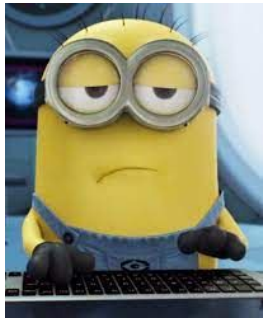


Introduction Bioinformatics & Sequencing

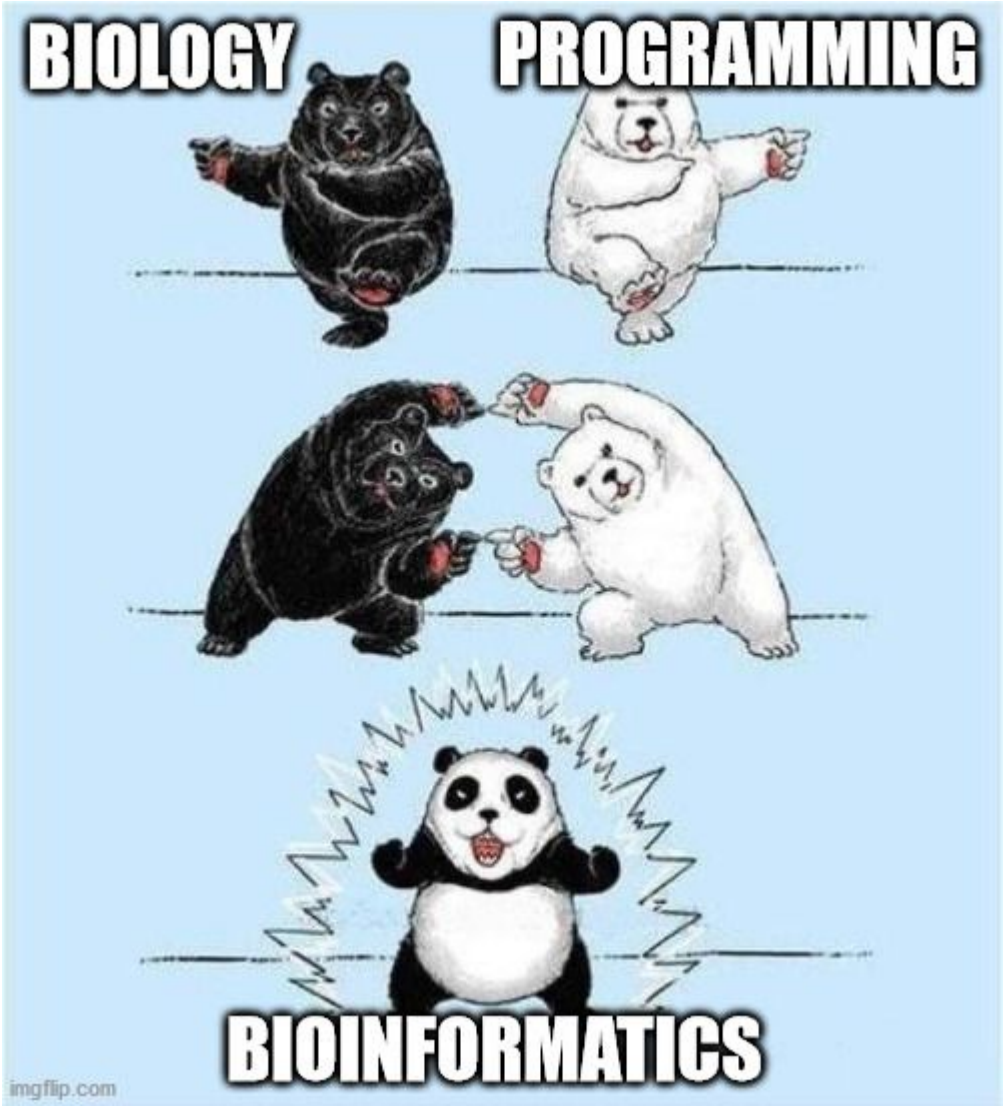
Rabat, Novembre 2022

Institut de Recherche pour le Développement FRANCE

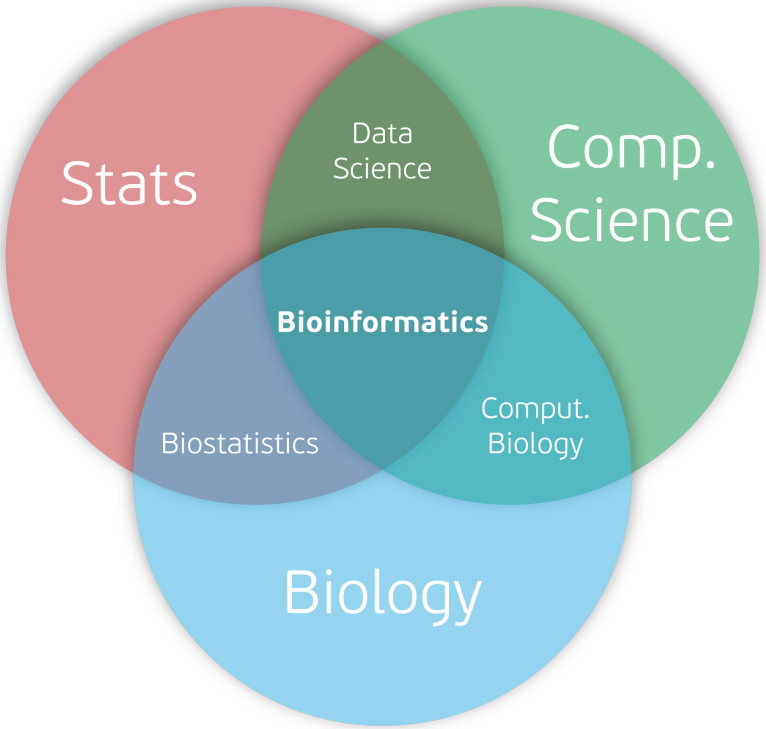
Université Mohammed V
Faculté des Sciences
Rabat

slides par C. Tranchant & A. Dereeper, J.Orjuela, F. Sabot

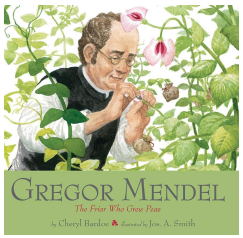
What is bioinformatics ?



An interdisciplinary science



De la génétique à la bioinformatique...



1866

Lois de l'hérédité

1944

Nature chimique de l'ADN, matériel héréditaire

O. Avery, C. McLeod & McCarthy

1954

Structure en double hélice de l'ADN

J. Watson & F. Cricks & franklin

1961

Code génétique et règle de correspondance gènes-protéines

M. Nirenberg & H. Matthaei

1965

Mécanismes de la régulation génétique

André Lwoff, F. Jacob & J. Monod

De la génétique à la bioinformatique...

1970

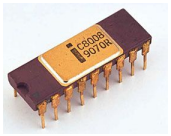
Algo *Alignement global de séquence*

Needman, & Wunsh

1977

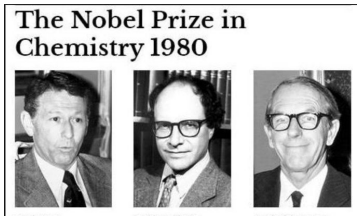
Micro-ordinateurs

1972 **1er microprocesseur intel 8008**



Séquençage ADN

P. Berg, W. Gilbert & F. Sanger



1980

Banque EMBL, GenBank, PIR

1984

Amplification ADN - PCR

Karry Mullis

1985

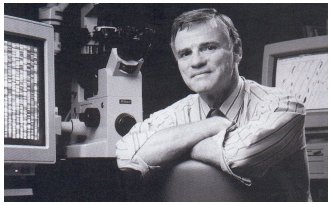
Algo Alignement local de séquence FASTA

Person & Lipman

1987

1er séquenceur automaisé

L. Hood Société Applied Biosystems



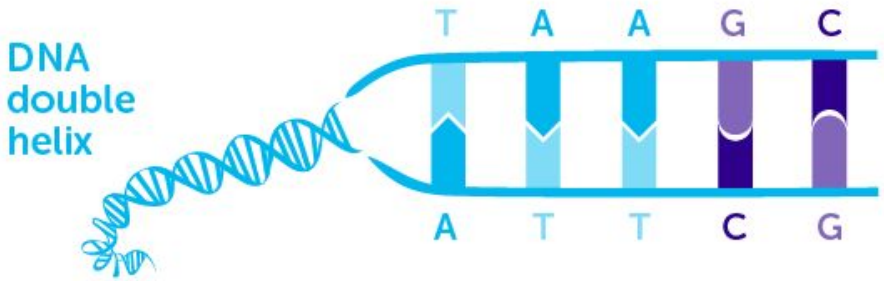
1990

Algo Alignement local de séquence BLAST

Altschul & al.

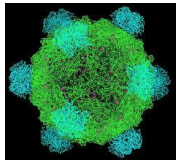
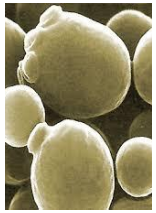
A little history of sequencing...

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



A little history of sequencing...

Next-generation sequencing

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

Several sequencing technology

SEQUENCING TECHNOLOGY

1 **FIRST GENERATION**
From 1977



sanger

2 **SECOND GENERATION**
From 2007



454, Illumina, IonTorrent

3 **THIRD GENERATION**
From 2011


PacBio, ONT

Several sequencing technology

SEQUENCING TECHNOLOGY

1 **FIRST GENERATION** From 1977



sanger

2 **SECOND GENERATION** From 2007



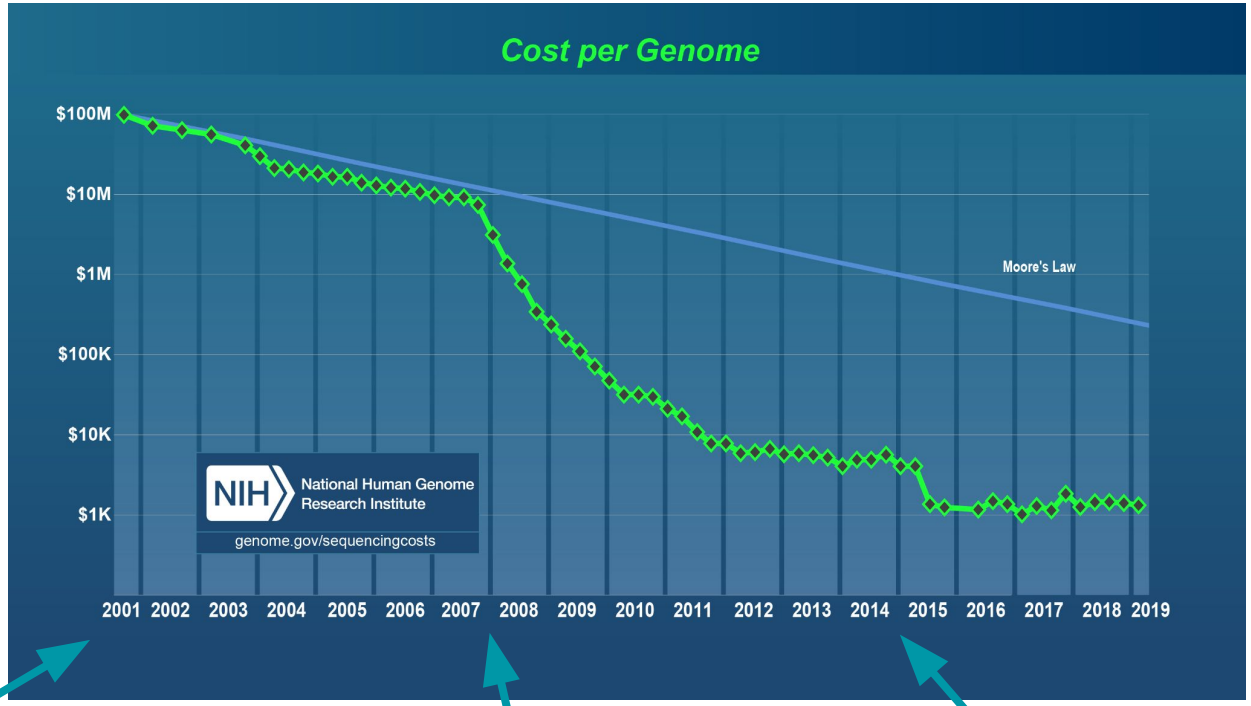
454, Illumina, IonTorrent

3 **THIRD GENERATION** From 2011


PacBio, ONT

Sequencing output, price, reads size, sequencing quality

From Sanger to 3rd sequencing technology



1 FIRST GENERATION From 1977

sanger

2 SECOND GENERATION From 2007

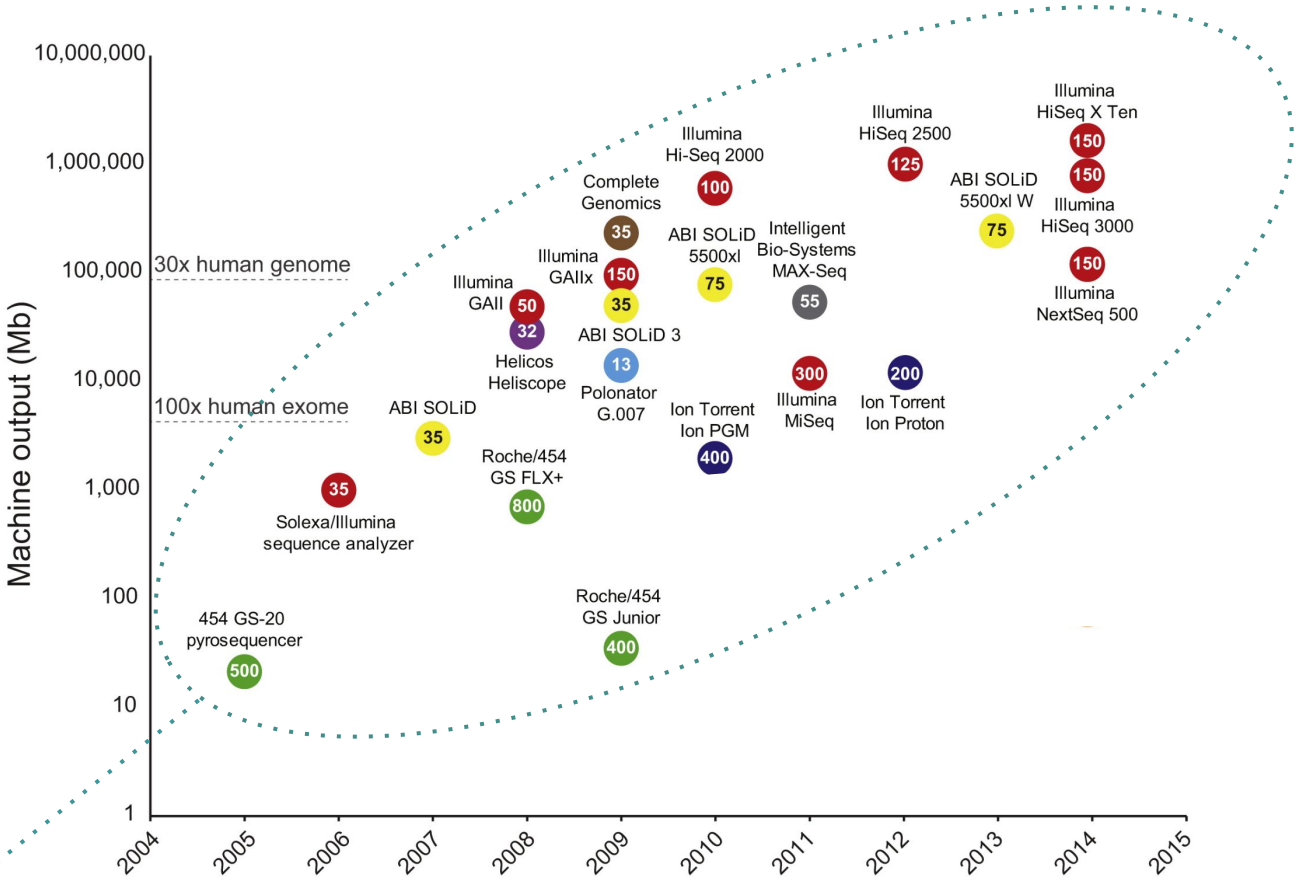
Illumina

3 THIRD GENERATION From 2011

PacBio, ONT

From Sanger to 3rd sequencing technology

Une augmentation du débit de séquençage

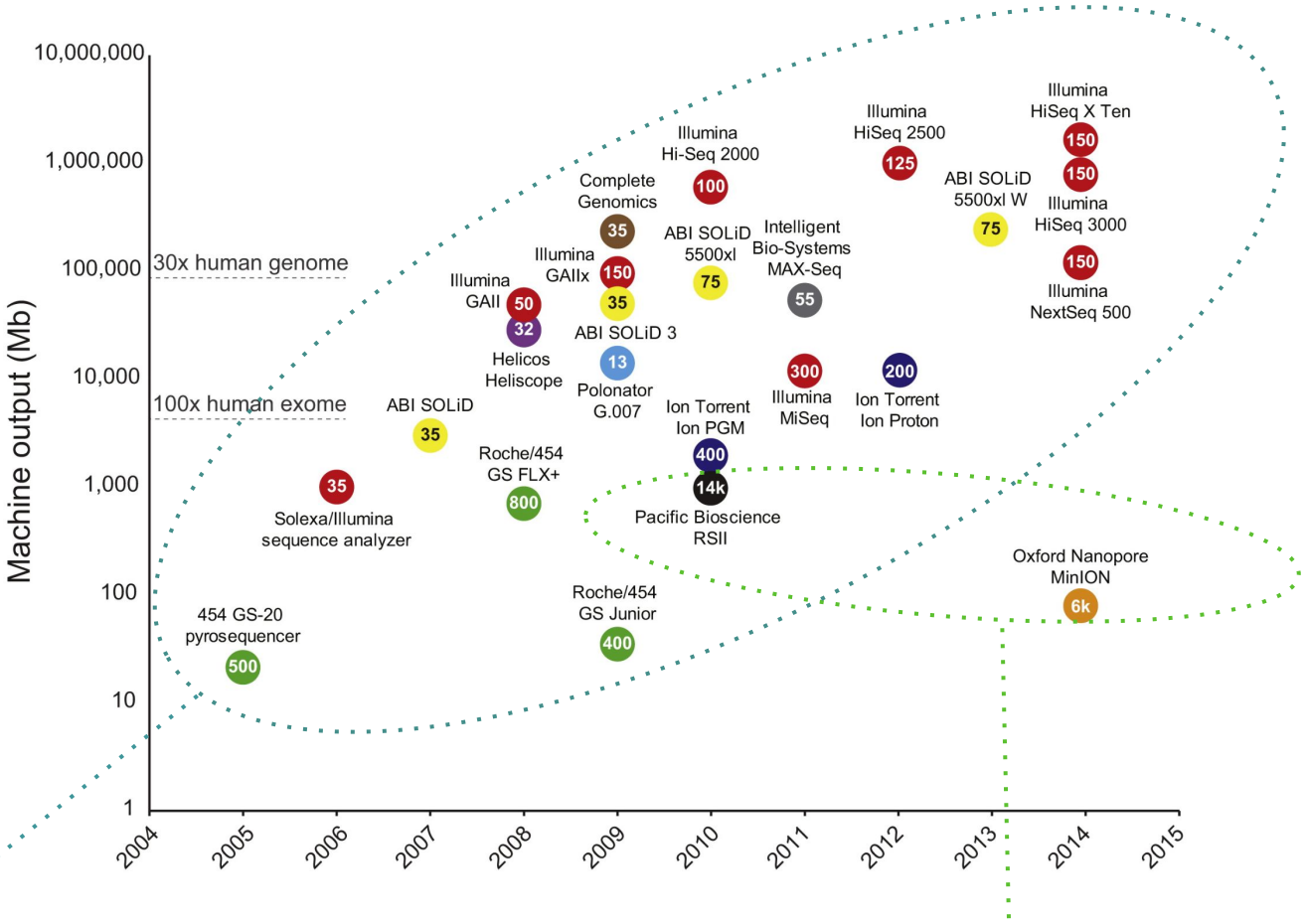


2 **SECOND GENERATION**
From 2007

Illumina

From Sanger to 3rd sequencing technology

Une augmentation du débit de séquençage



2 SECOND GENERATION
From 2007

Illumina

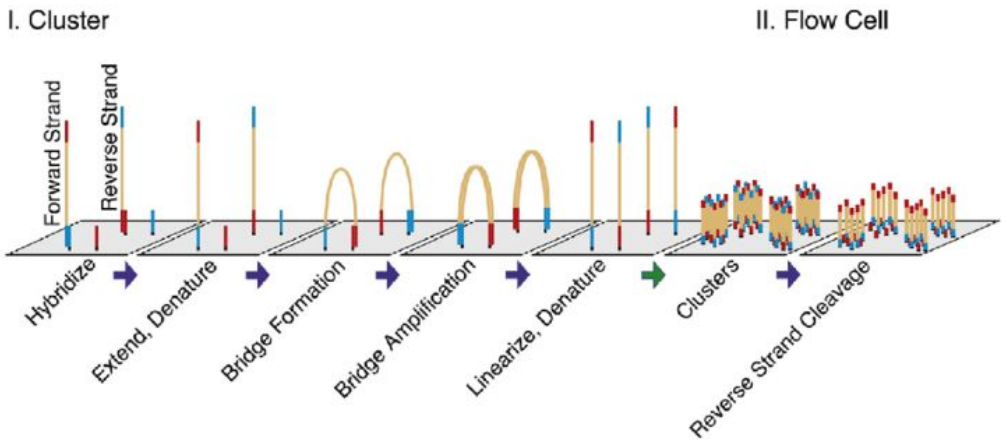
3 THIRD GENERATION
From 2011

Short Reads ?
Long Reads ?

