

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

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 SNP detection
phylogeny structural variation
comparative genomics transcriptome assembly differential expression
GWASpangenomics
population genetics metagenomics
polyploidy



Rice



Banana



Palm



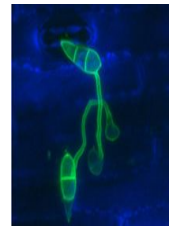
Sorghum



Coffee



Cassava



Magnaporthe



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

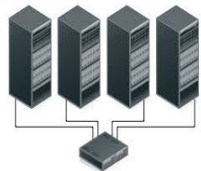
Workflow manager

TOOLBOX
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

HPC and trainings....



37 courses organized last 7 years



Genome Hubs & Information System



Gigwa

SNPs and Indels

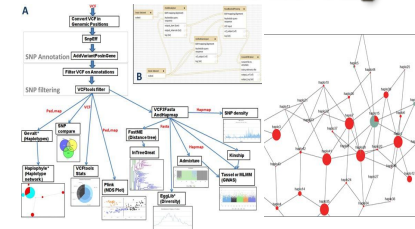
GreenPhyl

Family Id	Family Name	Number of sequences	Status
GP000010	Cytochrome P450 superfamily	6942	●
GP000017	AP2/ERF transcription factor family: ERF/ERF1 group (partial)	5142	●
GP000020	NAC transcription factor family	4574	●
GP000028	MADS transcription factor family		
GP000018	Haem peroxidase superfamily		
GP000046	General substrate transporter superfamily		
GP000022	Subtilisin-like Serine Proteases family		
GP000019	NIP, NRT1/PTR FAMILY		

Gene families



SNiPlay



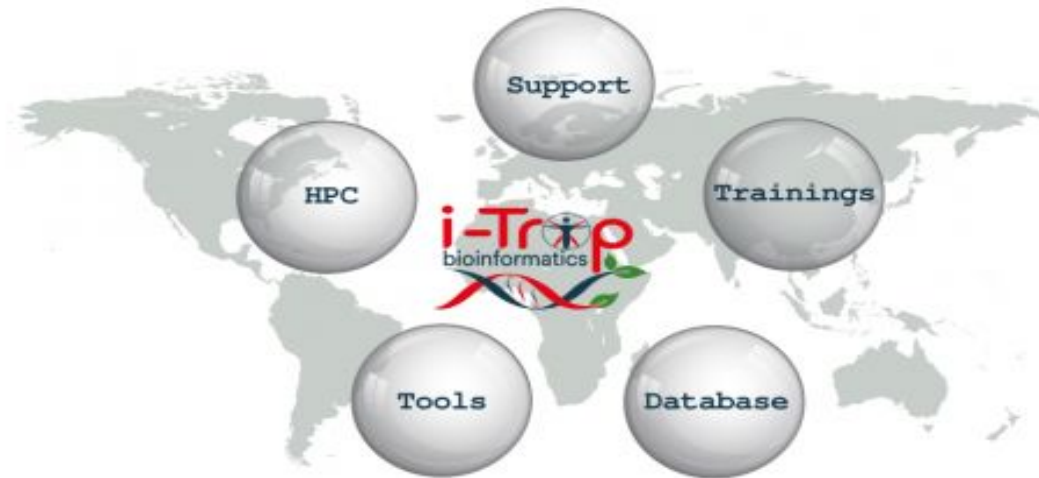
<https://github.com/SouthGreenPlatform>



@green_bioinfo

I-Trop

Plant & Health Bioinformatics Platform



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



AUORE
COMTE

IE bioinfo



ALEXIS
DEREEPER

IE bioinfo



BRUNO
GRANOULLAC

IE systèmes
d'information



JULIE
ORJUELA

IE bioinfo



NDOMASSI
TANDO

IE systèmes



CHRISTINE
TRANCHANT

IR bioinfo

bioinfo@ird.fr



[@ltropBioinfo](https://twitter.com/ltropBioinfo)

Formations 2022
Montpellier

4-5 Avril

Guide de survie à linux
Agropolis, salle Badiane

19-20 Avril

Linux avancé
Agropolis, salle Badiane

18-19 Mai

**Utilisation avancée
d'un cluster de calcul**
IRD, amphi capmeditrop

14 Juin

**Génomique bactérienne
comparative**
Agropolis, salle Badiane

10 Juin

**Initiation à l'analyse de
données RNAseq**
Agropolis, salle Badiane

30 Mai - 2 Juin

Python
Agropolis, salle Badiane

21-24 Juin

**Analyse de variants
à partir de short and long reads**
Agropolis, salle Bambou

Métagénomique

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
- **tablet**

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

www.southgreen.fr

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



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



Applications :

- Mapper des reads contre un génome *bwa, minimap2*
- Détecter 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*

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



Applications :

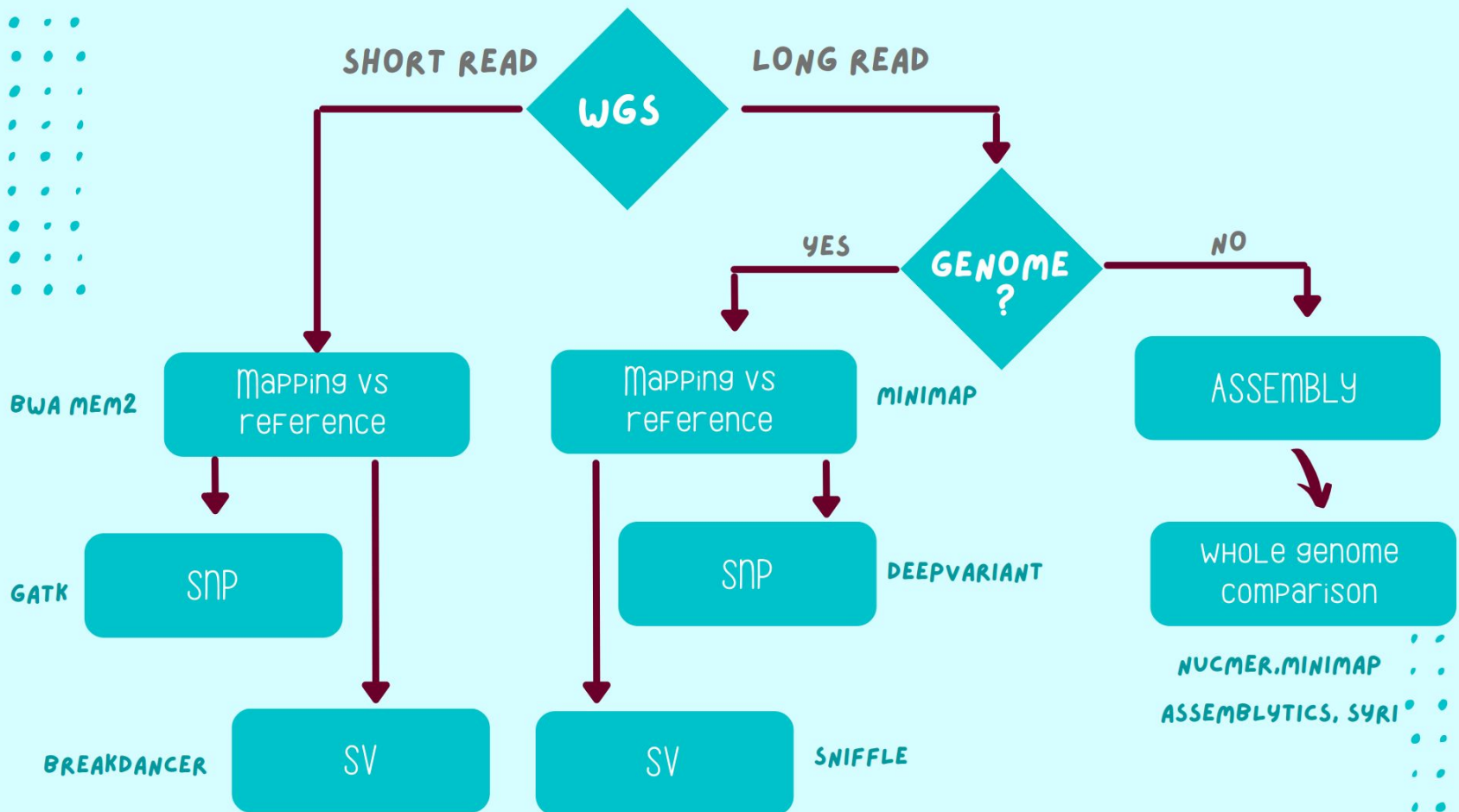
- Mapper des reads contre un génome *bwa, minimap2*
- Détecter des SNPs à partir du mapping de reads - *GATK, deepvariants*
- Analyser les données SNPs brutes (ex: stats, filtres) - *vcftools, bcftools*
- Exemples d'études possibles à partir de SNPs - *SNIPplay*
- Détecter des variants structuraux (SV)



Avec jupyter book : lancer les commandes + analyser les résultats

=> Avoir un plan de bataille opérationnel

SV DETECTION





Let's discover Jupyter through the IFB cloud

Working environment

What is jupyter book ?

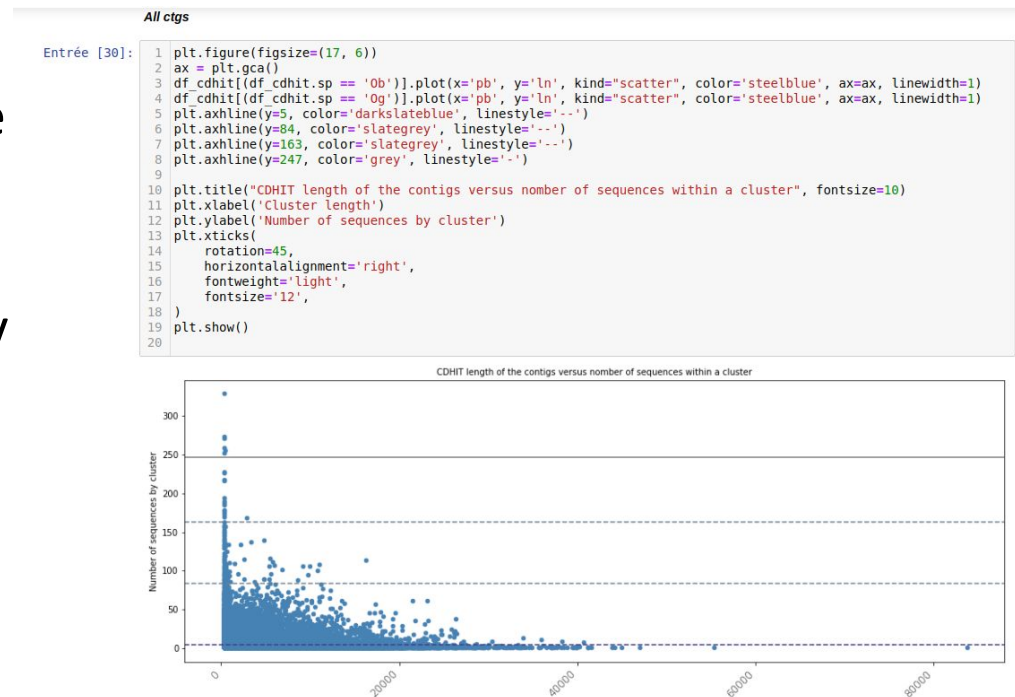
- One of the most popular tool among data scientists to perform data analysis
- Provides a complete environment in which numerous programming languages can be use through a simple web browser

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



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



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 titled 'parseCistr-Copy1'. The interface includes a menu bar (Fichier, Édition, Affichage, Insérer, Cellule, Noyau, Widgets, Aide) and a toolbar with icons for file operations, execution, and navigation. The notebook content is as follows:

Anchoring 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` create the csv file `allCtgsIRIGIN_T0G5681.dedup8095.PANDAS.csv`

Load csv file into a pandaframe

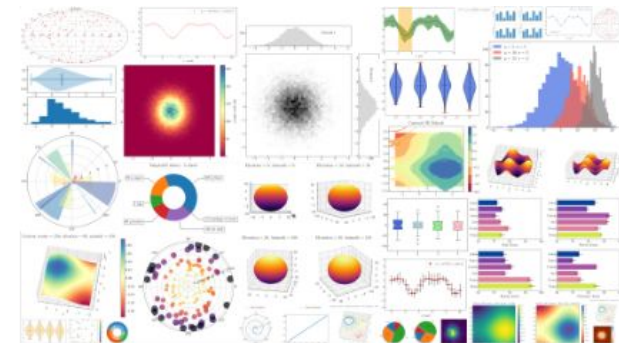
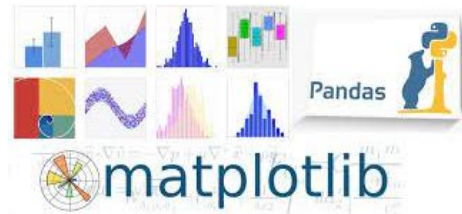
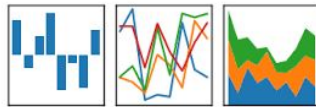
```
Entrée [1]: 1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
4
5 csv_cdhit_file = "/home/christine/Documents/These/Data/CDHIT/ALL_CTGS_MERGE/allCtgsIRIGIN_T0G5681.dedup8095.PAN
6 df_cdhit= pd.read_csv(csv_cdhit_file,names=['ctg','sp','ctg-list','sp-list'], header=0)
7 #print(df_cdhit)
8
```



- 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

- 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”

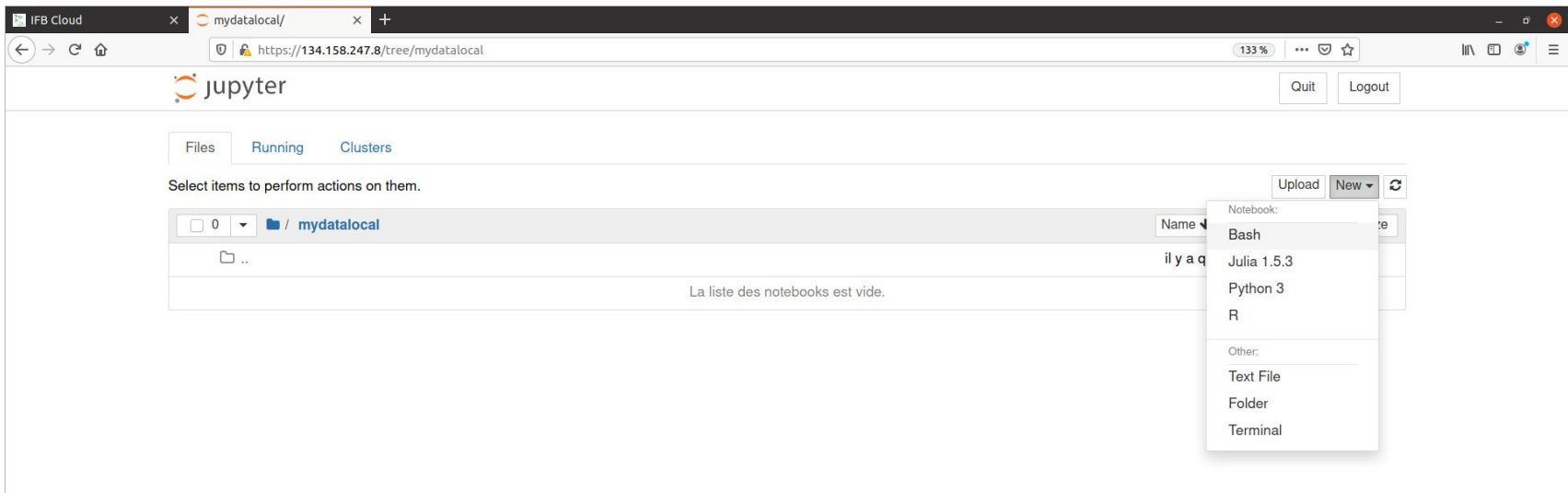


How will you use Jupyter Notebook ?

- Launch our analyses through a jupyter book within a virtual machine launched via the IFB cloud “BIOSPHERE”

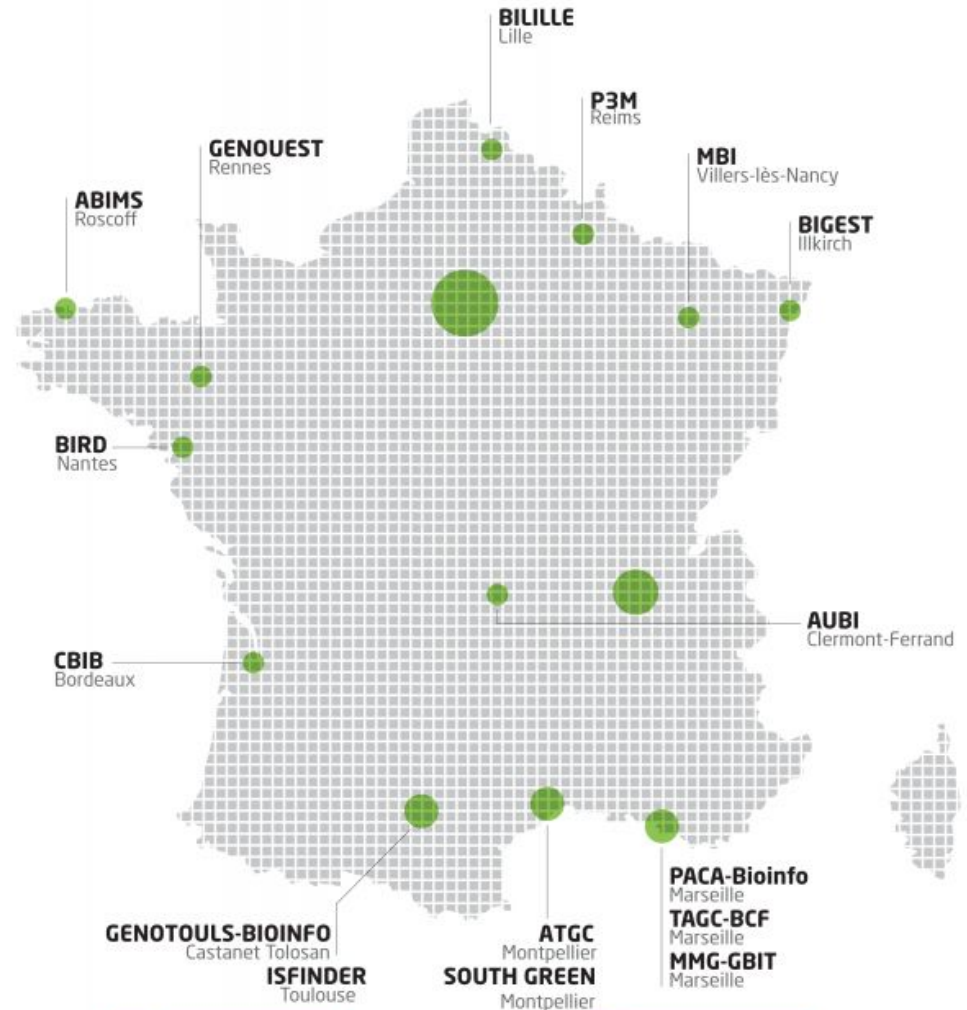


- Through this virtual machine, we will create jupyter books and execute all our analysis





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



RÉGION PARISIENNE

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

C3BI

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

RÉGION LYONNAISE

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

- A federation of clouds, which relies on interconnected IFB's infrastructures, providing distributed services to analyze life science data
- .Access to a large set of virtual machines (computing ressources, bioinformatics tool)
- Used for scientific production in the life sciences, developments, and also to support events like cloud and scientific training sessions, hackathons or workshops.

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

biosphere.france-bioinformatique.fr

https://biosphere.france-bioinformatique.fr/cloudweb/login/?next=/
ifb Biosphere RAINBio myVM DATA Support [en] Sign in

SIGN IN

 **ifb**
INSTITUT FRANÇAIS
DE BIOINFORMATIQUE

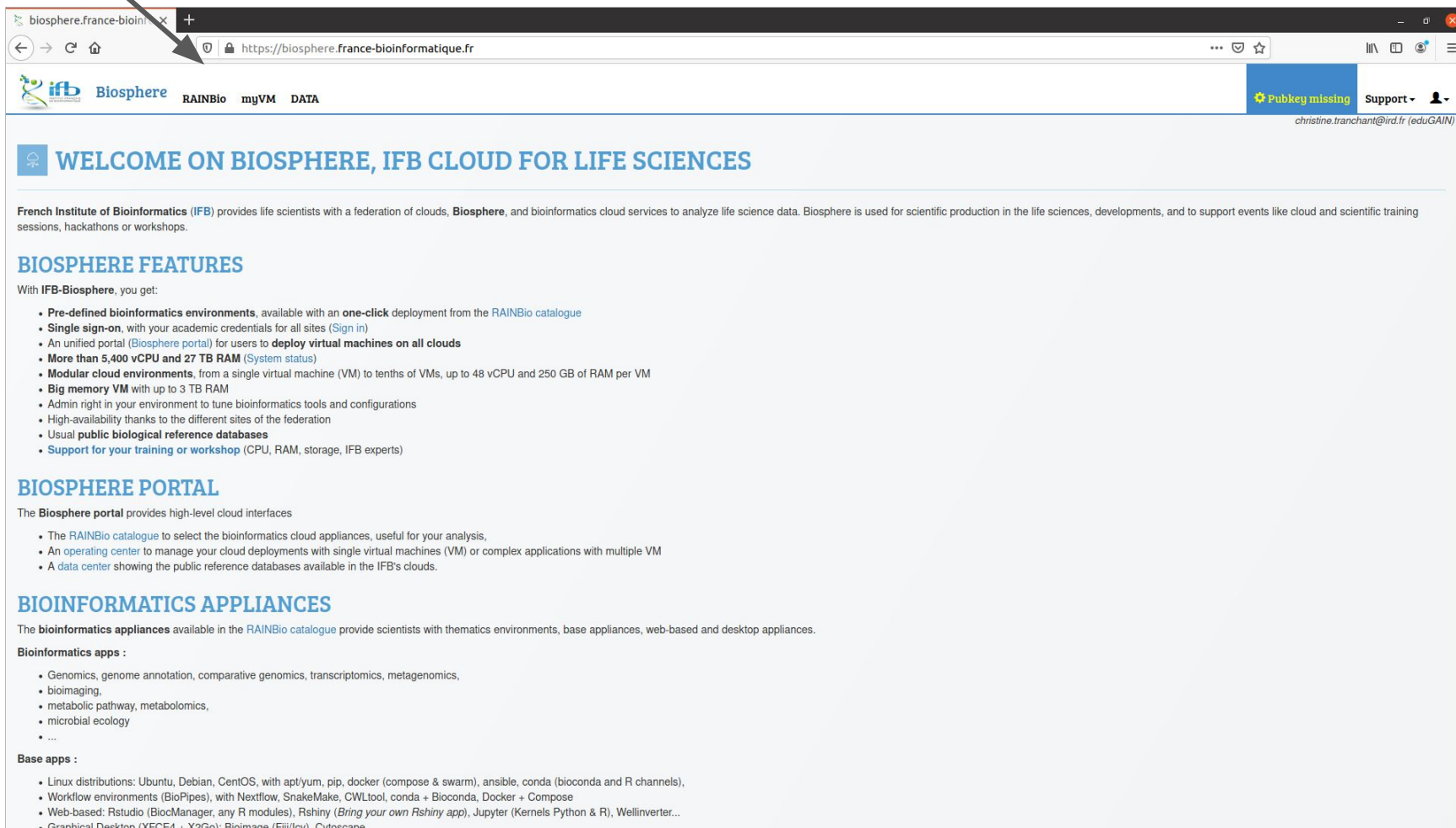
Use your academic credentials (CNRS, INRAE, Inserm, Universities...)

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 URL <https://biosphere.france-bioinformatique.fr>. The page features a navigation bar with links for **RAINBio**, **myVM**, and **DATA**. A blue arrow points to the **RAINBio** link. The main content area includes a welcome message, a list of features, and sections for the portal and appliances.

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/ivd), Cytoscape

vm's name : **analysesSV**

IFB Biosphere RAINBio myVM DATA

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

RAINBIO - APPLIANCES BIOINFORMATIQUES DANS LE CLOUD

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

App Store (4) Appliances Outils Topics Appliance éditables Ajouter

AnalysesSV

- DEV
- bcftools, BEDTools, BWA, Jupyter, Matplotlib, pandas,
- DNA polymorphism, Genetic variation, Genomics

CoursAnalysesNanoporeSG

- bandage, Jupyter
- Data architecture, analysis and design, Mathematics, Statistics and probability

NGSanalysisJupyter

- BEDTools, BWA, Jupyter, SAMtools
- Data architecture, analysis and design, Mathematics, Statistics and probability

REPET

- Repet
- Bioinformatics

analyses

Le code couleur reste le même pour une même appliance.

cea cnrs INRAE Inserm INVESTISSEMENTS D'AVENIR

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

Developpers

- Francois Sabot [SouthGreen Platform](#)
- Julie Orjuela-Bouniol [SouthGreen Platform](#)

App data

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

Licence

Licensed under GPLv3

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

Outils

[bcftools](#) [BEDTools](#) [BWA](#) [Jupyter](#) [Matplotlib](#) [pandas](#) [SAMtools](#)

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

Caractéristiques

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

Crédits

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

LANCER

⚡ LANCER

🚀 DÉPLOIEMENT AVANCÉ

Appliance AnalysesSV

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 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

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

App data

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

Licence

Licensed under GPLv3

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

Outils

LANCER

LANCER

DÉPLOIEMENT AVANCÉ

Configurer le déploiement d'une appliance

Déploiement de l'appliance "AnalysesSV"

Name

CTranchant

Groupe à utiliser

DIADÉ (Diversité, Adaptation)

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

vCPU.h

Cloud

ifb-core-cloudbis

Gabarit d'image cloud

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

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

ifb.m4.large (2 vCPU, 8Go GB RAM, 50Go GB local disk)

ifb.m4.xlarge (4 vCPU, 16Go GB RAM, 100Go GB local disk)

ifb.m4.2xlarge (8 vCPU, 32Go GB RAM, 200Go GB local disk)

ifb.m4.4xlarge (16 vCPU, 64Go GB RAM, 400Go GB local disk)

ifb.x1e.4xlarge (BigMem) (16 vCPU, 384Go GB RAM, 600Go GB local disk)

ifb.m4.6xlarge (24 vCPU, 96Go GB RAM, 600Go GB local disk)

ifb.m4.8xlarge (32 vCPU, 128Go GB RAM, 800Go GB local disk)

ifb.x1e.8xlarge (BigMem) (32 vCPU, 768Go GB RAM, 600Go GB local disk)

ifb.m4.12xlarge (48 vCPU, 192Go GB RAM, 1.2To GB local disk)

ifb.x1e.12xlarge (BigMem) (48 vCPU, 1.1To GB RAM, 50Go GB local disk)

ifb.m4.14xlarge (56 vCPU, 240Go GB RAM, 1.4To GB local disk)

ifb.x1e.16xlarge (BigMem) (62 vCPU, 1.5To GB RAM, 1.5To GB local disk)

ifb.x1e.32xlarge (BigMem) (124 vCPU, 2.9To GB RAM, 2.9To GB local disk)

Annuler

Loading...

CLOUD

Déploiements



☐	ID	Nom		Début	Groupes	Spécification	Broker	Cloud	Accès
☐	19435	AnalysesSV (1.0) DEV CTranchant		↑ Jui 15 2022, 16h54	DIADÉ	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 !

CLOUD

Déploiements

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

Arrêter les déploiements

Tout voir (1)

get the url... link "https"

CLOUD

Déploiements

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

Arrêter les déploiements

Tout voir (1)

Get the token identifiant... link “Params”

The screenshot shows the SouthGreen bioinformatics platform interface. A modal window titled "Paramètres" is open, displaying a table with two columns: "nom" and "valeur". The table contains one row with the name "JUPYTER_TOKEN" and the value "7b708d2b24d1947f0a48baf37f0986c9ca4774d8157a8825".

In the background, the main interface is visible. The top navigation bar includes "IFB Biosphere", "RAINBio", and "myVM". The right side shows a "Clé publique (PubKey) absente" status and a "Support" link. The user's email "christine.tranchant@ird.fr (eduGAIN)" is displayed.

The main content area is titled "CLOUD" and features a "Déploiements" section. Below this, there is a table of deployments with columns: ID, Nom, Début, Groupes, Spécification, Broker, Cloud, and Accès. The table shows one deployment with ID 19435, named "AnalysesSV (1.0) DEV" by "CTranchant". The deployment started on "Jui 15 2022, 16h54" under the "DIADE" group. The specification is 64x16, and the broker is "1e82". The cloud provider is "ifb-core-cloudbis". The access URL is "https Params 134.158.248.237".

Below the deployment table, there is a button "Arrêter les déploiements" and a link "Tout voir (1)".

At the bottom, there is a section for "Appliances et déploiements favoris" with tabs for "Déploiements récemment terminés" and "Quota". Below this is a table with columns: ID, Broker, Nom, Der. dém., and Paramétrage.

