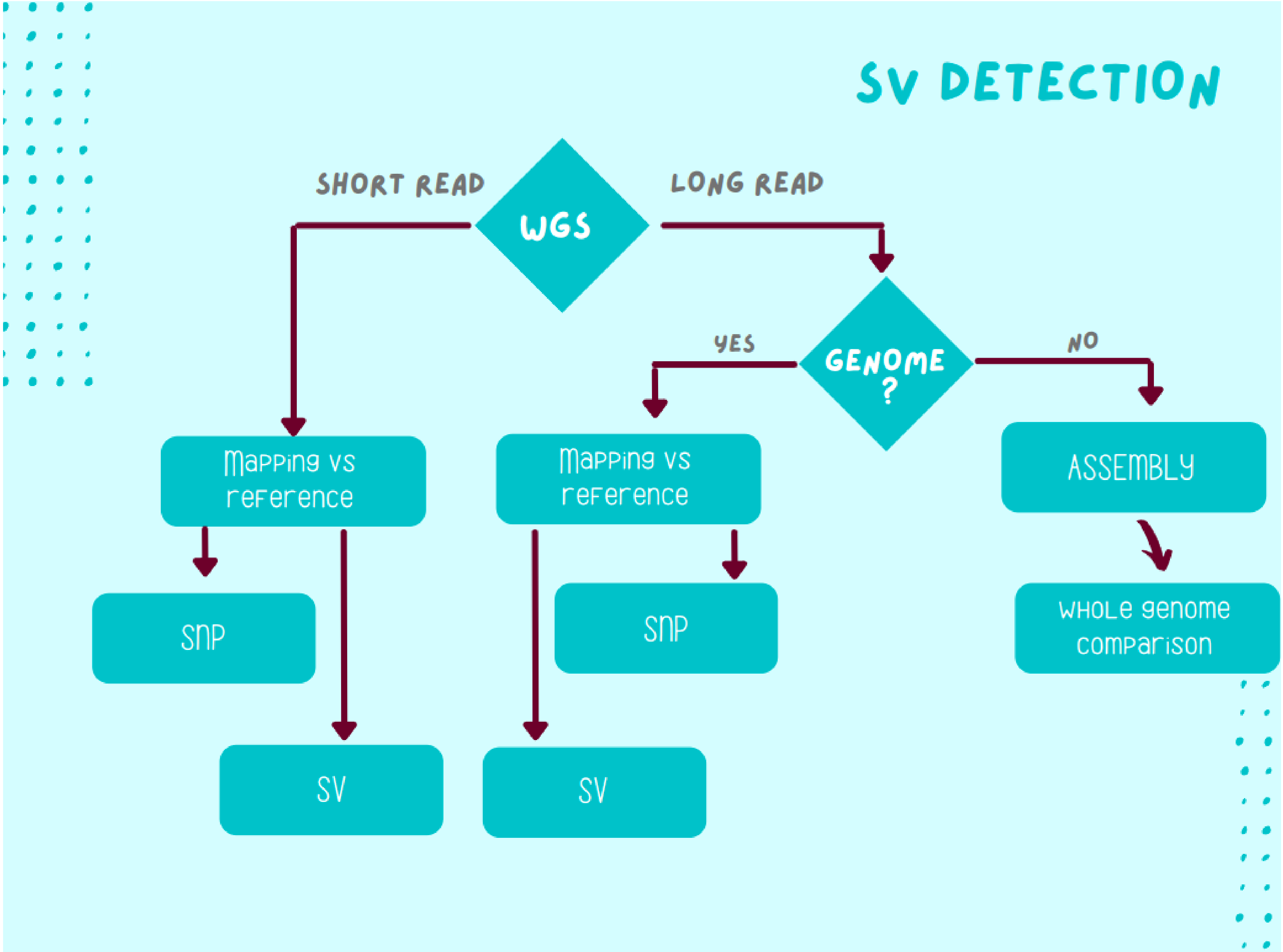
Détection de variants à partir de données de séquençage short & long reads

Alexis Dereeper - UMR PHIM

Julie Orjuela - UMR DIADE

Christine Tranchant-Dubreuil - UMR DIADE

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



SV DETECTION

Objectifs

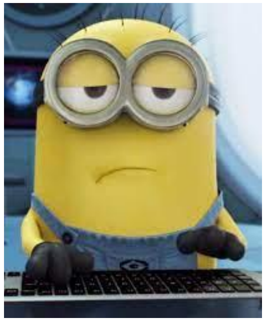
Détecter des variants (SNP, variants structuraux) à partir de données de séquençage short et long reads.

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***

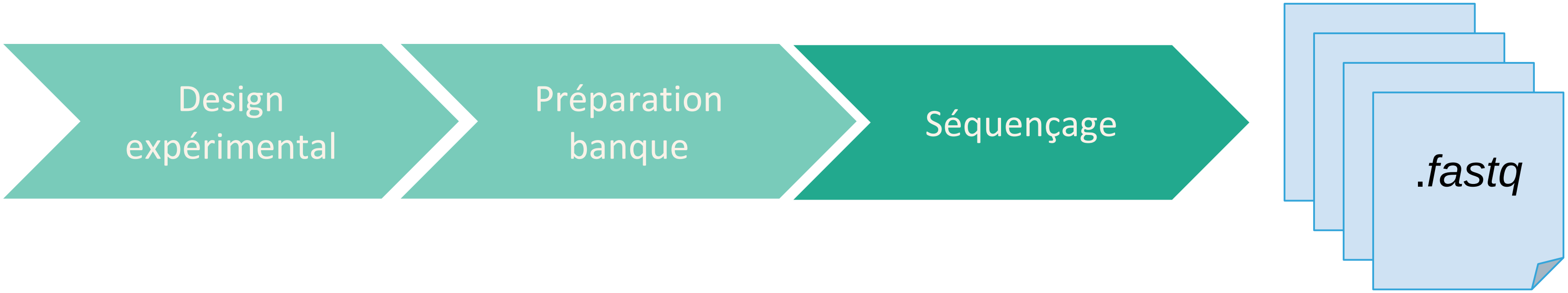
Avec jupyter book : lancer les commandes + analyser les résultats

=> Avoir un plan de bataille opérationnel



RAW SEQUENCING DATA

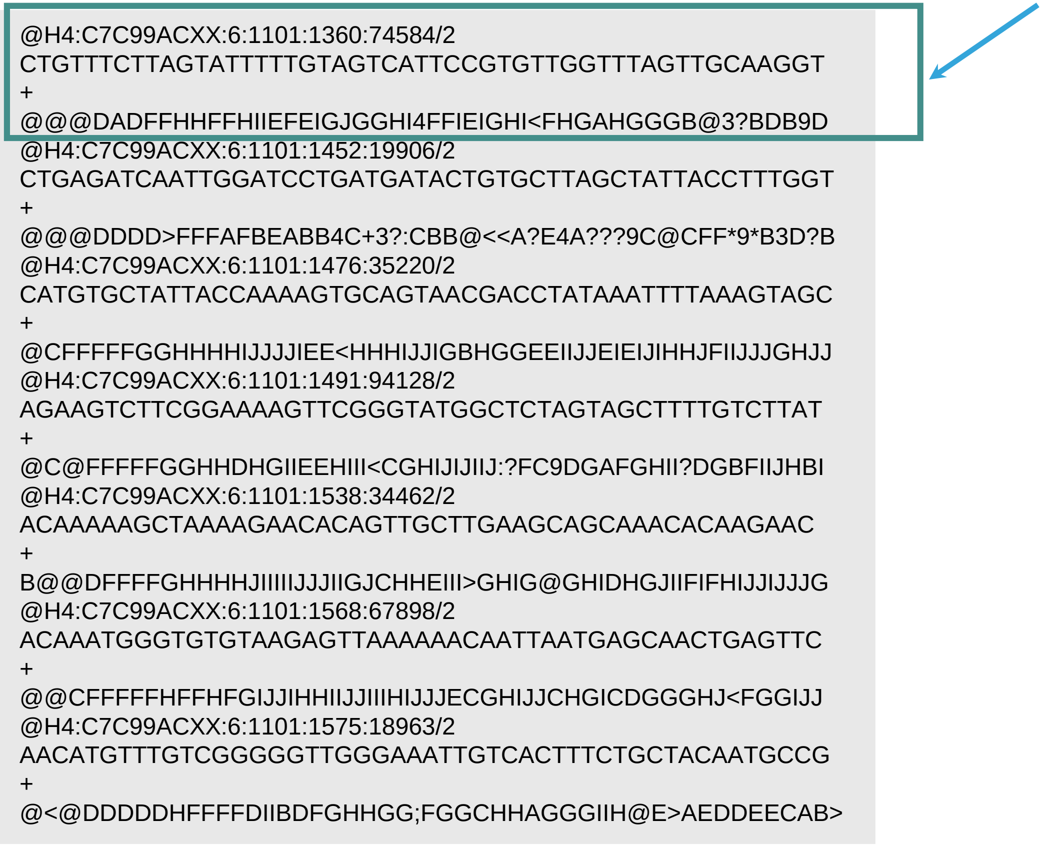
OVERVIEW OF DNA SEQUENCING PROJECT



- 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 ?

