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

www.southgreen.fr

https://southgreenplatform.github.io/trainings

SouthGreen bioinformatics platform

Modules de formation 2022

agap

Bioversity International

diade

PHIM

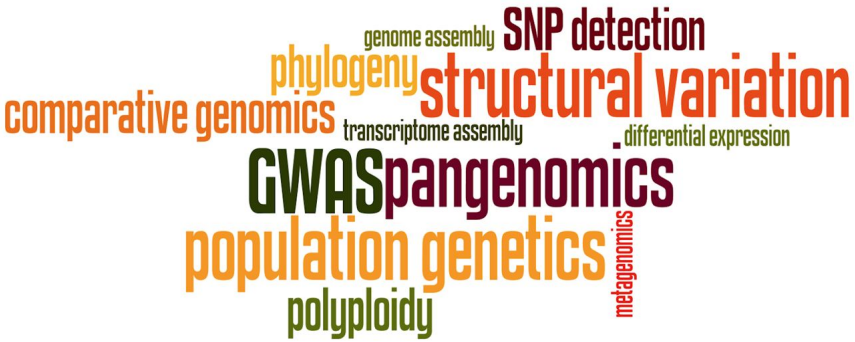
intertryp

MiVEGEC

TransVIHMI

SouthGreen bioinformatics platform

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




Rice

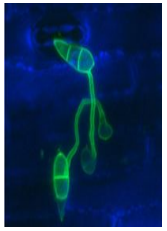

Banana


Palm


Sorghum


Coffee


Cassava


Magnaporthe

www.southgreen.fr

SouthGreen bioinformatics platform

Charte

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

SouthGreen bioinformatics platform

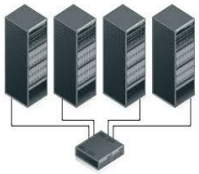
Workflow manager

TOGGLE
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

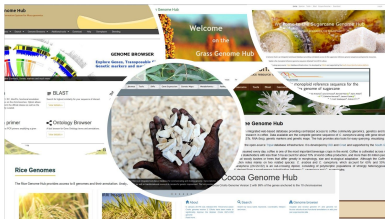
HPC and trainings….



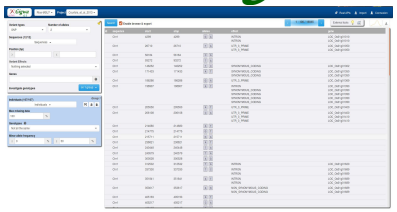
IRD Institut de Recherche pour le Développement

cirad

Genome Hubs & Information System

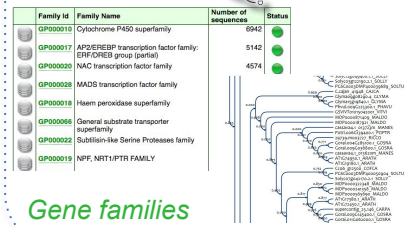


Gigwa



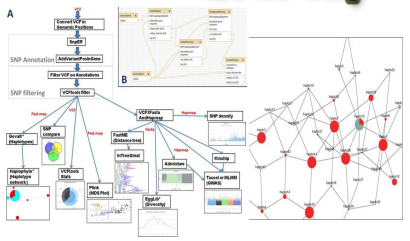
SNPs and Indels

GreenPhyl



Gene families

SNiPlay



https://github.com/SouthGreenPlatform

@green_bioinfo

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

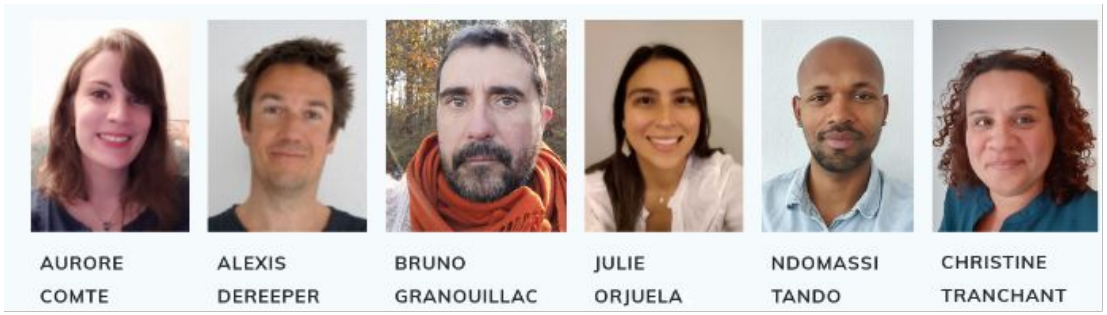
I-Trop

Plant & Health Bioinformatics Platform

Institut de Recherche pour le Développement FRANCE

EURO-QUALITY SYSTEM ISO 9001

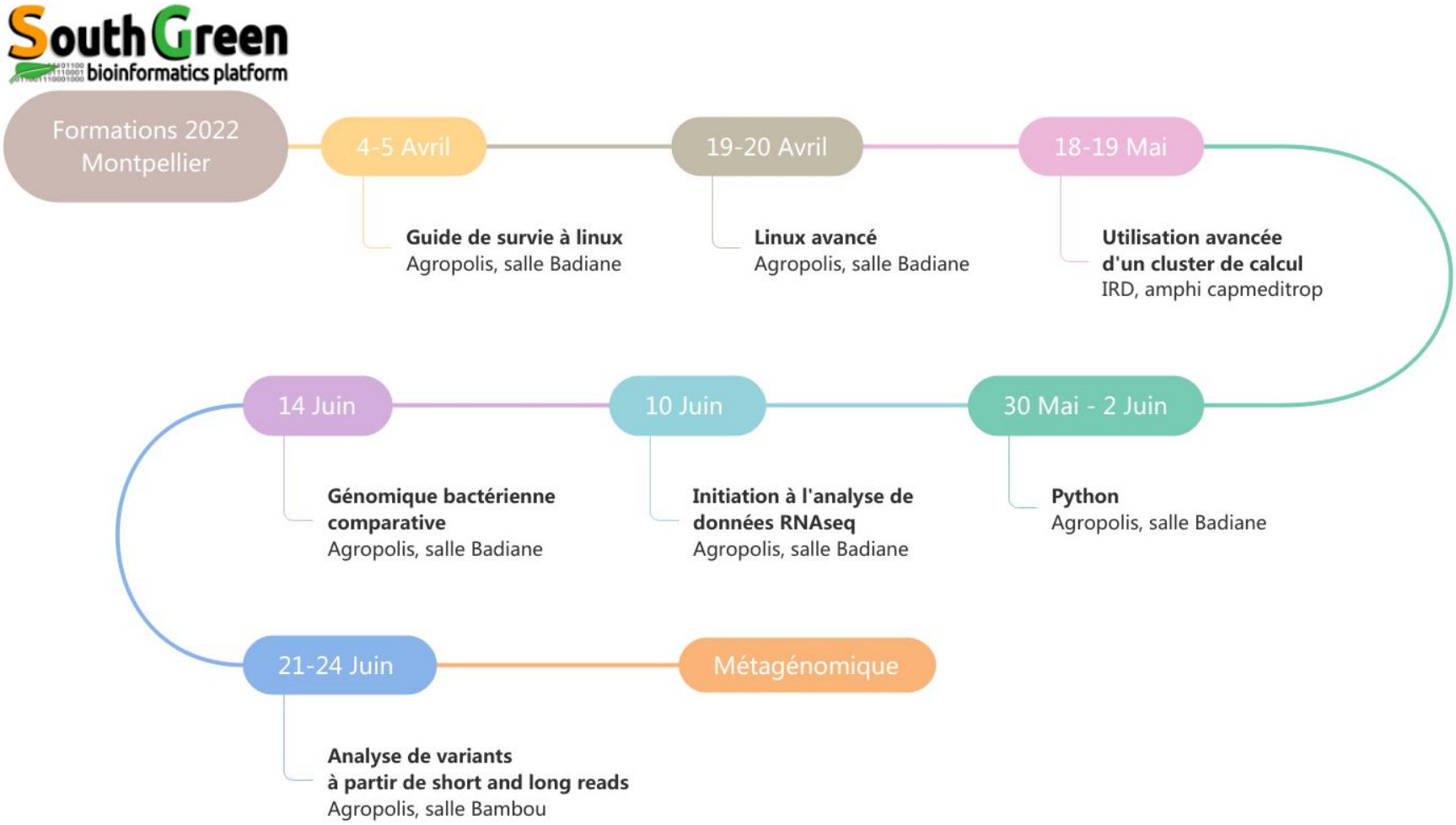
https://bioinfo.ird.fr/



AURORE COMTE	ALEXIS DEREEPER	BRUNO GRANOUILLAC	JULIE ORJUELA	NDOMASSI TANDO	CHRISTINE TRANCHANT
IE bioinfo	IE bioinfo	IE systèmes d'information	IE bioinfo	IE systèmes	IR bioinfo

bioinfo@ird.fr

@ItropBioinfo


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
30 Mai - 2 Juin
Python
Agropolis, salle Badiane
10 Juin
Initiation à l'analyse de
données RNAseq
Agropolis, salle Badiane
14 Juin
Génomique bactérienne
comparative
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

Objectifs

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***

Objectifs

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

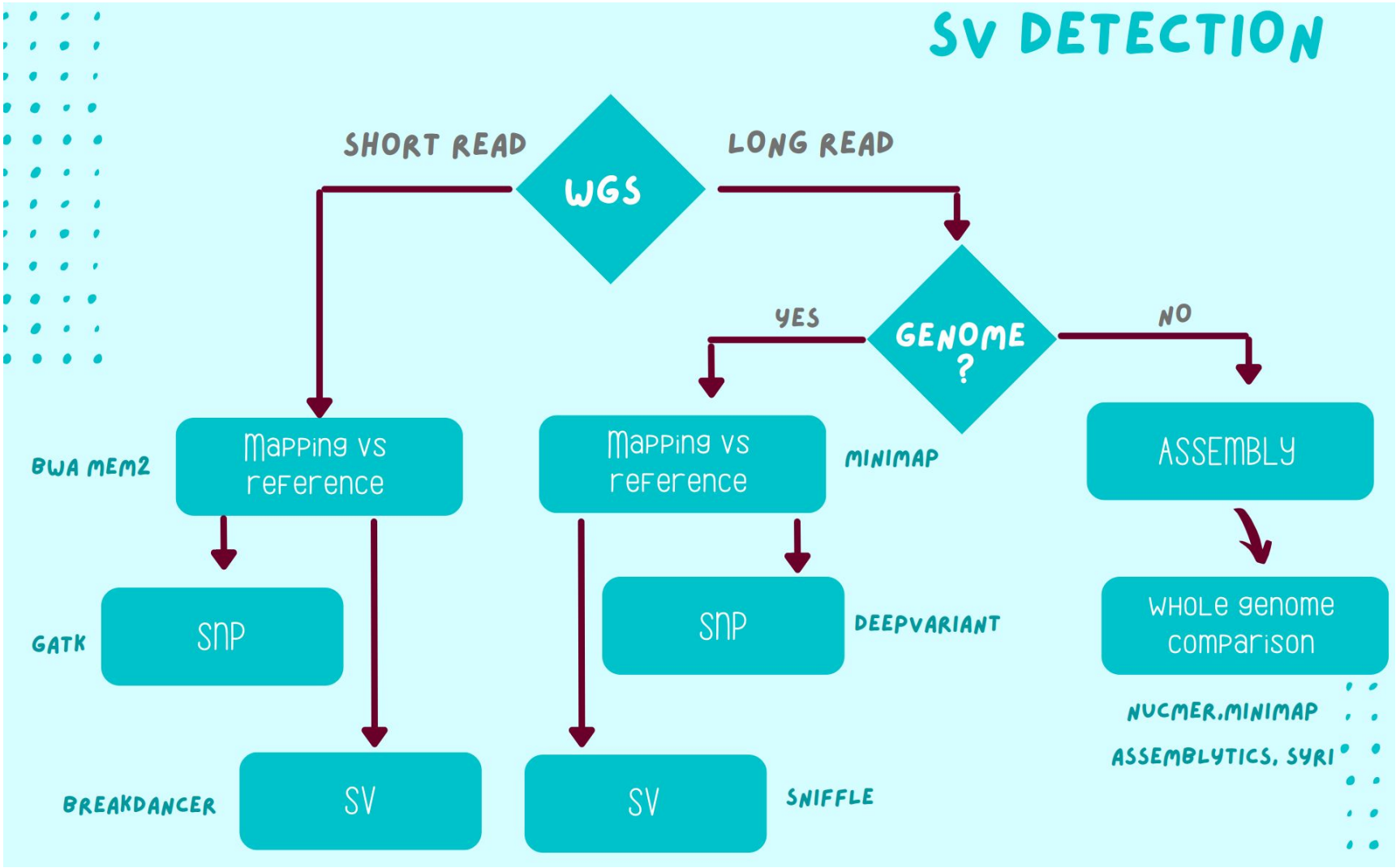
Applications :

- 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

Plan de bataille !!!



Let’s discover Jupyter through the IFB cloud

Working environment

What is jupyter book ?

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

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

Why use jupyter book ?

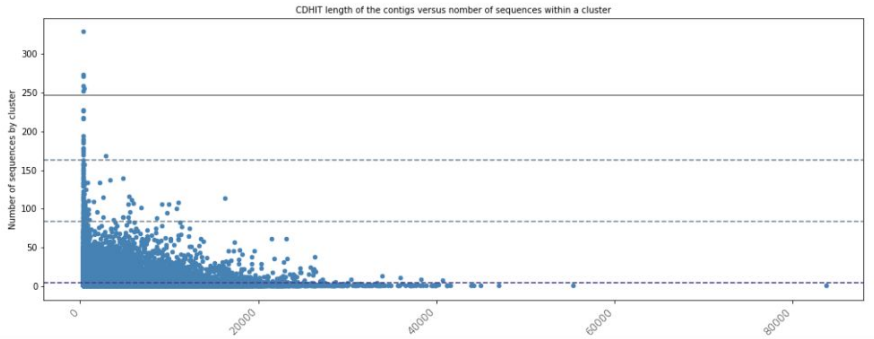
An unique interface/file where text,code and output codes can be mixed :

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

All ctgs

Entrée [30]:

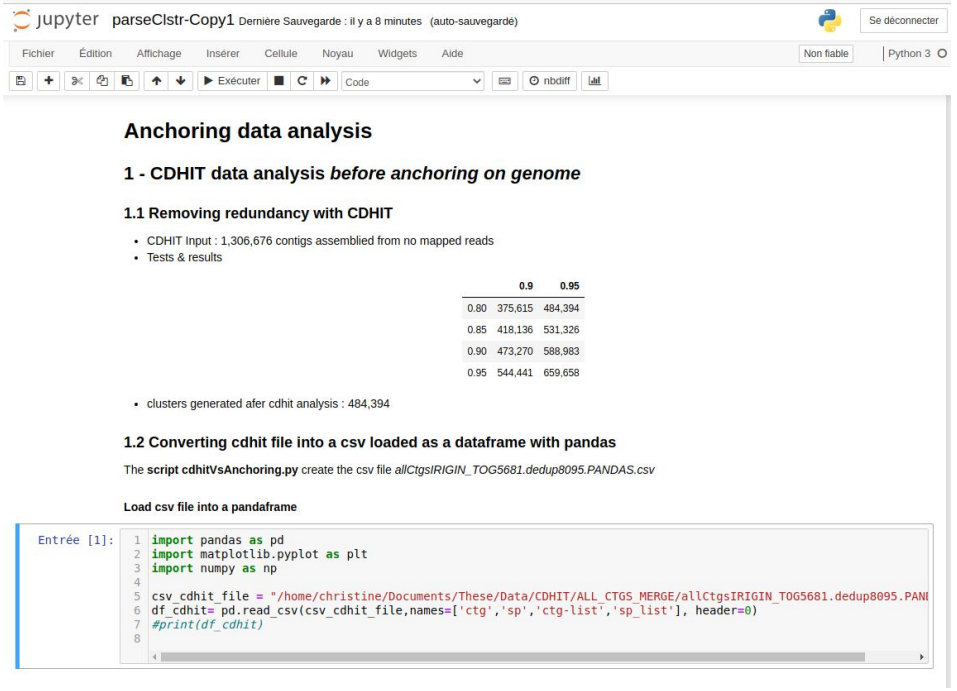
```
plt.figure(figsize=(17, 6))
ax = plt.gca()
df_cdhit[(df_cdhit.sp == 'Ob')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
df_cdhit[(df_cdhit.sp == 'Og')].plot(x='pb', y='ln', kind="scatter", color='steelblue', ax=ax, linewidth=1)
plt.axhline(y=5, color='darkslateblue', linestyle='--')
plt.axhline(y=84, color='slategrey', linestyle='--')
plt.axhline(y=163, color='slategrey', linestyle='--')
plt.axhline(y=247, color='grey', linestyle='-')

plt.title("CDHIT length of the contigs versus nomber of sequences within a cluster", fontsize=10)
plt.xlabel('Cluster length')
plt.ylabel('Number of sequences by cluster')
plt.xticks(
    rotation=45,
    horizontalalignment='right',
    fontweight='light',
    fontsize='12',
)
plt.show()
```



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



Lab notebook for science data ?

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

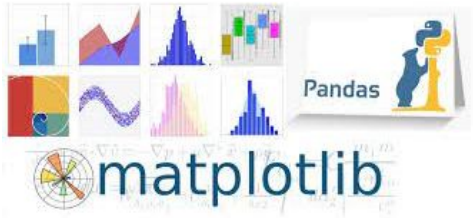
How to become a super datascientist ?

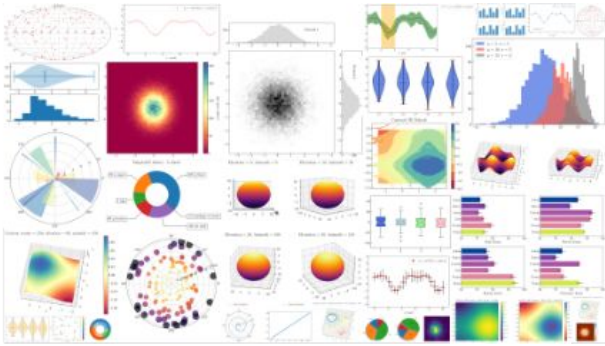
- facilement importer des fichiers tabulés dans des dataframes, similaires aux dataframes sous R.

  (et exporter)

- manipuler ces tableaux de données / DataFrames

- facilement tracer des graphes à partir de ces DataFrames grâce à matplotlib





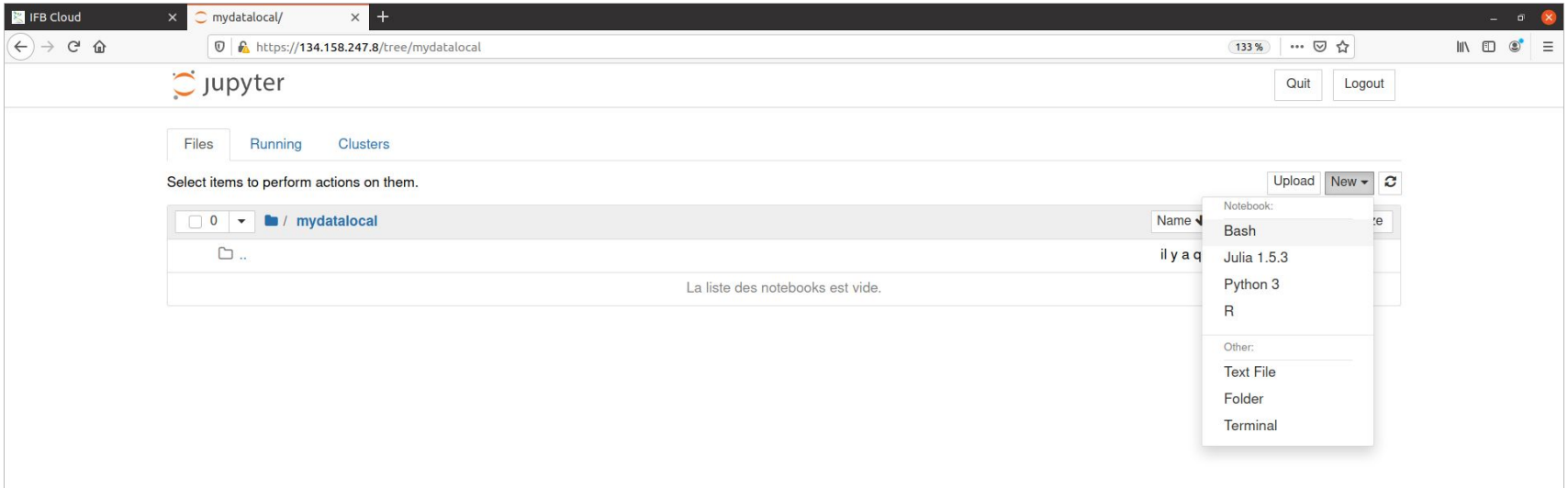
How will you use Jupyter Notebook ?

- 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”

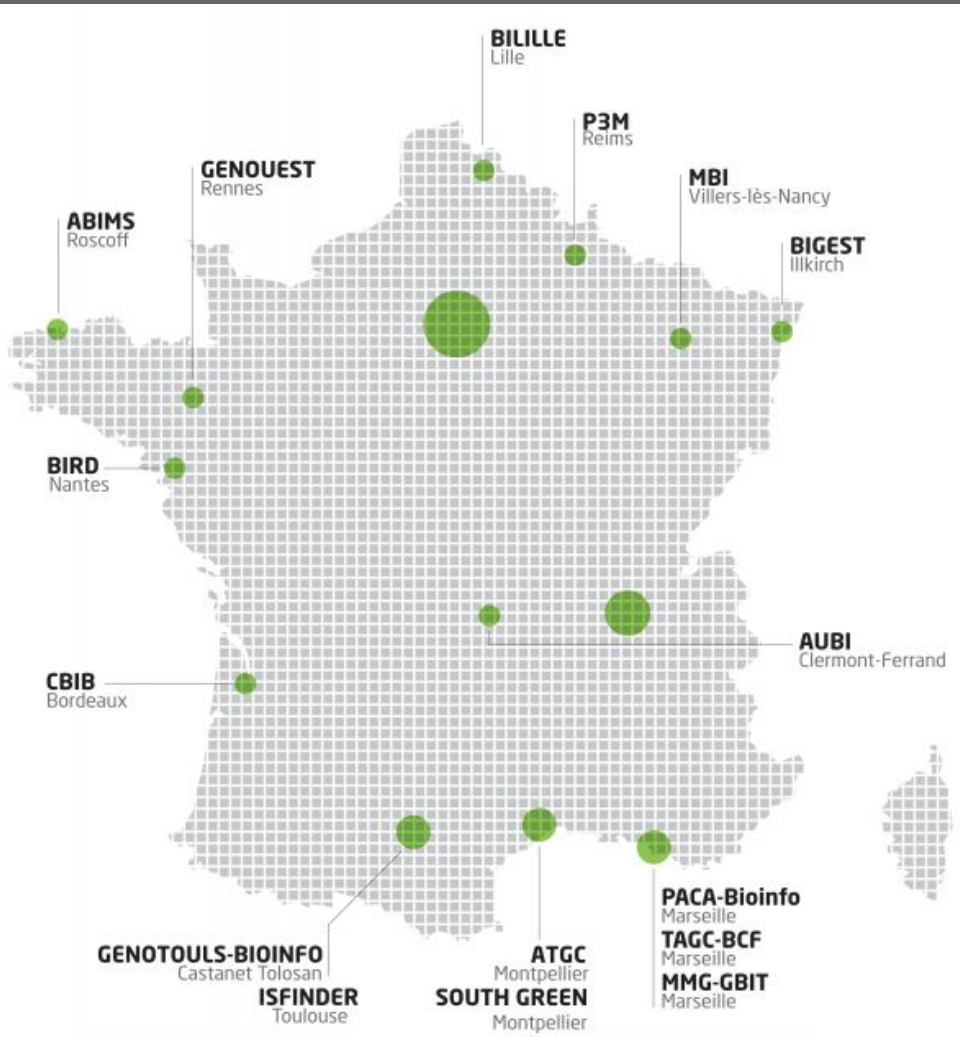
- Through this virtual machine, we will create jupyter books and execute all our analysis



IFB ?

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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

Biosphere, IFB CLOUD FOR LIFE SCIENCES

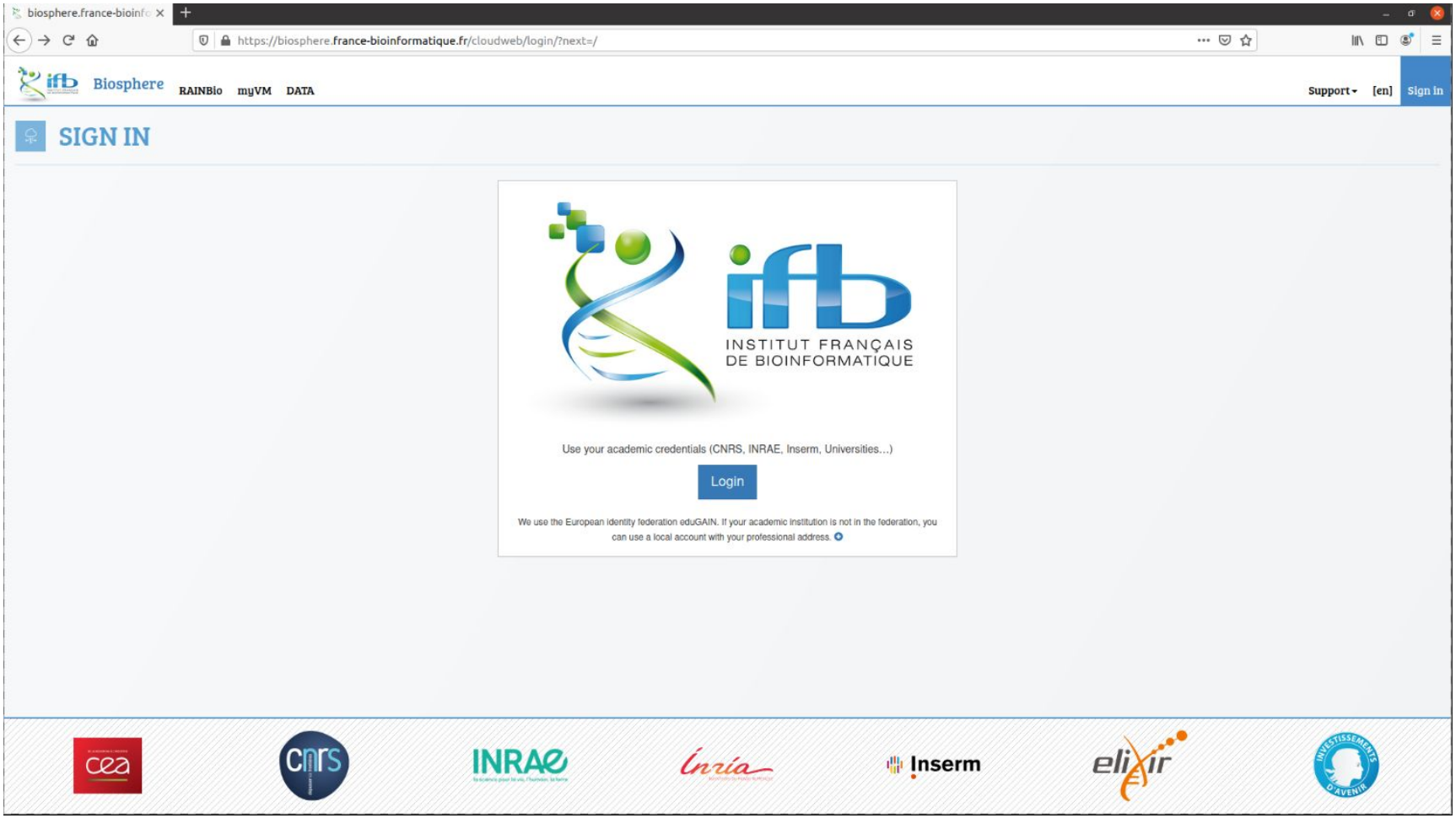
- A federation of clouds, which relies on interconnected IFB's infrastructures, providing distributed services to analyze life science data

- .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.