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

IFB Cloud x Accueil x +

← → ↻ 🏠 🔒 https://134.158.247.8/tree 133 % ... ☆

jupyter Quit Logout

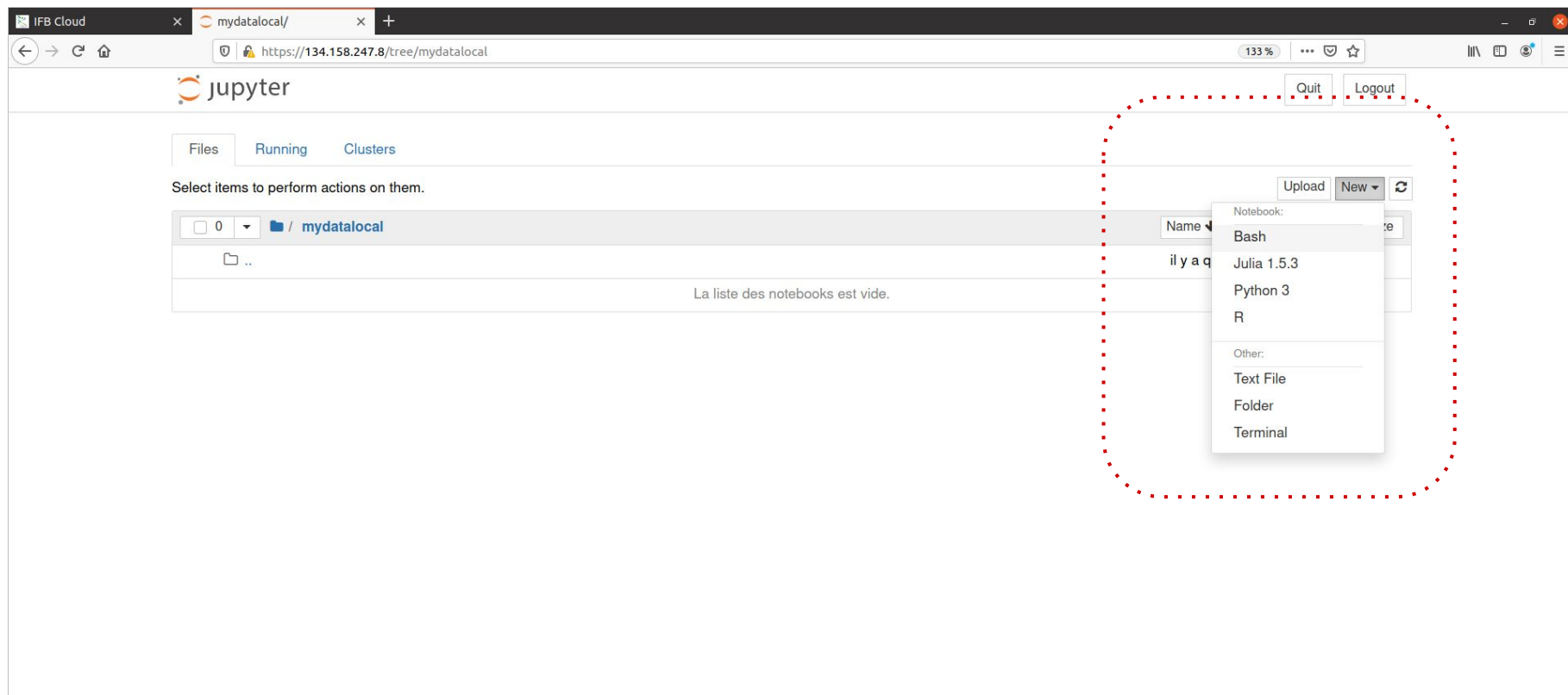
Files Running Clusters

Select items to perform actions on them. Upload New ↻

☐	0 ▾	📁 /	Name ▾	Last Modified	File size
☐		📁 mydatalocal		il y a un jour	
☐		📁 public		il y a 3 mois	

1079760dd85a9bbc43e10354 ^ ▾ Tout surligner Respecter la casse Respecter les accents et diacritiques Mots entiers Expression non trouvée x

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



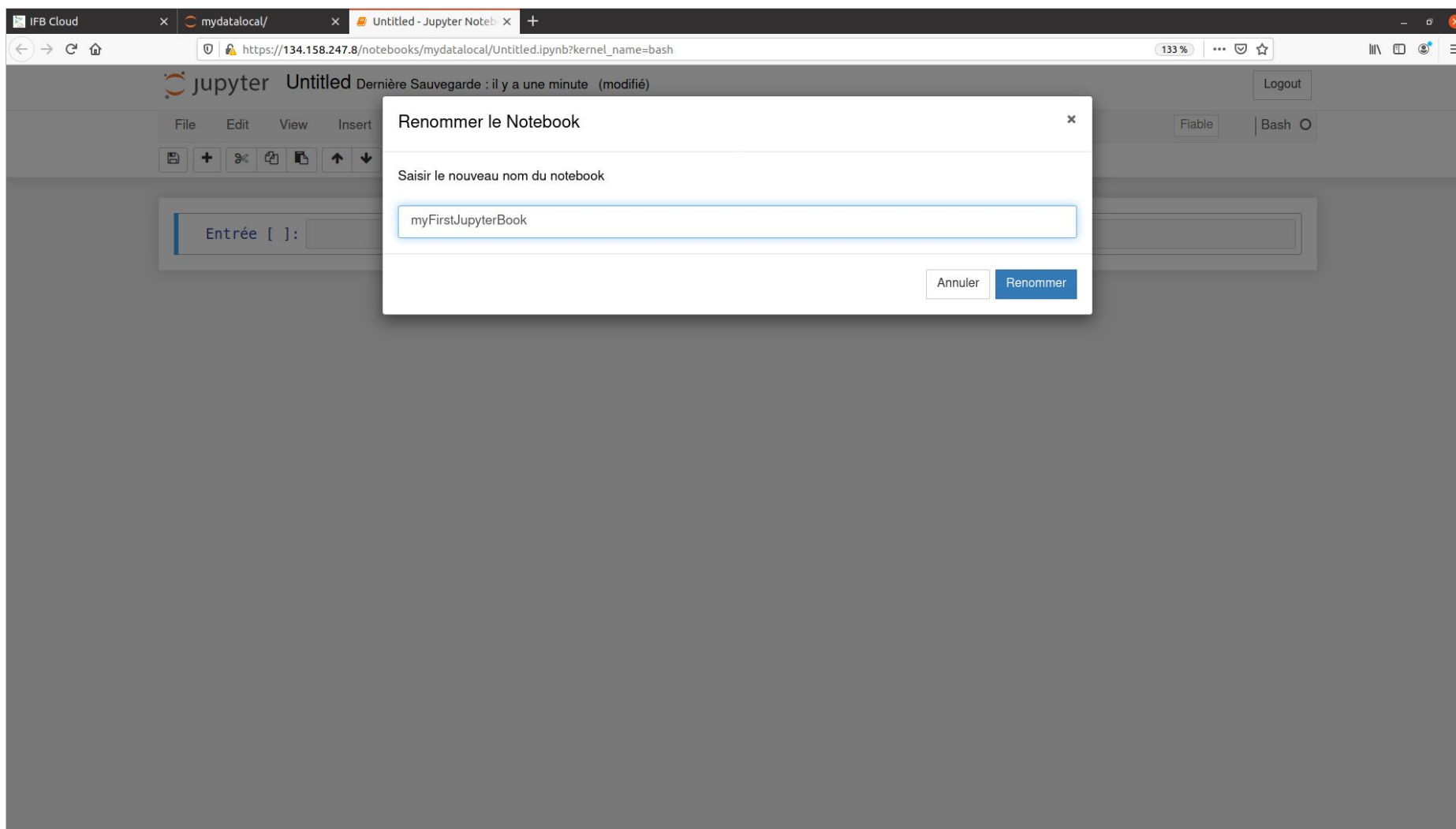
The screenshot shows the JupyterLab web interface in a browser window. The address bar displays the URL `https://134.158.247.8/tree/mydatalocal`. The interface includes a top bar with the Jupyter logo and navigation tabs for 'Files', 'Running', and 'Clusters'. Below the tabs, there is a message 'Select items to perform actions on them.' and a file browser showing the directory structure. A red dashed rectangle highlights the 'New' dropdown menu, which is open and shows the following options:

- Notebook:
 - Bash
 - Julia 1.5.3
 - Python 3
 - R
- Other:
 - Text File
 - Folder
 - Terminal

The 'Bash' option is highlighted, indicating it is the selected kernel for creating a new notebook.

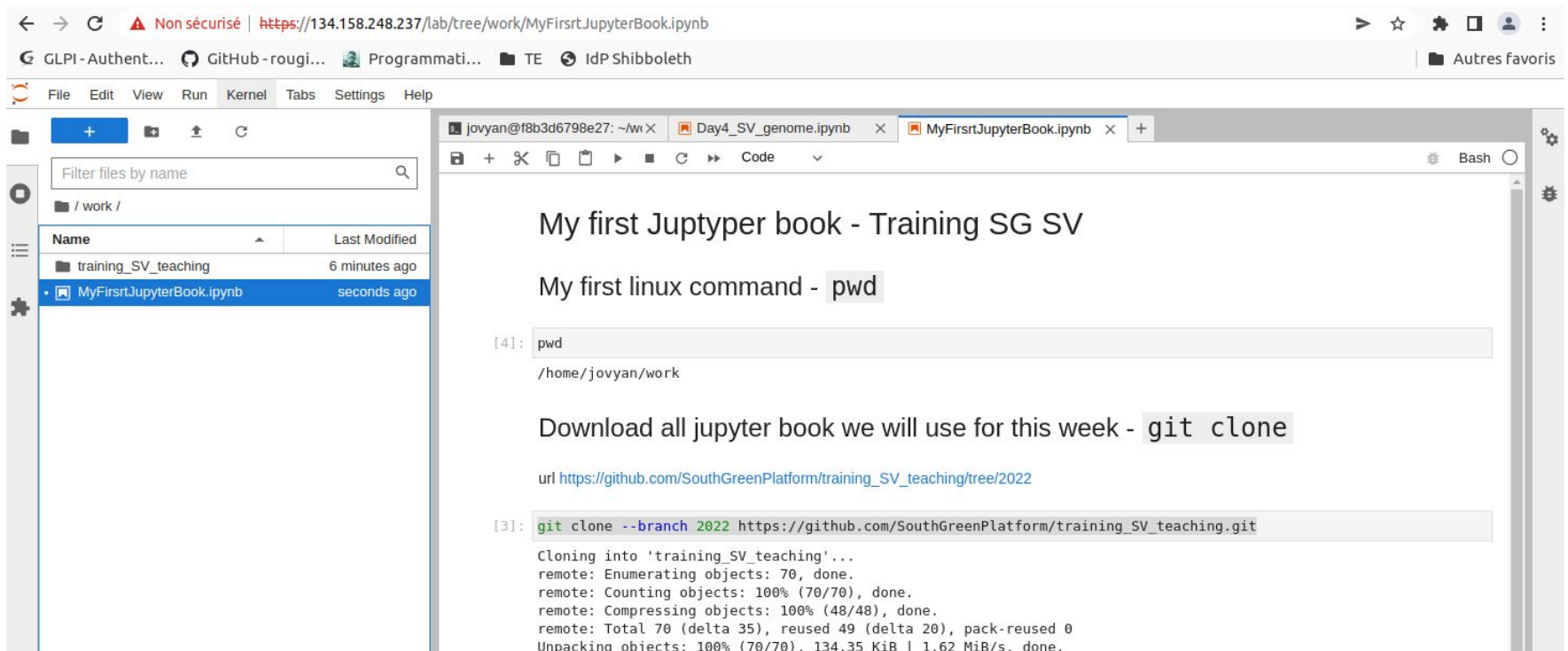
Rename your first jupyter book

myFirstJupyterBook



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

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



Non sécurisé | <https://134.158.248.237/lab/tree/work/MyFirstJupyterBook.ipynb>

GLPI - Authent... GitHub - rougi... Programmati... TE IdP Shibboleth

Autres favoris

File Edit View Run Kernel Tabs Settings Help

Filter files by name

/ work /

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

My first Juptyper book - Training SG SV

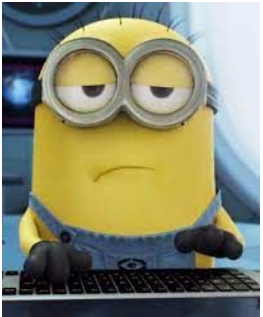
My first linux command - `pwd`

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

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

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

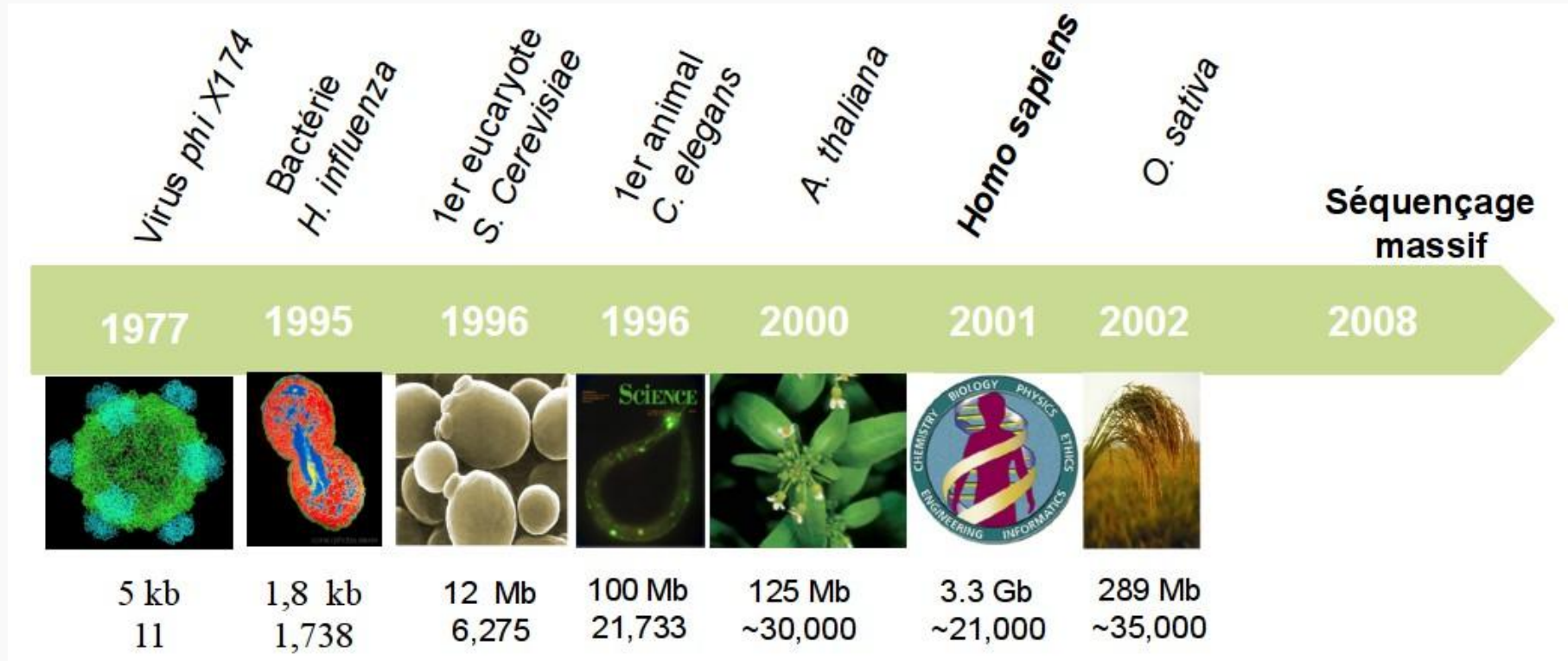
```
[3]: git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git
Cloning into 'training_SV_teaching'...
remote: Enumerating objects: 70, done.
remote: Counting objects: 100% (70/70), done.
remote: Compressing objects: 100% (48/48), done.
remote: Total 70 (delta 35), reused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



Introduction & NGS method

The NGS in themselves

A little history of sequencing...



...From Data Rarity to Data Deluge

Decoding for cents

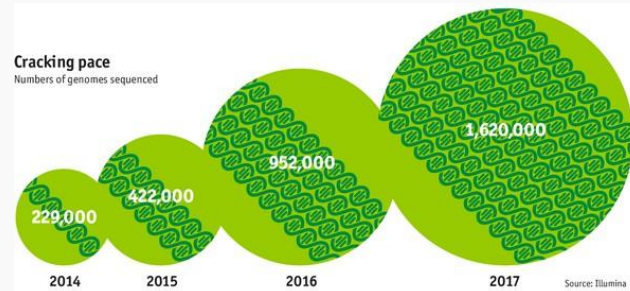
Cost of determining 1m bases of a DNA sequence
Log scale, \$



Source: National Human Genome Research Institute

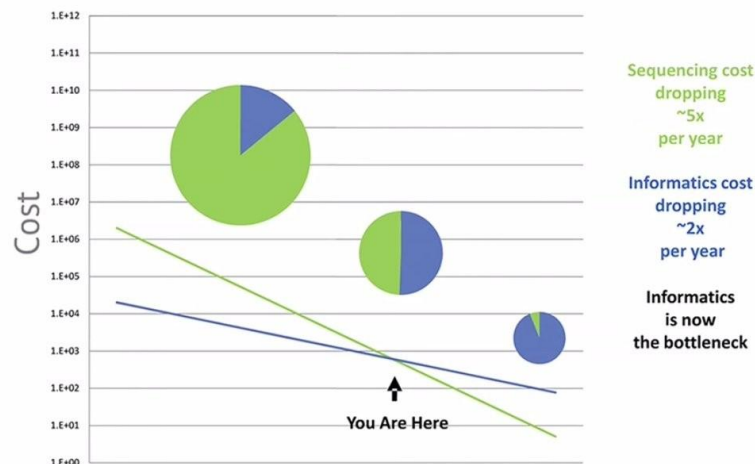
Cracking pace

Numbers of genomes sequenced



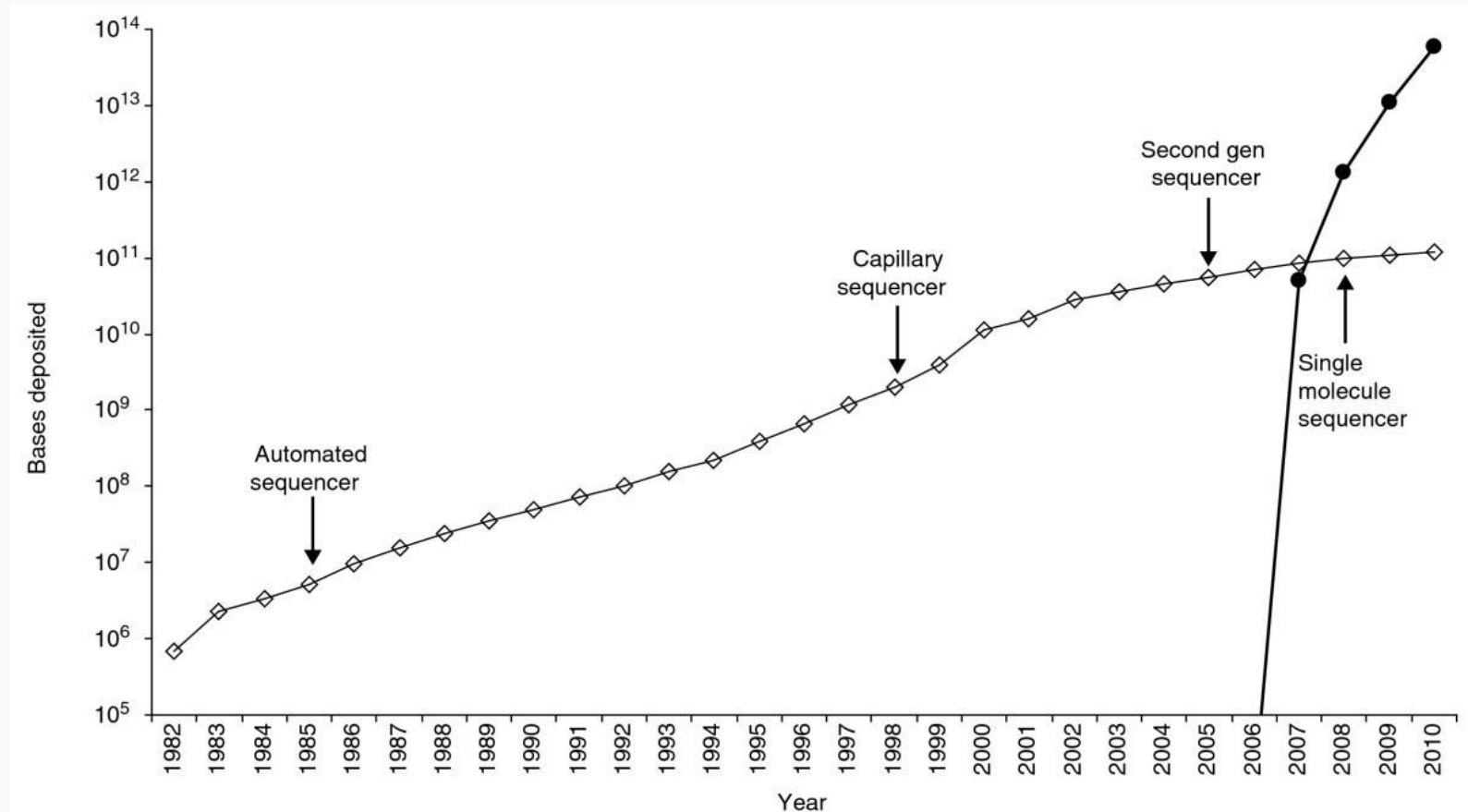
From The economist

DNA Sequencing Economics



From Bussiness
Insider

...From Data Rarity to Data Deluge



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

Methods

2nd Generation Sequencing

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

2nd Generation Sequencing

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

3rd Generation Sequencing

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

2nd Generation Sequencing

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

454

IonTorrent

Illumina

3rd Generation Sequencing

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

2nd Generation Sequencing

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

454

IonTorrent

Illumina

3rd Generation Sequencing

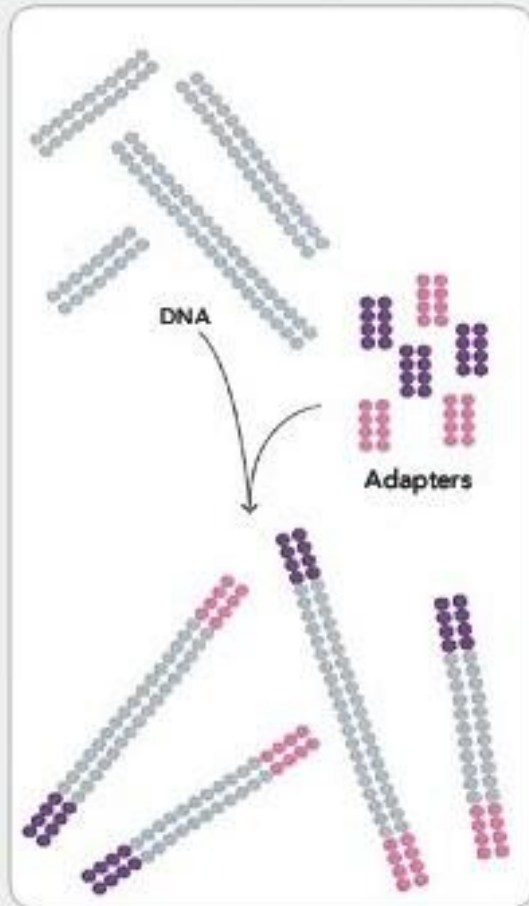
PacificBiosciences
Oxford
Nanopore

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



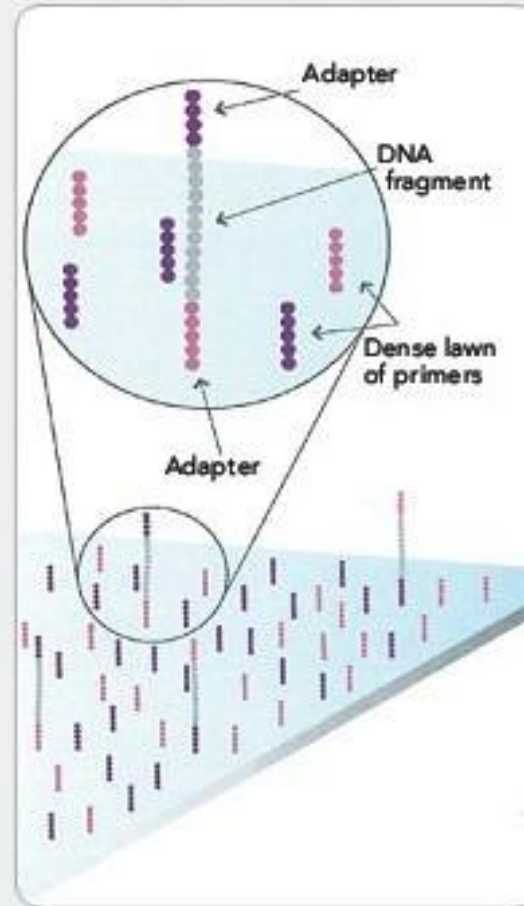


1. PREPARE GENOMIC DNA SAMPLE



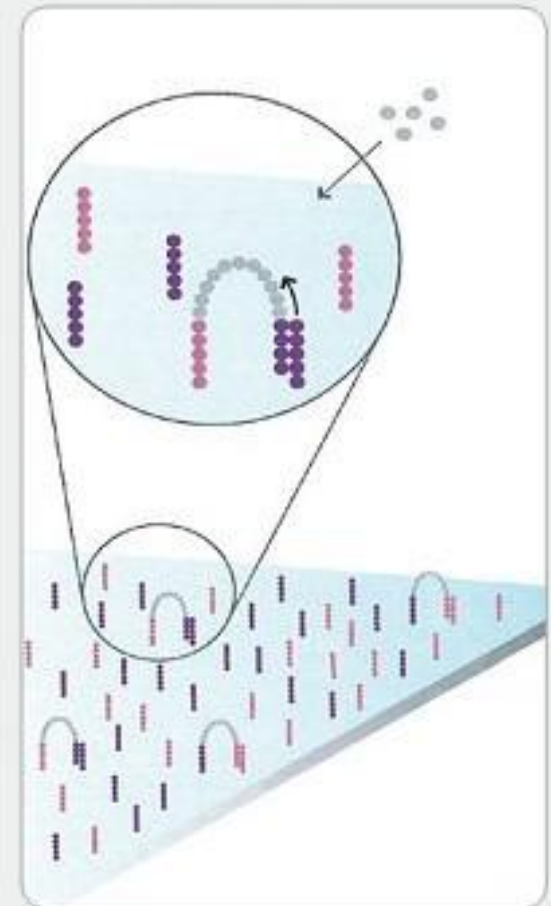
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



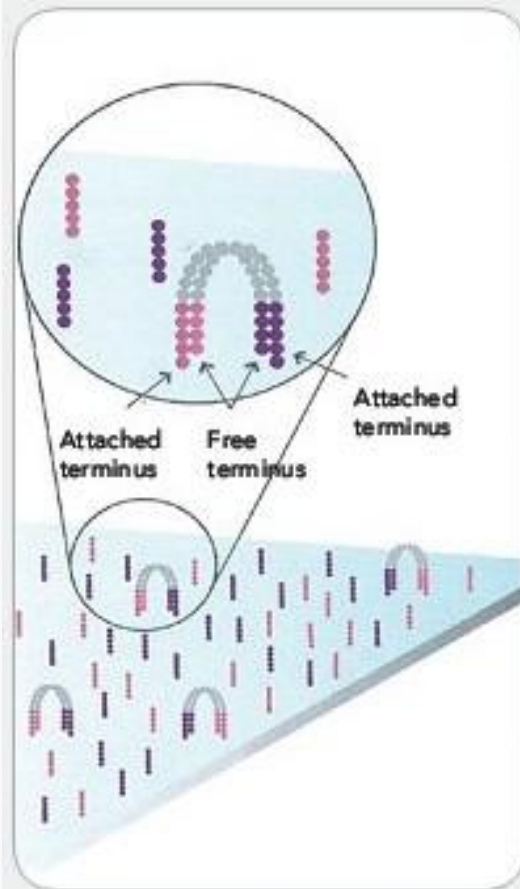
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION



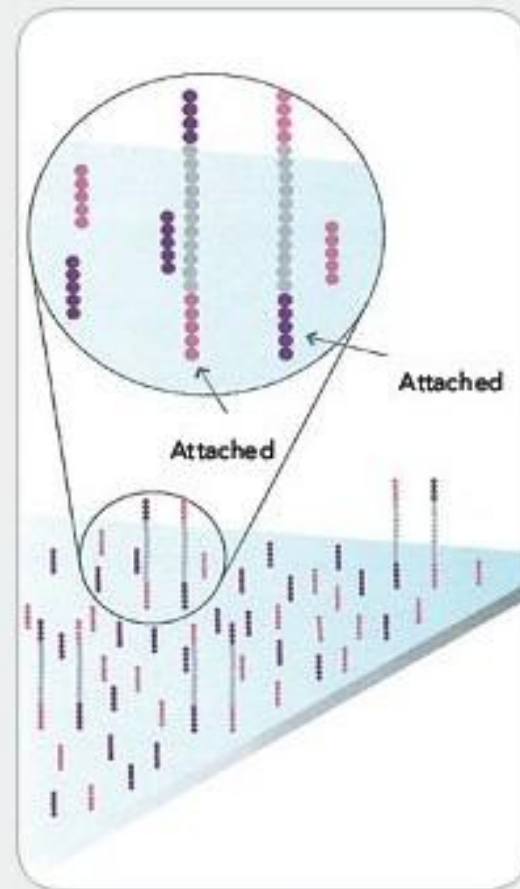
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED



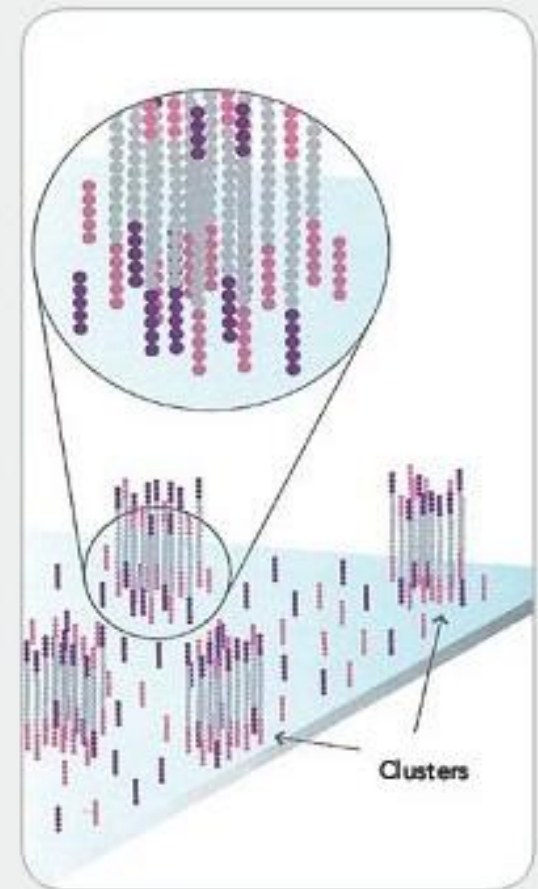
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



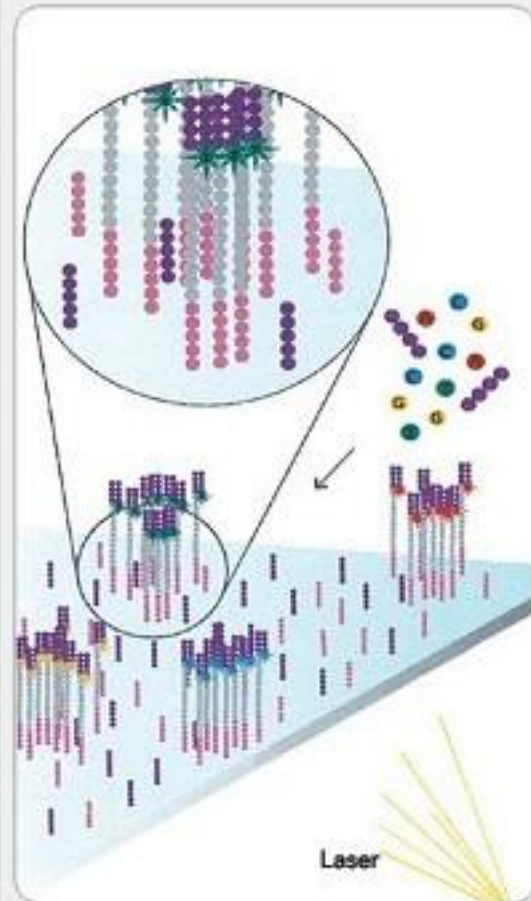
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



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

7. DETERMINE FIRST BASE



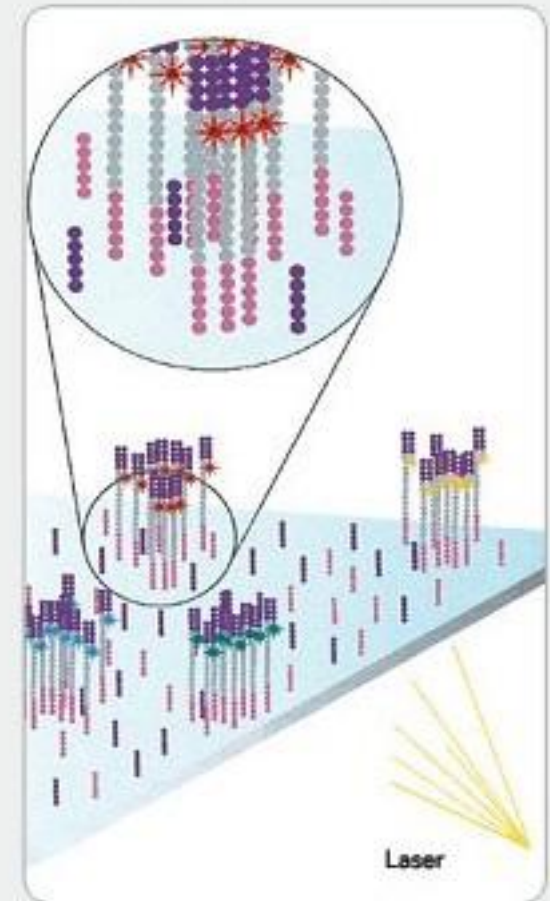
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



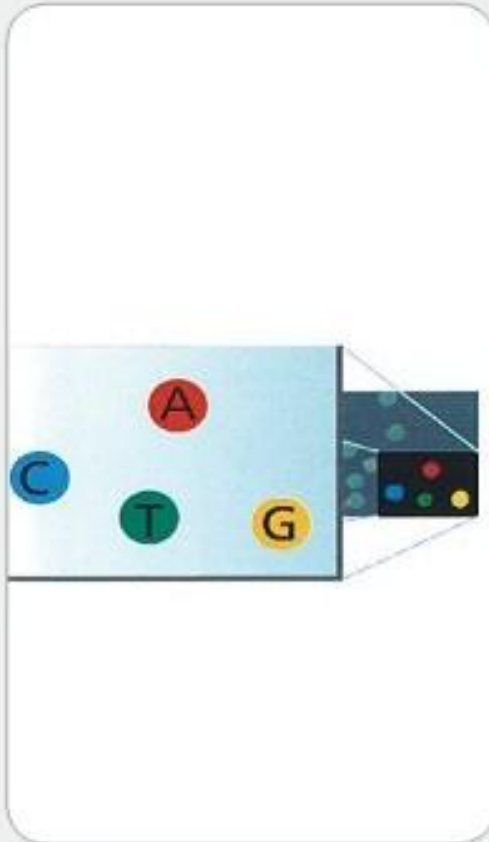
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



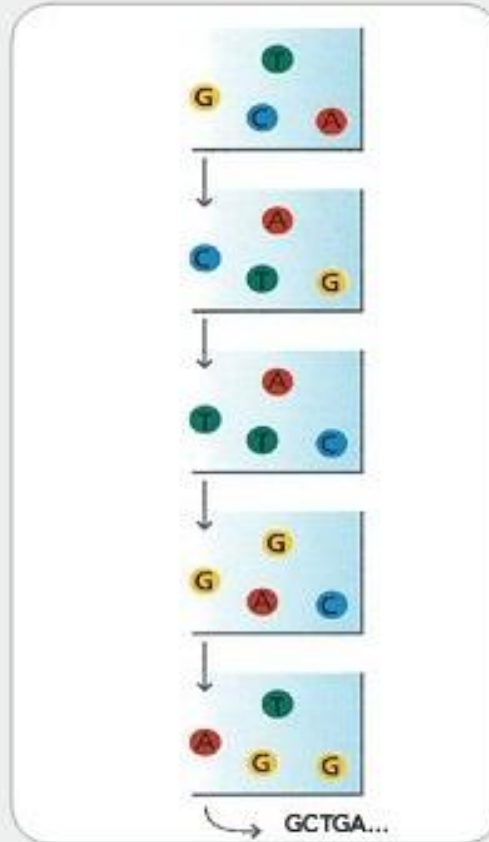
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



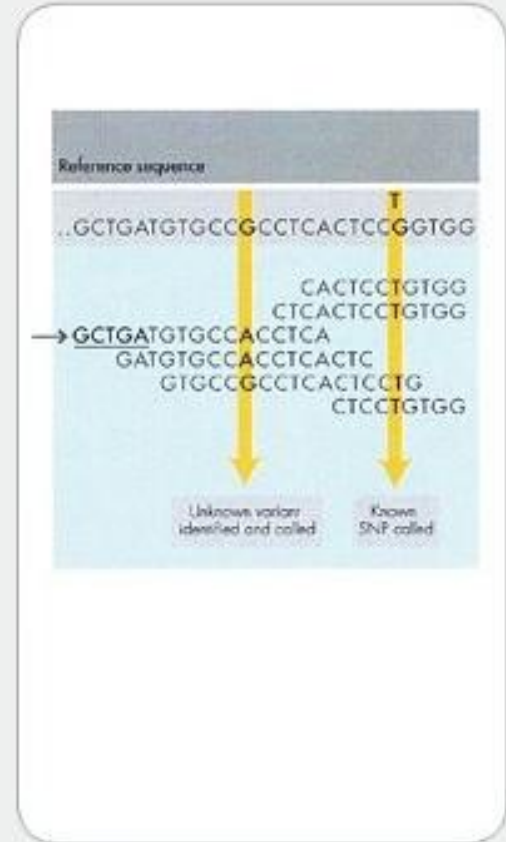
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

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)

