



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

[www.southgreen.fr](http://www.southgreen.fr)

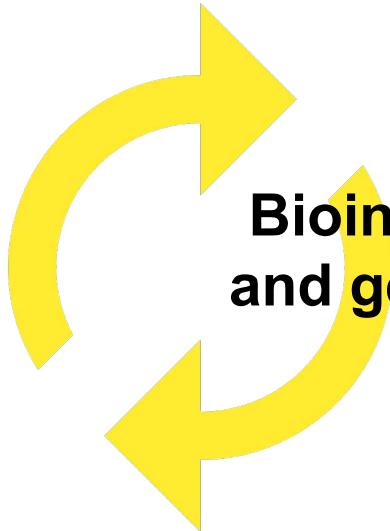
<https://southgreenplatform.github.io/trainings>





# Modules de formation 2022





Bioinformatics platform dedicated to the genetics  
and genomics of tropical and Mediterranean plants  
and their pathogens

comparative genomics  
phylogeny  
GWAS  
population genetics  
polyploidy

genome assembly  
transcriptome assembly  
metagenomics

SNP detection  
structural variation  
differential expression



Rice



Banana



Palm



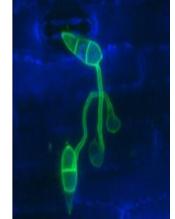
Sorghum



Coffee



Cassava



Magnaporthe

# South Green

bioinformatics platform



Larmande Pierre  
**Orjuela-Bouniol Julie**  
Sabot François  
Tando Ndomassi  
**Tranchant-Dubreuil Christine**



Comte Aurore  
Dereeper Alexis  
**Ravel Sébastien**



Bocs Stephanie  
Boizet Alice  
De Lamotte Frédéric  
**Droc Gaetan**  
Dufayard Jean-François  
Hamelin Chantal  
Martin Guillaume  
Pitollat Bertrand  
**Ruiz Manuel**  
**Sarah Gautier**  
Summo Marilyne



**Rouard Mathieu**  
Guignon Valentin  
Catherine Breton



Sempere Guilhem



# South Green bioinformatics platform

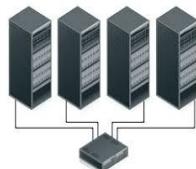
## Workflow manager

TOGGLE  
Toolbox for generic NGS analyses



Galaxy

## HPC and trainings....



## Genome Hubs & Information System



Gigwa

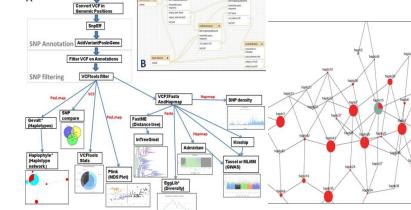
A screenshot of the Gigwa user interface showing a table of SNP and Indel data with various filters and search options.

SNPs and Indels

GreenPhyl

Family Id	Family Name	Number of sequences	Status
GP000010	Cytochrome P450 superfamily	6942	green
GP000017	AP0/DERB1 transcription factor family; ERF/DRB group (partial)	5142	green
GP000020	NAC transcription factor family	4574	green
GP000028	MADS transcription factor family		
GP000018	Haem peroxidase superfamily		
GP000066	General substrate transporter superfamily		
GP000022	Subtilisin-like Serine Proteases family		
GP000019	NPF, NRT1/PTR FAMILY		

Gene families



SNiPlay



<https://github.com/SouthGreenPlatform>

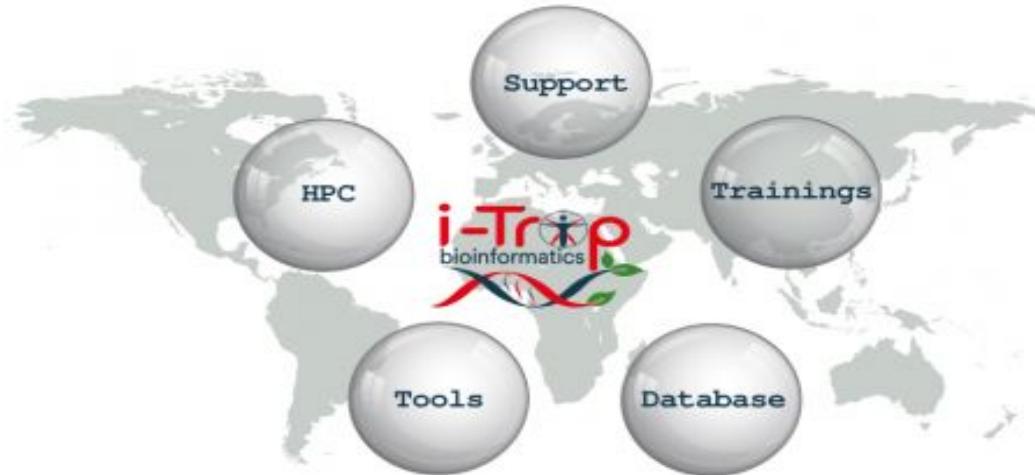


@green\_bioinfo

*The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics*, Current Plant Biology, 2016

# i-Trop

Plant & Health Bioinformatics Platform



<https://bioinfo.ird.fr/>



AURORE  
COMTE

ALEXIS  
DEREPPER

BRUNO  
GRANOUILAC

JULIE  
ORJUELA

NDOMASSI  
TANDO

CHRISTINE  
TRANCHANT

IE bioinfo

IE bioinfo

IE systèmes  
d'information

IE bioinfo

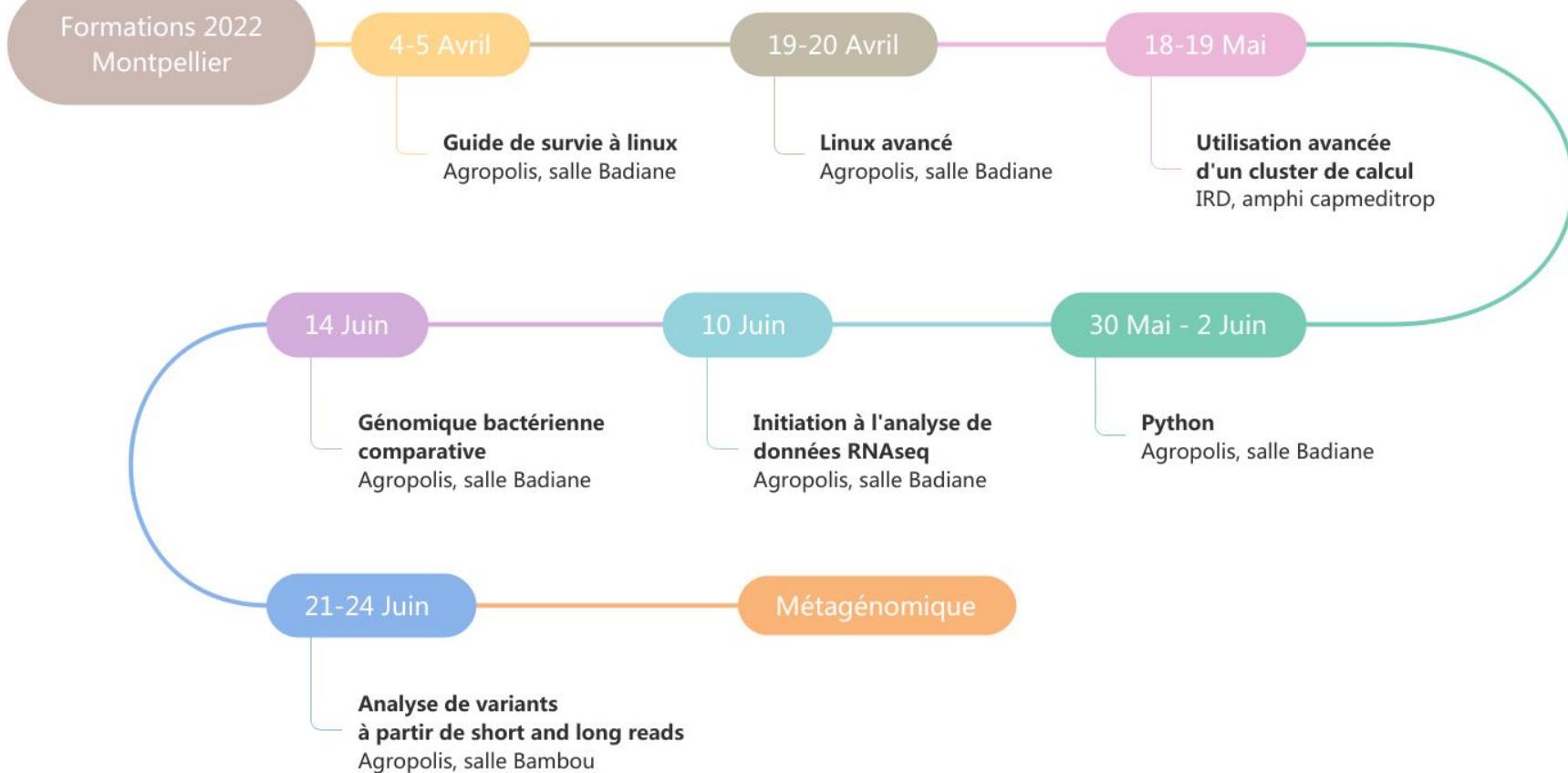
IE systèmes

IR bioinfo

[bioinfo@ird.fr](mailto:bioinfo@ird.fr)



@ItropBioinfo





# Modules de formation 2022

- Toutes nos formations :  
<https://southgreenplatform.github.io/trainings/>
- Topo & TP : <https://southgreenplatform.github.io/trainings/sv>  
[https://github.com/SouthGreenPlatform/training\\_SV\\_teaching/tree/2022](https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022)
- tablet



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

[www.southgreen.fr](http://www.southgreen.fr)

<https://southgreenplatform.github.io/trainings>



# 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, minimap2*
- Déetecter des SNPs à partir du mapping de reads - *GATK, deepvariants*
- Analyser les données SNPs brutes (ex: stats, filtres) - *vcftools, bcftools*
- Exemples d'études possibles à partir de SNPs - *SNIPPlay*
- DéTECTer des variants structuraux (SV) à partir de :
  - reads mappées contre un génome - *breakdancer, sniffle*
  - génomes entiers - *nucmer, assemblytics, siry*

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



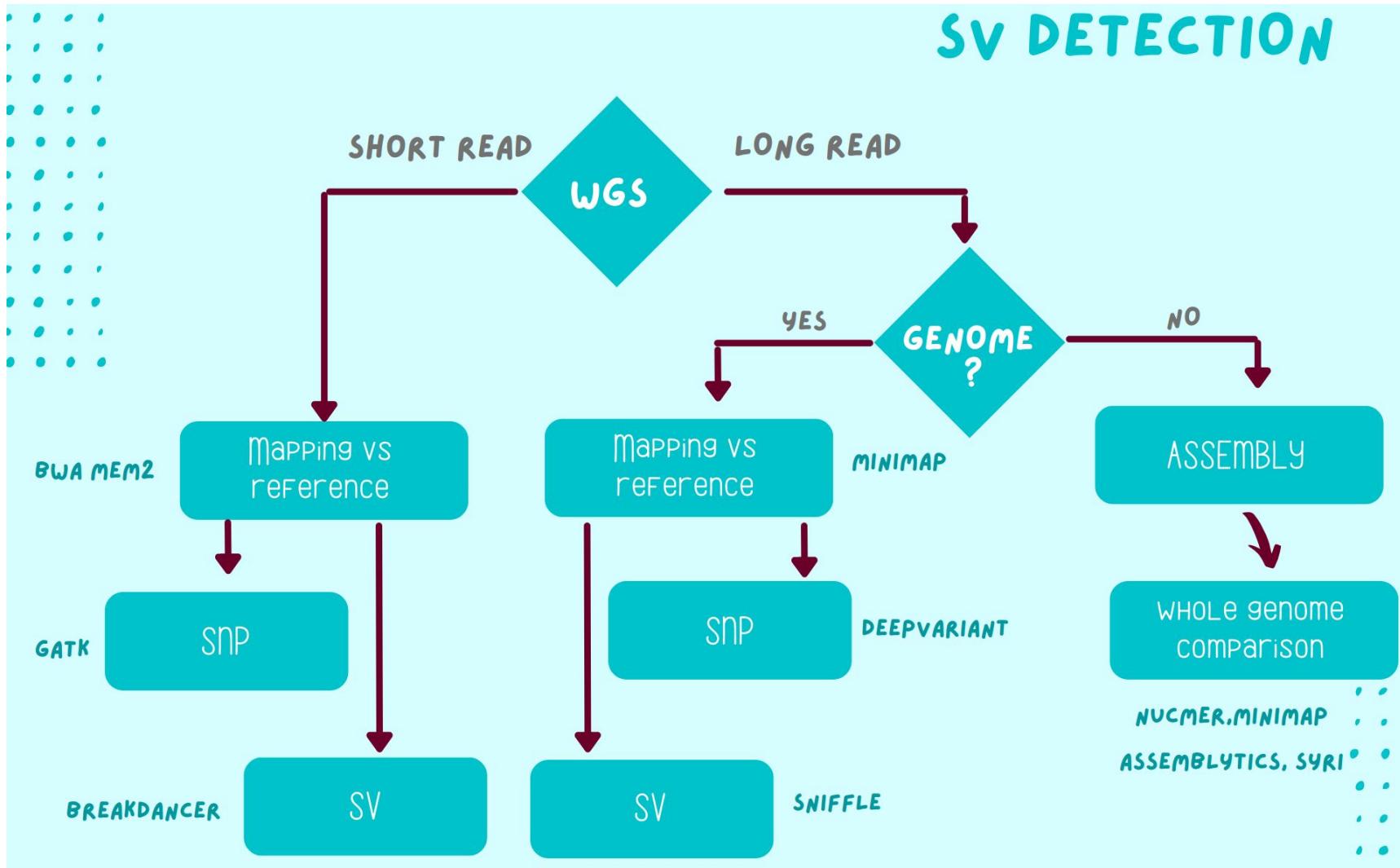
## Applications :

- Mapper des reads contre un génome *bwa, minimap2*
- Déetecter des SNPs à partir du mapping de reads - *GATK, deepvariants*
- Analyser les données SNPs brutes (ex: stats, filtres) - *vcftools, bcftools*
- Exemples d'études possibles à partir de SNPs - *SNIPPlay*
- DéTECTer des variants structuraux (SV)



**Avec jupyter book** : lancer les commandes + analyser les résultats  
=> Avoir un plan de bataille opérationnel

# Plan de bataille !!!





# Let's discover Jupyter through the IFB cloud

*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

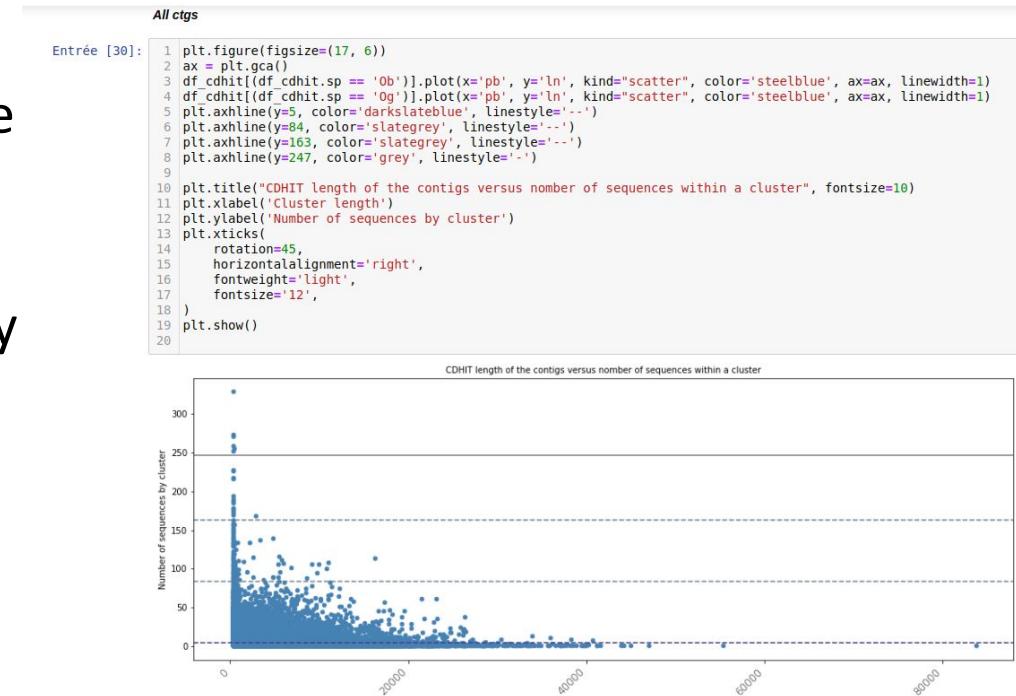
ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala



# Why use jupyter book ?

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

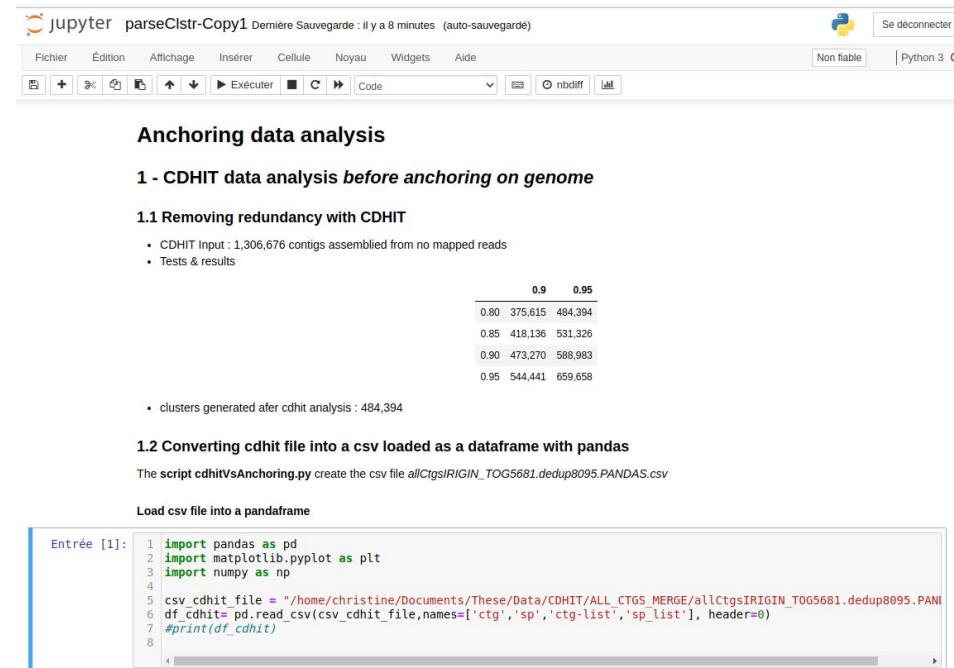
- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



# 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 parseCistr-Copy1 Dernière Sauvegarde : il y a 8 minutes (auto-sauvegarde) | Se déconnecter | Python 3 O
- Toolbar:** Fichier, Édition, Affichage, Insérer, Cellule, Noyau, Widgets, Aide.
- Cell Content:**
  - Section Header:** Anchoring data analysis
  - Section 1:** 1 - CDHIT data analysis before anchoring on genome
  - Section 1.1:** 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:** 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
  - Section 1.3:** Load csv file into a pandasframe
  - Code Cell [1]:**

```
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)
              8
              9
```

# Lab notebook for science data ?

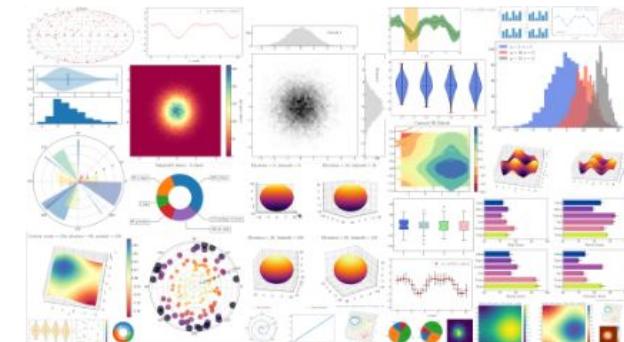
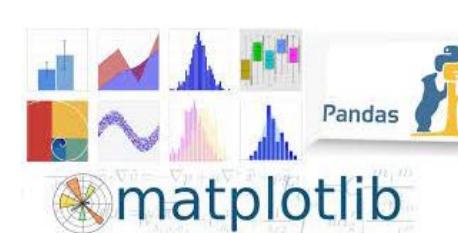
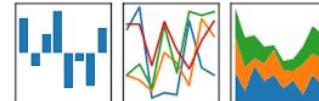


- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

# 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

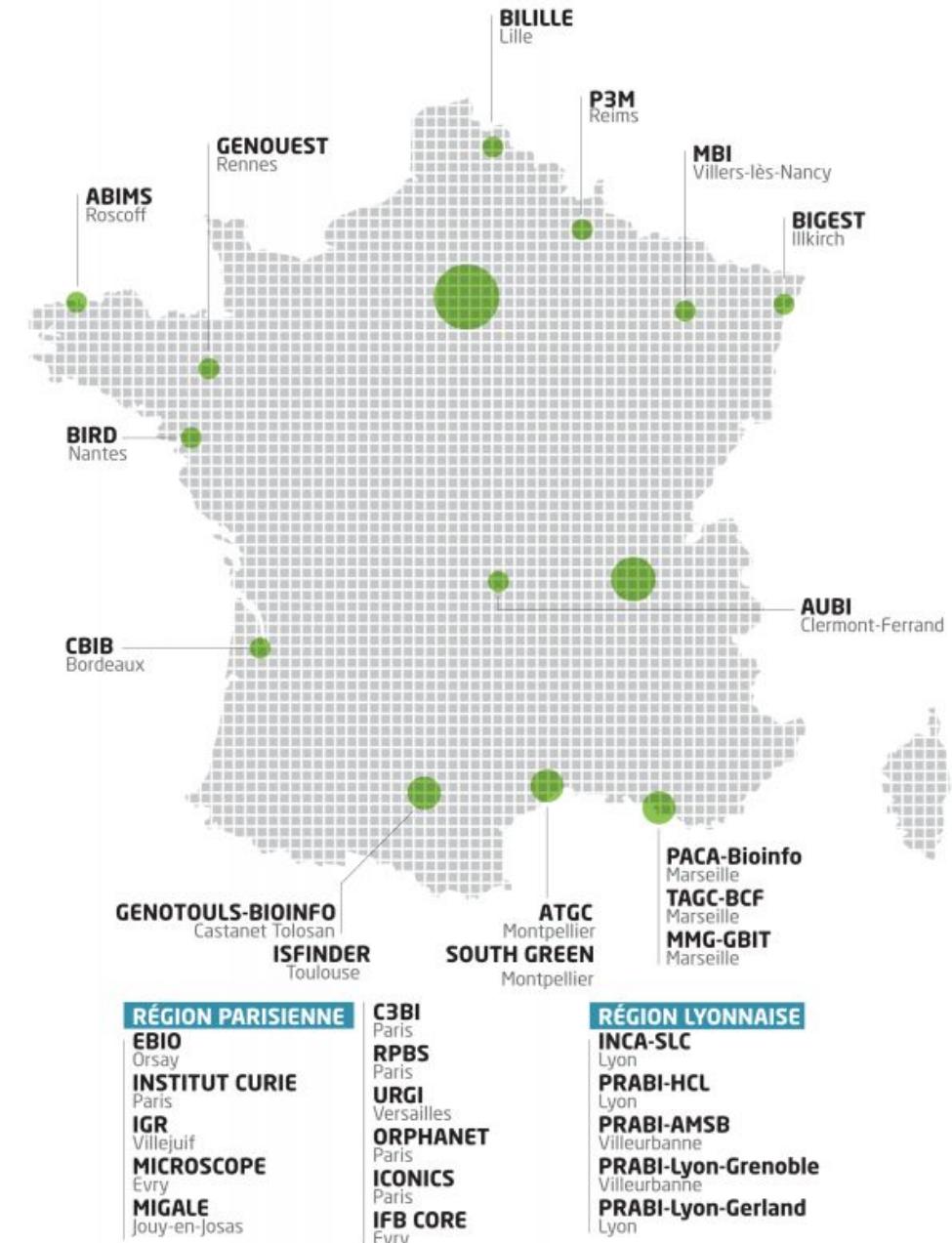


The screenshot shows the IFB Cloud web interface. The top navigation bar includes tabs for "IFB Cloud", "mydatalocal", and a "+" button. The address bar shows the URL <https://134.158.247.8/tree/mydatalocal>. The main content area is titled "jupyter" and contains three tabs: "Files", "Running", and "Clusters". The "Files" tab is selected, showing a file tree with a single folder named "mydatalocal". A message below the tree states "La liste des notebooks est vide." (The list of notebooks is empty). To the right, there is a "New" button with a dropdown menu open. The dropdown menu lists several options under "Notebook": "Bash", "Julia 1.5.3", "Python 3", and "R". Below these, under "Other:", are "Text File", "Folder", and "Terminal".



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

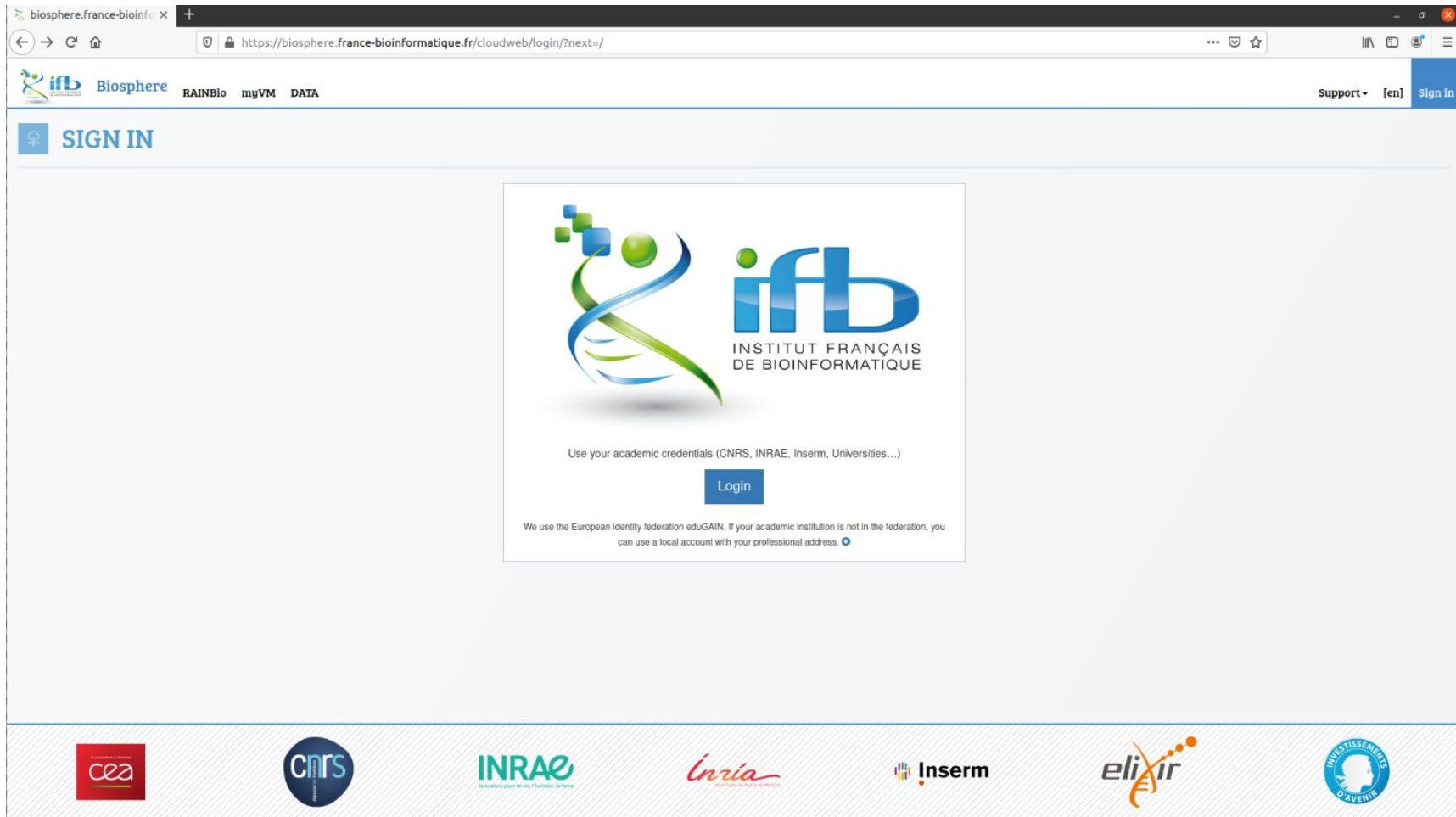
22 plateformes-membres  
 7 plateformes contributrices  
 8 équipes associées  
 >400 experts (~200 FTE)



- 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.

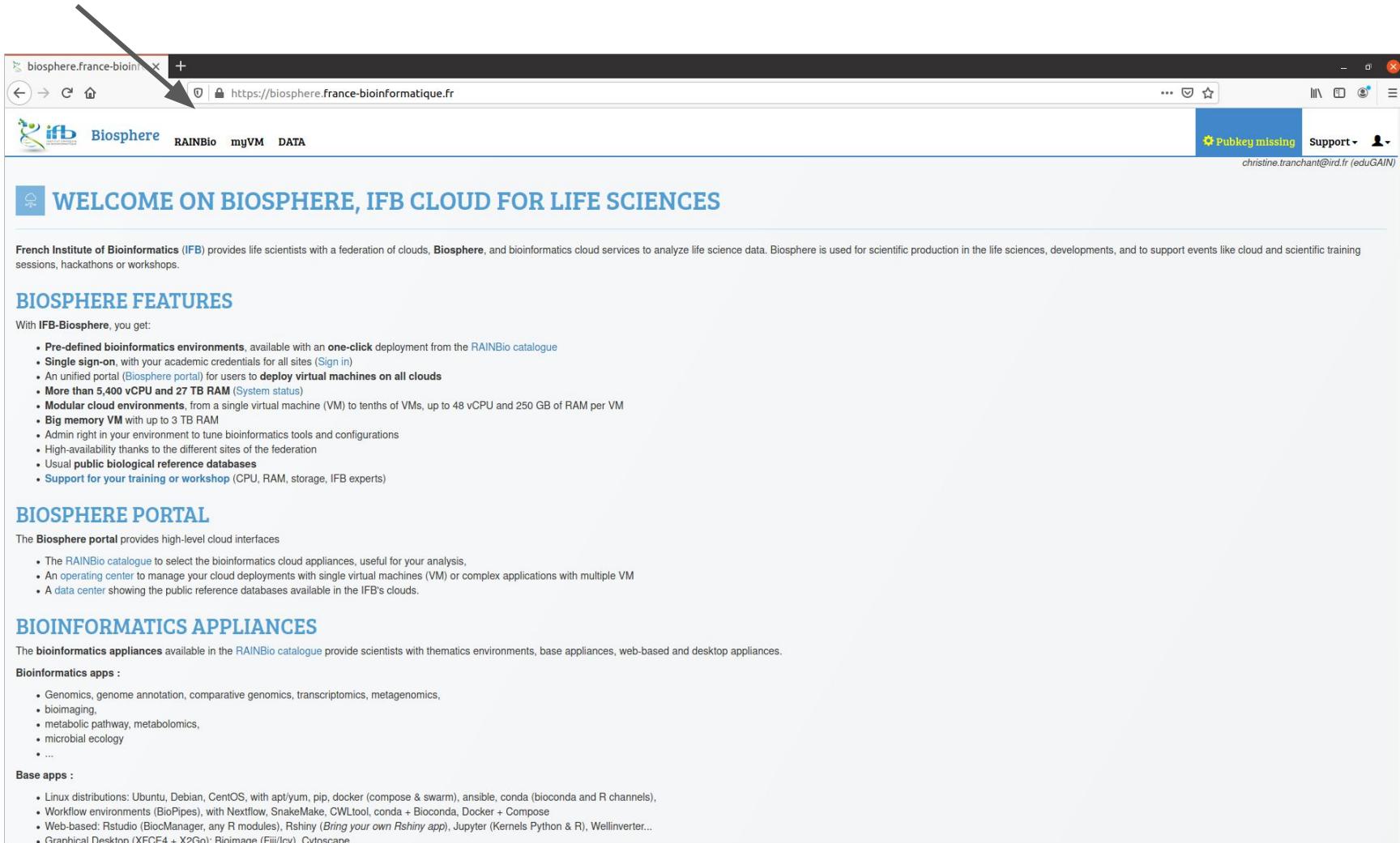
# Let's start with biosphere

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



The screenshot shows a web browser window for the Biosphere platform. The address bar displays the URL <https://biosphere.france-bioinformatique.fr/cloudweb/login/?next=/>. The page header includes the ifb logo, a "Biosphere" link, and navigation tabs for RAINBio, myVM, and DATA. On the right, there are "Support", language ("en"), and "Sign in" buttons. A large "SIGN IN" button is visible on the left. The main content area features the ifb logo and the text "INSTITUT FRANÇAIS DE BIOINFORMATIQUE". Below the logo, it says "Use your academic credentials (CNRS, INRAE, Inserm, Universities...)" and has a "Login" button. A note at the bottom states: "We use the European identity federation eduGAIN. If your academic institution is not in the federation, you can use a local account with your professional address." Logos for CEA, CNRS, INRAE, Inria, Inserm, elixir, and Investissements d'Avenir are displayed along the bottom.

## RAINBIO catalog to access our Virtual Machine (VM)



The screenshot shows a web browser window with the following details:

- Title Bar:** biosphere.france-bioinformatique.fr
- Address Bar:** https://biosphere.france-bioinformatique.fr
- Header:** Biosphere (with IFB logo), RAINBio, myVM, DATA
- Right Side:** Pubkey missing, Support, christine.tranchant@ird.fr (eduGAIN)

**Content Area:**

### 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 Desktop (XFCE4 + X2Go), Bioimage (Fiji/ICV), Cytoscape

# 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 (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)).

The main content area is titled "RAINBIO - APPLIANCES BIOINFORMATIQUES DANS LE CLOUD". Below the title, it says "Catalogue des appliances bioinformatiques dans le cloud, filtrez-les en utilisant les termes présents dans l'ontologie EDAM, ou en langage naturel." A search bar on the right contains the text "analyses".

The interface displays a grid of 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/datascience-notebook` but users can choose any other existing stack with an Advanced deployment in Biosphere portal.  
In addition, we integrated various tools to perform the SV detection

### Tools

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

### Contact

- Support Cloud IFB

### Developpers

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

### App data

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

### Licence

Licensed under GPLv3

Site web	<a href="https://hub.docker.com/r/francoissabot/trainingontvm">https://hub.docker.com/r/francoissabot/trainingontvm</a>
----------	---

 Clé publique (PubKey) absente

Support ▾

christine.tranchant@ird.fr (eduGAIN)

LANCER ▾

▶ LANCER

▶ 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](https://github.com/SouthGreenPlatform/training_SV_VM)

App de base Jupyter

### Caractéristiques

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

### Crédits

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

# Let's run your vm through the cloud

IFB Biosphère RAINBio myVM DATA

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

LANCEUR 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, Adaptation)

Cloud: ifb-core-cloudbis

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

vCPU.h

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

# 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

ready !

