



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

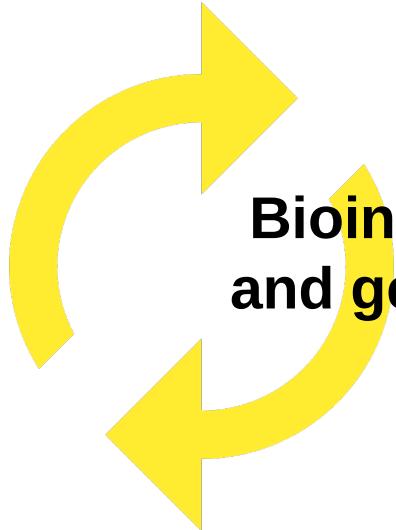
www.southgreen.fr

[https://southgreenplatform.github.io/
trainings](https://southgreenplatform.github.io/trainings)



Modules de formation 2022





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

comparative genomics
GWAS
population genetics
phylogeny
metagenomics
genome assembly
transcriptome assembly
SNP detection
structural variation
differential expression
polypliody



Rice



Banana



Palm



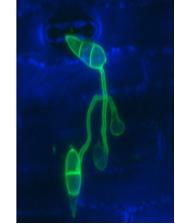
Sorghum



Coffee



Cassava



Microscopy

11101100
01110001
011001110001000



Larmande Pierre
Orjuela-Bouniol Julie

Sabot François
Tando Ndomassi
Tranchant-Dubreuil Christine



Comte Aurore
Dereepper 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

11101100
01110001
011001110001000

Workflow manager

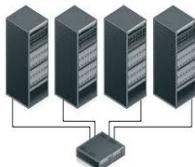
TOGGLE
Toolbox for generic NGS analyses



Galaxy

SNAKEMAKE

HPC and trainings....



Genome Hubs & Information System



Gigwa

Sample	Marker	Allele	Genotype	Quality
S1	M1	A	AA	Q1
S1	M2	B	AB	Q2
S1	M3	C	BC	Q3
S1	M4	D	CD	Q4
S2	M1	A	AA	Q1
S2	M2	B	AB	Q2
S2	M3	C	BC	Q3
S2	M4	D	CD	Q4
S3	M1	A	AA	Q1
S3	M2	B	AB	Q2
S3	M3	C	BC	Q3
S3	M4	D	CD	Q4

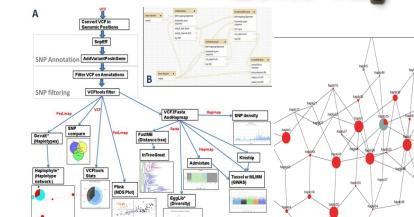
SNPs and Indels

GreenPhyl

Family Id	Family Name	Number of sequences	Status
GP000010	Cytochrome P450 superfamily	6042	Green
GP000017	AP2EDEBP transcription factor family; ERF/FOER group (partial)	5142	Green
GP000020	NAC transcription factor family	4574	Green
GP000028	MADS transcription factor family		
GP000018	Homeobox peroxisome superfamily		
GP000006	General substrate transporter		
GP000022	Sulfotransfase Serine Proteases family		
GP000019	NPF, NRT1/PTR FAMILY		

Gene families

SNiPlay



<https://github.com/SouthGreenPlatform>



@green_bioinfo

The South Green portal: a comprehensive resource for tropical and Mediterranean crops

I-Trop

Plant & Health Bioinformatics Platform



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



AURORE
COMTE



ALEXIS
DEREPEPER



BRUNO
GRANOUILLAGC



JULIE
ORJUELA



NDOMASSI
TANDO



CHRISTINE
TRANCHANT

bioinfo@ird.fr

IE bioinfo

IE bioinfo

IE systèmes
d'information

IE bioinfo

IE systèmes

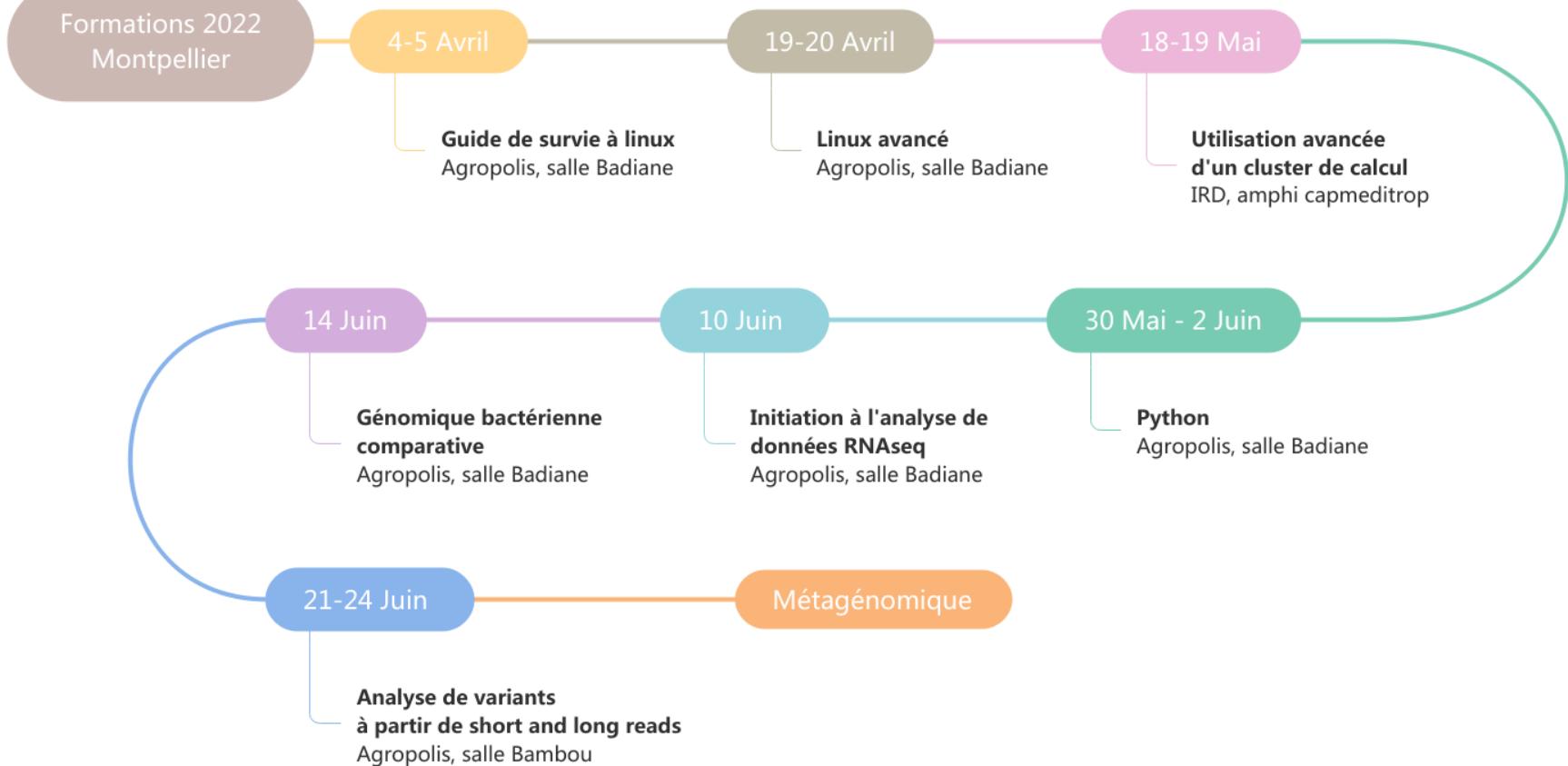
IR bioinfo



@ItropBioinfo

SouthGreen

bioinformatics platform





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](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*
- 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) à partir de :
 - reads mappées contre un génome - *breakdancer, sniffle*
 - génomes entiers - *nucmer, assemblytics, siry*

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



Applications :

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

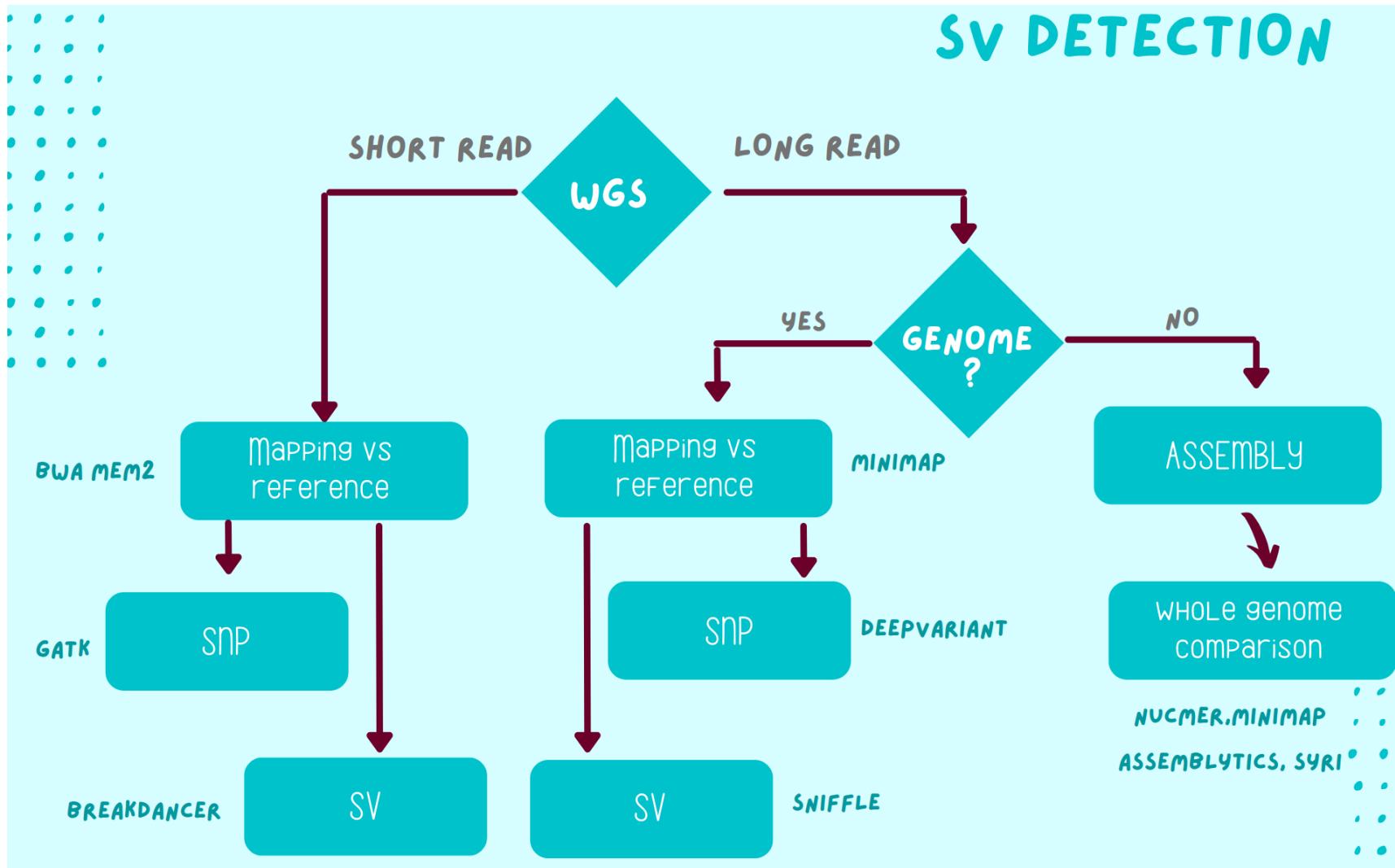
des variants structuraux (SV)

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

Working environment

What is jupyter book ?

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

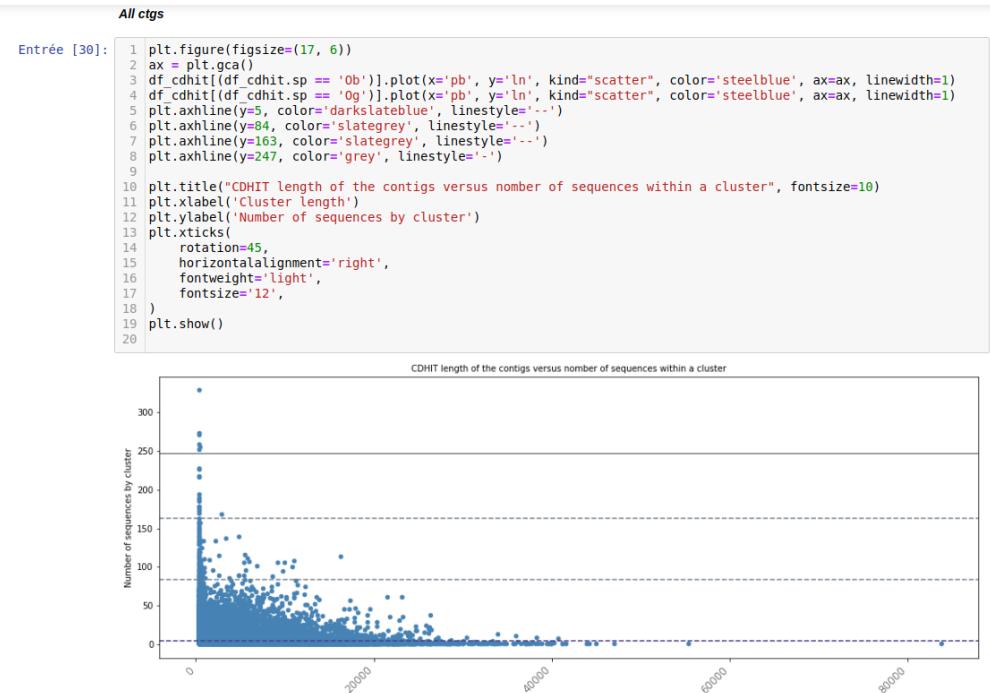
ex : Bash (Linux), Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala



Why use jupyter book ?

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

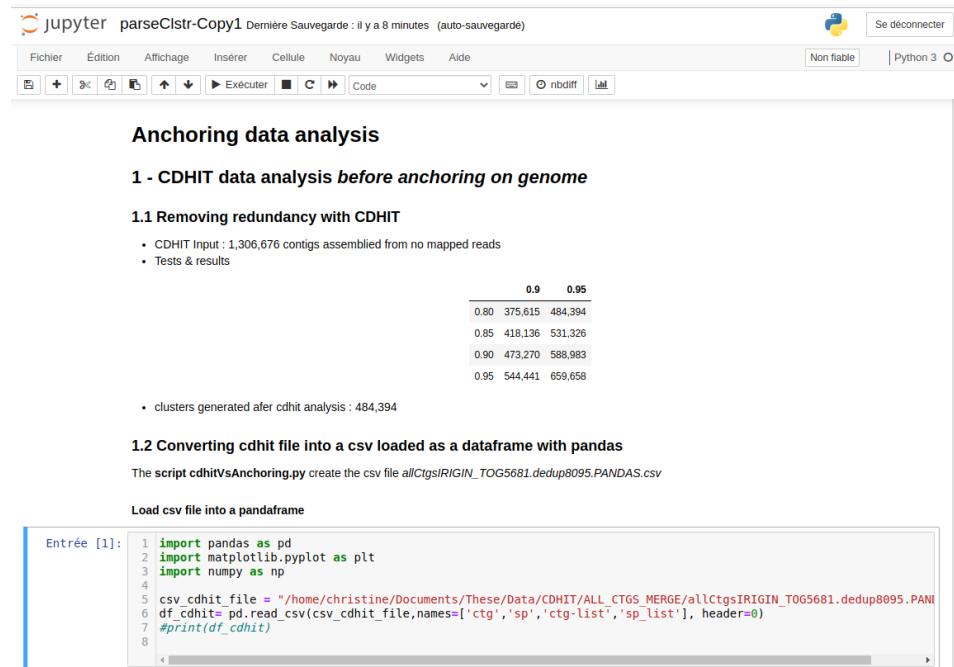
- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook



Why use jupyter book ?

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

- code can be executed inside each cell of the notebook
- code output is directly displayed in the notebook
- explanations, formulas, charts can be added



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

- Title Bar:** jupyter parseCstr-Copy1 Dernière Sauvegarde : il y a 8 minutes (auto-sauvegardé) | Se déconnecter | Python 3
- Toolbar:** Fichier, Édition, Affichage, Insérer, Cellule, Noyau, Widgets, Aide, Exécuter, Code, nbdiff, Lab
- Section Headers:** Anchoring data analysis, 1 - CDHIT data analysis before anchoring on genome, 1.1 Removing redundancy with CDHIT, 1.2 Converting cdhit file into a csv loaded as a dataframe with pandas.
- Text Output:** 1 - CDHIT data analysis before anchoring on genome
- Data Output:** A table showing cluster statistics:

	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
- Code Output:** Entrée [1]:

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
4
5 csv_cdhit_file = "/home/christine/Documents/These/Data/CDHIT/ALL_CTGS_MERGE/allCtgsIRIGIN_TOG5681.dedup8095.PAN
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 ?

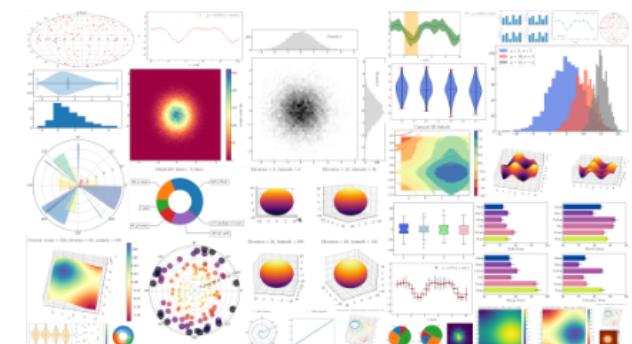
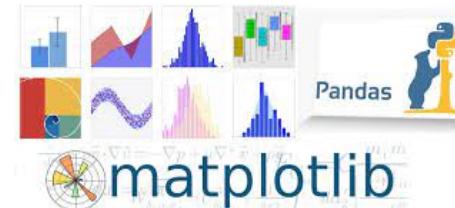
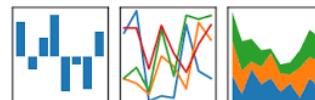


- One file to analyze data and generate reports
- Can be exported to many formats, including PDF and HTML, which makes it easy to share your project with anyone.
- Analysis are more transparent, repeatable and shareable

How to become a super

- 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

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



How will you use Jupyter

- 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

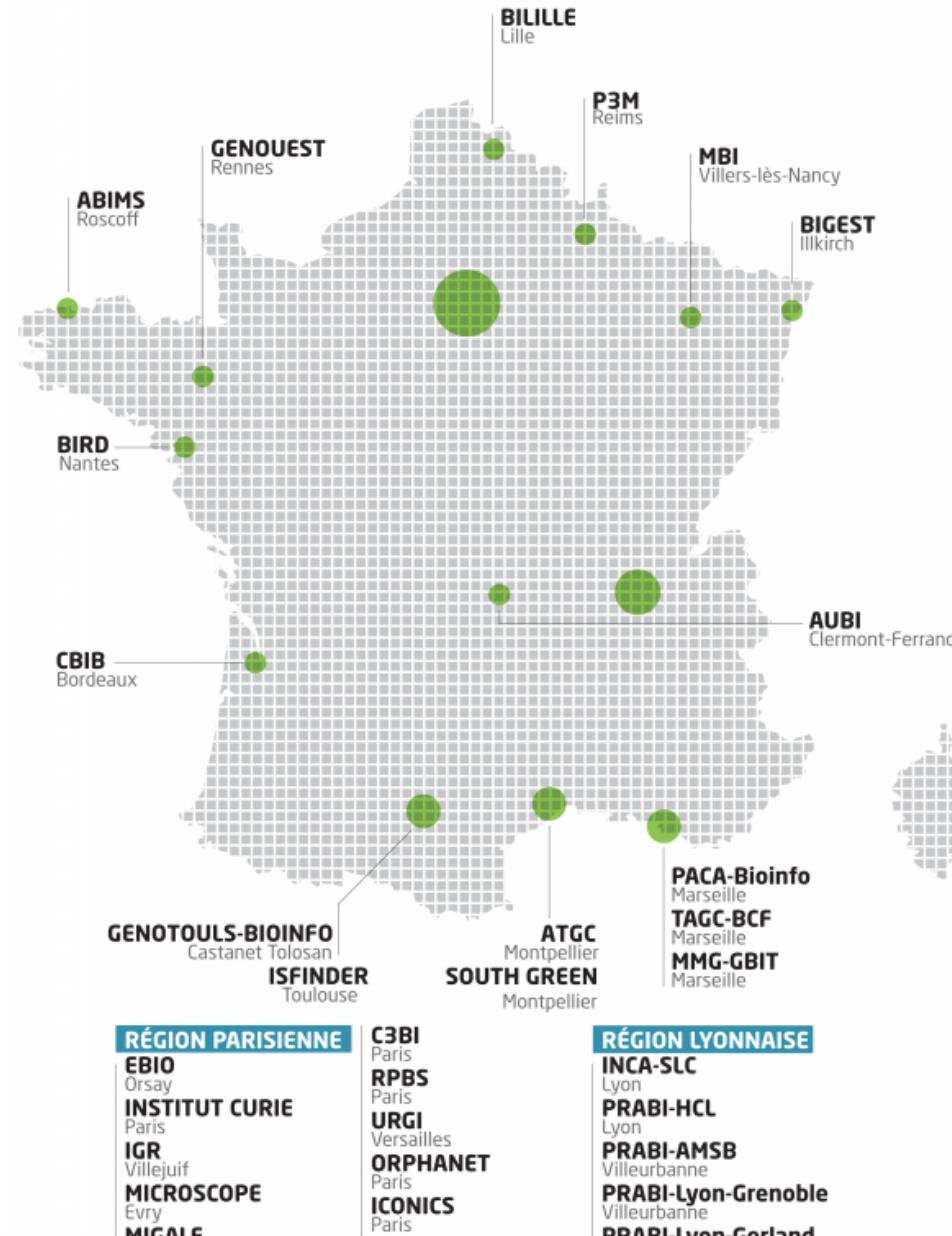


The screenshot shows a web-based interface for the IFB Cloud. The top navigation bar includes tabs for "IFB Cloud", "mydatalocal", and a "+" button. The address bar shows the URL <https://134.158.247.8/tree/mydatalocal>. The main content area features a "jupyter" logo and three tabs: "Files", "Running", and "Clusters". Below these tabs, a message says "Select items to perform actions on them." A file list shows "0" files in the "mydatalocal" folder. A message at the bottom states "La liste des notebooks est vide." On the right side, there is a sidebar with "Upload" and "New" buttons, and a dropdown menu for "Notebook" containing options like "Bash", "Julia 1.5.3", "Python 3", and "R". Another section labeled "Other:" includes "Text File", "Folder", and "Terminal".



