



# 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**

genome assembly  
**phylogeny**  
comparative genomics transcriptome assembly  
**GWAS**  
population genetics polyplodiy  
**pangenomics**  
metagenomics

SNP detection  
**structural variation**  
differential expression



Rice



Banana



Palm



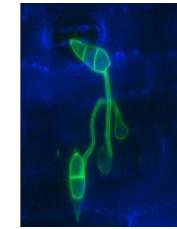
Sorghum



Coffee



Cassava



Magnaporthe

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

# SouthGreen

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 Fredéric  
**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

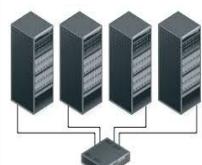
## Workflow manager

**TOGGLE**  
Toolbox for generic NGS analyses



**Galaxy**

## HPC and trainings....



## Genome Hubs & Information System



**Gigwa**

A screenshot of the Gigwa software interface, showing a table of SNP and Indel data with columns for ID, Position, Reference, and Variant.

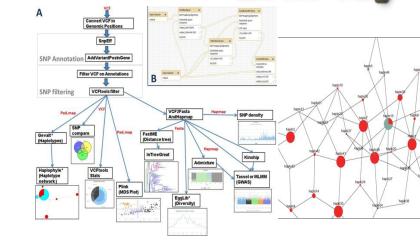
SNPs and Indels

**GreenPhyl**

A screenshot of the GreenPhyl software interface, showing a table of gene family information with columns for Family Id, Family Name, Number of sequences, and Status.

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

**bioinfo@ird.fr**



@ItropBioinfo

IE bioinfo

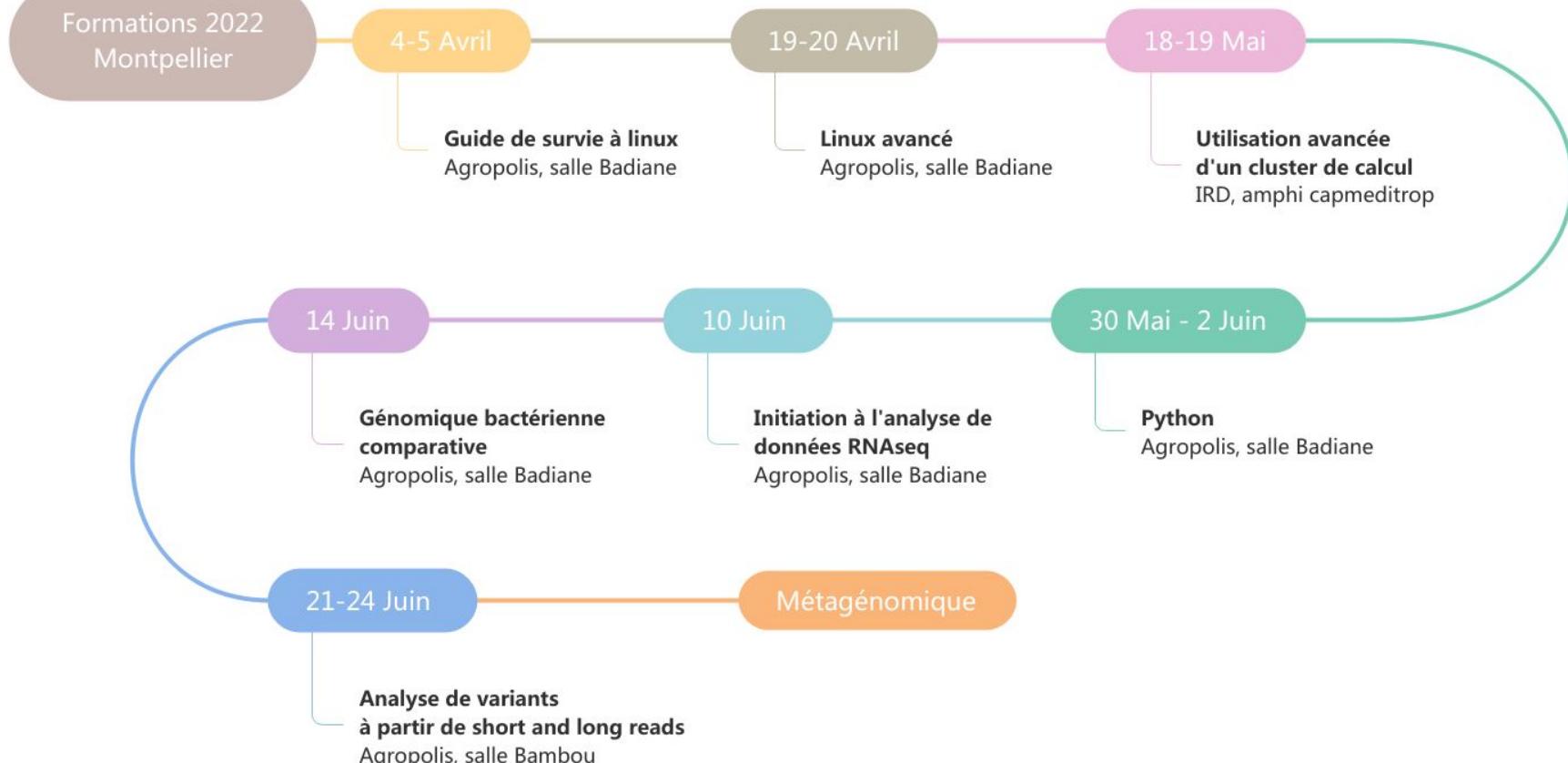
IE bioinfo

IE systèmes  
d'information

IE bioinfo

IE systèmes

IR bioinfo





# 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>



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 - *SNIPlay*
- DéTECTer des variants structuraux (SV) à partir de :
  - reads mappées contre un génome - *breakdancer, sniffle*
  - génomes entiers - *nucmer, assemblytics, siry*

# Objectifs

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



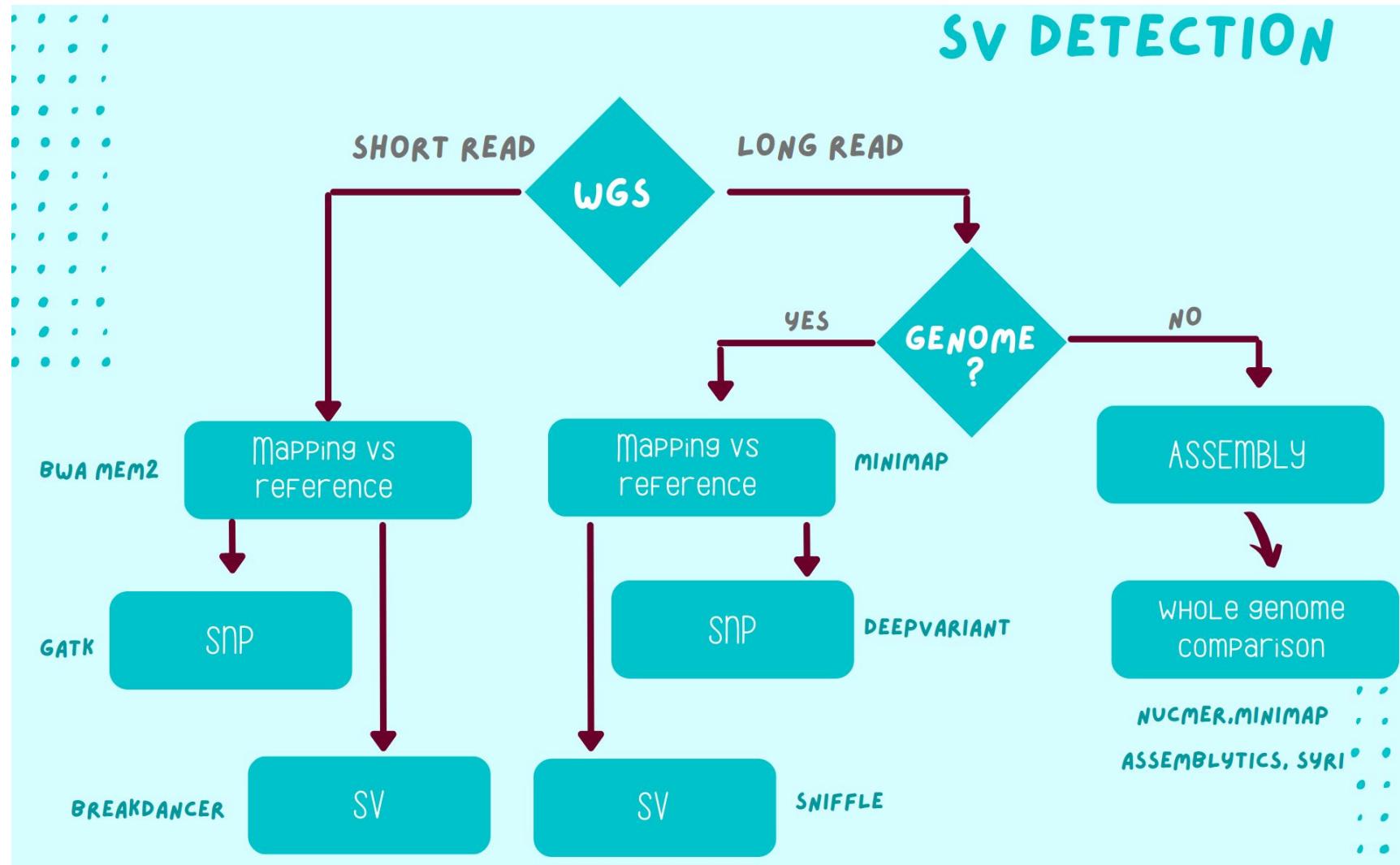
## Applications :

- 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 - *SNIPlay*
- Déetecter 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

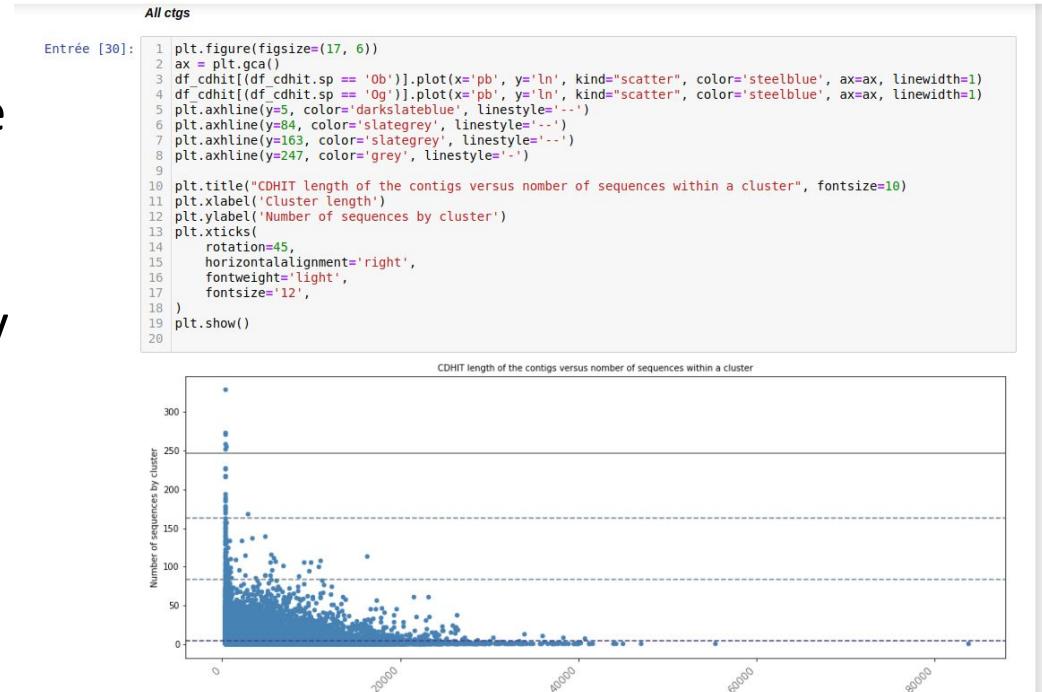
ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala



# Why use jupyter book ?

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

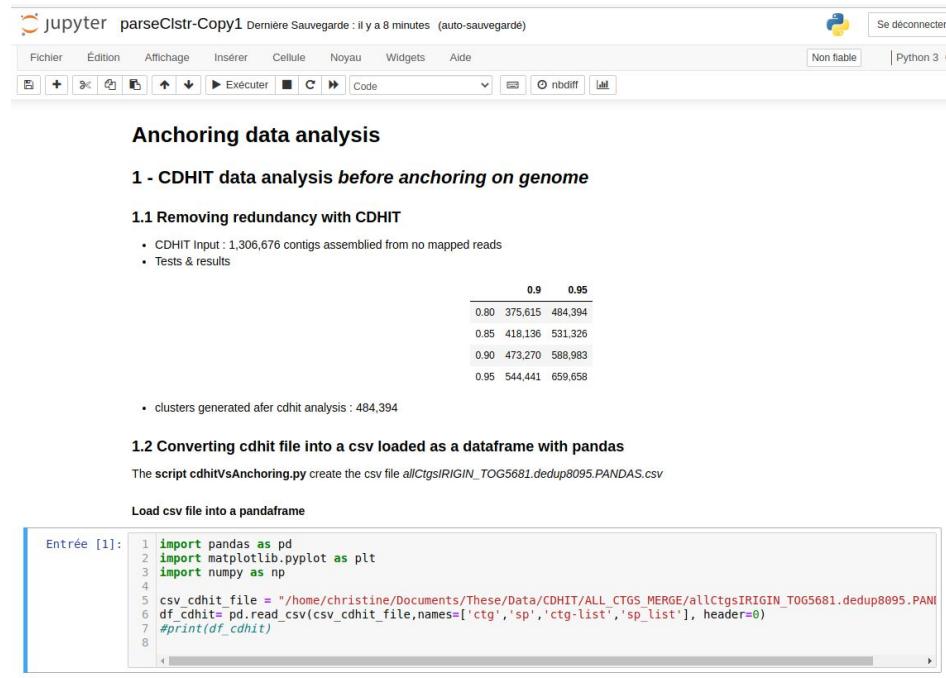
- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



# 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 content:

**Anchor data analysis**

**1 - CDHIT data analysis before anchoring on genome**

**1.1 Removing redundancy with CDHIT**

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

	0.9	0.95
0.80	375.615	484.394
0.85	418.136	531.326
0.90	473.270	588.983
0.95	544.441	659.658

clusters generated after cdhit analysis : 484,394

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

The script `cdhitVsAnchoring.py` creates the csv file `allCtgIRIGIN_TOG5681.dedup8095.PANDAS.csv`

Load csv file into a pandasframe

```
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/allCtgIRIGIN_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
```

# Lab notebook for science data ?



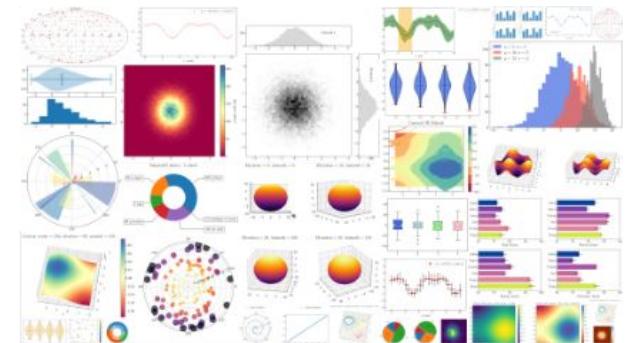
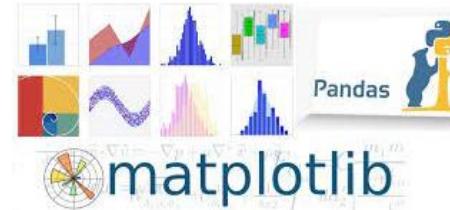
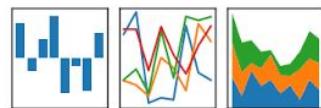
- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