IFB Biosphere RAINBio myVM DATA

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

## CLOUD

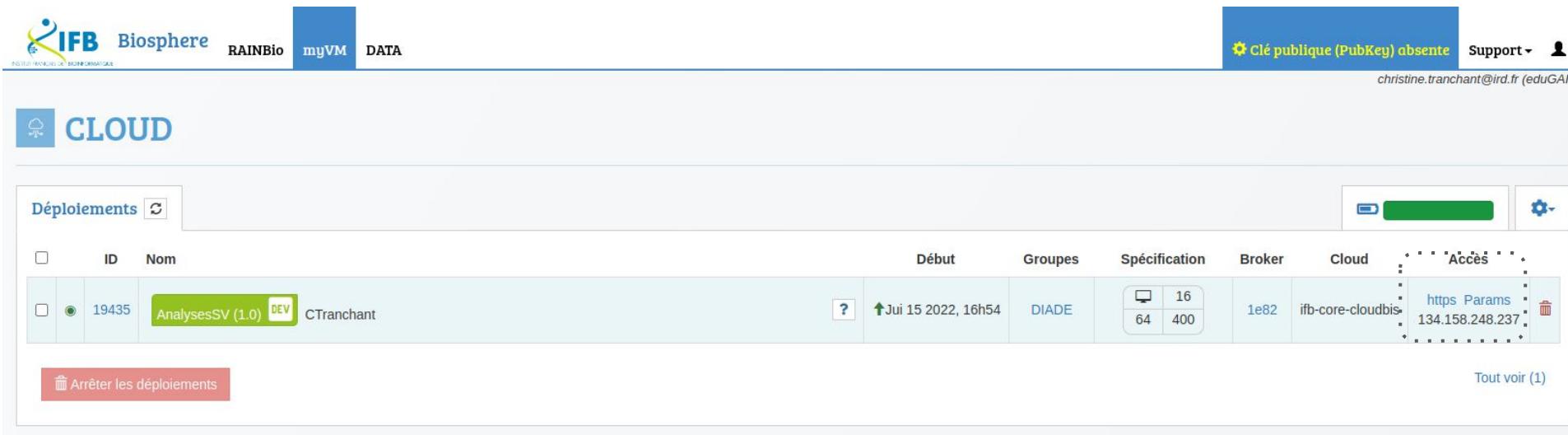
Déploiements

ID	Nom	Début	Groupes	Spécification	Broker	Cloud	Accès				
19435	AnalysesSV (1.0) DEV CTranchant	Jui 15 2022, 16h54	DIADE	<table border="1"><tr><td>64</td><td>16</td></tr><tr><td>400</td><td></td></tr></table>	64	16	400		1e82	ifb-core-cloudbis	<a href="https://134.158.248.237">https Params</a>
64	16										
400											

Arrêter les déploiements Tout voir (1)

# Let's run your vm through the cloud

get the url... link “https”



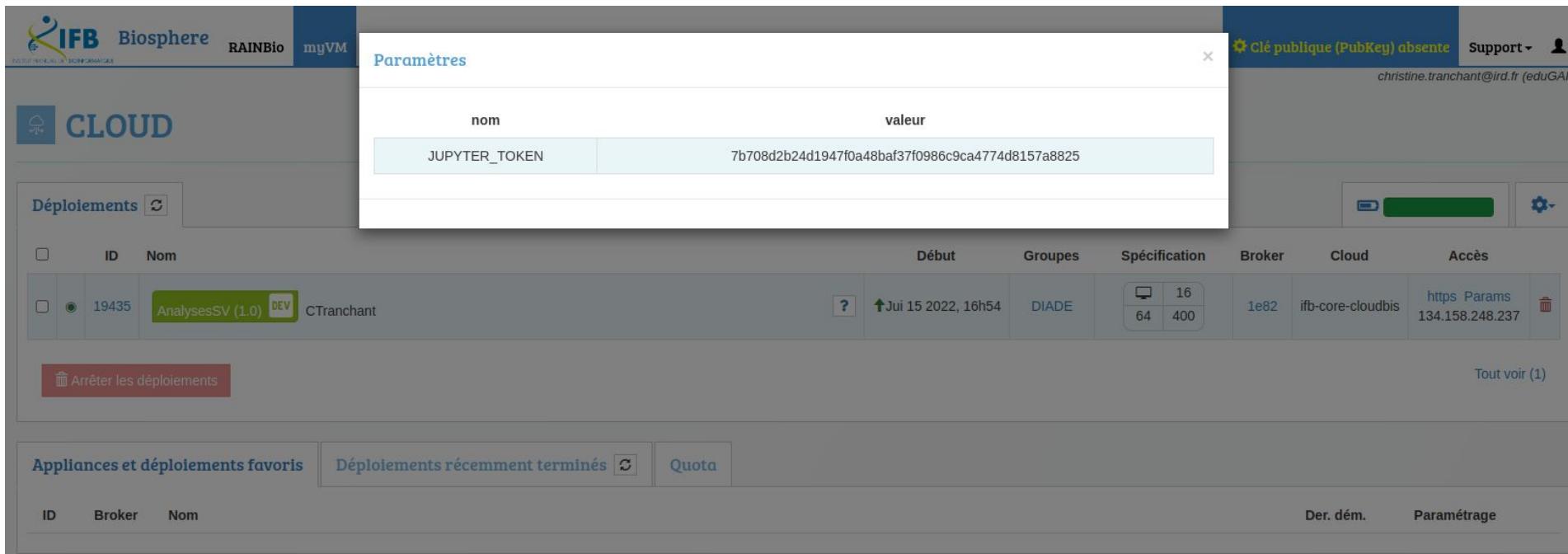
The screenshot shows the SouthGreen cloud interface. At the top, there is a navigation bar with the IFB Biosphere logo, RAINBio, myVM, DATA, and a support section indicating a public key is absent. The main area is titled "CLOUD". Under "Déploiements", a table lists one deployment:

ID	Nom	Début	Groupes	Spécification	Broker	Cloud	Accès			
19435	AnalysesSV (1.0) DEV CTranchant	Jui 15 2022, 16h54	DIADE	<table border="1"><tr><td>16</td></tr><tr><td>64</td></tr><tr><td>400</td></tr></table>	16	64	400	1e82	ifb-core-cloudbis	<a href="https://134.158.248.237">https Params</a>
16										
64										
400										

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

# Let's run our vm through the cloud

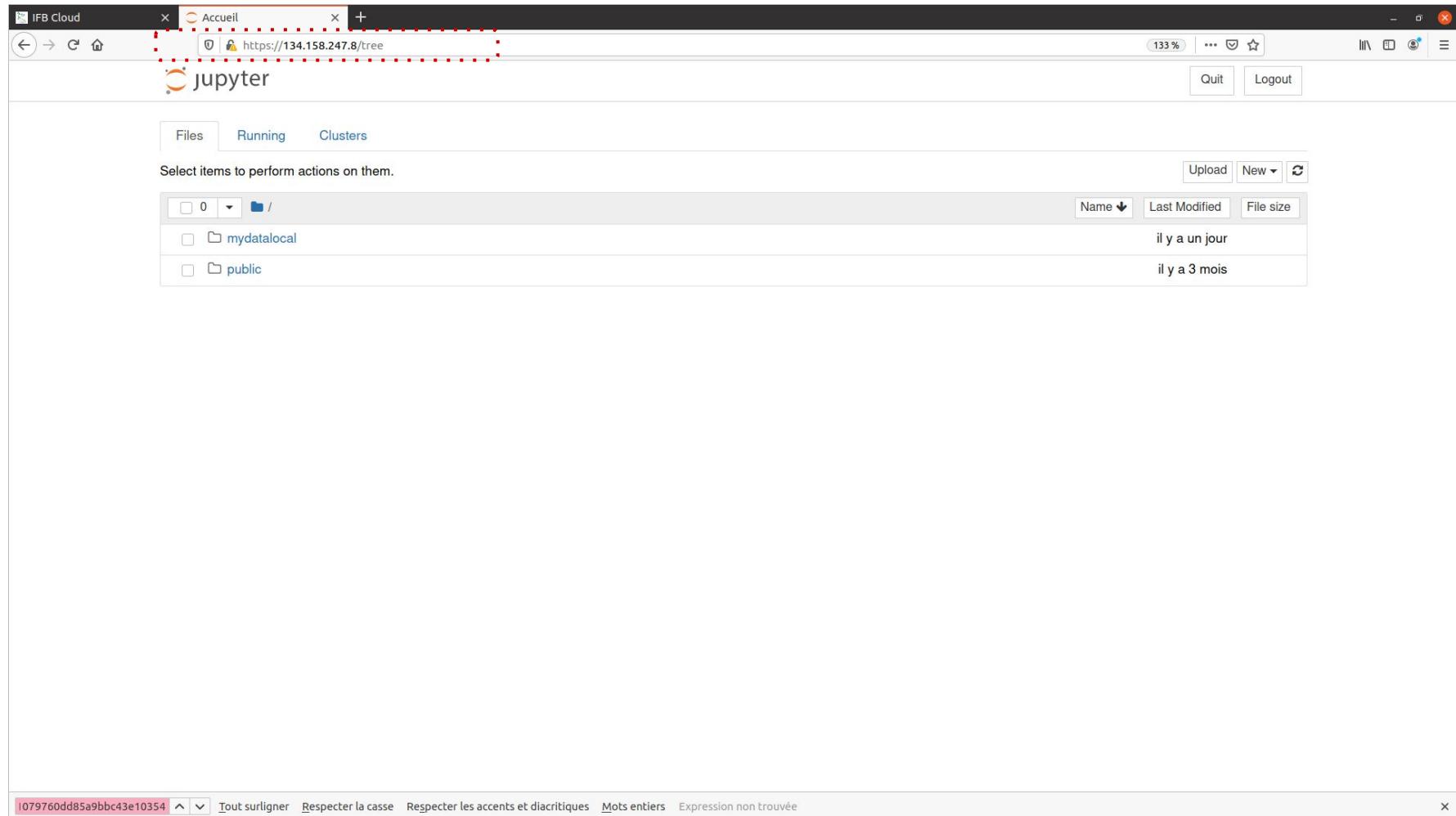
Get the token identifiant... link “Params”



The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there are navigation tabs: IFB Biosphère, RAINBio, myVM, CLOUD, and Déploiements. The CLOUD tab is active, showing a list of deployments. One deployment is selected: AnalysesSV (1.0) DEV by CTranchant, which was started on Juillet 15 2022, 16h54, with a DIADE specification, 1e82 broker, ifb-core-cloudbis cloud, and https://Params 134.158.248.237 access. Below the deployment list is a red button labeled "Arrêter les déploiements". A modal dialog titled "Paramètres" is open, showing a single parameter: JUPYTER\_TOKEN with the value 7b708d2b24d1947f0a48baf37f0986c9ca4774d8157a8825. The bottom of the screen shows sections for Appliances et déploiements favoris, Déploiements récemment terminés, and Quota.

# 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 the IFB Cloud web interface. At the top, there is a navigation bar with tabs for "IFB Cloud", "Accueil", and a "+" button. The URL in the address bar is <https://134.158.247.8/tree>. On the right side of the header are "Logout" and other user icons. Below the header, there is a search bar and a "jupyter" logo.

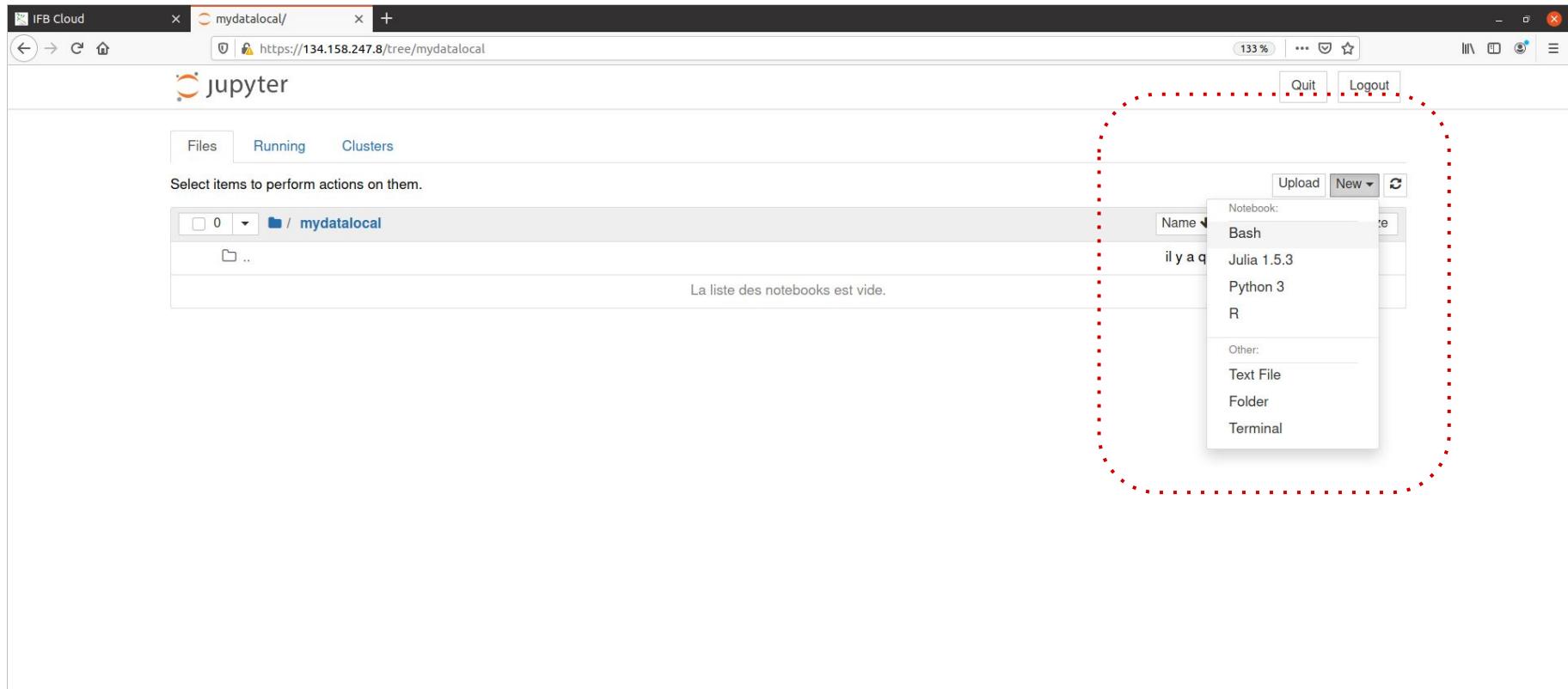
The main content area has three tabs: "Files" (selected), "Running", and "Clusters". A message says "Select items to perform actions on them." Below this is a file browser table:

	Name	Last Modified	File size
<input type="checkbox"/>	0		
<input type="checkbox"/>	mydatalocal	il y a un jour	
<input type="checkbox"/>	public	il y a 3 mois	

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

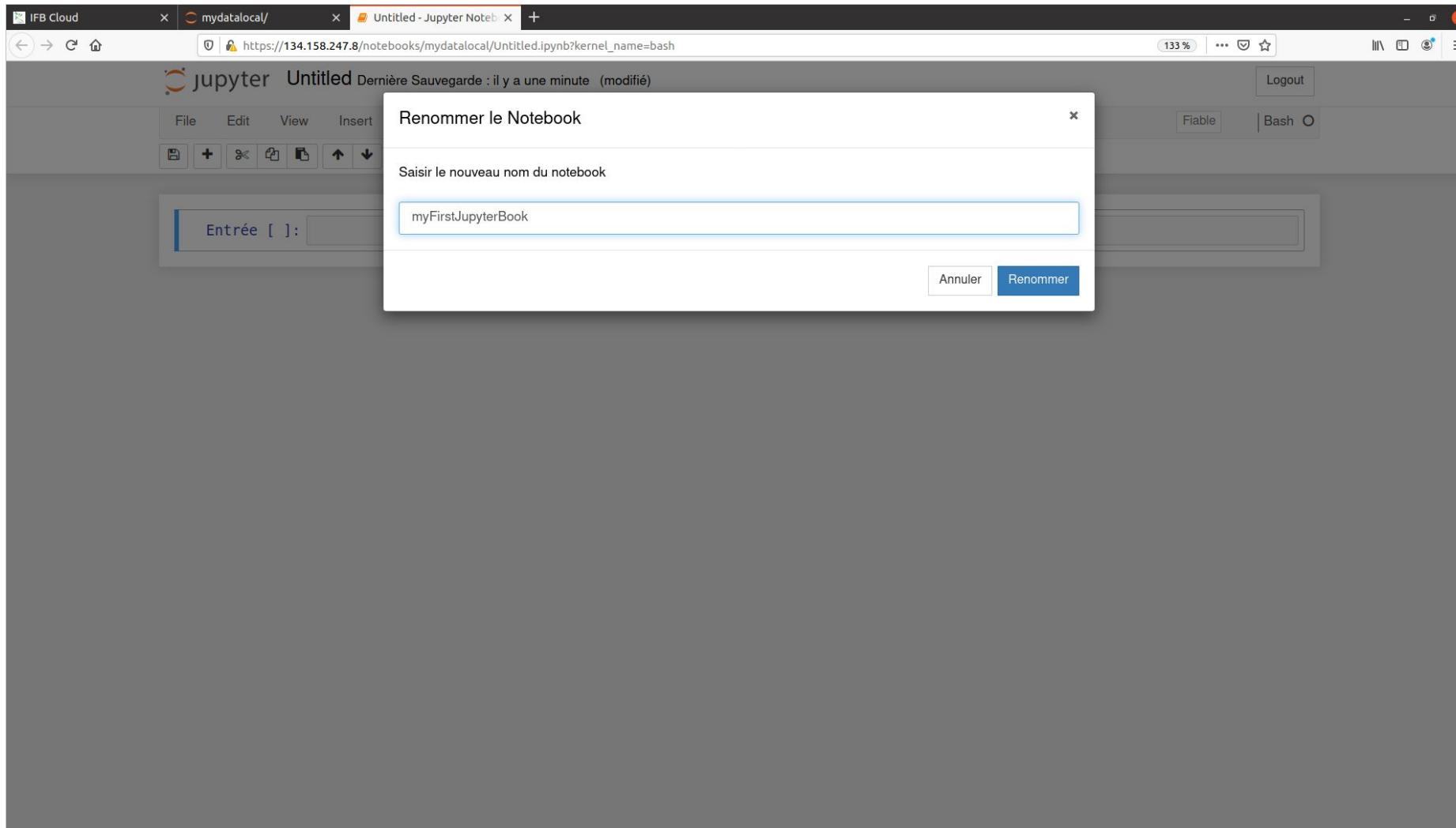
# Create your first jupyter book

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



# Rename your first jupyter book

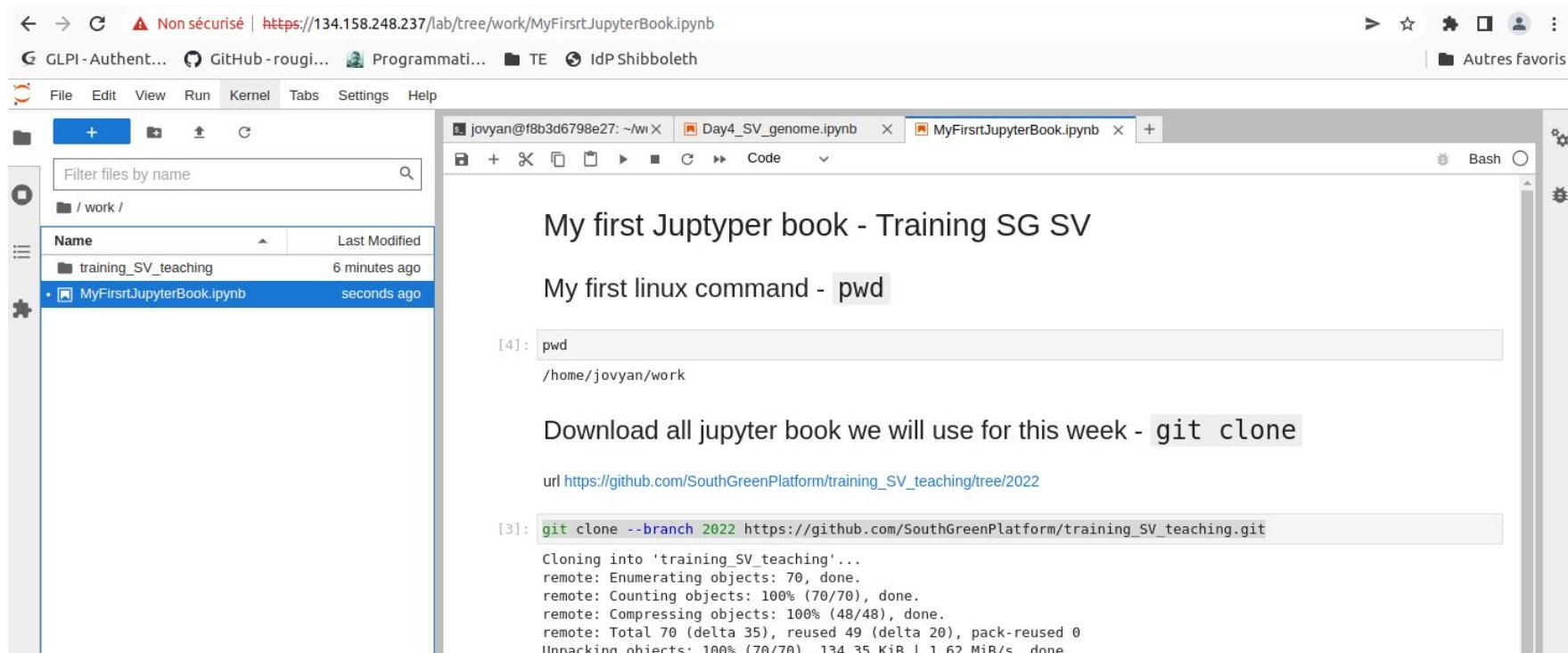
myFirstJupyterBook



# Run your first bask command - *git clone*

- All jupyterbook used for practice are here :  
[https://github.com/SouthGreenPlatform/training\\_SV\\_teaching/tree/2022](https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022)
- Download all the jupyter books with the command *git clone*

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



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

- File Browser:** On the left, there is a sidebar with a file browser. It shows a directory structure under "/work/". The "Name" column lists "training\_SV\_teaching" (modified 6 minutes ago) and "MyFirstJupyterBook.ipynb" (modified seconds ago).
- Code Cell Output:** In the main area, there are two code cells. The first cell contains the text:

```
My first Juptyer book - Training SG SV
```

```
My first linux command - pwd
```

```
[4]: pwd
```

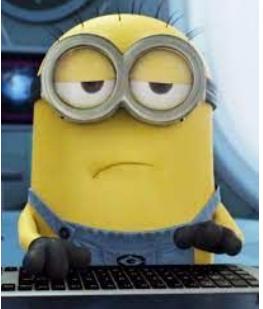
```
/home/jovyan/work
```
- Terminal Output:** The second code cell shows the output of the `git clone` command:

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

```
url https://github.com/SouthGreenPlatform/training\_SV\_teaching/tree/2022
```

```
[3]: git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git
```

```
Cloning into 'training_SV_teaching'...
remote: Enumerating objects: 70, done.
remote: Counting objects: 100% (70/70), done.
remote: Compressing objects: 100% (48/48), done.
remote: Total 70 (delta 35), reused 49 (delta 20), pack-reused 0
Unpacking objects: 100% (70/70), 134.35 KiB | 1.62 MiB/s, done.
```

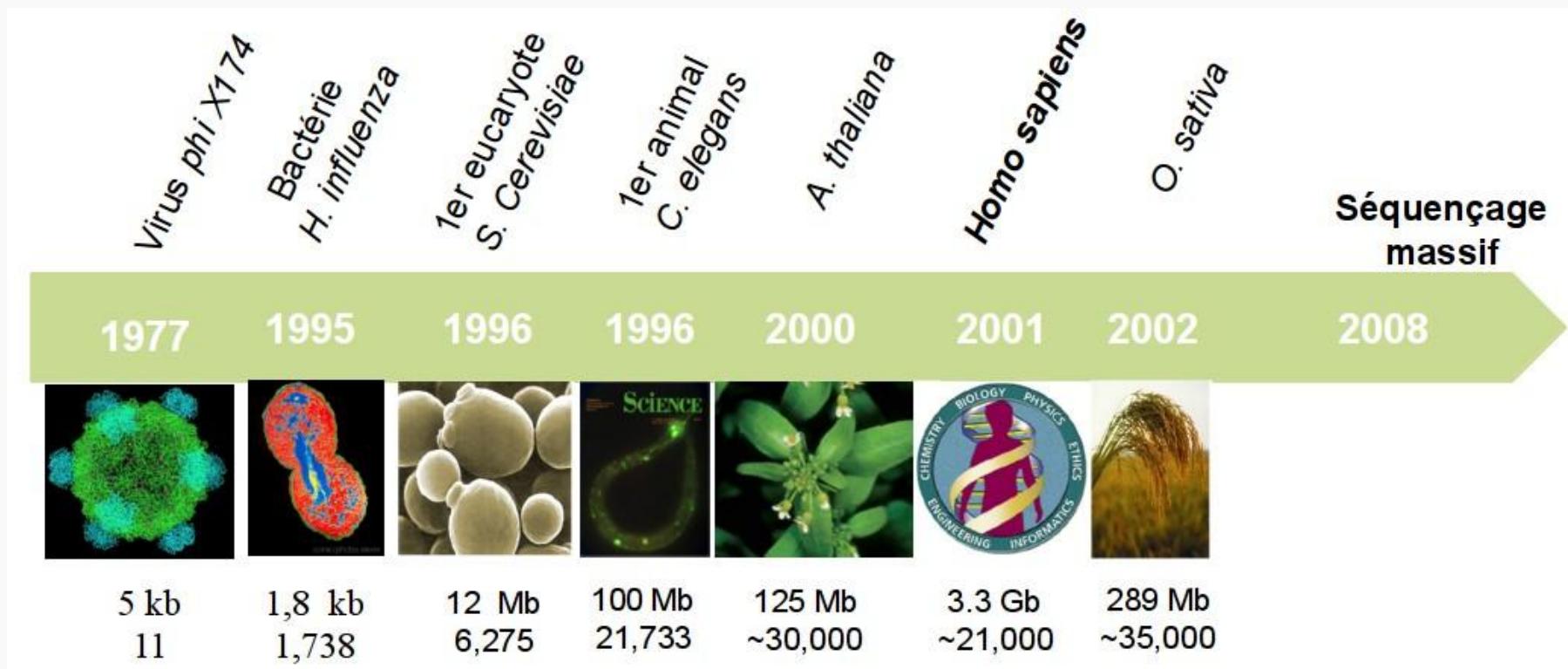
A small image of a yellow Minion character wearing blue overalls and large round glasses, sitting at a computer keyboard.

# Introduction & NGS method

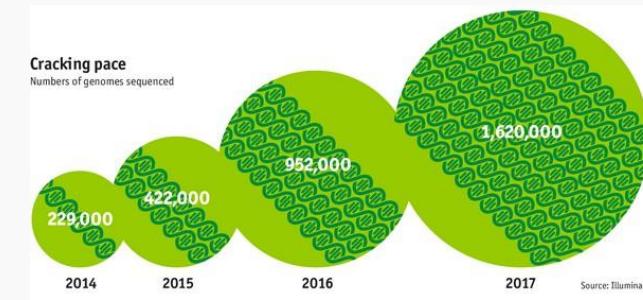
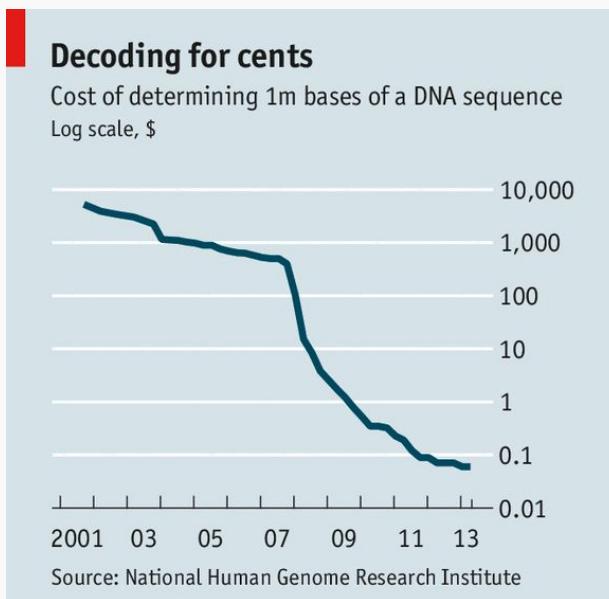
## **The NGS in themselves**

---

# A little history of sequencing...

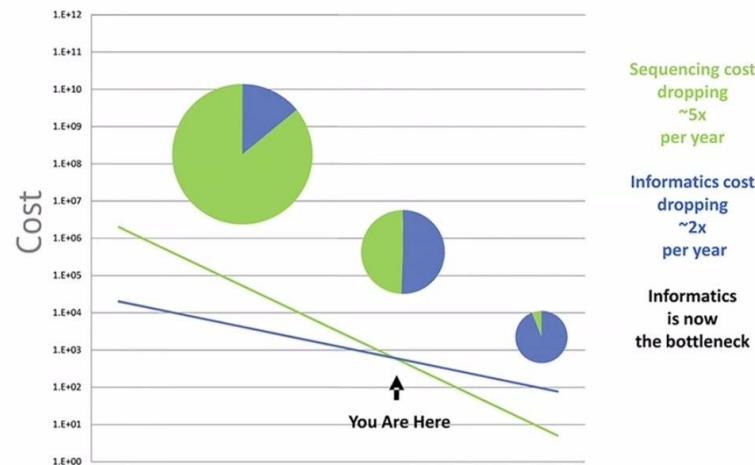


# ...From Data Rarity to Data Deluge



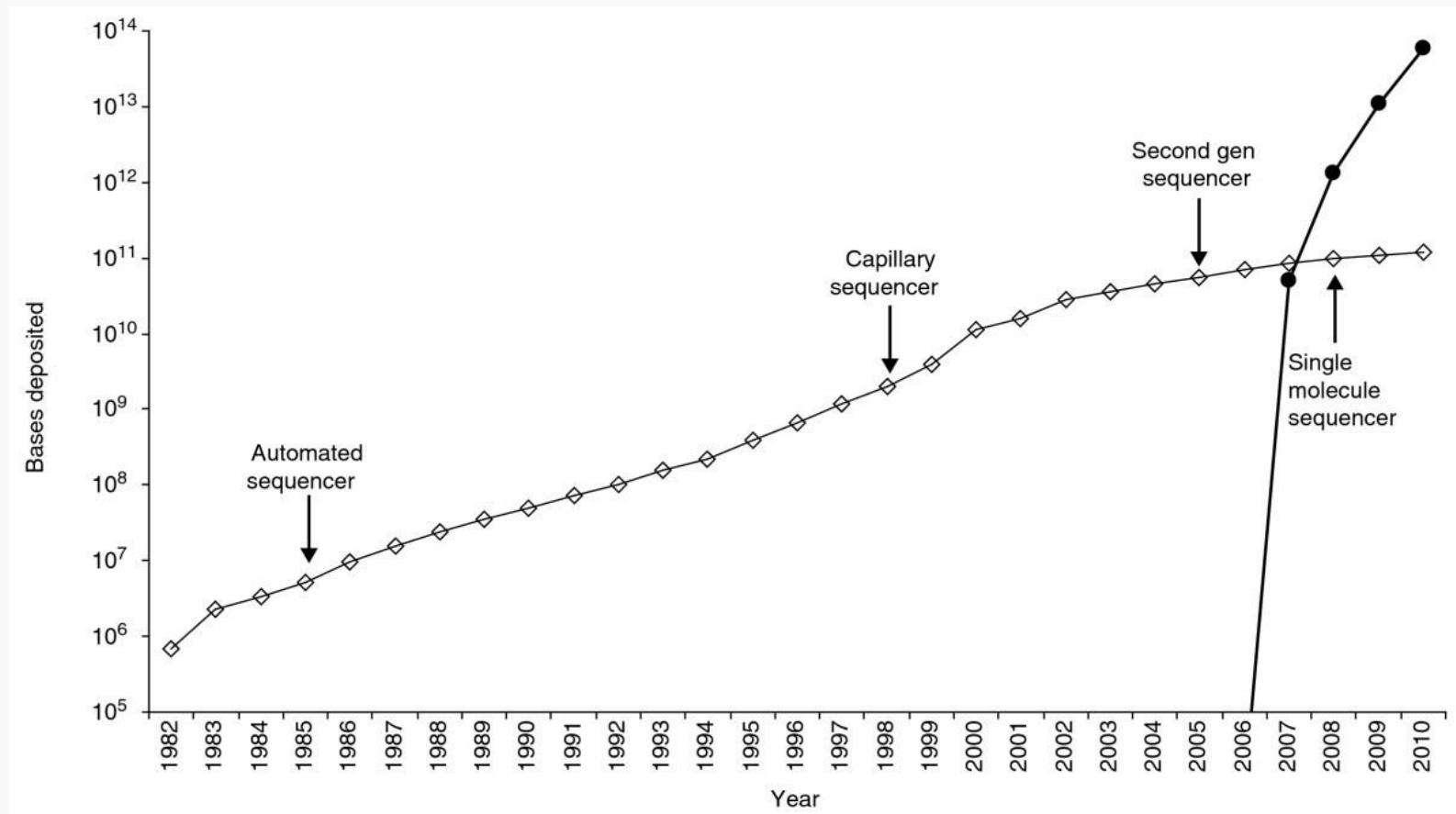
From The economist

## DNA Sequencing Economics



From Business Insider

# ...From Data Rarity to Data Deluge



# What can we do with it ?

- Genetic diversity
- Gene discovery
- Genomic structure
- Contamination/pathogen detection
- Metagenomic
- Pangenomic
- And many other things...