```
@H4:C7C99ACXX:6:1101:1360:74584/2
CTGTTTCTTAGTATTTTGTAGTCATTCCGTGTTGGTTTAGTTGCAAGGT
+
@@@DADFFHHFFHHIEFEIGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTTGGT
+
@@@DDDD>FFFAFBEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTGCAGTAACGACCTATAAATTTTAAAGTAGC
+
@CFFFFFGGHHHHIJJJIEE<HHHIJJIGBHGGEEIJJIEIJIHHJFIJJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAGATTGCGGTATGGCTCTAGTAGCTTTTGTCTTAT
+
@C@FFFFFGGHHHDHGIIEEHIII<CGHIJJIIJ:?FC9DGAFGHII?DGBFIIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFGHHHHJJIIJJJIGJCHHEIII>GHIG@GHIDHGJIIFHIJJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGTAAGAGTTAAAAACAATTAATGAGCAACTGAGTTC
+
@@CFFFFFHFFHFGIJJIIHHIJJIIHHIJJECGHIJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTTGTGCGGGGTTGGGAAATTGTCACTTTCTGCTACAATGCCG
+
@<@DDDDDHFFFDIIBDFGHHGG;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

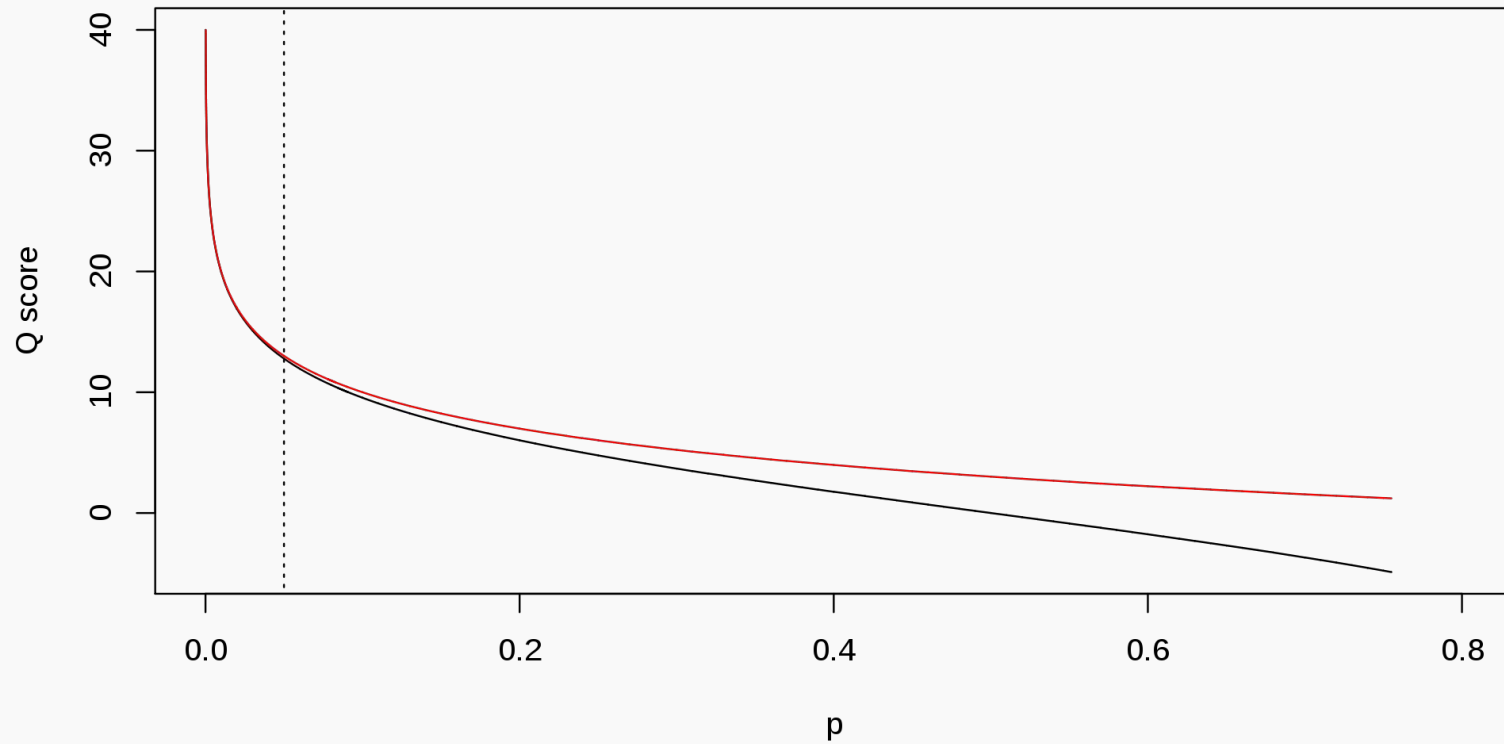
Phred
Quality
Score

0 .. 50

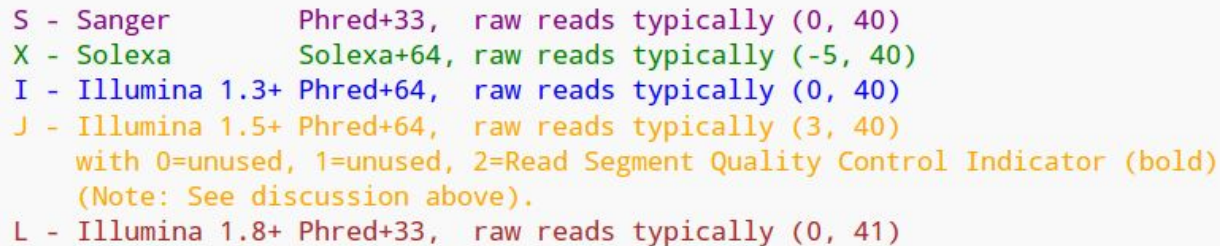
Code ASCII



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



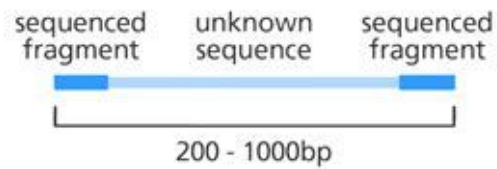
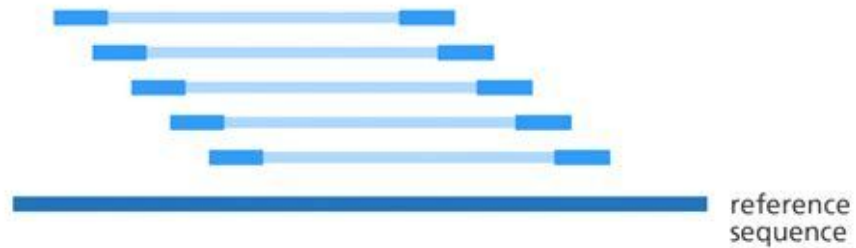
**Institut de Recherche
pour le Développement**
FRANCE



Single-end reads



Paired-end reads



```
@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> .
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

```
@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>

```
@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
```

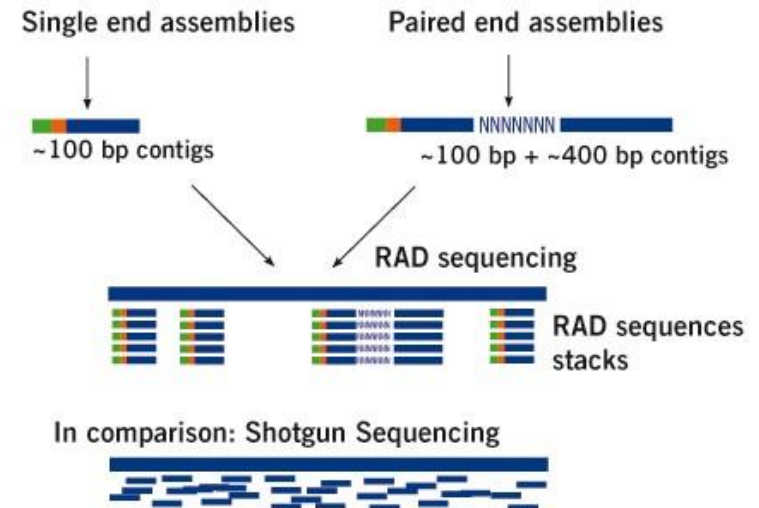
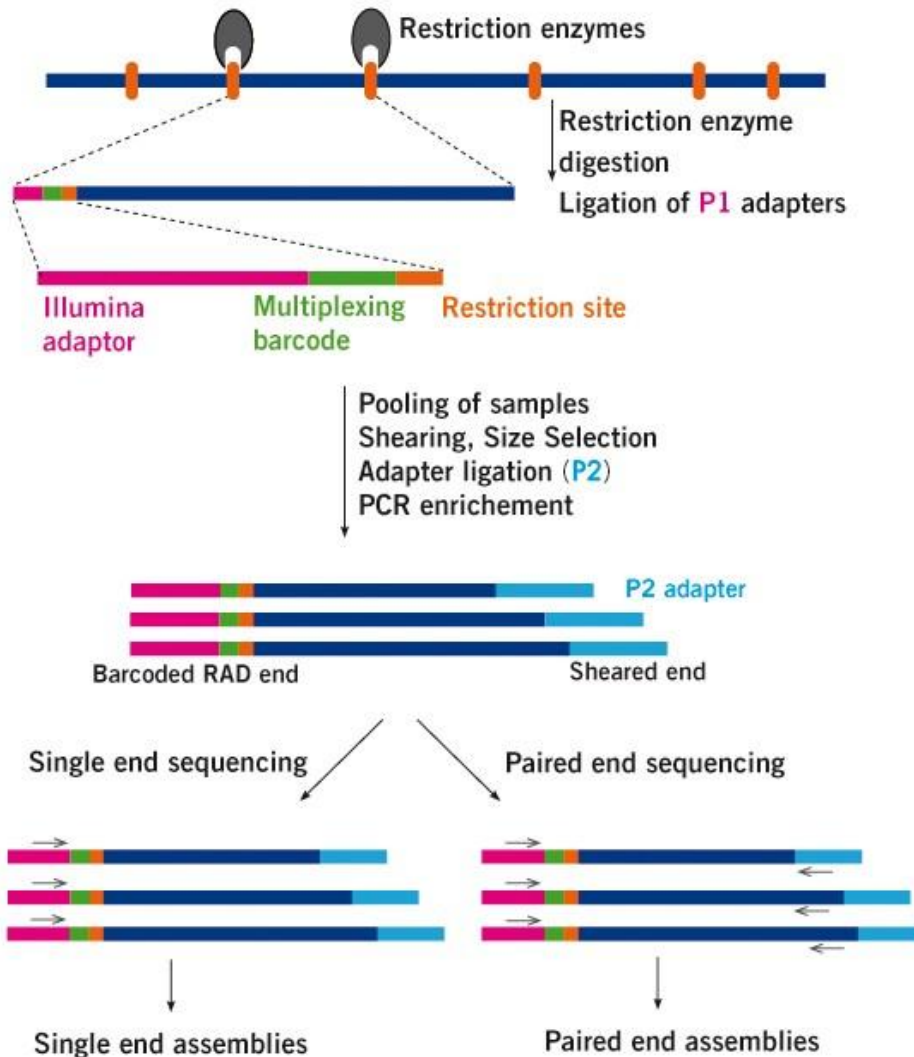
```
r003 16 ref 29 30 6H5M * 0 0 TAGGC * N
```

```
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT
```

alignement

Format tabulé

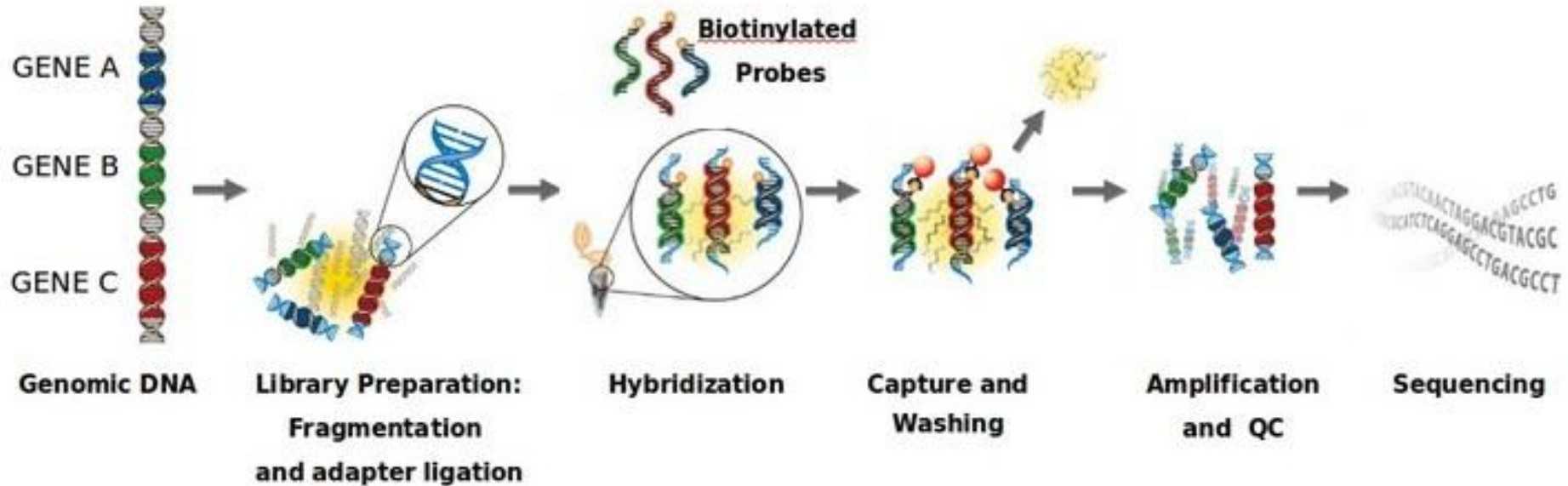
Col	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string (operations: MIDNSHP)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



DNA is digested with a specific set of restriction enzymes and two adaptors ligated to the fragments afterwards.

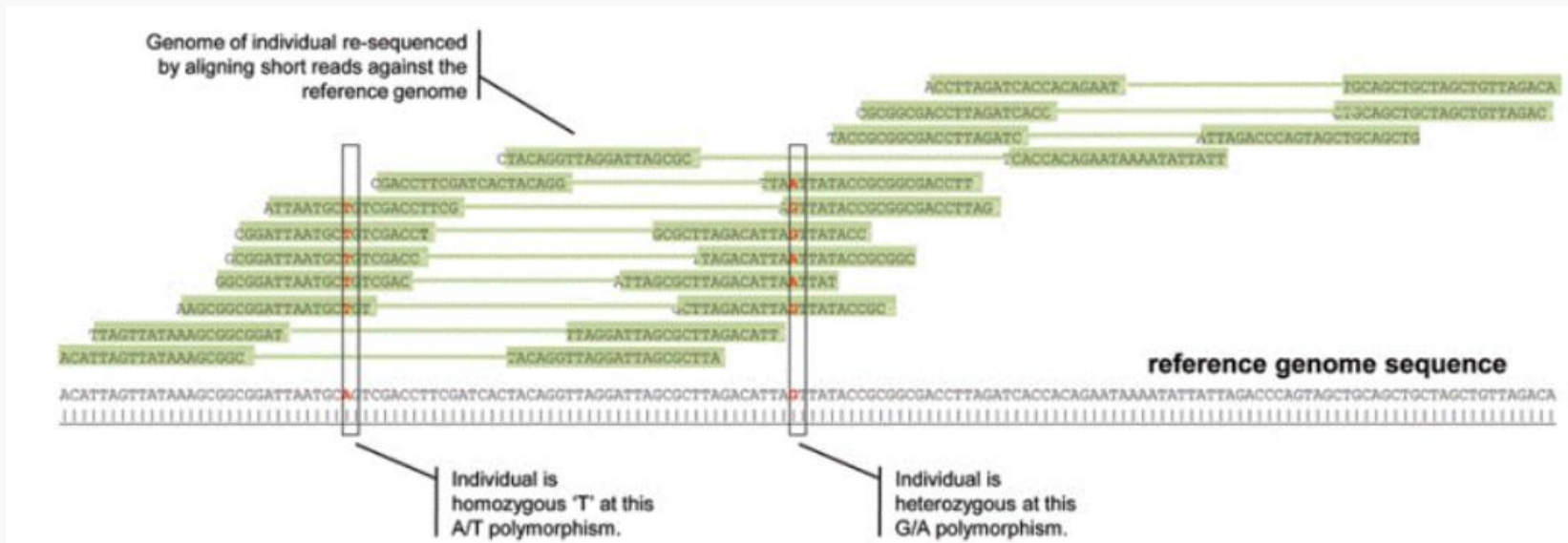
Sequencing the fragments from either one side or both sides results in RAD sequences stacks.

The stacks are the perfect starting point to identify and analyse genetic markers.



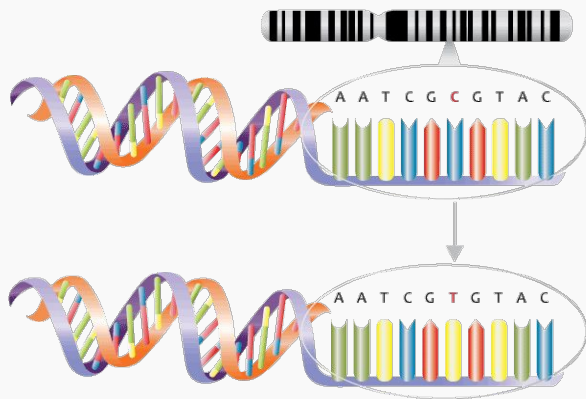
From CGFB, Bordeaux, France

SNP and InDel Detection



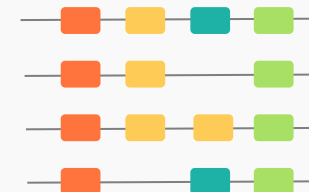
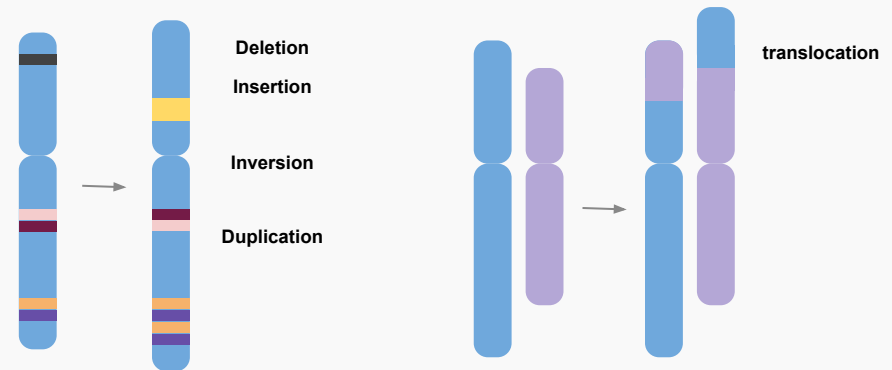
Mutations & Variations as main source of genetic diversity