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

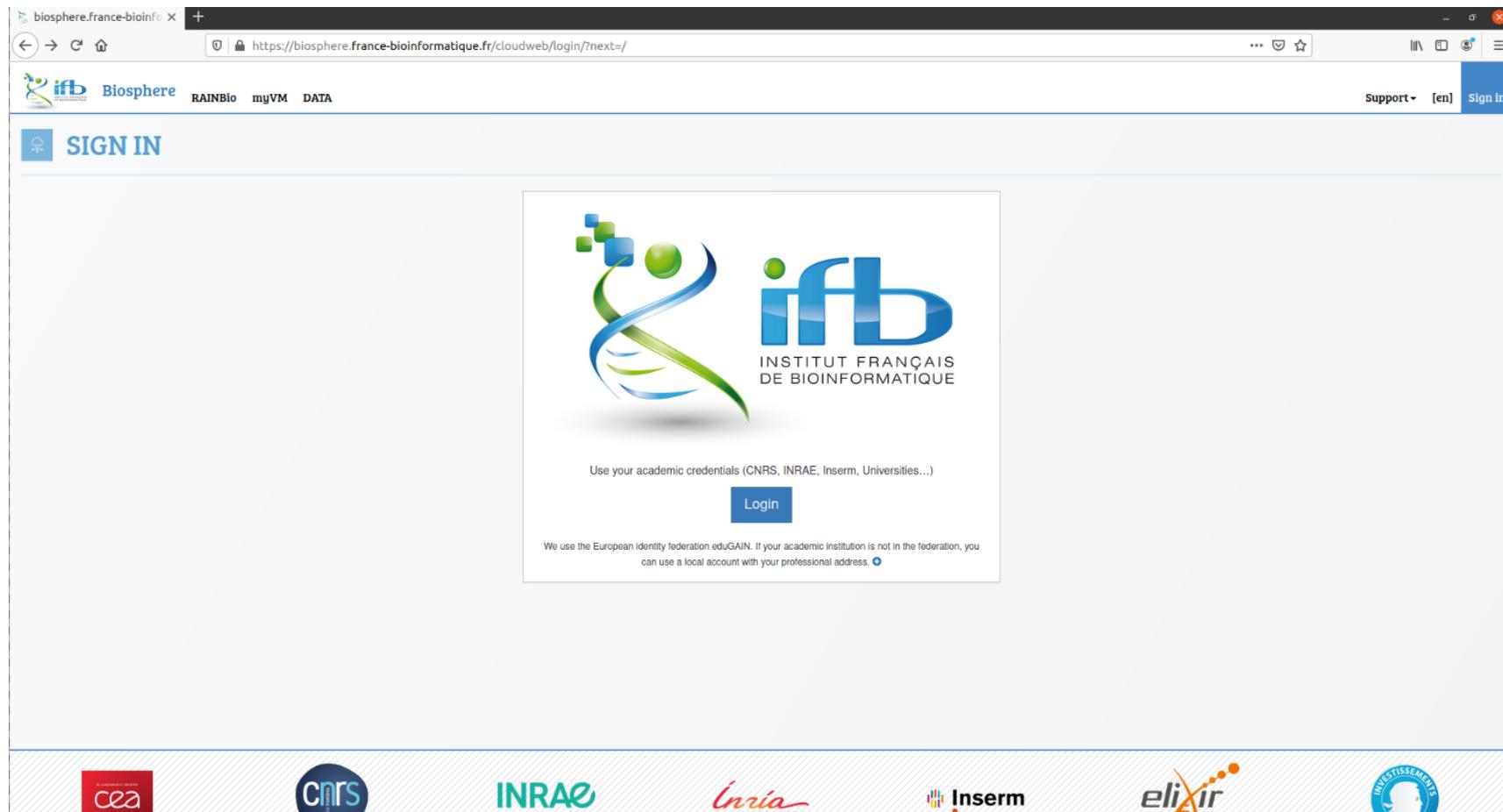
22 plateformes-membres
7 plateformes contributrices
8 équipes associées
>400 experts (~200 FTE)



- A federation of clouds, which relies on interconnected IFB's infrastructures, providing distributed services to analyze life science data
- Access to a large set of virtual machines (computing 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



The screenshot shows a web browser window for the URL <https://biosphere.france-bioinformatique.fr/cloudweb/login?next=/>. The page features a header with the IFB logo and navigation links for RAINBio, myVM, and DATA. A "SIGN IN" button is visible on the left. The main content area displays the IFB logo and the text "INSTITUT FRANÇAIS DE BIOINFORMATIQUE". Below the logo, there is a link to use academic credentials from CNRS, INRAE, Inserm, and Universities. A "Login" button is centered. A note at the bottom states: "We use the European identity federation eduGAIN. If your academic institution is not in the federation, you can use a local account with your professional address." Logos for CEA, CNRS, INRAE, Inria, Inserm, elixir, and INVESTISSEMENTS are displayed along the bottom.

RAINBIO catalog to access our Virtual Machine (VM)



WELCOME ON BIOSPHERE, IFB CLOUD FOR LIFE SCIENCES

French Institute of Bioinformatics (IFB) provides life scientists with a federation of clouds, **Biosphere**, and bioinformatics cloud services to analyze life science data. Biosphere is used for scientific production in the life sciences, developments, and to support events like cloud and scientific training sessions, hackathons or workshops.

BIOSPHERE FEATURES

With IFB-Biosphere, you get:

- Pre-defined bioinformatics environments, available with an one-click deployment from the RAINBio catalogue
- Single sign-on, with your academic credentials for all sites (Sign in)
- An unified portal (Biosphere portal) for users to deploy virtual machines on all clouds
- More than 5,400 vCPU and 27 TB RAM (System status)
- Modular cloud environments, from a single virtual machine (VM) to tenths of VMs, up to 48 vCPU and 250 GB of RAM per VM
- Big memory VM with up to 3 TB RAM
- Admin right in your environment to tune bioinformatics tools and configurations
- High-availability thanks to the different sites of the federation
- Usual public biological reference databases
- Support for your training or workshop (CPU, RAM, storage, IFB experts)

BIOSPHERE PORTAL

The Biosphere portal provides high-level cloud interfaces

- The RAINBio catalogue to select the bioinformatics cloud appliances, useful for your analysis.
- An operating center to manage your cloud deployments with single virtual machines (VM) or complex applications with multiple VM
- A data center showing the public reference databases available in the IFB's clouds.

BIOINFORMATICS APPLIANCES

The bioinformatics appliances available in the RAINBio catalogue provide scientists with thematic environments, base appliances, web-based and desktop appliances.

Bioinformatics apps :

- Genomics, genome annotation, comparative genomics, transcriptomics, metagenomics,
- bioimaging,
- metabolic pathway, metabolomics,
- microbial ecology
- ...

Base apps :

- Linux distributions: Ubuntu, Debian, CentOS, with apt/yum, pip, docker (compose & swarm), ansible, conda (bioconda and R channels),
- Workflow environments (BioPipes), with Nextflow, SnakeMake, CWLtool, conda + Bioconda, Docker + Compose
- Web-based: Rstudio (BioManager, any R modules), Rshiny (Bring your own Rshiny app), Jupyter (Kernels Python & R), Wellinverter...
- Graphical Desktop (XFCE4 + X2Go), Biominer (Fiji/Icv), Cytoscape

Searching for the vm we will

vm's name : **analysesSV**



The screenshot shows the RAINBIO interface for managing cloud appliances. At the top, there are navigation tabs: IFB Biosphere, RAINBio (selected), myVM, and DATA. On the right, there is a user profile for christine.tranchant@ird.fr (eduGAIN) and a support link. A search bar on the right contains the text "analyses". Below the tabs, a search result for "AnalysesSV" is highlighted in green. The result shows the following details:

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

Other results listed include "CoursAnalysesNanoporeSG", "NGSanalysisJupyter", and "REPET". A note at the bottom states: "Le code couleur reste le même pour une même appliance."

Let's run your vm through the

Appliance AnalysesSV ☆ DEV

 Exporter en md

Description

This IFB cloud appliance provides both the Jupyter Notebook and Lab environment (see [explanations](#)) to work on the structural variants detections on short and long reads.

This Jupyter app is based on the Jupyter Docker Stacks (see [details](#)). By default, this Biosphere app uses the stack `jupyter/datascience-notebook` but users can choose any other existing stack with an Advanced deployment in Biosphere portal.
In addition, we integrated various tools to perform the SV detection

Tools

- Bash kernel for jupyter
- Pandas
- Matplotlib
- Jupyter notebook/lab
- seqtk
- Minimap2
- BWA-MEM2
- Samtools/BCFtools
- BEDtools
- VCFtools
- GATK
- 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	https://hub.docker.com/r/francoissabot/trainingontvm
----------	---

Support

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) https://github.com/SouthGreenPlatform/training_SV_VM

App de base Jupyter

Caractéristiques

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

Crédits

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

Let's run your vm through the

IFB Biosphere RAINBio myVM DATA

christine.tranchant@ird.fr (eduGAIN)

LANCEUR DÉPLOIEMENT AVANCÉ

Appliance AnalysesSV ★ DEV

Exporter en md

Description

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

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

- François Sabot SouthGreen Platform
- Julie Orjuela-Bouniol SouthGreen Platform

App data

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

Licence

Licensed under GPLv3

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

Configurer le déploiement d'une appliance

Déploiement de l'appliance "AnalysesSV"

Name: CTranchant

Groupe à utiliser: DIADE (DIversité, Adaptation)

Cloud: ifb-core-cloudbis

Gabarit d'image cloud: vCPU.h

Quelle gabarit d'image doit être utilisé sur ce 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...

IFB Biosphère RAINBio myVM DATA

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

CLOUD

Déploiements [?](#)

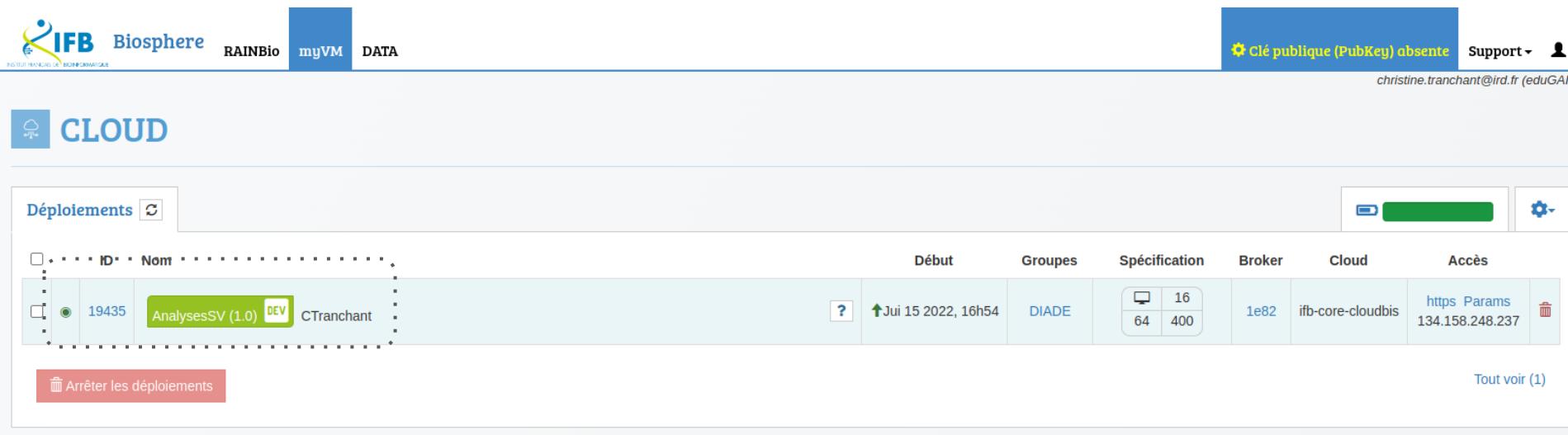
ID	Nom	Début	Groupes	Spécification	Broker	Cloud	Accès		
19435	AnalysesSV (1.0) DEV CTranchant	↑Jui 15 2022, 16h54	DIADE	<table border="1"><tr><td>16</td></tr><tr><td>64</td></tr></table>	16	64	1e82	ifb-core-cloudbis	Arrêter les déploiements Tout voir (1)
16									
64									

Arrêter les déploiements Tout voir (1)

Appliances et déploiements favoris Déploiements récemment terminés [?](#) Quota

ID	Broker	Nom	Der. dém.	Paramétrage

ready !



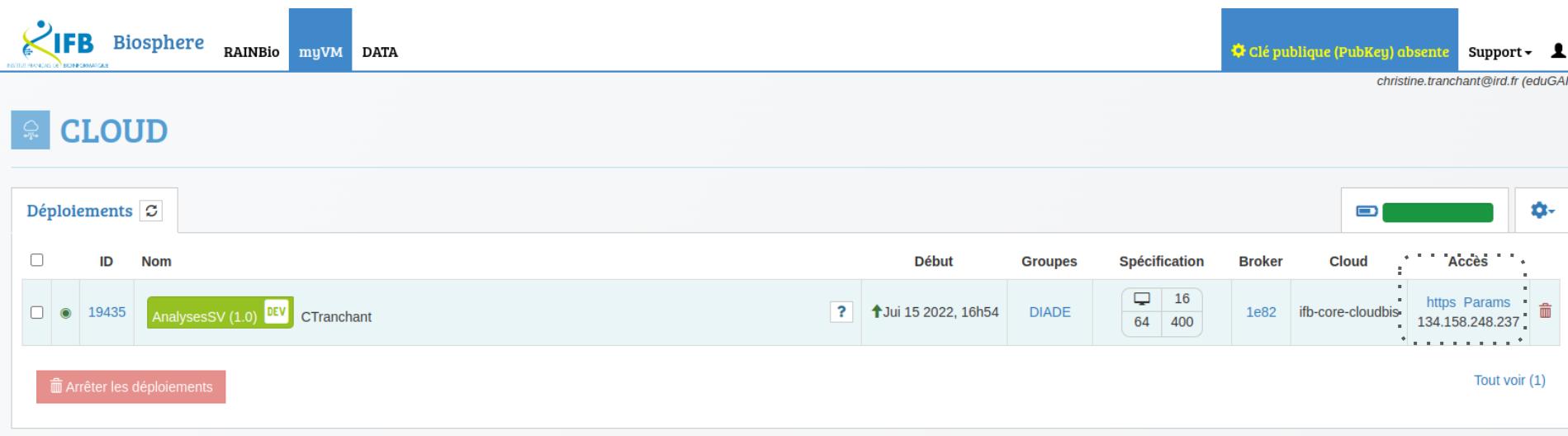
The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there is a navigation bar with the IFB Biosphere logo, RAINBio, myVM (which is selected), DATA, and a support section indicating a public key (PubKey) is absent. The email christine.tranchant@ird.fr (eduGAI) is also shown.

The main area is titled "CLOUD". Under "Déploiements", a table lists a single deployment:

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

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

get the url... link “https”



The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there is a navigation bar with the IFB Biosphere logo, RAINBio, myVM (which is highlighted in blue), DATA, and a message indicating a "Clé publique (PubKey) absente". The user email "christine.tranchant@ird.fr (eduGAI)" is also visible. Below the navigation bar, there is a "CLOUD" section with a "Déploiements" 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>64</td><td>16</td></tr><tr><td>400</td><td></td></tr></table>	64	16	400		1e82	ifb-core-cloudbis	https Params 134.158.248.237
64	16										
400											

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

Get the token identifiant... link “Params”

Screenshot of the SouthGreen bioinformatics platform interface showing the "myVM" section.

The main navigation bar includes: IFB Biosphère, RAINBio, myVM, CLOUD, Déploiements, and a user icon.

A modal window titled "Paramètres" is open, displaying a single parameter:

nom	valeur
JUPYTER_TOKEN	7b708d2b24d1947f0a48baf37f0986c9ca4774d8157a8825

The "Déploiements" section shows one deployment entry:

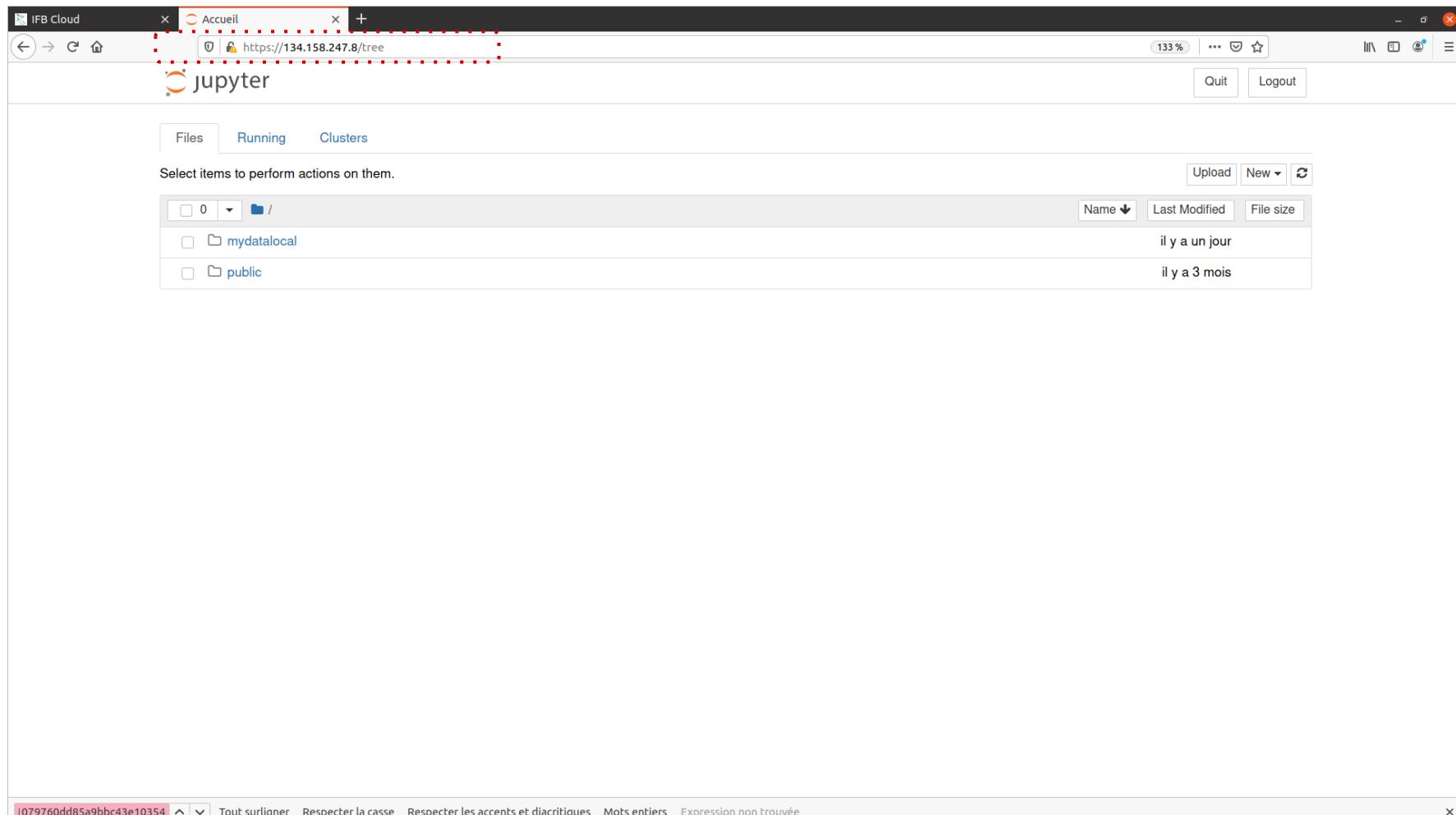
ID	Nom	Début	Groupes	Spécification	Broker	Cloud	Accès				
19435	AnalysesSV (1.0) DEV CTranchant	↑Jui 15 2022, 16h54	DIADE	<table border="1"><tr><td>64</td><td>16</td></tr><tr><td>400</td><td></td></tr></table>	64	16	400		1e82	ifb-core-cloudbis	https Params 134.158.248.237
64	16										
400											

Buttons at the bottom of the deployment table include "Arrêter les déploiements" and "Tout voir (1)".

The "Appliances et déploiements favoris" section is visible at the bottom.