# Methods

---

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short)
- Matrix amplification
- Short reads
- Limited error rate
- High throughput

## 3<sup>rd</sup> Generation Sequencing

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short) 454
- Matrix amplification IonTorrent
- Short reads Illumina
- Limited error rate
- High throughput

## 3<sup>rd</sup> Generation Sequencing

- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short) 454
- Matrix amplification IonTorrent
- Short reads Illumina
- Limited error rate
- High throughput

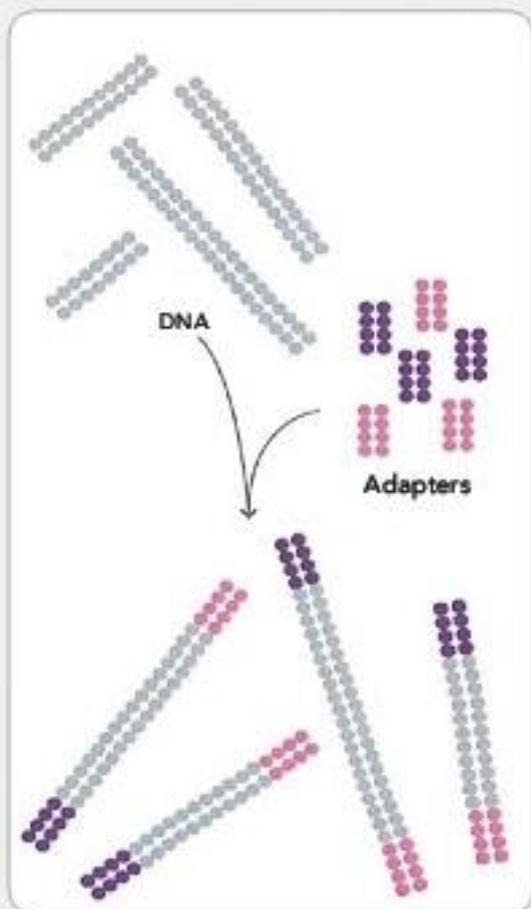
PacificBiosciences  
**Oxford**  
**Nanopore**

## 3<sup>rd</sup> Generation Sequencing

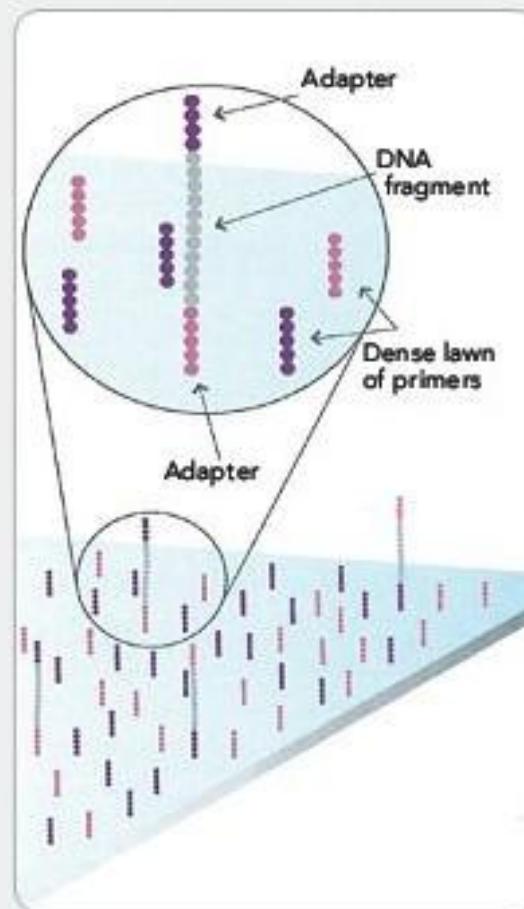
- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput



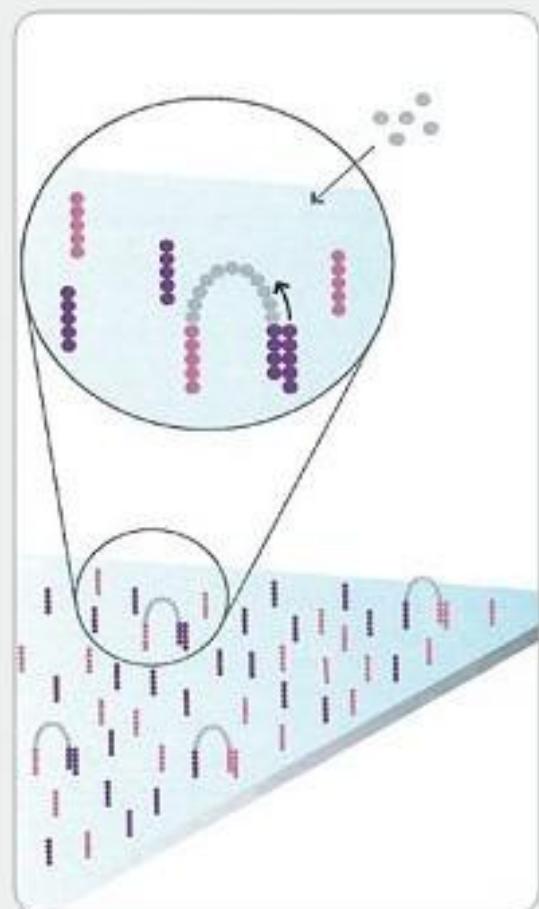


**1. PREPARE GENOMIC DNA SAMPLE**

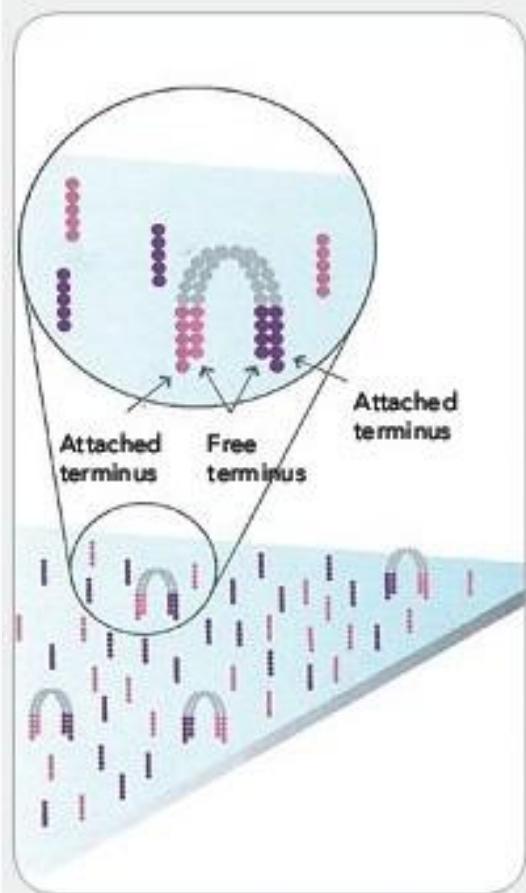
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

**2. ATTACH DNA TO SURFACE**

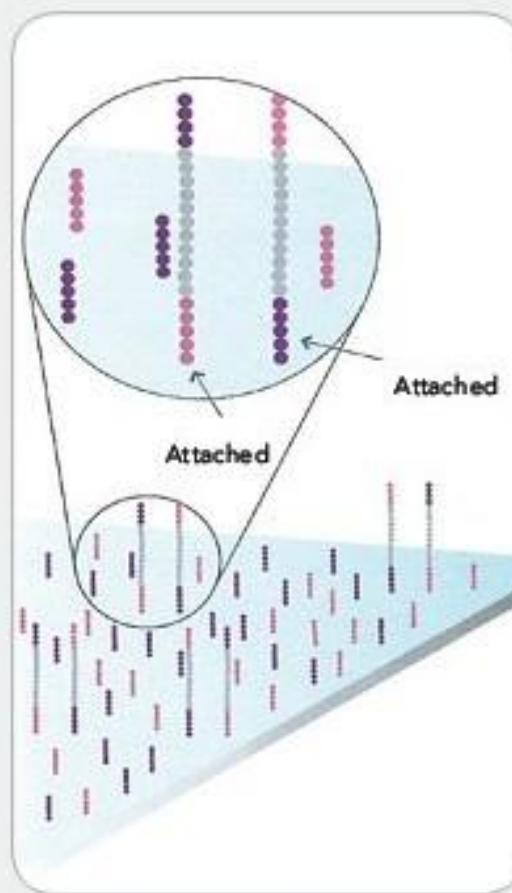
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

**3. BRIDGE AMPLIFICATION**

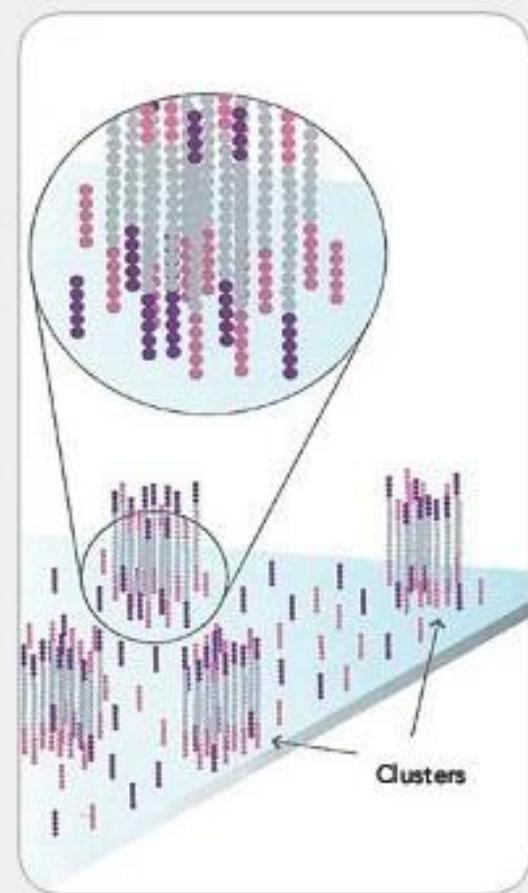
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

**4. FRAGMENTS BECOME DOUBLE STRANDED**

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

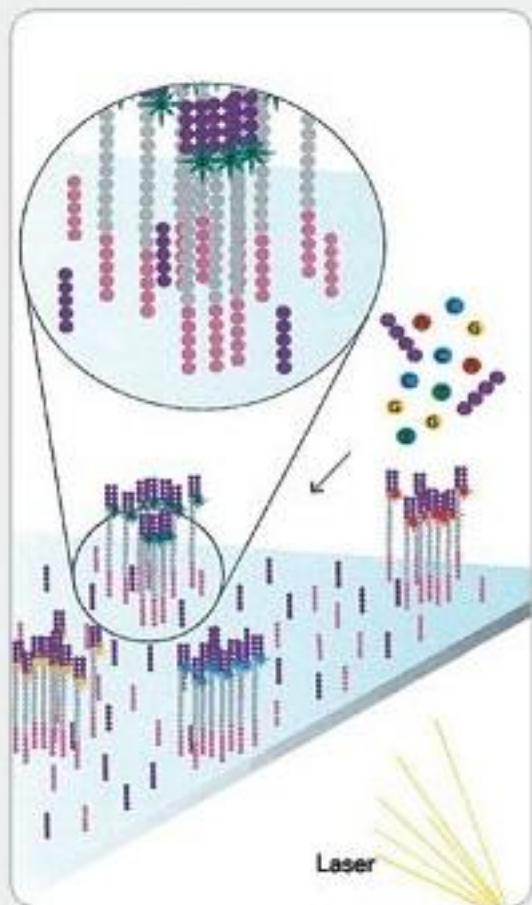
**5. DENATURE THE DOUBLE-STRANDED MOLECULES**

Denaturation leaves single-stranded templates anchored to the substrate.

**6. COMPLETE AMPLIFICATION**

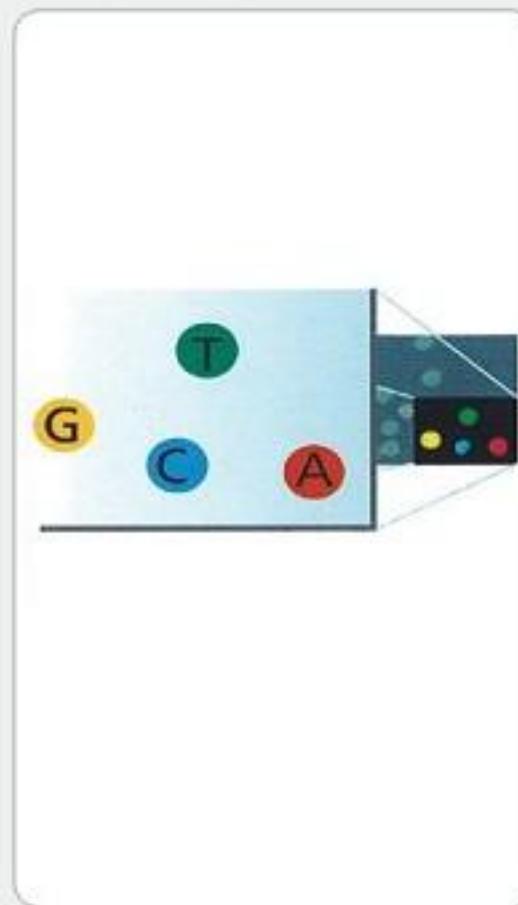
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

## 7. DETERMINE FIRST BASE



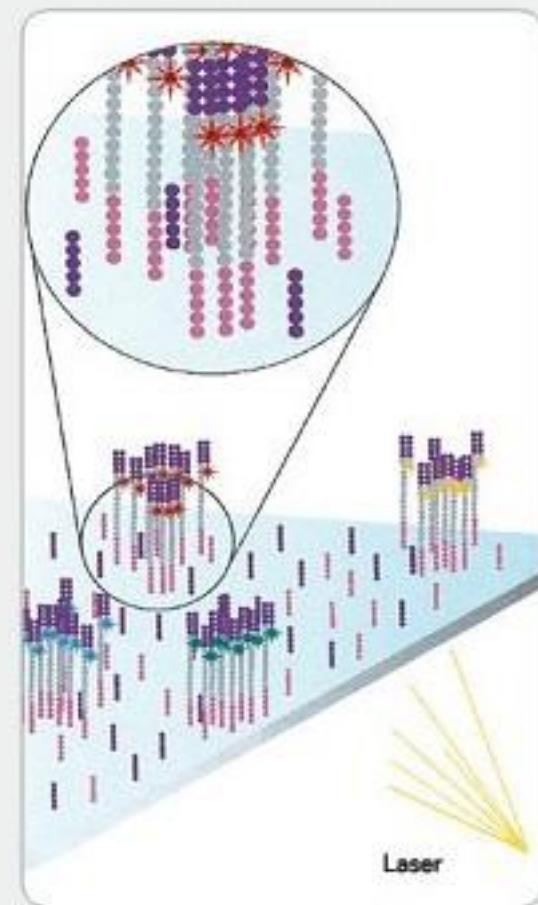
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

## 8. IMAGE FIRST BASE



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

## 9. DETERMINE SECOND BASE



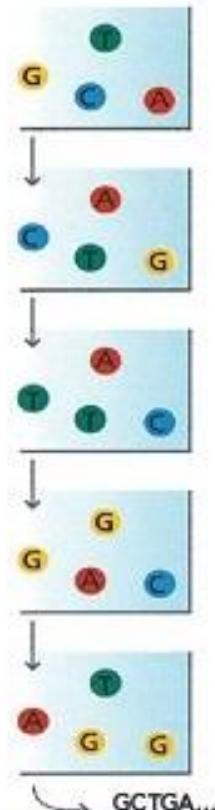
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

## 10. IMAGE SECOND CHEMISTRY CYCLE



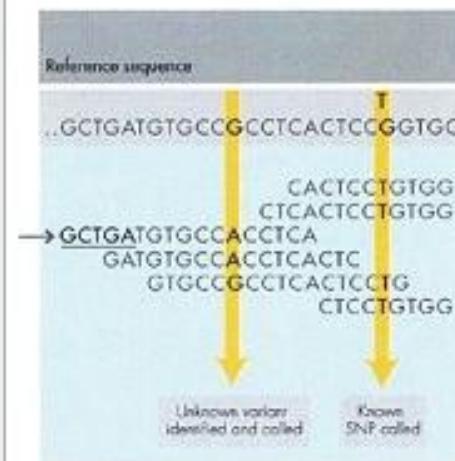
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

## 11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



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

## 12. ALIGN DATA



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

## Advantages :

- + Output volume (20 billions of 150b reads/6Tb, NovaSeq6000)
- + Accuracy (99.99 % - but questionable)
- + Run is cheap
- + MySeq is cheap (around 60 000 USD per machine)

Limits : Size (150 + 150 in NovaSeq, but 400 for MySeq)

# The FASTQ Format

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTCTTAGTATTTTAGTCATTCCGTGTTGGTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIIEFEGJGGHI4FFIEIGHI<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
CATGTGCTATTACCAAAAGTGCAGTAACGACCTATAAATTTAAAGTAGC
+
@CFFFFFFGGHHHHIJJJIEE<HHHIJIGBHGGEEIIJEIEIJIHHJFIIJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTTCGGGTATGGCTCTAGTAGCTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIEEHII<CGHIJIIJ?:FC9DGAFGHII?DGBFIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFFGHHHHJIIIIJJIIIGJCHHEIII>GHIG@GHIDHGJIIFHIIJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGAAGAGTTAAAAAACATTAAATGAGCAACTGAGTTC
+
@@CFFFFFHFFHFGIJJIIHHIIJJIIJJECGHIIJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTGTCGGGGTTGGGAAATTGTCACTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```

1 séquence = 4 lignes



- @identifiant de la séquence
- Séquence
- + (id séquence).
- Qualité de la séquence = un caractère ASCII pour chaque base

## PHRED SCORE

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

*Ewing 1998*

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

or

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

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99,99%
50	1 in 100000	99.999 %

# The QPHRED Value

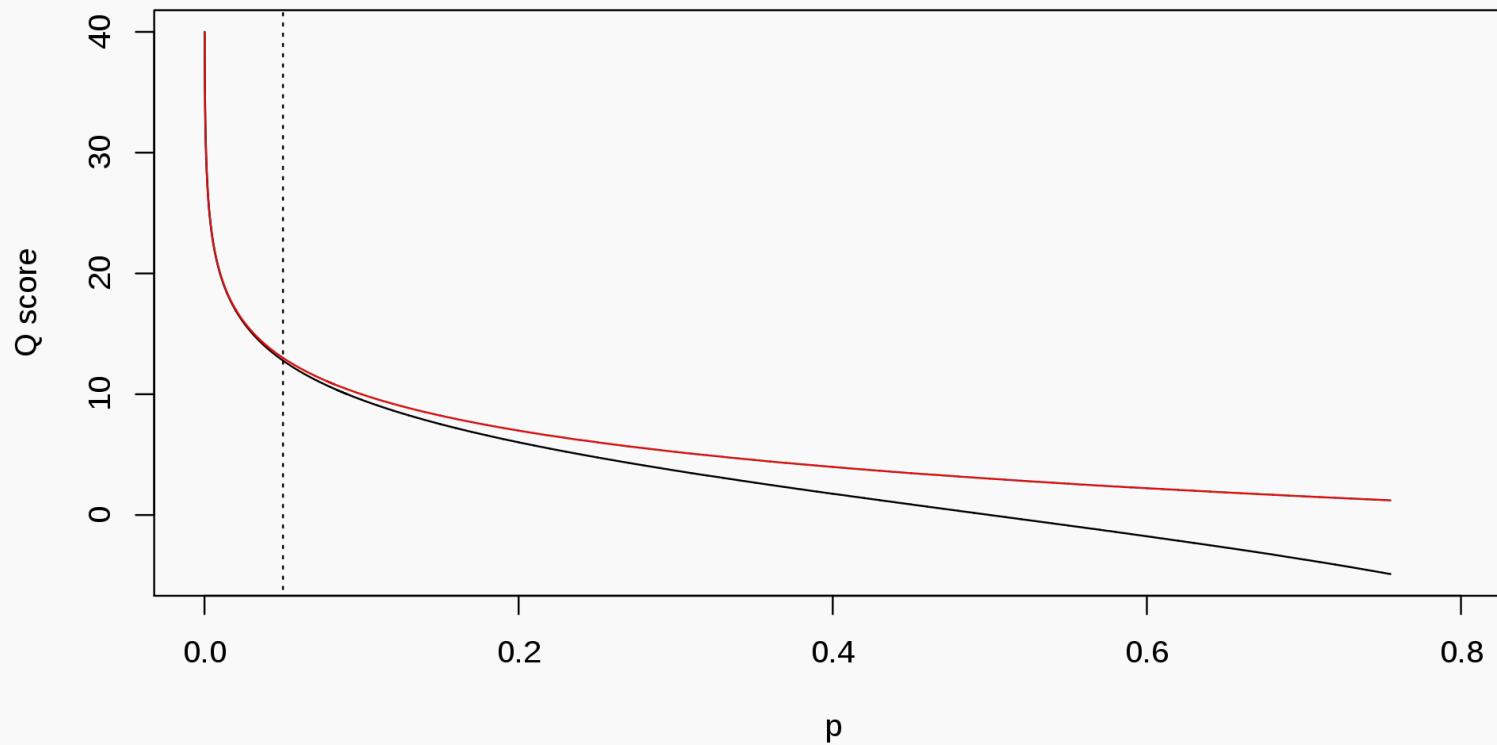
Phred Quality Score
0 .. 50

Code ASCII



Dec	Hex	Oct	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr
0 0	000	NULL		32 20	040	&#032;	Space		64 40	100	&#064;	@		96 60	140	&#096;	`	
1 1	001	SoH		33 21	041	&#033;	!		65 41	101	&#065;	A		97 61	141	&#097;	a	
2 2	002	SoTxt		34 22	042	&#034;	"		66 42	102	&#066;	B		98 62	142	&#098;	b	
3 3	003	EoTxt		35 23	043	&#035;	#		67 43	103	&#067;	C		99 63	143	&#099;	c	
4 4	004	EoT		36 24	044	&#036;	\$		68 44	104	&#068;	D		100 64	144	&#100;	d	
5 5	005	Enq		37 25	045	&#037;	%		69 45	105	&#069;	E		101 65	145	&#101;	e	
6 6	006	Ack		38 26	046	&#038;	&		70 46	106	&#070;	F		102 66	146	&#102;	f	
7 7	007	Bell		39 27	047	&#039;	'		71 47	107	&#071;	G		103 67	147	&#103;	g	
8 8	010	Bsp		40 28	050	&#040;	(		72 48	110	&#072;	H		104 68	150	&#104;	h	
9 9	011	HTab		41 29	051	&#041;	)		73 49	111	&#073;	I		105 69	151	&#105;	i	
10 A	012	LFeed		42 2A	052	&#042;	*		74 4A	112	&#074;	J		106 6A	152	&#106;	j	
11 B	013	VTab		43 2B	053	&#043;	+		75 4B	113	&#075;	K		107 6B	153	&#107;	k	
12 C	014	FFeed		44 2C	054	&#044;	,		76 4C	114	&#076;	L		108 6C	154	&#108;	l	
13 D	015	CR		45 2D	055	&#045;	-		77 4D	115	&#077;	M		109 6D	155	&#109;	m	
14 E	016	SOut		46 2E	056	&#046;	.		78 4E	116	&#078;	N		110 6E	156	&#110;	n	
15 F	017	SIn		47 2F	057	&#047;	/		79 4F	117	&#079;	O		111 6F	157	&#111;	o	
16 10	020	DLE		48 30	060	&#048;	0		80 50	120	&#080;	P		112 70	160	&#112;	p	
17 11	021	DC1		49 31	061	&#049;	1		81 51	121	&#081;	Q		113 71	161	&#113;	q	
18 12	022	DC2		50 32	062	&#050;	2		82 52	122	&#082;	R		114 72	162	&#114;	r	
19 13	023	DC3		51 33	063	&#051;	3		83 53	123	&#083;	S		115 73	163	&#115;	s	
20 14	024	DC4		52 34	064	&#052;	4		84 54	124	&#084;	T		116 74	164	&#116;	t	
21 15	025	NAck		53 35	065	&#053;	5		85 55	125	&#085;	U		117 75	165	&#117;	u	
22 16	026	Syn		54 36	066	&#054;	6		86 56	126	&#086;	V		118 76	166	&#118;	v	
23 17	027	EoTB		55 37	067	&#055;	7		87 57	127	&#087;	W		119 77	167	&#119;	w	
24 18	030	Can		56 38	070	&#056;	8		88 58	130	&#088;	X		120 78	170	&#120;	x	
25 19	031	EoM		57 39	071	&#057;	9		89 59	131	&#089;	Y		121 79	171	&#121;	y	
26 1A	032	Sub		58 3A	072	&#058;	:		90 5A	132	&#090;	Z		122 7A	172	&#122;	z	
27 1B	033	Esc		59 3B	073	&#059;	;		91 5B	133	&#091;	[		123 7B	173	&#123;	{	
28 1C	034	FSep		60 3C	074	&#060;	<		92 5C	134	&#092;	\		124 7C	174	&#124;		
29 1D	035	GSep		61 3D	075	&#061;	=		93 5D	135	&#093;	]		125 7D	175	&#125;	}	
30 1E	036	RSep		62 3E	076	&#062;	>		94 5E	136	&#094;	^		126 7E	176	&#126;	~	
31 1F	037	USep		63 3F	077	&#063;	?		95 5F	137	&#095;	_		127 7F	177	&#127;	Del	

# The QPHRED Value



# The QPHRED Scale



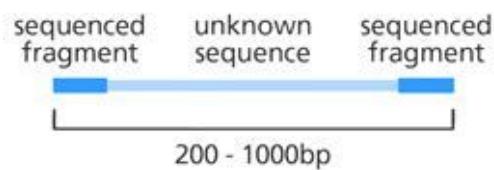
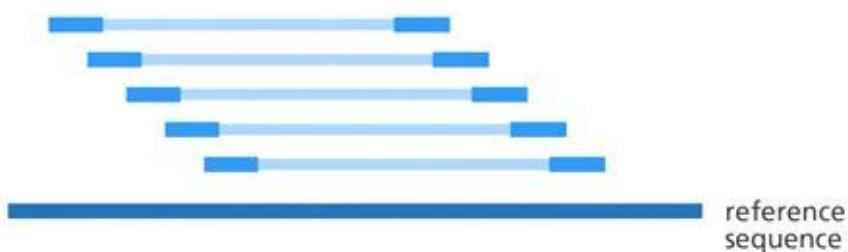
S - Sanger      Phred+33, raw reads typically (0, 40)  
X - Solexa      Solexa+64, raw reads typically (-5, 40)  
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)  
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)  
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (**bold**)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