# 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

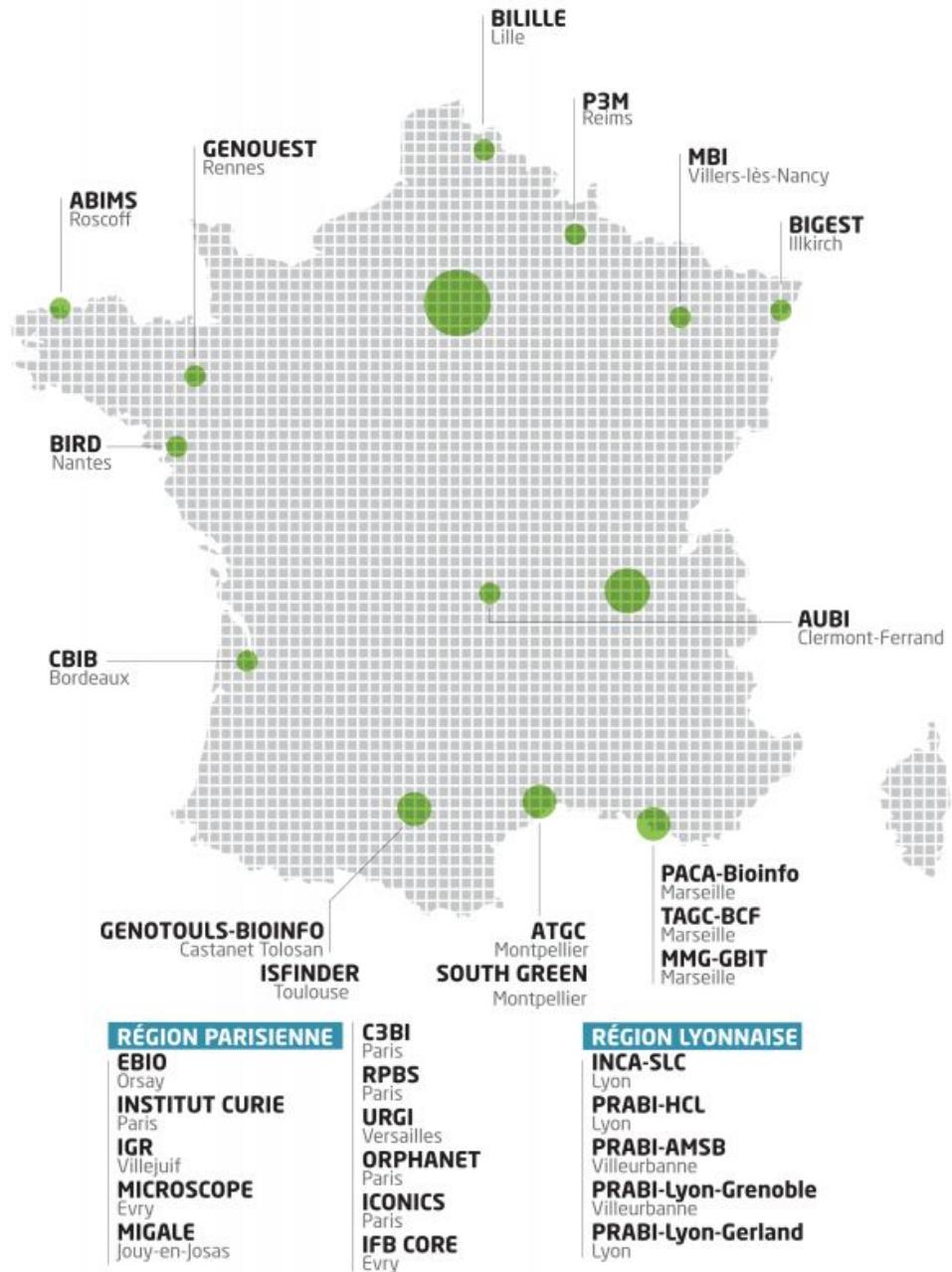


The screenshot shows the IFB Cloud web interface. The top navigation bar includes tabs for "IFB Cloud", "mydatalocal/", and a "+" button. The address bar shows the URL <https://134.158.247.8/tree/mydatalocal>. The main content area features a "jupyter" logo and three tabs: "Files", "Running", and "Clusters". Below these tabs, a message says "Select items to perform actions on them." A file browser shows a single folder named "mydatalocal" with a count of "0" files. To the right, there is a "New" dropdown menu with options like "Upload", "New", "Notebook", "Bash", "Julia 1.5.3", "Python 3", "R", "Text File", "Folder", and "Terminal". A status message at the bottom left says "La liste des notebooks est vide." (The list of notebooks is empty).



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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

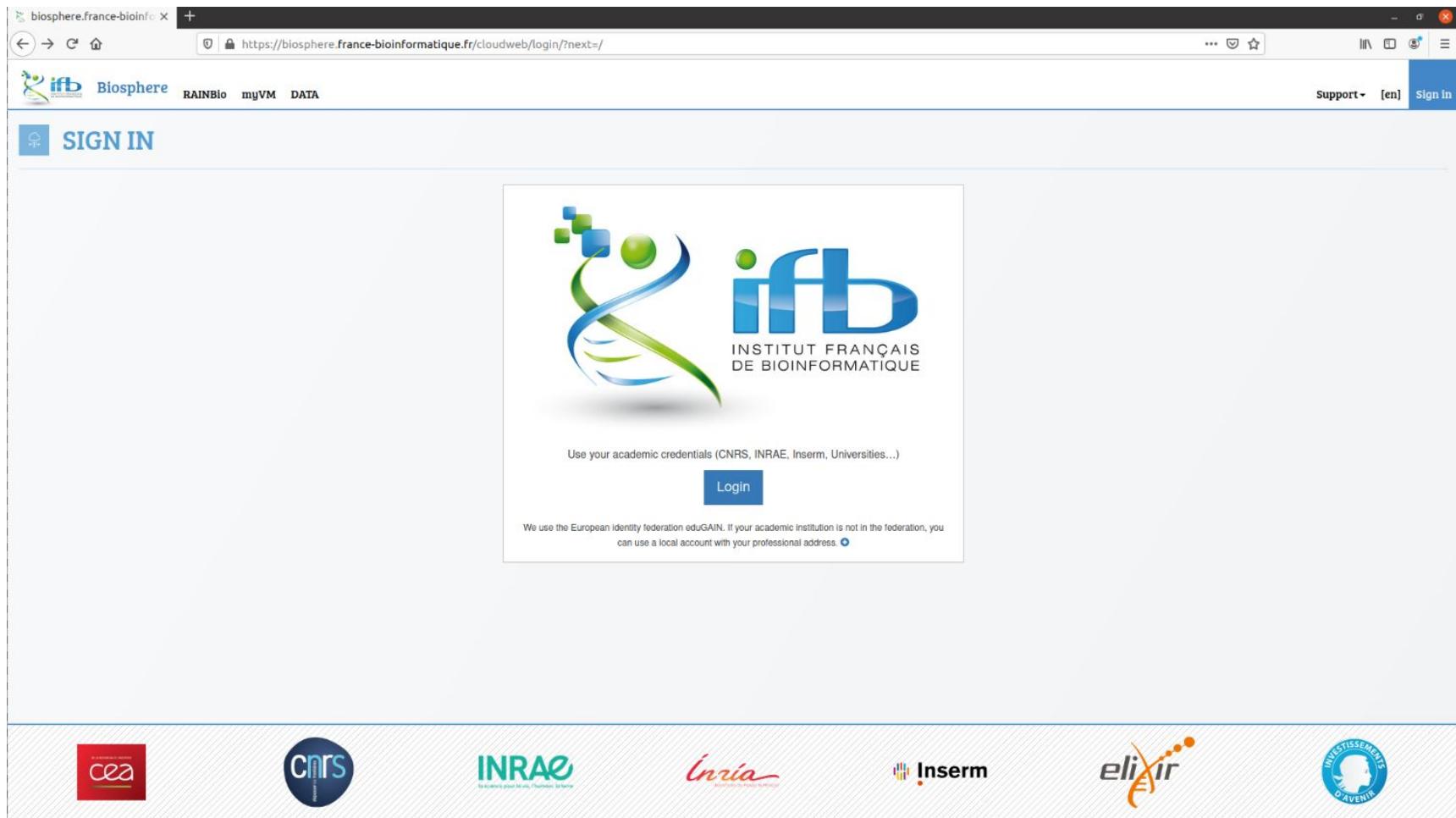


- A federation of clouds, which relies on interconnected IFB's infrastructures, providing distributed services to analyze life science data
- Access to a large set of virtual machines (computing resources, bioinformatics tool)
- Used for scientific production in the life sciences, developments, and also 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 login page at <https://biosphere.france-bioinformatique.fr/cloudweb/login?next=/>. The page features the IFB logo (Institut Français de Bioinformatique) and a "SIGN IN" button. It also includes a note about using academic credentials from institutions like CNRS, INRAE, Inserm, and Universities, and a link to the European identity federation eduGAIN.

bioINFORMATICs biosphere.france-bioinfo X +  
https://biosphere.france-bioinformatique.fr/cloudweb/login?next=/

ifb Biosphere RAINBio myVM DATA Support [en] Sign In

SIGN IN

ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

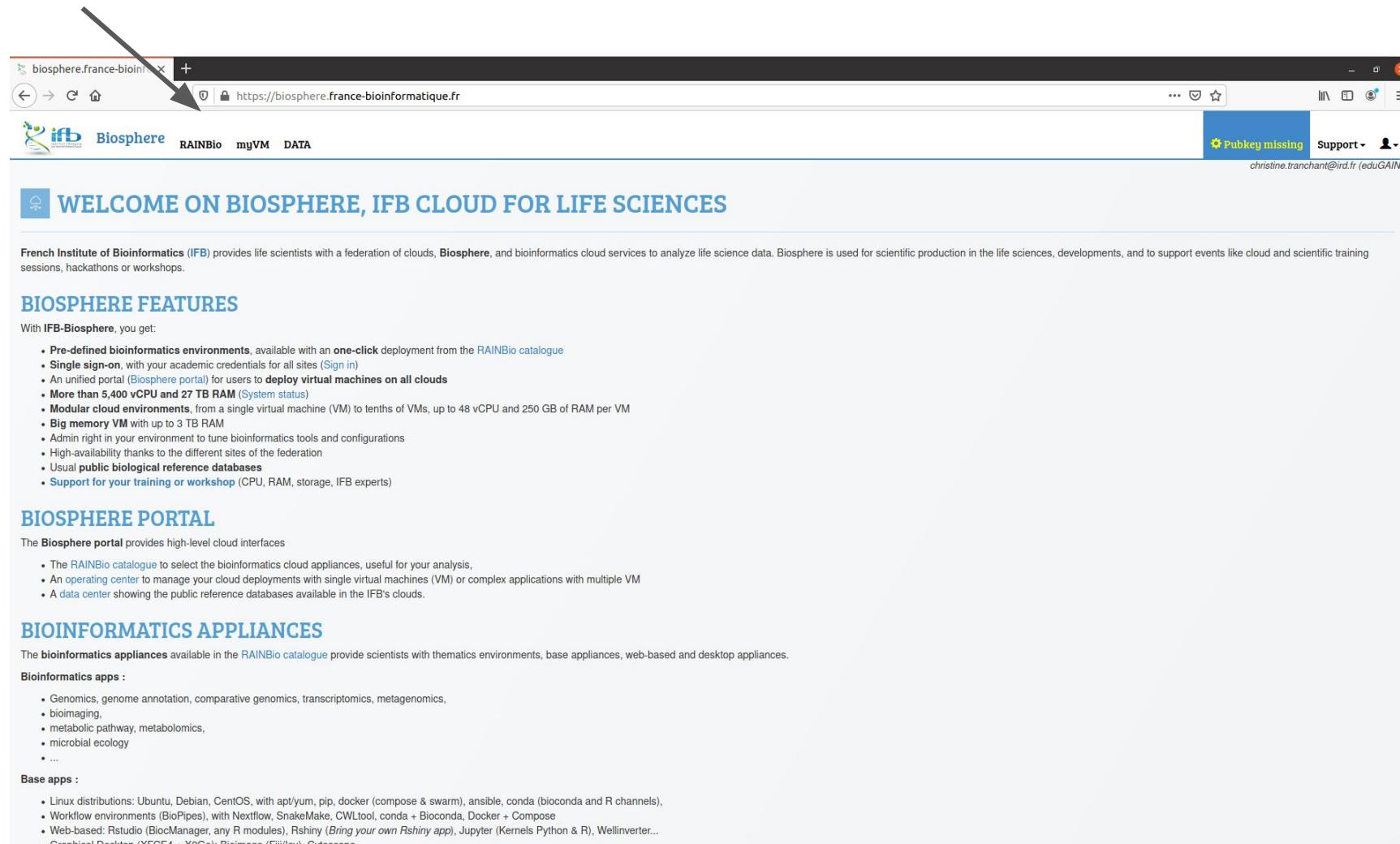
Use your academic credentials (CNRS, INRAE, Inserm, Universities...)

Login

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. [?](#)

cea cnrs INRAE Inria Inserm elixir INVESTISSEMENTS D'AVENIR

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

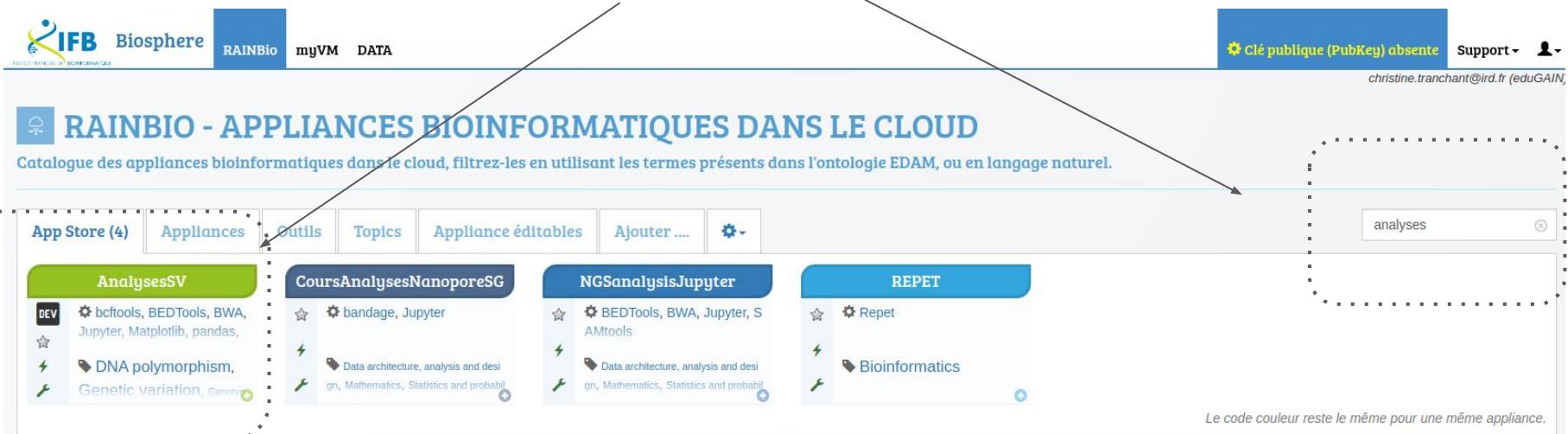


The screenshot shows a web browser window with the following details:

- Address Bar:** biosphere.france-bioinformatique.fr
- Page Title:** WELCOME ON BIOSPHERE, IFB CLOUD FOR LIFE SCIENCES
- Content Summary:** 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.
- Section: 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)
- Section: 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.
- Section: 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,
    - biomimaging,
    - 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/lev), Cytoscape

# Searching for the vm we will use

vm's name : **analysesSV**



The screenshot shows the RAINBIO interface for managing appliances in the cloud. At the top, there are navigation links: IFB Biosphere, RAINBio, myVM, DATA, Clé publique (PubKey) absente, Support, and a user profile. A search bar on the right contains the text "analyses". Below the header, a title reads "RAINBIO - APPLIANCES BIOINFORMATIQUES DANS LE CLOUD" with a subtitle: "Catalogue des appliances bioinformatiques dans le cloud, filtrez-les en utilisant les termes présents dans l'ontologie EDAM, ou en langage naturel." The main content area displays a grid of appliance cards. One card is highlighted in green and labeled "AnalysesSV", which corresponds to the search term. Other visible cards include "CoursAnalysesNanoporeSG", "NGSanalysisJupyter", and "REPET". Each card lists its features and associated technologies. A note at the bottom right states: "Le code couleur reste le même pour une même appliance." (The color code remains the same for the same appliance).