Let's run our vm through the

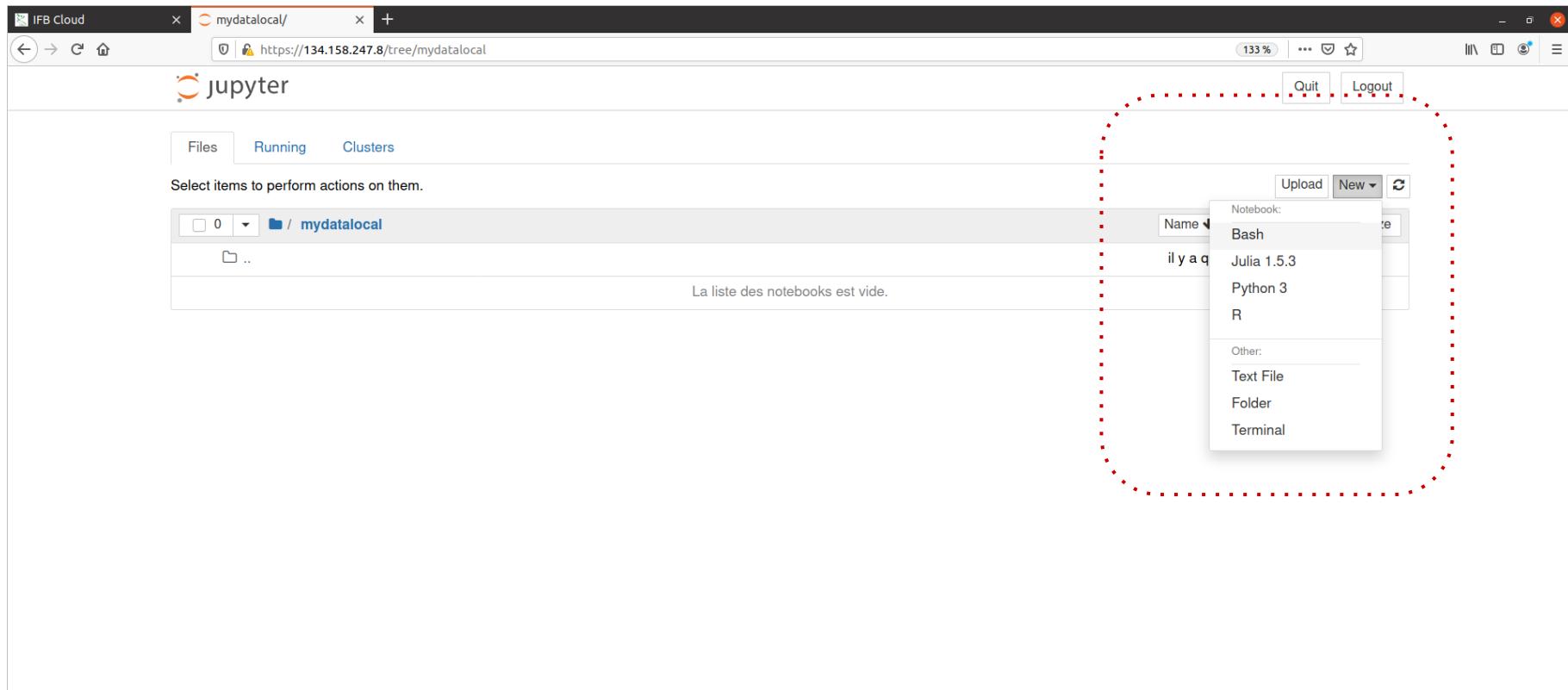
Open your vm (https link) to access to your own jupyter lab



Create your first jupyter book

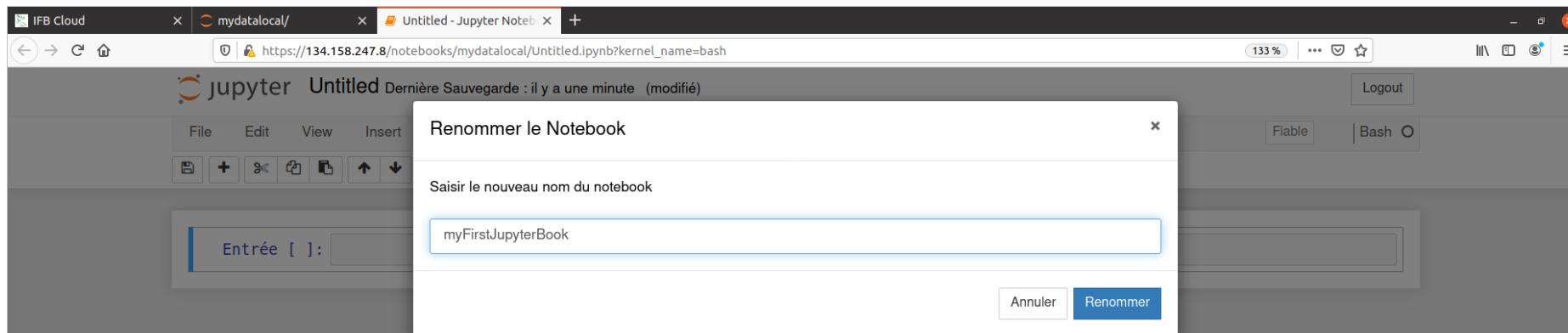
Go into the directory “work” and create a new jupyter book

-> kernel : bash



Rename your first jupyter book

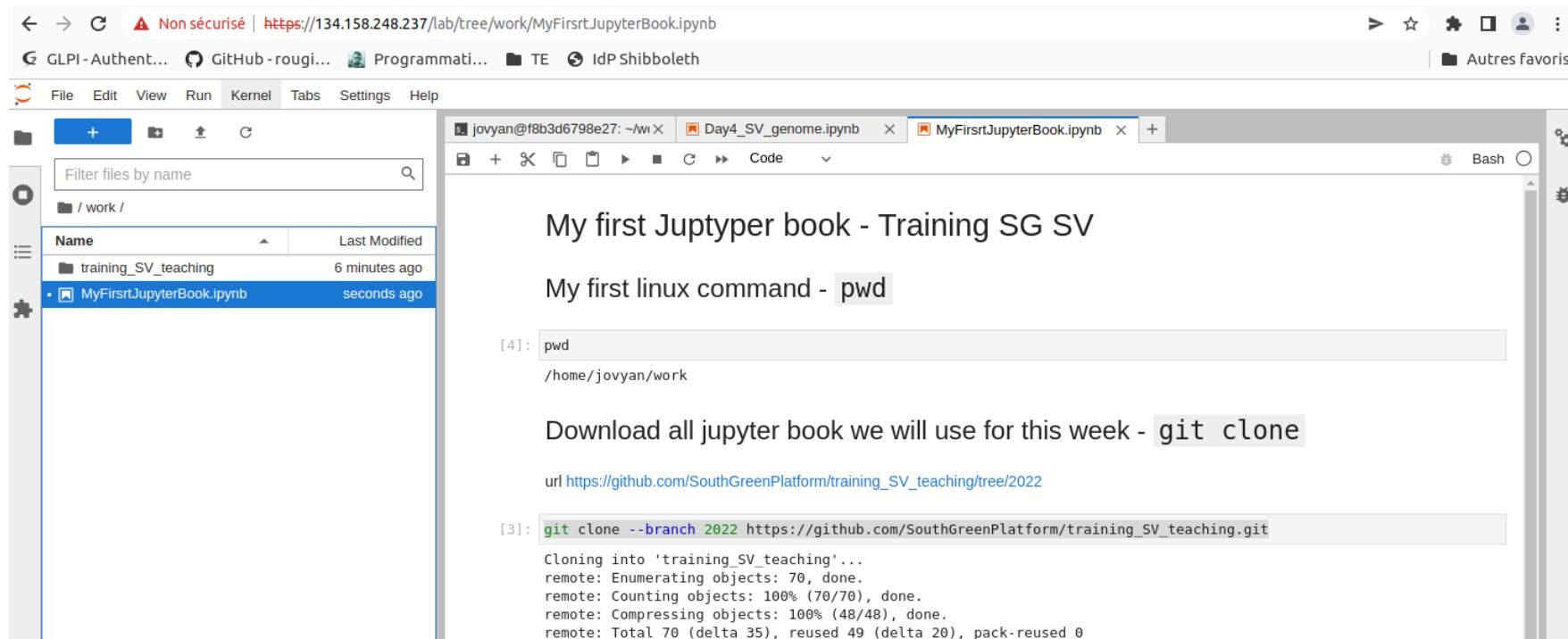
myFirstJupyterBook



Run your first bask command -

- 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*

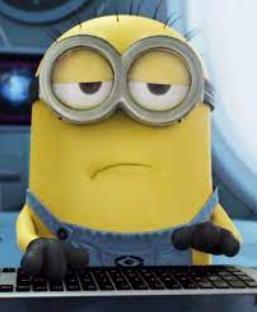
```
git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git
```



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

- File Browser:** On the left, there is a sidebar with a file browser. It shows a folder named "work" containing two items: "training_SV_teaching" (modified 6 minutes ago) and "MyFirstJupyterBook.ipynb" (modified seconds ago).
- Code Cell Output:** In the main area, there are several code cells and their outputs:
 - Cell 1: "My first Juptyer book - Training SG SV"
 - Cell 2: "My first linux command - pwd"
Output: [4]:
/home/jovyan/work
 - Cell 3: "Download all jupyter book we will use for this week - git clone"
Output: url https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022
 - Cell 4: "git clone --branch 2022 https://github.com/SouthGreenPlatform/training_SV_teaching.git"
Output:

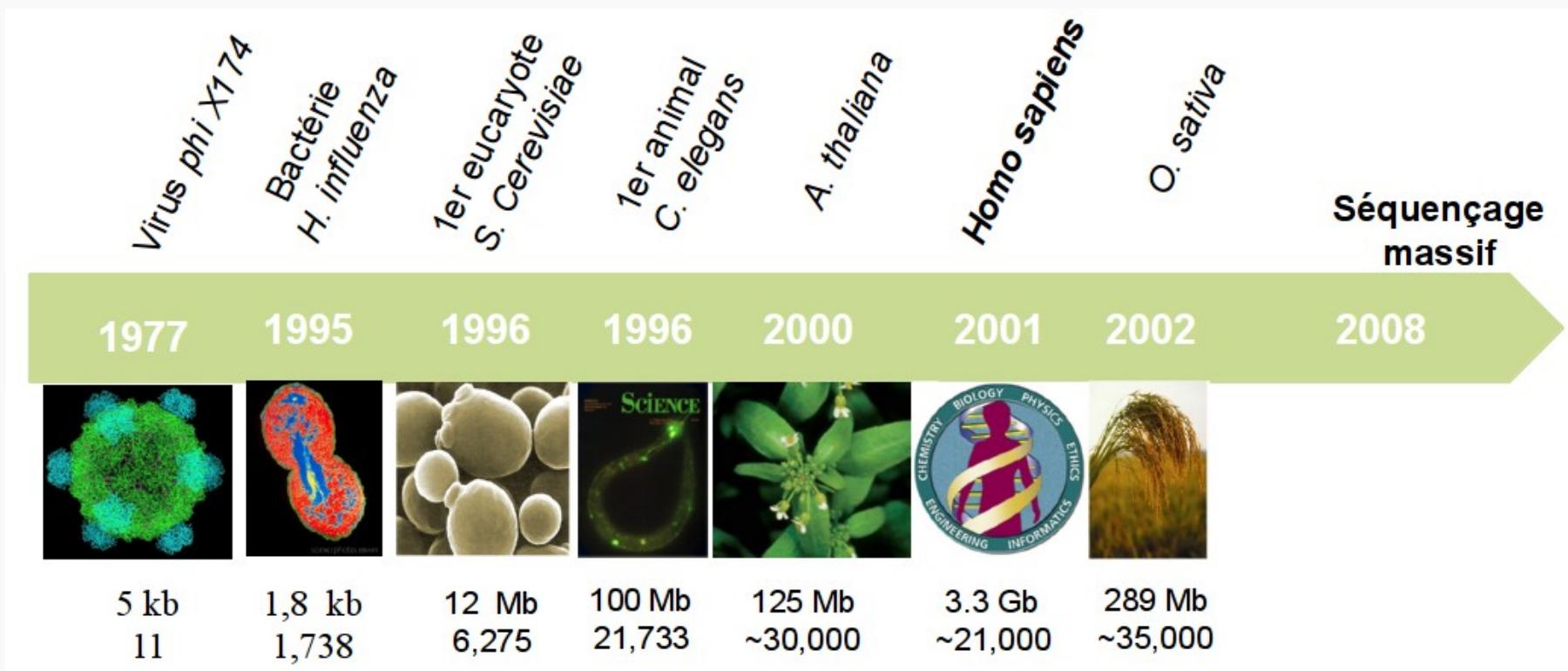
```
Cloning into 'training_SV_teaching'...
remote: Enumerating objects: 70, done.
remote: Counting objects: 100% (70/70), done.
remote: Compressing objects: 100% (48/48), done.
remote: Total 70 (delta 35), reused 49 (delta 20), pack-reused 0
```

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

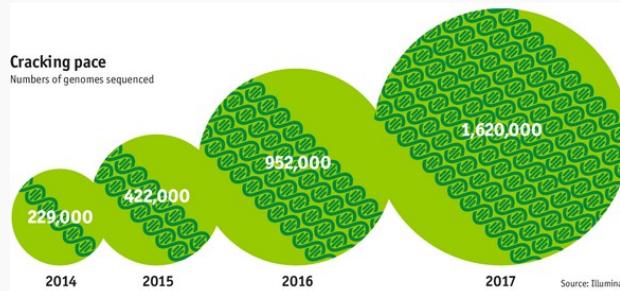
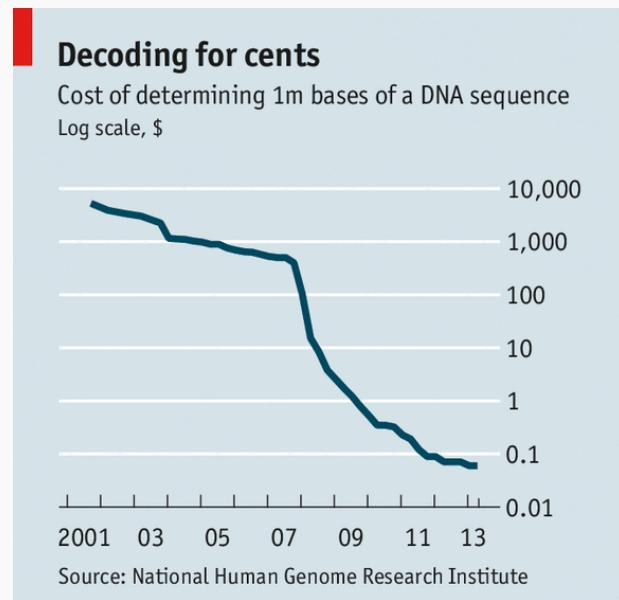
Introduction & NGS method

The NGS in themselves

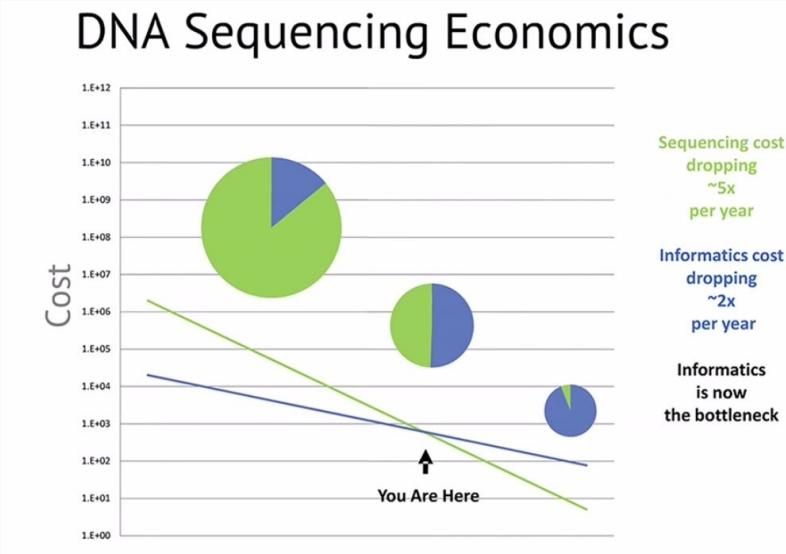
A little history of sequencing...



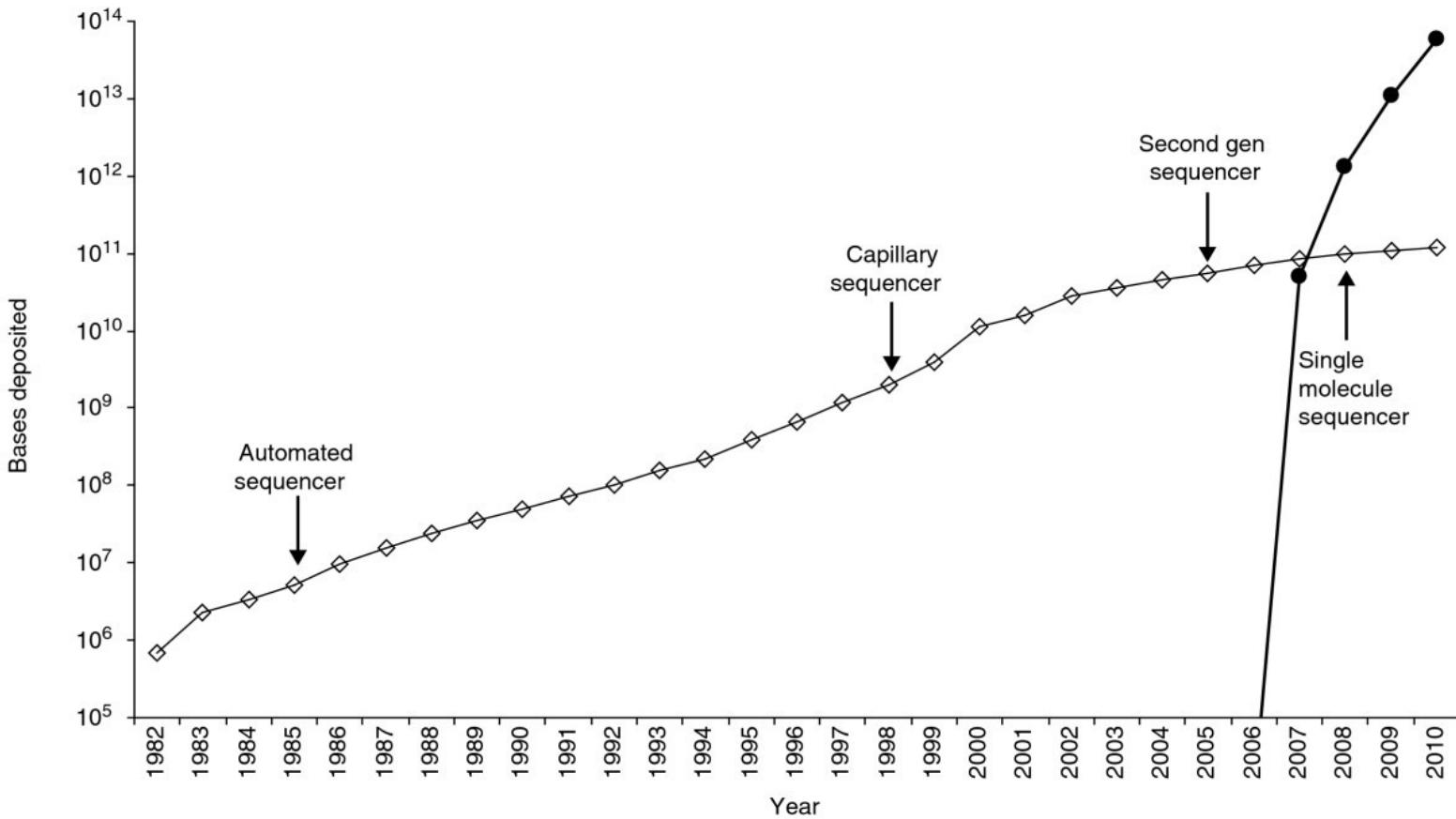
...From Data Rarity to Data Deluge



From The
economist



...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

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) 454
- Matrix amplification IonTorrent
- Short reads Illumina
- 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) 454
- Matrix amplification IonTorrent
- Short reads Illumina
- Limited error rate
- High throughput