# Mapping

Single-end reads



Paired-end reads



# SAM Format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
```

```
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAA
r003 0 ref 9 30 5H6M * 0 0 AGCT
r004 0 ref 16 30 6M14N5M * 0 0 ATAG
r003 16 ref 29 30 6H5M * 0 0 TAGG
r001 83 ref 37 30 9M = 7 -39 CAGC
```

## Header

- Ligne commençant par @

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: <i>unsorted</i> , <i>queryname</i> or <i>coordinate</i> .
	GO	Group order (full sorting is not imposed in a group). Valid values are: <i>none</i> , <i>query</i> or <i>reference</i> .
Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
read group	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
	PG	Program name
Program	VN	Program version
	CL	Command line
CO - comment		One-line text comments

# SAM Format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

alignement

Format tabulé

SAM format : <http://samtools.sourceforge.net/samtools.shtml>

# SAM Format

@HD VN:1.3 SO:coordinate

@SQ SN:ref LN:45

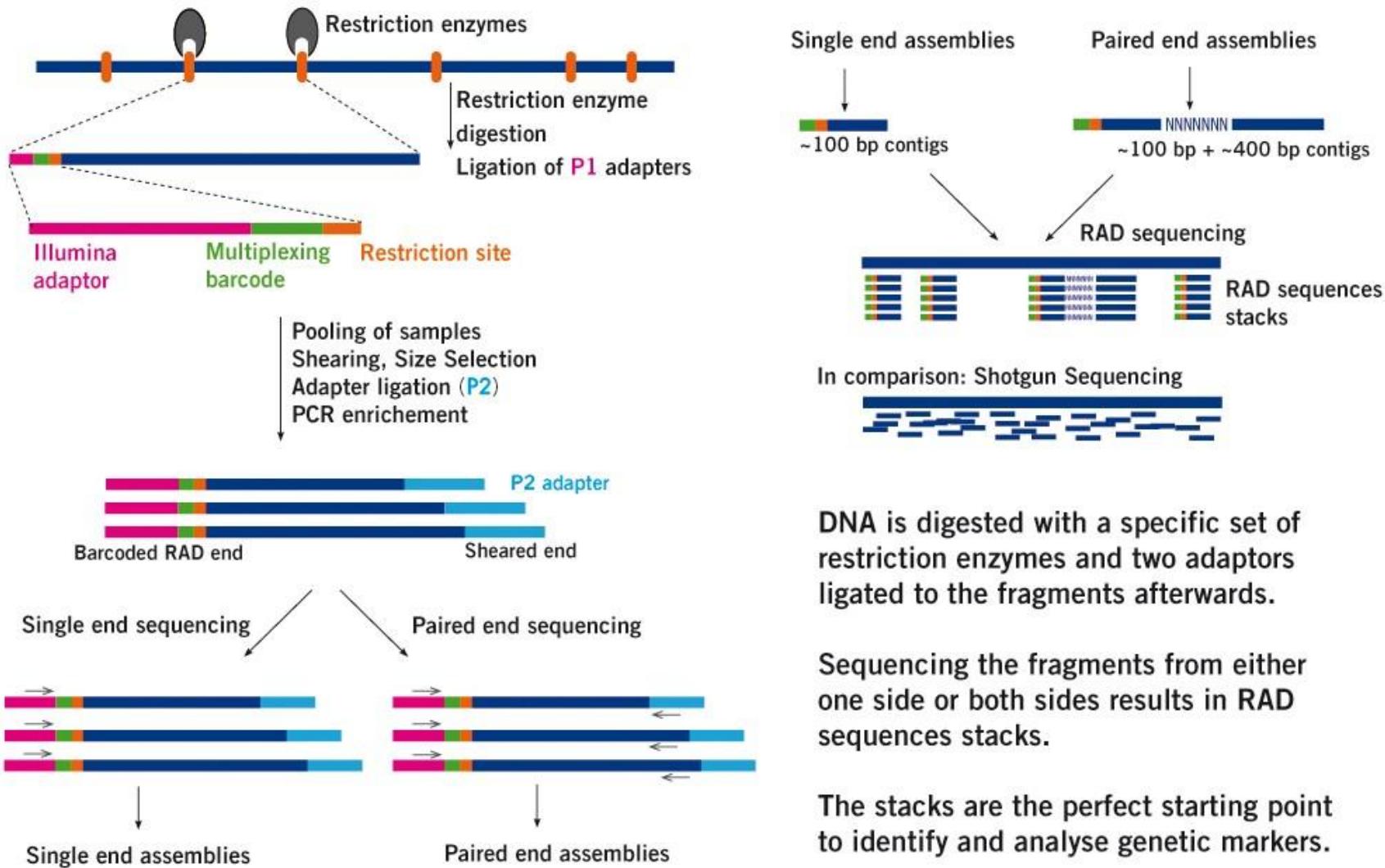
```
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC
r003 16 ref 29 30 6H5M * 0 0 TAGGC * N
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT
```

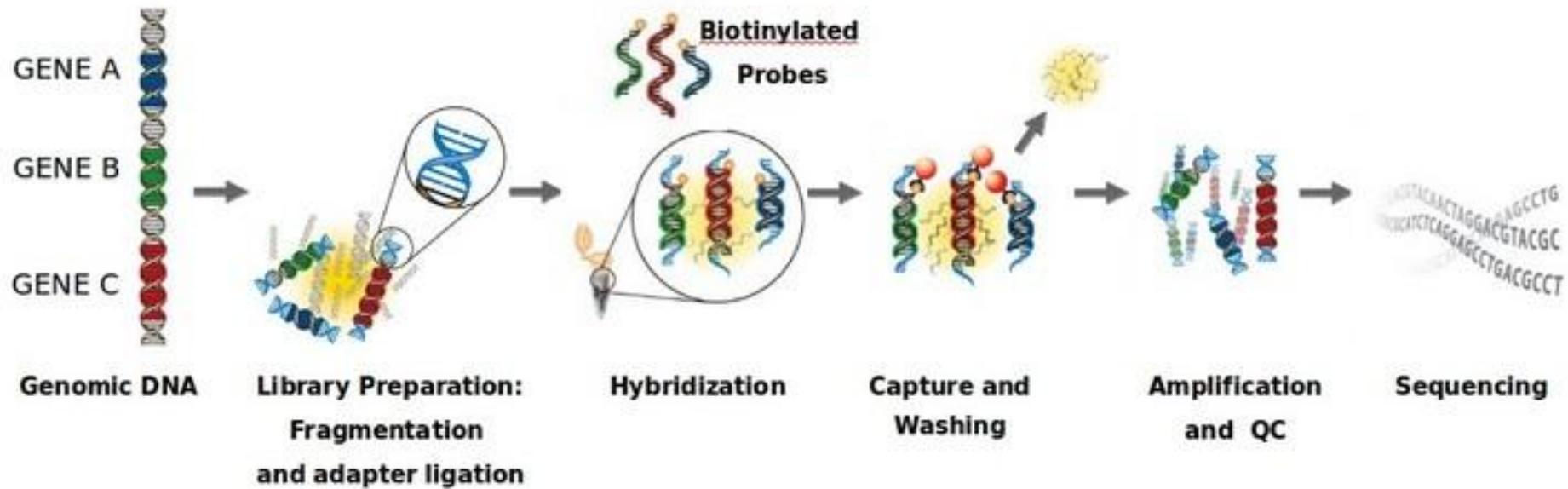
alignement

Format tabulé

Col	Name	Description
1	<b>QNAME</b>	Query NAME of the read or the read pair
2	<b>FLAG</b>	bitwise FLAG (pairing, strand, mate strand, etc.)
3	<b>RNAME</b>	Reference sequence NAME
4	<b>POS</b>	1-based leftmost POSition of clipped alignment
5	<b>MAPQ</b>	MAPping Quality (Phred-scaled)
6	<b>CIGAR</b>	extended CIGAR string (operations: M I D N S H P)
7	<b>NRNM</b>	Mate Reference NaMe ('=' if same as RNAME)
8	<b>MPOS</b>	1-based leftmost Mate POSition
9	<b>ISIZE</b>	inferred Insert SIZE
10	<b>SEQ</b>	query SEQuence on the same strand as the reference
11	<b>QUAL</b>	query QUALity (ASCII-33=Phred base quality)

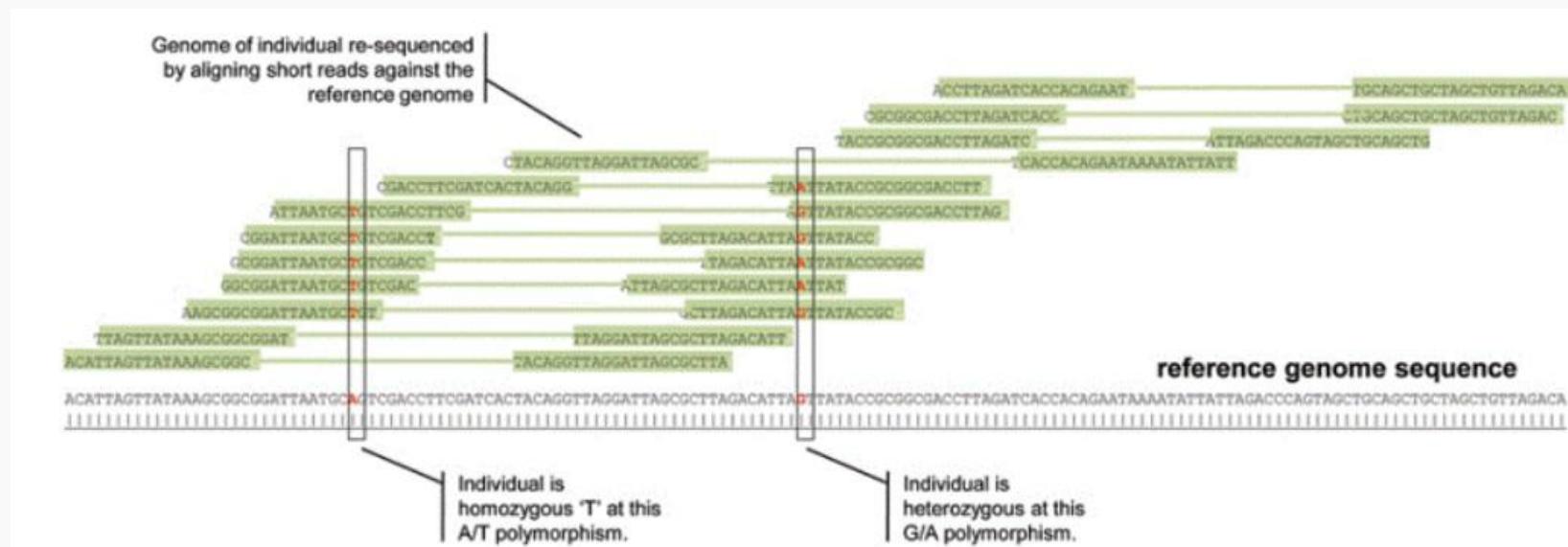
# RAD-Seq





From CGFB, Bordeaux, France

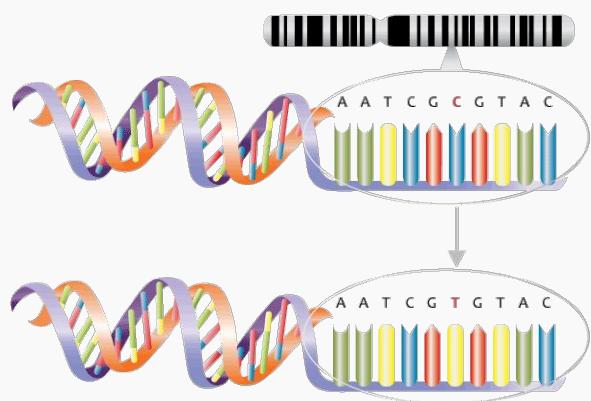
# SNP and InDel Detection



# SNP and InDel Detection

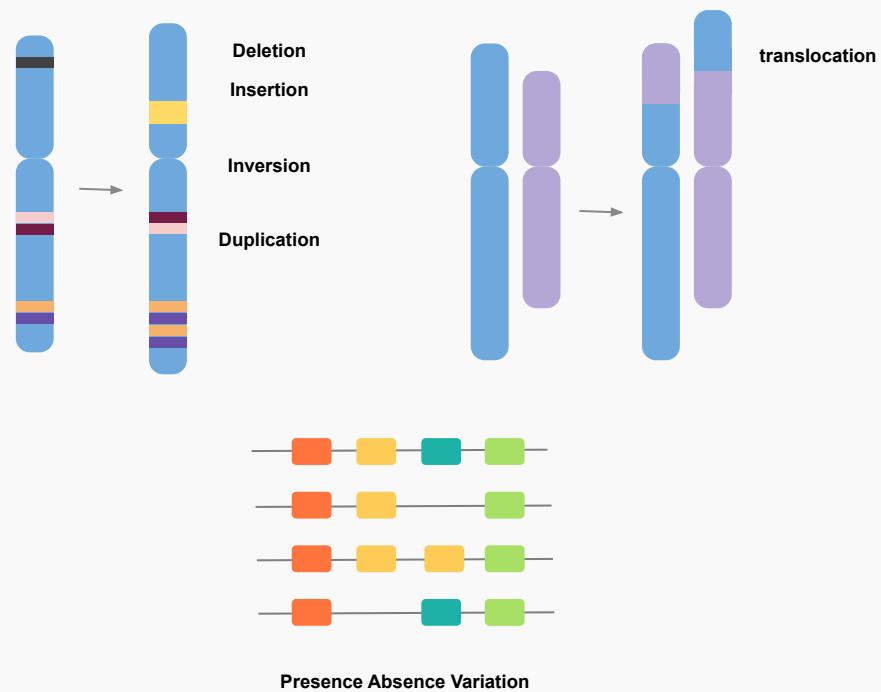
## Mutations & Variations as main source of genetic diversity

SNP



From NHS National Genetics and Genomics Education Centre, CC BY 2.0, via Wikimedia Commons

Structural Variations



# Common File for all Variations, the VCF

## Example

VCF header		Body										
		<pre>##fileformat=VCFv4.0 ##fileDate=20100707 ##source=VCFtools ##reference=NCBI36 ##INFO=&lt;ID=AA,Number=1,Type=String,Description="Ancestral Allele"&gt; ##INFO=&lt;ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"&gt; ##FORMAT=&lt;ID=GT,Number=1,Type=String,Description="Genotype"&gt; ##FORMAT=&lt;ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)"&gt; ##FORMAT=&lt;ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)"&gt; ##FORMAT=&lt;ID=DP,Number=1,Type=Integer,Description="Read Depth"&gt; ##ALT=&lt;ID=DEL,Description="Deletion"&gt; ##INFO=&lt;ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"&gt; ##INFO=&lt;ID=END,Number=1,Type=Integer,Description="End position of the variant"&gt; #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2</pre>										
		<pre>#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2</pre>										
		<pre>1 1 . 1 2 rs1 C A,AT 1 5 . 1 100 T,CT G &lt;DEL&gt;</pre>										
		<pre>REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2</pre>										
		<pre>ACG A,AT . C T,CT H2;AA=T A G PASS . T &lt;DEL&gt; PASS SVTYPE=DEL;END=300</pre>										
		<pre>FORMAT GT:DP 1/2:13 0/0:29 FORMAT GT:GQ 0 1:100 2/2:70 FORMAT GT:GQ 1 0:77 1/1:95 FORMAT GT:GQ:DP 1/1:12:3 0/0:20</pre>										
		<p>Phased data (G and C above are on the same chromosome)</p>										
		<p>Reference alleles (GT=0)</p>										
		<p>Alternate alleles (GT&gt;0 is an index to the ALT column)</p>										
		<p>Deletion SNP Insertion Other event Large SV</p>										

VCF = Variant Call Format From 1000 Genomes  
Project

- Amount of original samples
- Choice of Sample
- Purity of Sample
- Size of sequenced unit
- Error rate
- Volume of Outputted data

All linked to technical constraints

- Cleaning data level
- Mapping Conditions
- Mapping Cleaning Conditions
- Variation Calling level

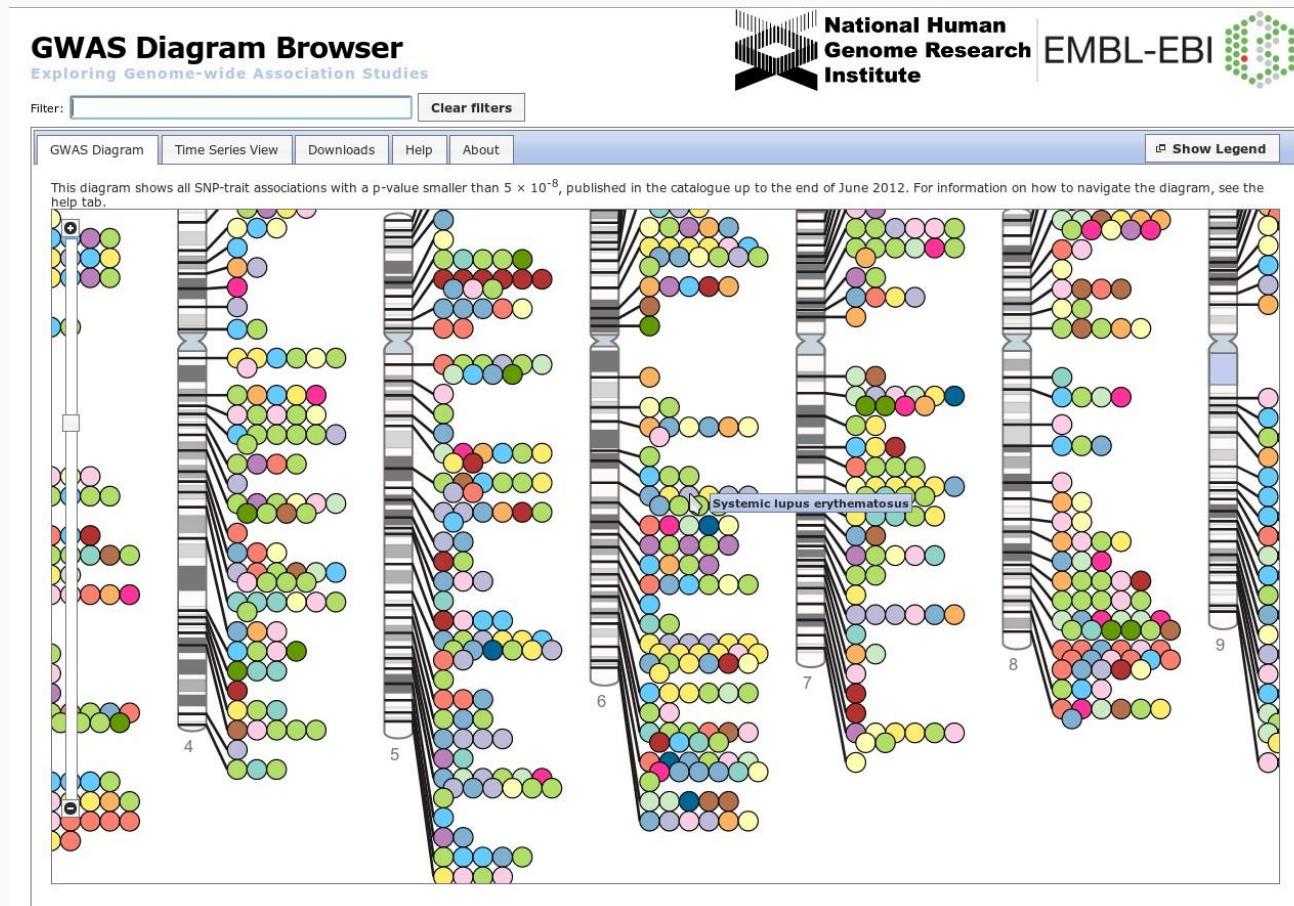
All linked to the Specificity/Sensitivity Informatics Paradox

# Applications

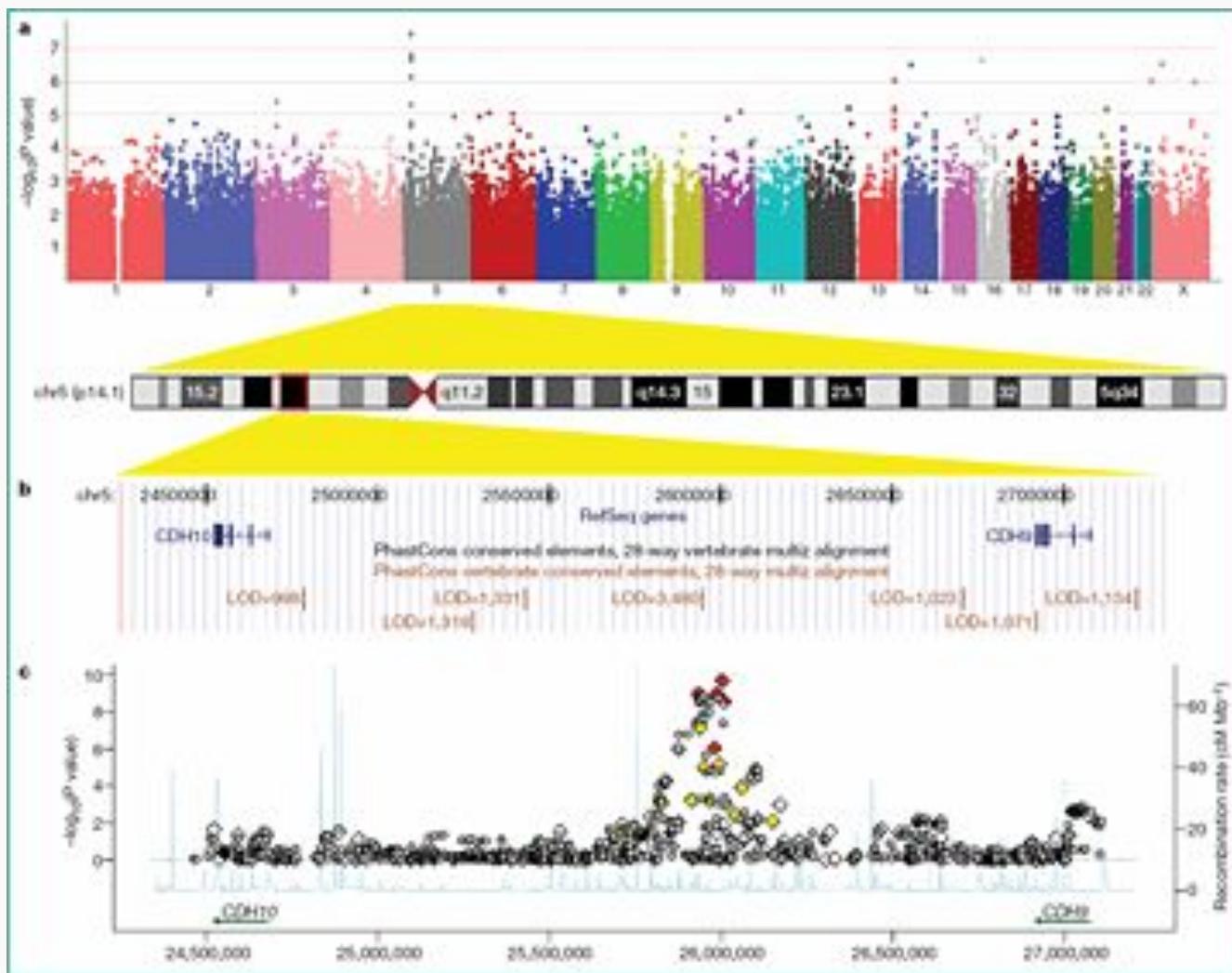
---

- Gene discovery/GWAs
- Species Definition
- Subspecies/specific subgroup definition
- Global genotyping (for breeding in agriculture e.g.)
- Genomic Ecology (Transposable elements, etc...)

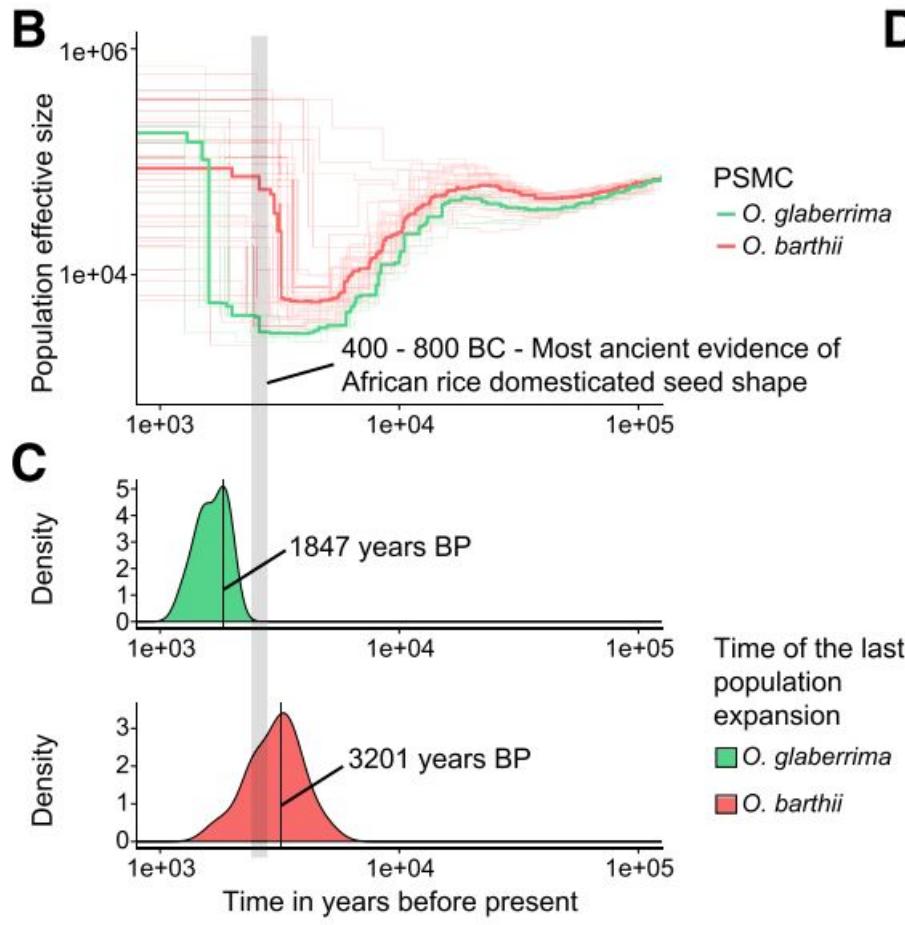
# Example in GWAs & Population Genomics



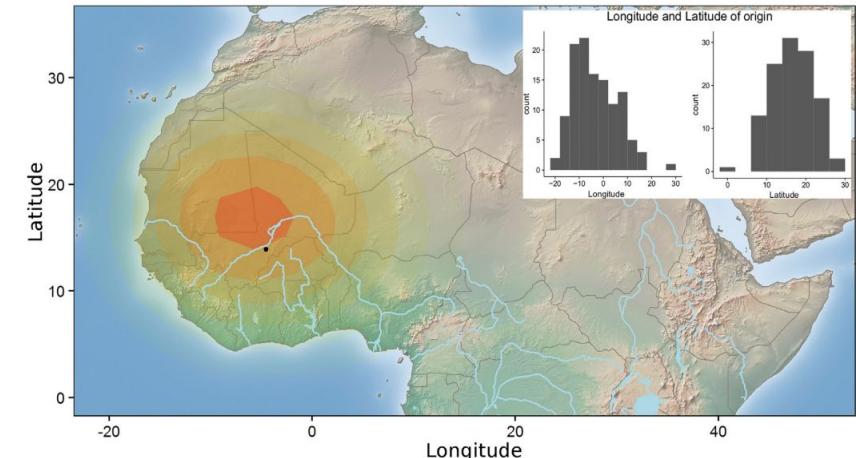
# Example in GWAs & Population Genomics



# Example in Global Genotyping & Population Genomics



**D**



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

From Cubry et al, 2018

# Large Projects

The image displays three side-by-side screenshots of scientific databases:

- 1000 Genomes**: A Deep Catalog of Human Genetic Variation. The screenshot shows a dark header with the project name and a banner featuring a microscopic view of chromosomes. Below the header is a navigation bar with links like Home, About, Data, Analysis, Participants, Contact, and Help desk. A "LATEST ANNOUNCEMENTS" section highlights a "February 2011 Data Up" and a "Full Project Indel Release".
- 1001 Genomes**: A Catalog of *Arabidopsis thaliana* Genetic Variation. The header features a banner with a blue flower. The navigation bar includes Home, Collaborators, Accessions, Tools, Software, Data Center, Gallery, About, and Help desk.
- Genome 10K**: Unveiling animal diversity. The header has a banner with a DNA helix and the text "GENOME 10K® Unveiling animal diversity". The navigation bar includes Database & Species lists, News, Events, Publications, Participants, and For G10K Organizers (restricted). A search bar is also present.

# Sample types

- DNA from plant, animal, microbial...
- Organite DNA (mitochondria, chloroplast)
- Subsample DNA (exon capture, 16S capture for Barcoding)
- Viral sample from infected tissue
- Environmental sample: water, feces, cloud...

# Possibilities in the next 5-10 years (From a presentation in 2013)



- Real-time Transcriptomics

# Possibilities in the next 5-10 years (From a presentation in 2013)



- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014

# Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015

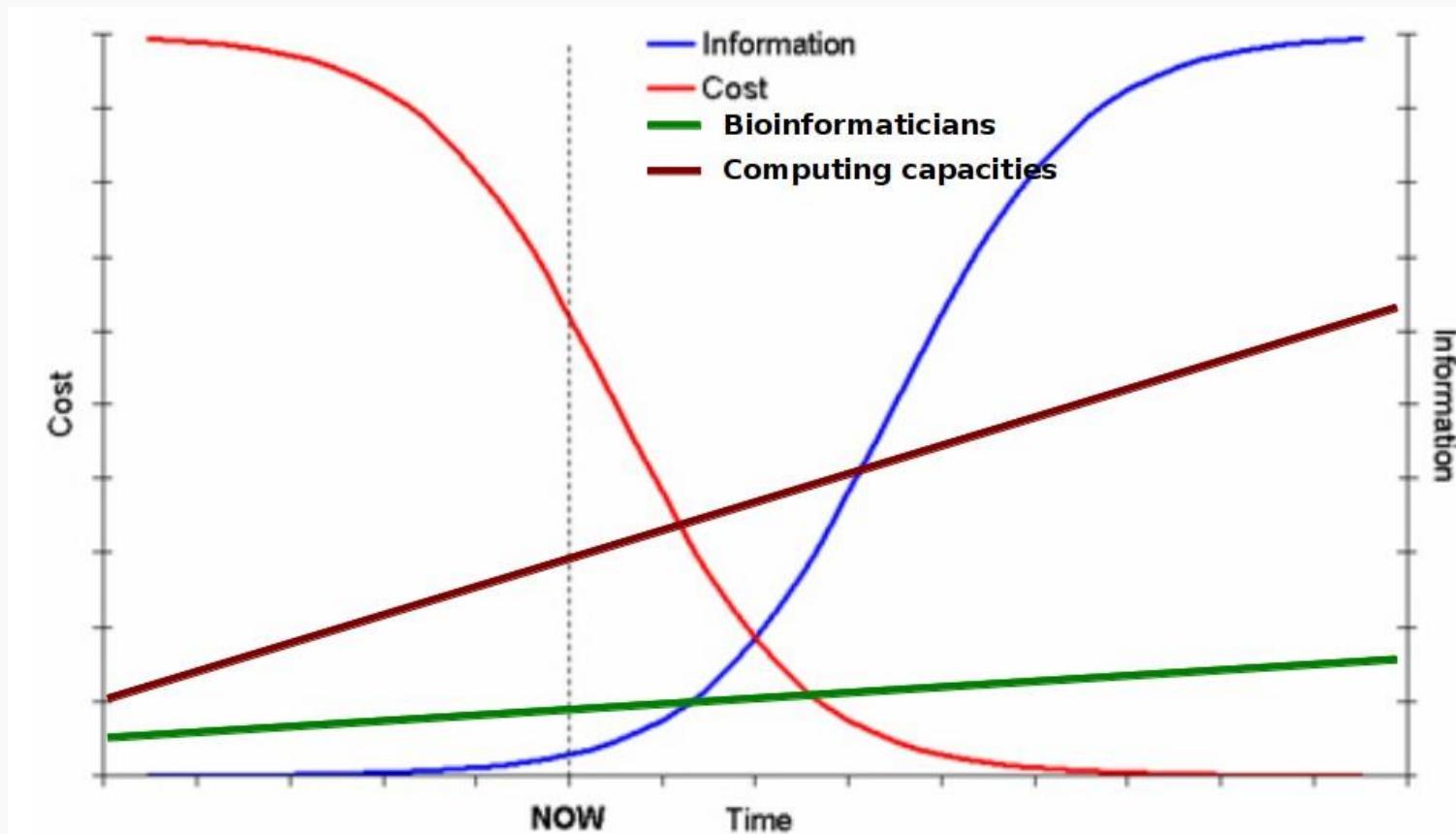
# Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available

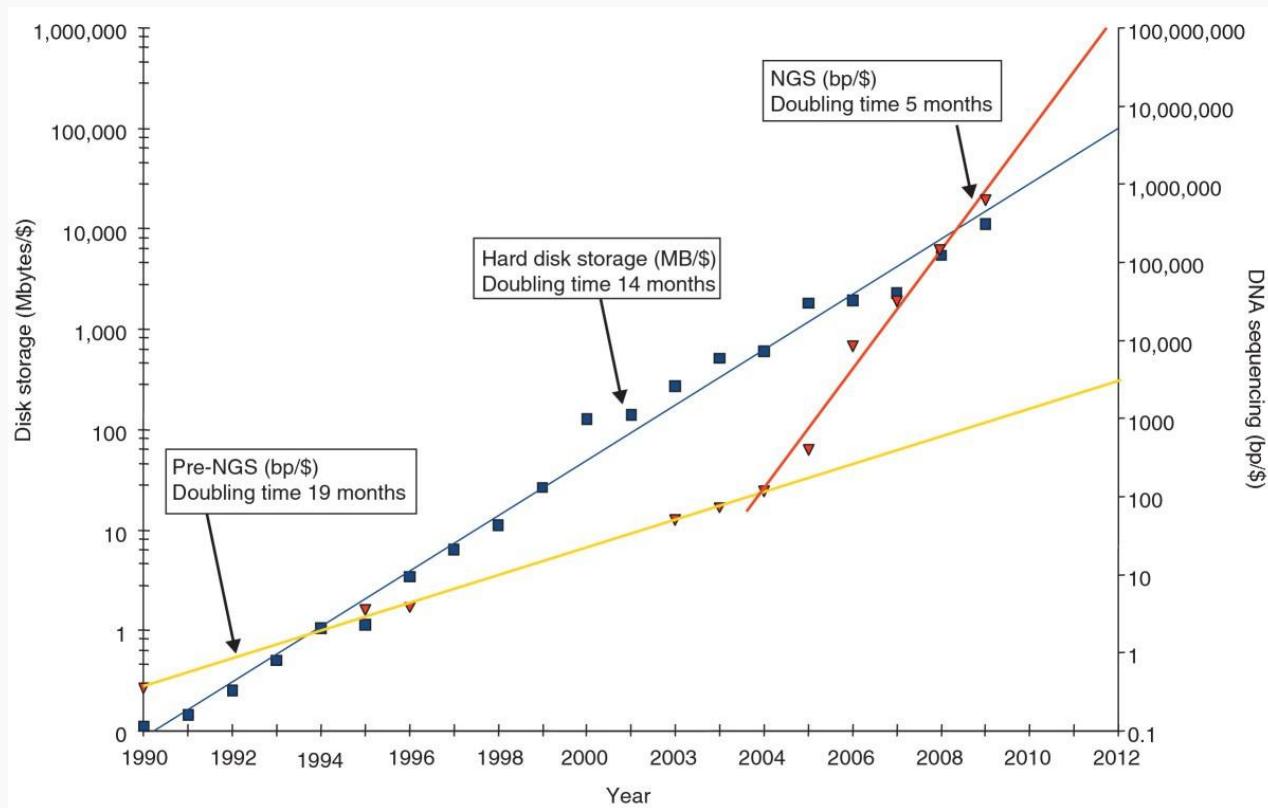
# Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available
- And any new ideas you will have...

# Keep in mind!

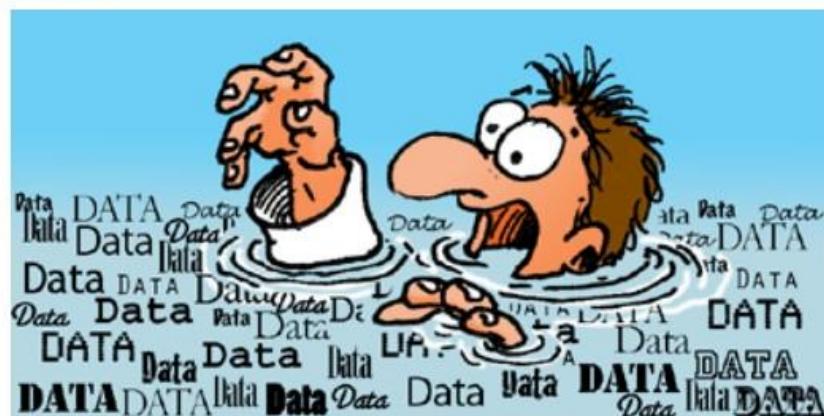
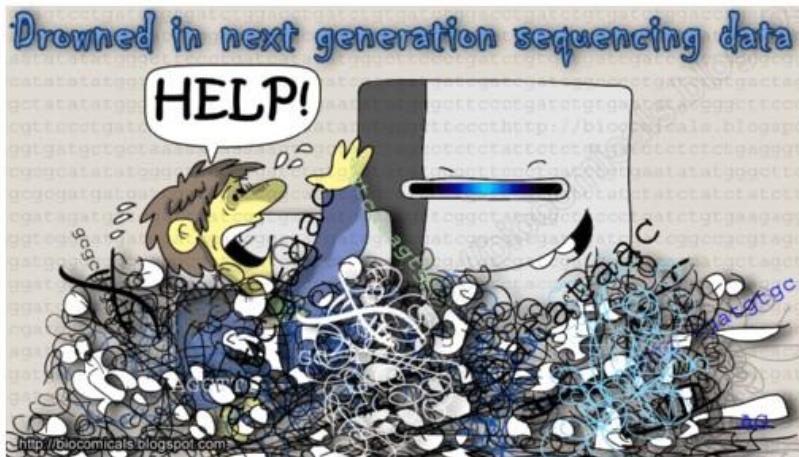


# ...From Data Rarity to Data Deluge



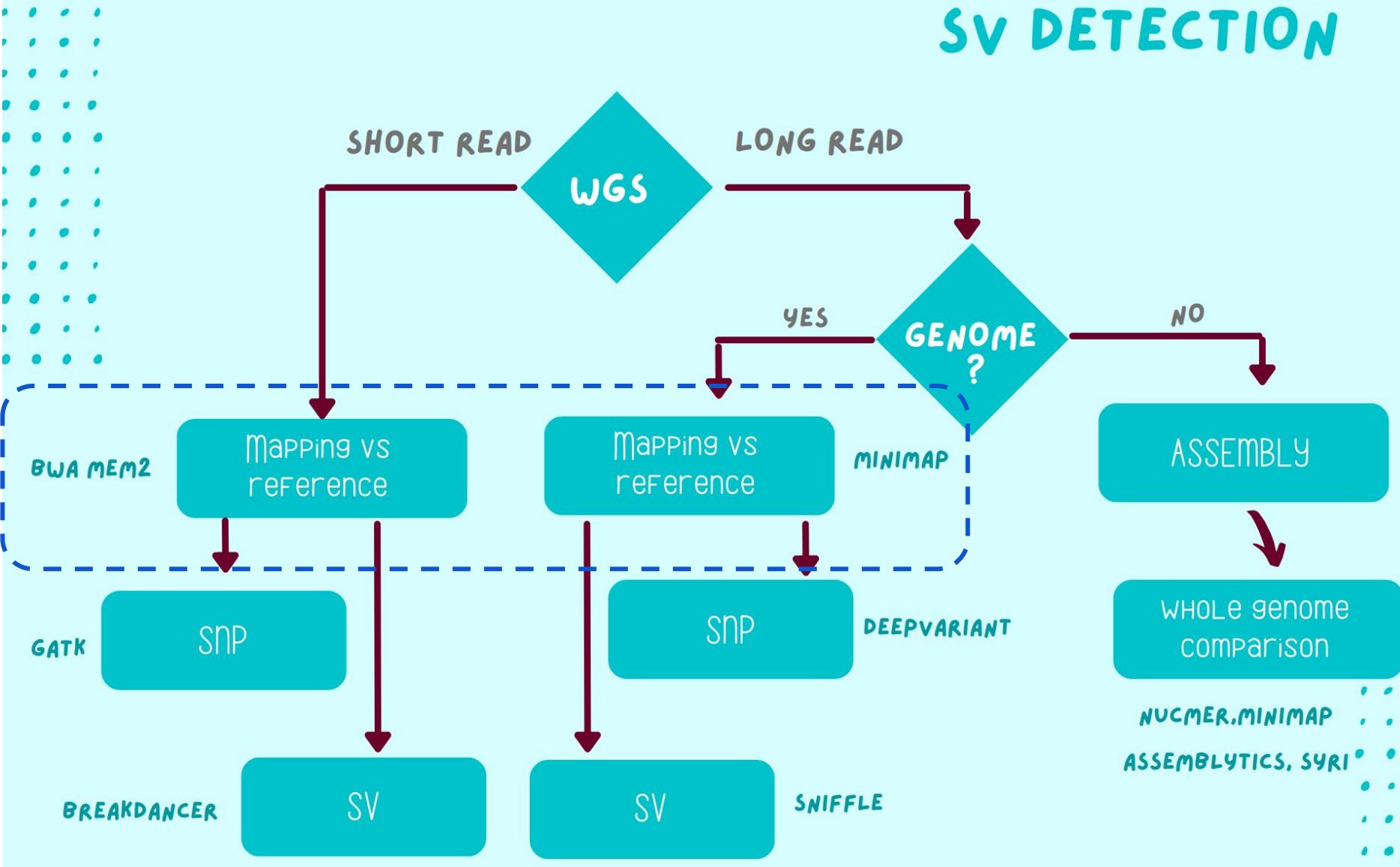
From L. Stein, 2010

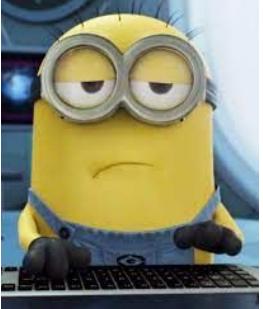
# Be Careful to data drowning!



# Training plan - day 1

## SV DETECTION



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

# Mapping and SNP

- Quality control of NGS data
- Learn to manipulate NGS data
- Having a critical look on *Mapping*
- Learn to launch a *Calling* and having a critical look

# The data

Diploid Asian Rice, *Oryza*



From  
Wikimedia

Diploid Asian Rice, *Oryza*



From  
Wikimedia

1. Select/Cut 1 Mb on Chromosome 10

# The data

Diploid Asian Rice, *Oryza*



From  
Wikimedia

1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones

Diploid Asian Rice, *Oryza*



From  
Wikimedia

1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones
3. Introduce
  - SNP (1-10%),
  - indel (10b-10kb),
  - duplications...

Diploid Asian Rice, *Oryza*



From  
Wikimedia

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...

Diploid Asian Rice, *Oryza*



From  
Wikimedia

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~~

# The FASTQ Format

The diagram illustrates the structure of a FASTQ file. It shows four lines of FASTQ data with blue arrows pointing from the right to specific fields. The first arrow points to the sequencing information line, which starts with '@'. The second arrow points to the nucleotide sequence line, which contains the actual DNA sequence. The third arrow points to the quality score line, which consists of ASCII characters representing the quality of each base. The fourth arrow points to the trailing line, which ends with a double quote.