SNP



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

Structural Variations



Presence Absence Variation

Common File for all Variations, the VCF

Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##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">
```

Body

#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		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Annotations:

- Mandatory header lines:** ##fileformat=VCFv4.0
- Optional header lines (meta-data about the annotations in the VCF body):** ##fileDate=20100707, ##source=VCFtools, ##reference=NCBI36, ##INFO=, ##FORMAT=, ##ALT=, ##INFO=
- Reference alleles (GT=0):** A, T, G, T
- Alternate alleles (GT>0 is an index to the ALT column):** AT, CT, G,
- Deletion:** T to
- SNP:** C to T
- Large SV:** A to G
- Insertion:** T to CT
- Other event:** A to AT
- Phased data (G and C above are on the same chromosome):** 0|1:100, 1|0:77

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

GWAS Diagram Browser

Exploring Genome-wide Association Studies



National Human
Genome Research
Institute

EMBL-EBI



Filter:

Clear filters

GWAS Diagram

Time Series View

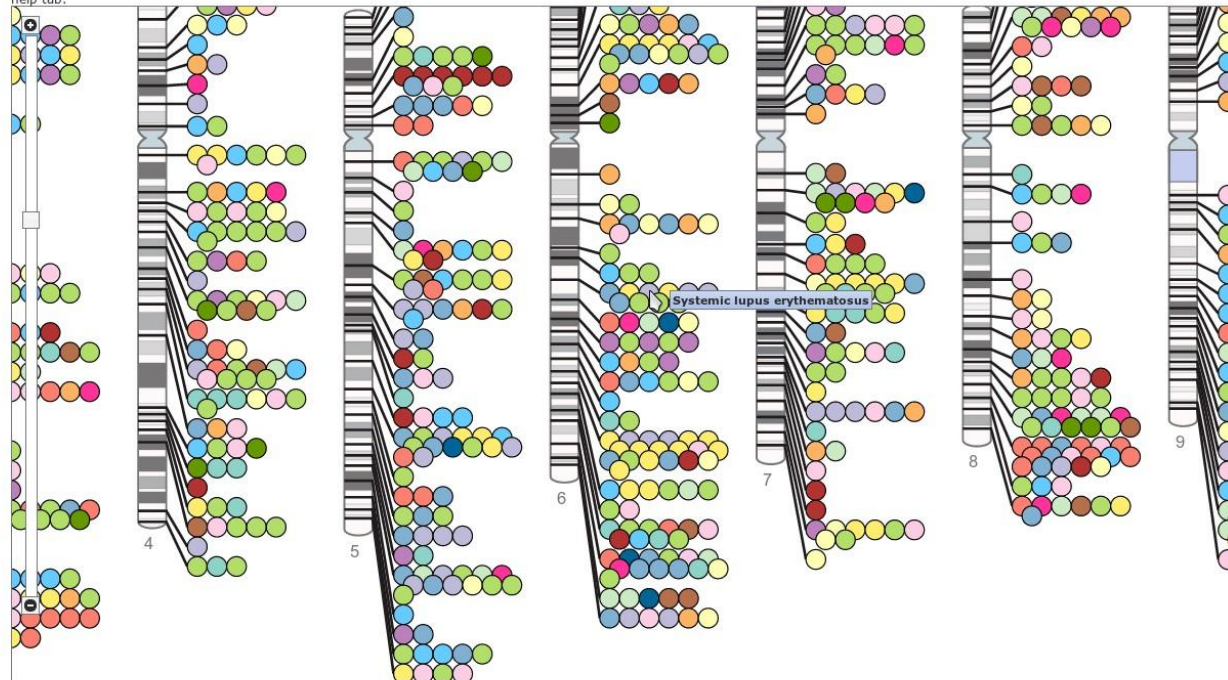
Downloads

Help

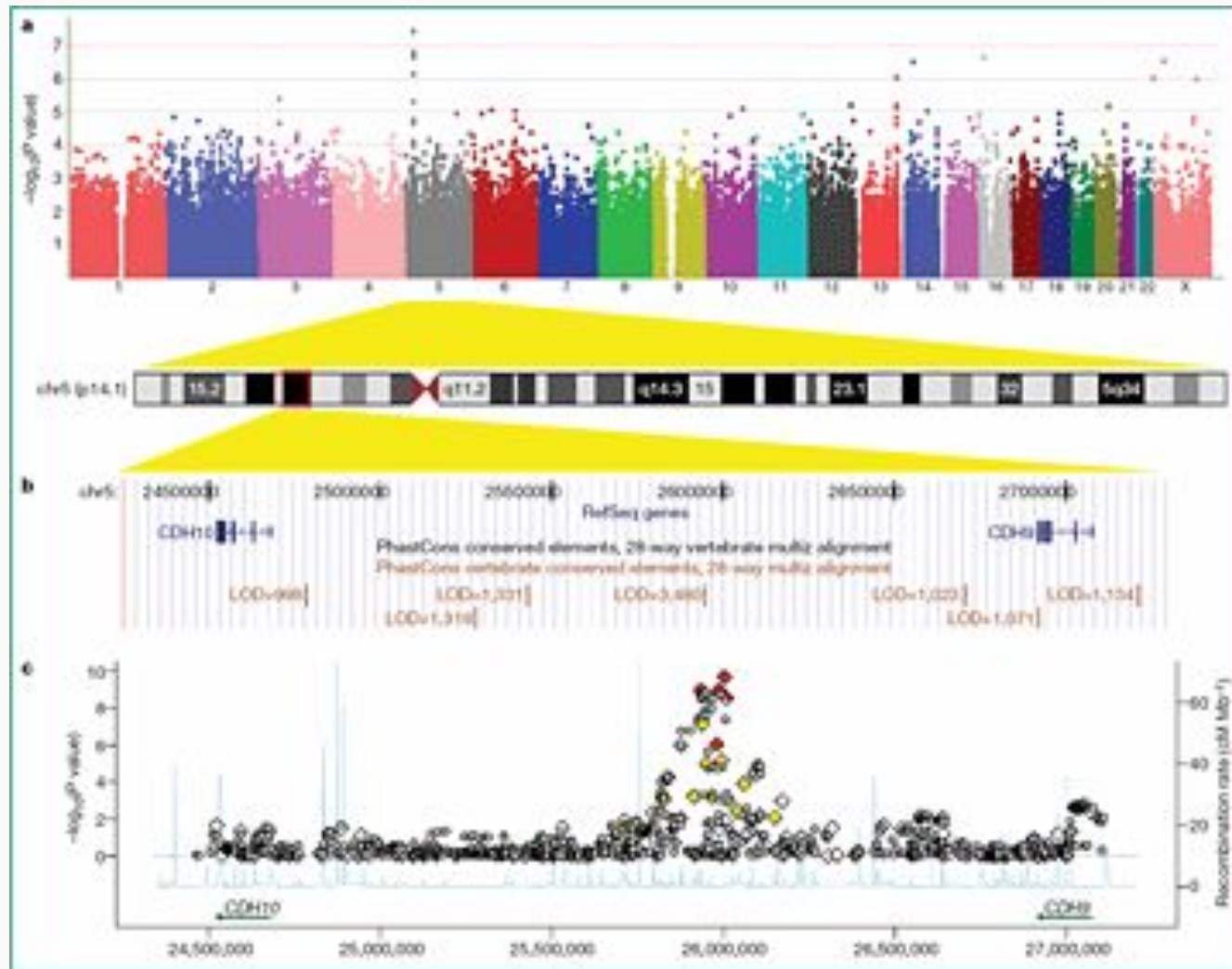
About

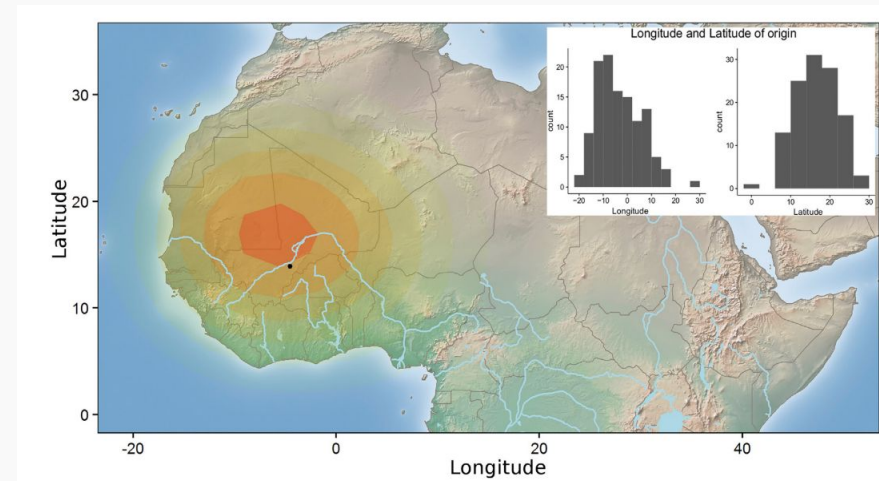
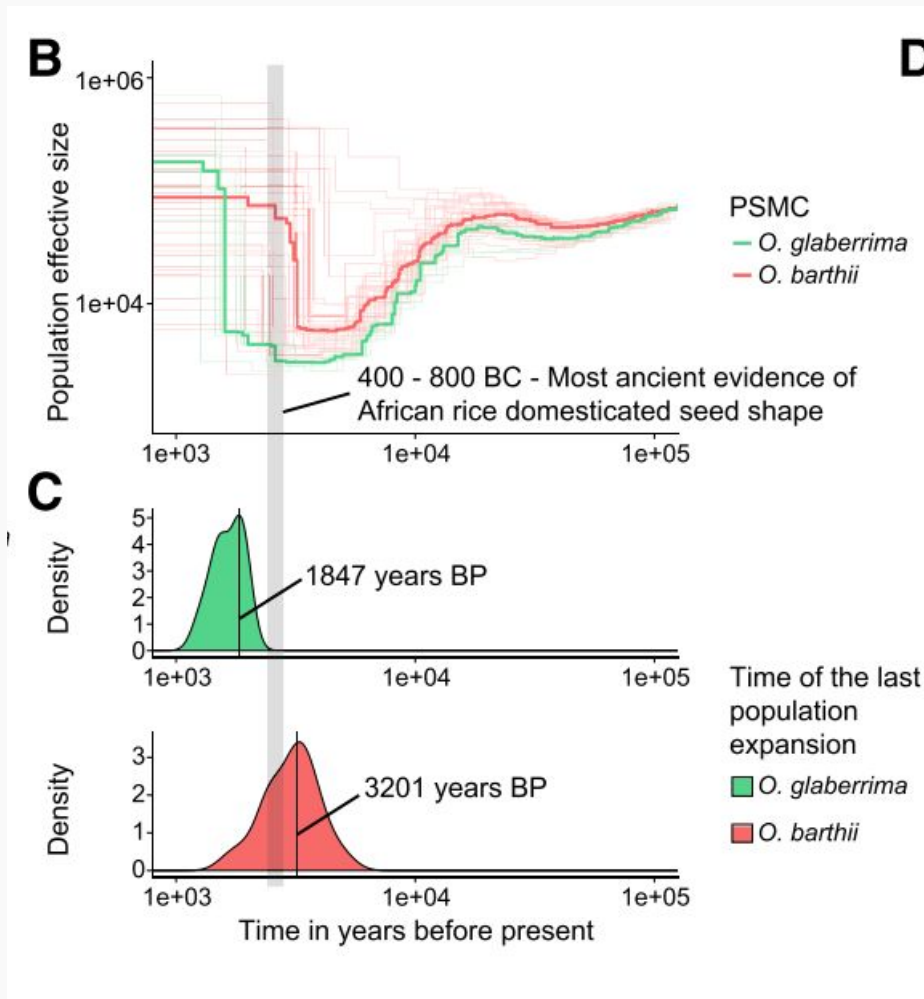
Show Legend

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



Example in GWAs & Population Genomics





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

From Cubry et al, 2018



1000 Genomes
A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants Contact

LATEST ANNOUNCEMENTS

WEDNESDAY FEBRUARY 16, 2011
February 2011 Data Up
Full Project Indel Release
Indels calls from **Dindel**. These calls genomes project. This release is ba
Data access links: [EBI](#) / [NCBI](#)

1001 Genomes
A Catalog of *Arabidopsis thaliana* Genetic Variation

Home Collaborators Accessions Tools Software Data Center Gallery About Help desk

Welcome to the 1001 Genomes Project

Links

 Database & Species lists News Events Publications Participants For G10K Organizers (restricted)

Search: Go

GENOME 10K®
Unveiling animal diversity

Genome 10K Project

To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet.

April 2009—The Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. The trajectory of cost reduction in DNA sequencing suggests that this project will be feasible within a few

Join us
Become a G10K affiliate

Genome assembly

- 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)

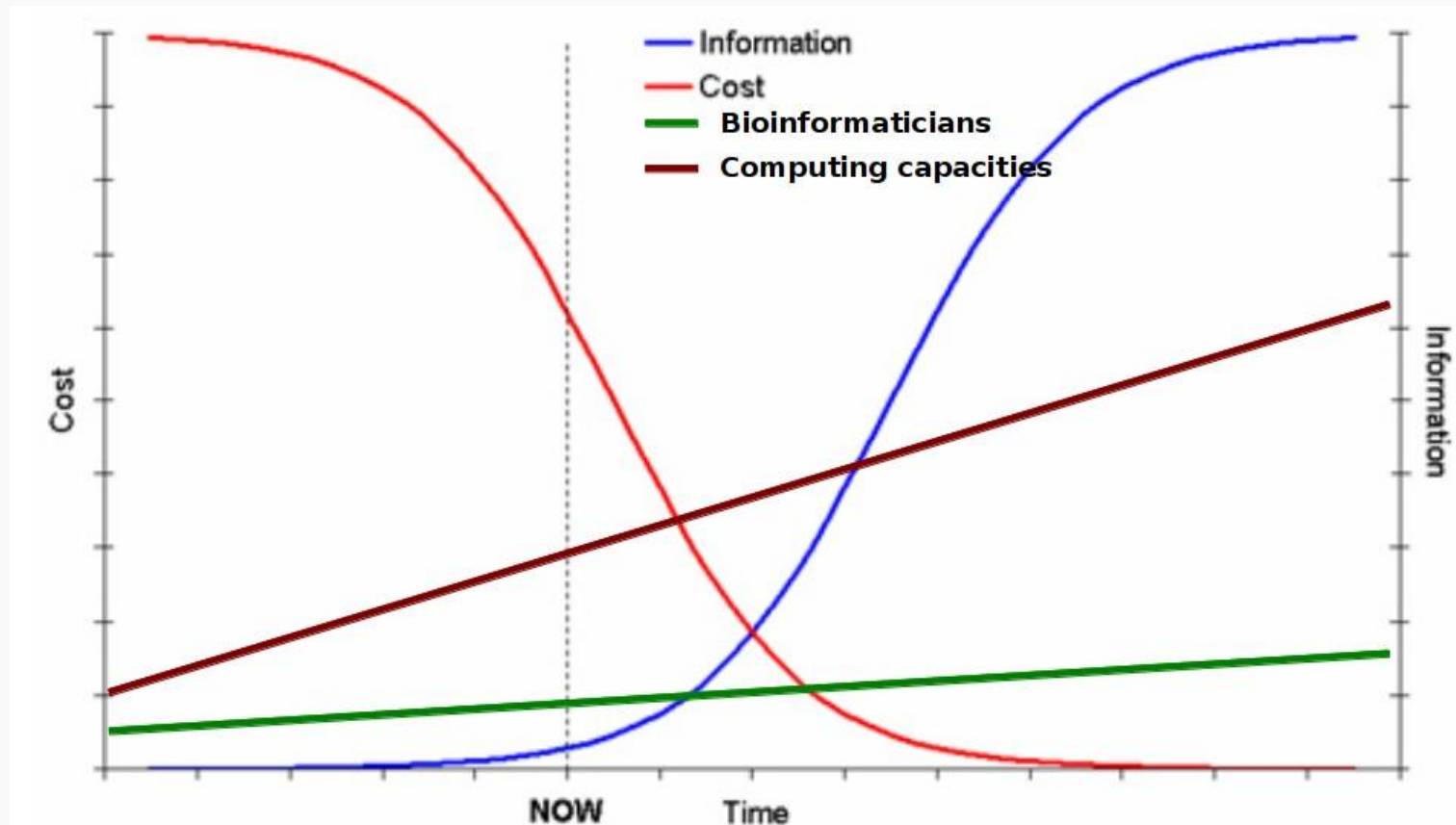
- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015

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

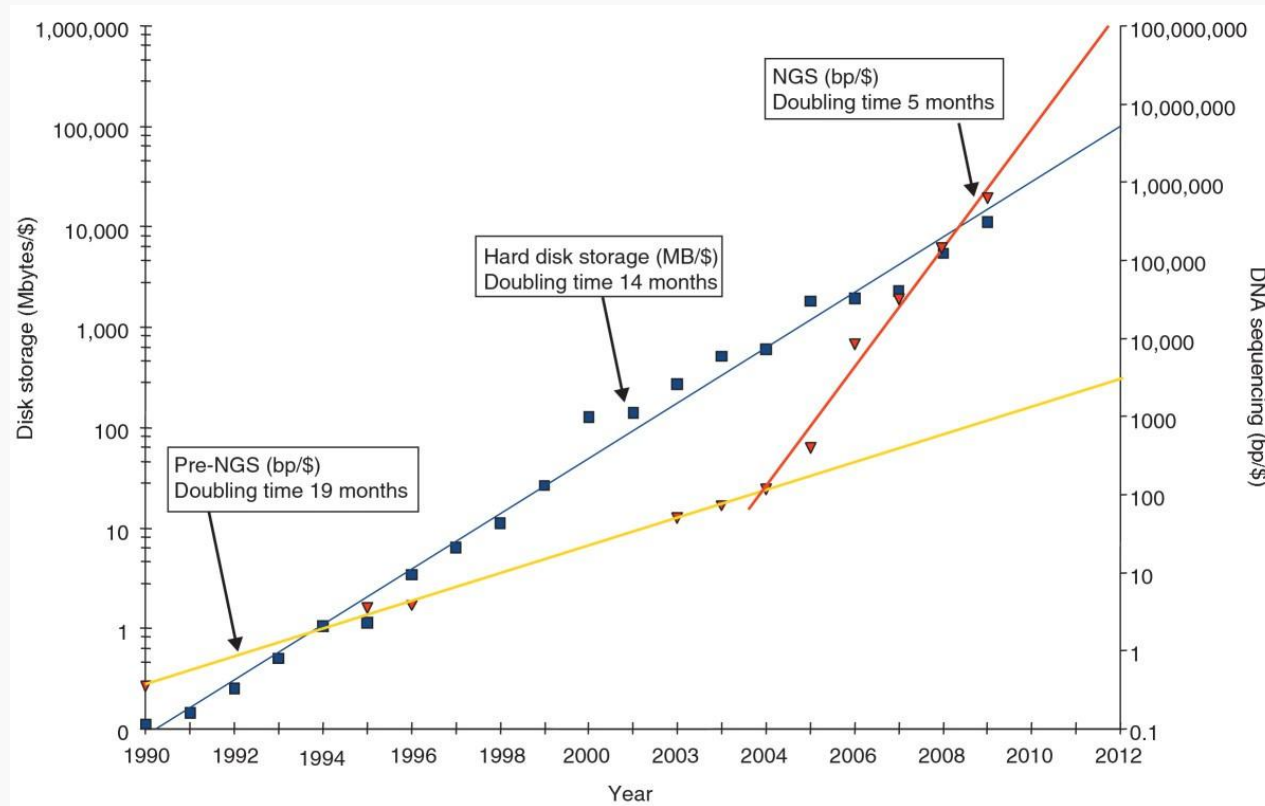
- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available

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...

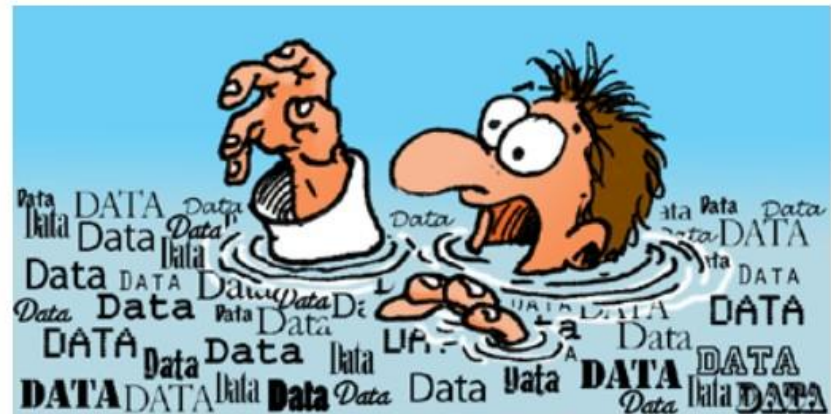


...From Data Rarity to Data Deluge



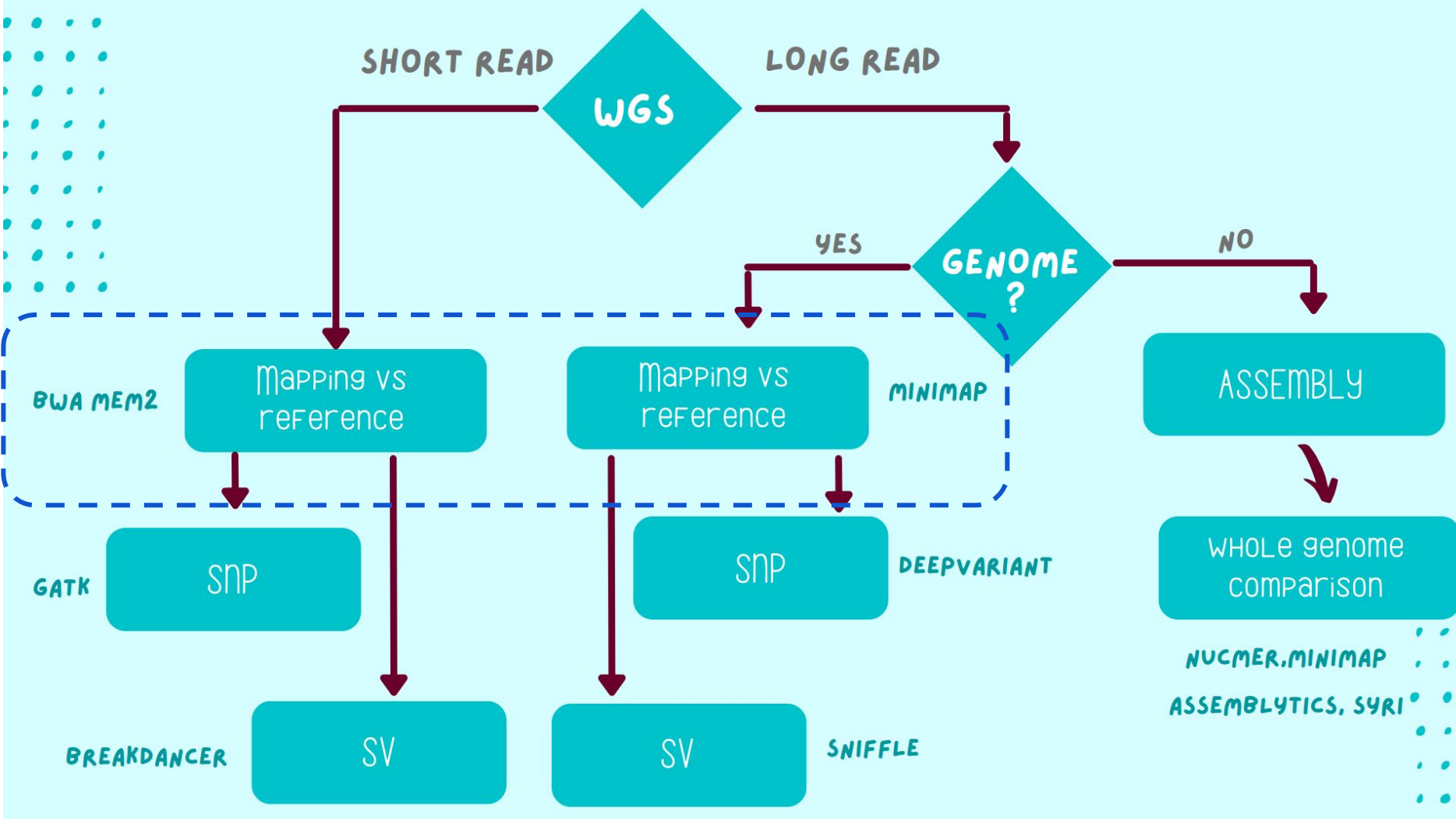
From L. Stein, 2010

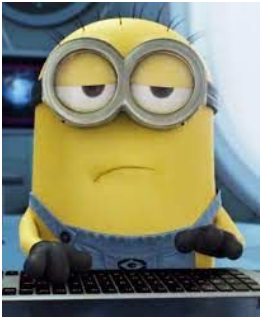
Be Careful to data drowning!



Training plan - day 1

SV DETECTION





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

Diploid Asian Rice, *Oryza*



From
Wikimedia

Diploid Asian Rice, *Oryza*



From
Wikimedia

1. Select/Cut 1 Mb on Chromosome 10

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

```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTTGTTAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGAGTGTTTGGAAAGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
TGGTTTATGCAGAAATTTCTAGAATAAGGGTAACTT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
TCTCAGAACTTGTTTGTGATGTGTGTATTCAACTA
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
TTGATTTAACTCTGACAAAATAAACAAAGTCTTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

Sequencing info

Nucleotide sequence

Quality score in ASCII

IRD
Institut de Recherche
pour le Développement
FRANCE

30	1 pour 1000
40	1 pour 10000
50	1 pour 100000

```

SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
...XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789;<=>?@ABCDEFGHIJKLMN
OPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                                     |         |           |
33                               59   64       73             104               126
0.....26...31.....40
                        -5....0.....9.....40
                           0.....9.....40
                             3.....9.....40
0.2.....26...31.....41

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 (b)
    (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...

We will

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

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

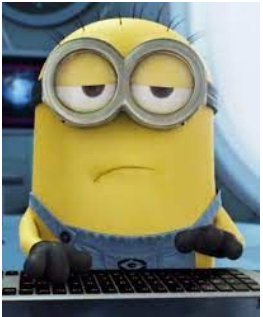


Practice

Let's work with the jupyter book :

Day1_Mapping_Practice_EMPTY.ipynb





Day2 - Let's start slowly

Training plan - day 1 & 2 (a little)

SV DETECTION

Day1

SHORT READ

LONG READ

WGS

GENOME?

YES

NO

BWA MEM2

Mapping vs
reference

Mapping vs
reference

MINIMAP

ASSEMBLY

GATK

SNP

SNP

DEEPVARIANT

WHOLE genome
comparison

BREAKDANCER

SV

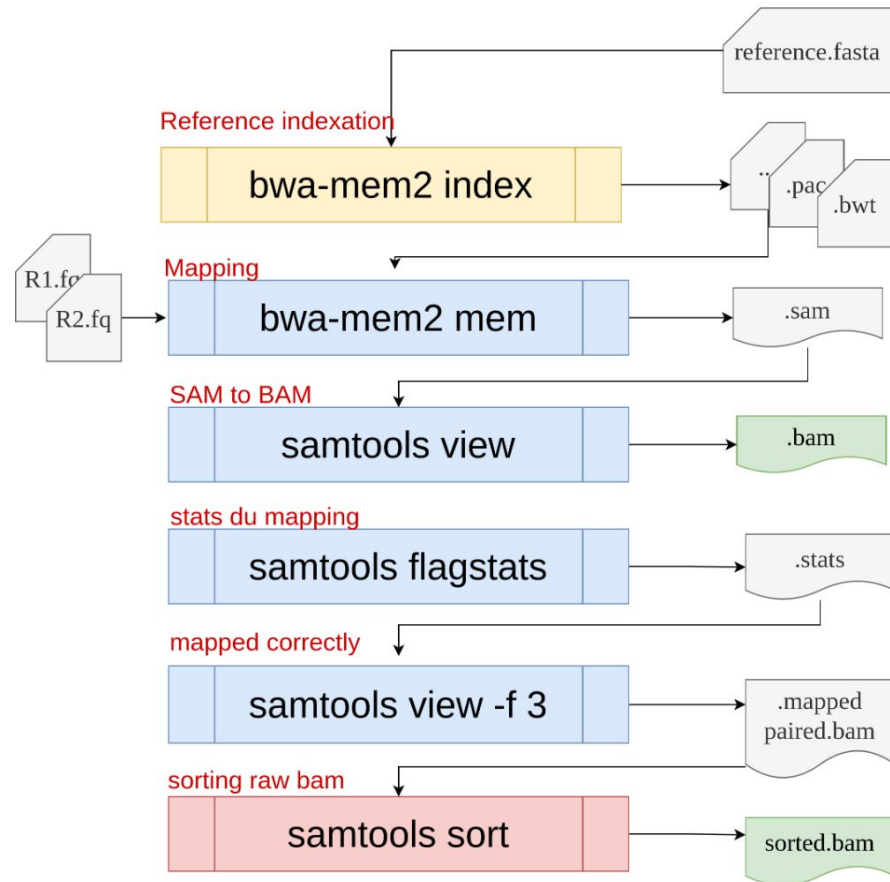
SV

SNIFFLE

NUCMER.MINIMAP

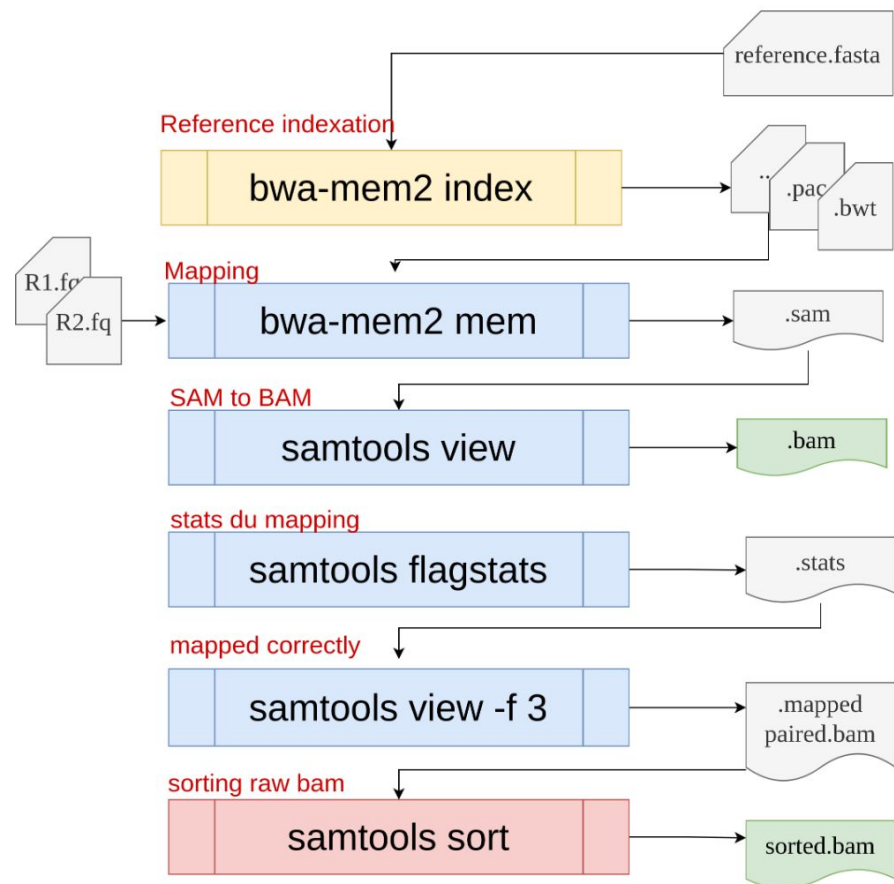
ASSEMBLYTICS, SYRI

..
..
..
..
..
..

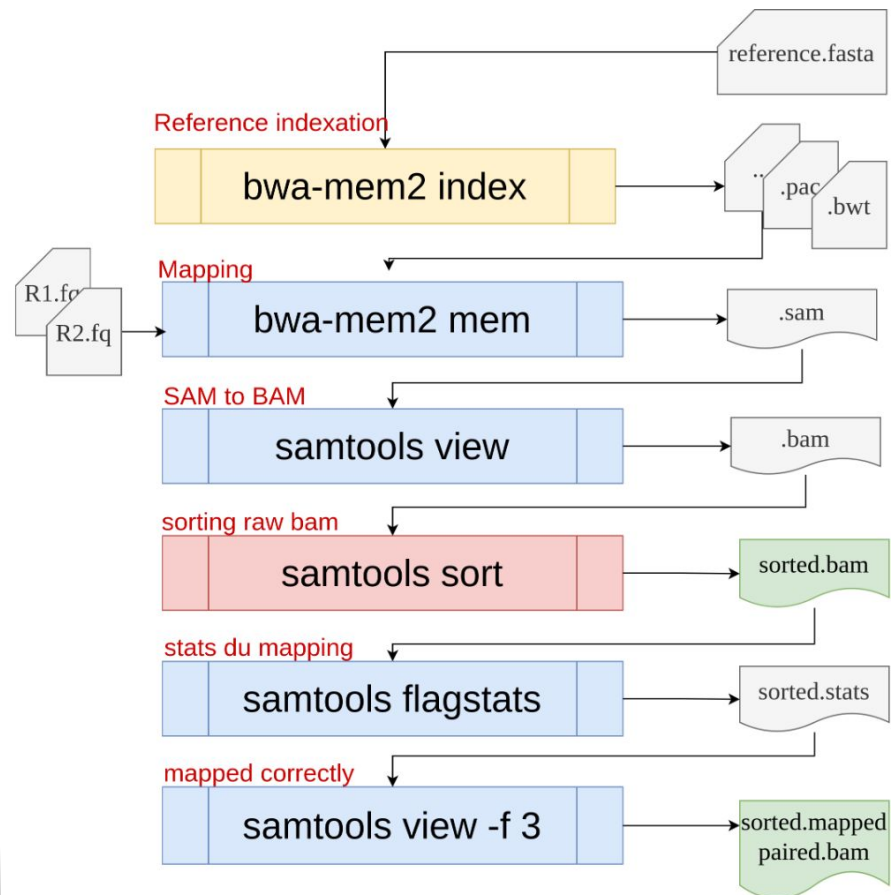


En cours Day1

Plan de bataille du mapping !



En cours Day1



ça marche mieux !

```
: 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
```

```
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

- 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



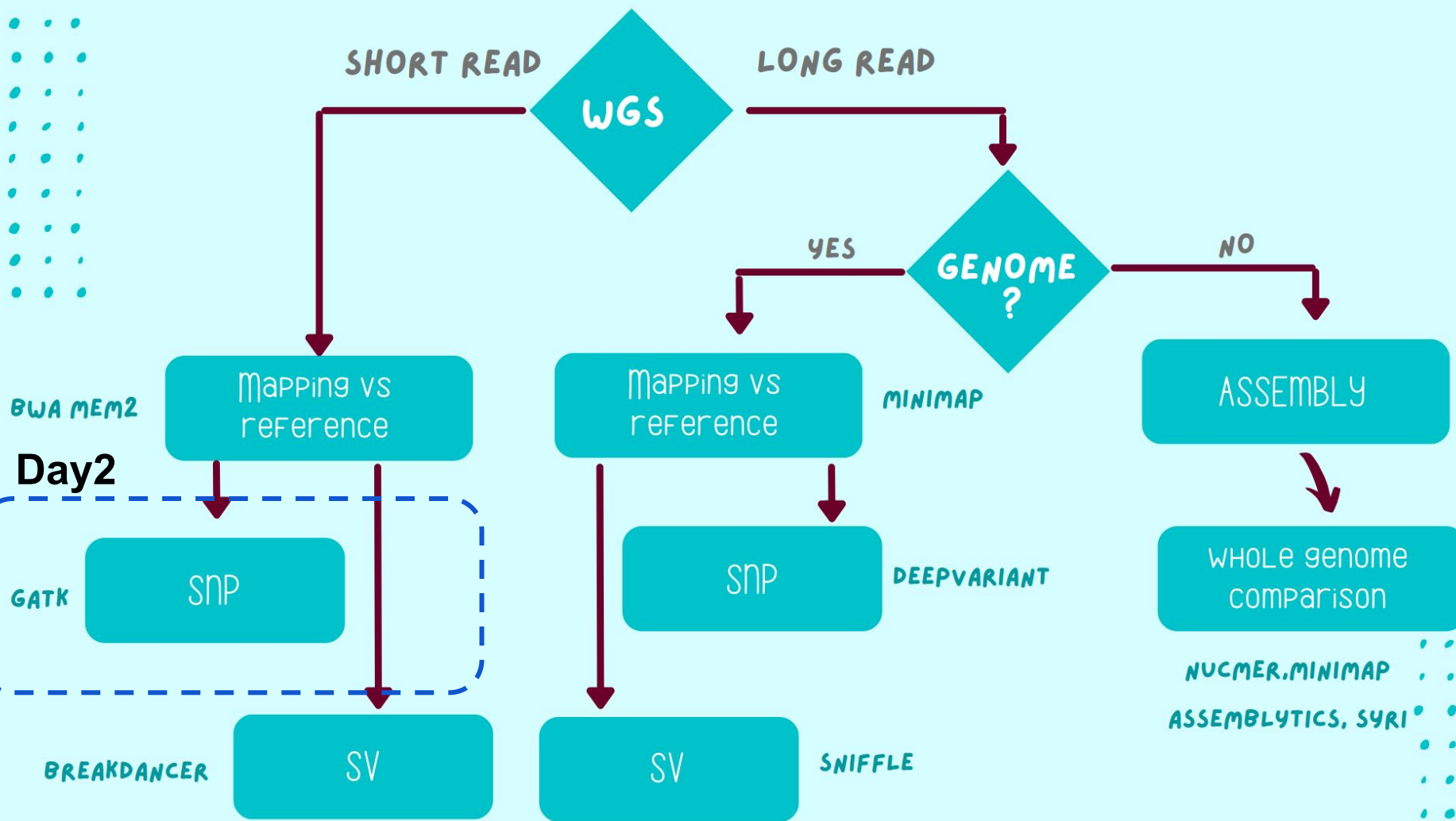
Practice

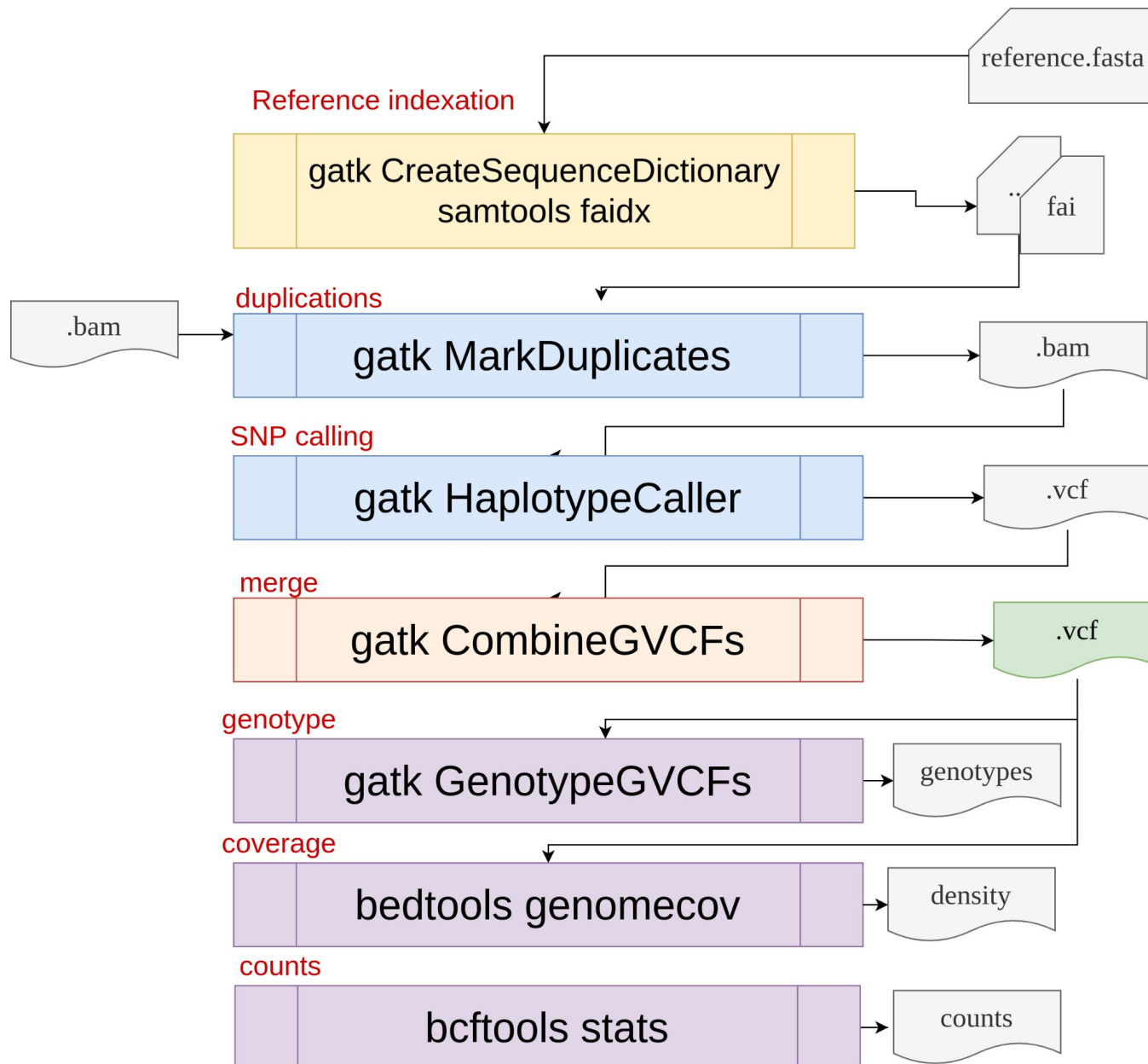
Let's work with the jupyter book :