PacificBiosciences
Oxford
Nanopore

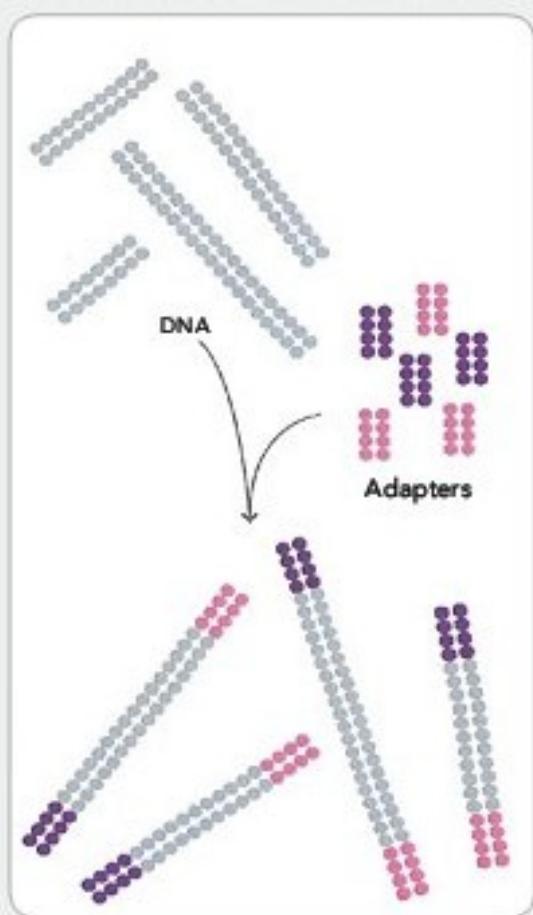
3rd Generation Sequencing

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



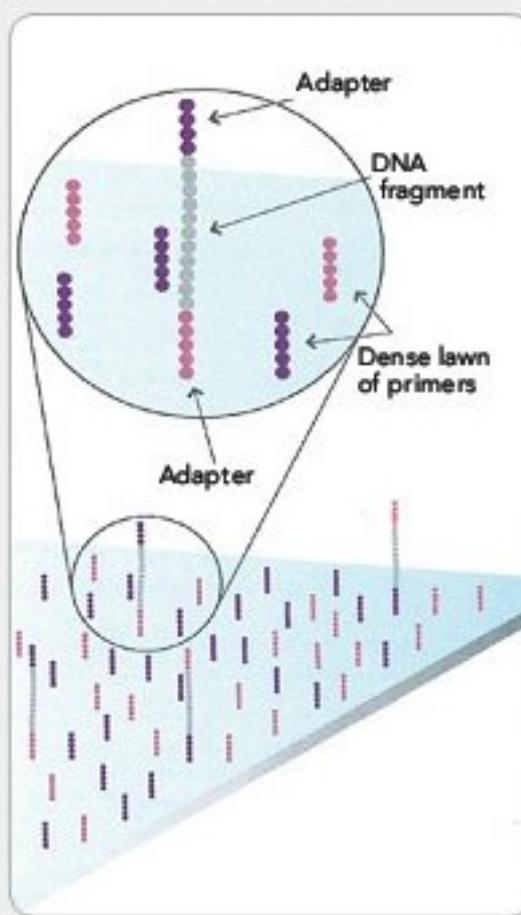


1. PREPARE GENOMIC DNA SAMPLE



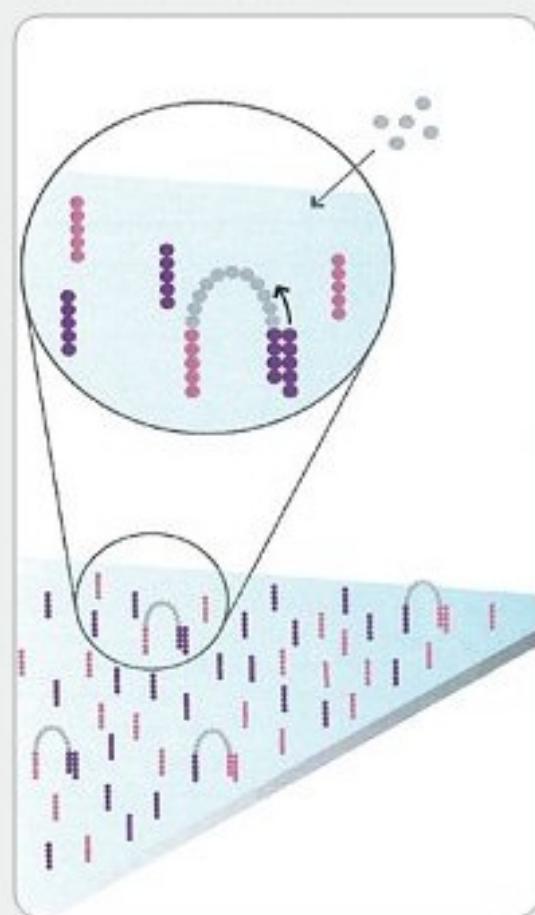
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

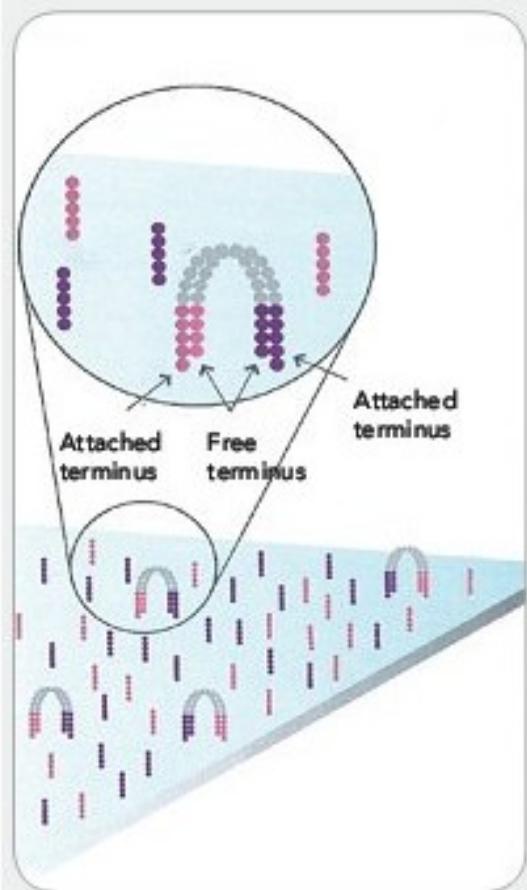
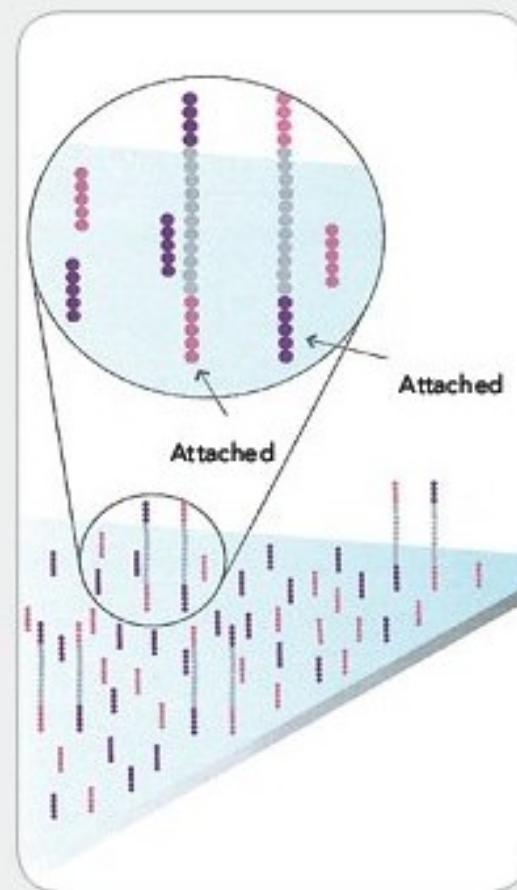
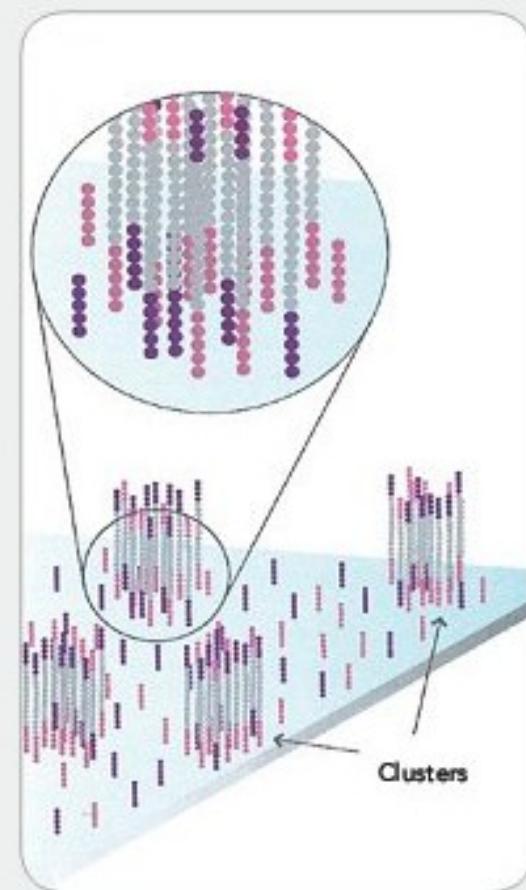


Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION



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

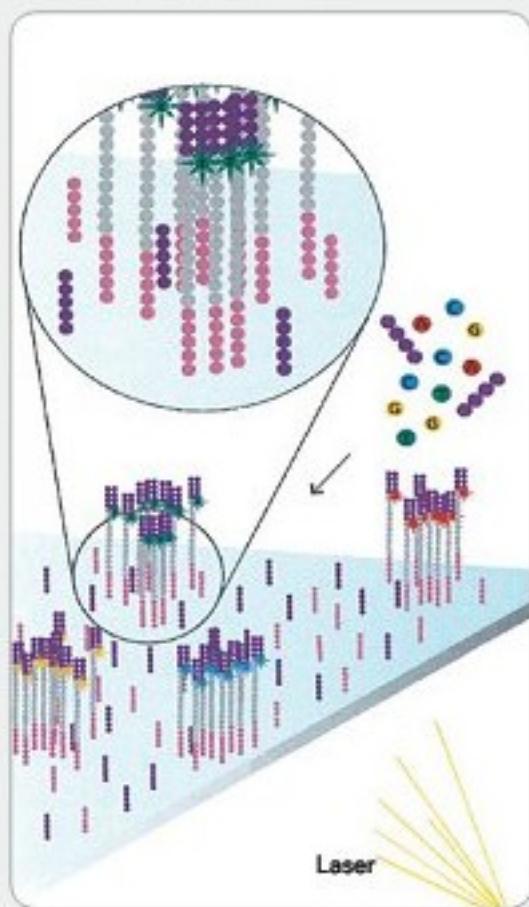
4. FRAGMENTS BECOME DOUBLE STRANDED**5. DENATURE THE DOUBLE-STRANDED MOLECULES****6. COMPLETE AMPLIFICATION**

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

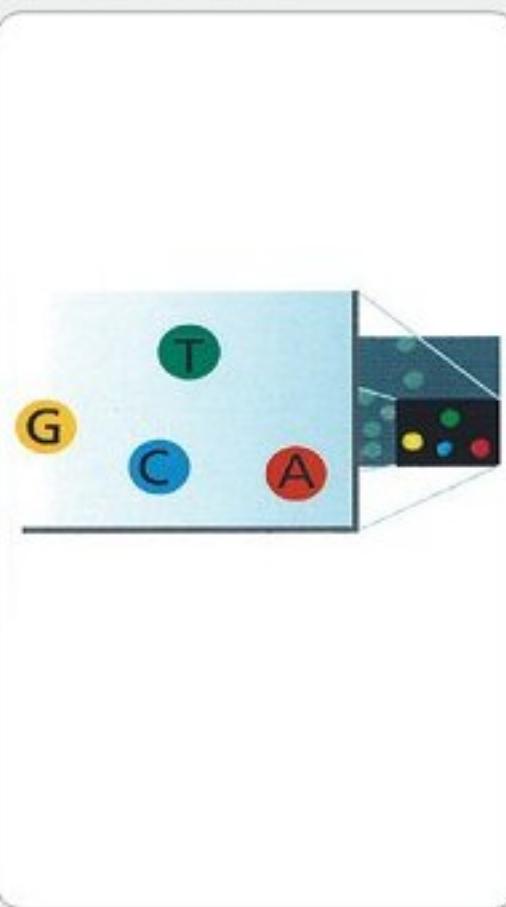
Denaturation leaves single-stranded templates anchored to the substrate.

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

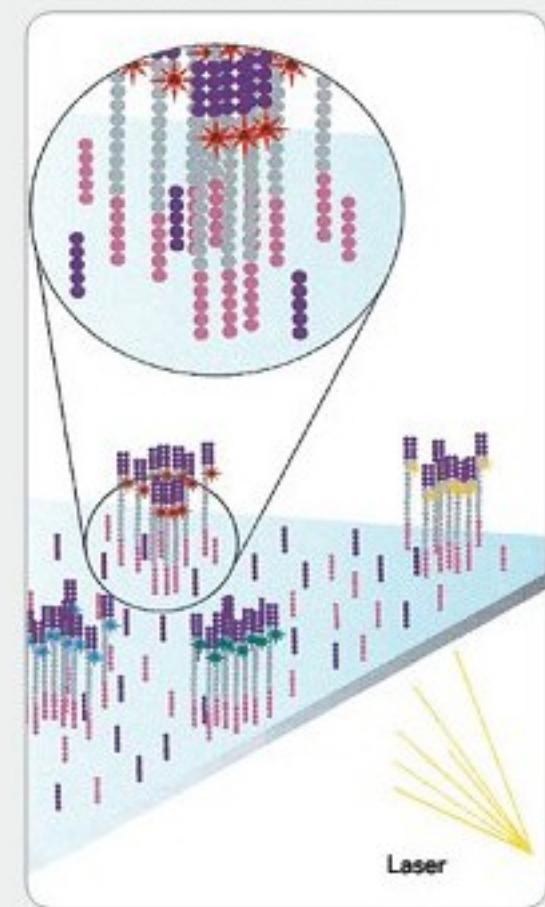
7. DETERMINE FIRST BASE



8. IMAGE FIRST BASE



9. DETERMINE SECOND BASE

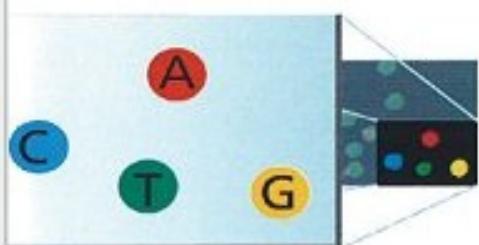


First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

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.

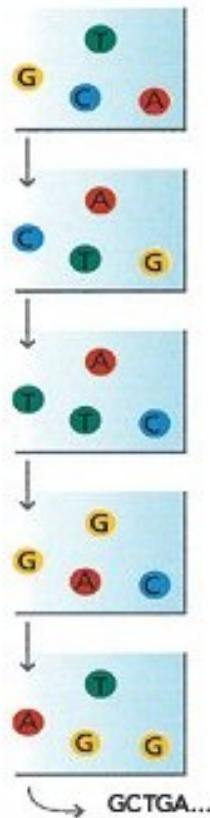
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



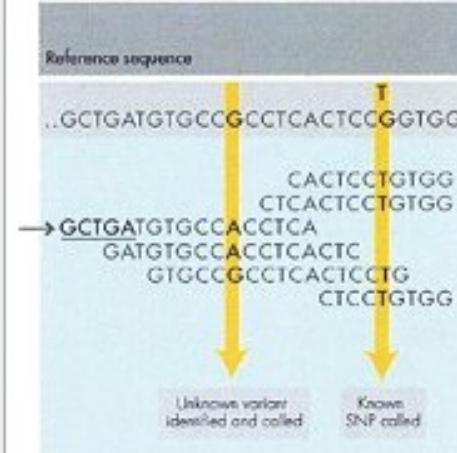
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

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
CTGTTCTTAGTATTTTGTTAGTCATTCCGTGGTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIIEFEGJGGHI4FFIEIGHI<FHGAHGGGB@3?BDB9D
@H4:C7C99ACXX:6:1101:1452:19906/2
CTGAGATCAATTGGATCCTGATGATACTGTGCTTAGCTATTACCTTGGT
+
@@@DDDD>FFFABEABB4C+3?:CBB@<<A?E4A???9C@CFF*9*B3D?B
@H4:C7C99ACXX:6:1101:1476:35220/2
CATGTGCTATTACCAAAAGTCAGTAACGACCTATAAATTTAAAGTAGC
+
@CFFFFFGGHHHHJJJJIEE<HHHIJIGBHGGEEIIJEIEIJHHJFIIJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTTGGGTATGGCTCTAGTAGCTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIIEHIII<CGHIJIIJ?:FC9DGAFGHII?DGBFIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@@DFFFFGHHHHJIIIIJJJIIGJCHHEIII>GHIG@GHIDHGJIIFHJJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGAAGAGTTAAAAAACATTAAATGAGCAACTGAGTTC
+
@@@CFFFFFHFFHFGIJJIHIIJJIIHJJECGHIIJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTGTCGGGGTTGGGAAATTGTCACTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```

1 séquence = 4 lignes

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



PHRED SCORE

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

*Ewing
1998*

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

or

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

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

The QPHRED Value

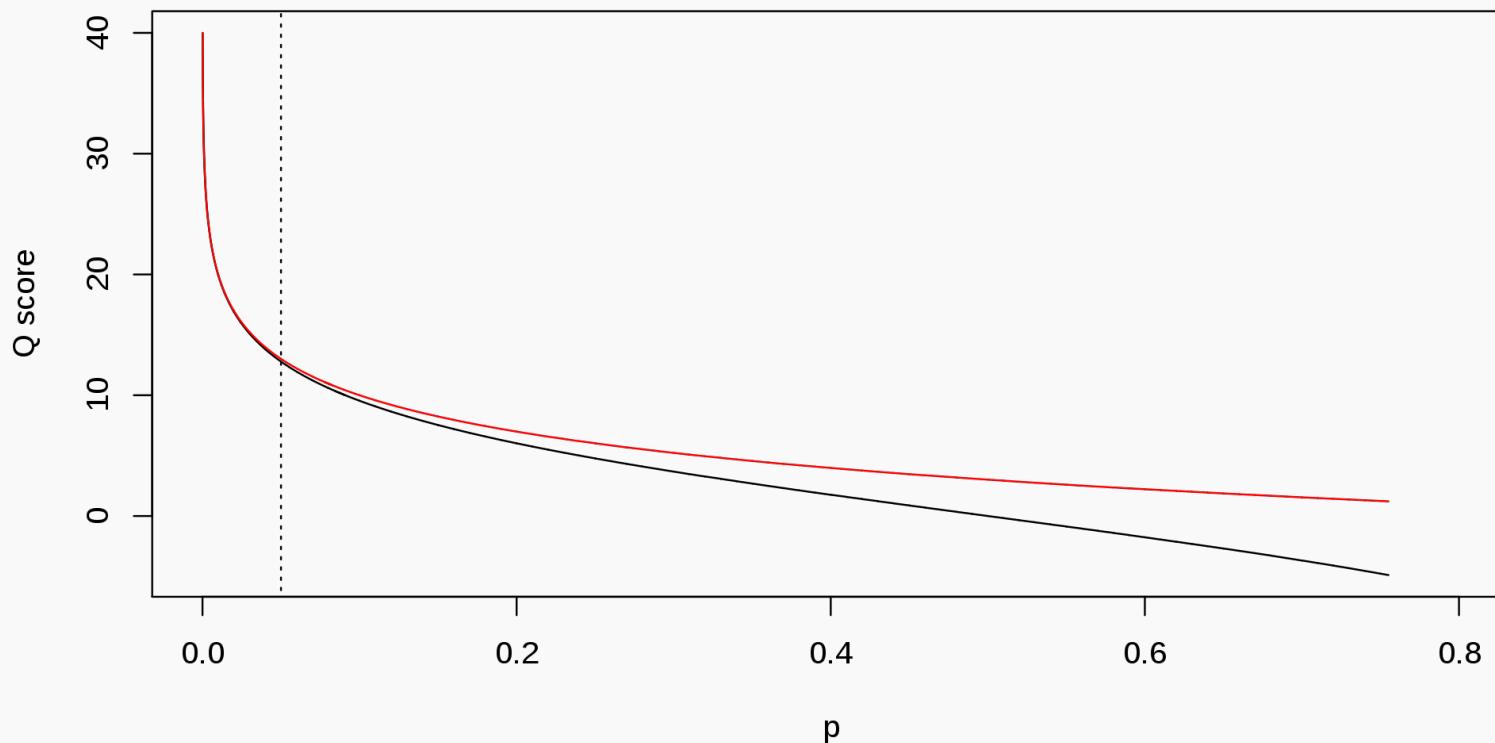
Phred Quality Score
0 .. 50

Code ASCII



Dec	Hex	Oct	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr
0 0	000	NULL		32 20	040	 	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	_	Del		127 ZF	177		Del	

The QPHRED Value



The QPHRED Scale



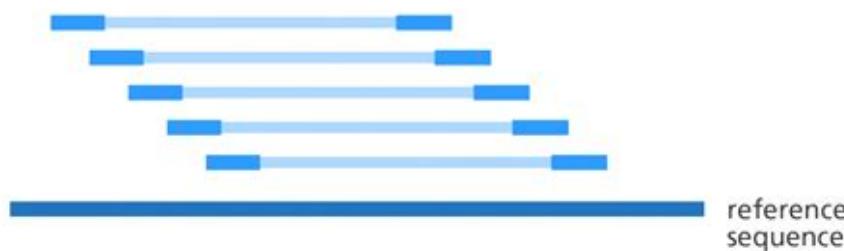
- S - Sanger Phred+33, raw reads typically (0, 40)
- X - Solexa Solexa+64, raw reads typically (-5, 40)
- I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
- J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See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

Mapping

Single-end reads



Paired-end reads



sequenced fragment unknown sequence sequenced fragment

A scale bar indicating the length of the sequencing fragments. It shows two blue segments on a horizontal line, with the text "200 - 1000bp" written below the line.

```

@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	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
-	LN*	Sequence length.
Sequence dictionary	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	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)
read group	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	ID*	Program name
-	VN	Program version
Program	CL	Command line
CO - comment		One-line text comments

SAM Format

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

alignement

Format tabulé

SAM format :

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