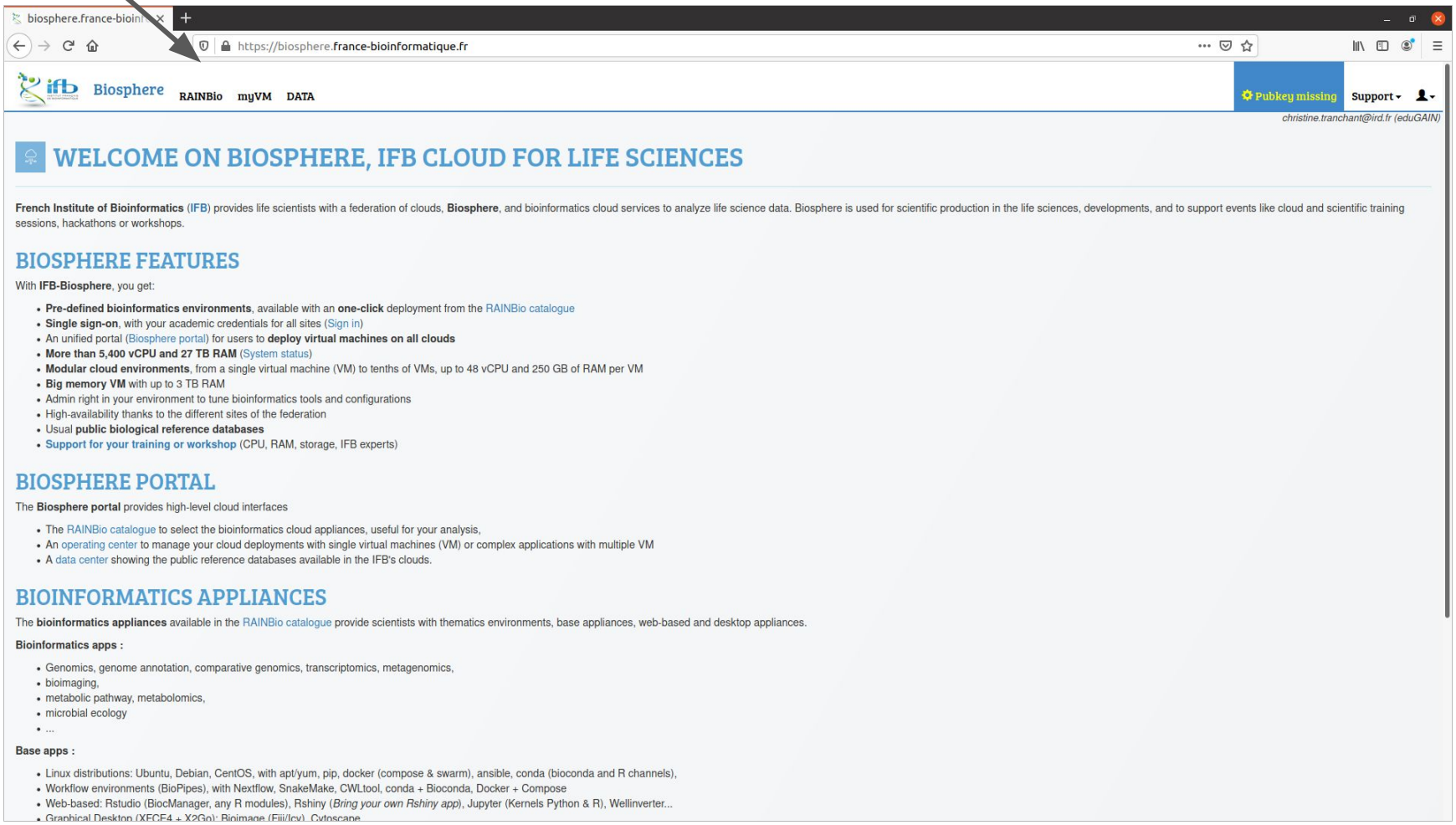
Let's start with biosphere

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



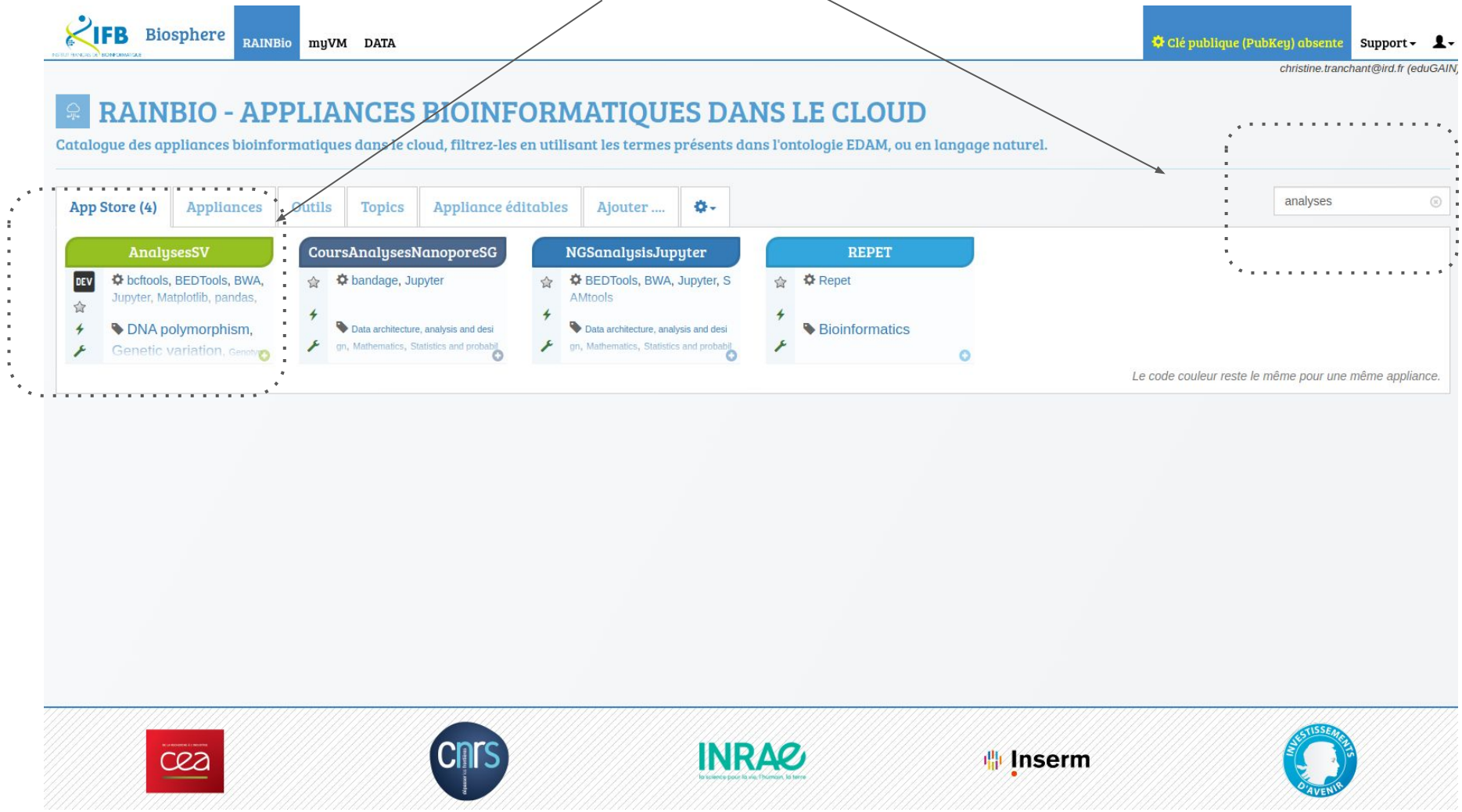
Connected / here we are

RAINBIO catalog to access our Virtual Machine (VM)

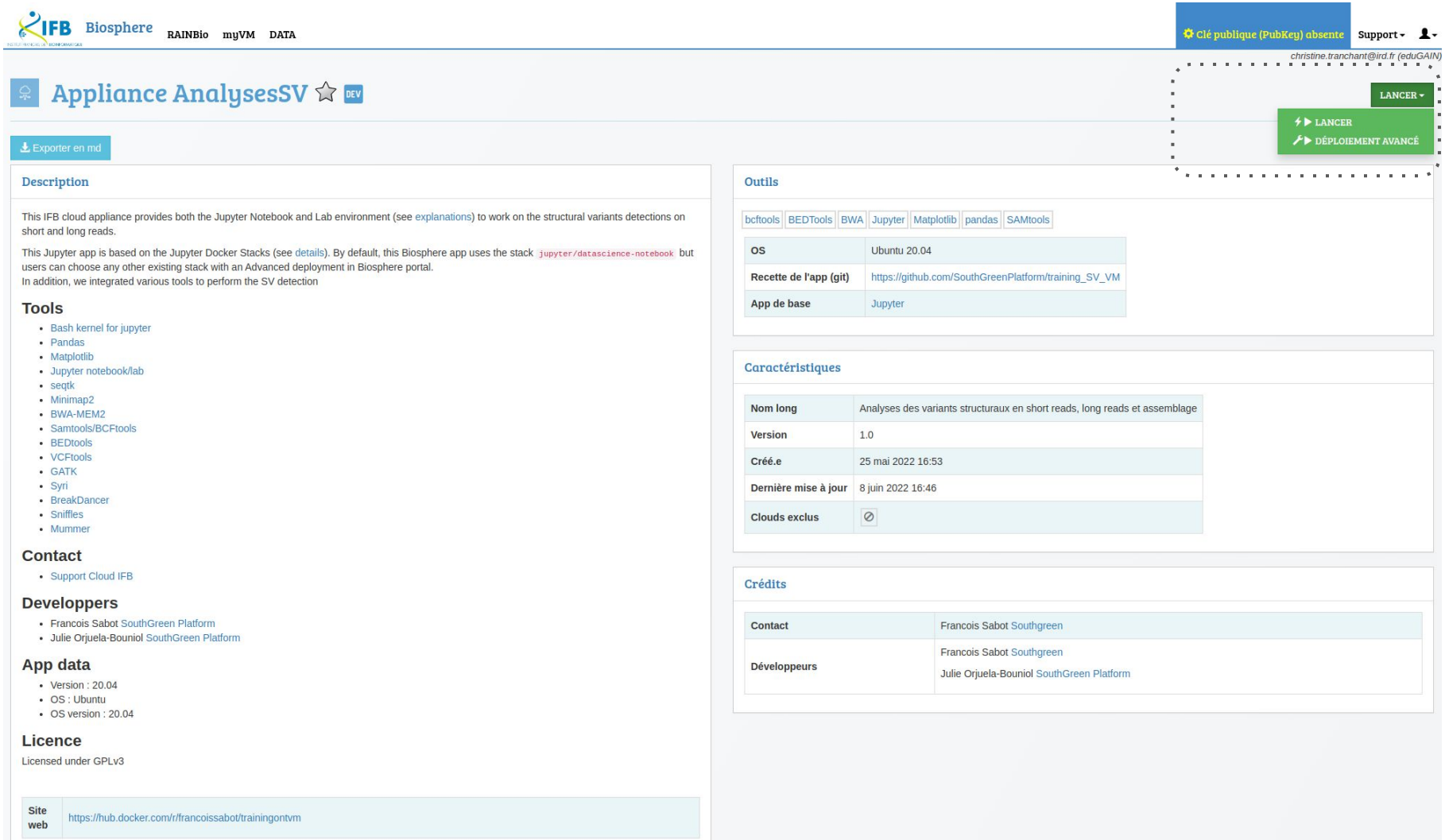


Searching for the vm we will use

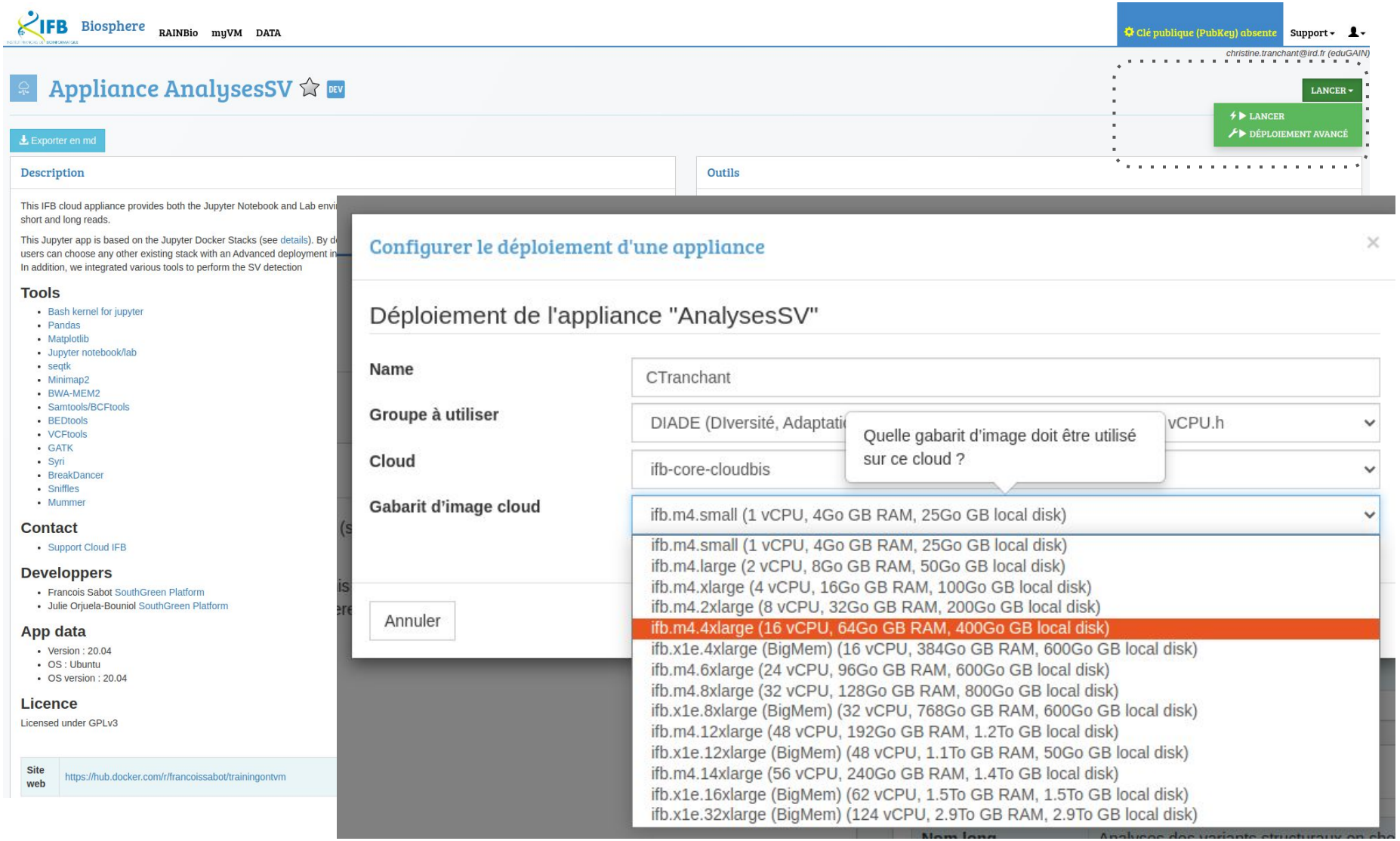
vm's name : **analysesSV**



Let’s run your vm through the cloud

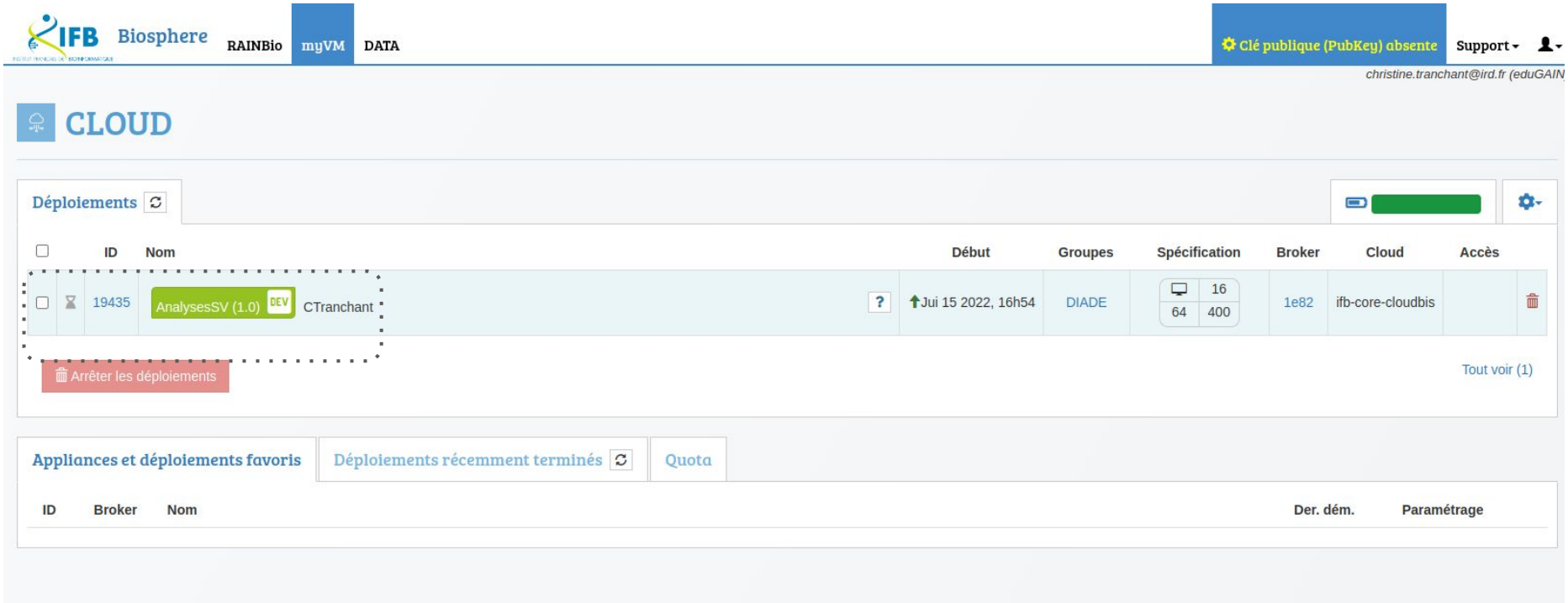


Let’s run your vm through the cloud



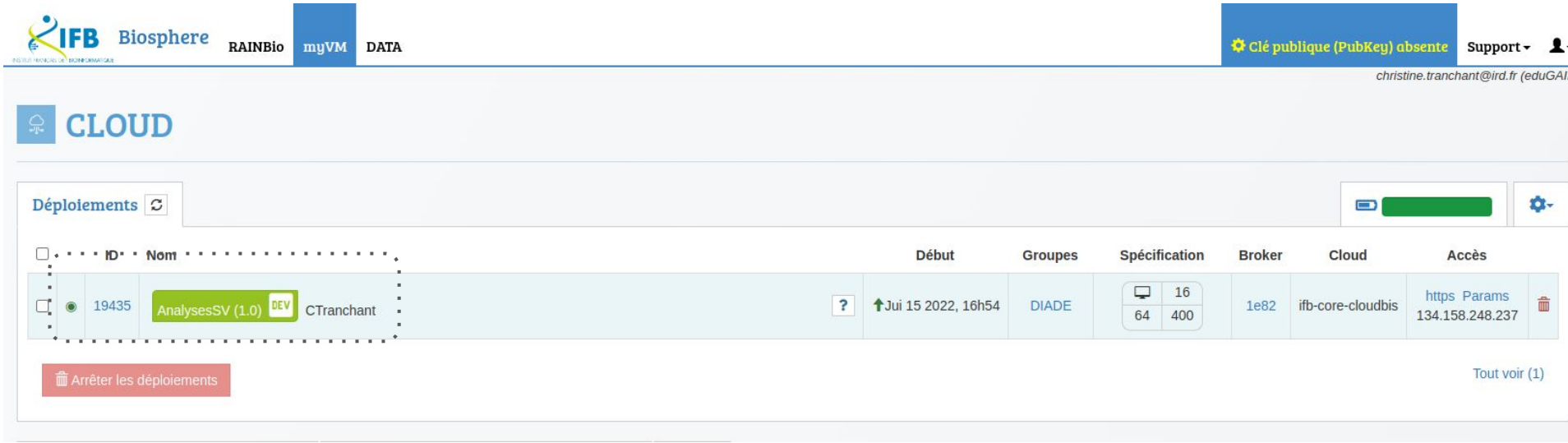
Let’s run your vm through the cloud

Loading…



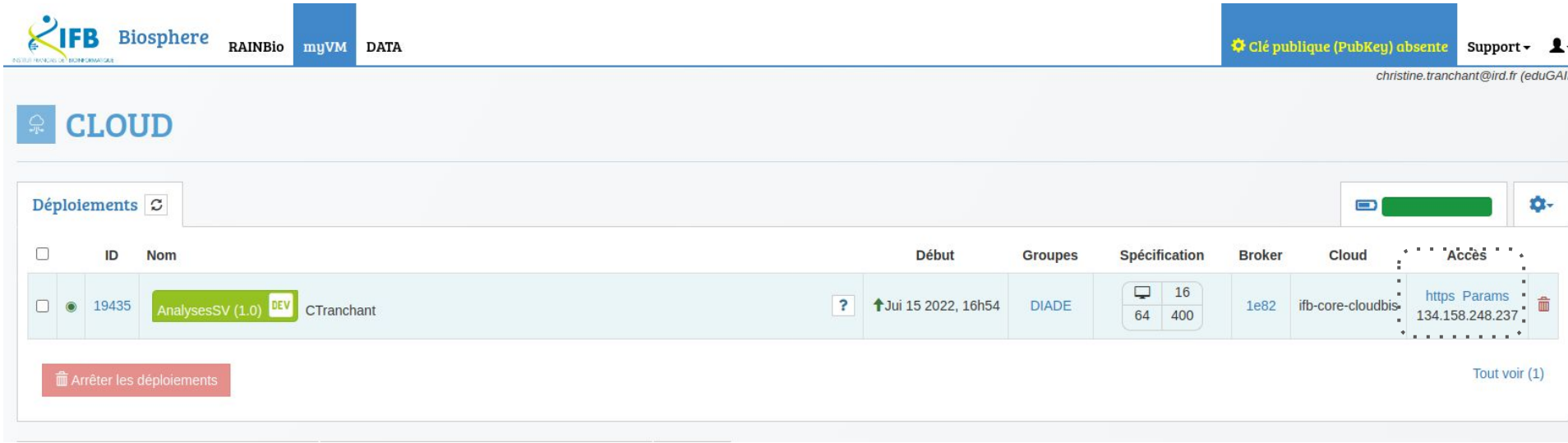
Let’s run your vm through the cloud

ready !



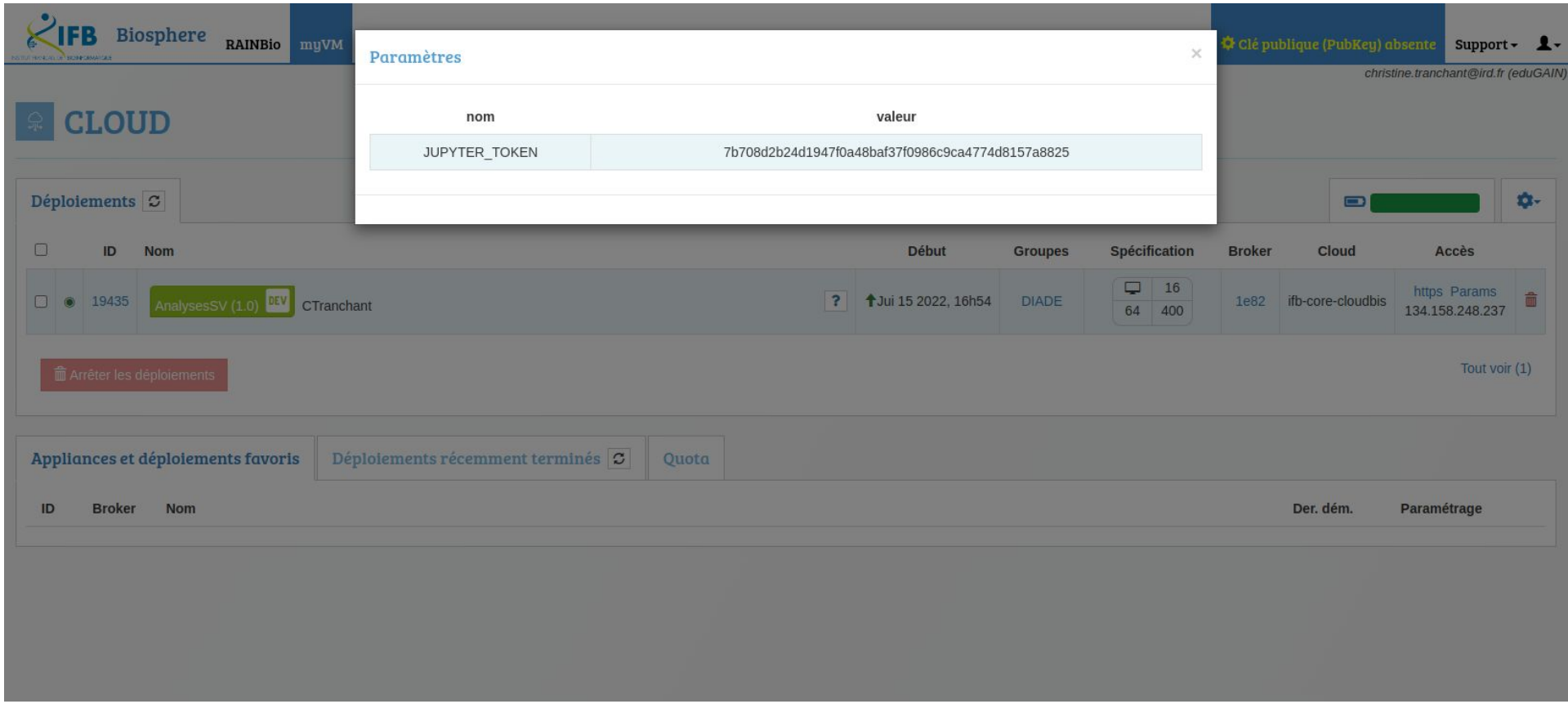
Let’s run your vm through the cloud

get the url… link “https”



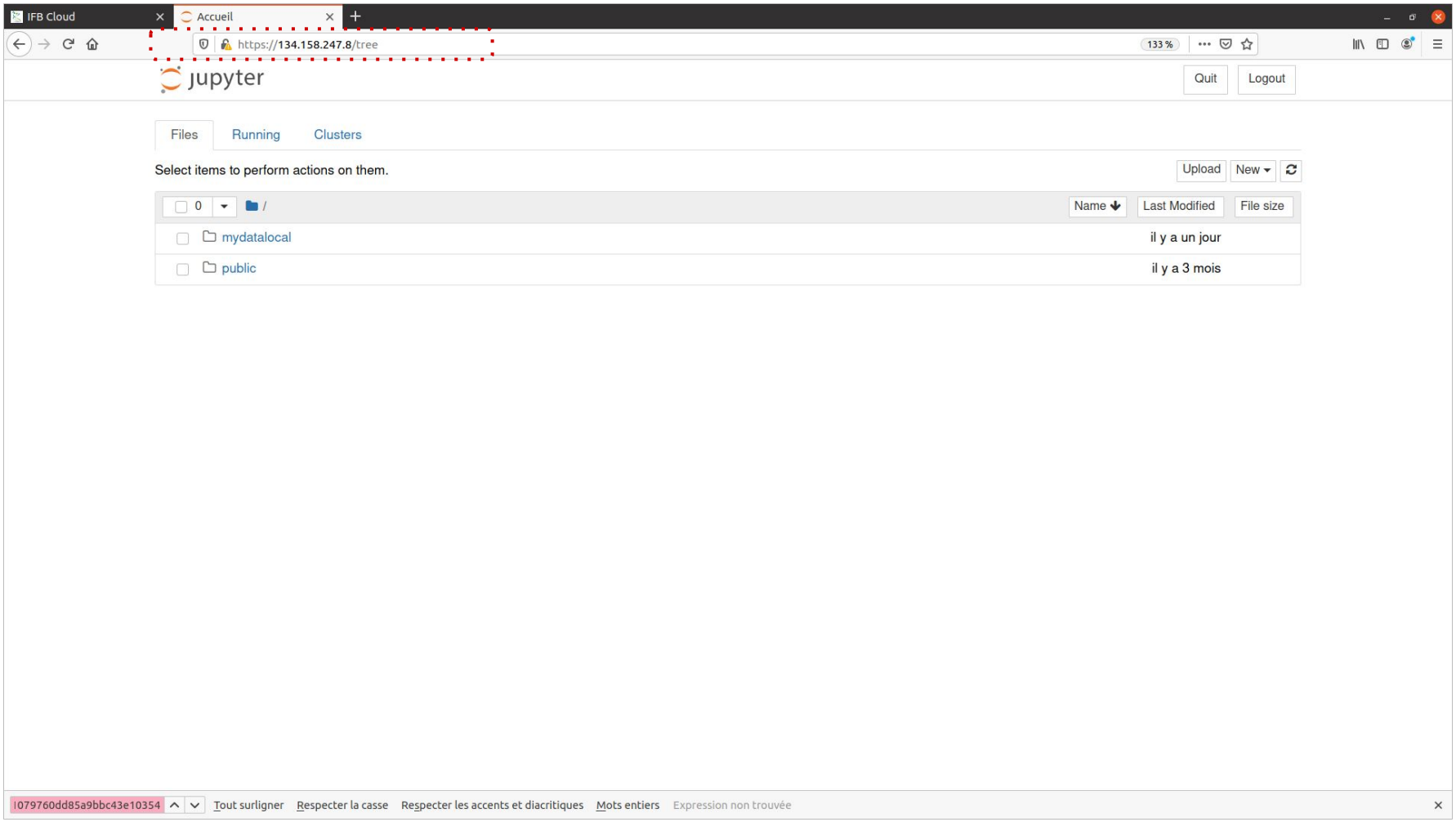
Let’s run our vm through the cloud

Get the token identifiant… link “Params”