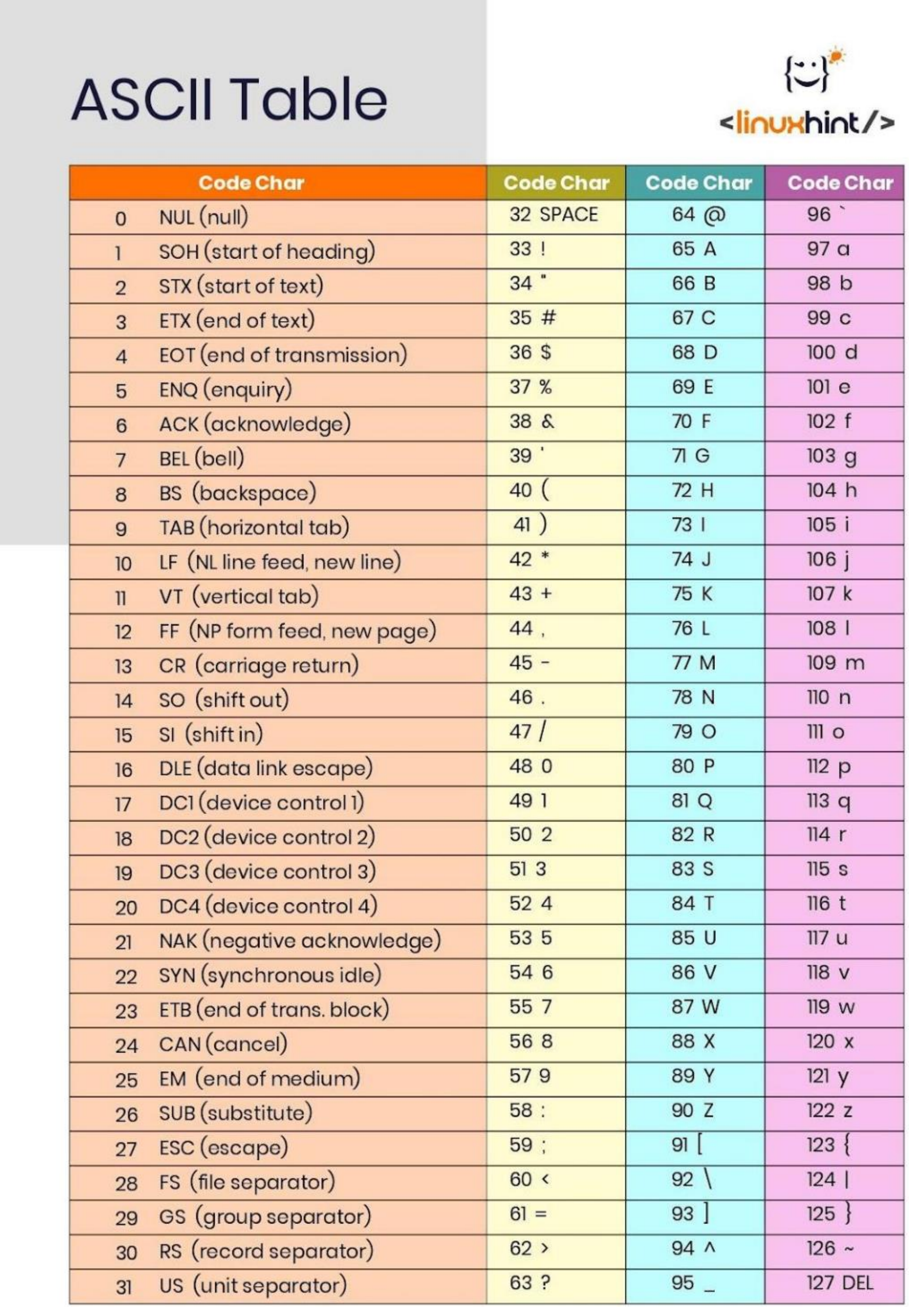
```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTTGTAGTCATTCCGTGTTGGTTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTTGGT
+
@@@DDDD>FFFAFBEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
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



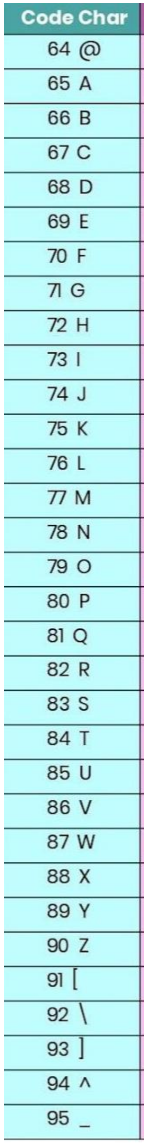
ASCII Table

<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

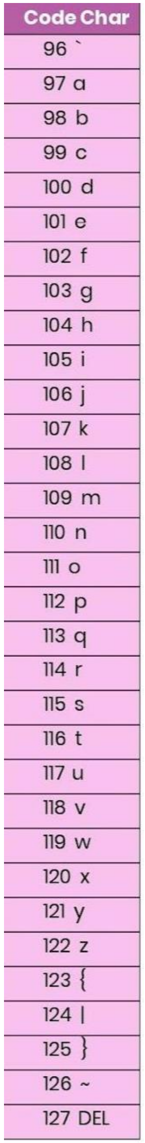
How to code quality score for each base with one letter ?

Code ASCII



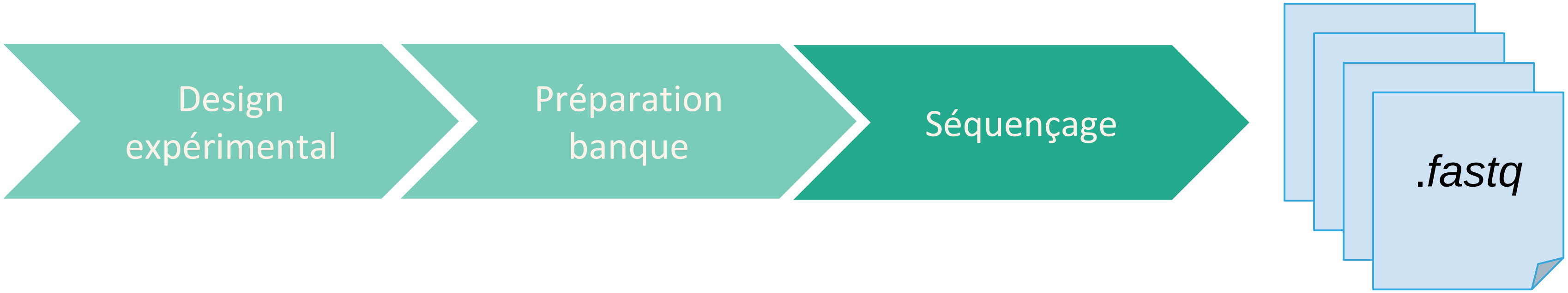
Code	Char
64	@
65	A
66	B
67	C
68	D
69	E
70	F
71	G
72	H
73	I
74	J
75	K
76	L
77	M
78	N
79	O
80	P
81	Q
82	R
83	S
84	T
85	U
86	V
87	W
88	X
89	Y
90	Z
91	[
92	\
93	]
94	^
95	_

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

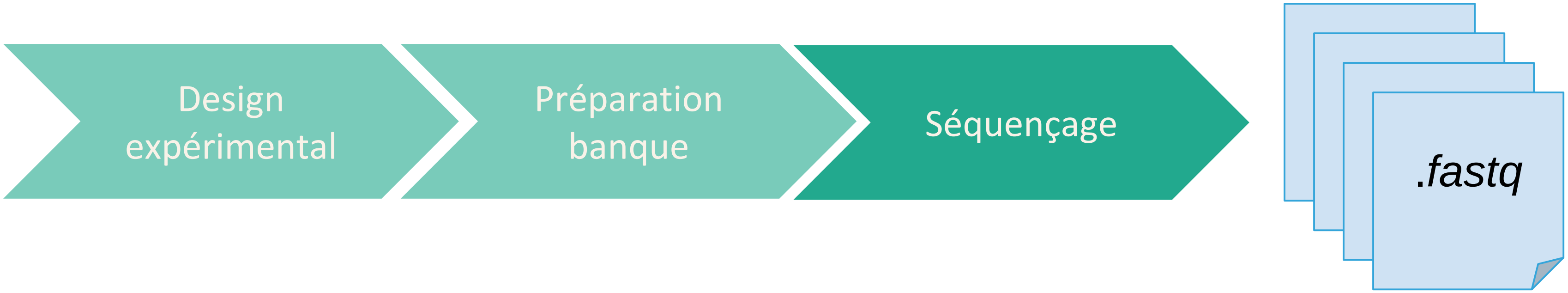
OVERVIEW OF DNA SEQUENCING PROJECT



- Statistics
- Sequencing quality ? Adaptators ?

  Contaminants ?

OVERVIEW OF DNA SEQUENCING PROJECT



- Statistics
- Sequencing quality ? Adaptators ?

  Contaminants ?

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

fastqc

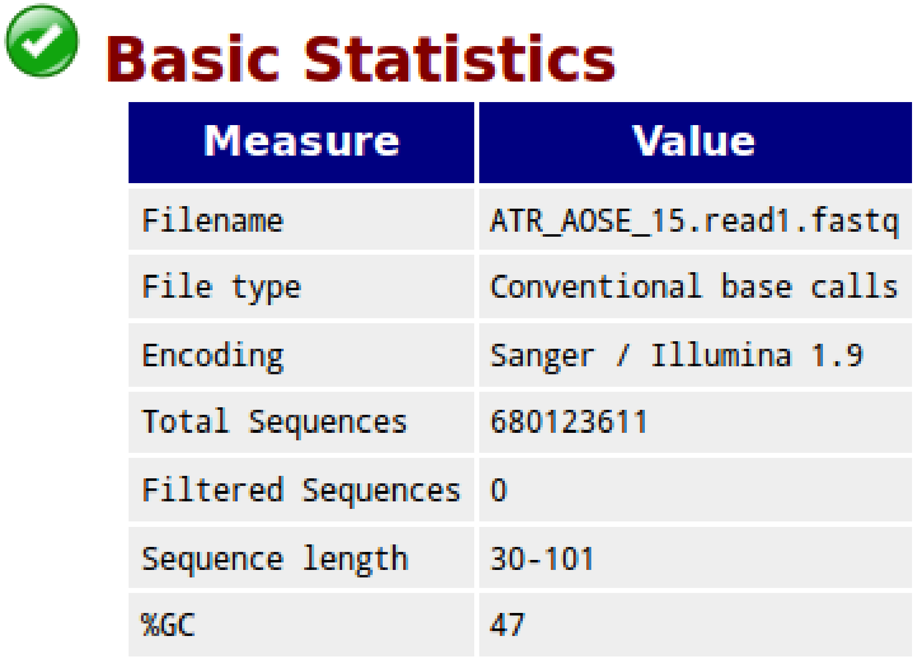
fastqc to get some basic statistics and to do some quality control checks