```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTTGTAAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGAGTGTGAAAGGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
TGGTTTATGCAGAAATTCTAGAATAAGGGTAACCTT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
TCTCAGAAAATTGTTGTGATGTGTATTCAACTA
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
TTGATTTAACTCTGACAAAATAAACAAAAGTCTTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

# The QPHRED Scale

Score qualité phred	Proba. d'une identification incorrecte	Précision de l'identification
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	S.....XXXXXXXXXXXXXX..... .....I.....J.....L.....!#%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNPQRSTUVWXYZ[\]^_`abcdefghijklmnoprstuvwxyz{ }~   59 64 73   104   33 0.....26....31.....40 -5....0.....9.....40 0.....9.....40 3.....9.....40 0.2.....26....31.....41
40	1 pour 10000	
50	1 pour 100000	

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

## “Classic” launch

1. *Mapping:* bwa aln/sampe, bwa mem, bowtie2, ...

## “Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2,  
...
2. *Cleaning mapping*: samtools, picard-tools,...

## “Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2,  
...
2. *Cleaning mapping*: samtools, picard-tools,...
3. *Realigning and Duplicates*: GATK,  
picard-tools,...

OPTIONAL!

## “Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2,  
...
2. *Cleaning mapping*: samtools, picard-tools,...
3. *Realigning and Duplicates*: GATK,  
picard-tools,...
- OPTIONAL!
4. *SNP calling and Cleaning*: GATK,...

## “Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bowtie2,  
...
2. *Cleaning mapping*: samtools, picard-tools,...
3. *Realigning and Duplicates*: GATK,  
picard-tools,...
- OPTIONAL!
4. *SNP calling and Cleaning*: GATK,...

Between 8 and 15 different commands...

# Let's do it by hands...

We will

1. Map the data of Clone 1 on *reference.fasta* using *bwa*

# Let's do it by hands...

We will

1. Map the data of Clone 1 on *reference.fasta* using *bwa*
2. Look at the SAM file
- 3.
- 4.
- 5.
- 6.

## SAM format :

<http://samtools.sourceforge.net/samtools.shtml>

Type	Tag	Description
HD – header	VN*	File format version.
	SO	Sort order. Valid values are: <i>unsorted</i> , <i>queryname</i> or <i>coordinate</i> .
	GO	Group order (full sorting is not imposed in a group). Valid values are: <i>none</i> , <i>query</i> or <i>reference</i> .
SQ – Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
RG – read group	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
	ID*	Program name
PG – Program	VN	Program version
	CL	Command line
CO – comment		One-line text comments

## SAM format :

<http://samtools.sourceforge.net/samtools.shtml>

Col	Name	Description
1	<b>QNAME</b>	Query NAME of the read or the read pair
2	<b>FLAG</b>	bitwise FLAG (pairing, strand, mate strand, etc.)
3	<b>RNAME</b>	Reference sequence NAME
4	<b>POS</b>	1-based leftmost POSition of clipped alignment
5	<b>MAPQ</b>	MAPping Quality (Phred-scaled)
6	<b>CIGAR</b>	extended CIGAR string (operations: MIDNSHP)
7	<b>NRNM</b>	Mate Reference NaMe (`=' if same as RNAME)
8	<b>MPOS</b>	1-based leftmost Mate POSition
9	<b>ISIZE</b>	inferred Insert SIZE
10	<b>SEQ</b>	query SEQuence on the reference
11	<b>QUAL</b>	query QUALity (ASCII-33)

```

@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *

```

## SAM format: FLAG field

numeric	binary	description
1	00000001	template has multiple fragments in sequencing
2	00000010	each fragment properly mapped according to aligner
4	00000100	fragment is unmapped
8	00001000	mate is unmapped
16	00010000	sequence is reverse complemented
32	00100000	sequence of mate is reversed
64	01000000	is first fragment in template
128	10000000	is second fragment in template

We will

1. Map the data of Clone 1 on *reference.fasta* using *bwa*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it

We will

1. Map the data of Clone 1 on *reference.fasta* using *bwa*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it
4. Remove wrong mapping

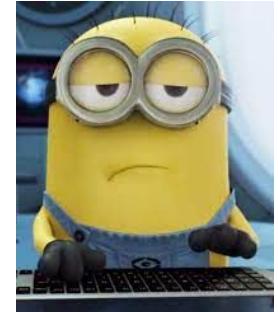


# Practice

Let's work with the jupyter book :

**Day1\_Mapping\_Practice\_EMPTY.ipynb**

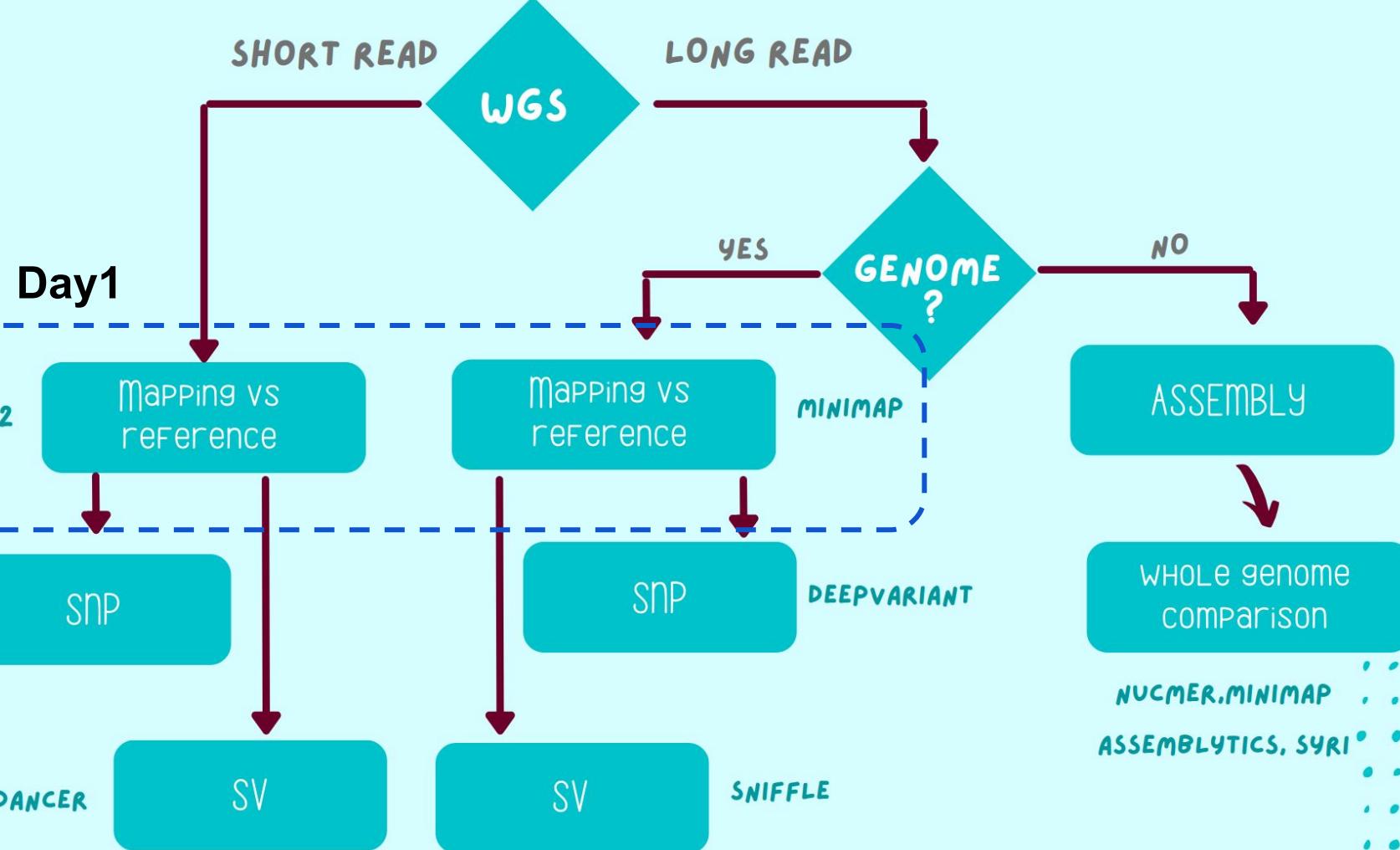


A small image of a yellow Minion character wearing blue overalls and glasses, sitting at a keyboard.

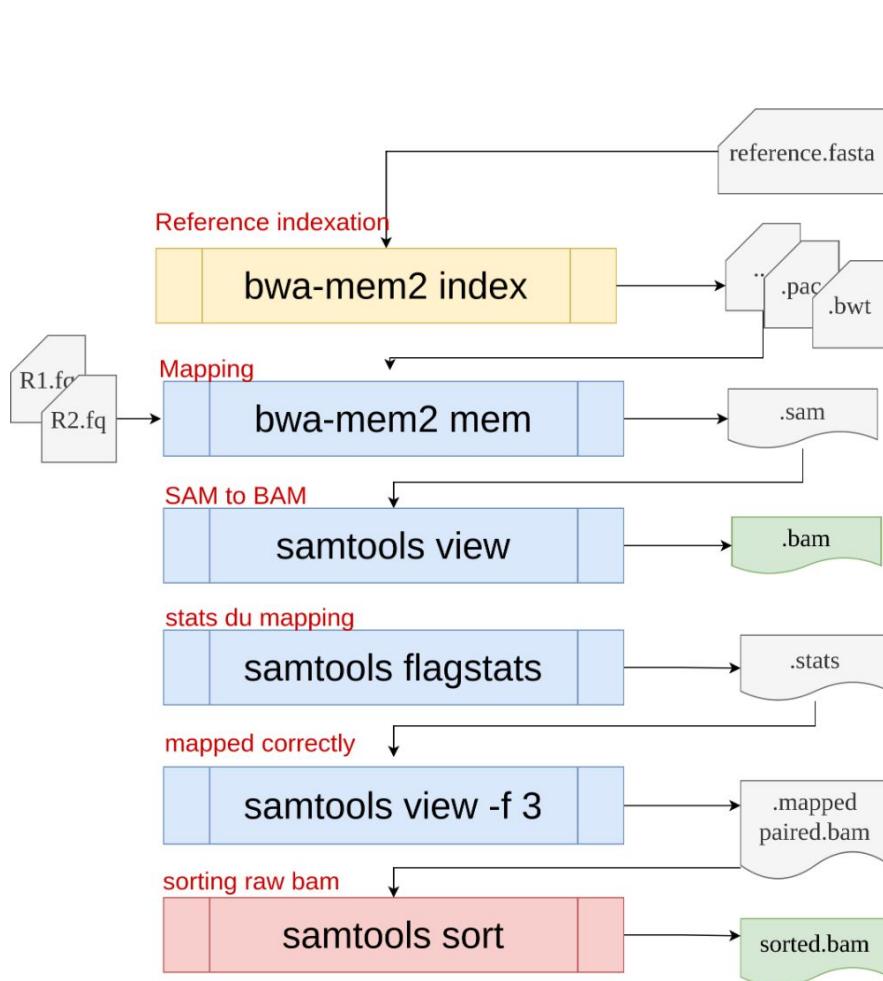
# Day2 - Let's start slowly

# Training plan - day 1 & 2 (a little)

## SV DETECTION

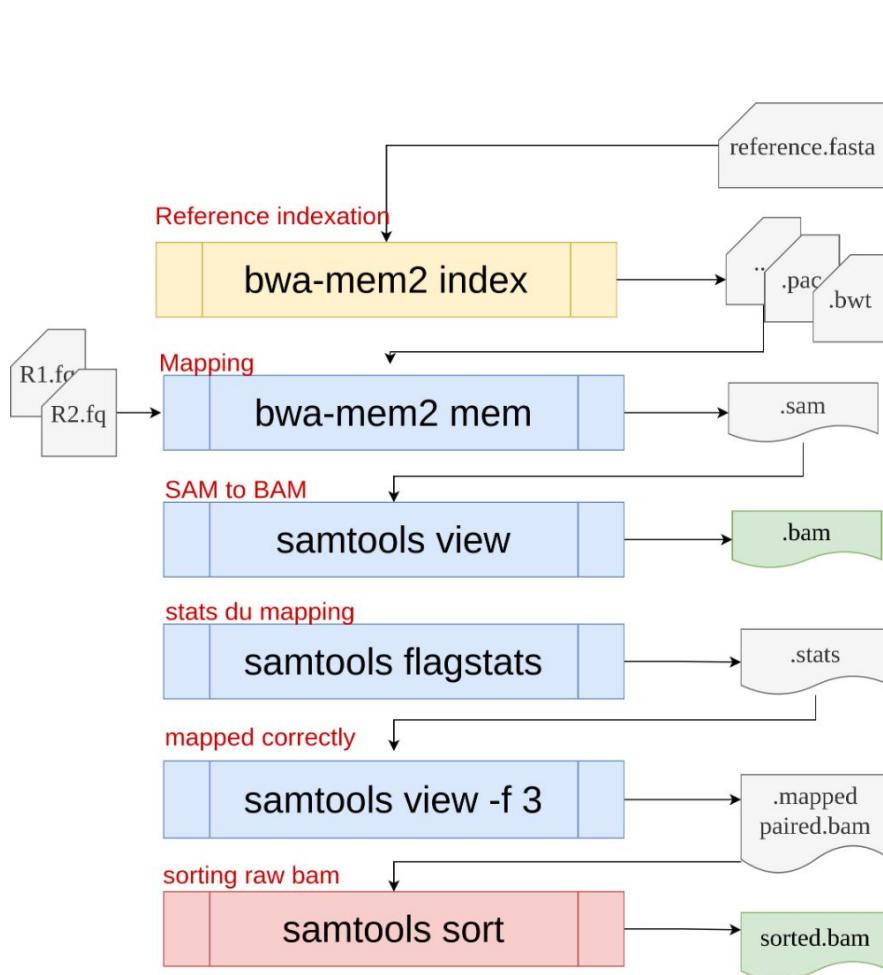


# Plan de bataille du mapping !

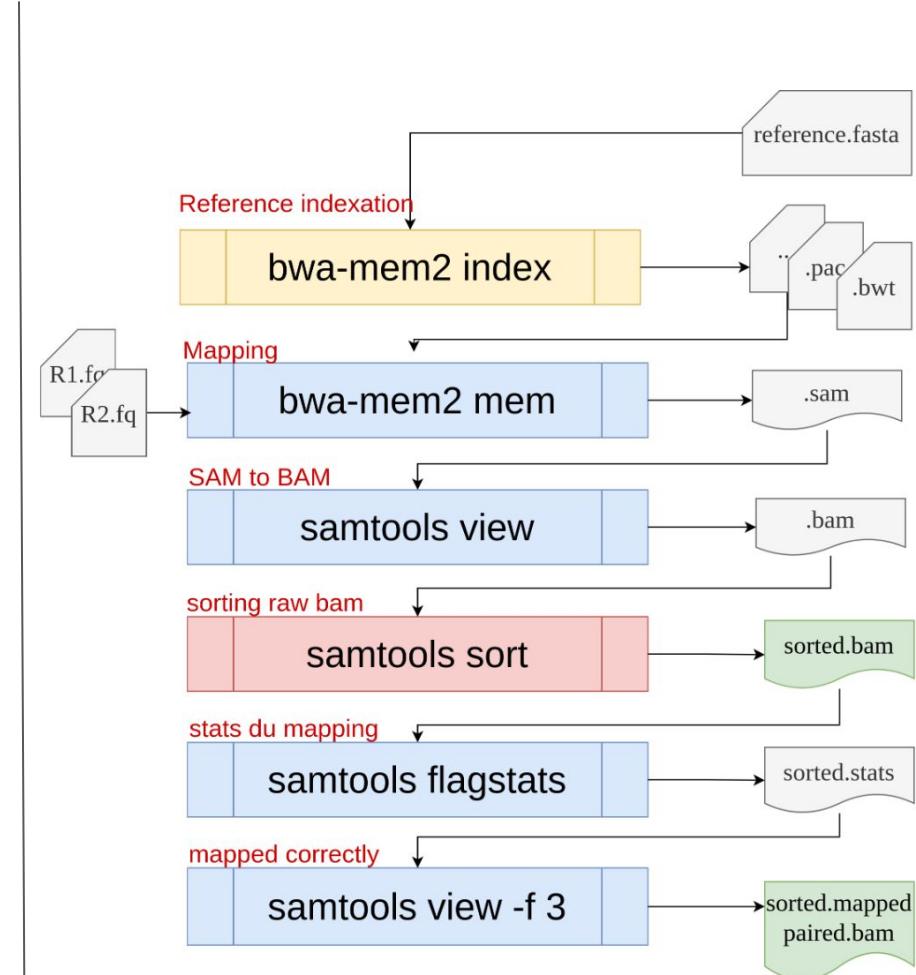


En cours Day1

# Plan de bataille du mapping !



En cours Day1



ça marche mieux !

# Plan de bataille du mapping - bwa-mem2 !

=> Let's map with data from all clones using a loop for mapping, with a single folder per sample

```
for i in {1..20}
do
  cd ~/work/MAPPING-ILL
  echo -e "\n\n>>>>>>> Creation directory for Clone${i}"
  mkdir -p dirClone${i}
  cd dirClone${i}

  echo -e "\n>>> Declare variables${i}"
  REF="/home/jovyan/work/SV_DATA/REF/reference.fasta"
  ILL_R1="/home/jovyan/work/SV_DATA/SHORT_READS/Clone${i}_R1.fastq.gz"
  ILL_R2="/home/jovyan/work/SV_DATA/SHORT_READS/Clone${i}_R2.fastq.gz"

  echo -e "\n>>> Mapping Clone${i}\n"
  bwa-mem2 mem -M -t 8 $REF $ILL_R1 $ILL_R2 > Clone${i}.sam

  echo -e "\n>>> convert sam to bam for Clone${i}"
  samtools view -@4 -bh -S -o Clone${i}.bam Clone${i}.sam
  rm Clone${i}.sam

  echo -e "\n>>> Sort raw bam file ${i}"
  samtools sort -@4 Clone${i}.bam Clone${i}.sorted
  rm Clone${i}.bam

  echo -e "\n>>> Flagstats from all reads using by default sorted bam ${i}"
  samtools flagstat Clone${i}.sorted.bam >Clone${i}.sorted.flagstat

  echo -e "\n>>> Extract only correctly mapped ${i}"
  samtools view -bh -@4 -f 0x02 -o Clone${i}.mappedpaired.bam Clone${i}.sorted.bam

  echo -e "\n>>> index mappedpaired bam ${i}"
  samtools index Clone${i}.sorted.mappedpaired.bam
  samtools index Clone${i}.sorted.bam

done
```

# Plan de bataille du mapping - minimap2 !

=> Let's map with data from all clones using a loop for mapping, with a single folder per sample and ONT reads

```
for i in {1..20}
do
    ONT="/home/jovyan/work/SV_DATA/LONG_READS/Clone${i}.fastq.gz"
    mkdir -p ~/work/MAPPING-ONT
    cd ~/work/MAPPING-ONT
    echo -e "\n>>>>>>>> Creation directory for Clone${i}\n"
    mkdir -p dirClone${i}
    cd dirClone${i}

    echo -e ">>>> Mapping Clone${i} minimap2\n"
    minimap2 -ax map-ont -t 12 ${REF} ${ONT} > Clone${i}_ONT.sam

    echo -e "\nConvert samtobam \n"
    samtools view -@8 -bh -S -o Clone${i}_ONT.bam Clone${i}_ONT.sam
    rm Clone${i}_ONT.sam

    echo -e "\nSort and index raw bam \n"
    samtools sort -@8 Clone${i}_ONT.bam Clone${i}_ONT.sorted

    echo -e "\nCalculate stats from mapping\n"
    samtools flagstat Clone${i}_ONT.sorted.bam >Clone${i}_ONT.sorted.flagstats

    echo -e "\nFilter raw bam \n"
    samtools view -@8 -bh -F 0x904 -o Clone${i}_ONT.sorted.correctlymapped.bam Clone${i}_ONT.sorted.bam

    echo -e "\nindex raw and filtered bam files \n"
    samtools index Clone${i}_ONT.sorted.bam
    samtools index Clone${i}_ONT.sorted.correctlymapped.bam

done
```

- Download Tablet (**use Google and  
Tablet+NGS**)

- Download Tablet (**use Google and Tablet+NGS**)
- Transfer the BAM from Clone 10 and the reference from the machine to your local computer (use scp or direct download from the browser)

- Download Tablet (**use Google and Tablet+NGS**)
- Transfer the BAM from Clone 10 and the reference from the machine to your local computer (use scp or direct download from the browser)

- Download Tablet (**use Google and Tablet+NGS**)
- Transfer the BAM from Clone 10 and the reference from the machine to your local computer (use scp or direct download from the browser)
- Open Tablet, load an assembly

# Observing mapping

- Download Tablet (**use Google and Tablet+NGS**)
- Transfer the BAM from Clone 10 and the reference from the machine to your local computer (use scp or direct download from the browser)
- Open Tablet, load an assembly
- Look at the mapping and try to find SNPs



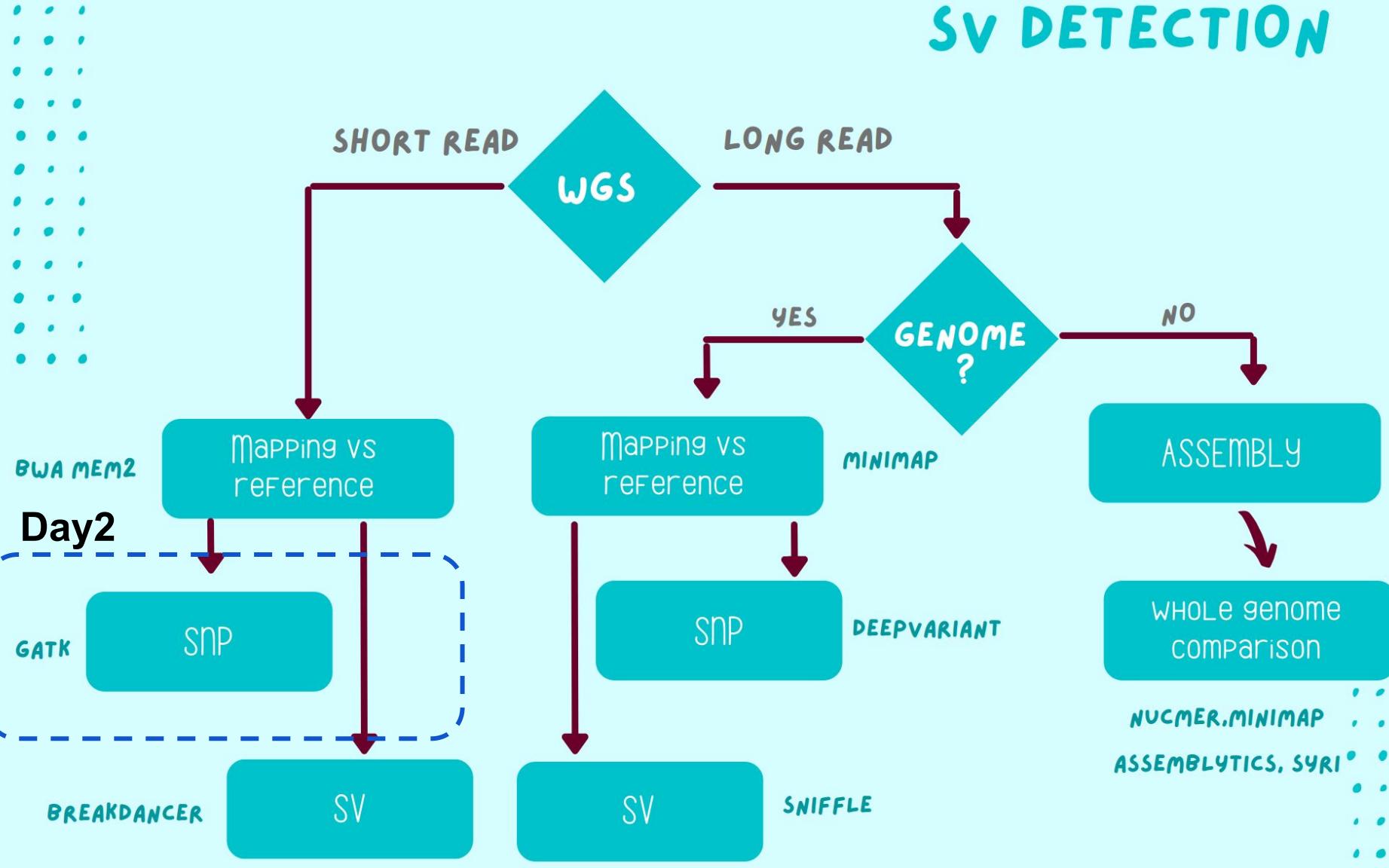
# Practice

Let's work with the jupyter book :

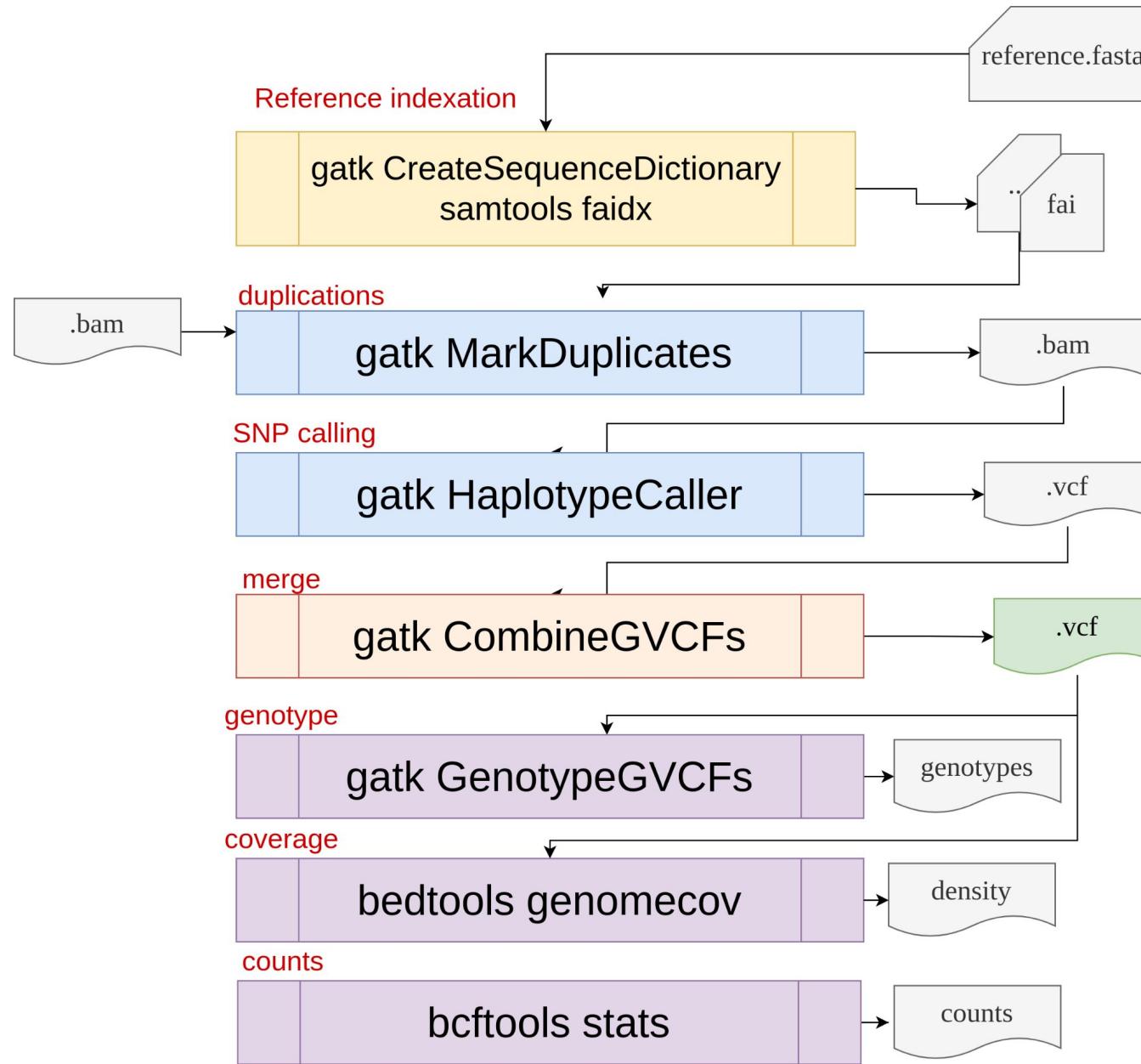
**Day2a\_Mapping\_analysis-EMPTY.ipynb**

# Training plan - day 2

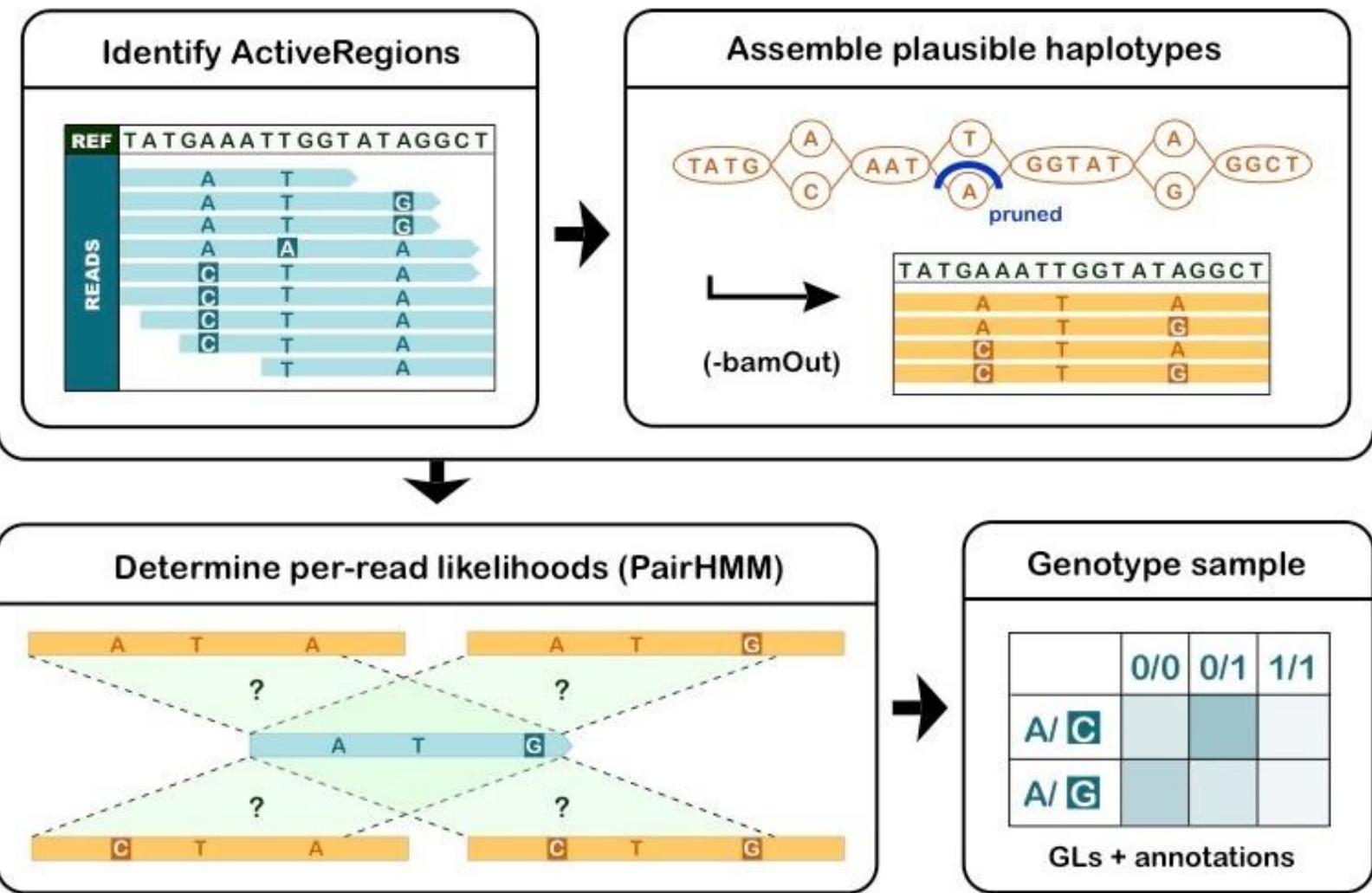
## SV DETECTION



# Plan de bataille du “SNP calling” !



# The HaplotypeCaller



From UniBe.ch  
Training

## The Variant Call Format (VCF) used in bioinformatics for storing gene sequence variations

<b>VCF header</b>	<pre>##fileformat=VCFv4.0 ##fileDate=20100707 ##source=VCFtools ##reference=NCBI36 ##INFO=&lt;ID=AA,Number=1,Type=String,Description="Ancestral Allele"&gt; ##INFO=&lt;ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"&gt; ##FORMAT=&lt;ID=GT,Number=1,Type=String,Description="Genotype"&gt; ##FORMAT=&lt;ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)"&gt; ##FORMAT=&lt;ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)"&gt; ##FORMAT=&lt;ID=DP,Number=1,Type=Integer,Description="Read Depth"&gt; ##ALT=&lt;ID=DEL,Description="Deletion"&gt; ##INFO=&lt;ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"&gt; ##INFO=&lt;ID=END,Number=1,Type=Integer,Description="End position of the variant"&gt;</pre>																																																																	
											<b>Optional header lines</b> (meta-data about the annotations in the VCF body)																																																							
<b>Body</b>	<table border="1"> <thead> <tr> <th>#CHROM</th><th>POS</th><th>ID</th><th>REF</th><th>ALT</th><th>QUAL</th><th>FILTER</th><th>INFO</th><th>FORMAT</th><th>SAMPLE1</th><th>SAMPLE2</th></tr> </thead> <tbody> <tr> <td>1</td><td>1</td><td>.</td><td>ACG</td><td>A,AT</td><td>.</td><td>PASS</td><td>.</td><td>GT:DP</td><td>1/2:13</td><td>0/0:29</td></tr> <tr> <td>1</td><td>2</td><td>rs1</td><td>C</td><td>T,CT</td><td>.</td><td>PASS</td><td>H2;AA=T</td><td>GT:GQ</td><td>0 1:100</td><td>2/2:70</td></tr> <tr> <td>1</td><td>5</td><td>.</td><td>A</td><td>G</td><td>.</td><td>PASS</td><td>.</td><td>GT:GQ</td><td>1 0:77</td><td>1/1:95</td></tr> <tr> <td>1</td><td>100</td><td></td><td>T</td><td>&lt;DEL&gt;</td><td>.</td><td>PASS</td><td>SVTYPE=DEL;END=300</td><td>GT:GQ:DP</td><td>1/1:12:3</td><td>0/0:20</td></tr> </tbody> </table>											#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2	1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29	1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70	1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95	1	100		T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2																																																								
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29																																																								
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70																																																								
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95																																																								
1	100		T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20																																																								
<p>Deletion      SNP      Large SV      Insertion      Other event</p>										<b>Reference alleles</b> (GT=0)																																																								
<p>Phased data (G and C above are on the same chromosome)</p>										<b>Alternate alleles</b> (GT>0 is an index to the ALT column)																																																								



# Practice

Let's work with the jupyter book :

**Day2b\_Appel\_variants\_SNP\_EMPTY.ipynb**

# Plan de bataille du “SNP filtering” !

- Using GATK Variant Filtration, a flag per filter

## Using GATK Variant Filtration, a flag per filter

- Depth filter:  $DP < ?$  or  $DP > ?$
- QUAL filter:  $QUAL < ?$
- SNPcluster filter: *more than 3 SNP per 10b*



# Practice

Let's work with the jupyter book :

**Day2c\_SNP\_analysis\_EMPTY.ipynb**

## Problems with manual launches

- Long
- Fastidious
- Error prone
- Tracability and reproducibilty not ensured

## Problems with manual launches

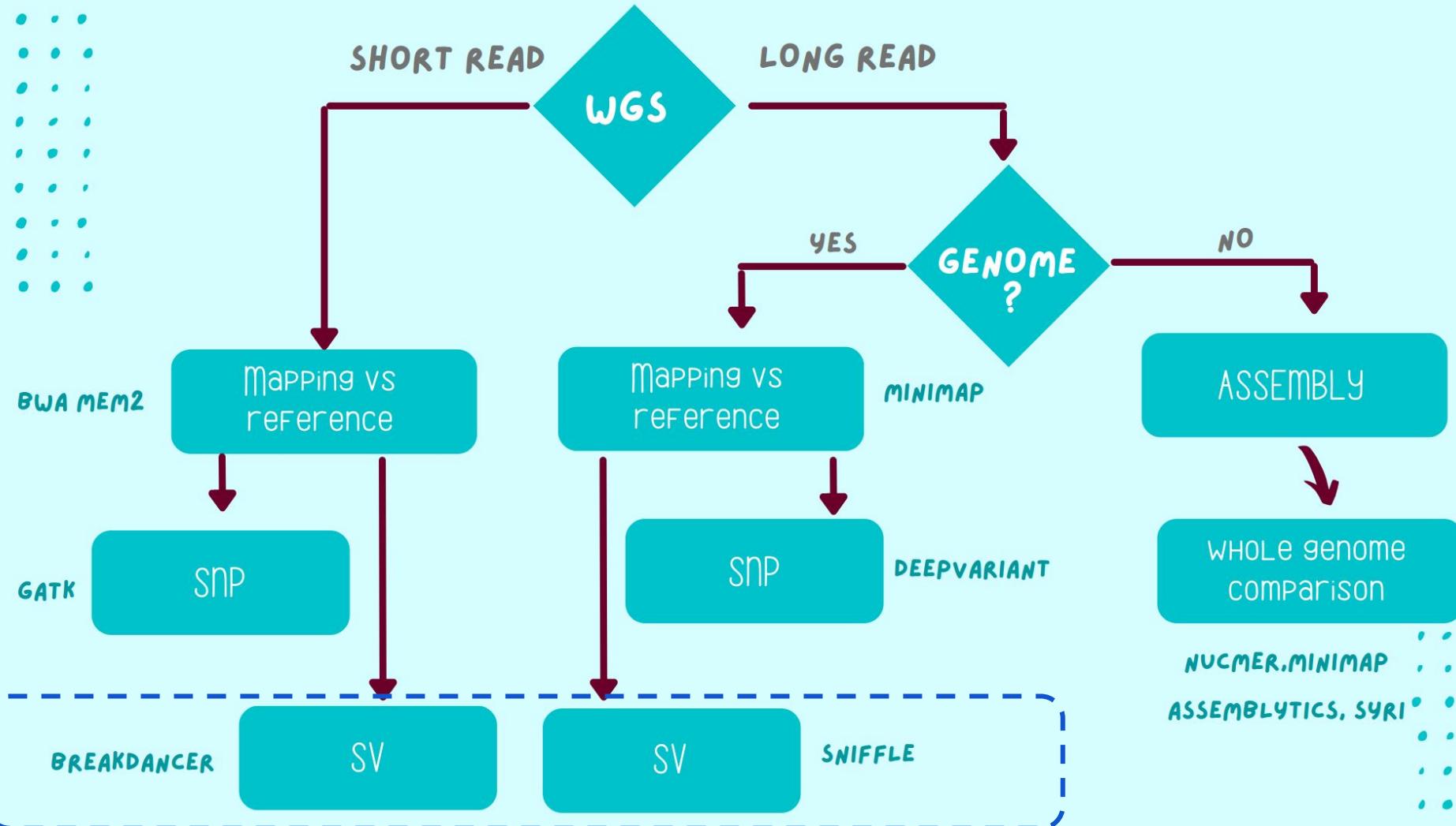
- Long
- Fastidious
- Error prone
- Tracability and reproducibility not ensured

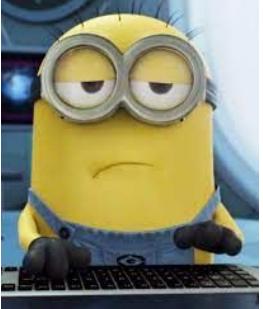
Solution  $\Rightarrow$  Workflow Manager (or scripts...):  
SnakeMake, TOGGLE, NextFlow



# Training plan - day 3

## SV DETECTION



A small image of a yellow Minion character wearing black-rimmed glasses and a blue vest, sitting at a computer keyboard.