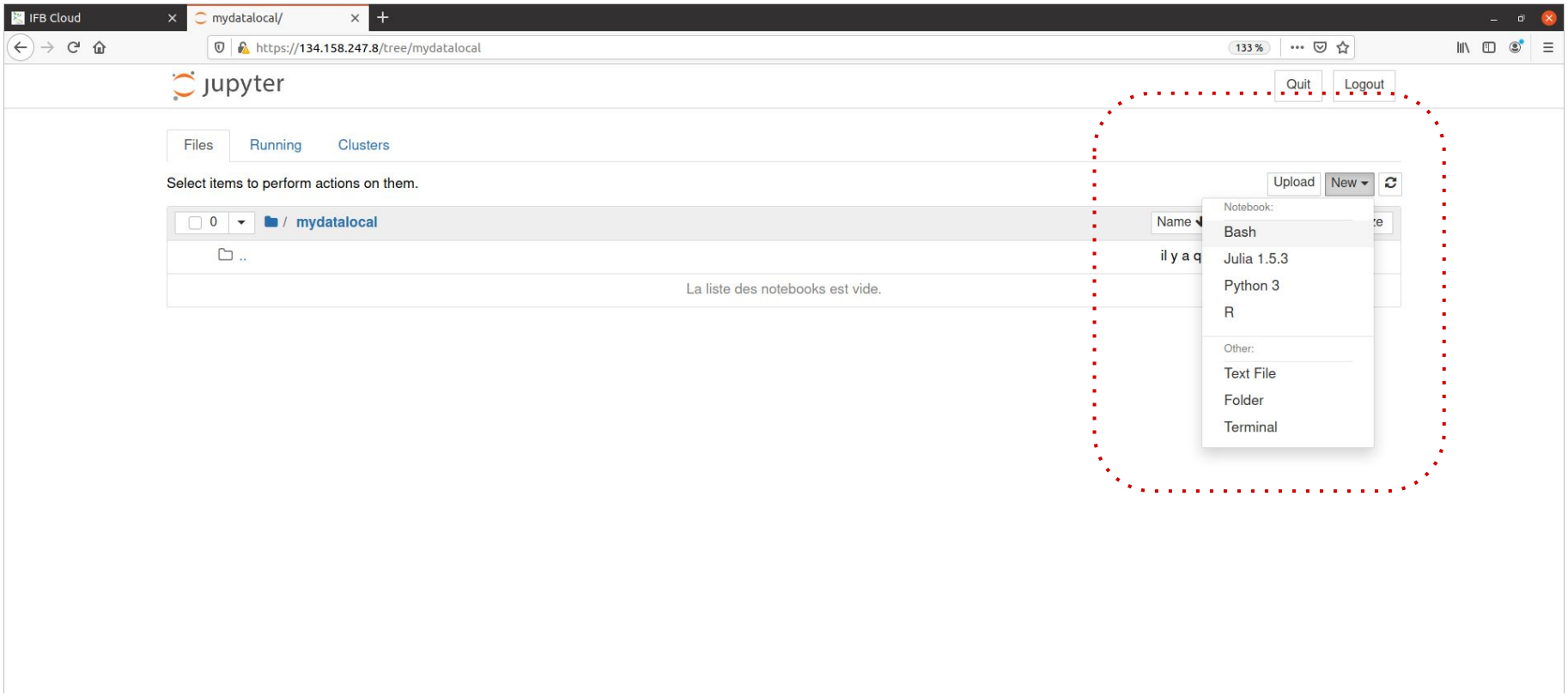
Let’s run our vm through the cloud

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



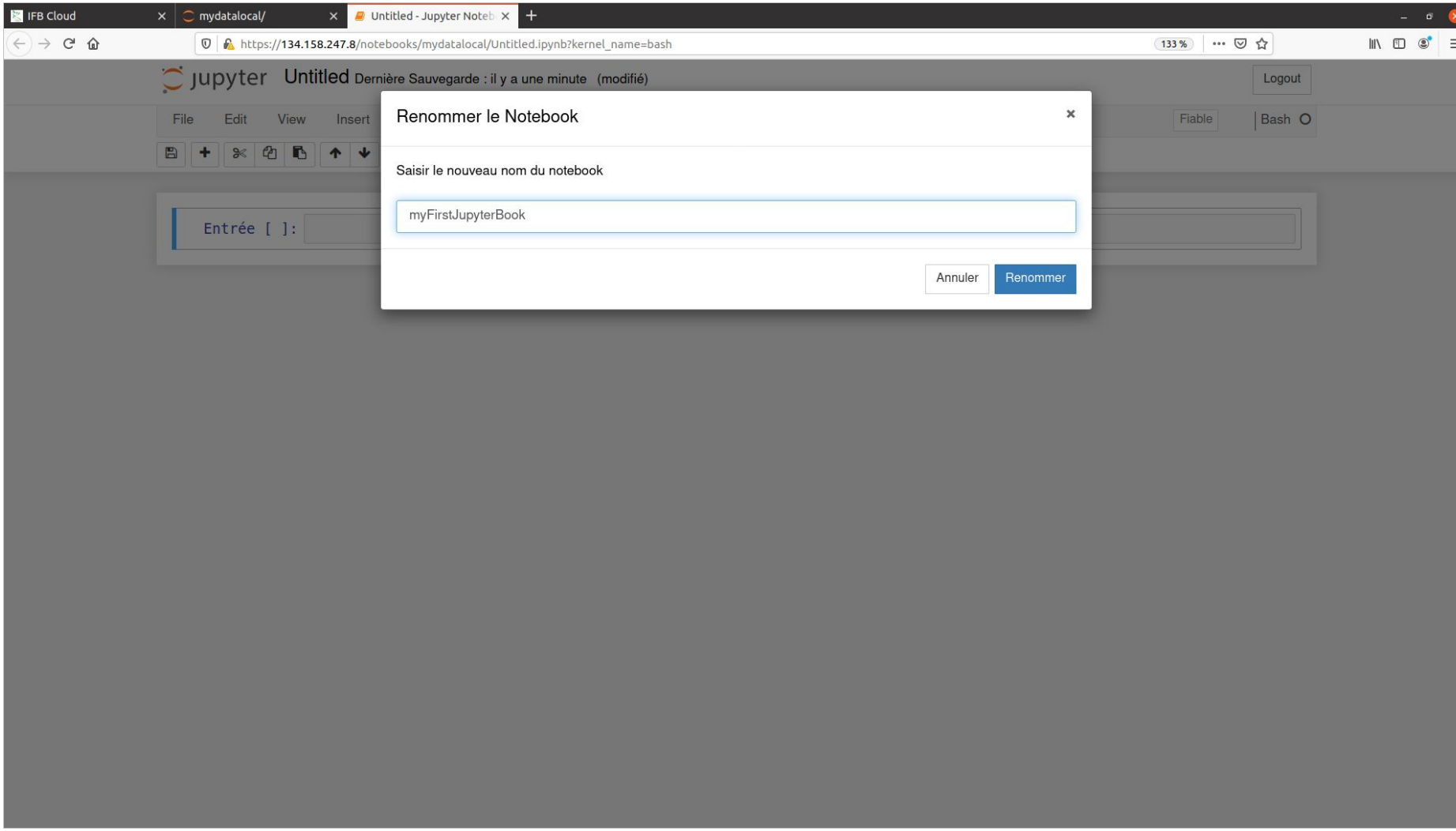
Create your first jupyter book

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



Rename your first jupyter book

myFirstJupyterBook

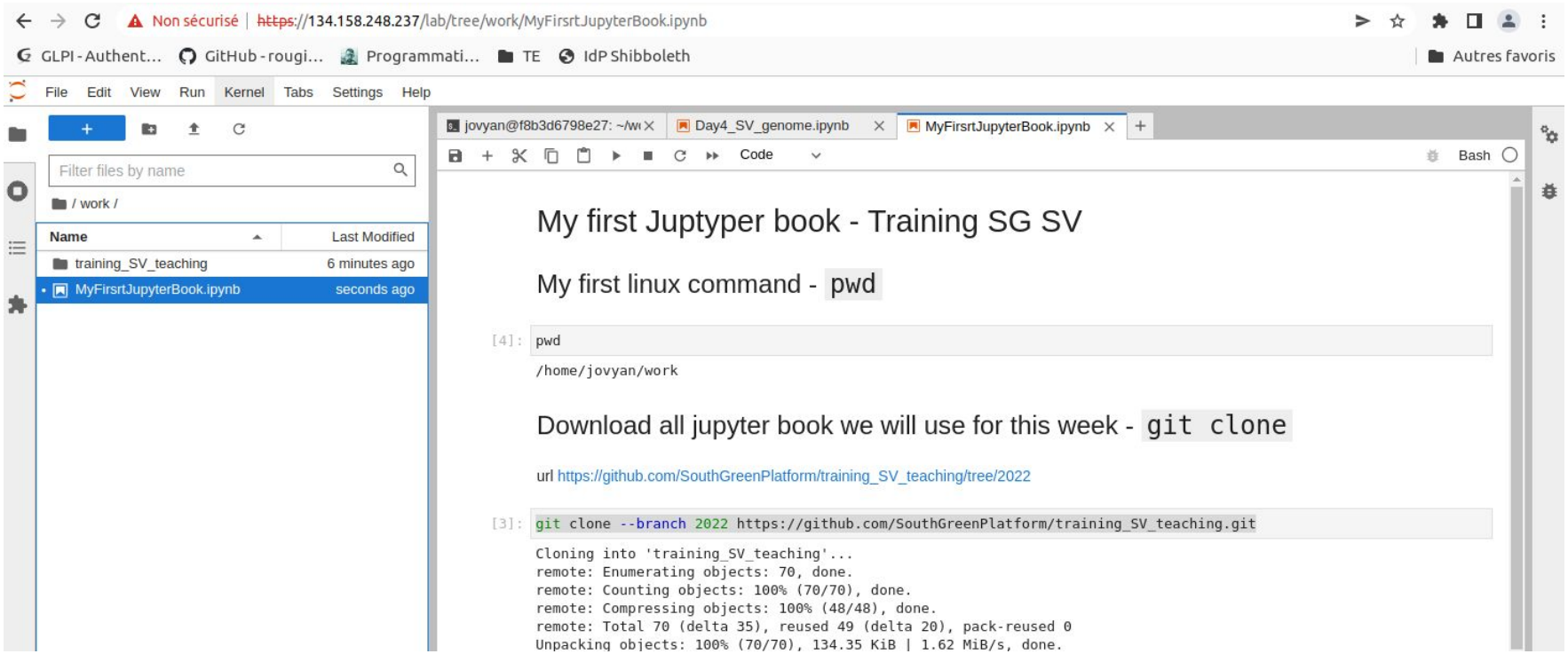


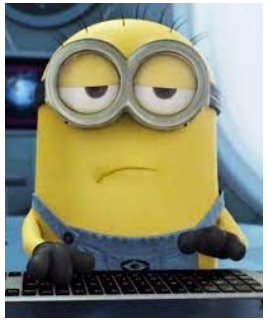
Run your first bask command - *git clone*

- All jupyterbook used for practice are here :
  https://github.com/SouthGreenPlatform/training_SV_teaching/tree/2022

- Download all the jupyter books with the command *git clone*

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

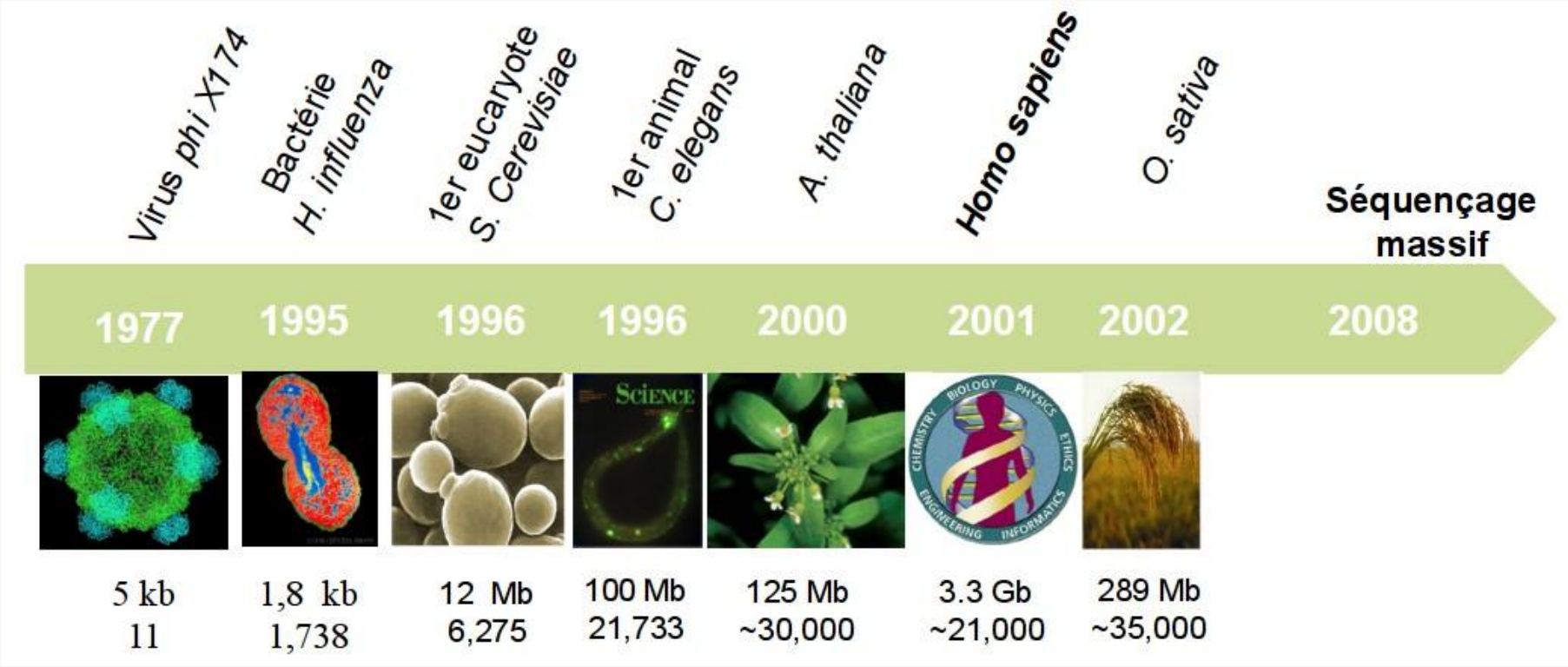




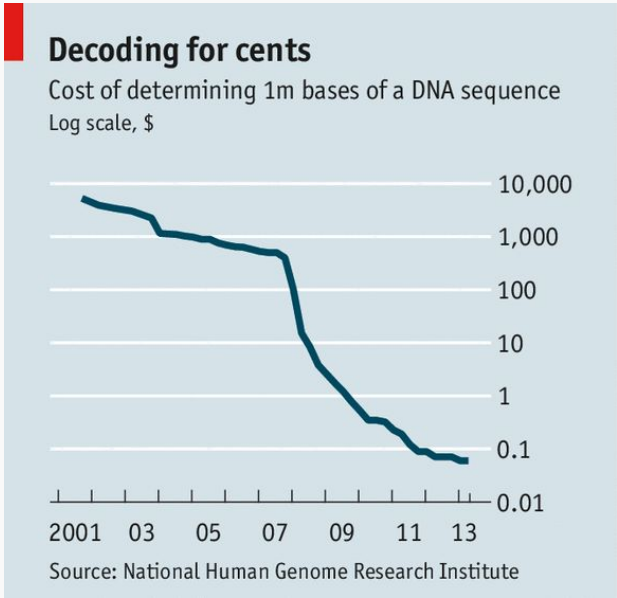
Introduction & NGS method

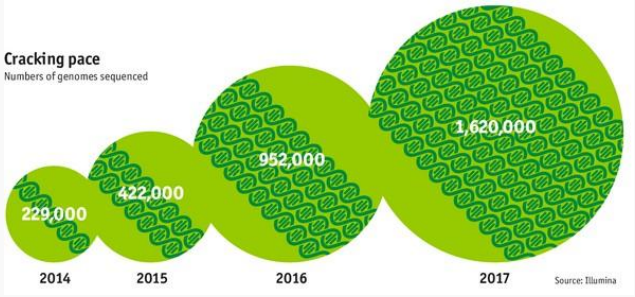
The NGS in themselves

A little history of sequencing...

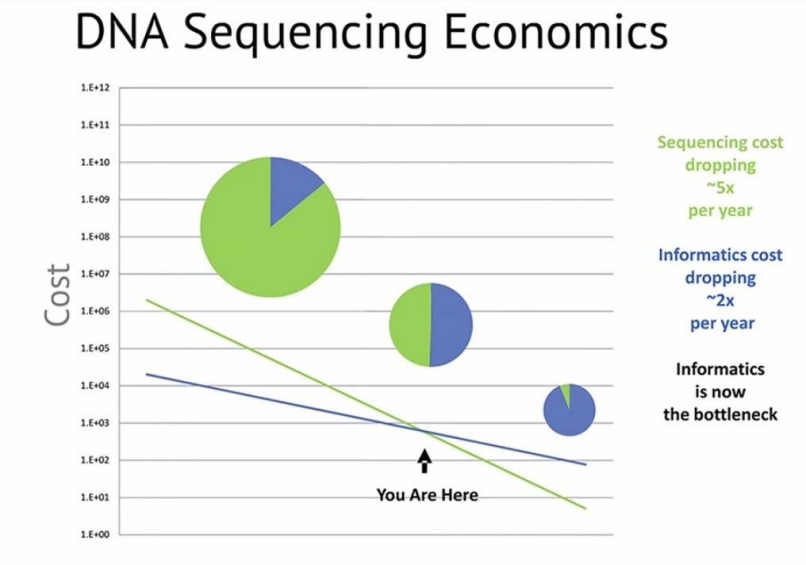


...From Data Rarity to Data Deluge



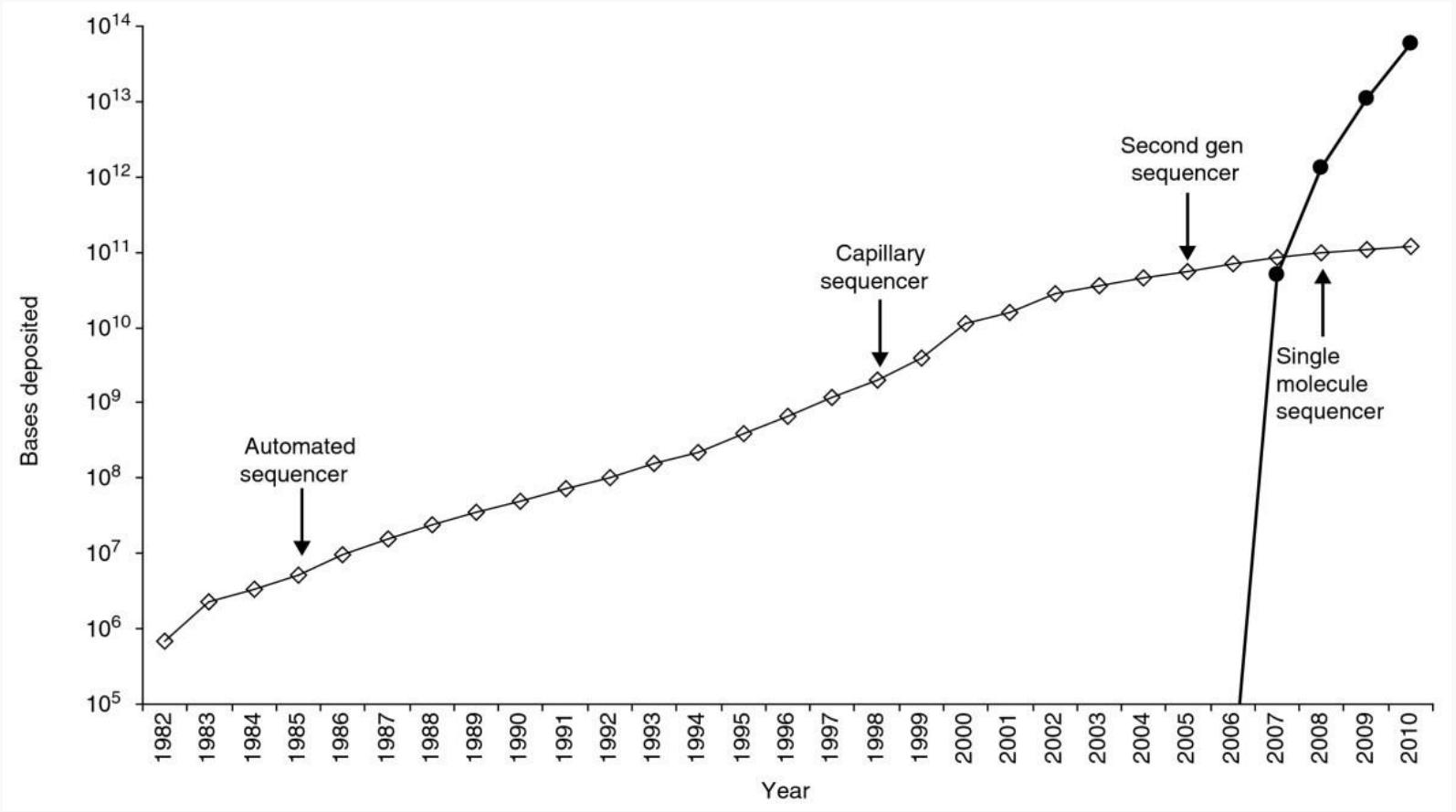


From The economist



From Bussiness Insider

...From Data Rarity to Data Deluge



What can we do with it ?

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

Methods

Current Technologies

2nd Generation Sequencing

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

Current Technologies

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

Current Technologies

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

Current Technologies

2nd Generation Sequencing

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

454
IonTorrent
Illumina

3rd Generation Sequencing

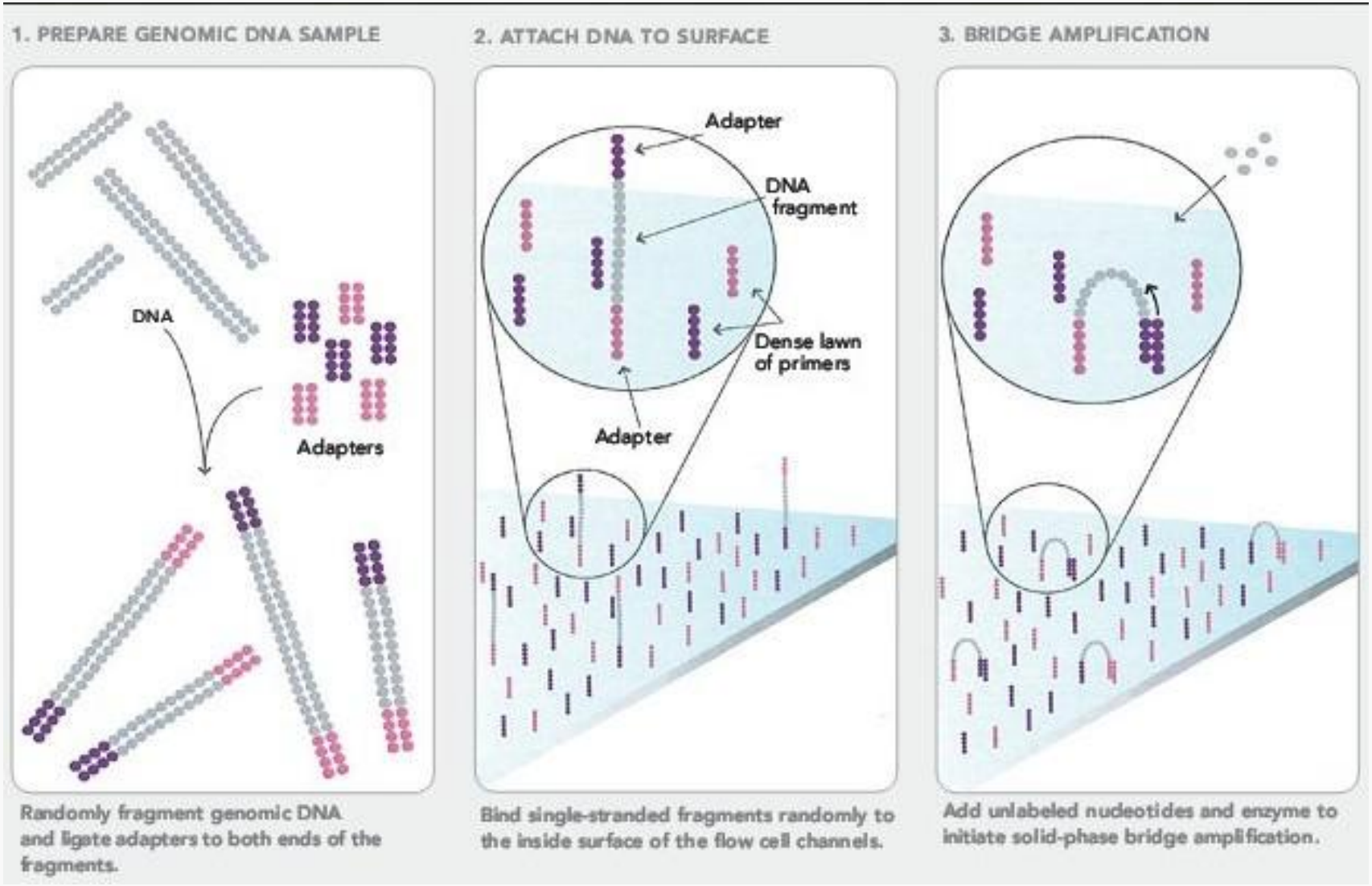
- DNA fragmentation (long)
- NO MATRIX AMPLIFICATION
- Long reads
- Important error rate
- Medium throughput

PacificBiosciences
Oxford Nanopore

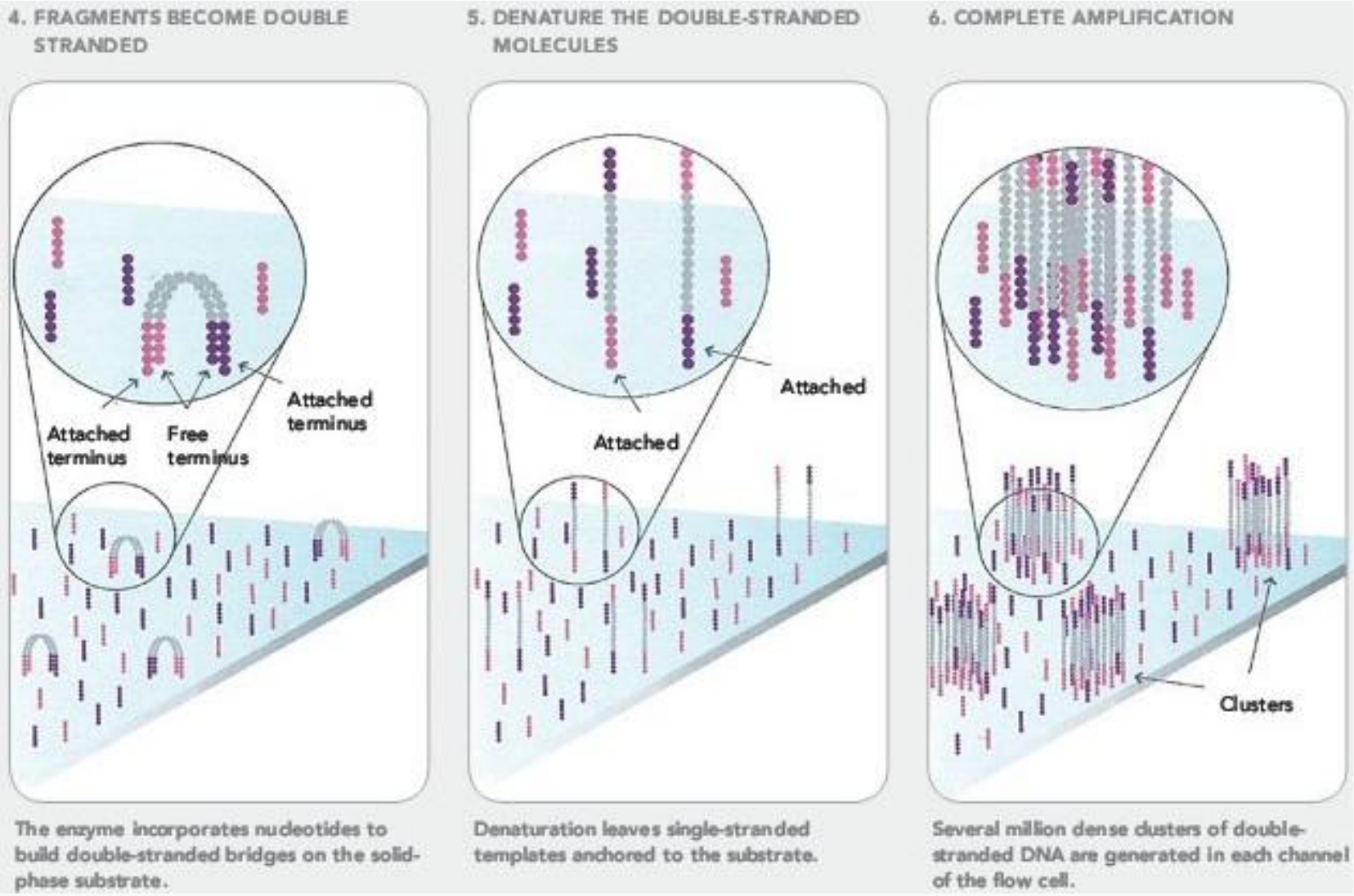
Illumina

Illumina

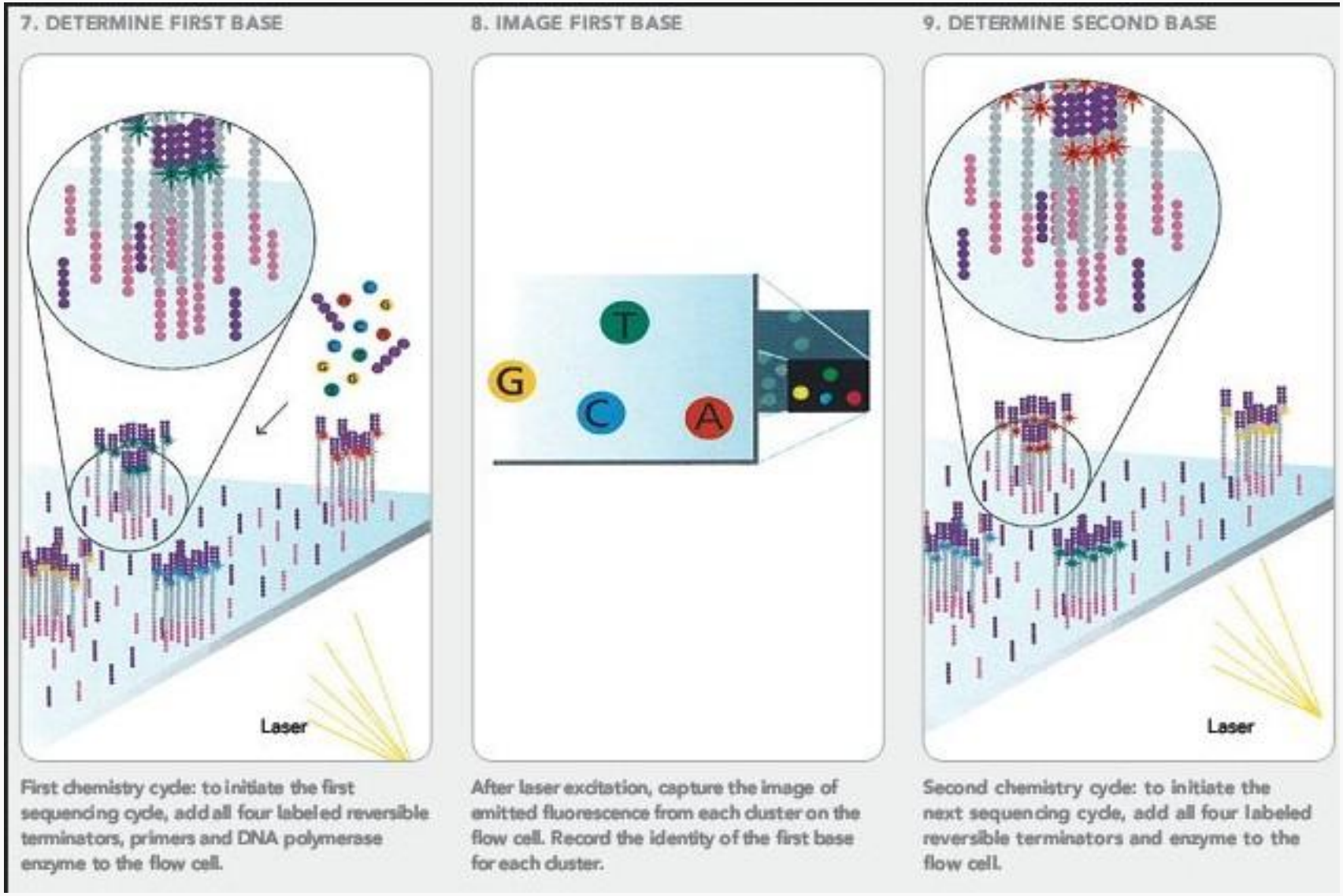
Illumina


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
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.

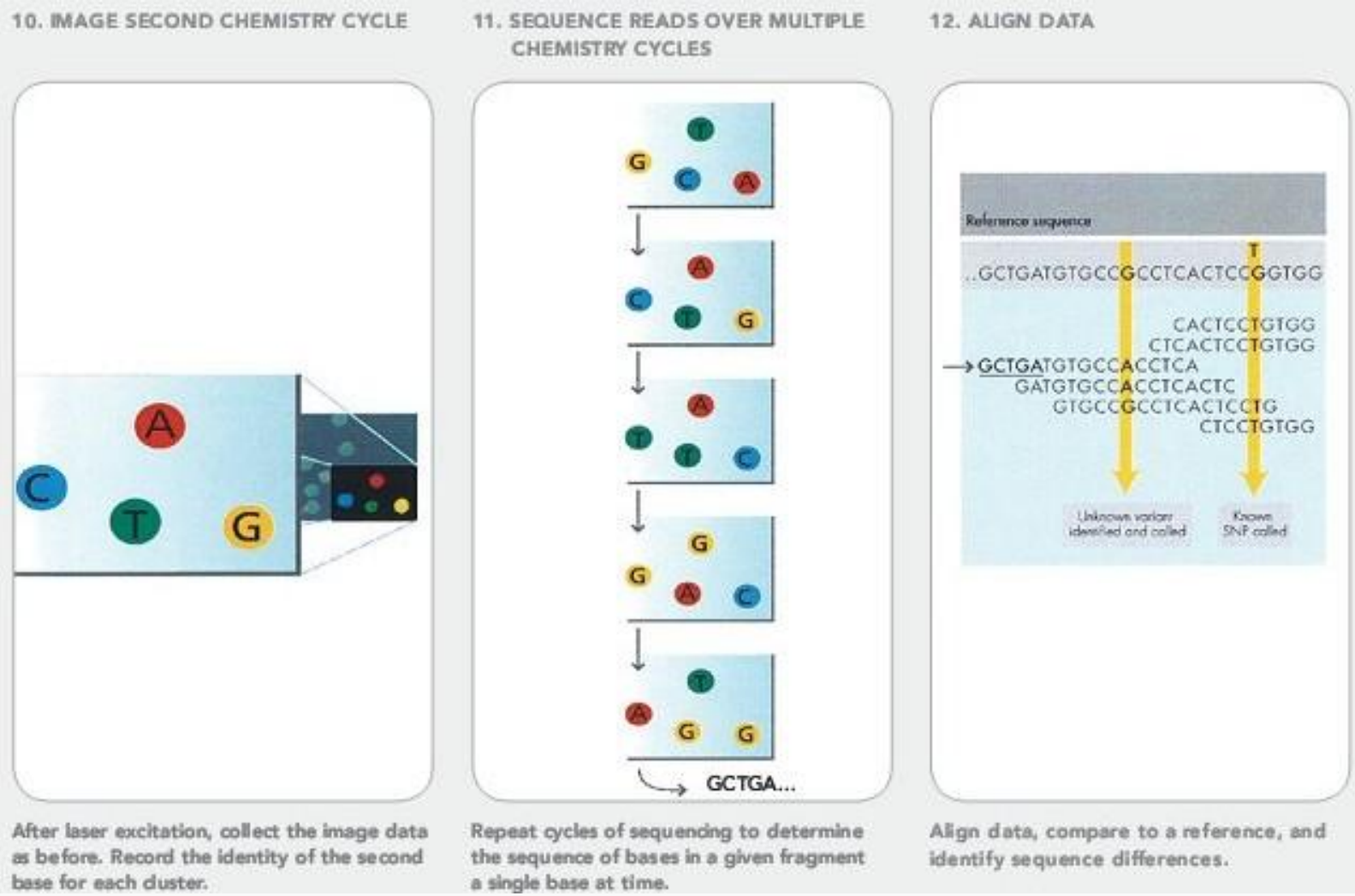
Illumina


4. FRAGMENTS BECOME DOUBLE STRANDED
5. DENATURE THE DOUBLE-STRANDED MOLECULES
6. COMPLETE AMPLIFICATION
Attached terminus
Free terminus
Attached terminus
Attached
Attached
Clusters
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.