# Let's run your vm through the cloud

**IFB Biosphere** RAINBio myVM DATA

 **Appliance AnalysesSV**  

[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
- Syri
- BreakDancer
- Sniffles
- Mummer

**Contact**

- [Support Cloud IFB](#)

**Developers**

- 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/trainingonvm">https://hub.docker.com/r/francoissabot/trainingonvm</a>
----------	---

 Clé publique (PubKey) absente

christine.tranchant@ird.fr (eduGAIN)

 LANCER  DÉPLOIEMENT AVANCÉ

**Outils**

bcftools | BEDTools | BWA | Jupyter | Matplotlib | pandas | SAMtools

OS	Ubuntu 20.04
Recette de l'app (git)	<a href="https://github.com/SouthGreenPlatform/training_SV_VM">https://github.com/SouthGreenPlatform/training_SV_VM</a>
App de base	Jupyter

**Caractéristiques**

Nom long	Analyses des variants structuraux en short reads, long reads et assemblage
Version	1.0
Créée	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 Biosphere** RAINBio myVM DATA

Clé publique (PubKey) absente 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 environment 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 in the dropdown menu. 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/trainingonvm>

### Configurer le déploiement d'une appliance

Déploiement de l'appliance "AnalysesSV"

Name: CTranchant

Groupe à utiliser: DIADE (Diversité, Adaptation)

Cloud: ifb-core-cloudbis

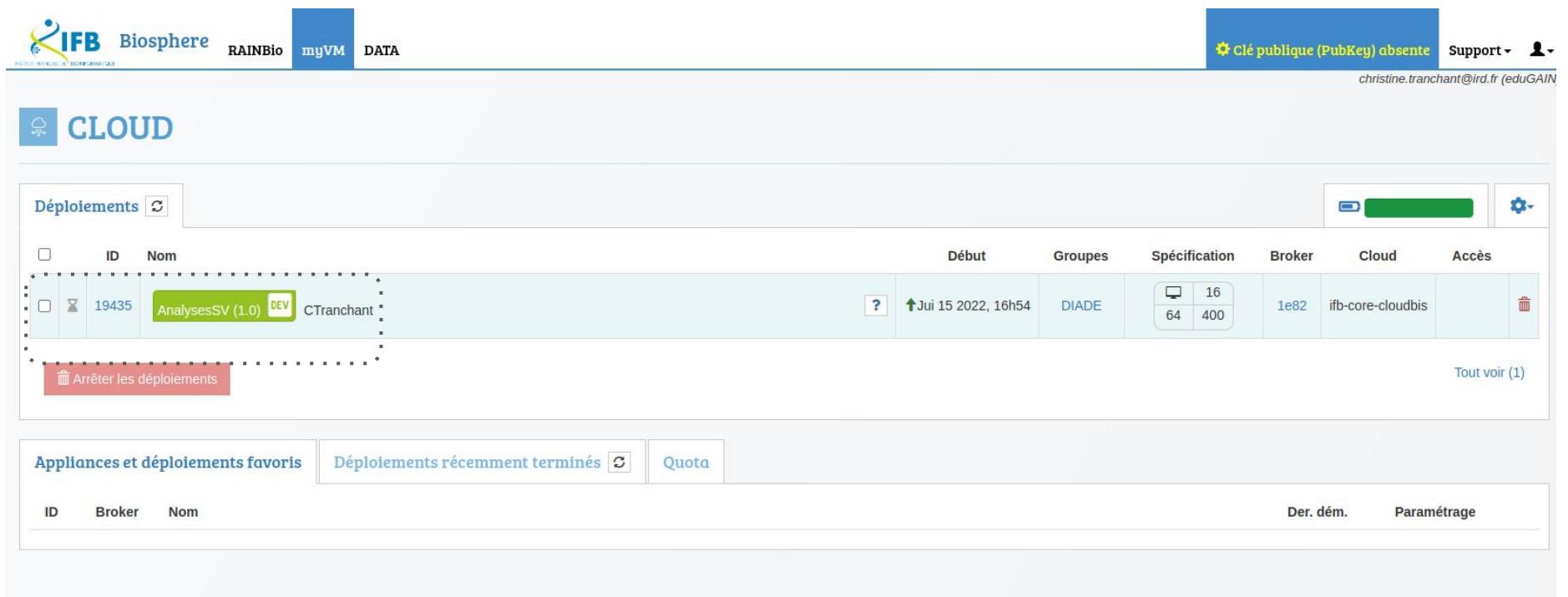
Quelle gabarit d'image doit être utilisé sur ce cloud ? vCPU.h

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)

# Let's run your vm through the cloud

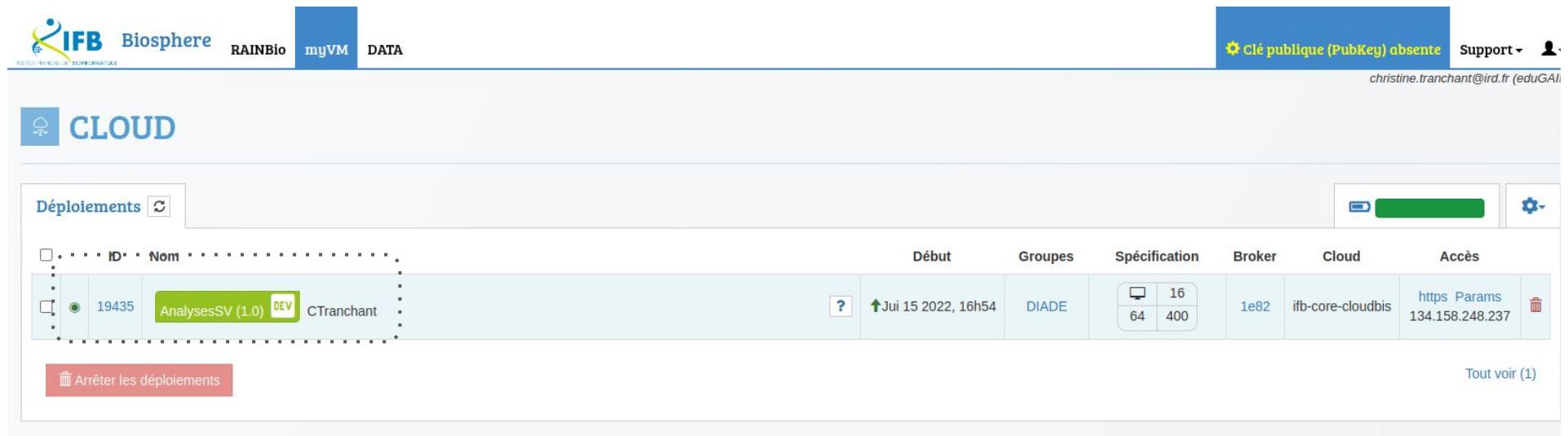
Loading...



The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there is a navigation bar with tabs: IFB Biosphere, RAINBio, myVM (which is selected and highlighted in blue), 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 (Support). Below the header, the main content area is titled "CLOUD". It displays a table of "Déploiements" (Deployments). The table has columns for ID, Nom (Name), Début (Start), Groupes (Groups), Spécification (Specification), Broker, Cloud, and Accès (Access). One deployment is listed: ID 19435, Name AnalysesSV (1.0) DEV, Started at Jui 15 2022, 16h54, DIADE broker, ifb-core-cloudbis cloud, and 1e82 access. There is a red button labeled "Arrêter les déploiements" (Stop deployments) and a link "Tout voir (1)". Below this, there are sections for "Appliances et déploiements favoris" (Favorite appliances and deployments), "Déploiements récemment terminés" (Recently completed deployments), and "Quota".

# Let's run your vm through the cloud

ready !



The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there is a navigation bar with links: IFB Biosphere, RAINBio, myVM (which is highlighted in blue), 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 (eduGAL)).

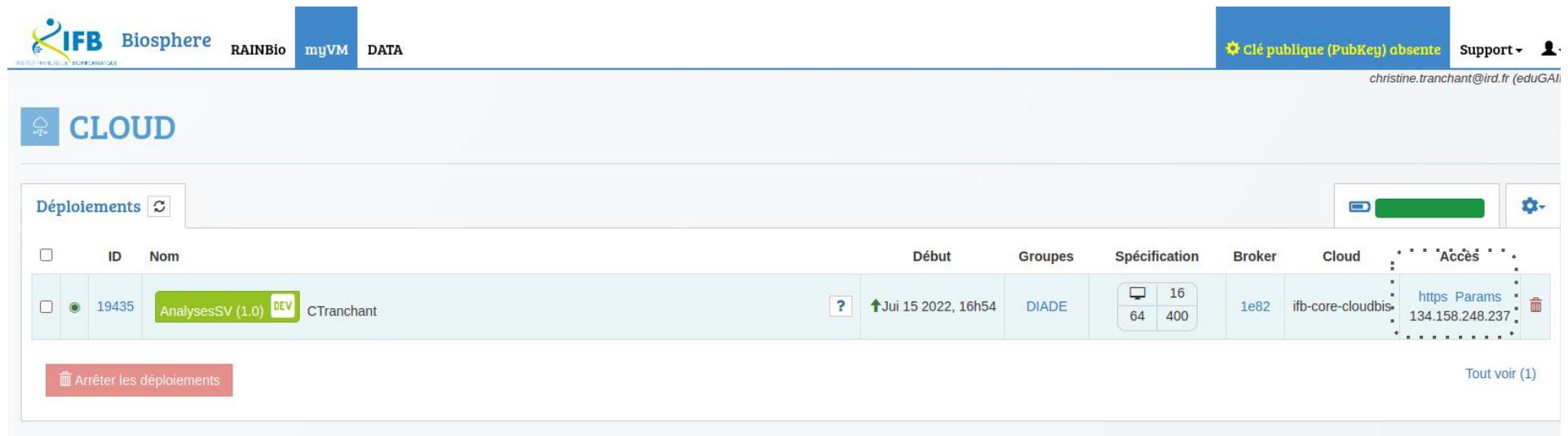
The main area is titled "CLOUD". It contains a table titled "Déploiements" (Deployments). The table has columns for ID, Nom (Name), Début (Start), Groupes (Groups), Spécification (Specification), Broker, Cloud, and Accès (Access). One row is visible in the table:

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><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 of the table area, there is a red button labeled "Arrêter les déploiements" (Stop deployments). At the bottom right, there is a link "Tout voir (1)" (View all 1).

# Let's run your vm through the cloud

get the url... link “https”



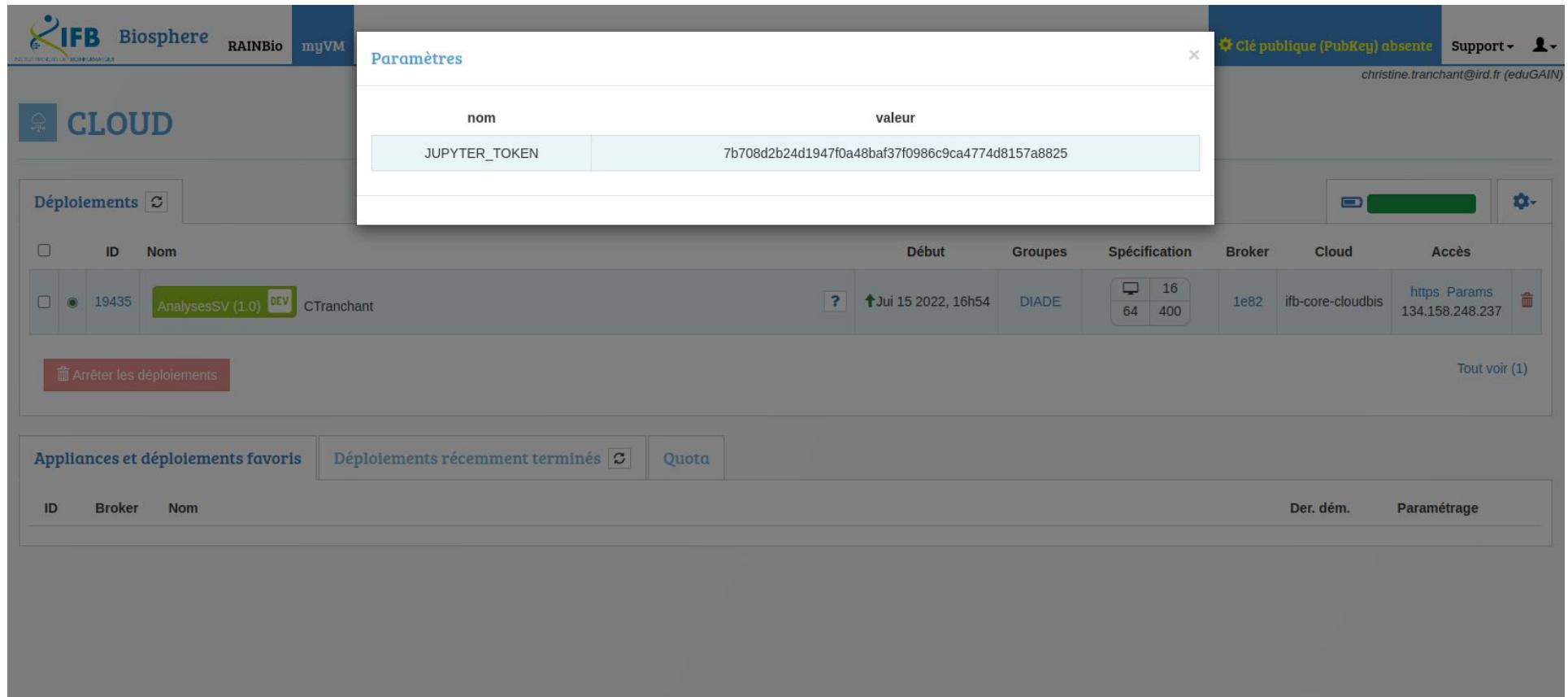
The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there is a navigation bar with tabs: IFB Biosphere, RAINBio, myVM (which is selected), and DATA. On the right side of the header, there is a message about a missing public key and an email address: christine.tranchant@ird.fr (eduGAL).

The main area is titled "CLOUD". It displays a table of "Déploiements" (Deployments). The columns include: checkbox, ID (19435), Nom (AnalysesSV (1.0) DEV), Début (Jui 15 2022, 16h54), Groupes (DIADE), Spécification (16 cores, 64 GB RAM, 400 GB disk), Broker (1e82), Cloud (ifb-core-cloudbis), and Accès (https://134.158.248.237/). There is also a "Params" column and a delete icon.

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

# 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, and CLOUD. The CLOUD tab is active, showing a deployment named "AnalysesSV (1.0) DEV" with ID 19435, created by user CTranchant on July 15, 2022, at 16h54. The deployment is associated with the DIADE broker and has 16 cores, 64 GB of memory, and 400 GB of storage. It is connected to the ifb-core-cloudbis Cloud and has an https Params URL: 134.158.248.237. A red button labeled "Arrêter les déploiements" (Stop deployments) is visible.

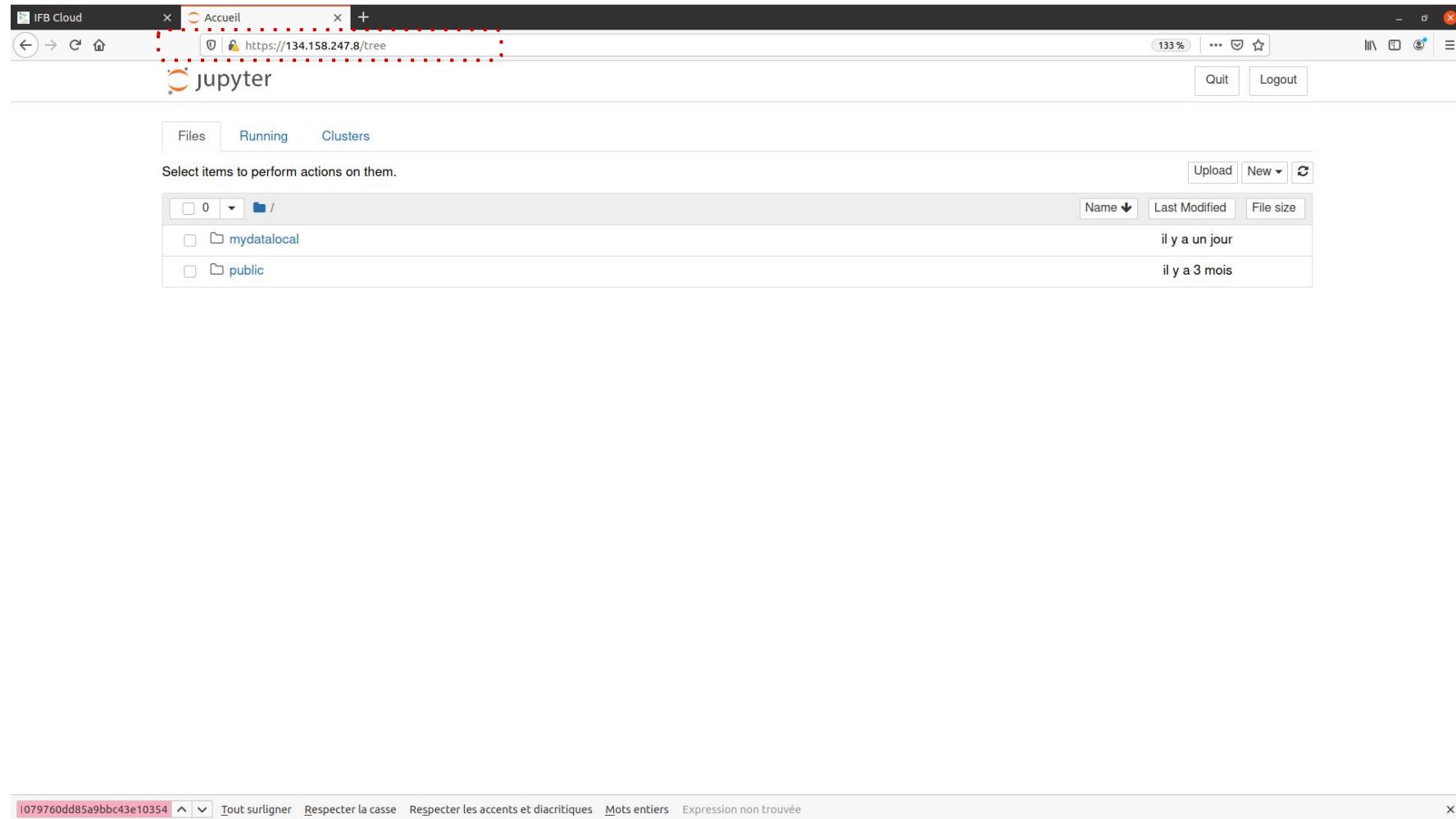
A modal dialog titled "Paramètres" (Parameters) is open, displaying a single parameter entry:

nom	valeur
JUPYTER_TOKEN	7b708d2b24d1947f0a48baf37f0986c9ca4774d8157a8825

At the bottom of the interface, there are sections for "Appliances et déploiements favoris" (Favorites), "Déploiements récemment terminés" (Recently completed deployments), and "Quota".