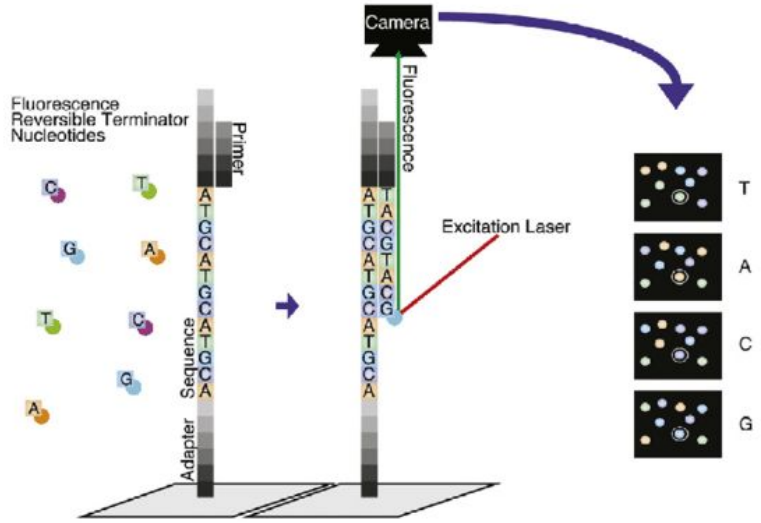
Short Reads - Illumina technology

2 **SECOND GENERATION**
From 2007

A. Clustering



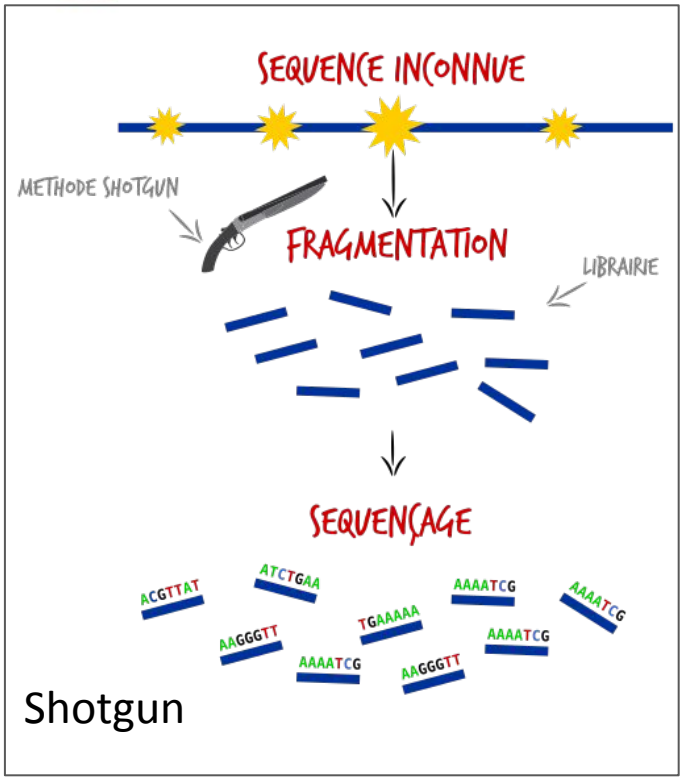
B. High-throughput sequencing



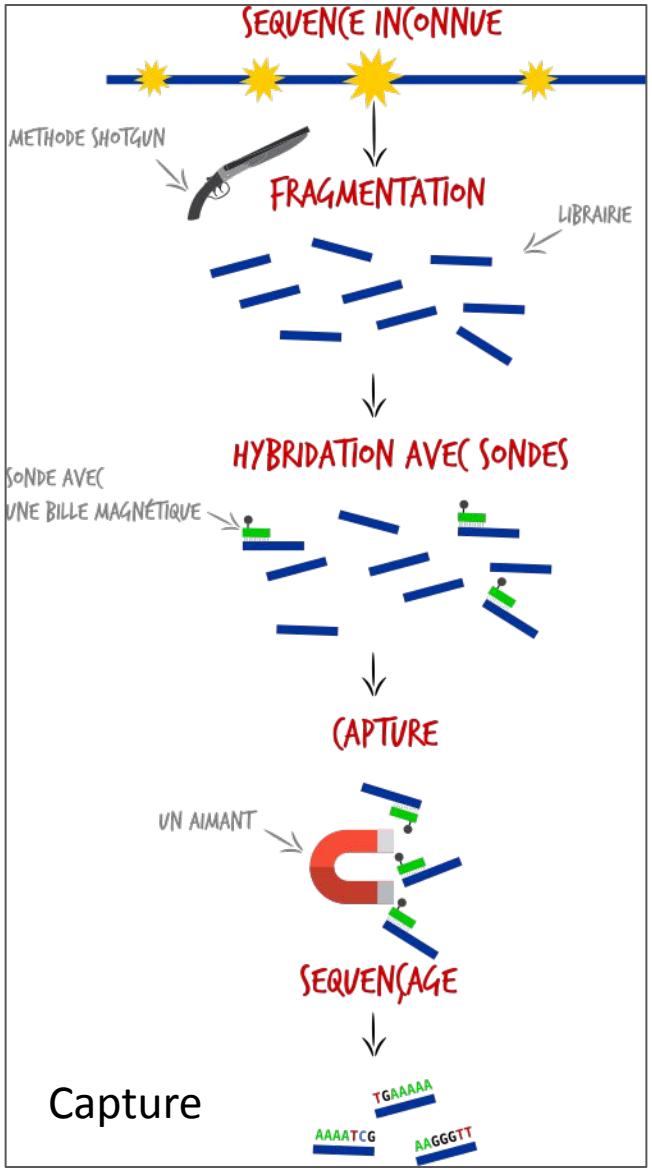
Vijender Chaitankar et al 2016

Short Reads - Illumina technology

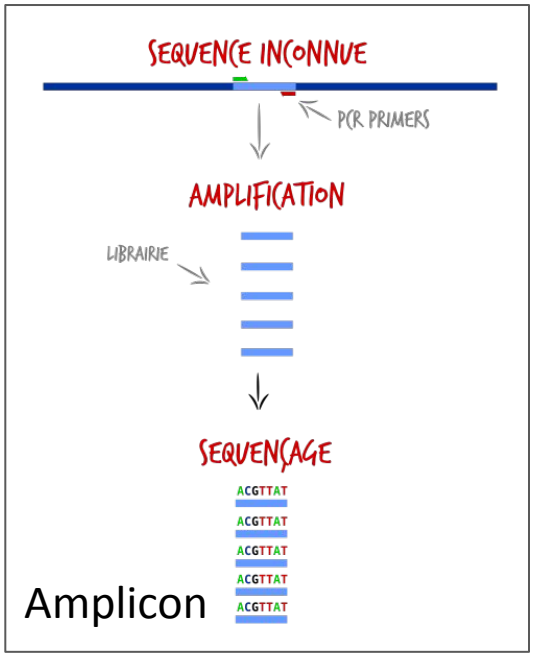
2 **SECOND GENERATION**
From 2007



Shotgun



Capture



Amplicon

https://dridk.me/ngs.html

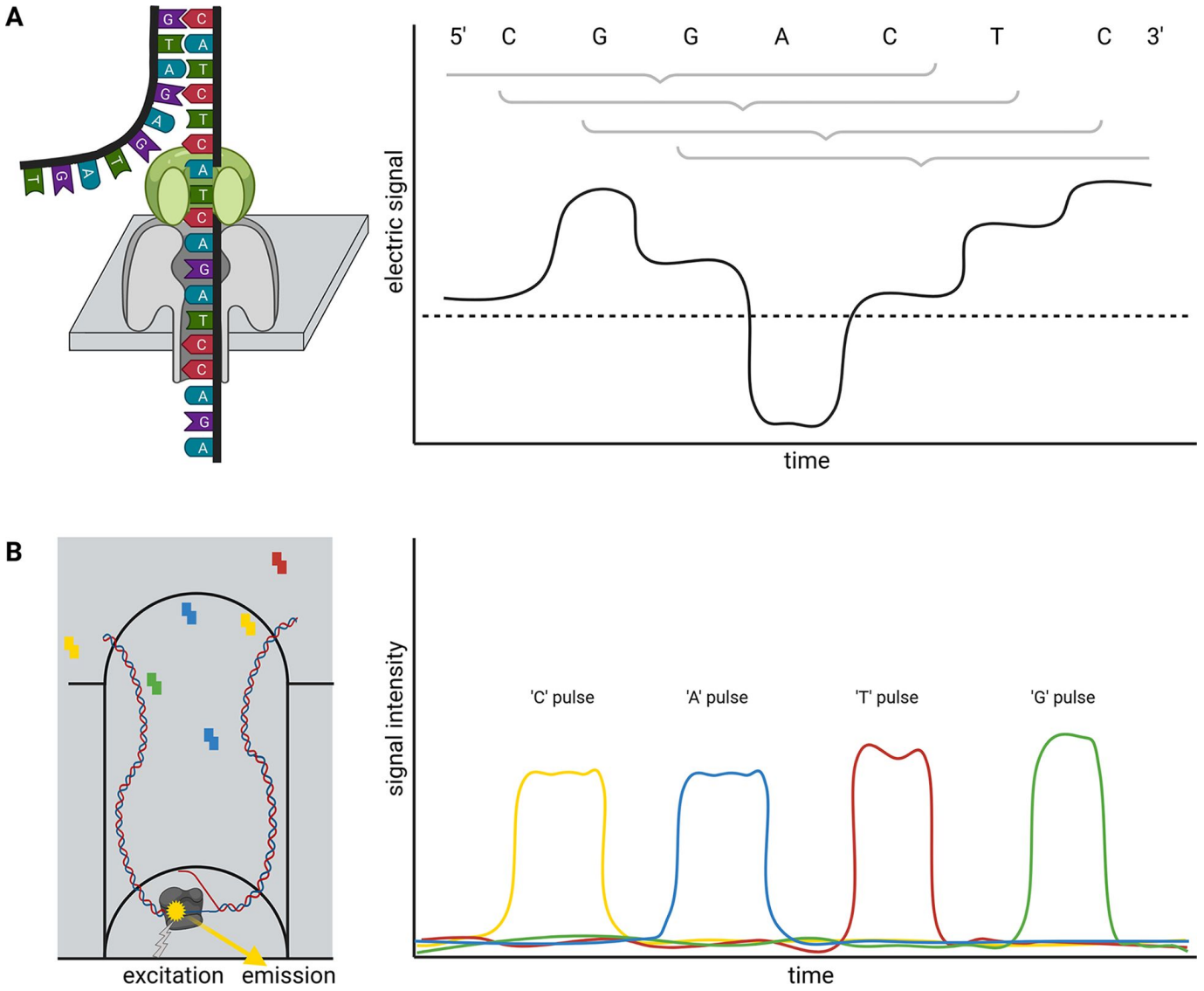
Short Reads - Illumina - PROS & CONS

2 SECOND GENERATION
From 2007

✓	**Output volume**	20 billions of 150b reads, 6T	*NovaSeq6000*
✓	**Accuracy**	~99 %	
✓	**Run is cheap**		
✓	**MySeq is cheap**	~60 000 USD per machine	

- **Size** 150 + 150, *NovaSeq*
  but 400-600 pb, *MySeq*

3 **THIRD GENERATION**
From 2011



Boas Pucker et al. 2022

3 **THIRD GENERATION**
From 2011

Two technologies

Oxford Nanopore

MinION

GridION

PromethION

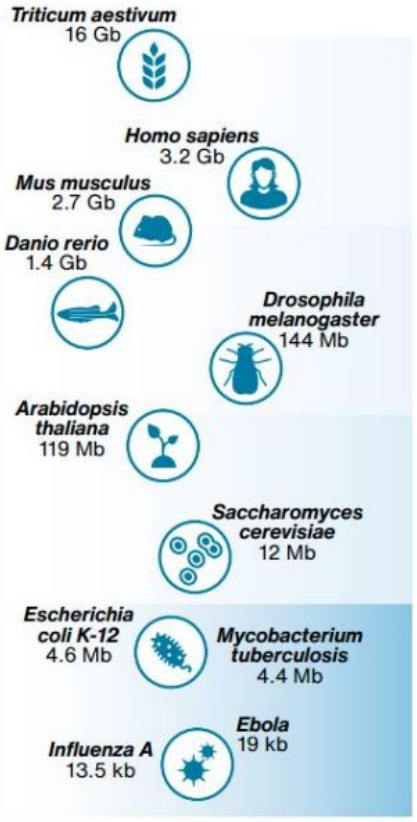
Pacific BioScience

RSII

Sequel

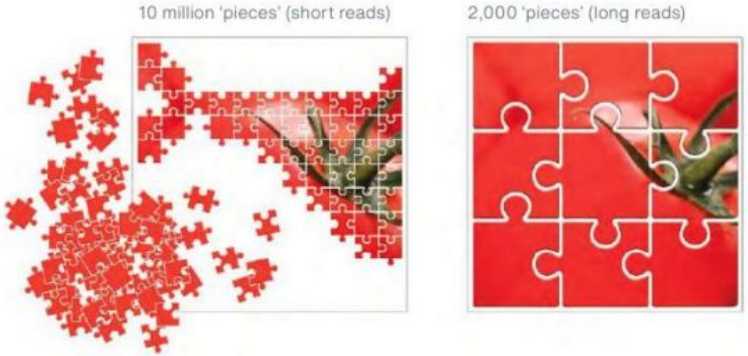
from Elixir GAAS 2018

3 **THIRD GENERATION**
From 2011



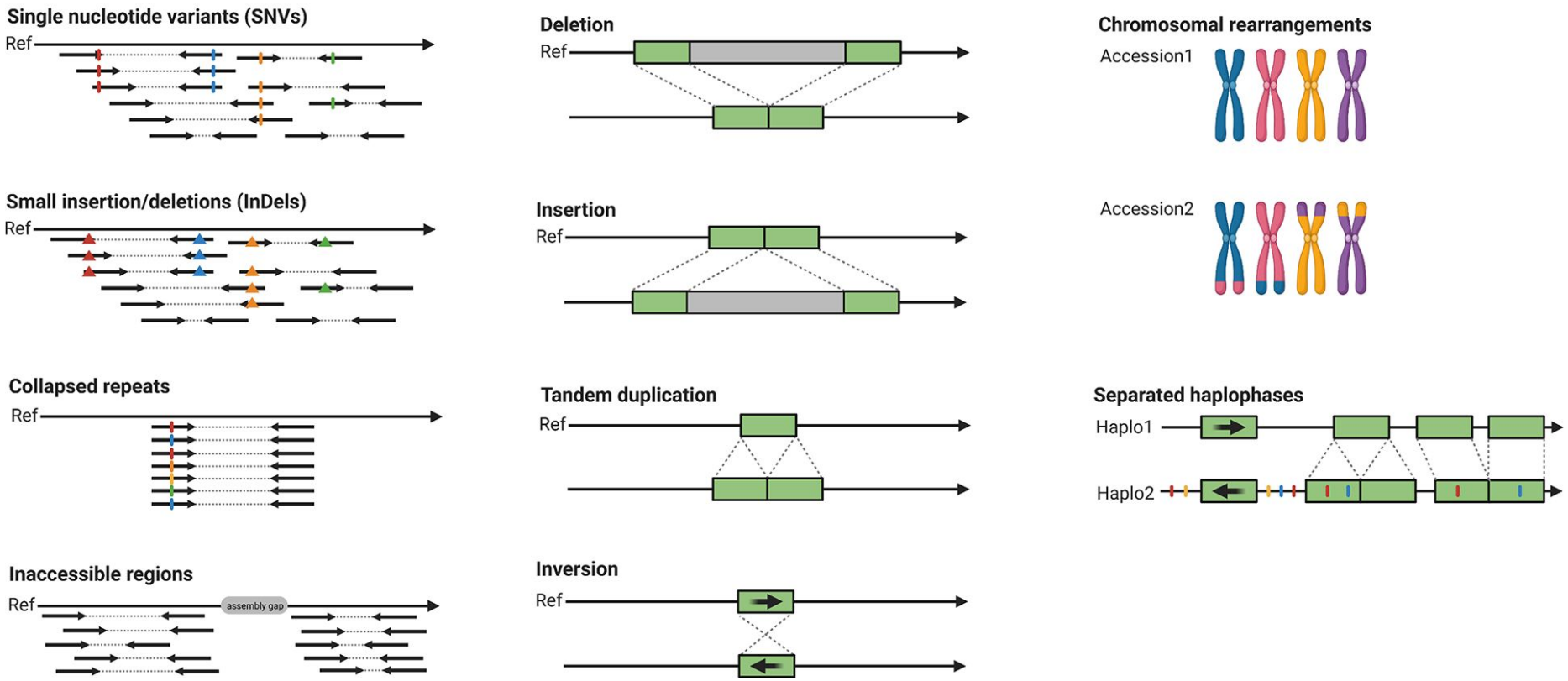
Microbial genomes | Human genomes | Animal genomes | Plant genomes