Illumina


7. DETERMINE FIRST BASE
8. IMAGE FIRST BASE
9. DETERMINE SECOND BASE
Laser
T
G
C
A
Laser
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.

Illumina


10. IMAGE SECOND CHEMISTRY CYCLE
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
12. ALIGN DATA
A
C
T
G
GCTGA...
Reference sequence
..GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
Align data, compare to a reference, and identify sequence differences.

Illumina - Advantages and Limits

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
CTGTTTCTTAGTATTTTTGTAGTCATTCCGTGTTGGTTTAGTTGCAAGGT
+
@@@DADFFHHFFHIIEFEIGJGGHI4FFIEIGHI<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<HHHIJJIGBHGGEEIIJJEIEIJIHHJFIIJJJGHJJ
@H4:C7C99ACXX:6:1101:1491:94128/2
AGAAGTCTTCGGAAAAGTTCGGGTATGGCTCTAGTAGCTTTTGTCTTAT
+
@C@FFFFFGGHHDHGIIEEHIII<CGHIJIJIIJ:?FC9DGAFGHII?DGBFIIJHBI
@H4:C7C99ACXX:6:1101:1538:34462/2
ACAAAAAGCTAAAAGAACACAGTTGCTTGAAGCAGCAAACACAAGAAC
+
B@@DFFFFGHHHHJIIIIIJJJIIGJCHHEIII>GHIG@GHIDHGJIIFIFHIJJIJJJG
@H4:C7C99ACXX:6:1101:1568:67898/2
ACAAATGGGTGTGTAAGAGTTAAAAAACAATTAATGAGCAACTGAGTTC
+
@@CFFFFFHFFHFGIJJIHHIIJJIIIHIJJJECGHIJJCHGICDGGGHJ<FGGIJJ
@H4:C7C99ACXX:6:1101:1575:18963/2
AACATGTTTGTCGGGGGTTGGGAAATTGTCACTTTCTGCTACAATGCCG
+
@<@DDDDDHFFFFDIIBDFGHHGG;FGGCHHAGGGIIH@E>AEDDEECAB>
```

1 séquence = 4 lignes

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

PHRED SCORE

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

Ewing 1998

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

or

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

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

The QPHRED Value

Phred Quality Score

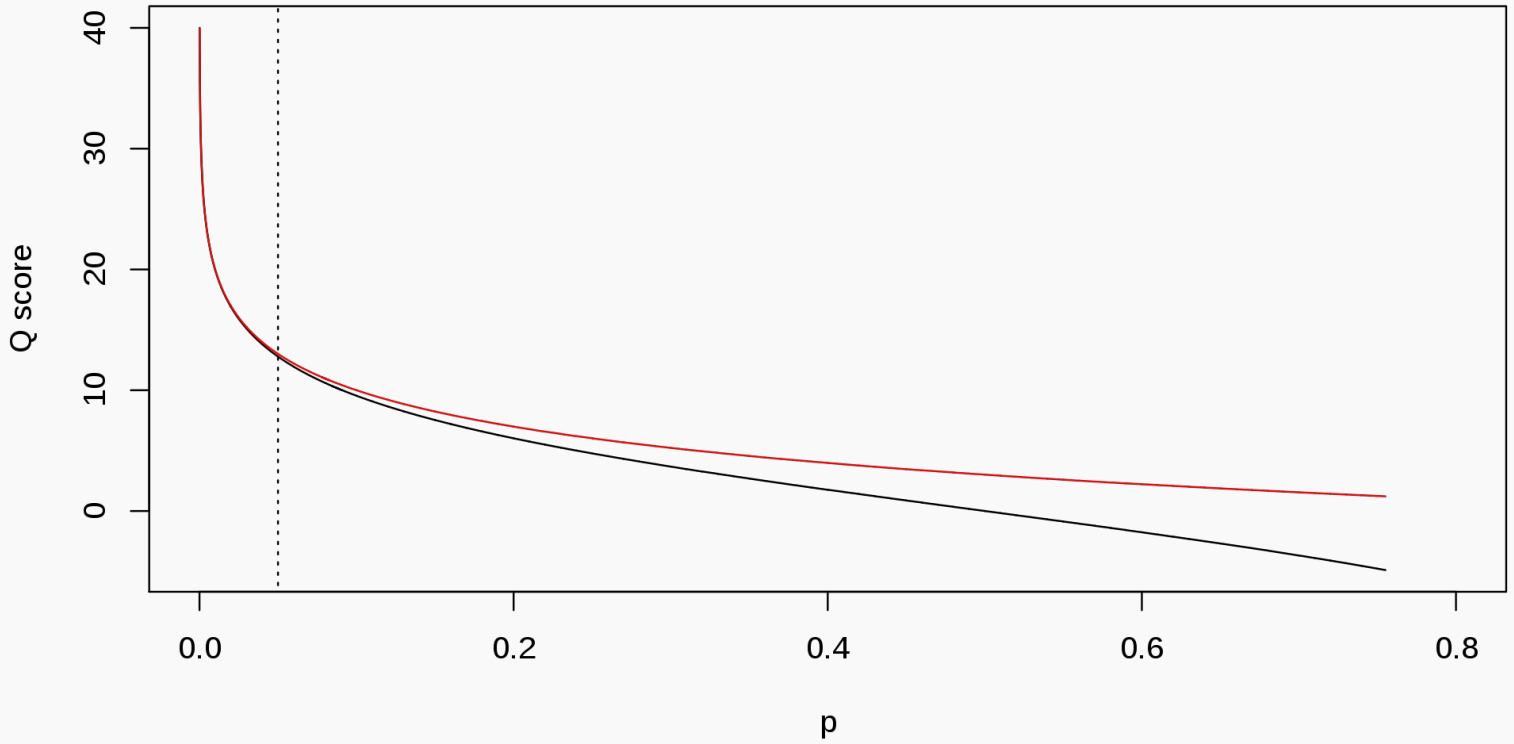
0 .. 50

Code ASCII

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

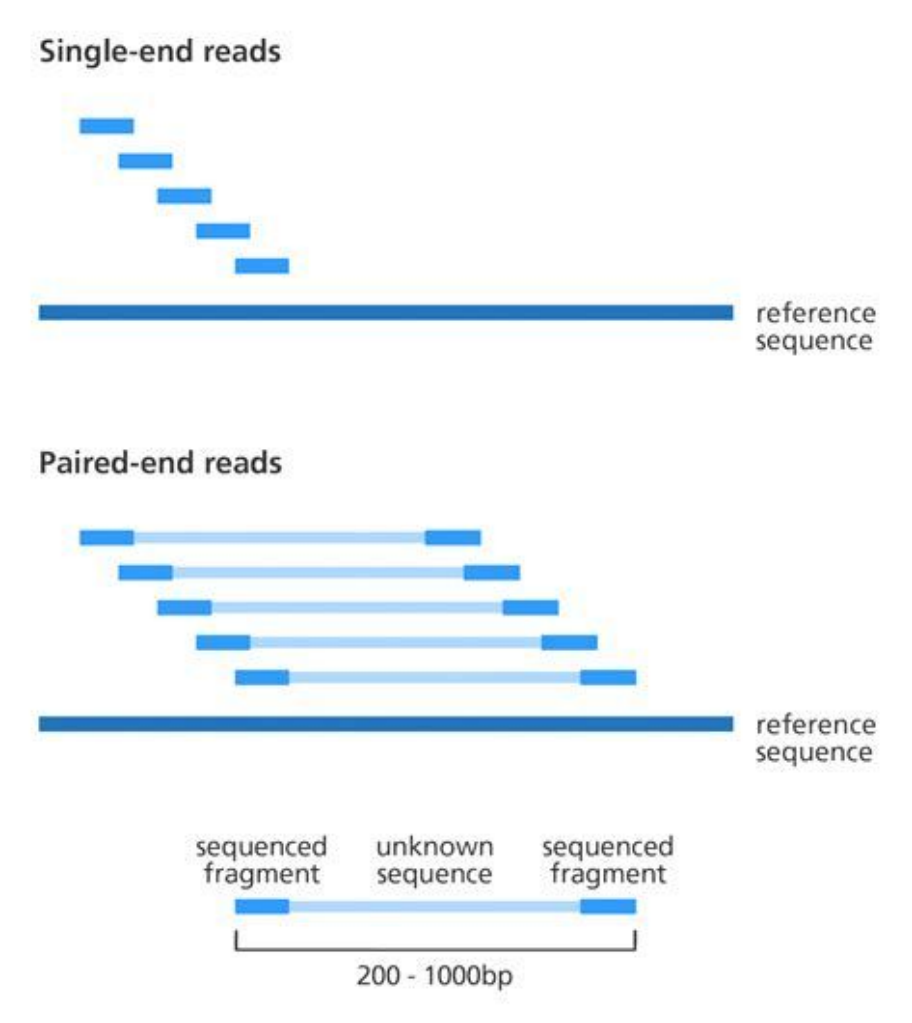
charstable.com

The QPHRED Value



The QPHRED Scale

Mapping


Single-end reads
reference sequence
Paired-end reads
reference sequence
sequenced fragment
unknown sequence
sequenced fragment
200 - 1000bp

SAM Format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAA
r003   0 ref  9 30 5H6M       *  0   0 AGCT
r004   0 ref 16 30 6M14N5M    *  0   0 ATAG
r003  16 ref 29 30 6H5M       *  0   0 TAGG
r001  83 ref 37 30 9M         =  7 -39 CAGC
```

Header

- Ligne commençant par @

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
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

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

alignement

Format tabulé

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

SAM Format

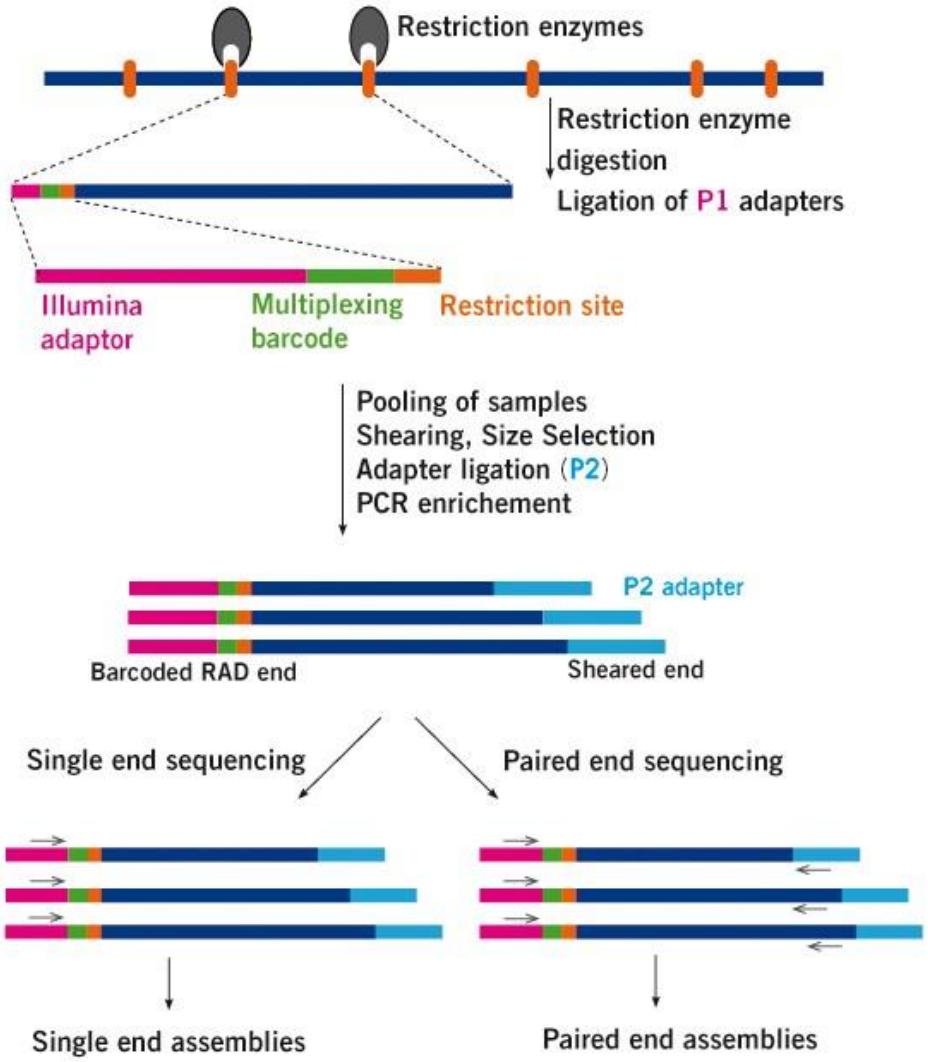
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC
r003  16 ref 29 30 6H5M  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

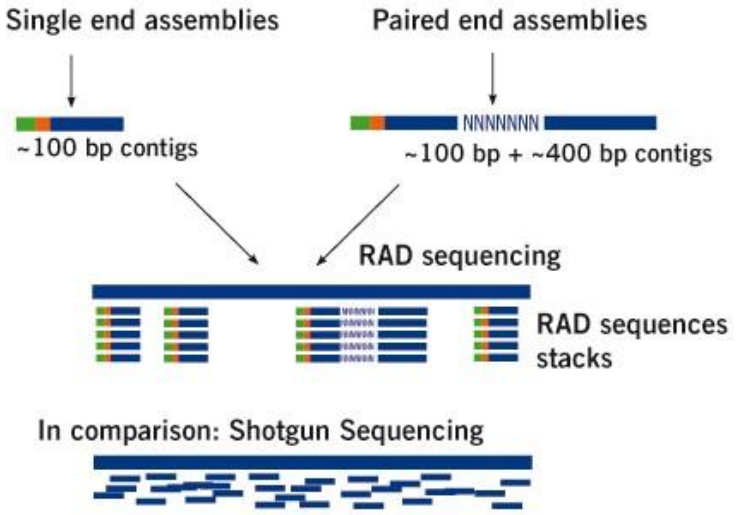
alignement

Format tabulé

Col	Name	Description
1	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)

RAD-Seq





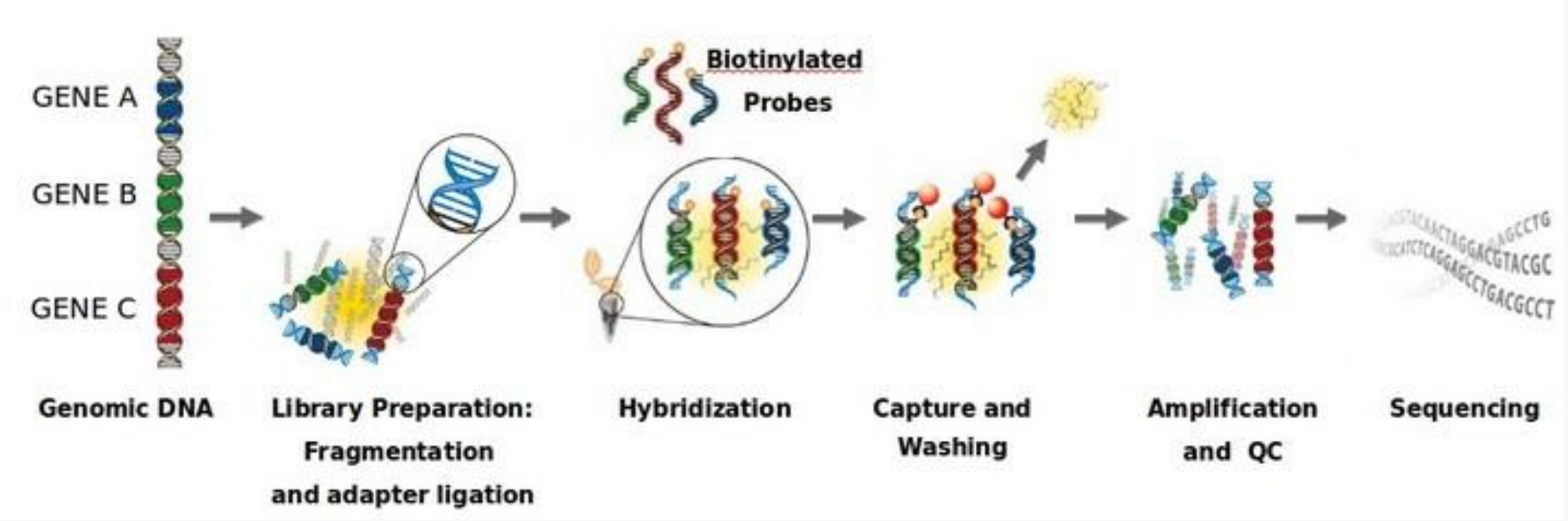
DNA is digested with a specific set of restriction enzymes and two adaptors ligated to the fragments afterwards.

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.

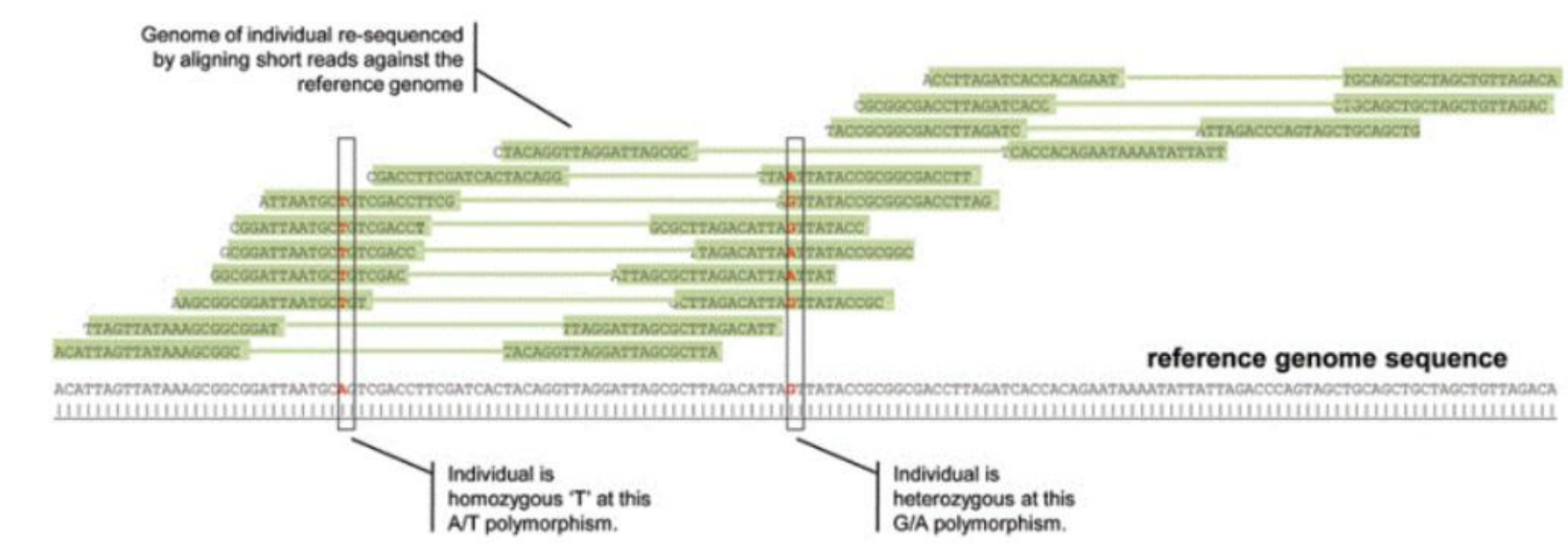
From Eurofins

CaptureSeq



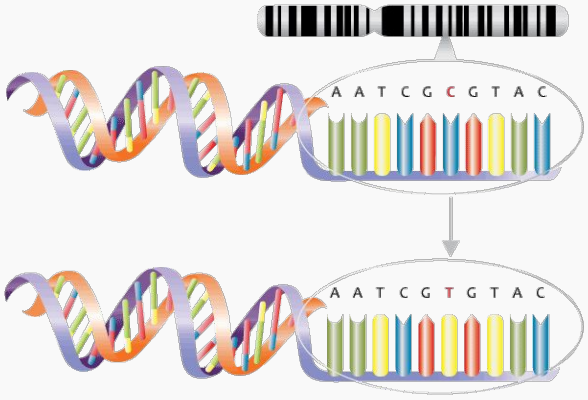
From CGFB, Bordeaux, France

SNP and InDel Detection



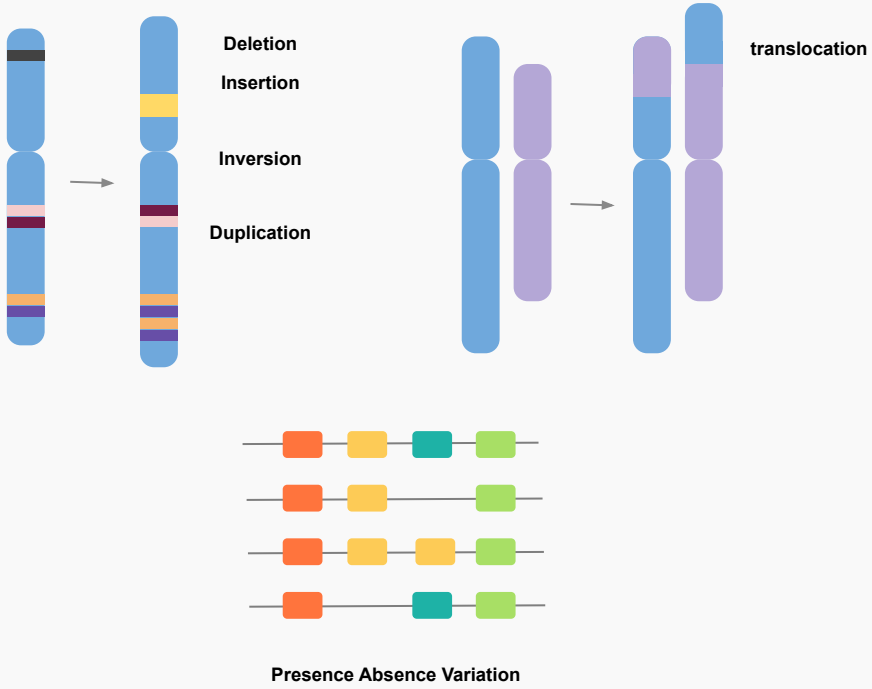
Mutations & Variations as main source of genetic diversity

SNP

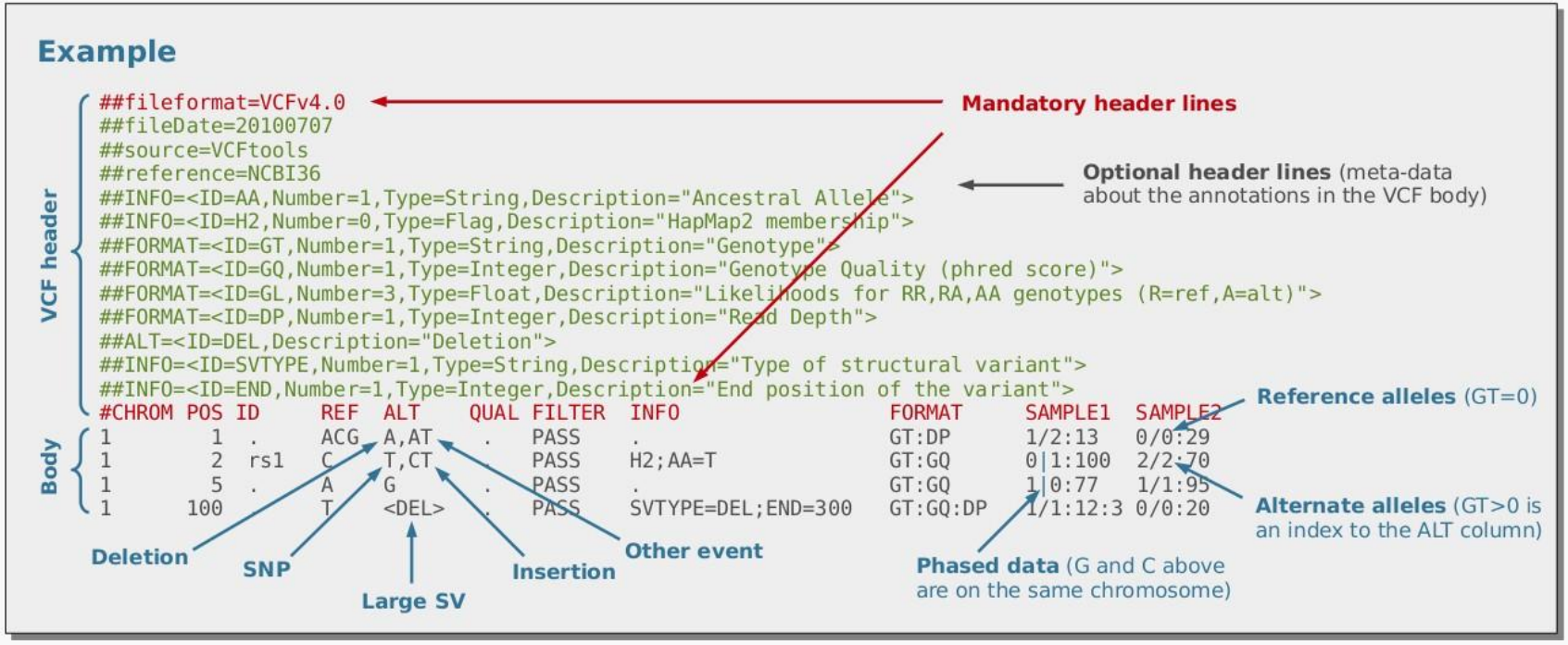


From NHS National Genetics and Genomics Education Centre, CC BY 2.0, via Wikimedia Commons

Structural Variations



Common File for all Variations, the VCF



VCF = Variant Call Format From 1000 Genomes Project

Limits, Wetlab

- 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

Limits, Technical

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

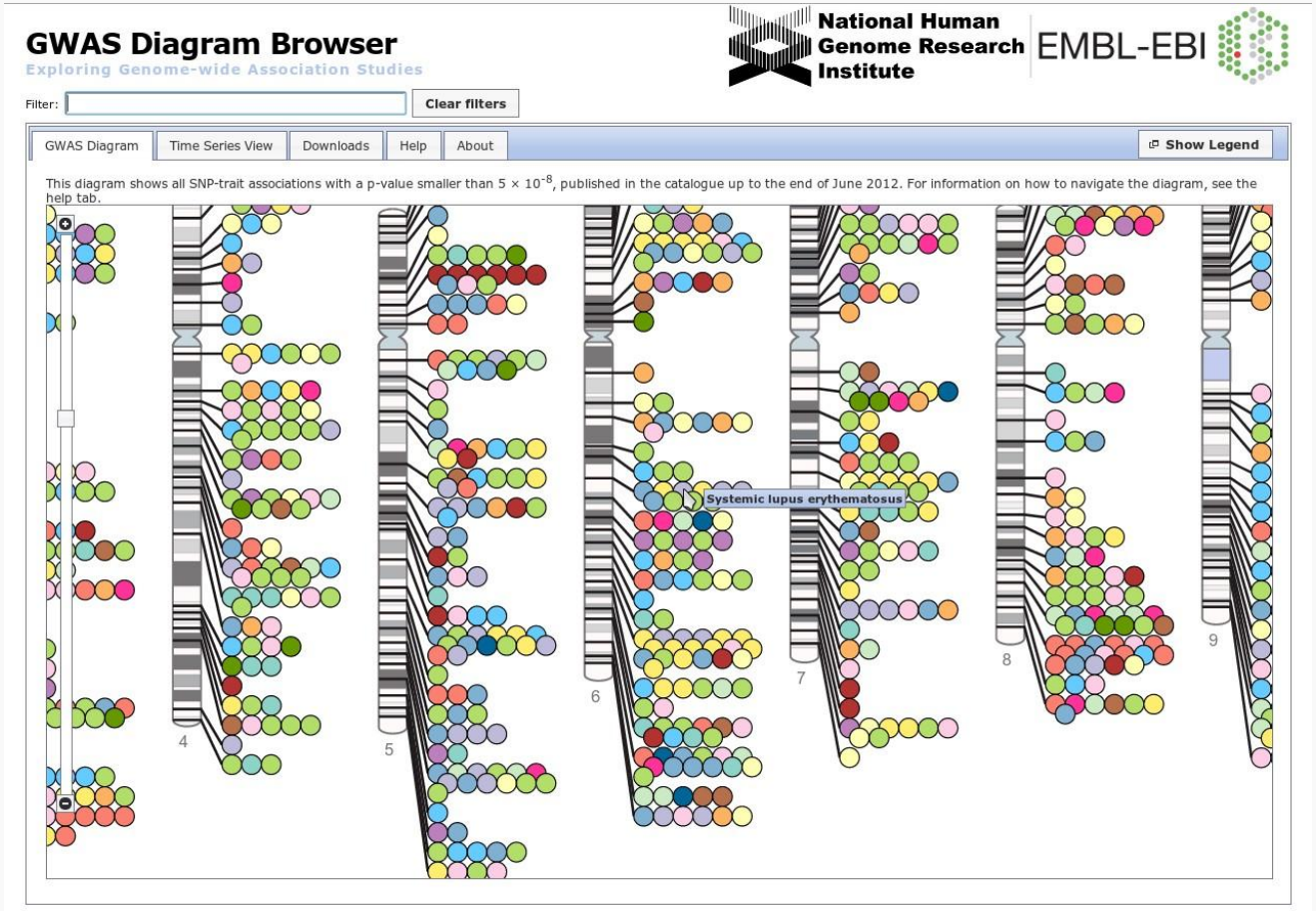
All linked to the Specificity/Sensitivity Informatics Paradox