```

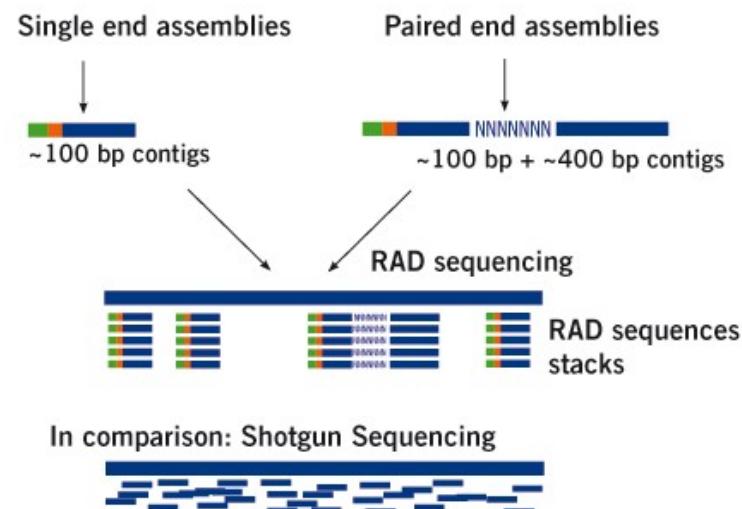
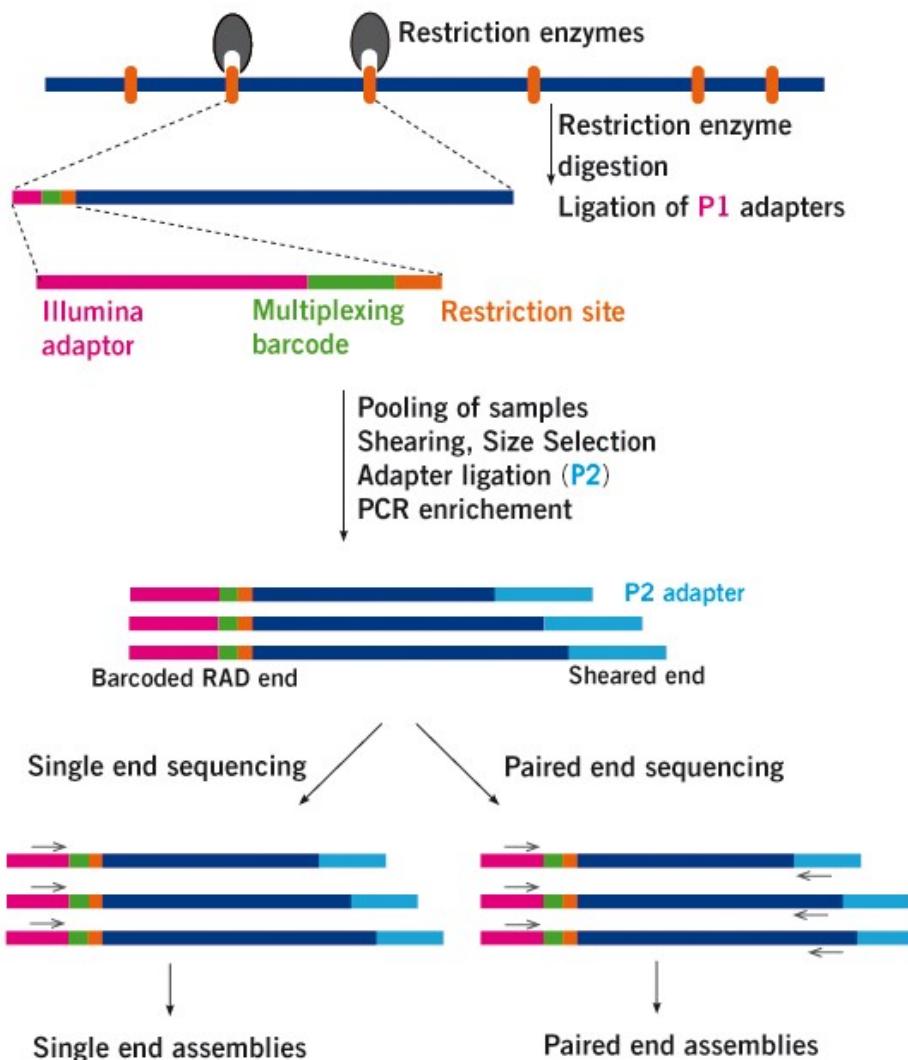
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC
r003 16 ref 29 30 6H5M * 0 0 TAGGC * N
r001 83 ref 37 30 9M = 7 -39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

```

alignement

Format tabulé

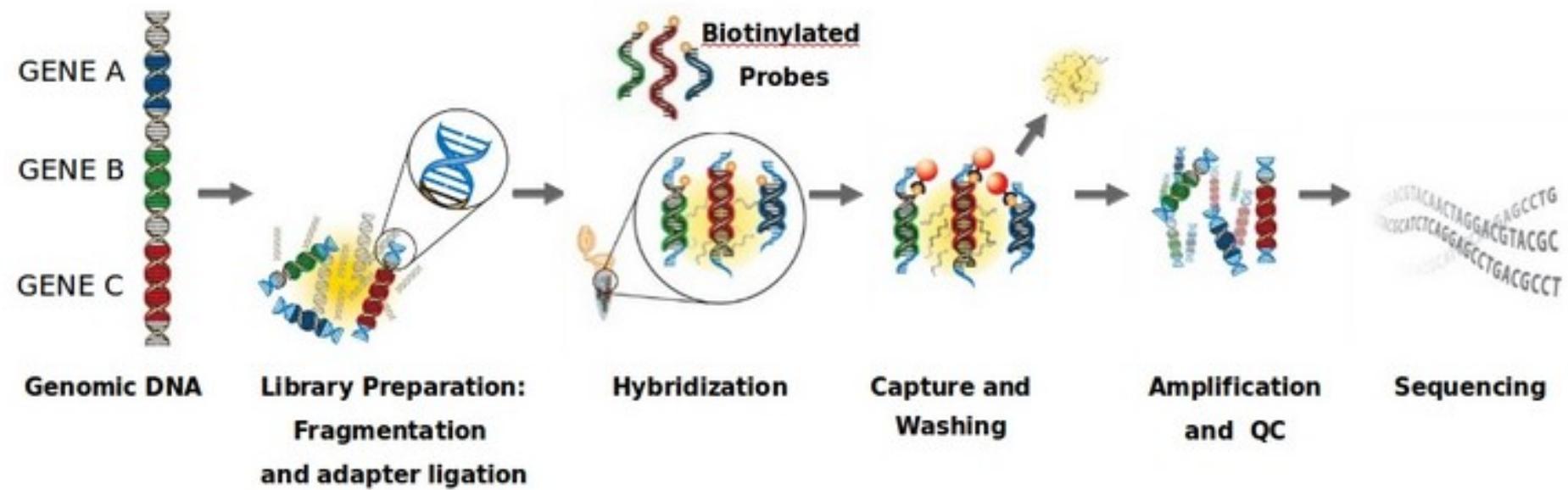
Col	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string (operations: M I D N S H P)
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



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

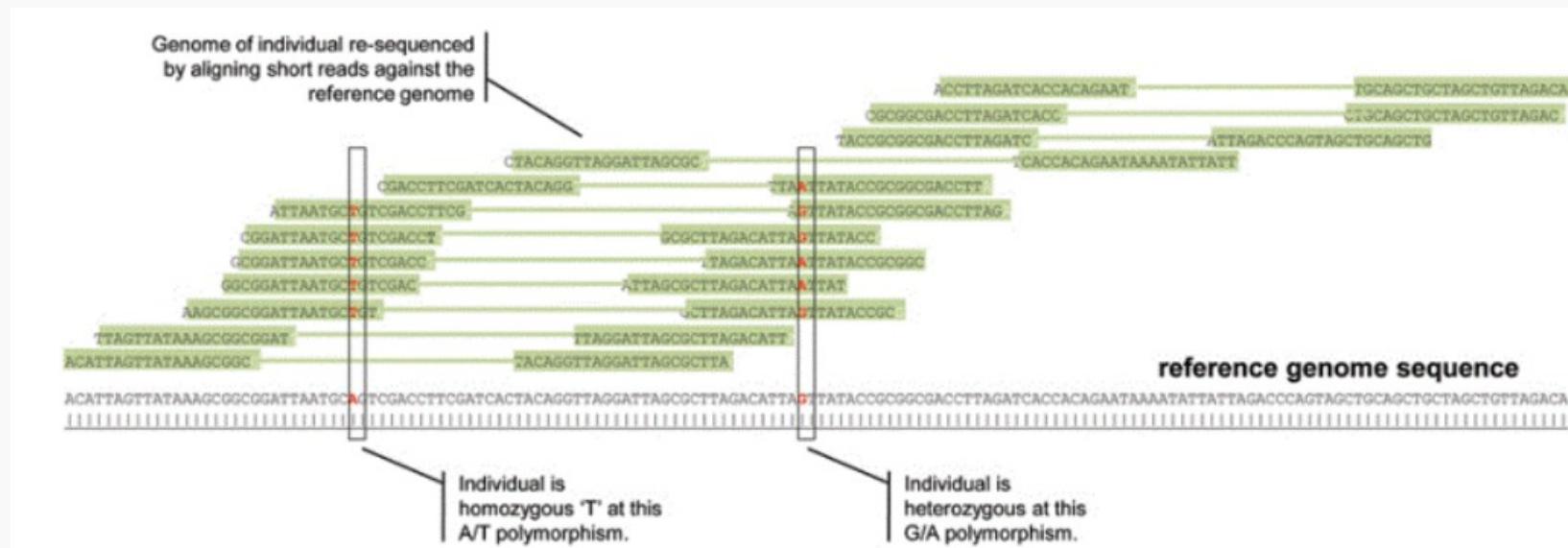
Sequencing the fragments from either one side or both sides results in RAD sequences stacks.

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



From CGFB, Bordeaux,
France

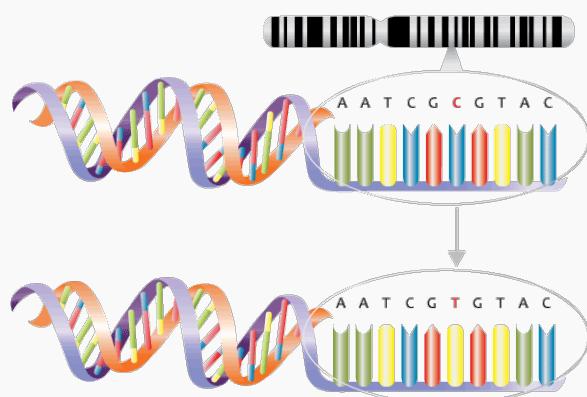
SNP and InDel Detection



SNP and InDel Detection

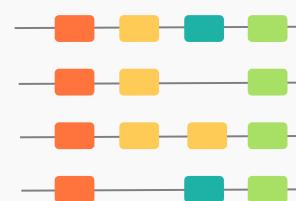
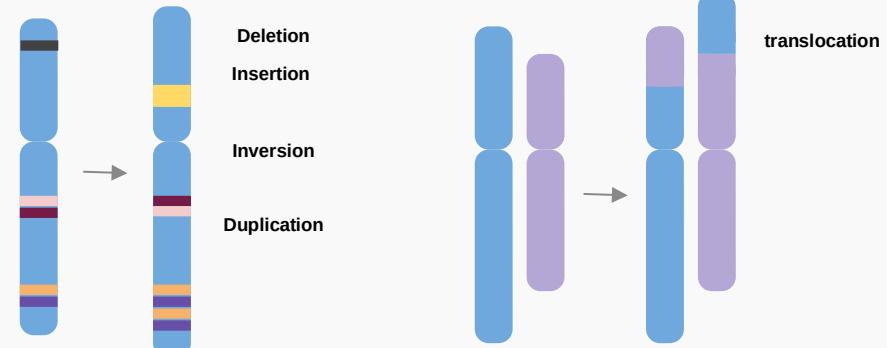
Mutations & Variations as main source of genetic diversity

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									Mandatory header lines
##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
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77
1	100		T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3
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)									

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

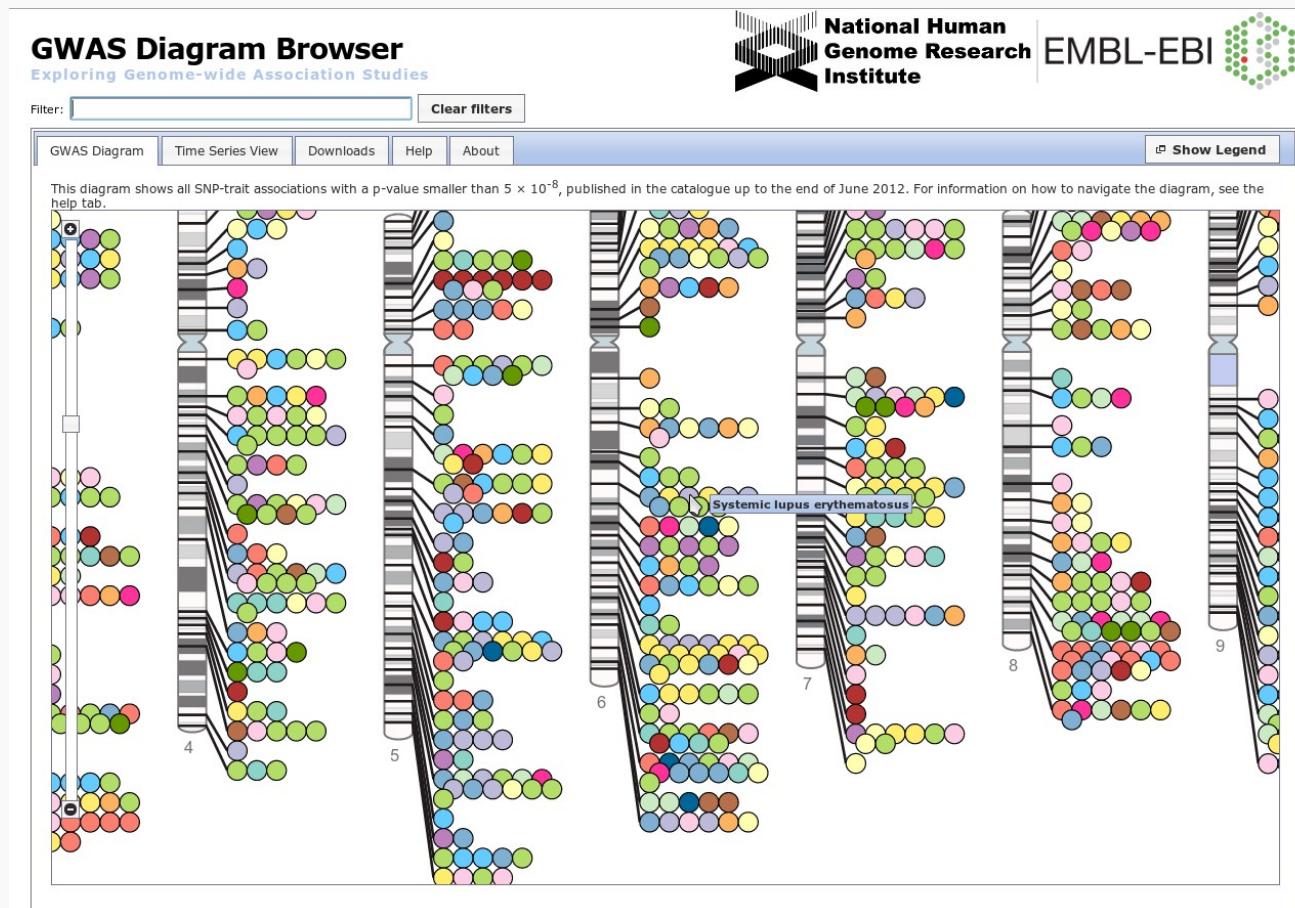
- Cleaning data level
- Mapping Conditions
- Mapping Cleaning Conditions
- Variation Calling level

All linked to the Specificity/Sensitivity Informatics Paradox

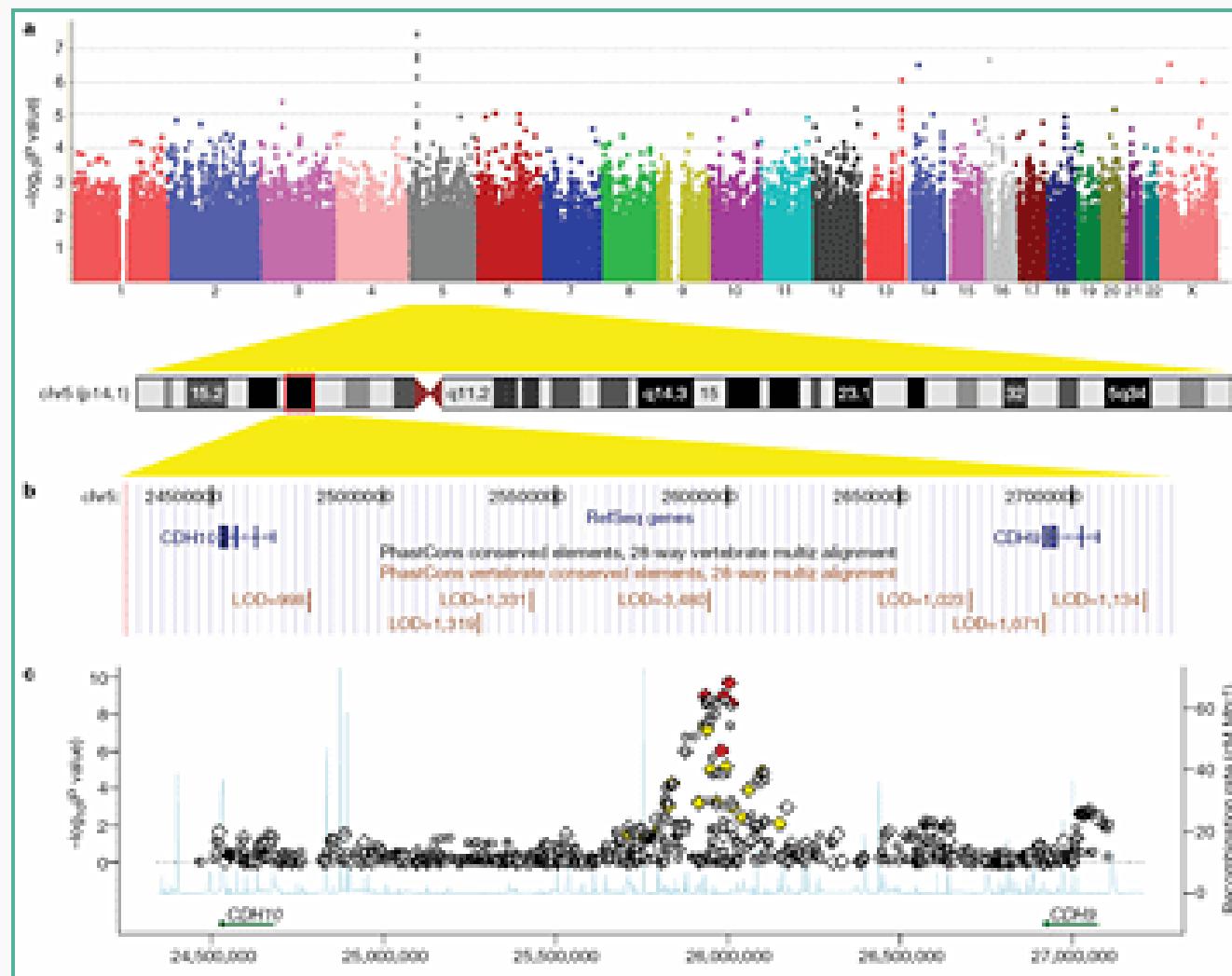
Applications

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

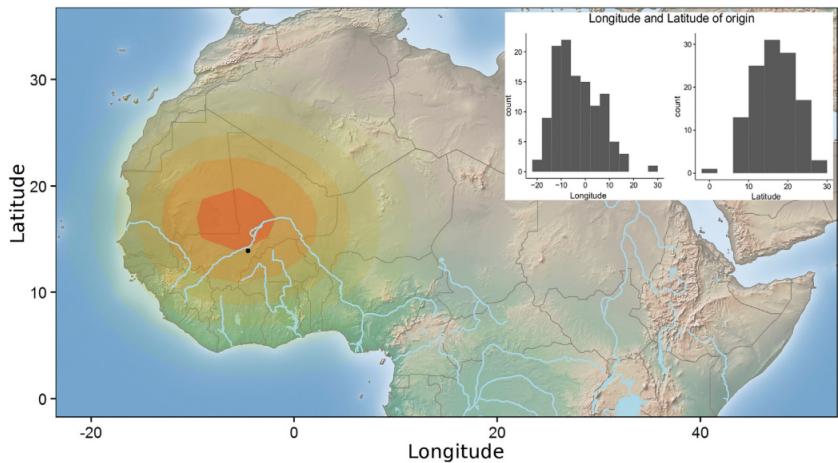
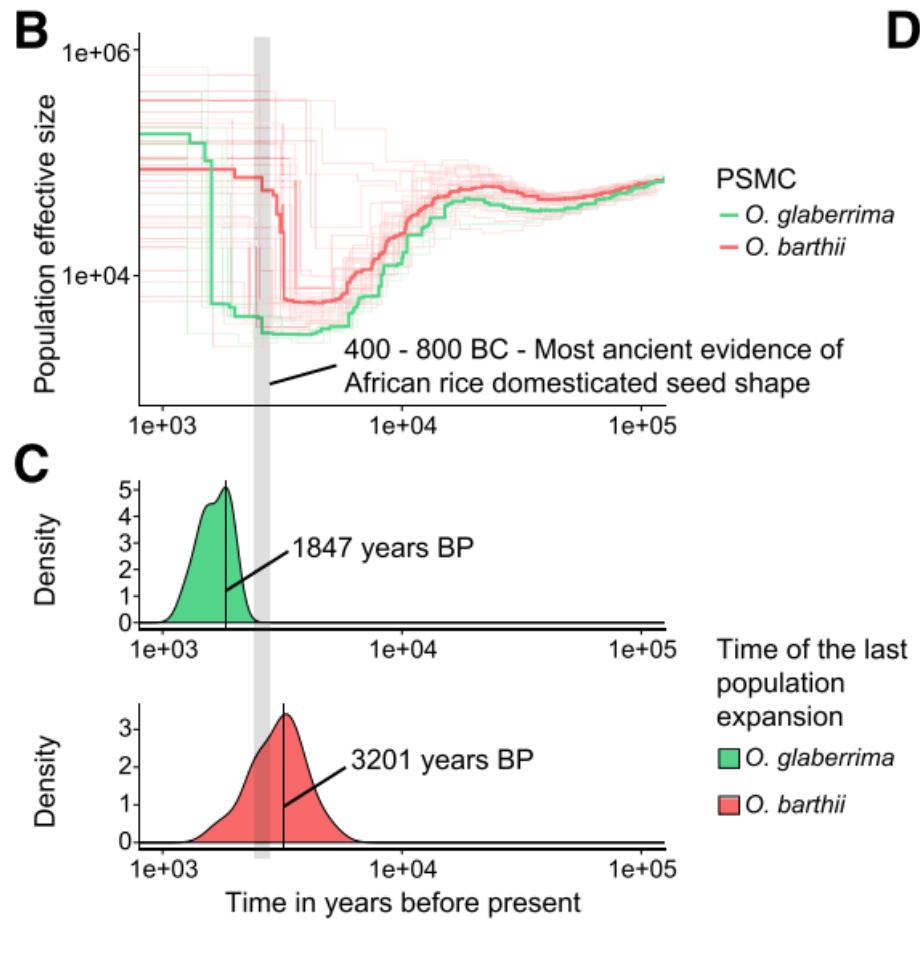
Example in GWAs & Population Genomics



Example in GWAs & Population Genomics



Example in Global Genotyping & Population Genomics



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

From Cubry et al,

Large Projects

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

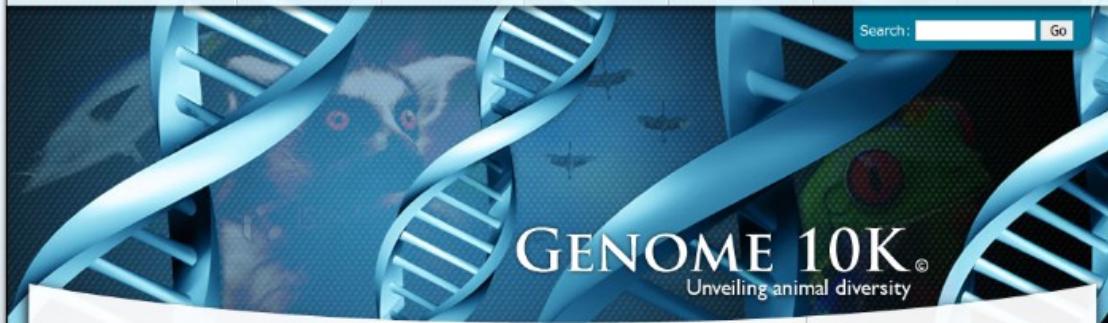
Home Collaborators Accessions Tools Software Data Center Gallery About Help desk

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

Sample types

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

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



- Real-time Transcriptomics

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



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

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

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

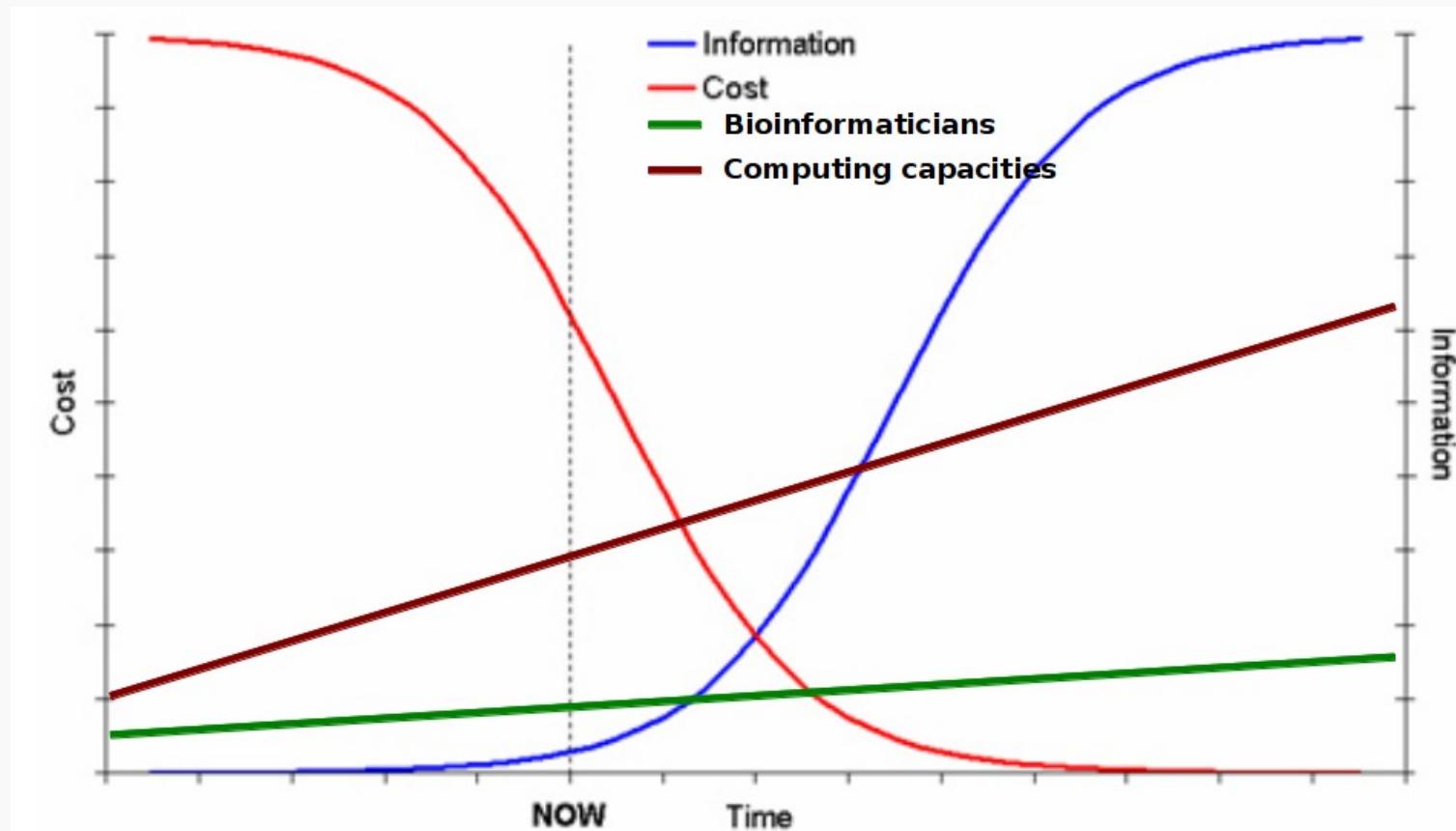
Possibilities in the next 5-10 years (From a presentation in 2013)

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

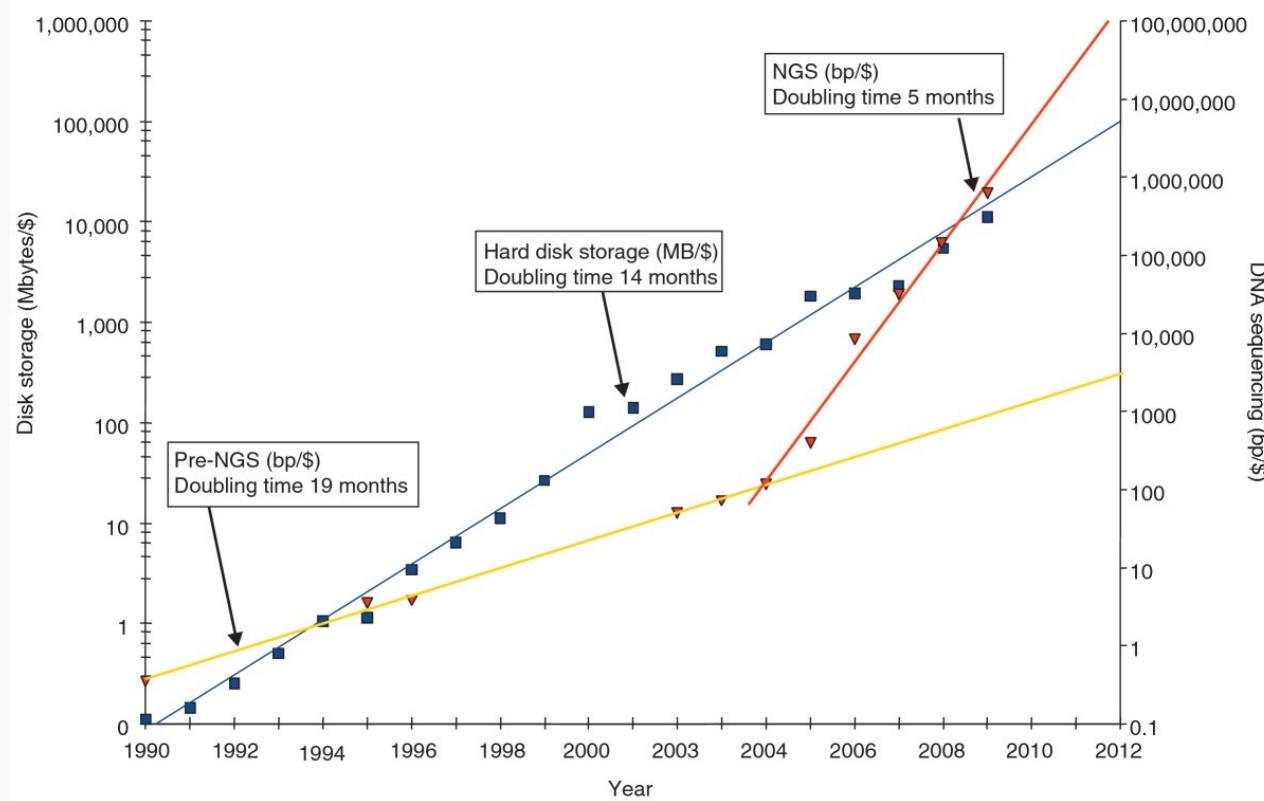
Possibilities in the next 5-10 years (From a presentation in 2013)

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

Keep in mind!

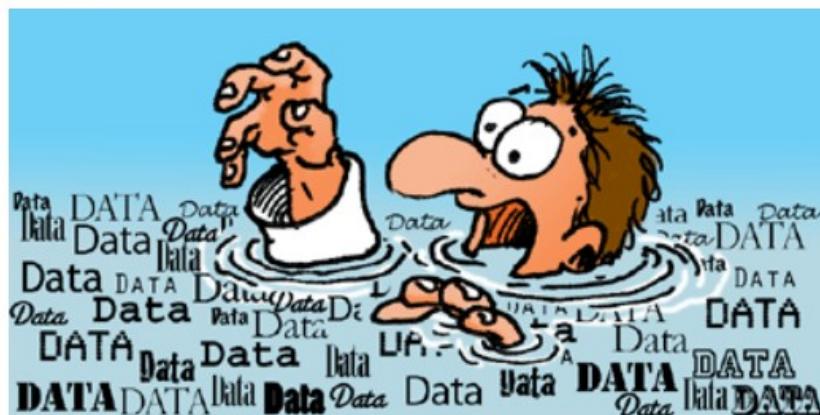


...From Data Rarity to Data Deluge



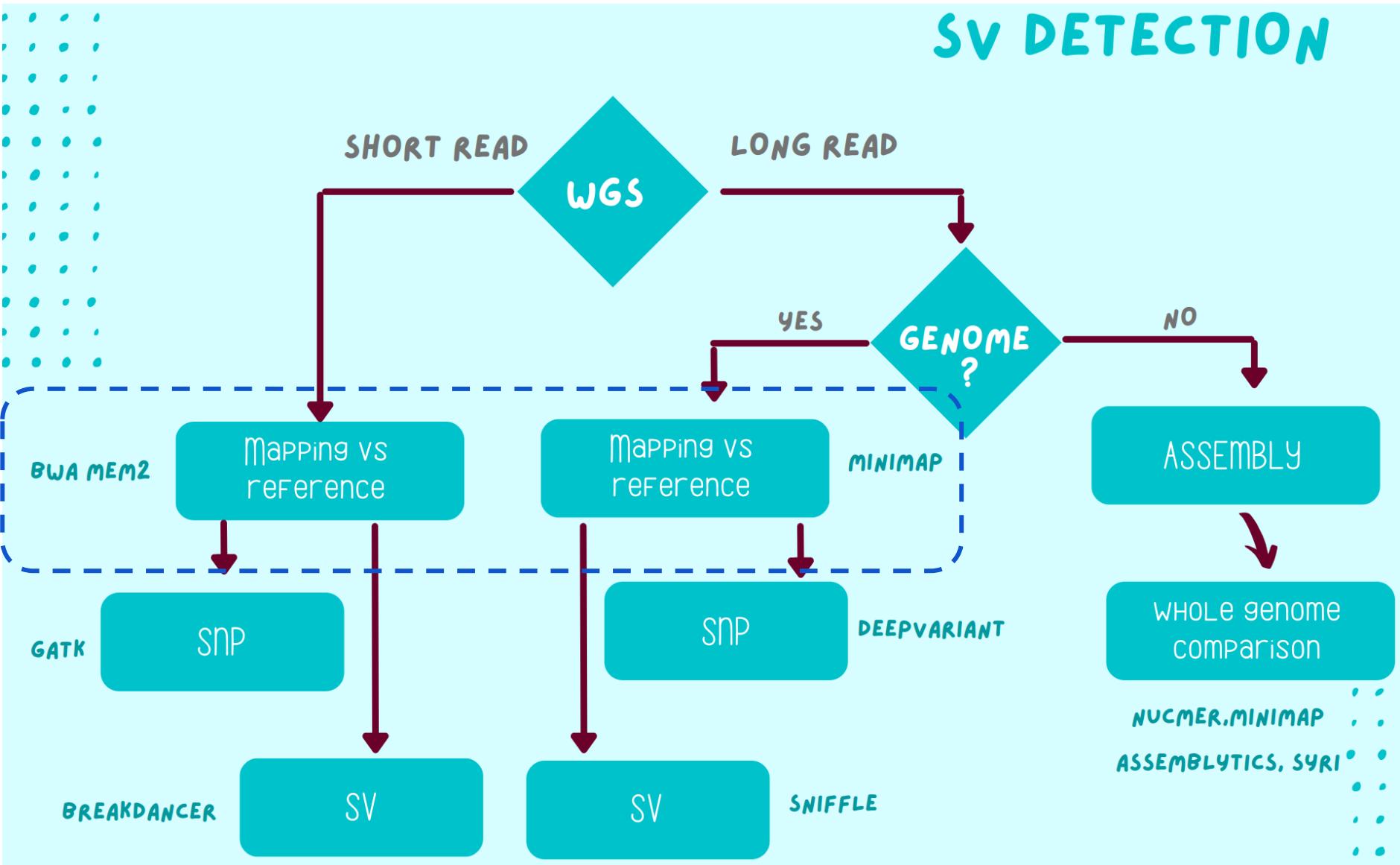
From L. Stein,
2010

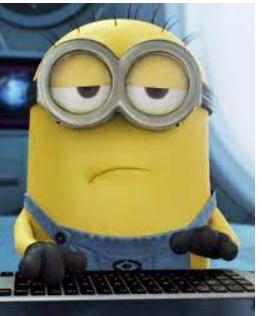
Be Careful to data drowning!



Training plan - day 1

Sv DETECTION



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

Mapping and SNP

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

The data

Diploid Asian Rice, *Oryza*



From
Wikimedia

Diploid Asian Rice, *Oryza*



From
Wikimedia

1. Select/Cut 1 Mb on Chromosome 10

The data

Diploid Asian Rice, *Oryza*



From
Wikimedia

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

Diploid Asian Rice, *Oryza*



From
Wikimedia

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

Diploid Asian Rice, *Oryza*



From
Wikimedia

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

Diploid Asian Rice, *Oryza*



From
Wikimedia

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

The FASTQ Format

The diagram illustrates the structure of a FASTQ file. It shows four lines of FASTQ data with blue arrows pointing from the right to specific parts of each line, labeled with their respective components:

- Sequencing info:** Points to the first line of each group, which contains sequencing parameters like the flowcell ID (@HWI-EAS236_3), lane number (FC_20), and sequencing run details (BTNAAXX:2:1).
- Nucleotide sequence:** Points to the second line of each group, which displays the actual DNA sequence (e.g., GAGAAGTTCAACAGCTGGTATTATTTGTAAACAT).
- Quality score in ASCII:** Points to the third line of each group, which shows quality scores in ASCII format (e.g., hhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE).
- Line separator:** Points to the fourth line of each group, which is a plus sign (+) used to separate the sequence from the quality scores.

