

Projet Burkinabioinfo



Overview - Actions de formation



2018 CERMAS
2019 Institut Pasteur du Congo

48h, 24 personnes

2 x 32h, 25 personnes



Programmes formation et organisation



Programme & Logistique

	Débutant	Veteran SNP / diversité 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



 **AHONON**
Awovi Selom



 **Gbekley**
Efui Holaly



 **TUINA** Séverin



 **DOSSIM**
Sika



 **BA** Aminata
Hamidou

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



 **TONDE**
Ignace



 **SIRIMA**
Constant



 **BADOUM** Emilie
Salimata



 **ZOURE** Abdou
Azaque



 **ADAMOU IBRAHIM**
Maman Laouali

Solenostemon
rotundifolius,
interactions génotype x
environnement , profil
génétique,

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

Microbiome intestinal du moustique
(Illumina) , Gène BRCA (Sanger)

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



 **SANOU** Estèle
Pélagie



 **PALANGA** Essowè

Metagenomique, virus, interaction plante-
parasite, phytopathologie

13 Apprenants

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



OUEDRAOGO
Jacques



DANOU-KODJO
Kodjovi Atassé



SAGNON
Adama



SORY Siedou

Métagénomique-Variabilité
génétique-Phytopathologie

phosphate, solubilisation,
bactéries, champignons

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



NAME Pakyendou
Estel



ZONGO
Saïdou

Epidémiologie; Virus; ADN; CRESS;
Séquençage

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



SAWADOGO
Seydou



LALLOGO P. Doriane
Tatiâna

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

Surveillance participative, maladies
virales, racines et tubercules,
séquençage

8 Apprenants



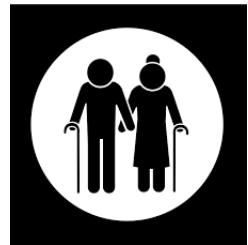
Dominique



Admins : Seydou, Ousmanne Ndomassi

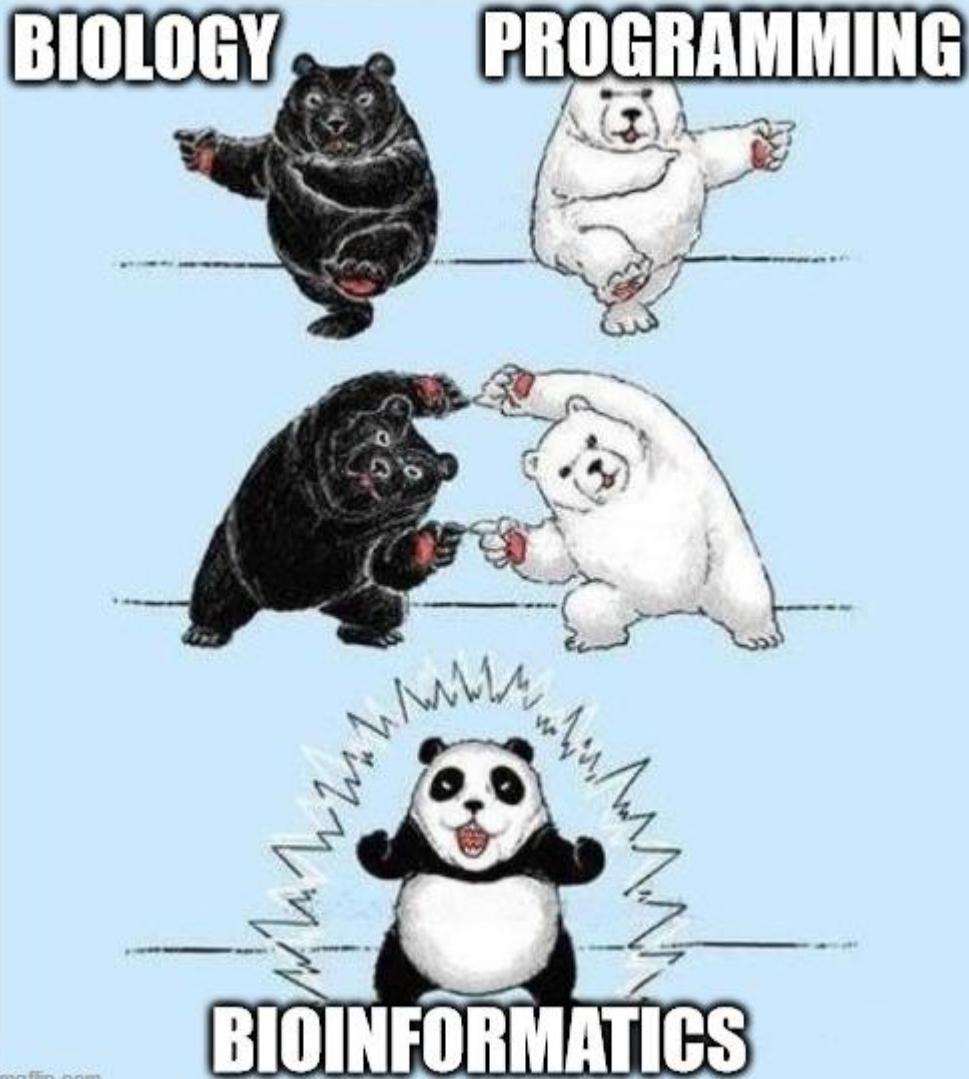
Fidele, romaric, Isidore
Le comité d organisation ... ?
et bien d autres encore.... Neyra?



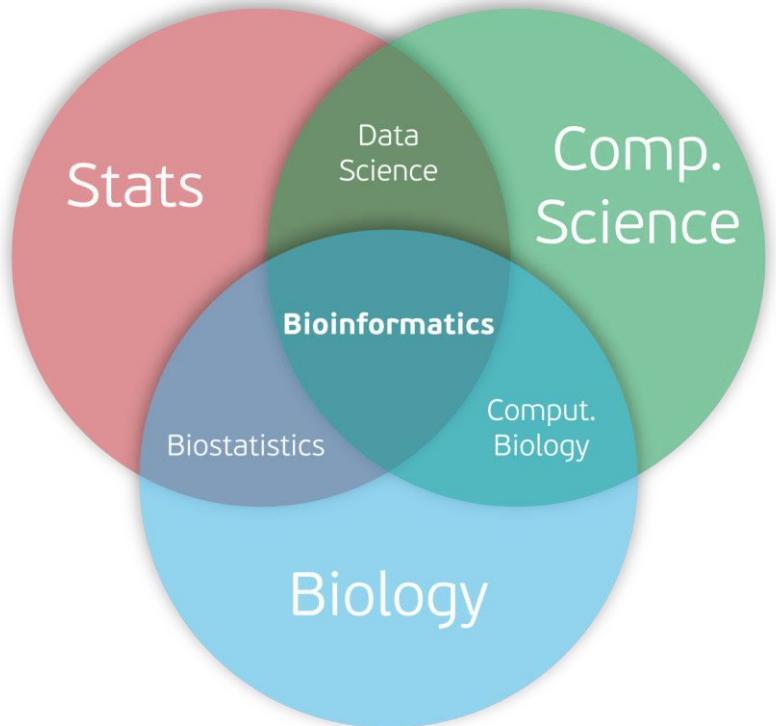


Introduction Bioinformatics & Sequencing

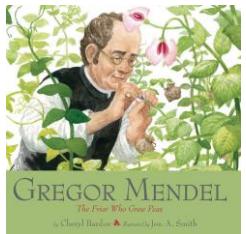
What is the bioinformatics ?



A interdisciplinary science



De la génétique à la bioinformatique...



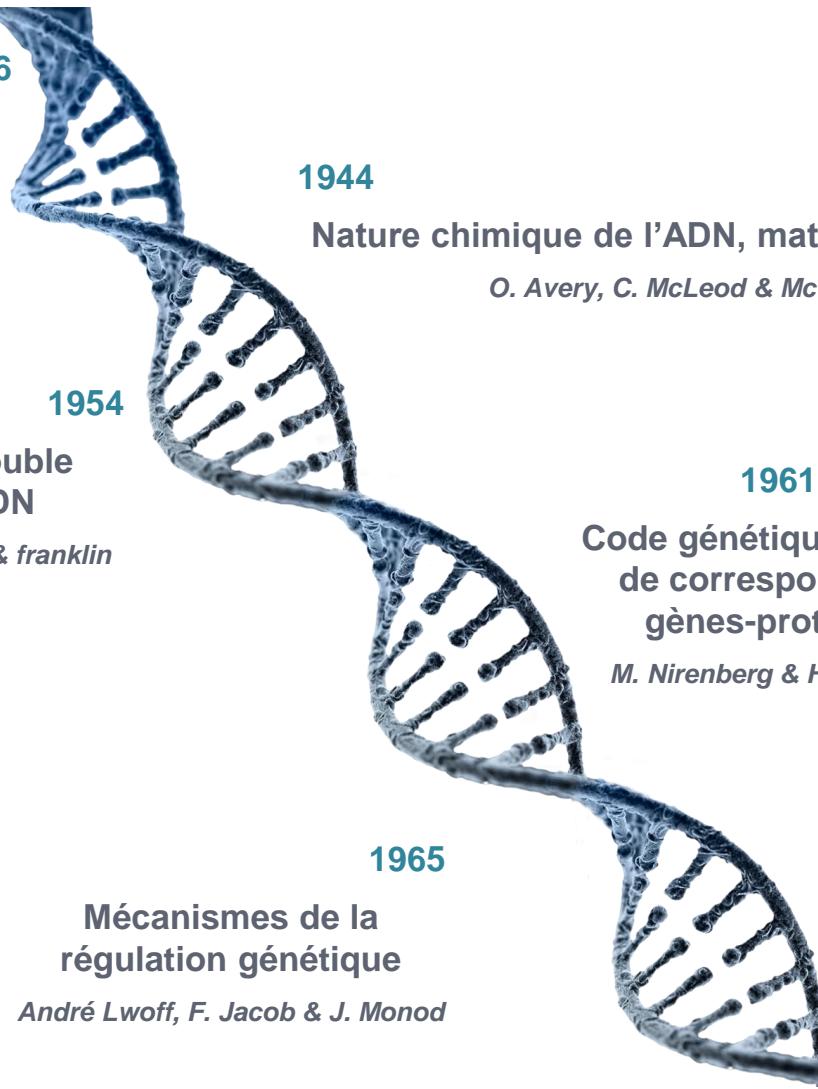
1866
Lois de l'hérédité



1954
Structure en double hélice de l'ADN
J. Watson & F. Cricks & franklin



1965
Mécanismes de la régulation génétique
André Lwoff, F. Jacob & J. Monod



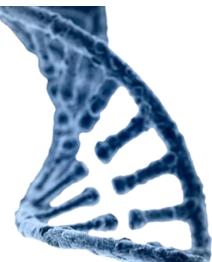
1944
Nature chimique de l'ADN, matériel héréditaire
O. Avery, C. McLeod & McCarthy



De la génétique à la bioinformatique...

1970

Algo Alignement
global de séquence
Needman, & Wunsh



1972
8008

1er microprocesseur intel



1977 Micro-ordinateurs



1980

Banque EMBL, GenBank, PIR

1985

Algo Alignement local de séquence
FASTA
Person & Lipman

1990

Algo Alignement local de séquence
BLAST
Altschul & al.

Séquençage ADN

P. Berg, W. Gilbert & F. Sanger

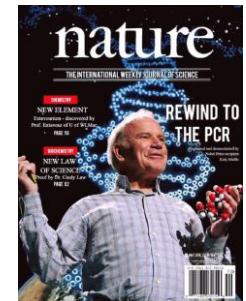
The Nobel Prize in
Chemistry 1980



1984

Amplification ADN - PCR

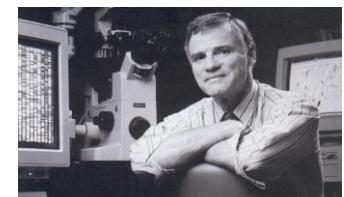
Karry Mullis



1987

1er séquenceur automatisé

L. Hood Société Applied Biosystems

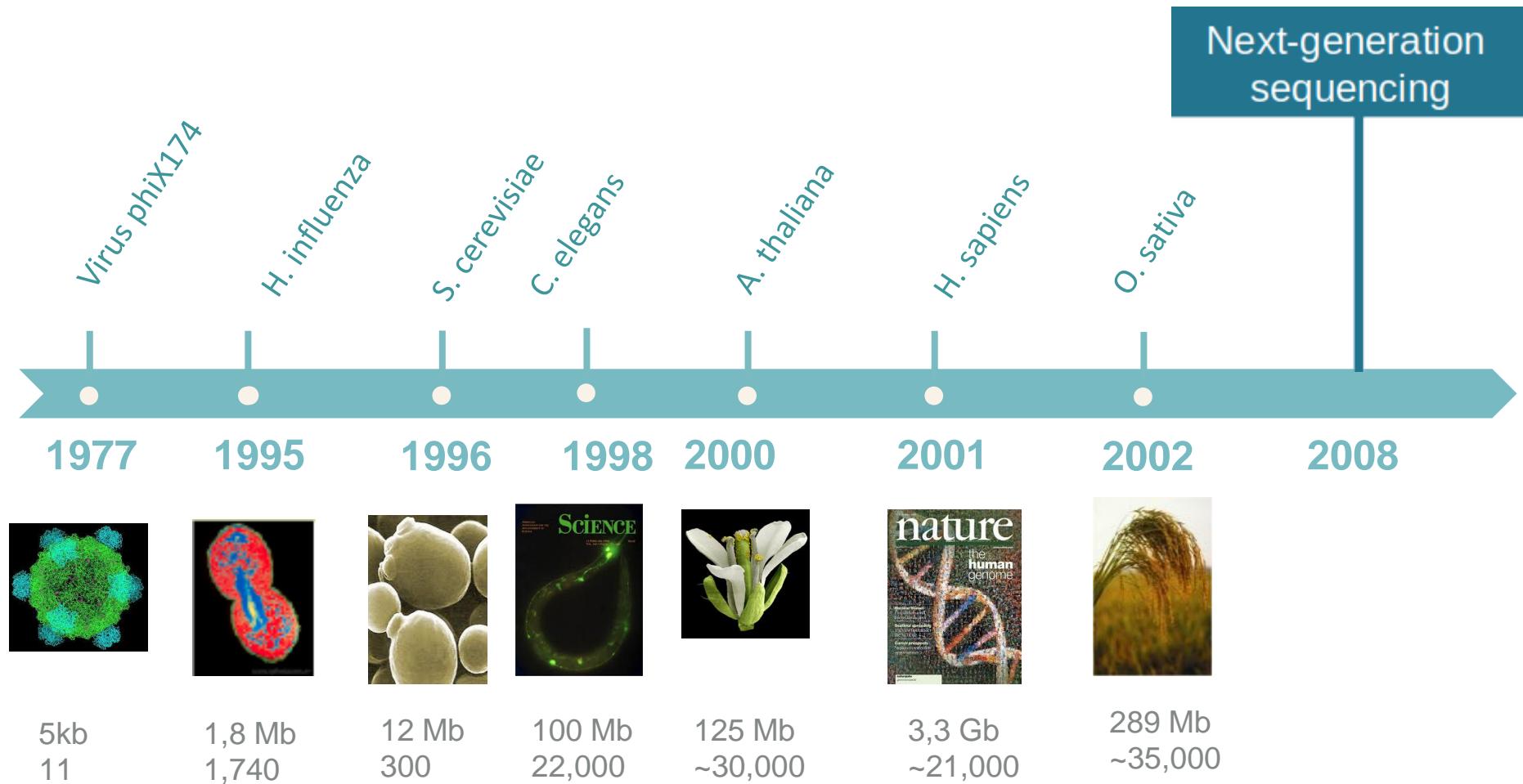


A little history of sequencing...

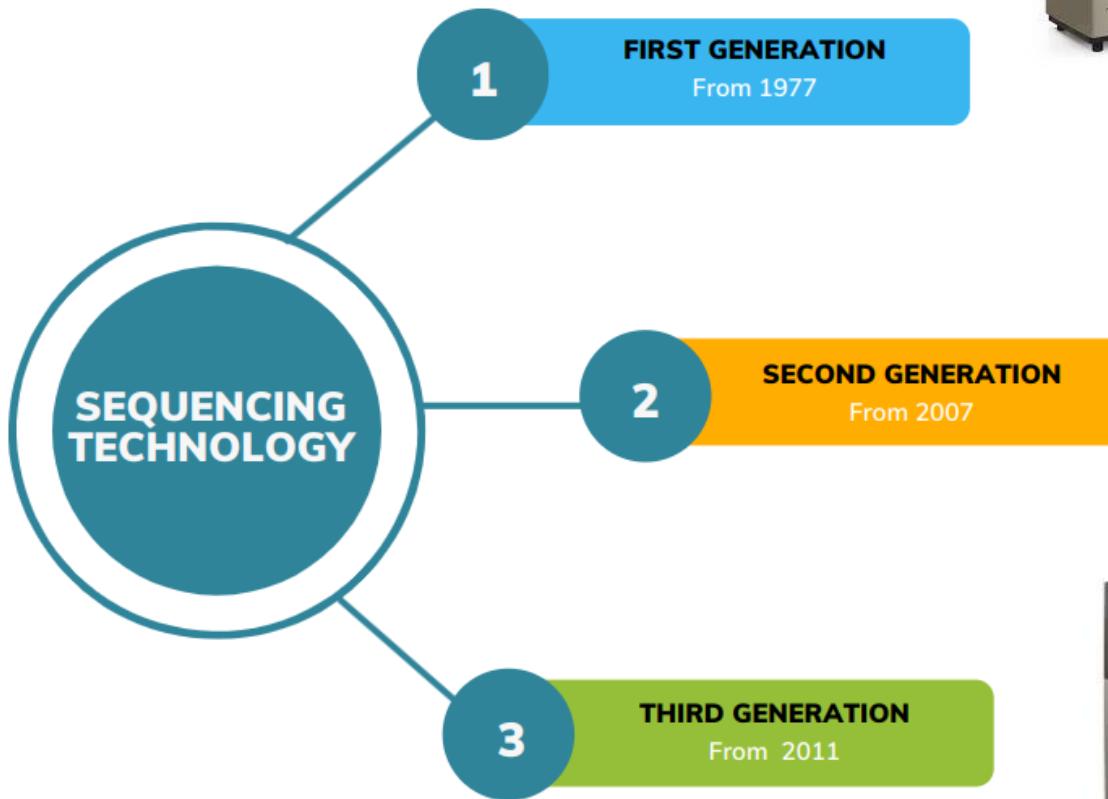
*DNA sequencing : determining the order of the four bases or nucleotides that make up a given molecule of DNA



A little history of sequencing...



Several sequencing technology



sanger

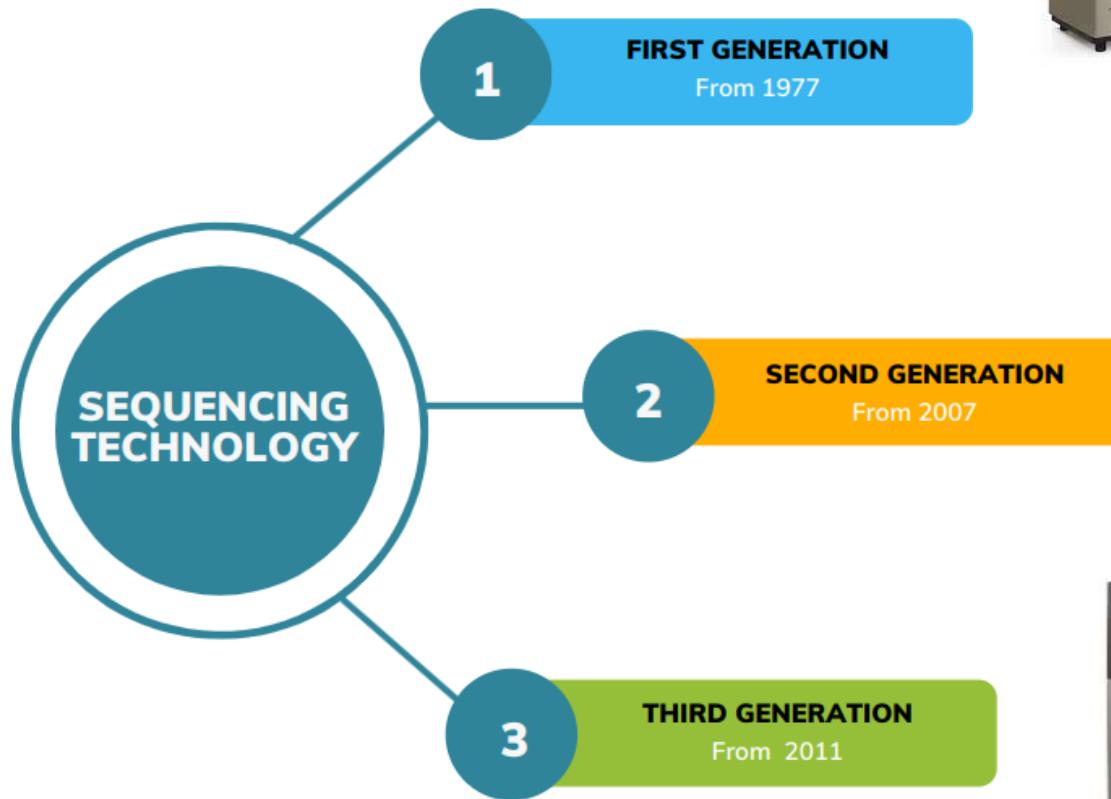


solexa, 454, illumina



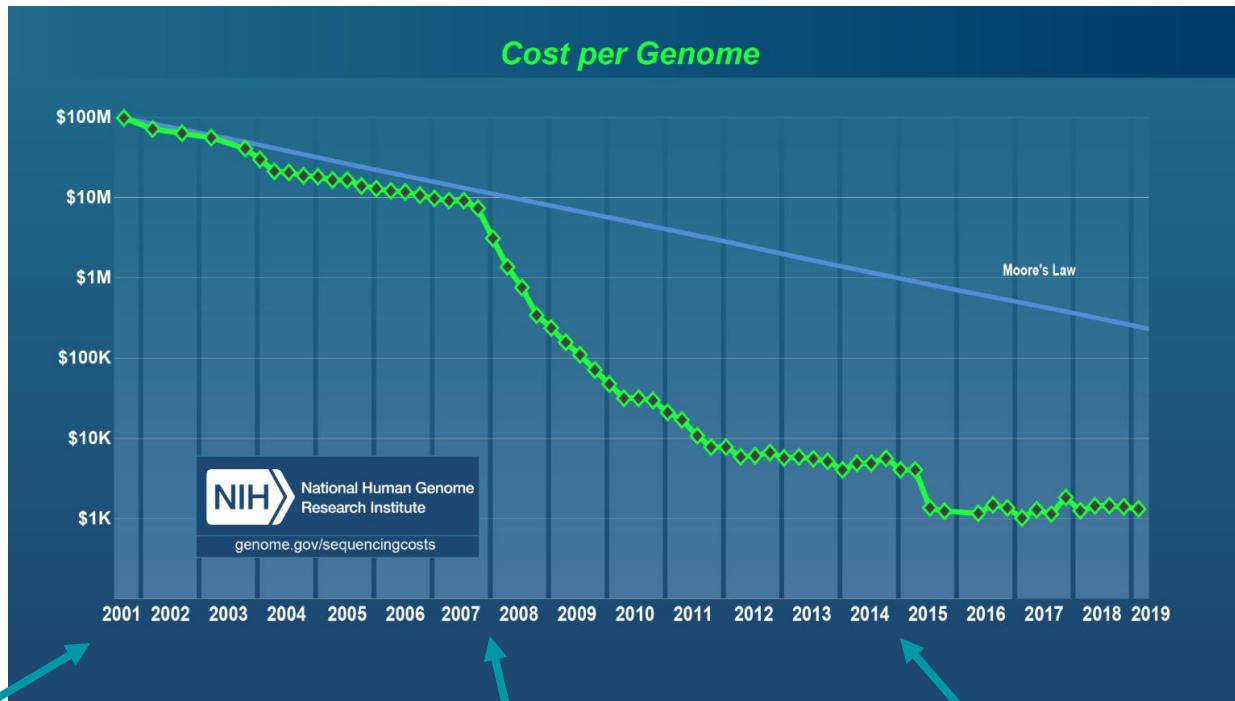
PacBio, oxford
nanopore

Several sequencing technology



Sequencing output, price, reads size, sequencing quality

From Sanger to 3rd sequencing technology



1

FIRST GENERATION
From 1977

sanger

SECOND GENERATION
From 2007

Illumina

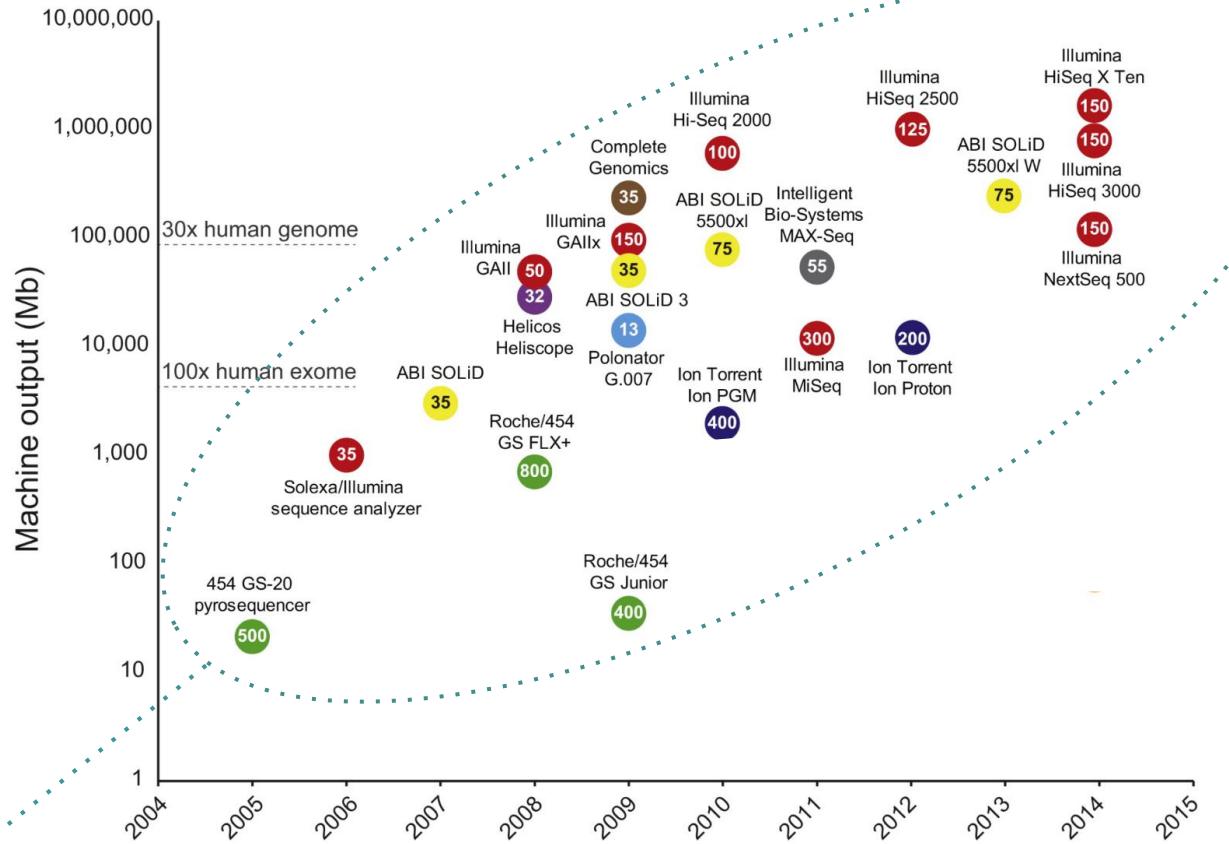
2

THIRD GENERATION
From 2011

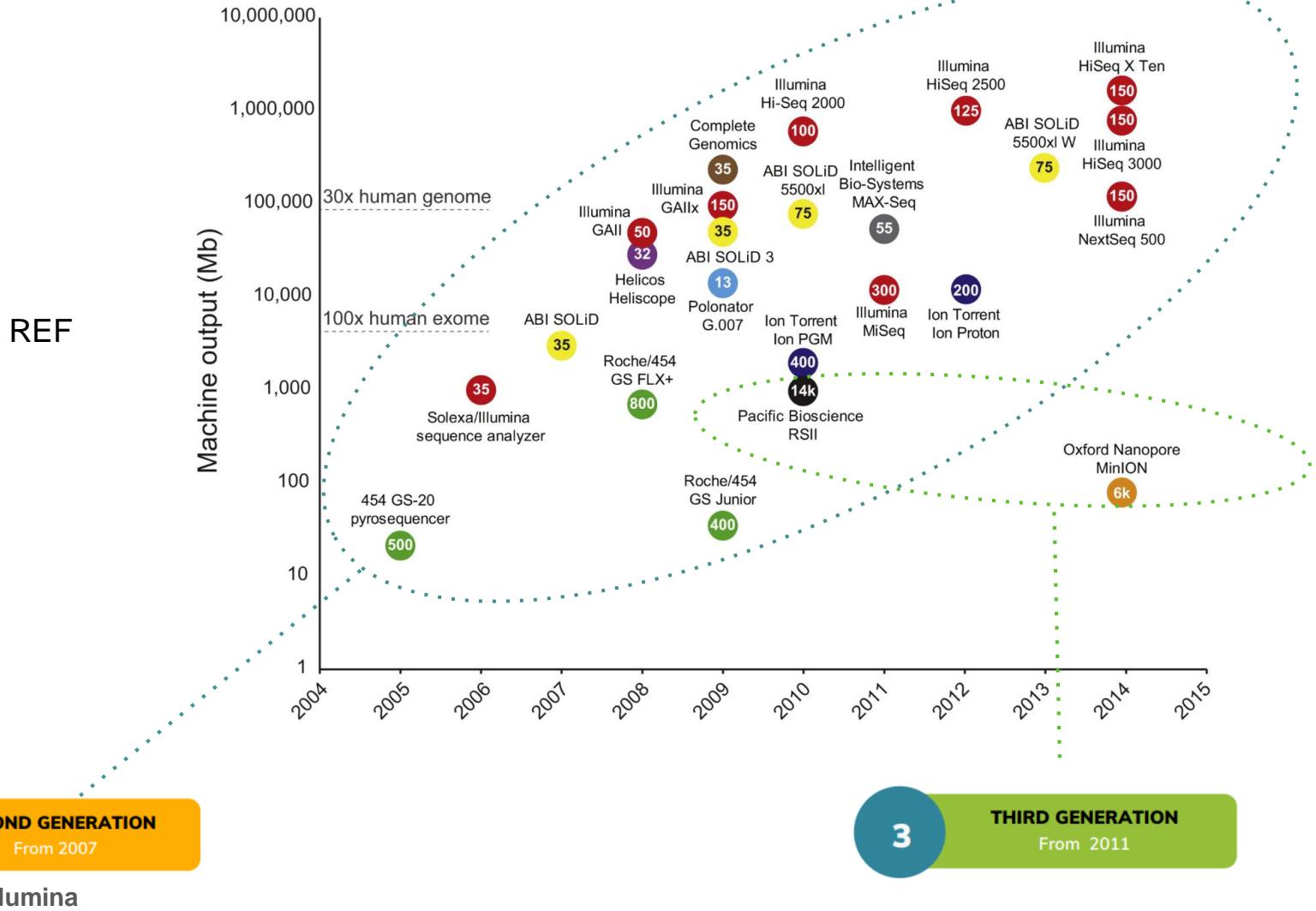
PacBio, ONT



Une augmentation du débit de séquençage



Une augmentation du débit de séquençage





Short Reads ? Long Reads ?