Applications

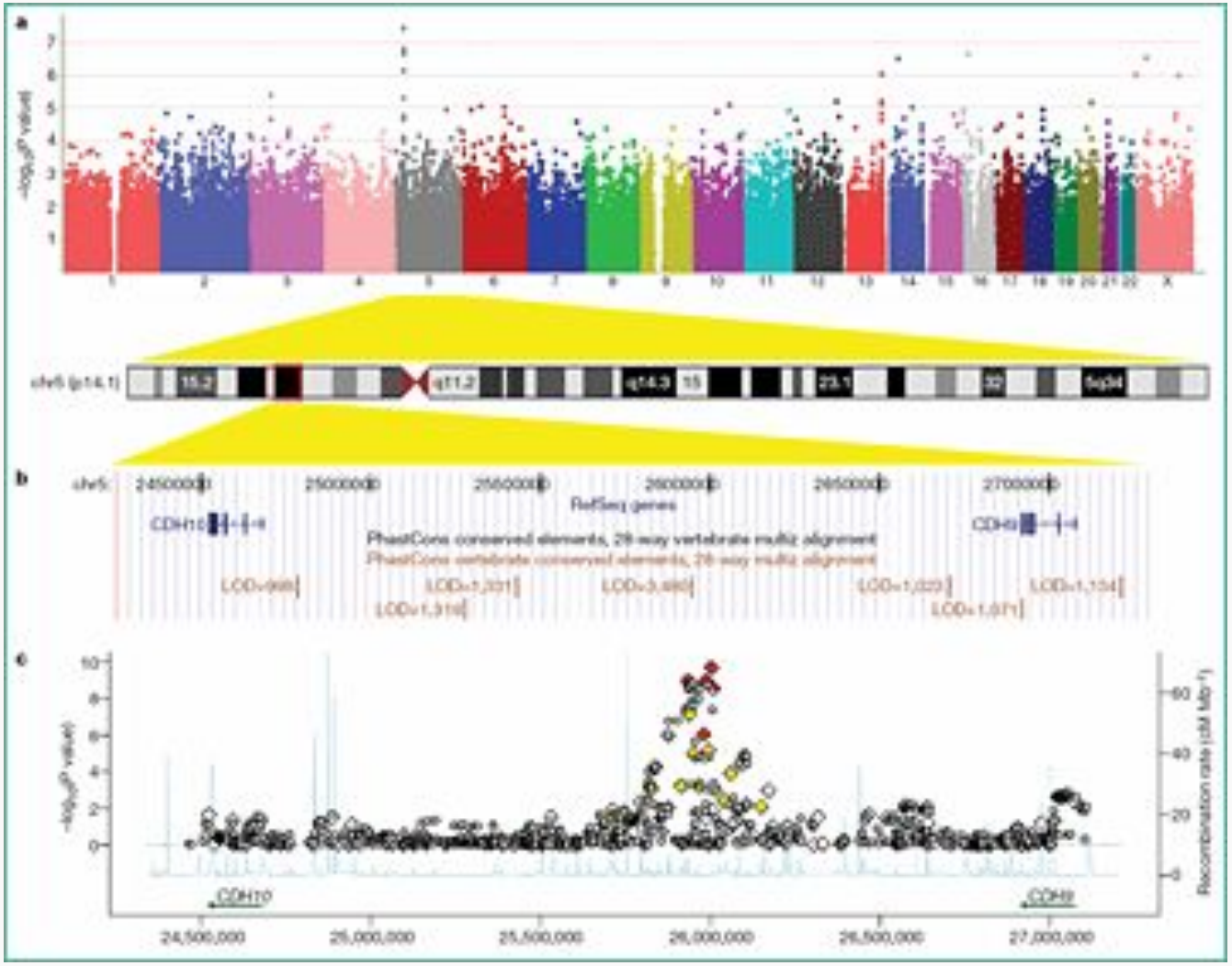
Population Genomics/Genomics

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

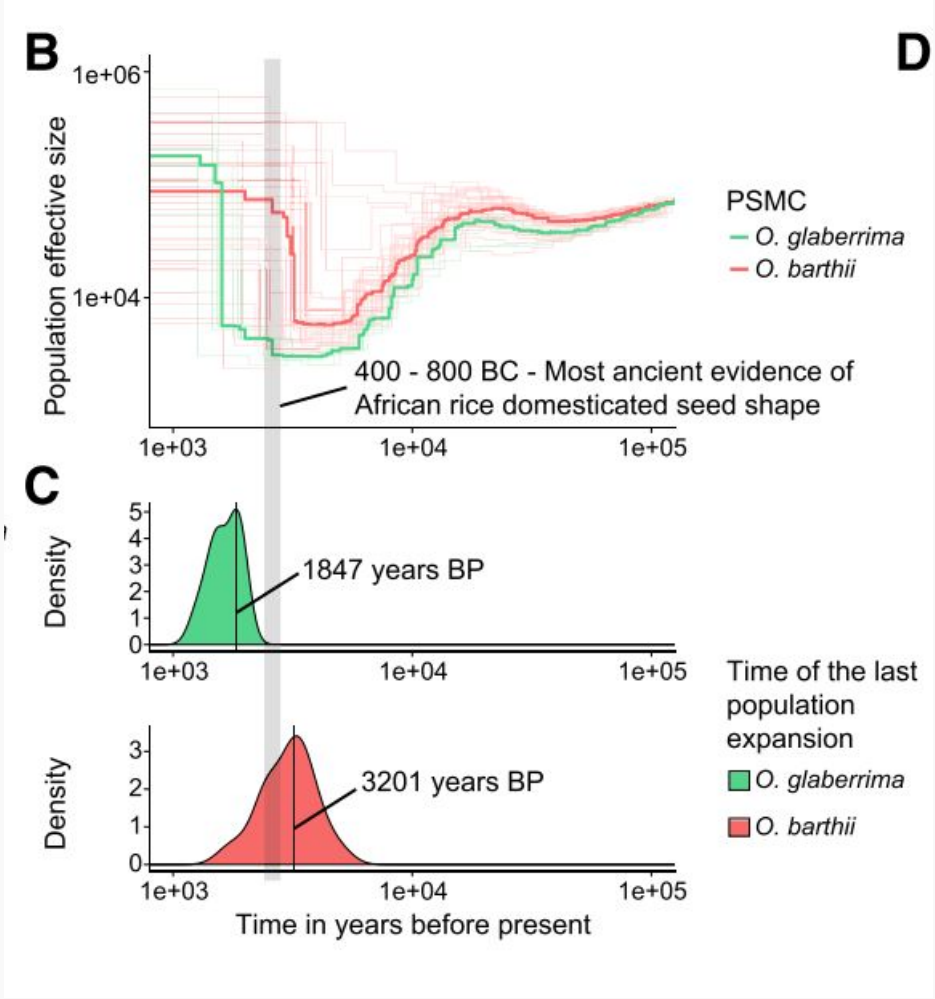
Example in GWAs & Population Genomics

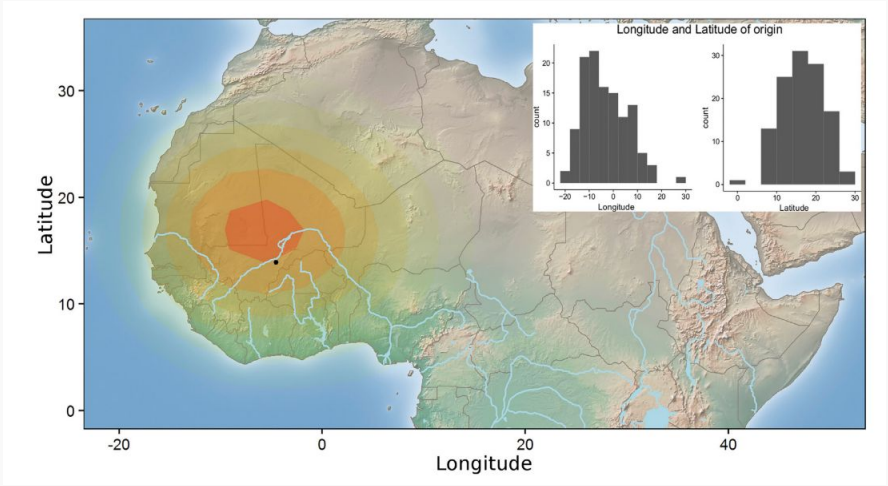


Example in GWAs & Population Genomics



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, 2018

Large Projects

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

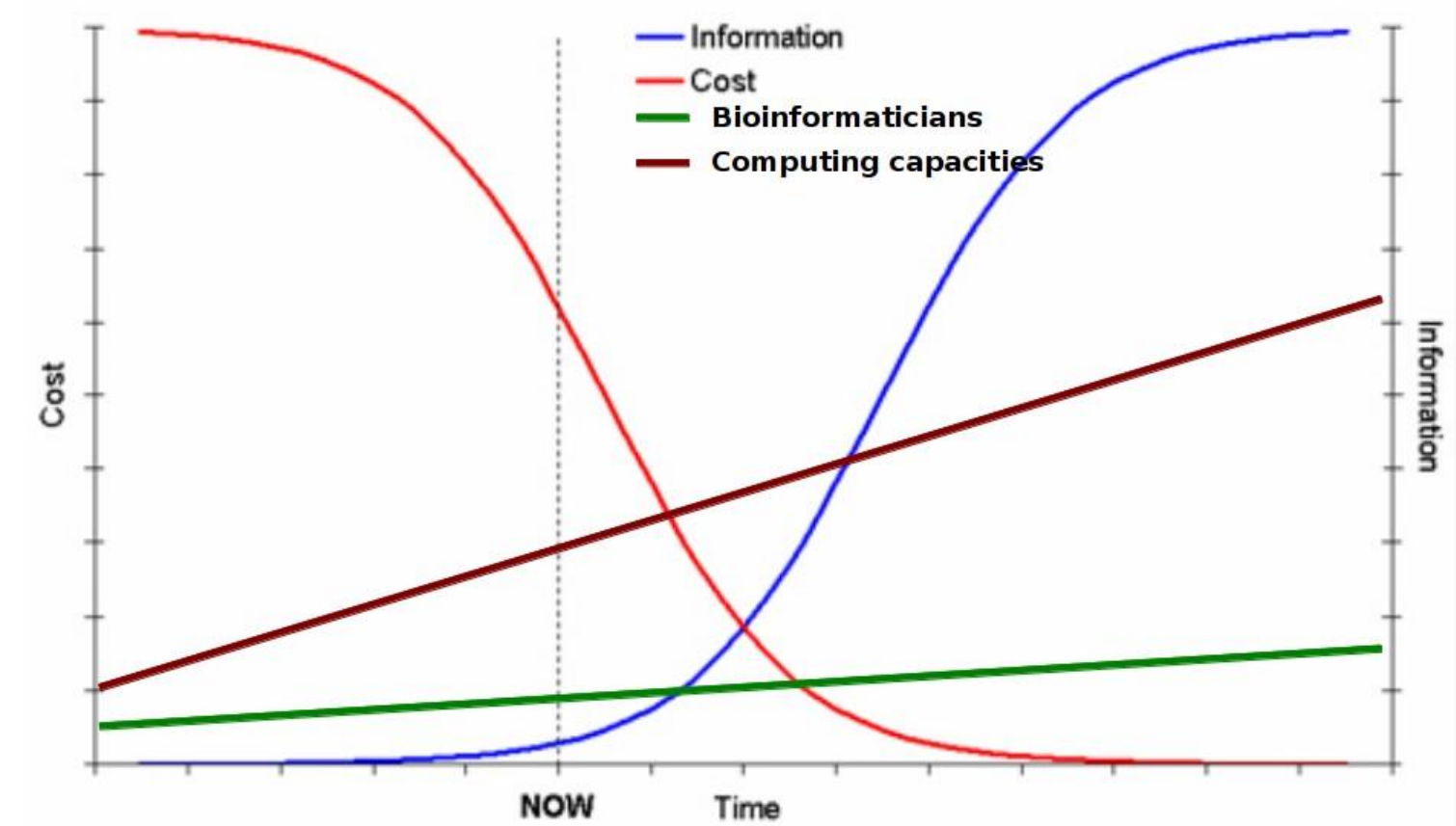
Possibilities in the next 5-10 years (From a presentation in 2013)

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

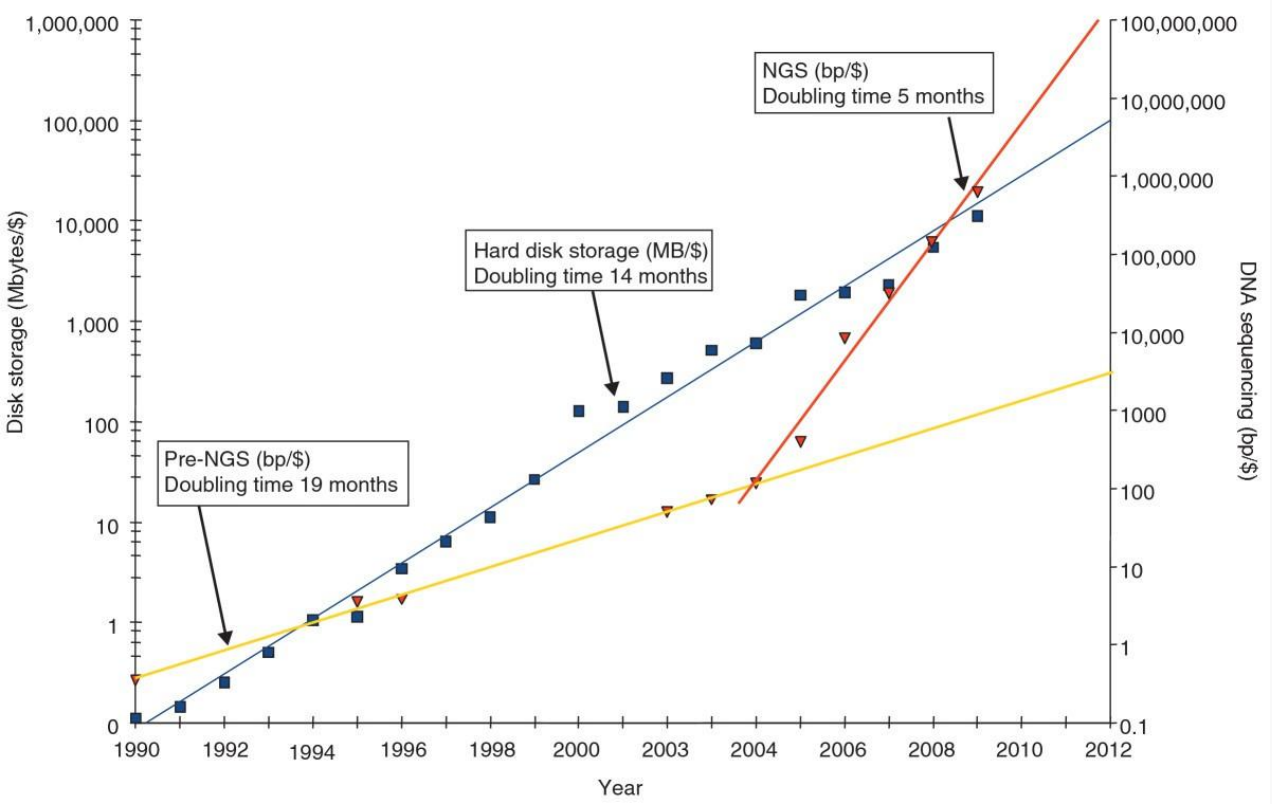
Possibilities in the next 5-10 years (From a presentation in 2013)

- 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!

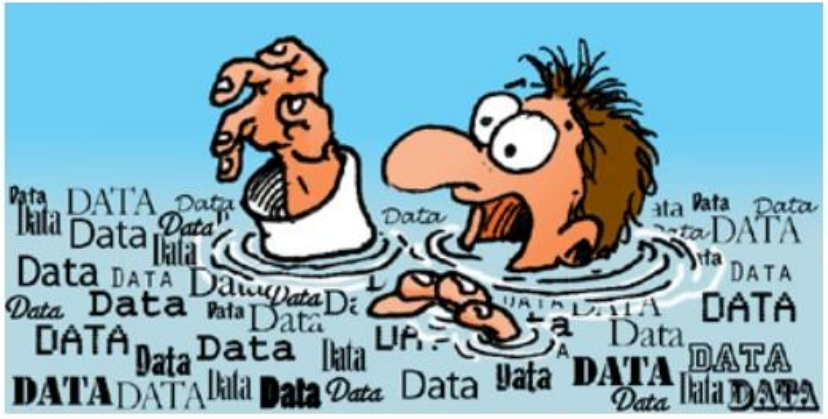

Information
Cost
Bioinformaticians
Computing capacities
Cost
Information
NOW
Time

...From Data Rarity to Data Deluge

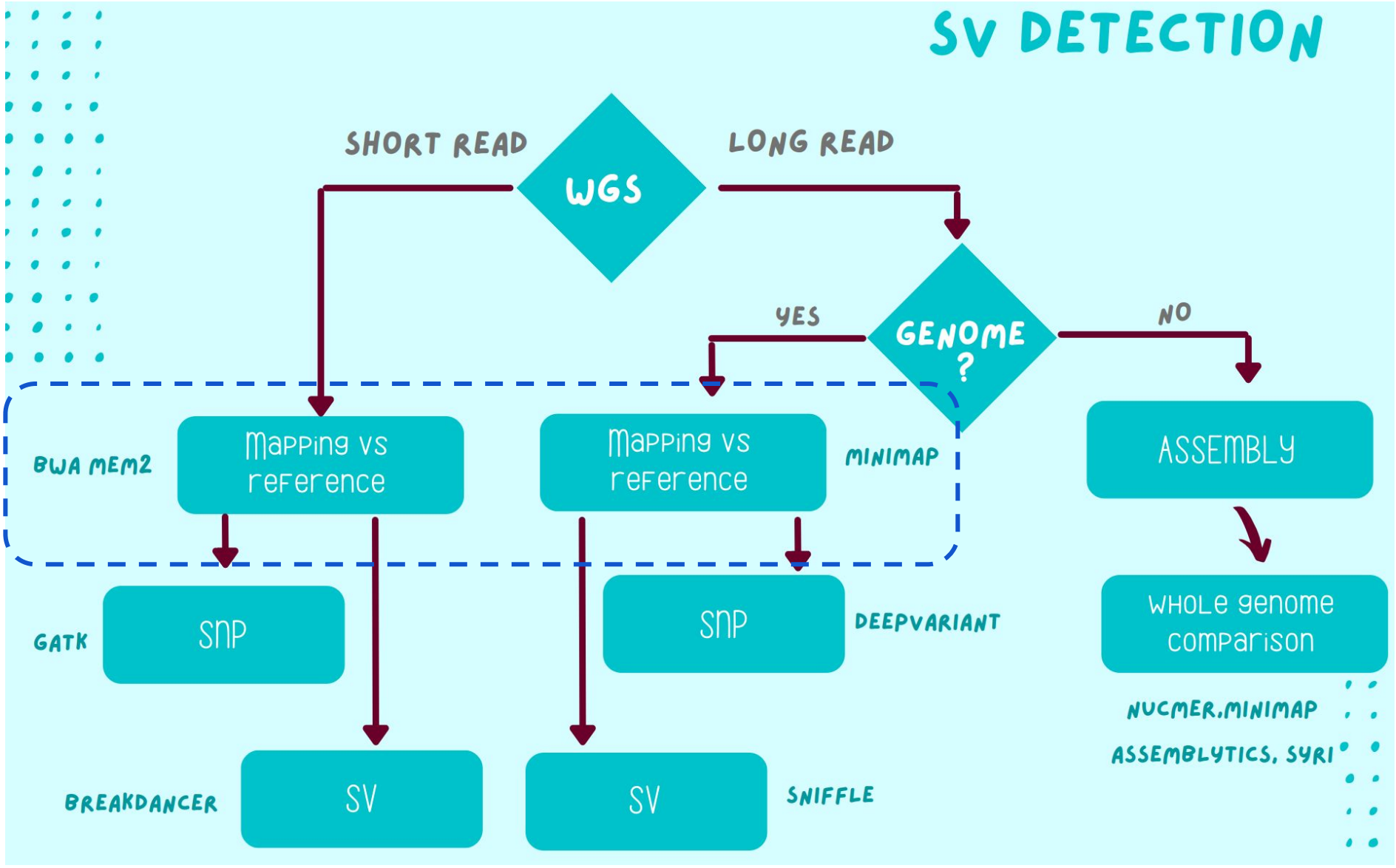


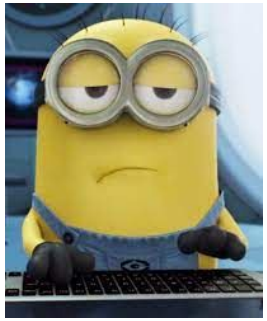
From L. Stein, 2010

Be Careful to data drowning!



Training plan - day 1





Mapping and SNP

TP goal

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

The data

Diploid Asian Rice, *Oryza*

From
Wikimedia

The data

Diploid Asian Rice, *Oryza*

From
Wikimedia

1. Select/Cut 1 Mb on Chromosome 10

The data

Diploid Asian Rice, *Oryza*

From Wikimedia

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

The data

Diploid Asian Rice, *Oryza*

From Wikimedia

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

The data

Diploid Asian Rice, *Oryza*

From Wikimedia

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

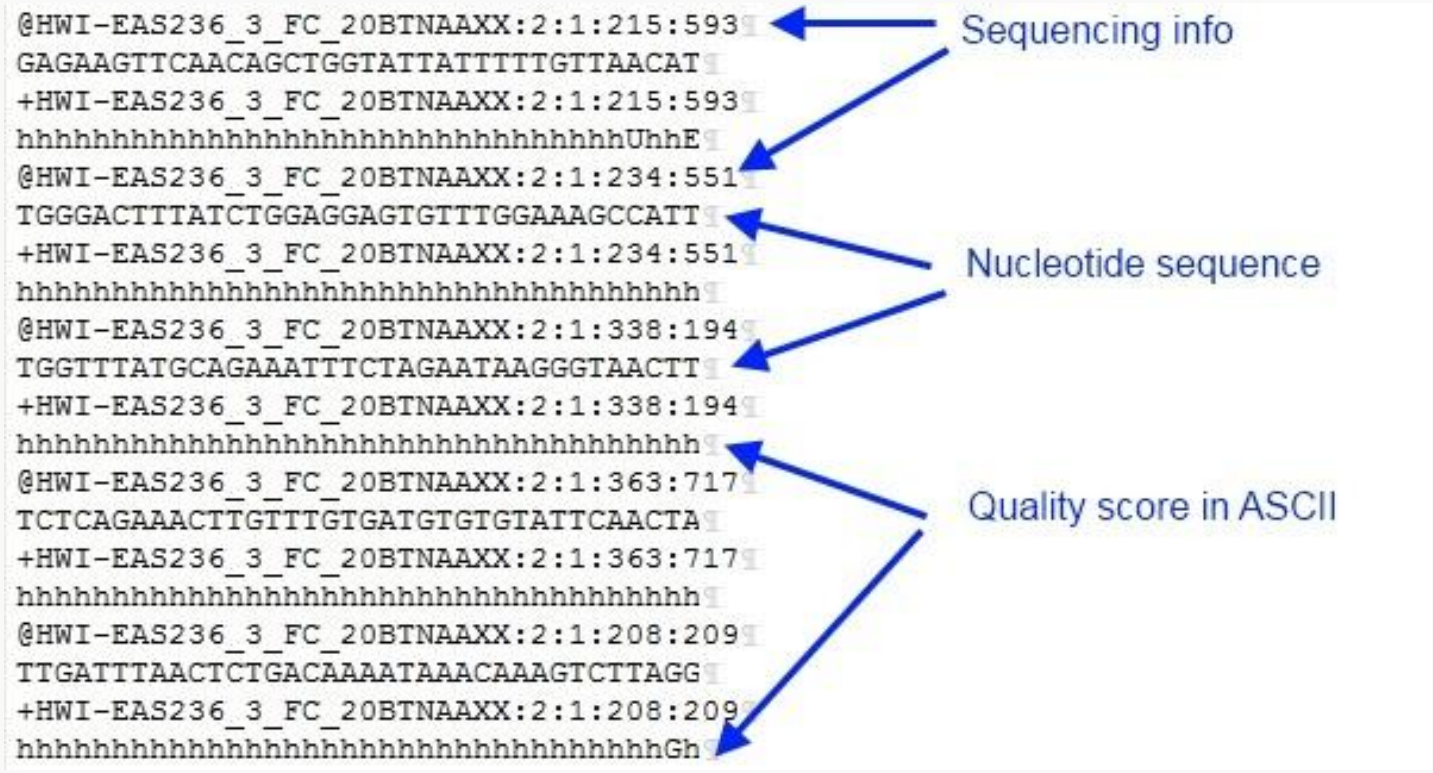
The data

Diploid Asian Rice, *Oryza*

From Wikimedia

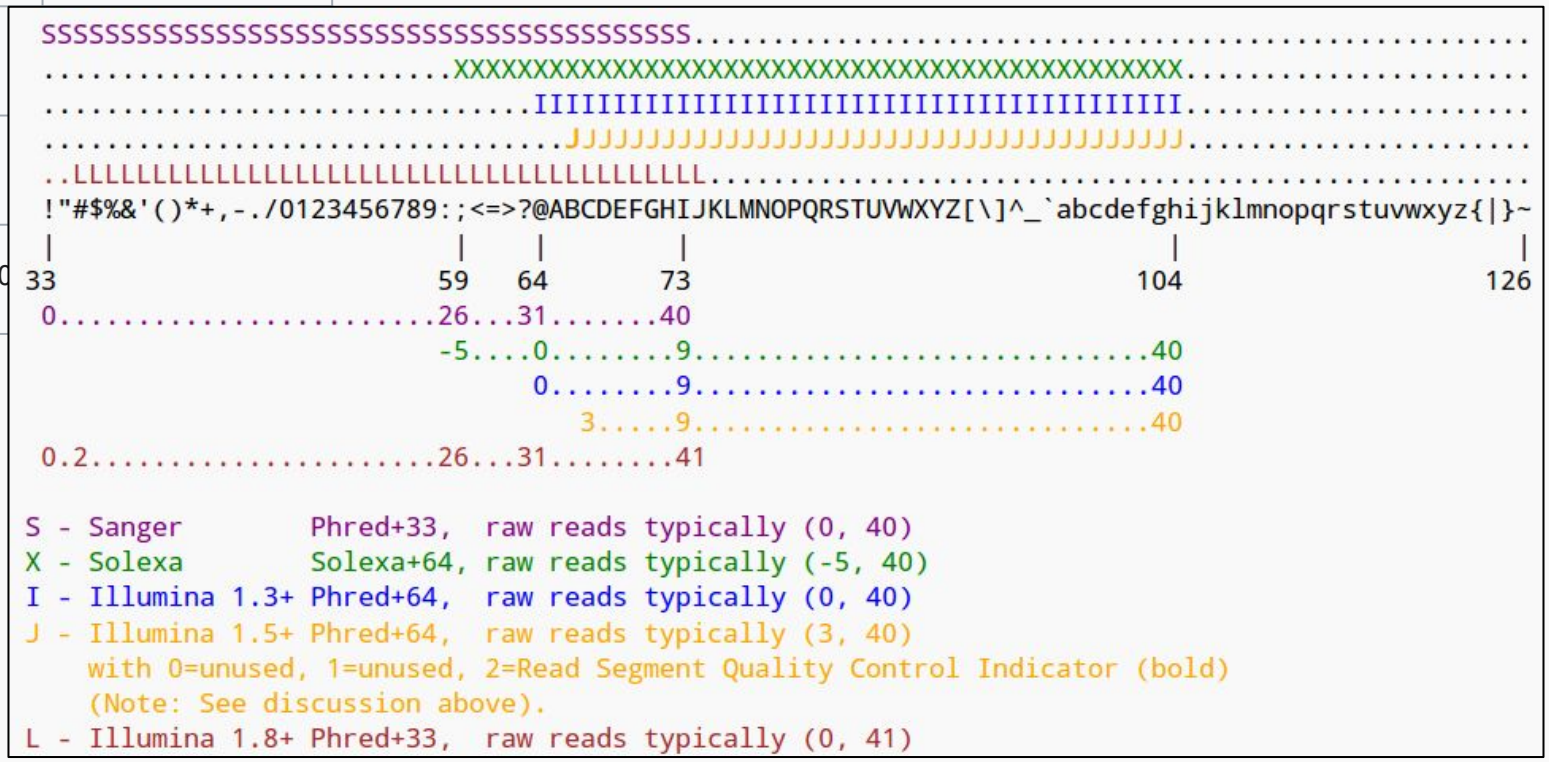
1. Select/Cut 1 Mb on Chromosome 10
2. Create 20 exact clones
3. Introduce
    - SNP (1-10%),
    - indel (10b-10kb),
    - duplications...
4. Generate short & long reads for each clone...
5. ~~Torturing students with these data~~

The FASTQ Format



The 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	
40	1 pour 10000	
50	1 pour 100000	



List of analyses

"Classic" launch

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

List of analyses

"Classic" launch

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

List of analyses

"Classic" launch

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

List of analyses

"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,...

List of analyses

"Classic" launch

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

Between 8 and 15 different commands...

Let’s do it by hands...

We will

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

Let's do it by hands...