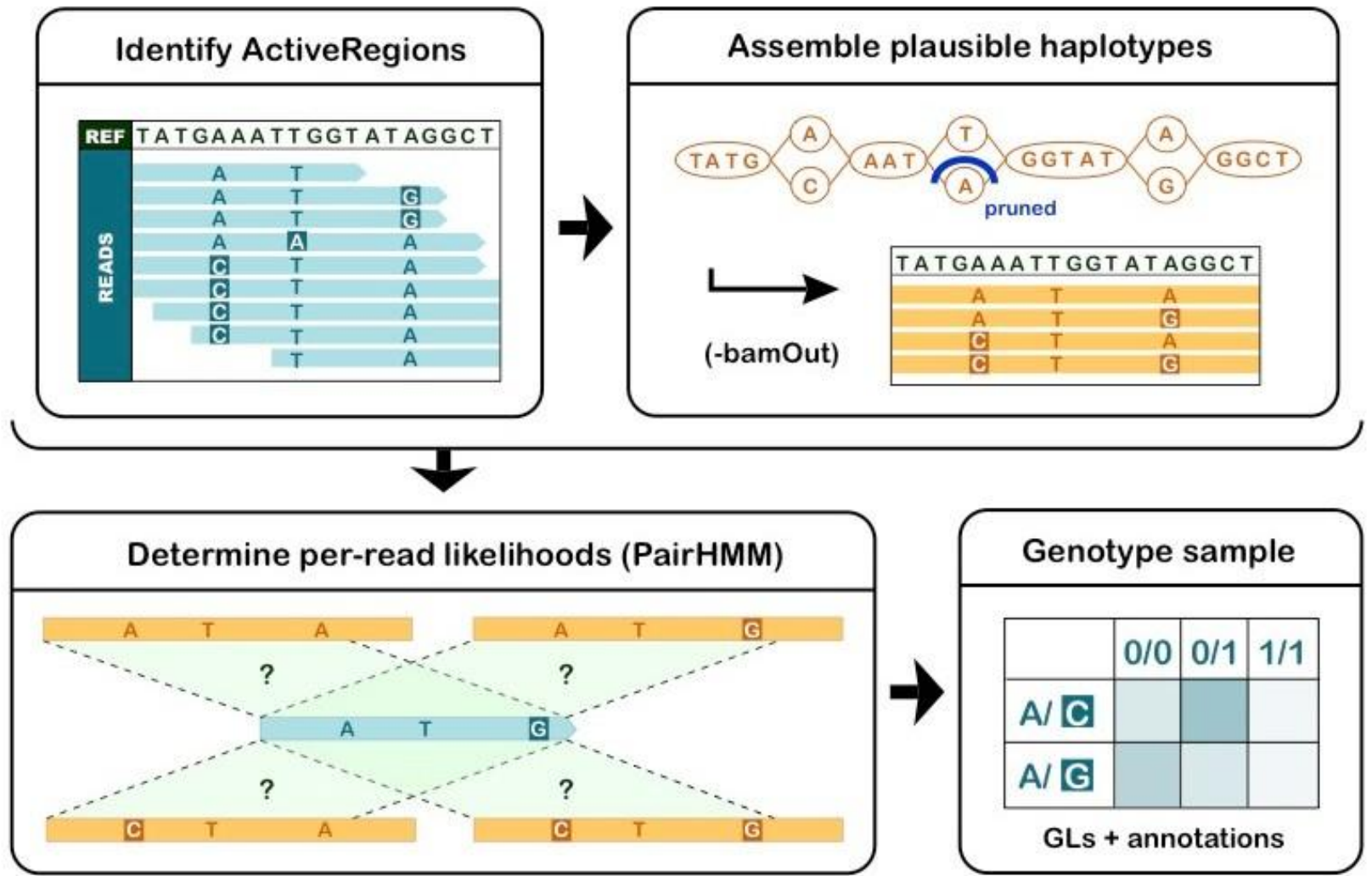
Day2a_Mapping_analysis-EMPTY.ipynb

Training plan - day 2

SV DETECTION







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

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##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">
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Body

#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		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Deletion

SNP

Large SV

Insertion

Other event

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)



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 reproducibility not ensured

Problems with manual launches

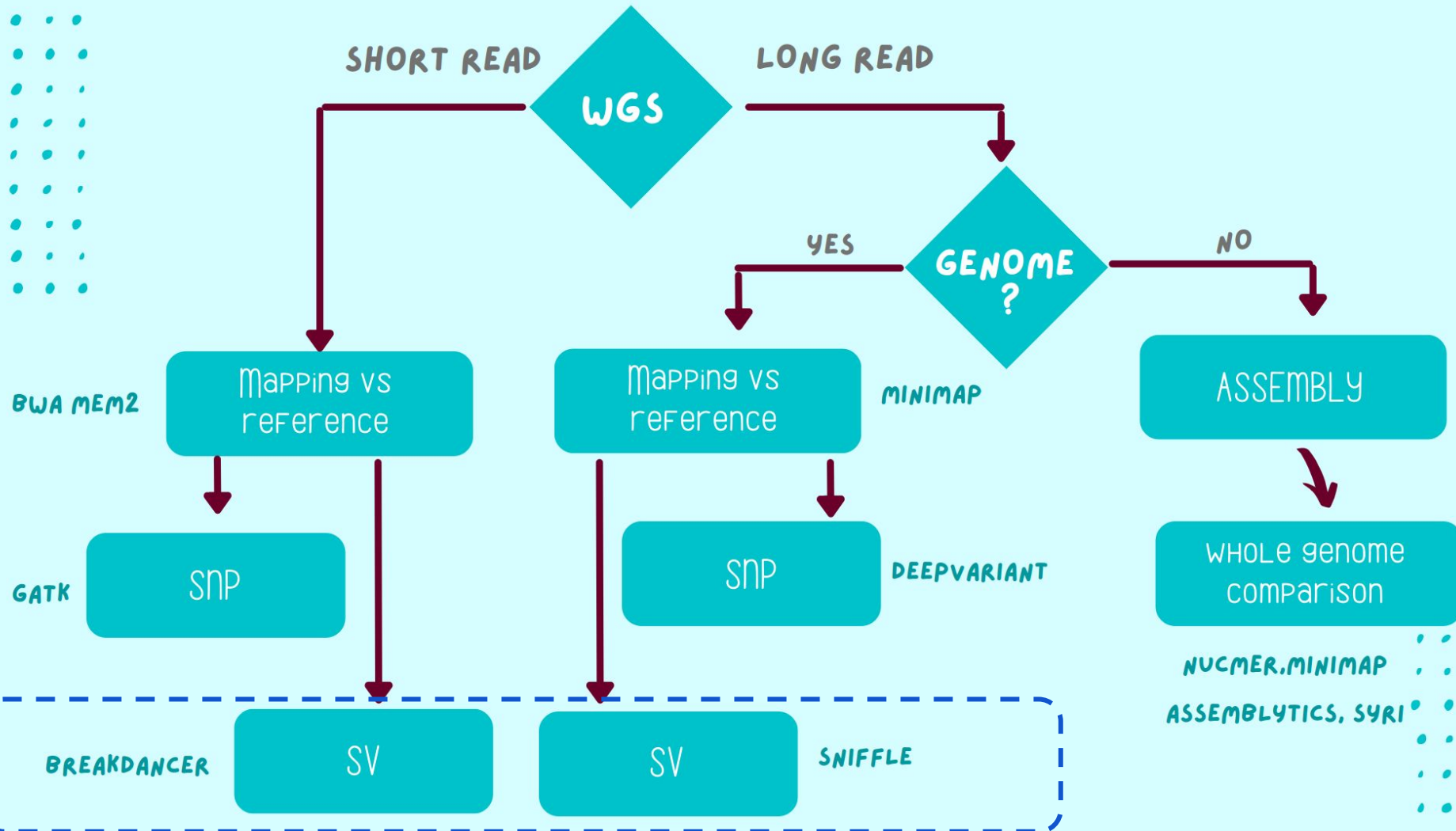
- Long
- Fastidious
- Error prone
- Tracability and reproducibility not ensured

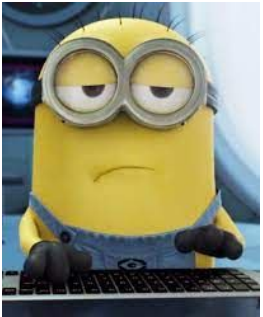
Solution $\Rightarrow$ Workflow Manager (or scripts...):
SnakeMake, TOGGLE, NextFlow



Training plan - day 3

SV DETECTION





Long Read Mapping

Two technologies

Oxford Nanopore



MinION

GridION



PromethION

Pacific BioScience

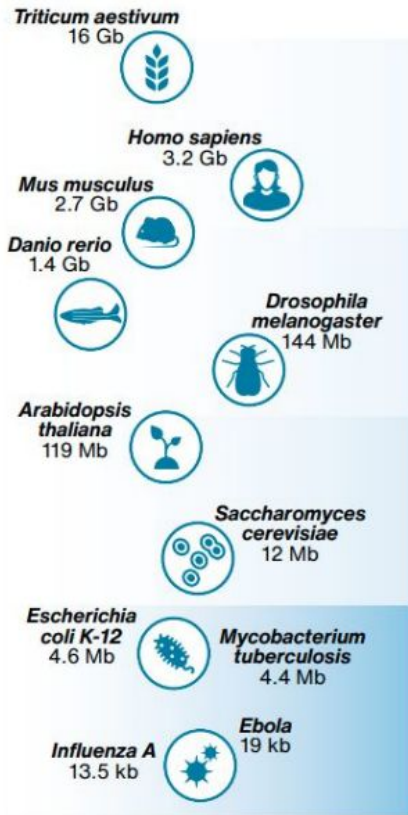


RSII



Sequel

from Elixir GAAS 2018



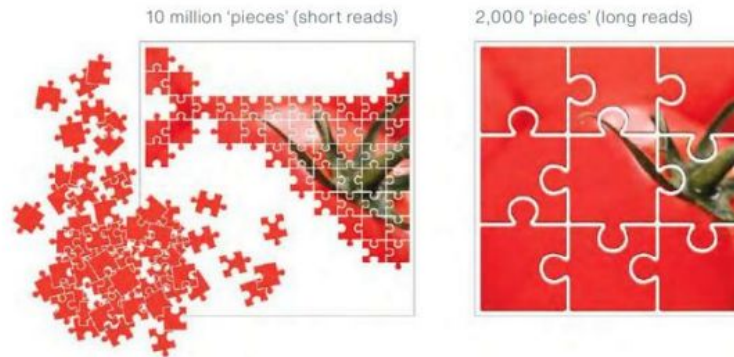
Microbial genomes

Human genomes

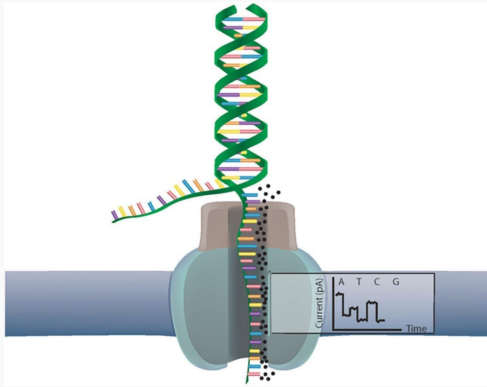
Animal genomes

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

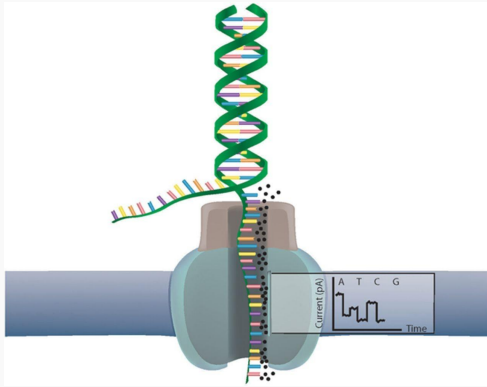






From Circulation
Research

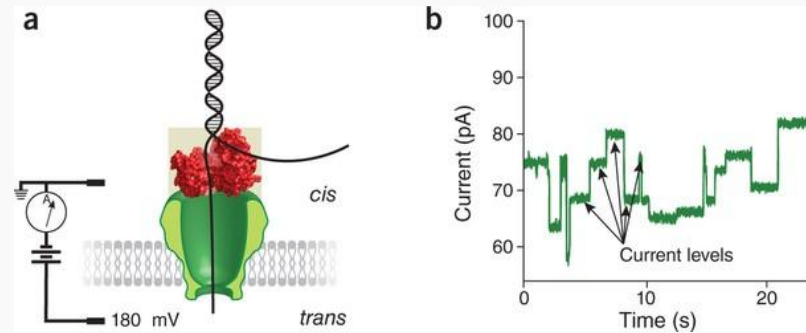
- No Amplification
- NO SYNTHESIS
- Very Long Length



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 |  Human genomics |
|  Microbiome |  Clinical research |
|  Environmental |  Cancer |
|  Plant |  Transcriptome |
|  Animal |  Populations genomics |

What you can do with it ?

Research areas

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

Investigations

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

What you can do with it ?

Research areas

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

Investigations

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

Techniques

-  Whole genome
-  Targeted
-  Whole transcriptome
-  Metagenomics

2nd Generation Sequencing

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

454

IonTorrent

Illumina

3rd Generation Sequencing

PacificBiosciences
Oxford
Nanopore

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

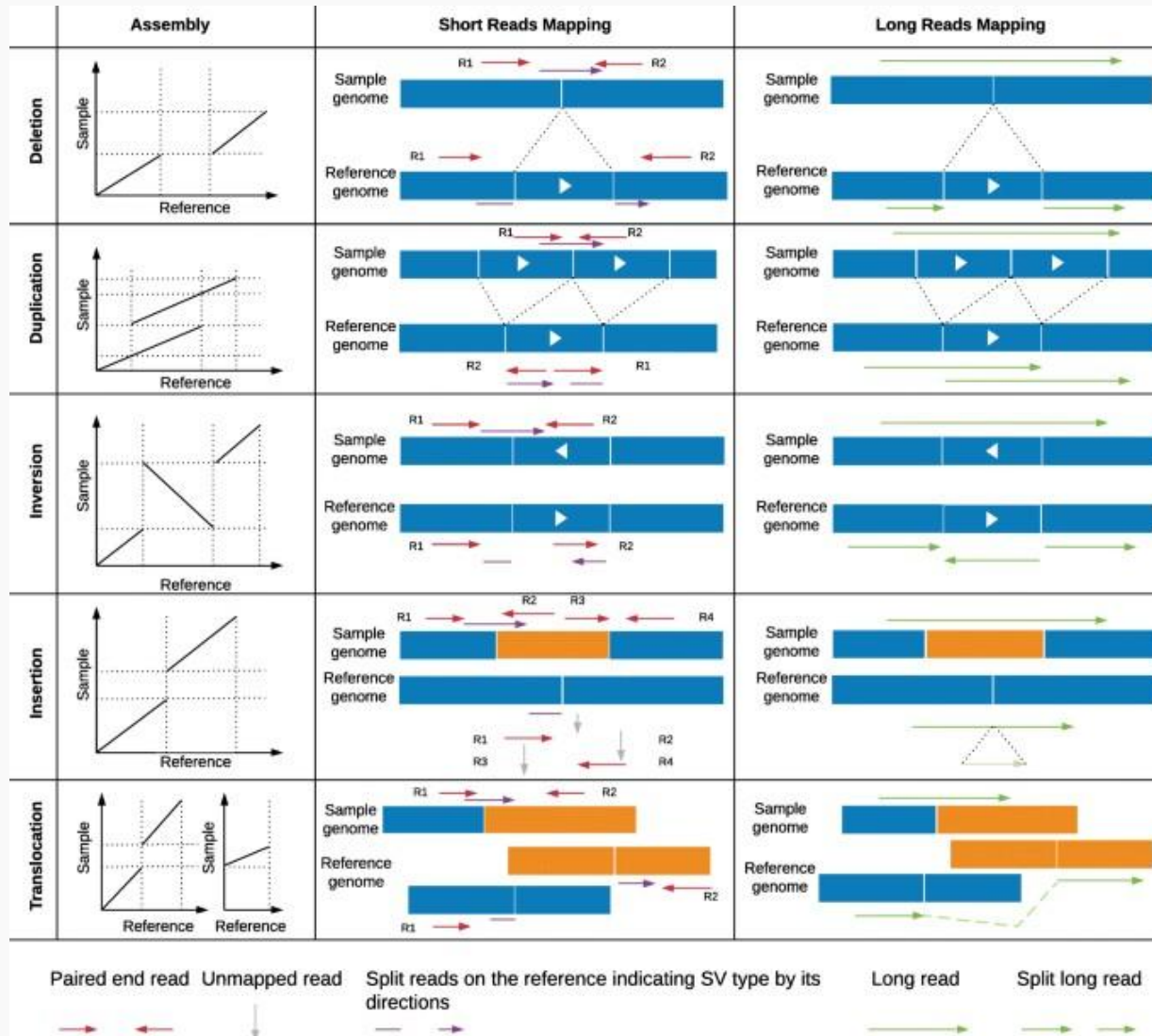
NGS technologies comparison

NGS platforms

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

From wikipedia website

Structural Variant Detection





Practice

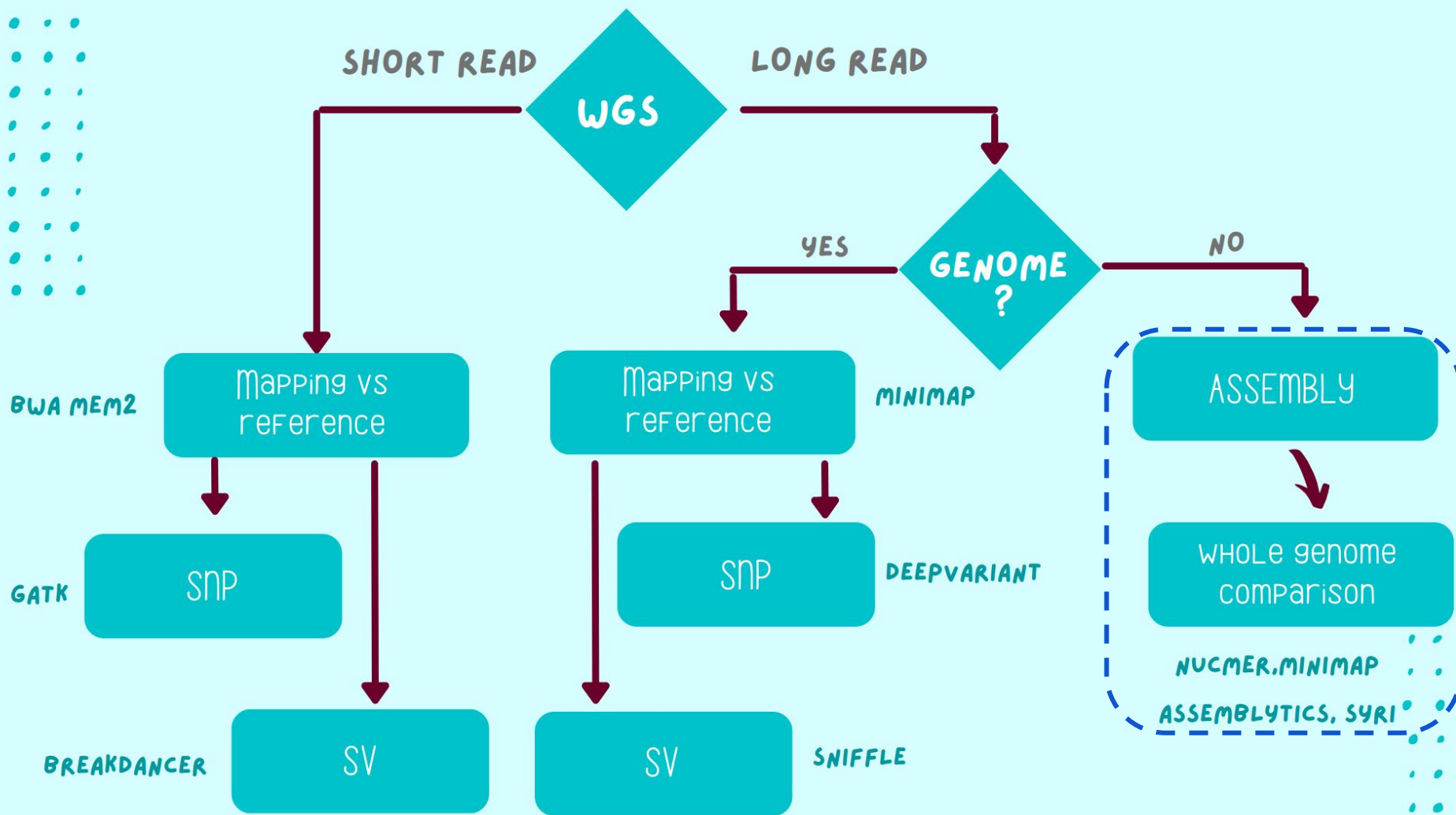
Let's work with the jupyter book :

Day3_Variants_structuraux_EMPTY.ipynb



Training plan - day 4

SV DETECTION



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

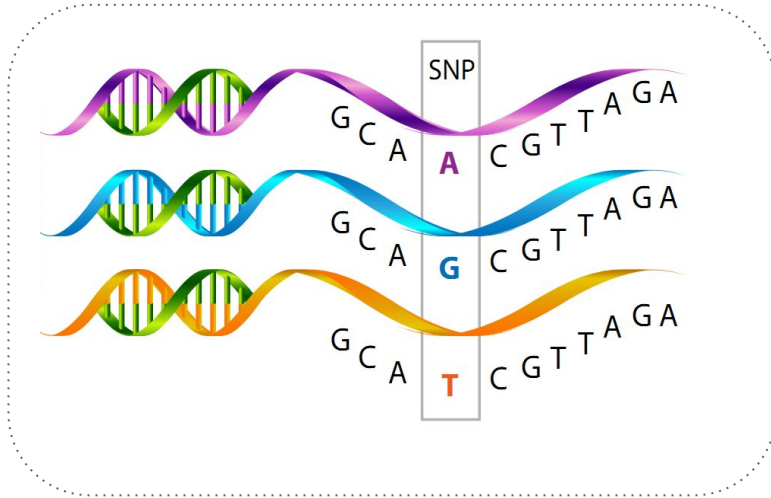


So let's summarize...

Remember why we started...

Mutations & Variations as main source of genetic diversity

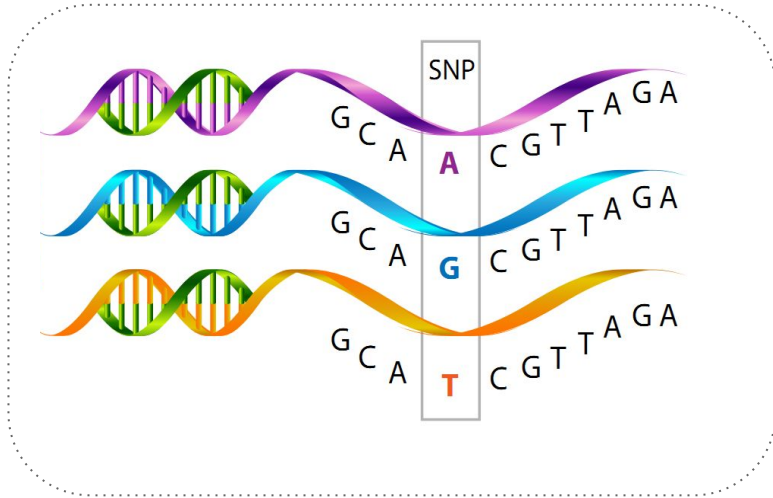
Single Nucleotide Polymorphism



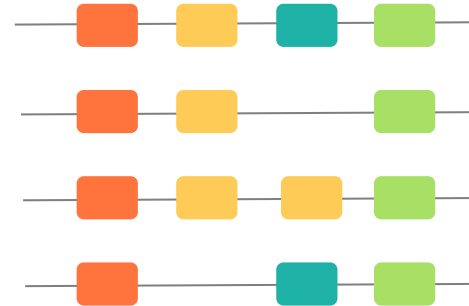
Remember why we started...

Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism



Structural Variations



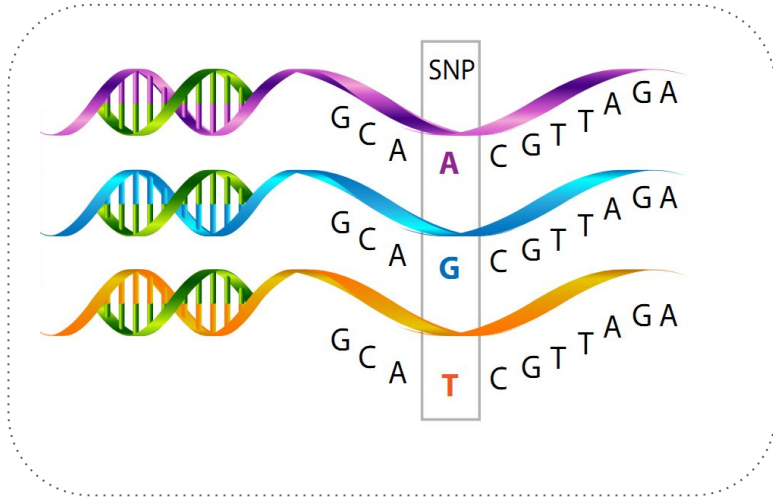
Presence Absence Variation (PAV)

Deletion, duplication, copy number variation, mobile element insertion

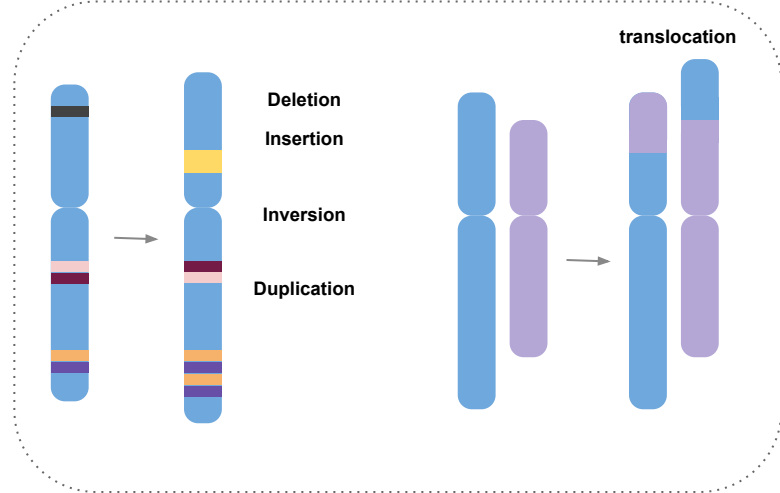
Remember why we started...

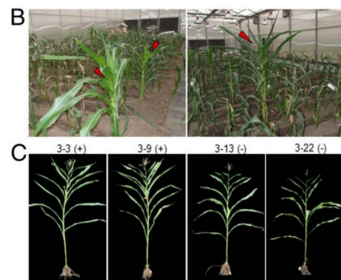
Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism



Structural Variations

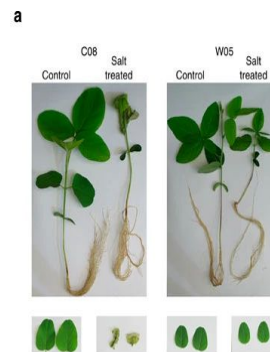




From Yang et al., 2013



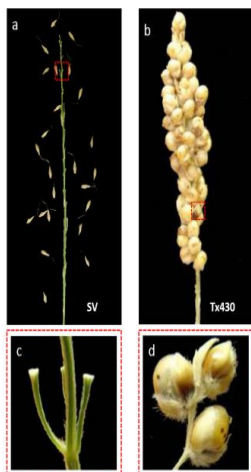
From Li et al. 2012



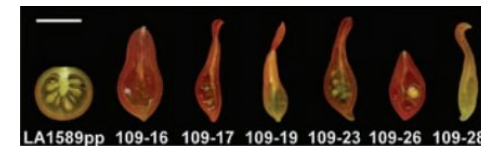
From Qi et al. 2014



From Yang et al., 2014



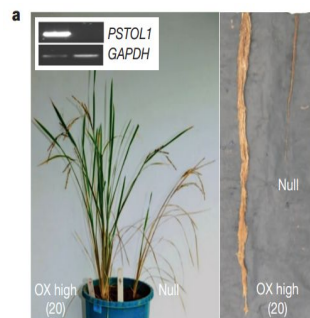
From Lin et al. 2012



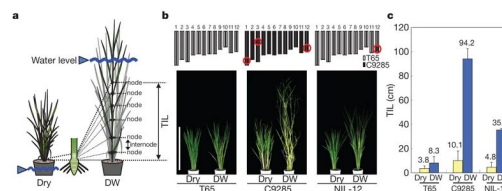
From Xiao et al. 2008



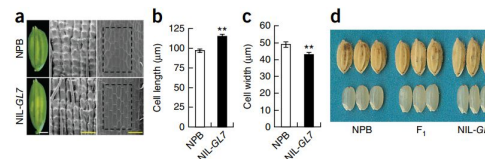
From Xu et al. 2006



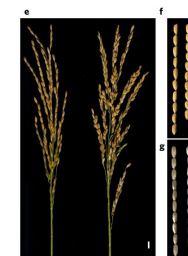
From Gamuyao et al. 2012



From Hattori et al. 2009

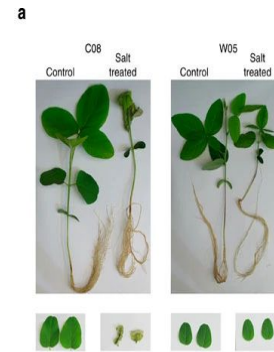
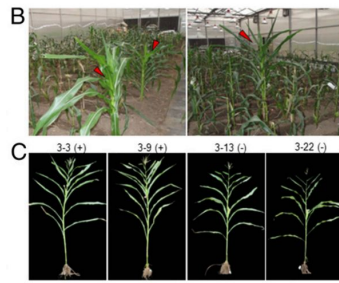


From Wang et al. 2015



From Bai et al. 2017





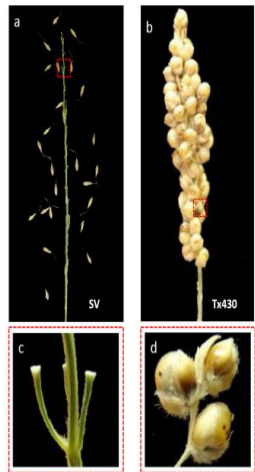
From Yang et al., 2013

From Li et al. 2012

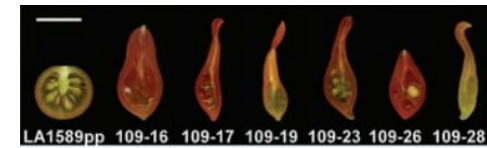
From Qi et al. 2014

From Yang et al., 2014

Is One Reference genome enough to capture all genetic diversity ?