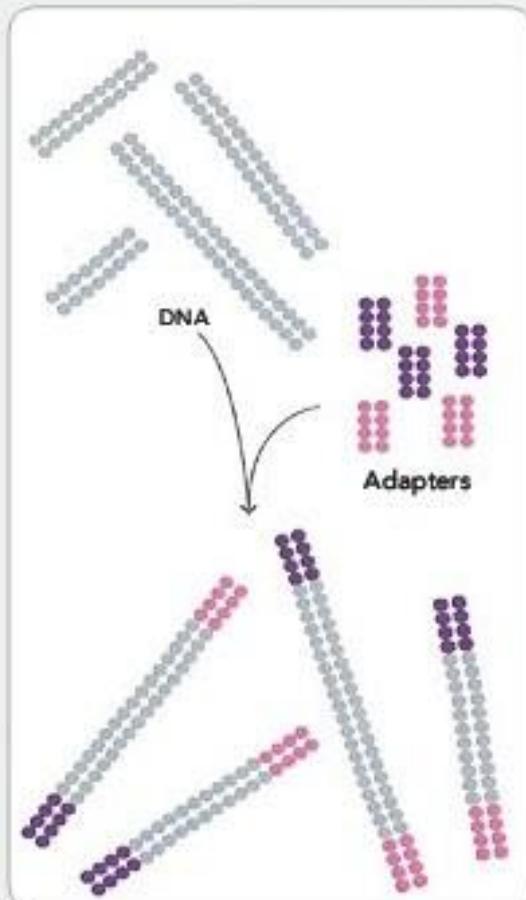
Short Reads - Illumina technology

2

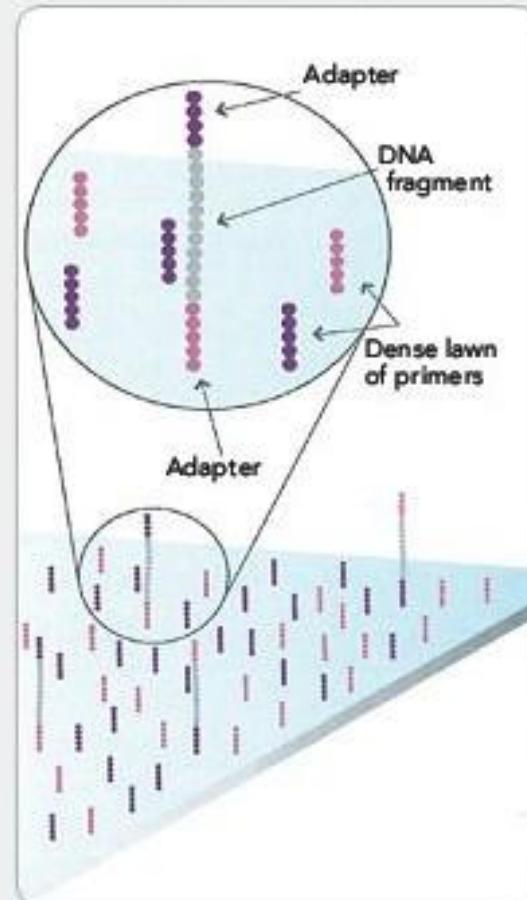
SECOND GENERATION

From 2007

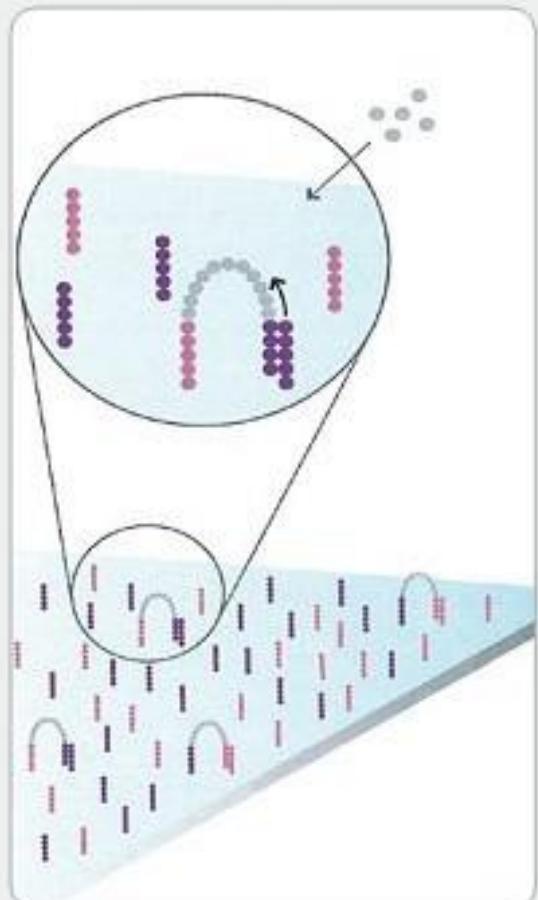
1. PREPARE GENOMIC DNA SAMPLE



2. ATTACH DNA TO SURFACE



3. BRIDGE AMPLIFICATION



Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

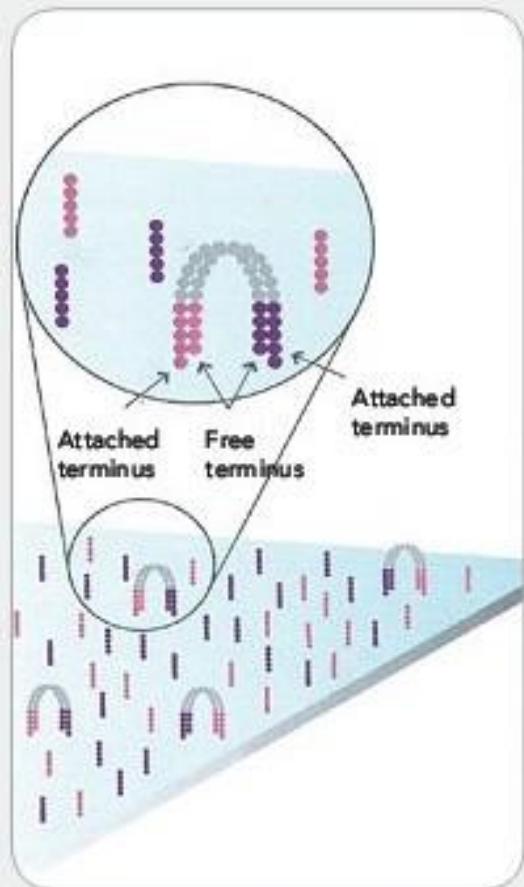
Short Reads - Illumina technology

2

SECOND GENERATION

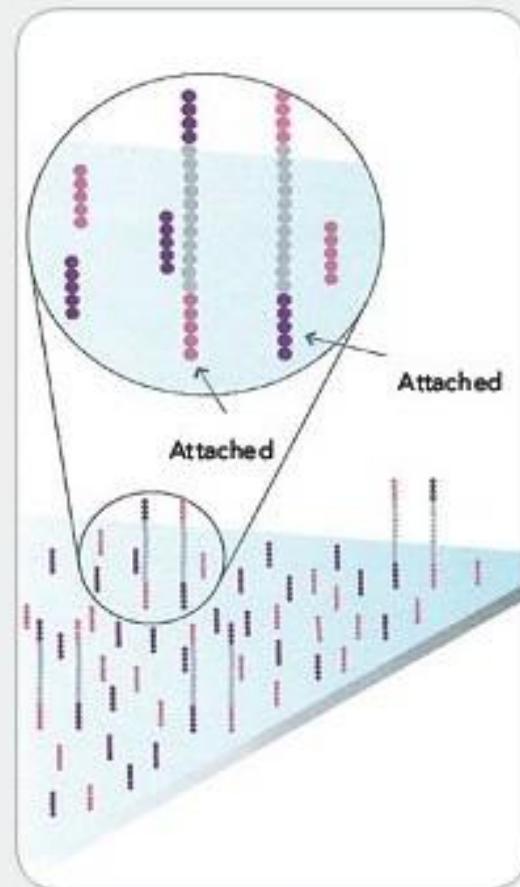
From 2007

4. FRAGMENTS BECOME DOUBLE STRANDED



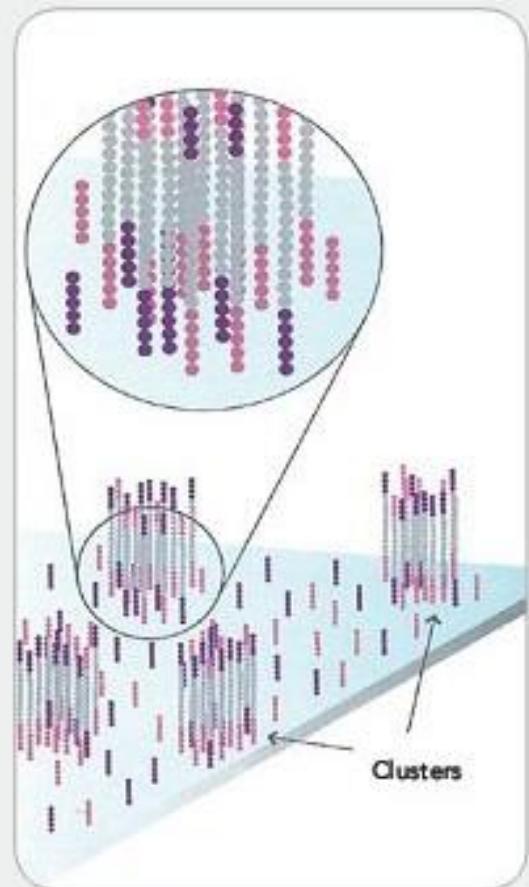
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



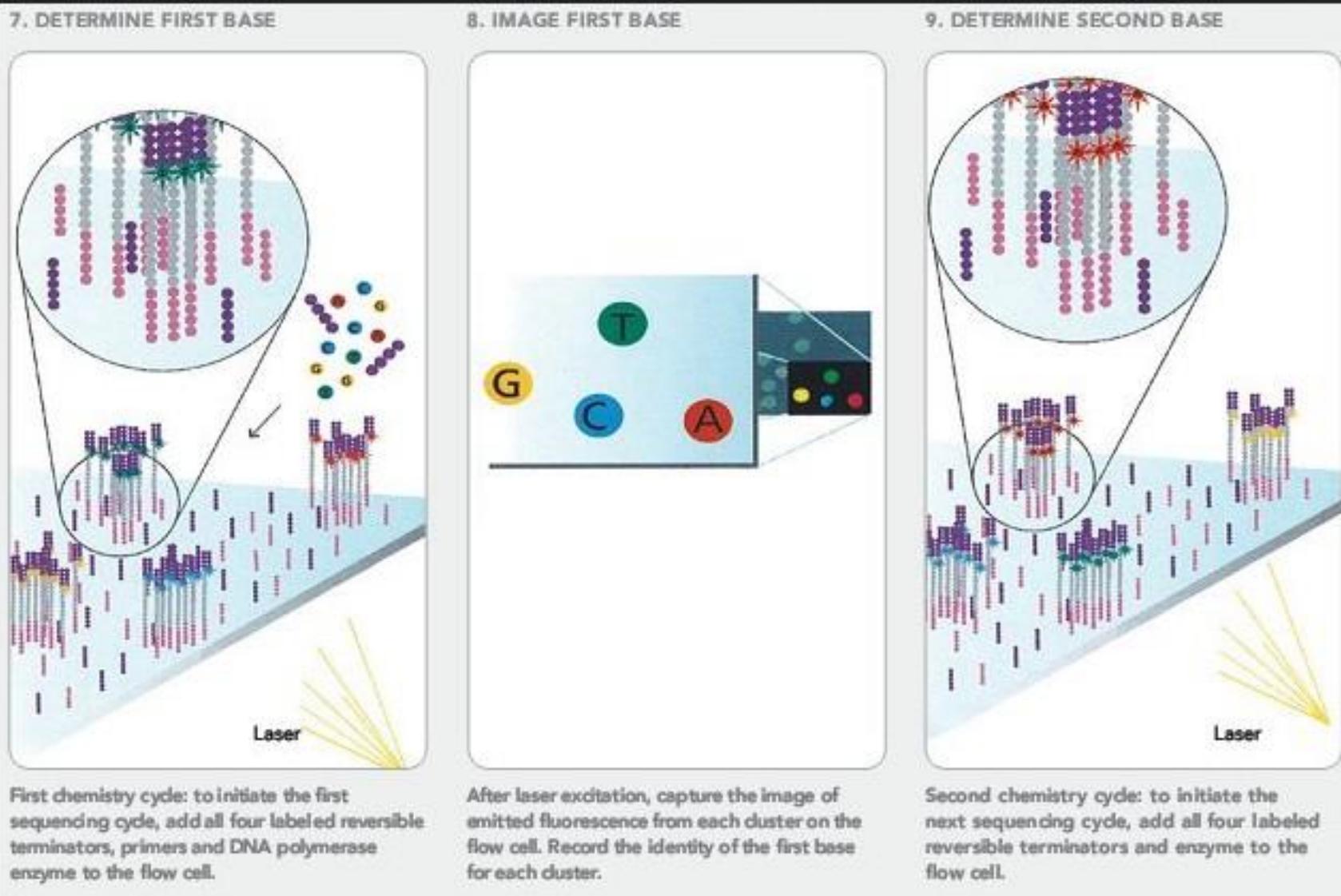
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Short Reads - Illumina technology

2

SECOND GENERATION

From 2007



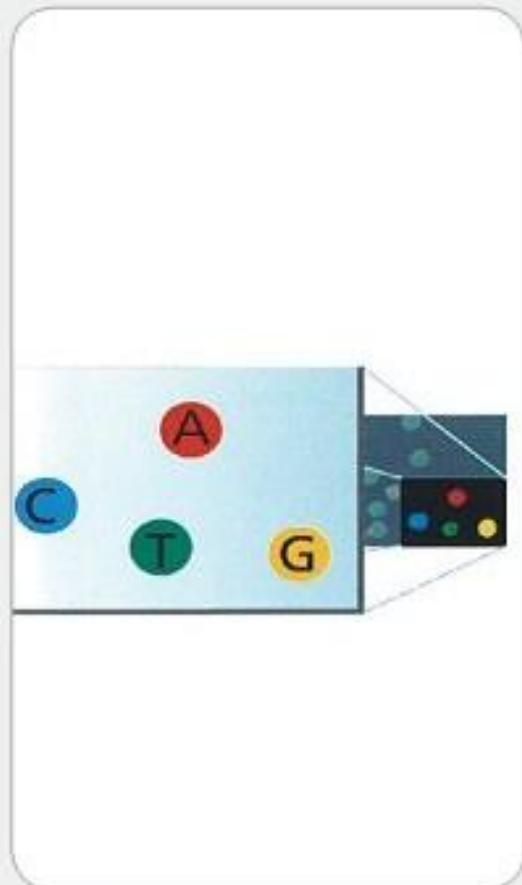
Short Reads - Illumina technology

2

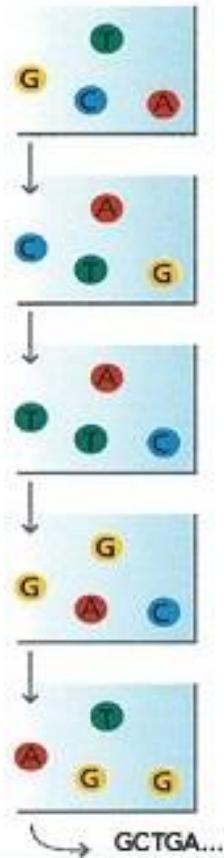
SECOND GENERATION

From 2007

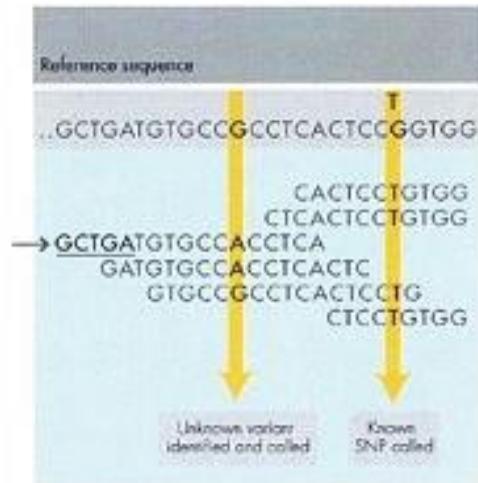
10. IMAGE SECOND CHEMISTRY CYCLE



11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



12. ALIGN DATA



After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

Align data, compare to a reference, and identify sequence differences.

2

SECOND GENERATION
From 2007



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

NovaSeq6000

- ✓ **Accuracy** 99.99 % - but questionable
- ✓ **Run is cheap**
- ✓ **MySeq is cheap** ~60 000 USD per machine



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

3

THIRD GENERATION

From 2011

Two technologies

Oxford Nanopore



MinION



GridION



PromethION

Pacific BioScience



RSII



Sequel

from Elixir GAAS 2018

Long Reads - Oxford nanopore

3

THIRD GENERATION
From 2011



Long Reads - Oxford nanopore

3

THIRD GENERATION

From 2011

<i>Triticum aestivum</i> 16 Gb	
<i>Homo sapiens</i> 3.2 Gb	
<i>Mus musculus</i> 2.7 Gb	
<i>Danio rerio</i> 1.4 Gb	
<i>Drosophila melanogaster</i> 144 Mb	
<i>Arabidopsis thaliana</i> 119 Mb	
<i>Saccharomyces cerevisiae</i> 12 Mb	
<i>Escherichia coli K-12</i> 4.6 Mb	
<i>Mycobacterium tuberculosis</i> 4.4 Mb	
<i>Influenza A</i> 13.5 kb	
<i>Ebola</i> 19 kb	

Microbial genomes

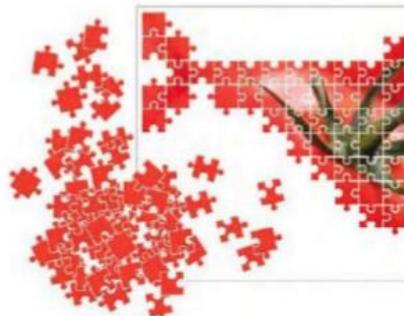
Human genomes

Animal genomes

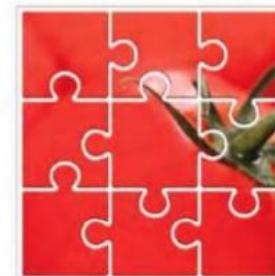
Plant genomes

- 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)



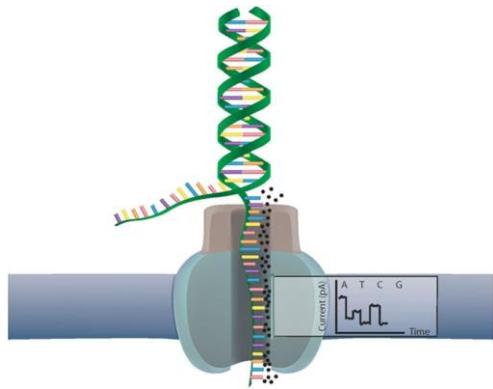
2,000 'pieces' (long reads)



Long Reads - Oxford nanopore technology

3

THIRD GENERATION
From 2011



From Circulation
Research

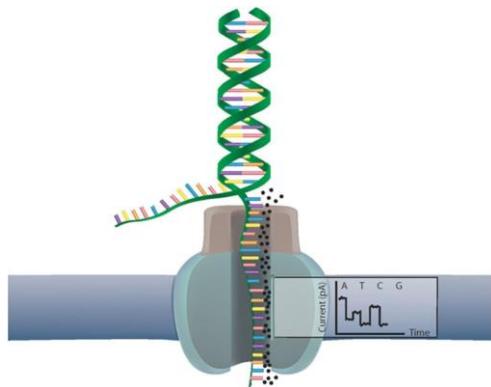
- No Amplification
- NO SYNTHESIS
- Very Long Length

Long Reads - Oxford nanopore technology

3

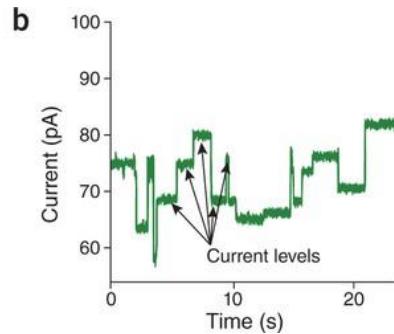
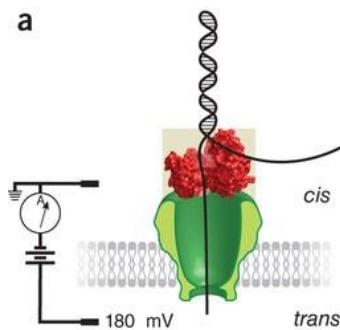
THIRD GENERATION

From 2011



From Circulation
Research

- No Amplification
- NO SYNTHESIS
- Very Long Length



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

3

THIRD GENERATION
From 2011



- ✓ Single strand direct sequencing
 - ✓ Bases Modification detection in real-time
 - ✓ Native RNA!
 - ✓ **Read length** ~ 10-50kb more than 2Mb rep
 - ✓ **Run cheap** 1,000 USD for 30Gb by now minimum
 - ✓ **Machine cheap** 1,000 USD for Minion
 - ✓ **Fast** 15mn library, 48-72h run
-
- Error Rate 3-8%, can be corrected, 1-2% in tests
 - Quality of DNA/RNA limits the sequencing



3

THIRD GENERATION

From 2011

Research areas

✿ Microbiology

👤 Human genomics

_MICROBIOME

👤 Clinical research

悱 Environmental

♋ Cancer

♣ Plant

TRANSCRIPTOME

.MOUSE Animal

POPULATIONS
genomics

From Nanopore website

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

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

Comparison

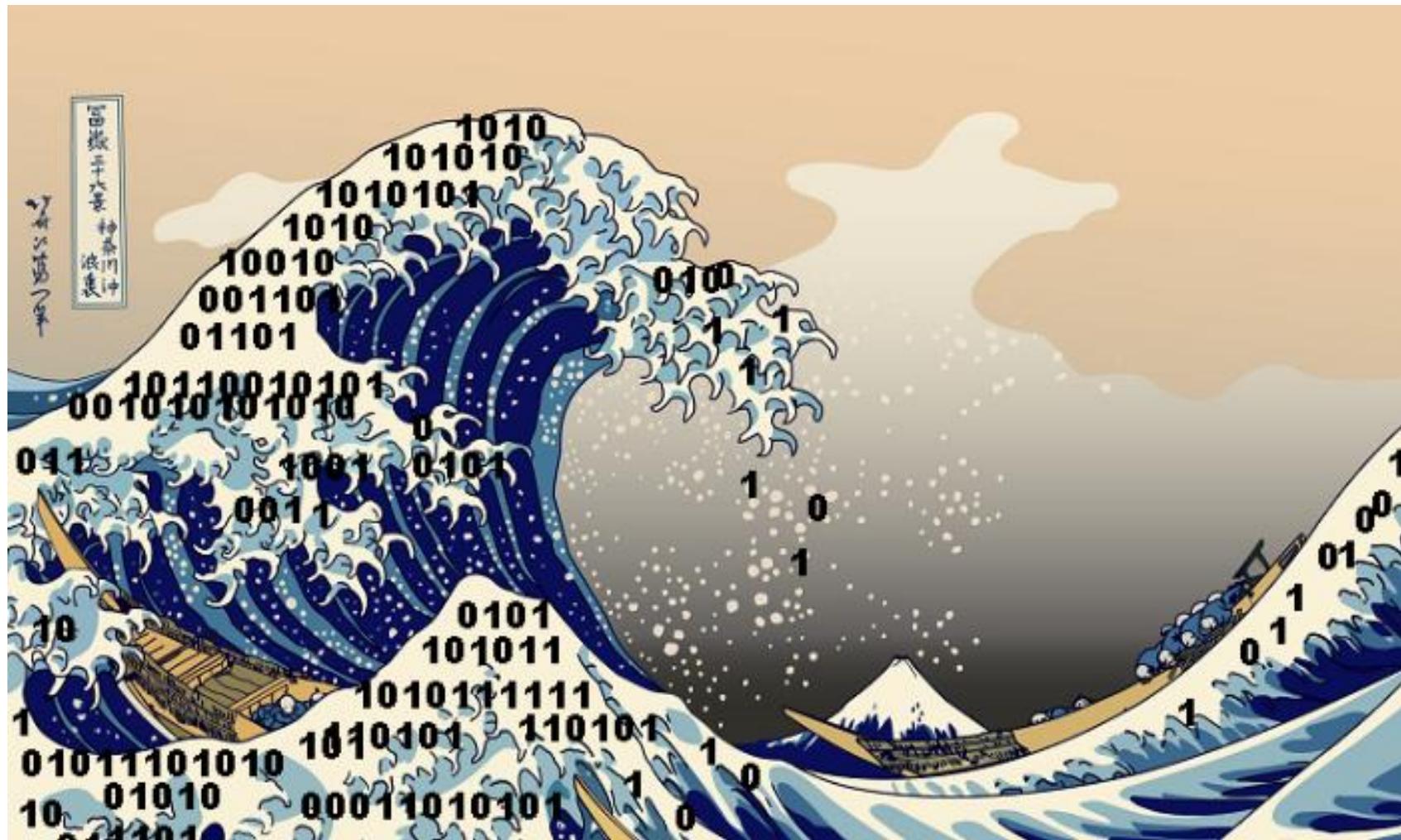
NGS platforms

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

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

From wikipedia website

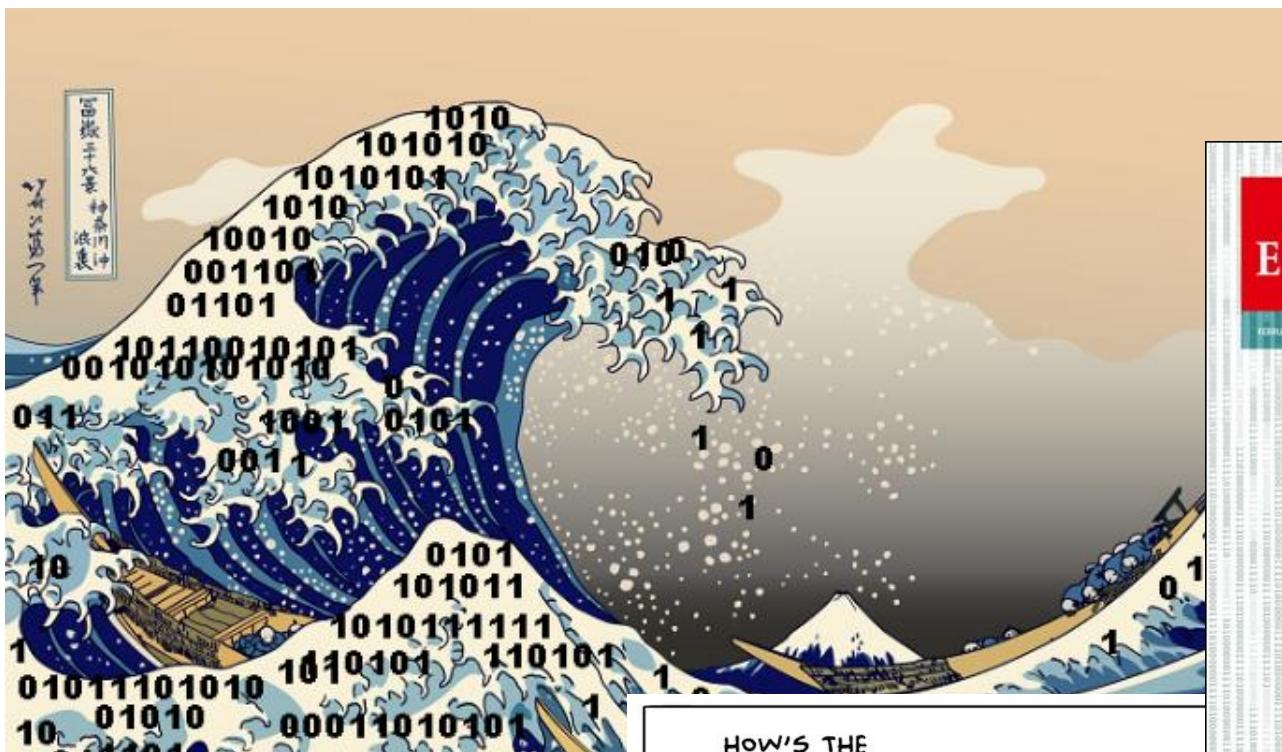
From data rarity to data deluge



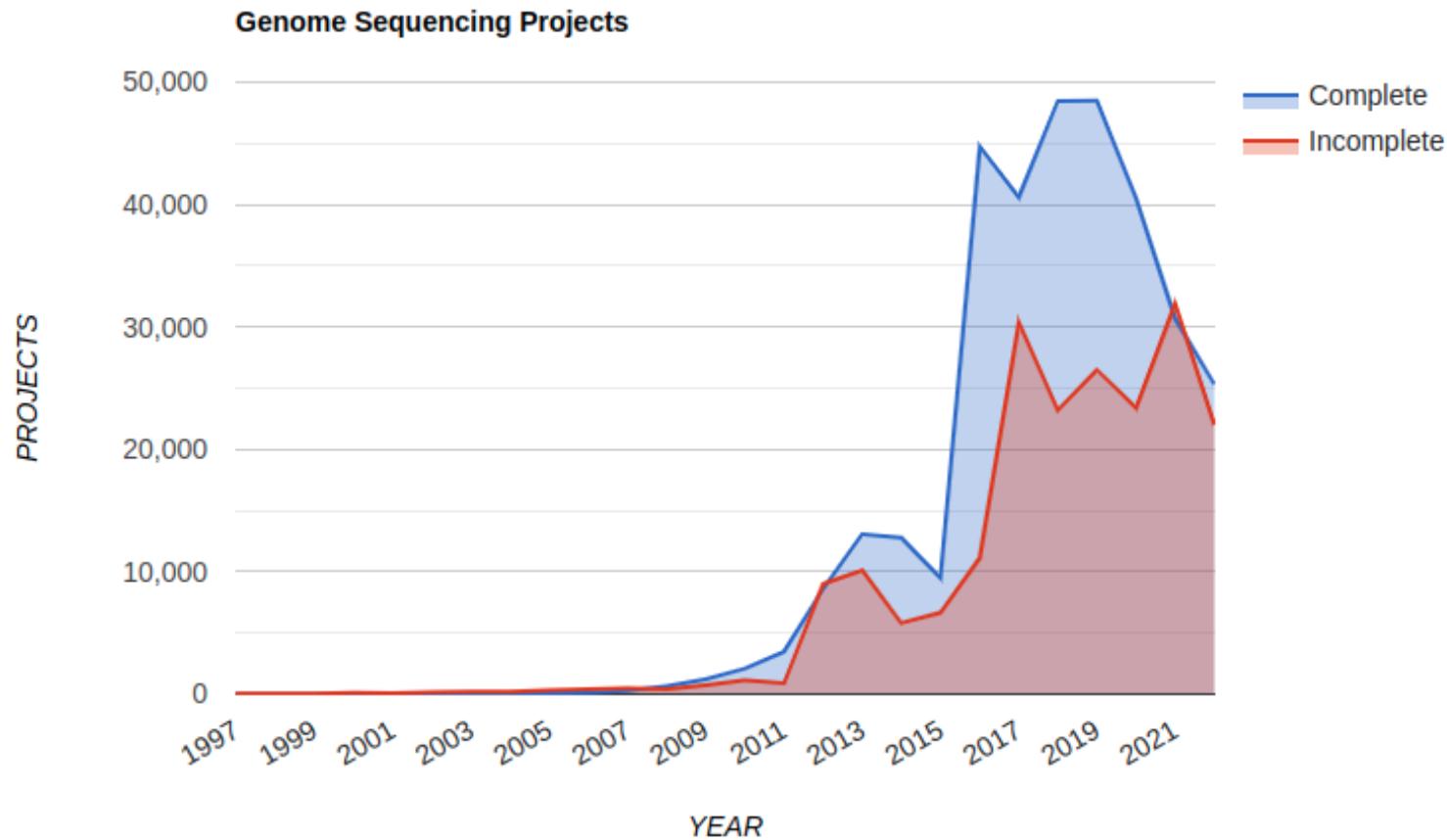
The Great Wave off Kanagawa, Hokusai

@amitechsolutions.com

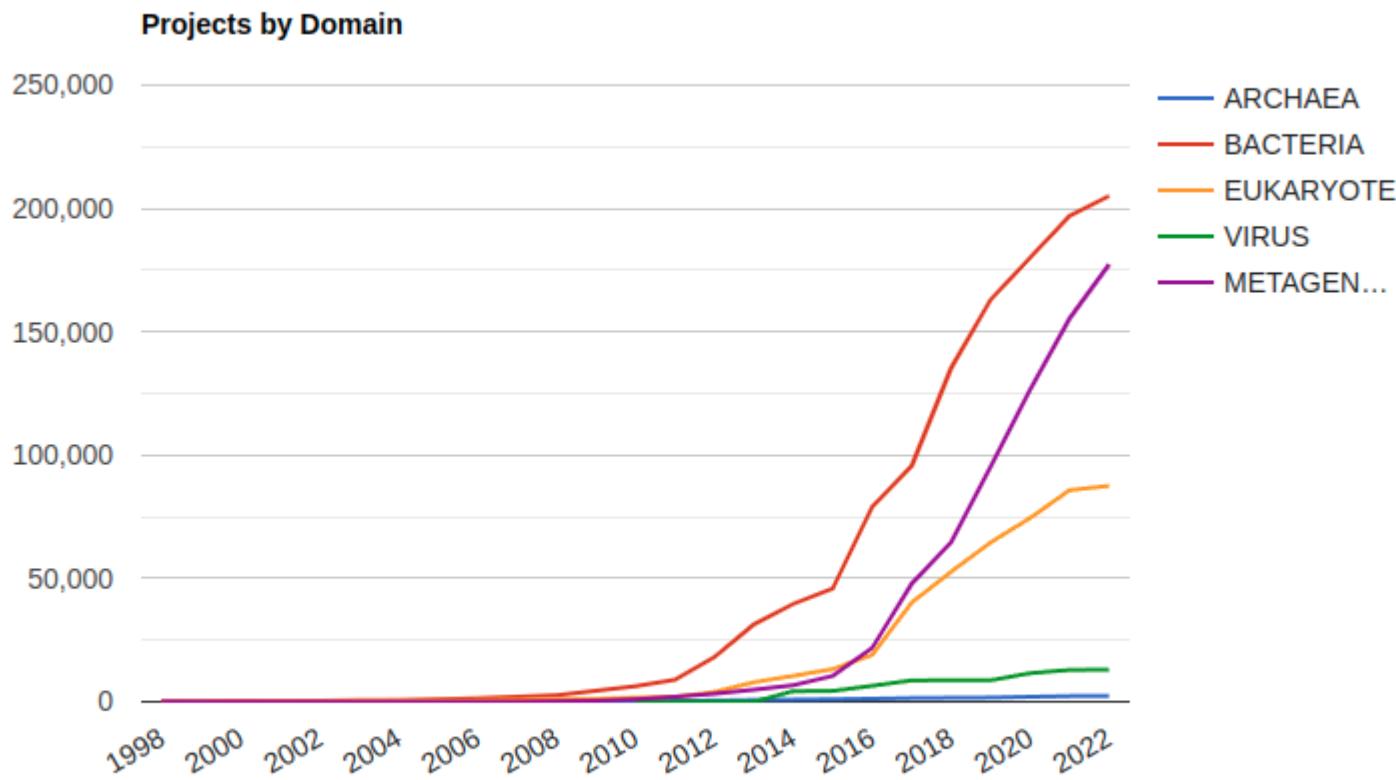
From data rarity to data deluge



Genome Totals by year and status



Project Totals by year and domain group



Phylogenetic distribution of Bacterial Genome Projects

Biological Databases

✓ Sequence

- Nucleic :
- Proteic :