We will

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

SAM file format for Sequence Alignment Map

SAM format :

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

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
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 file format for Sequence Alignment Map

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         *
```

the SAM bitwise FLAG

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

From J. Aerts, in SlideShare

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

Let’s do it by hands...

We will

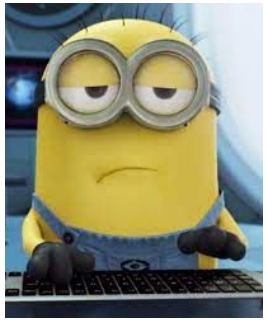
1. Map the data of Clone 1 on *reference.fasta* using *bwa*
2. Look at the SAM file
3. Compress SAM in BAM and reorder it
4. Remove wrong mapping

Practice

Let’s work with the jupyter book :

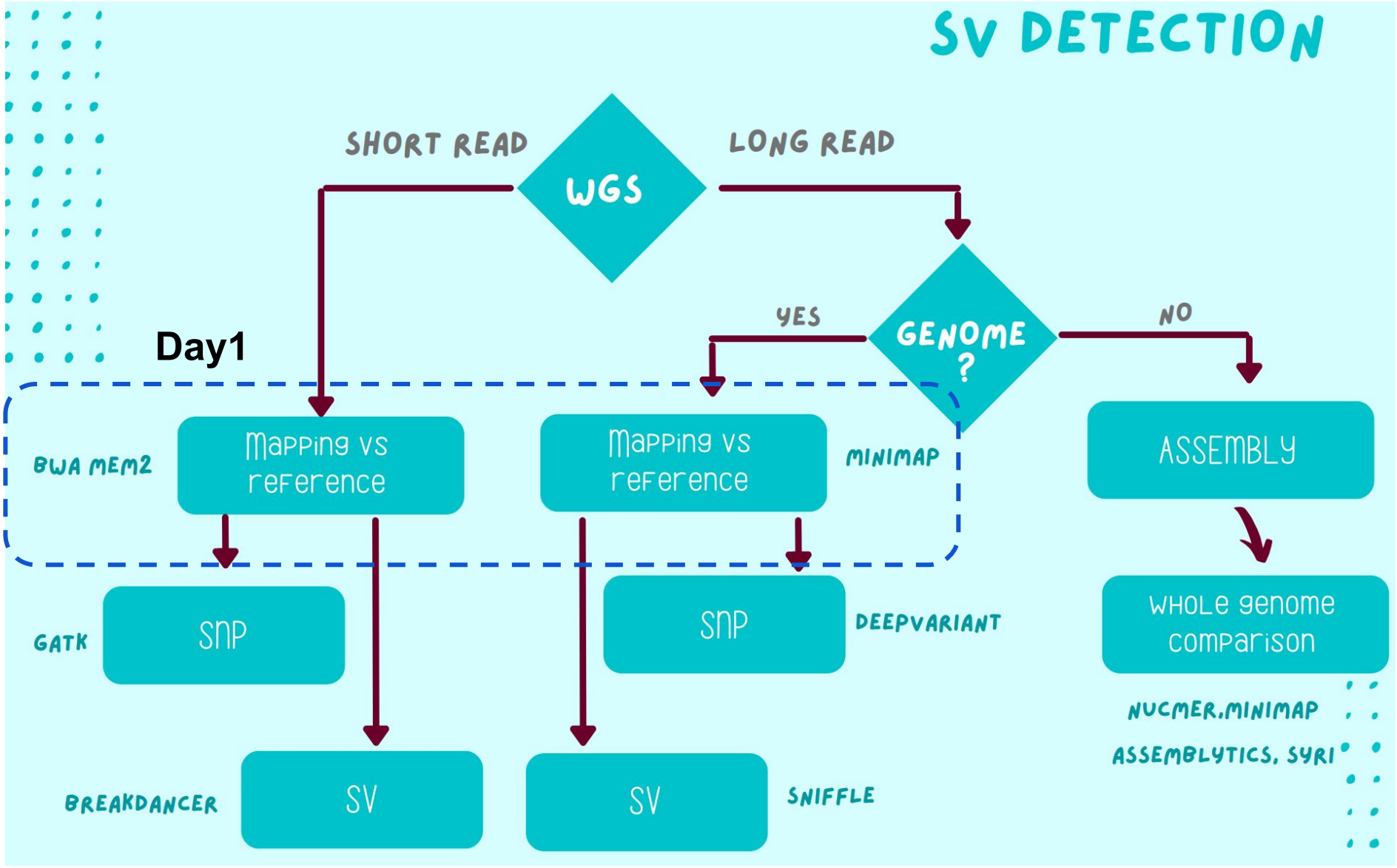
Day1_Mapping_Practice_EMPTY.ipynb


THAT'S ENOUGH
FOR NOW

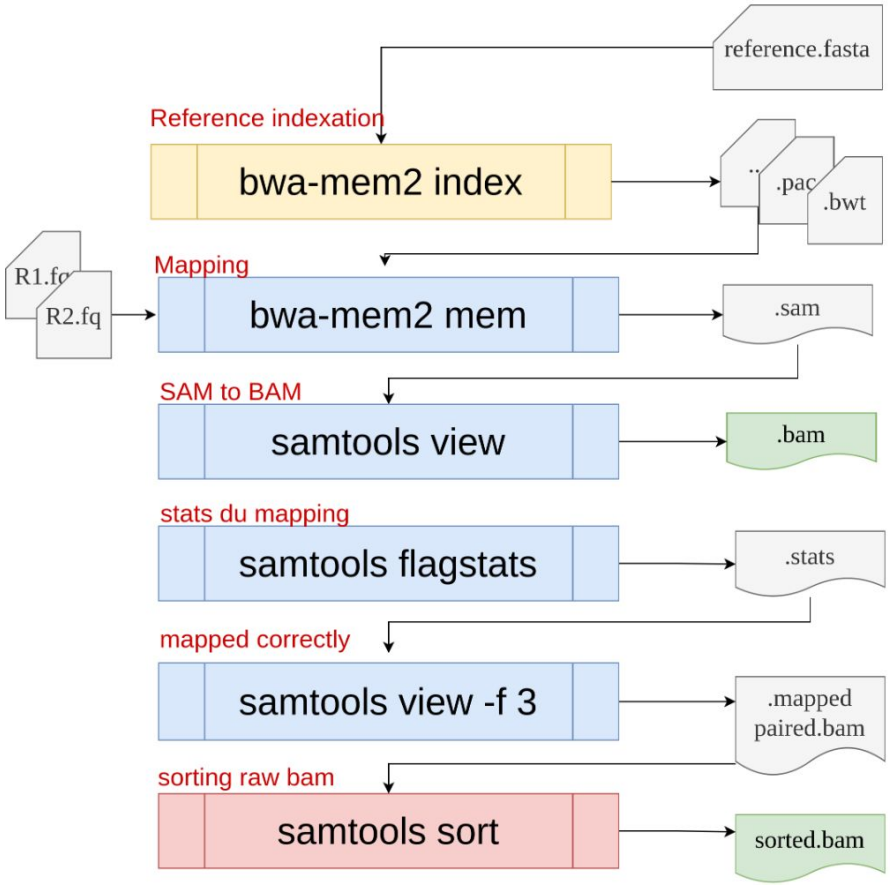


Day2 - Let’s start slowly

Training plan - day 1 & 2 (a little)

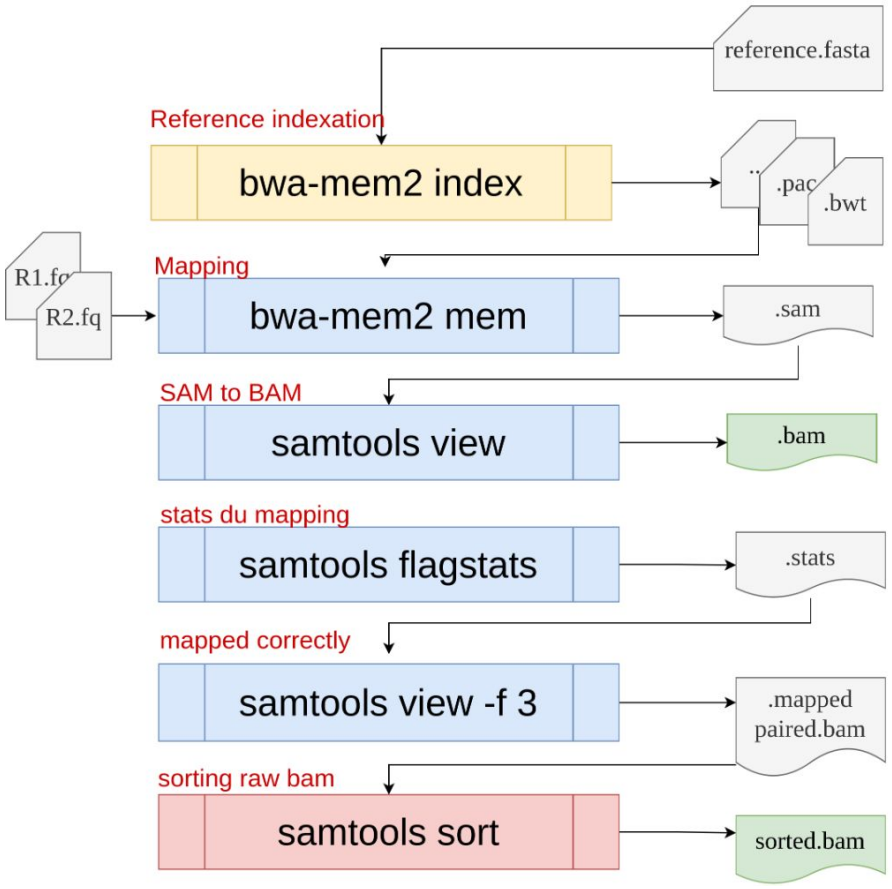


Plan de bataille du mapping !

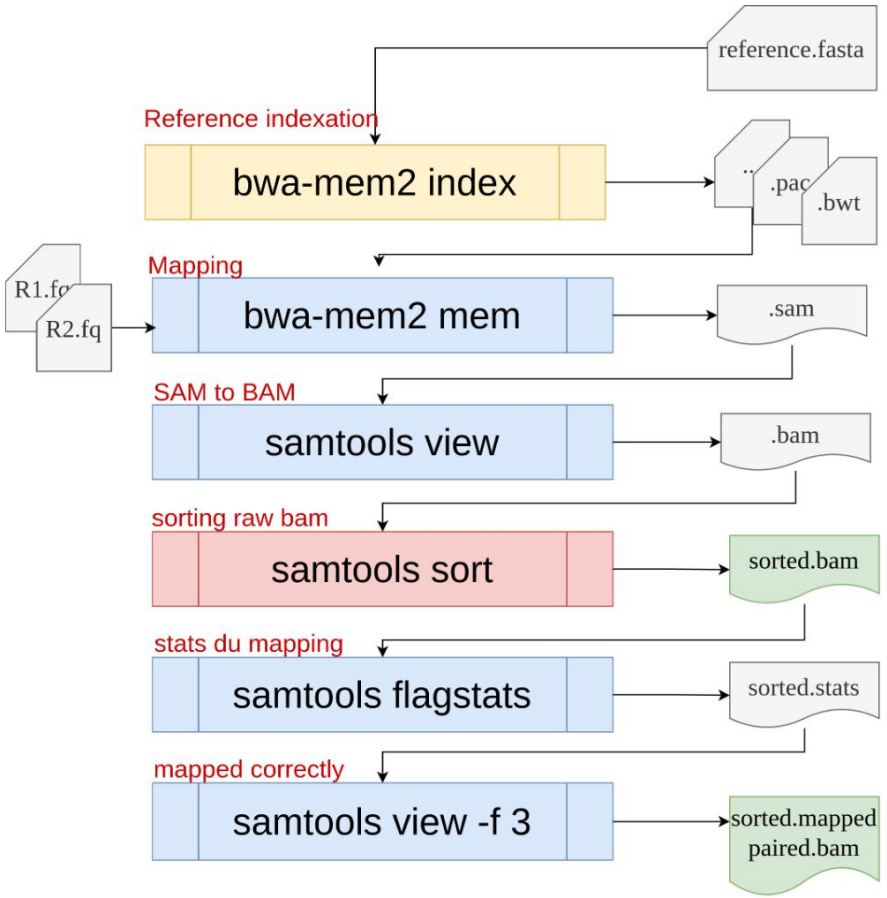


En cours Day1

Plan de bataille du mapping !



En cours Day1



ça marche mieux !

Plan de bataille du mapping - bwa-mem2 !

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

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

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

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

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

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

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

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

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

    done
```

Plan de bataille du mapping - minimap2 !