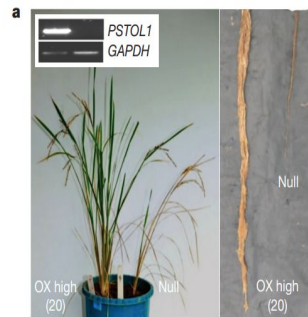
From Lin et al. 2012



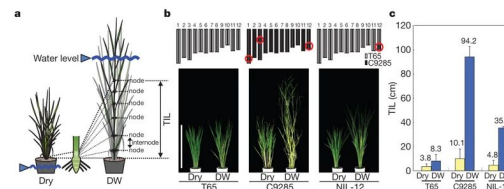
From Xiao et al. 2008



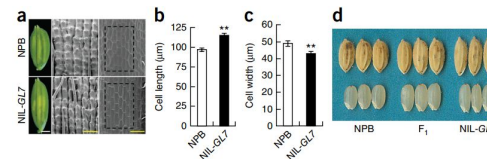
From Xu et al. 2006



From Gamuyao et al. 2012



From Hattori et al. 2009



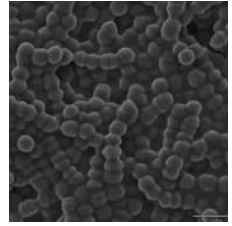
From Wang et al. 2015



From Bai et al. 2017



Gene number variations within a species



Streptococcus agalactiae



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

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

Tettelin et al., 2005

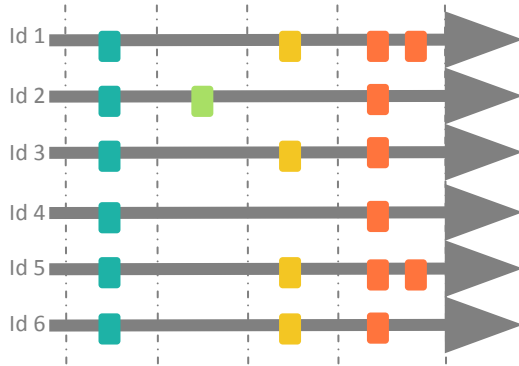
- ▶ 8 strains sequenced
- ▶ SNP variations



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

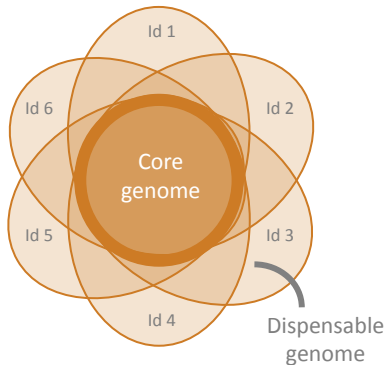
Pangenome concept

Pangenome concept

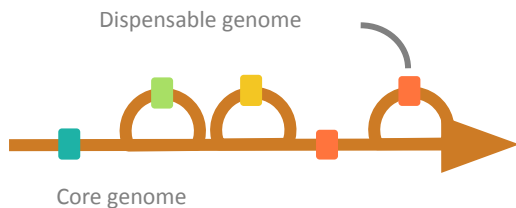


Pangenome

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



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

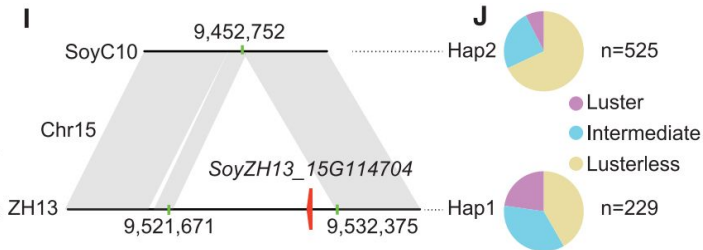
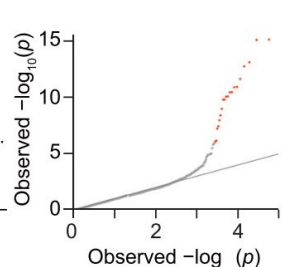
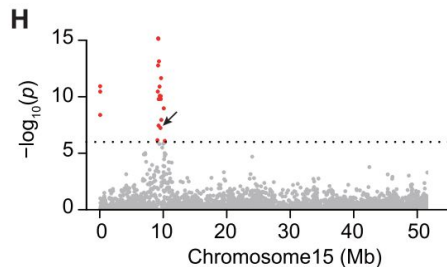


What's else ...



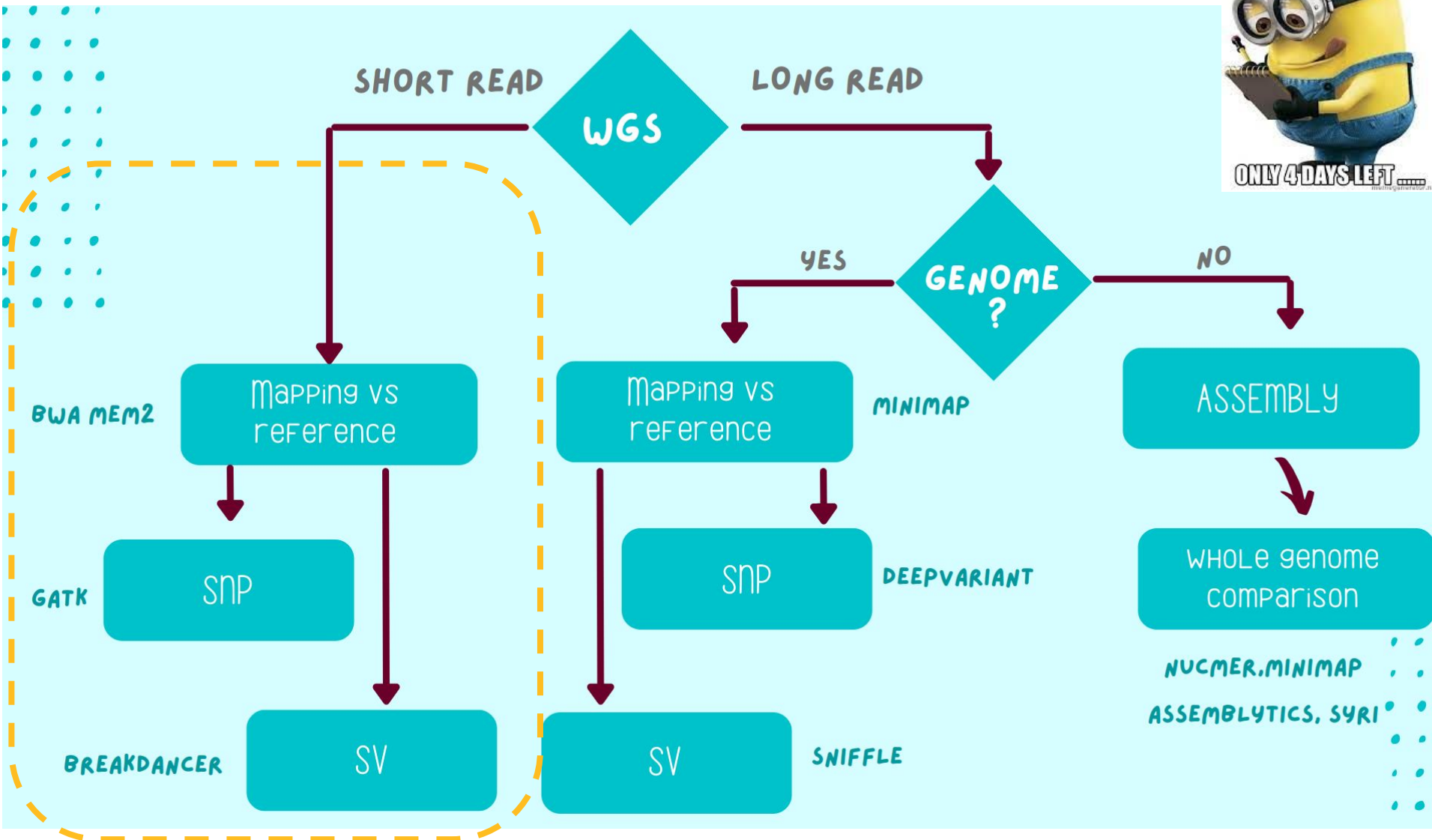
- ▶ 12,150 genes absent from the reference (18 cultivars)

- ▶ 27,175 genes absent from the reference (26 cultivars)

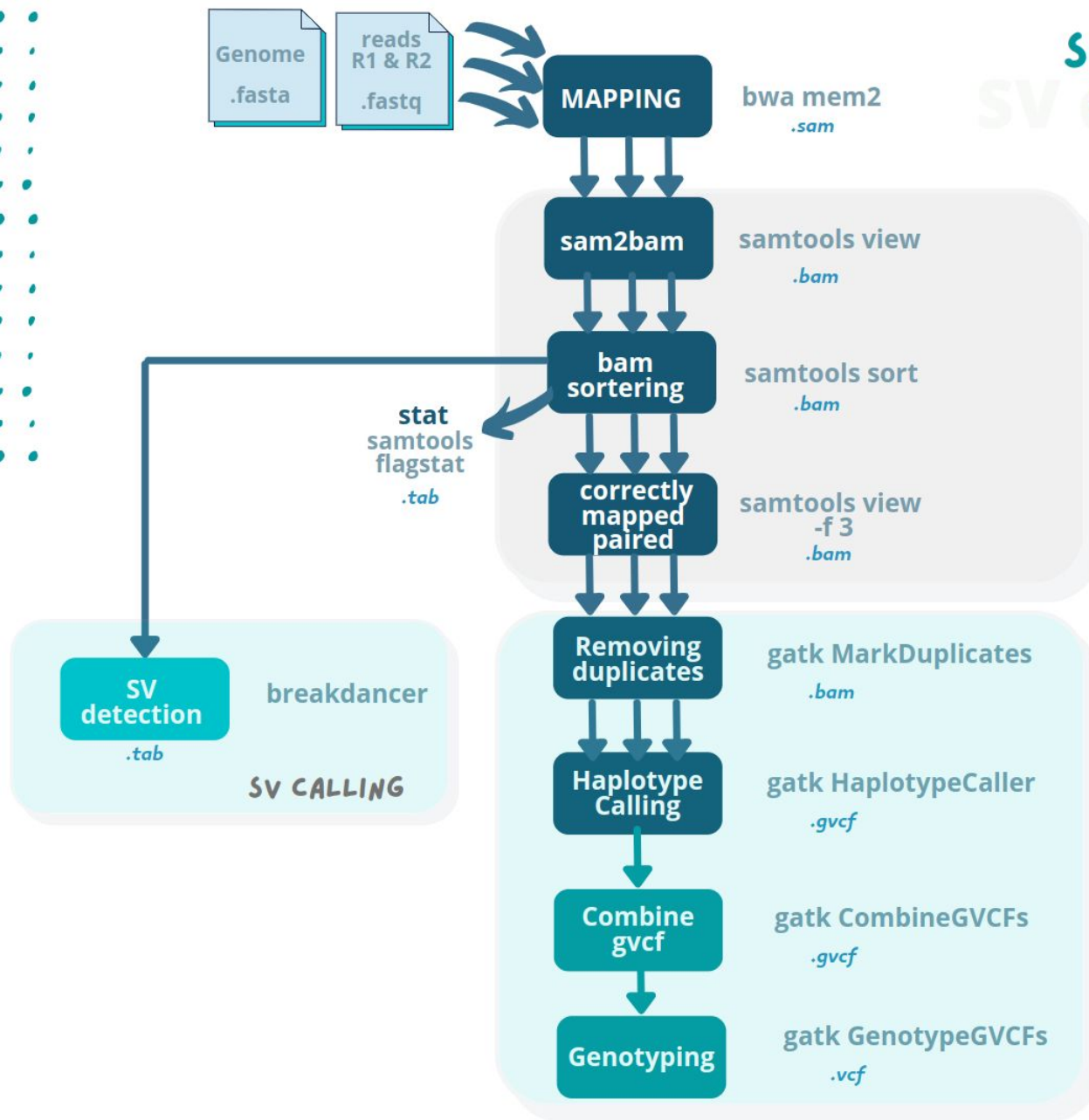


How to detect SVs and analyse them ?

REMEMBER FOLKS



From short reads....



**SNP DETECTION
SHORT READS**

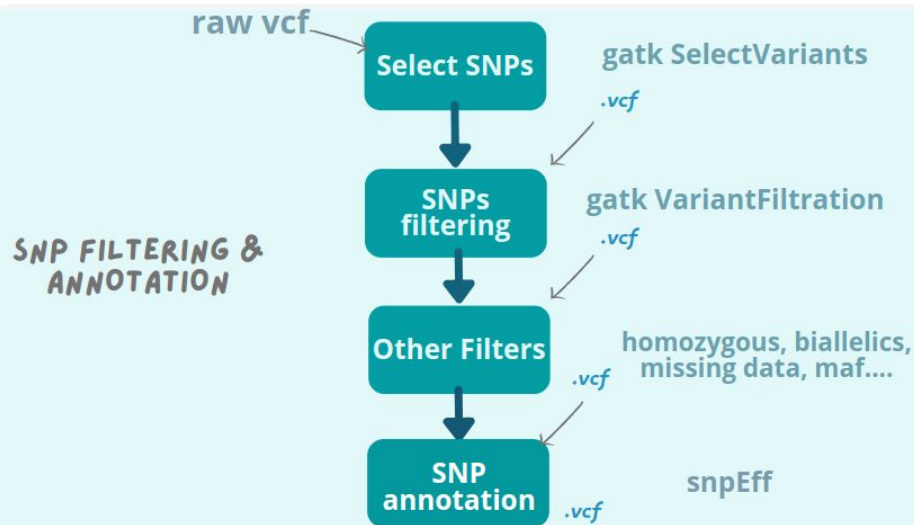
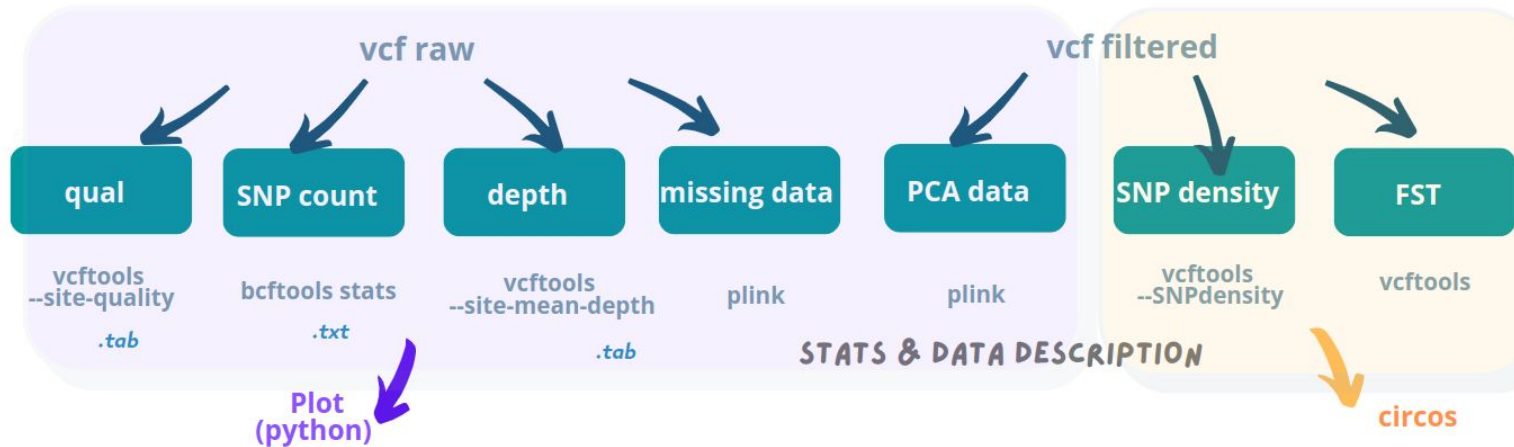
**SAM/BAM
MANIPULATION**

SNP CALLING



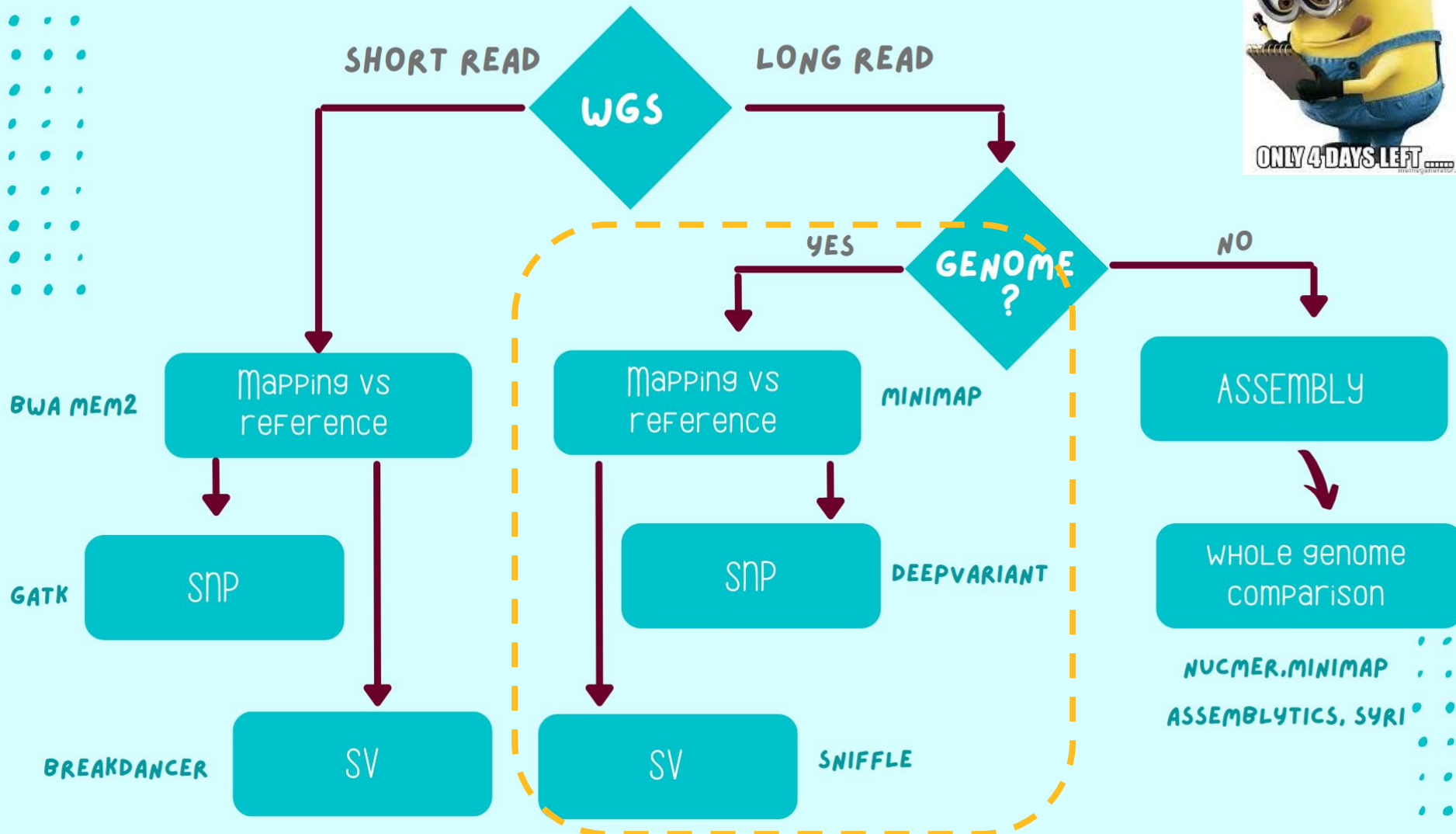
From short reads....

SNP ANALYSIS



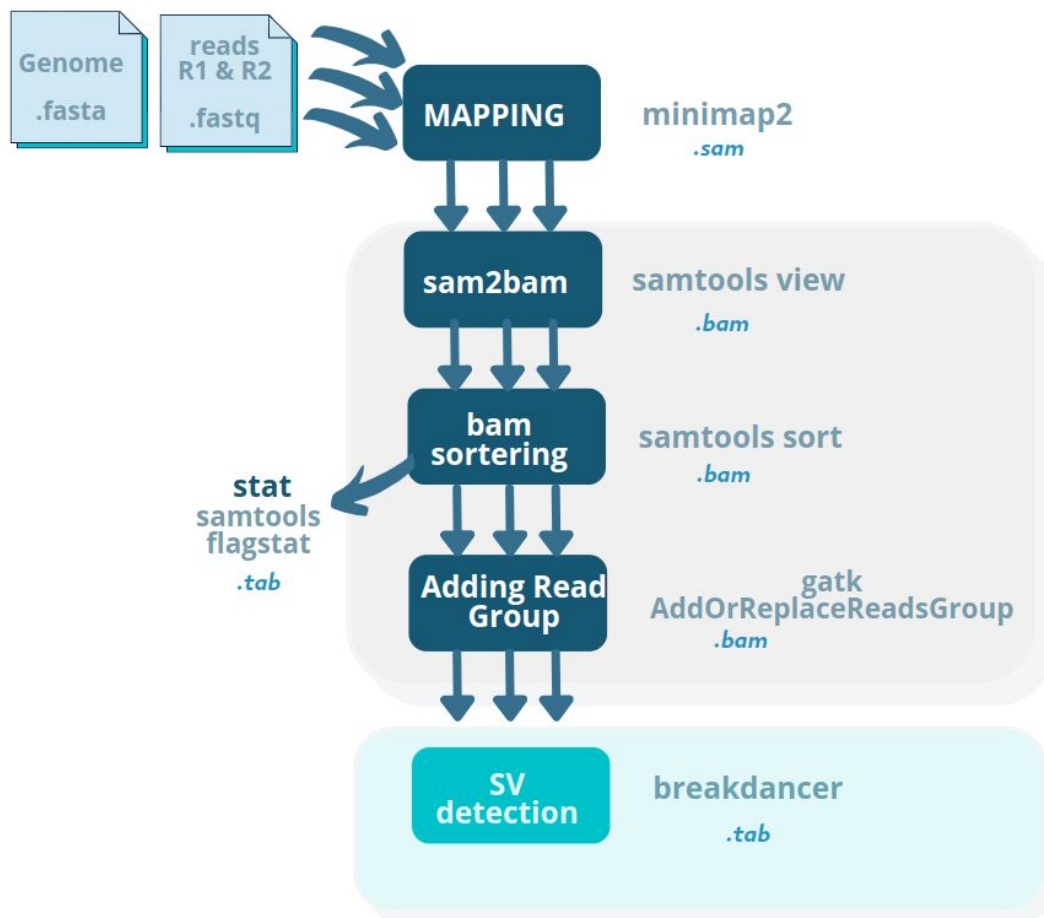
How to detect SVs and analyse them ?

REMEMBER FOLKS



From long reads...

SV DETECTION LONG READS



REMEMBER FOLKS



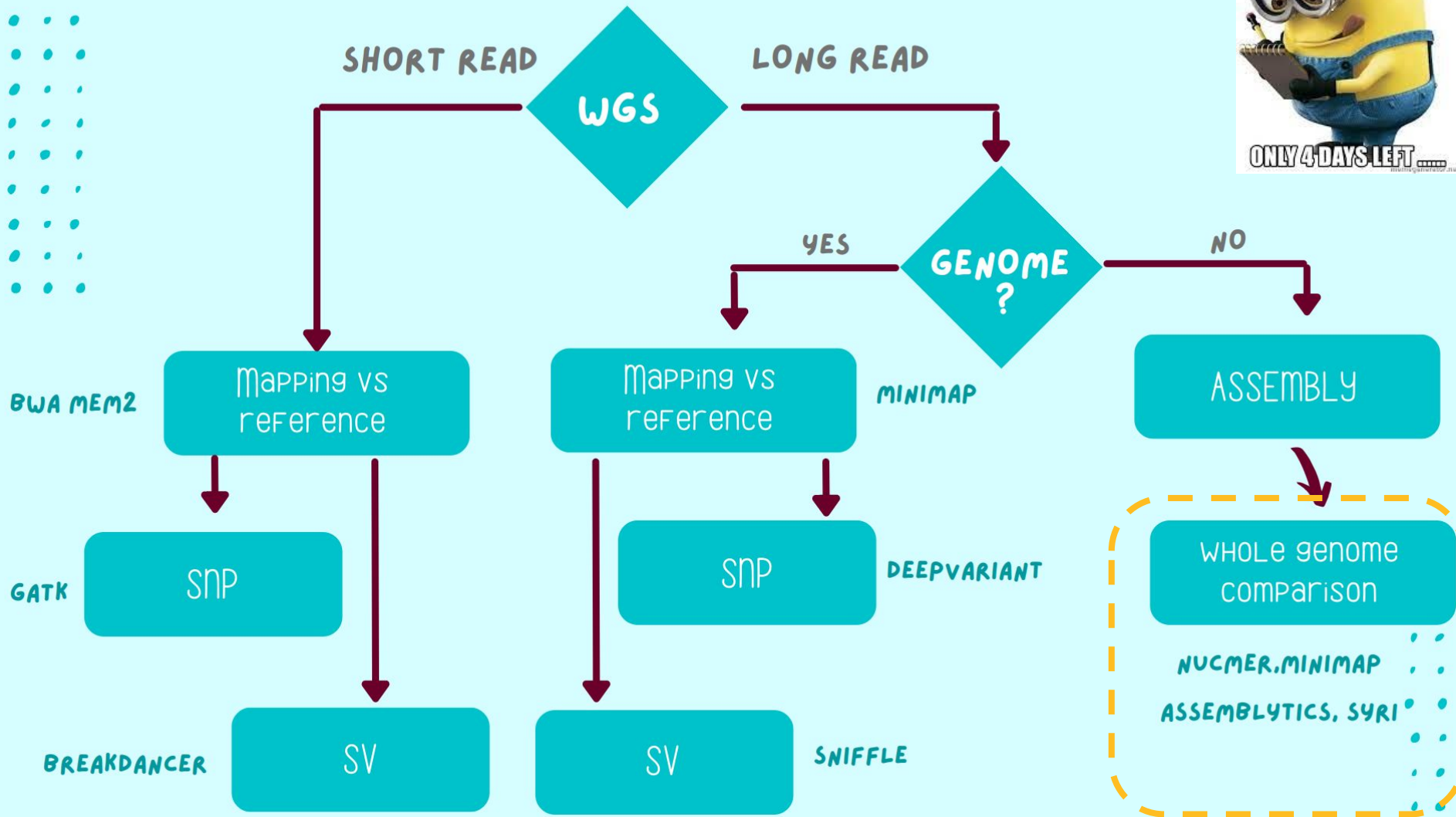
ONLY 4 DAYS LEFT

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 ?