PIR, Pfam, Prosite

✓ Structure

PDB

SCOP

CATH

✓ Specialized

by organism, by sequence type

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

NIH National Library of Medicine
National Center for Biotechnology Information

All Databases

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

THE 2022 VIRTUAL OUTREACH EVENT SERIES

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

National Library of Medicine
National Center for Biotechnology Information

All Databases

[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

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

National Library of Medicine
 National Center for Biotechnology Information

All Databases

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

Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy BioCollections

Search for as lock

Display levels using filter:

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](#); [Petrosavidae](#); [commelinids](#); [Poales](#); [Poaceae](#); [BOP clade](#); [Oryzoideae](#); [Oryzeae](#); [Oryzinae](#); [Oryza](#)

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.org/10.1186/2047-217X-3-7>

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	26,120
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

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

National Library of Medicine

National Center for Biotechnology Information

Genome txid4530[Organism:exp]

[Create alert](#) [Limits](#) [Advanced](#)

[Help](#)

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: [ID: 10](#)

Organism Overview ; [Genome Assembly and Annotation report \[95\]](#) ; [Organelle Annotation Report \[8\]](#)

Oryza sativa (Asian cultivated rice)
 Oryza sativa Organism overview

Lineage: Eukarya[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:

- Or 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_020256.1	AP014967.1		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

[See more...](#)

NCBI Resources

[Genome Data Viewer](#)

Tools

[BLAST Genome](#)

Related information

Assembly
BioProject
Gene
Components
Protein
PubMed
Taxonomy

Search details

txid4530[Organism:exp]

[See more...](#)

Recent activity

Turn Off Clear

Oryza sativa
txid4530[Organism:exp] (1)
embryophyta AND ((refseq[filter] OR swissprot[filter])) (7447863)
embryophyta AND (refseq[filter]) (7408233)
(oryza) AND "Oryza sativa"[orgn] (444207)

[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 variants to gene and position! ClinVar is accessible public archive of variants

Celebrating 1 Year of NCBI Virtual Outreach Events

We launched the NCBI Virtual Event series in the fall of 2022

[See more...](#)

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

National Library of Medicine
National Center for Biotechnology Information

All Databases

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Nucleotide

Species Summary Sort by Default order [Filters: Manage Filters](#)

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
[Oryza sativa](#)
Asian cultivated rice (*Oryza sativa*) is a species of monocot in the family Poaceae (grass family).
Taxonomy ID: 4530
[Genomes](#) [Genes](#) [BLAST](#)

Was this helpful?

Items: 1 to 20 of 2291284
 << First < Prev Page of 114565 Next > Last >>
 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
[GenBank](#) [FASTA](#) [Graphics](#)

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:

Search details
 txid4530[Organism:exp]

Popular Resources

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [BLAST](#)
- Nucleotide** (circled in red)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

NCBI News & Blog

- [New ClinVar graphical display](#)
- [Maps clinically significant variant gene and position! ClinVar is accessible public archive of variants](#)
- [Celebrating 1 Year of NCBI Virtual Outreach Events](#)
- [We launched the NCBI Virtual Event series in the fall of 2022](#)

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

National Library of Medicine
National Center for Biotechnology Information

All Databases

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

Log in

PubMed.gov

african rice domestication population genomics Advanced Create alert Create RSS

Save Email Send to Sorted by: Best match Display options

MY NCBI FILTERS

RESULTS BY YEAR

2014 2022

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

Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*).
Cite Nabholz B, Sarah G, Sabot F, Ruiz M, Adam H, Nidelet S, Ghesquière A, Santoni S, David J, Glémén S. Mol Ecol. 2014 May;23(9):2210-27. doi: 10.1111/mec.12738. Epub 2014 Apr 18.
Share 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 ...

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

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

SRA / ENA (European Nucleotide Archive)

Phytozome: Ressources génomiques de plantes

Sequencing project

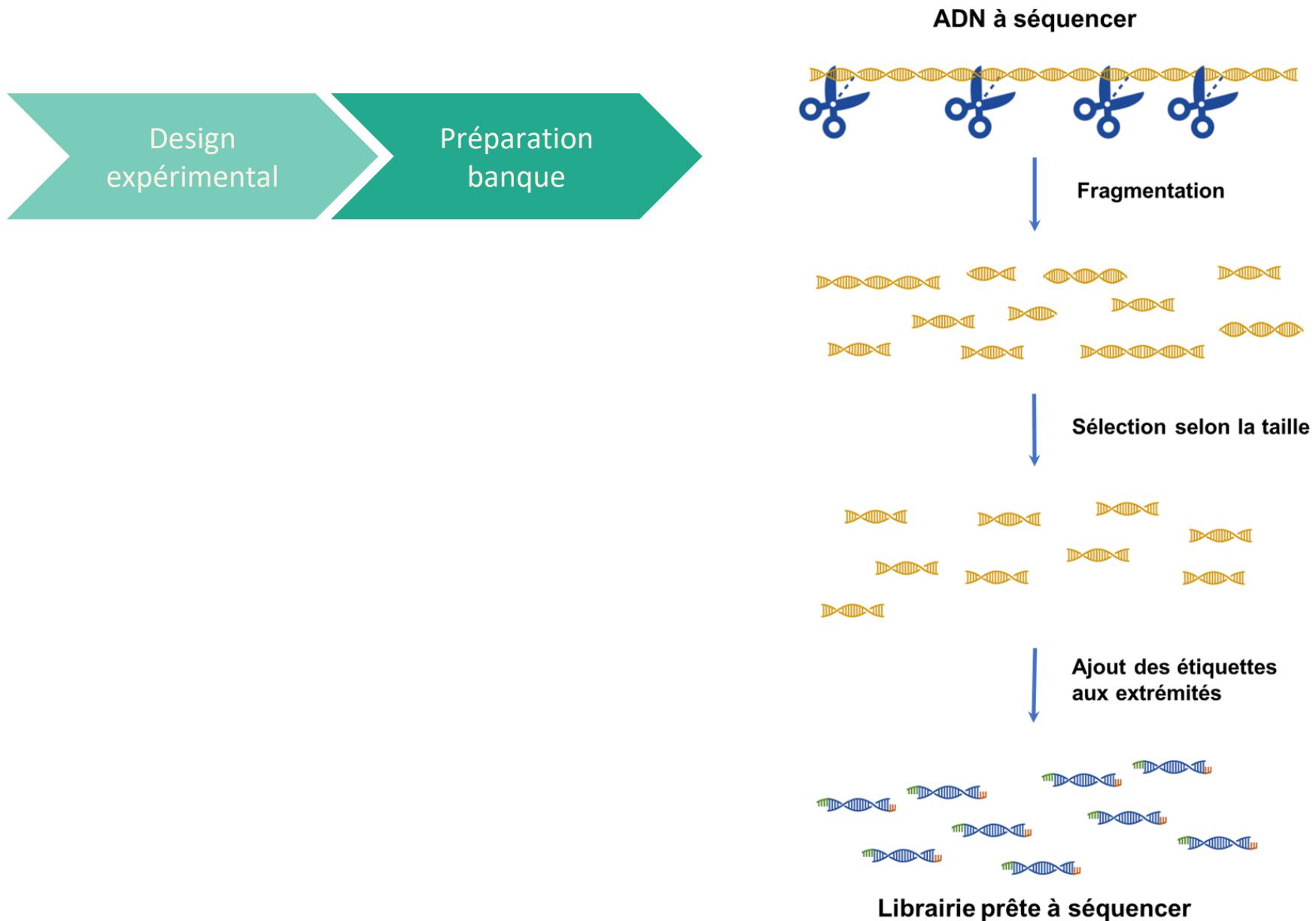


OVERVIEW OF DNA SEQUENCING PROJECT

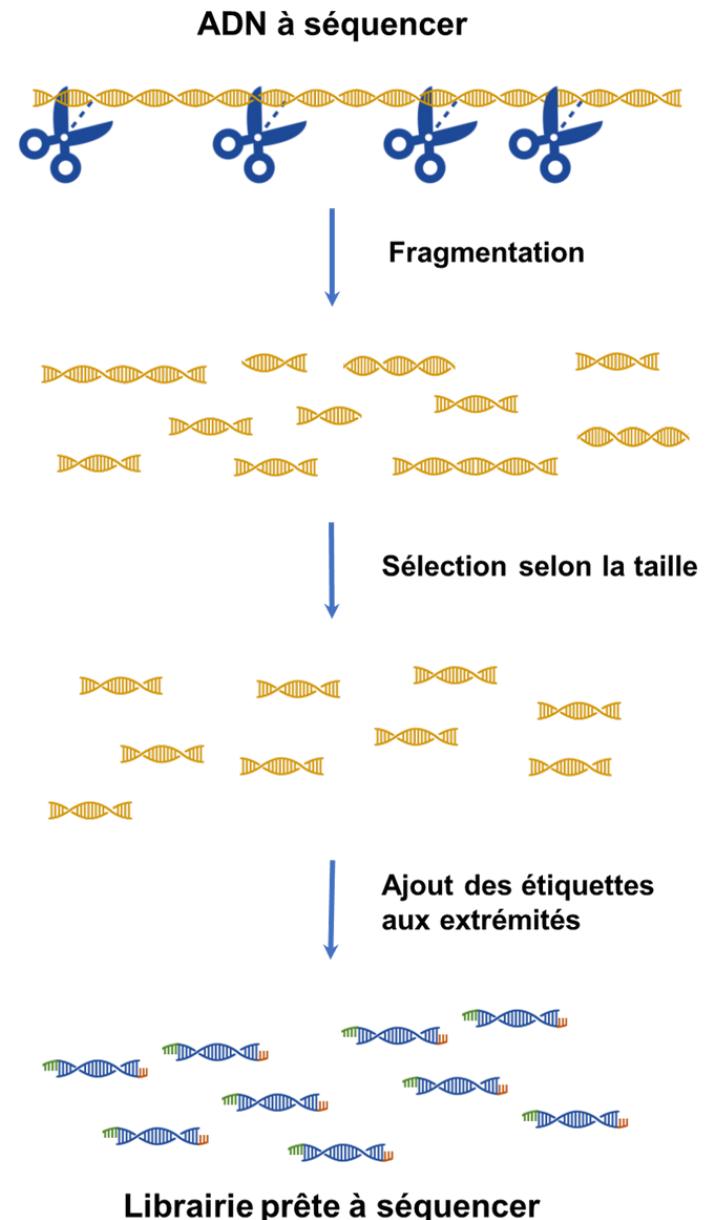
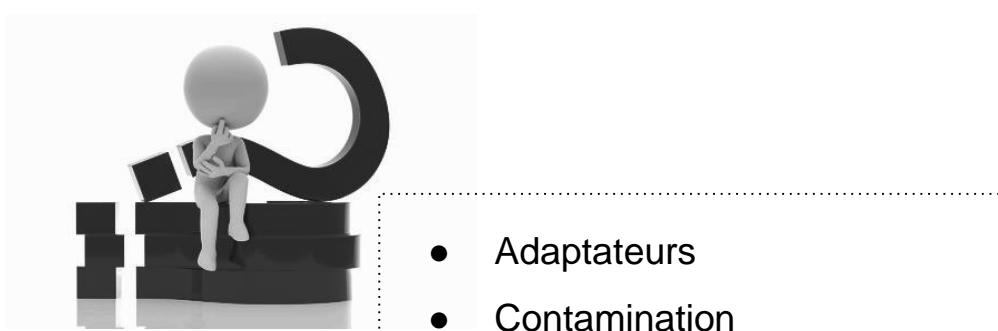
Design expérimental

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

OVERVIEW OF DNA SEQUENCING PROJECT



OVERVIEW OF DNA SEQUENCING PROJECT



OVERVIEW OF DNA SEQUENCING PROJECT

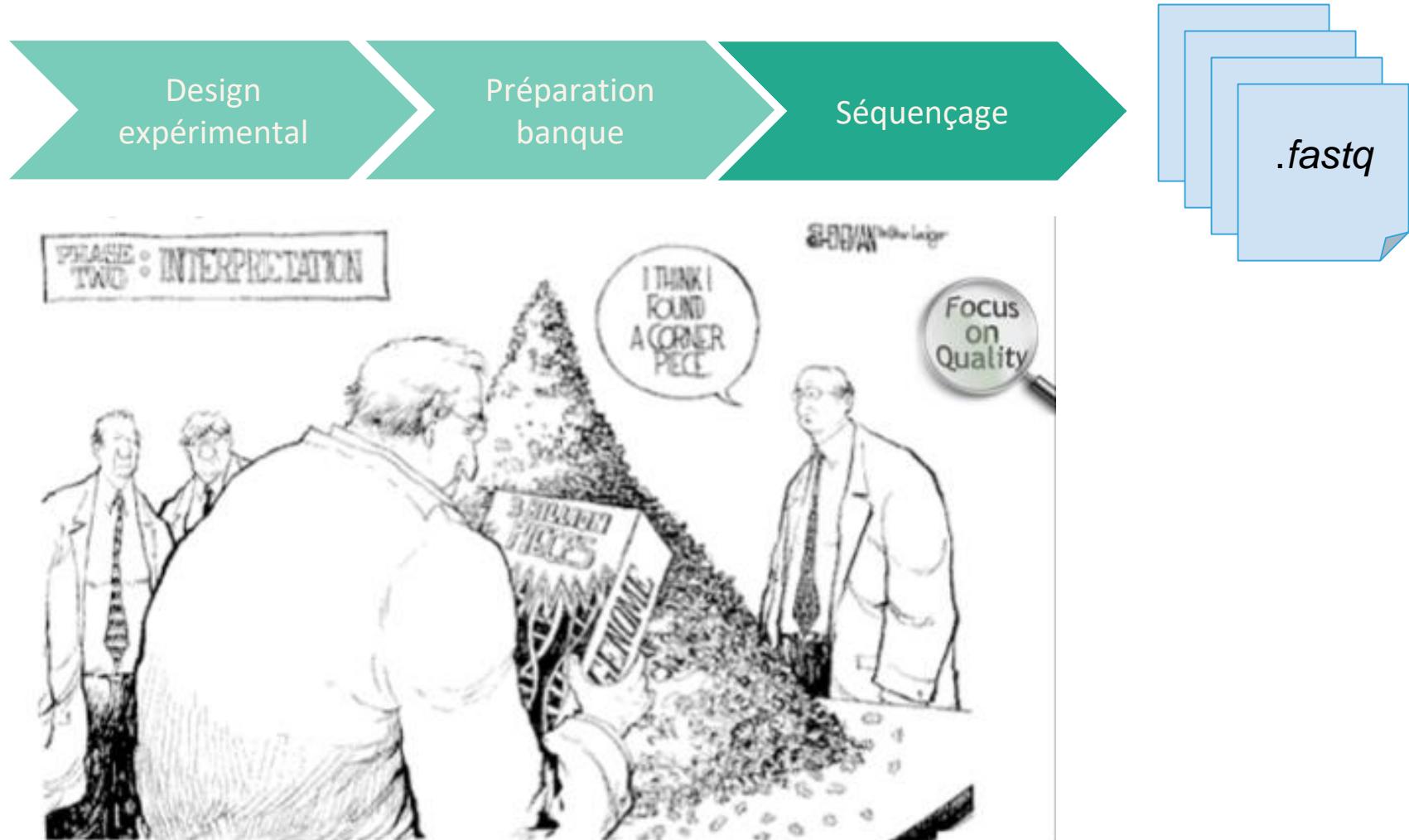


OVERVIEW OF DNA SEQUENCING PROJECT



- Qualité de séquençage
- Profondeur de séquençage

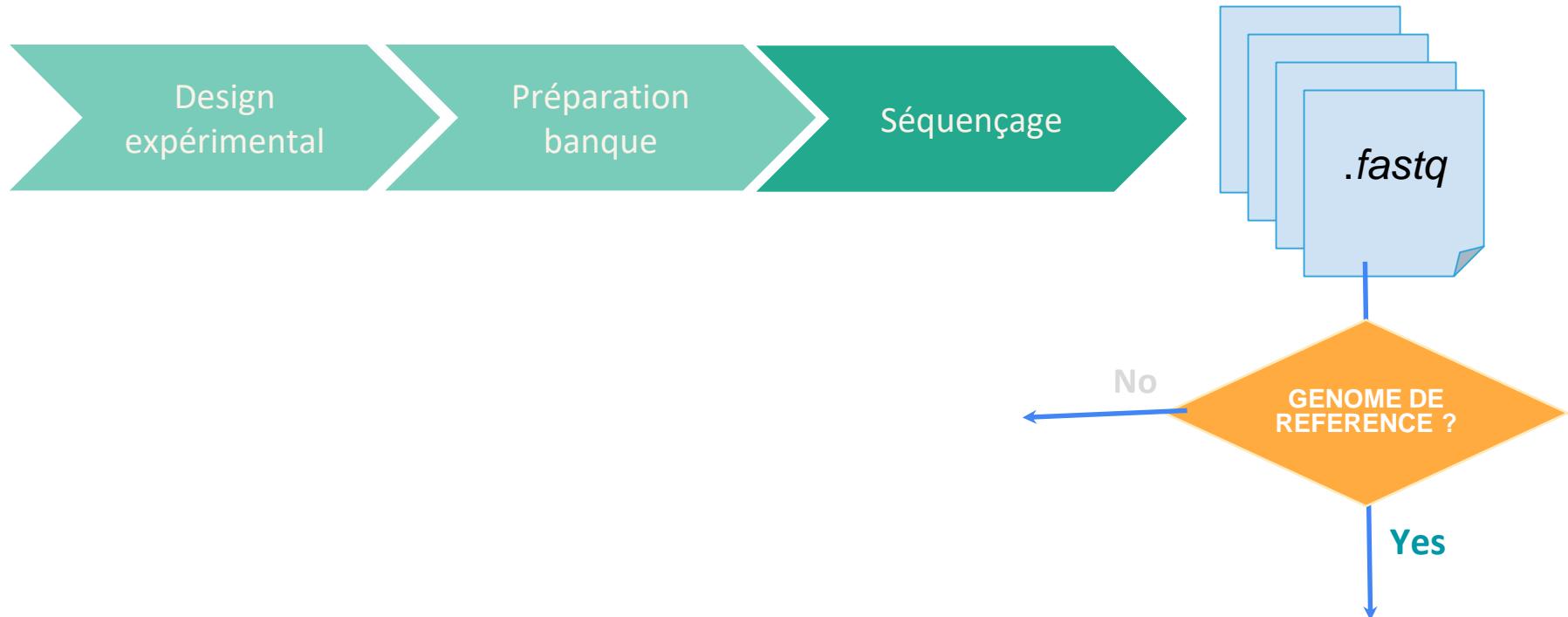
OVERVIEW OF DNA SEQUENCING PROJECT



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



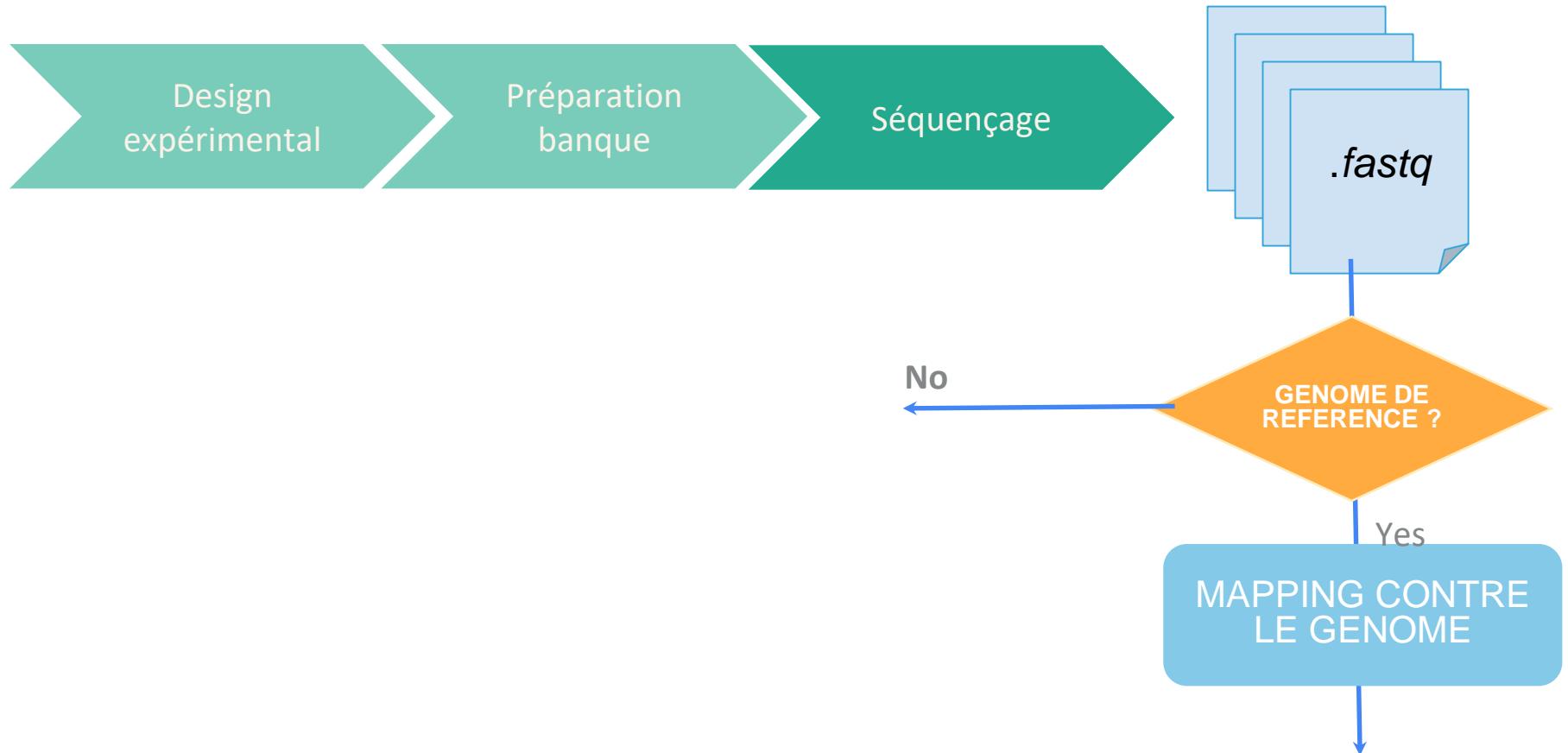
OVERVIEW OF DNA SEQUENCING PROJECT



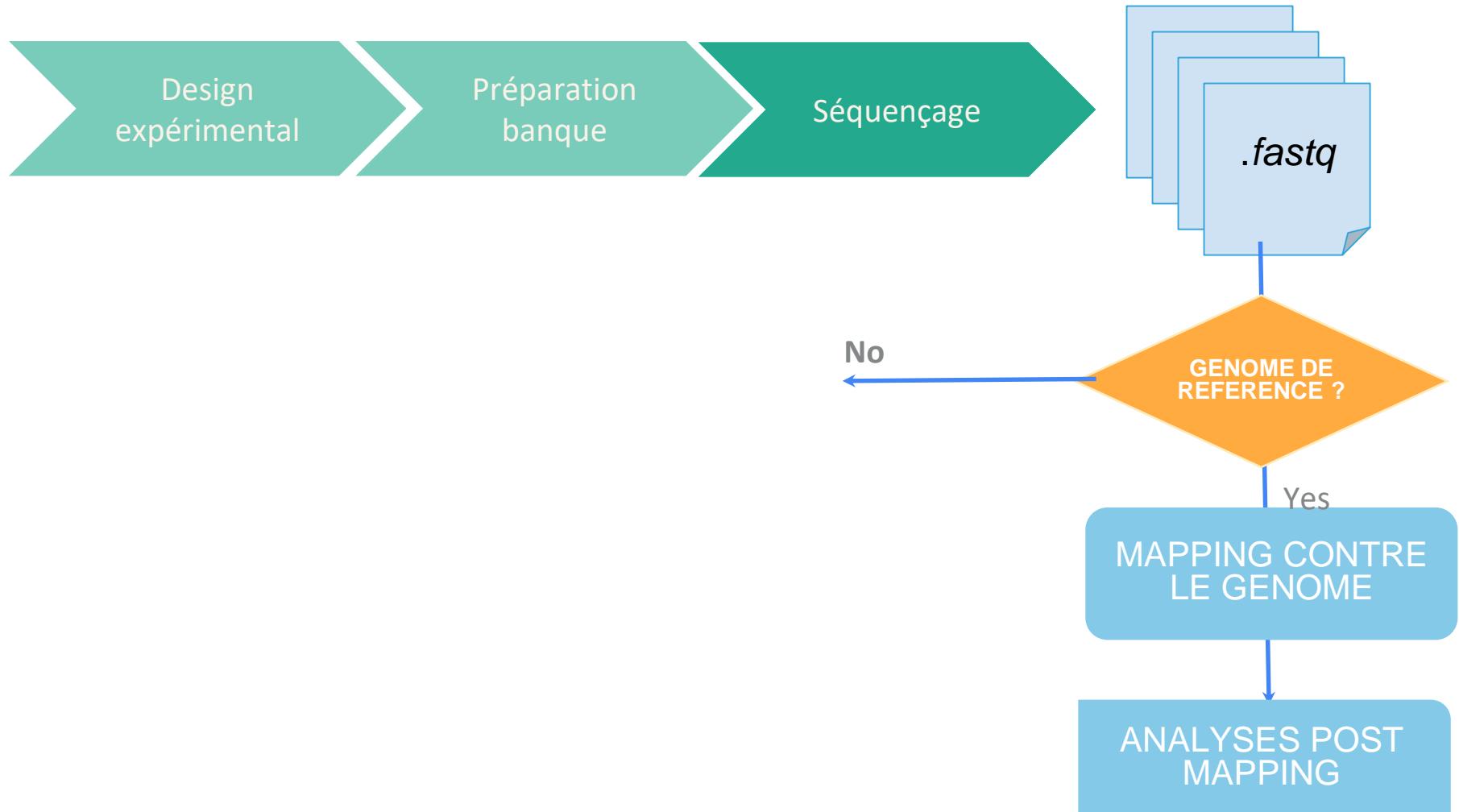
OVERVIEW OF DNA SEQUENCING PROJECT



OVERVIEW OF DNA SEQUENCING PROJECT



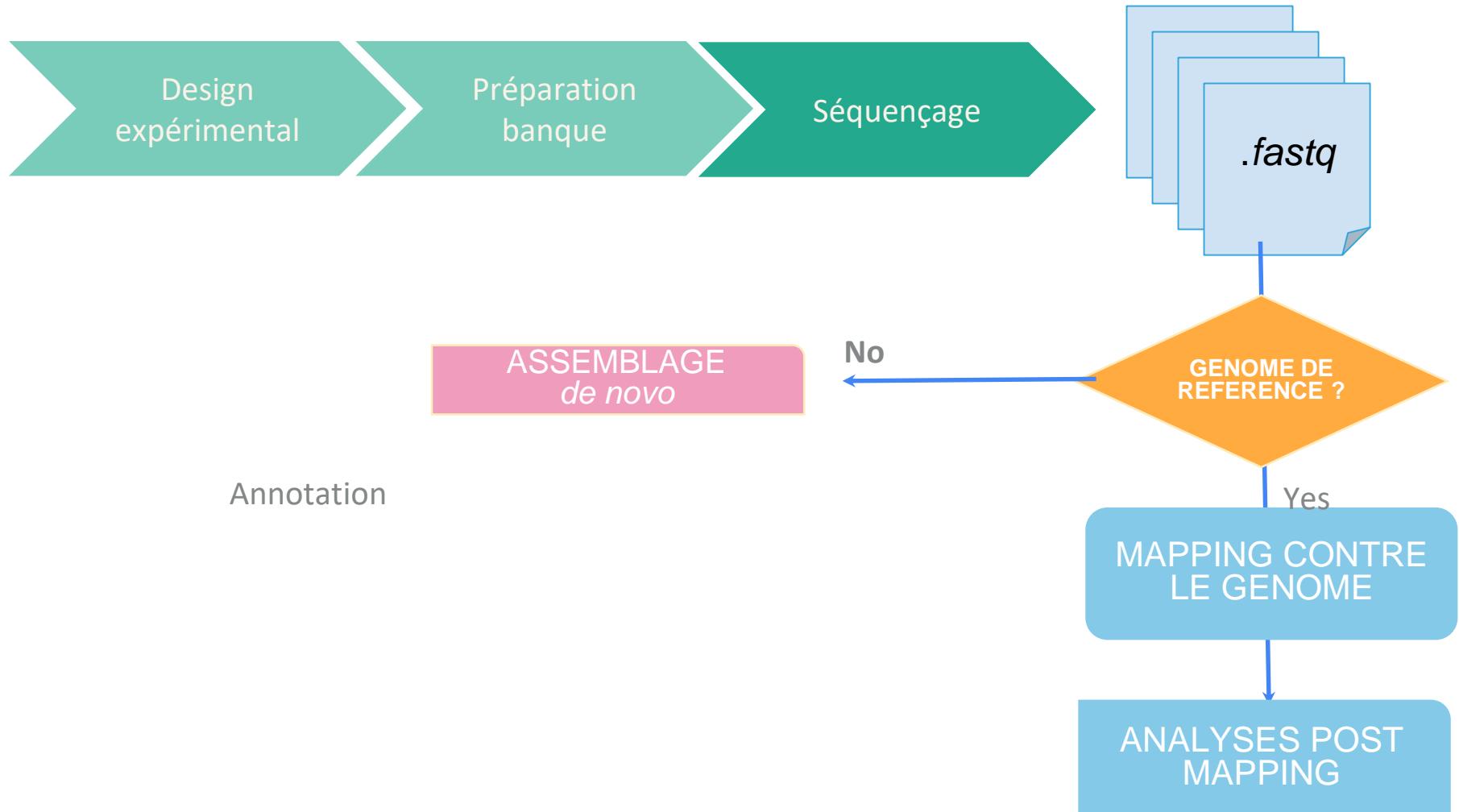
OVERVIEW OF DNA SEQUENCING PROJECT



Adapted from Ross Whetten...