```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTGTAAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGGAGTGGAAAGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
TGGTTTATGCAGAAATTCTAGAATAAGGGTAACCTT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
TCTCAGAAAATTGTTGTGATGTGTATTCAAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
TTGATTTAACTCTGACAAAATAAACAAAGTCTTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

The QPHRED Scale

Score qualité phred	Proba. d'une identification incorrecte	Précision de l'identification
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	99.9 %
40	1 pour 10000	
50	1 pour 100000	



S - Sanger Phred+33, raw reads typically (0, 40)

X - Solexa Solexa+64, raw reads typically (-5, 40)

I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)

J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
 with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
 (Note: See discussion above).

L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

“Classic” launch

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

“Classic” launch

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

“Classic” launch

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

OPTIONAL!

“Classic” launch

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

“Classic” launch

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

Between 8 and 15 different commands...

Let's do it by hands...

We will

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

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: M ID NSHP)
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

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

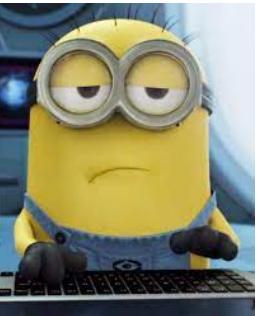
1. Map the data of Clone 1 on *reference.fasta* using *bwa*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it
4. Remove wrong mapping



Practice

Let's work with the jupyter book :