# Long Read Mapping

## Two technologies

Oxford Nanopore



MinION

GridION

PromethION

Pacific BioScience



RSII



Sequel

from Elixir GAAS 2018

# Long Reads

<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

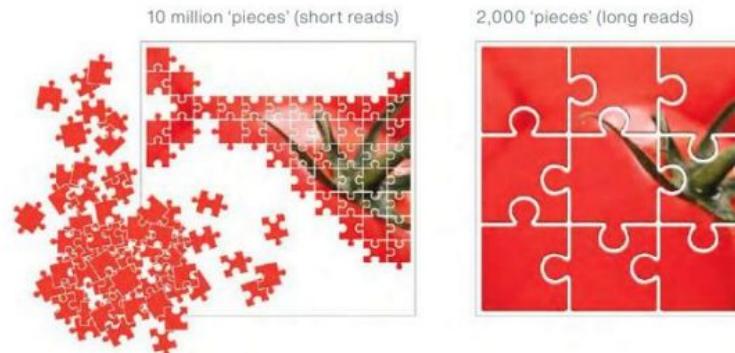
Microbial genomes

Human genomes

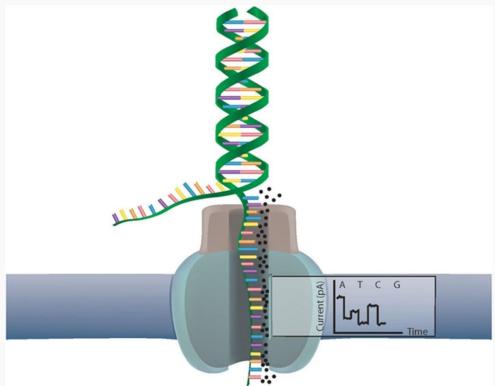
Animal genomes

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

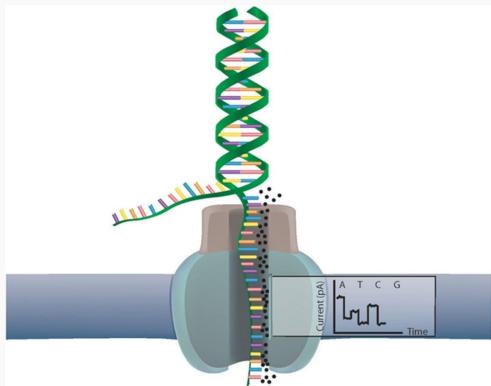






From Circulation  
Research

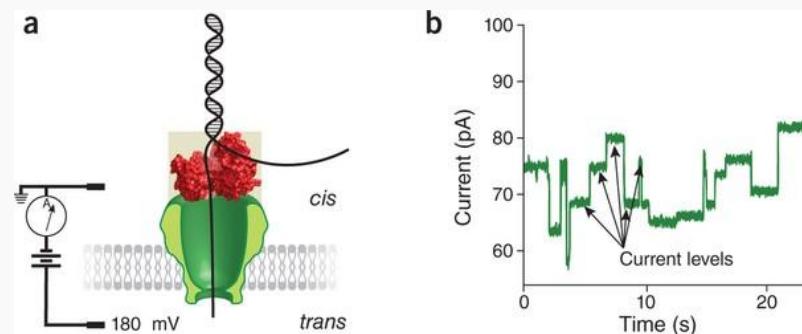
- No Amplification
- NO SYNTHESIS
- Very Long Length



From Circulation  
Research

- No Amplification
- NO SYNTHESIS
- Very Long Length

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



From Nature Biotechnology

## Advantages :

- Length (mean 10-50kb, more than 2Mb reported)
- Bases Modification detection in real-time
- Native RNA!
- Single strand direct sequencing
- Machine cheap (1,000 USD for Minion)
- Run cheap (1,000 USD for 30Gb by now minimum)
- Fast: 15mn library, 48-72h run

## Advantages

- Length (mean 10-50kb, more than 2Mb reported)
- Bases Modification detection in real-time
- Native RNA!
- Single strand direct sequencing
- Machine cheap (1,000 USD for Minion)
- Run cheap (1,000 USD for 30Gb by now minimum)
- Fast: 15mn library, 48-72h run

## Limits :

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

# What you can do with it ?

## Research areas

 Microbiology

 Microbiome

 Environmental

 Plant

 Animal

 Human genomics

 Clinical research

 Cancer

 Transcriptome

 Populations  
genomics

# What you can do with it ?

## Research areas

- Microbiology
- Microbiome
- Environmental
- Plant
- Animal

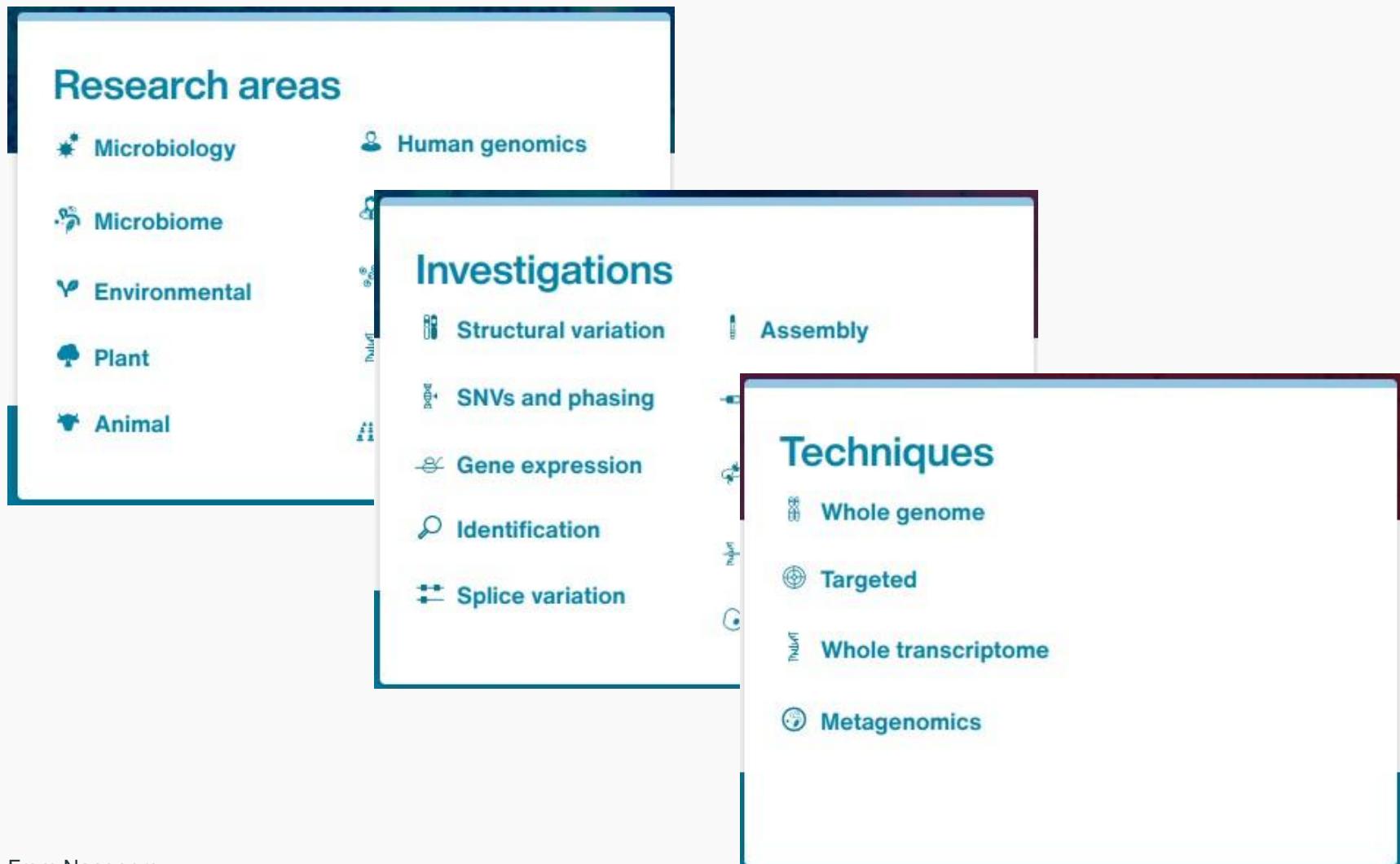
## Human genomics

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

## Investigations

- Assembly
- Fusion transcripts
- Chromatin conformation
- Epigenetics
- Single cell

# What you can do with it ?



## 2<sup>nd</sup> Generation Sequencing

- DNA fragmentation (short) 454
- Matrix amplification IonTorrent
- Short reads Illumina
- Limited error rate
- High throughput

PacificBiosciences  
**Oxford**  
**Nanopore**

## 3<sup>rd</sup> Generation Sequencing

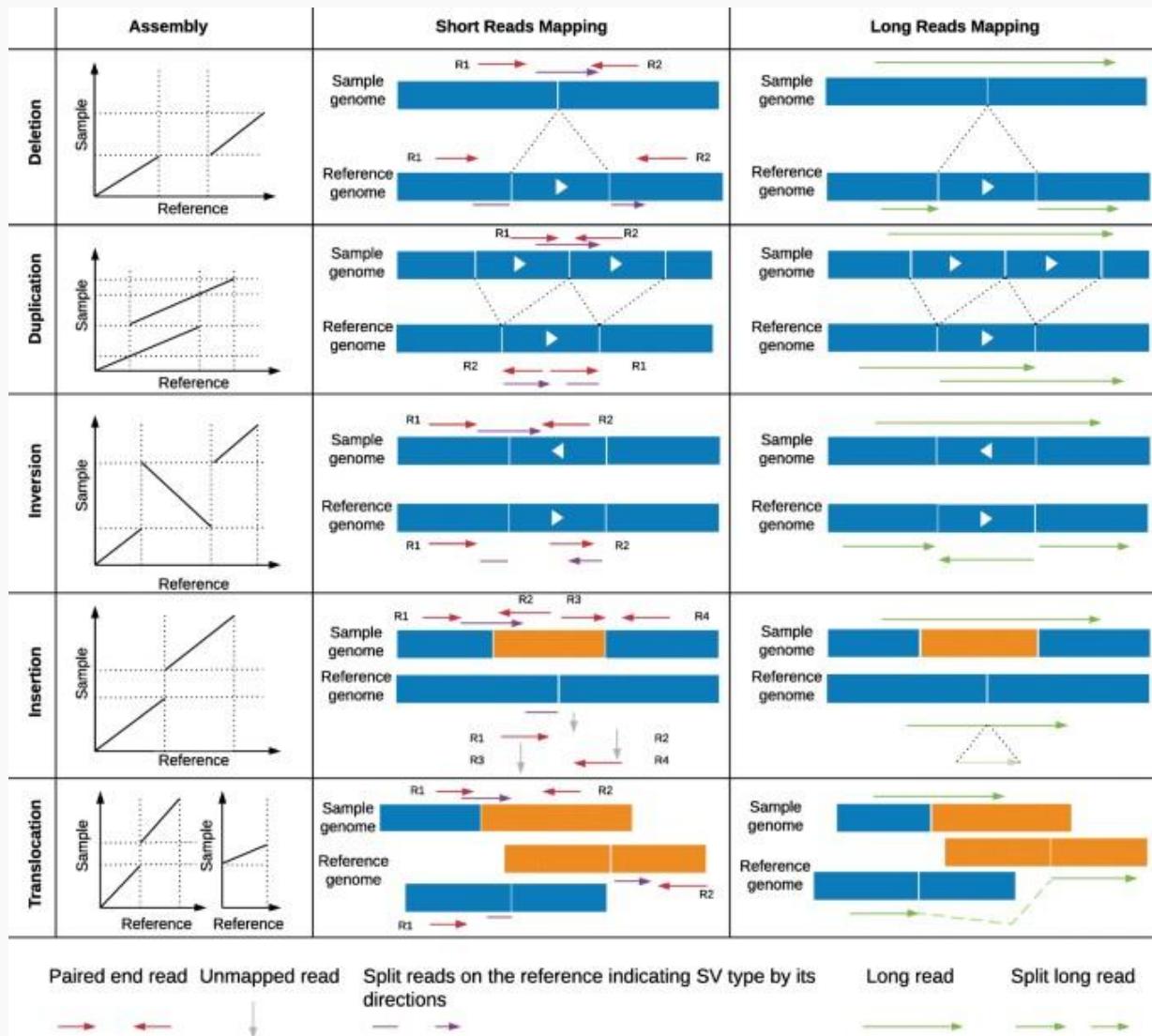
- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

# NGS technologies comparison

NGS platforms					
Platform	Template preparation	Chemistry	Max read length (bases)	Run times (days)	Max Gb per Run
<b>Roche 454</b>	Clonal-emPCR	Pyrosequencing	400‡	0.42	0.40-0.60
<b>GS FLX Titanium</b>	Clonal-emPCR	Pyrosequencing	400‡	0.42	0.035
<b>Illumina MiSeq</b>	Clonal Bridge Amplification	Reversible Dye Terminator	2x300	0.17-2.7	15
<b>Illumina HiSeq</b>	Clonal Bridge Amplification	Reversible Dye Terminator	2x150	0.3-11 <sup>[10]</sup>	1000 <sup>[11]</sup>
<b>Illumina Genome Analyzer IIx</b>	Clonal Bridge Amplification	Reversible Dye Terminator <sup>[12][13]</sup>	2x150	2-14	95
<b>Life Technologies SOLiD4</b>	Clonal-emPCR	Oligonucleotide 8-mer Chained Ligation <sup>[14]</sup>	20-45	4-7	35-50
<b>Life Technologies Ion Proton<sup>[15]</sup></b>	Clonal-emPCR	Native dNTPs, proton detection	200	0.5	100
<b>Complete Genomics</b>	Gridded DNA-nanoballs	Oligonucleotide 9-mer Unchained Ligation <sup>[16][17][18]</sup>	7x10	11	3000
<b>Helicos Biosciences Heliscope</b>	Single Molecule	Reversible Dye Terminator	35‡	8	25
<b>Pacific Biosciences SMRT</b>	Single Molecule	Phospholinked Fluorescent Nucleotides	10,000 (N50); 30,000+ (max) <sup>[19]</sup>	0.08	0.5 <sup>[20]</sup>

From wikipedia website

# Structural Variant Detection





# Practice

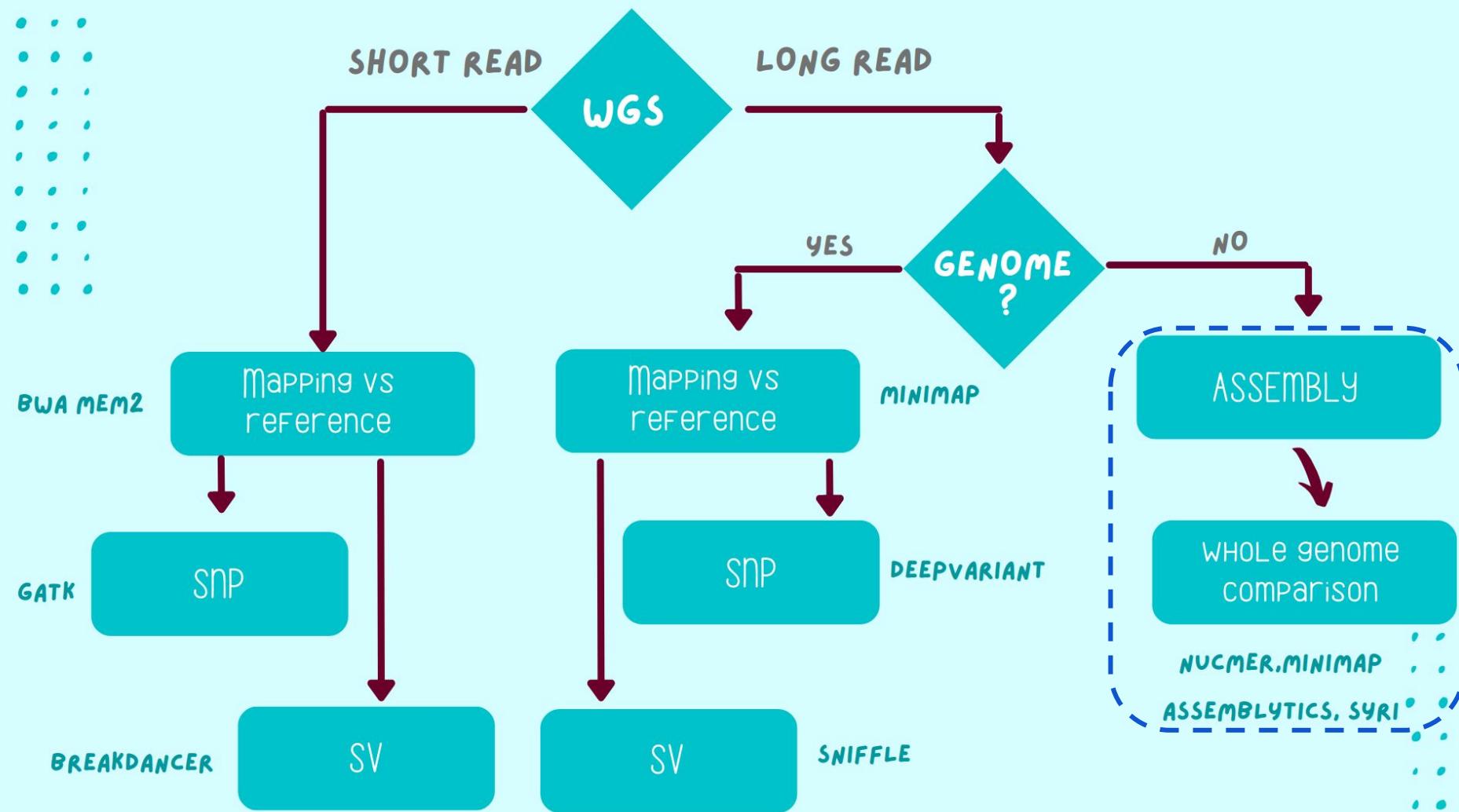
Let's work with the jupyter book :

**Day3\_Variants\_structuraux\_EMPTY.ipynb**

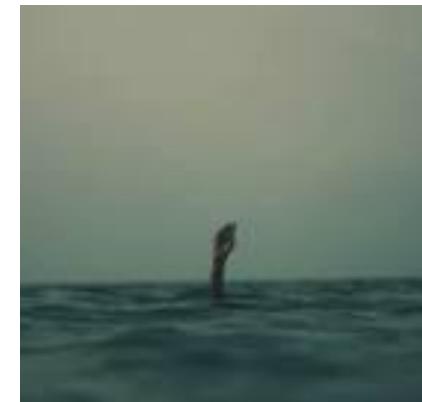


# Training plan - day 4

## SV DETECTION



# If you are lost after these three days of training...

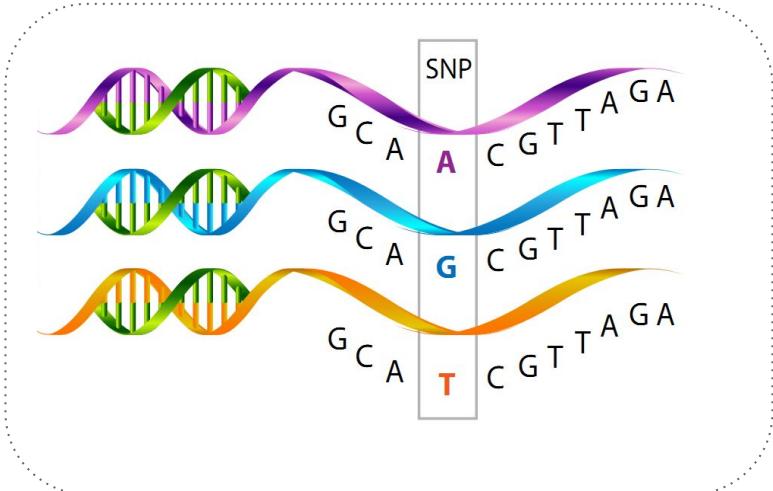


*So let's summarize...*

# Remember why we started...

Mutations & Variations as main source of genetic diversity

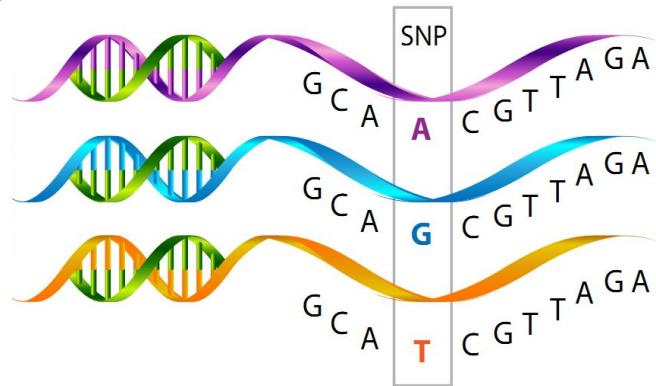
## Single Nucleotide Polymorphism



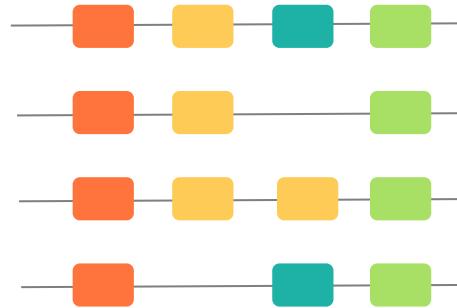
# Remember why we started...

Mutations & Variations as main source of genetic diversity

## Single Nucleotide Polymorphism



## Structural Variations



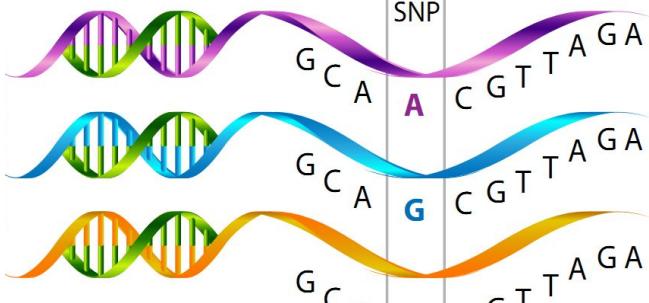
## Presence Absence Variation (PAV)

Deletion, duplication, copy number variation, mobile element insertion

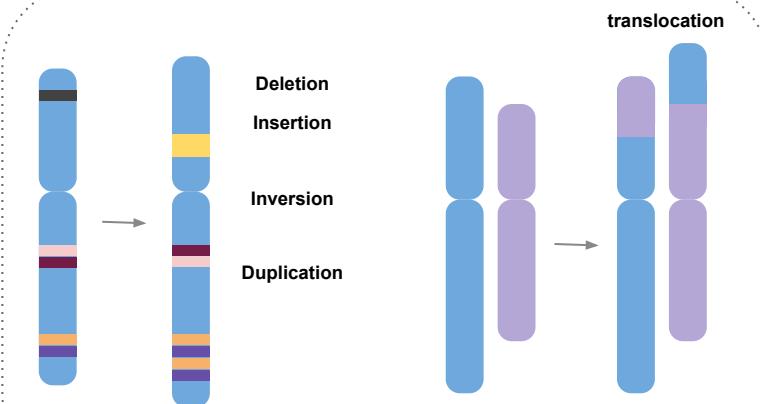
# Remember why we started...

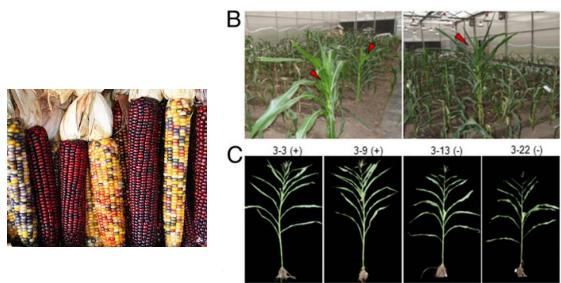
Mutations & Variations as main source of genetic diversity

## Single Nucleotide Polymorphism



## Structural Variations

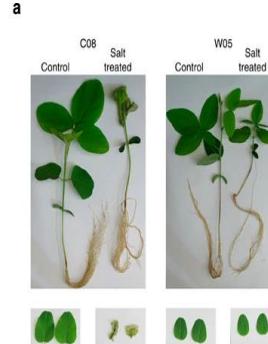




From Yang et al., 2013



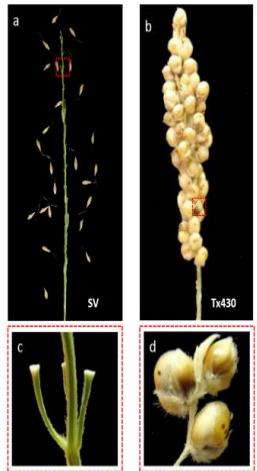
From Li et al. 2012



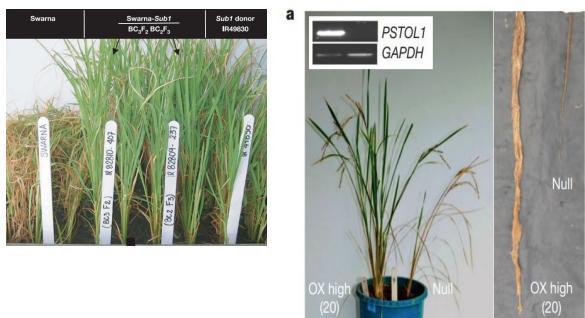
From Qi et al. 2014



From Yang et al., 2014

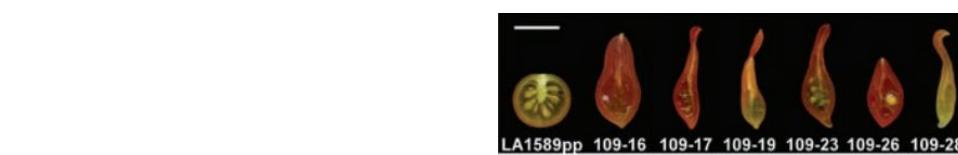


From Lin et al. 2012

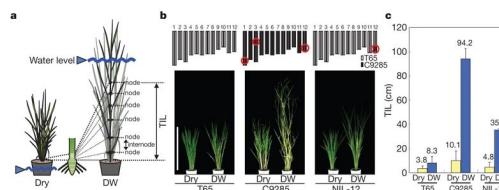


From Xu et al. 2006

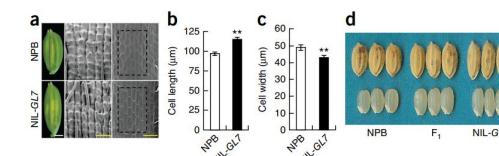
From Gamuyao et al. 2012



From Xiao et al. 2008



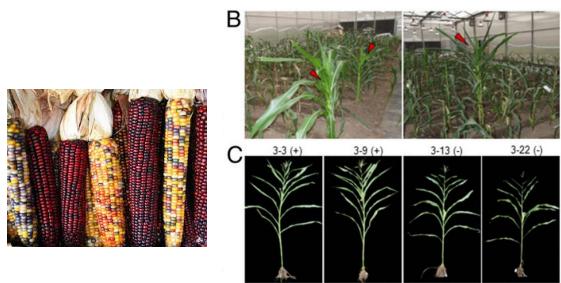
From Hattori et al. 2009



From Wang et al. 2015



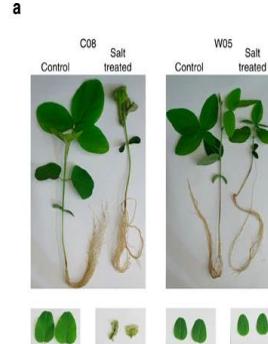
From Bai et al. 2017



From Yang et al., 2013



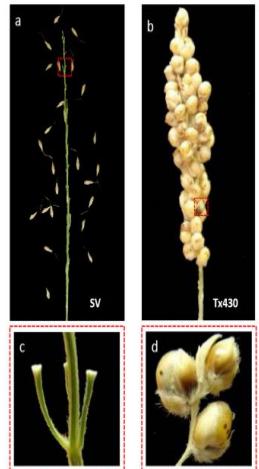
From Li et al. 2012



From Qi et al. 2014

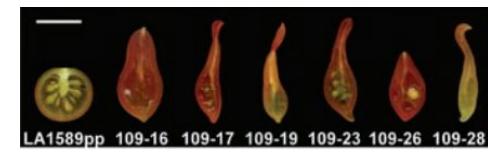


From Yang et al., 2014

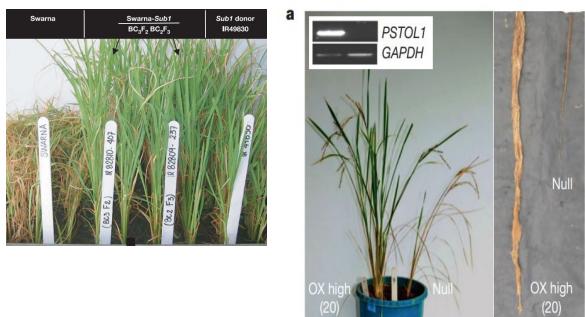


From Lin et al. 2012

## Is One Reference genome enough to capture all genetic diversity ?

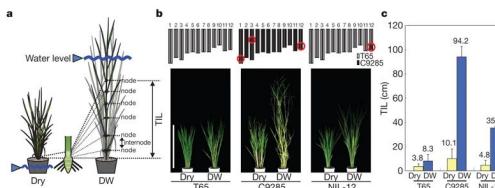


From Xiao et al. 2008

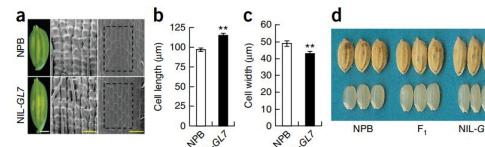


From Xu et al. 2006

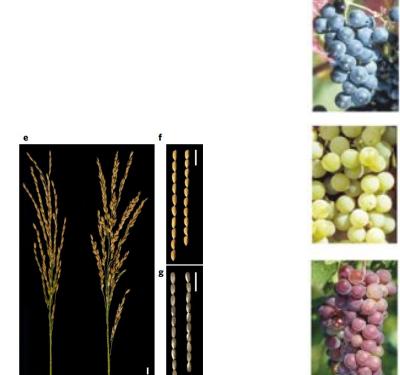
From Gamuyao et al. 2012



From Hattori et al. 2009



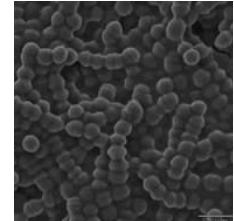
From Wang et al. 2015



From Bai et al. 2017



# Gene number variations within a species



*Streptococcus agalactiae*

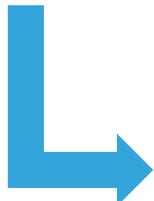


## Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”

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

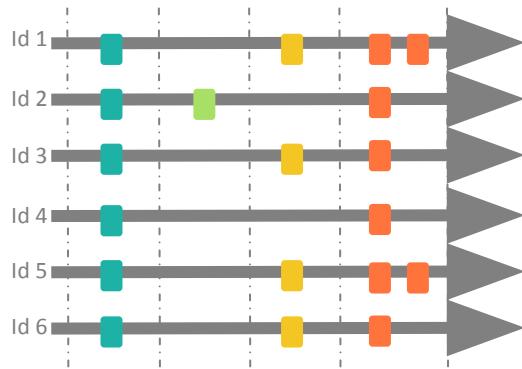
*Tettelin et al., 2005*

- ▶ 8 strains sequenced
- ▶ SNP variations



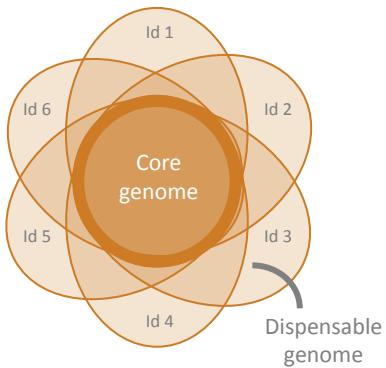
Large number of genes not shared between isolates  
20% genome variability and 80 % shared by all isolates  
*Pangenome concept*

# Pangenome concept

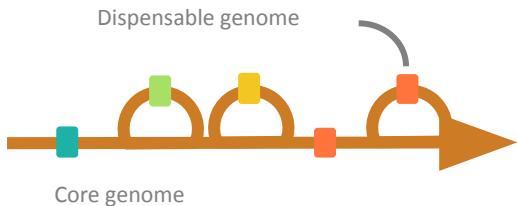


## Pangenome

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



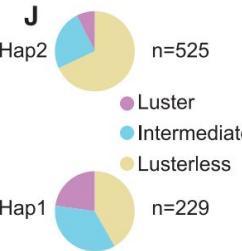
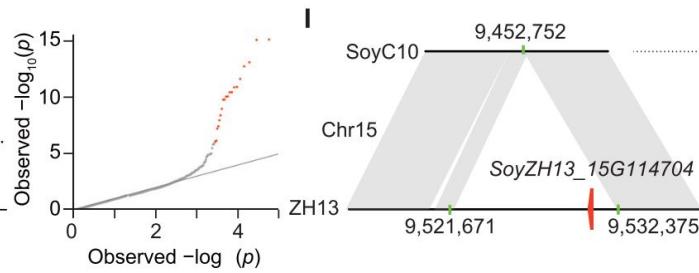
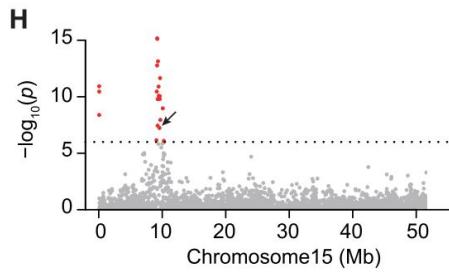
- ▶ **Core genome** : present in all individuals
- ▶ **Disposable genome** : absent from one or several individuals (also called variable, accessory,...)