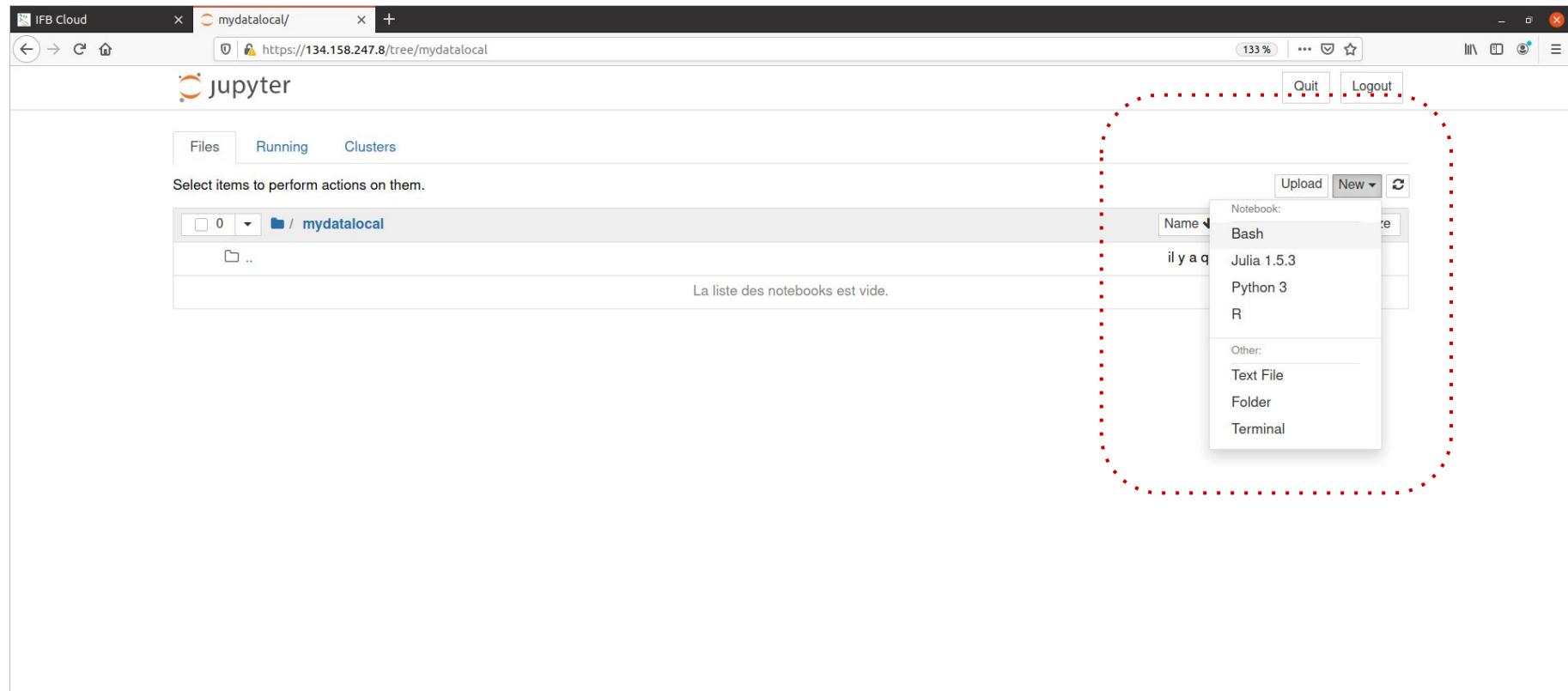
# Let's run our vm through the cloud

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



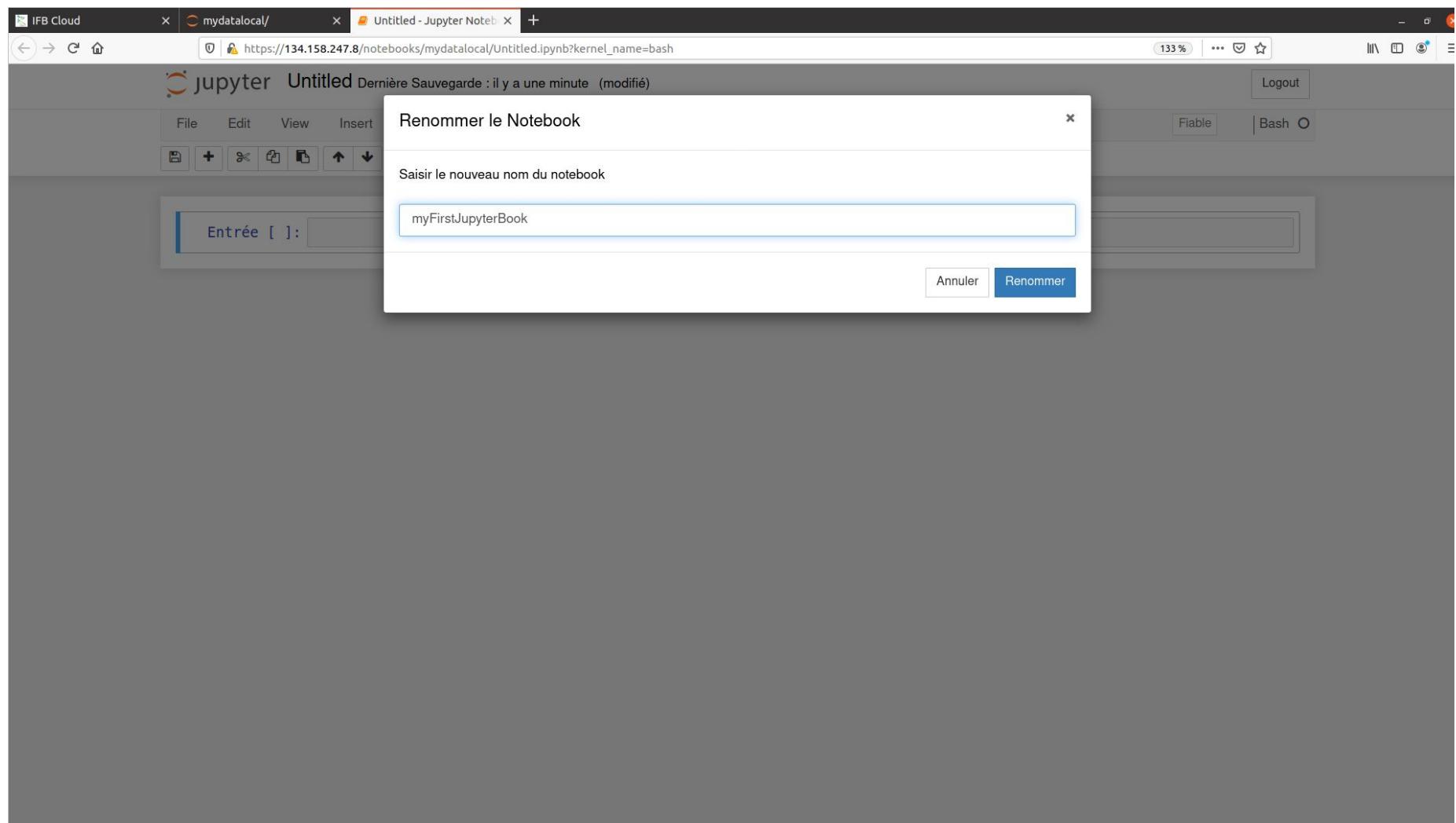
# Create your first jupyter book

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



# Rename your first jupyter book

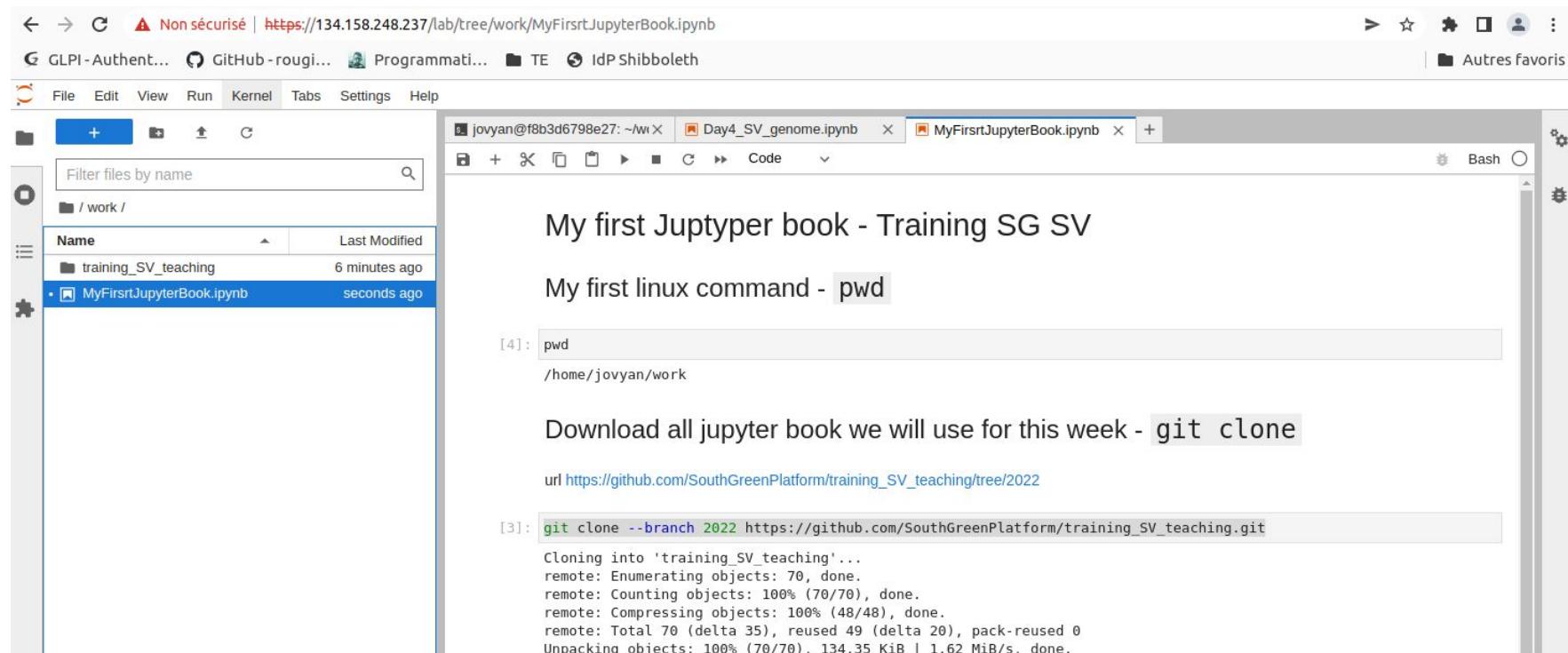
## myFirstJupyterBook



# Run your first bask command - *git clone*

- All jupyterbook used for practice are here :  
[https://github.com/SouthGreenPlatform/training\\_SV\\_teaching/tree/2022](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 two tabs open: "Day4\_SV\_genome.ipynb" and "MyFirstJupyterBook.ipynb". The right pane displays the content of "MyFirstJupyterBook.ipynb". The code cell contains the command:

```
My first Juptyper book - Training SG SV
My first linux command - pwd
[4]: pwd
/home/jovyan/work
```

Below the code cell, the output shows the current working directory:

```
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.
```

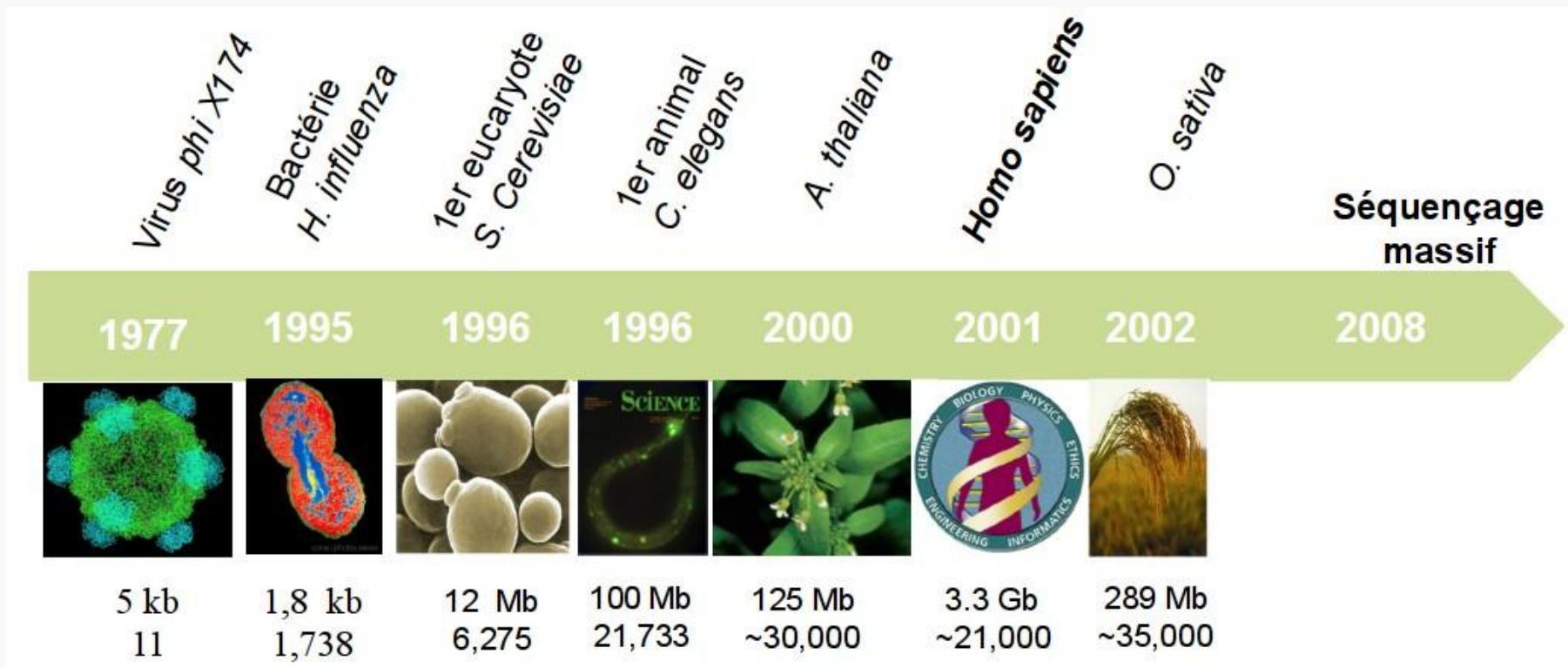


# Introduction & NGS method

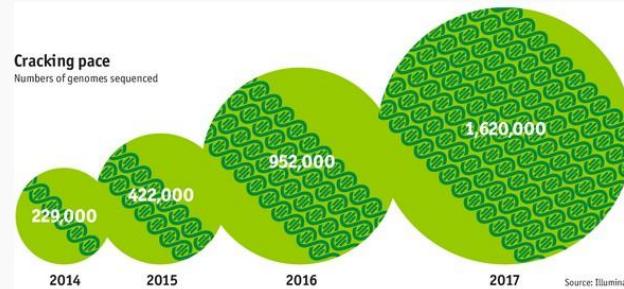
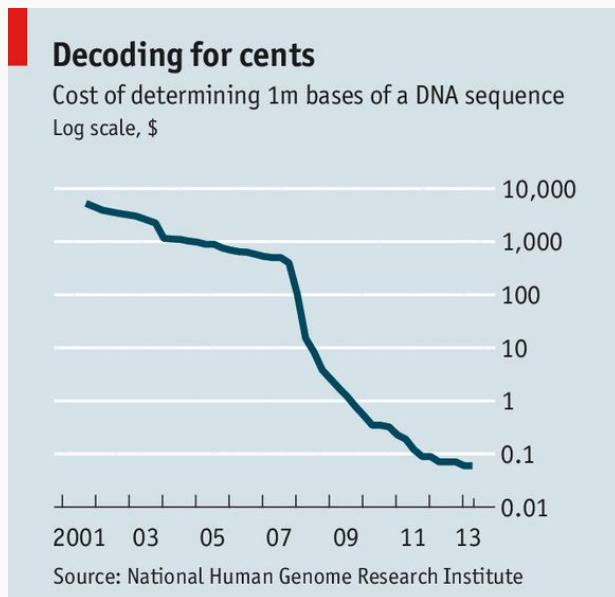
## **The NGS in themselves**