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



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

3 **THIRD GENERATION**
From 2011



Boas Pucker et al. 2022

Long Reads - Oxford nanopore - PROS & CONS

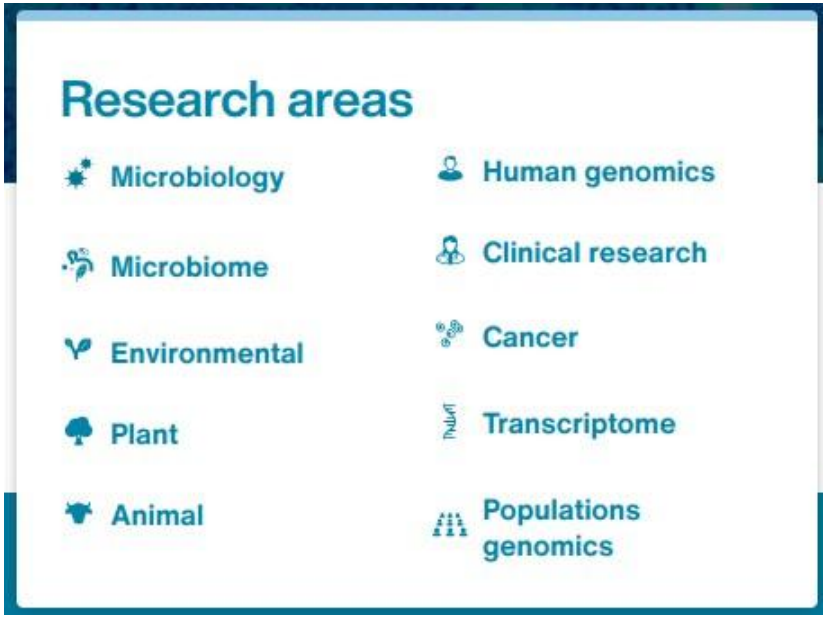
3 **THIRD GENERATION**
From 2011

- ✓ **No Amplification, NO SYNTHESIS, Very Long Length**
- ✓ Single strand direct sequencing
- ✓ Bases Modification detection in real-time
- ✓ Native RNA!
- ✓ **Read length** ~ 10-450kb more than 4Mb reported
- ✓ **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, 0.5-2% in tests
- Quality of DNA/RNA limits the sequencing

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

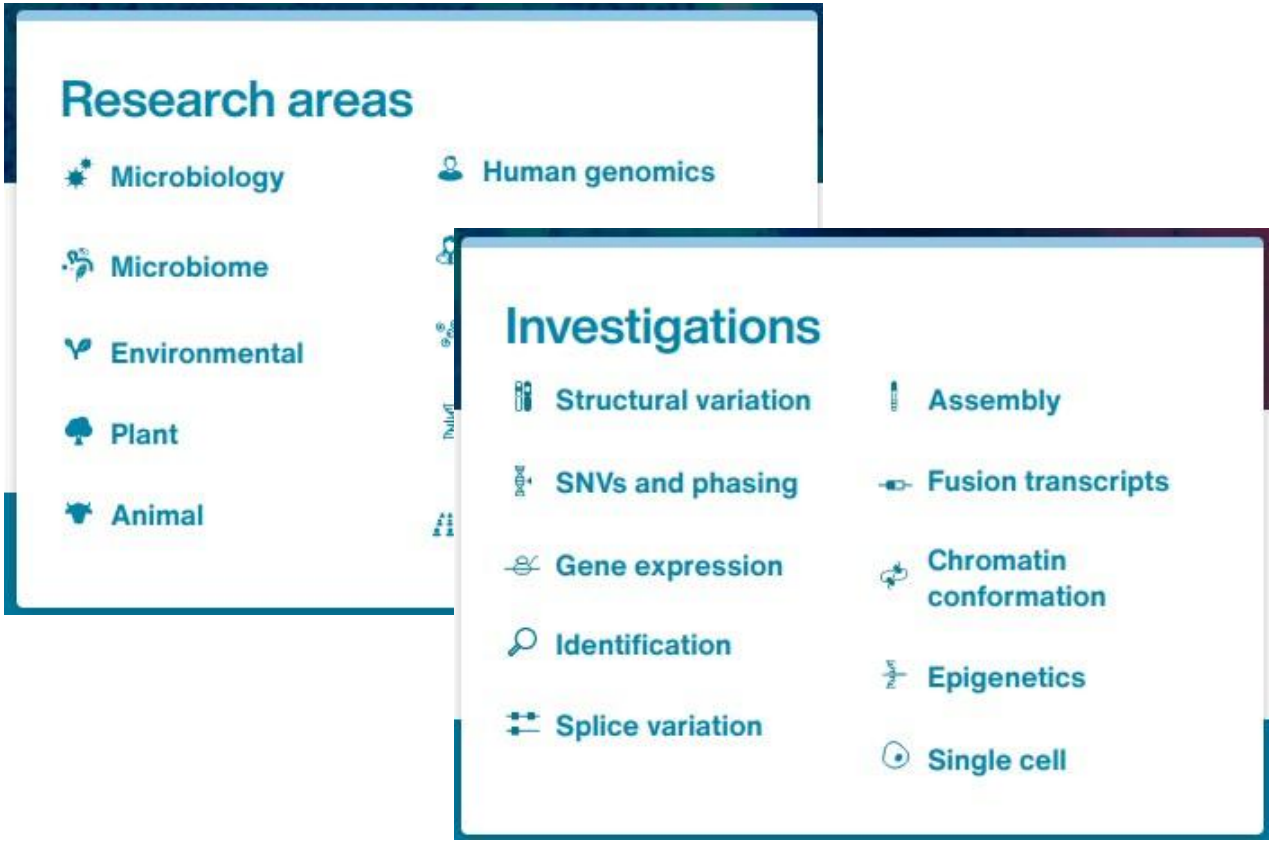
3 **THIRD GENERATION**
From 2011



From Nanopore website

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

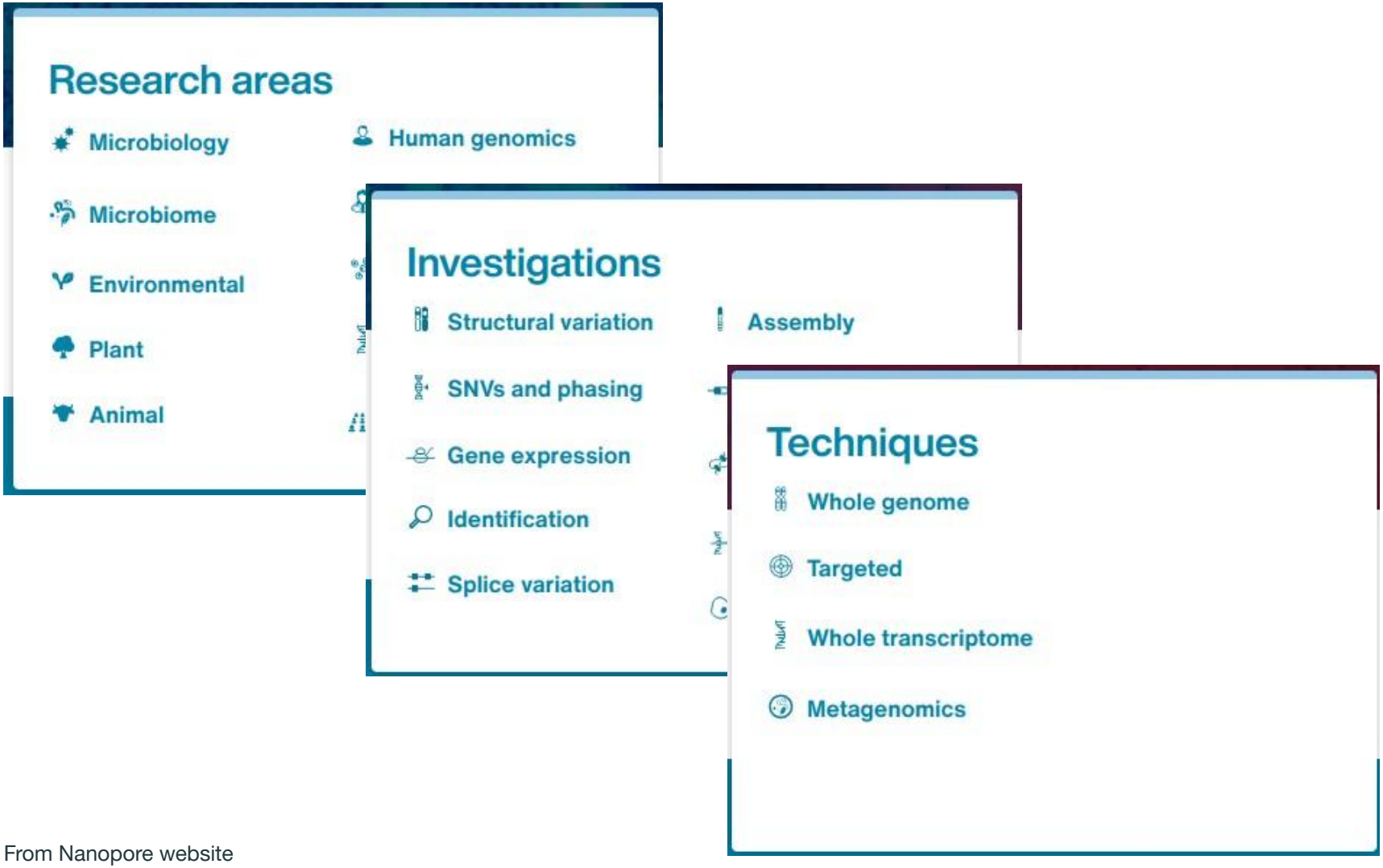
3 **THIRD GENERATION**
From 2011



From Nanopore website

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

3 **THIRD GENERATION**
From 2011



From Nanopore website

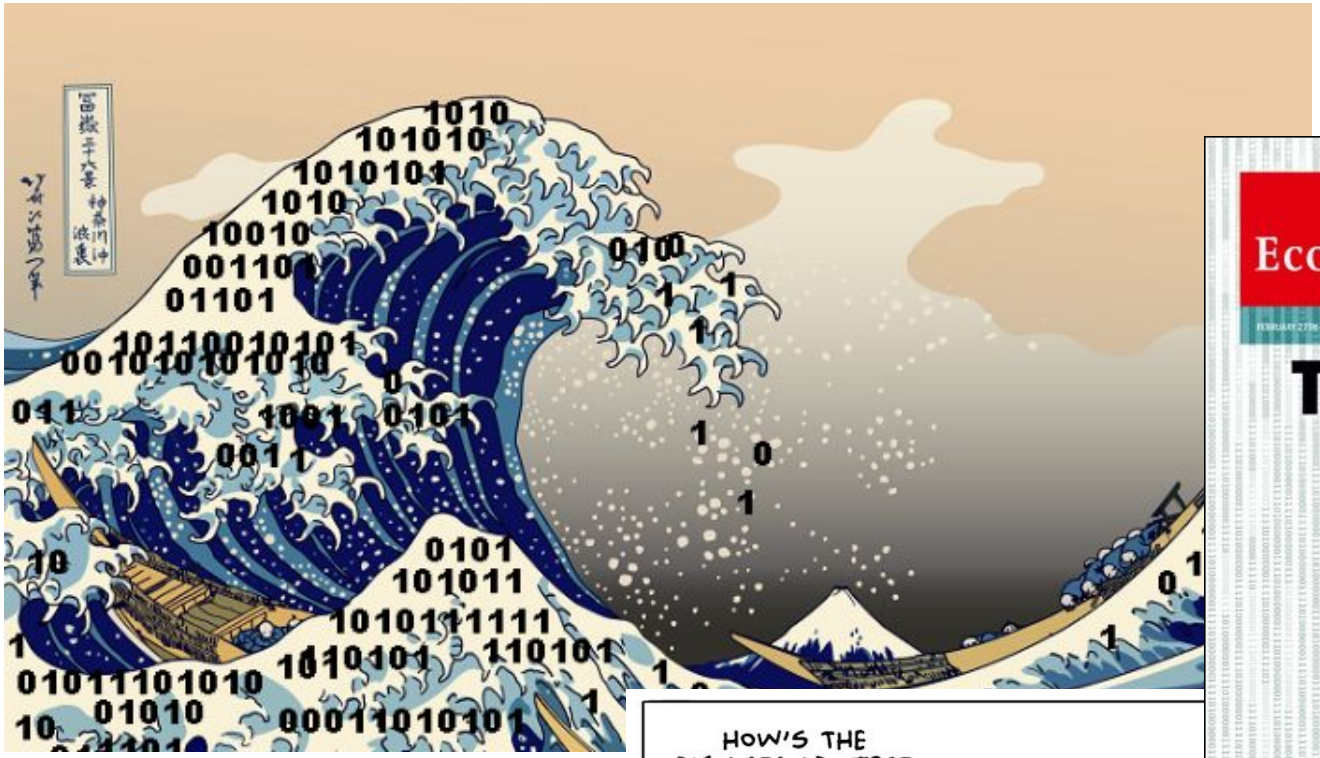
Comparison

Comparison of high-throughput sequencing methods[78][79]

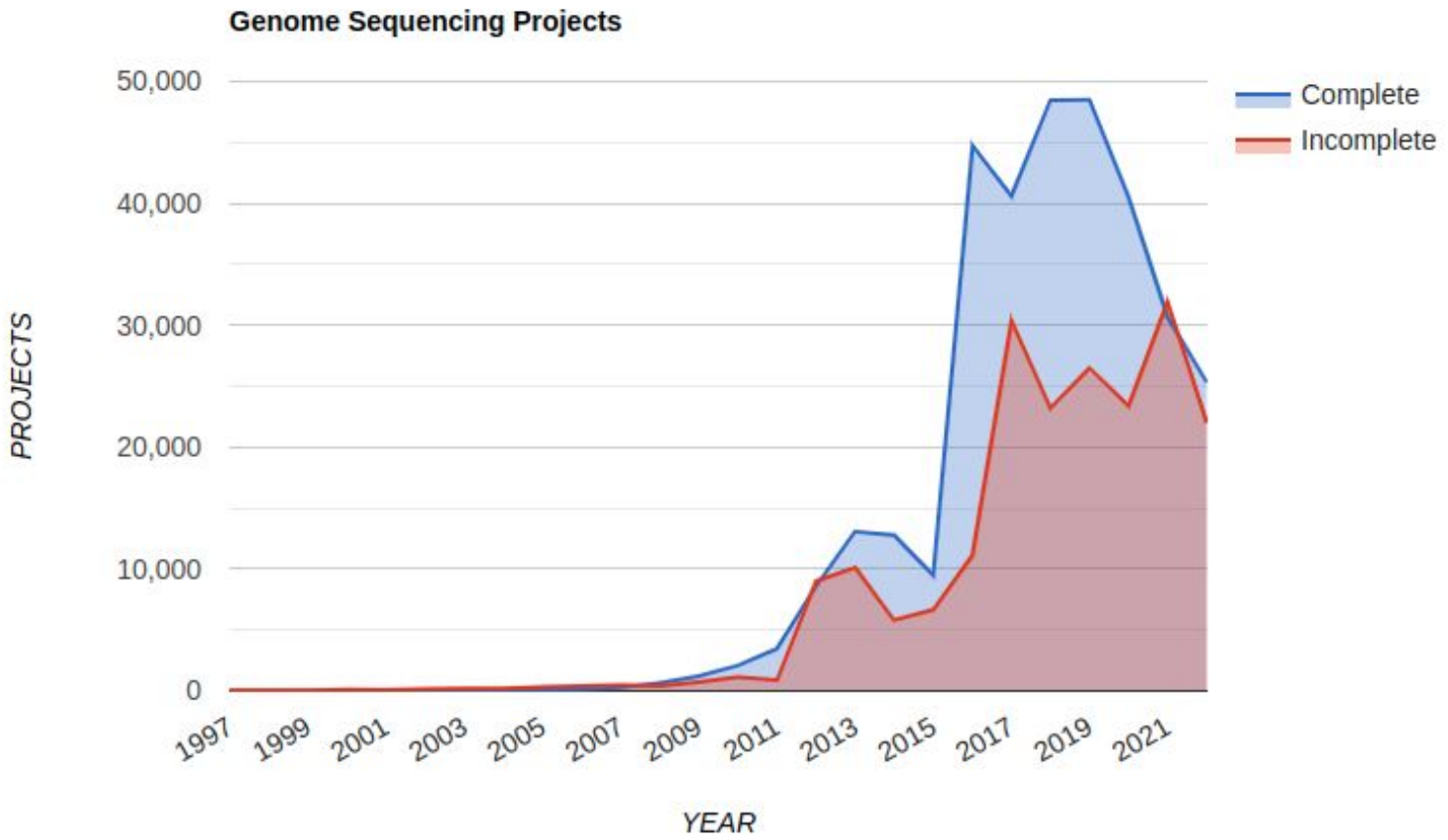
Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 billion bases (in US$)
Single-molecule real-time sequencing (Pacific Biosciences)	30,000 bp (N50); maximum read length >100,000 bases[80][81][82]	87% raw-read accuracy[83]	4,000,000 per Sequel 2 SMRT cell, 100–200 gigabases[80][84][85]	30 minutes to 20 hours[80][86]	$7.2-$43.3
Ion semiconductor (Ion Torrent sequencing)	up to 600 bp[88]	99.6%[89]	up to 80 million	2 hours	$66.8-$950
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	$10,000
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75–300 bp; MiSeq: 50–600 bp; HiSeq 2500: 50–500 bp; HiSeq 3/4000: 50–300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1–25 Million; NextSeq: 130-00 Million; HiSeq 2500: 300 million – 2 billion; HiSeq 3/4000 2.5 billion; HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length[90]	$5 to $150
Combinatorial probe anchor synthesis (cPAS- BGI/MGI)	BGISEQ-50: 35-50bp; MGISEQ 200: 50-200bp; BGISEQ-500, MGISEQ-2000: 50-300bp[91]	99.9% (Phred30)	BGISEQ-50: 160M; MGISEQ 200: 300M; BGISEQ-500: 1300M per flow cell; MGISEQ-2000: 375M FCS flow cell, 1500M FCL flow cell per flow cell.	1 to 9 days depending on instrument, read length and number of flow cells run at a time.	$5– $120
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	$60–130
Nanopore Sequencing	Dependent on library preparation, not the device, so user chooses read length (up to 2,272,580 bp reported[93]).	~92–97% single read	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	$7–100
GenapSys Sequencing	Around 150 bp single-end	99.9% (Phred30)	1 to 16 million	Around 24 hours	$667
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2,400,000