The Pan-Genome of the cosmopolitan picophytoplankton *Bathycoccus* : a first step towards understanding adaptation to latitude and season

**François-Yves Bouget
Martine Devic
Louis Denu**

Observatoire océanologique de Banyuls sur mer
Laboratoire d'océanographie microbienne



Structure of Bathycoccus genome

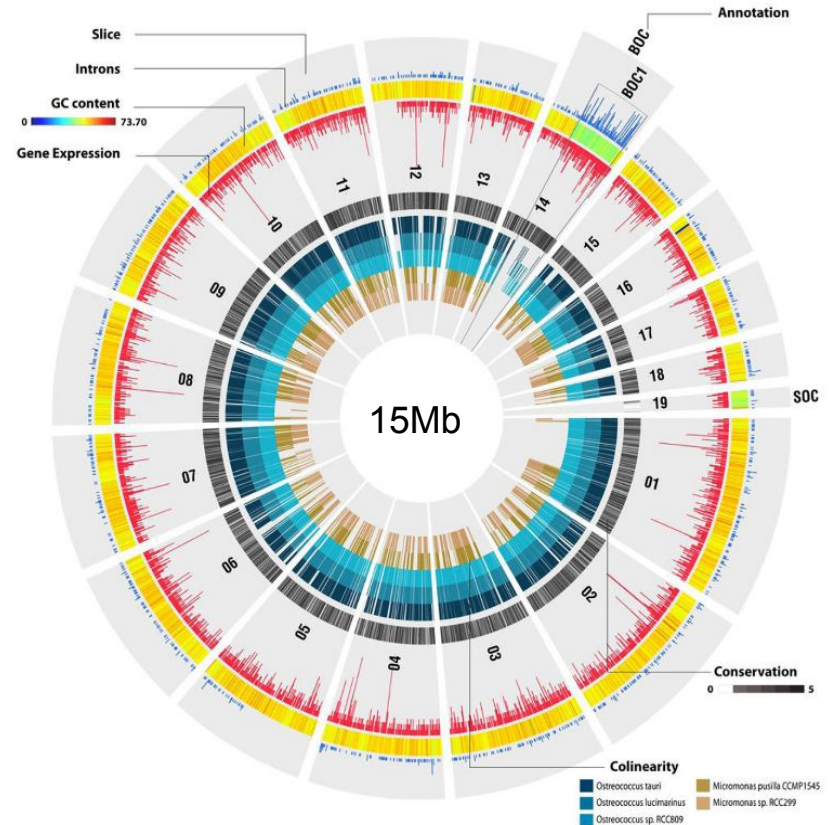
Reference genome
strain RCC1105
Banyuls bay in 2006

15Mb haploid genome

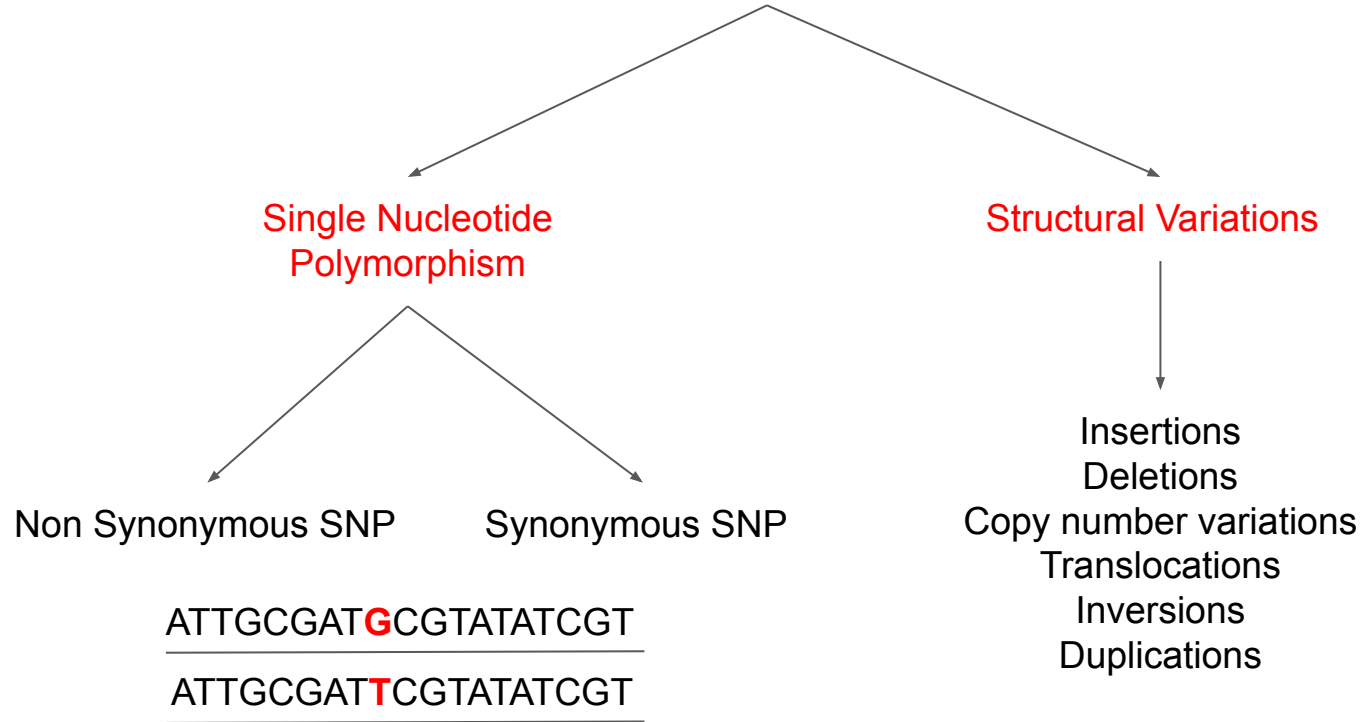
19 nuclear chromosomes

2 Outliers regions with irregular GC content and
gene composition:

- Big outlier region on **chromosome 14** (BOC)
- Small outlier **chromosome 19** (SOC)



Genetic variations

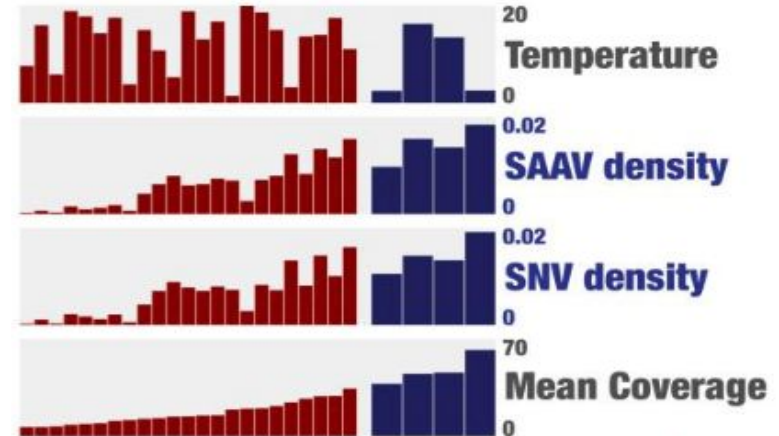
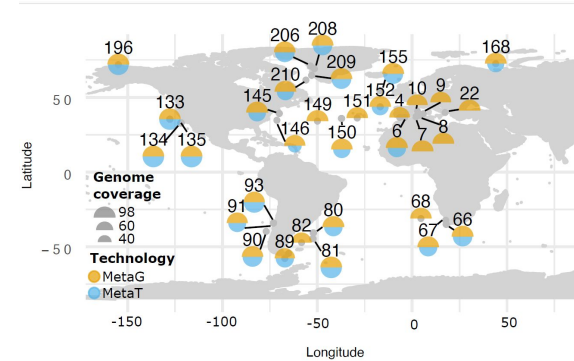


Bathycoccus SNP in metagenomic datasets

Mapping of metagenomic reads
2% of coding sequences are variable

**Non synonymous SNP not correlated
with temperature**

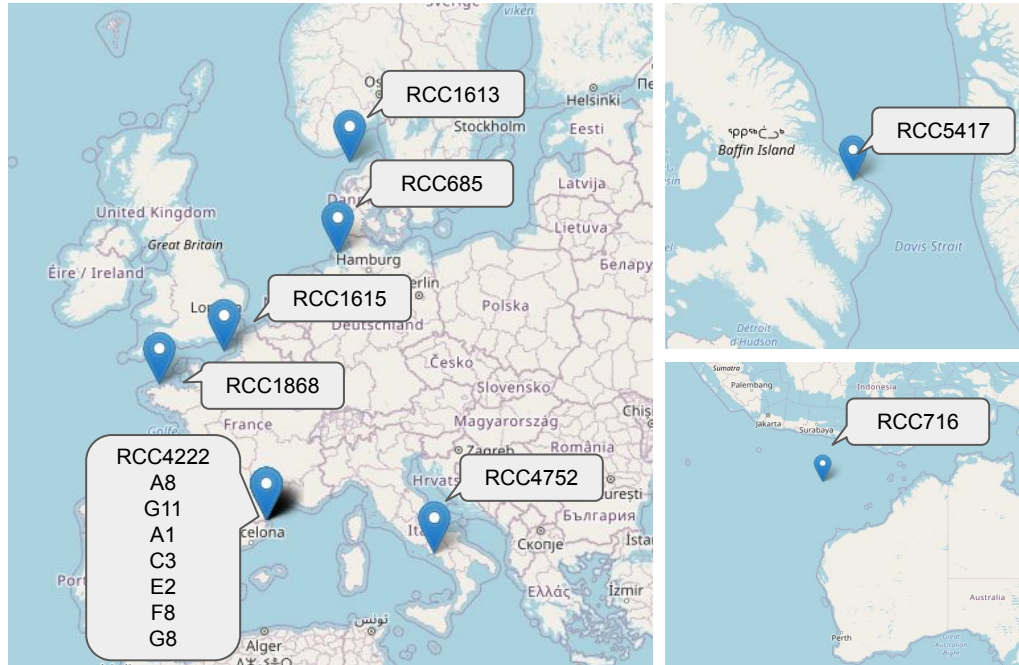
**Metagenomic only approaches
is more revealing of population
diversity than adaptation
mechanism**



Metagenomic datasets used with only one reference genome aren't sufficient to describe genetic variations and to study adaptation

To efficiently use these dataset to describe genetic diversity and study adaptation we need better and more diverse references

Available worldwide and local Bathycoccus strains



Souche	Latitude
RCC5417	67.48
RCC1613	57.57
RCC685	54.18
RCC1615	50.2
RCC1868	48.75
RCC4222	42.48
A8	42.48
G11	42.48
A1	42.48
C3	42.48
E2	42.48
F8	42.48
G8	42.48
RCC4752	40.75
RCC716	-14.48

Banyuls strains

Identification and description of genetic diversity through pangenomic approach to study role of diversity on adaptation

Quantification and description of genomic diversity through analysis of *Bathycoccus* pangenome

Pangenome

Repertory of all sequences/genes found in a species

Number of new sequences discovered after each new genome sequenced follow a decreasing power law

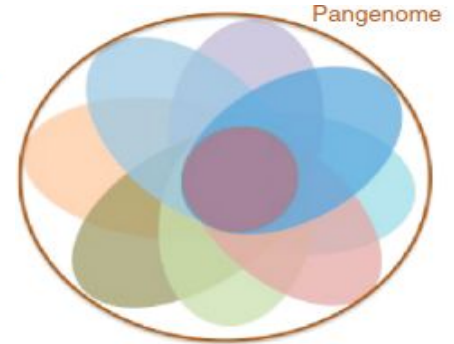
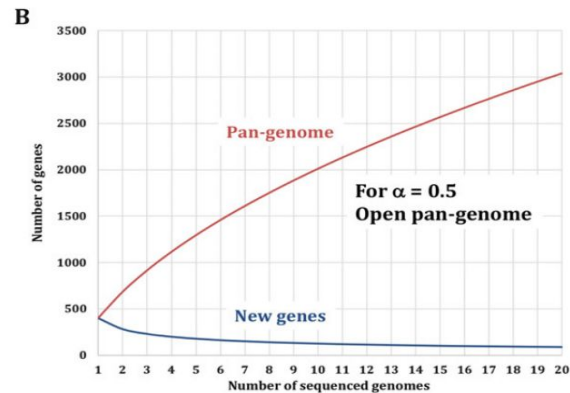
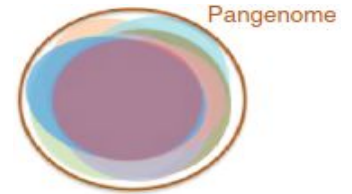
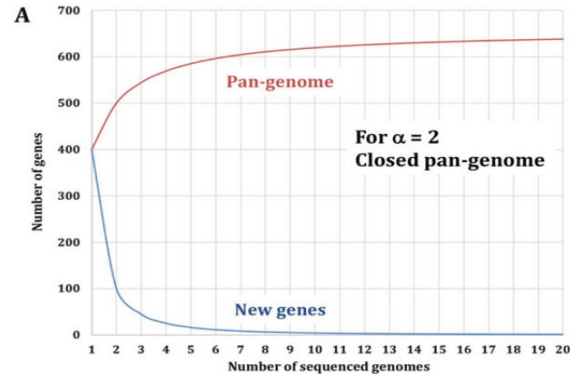
$$n = \kappa N^{-\alpha}$$

Closed Pangenome

Limited diversity

Open Pangenome

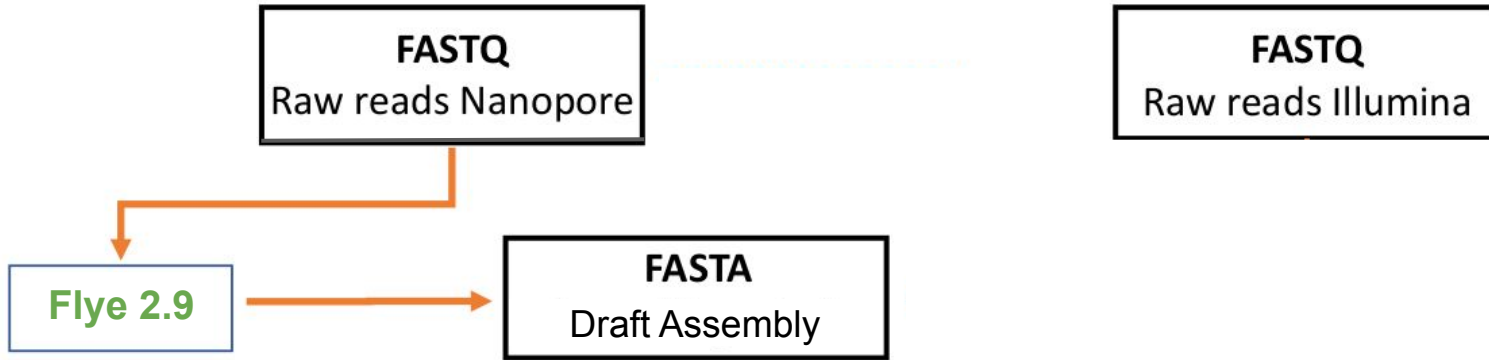
Boundless diversity



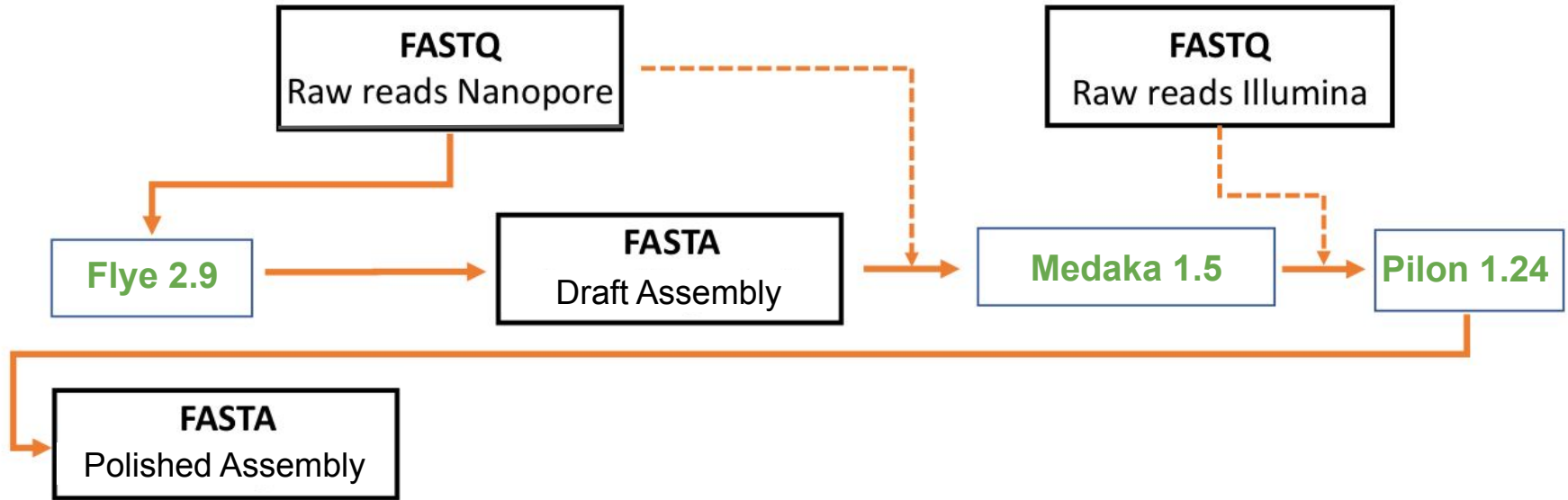
De novo assembly pipeline for Bathycoccus



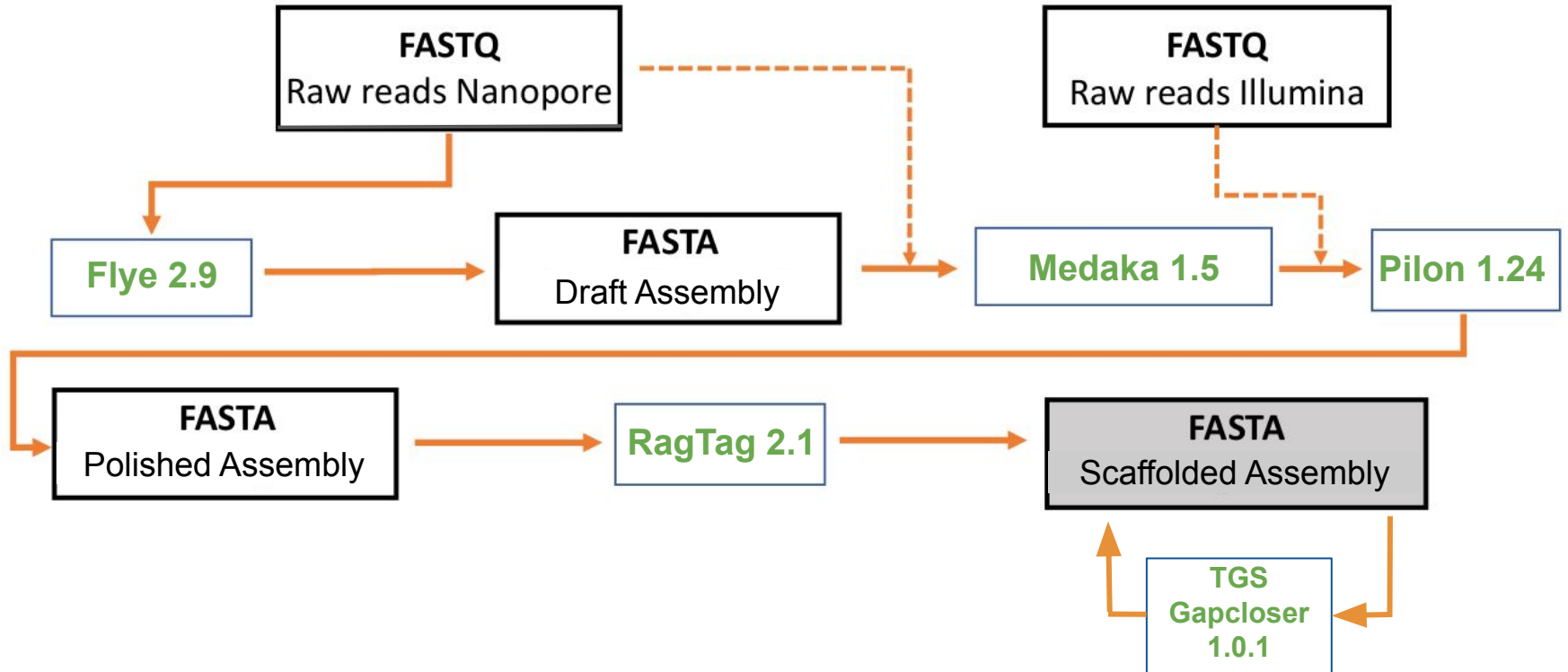
De novo assembly pipeline for Bathycoccus



De novo assembly pipeline for Bathycoccus



De novo assembly pipeline for Bathycoccus



De novo assembly of 15 strains of Bathycoccus

Assembly size ranging from
15.13 Mb to 15.93 Mb

Average BUSCO completion score
95.34%

Average number of gap in assembly
3

Average quality score
40.17

**15 High Continuity, High Quality
assemblies usable for
pangenome construction**

Strain	Genome size	BUSCO	Gap	Quality value
E2	15.40 Mb	97.10%	1	47.7337
1613	15.47 Mb	97.00%	1	45.4014
A8	15.33 Mb	97.00%	0	47.3761
4222	15.19 Mb	96.80%	0	41.3156
5417	15.37 Mb	96.70%	4	42.3836
685	15.93 Mb	96.70%	0	40.0689
G11	15.28 Mb	96.70%	0	45.544
716	15.35 Mb	95.80%	7	33.0739
1868	15.68 Mb	95.60%	2	29.8042
4752	15.39 Mb	95.30%	5	38.052
1615	15.40 Mb	94.10%	21	31.142
A1	15.13 Mb	93.50%	2	*
G8	15.68 Mb	93.00%	4	*
F8	15.20 Mb	92.70%	0	*
C3	15.33 Mb	92.10%	4	*

* Illumina
unavailable

Structure of Bathycoccus Pangenome

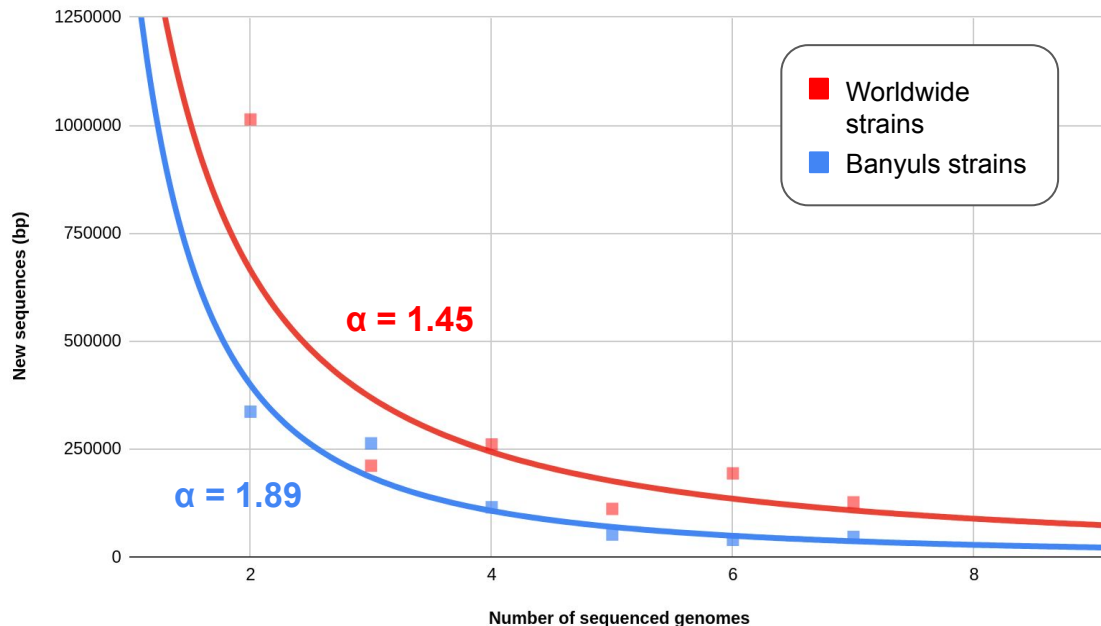
Worldwide diversity and Local diversity

> 2Mb of new accessory sequences
(15% of genome size)

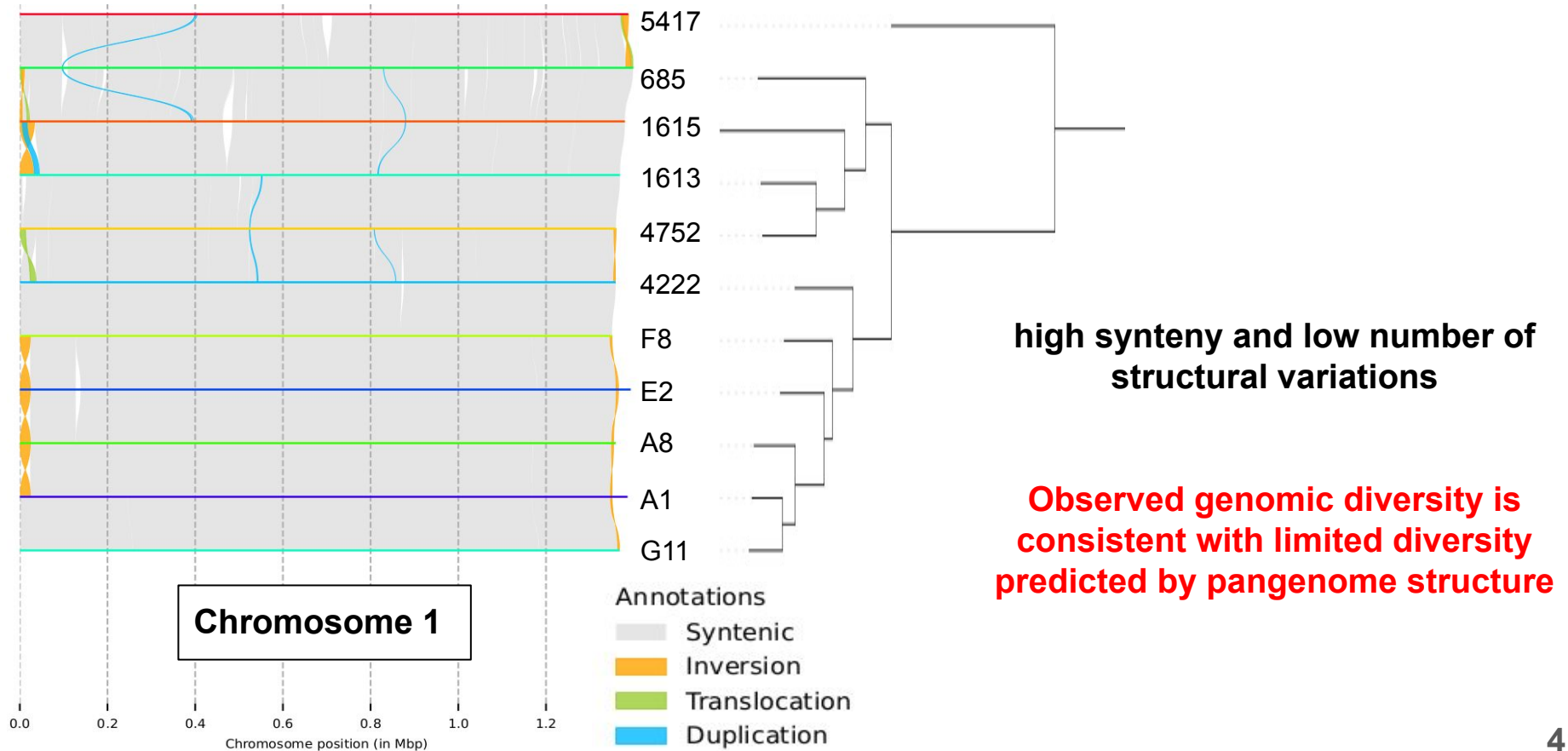
$$\alpha > 1$$

**Both worldwide and local
pangenome are closed**

**Sequence diversity seems limited
in Bathycoccus even though it's
a cosmopolitan organism with
high expected population size**



Structural variations in representative chromosome



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



Applications :

- Mapper des reads contre un génome *bwa, minimap2*
- Détecter 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*

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



Applications :

- Mapper des reads contre un génome *bwa, minimap2*
- Détecter 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)



Avec jupyter book : lancer les commandes + analyser les résultats

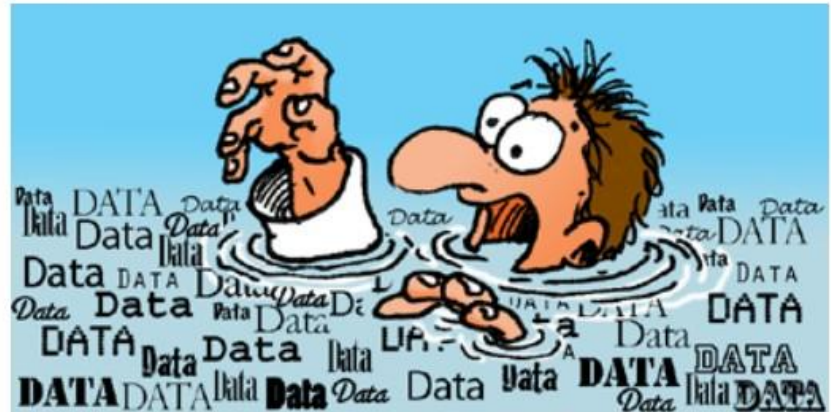
=> Avoir un plan de bataille opérationnel



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 : 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/>



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

Service formation IRD et CIRAD

- **Christel Gruau**
- **Isabelle Lecomte**

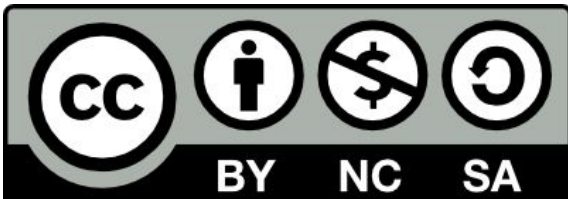
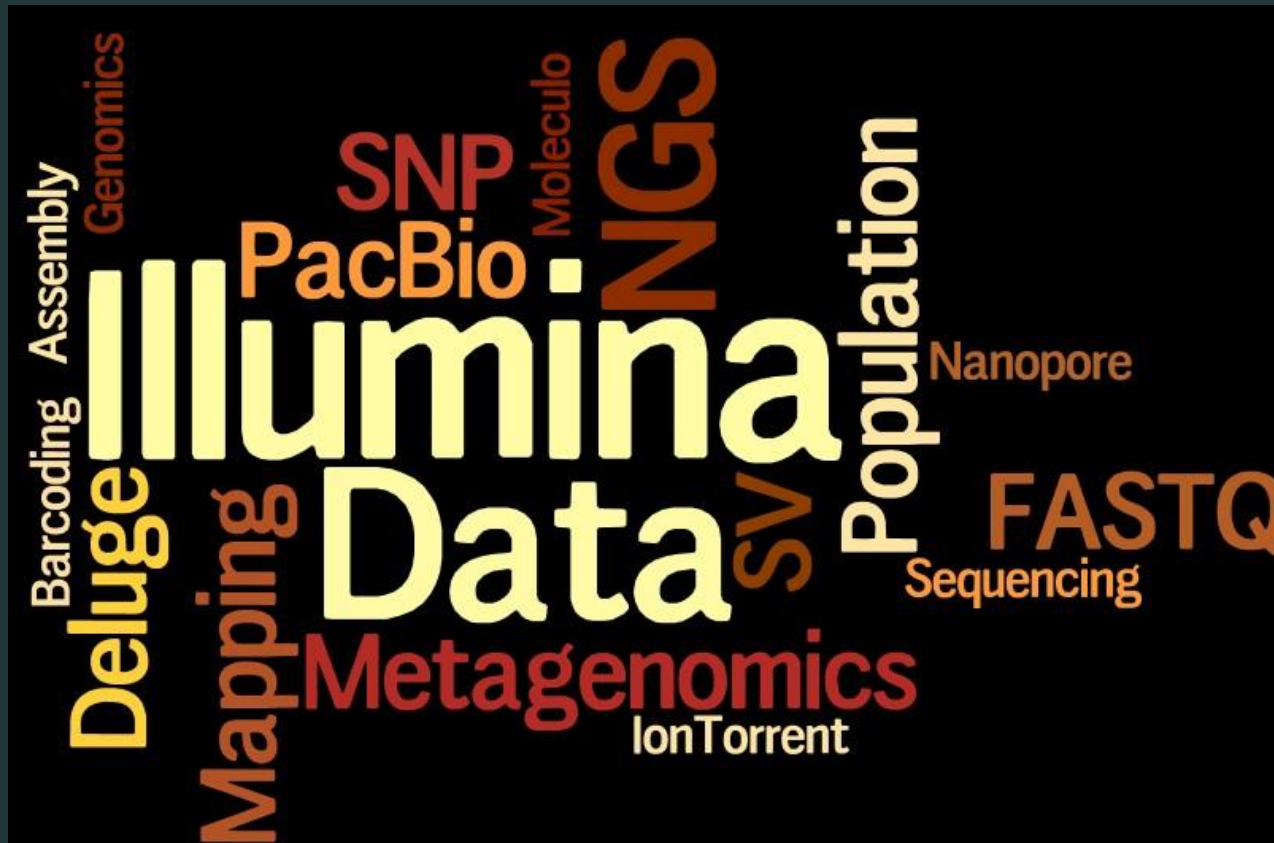


Core Cloud

- **Christophe Blanchet** et
toute l'équipe biosphere



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/>

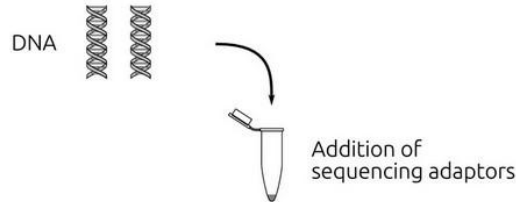
- **Alexis Dereeper**
- **Julie Orjuela-Bouniol**
- **François Sabot**
- **Christine Tranchant-Dubreuil**



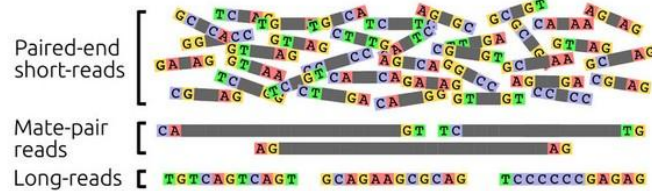
Variations around genotypes and SNP anthem