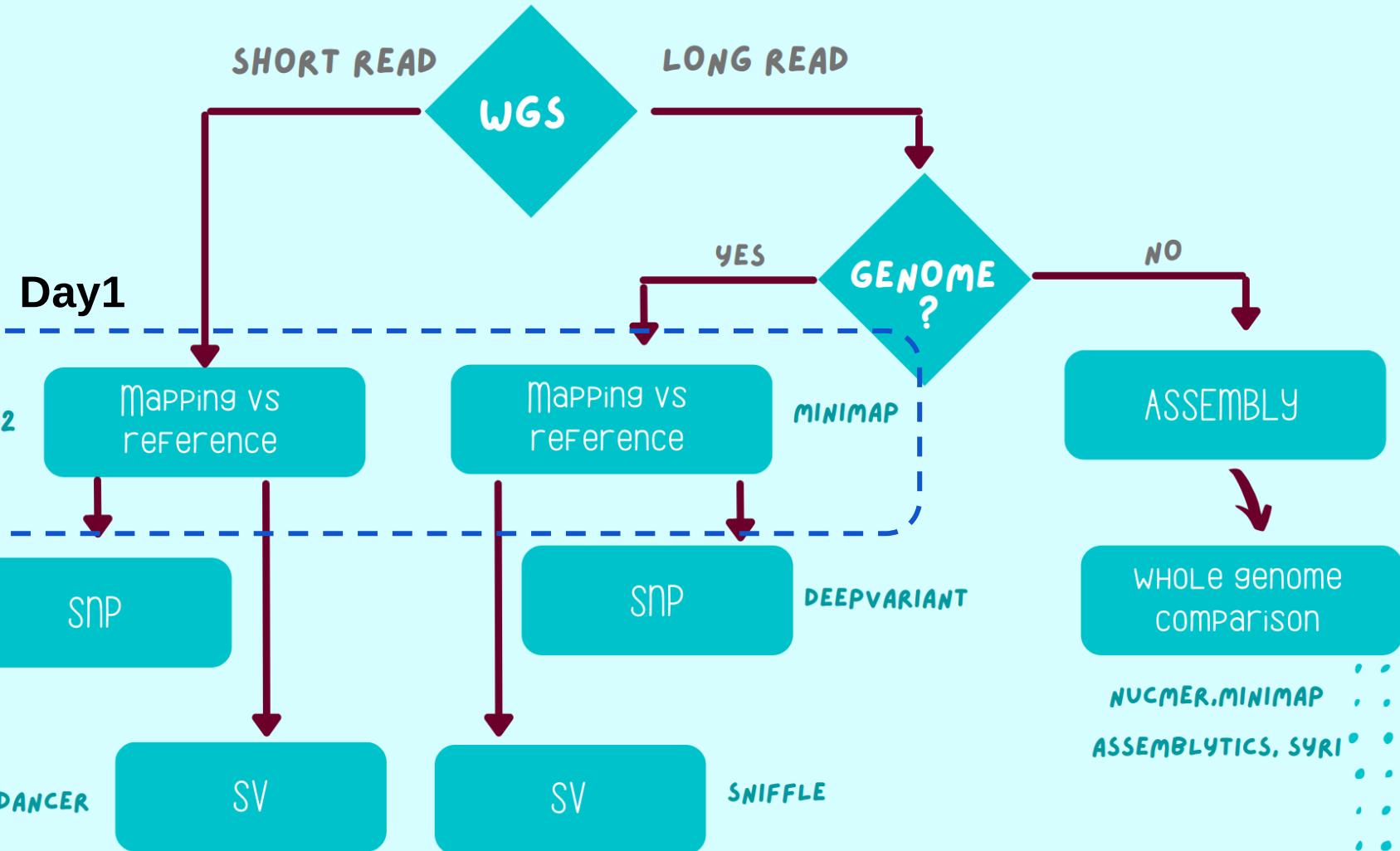
Day1_Mapping_Practice_EMPTY.ipynb

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

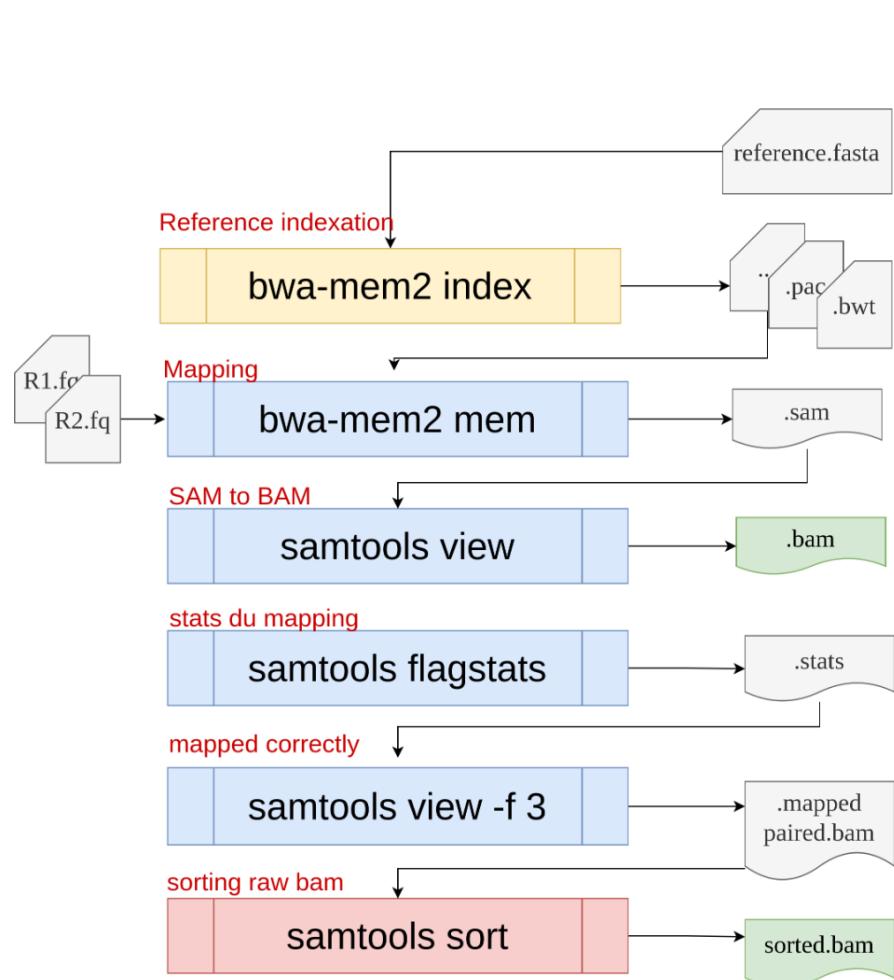
Day2 - Let's start slowly

Training plan

Sv DETECTION

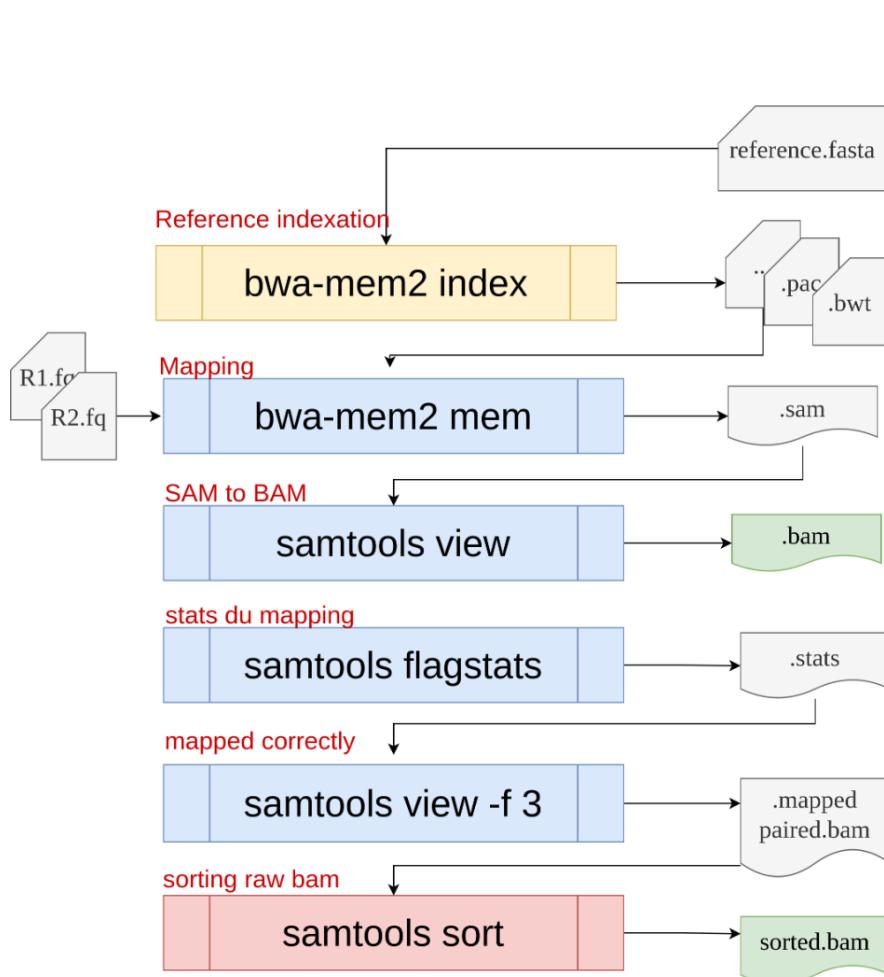


Plan de bataille du mapping !

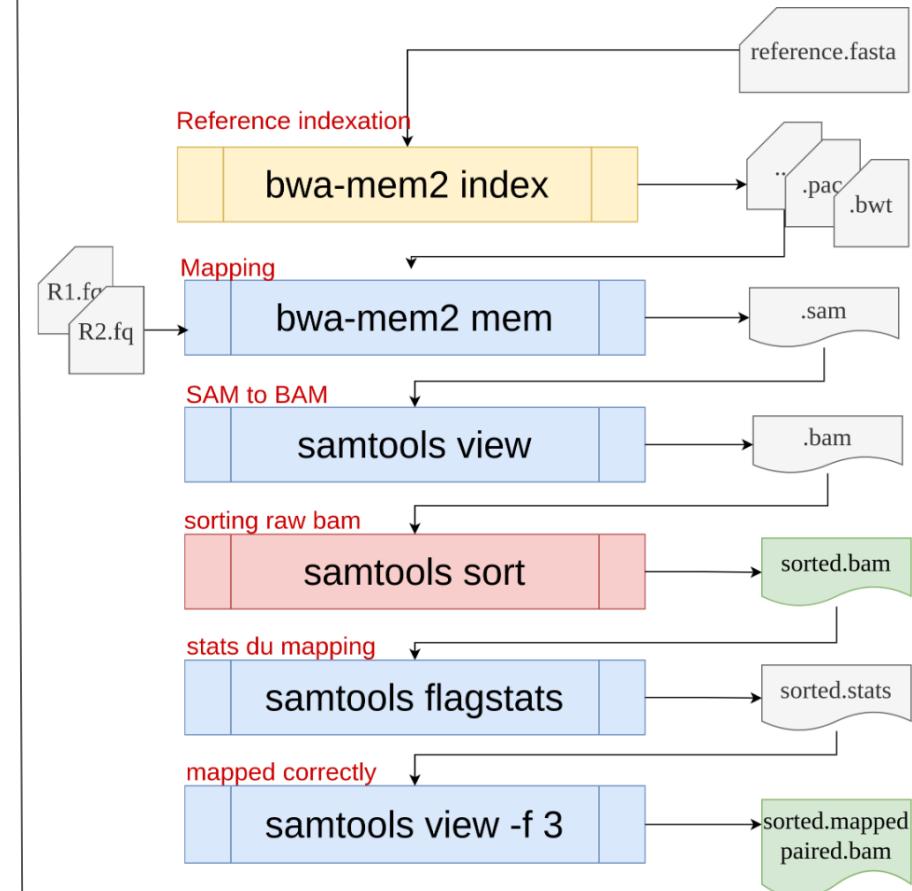


En cours Day1

Plan de bataille du mapping !



En cours Day1



ça marche mieux !

Plan de bataille du mapping -

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

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

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

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

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

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

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

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

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

done
```

Plan de bataille du mapping -

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

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

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

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

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

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

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

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

done
```

Observing mapping

- Download Tablet (use Google and Tablet+NGS)

Observing mapping

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

Observing mapping

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

Observing mapping

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

Observing mapping

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



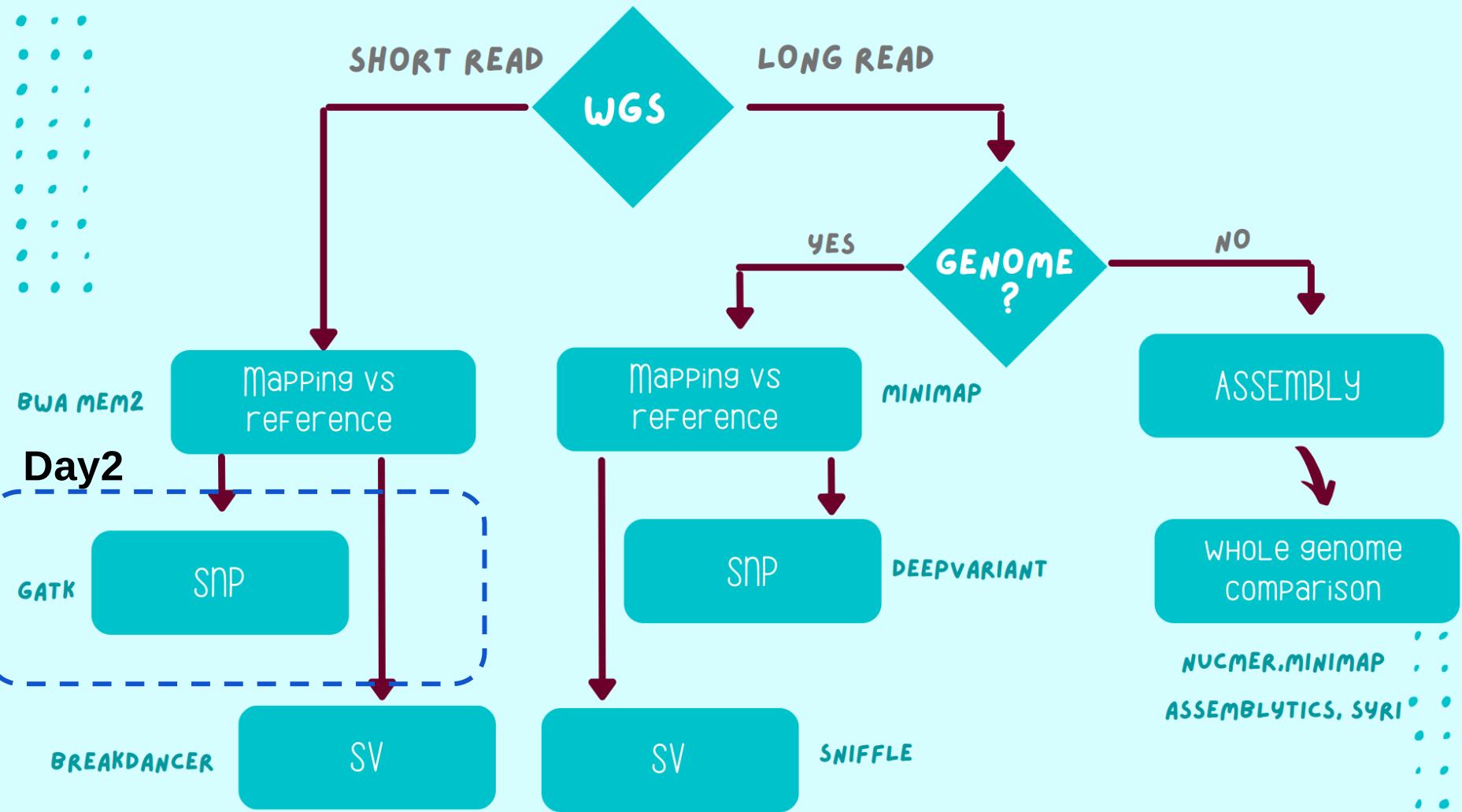
Practice

Let's work with the jupyter book :

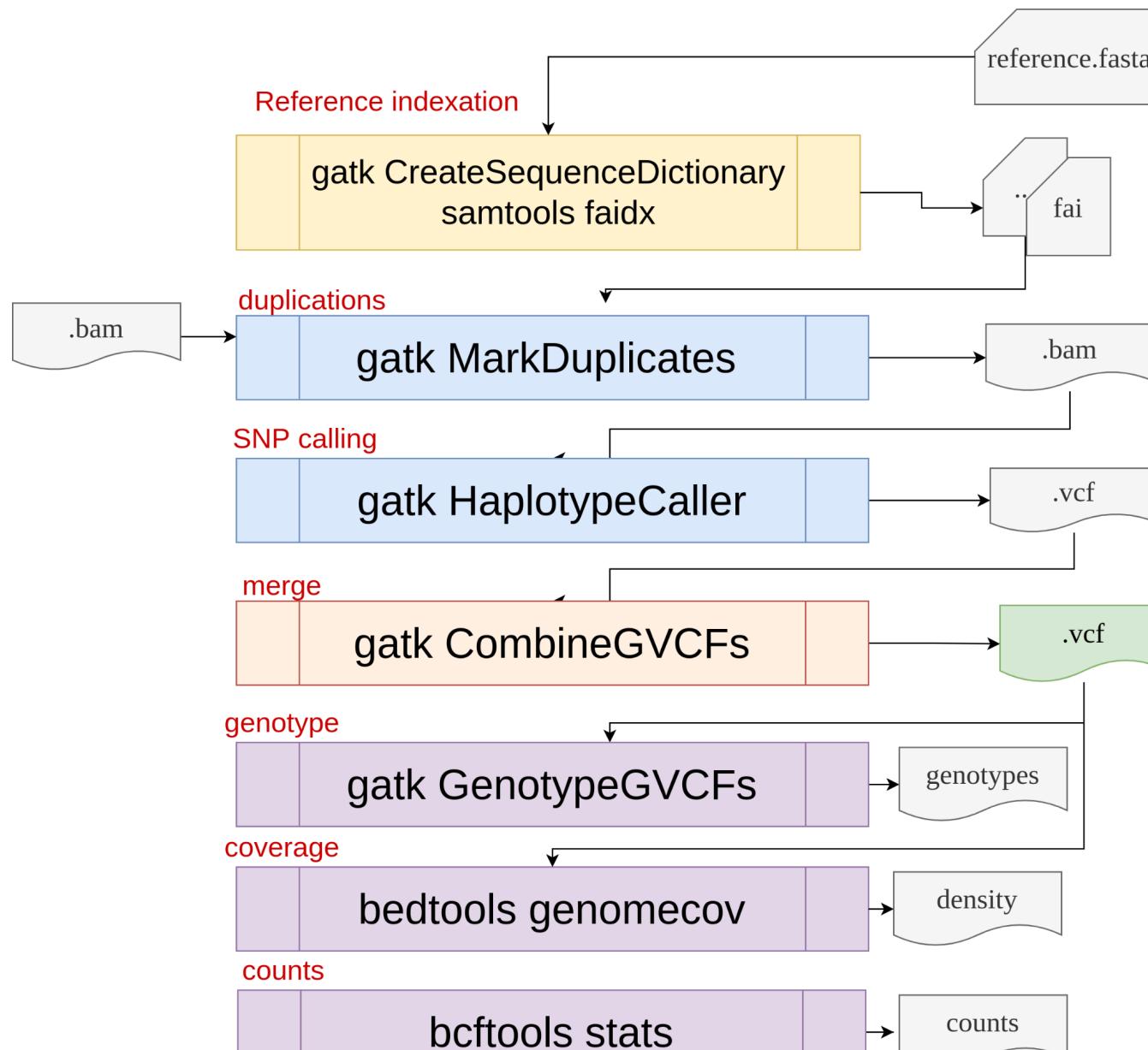
[Day2a_Mapping_analysis-EMPTY.ipynb](#)

Training plan

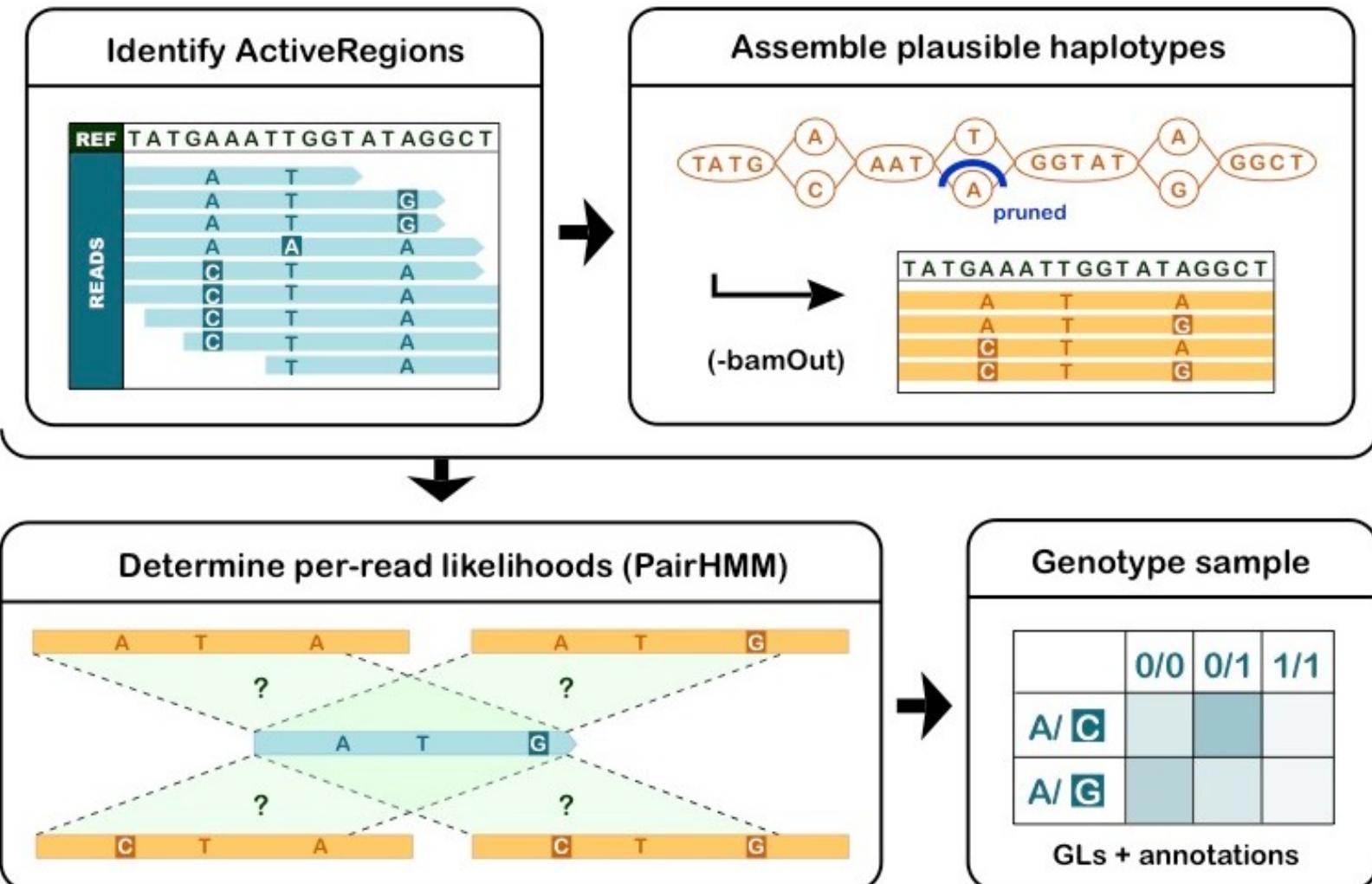
Sv DETECTION



Plan de bataille du “SNP”



The HaplotypeCaller



From UniBe.ch
Training

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

<pre> ##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"> #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2 </pre>									
VCF header					Mandatory header lines				
					Optional header lines (meta-data about the annotations in the VCF body)				
Body	1	1	.	ACG	A,AT	.	PASS	GT:DP	1 2:13
	1	2	rs1	C	T,CT	.	PASS	GT:GQ	0 1:100
	1	5	.	A	G	.	PASS	GT:GQ	1 0:77
	1	100		T		.	SVTYPE=DEL;END=300	GT:GQ:DP	1 1:12:3
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)									
Deletion SNP Large SV Other event Insertion									



Practice

Let's work with the jupyter book :

Day2b_Appel_variants_SNP_EMPTY.ipynb

Plan de bataille du “SNP

Basics filtering

- Using GATK Variant Filtration, a flag per filter

Using GATK Variant Filtration, a flag per filter

- Depth filter: $DP < ? \text{ or } DP > ?$
- QUAL filter: $QUAL < ?$
- SNPcluster filter: $\text{more than 3 SNP per 10b}$



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

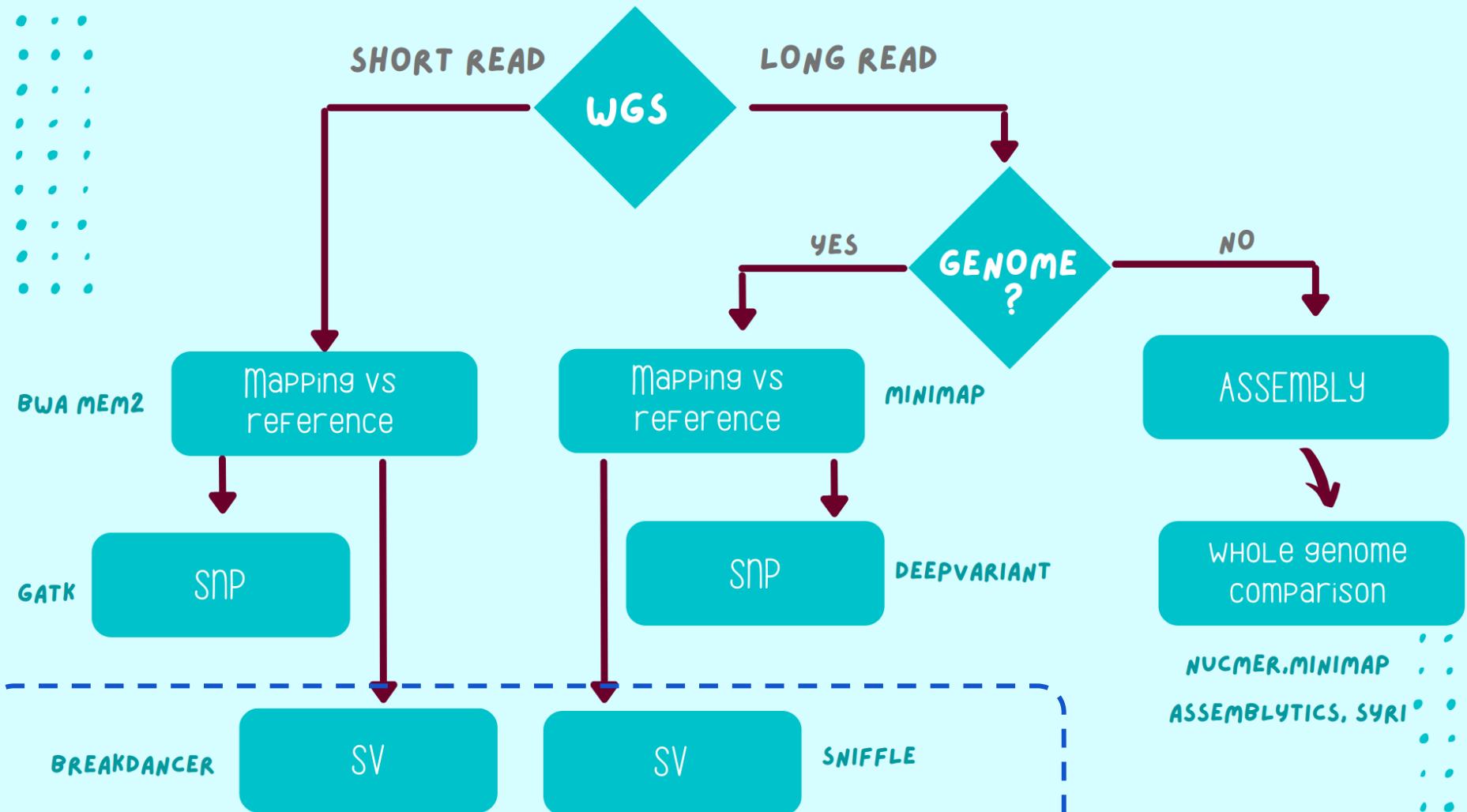
Problems with manual launches

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

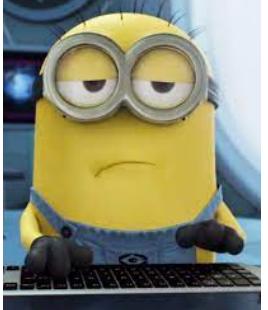
Solution ==> Workflow Manager (or scripts...):
SnakeMake, TOGGLE, NextFlow

Training plan - day 3

Sv DETECTION



Passer ça Jour3 !