From wikipedia website

From data rarity to data deluge



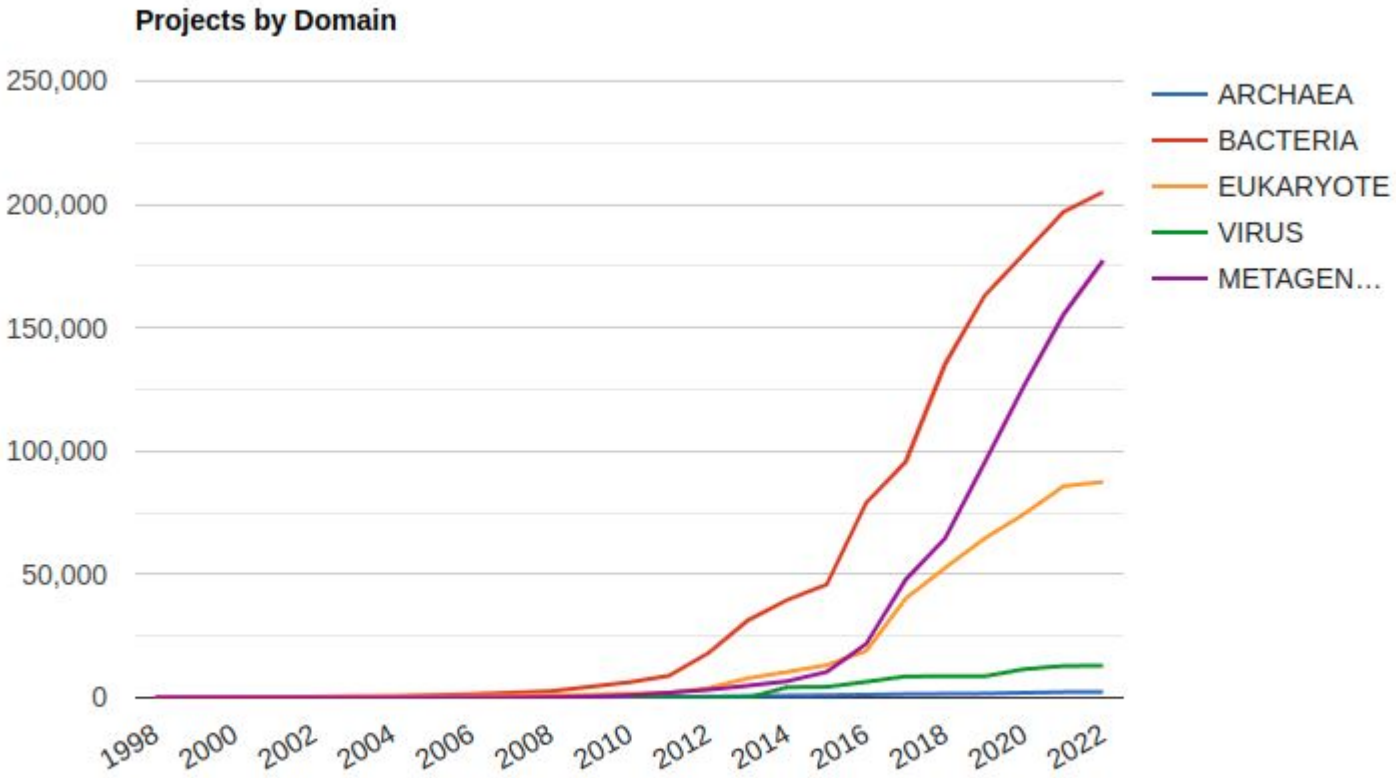
Genome Totals by year and status



GOLD
GENOMES ONLINE DATABASE

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

Project Totals by year and domain group

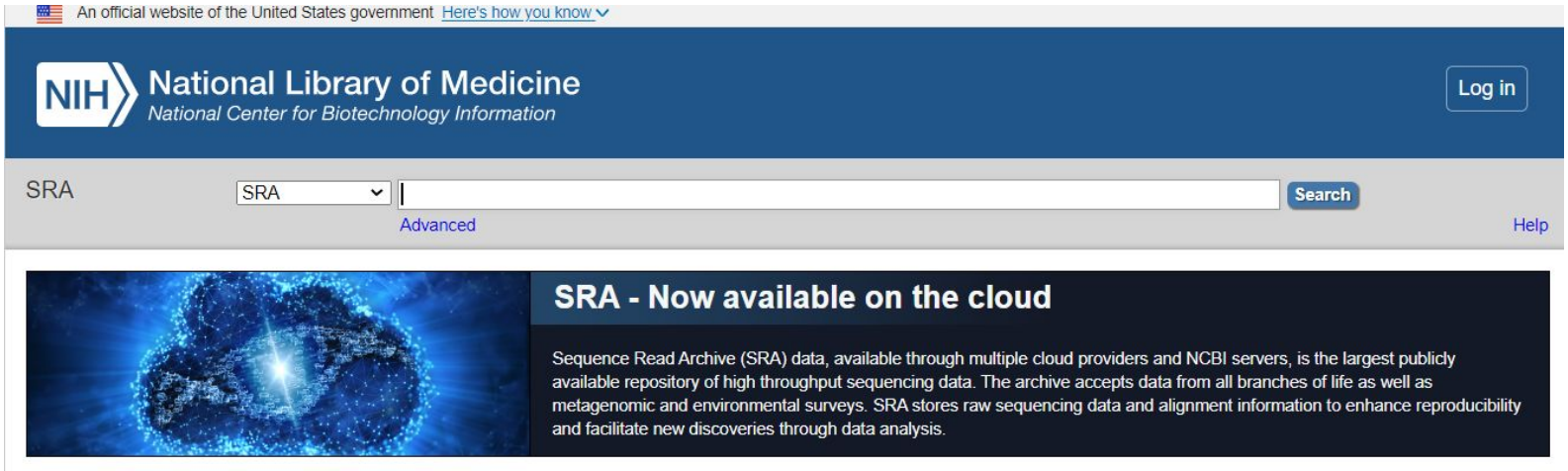


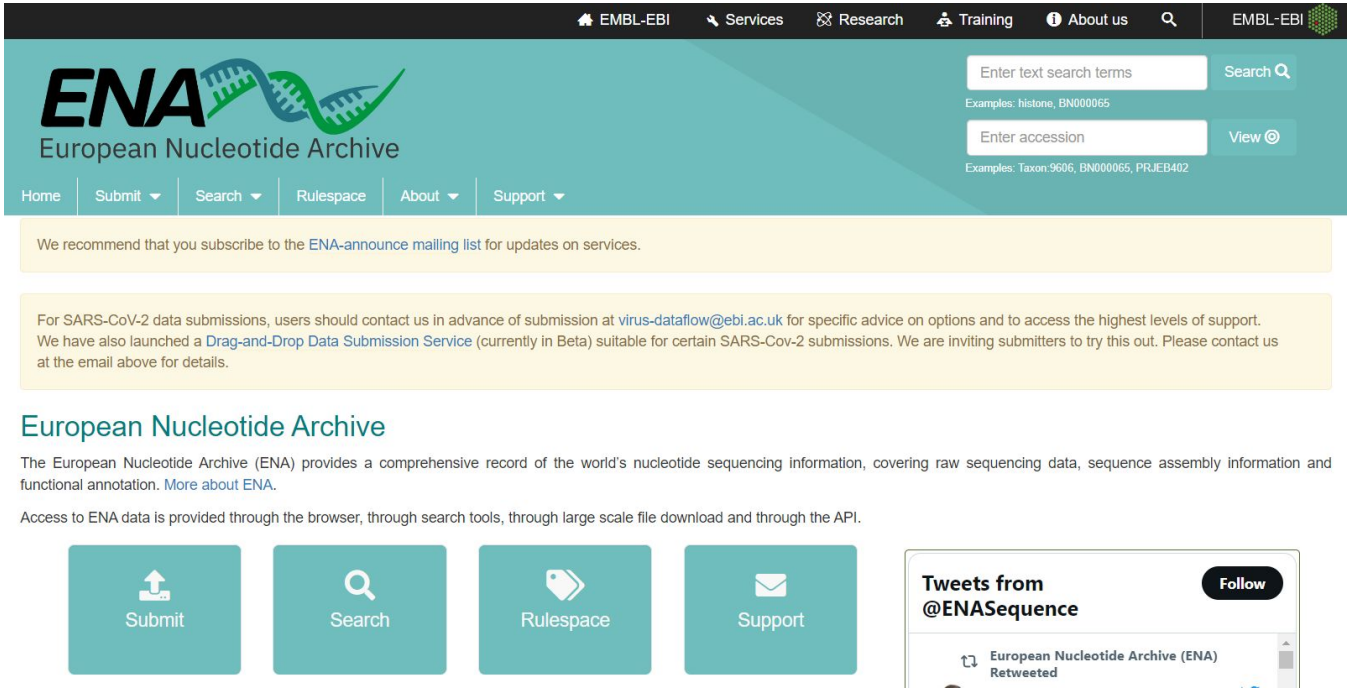
Phylogenetic distribution of Bacterial Genome Projects

GOLD
GENOMES ONLINE DATABASE

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

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





Sequencing project

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental

- Question scientifique => quelle stratégie ? Quel échantillonnage ?

  Quelle stratégie bioinfo ?

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental

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

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental

- Question scientifique => quelle stratégie ? Quel échantillonnage ? Quelle stratégie bioinfo ?
- Quel méthodo de séquençage ? Quelle couverture de séquençage ?
- Quel volume de données brut? Sur quel cluster les analyses bioinformatiques vont-elles être tournées ?

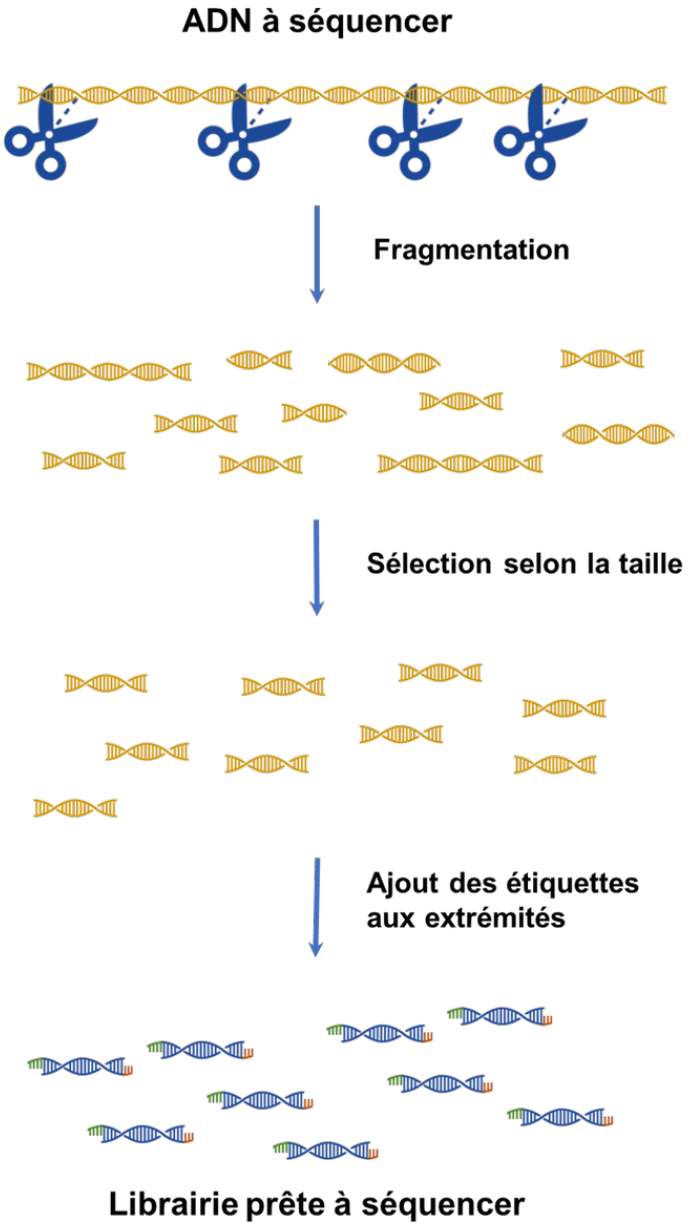
OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental

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

OVERVIEW OF DNA SEQUENCING PROJECT

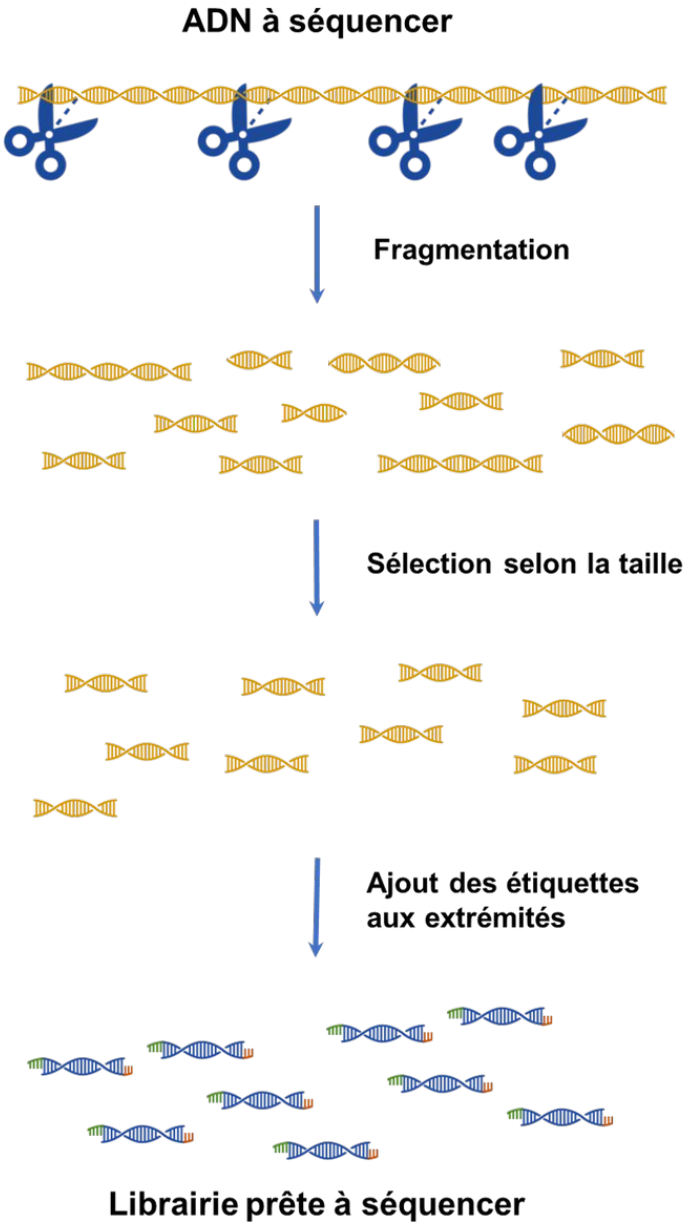
Design expérimental → Préparation banque



OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque

- Adaptateurs
- Contamination



OVERVIEW OF DNA SEQUENCING PROJECT

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

- Qualité de séquençage
- Profondeur de séquençage

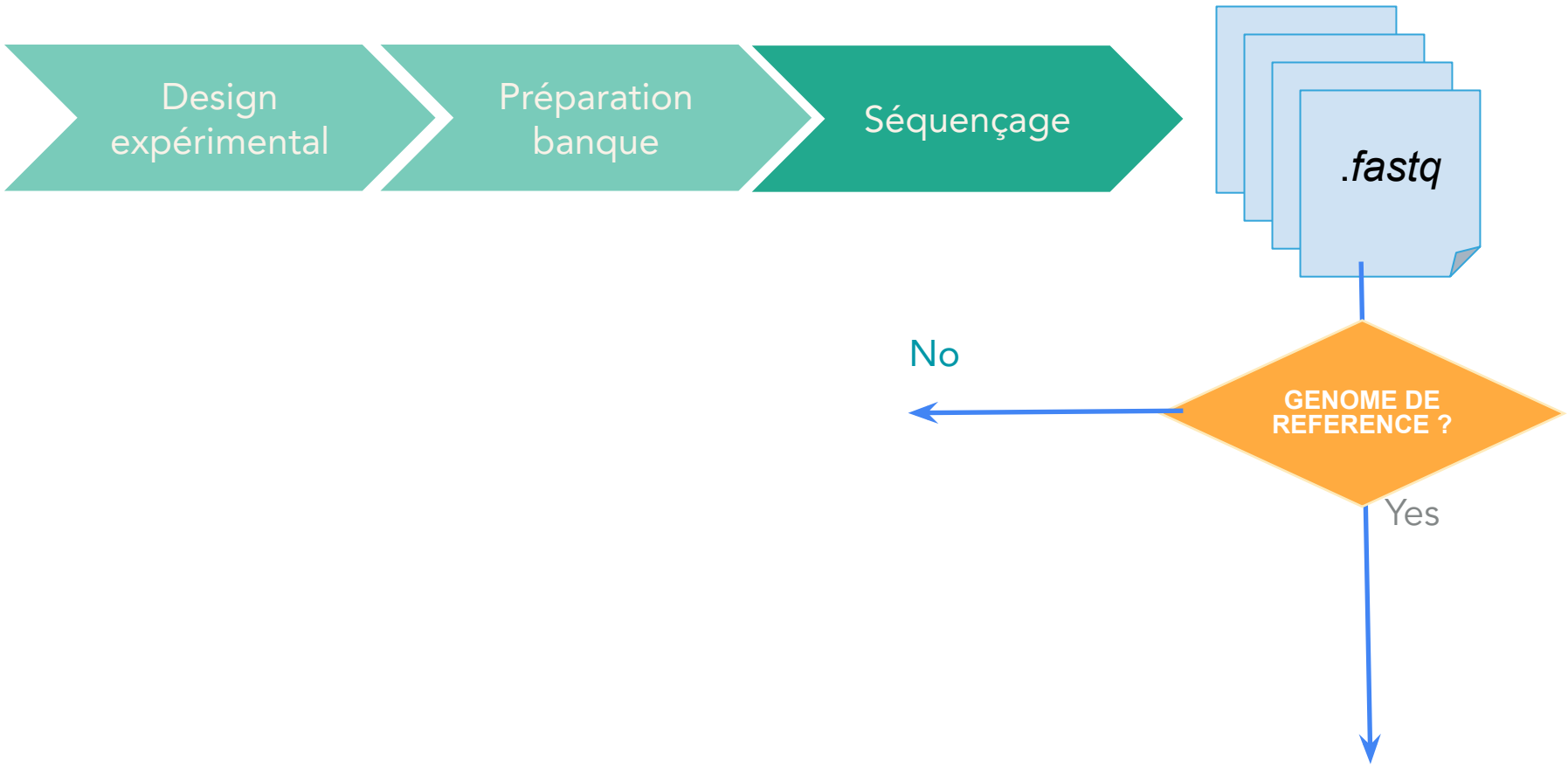
OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