```
# 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

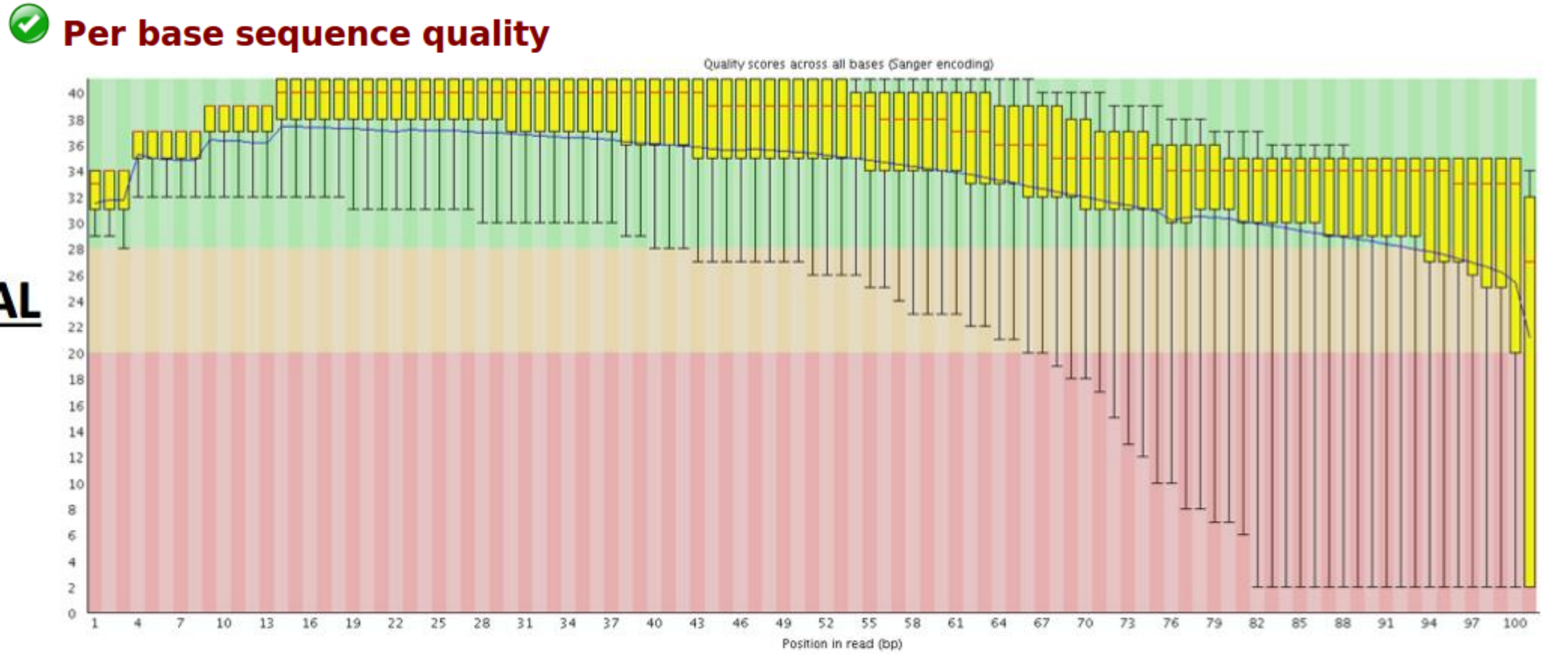
Measure	Value
Filename	ATR_AOSE_15.read1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	680123611
Filtered Sequences	0
Sequence length	30-101
%GC	47