SNP, GWAS? expression
différentielle

OVERVIEW OF DNA SEQUENCING PROJECT



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

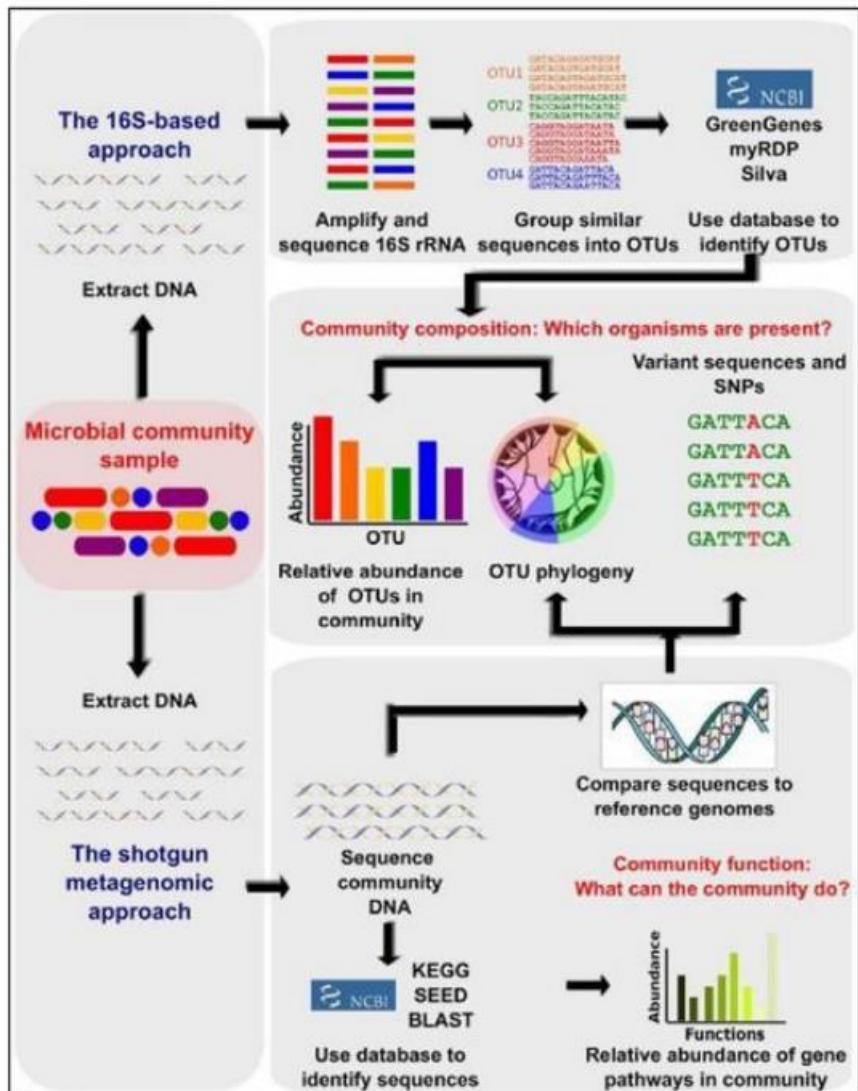
=> Organisms can be studied directly in their environments bypassing the need to isolate each species

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

Two main strategies in metagenomics

We can distinguish targeted metagenomics or shot-gun metagenomics :

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



Markers genes vs Shotgun metagenomics

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

Strategies in Diversity Characterisation

Technique	Advantages and challenges	Main applications
Metataxonomics using amplicon sequencing of the 16S or 18S rRNA gene or ITS	<ul style="list-style-type: none"> + Fast and cost-effective identification of a wide variety of bacteria and eukaryotes - Does not capture gene content other than the targeted genes - Amplification bias - Viruses cannot be captured 	<ul style="list-style-type: none"> * Profiling of what is present * Microbial ecology * rRNA-based phylogeny
Metagenomics using random shotgun sequencing of DNA or RNA	<ul style="list-style-type: none"> + No amplification bias + Detects bacteria, archaea, viruses and eukaryotes + Enables <i>de novo</i> assembly of genomes - Requires high read count - Many reads may be from host - Requires reference genomes for classification 	<ul style="list-style-type: none"> * Profiling of what is present across all domains * Functional genome analyses * Phylogeny * Detection of pathogens
Meta-transcriptomics using sequencing of mRNA	<ul style="list-style-type: none"> + Identifies active genes and pathways - mRNA is unstable - Multiple purification and amplification steps can lead to more noise 	<ul style="list-style-type: none"> * Transcriptional profiling of what is active

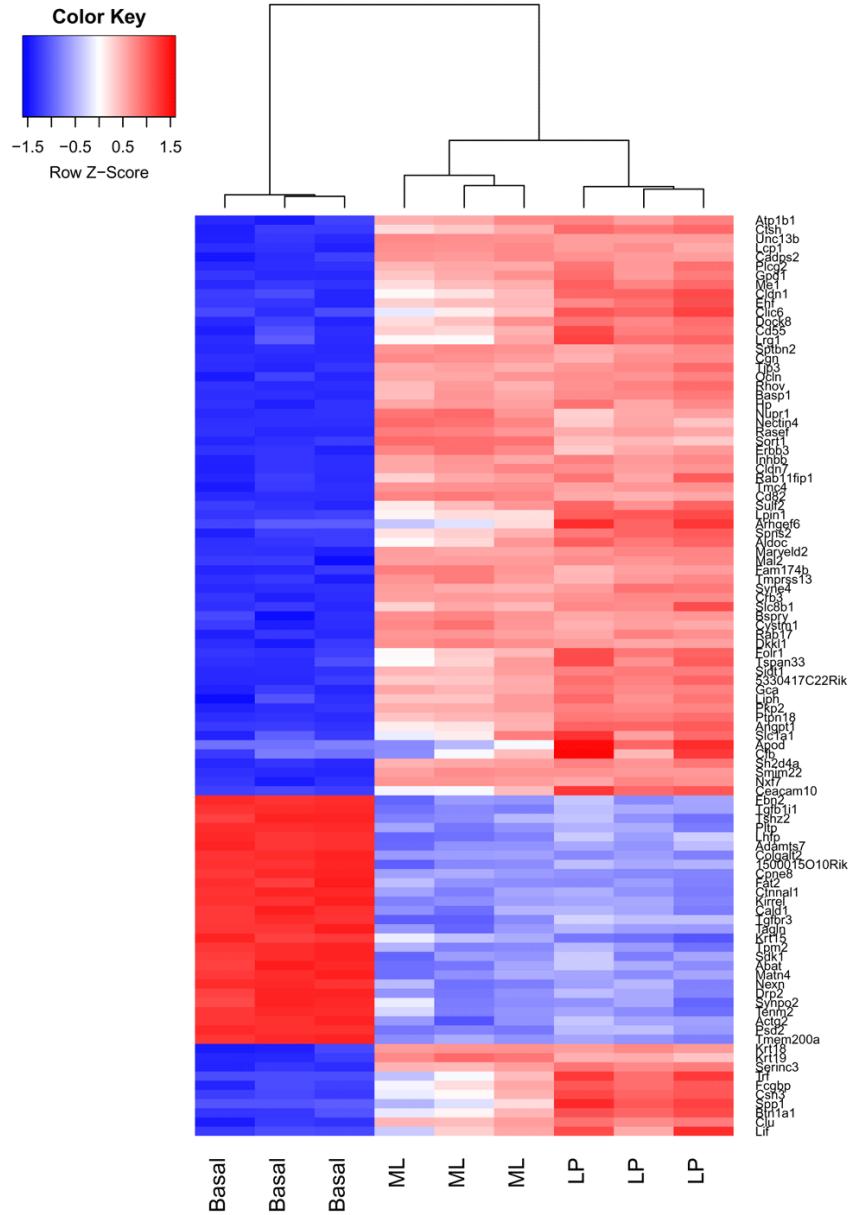
Projets métagénomiques

4900 projets sur NCBI (avril 2018)

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

Pourquoi faire du RNAseq ?

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



Application: Transcriptomics / RNASeq

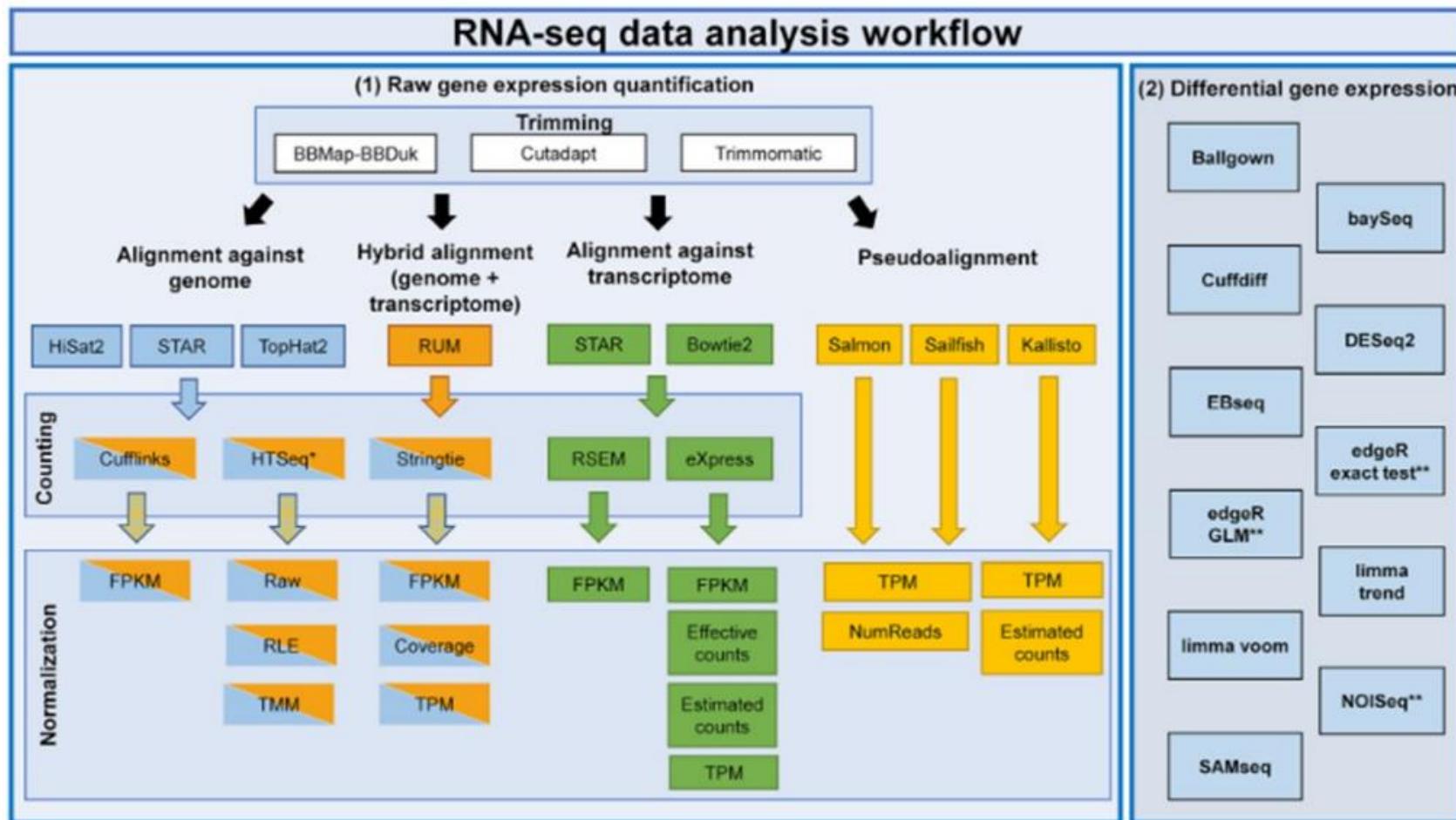


Figure 1. RNA-seq analysis workflow. Left panel (1) represents the raw gene expression quantification workflow. Every box contains the algorithms and methods used for the RNA-seq analysis at trimming, alignment, counting, normalization and pseudoalignment levels. The right panel (2) represents the algorithms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

⇒ 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 problématique



Studying Genetic Diversity using Single Nucleotidic Polymorphism (SNP)

Population Variation



Rice pathogene ?



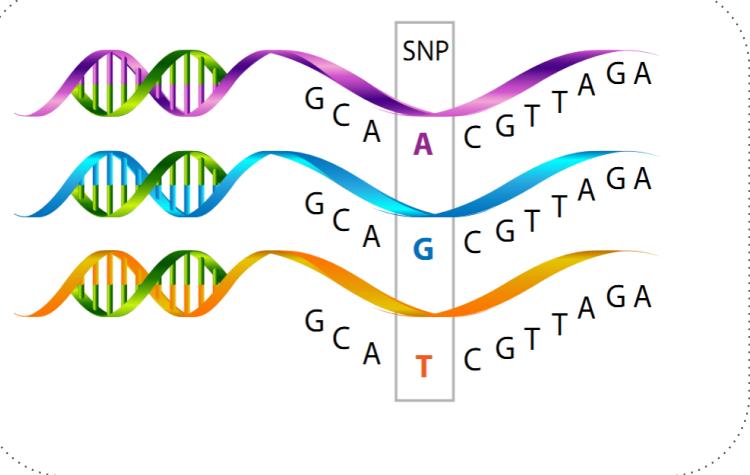
add
images

Understanding how individuals of a same species vary

- ✓ **Variations** between individuals
- ✓ **Natural selection** in a population
 - + Each individual = unique combination of traits
 - + Inherited variations that confer an advantage
(increasing an organism's chance of survival) will be passed to offspring

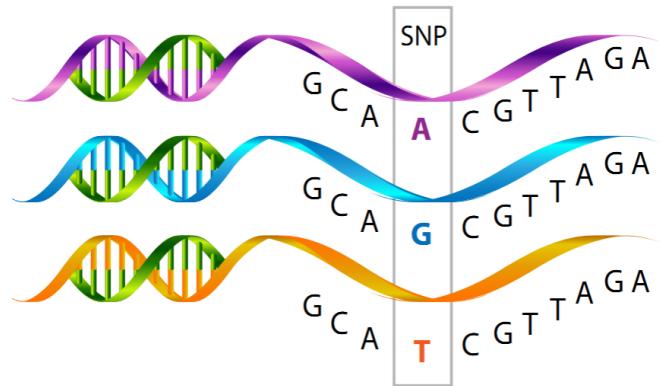
Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism

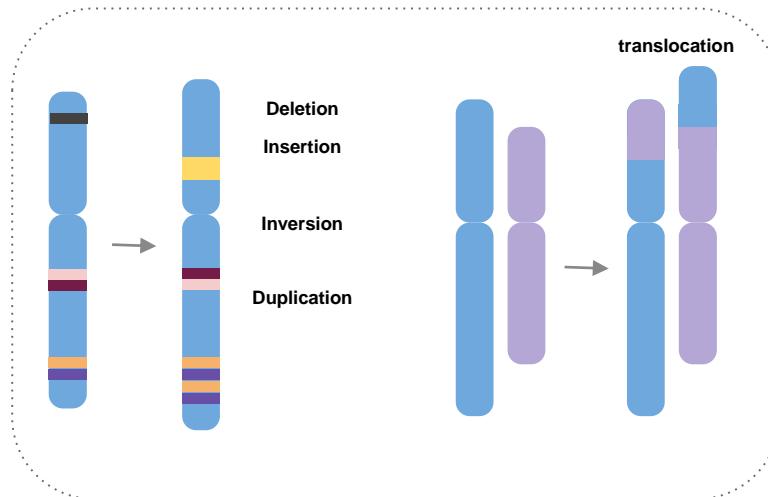


Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism

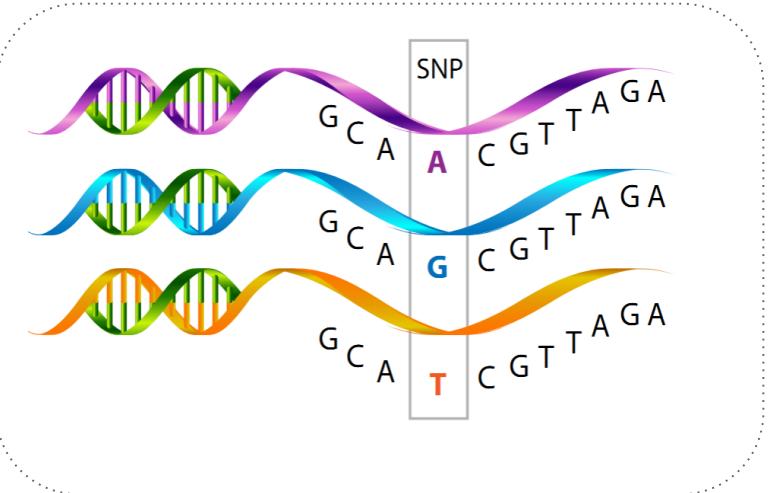


Structural Variations

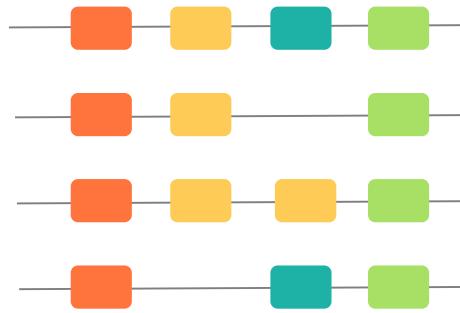


Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism



Structural Variations

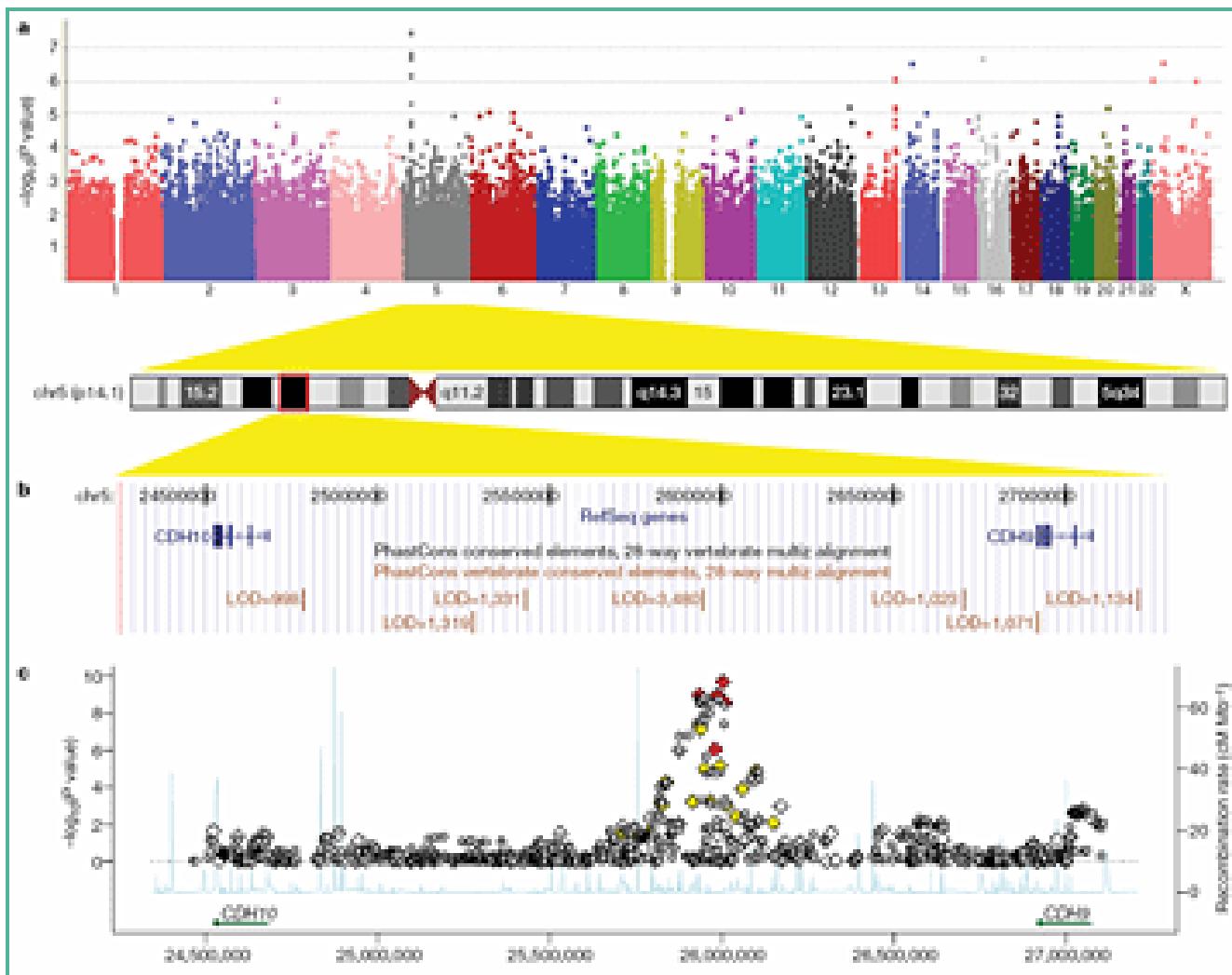


Presence Absence Variation (PAV)

Deletion, duplication, copy number variation, mobile element insertion

GWAS (Genome Wide Association Studies)

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



GWAS (Genome Wide Association Studies)

GWAS Diagram Browser

Exploring Genome-wide Association Studies

Filter:

[Clear filters](#)

[GWAS Diagram](#)

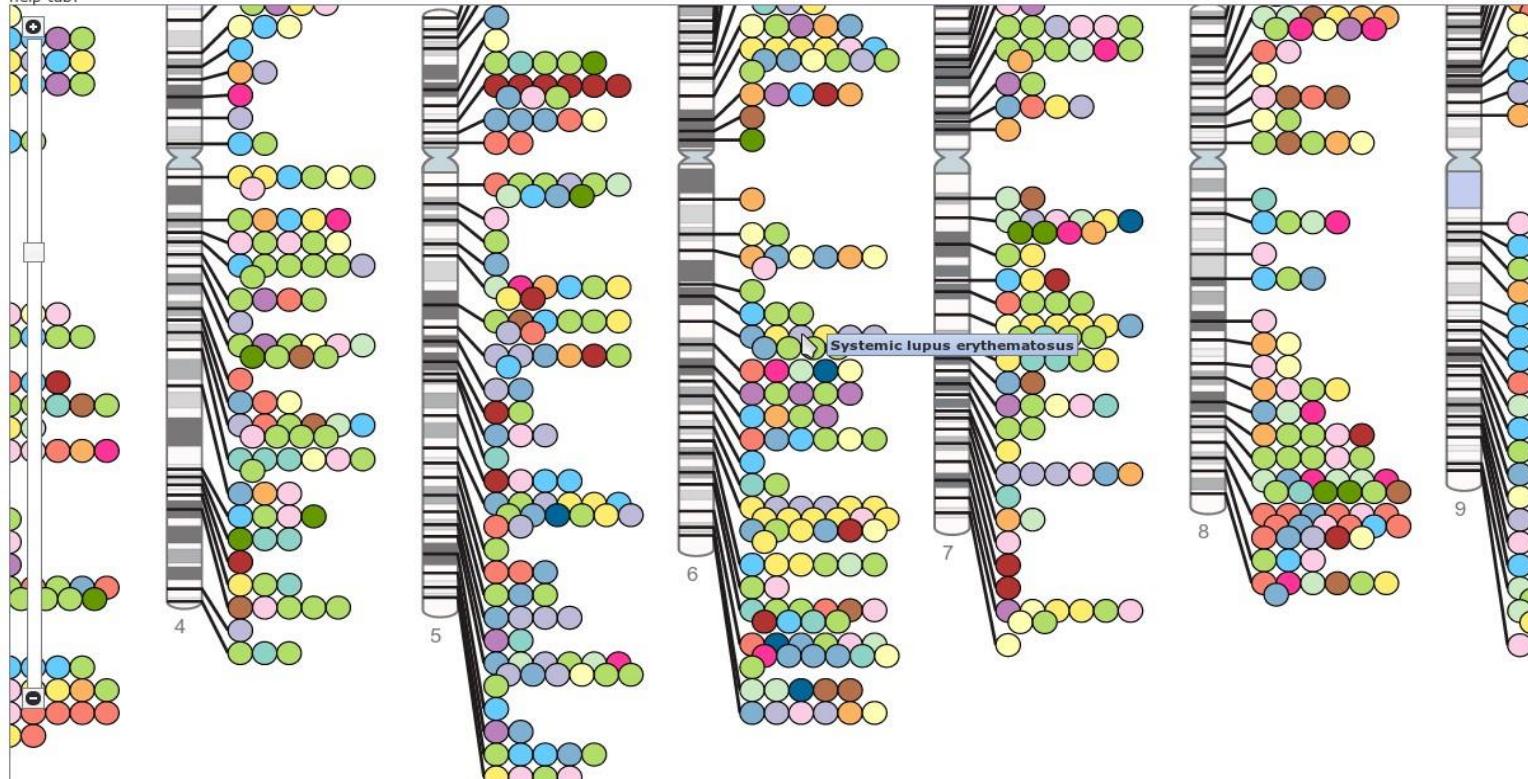
[Time Series View](#)

[Downloads](#)

[Help](#) [About](#)

[Show Legend](#)

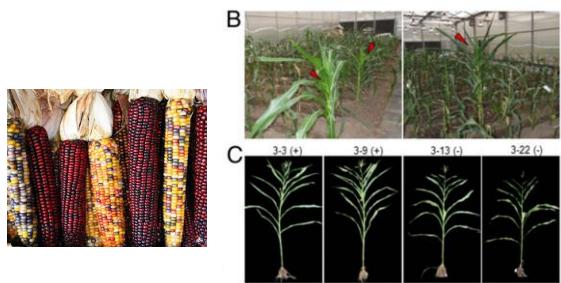
This diagram shows all SNP-trait associations with a p-value smaller than 5×10^{-8} , published in the catalogue up to the end of June 2012. For information on how to navigate the diagram, see the [help tab](#).



National Human
Genome Research
Institute

EMBL-EBI

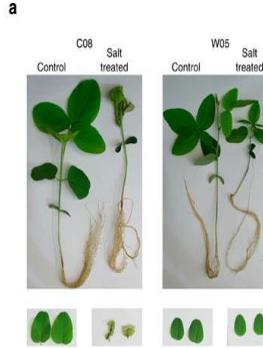




From Yang et al., 2013



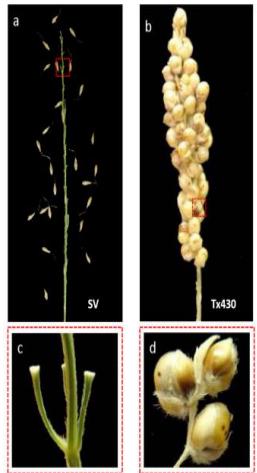
From Li et al. 2012



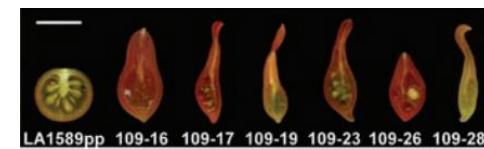
From Qi et al. 2014



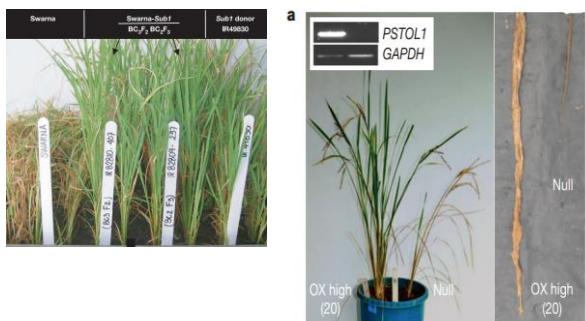
From Yang et al., 2014



From Lin et al. 2012

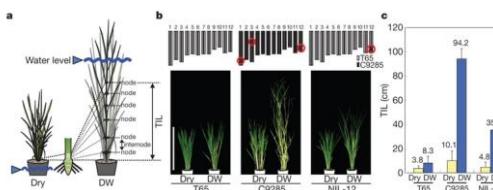


From Xiao et al. 2008

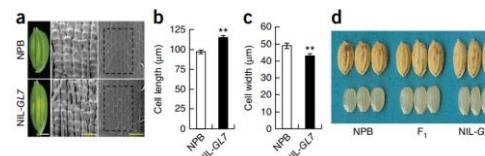


From Xu et al. 2006

From Gamuyao et al. 2012



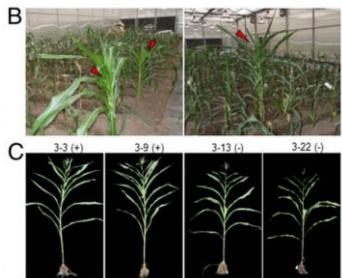
From Hattori et al. 2009



From Wang et al. 2015



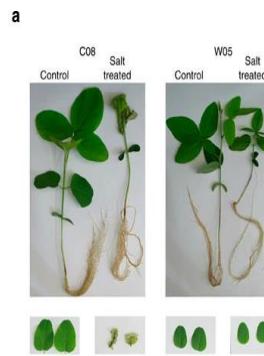
From Bai et al. 2017



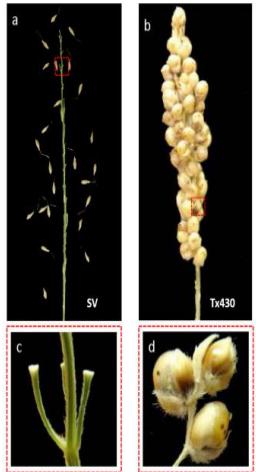
From Yang et al., 2013



From Li et al. 2012

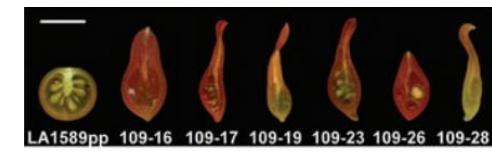


From Yang et al., 2014

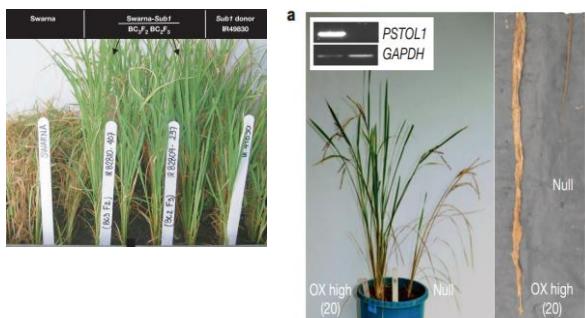


From Lin et al. 2012

Is One Reference genome enough to capture all genetic diversity ?