(a) Whole-genome *de novo* sequencing

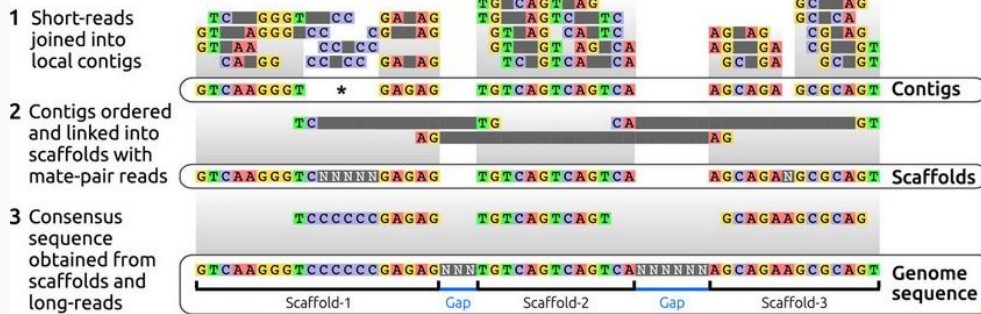
Library preparation



High-throughput sequencing

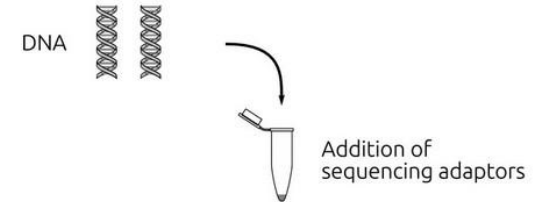


Read alignment and genome assembly



(b) Whole-genome resequencing

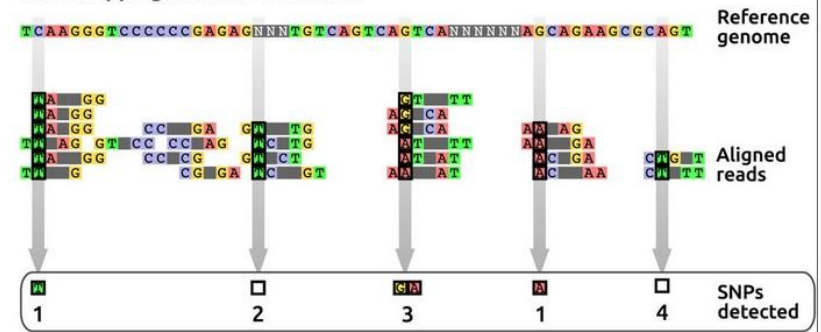
Library preparation



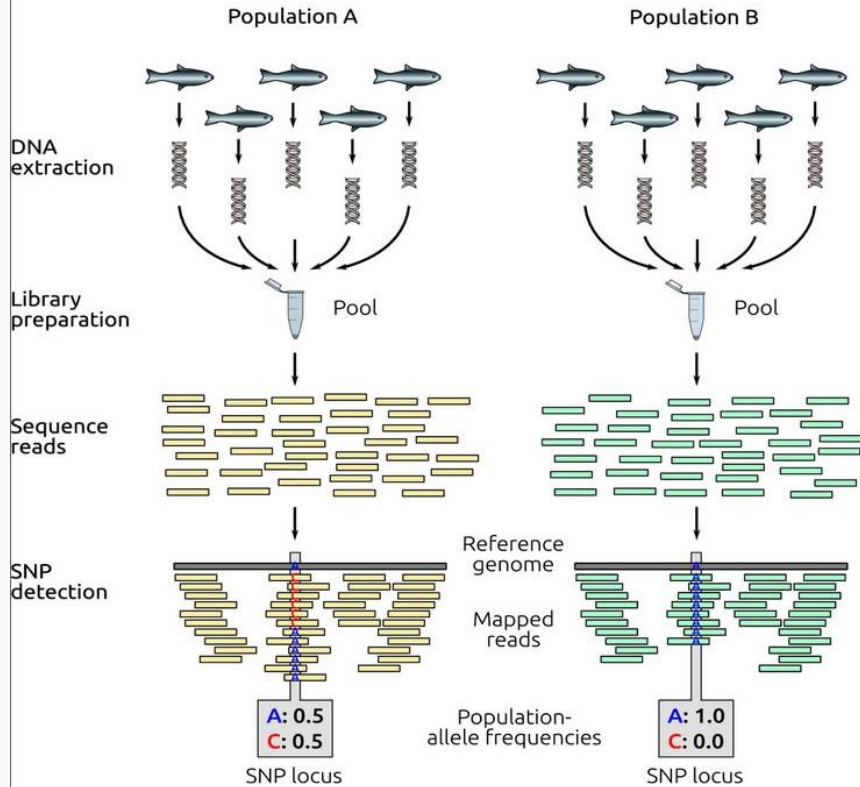
High-throughput sequencing



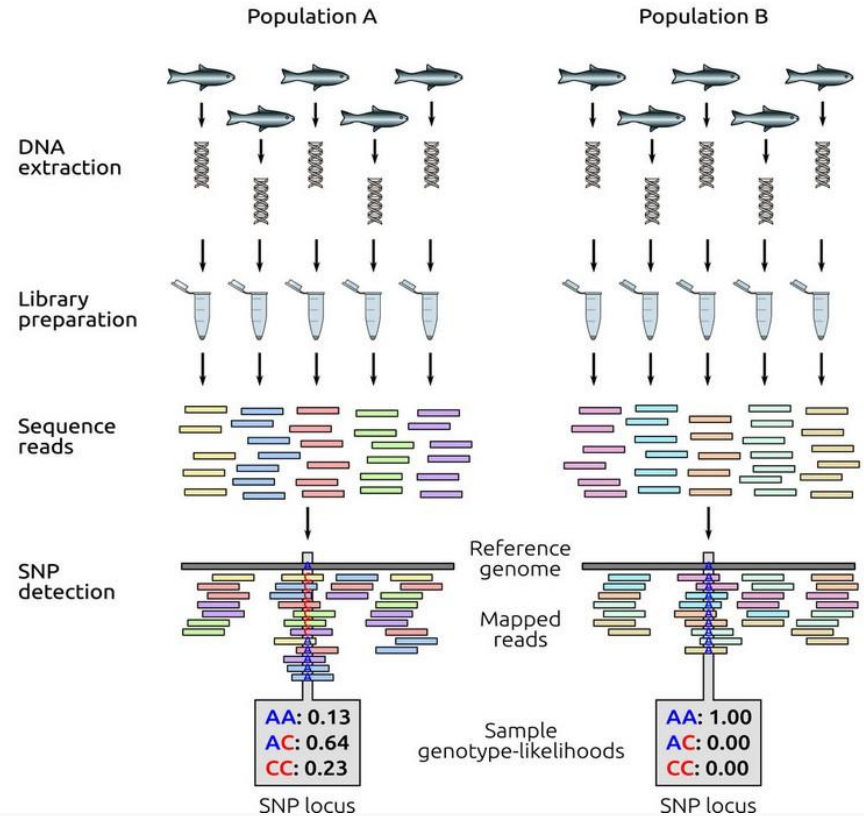
Read mapping and SNP detection



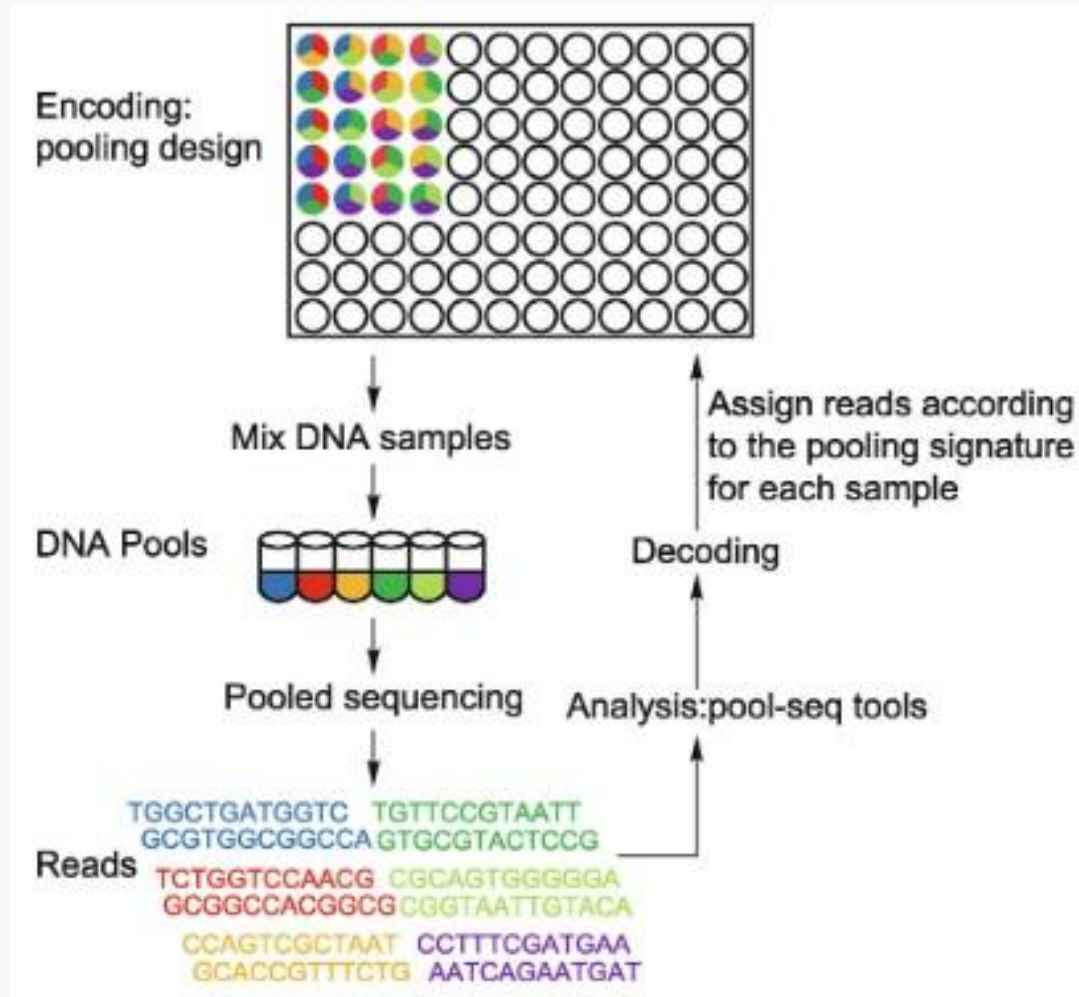
(a) High-coverage whole-genome resequencing of pooled DNA (Pool-seq)



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



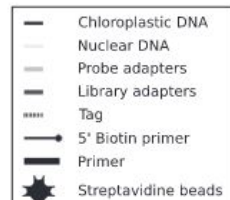
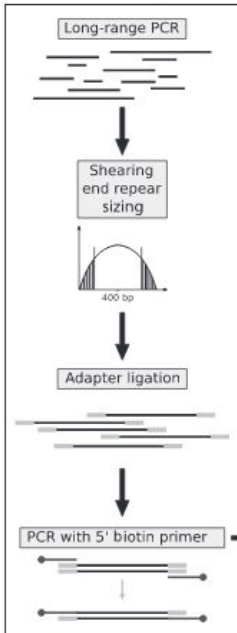
Fuentes-Pardo and Ruzzante, 2017



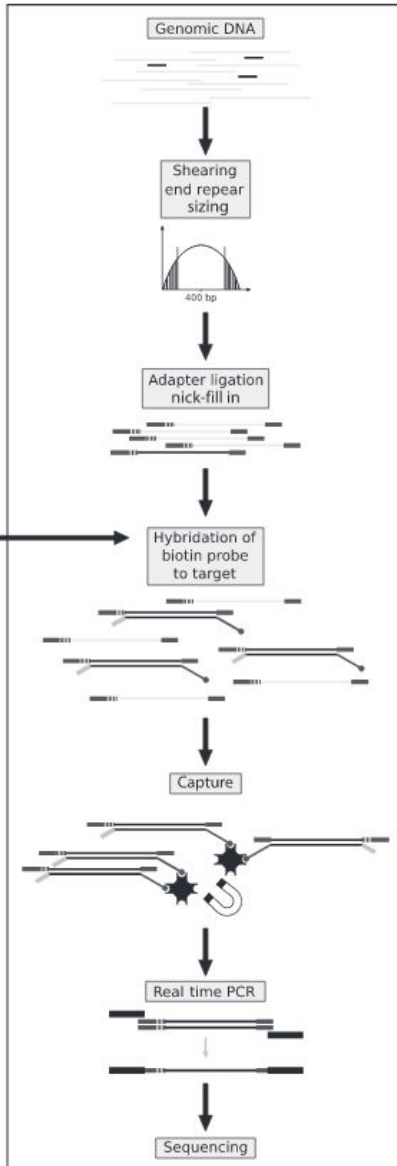
Cao-Sun et al, 2016

Phylogenomics using probes

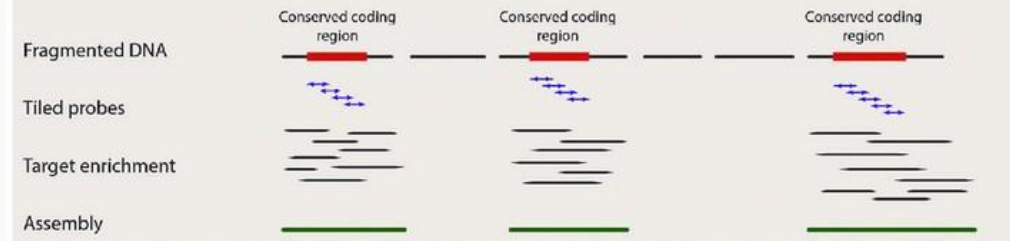
Production of probes



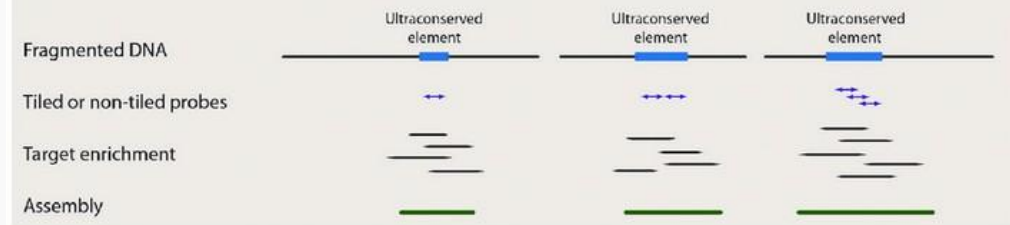
Production of enriched libraries



(A) Anchored hybrid enrichment



(B) Ultraconserved elements



Young and Gillung, 2019

Amplicon Generation Workflow

Create custom oligo capture probes
flanking each region of interest



CAT probes hybridize to flanking
regions of interest in **unfragmented** gDNA.



From Illumina

Extension/Ligation between Custom Probes
across regions of interest



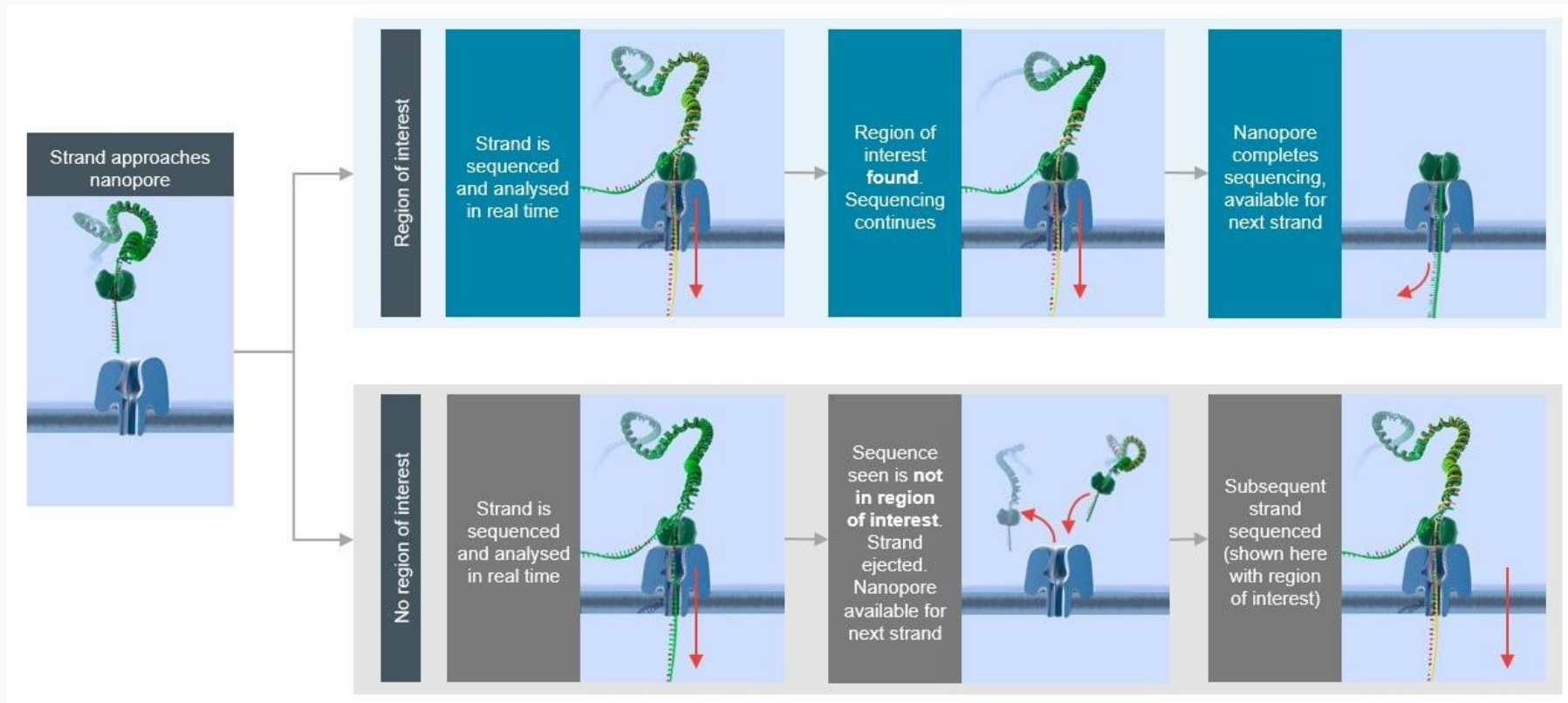
PCR adds indexes and sequencing primers



Uniquely tagged amplicon library ready
for cluster generation and sequencing

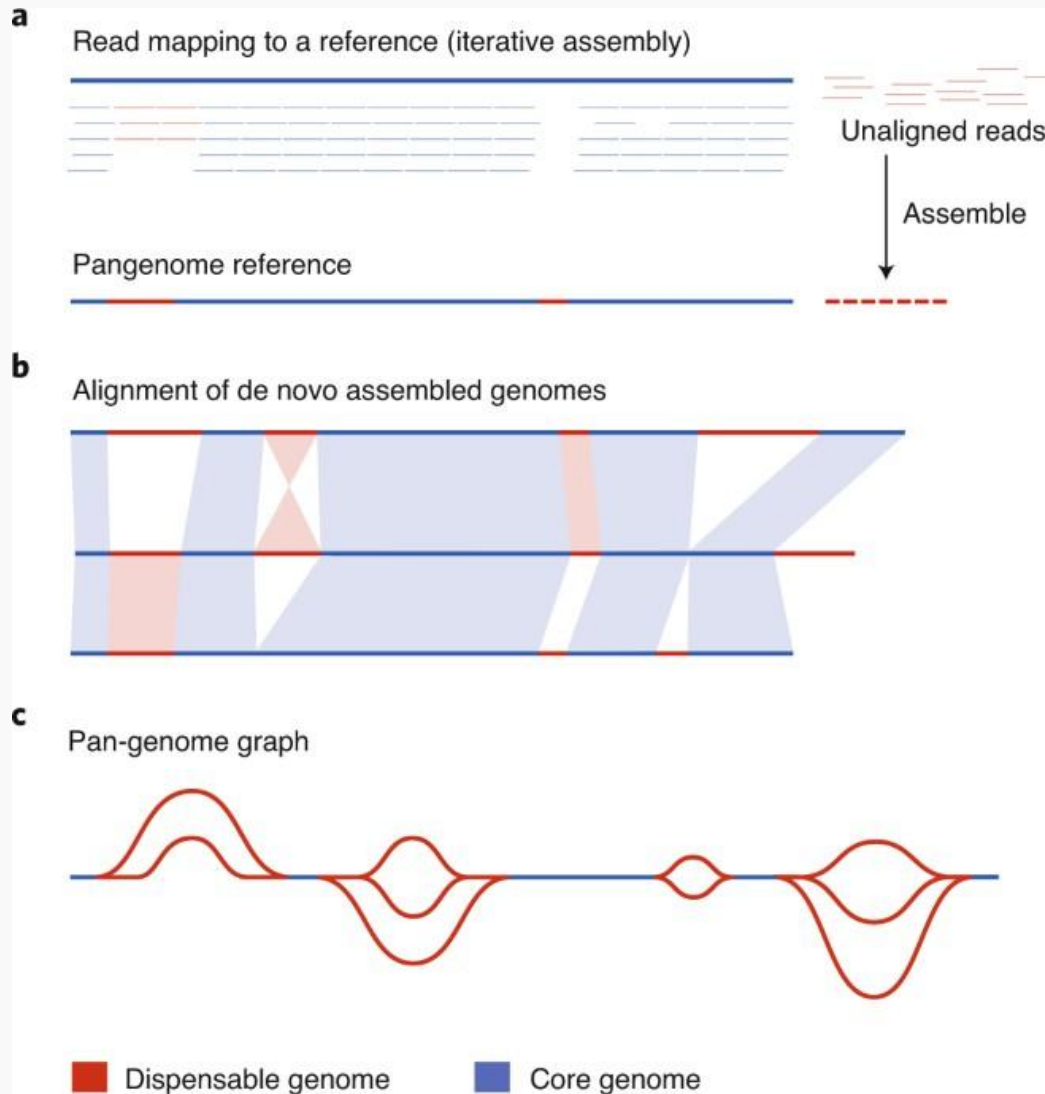


Adaptative sequencing

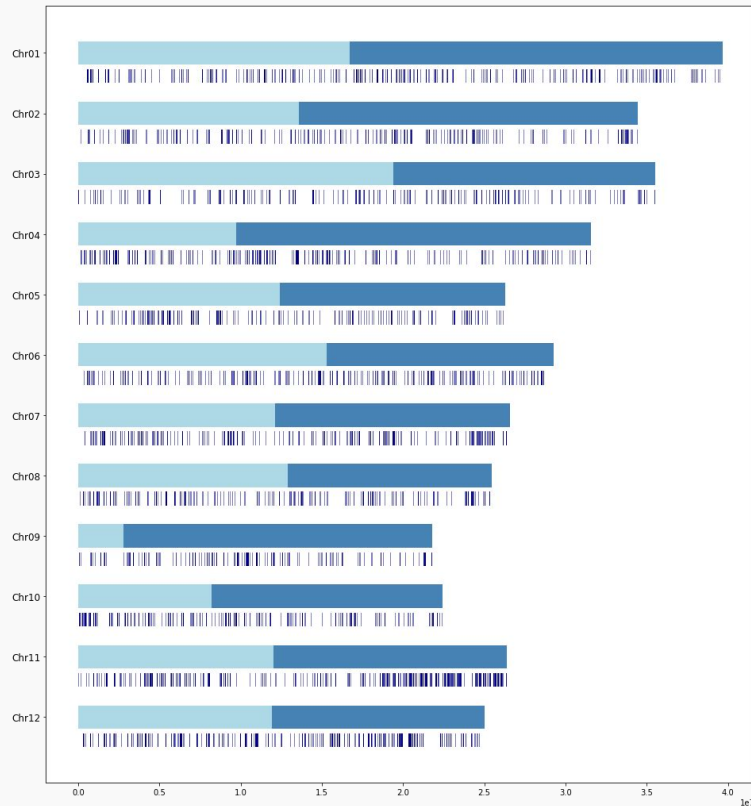


From Nanopore

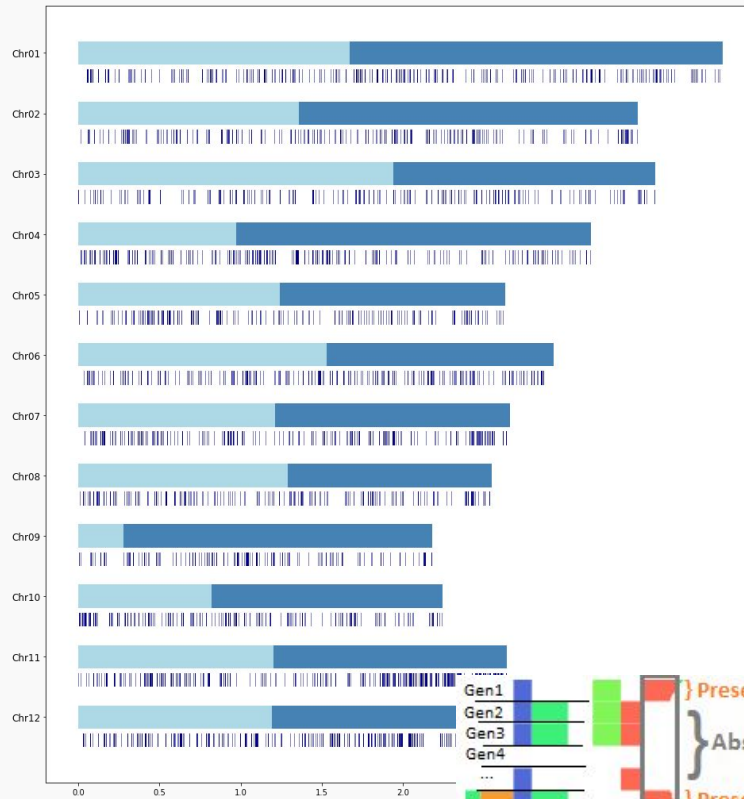
Other ways to analyse diversity...



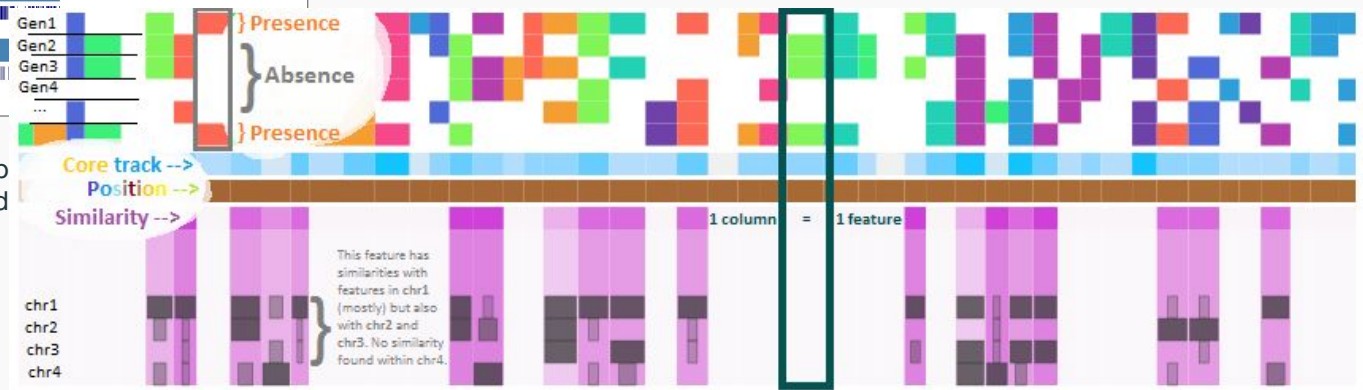
From Bayer et al, 2020



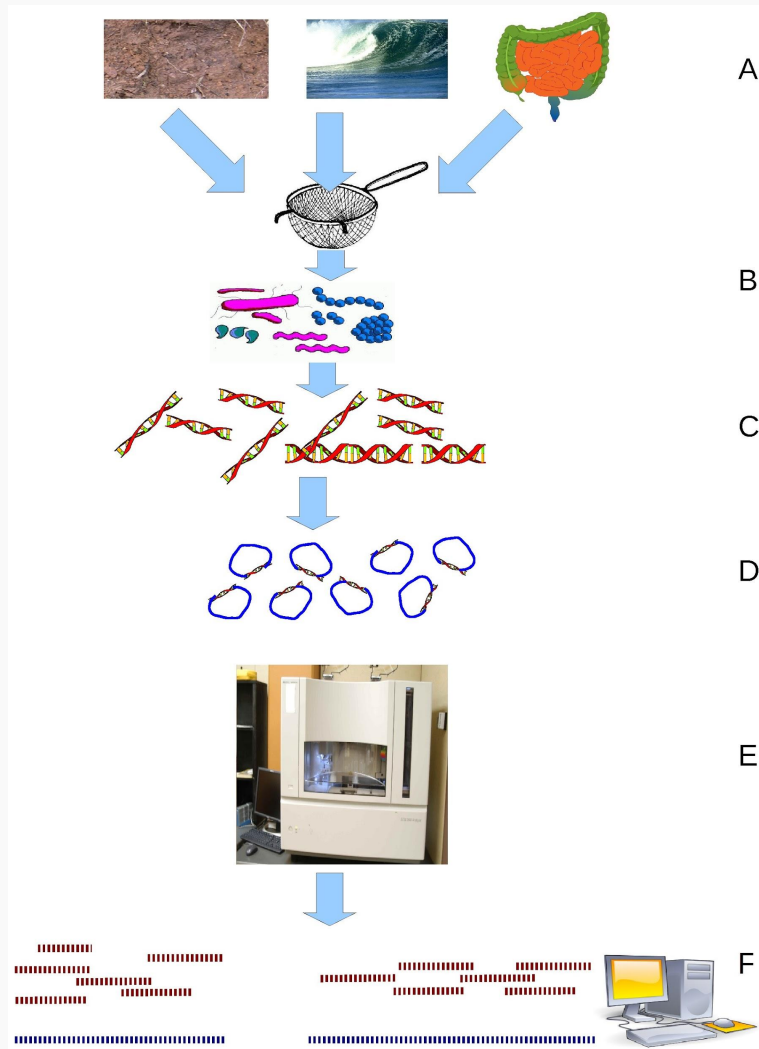
From Tranchant-Dubreuil et al, unpublished



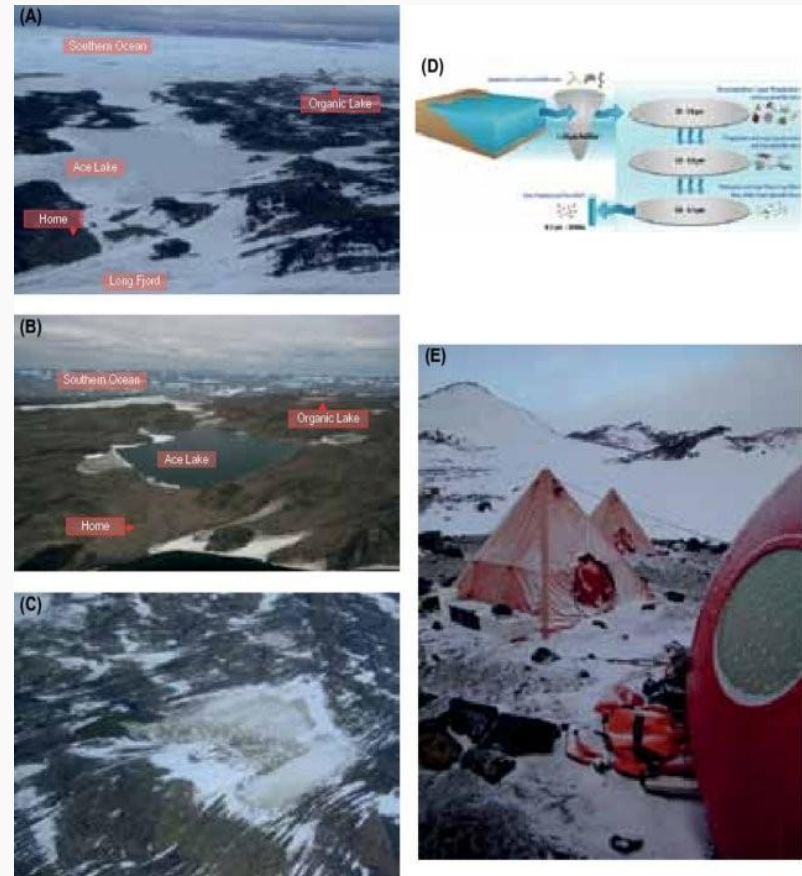
From Tranchant-Dubreuil et al, unpubl b d



From Durant et al, unpublished



From Wikipedia



From Yau and Cavicchioli, 2011