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

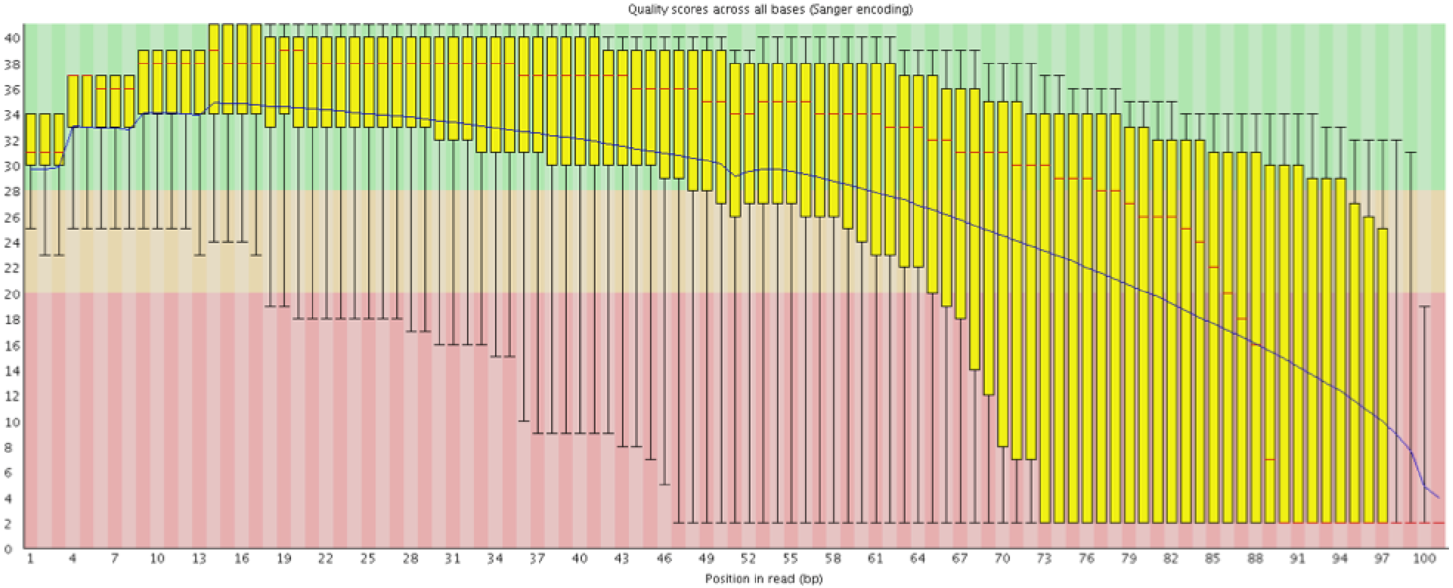
GOOD/NORMAL LANE



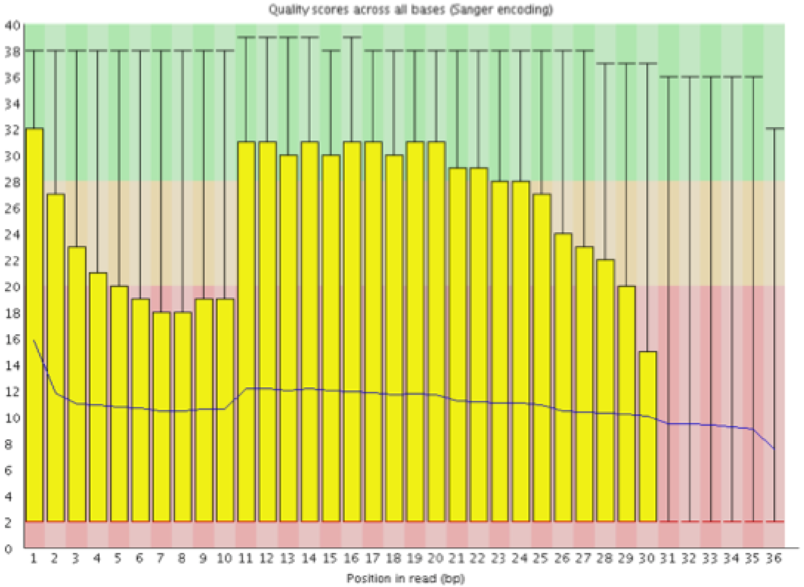
✓ Per base sequence quality

FastQC : Per base sequence quality

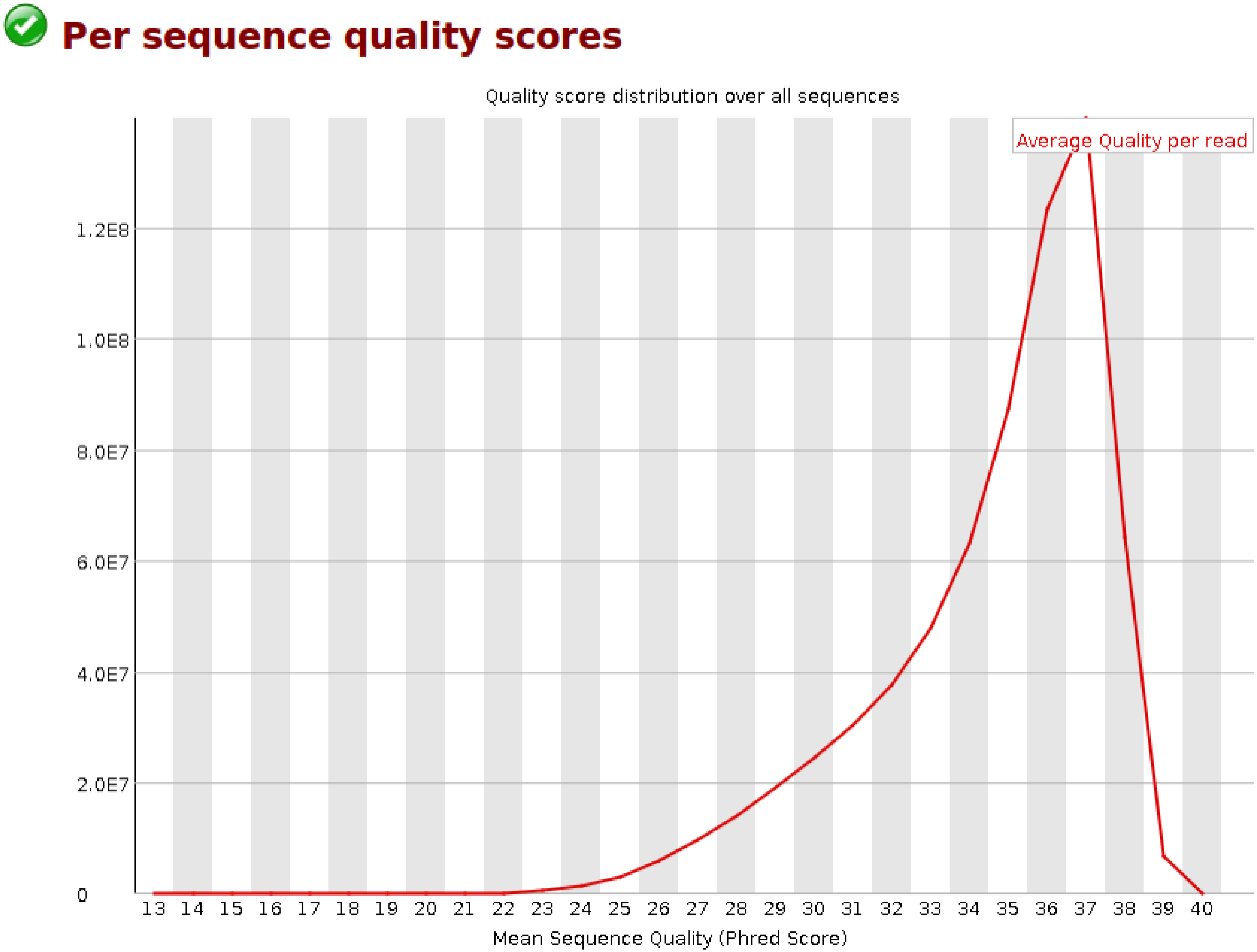
SALVAGEABLE LANE



FAILED LANE



FastQC: Per sequence quality scores



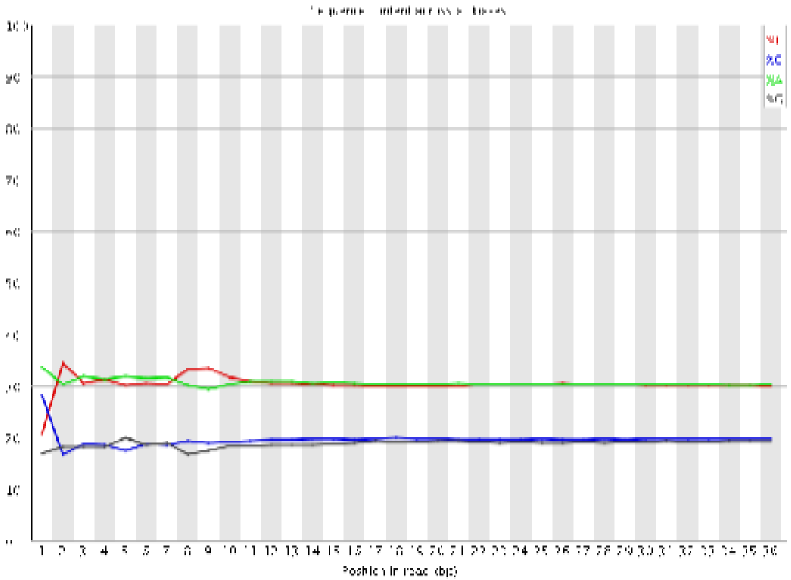
✓ Per sequence quality scores