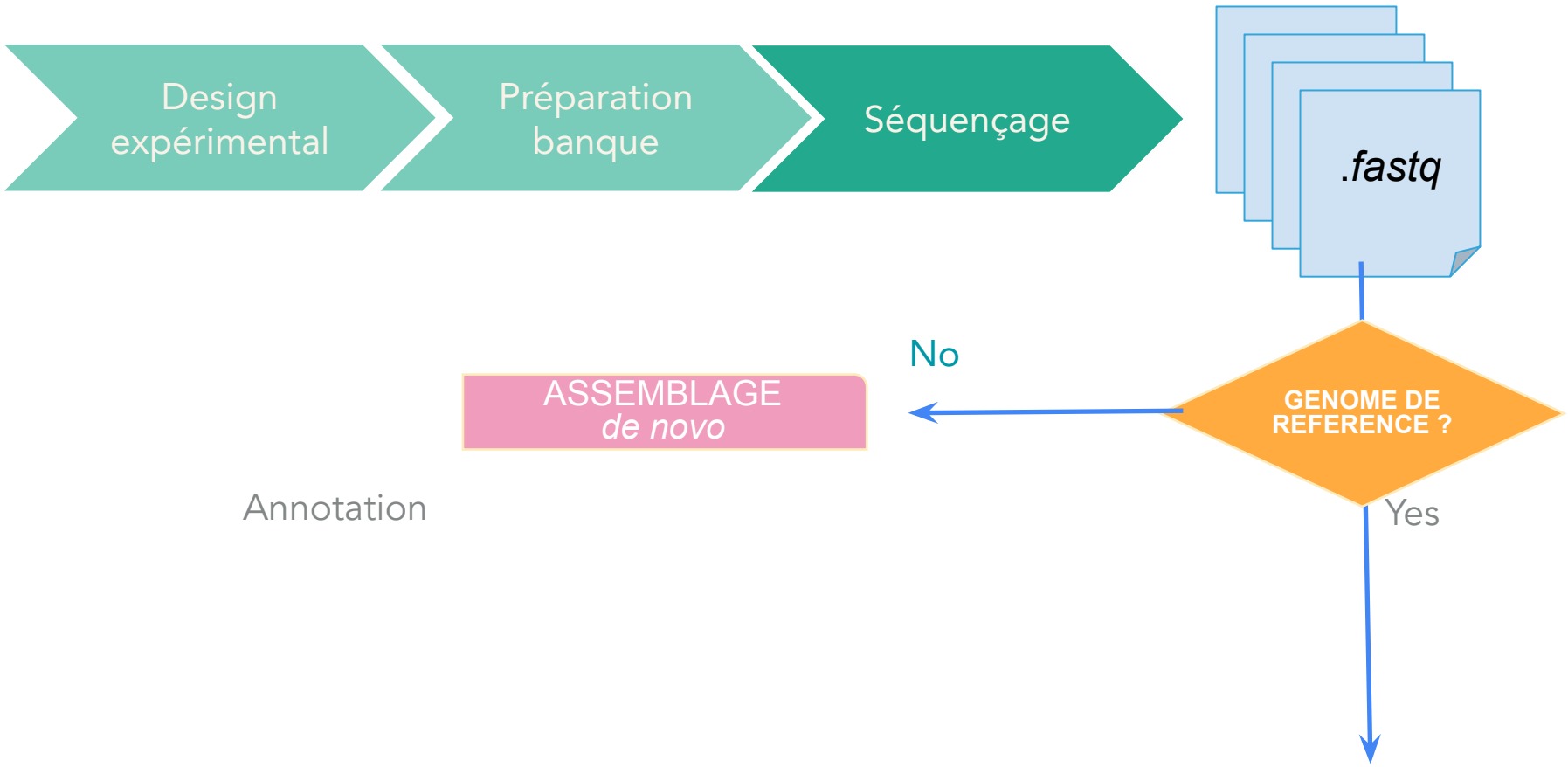
Genomic DNA is fragmented (not Nanopore) and sequenced -> millions of small sequences (reads) from random parts of the genome

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

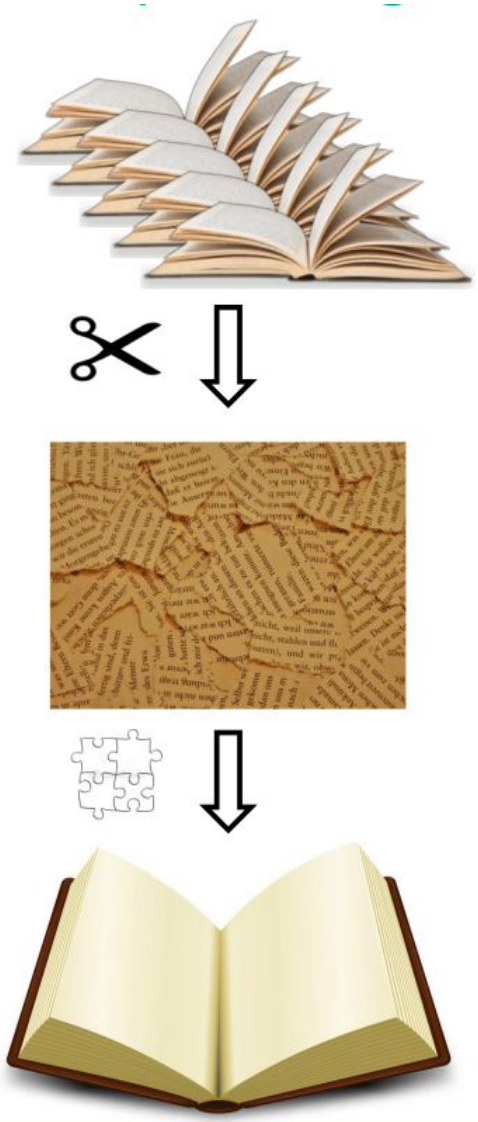
OVERVIEW OF DNA SEQUENCING PROJECT



OVERVIEW OF DNA SEQUENCING PROJECT

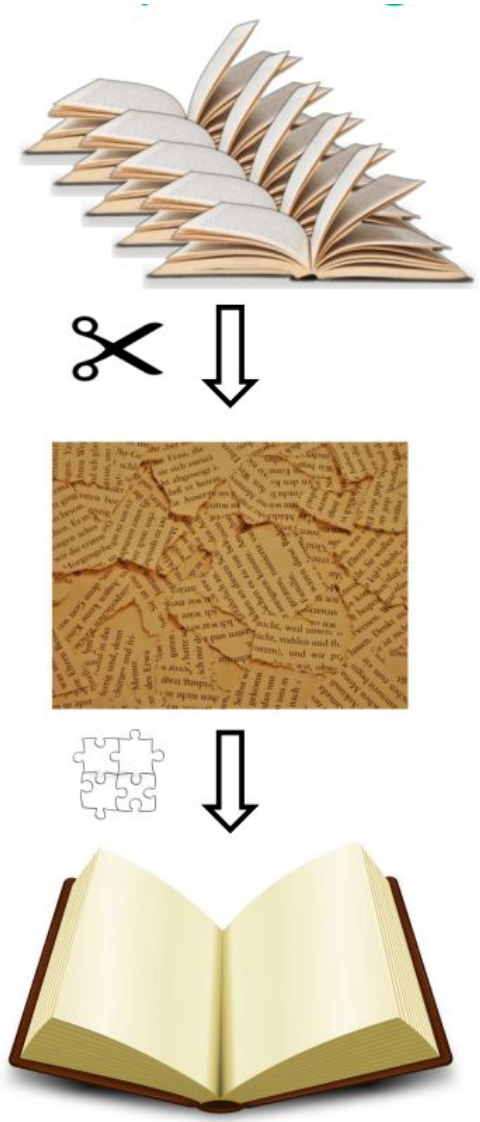


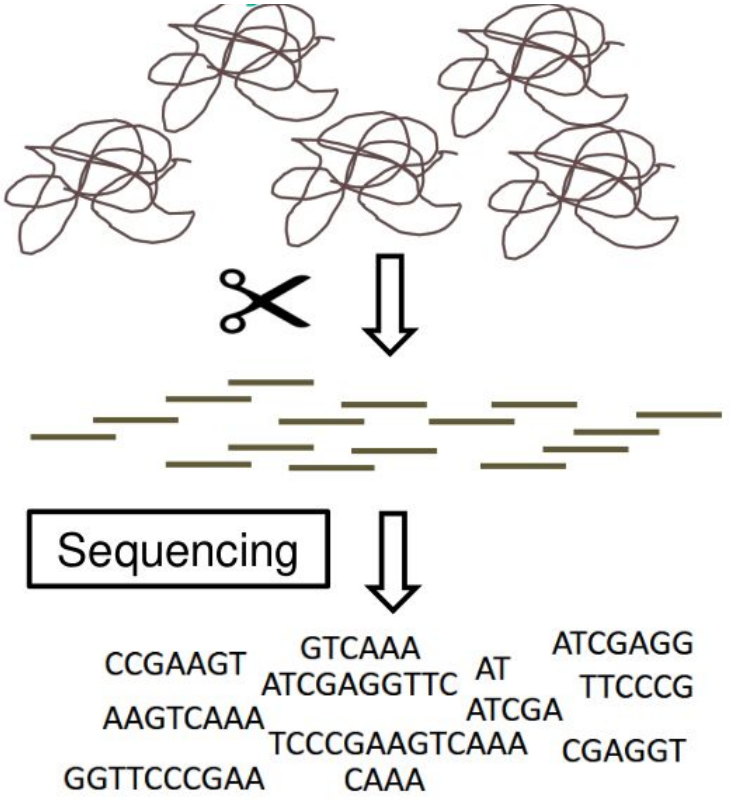
Sequencing and genome assembly



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

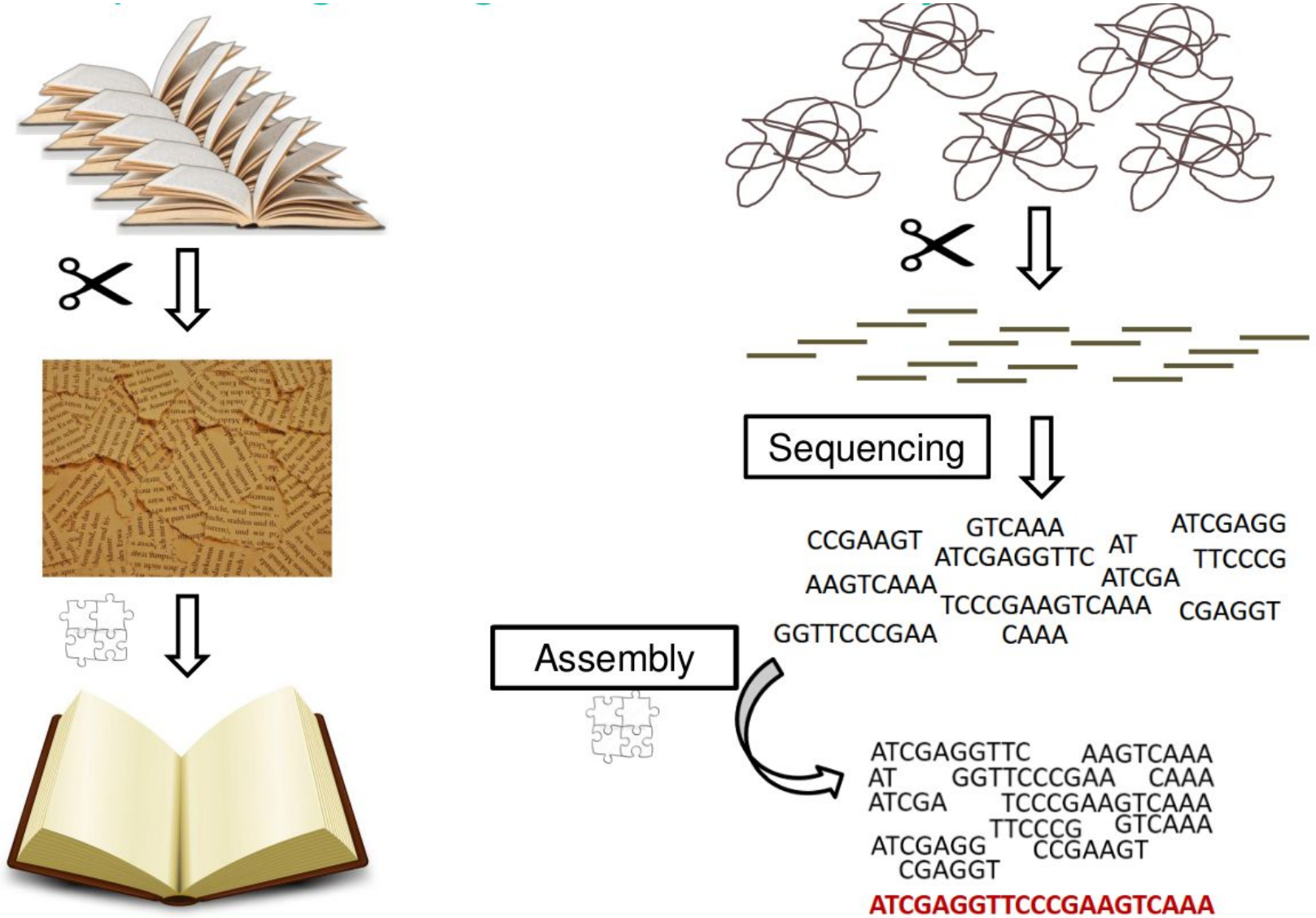
Sequencing and genome assembly





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

Sequencing and genome assembly

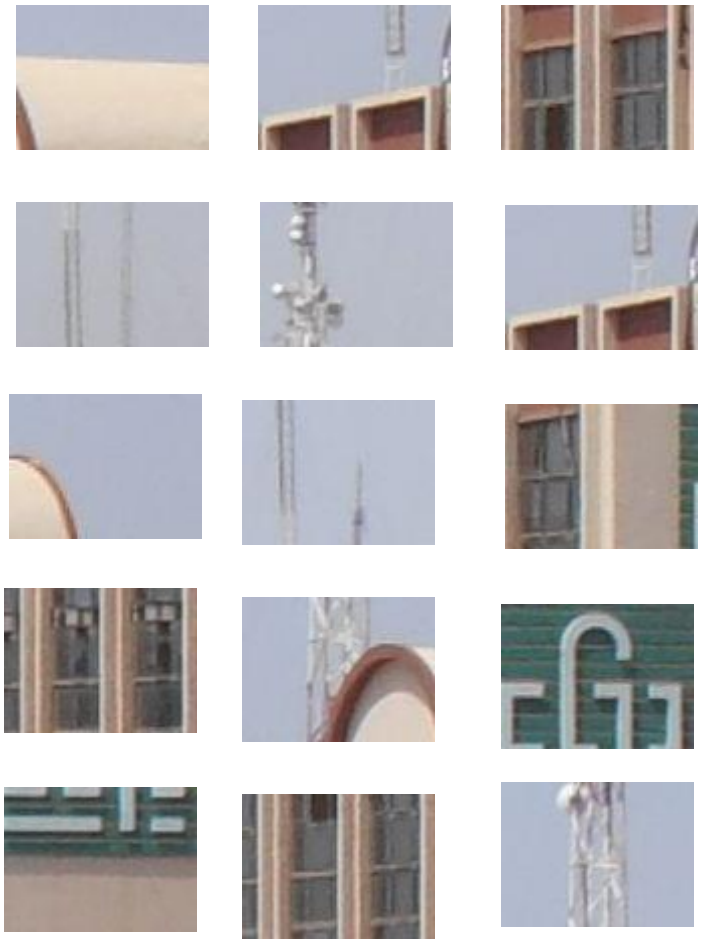


From Camille Rustenholz (Univ. Strasbourg, inrae) - Methods for plant genome assembly

Sequencing and genome assembly

Puzzle 400 pièces “petite taille”

Puzzle 400 pièces “petite taille”



+ 100 pièces “ciel” + …

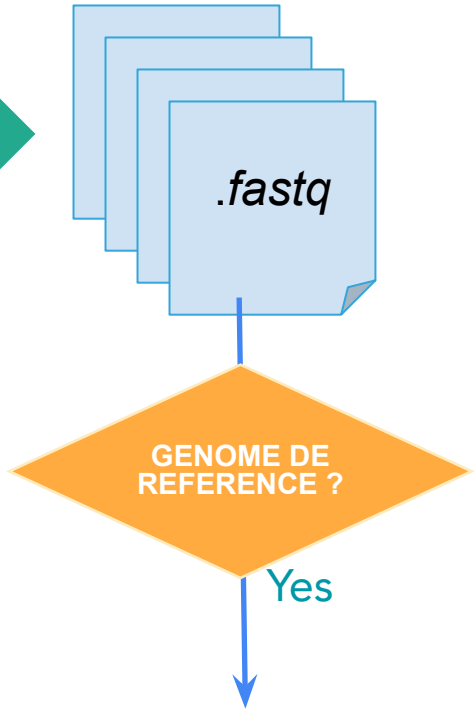
Puzzle 100 pièces - “grande taille”

Puzzle 100 pièces “grande taille”

\+ ~ 20 pièces “ciel”

OVERVIEW OF DNA SEQUENCING PROJECT

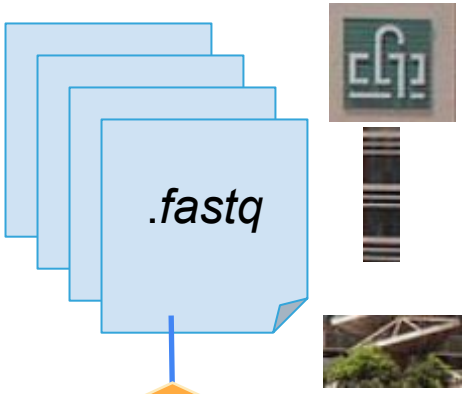
Design expérimental → Préparation banque → Séquençage



Adapted from Ross Whetten…

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage



Adapted from Ross Whetten…

OVERVIEW OF DNA SEQUENCING PROJECT

Design expérimental → Préparation banque → Séquençage → .fastq

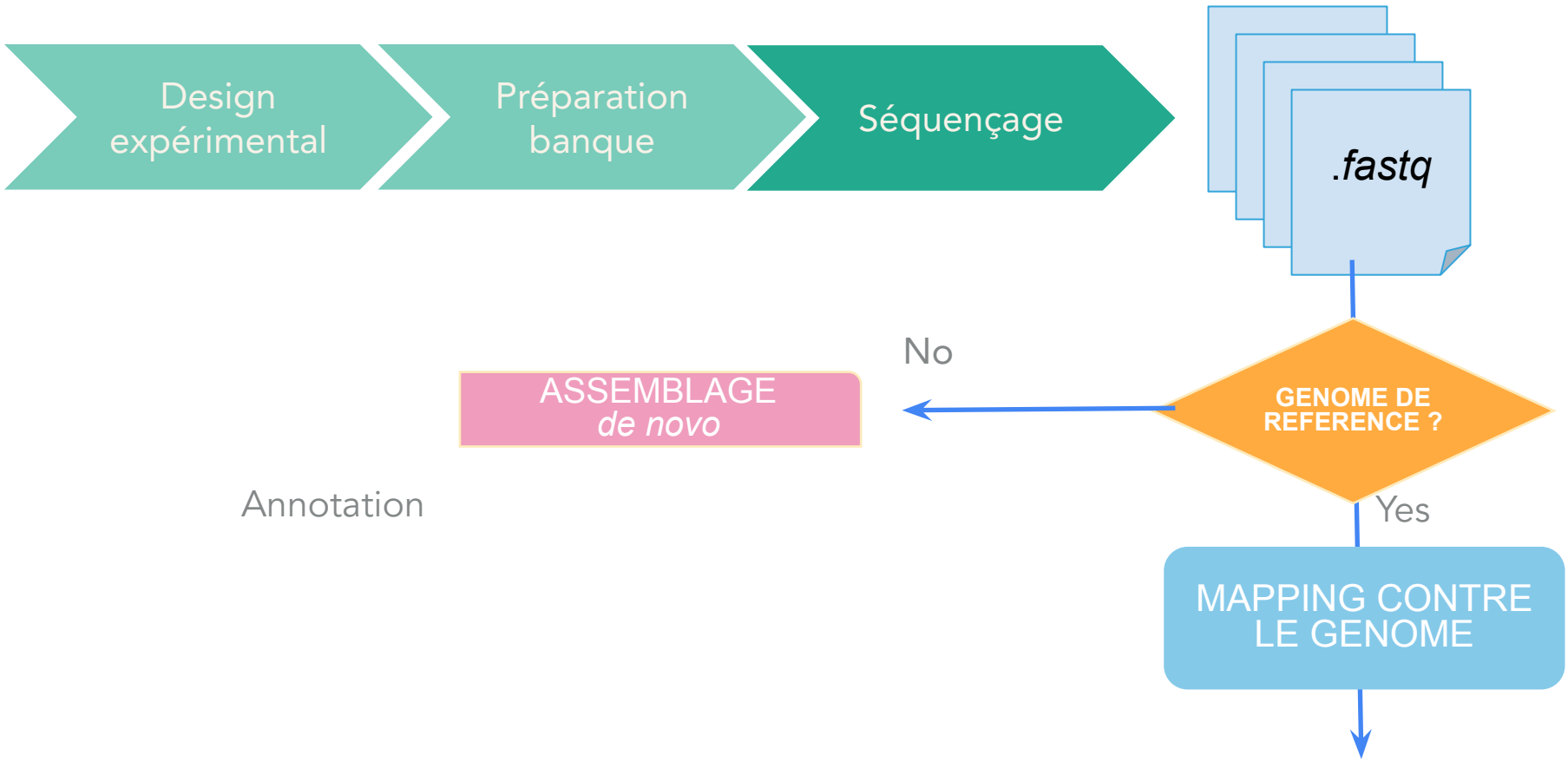
GENOME DE REFERENCE ?