---

# A little history of sequencing...

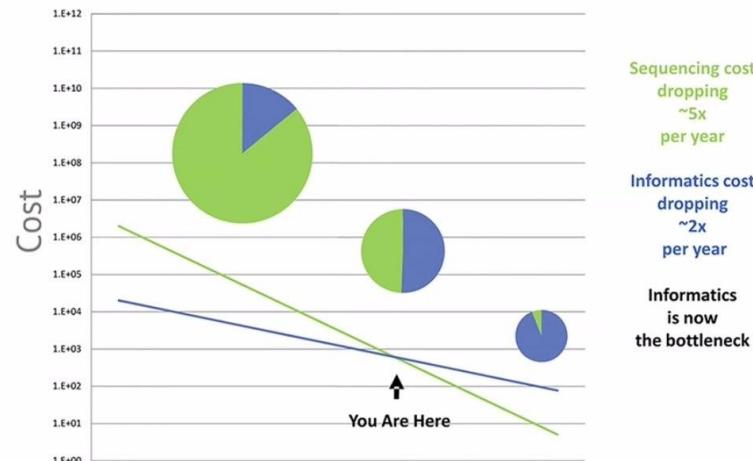


# ...From Data Rarity to Data Deluge



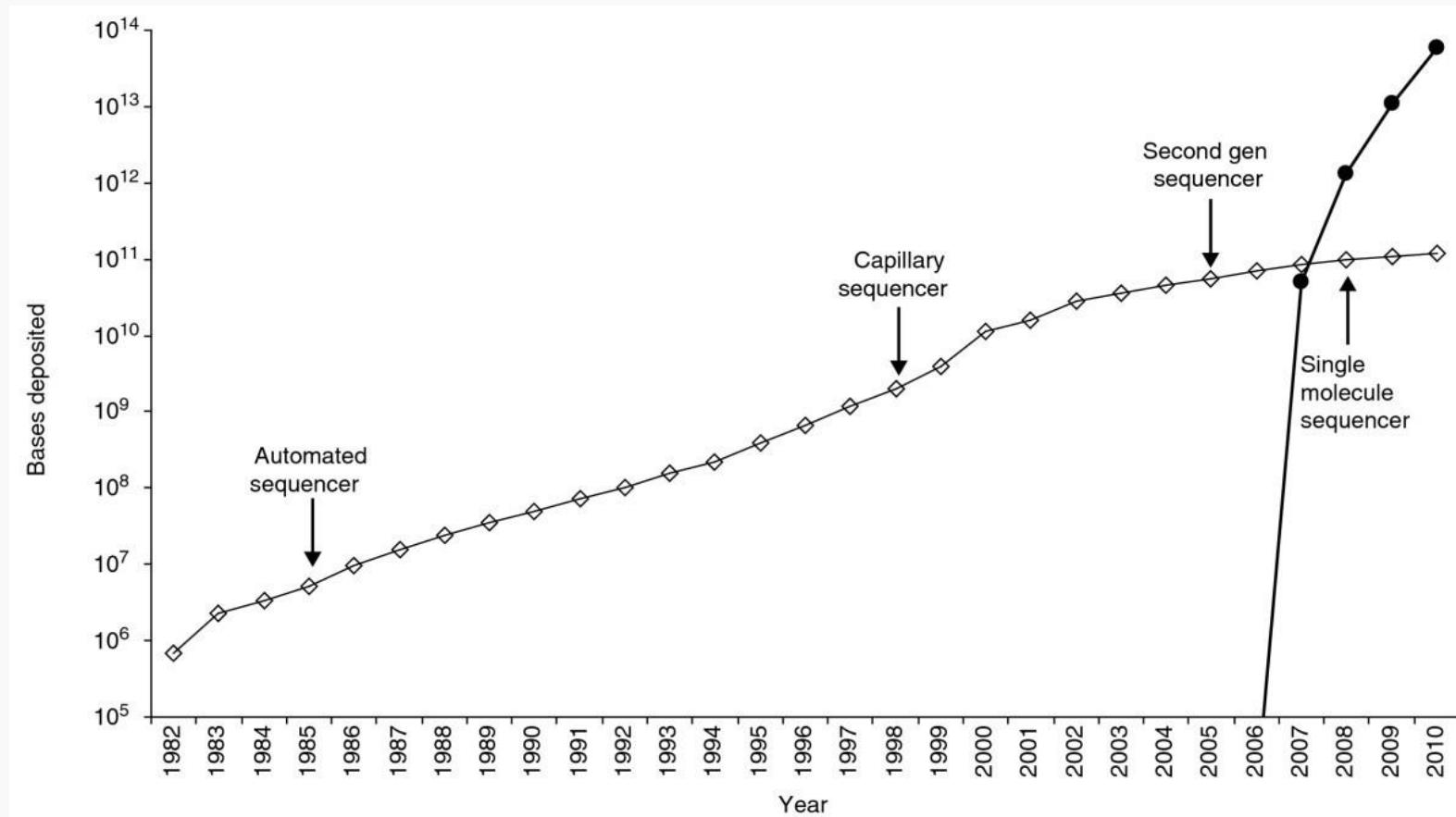
From The economist

## 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

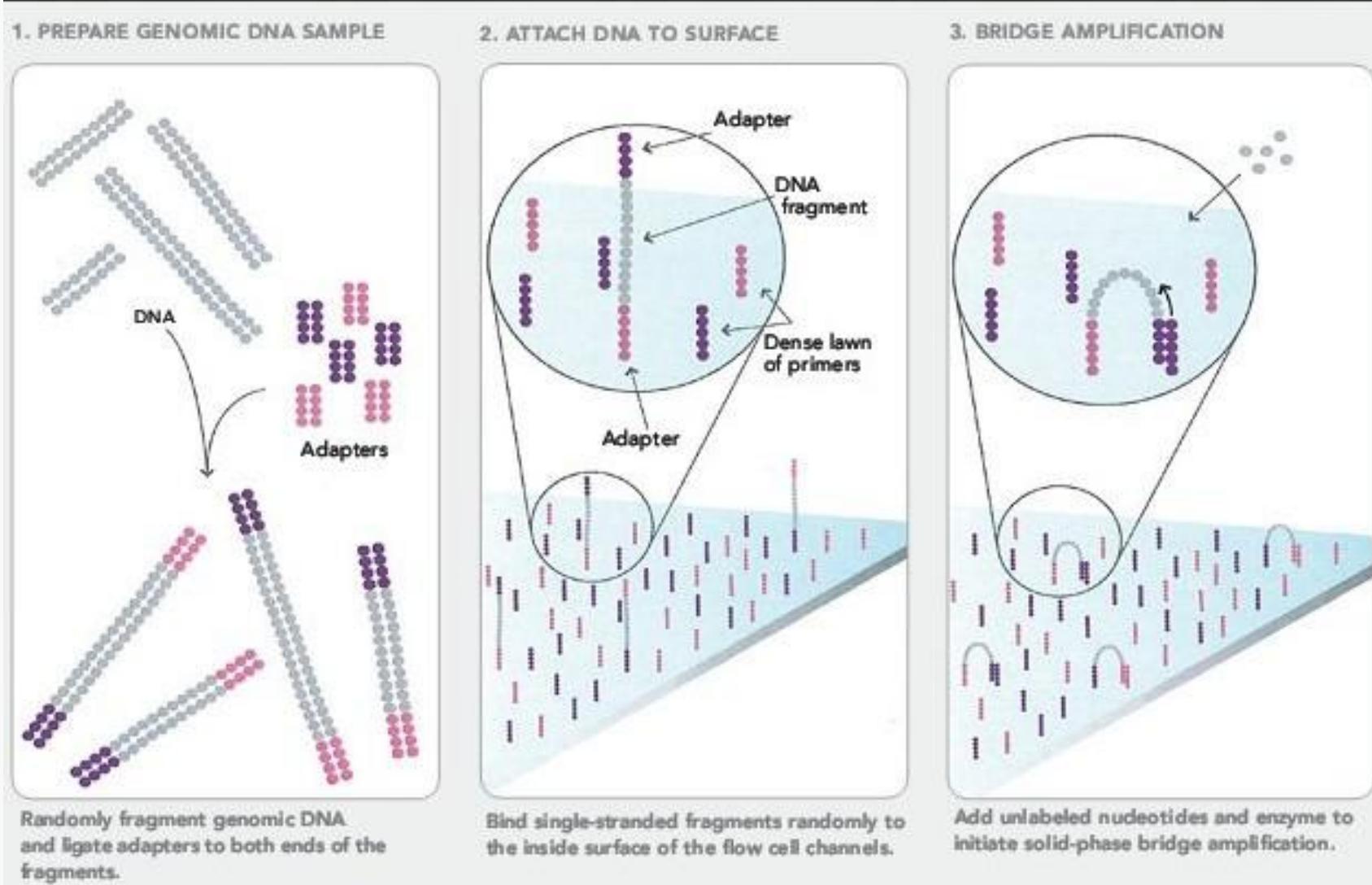
PacificBiosciences  
**Oxford**  
**Nanopore**

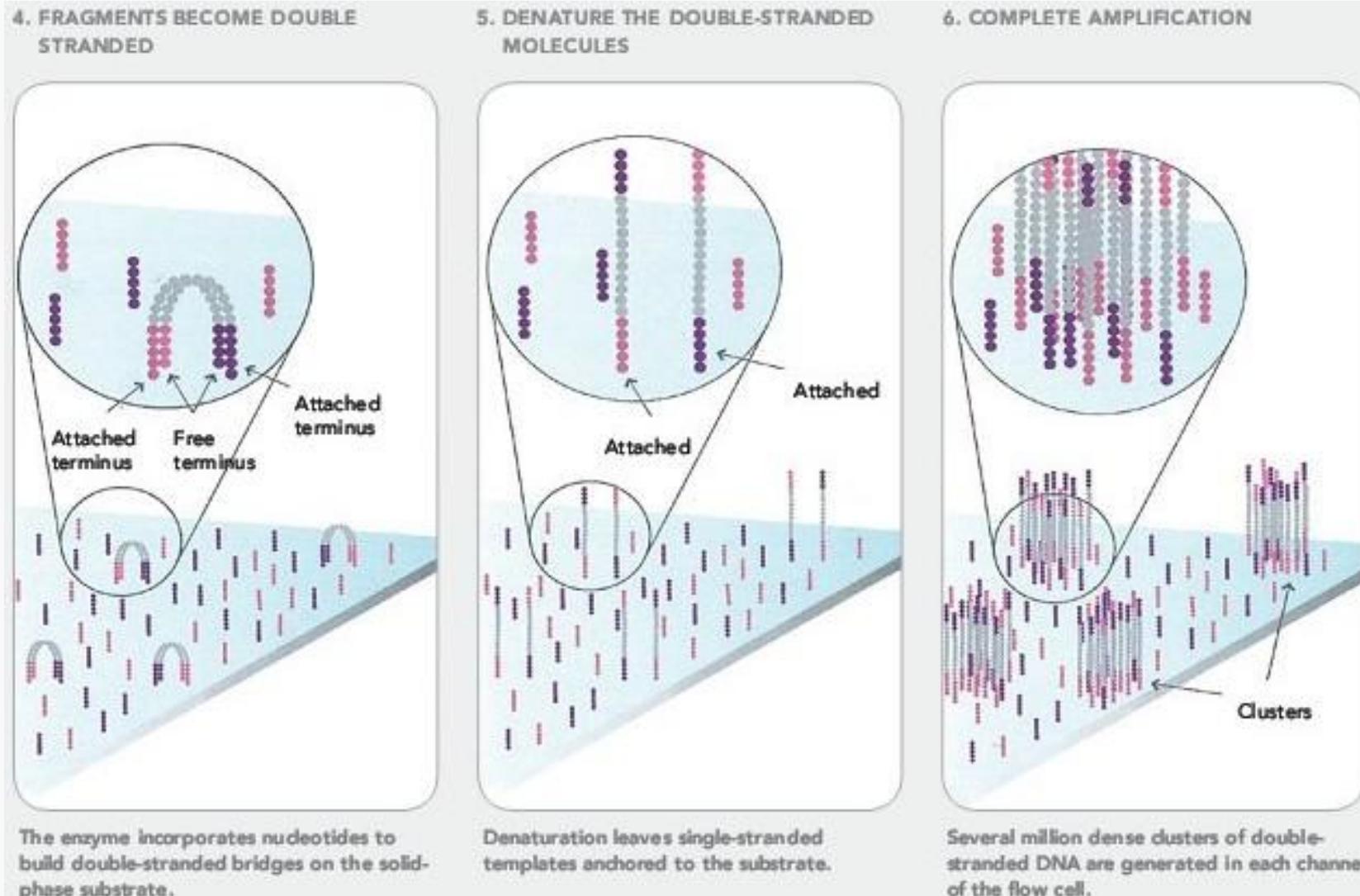
## 3<sup>rd</sup> Generation Sequencing

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

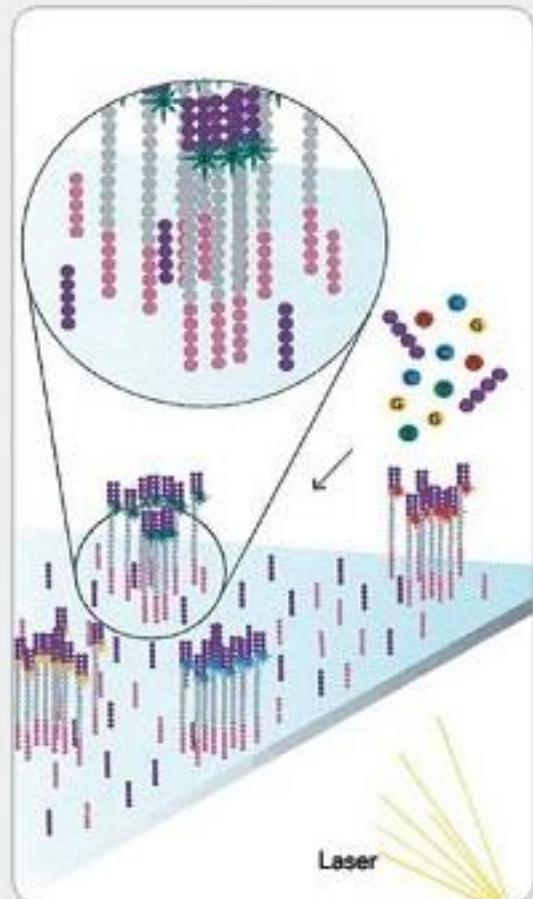






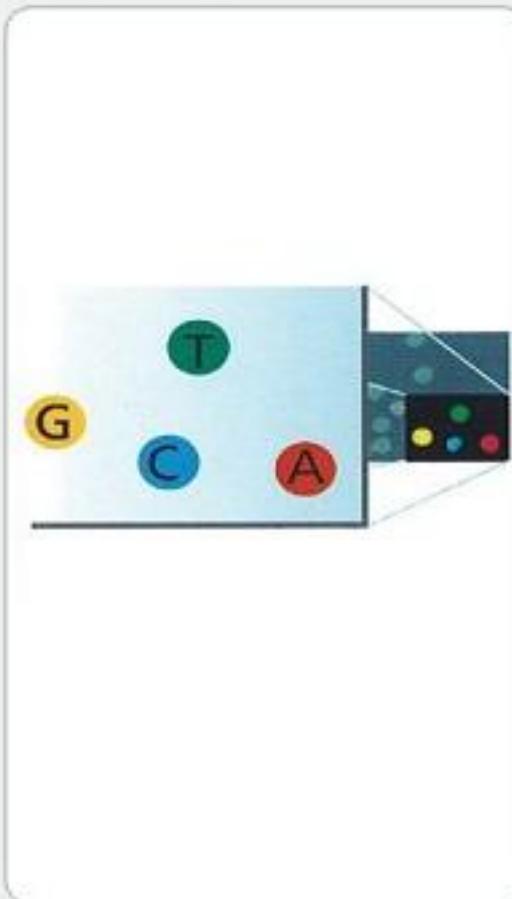


## 7. DETERMINE FIRST BASE



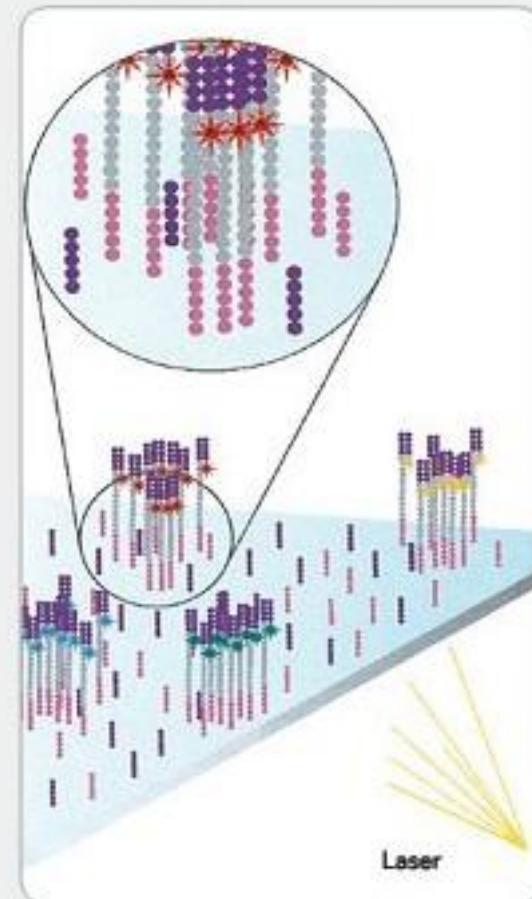
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

## 8. IMAGE FIRST BASE



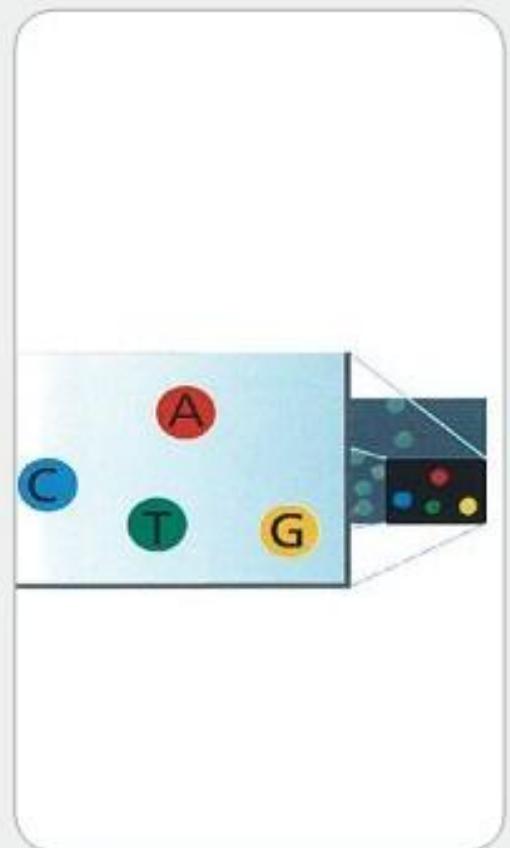
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