FastQC: Per base sequence content

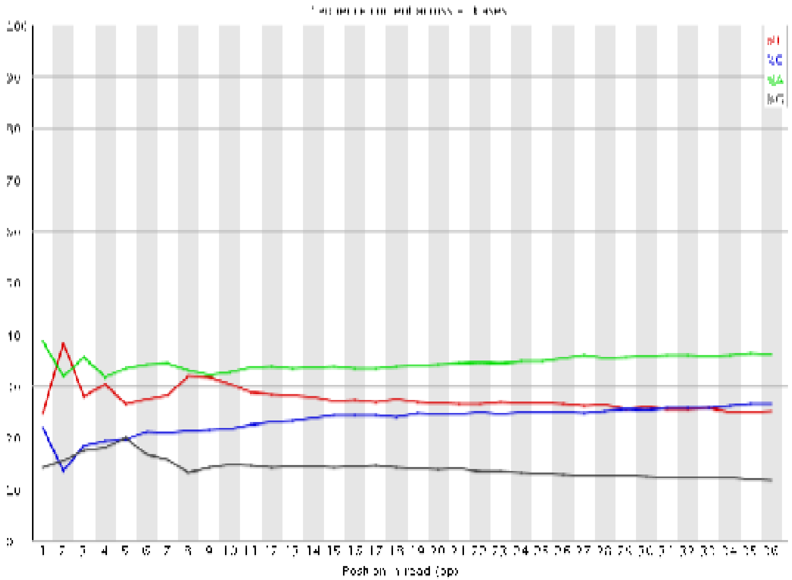
This plot shows the nucleotide distribution per read position for all reads in a lane.

- 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



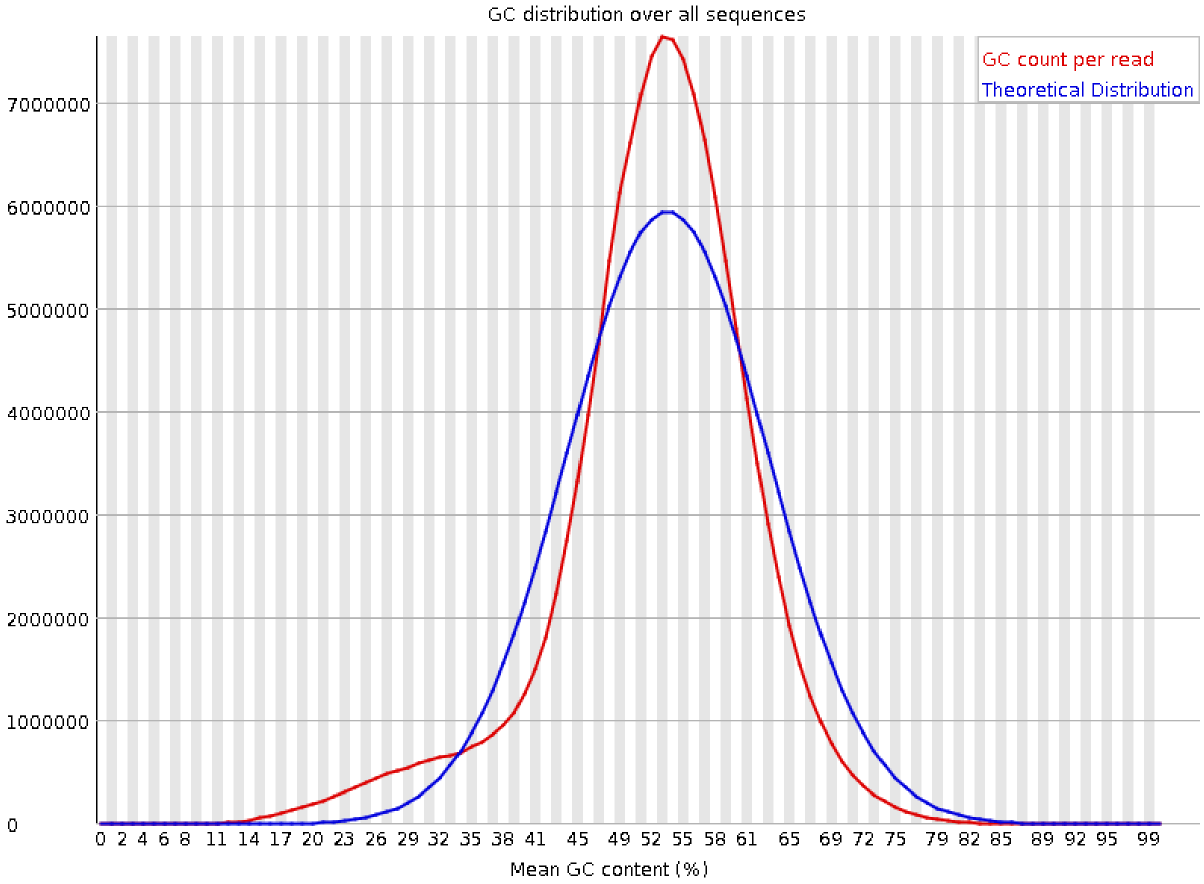
BAD LANE



Can this be fixed? **No.**

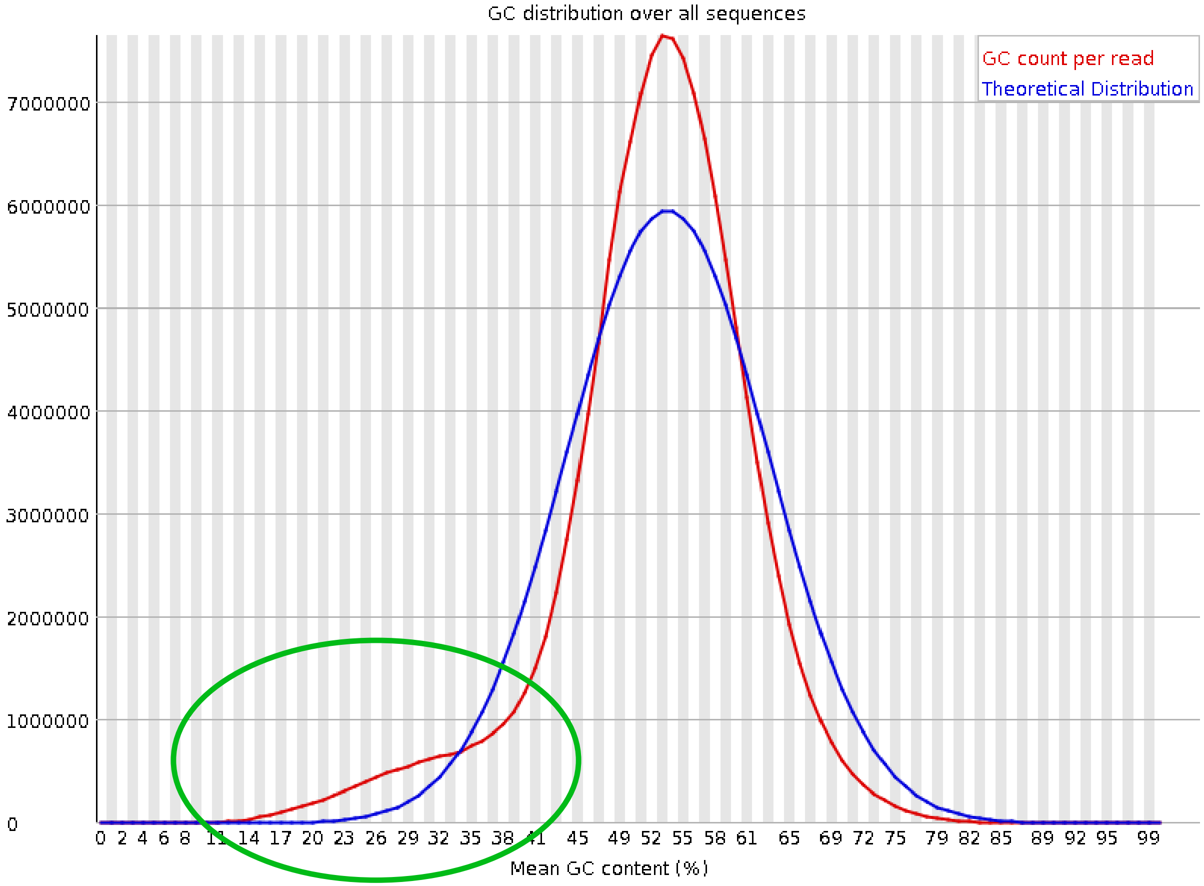
FastQC: Per sequence GC content

- A contamination ?