Yes

MAPPING CONTRE LE GENOME

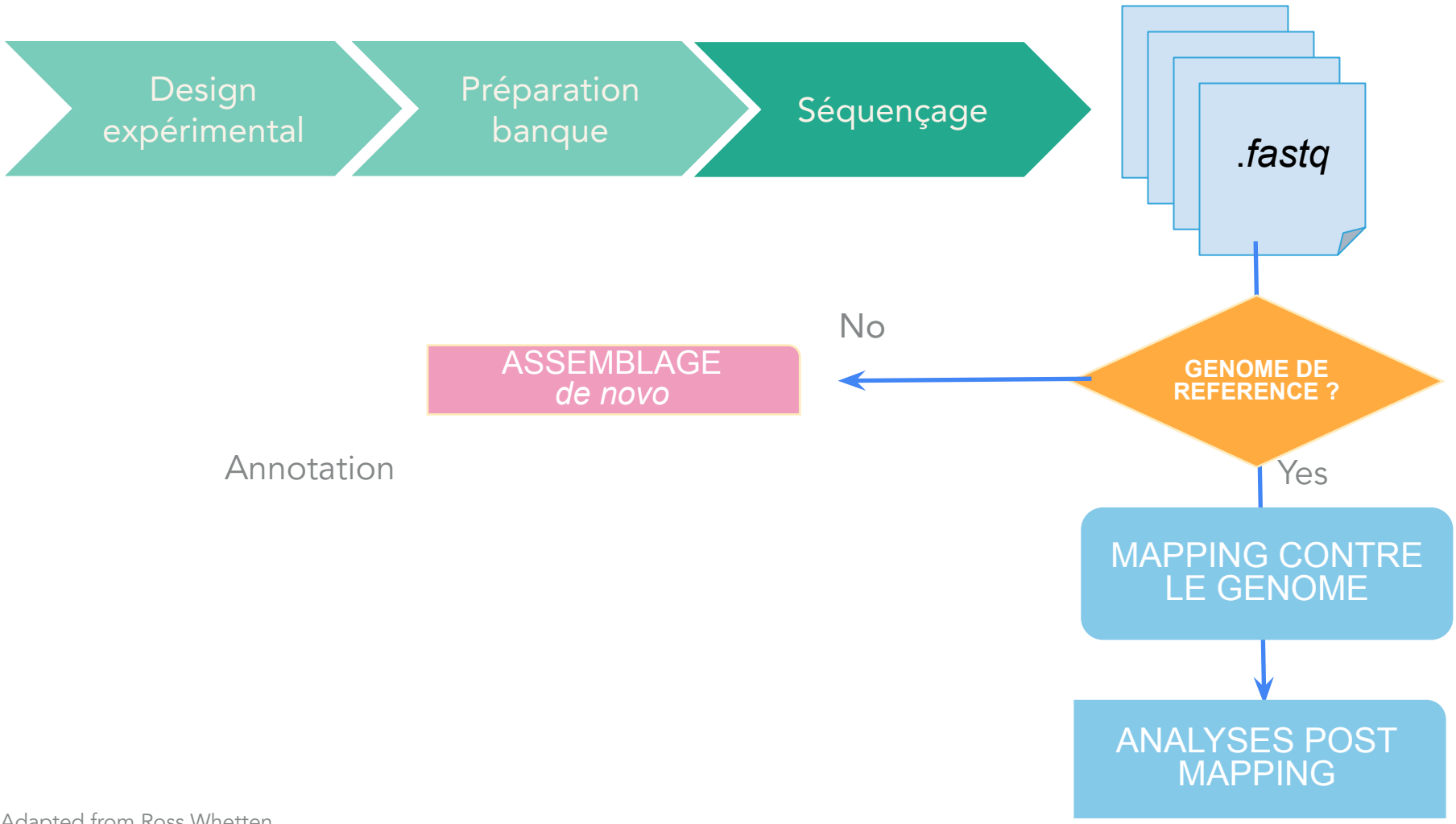
Adapted from Ross Whetten…

OVERVIEW OF DNA SEQUENCING PROJECT



Adapted from Ross Whetten…

OVERVIEW OF DNA SEQUENCING PROJECT



Adapted from Ross Whetten…

What metagenomics is ?

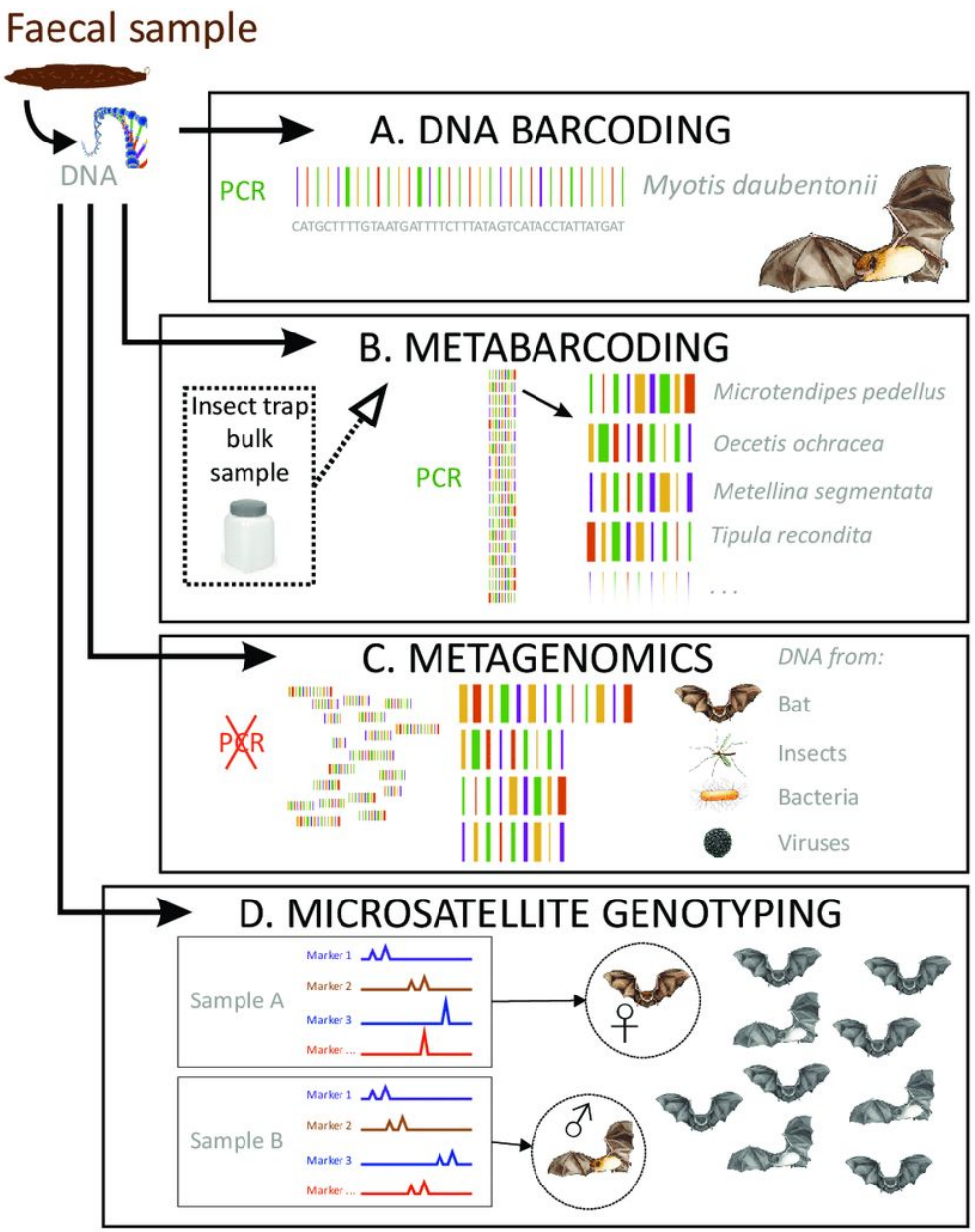
Metagenomics (Environmental Genomics or Community Genomics) is the study of genomes recovered from environmental samples without the need for culturing them

Metagenomics processes data using bioinformatics tools

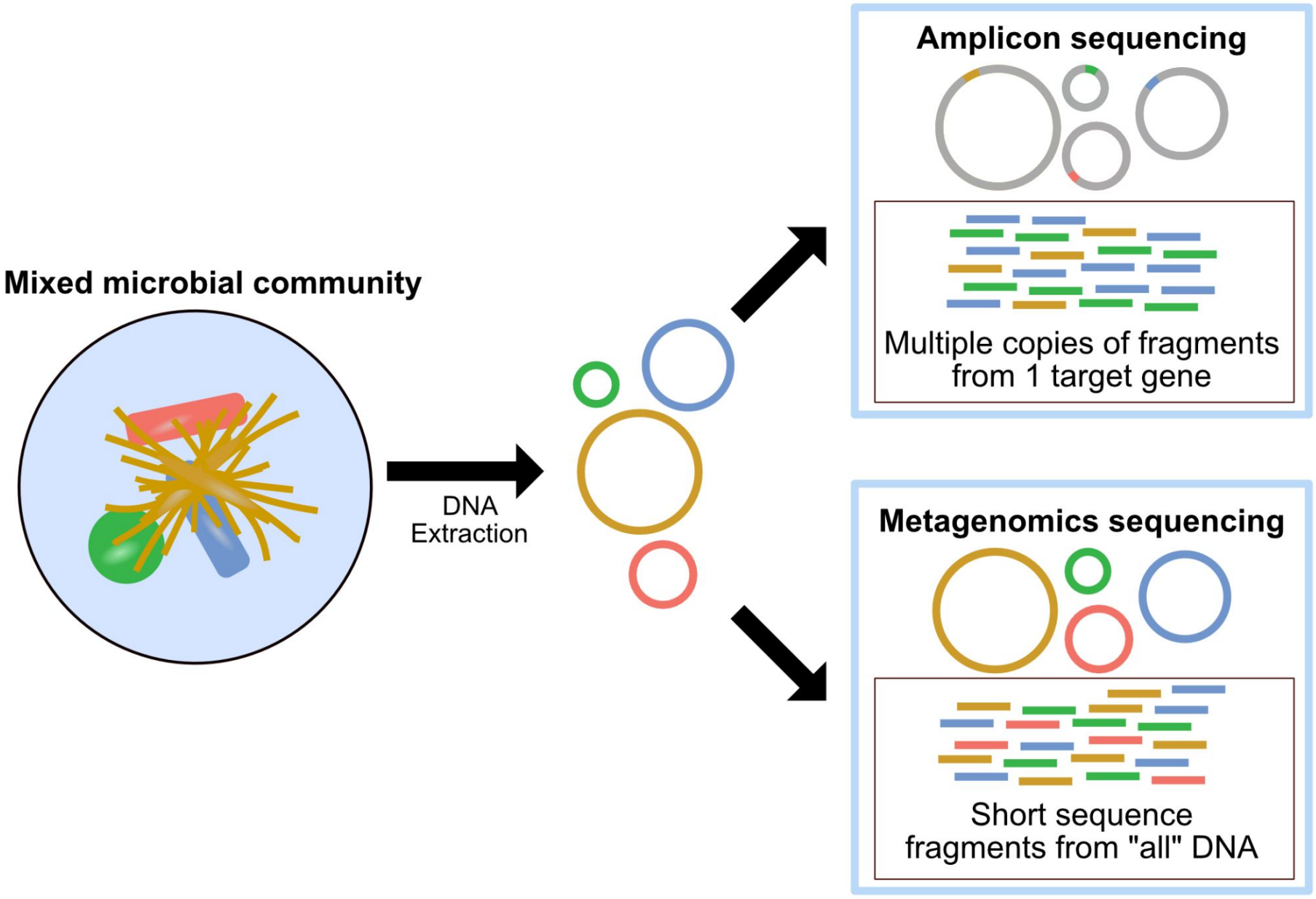
=> Organisms can be studied directly in their environments bypassing the need to isolate each species

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

Application: Metagenomics

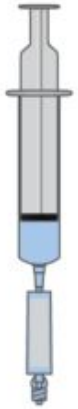


Application: Metagenomics

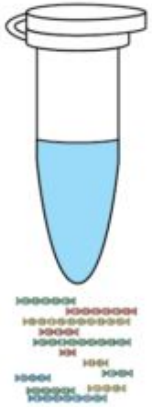


Application: Metabarcoding

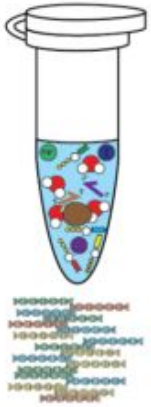
1 Collect an environmental sample



2 DNA extraction from environmental sample

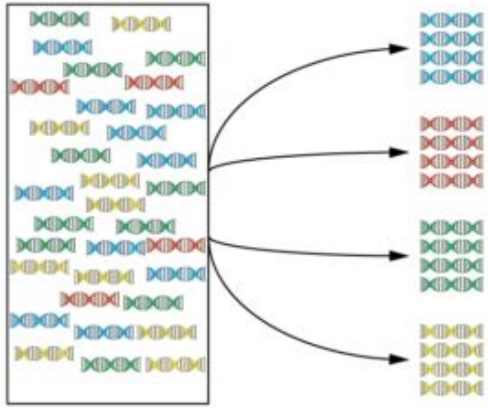


3 Amplify DNA markers



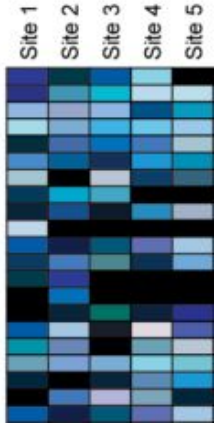
4 High-throughput sequencing

5 Bioinformatic processing



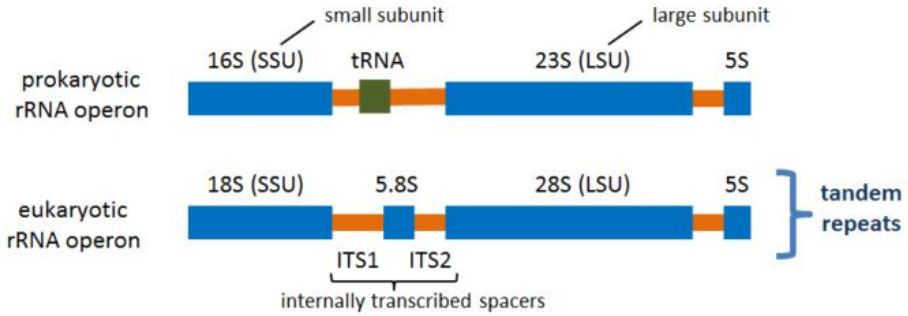
6 Species identification

7 Ecological analysis



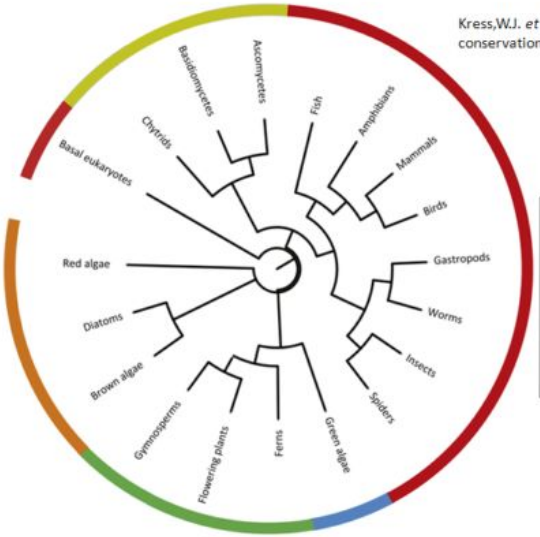
http://www.sixthresearcher.com/amplicon-sequencing-and-high-throughput-genotyping-metagenomics/

Application: Metabarcoding



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

Which barcode to choose?