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

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

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

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

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

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

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

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

    done
```

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

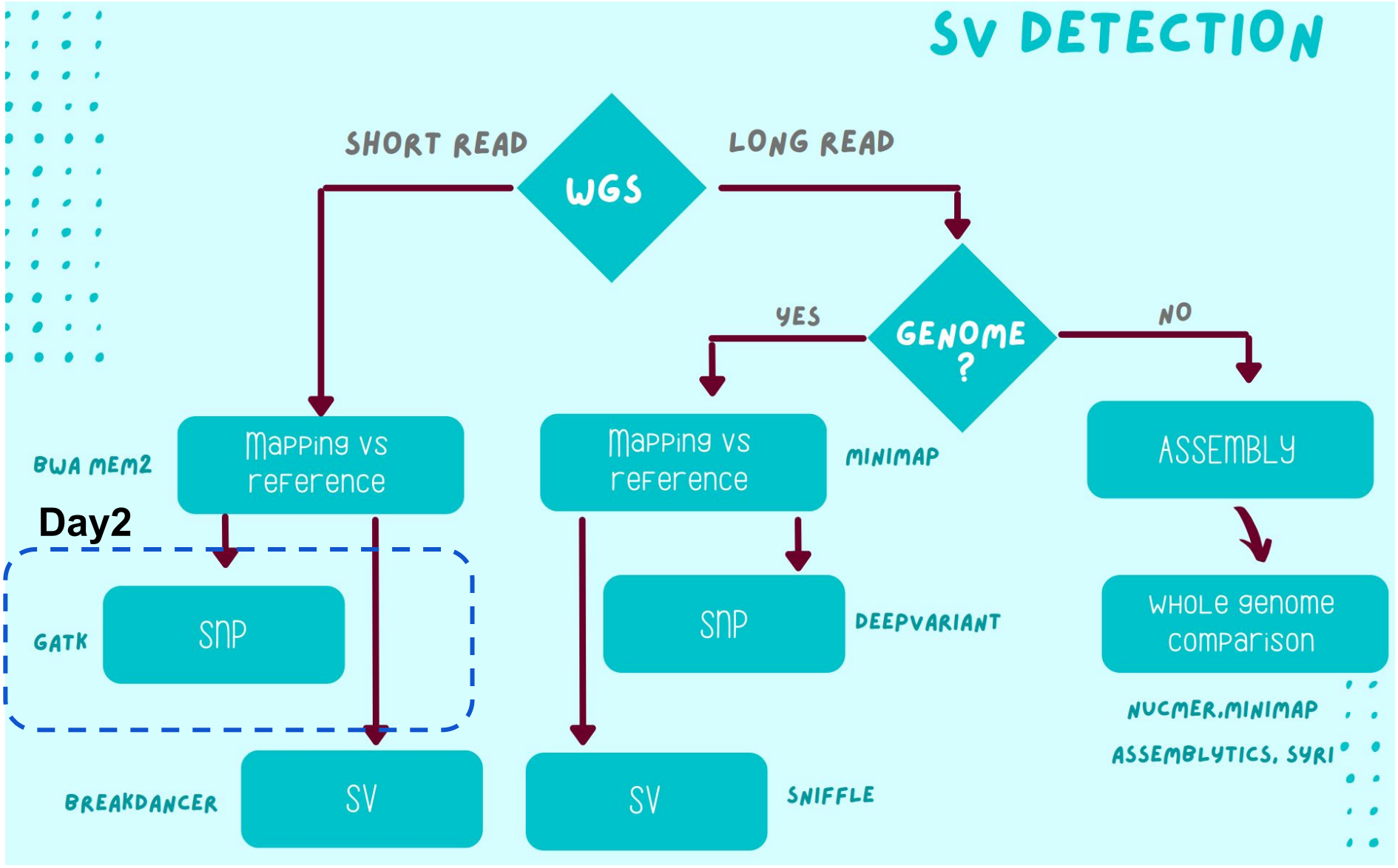
- Download Tablet (use Google and Tablet+NGS)
- Transfer the BAM from Clone 10 and the reference from the machine to your local computer (use scp or direct download from the browser)
- Open Tablet, load an assembly
- Look at the mapping and try to find SNPs

Practice

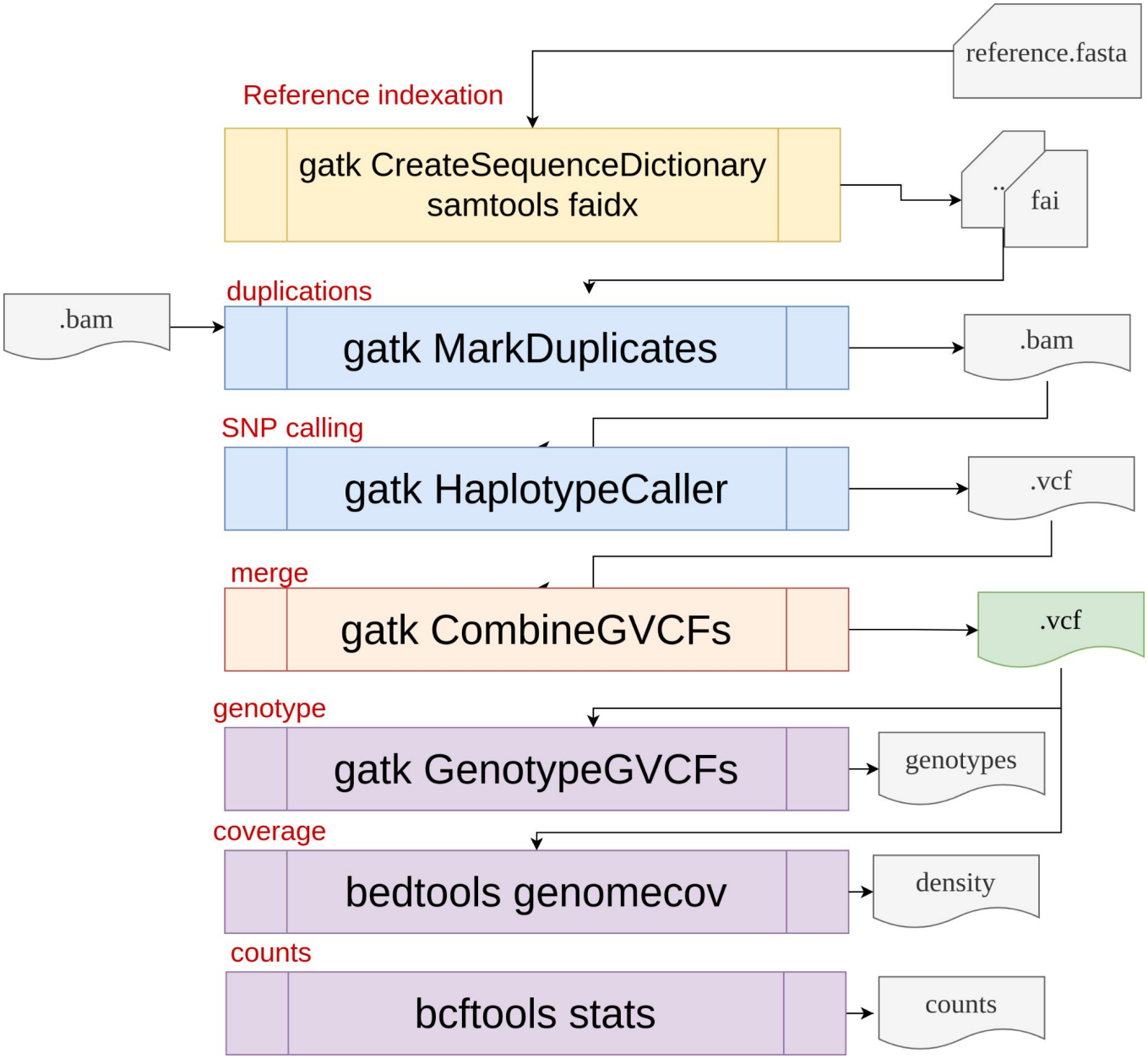
Let’s work with the jupyter book :

Day2a_Mapping_analysis-EMPTY.ipynb

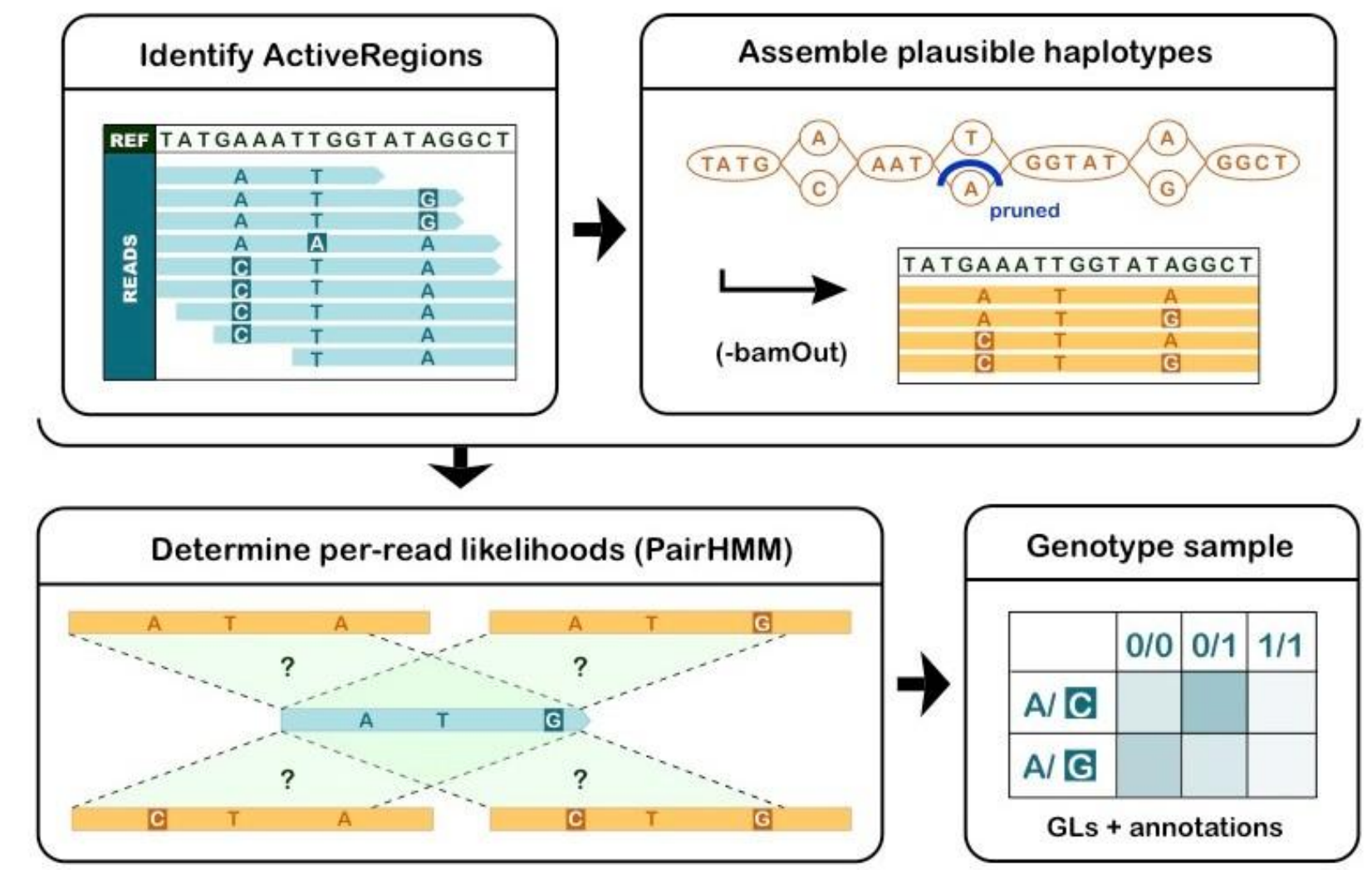
Training plan - day 2



Plan de bataille du “SNP calling” !



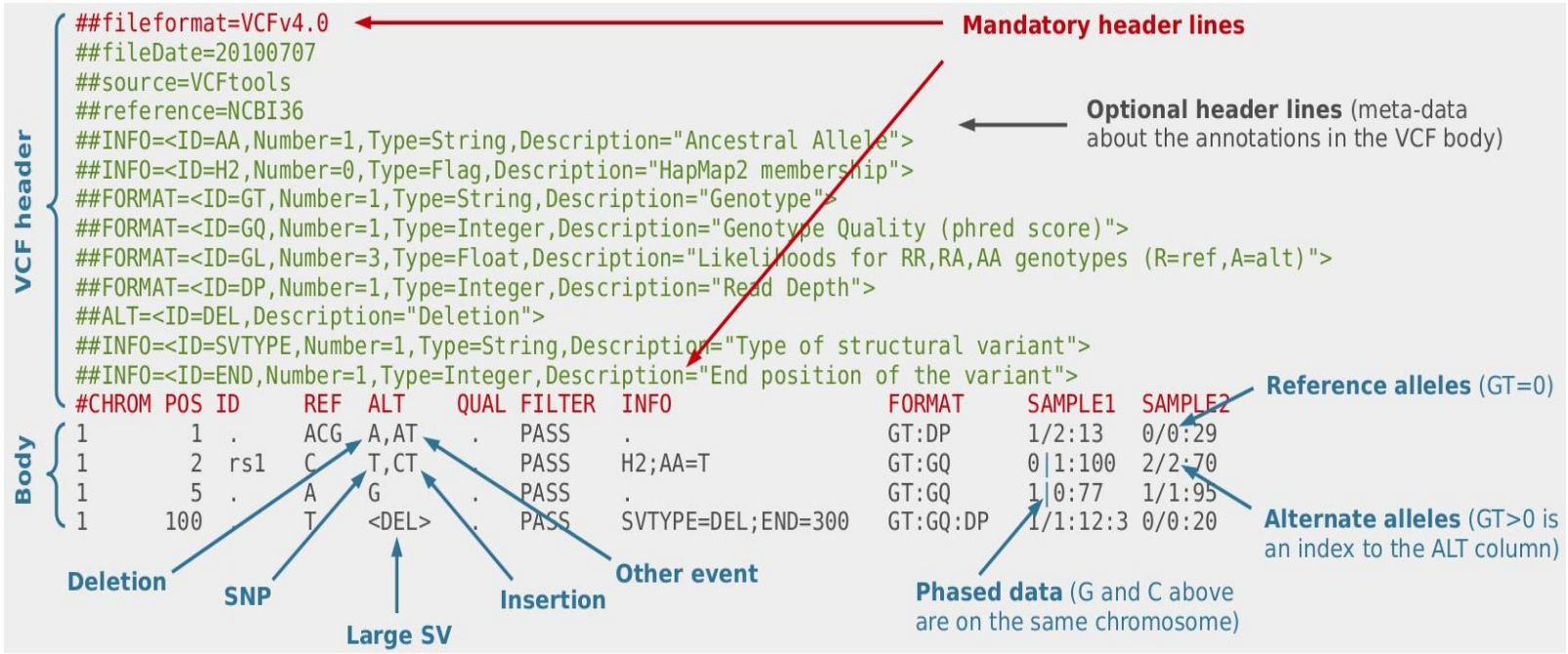
The HaplotypeCaller



From UniBe.ch
Training

VCF file format for Variant Call Format

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



Practice

Let’s work with the jupyter book :

Day2b_Appel_variants_SNP_EMPTY.ipynb

Plan de bataille du “SNP filtering” !

Basics filtering

- Using GATK Variant Filtration, a flag per filter

Basics filtering

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

Be “lazy” but clever

Problems with manual launches

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

Be “lazy” but clever

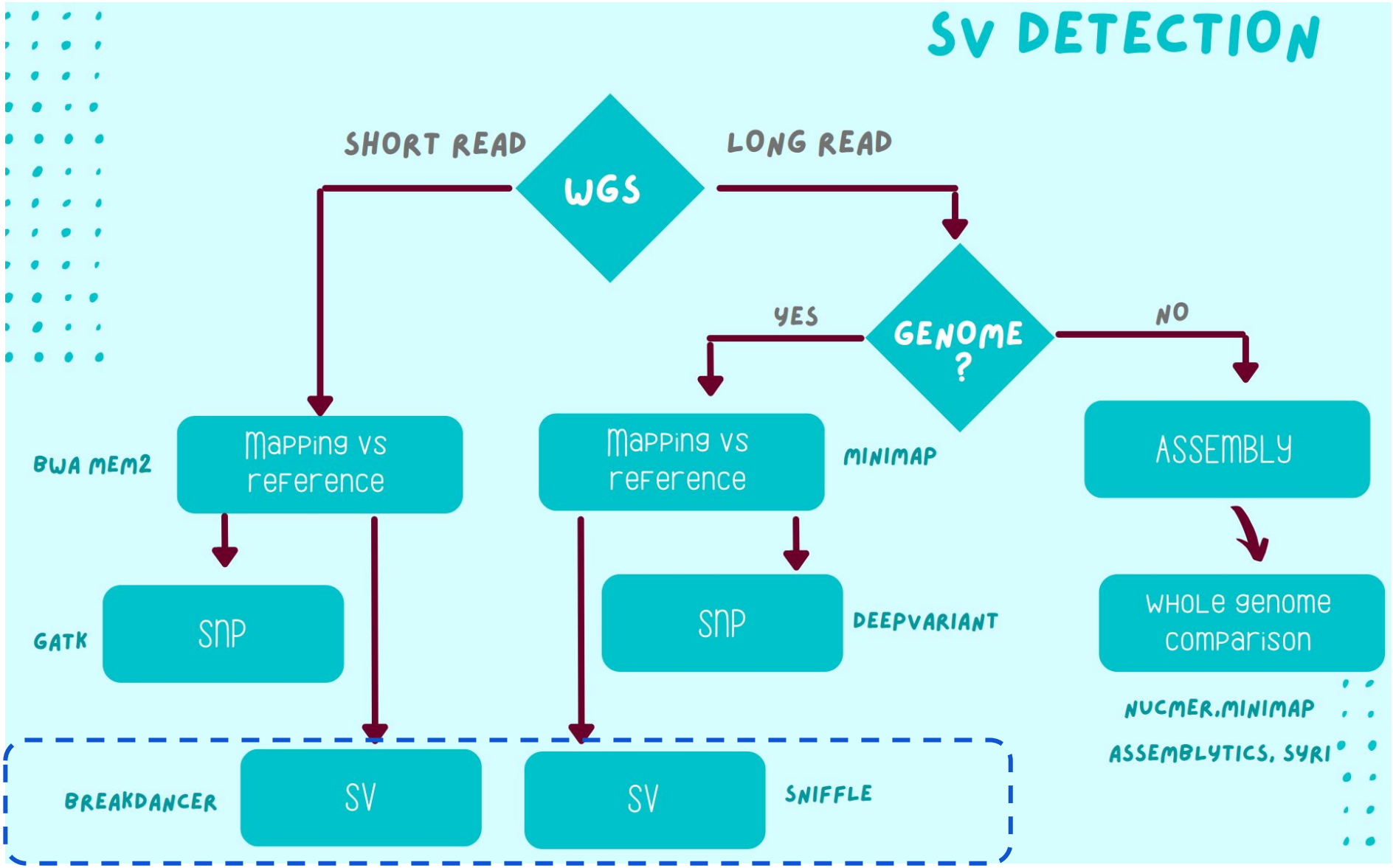
Problems with manual launches

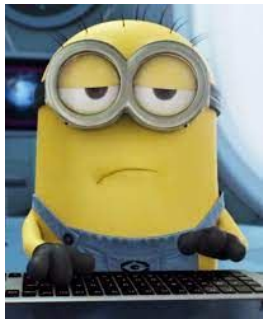
- Long
- Fastidious
- Error prone
- Tracability and reproducibilty not ensured

Solution $\Longrightarrow$ Workflow Manager (or scripts...):
SnakeMake, TOGGLe, NextFlow


THAT'S ENOUGH
FOR NOW

Training plan - day 3





Long Read Mapping

Long Reads

Two technologies

Oxford Nanopore

MinION

GridION

PromethION

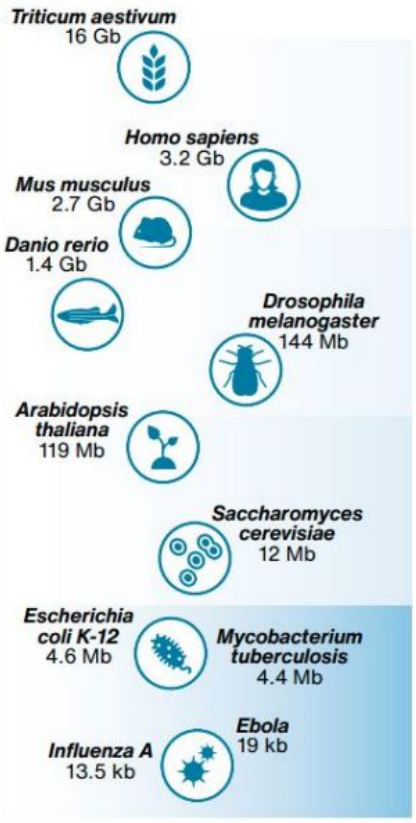
Pacific BioScience

RSII

Sequel

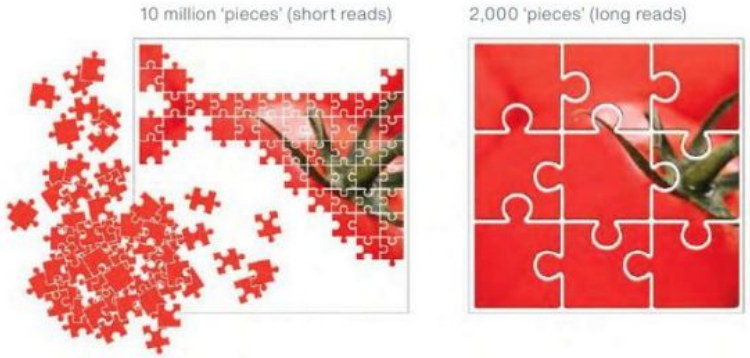
from Elixir GAAS 2018

Long Reads



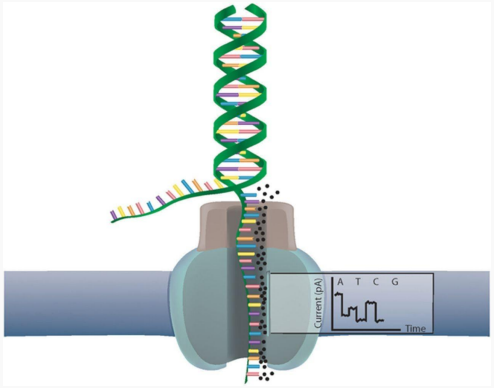
Microbial genomes | Human genomes | Animal genomes | Plant genomes

- 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



Oxford Nanopore

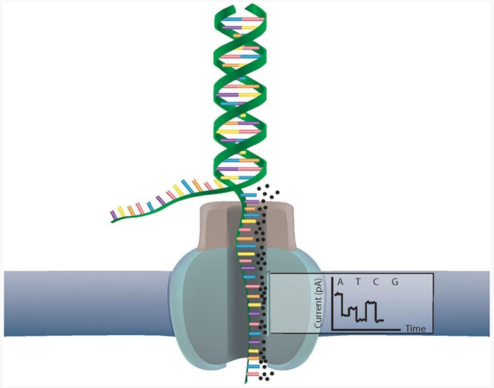
Oxford Nanopore



From Circulation Research

- No Amplification
- NO SYNTHESIS
- Very Long Length

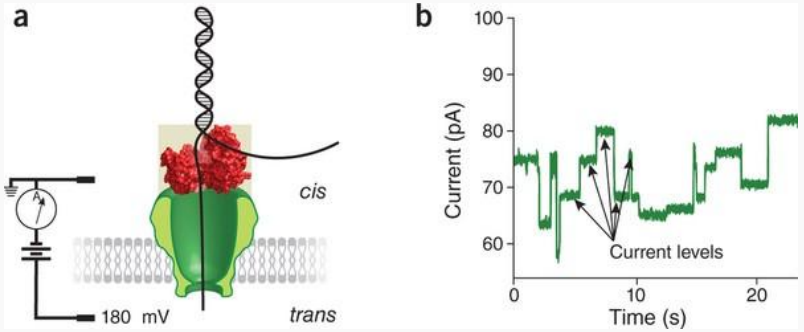
Oxford Nanopore



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

ONT - Advantages and Limits

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

ONT - Advantages and Limits

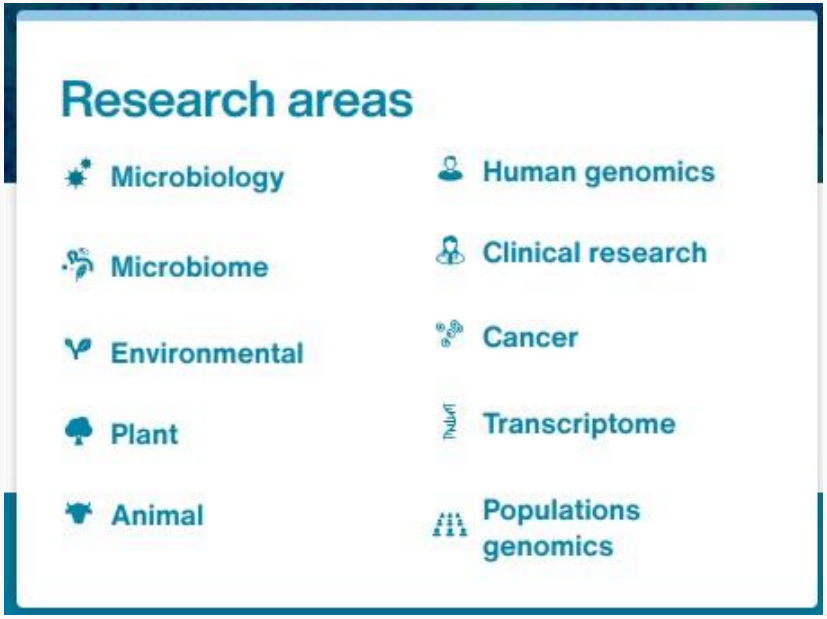
Advantages

- 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 :

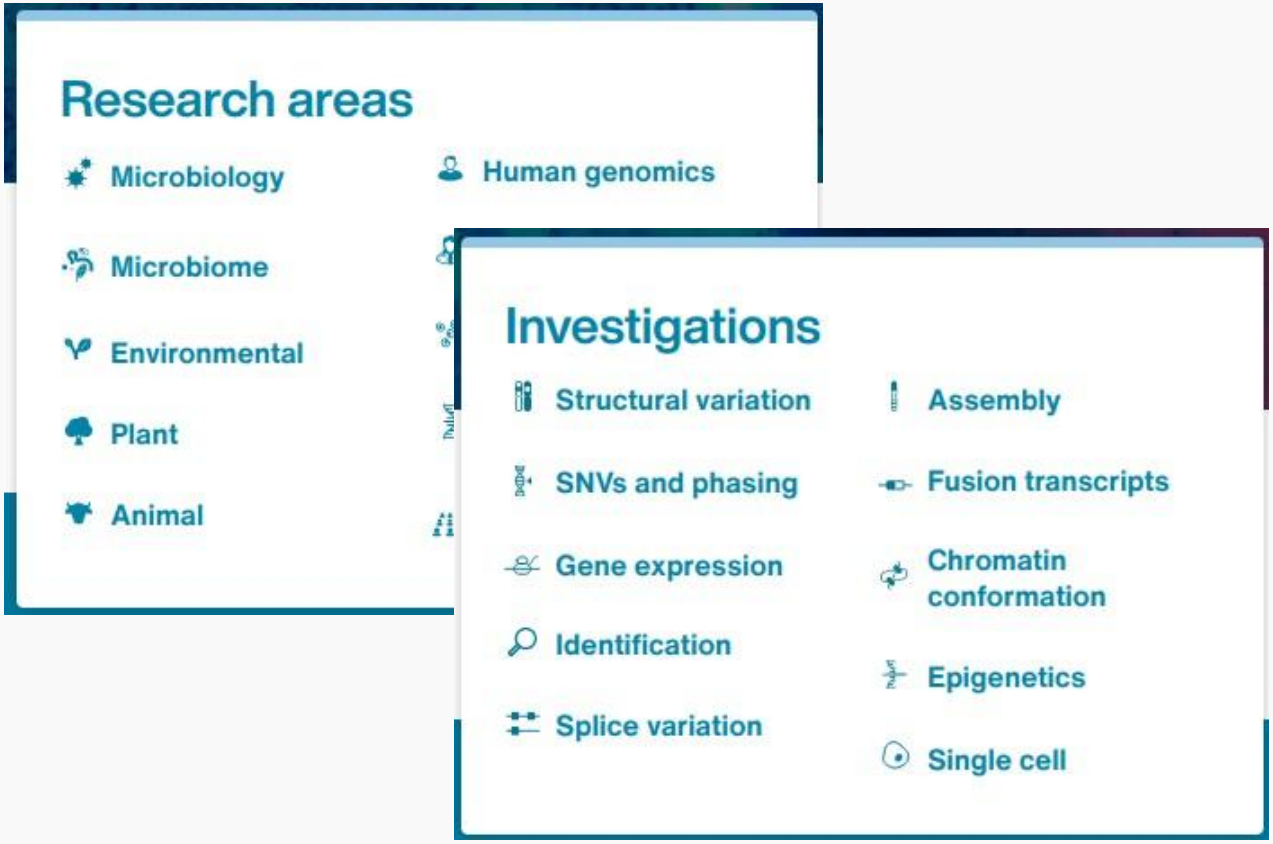
- Error Rate (3-8%, can be corrected, 1-2% in tests)
- Quality of DNA/RNA limits the sequencing
- Heu...

What you can do with it ?



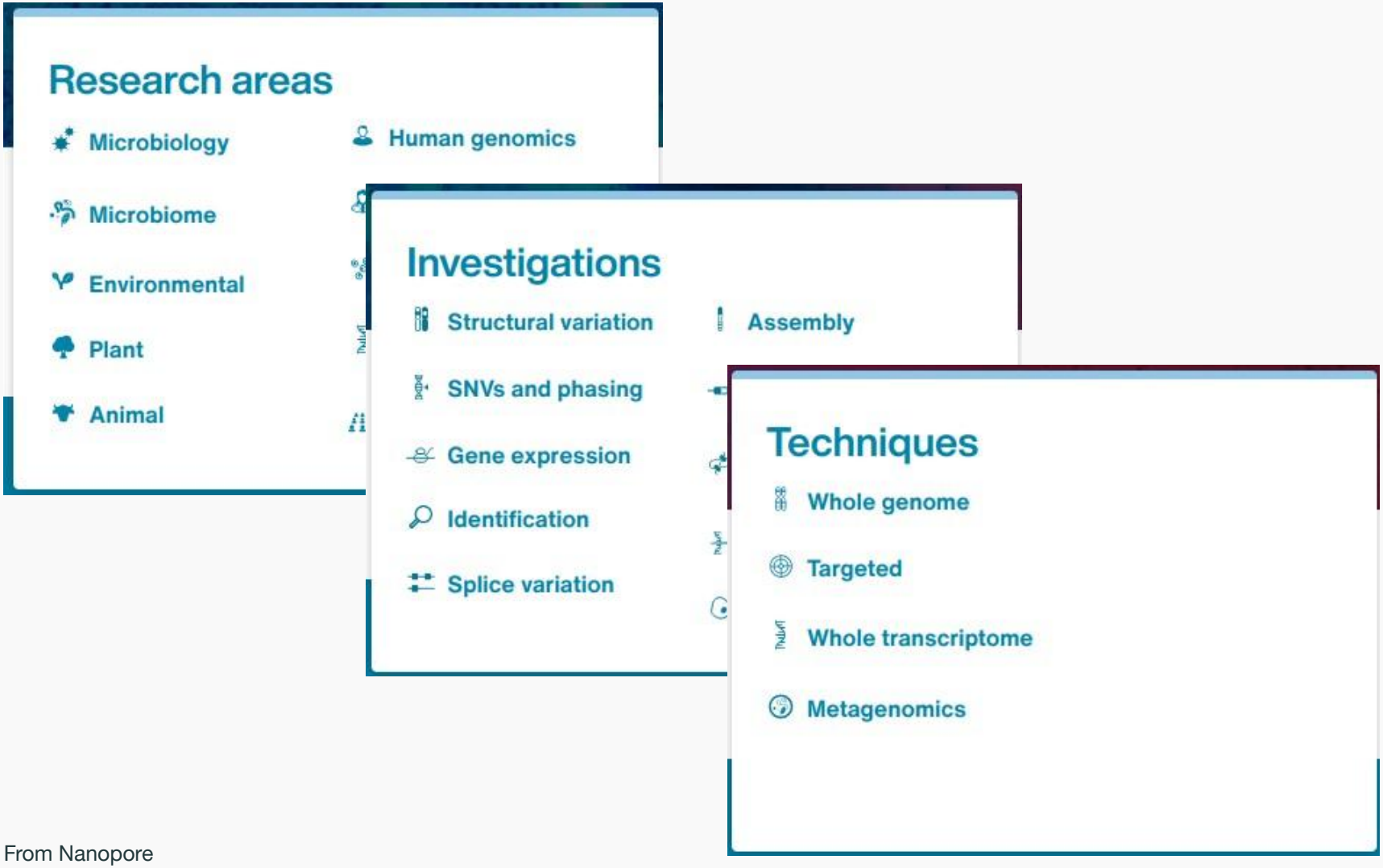
From Nanopore website

What you can do with it ?



From Nanopore website

What you can do with it ?



From Nanopore website

In brief

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

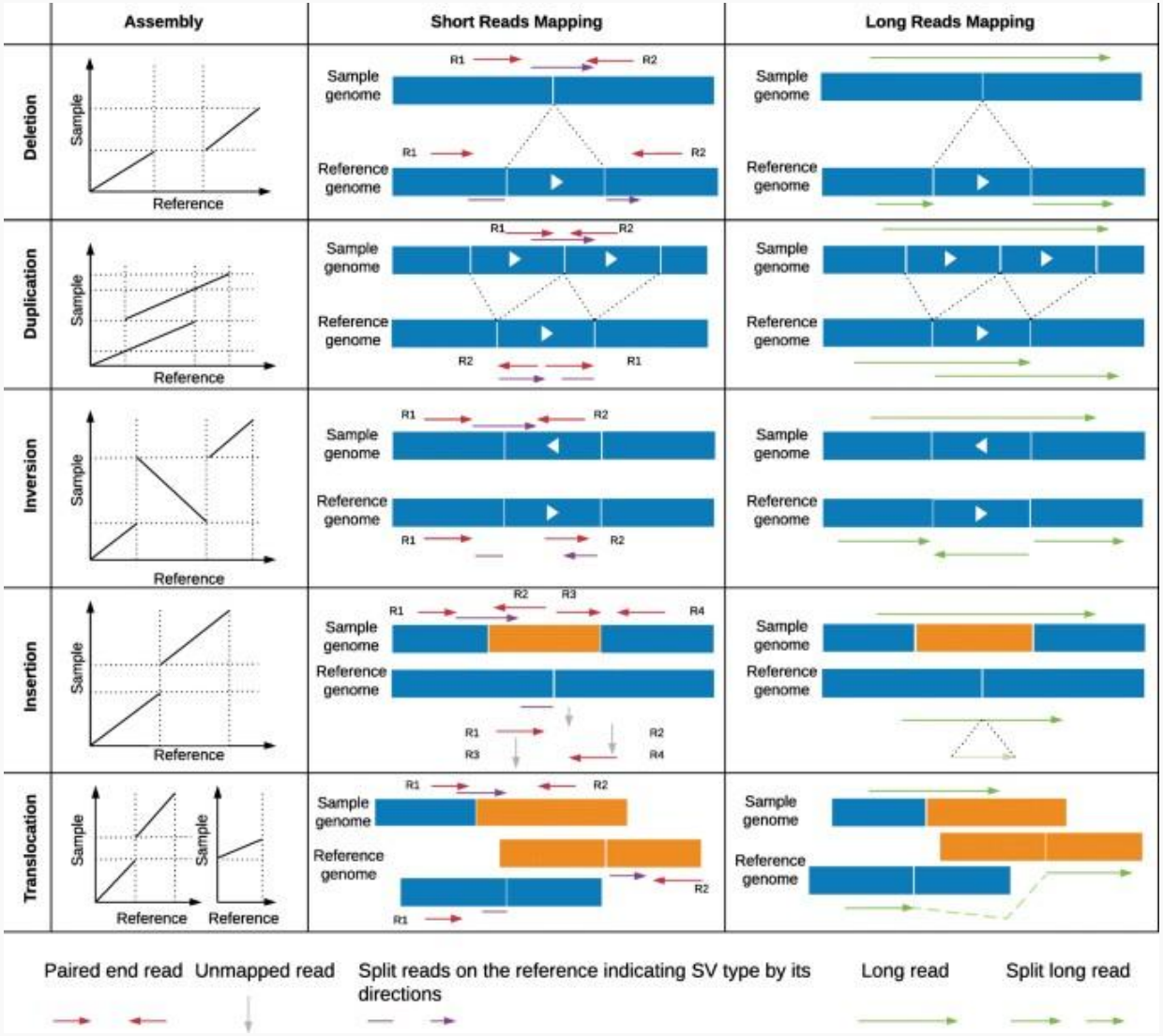
NGS technologies comparison

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



From Mahmoud et al, 2019

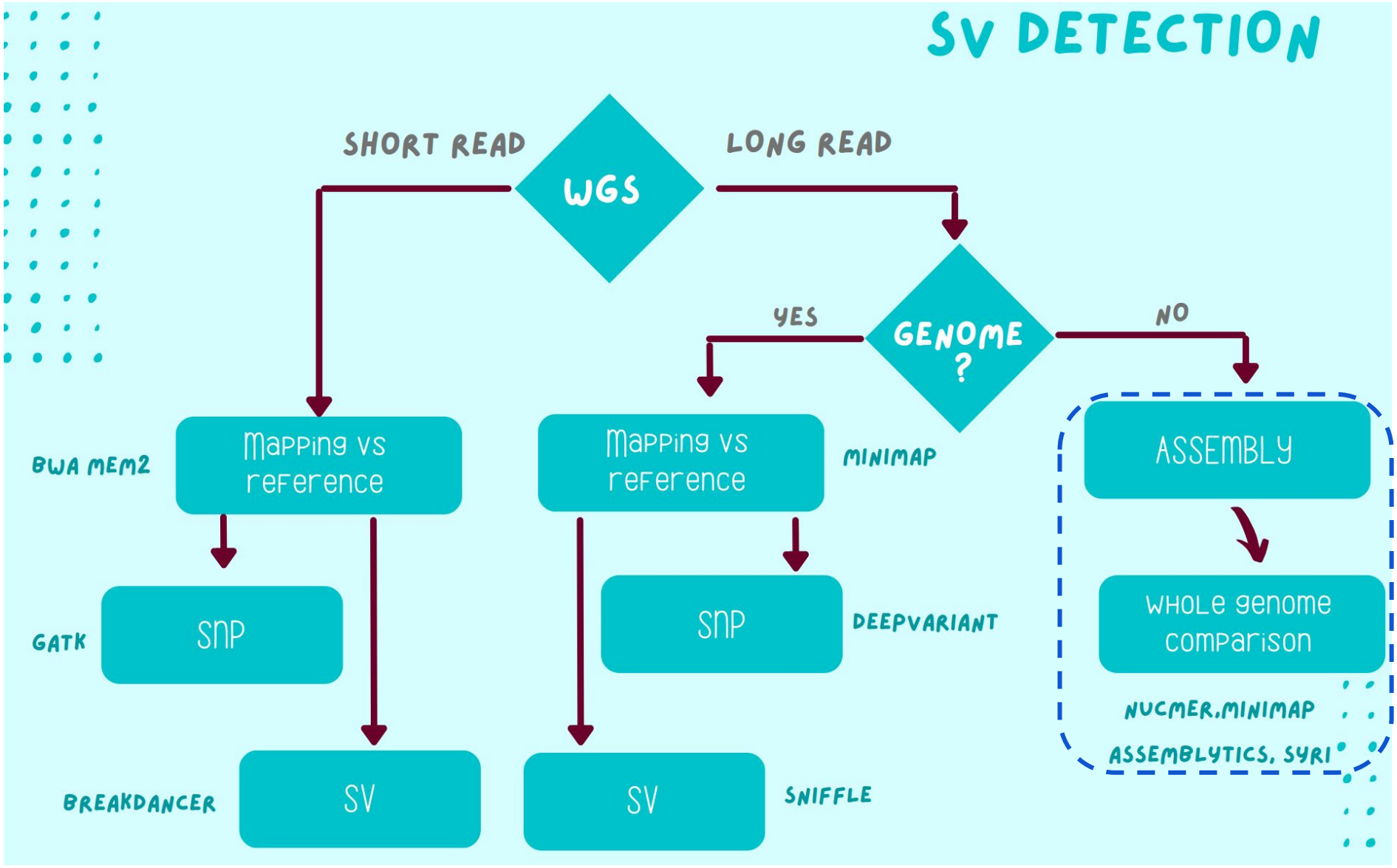
Practice

Let’s work with the jupyter book :

Day3_Variants_structuraux_EMPTY.ipynb


THAT'S ENOUGH
FOR NOW

Training plan - day 4



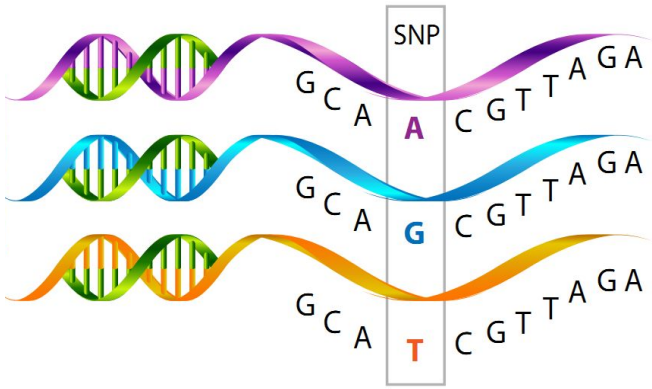
If you are lost after these three days of training…

So let's summarize…

Remember why we started…

Mutations & Variations as main source of genetic diversity

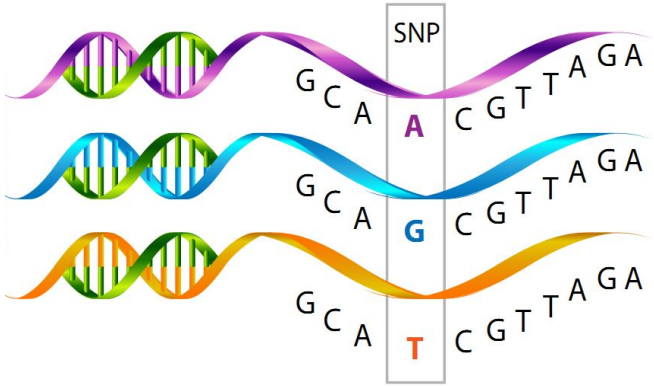
Single Nucleotide Polymorphism



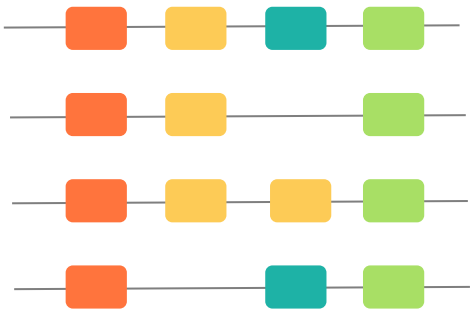
Remember why we started…

Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism



Structural Variations



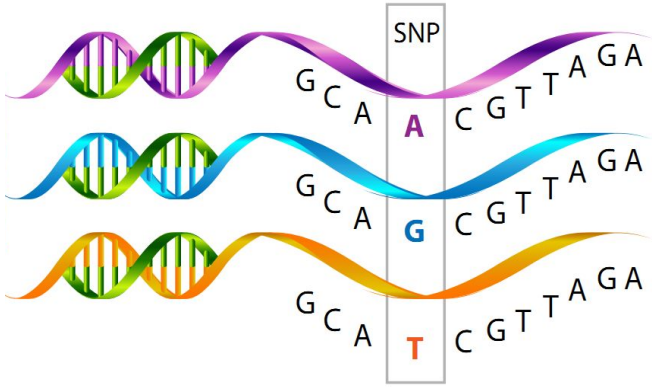
Presence Absence Variation (PAV)

Deletion, duplication, copy number variation, mobile element insertion

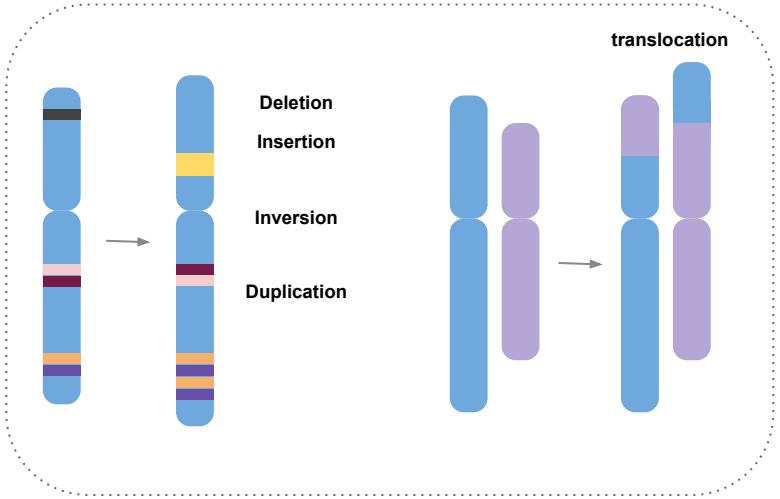
Remember why we started…

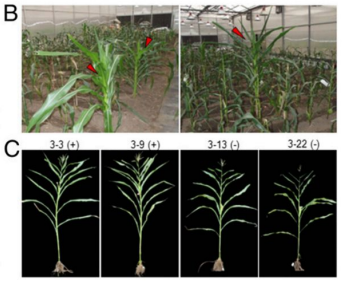
Mutations & Variations as main source of genetic diversity

Single Nucleotide Polymorphism



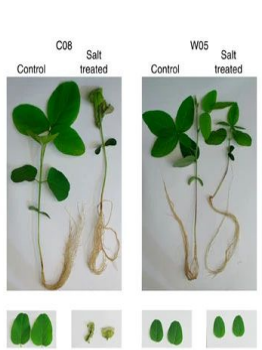
Structural Variations





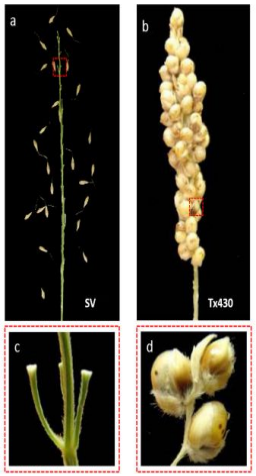