## 9. DETERMINE SECOND BASE



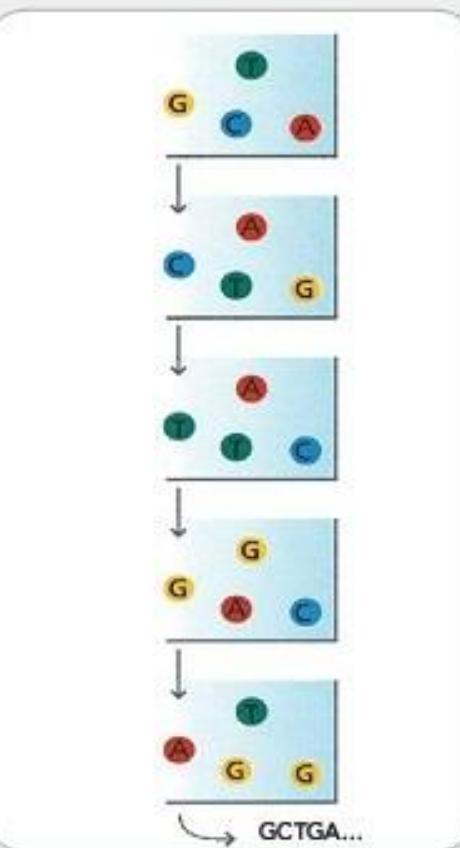
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

## 10. IMAGE SECOND CHEMISTRY CYCLE



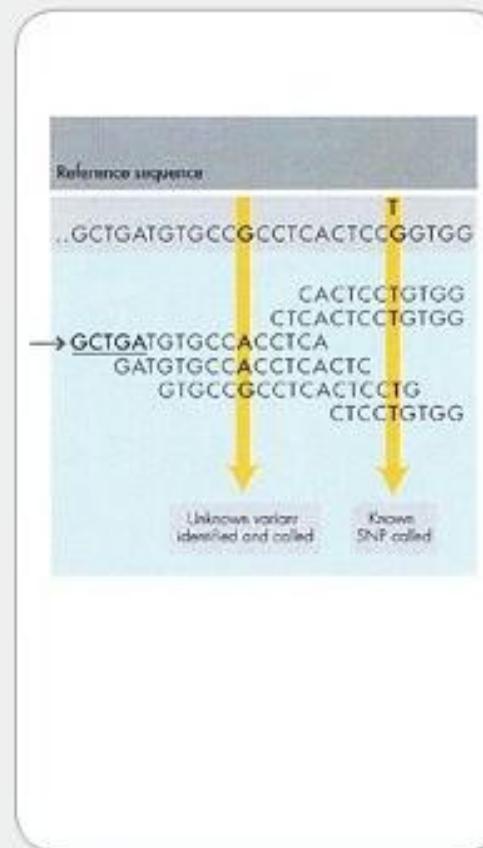
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

## 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
CTGTTTCTTAGTATTTTGATGTAGTCATTCCGTGTTGGTTAGTTGCAAGGT
+
@@@@DADFFHHFFHIIIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTTGGT
+
@@@@DDDD>FFFABEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTCAGTAACGACCTATAAATTTAAAGTAGC
+
@CFFFFFGHHHHJJIEE<HHHJJIGBHGGEEIJJEIEIJIHHJFIIJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTCGGAAAAGTCGGGTATGGCTCTAGTAGCTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIIEHIII<CGHIJIIJ:?:FC9DGAFGHII?DGBFIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAAGCTAAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@@DFFFFGHHHHJJIIIIJJIGJCHHEIII>GHIG@GHIDHGJIIFHIIJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTAAAGAGTTAAAAAACAAATTATGAGCAACTGAGTTC
+
@@@CFFFFFHFFFGIJJIHIIJJIIHJJECGHJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTGTCGGGGTTGGAAATTGTCACTTCTGCTACAATGCCG
+
@<@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

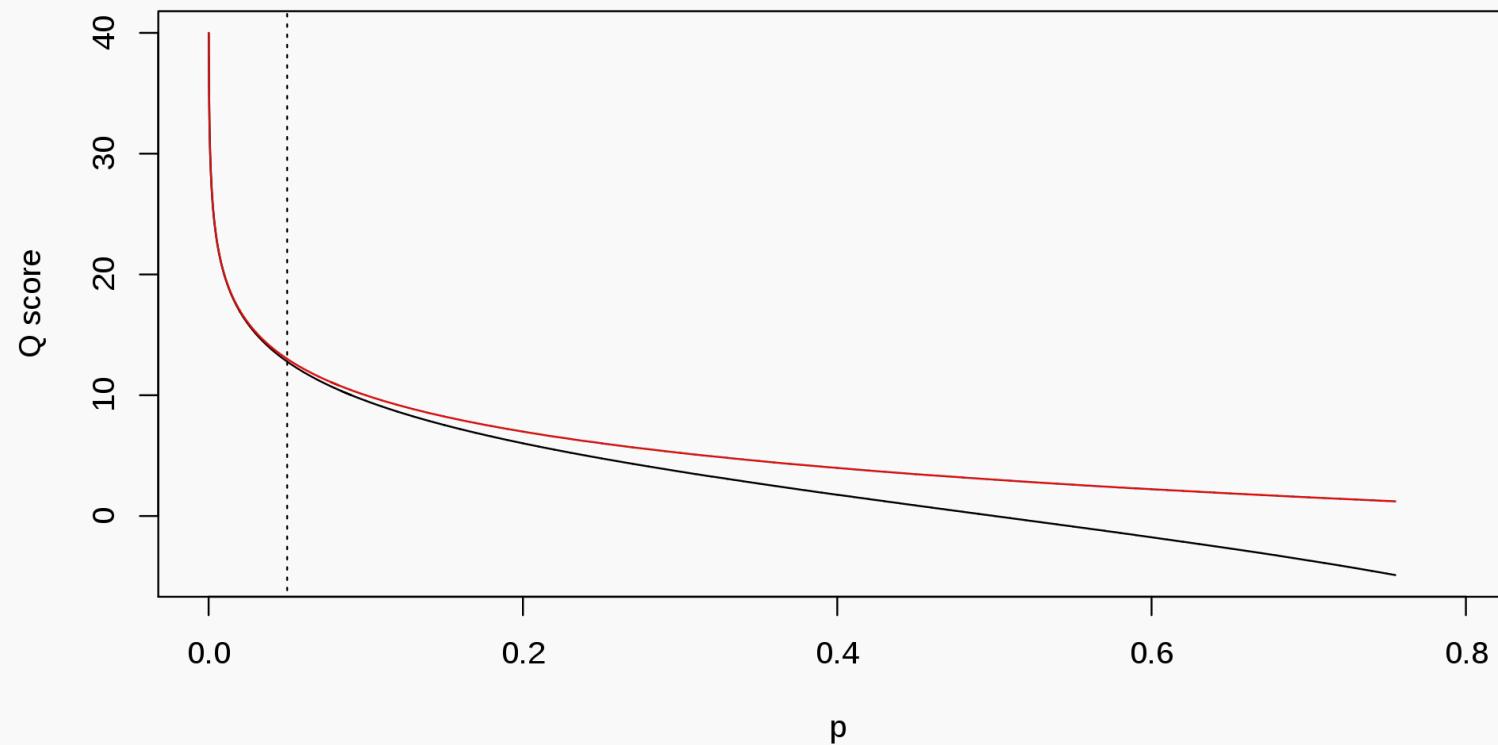
Phred Quality Score
0 .. 50

Code ASCII

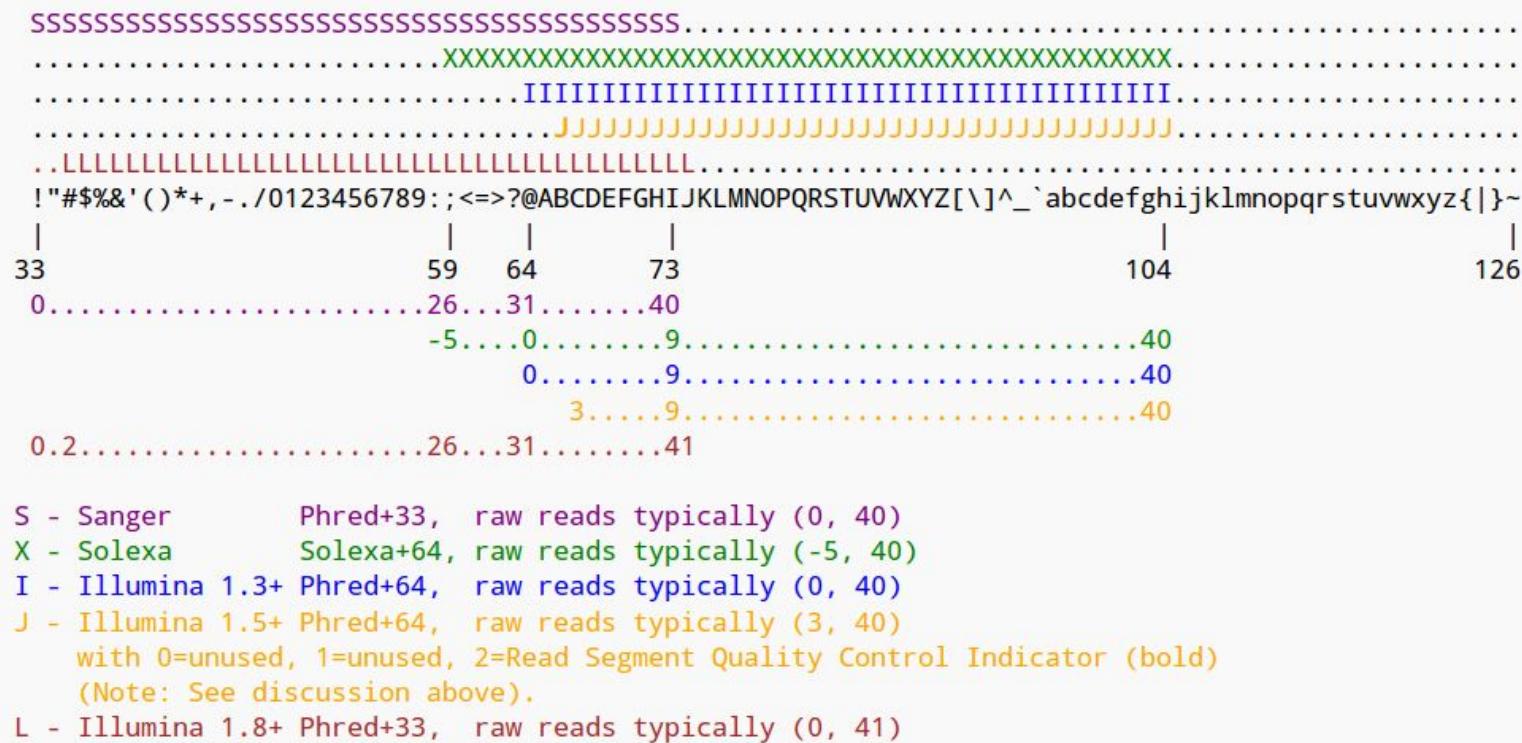


Dec	Hex	Oct	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr
0 0	000	NULL		32 20	040	&#032;	Space		64 40	100	&#064;	@		96 60	140	&#096;	`	
1 1	001	SoH		33 21	041	&#033;	!		65 41	101	&#065;	A		97 61	141	&#097;	a	
2 2	002	SoTxt		34 22	042	&#034;	"		66 42	102	&#066;	B		98 62	142	&#098;	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	