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

Tree of life

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

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

http://www.barcodeoflife.org/

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

Application: Metagenomic

Realtime analysis provides rapid answers

Detection & characterization of bacterial pathogens

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

Journal of Antimicrobial Chemotherapy, Volume 72, Issue 1, 1 January 2017, Pages 104–114

Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing

K. Schmidt D. M. Livermore

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

Journal of Clinical Microbiology - 19th December 2016

Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples

Antonina A. Votintseva

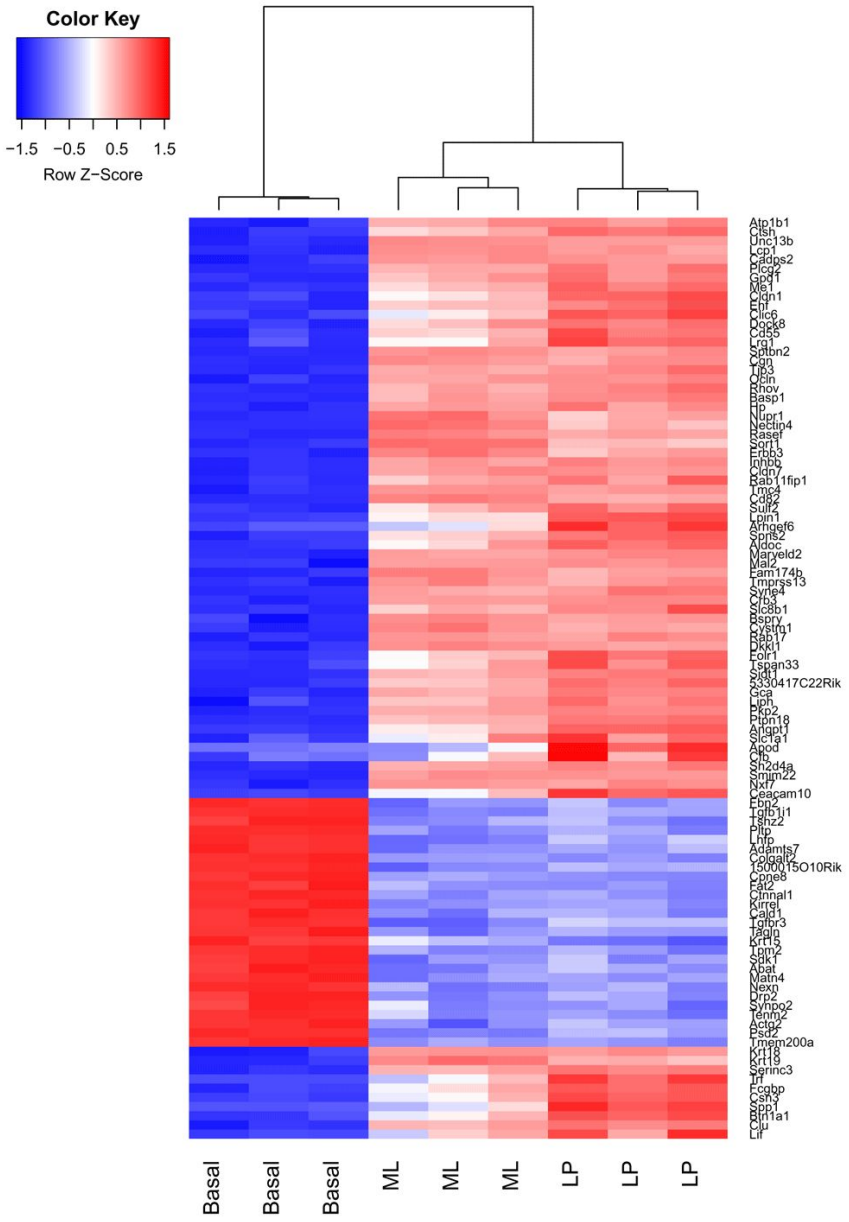
"the estimated turnaround time from patient to identification of BCG was 6 hours, with full susceptibility and surveillance results 2 hours later"

Oxford NANOPORE Technologies

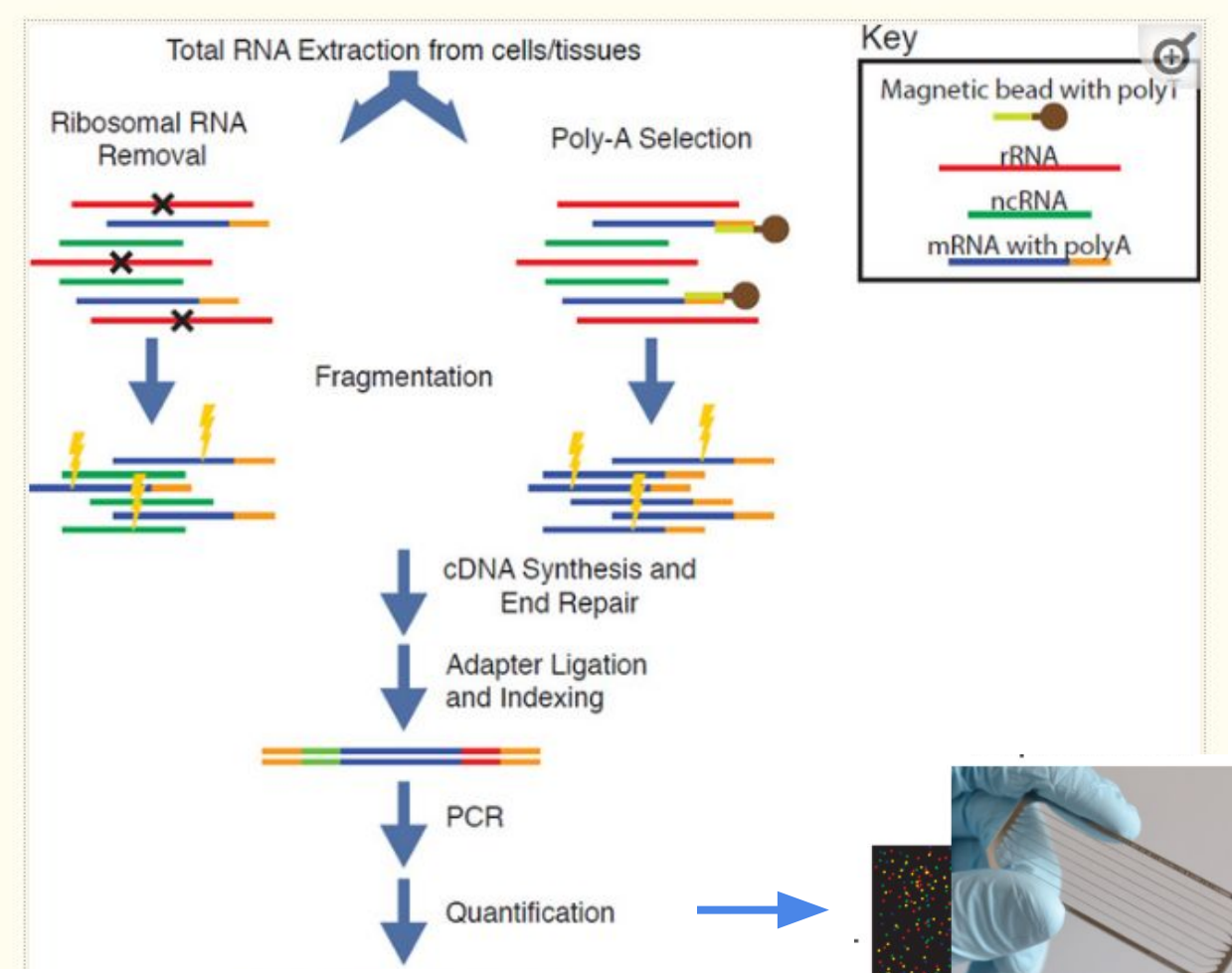
Application: Transcriptomics / RNASeq

Pourquoi faire du RNAseq ?

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

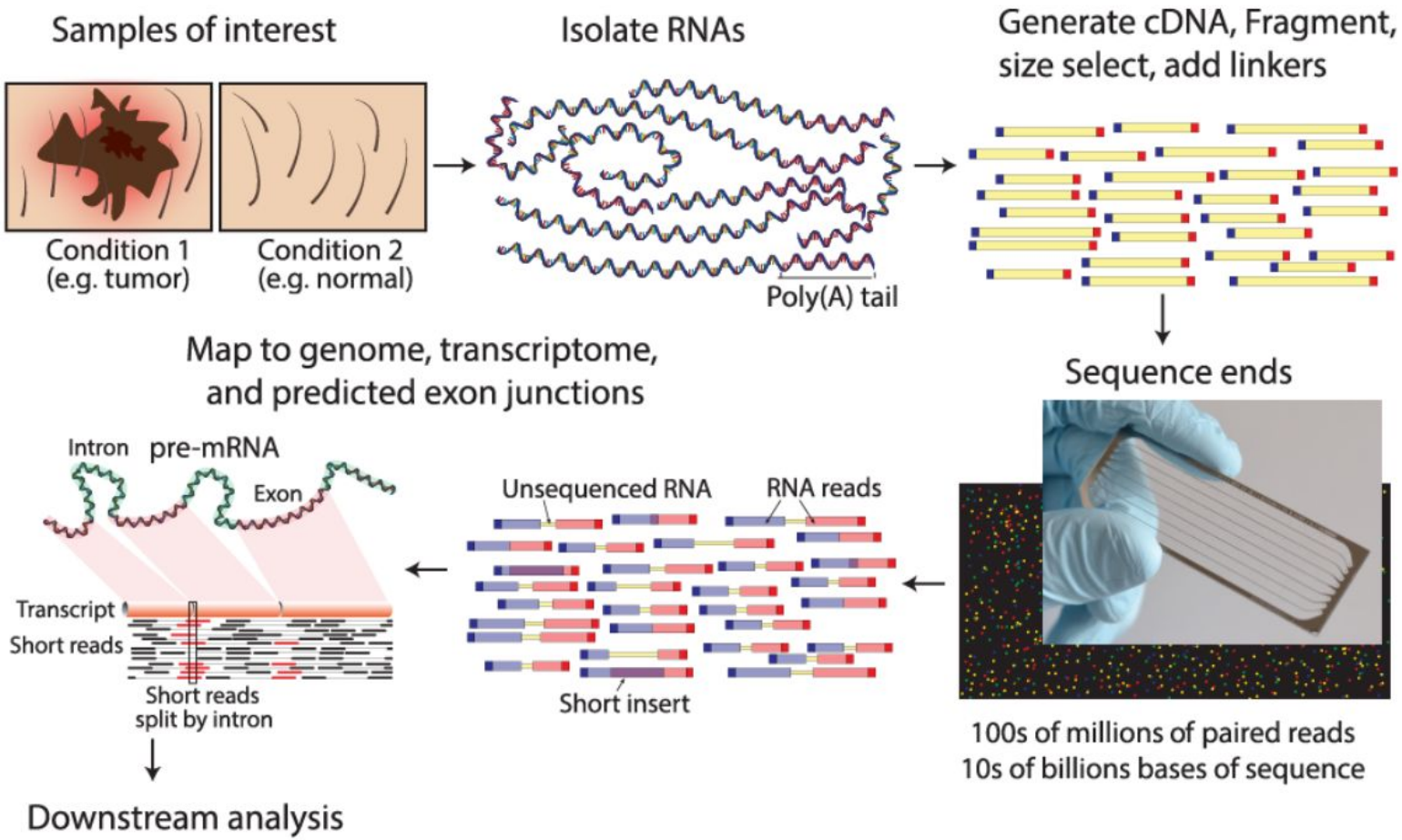


Application: Transcriptomics / RNASeq

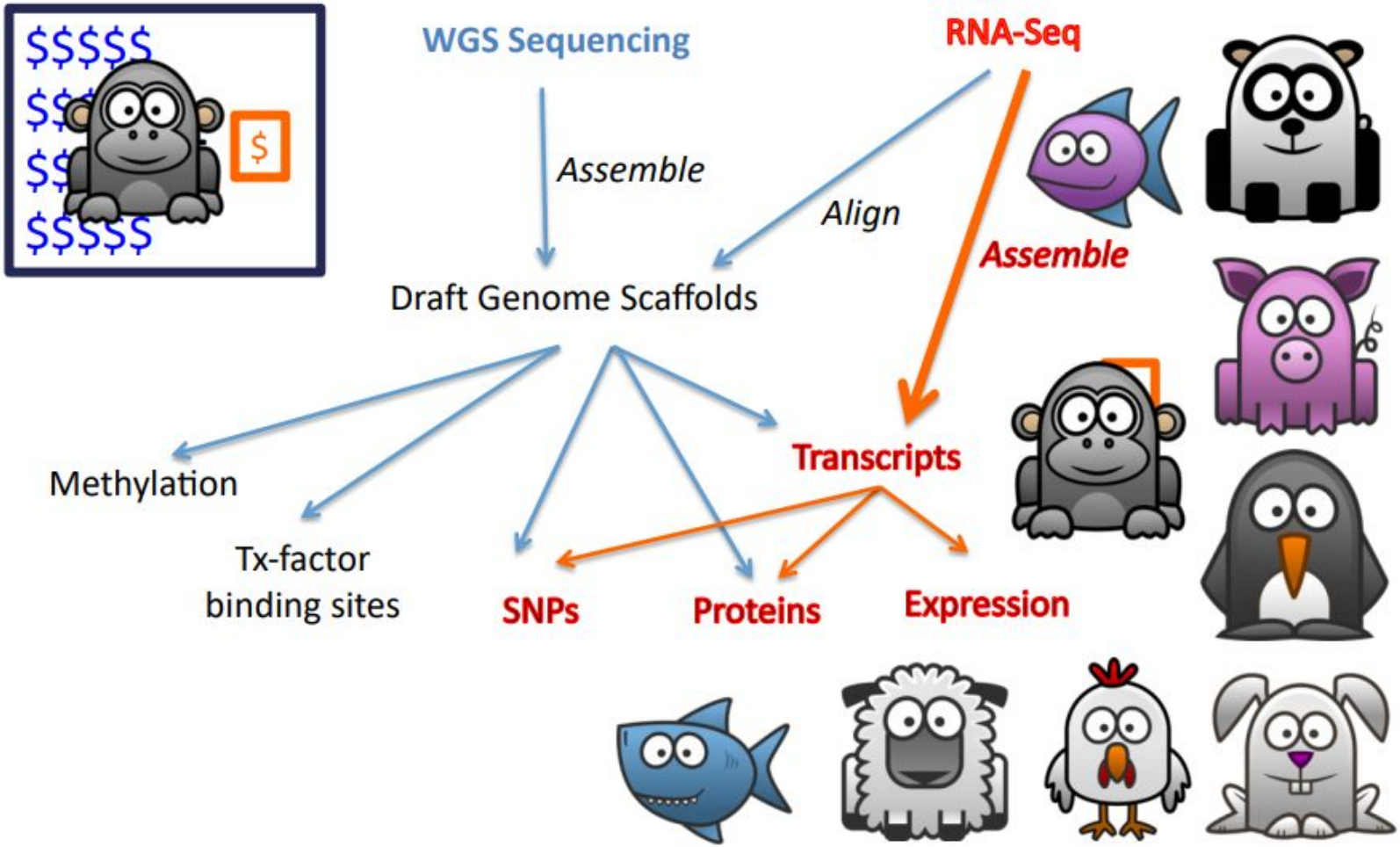


100s of millions of paired reads
10s of billions bases of sequence

RNA sequencing

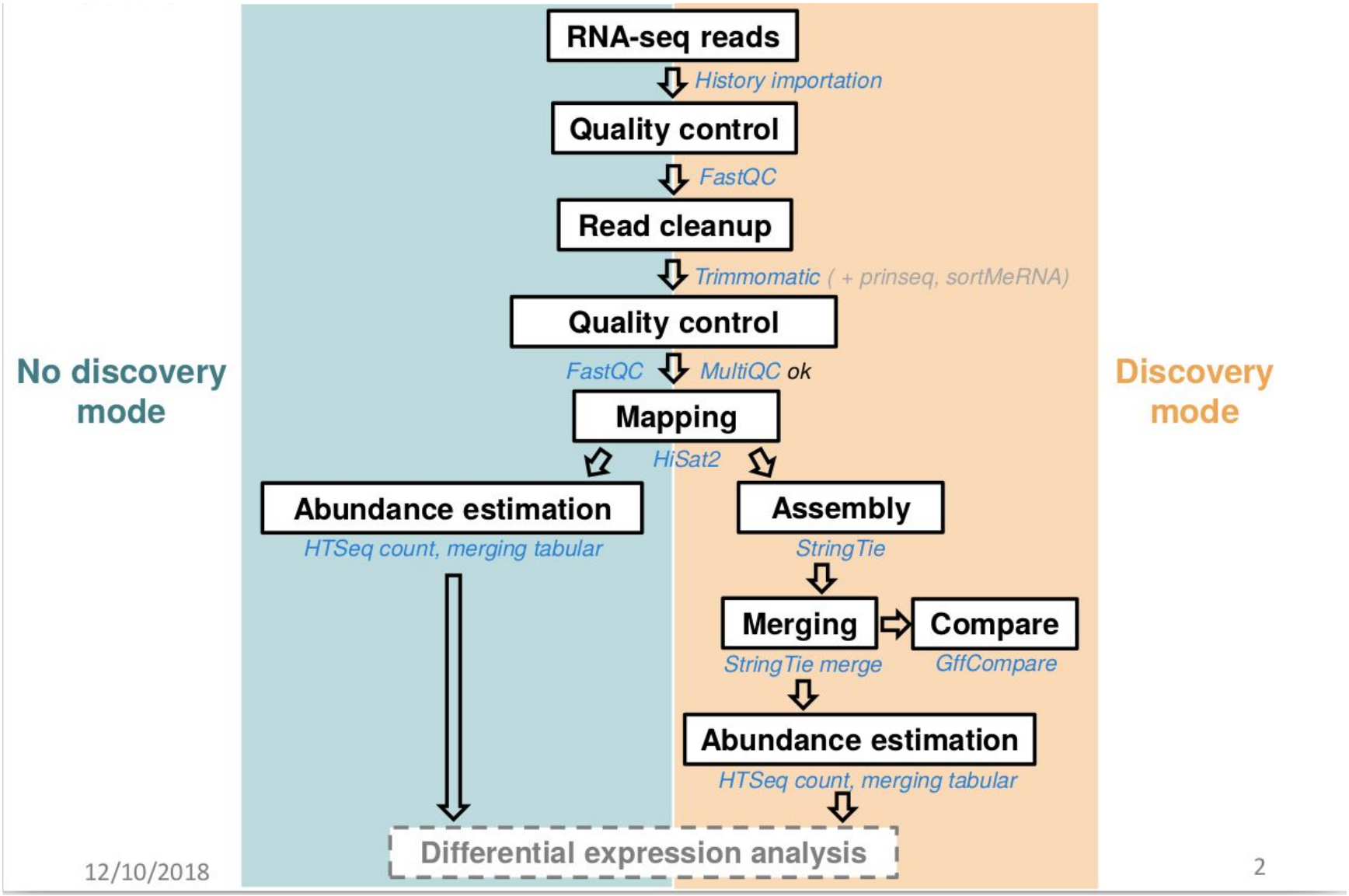


Application: Transcriptomics / RNASeq



From SG-ABiMS RNAseq Training Sept 2019

Application: Transcriptomics / RNASeq



From Abims RNAseq Training oct 2018

Application: Transcriptomics / RNASeq

⇒ Comparaison entre conditions expérimentales différentes

Ex:

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

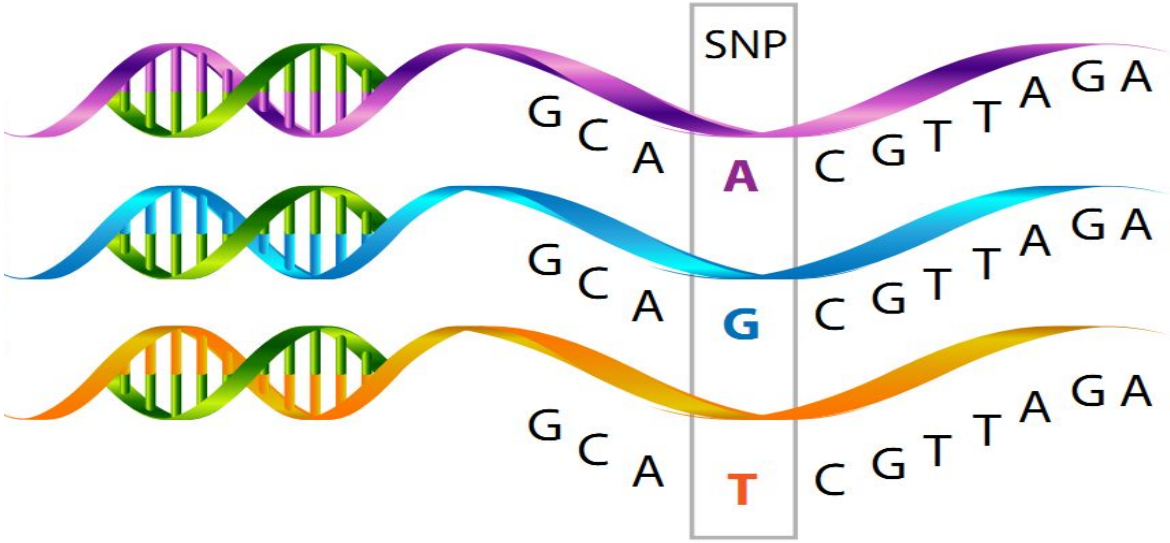
⇒ Comparaison dans le temps (time series): cinétique

Ex:

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

=> logiciels dédiés pour ce type de problèmatique

Single Nucleotide Polymorphism



Origin of domestication and evolutionary history of African crop?