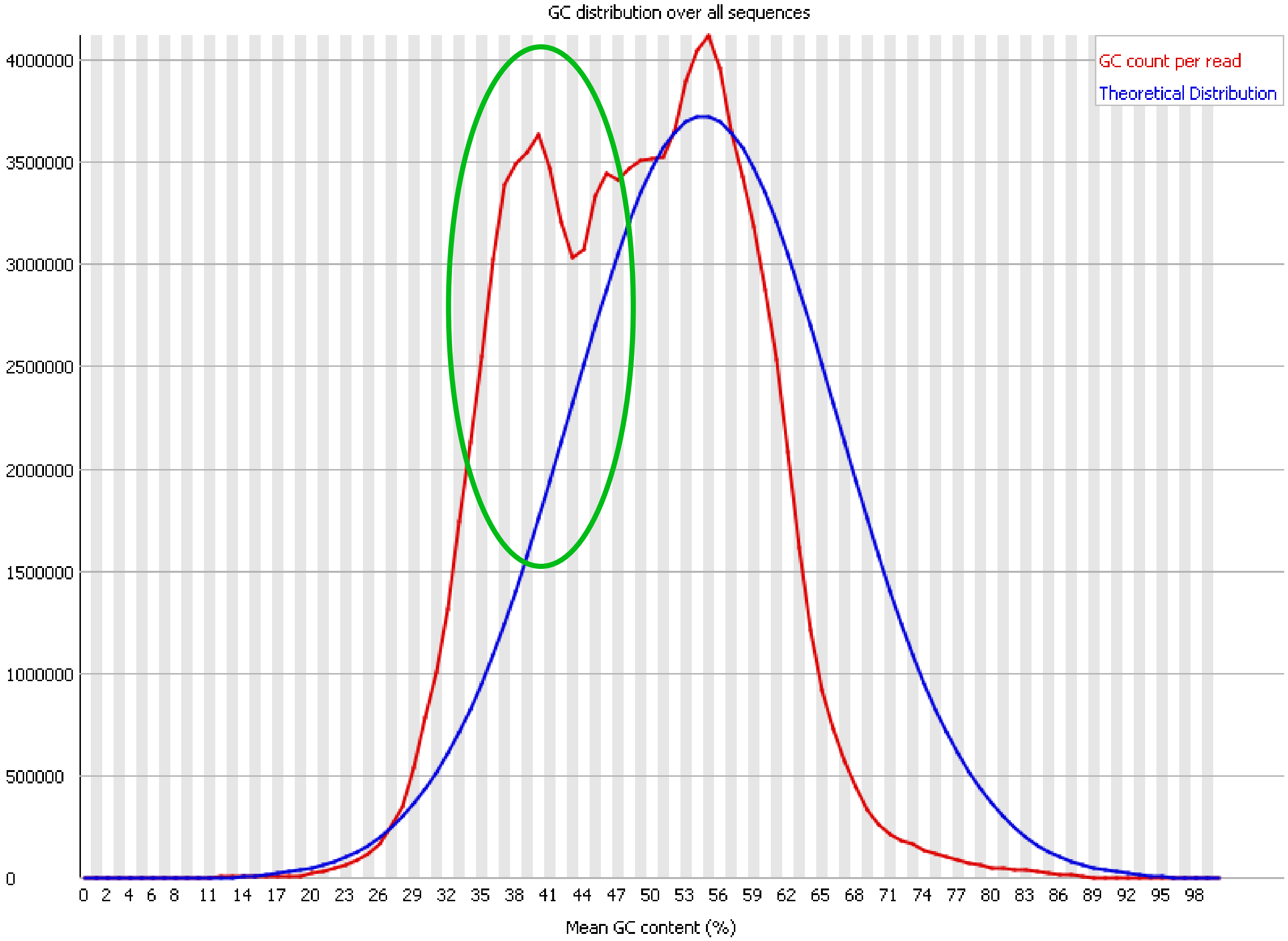
FastQC: Per sequence GC content

- A contamination ?