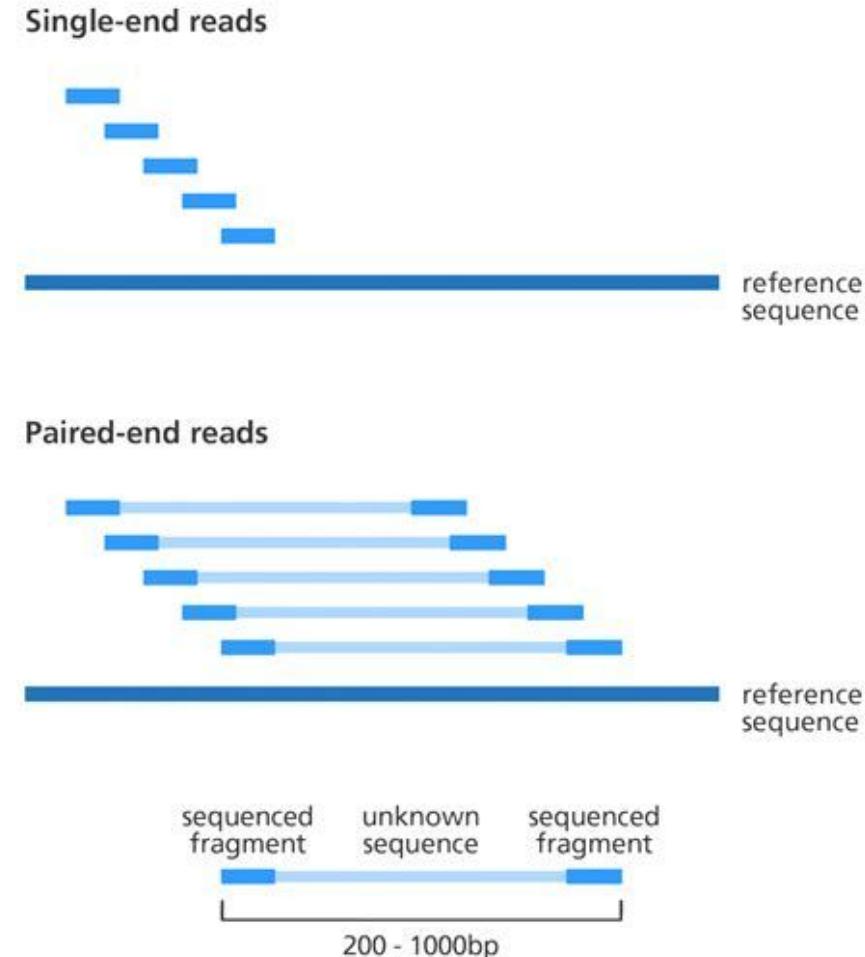
# The QPHRED Value



# The QPHRED Scale



# Mapping



```

@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.
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).
PG - Program	PL	Platform/technology used to produce the read.
	ID*	Program name
	VN	Program version
CO - comment	CL	Command line
		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>

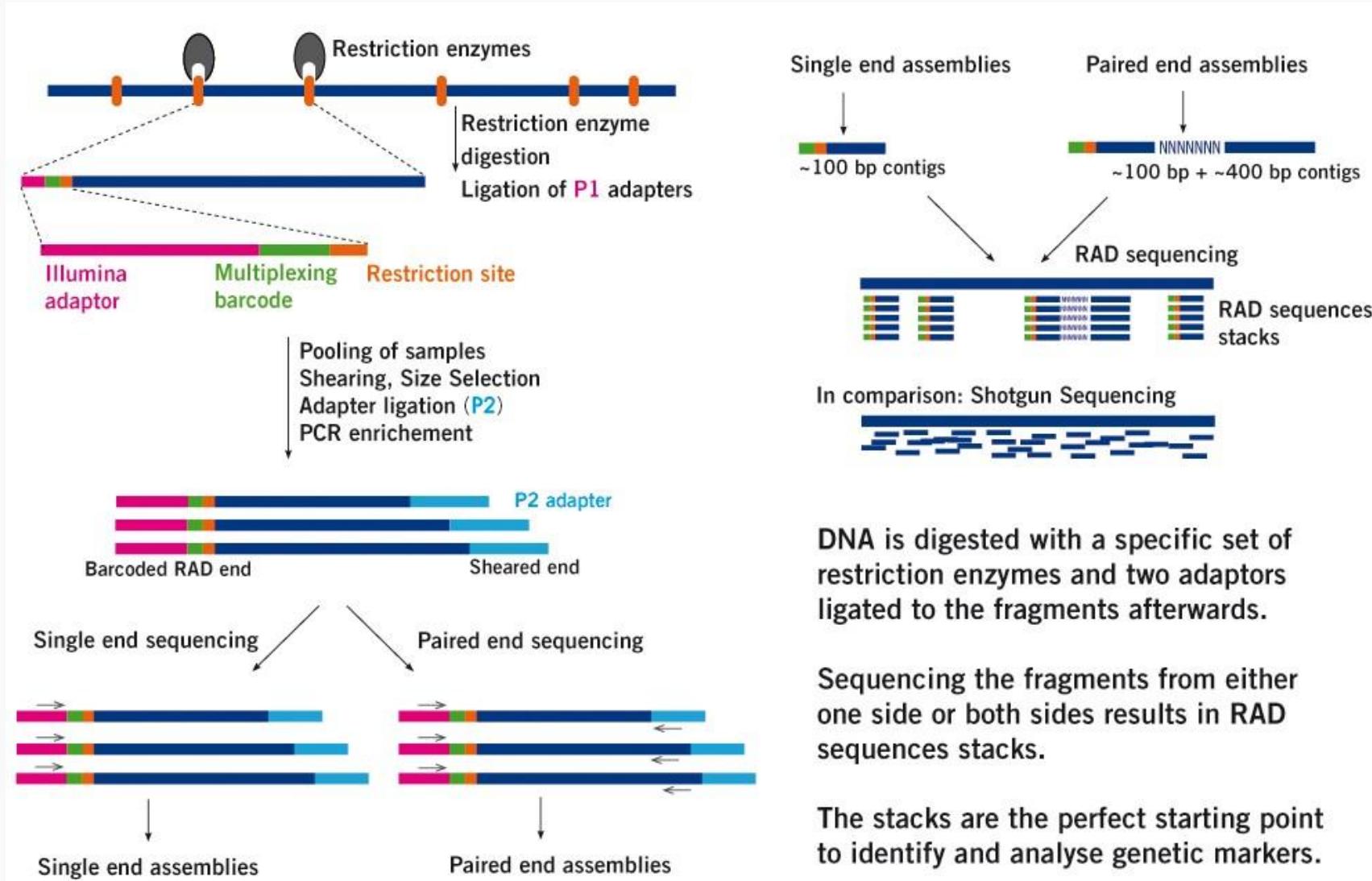
# SAM Format

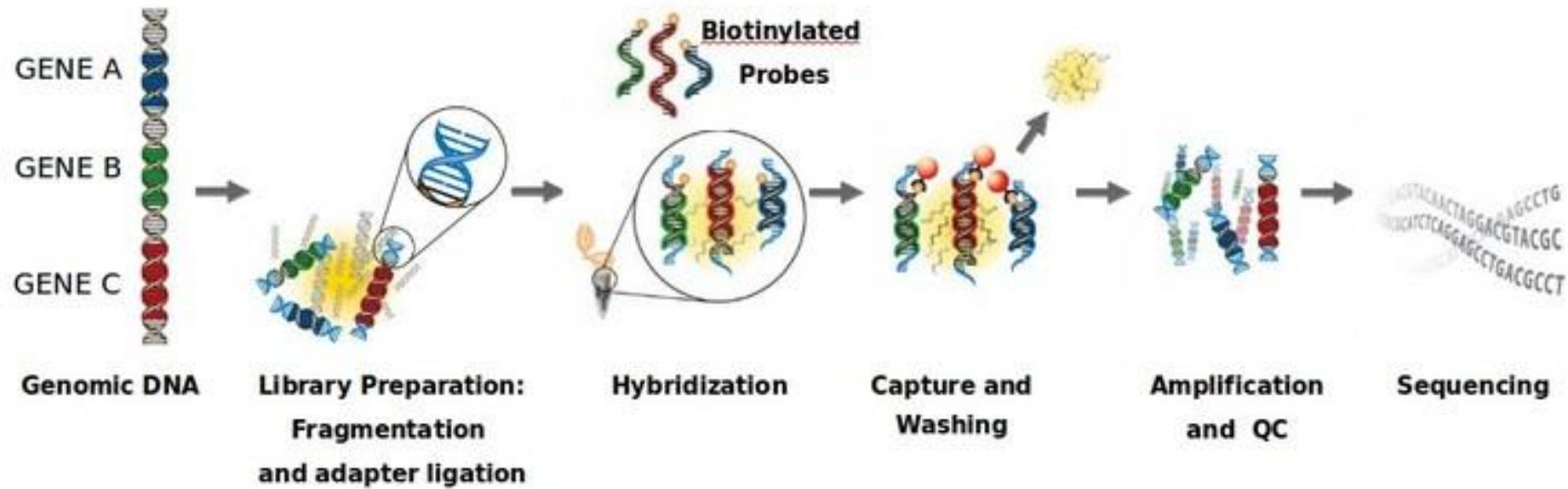
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC
r003 16 ref 29 30 6H5M * 0 0 TAGGC * N
r001 83 ref 37 30 9M =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

alignement

Format tabulé

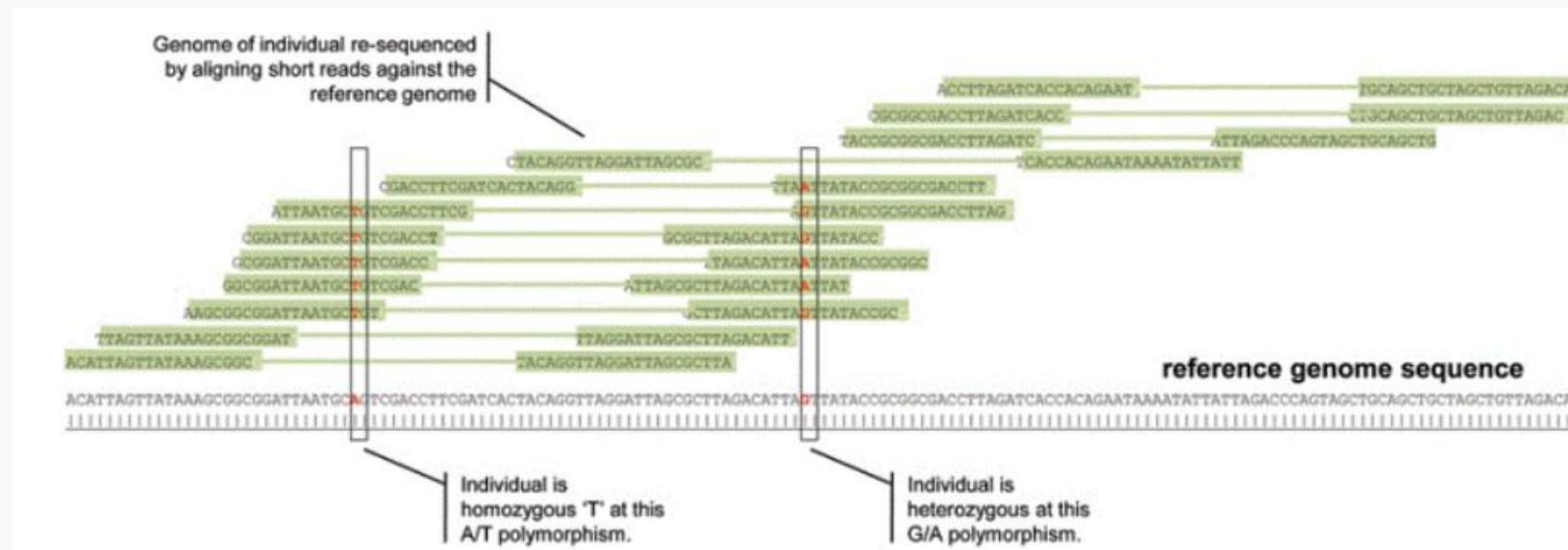
Col	Name	Description
1	<b>QNAME</b>	Query NAME of the read or the read pair
2	<b>FLAG</b>	bitwise FLAG (pairing, strand, mate strand, etc.)
3	<b>RNAME</b>	Reference sequence NAME
4	<b>POS</b>	1-based leftmost POSition of clipped alignment
5	<b>MAPQ</b>	MAPping Quality (Phred-scaled)
6	<b>CIGAR</b>	extended CIGAR string (operations: MIDNSHP)
7	<b>NNRM</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)



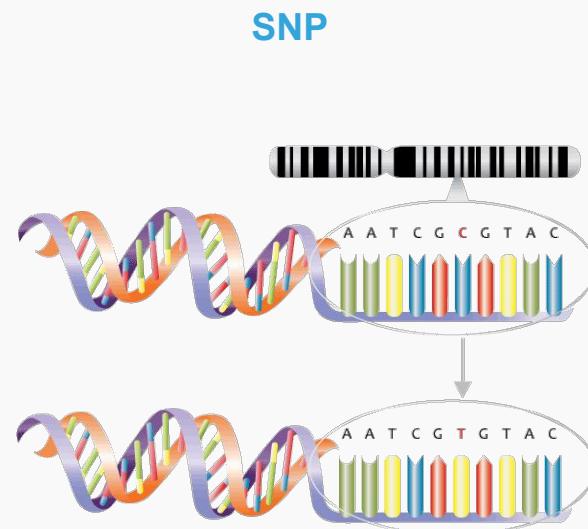


From CGFB, Bordeaux, France

# SNP and InDel Detection

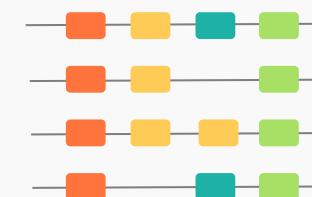
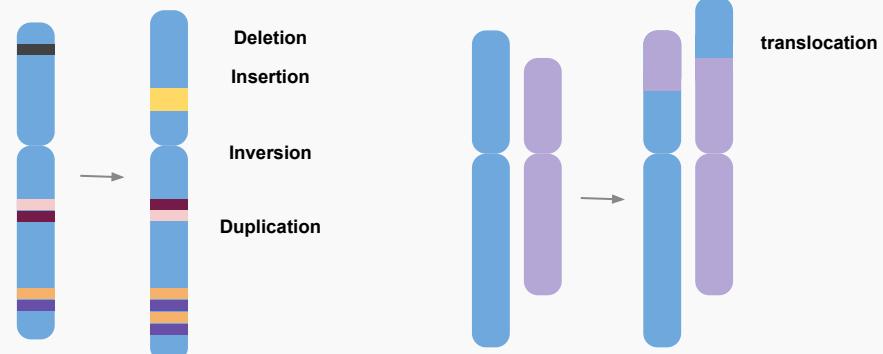


## Mutations & Variations as main source of genetic diversity



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		Mandatory header lines							
		##fileformat=VCFv4.0							
		##fileDate=20100707							
		##source=VCFtools							
		##reference=NCBI36							
Body		##INFO=<ID=AA,Number=1>Type=String>Description="Ancestral Allele">							
		##INFO=<ID=H2,Number=0>Type=Flag>Description="HapMap2 membership">							
		##FORMAT=<ID=GT,Number=1>Type=String>Description="Genotype">							
		##FORMAT=<ID=GQ,Number=1>Type=Integer>Description="Genotype Quality (phred score)">							
		##FORMAT=<ID=GL,Number=3>Type=Float>Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">							
		##FORMAT=<ID=DP,Number=1>Type=Integer>Description="Read Depth">							
		##ALT=<ID=DEL,Description="Deletion">							
		##INFO=<ID=SVTYPE,Number=1>Type=String>Description="Type of structural variant">							
		##INFO=<ID=END,Number=1>Type=Integer>Description="End position of the variant">							
		#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 . <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20							
		Deletion SNP Large SV Insertion Other event							
		Phased data (G and C above are on the same chromosome)							
		Reference alleles (GT=0)							
		Alternate alleles (GT>0 is an index to the ALT column)							

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

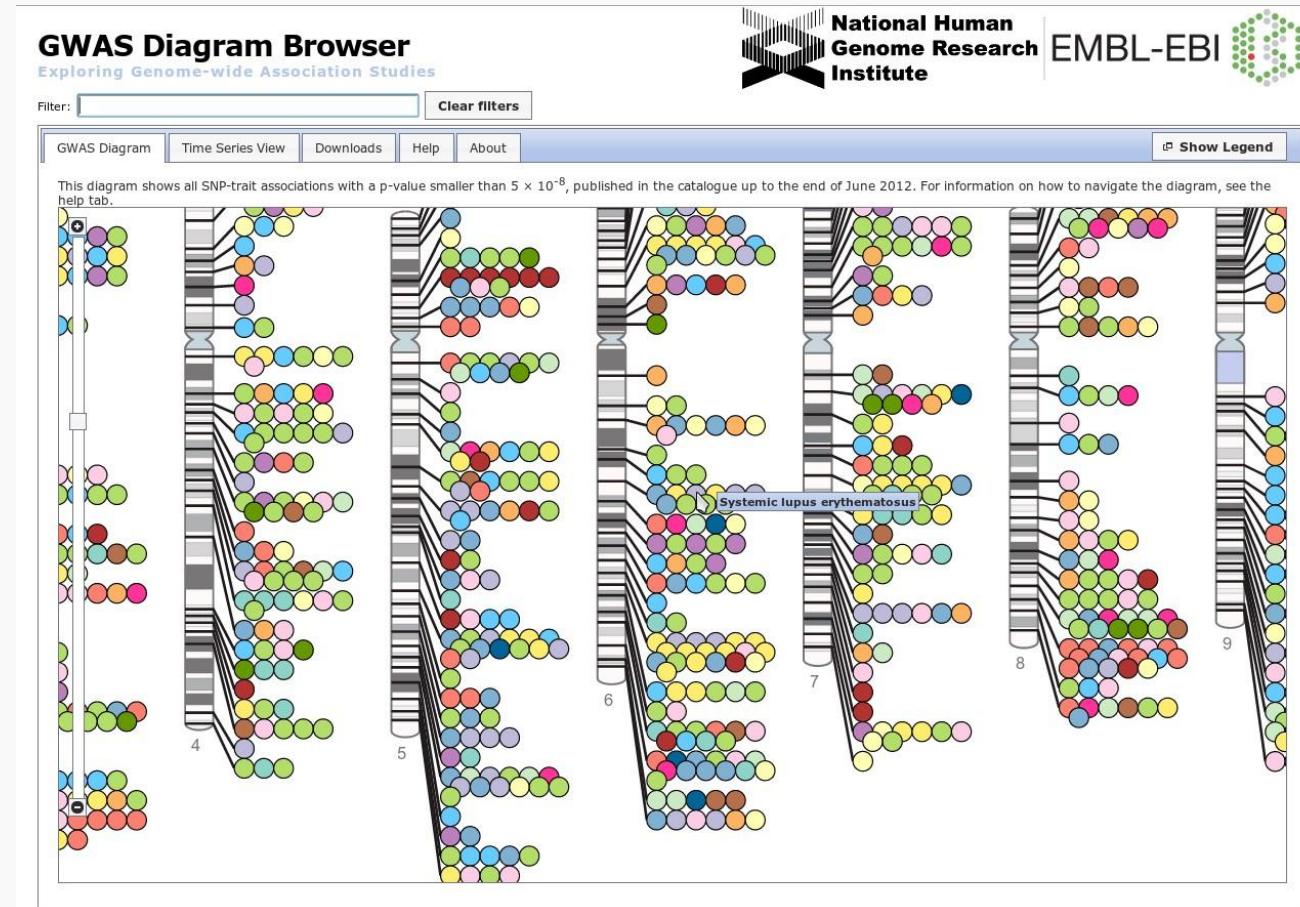
All linked to the Specificity/Sensitivity Informatics Paradox

# **Applications**

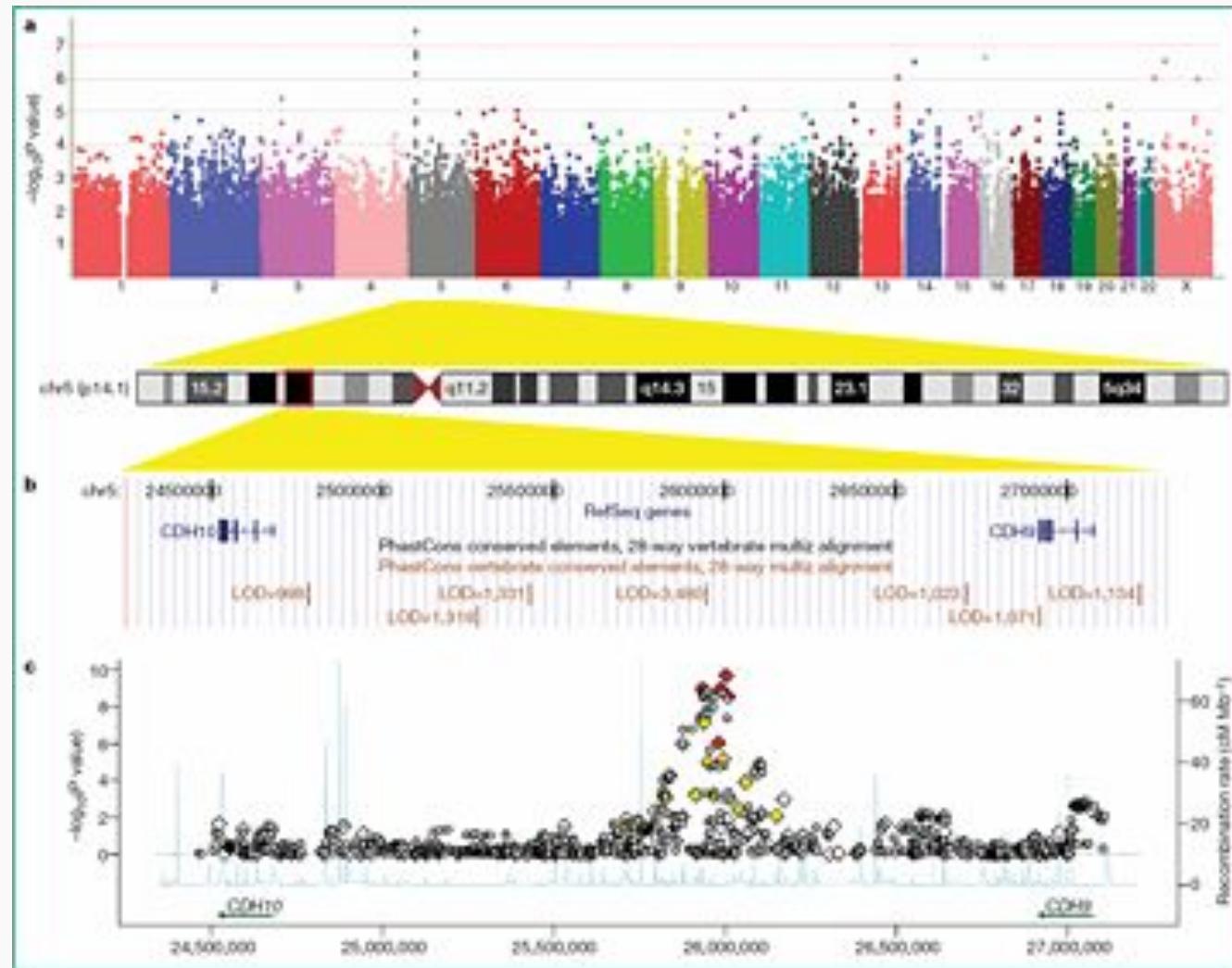
---

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

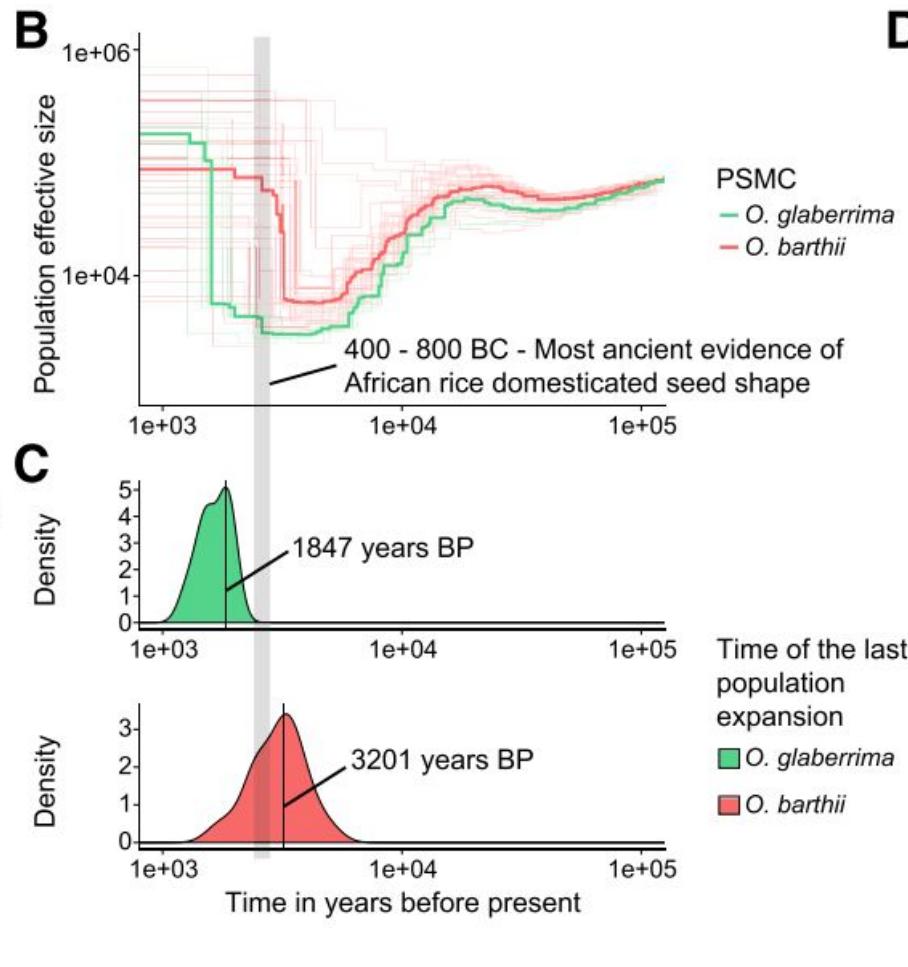
# Example in GWAs & Population Genomics



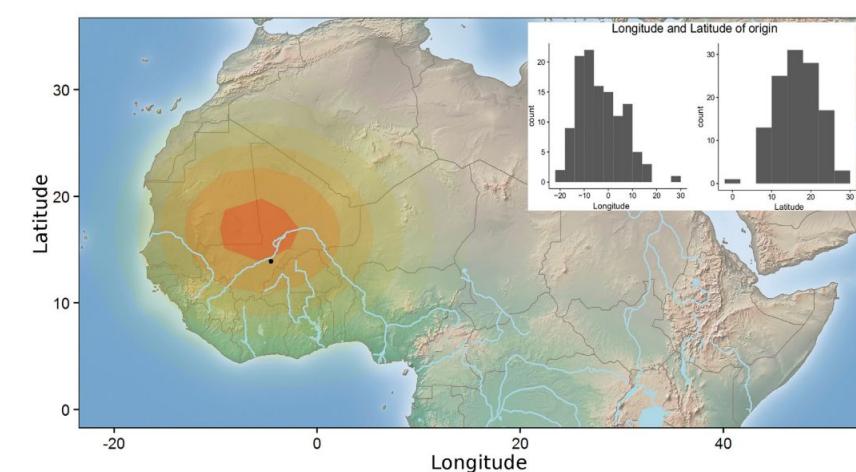