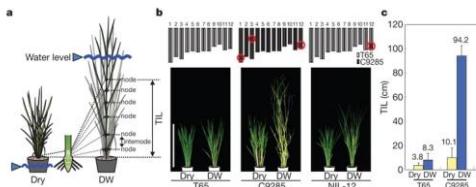


From Xiao et al. 2008

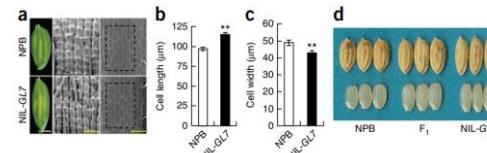


From Xu et al. 2006

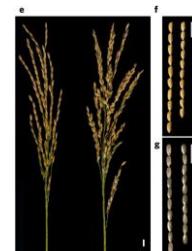
From Gamuyao et al. 2012



From Hattori et al. 2009



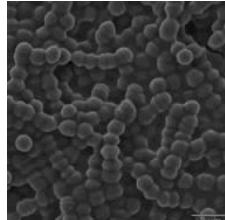
From Wang et al. 2015



From Bai et al. 2017

Gene number variations within a species

Streptococcus agalactiae



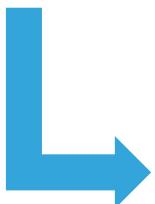
NAS

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

Hervé Tettelin^{a,b}, Vega Maignani^{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,

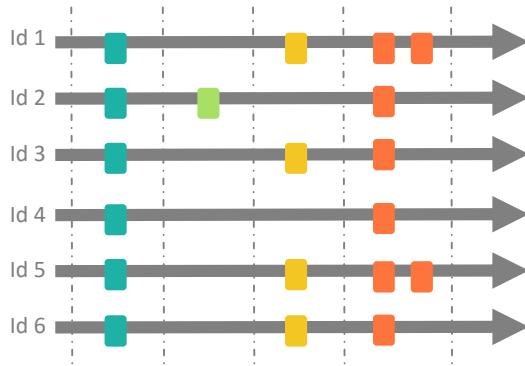
Tettelin et al., 2005

- ▶ 8 strains sequenced
- ▶ SNP variations



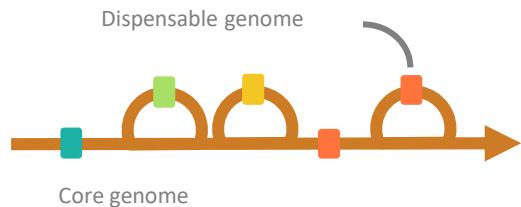
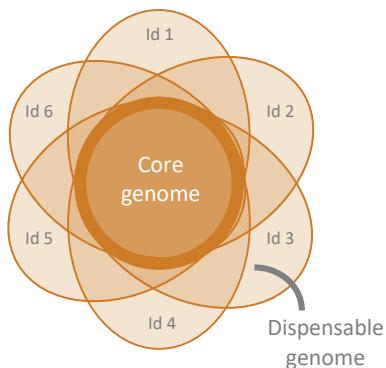
Large number of genes not shared between isolates
20% genome variability and 80 % shared by all isolates
Pangenome concept

Pangenome concept



Pangenome

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

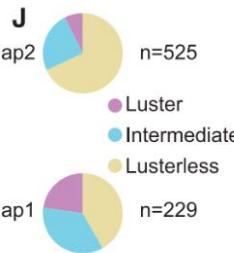
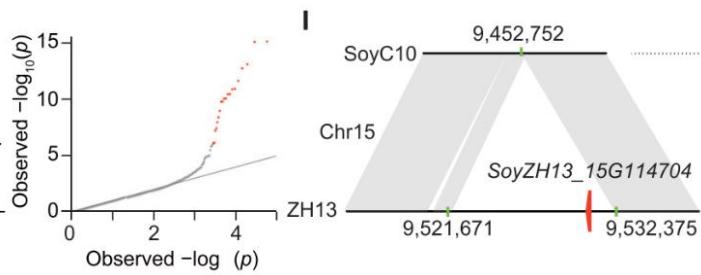
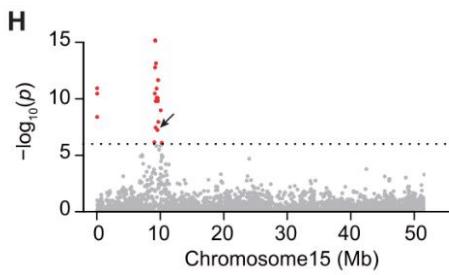


- ▶ **Core genome** : present in all individuals
- ▶ **Disposable genome** : absent from one or several individuals (also called variable, accessory,...)

What's else ?

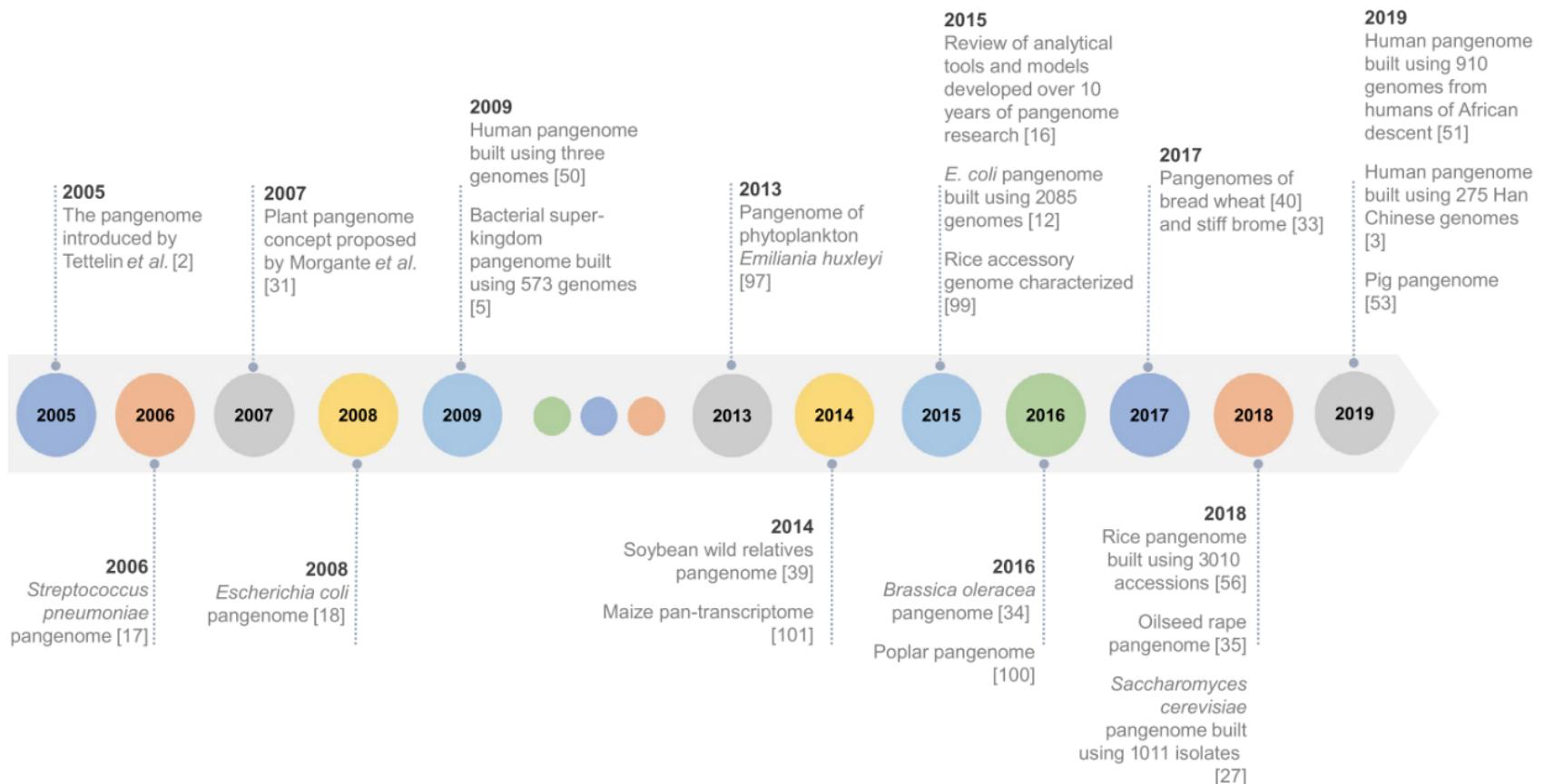


- ▶ 12,150 genes absent from the reference (18 cultivars)



From the first pangenome analyse by Tettelin & al.

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



Trends in Genetics

How to detect SV?



2 formations => 2 ambiances

Mode training mais bcp
de pratique par soi même
ou collectivement



Mode projet

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

How to detect SV?

C est le but de la formation

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



2 formations => 2 ambiances

Mode training mais bcp
de pratique par soi même
ou collectivement

Mode projet

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

How to detect SV?

C est le but de la formation

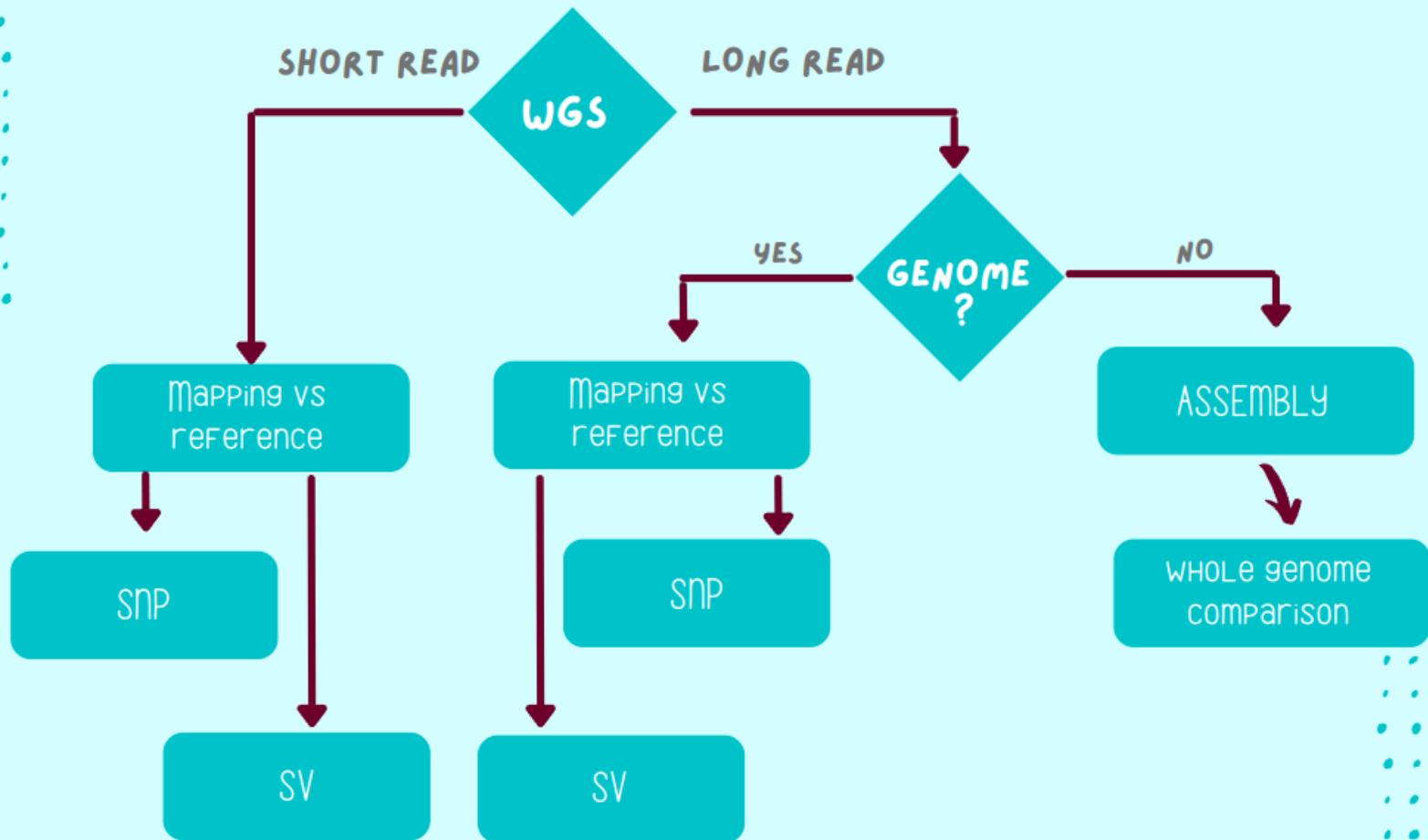
=> Protocole

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



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

SV DETECTION



Training



#data

What data will we use for our training ?



Diploid Asian Rice, *Oryza sativa*

Chara

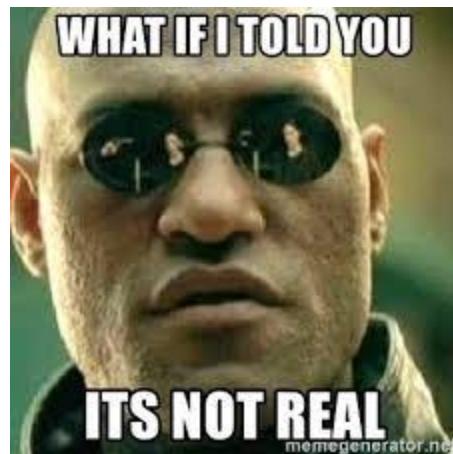
From
Wikimedia

What data will we use for our training ?

Diploid Asian Rice, *Oryza*



From
Wikimedia



What data will we use for our training ?

Diploid Asian Rice, *Oryza*



From
Wikimedia

add
images

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

illumina®

 Oxford
NANOPORETM
Technologies



Projet SNP



MISSION ~~IMPOSSIBLE~~
NOM DE CODE : "PROJET SNP"

Votre mission si vous l'acceptez...



#LIEU : Burkina Faso

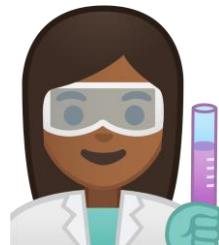




#MISSION :

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

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



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



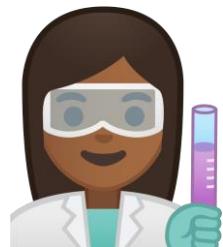
#MISSION :

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

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



Dominique

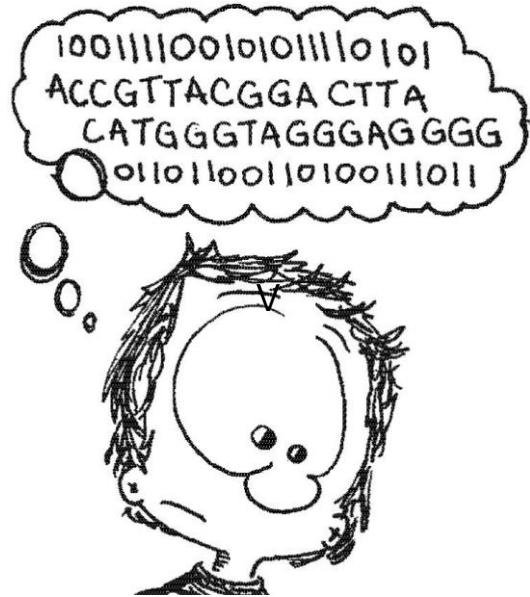


Je compte sur vous !!!!

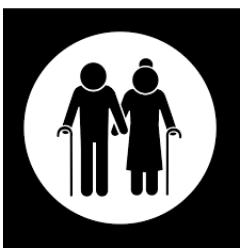


#DATA :

Décrire où ils trouvent les données



A vous de jouer !



Metagenomic



Bioinformatics resources

- 2 façons d'utiliser linux :

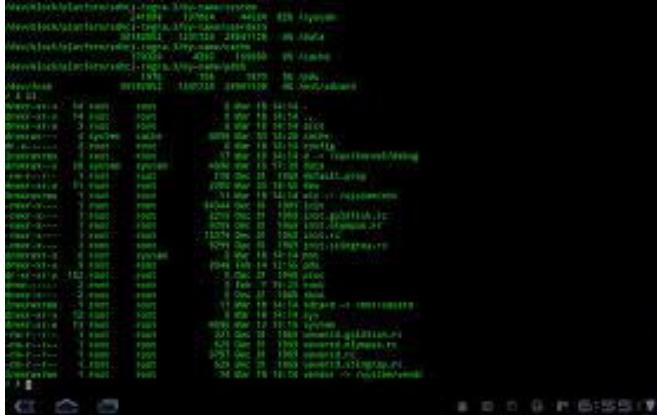
en *mode graphique*



- 2 façons d'utiliser linux :

en *mode graphique*

en *mode console*



A screenshot of a Linux desktop environment. In the foreground, there is a terminal window displaying a command-line interface with various commands and output. The desktop background is dark, and the taskbar at the bottom shows several icons.



Pourquoi utiliser Linux ?

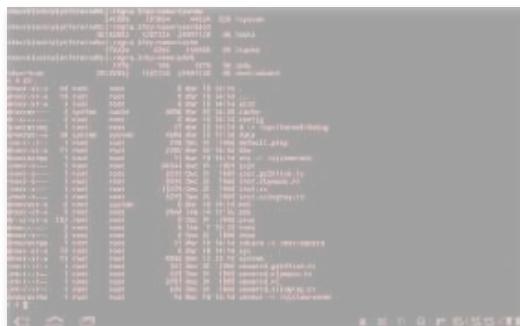


- Nombreux programmes rapides & puissants
- Facile de lier des commandes/programmes entre eux (workflow)
- Nombreux outils bioinformatique disponibles
- Pas besoin de ressources matérielles importantes
- 90% des serveurs fonctionnent sous Linux

Pourquoi utiliser Linux ?



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



Pas d'interfaces graphiques

Convivialité de la ligne de commande ?





Nécessité de la pratique et de l'expérience

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



Let's discover Jupyter !

Working environment

What is jupyter book ?

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

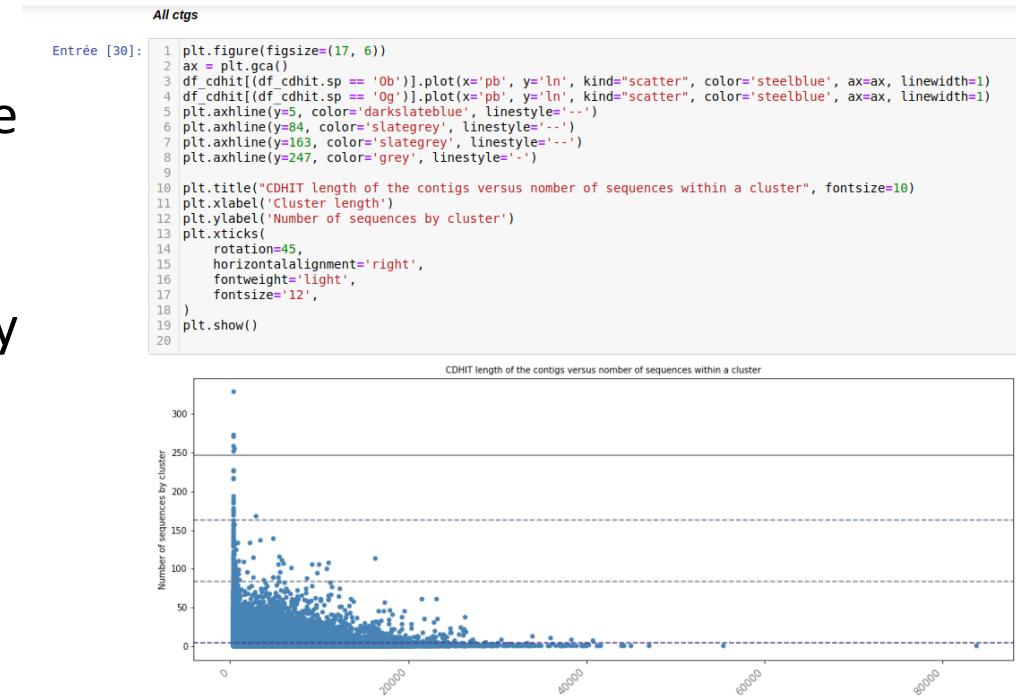
ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala



Why use jupyter book ?

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

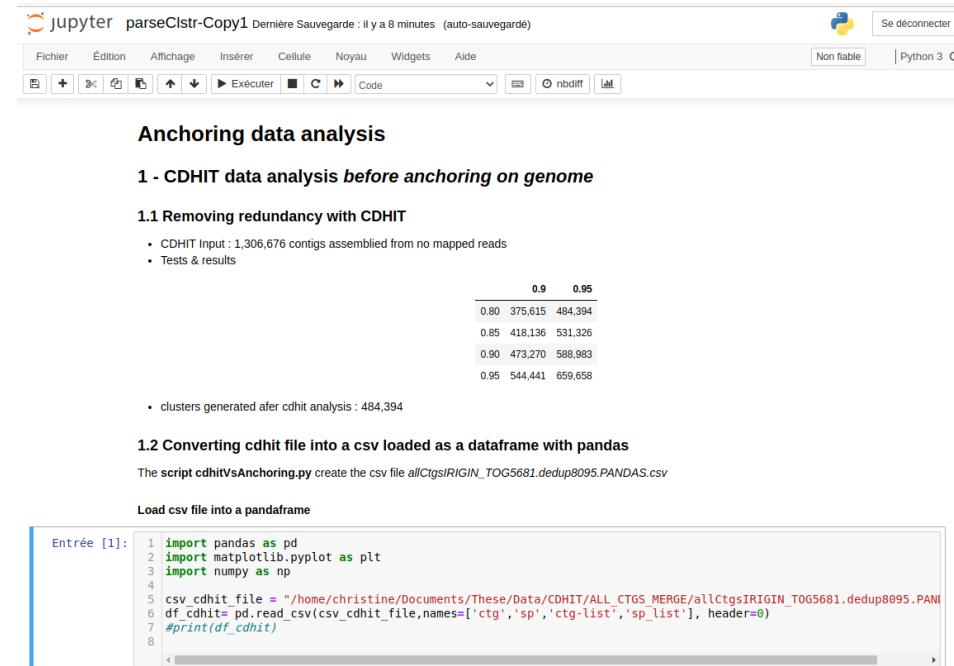
- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



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



The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** jupyter parseClstr-Copy1 Dernière Sauvegarde : il y a 8 minutes (auto-sauvegarde) | Se déconnecter | Non flable | Python 3 O
- Toolbar:** Fichier, Édition, Affichage, Insérer, Cellule, Noyau, Widgets, Aide.
- Cell Content:**
 - Section:** Anchoring data analysis
 - Section:** 1 - CDHIT data analysis before anchoring on genome
 - Section:** 1.1 Removing redundancy with CDHIT
 - CDHIT Input : 1,306,676 contigs assembled from no mapped reads
 - Tests & results
 - Data Table:**

	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 after cdhit analysis : 484,394
 - Section:** 1.2 Converting cdhit file into a csv loaded as a dataframe with pandas
 - Text:** The script cdhitVsAnchoring.py creates the csv file allCtgSIRIGIN_TOG5681.dedup8095.PANDAS.csv
 - Text:** Load csv file into a pandasframe
 - Code Cell:** Entrée [1]:

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
4
5 csv_cdhit_file = "/home/christine/Documents/These/Data/CDHIT/ALL_CTGS_MERGE/allCtgSIRIGIN_TOG5681.dedup8095.PANDAS.csv"
6 df_cdhit= pd.read_csv(csv_cdhit_file,names=['ctg','sp','ctg-list','sp_list'], header=0)
7 #print(df_cdhit)
```

Lab notebook for science data ?

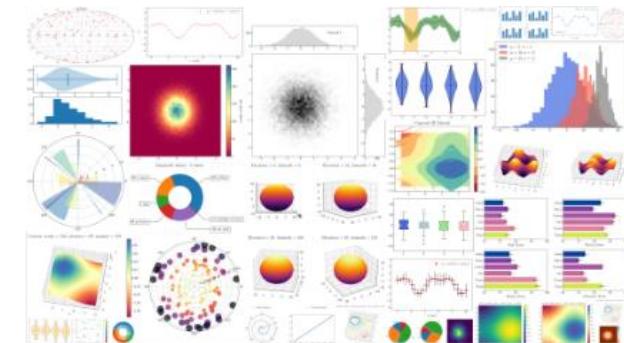
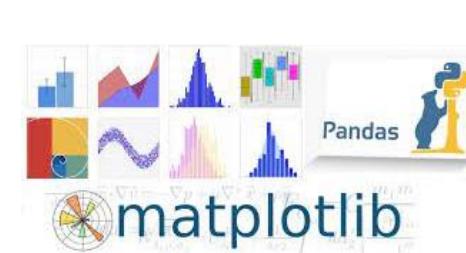
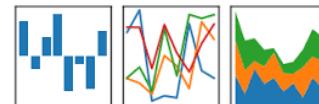


- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

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

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



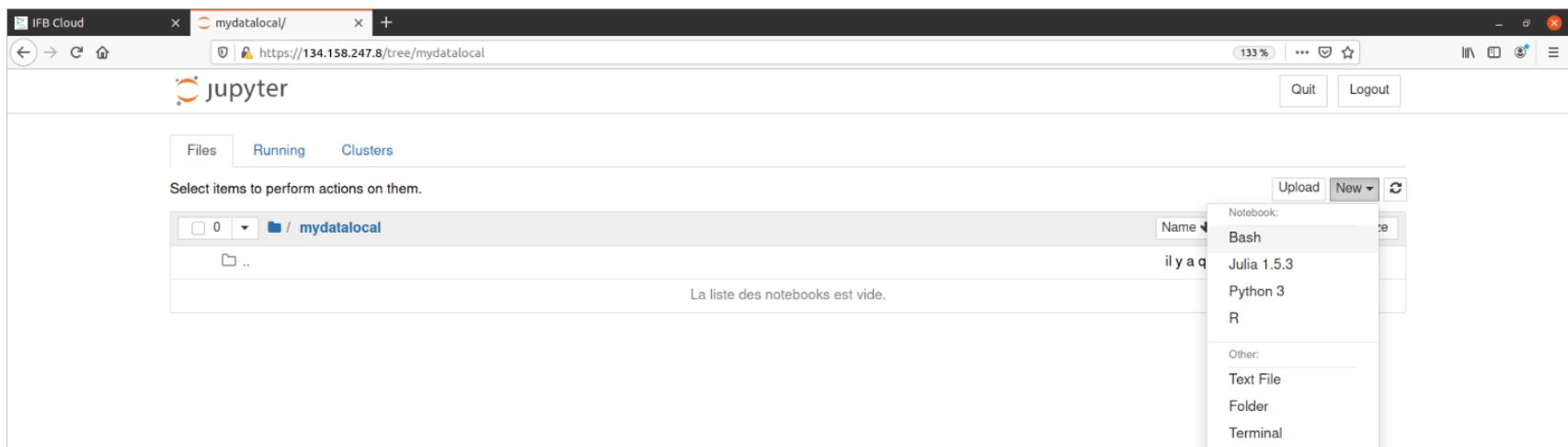
How will you use Jupyter Notebook ?

- 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

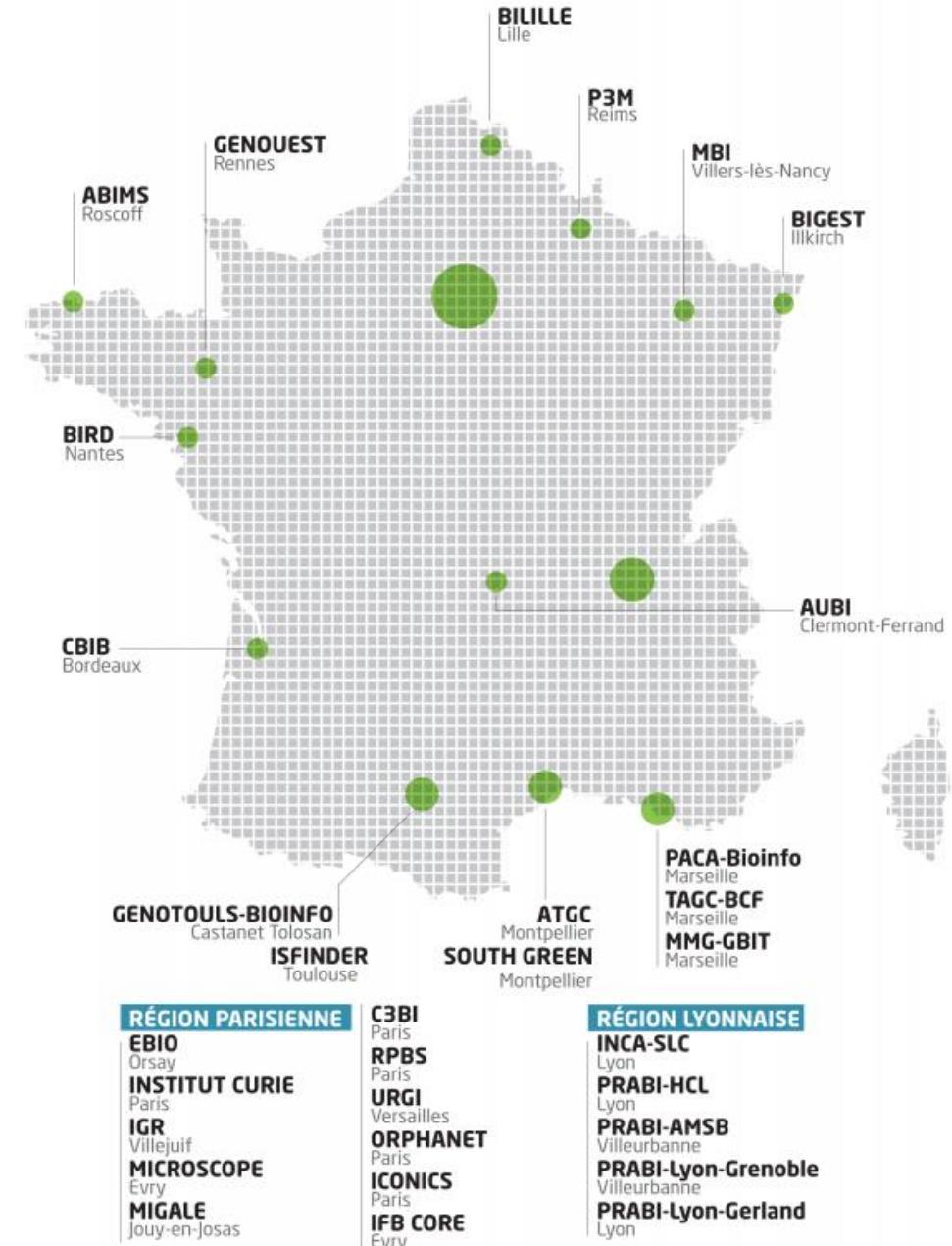


The screenshot shows the IFB Cloud web interface for managing Jupyter Notebooks. The browser window has a title bar "IFB Cloud" and a tab "mydatalocal". The address bar shows the URL "https://134.158.247.8/tree/mydatalocal". The main content area displays a Jupyter dashboard with three tabs: "Files", "Running", and "Clusters". Under the "Files" tab, there is a file tree showing a single folder named "mydatalocal". A message below the tree states "La liste des notebooks est vide." (The list of notebooks is empty). On the right side of the dashboard, there is a "New" button with a dropdown menu open. The dropdown menu lists several options for creating new notebooks: "Notebook:" (with "Bash", "Julia 1.5.3", "Python 3", and "R" listed), "Other:" (with "Text File", "Folder", and "Terminal" listed), and "Upload" (with a browse icon).



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

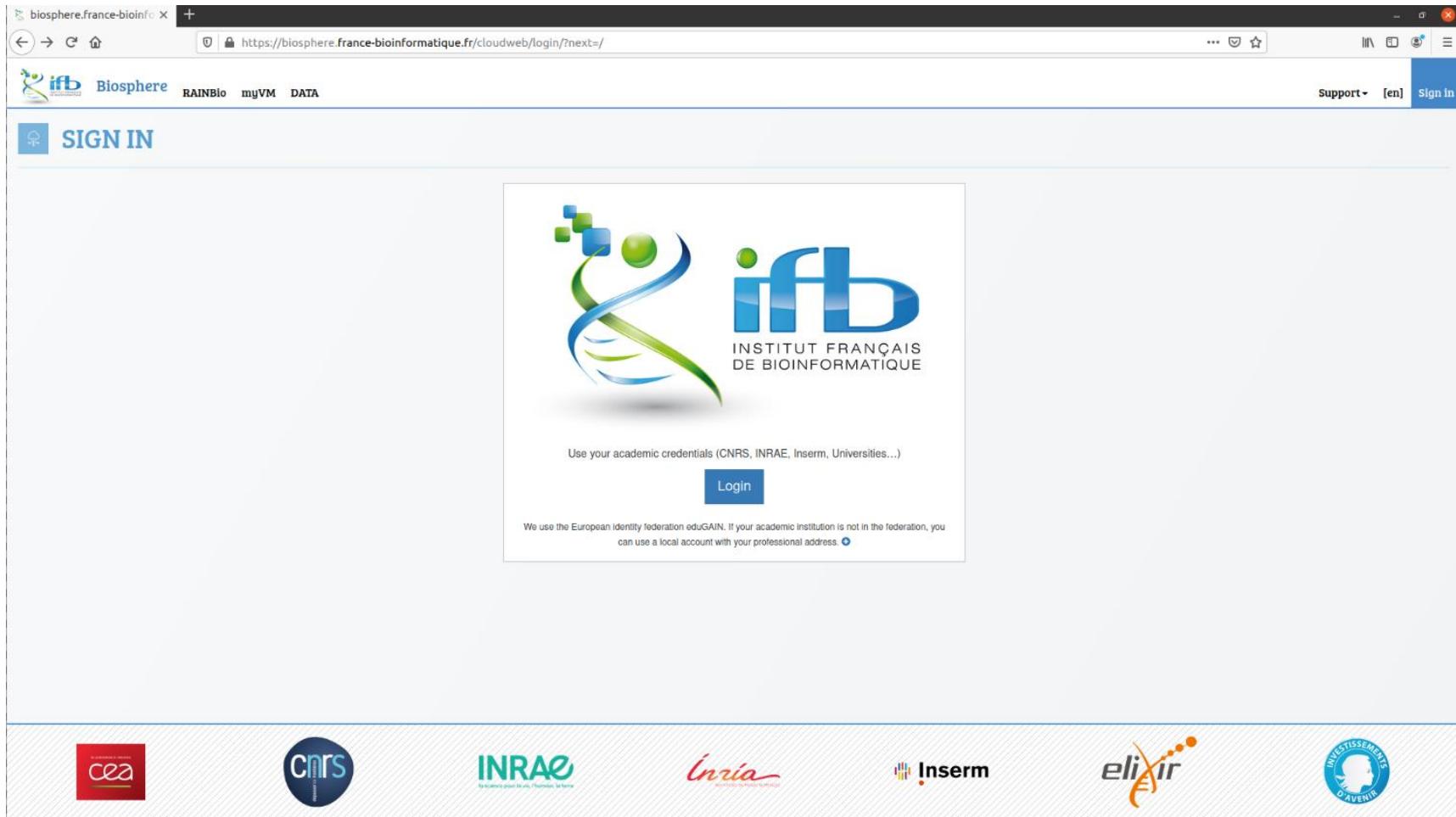
22 plateformes-membres
 7 plateformes contributrices
 8 équipes associées
 >400 experts (~200 FTE)



- 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 resources, 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.

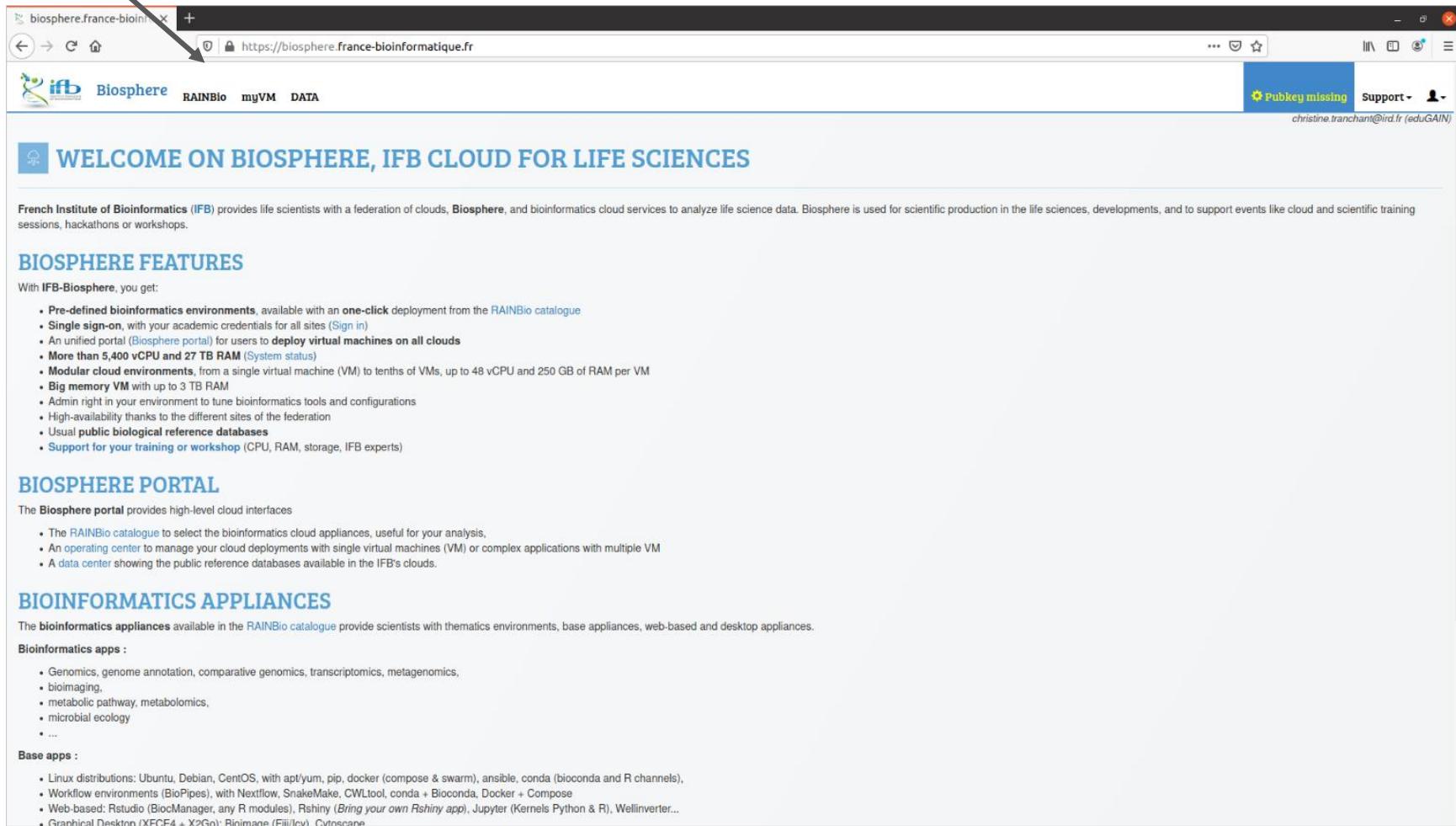
Let's start with biosphere

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



The screenshot shows a web browser window for the 'biosphere.france-bioinfo' website. The address bar displays the URL <https://biosphere.france-bioinformatique.fr/cloudweb/login/?next=/>. The page header includes the 'ifb' logo, the word 'Biosphere', and navigation links for 'RAINBio', 'myVM', and 'DATA'. On the right side of the header are 'Support' and 'Sign in' buttons. A large 'SIGN IN' button is prominently displayed on the left. In the center, there is a logo for 'ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE' featuring stylized green and blue shapes. Below the logo is a text box containing the instruction 'Use your academic credentials (CNRS, INRAE, Inserm, Universities...)'. A blue 'Login' button is centered below this text. At the bottom of the page, there is a note about using the European identity federation eduGAIN, followed by a link. The footer contains logos for various partners: CEA, CNRS, INRAE, Inria, Inserm, elixir, and Investissements d'Avenir.

RAINBIO catalog to access our Virtual Machine (VM)



RAINBIO catalog to access our Virtual Machine (VM)

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 thematic 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 (BioManager, any R modules), Rshiny (Bring your own Rshiny app), Jupyter (Kernels Python & R), Wellinverter...
- Graphical Deckton (XFCE4 + Xfce: Rainmane (Fiji/lev) Cythonize

Searching for the vm we will use

vm's name : analysesSV



The screenshot shows the RAINBio web interface. At the top, there is a navigation bar with tabs: IFB Biosphere, RAINBio (which is selected), myVM, and DATA. On the right side of the header, there is a message about a missing public key (Clé publique (PubKey) absente) and a support link (christine.tranchant@ird.fr (eduGAIN)). Below the header, a search bar contains the text "analyses". The main content area is titled "RAINBIO - APPLIANCES BIOINFORMATIQUES DANS LE CLOUD" and describes it as a catalogue of bioinformatics appliances in the cloud. A search filter on the right shows the term "analyses". Below the title, there are four appliance cards:

- AnalysesSV** (DEV): bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas, DNA polymorphism, Genetic variation.
- CoursAnalysesNanoporeSG**: bandage, Jupyter
- NGSanalysisJupyter**: BEDTools, BWA, Jupyter, SAMtools
- REPET**: Repet, Bioinformatics

A note at the bottom right states: "Le code couleur reste le même pour une même appliance."



Let's run your vm through the cloud

Appliance AnalysesSV ★ DEV

Exporter en md

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/datasience-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

- François 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>

Clé publique (PubKey) absente

christine.tranchant@ird.fr (eduGAIN)

LANCEZ

DÉPLOIEMENT AVANCÉ

Outils

bctools 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	François Sabot Southgreen
Développeurs	François Sabot Southgreen Julie Orjuela-Bouniol SouthGreen Platform

Let's run your vm through the cloud

IFB Biosphère RAINBio myVM DATA

Clé publique (PubKey) absente christine.tranchant@ird.fr (eduGAIN)

LANCER DÉPLOIEMENT AVANCÉ

Appliance AnalysesSV ★ DEV

Exporter en md

Description

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

This Jupyter app is based on the Jupyter Docker Stacks (see [details](#)). By default, users can choose any other existing stack with an Advanced deployment interface. 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

- François 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

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

vCPU.h

Cloud

ifb-core-cloudbis

Gabarit d'image cloud

ifb.m4.small (1 vCPU, 4Go GB RAM, 25Go GB local disk)

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 variantes structurales en cha

Let's run your vm through the cloud

Loading...

IFB Biosphère 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	

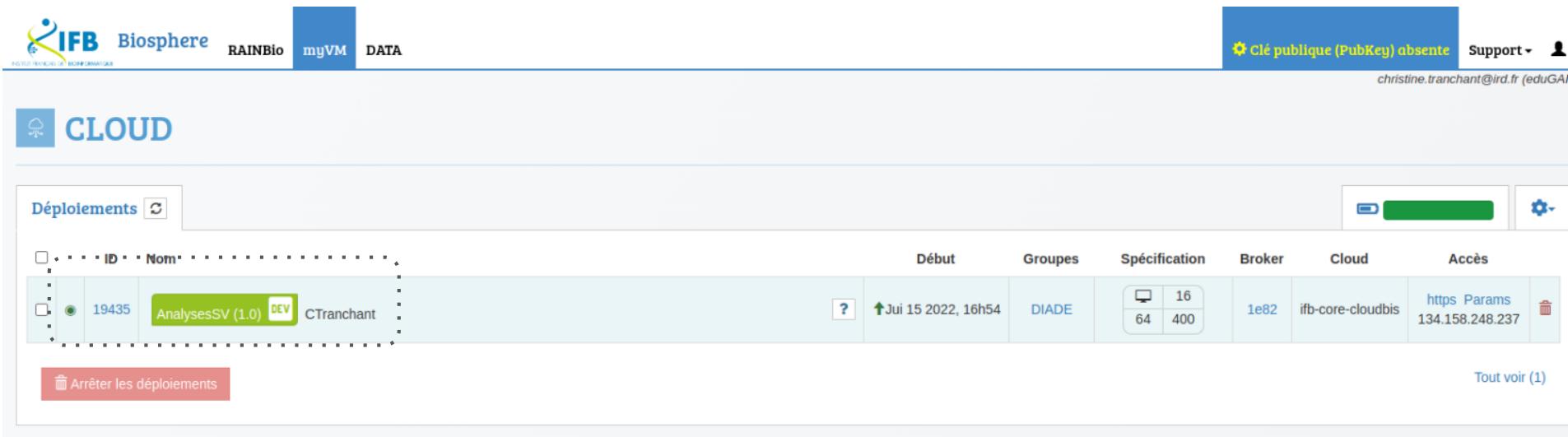
Arrêter les déploiements Tout voir (1)

Appliances et déploiements favoris Déploiements récemment terminés Quota

ID	Broker	Nom	Der. dém.	Paramétrage

Let's run your vm through the cloud

ready !



The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there is a navigation bar with the following items: IFB Biosphere (with a logo), RAINBio, myVM (selected), DATA, Clé publique (PubKey) absente (with a gear icon), and Support (with a user icon). The email address christine.tranchant@ird.fr (eduGAI) is also displayed.

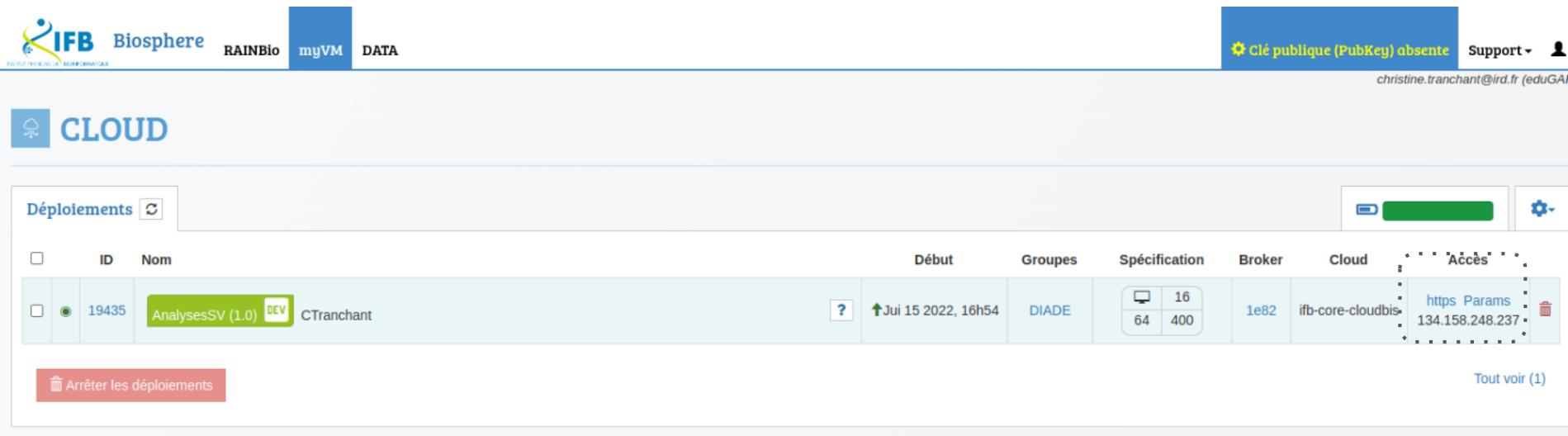
The main area is titled "CLOUD". It features a table for "Déploiements" (Deployments). The columns are: ID, Nom (Name), Début (Start), Groupes (Groups), Spécification (Specification), Broker, Cloud, and Accès (Access).

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

At the bottom left, there is a red button labeled "Arrêter les déploiements" (Stop deployments). At the bottom right, it says "Tout voir (1)" (View all 1).

Let's run your vm through the cloud

get the url... link “https”

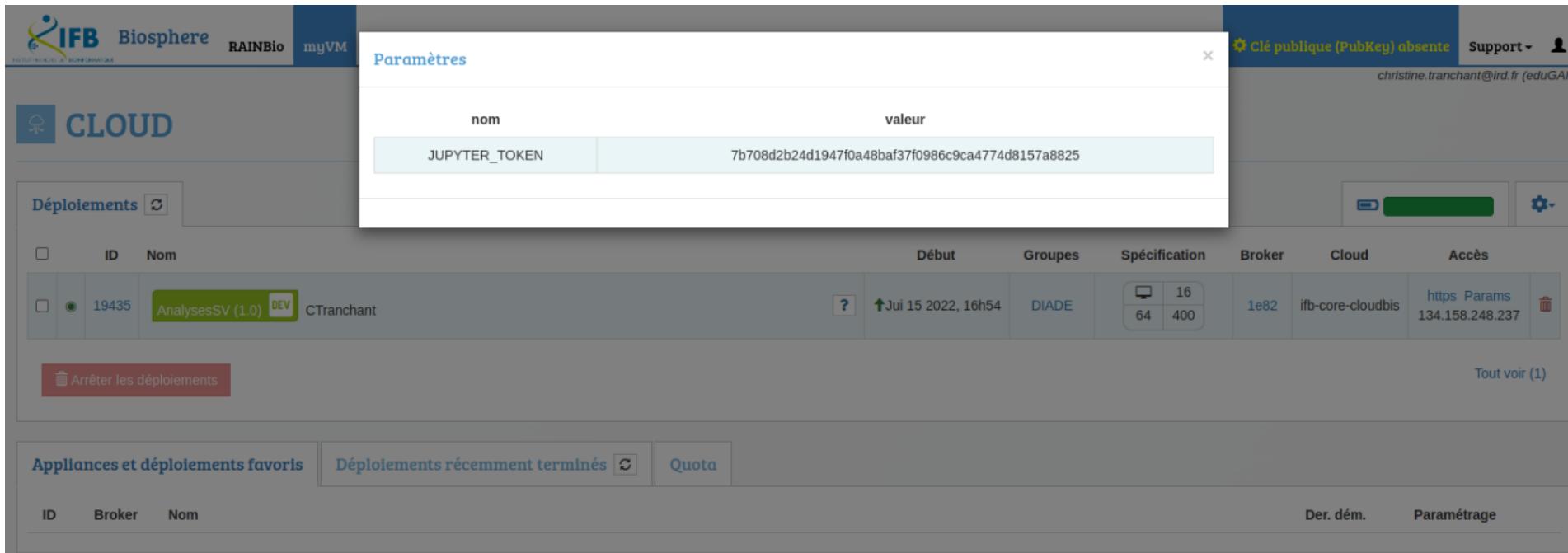


The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there is a navigation bar with the SouthGreen logo, followed by tabs for "IFB Biosphere", "RAINBio", "myVM", and "DATA". On the right side of the top bar, there is a message about a public key being absent ("Clé publique (PubKey) absente") and a support link ("christine.tranchant@ird.fr (eduGAI)"). Below the top bar, there is a large blue header with the word "CLOUD" in white. Underneath the header, there is a table titled "Déploiements" (Deployments). The table has columns for "ID", "Nom" (Name), "Début" (Start), "Groupes" (Groups), "Spécification" (Specification), "Broker", "Cloud", and "Accès" (Access). One deployment is listed: ID 19435, Name "AnalysesSV (1.0) DEV", Started on "Jui 15 2022, 16h54", Group "DIADE", Specification "16 64 400", Broker "1e82", Cloud "ifb-core-cloudbis", and Access "https Params 134.158.248.237". There is also a red button labeled "Arrêter les déploiements" (Stop deployments) and a link "Tout voir (1)" (View all 1).

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

Let's run our vm through the cloud

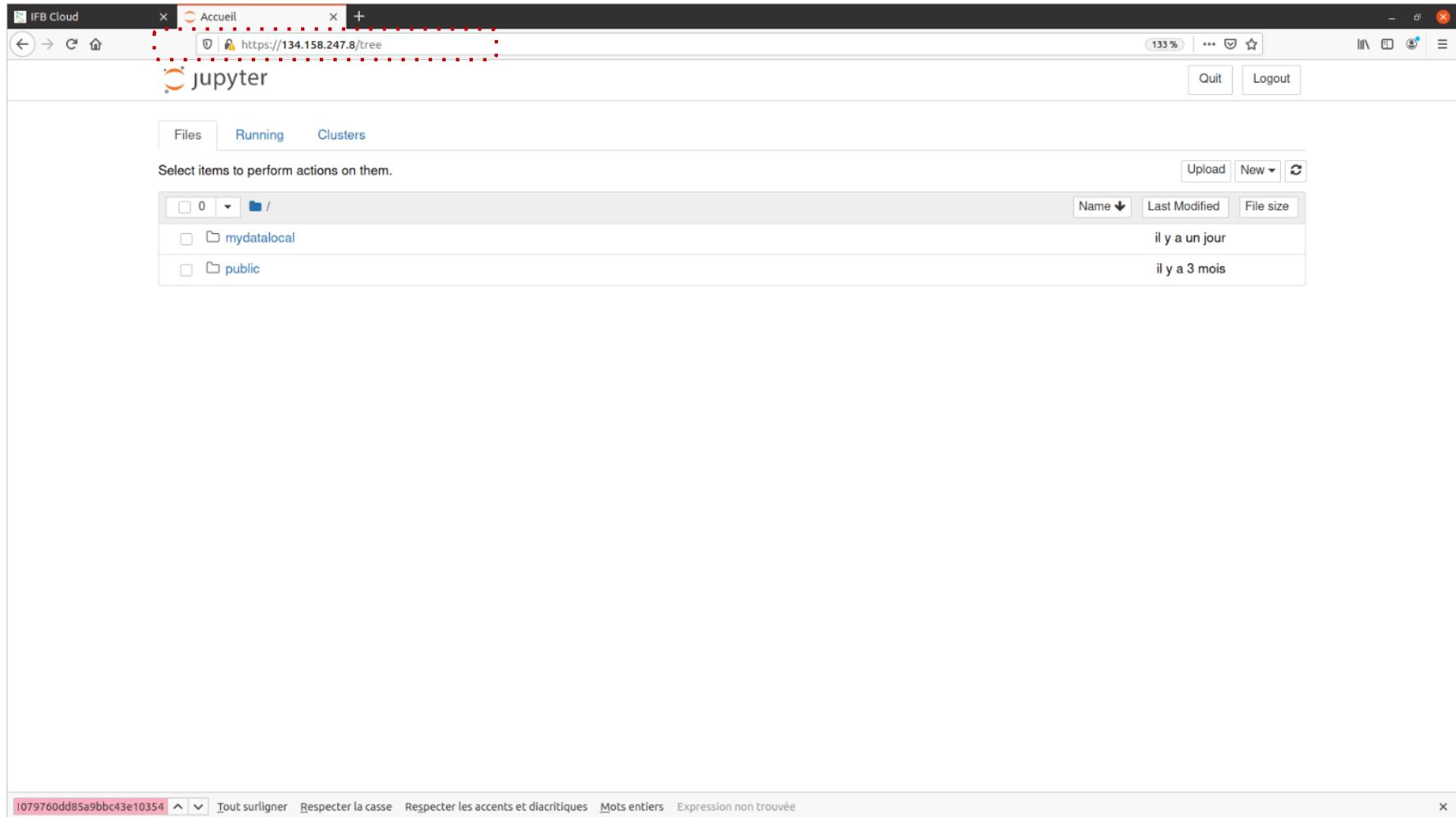
Get the token identifiant... link "Params"



The screenshot shows the RAINBio web interface. At the top, there are navigation tabs: IFB Biosphère, RAINBio, and myVM. On the left, there's a sidebar with a CLOUD icon. The main area displays deployment details for a job named 'AnalysesSV (1.0) DEV' with ID 19435, run by user CTranchant. A modal window titled 'Paramètres' is open, showing a single parameter entry: 'JUPYTER_TOKEN' with the value '7b708d2b24d1947f0a48baf37f0986c9ca4774d8157a8825'. The background shows other deployment lists and system status indicators.

Let's run our vm through the cloud

Open your vm ([https link](https://134.158.247.8/tree)) to access to your own jupyter lab



The screenshot shows a web-based interface for managing files in a Jupyter lab environment. The top navigation bar includes tabs for 'Files' (selected), 'Running', and 'Clusters'. The address bar shows the URL <https://134.158.247.8/tree>. On the right side of the header are 'Quit' and 'Logout' buttons. Below the header is a search bar with placeholder text 'Select items to perform actions on them.' and buttons for 'Upload', 'New', and a refresh icon.

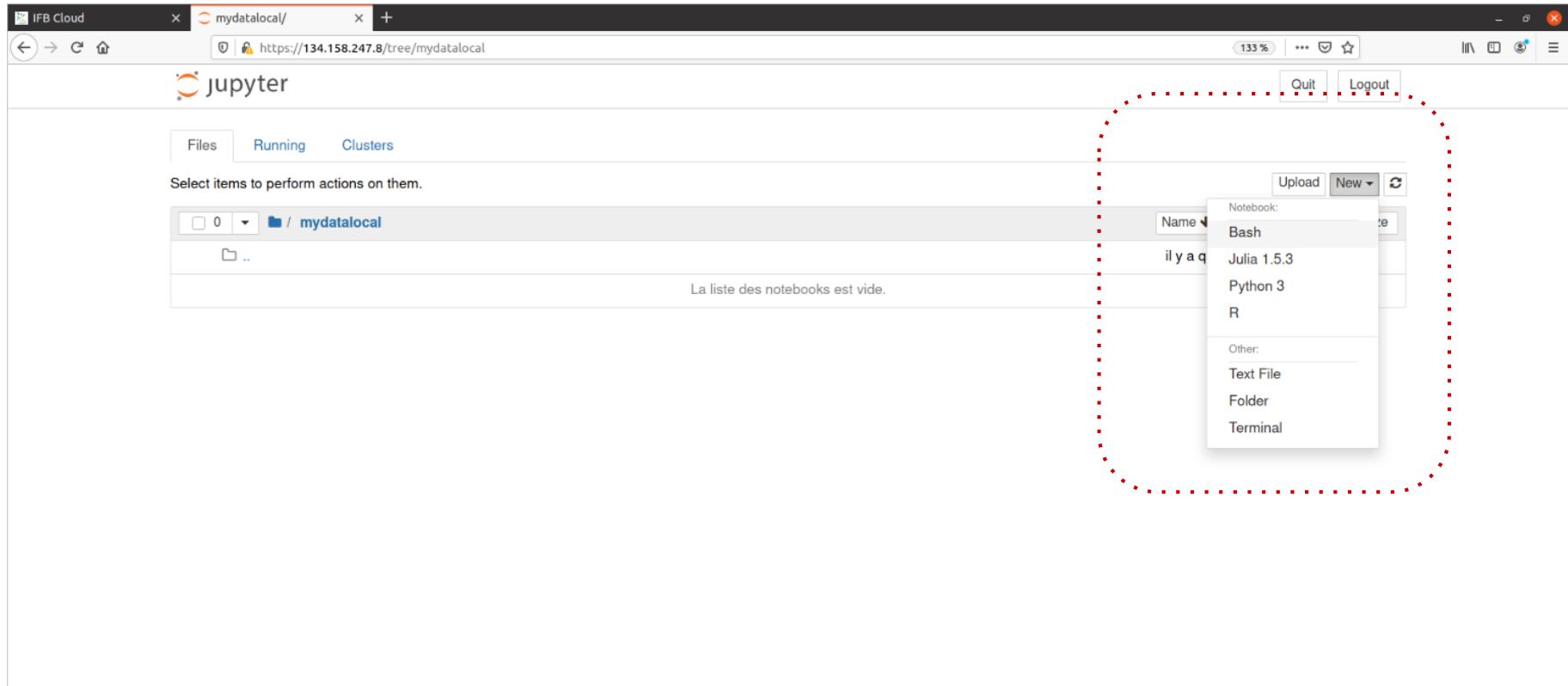
The main content area displays a file tree with two folders: 'mydatalocal' (modified 'il y a un jour') and 'public' (modified 'il y a 3 mois'). A dropdown menu shows '0' items selected. To the right of the file list are sorting options: 'Name' (sorted by name), 'Last Modified' (sorted by last modified date), and 'File size' (sorted by file size).

At the bottom of the page, there is a search bar containing the text '1079760dd85a9bbc43e10354' and several search filters: 'Tout surigner', 'Respecter la casse', 'Respecter les accents et diacritiques', 'Mots entiers', and 'Expression non trouvée'.

Create your first jupyter book

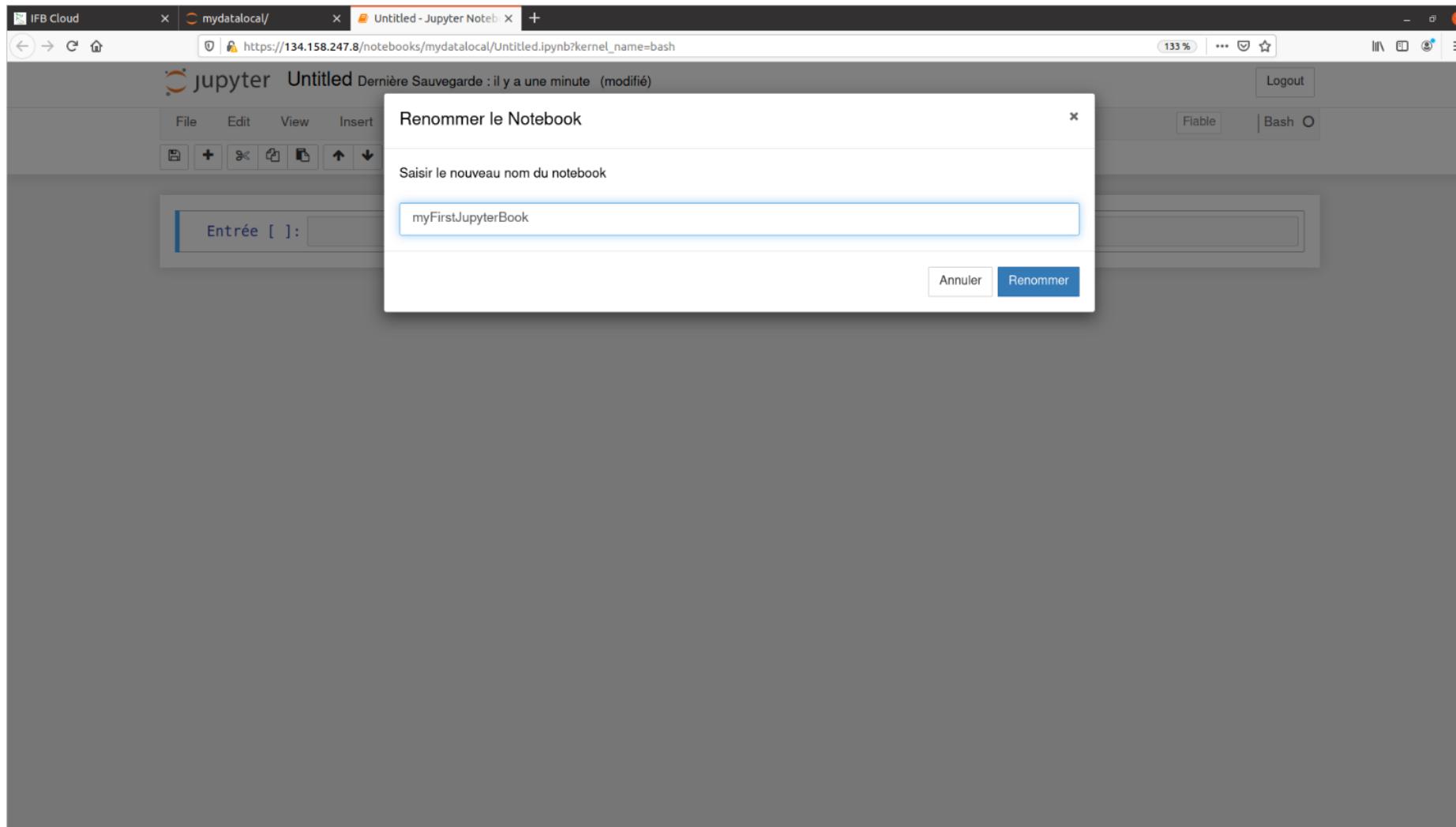
Go into the directory “work” and create a new jupyter book

-> kernel : bash



Rename your first jupyter book

myFirstJupyterBook

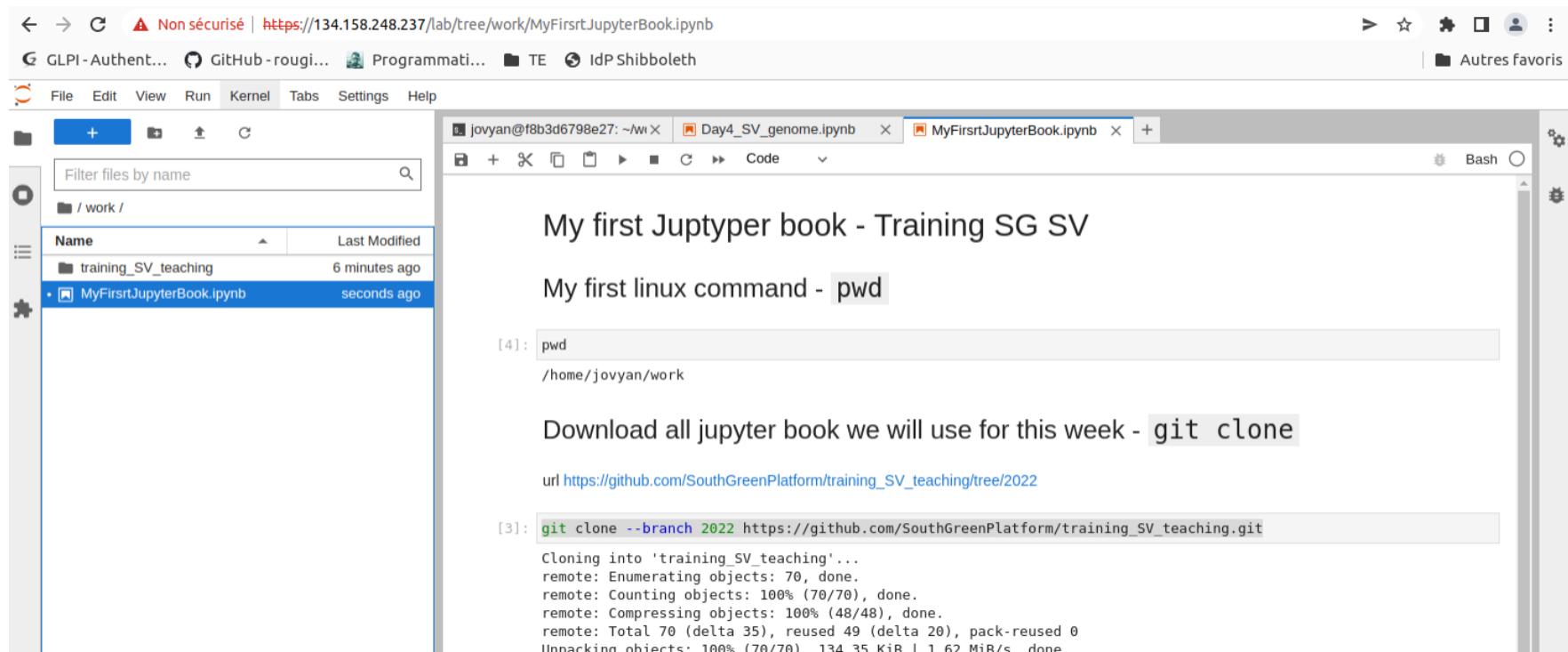


Run your first bask command - *git clone*

- All jupyterbook used for practice are here :
https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022
- Download all the jupyter books with the command *git clone*

```
git clone --branch 2022_burkina
```

```
https://github.com/SouthGreenPlatform/training\_SV\_teaching.git
```



The screenshot shows a Jupyter Notebook interface. On the left, there is a file browser window titled 'work' showing two files: 'training_SV_teaching.ipynb' (modified 6 minutes ago) and 'MyFirstJupyterBook.ipynb' (modified seconds ago). The main area displays a Jupyter notebook cell with the title 'My first Juptyer book - Training SG SV'. Below it, another cell has the title 'My first linux command - pwd'. A terminal window is open at the bottom, showing the command 'pwd' and its output '/home/jovyan/work'. A code cell at the bottom contains the command 'git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git'. The terminal output for this command shows the cloning process, including object enumeration, counting, compressing, and unpacking.