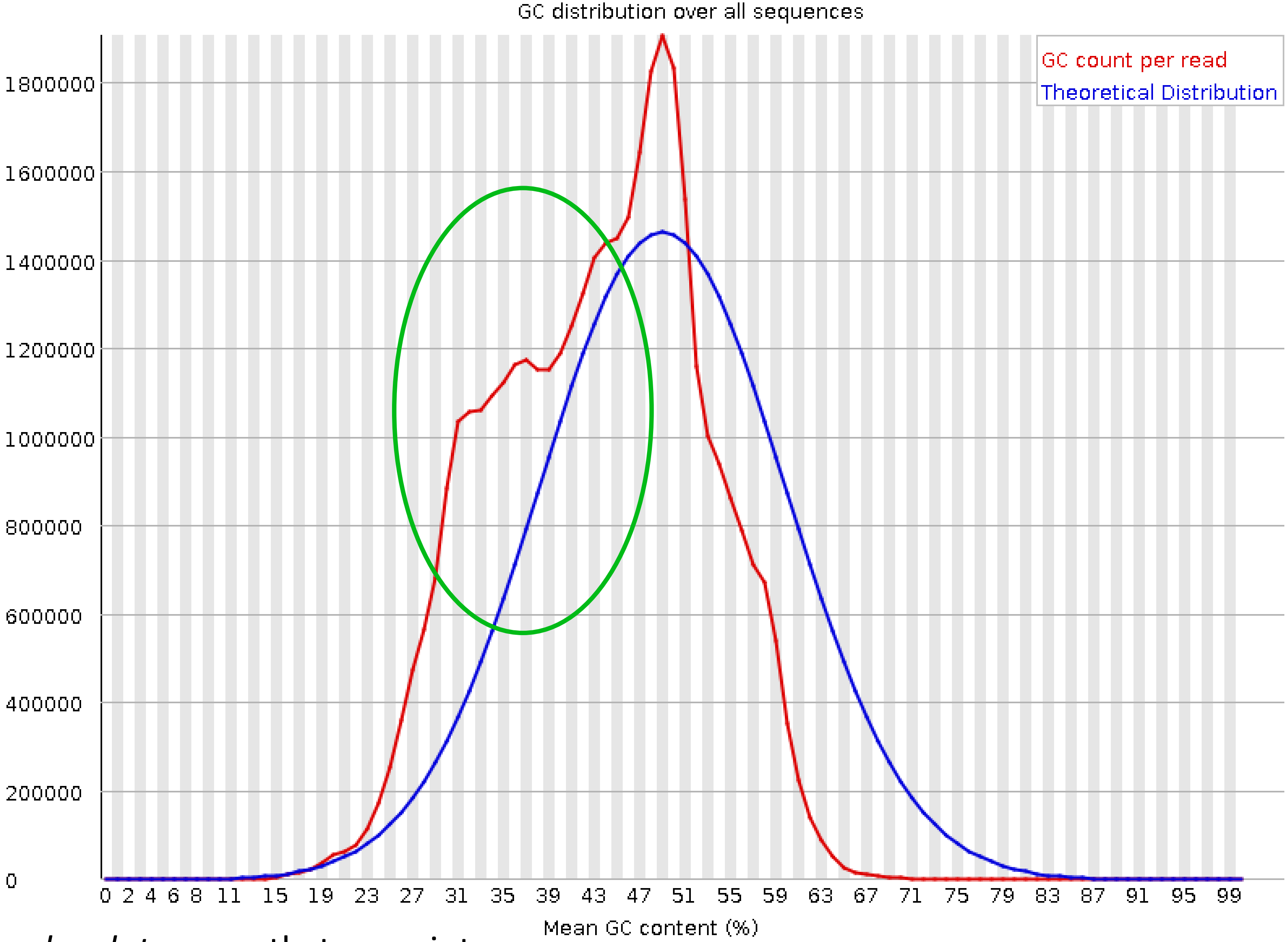


<u>Can this be fixed ?</u> Maybe…

FastQC: Per sequence GC content



Third-party contamination : detection



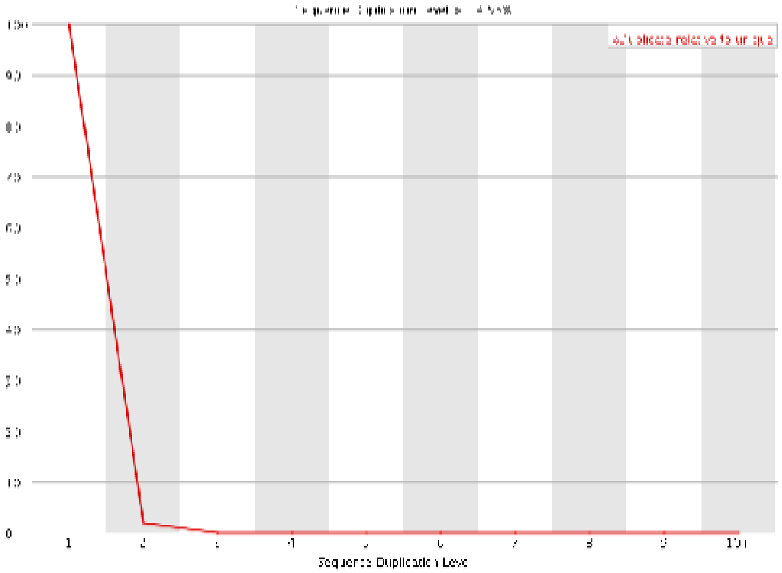
Sabellaria alveolata : mantle transcriptome

FastQC: Sequence Duplication Levels

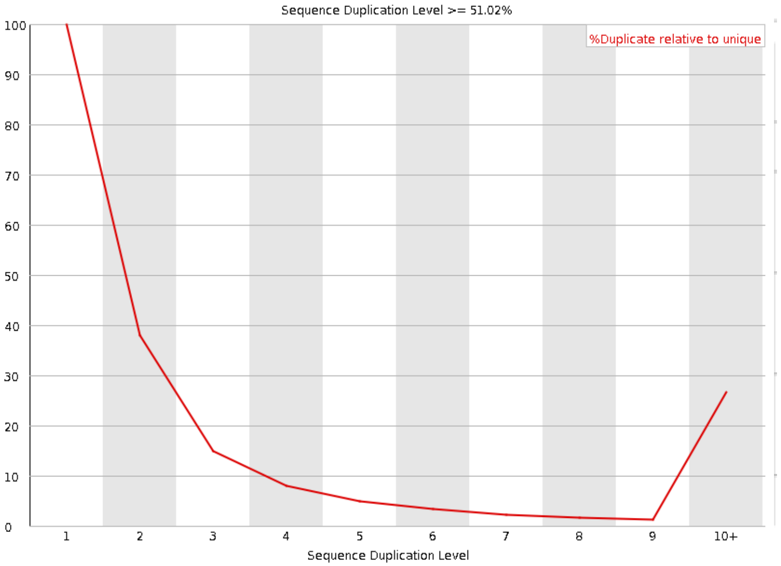
This plot shows the degree of duplication for a subset of reads in a lane.