# What's else ...

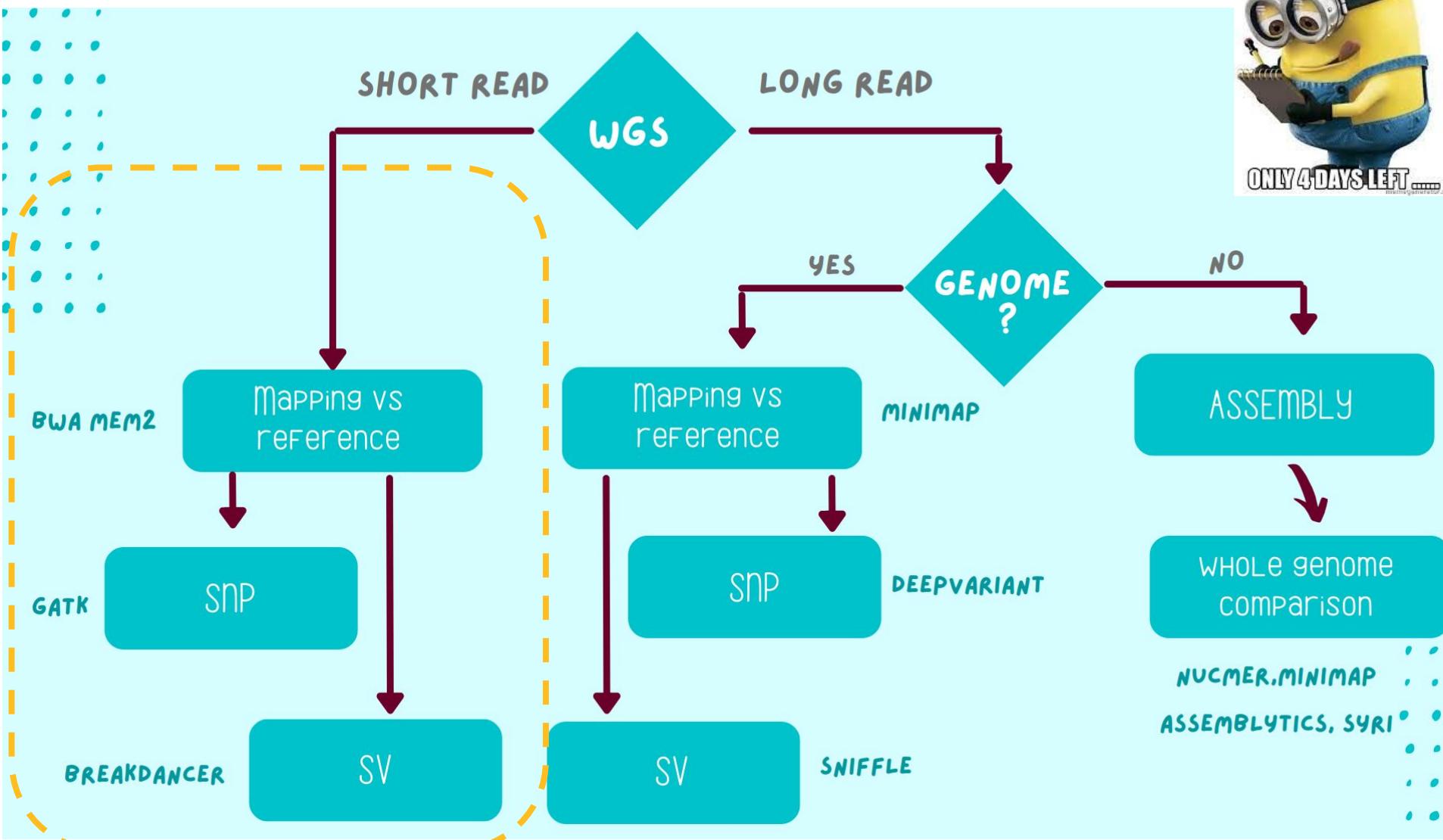


- ▶ 12,150 genes absent from the reference (18 cultivars)

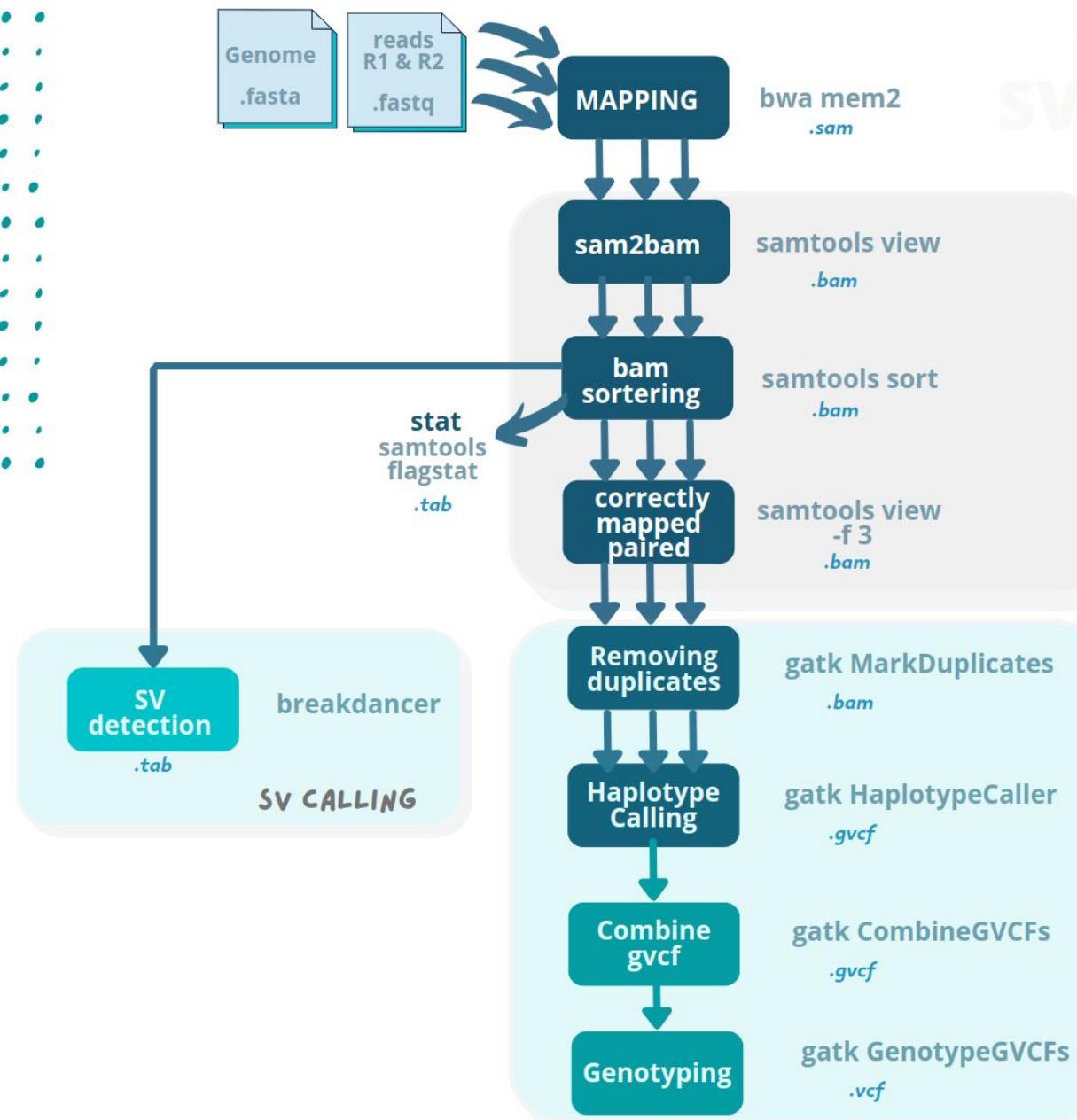


# How to detect SVs and analyse them ?

REMEMBER FOLKS



# From short reads....



**SNP DETECTION  
SHORT READS**

SV

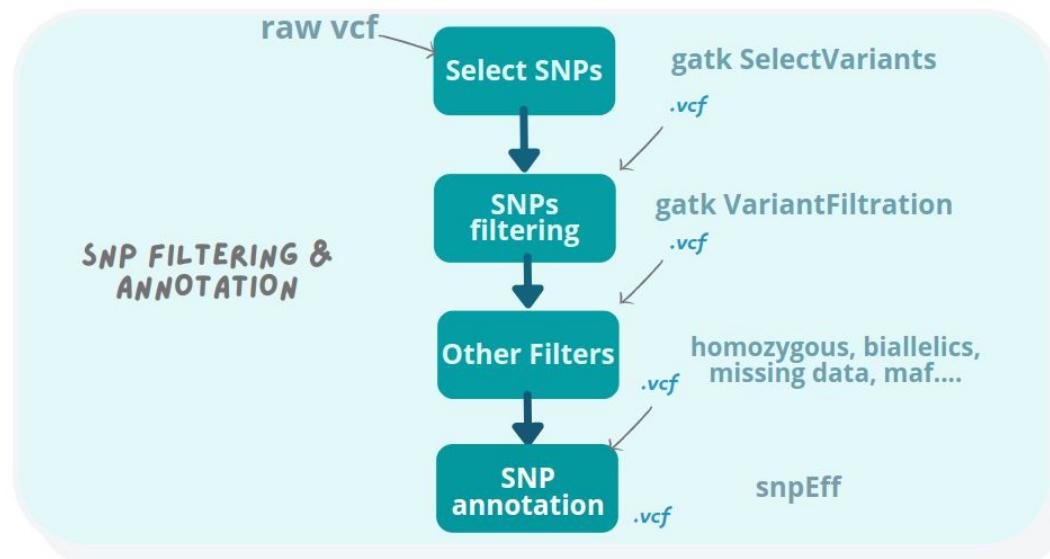
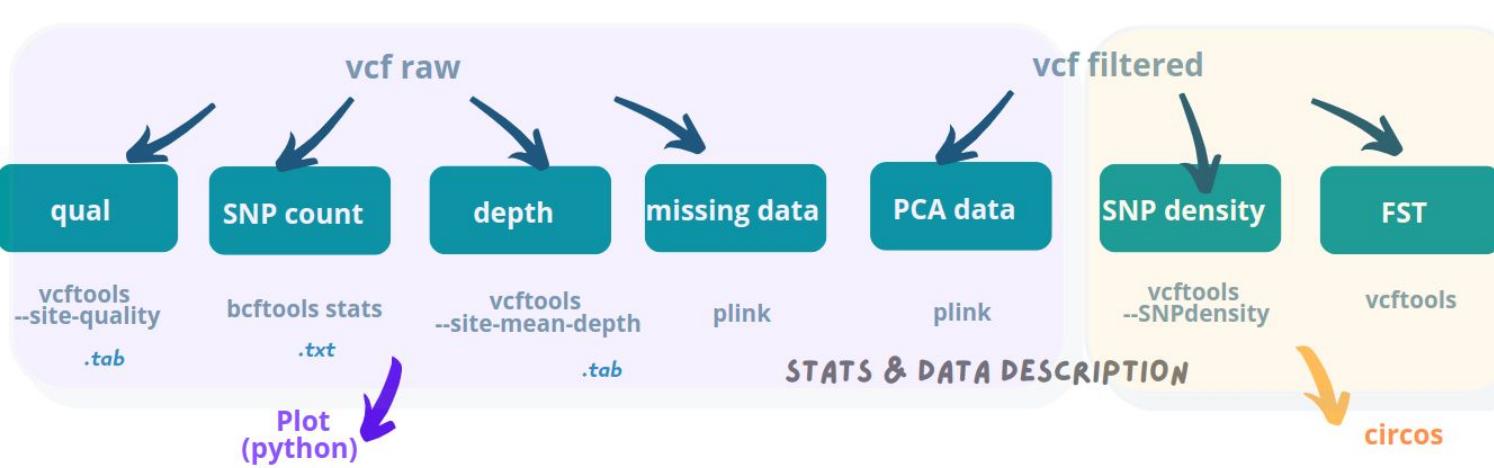


**SAM/BAM  
MANIPULATION**

**SNP CALLING**

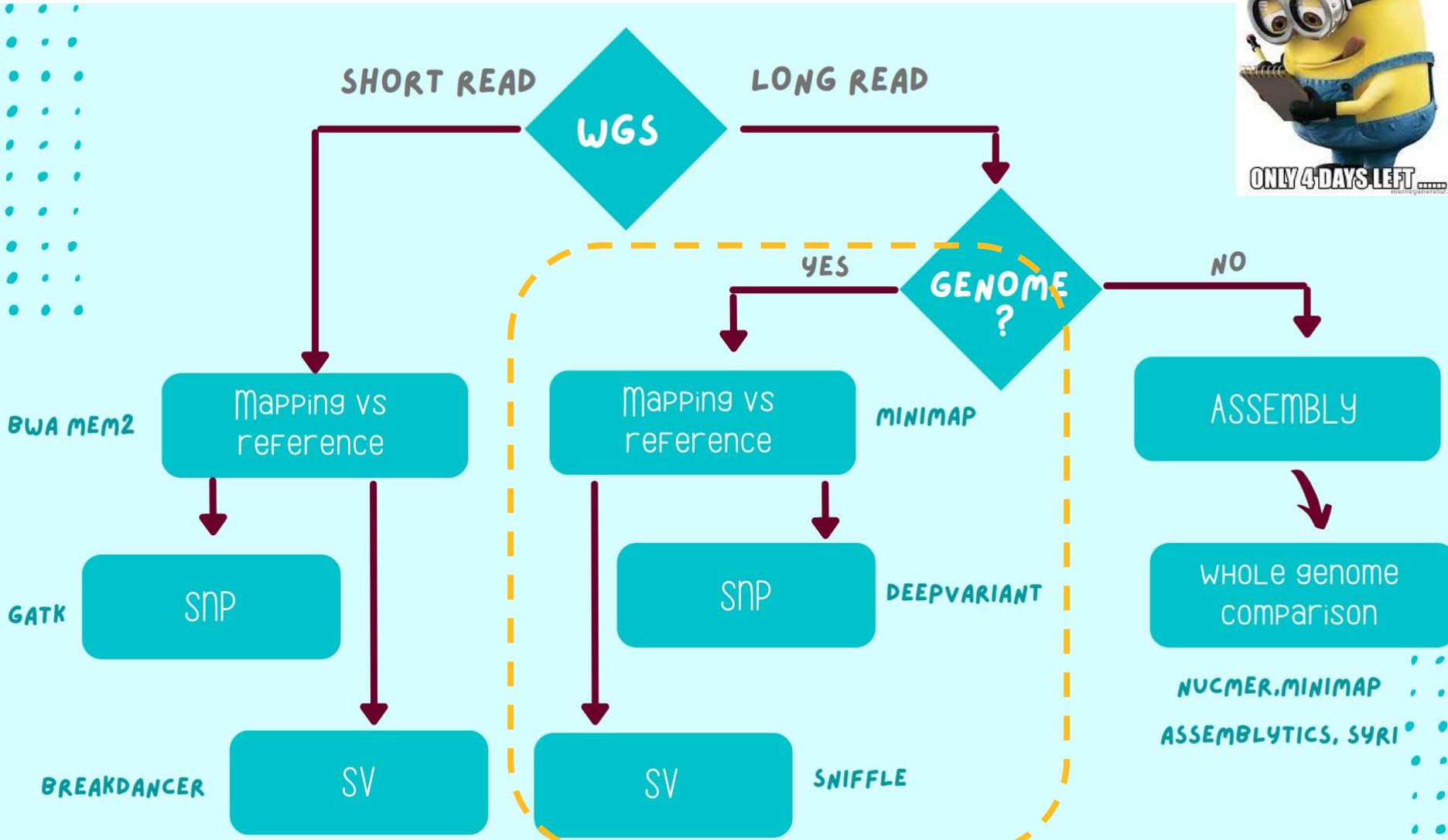
# From short reads....

## SNP ANALYSIS

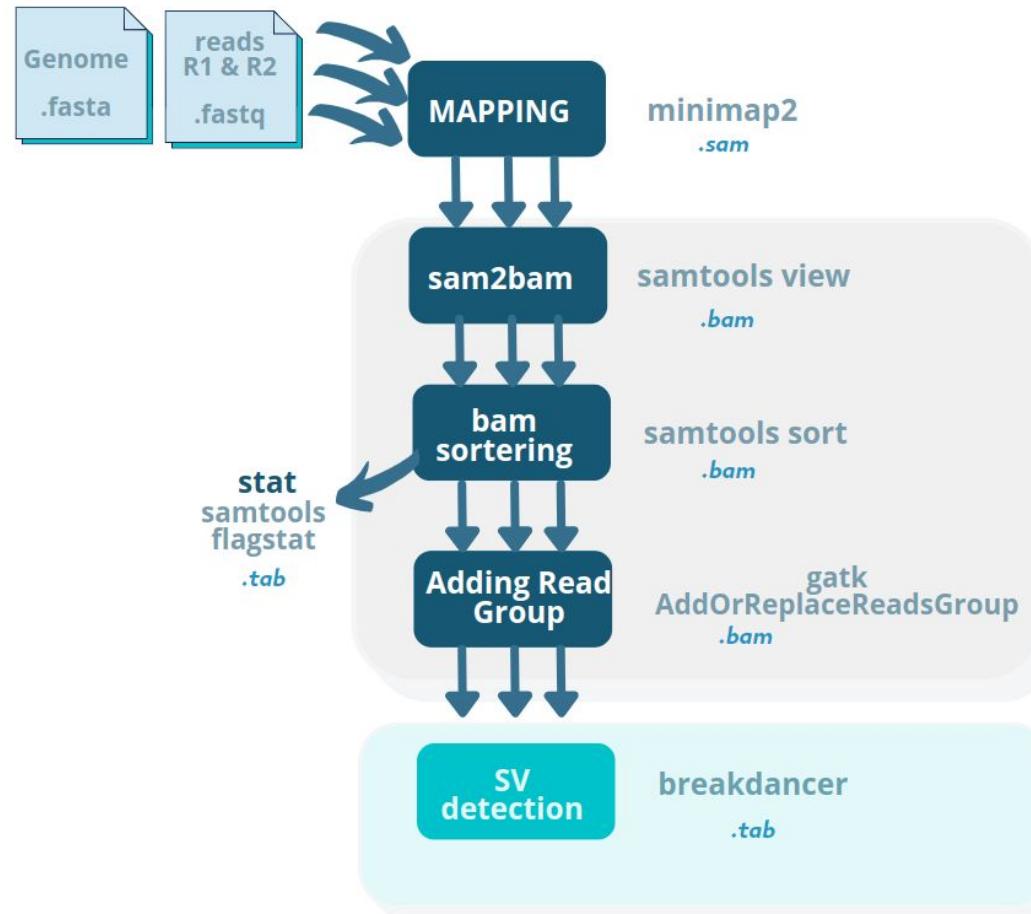


# How to detect SVs and analyse them ?

REMEMBER FOLKS



# From long reads...



SV DETECTION  
LONG READS

REMEMBER FOLKS



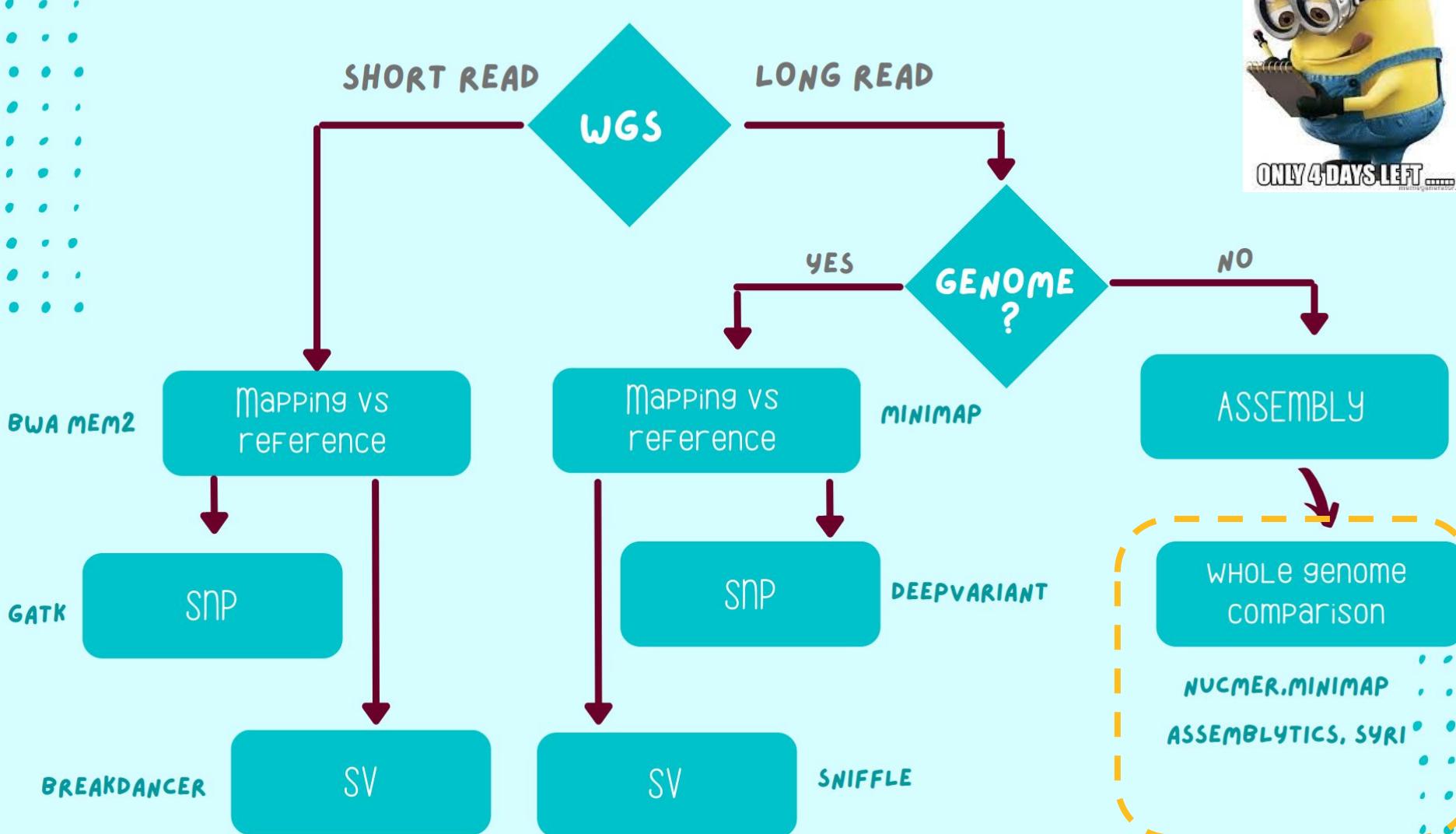
ONLY 4 DAYS LEFT.....

SAM/BAM  
MANIPULATION

SV CALLING

# How to detect SVs and analyse them ?

REMEMBER FOLKS





# Practice

Let's work with the jupyter book :

**Day4\_SV\_genome\_EMPTY.ipynb**



# What data will we use ?

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



## Applications :

- Mapper des reads contre un génome *bwa, minimap2*
- Déetecter des SNPs à partir du mapping de reads - *GATK, deepvariants*
- Analyser les données SNPs brutes (ex: stats, filtres) - *vcftools, bcftools*
- Exemples d'études possibles à partir de SNPs - *SNIPPlay*
- DéTECTer des variants structuraux (SV) à partir de :
  - reads mappées contre un génome - *breakdancer, sniffle*
  - génomes entiers - *nucmer, assemblytics, siry*

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



## Applications :

- Mapper des reads contre un génome *bwa, minimap2*
- Déetecter des SNPs à partir du mapping de reads - *GATK, deepvariants*
- Analyser les données SNPs brutes (ex: stats, filtres) - *vcftools, bcftools*
- Exemples d'études possibles à partir de SNPs - *SNIPPlay*
- DéTECTer des variants structuraux (SV)



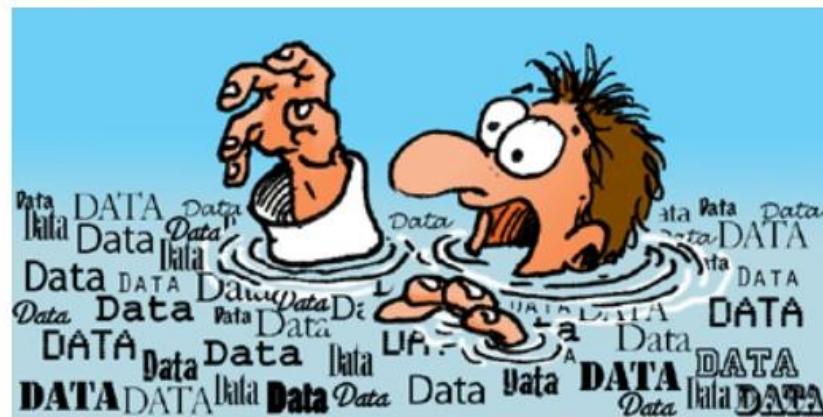
**Avec jupyter book** : lancer les commandes + analyser les résultats  
=> Avoir un plan de bataille opérationnel

## Quelques conseils



- Prototyper votre analyse complète sur quelques individus puis en augmentant le nombre d'individus
- Attention aux paramètres d'étapes clés comme le mapping ou l'alignement
- Analyser les données brutes et définir des filtres
- Persévérance et patience !!!!

# Be Careful to data drowning!



Si vous utilisez les ressources du plateau i-Trop.

Merci de nous citer avec:

“ The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (South Green Platform) at IRD montpellier for providing HPC resources that have contributed to the research results reported within this paper.

URL: <https://bioinfo.ird.fr/> - <http://www.southgreen.fr>”

- Pensez à inclure un budget ressources de calcul dans vos réponses à projets
- Besoin en disques dur, renouvellement de machines etc...
- Devis disponibles
- Contactez [bioinfo@ird.fr](mailto:bioinfo@ird.fr) : aide, définition de besoins, devis...

En informatique,  
la pensée magique ne fonctionne pas !

- Il faut pratiquer ... et ... *restez calme !*
- ... *à vous de jouer !*



Copyright © Randy Glasbergen - www.glasbergen.com



Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International: <http://creativecommons.org/licenses/by-nc-sa/4.0/>



thank you!

**Louis Dennu pour les données du “practice Day 4”**

## Service formation IRD et CIRAD

- Christel Gruau
- Isabelle Lecomte

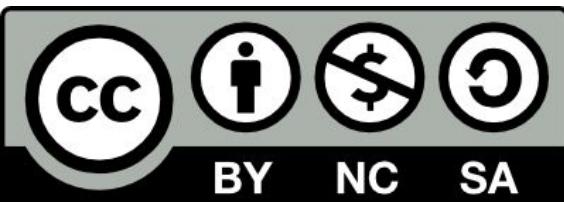
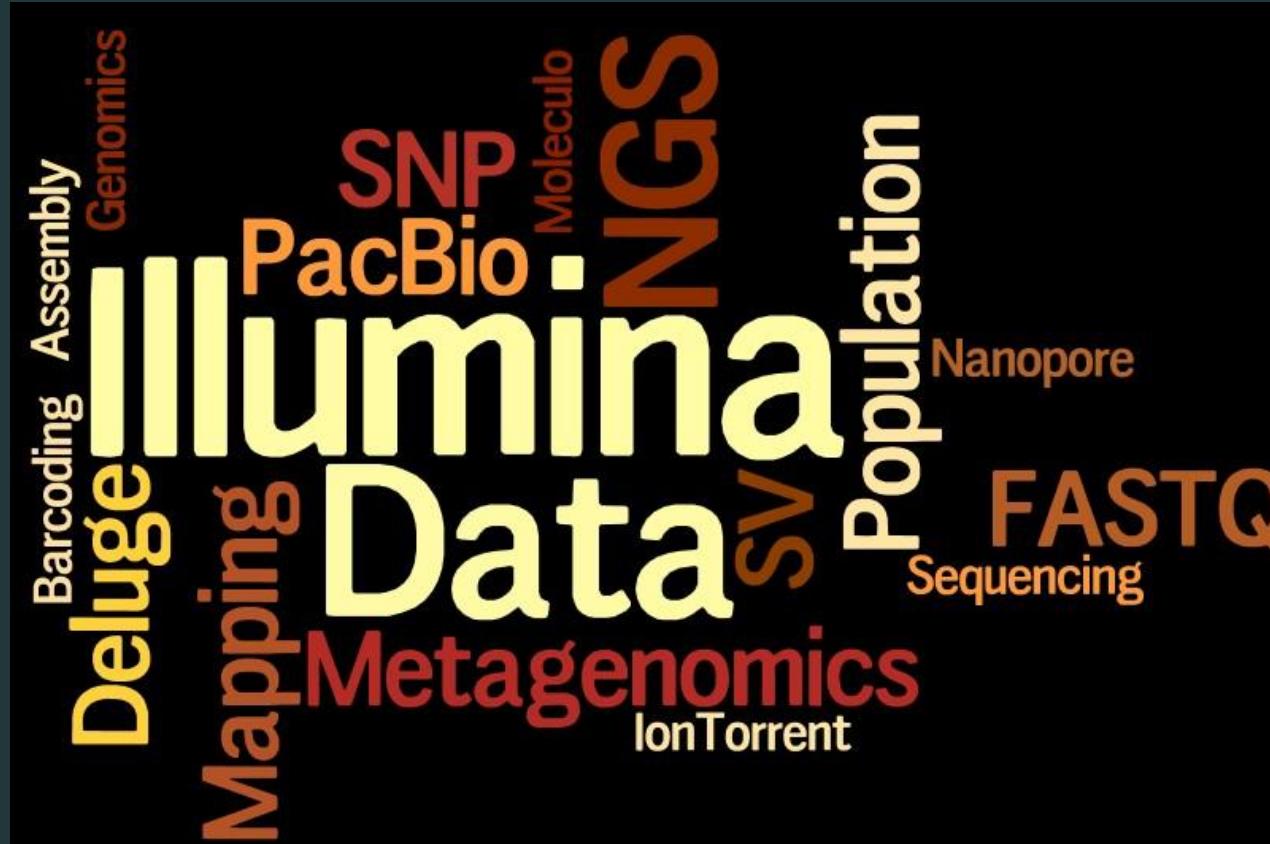


## Core Cloud

- Christophe Blanchet et toute l'équipe biosphère



# Thanks you all, and hoping the teaching was helpful

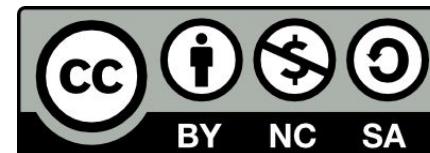
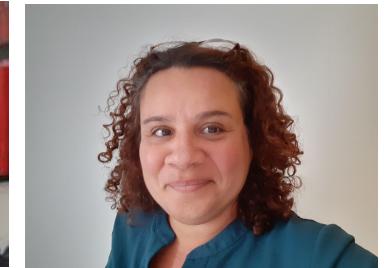


Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:

<http://creativecommons.org/licenses/by-nc-sa/4.0/>

# Formateurs

- **Alexis Dereeper**
- **Julie Orjuela-Bouniol**
- **François Sabot**
- **Christine Tranchant-Dubreuil**





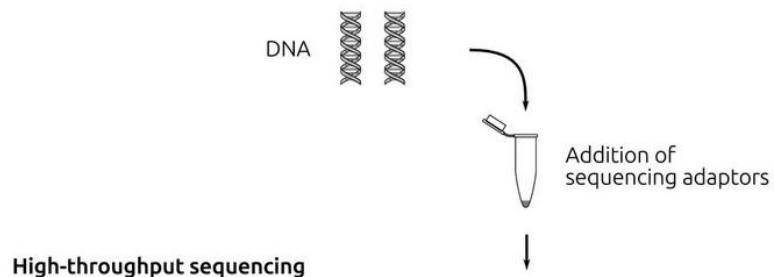
## Variations around genotypes and SNP anthem