Long Read Mapping

Two technologies

Oxford Nanopore



MinION

GridION

PromethION

Pacific BioScience



RSII



Sequel

from Elixir GAAS 2018

Long Reads

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

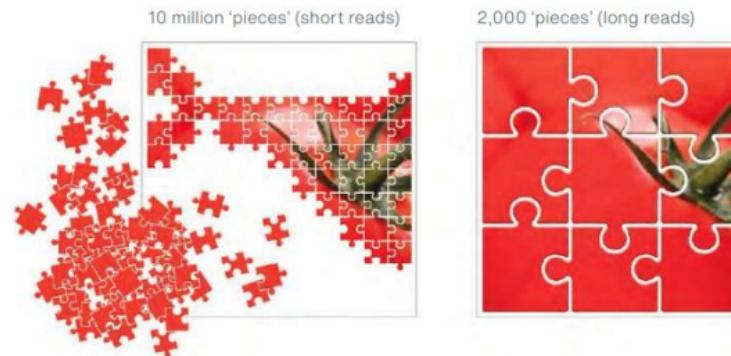
Microbial genomes

Human genomes

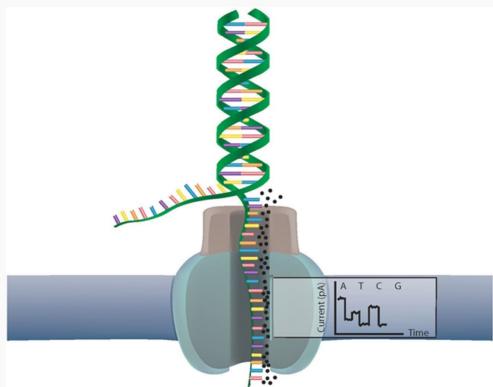
Animal genomes

Plant genomes

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

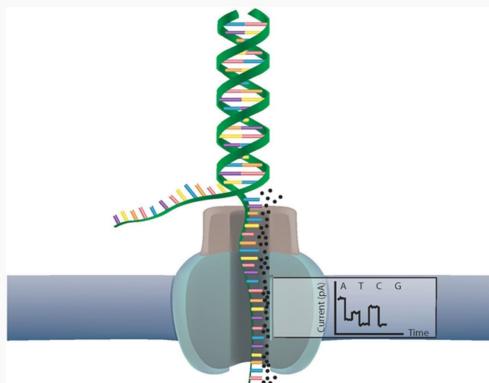






From Circulation
Research

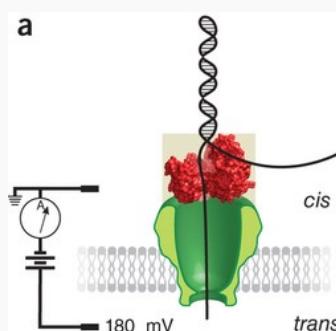
- No Amplification
- NO SYNTHESIS
- Very Long Length



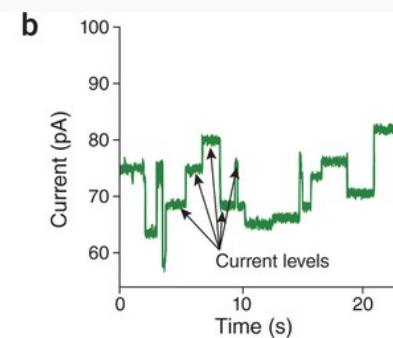
From Circulation Research

- No Amplification
- NO SYNTHESIS
- Very Long Length

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



From Nature Biotechnology



Advantages :

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

Advantages

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

Limits :

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

What you can do with it ?

Research areas

 Microbiology

 Human genomics

 Microbiome

 Clinical research

 Environmental

 Cancer

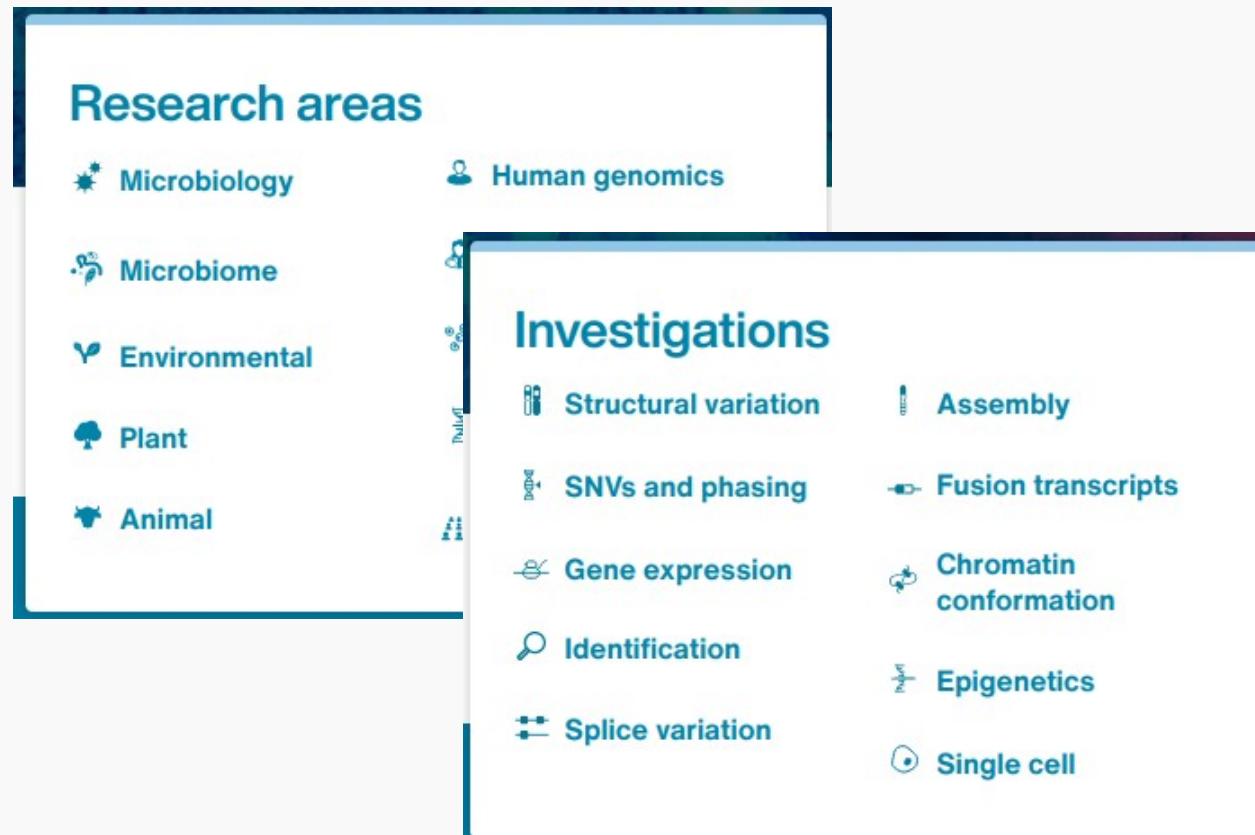
 Plant

 Transcriptome

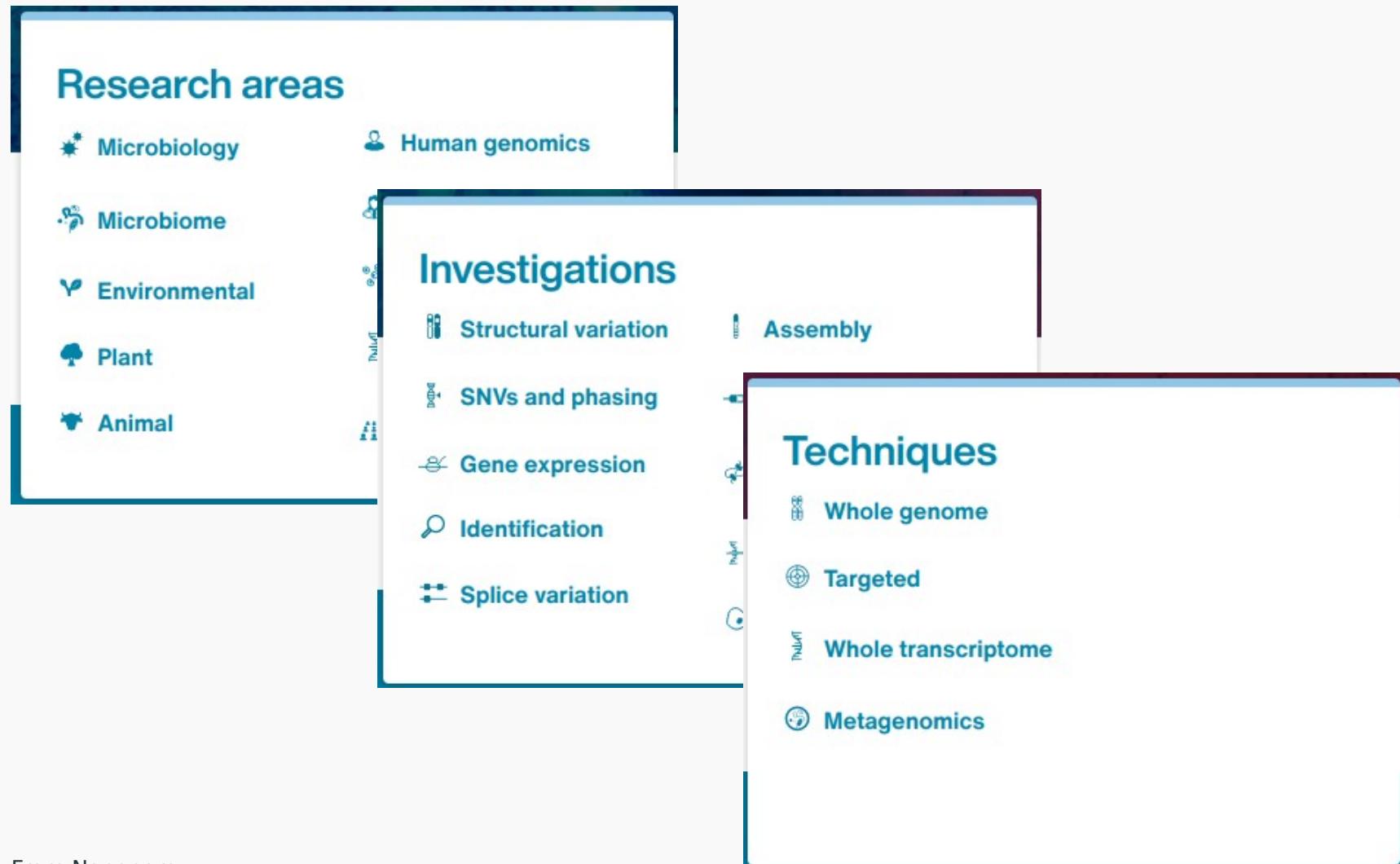
 Animal

 Populations
genomics

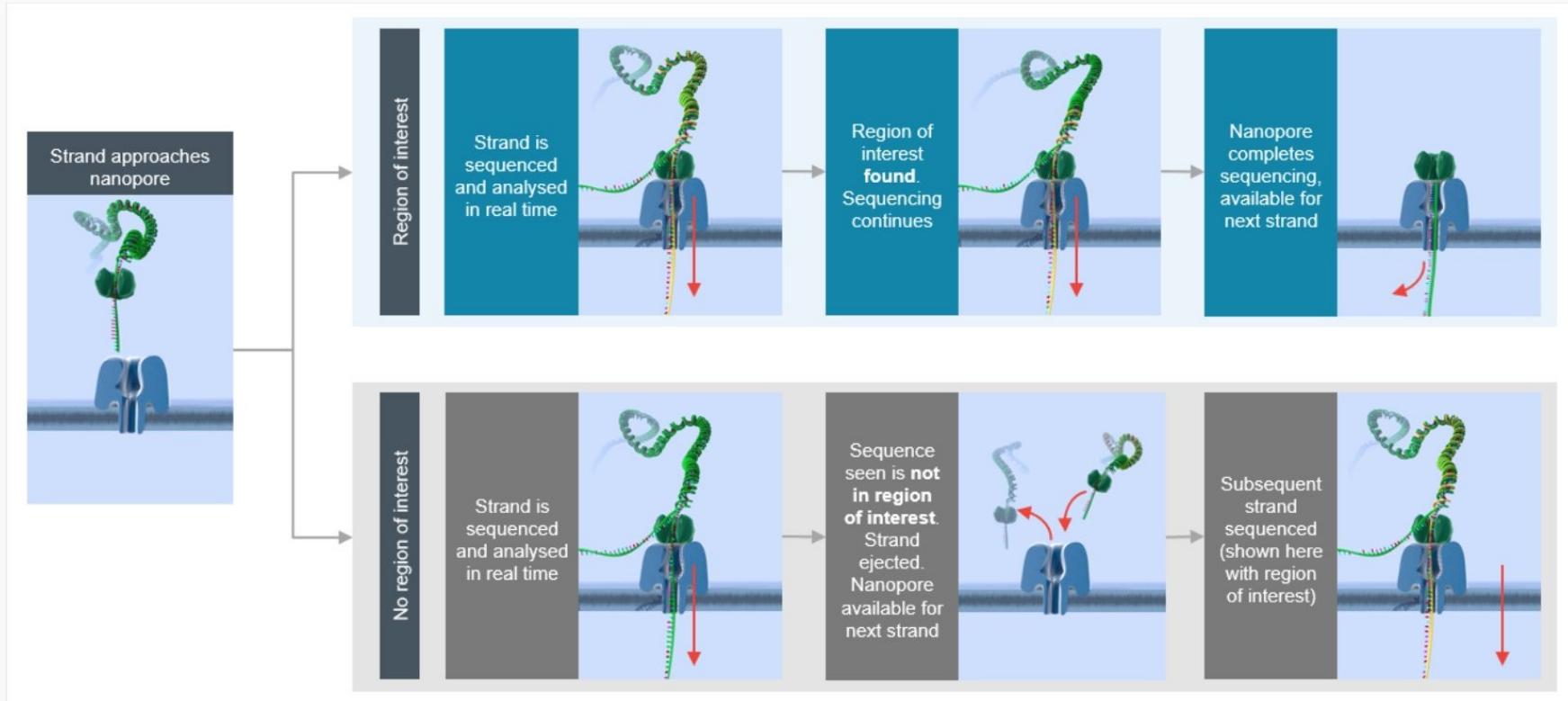
What you can do with it ?



What you can do with it ?



What you can do with it ?



2nd Generation Sequencing

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

PacificBiosciences
Oxford
Nanopore

3rd Generation Sequencing

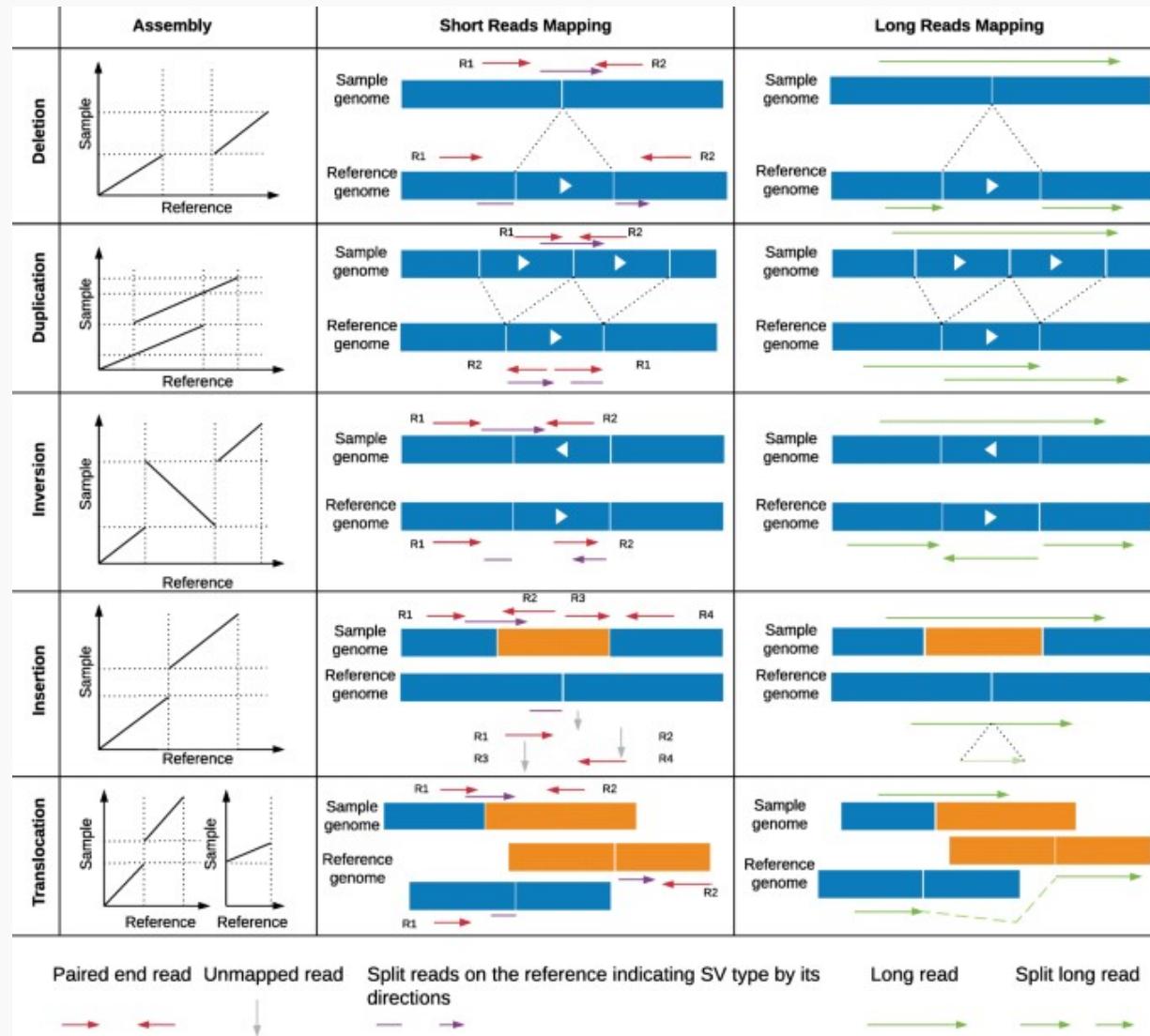
- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

NGS technologies comparison

NGS platforms

Platform	Template preparation	Chemistry	Max read length (bases)	Run times (days)	Max Gb per Run
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]

Structural Variant Detection





Practice

Let's work with the jupyter book :

[Day3 Variants structuraux EMPTY.ipynb](#)