Where, when, how, (why) ?

African rice

Pearl millet

Yam

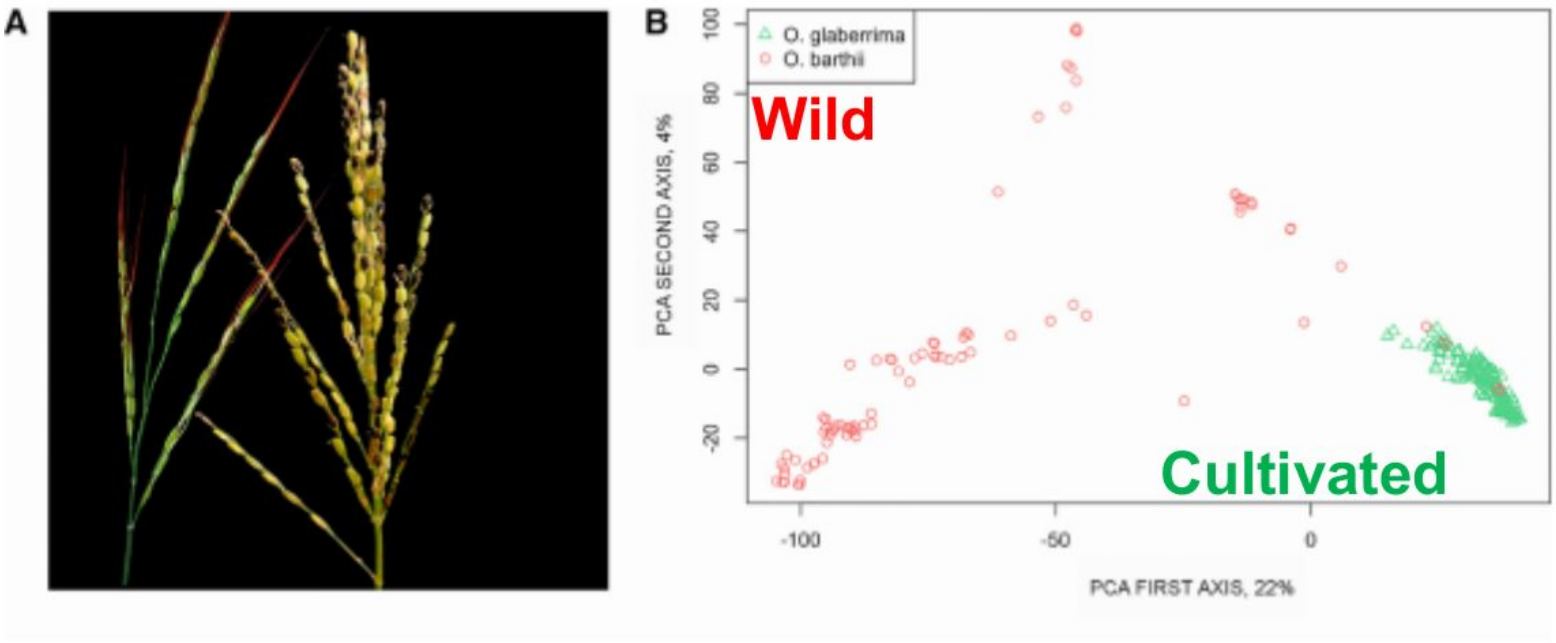
Fonio

Sorghum

From Y. Vigouroux

Application : African rice domestication

246 fully resequenced genomes
3 051 681 SNPs



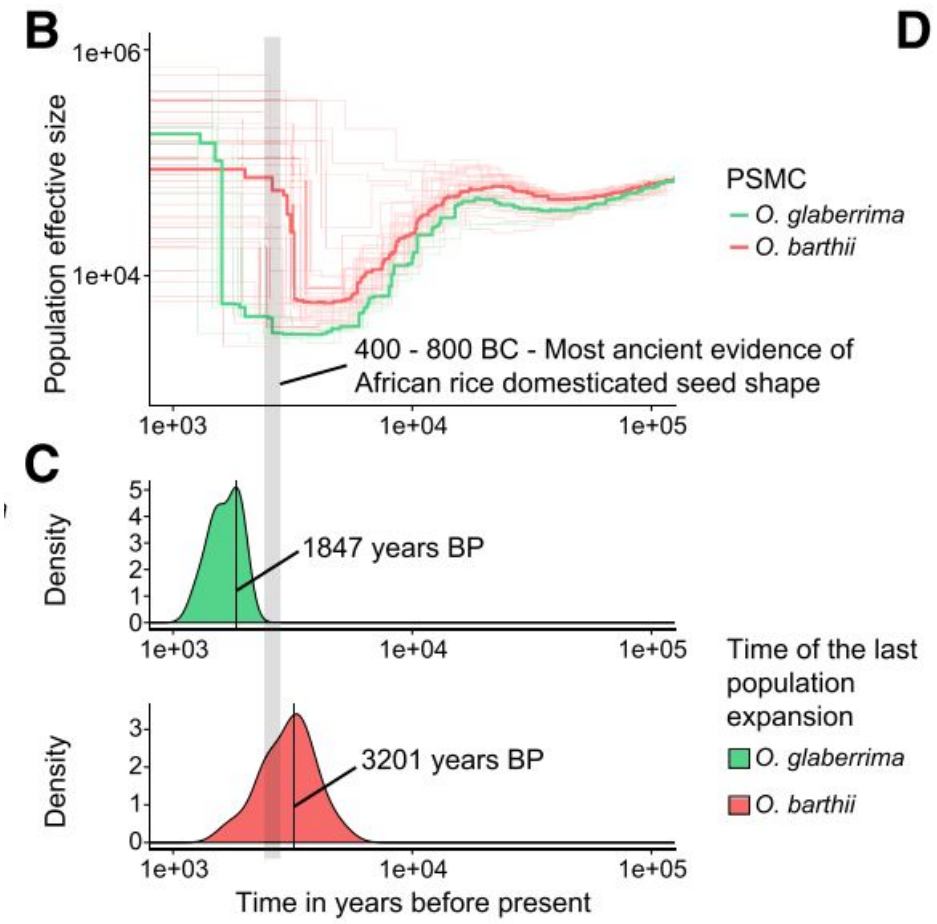
Cubry P, Tranchant-Dubreuil C, Thuillet AC, Monat C, *et al*. Current Biol 2018

From Y. Vigouroux

Application : Genomics help untangled past events

WHEN ?

Pairwise Sequentially Markovian Coalescent (PSMC)



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

From Cubry et al, 2018

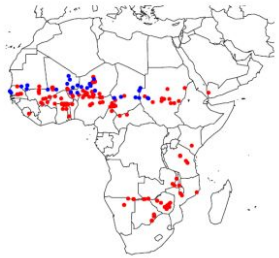
Application : Aproximate Bayesian spatial approach

WHERE ?

Simulated data

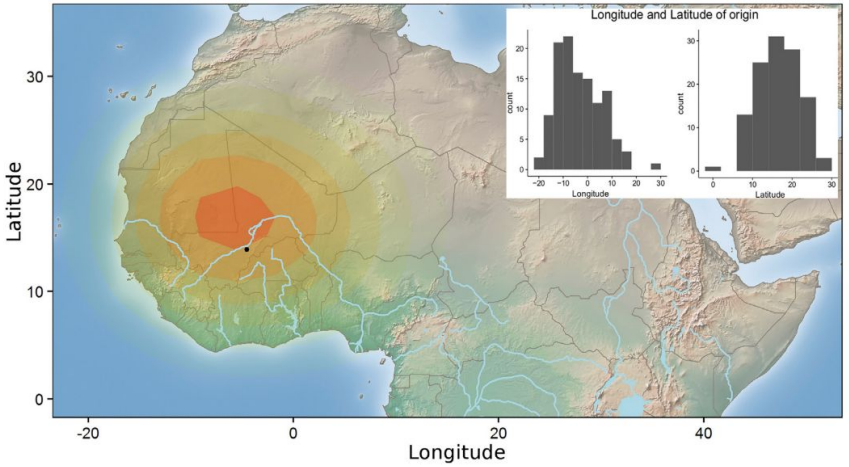
SPLATCH
Simulated diversity statistics
(SFS, rare variants)

Observed data



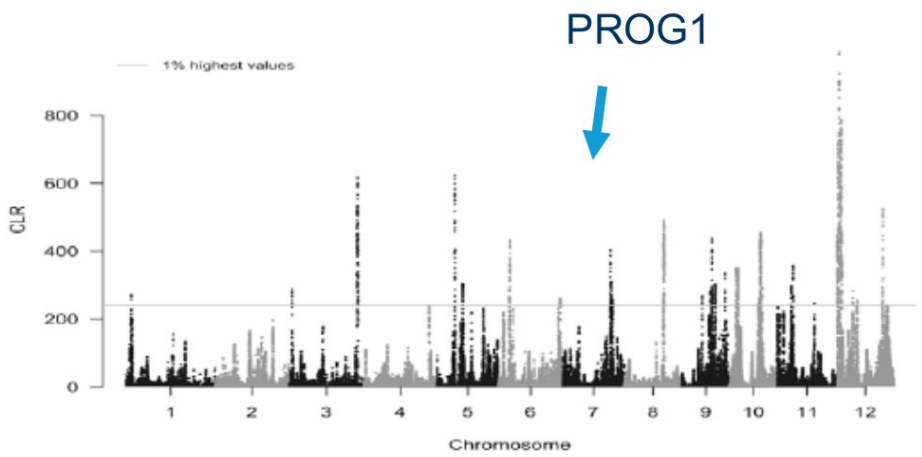
Observed diversity statistics
(SFS, rare variants)

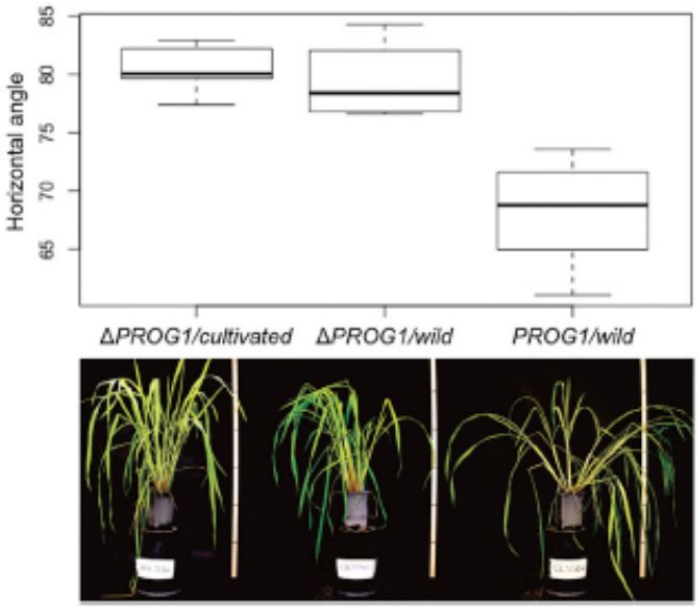
ABC estimation of probability of origin



From Cubry et al, 2018

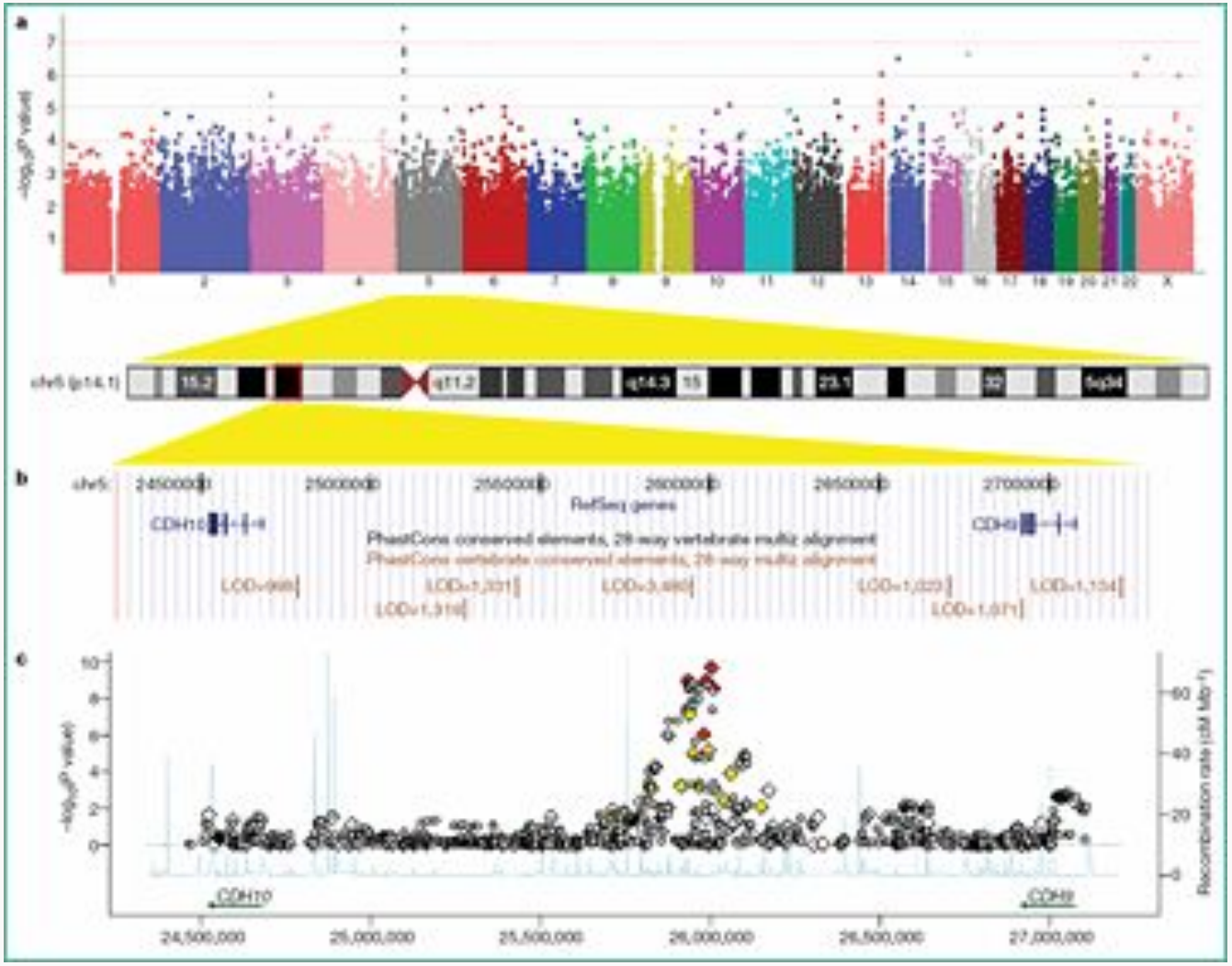
Prostrate growth 1



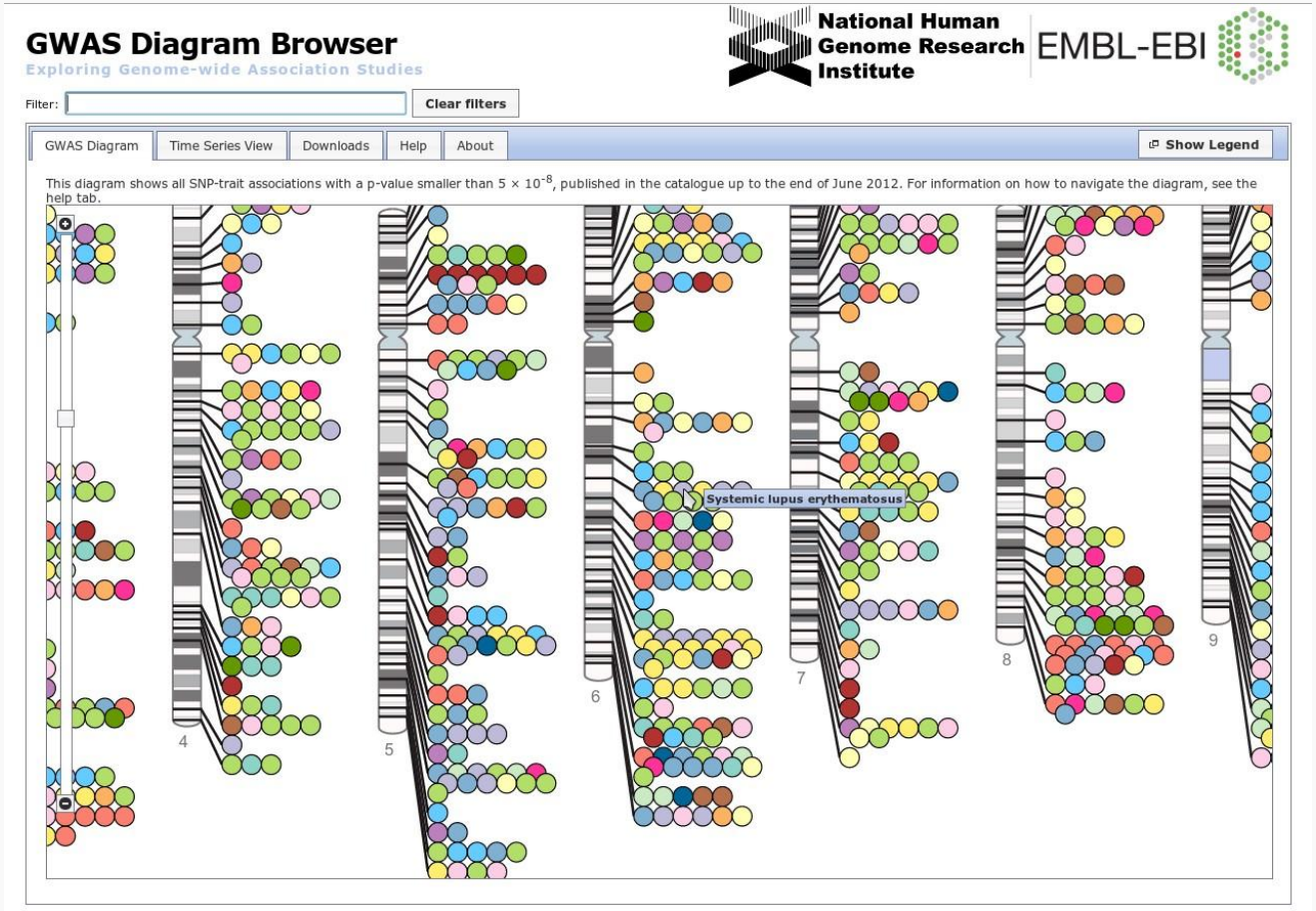


Prog1 deletion

Example in GWAs & Population Genomics



Example in GWAs & Population Genomics



Be Careful to data drowning!