# Example in GWAs & Population Genomics



# 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 menu bar below it. A "LATEST ANNOUNCEMENTS" section highlights the "February 2011 Data Up" release, featuring Indels calls from Dindel. It also includes links to EBI and NCBI.
- 1001 Genomes**: A Catalog of *Arabidopsis thaliana* Genetic Variation. This screenshot shows a light-colored header with the project name and a menu bar. It features a decorative image of a flower in the top right corner.
- Genome 10K**: Unveiling animal diversity. This screenshot shows a blue-themed header with the project name and a menu bar. The background features a collage of various animal DNA helixes and images. A search bar is visible in the top right.

- 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)



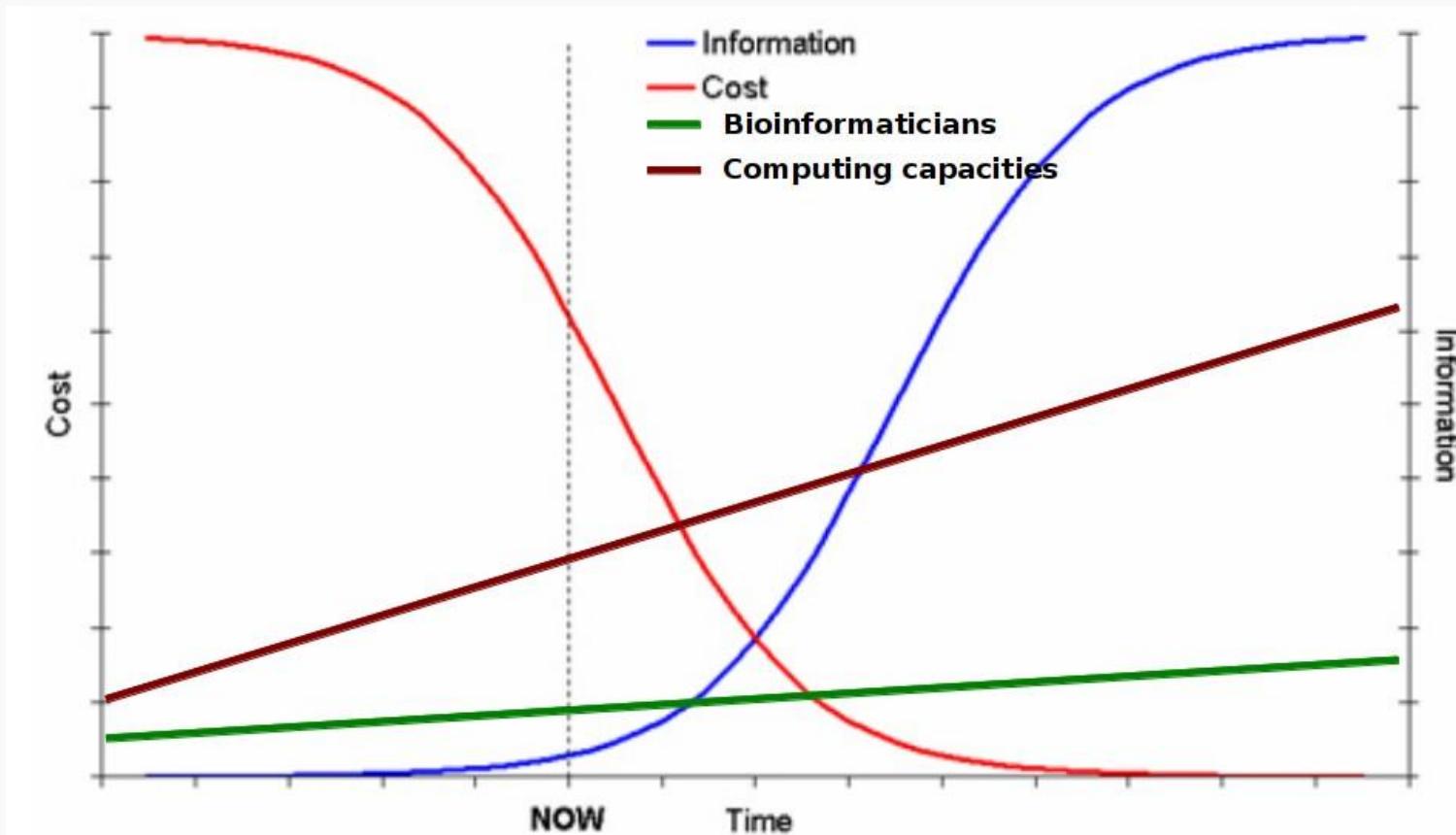
- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available

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

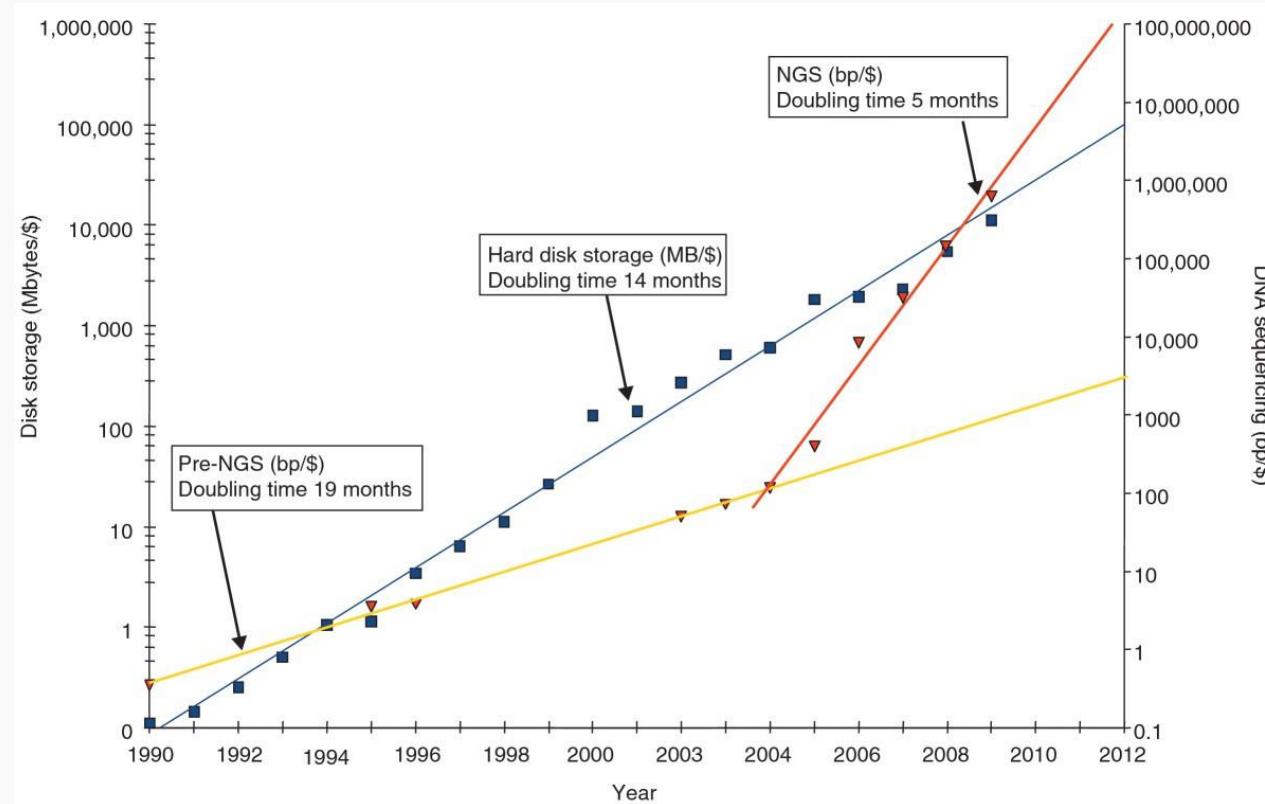


- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available
- And any new ideas you will have...

# Keep in mind!



# ...From Data Rarity to Data Deluge

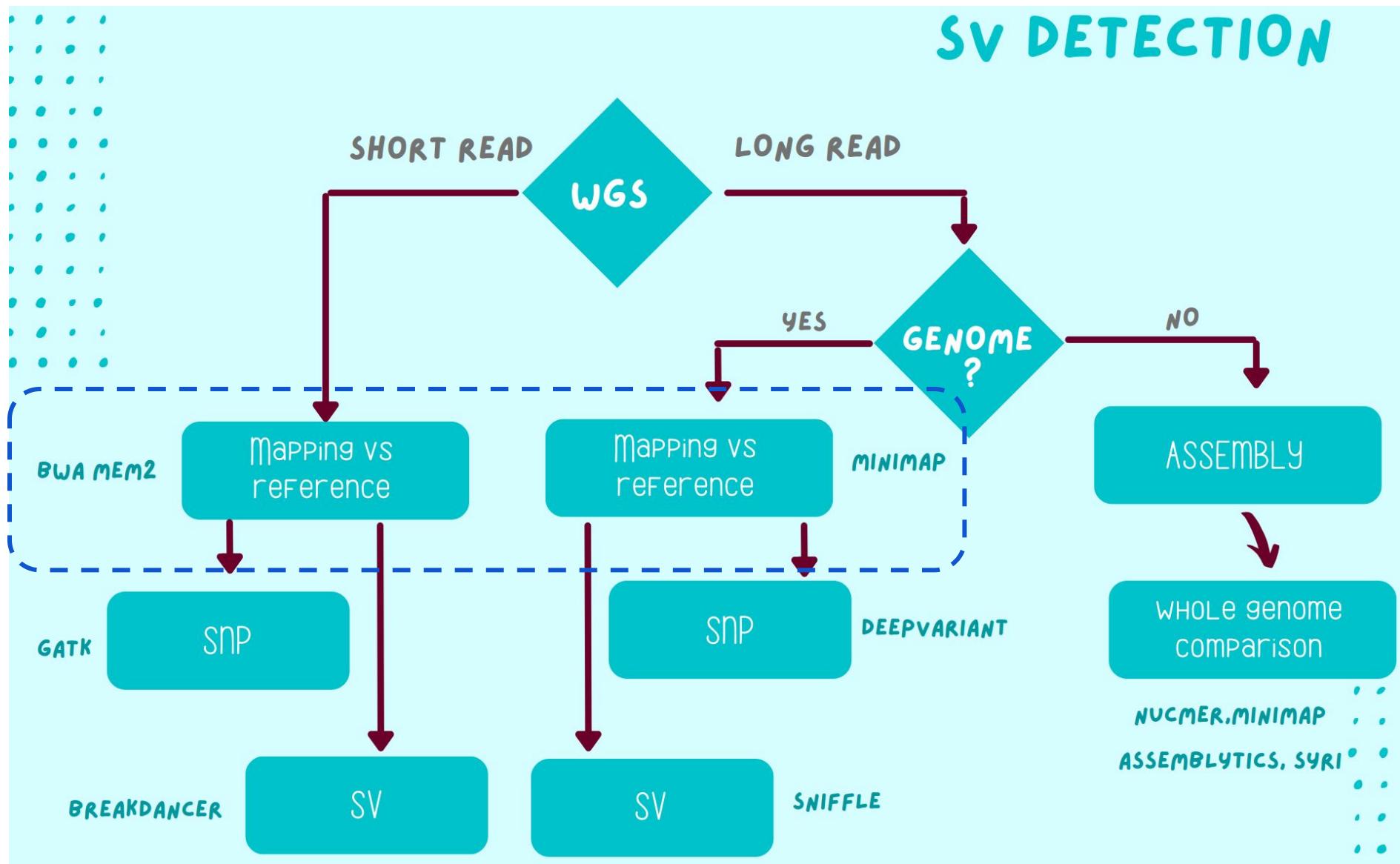


From L. Stein, 2010

# Be Careful to data drowning!



# Training plan - day 1





# 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

# The data

Diploid Asian Rice, *Oryza*



From  
Wikimedia

1. Select/Cut 1 Mb on Chromosome 10

# The data

Diploid Asian Rice, *Oryza*



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

From  
Wikimedia

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...

# The data

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 three lines of text with arrows pointing to specific parts:

- Sequencing info:** Points to the first line starting with '@HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:215:593'. This line contains sequencing parameters and a sample identifier.
- Nucleotide sequence:** Points to the second line starting with 'GAGAAGTTCAACAGCTGGTATTATTTGTTAACAT'. This line contains the actual DNA sequence.
- Quality score in ASCII:** Points to the third line starting with '+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:215:593'. This line contains the quality scores represented as ASCII characters.

```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTGTTAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGAGTGTGGAAAGGCCATT
+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
TCTCAGAAACTTGTGTGATGTGTATTCAAACCA
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
TTGATTTAACTCTGACAAAATAAACAAAGTCCTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

# The QPHRED Scale

## “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.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	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	ID*	Program name
	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

# 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. 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

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
5. Mark the duplicates

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
5. Mark the duplicates
6. Call SNP on this individual



# Practice

Let's work with the jupyter book :

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

- 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**)

- 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

- 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

# Training plan - day 1