```
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.
```




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

Alexis Dereeper - UMR PHIM

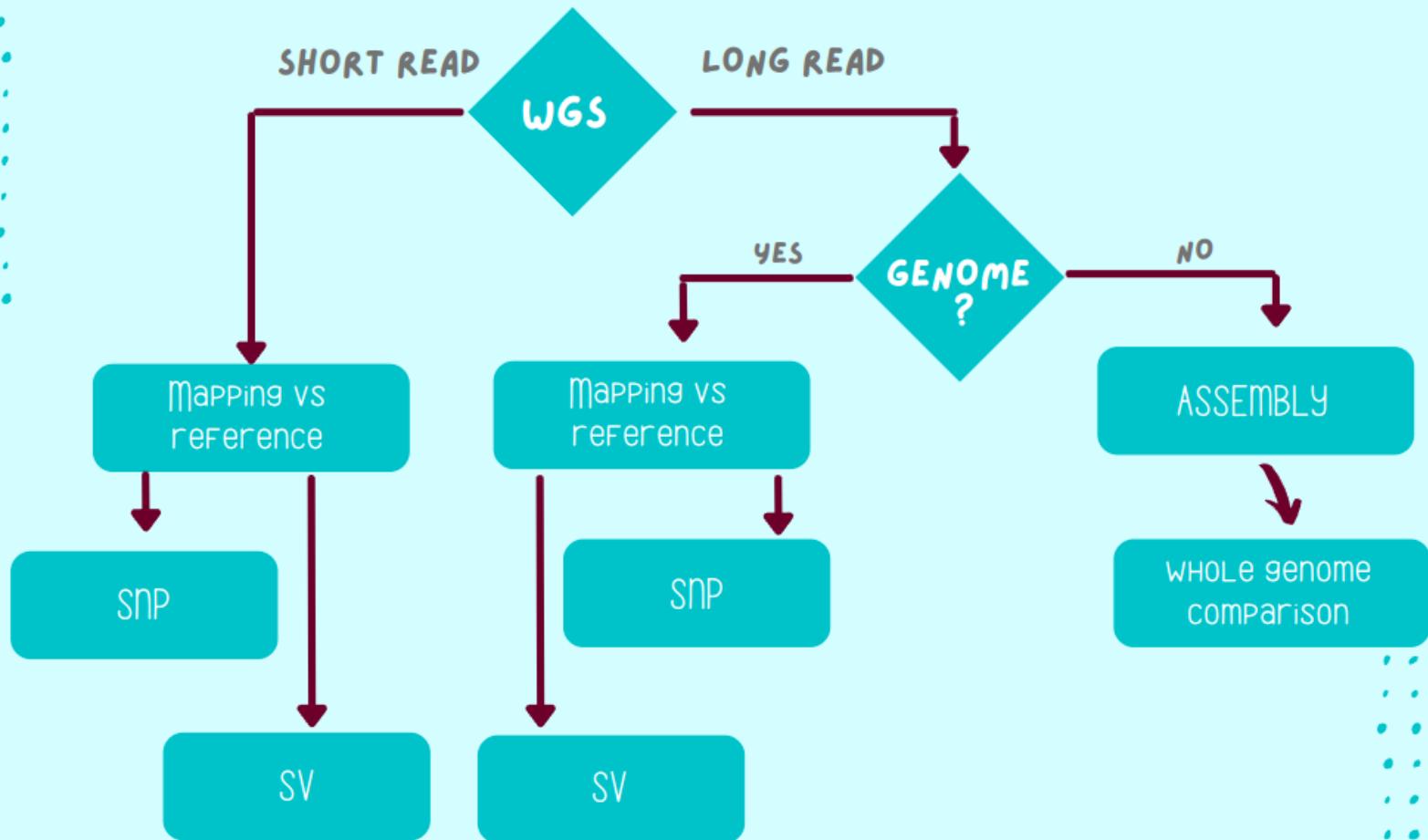
Julie Orjuela - UMR DIADE

Christine Tranchant-Dubreuil - UMR DIADE



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

SV DETECTION



Objectifs

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

Applications :

- Mapper des reads contre un génome *bwa*
- Déetecter 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

A small image of a Minion character from the movie Despicable Me, wearing a blue shirt and glasses, sitting at a computer keyboard.

RAW SEQUENCING DATA

OVERVIEW OF DNA SEQUENCING PROJECT



- Statistics
- Sequencing quality ?
- Adaptators ? Contaminants ?

fastq format

```

@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTCTTAGTATTTGTAGTCATTCCGTGTTGGTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIIEFEGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTG
+
@@@DDDD>FFFABEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTGCAGTAACGACCTATAAATTAAAGTAGC
+
@CCCCFFGGHHHHJJJJIEE<HHHJJIGBHGGEEIIJEIEIJIHHJFIIJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTCGGGTATGGCTCTAGTAGCTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIEEHIII<CGHIJIIJ?:FC9DGAFGHII?DGBFIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFGHHHHJJIIJJIIIGJCHHEIII>GHIG@GHIDHGJIIFHIIJJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGAAGAGTTAAAAACAAATTATGAGCAACTGAGTTC
+
@@@CCCCFFHFFHFGIJJIIHIIJJIIHJJECGHIIJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTGTCGGGGTTGGAAATTGTCACTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHGG;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 Ewing 1998

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

- 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
CTGTTCTTAGTATTTTGATGTCATTCCGTGGTTAGTTGCAAGGT
+
@@@DADFFHHFFHIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTGGT
+
@@@DDDD>FFFABEABB4C+3?:CBB@<>A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTCAGTAACGACCTATAAATTAAAGTAGC
+
@CFFFFFGHHHHJJJJIEE<HHHJJJIGBHGGEEIIJJIEIEIJHHJFJJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTCGGGTATGGCTCTAGTAGCTTTGTCTTAT
+
@C@FFFFFGHHHDHGIIEEHII<CGHIJIIJ?:FC9DGAFGHII?DGBFIIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@>DFFFFGHHHHJJJJJJIIIGJCHHEIII>GHIG@GHIDHGJIIFHJJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGTGTGAAGAGTTAAAAACAATTATGAGCAACTGAGTTC
+
@@@CFFFFFHFFHFGIJJHIIJJJJIIJJECGHJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTGTCGGGGTTGGAAATTGTCACTTCTGCTACAATGCCG
+
@<@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 _



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 ? Adapters ?
- 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



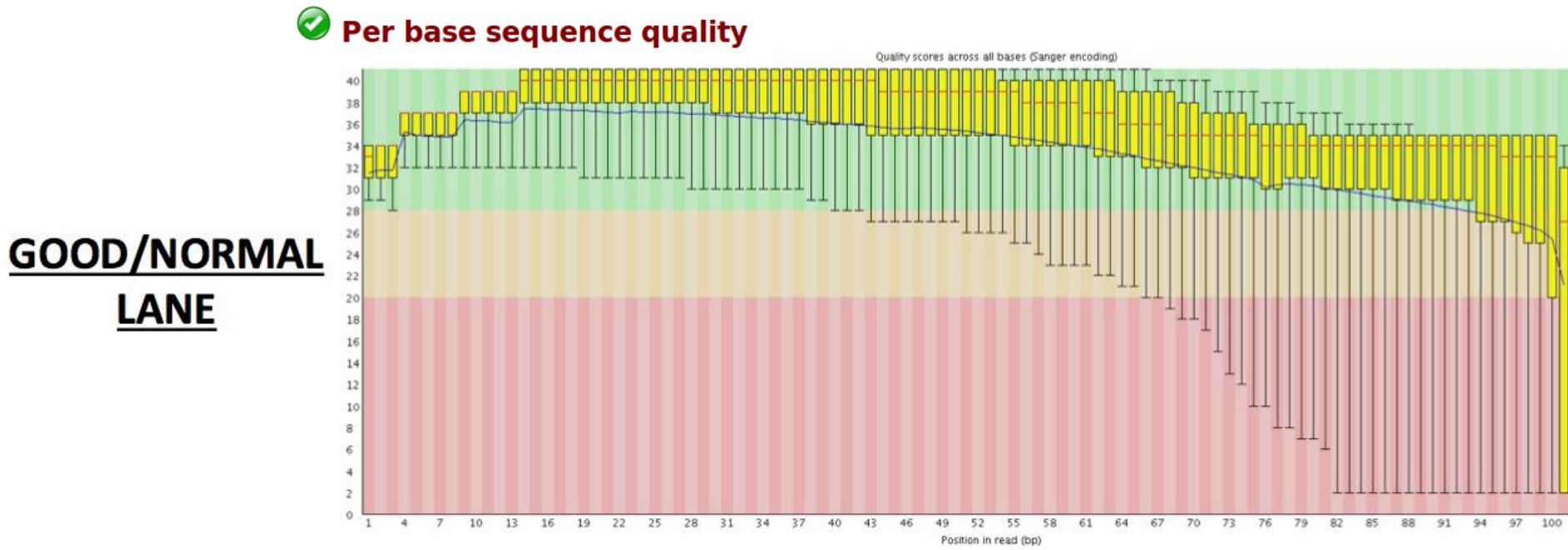
Basic Statistics

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

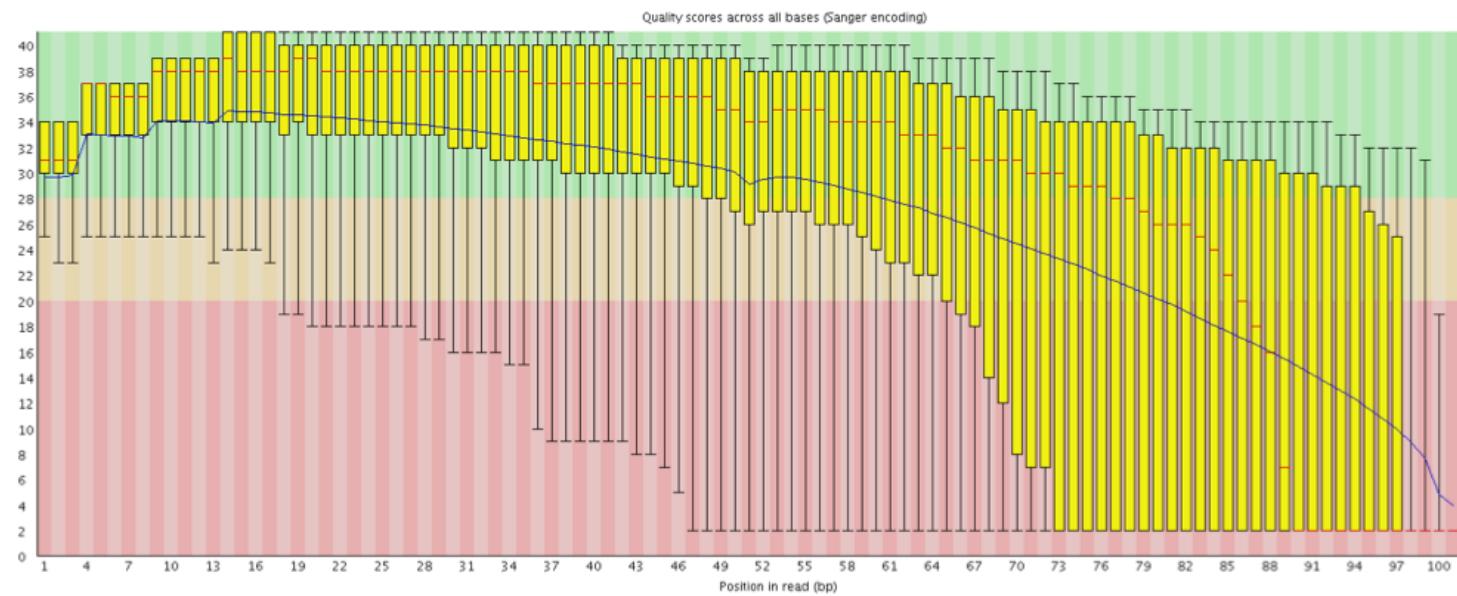
This plot shows the base quality score distribution for all reads in a lane, with each read position considered independently.

- 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

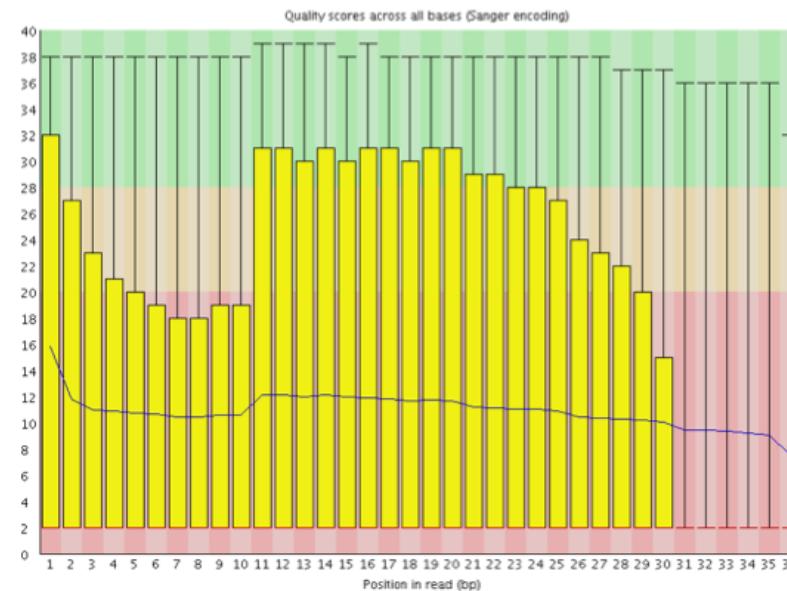


FastQC : Per base sequence quality

SALVAGEABLE
LANE



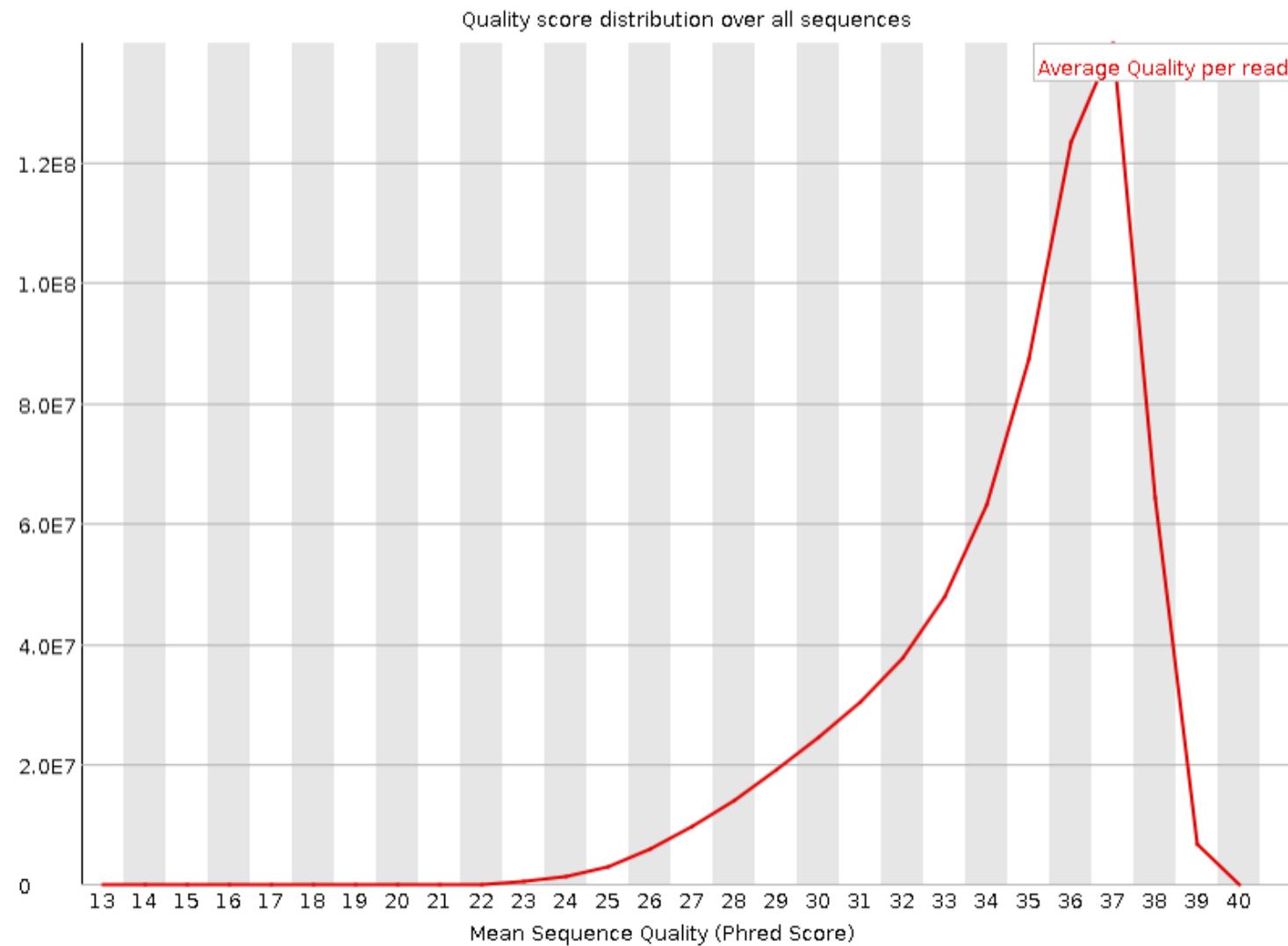
FAILED LANE



FastQC: Per sequence quality scores



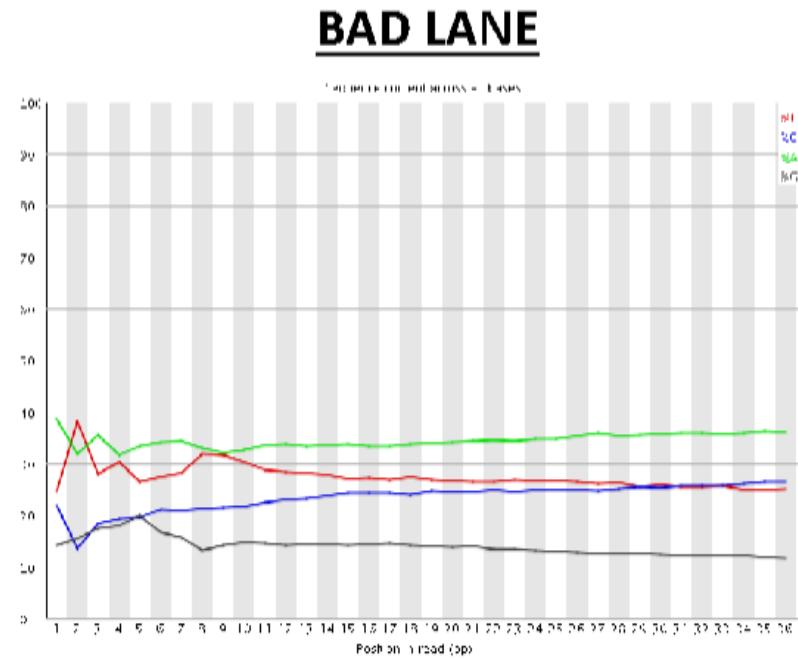
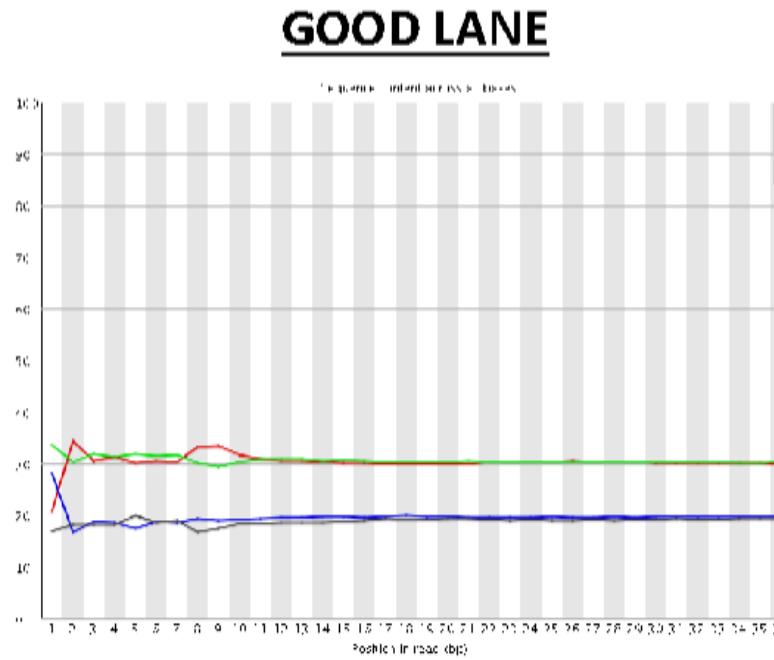
Per sequence quality scores



FastQC: Per base sequence content

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

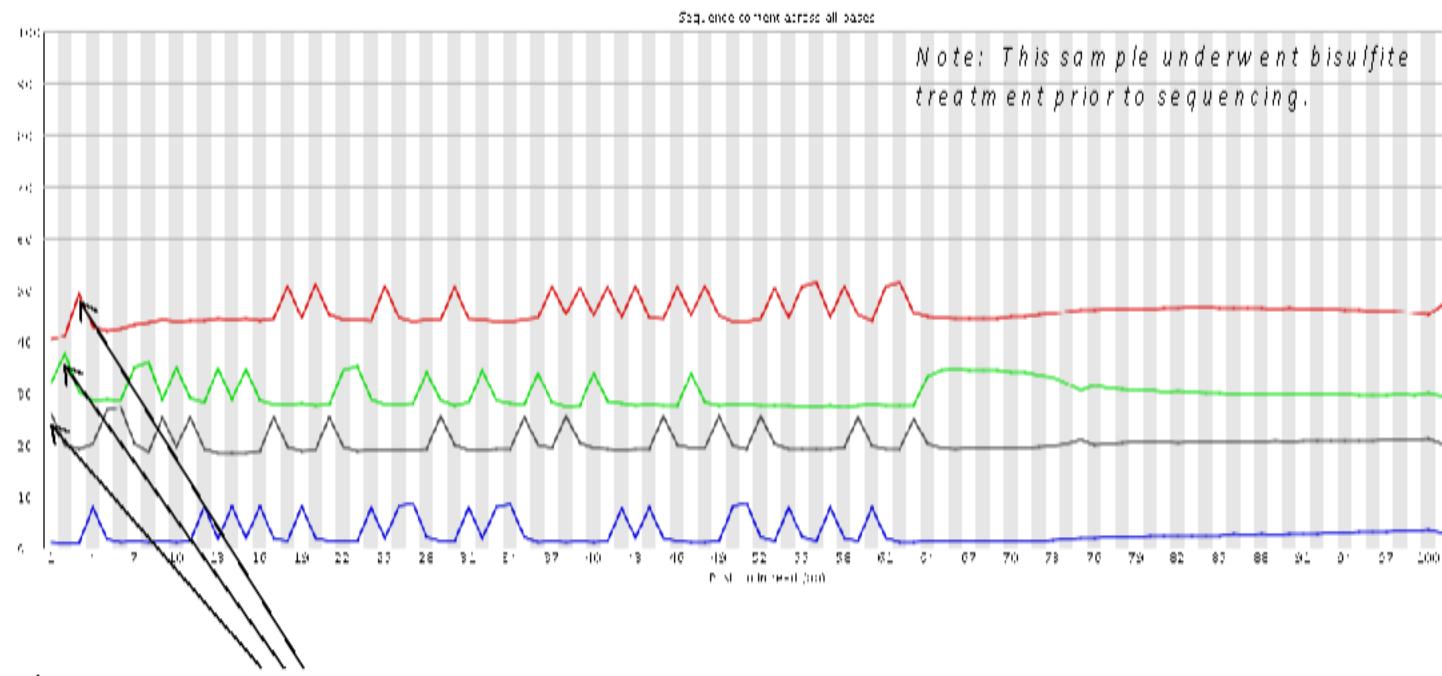
- x-axis = position in read (bp)
- y-axis = % of all reads in the lane
- colors refer to individual nucleotides: **A**, **C**, **G**, **T**



Can this be fixed? No.

FastQC: Per base sequence content

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



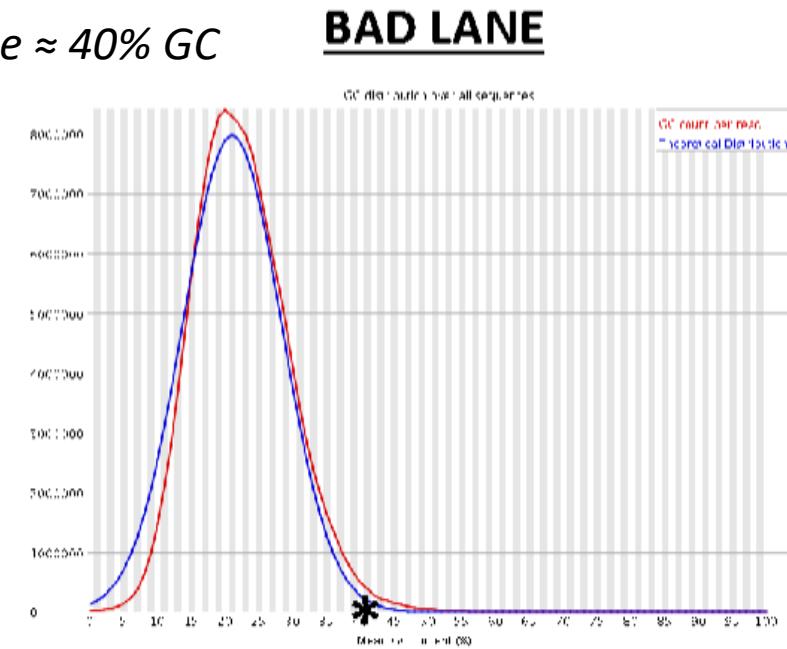
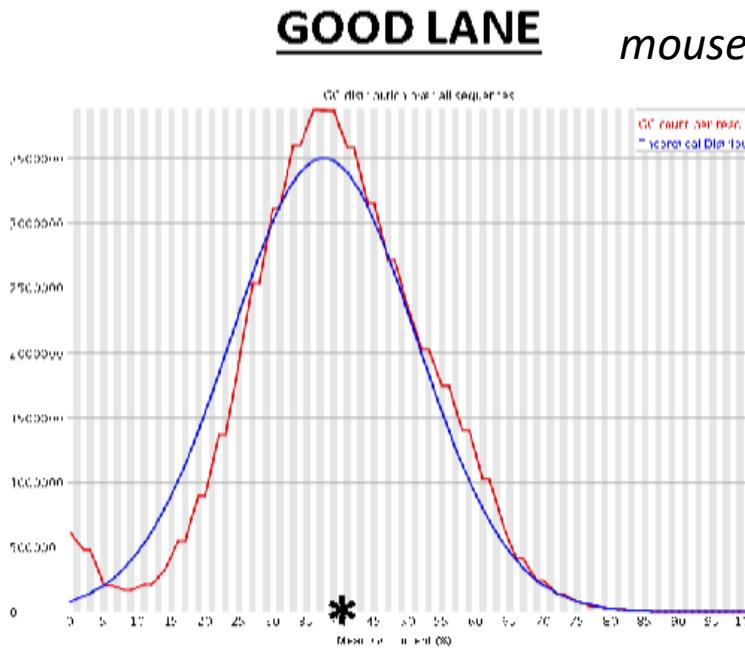
In this lane, ~10% of reads have the adapter sequence & the rest are normal.