---

# Standard approaches

## (a) Whole-genome *de novo* sequencing

### Library preparation

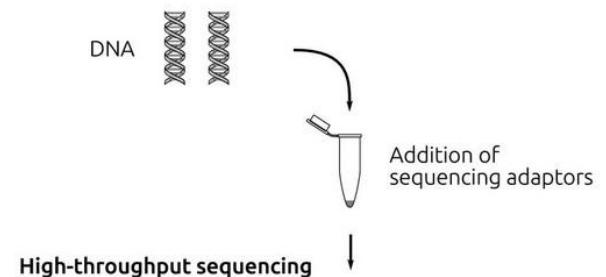


### Read alignment and genome assembly

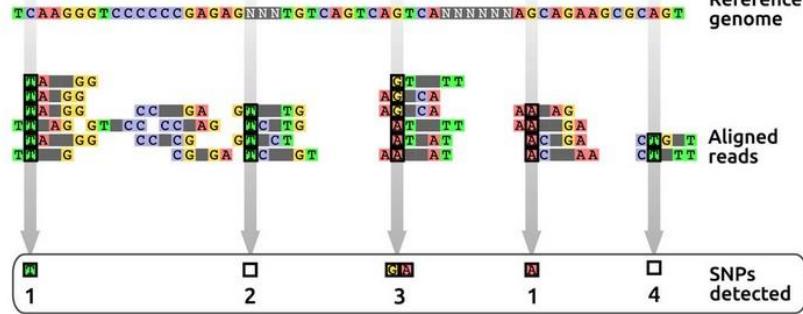
- 1 Short-reads joined into local contigs
  - 2 Contigs ordered and linked into scaffolds with mate-pair reads
  - 3 Consensus sequence obtained from scaffolds and long-reads
- Reference genome
- Aligned reads
- SNPs detected
- Genome sequence
- Scaffold-1      Gap      Scaffold-2      Gap      Scaffold-3

## (b) Whole-genome resequencing

### Library preparation

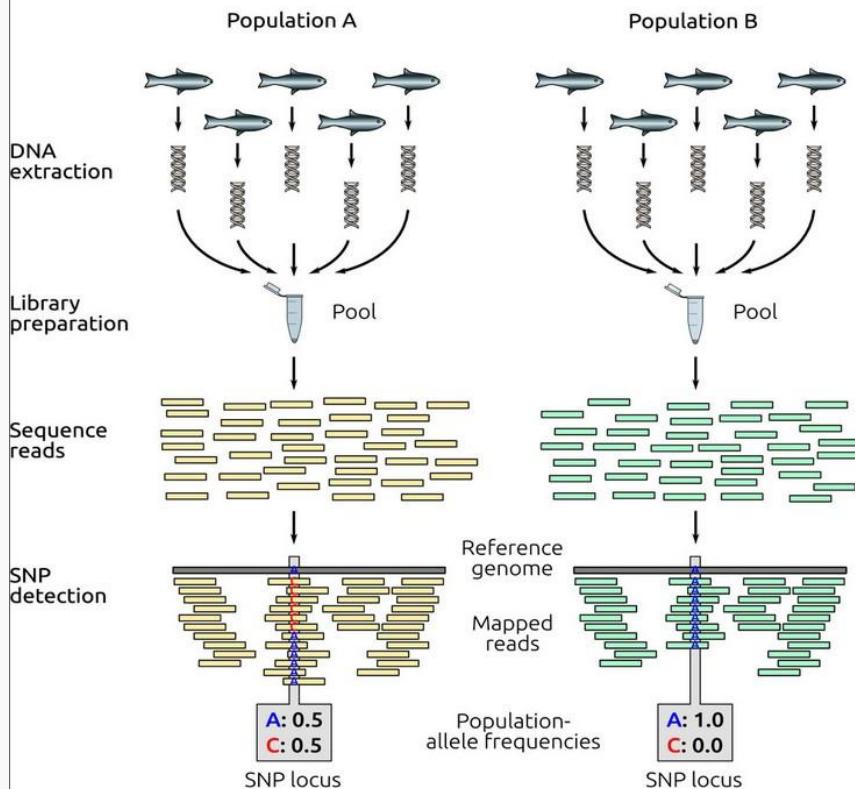


### Read mapping and SNP detection

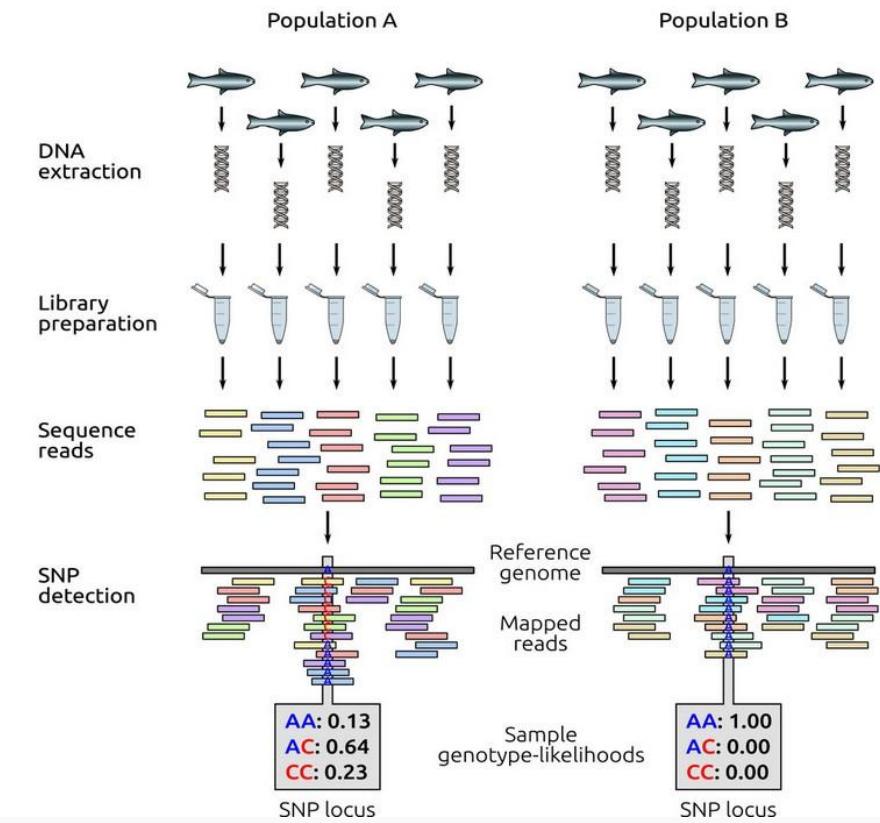


Fuentes-Pardo and Ruzzante,  
2017

**(a) High-coverage whole-genome resequencing of pooled DNA (Pool-seq)**

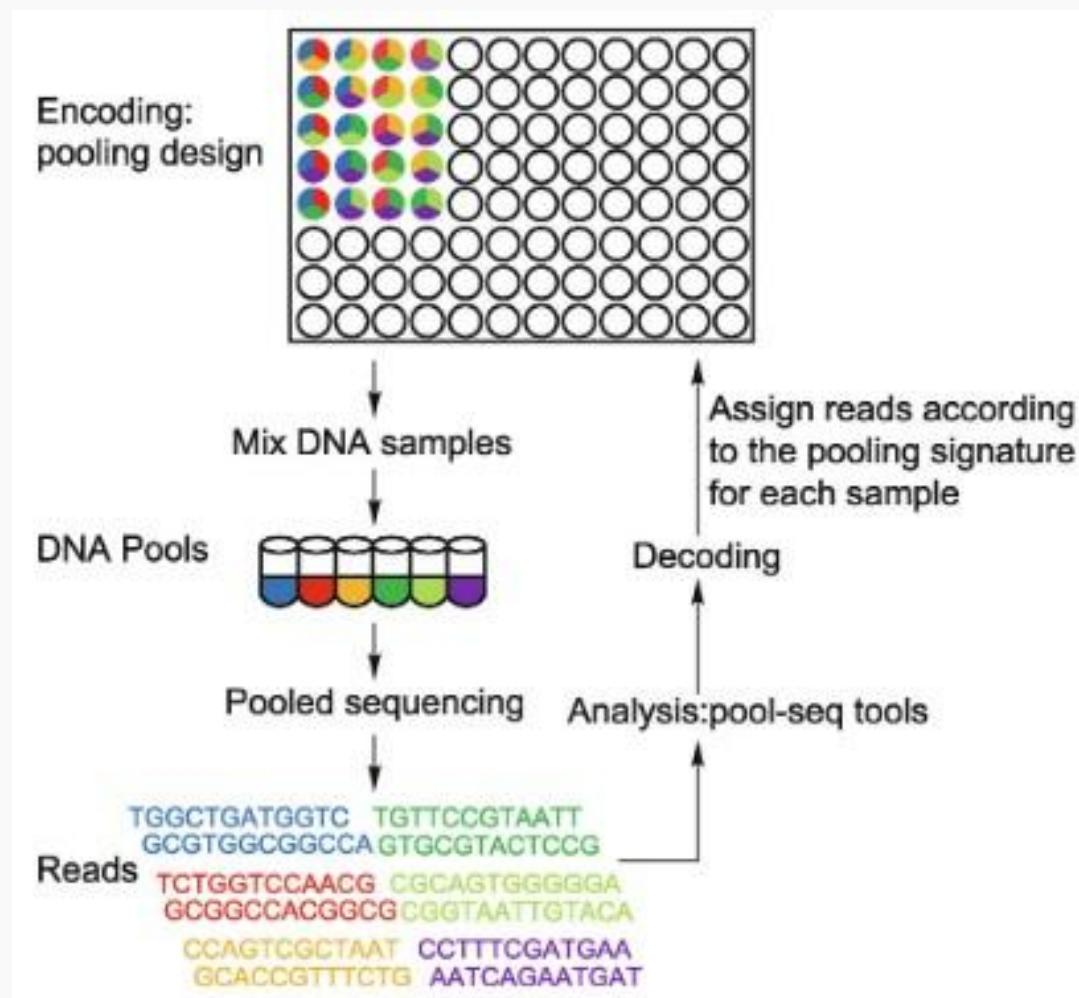


**(b) Low-coverage whole-genome resequencing of individuals from a population (lcWGR)**



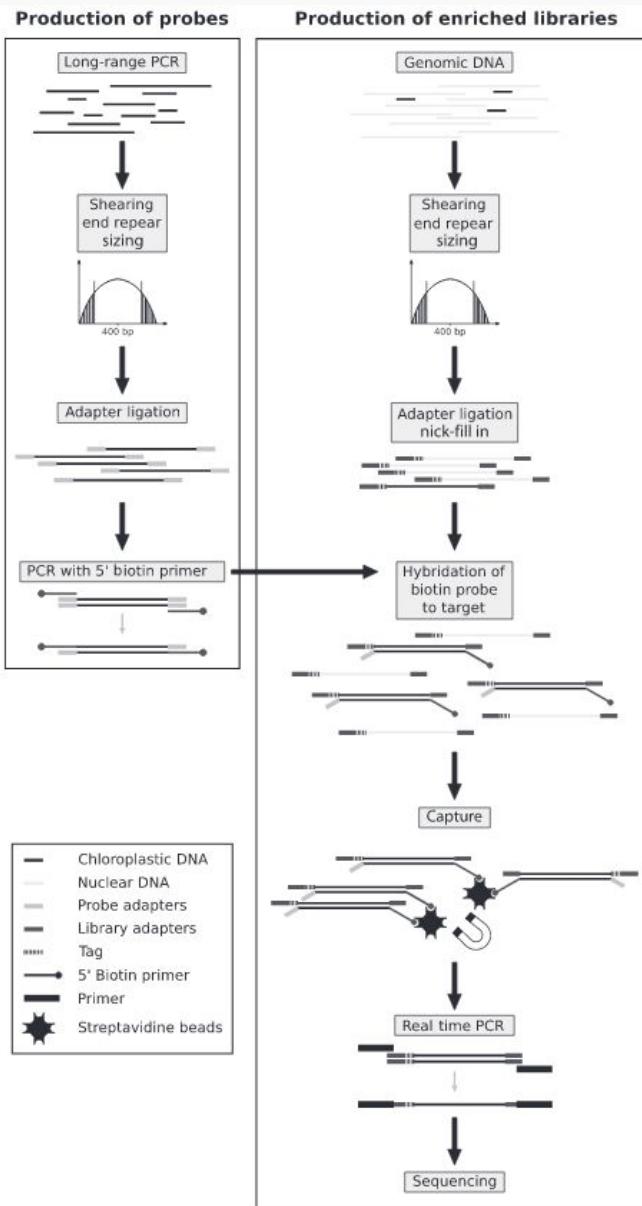
Fuentes-Pardo and Ruzzante,  
2017

# PoolSeq, barcoded

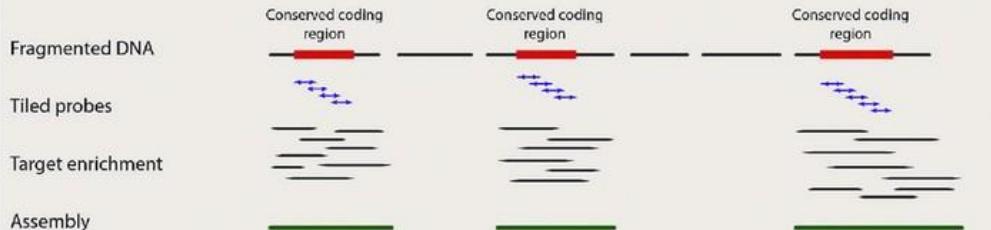


Cao-Sun et al, 2016

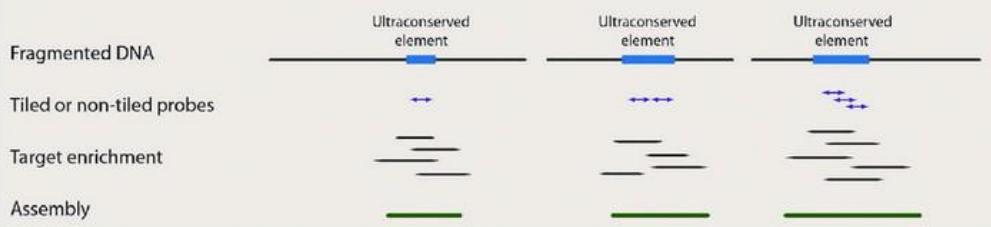
# Phylogenomics using probes



## (A) Anchored hybrid enrichment

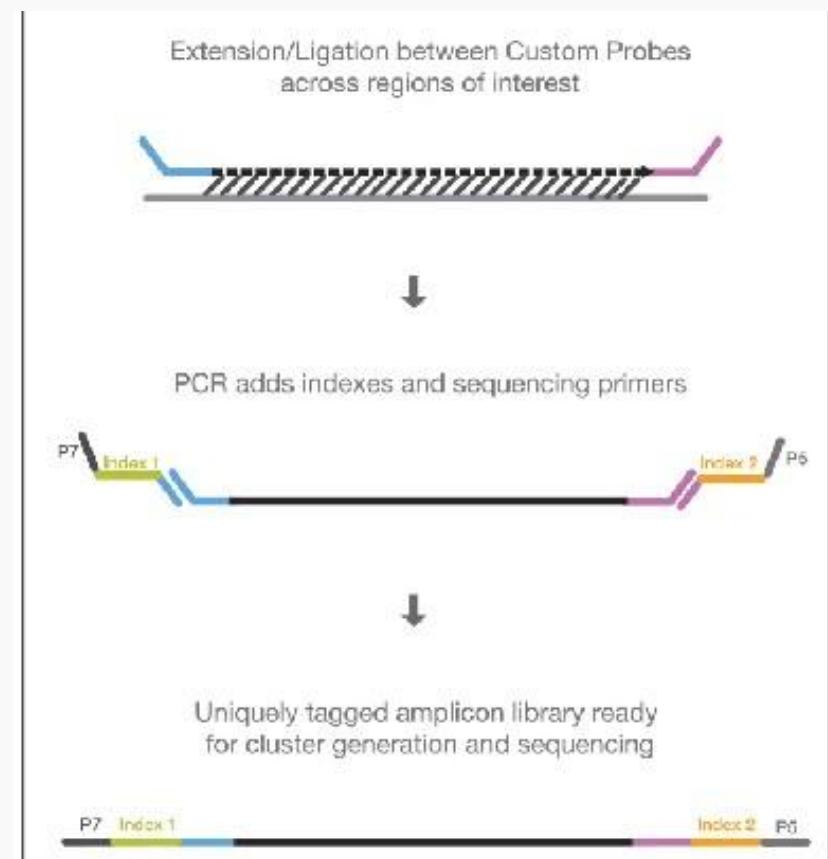
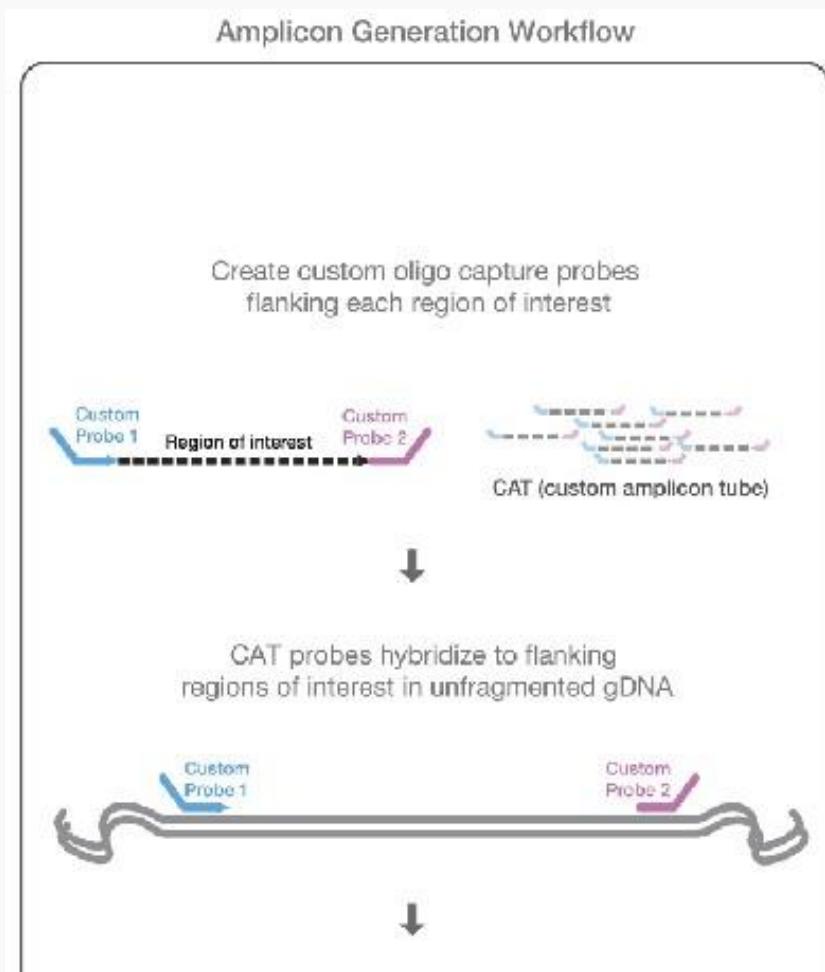


## (B) Ultraconserved elements

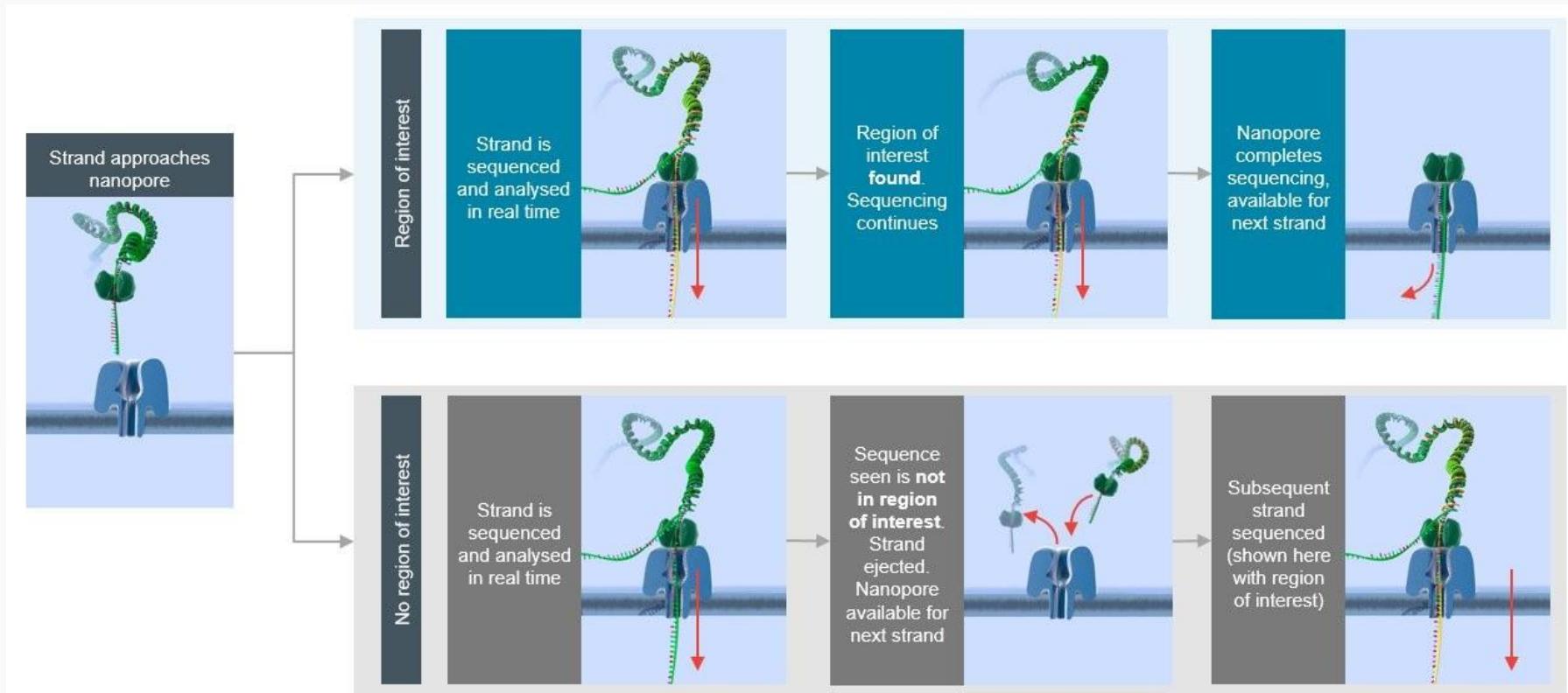


Young and Gillung, 2019

# Phylogenomics using amplicons



# Adaptative sequencing

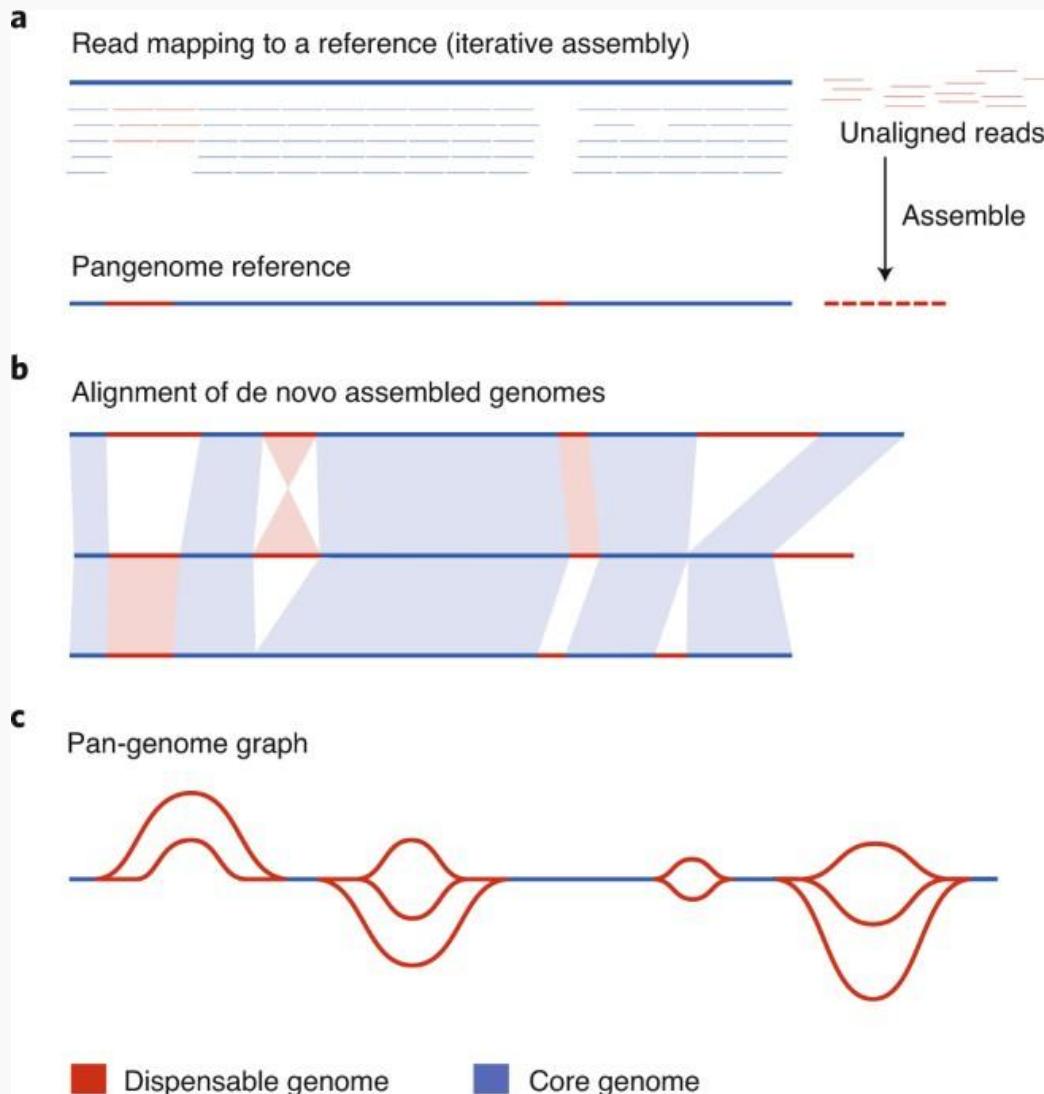


From Nanopore

## Other ways to analyse diversity...

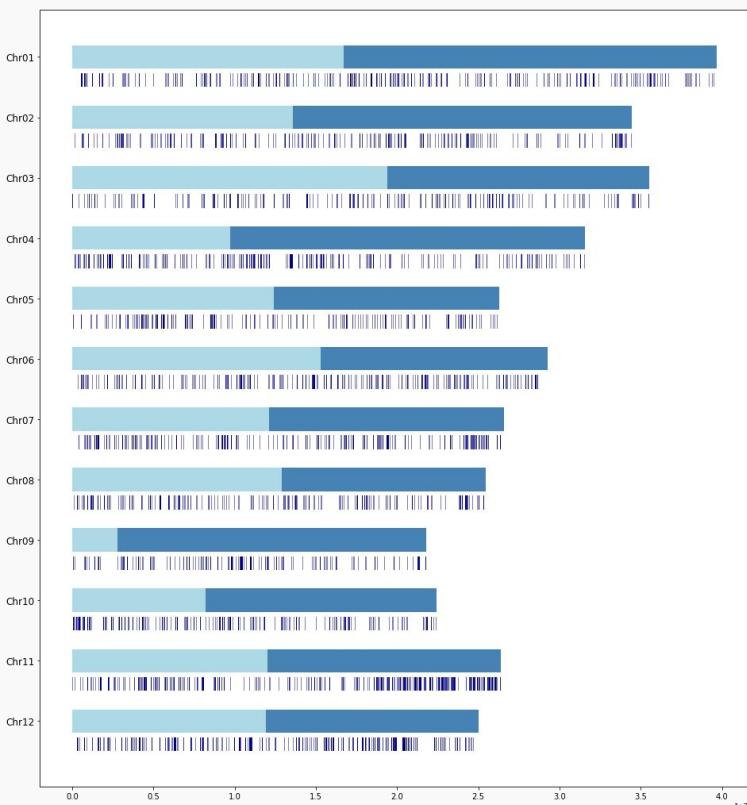
---

# Pangenomics



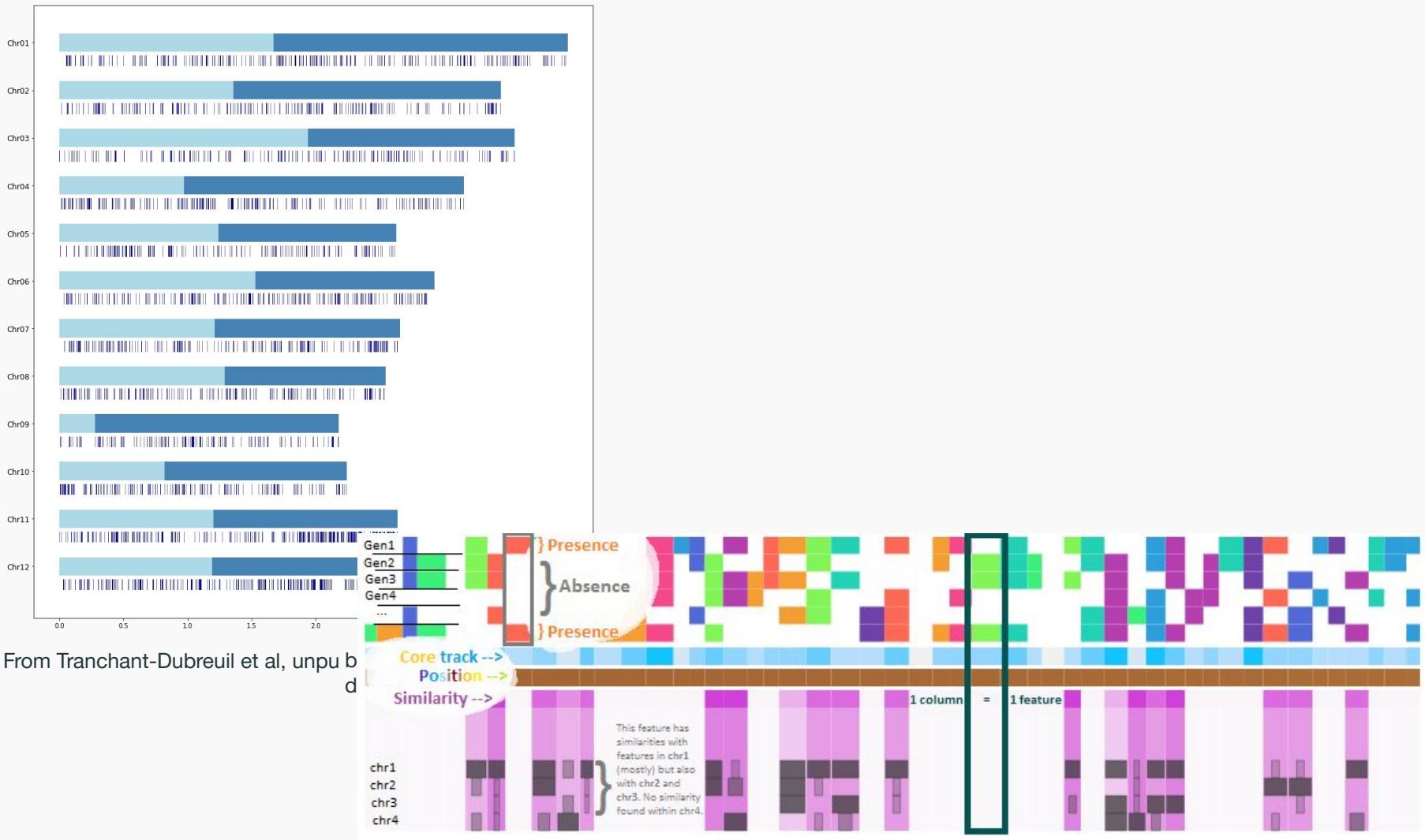
From Bayer et al, 2020

# Pangenomics & visualisation

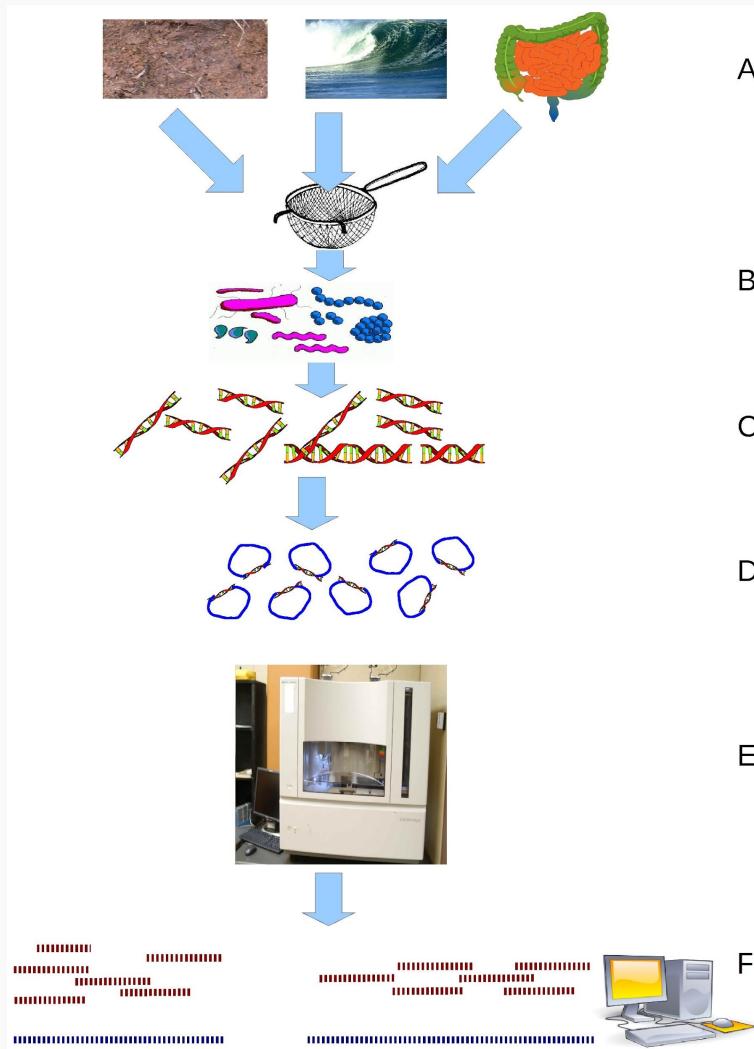


From Tranchant-Dubreuil et al, unpublished

# Pangenomics & visualisation



# Metagenomics



From Wikipedia



From Yau and Cavicchioli, 2011