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

GOOD LANE

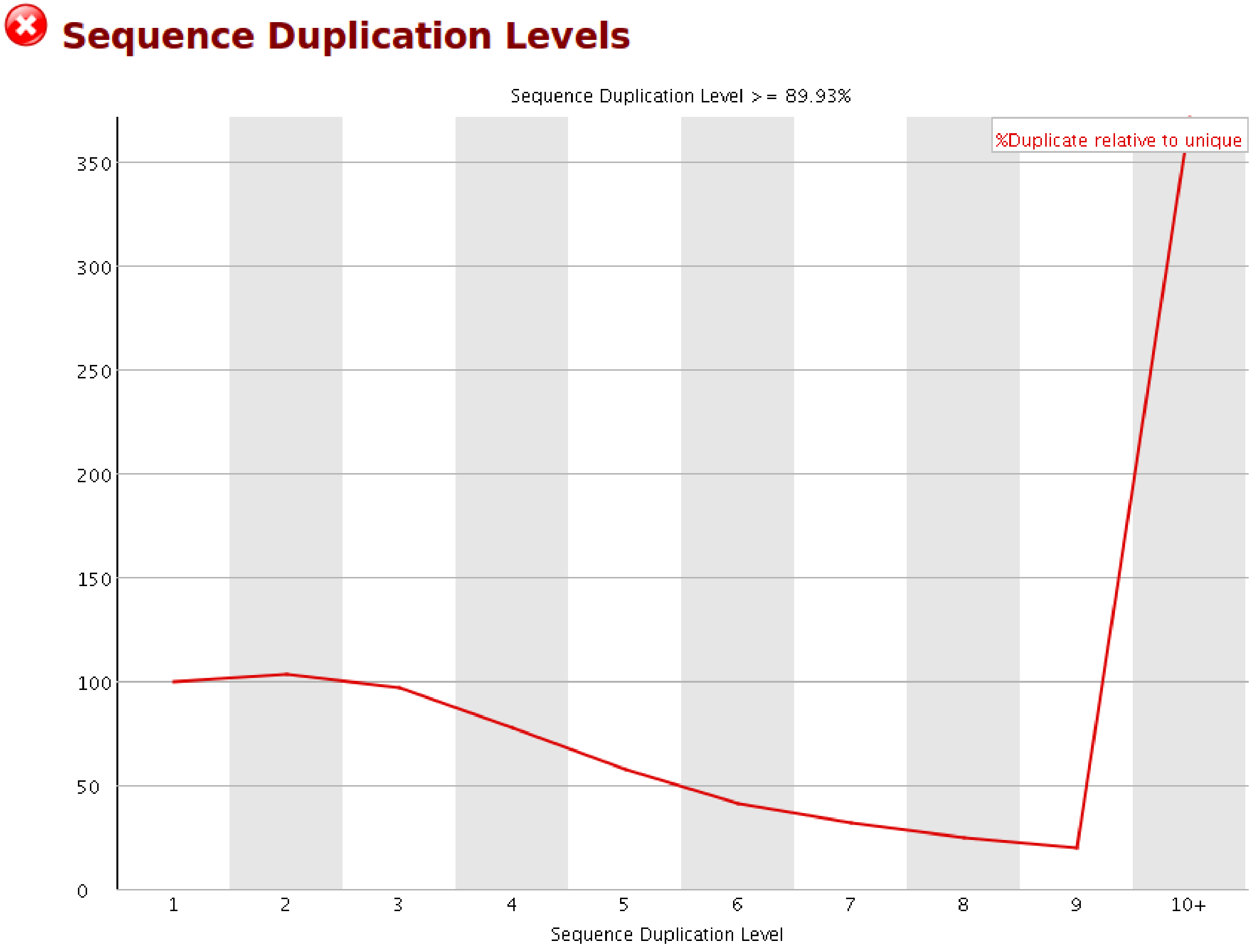


BAD LANE



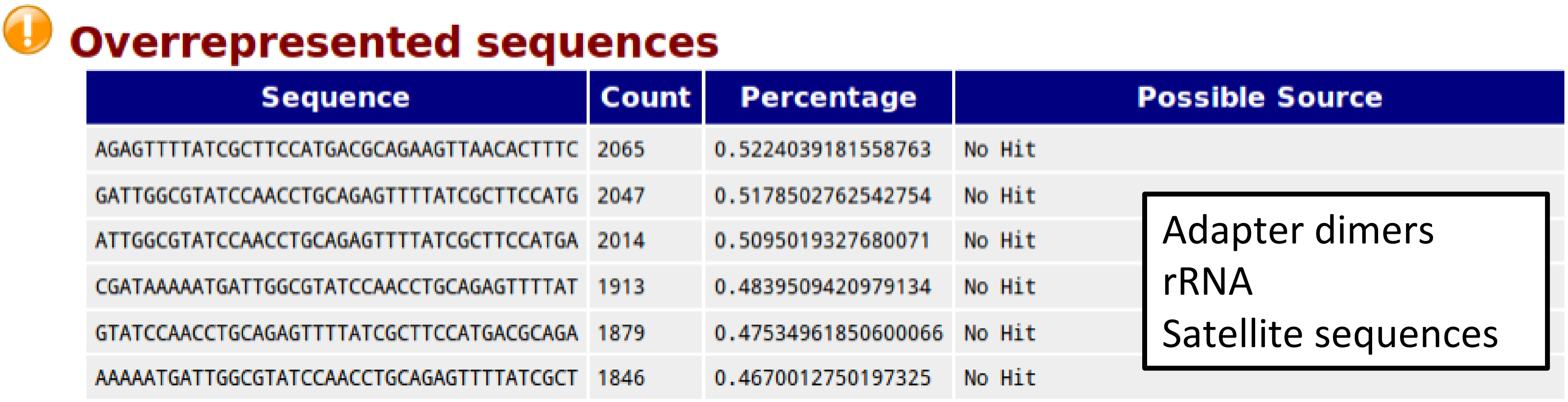
Can this be fixed? **Maybe.**

FastQC: Sequence Duplication Levels



Can this be fixed? **Hem…**

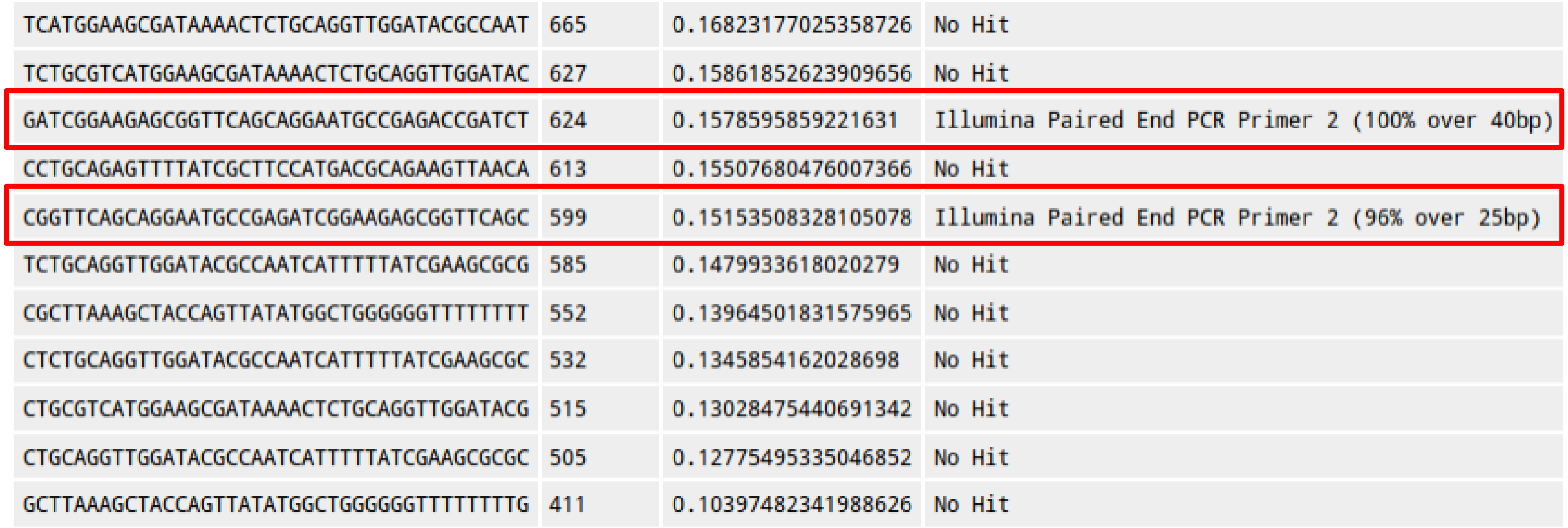
FastQC: Overrepresented sequences



Overrepresented sequences

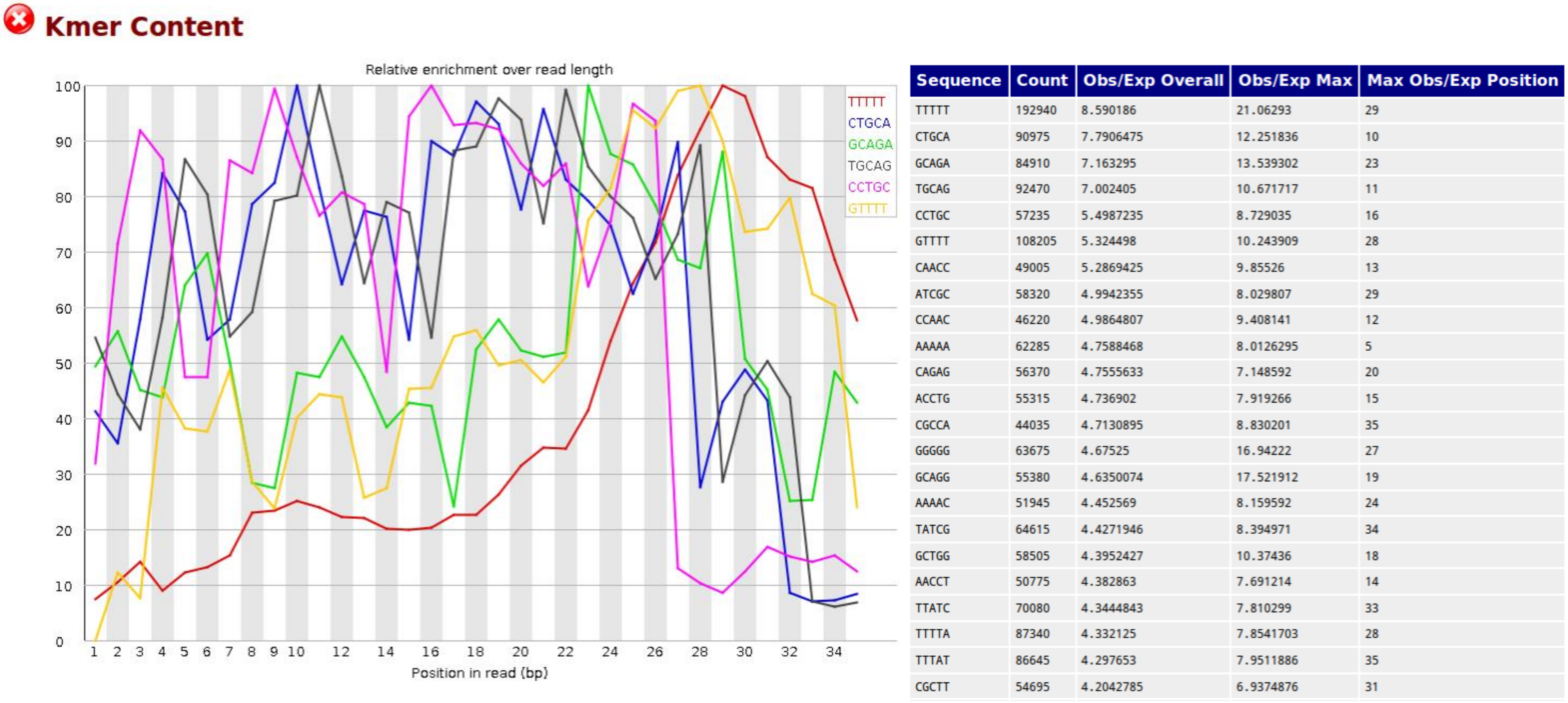
Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit

Adapter dimers
rRNA
Satellite sequences



Sequence	Count	Percentage	Possible Source
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

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

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

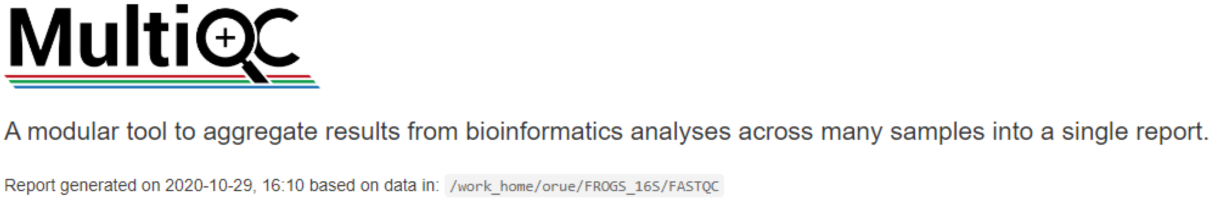
With numerous fastq and fastqc report => use MultiQC

MultiQC

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/



A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2020-10-29, 16:10 based on data in: `/work_home/orue/FROGS_16S/FASTQC`

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