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

FastQC: Per sequence GC content

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

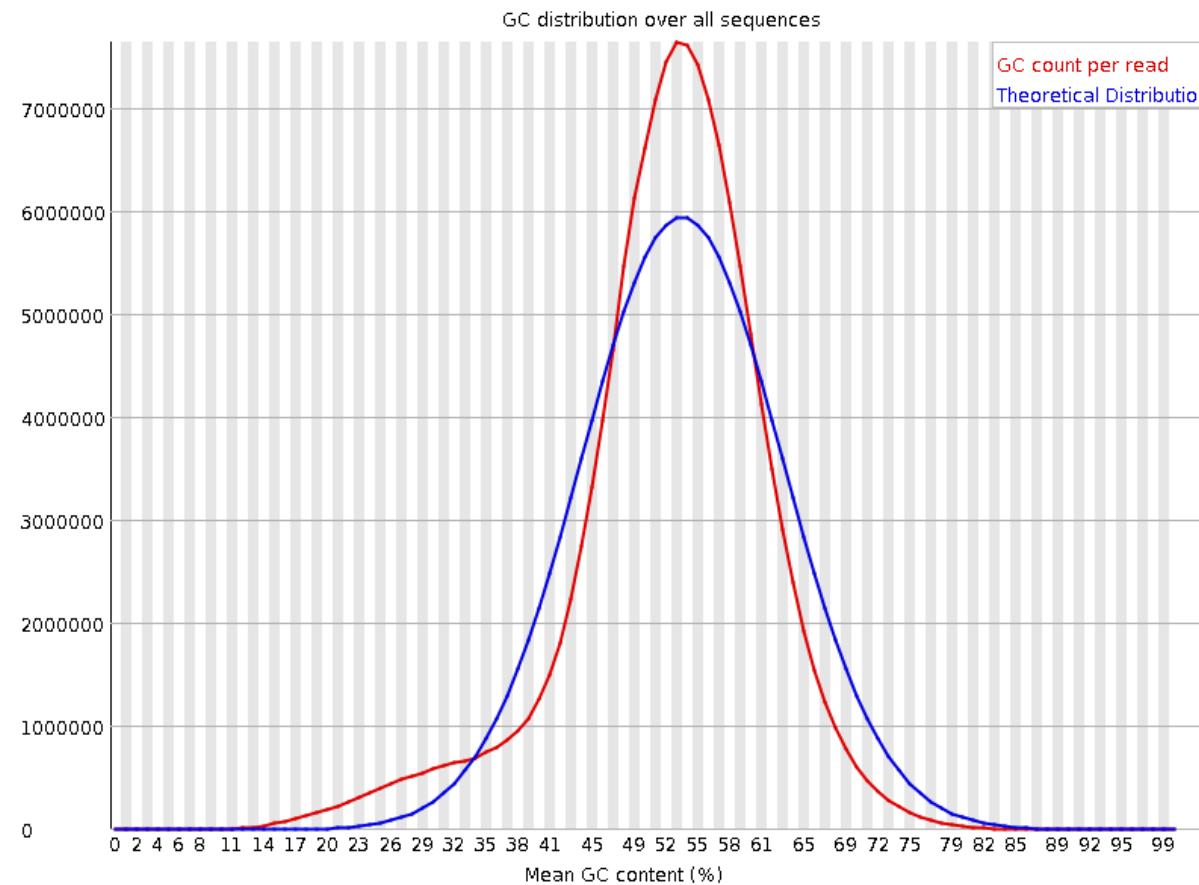
- x-axis = mean GC content (%)
- y-axis = # of reads
- red: observed read count, blue: theoretical distribution (given observed)



Can this be fixed? No.

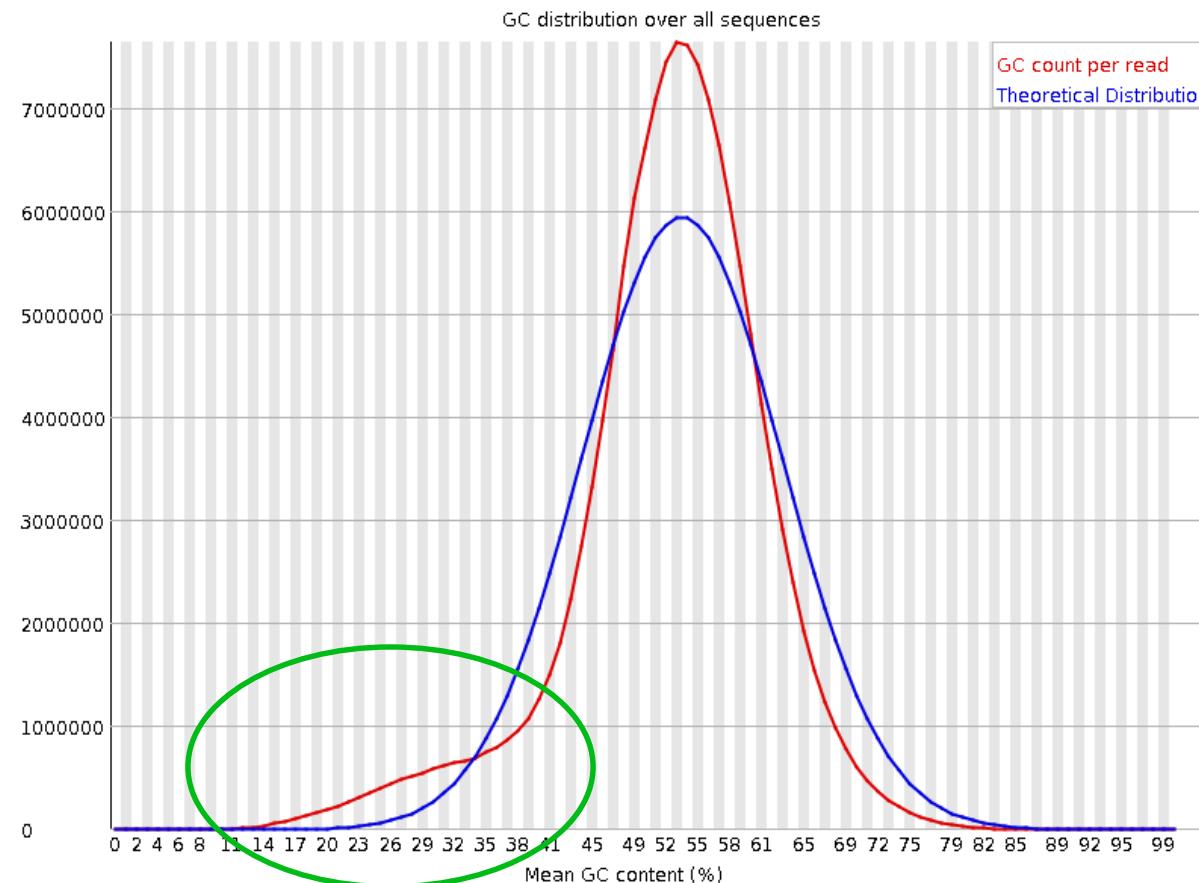
FastQC: Per sequence GC content

- A contamination ?



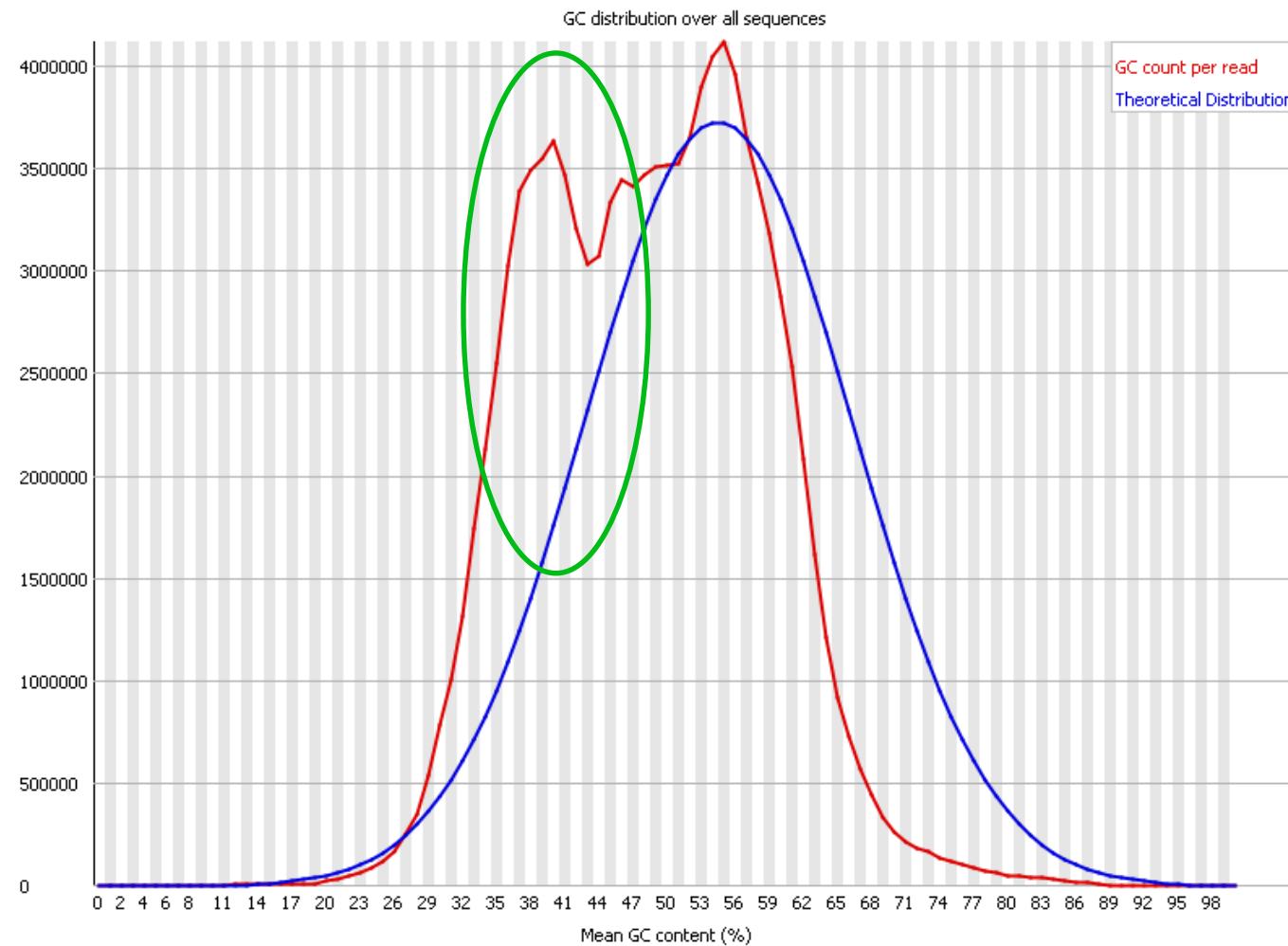
FastQC: Per sequence GC content

- A contamination ?

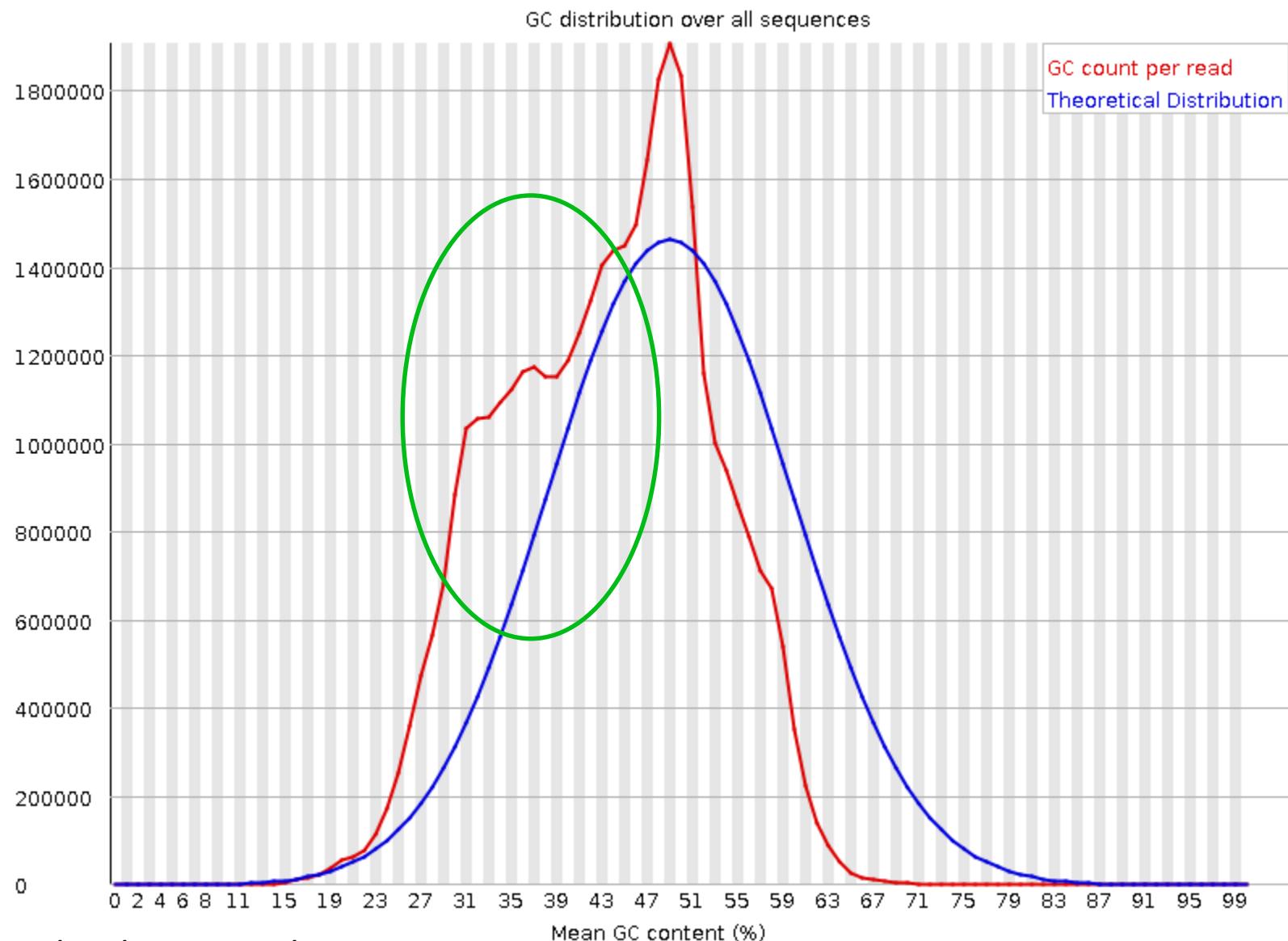


Can this be fixed ? Maybe...

FastQC: Per sequence GC content



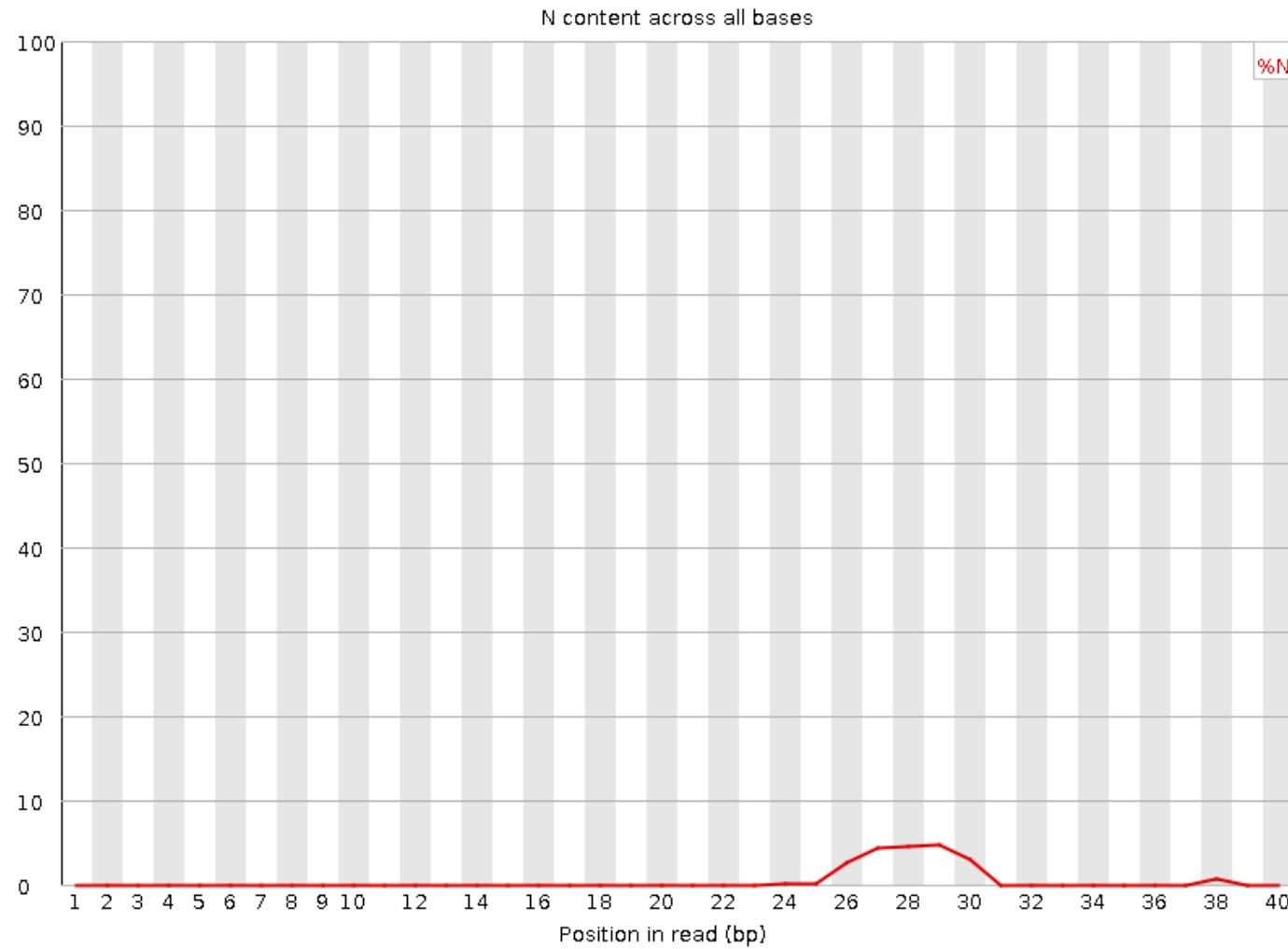
Third-party contamination : detection



FastQC: Per base N content



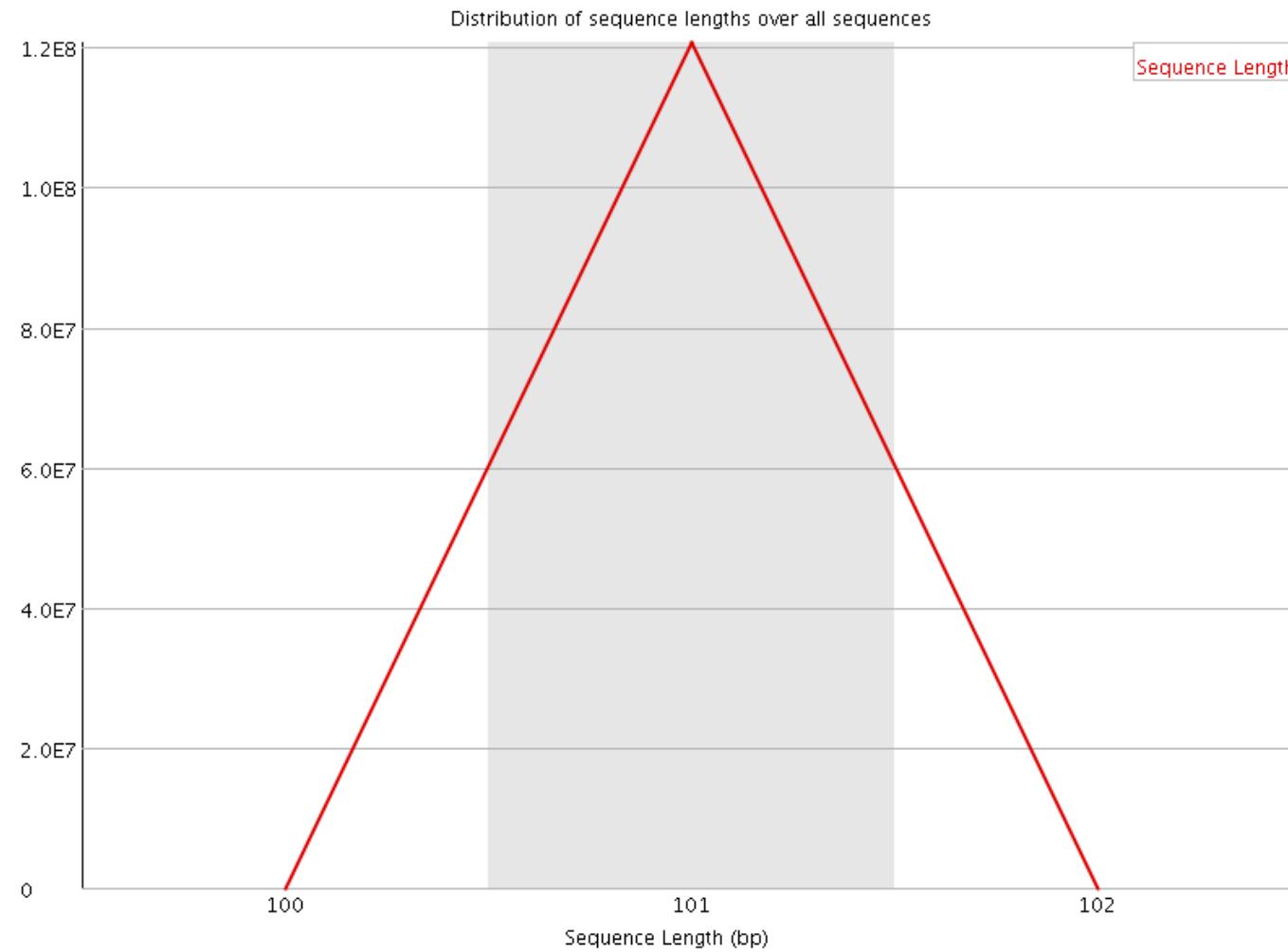
Per base N content



FastQC: Sequence Length Distribution



Sequence Length Distribution

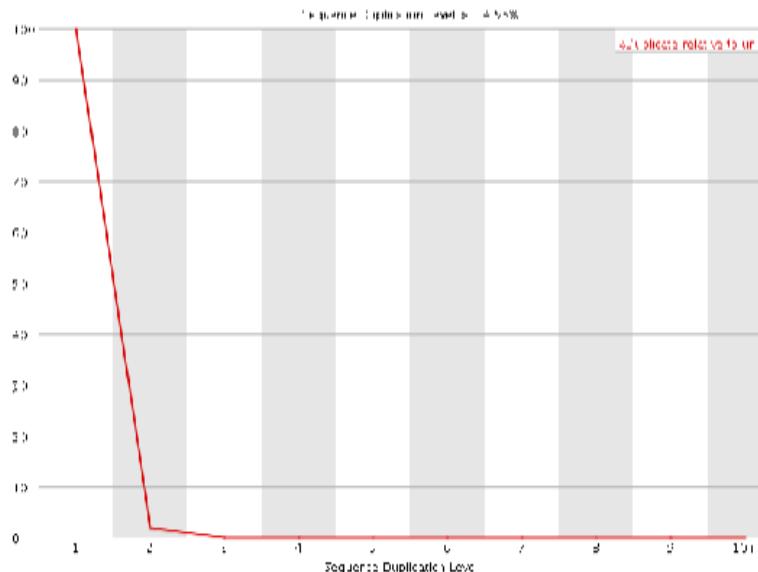


FastQC: Sequence Duplication Levels

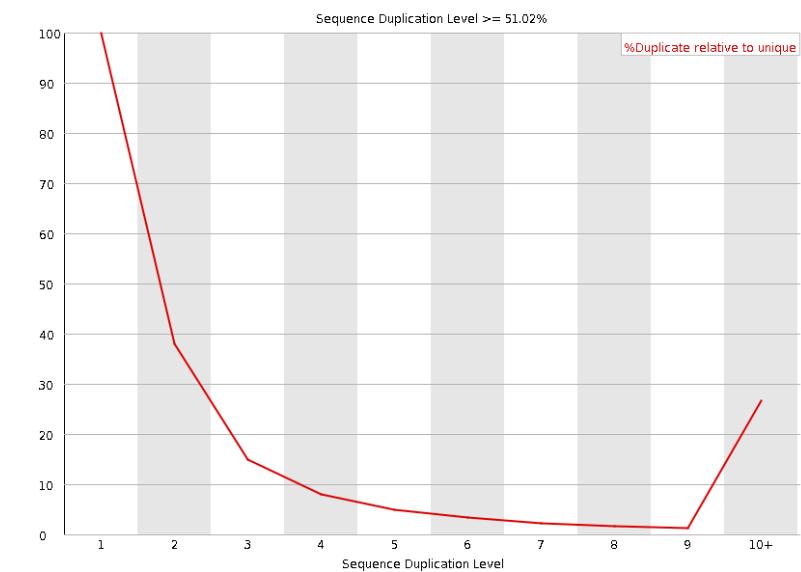
This plot shows the degree of duplication for a subset of reads in a lane.

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

GOOD LANE



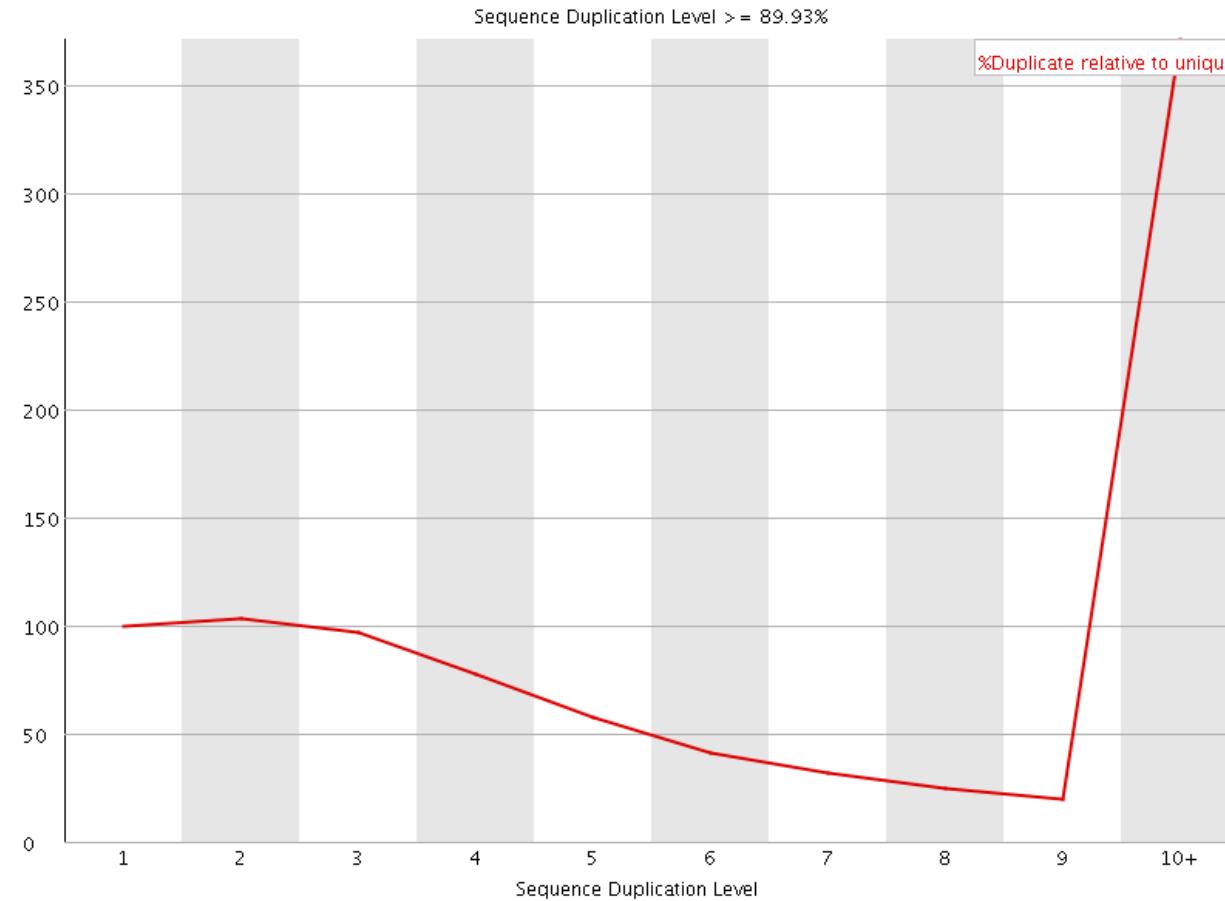
BAD LANE



Can this be fixed? Maybe.

FastQC: Sequence Duplication Levels

✖ Sequence Duplication Levels



Can this be fixed? Hem...

FastQC: Overrepresented sequences



Overrepresented sequences

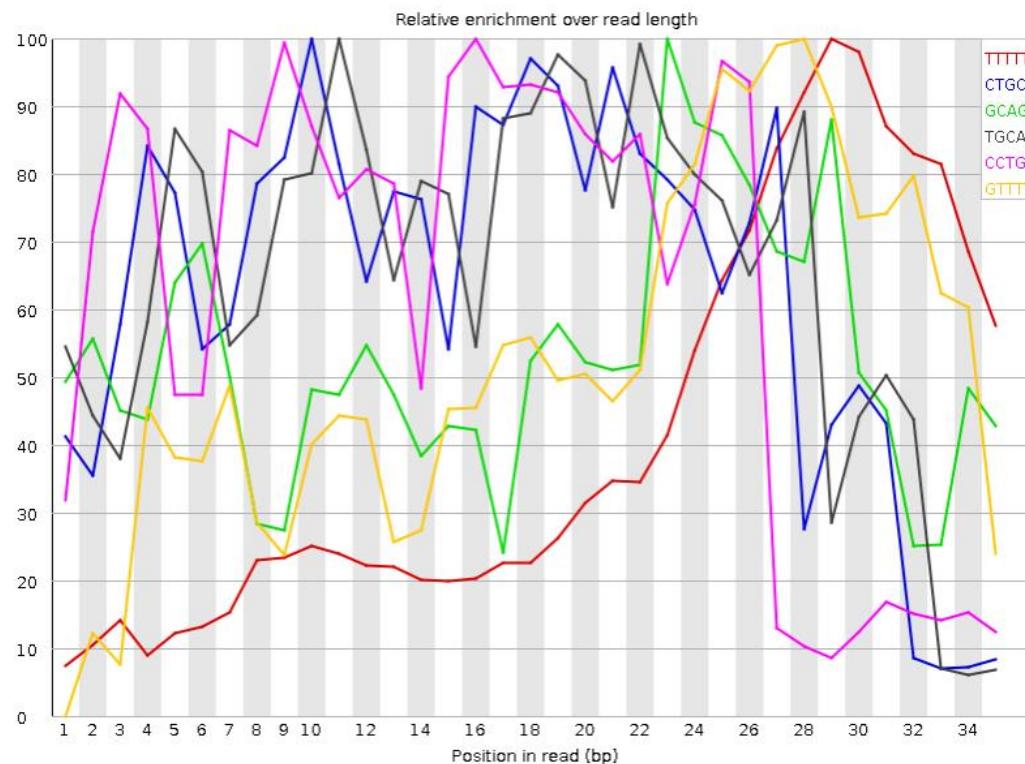
Sequence	Count	Percentage	Possible Source
AGAGTTTATCGCTTCATGACGCAGAAGTTAACACTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTATCGCTTCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTATCGCTTCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTATCGCTTCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTATCGCT	1846	0.4670012750197325	No Hit

Adapter dimers
rRNA
Satellite sequences

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)
CCTGCAGAGTTTATCGCTTCATGACGCAGAAGTTAAC	613	0.15507680476007366	No Hit
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTATCGAAGCGCG	585	0.1479933618020279	No Hit
CGCTTAAAGCTACCAGTTATGGCTGGGGGGTTTTTTT	552	0.13964501831575965	No Hit
CTCTGCAGGTTGGATACGCCAATCATTTTATCGAAGCGCG	532	0.1345854162028698	No Hit
CTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATACG	515	0.13028475440691342	No Hit
CTGCAGGTTGGATACGCCAATCATTTTATCGAAGCGCGC	505	0.12775495335046852	No Hit
GCTTAAAGCTACCAGTTATGGCTGGGGGGTTTTTTG	411	0.10397482341988626	No Hit

FastQC: Kmer Content

Kmer Content



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
TTTT	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

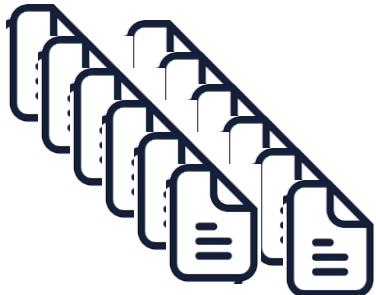
fastqc command

```
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 : 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 retrait des adaptateurs