FromYang et al., 2013

From Li et al. 2012

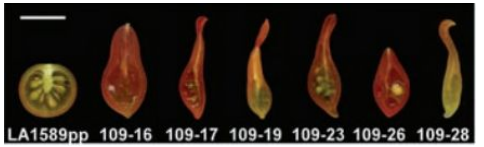


From Qi et al. 2014

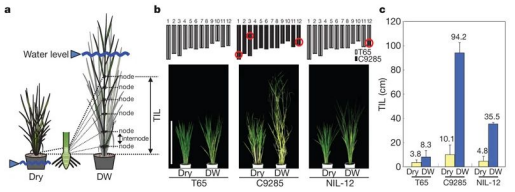
FromYang et al., 2014



From Lin et al. 2012

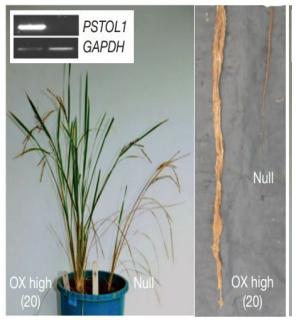


From Xiao et al. 2008

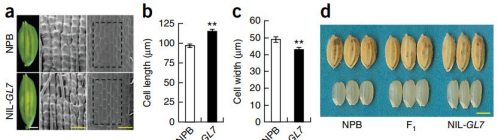


From Hattori et al. 2009

From Xu et al. 2006

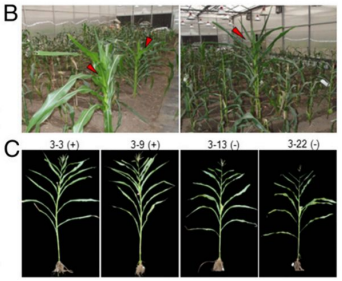


From Gamuyao et al. 2012



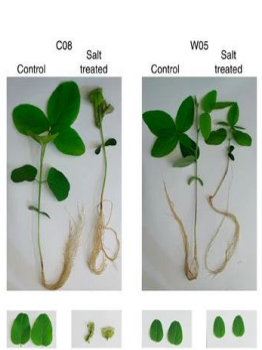
From Wang et al. 2015

From Bai et al. 2017



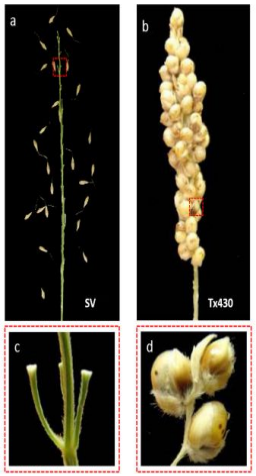
FromYang et al., 2013

From Li et al. 2012



From Qi et al. 2014

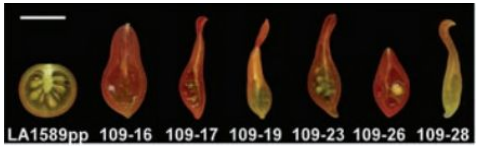
FromYang et al., 2014



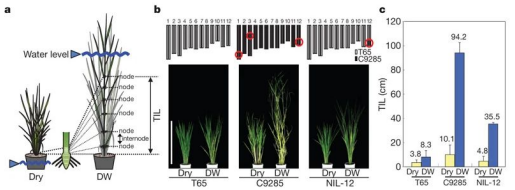
From Lin et al. 2012

Is One Reference genome enough

to capture all genetic diversity ?

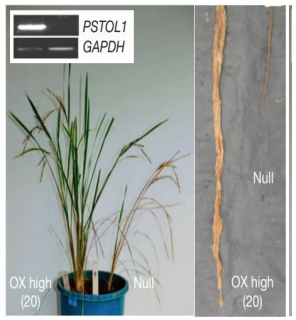


From Xiao et al. 2008

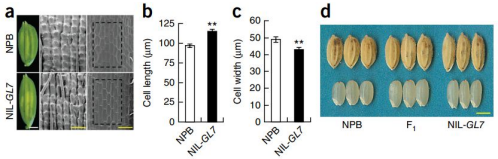


From Hattori et al. 2009

From Xu et al. 2006



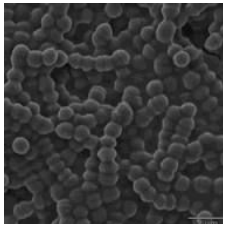
From Gamuyao et al. 2012



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 Masignani[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],

Tetellin et al., 2005

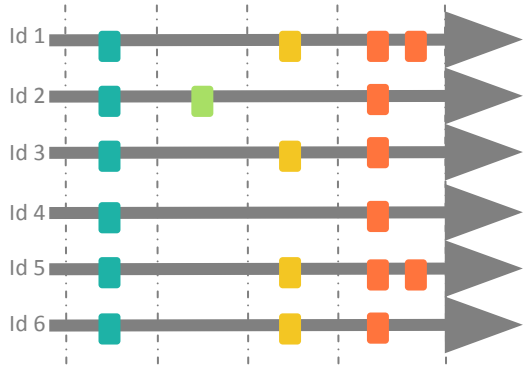
- 8 strains sequenced
- SNP variations

Large number of genes not shared between isolates

20% genome variability and 80 % shared by all isolates

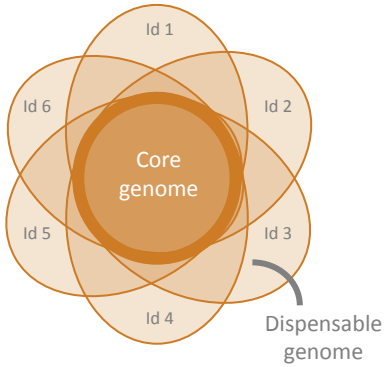
Pangenome concept

Pangenome concept

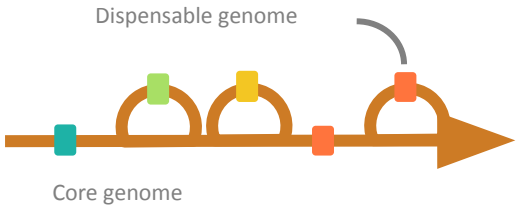


Pangenome

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



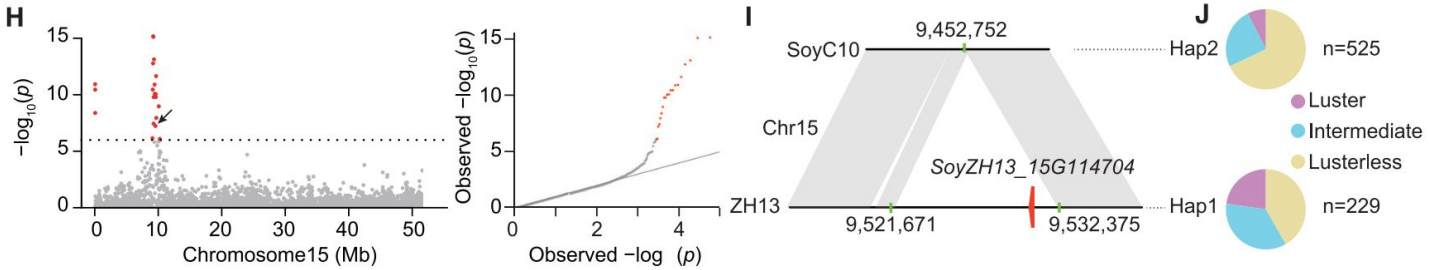
- **Core genome** : present in all individuals
- **Dispensable genome** : absent from one or several individuals (also called variable, accessory,...)



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 ?

WGS

SHORT READ

LONG READ

GENOME ?

YES

NO

BWA MEM2

Mapping vs reference

GATK

SNP

BREAKDANCER

SV

Mapping vs reference

MINIMAP

SNP

DEEPVARIANT

SV

SNIFFLE

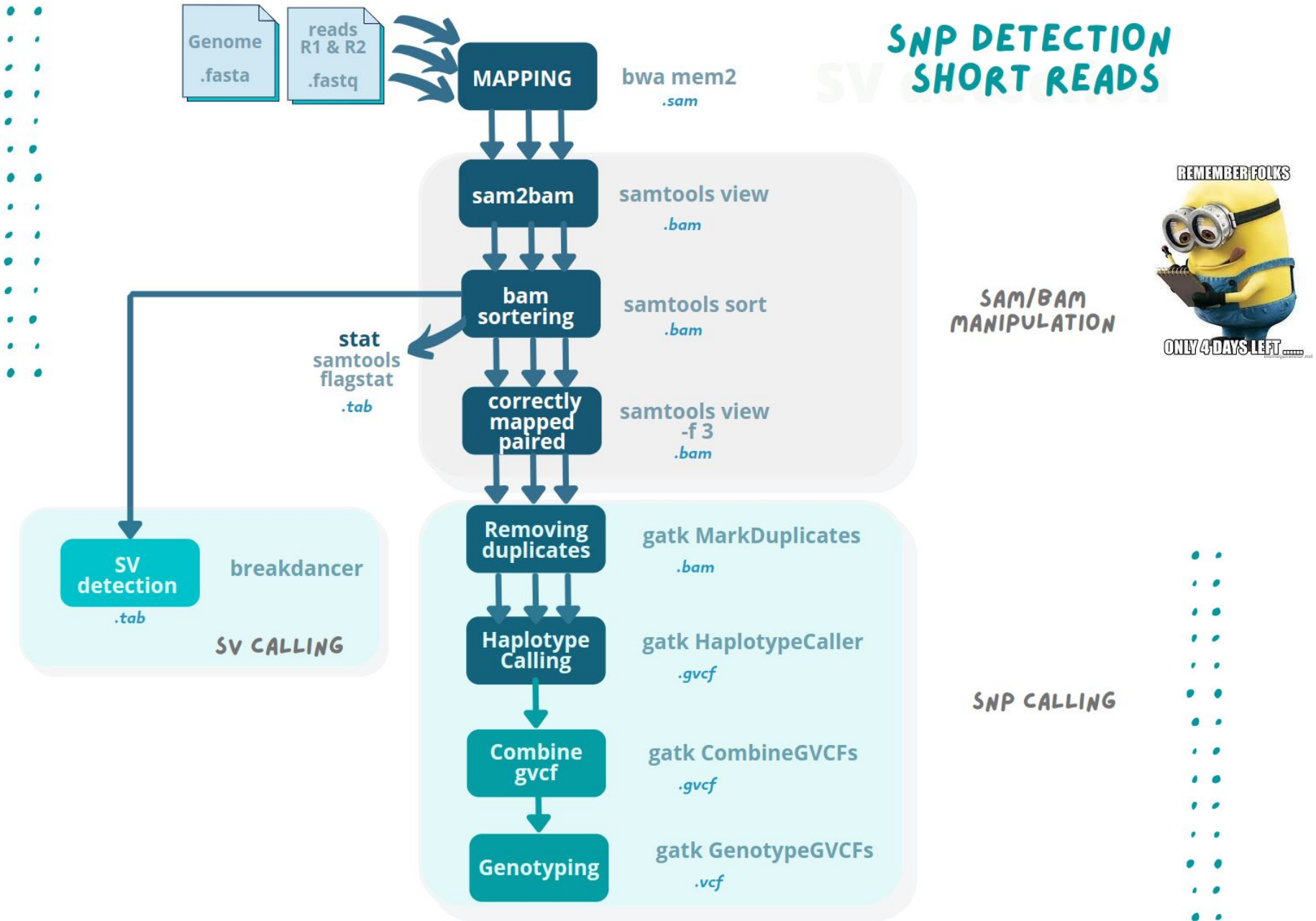
ASSEMBLY

WHOLE genome comparison

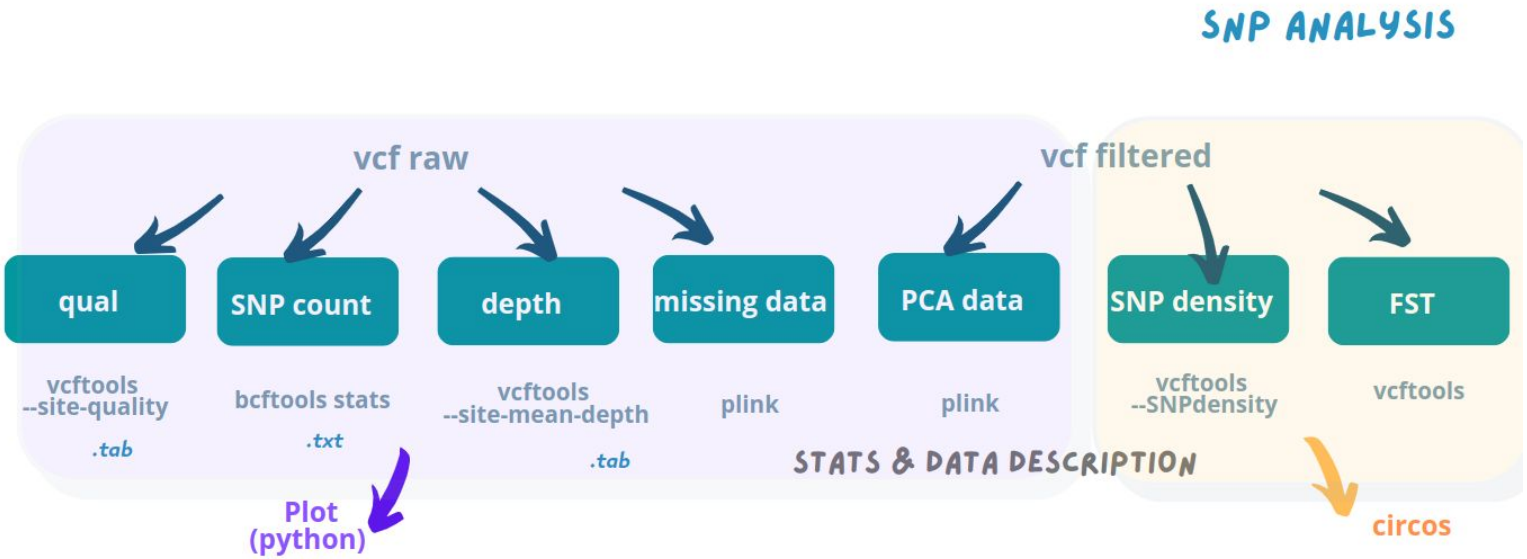
NUCMER,MINIMAP

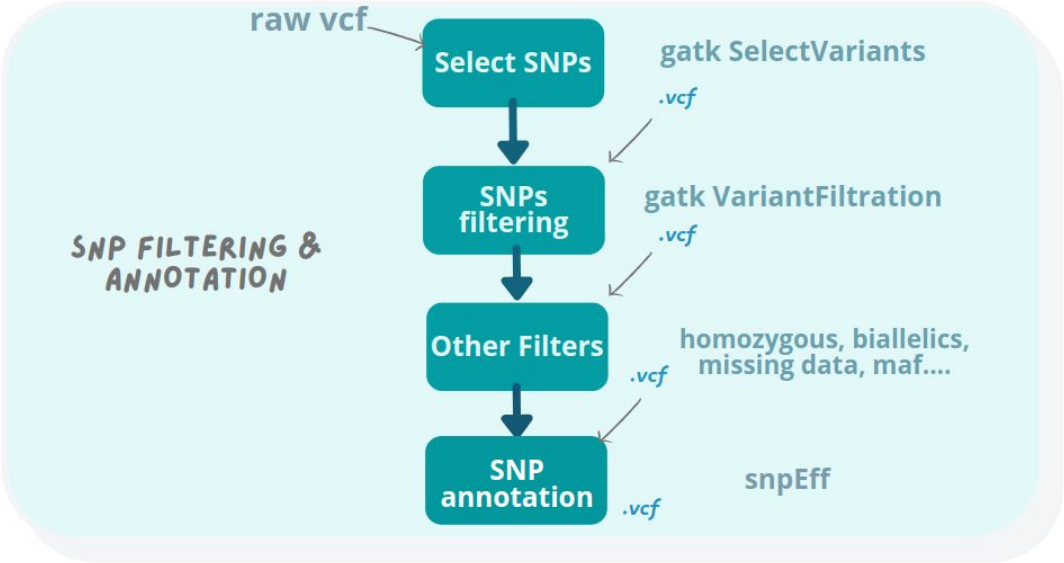
ASSEMBLYTICS, SYRI

From short reads….



From short reads….





How to detect SVs and analyse them ?

WGS

SHORT READ

LONG READ

GENOME ?

YES

NO

BWA MEM2

Mapping vs reference

GATK

SNP

BREAKDANCER

SV

Mapping vs reference

MINIMAP

SNP

DEEPVARIANT

SV

SNIFFLE

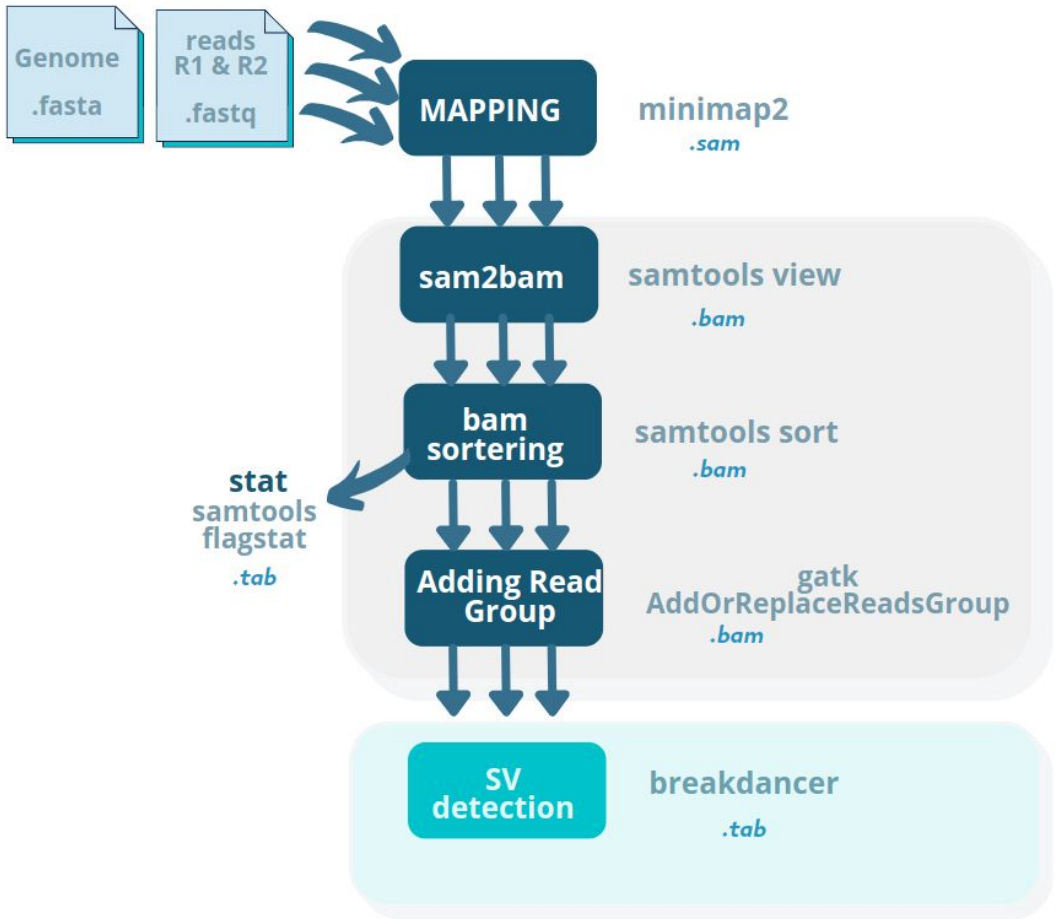
ASSEMBLY

WHOLE genome comparison

NUCMER,MINIMAP

ASSEMBLYTICS, SYRI

From long reads…



SV DETECTION
LONG READS

SAM/BAM
MANIPULATION

SV CALLING

How to detect SVs and analyse them ?

WGS

SHORT READ

LONG READ

GENOME ?

YES

NO

BWA MEM2

Mapping vs reference

GATK

SNP

BREAKDANCER

SV

Mapping vs reference

MINIMAP

SNP

DEEPVARIANT

SV

SNIFFLE

ASSEMBLY

WHOLE genome comparison

NUCMER,MINIMAP

ASSEMBLYTICS, SYRI

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 Dennu

Observatoire océanologique de Banyuls sur mer
Laboratoire d'océanographie microbienne

Structure of Bathycoccus genome

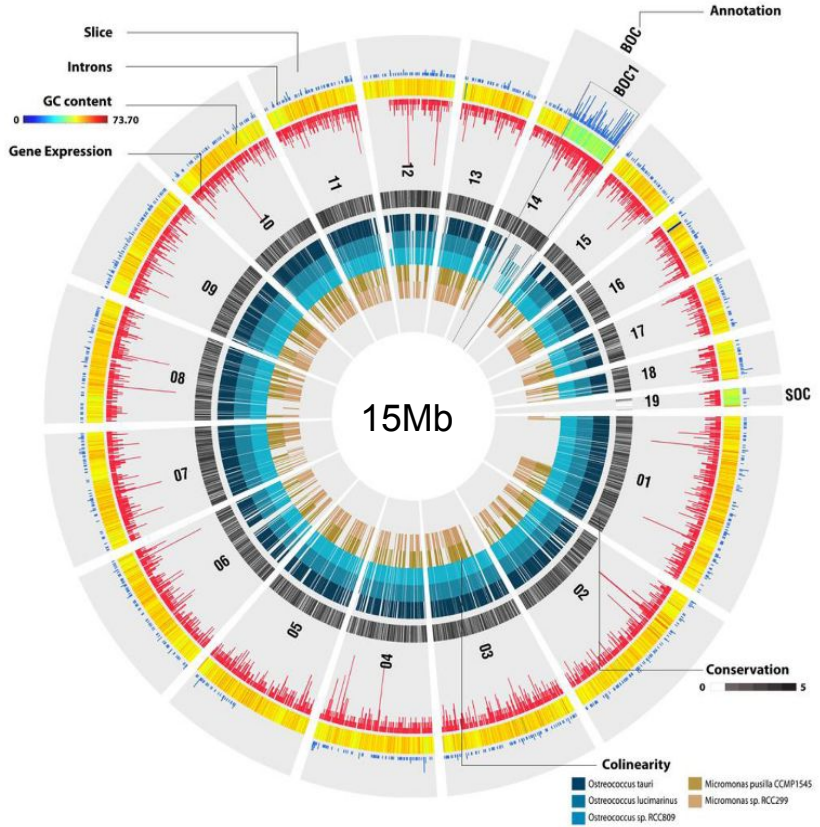
Reference genome
strain RCC1105
Banyuls bay in 2006

15Mb haploid genome

19 nuclear chromosomes

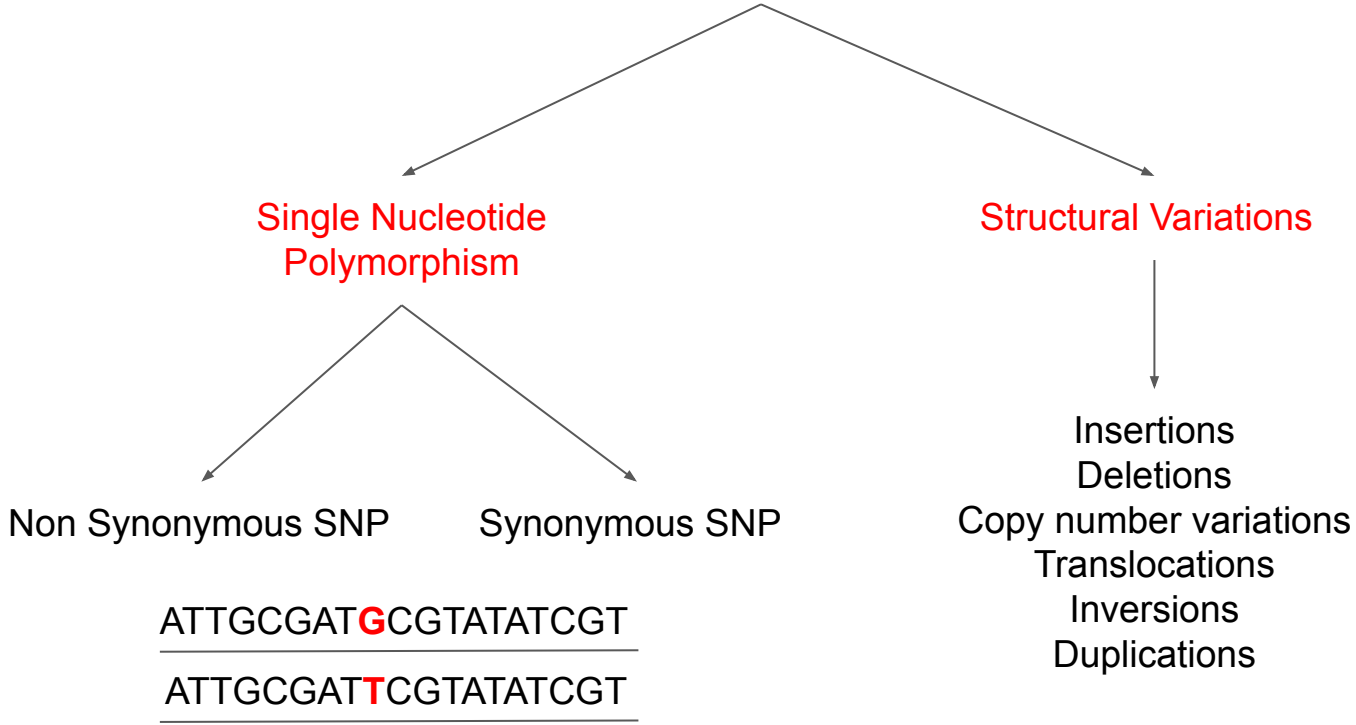
2 Outliers regions with irregular GC content and gene composition:

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



Moreau et al, 2012

Genetic variations

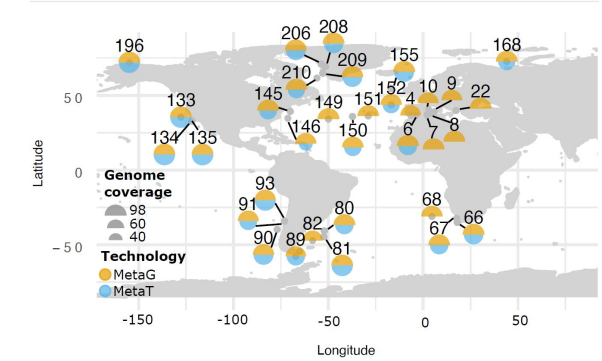


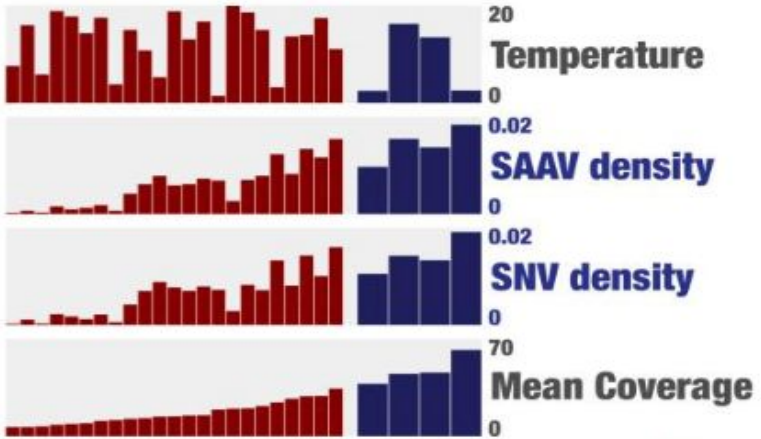
Bathycoccus SNP in metagenomic datasets

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



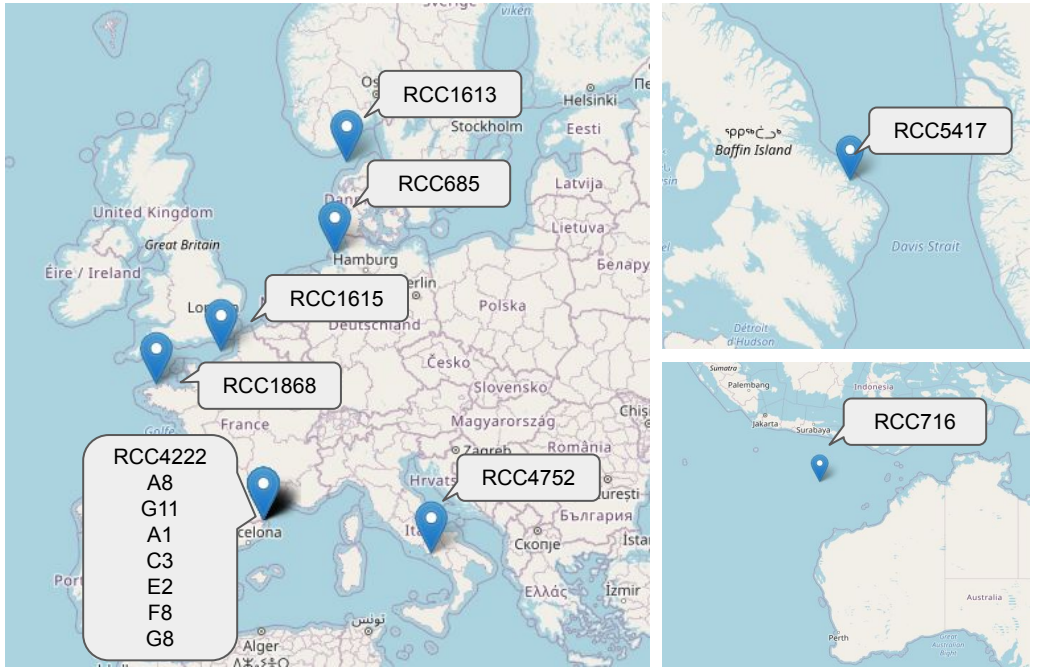


Leconte et al, 2021

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

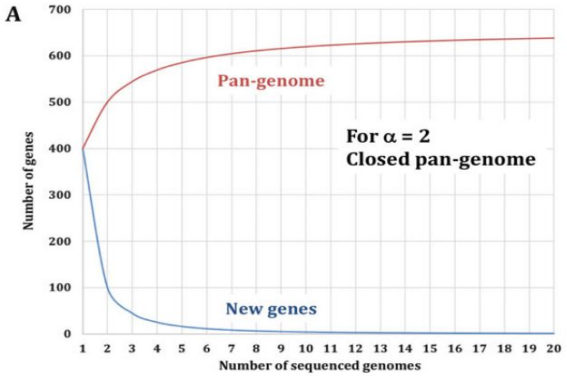
Pangenome
Repertory of all sequences/genes found in a species

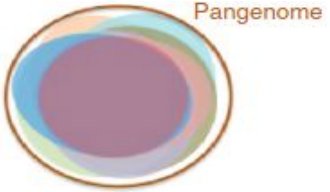
Number of new sequences discovered after each new genome sequenced follow a decreasing power law

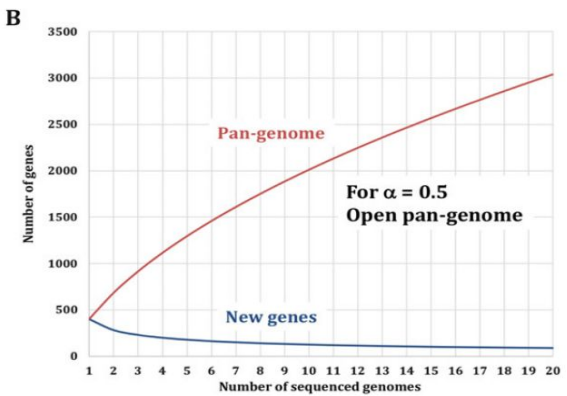
$$n = \kappa N^{-\alpha}$$

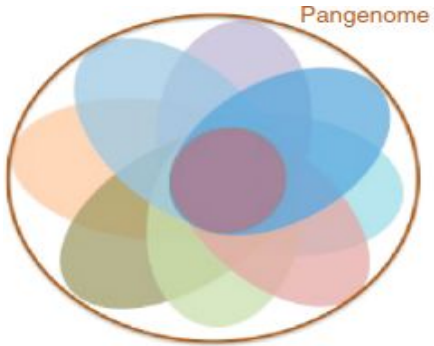
Closed Pangenome
Limited diversity

Open Pangenome
Boundless diversity





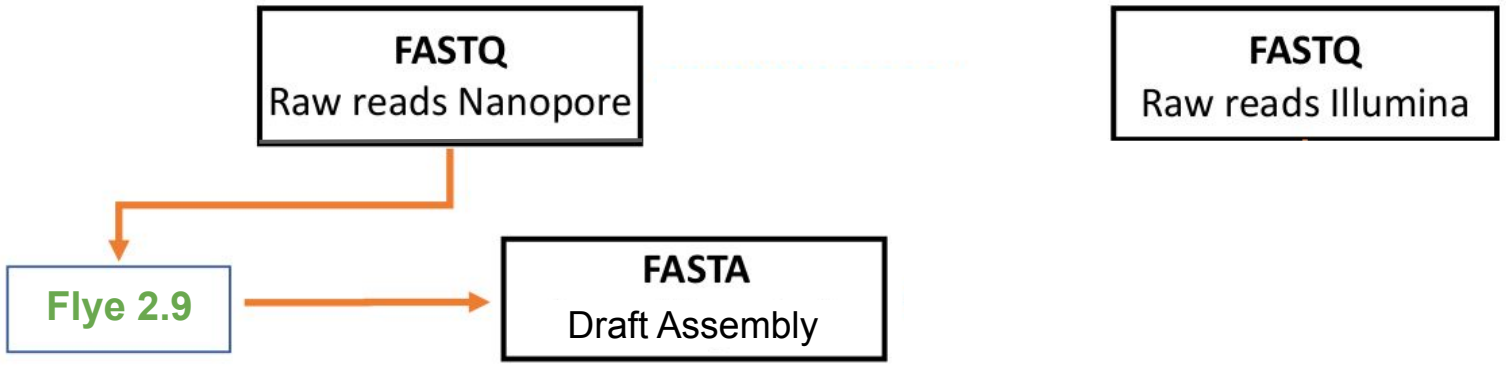




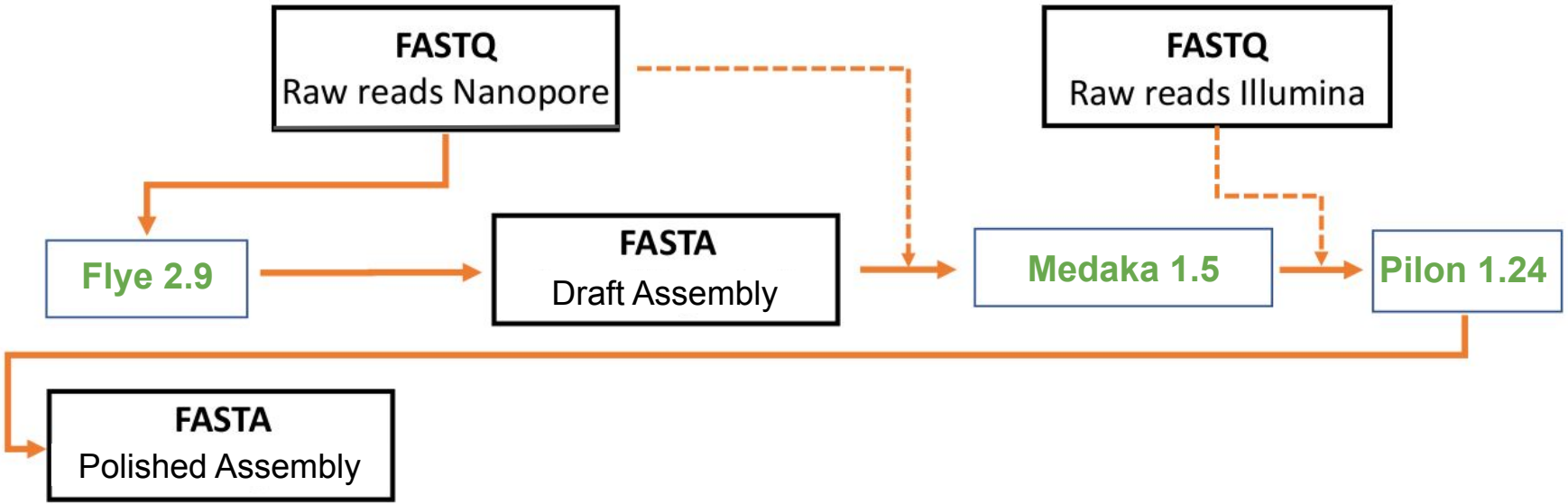
Richard, 2020
Tranchant-Dubreuil et al, 2019

De novo assembly pipeline for Bathycoccus

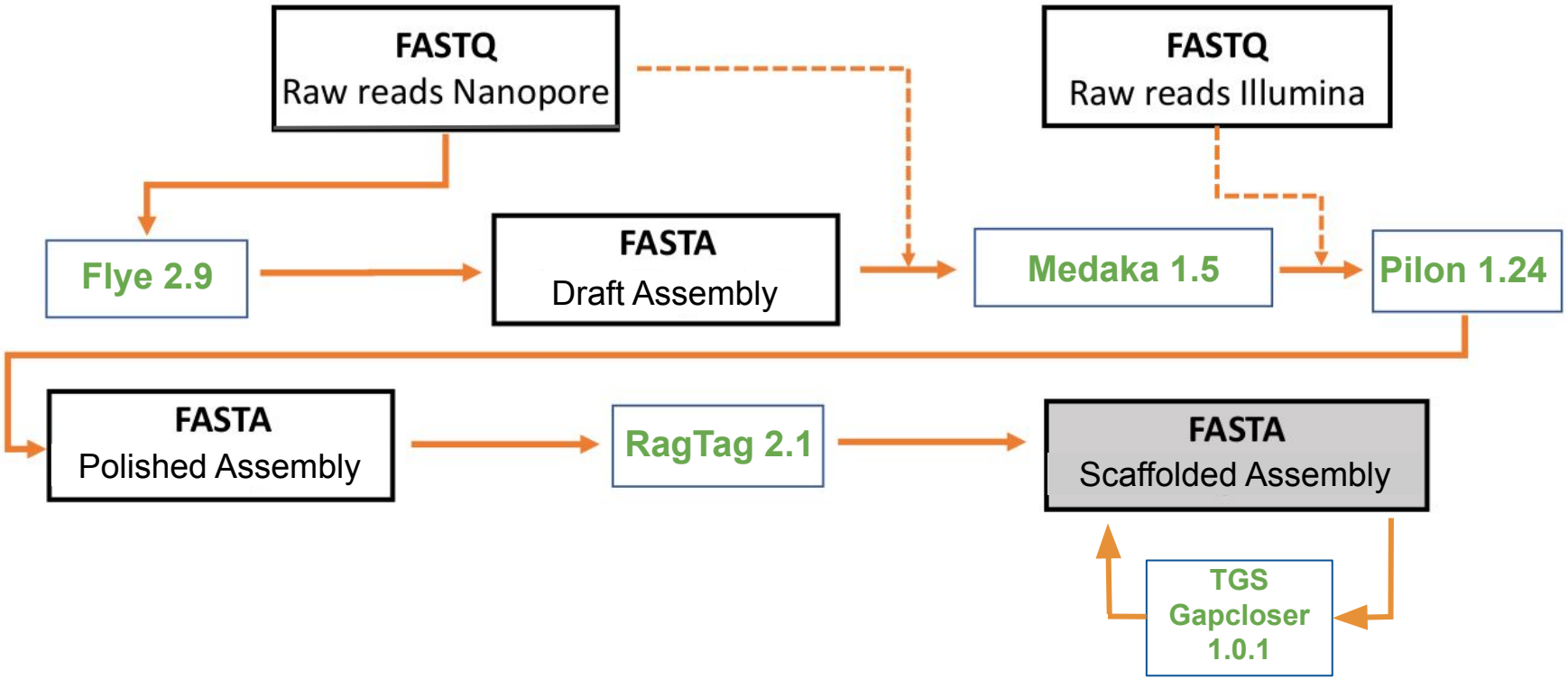
De novo assembly pipeline for Bathycoccus



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	*

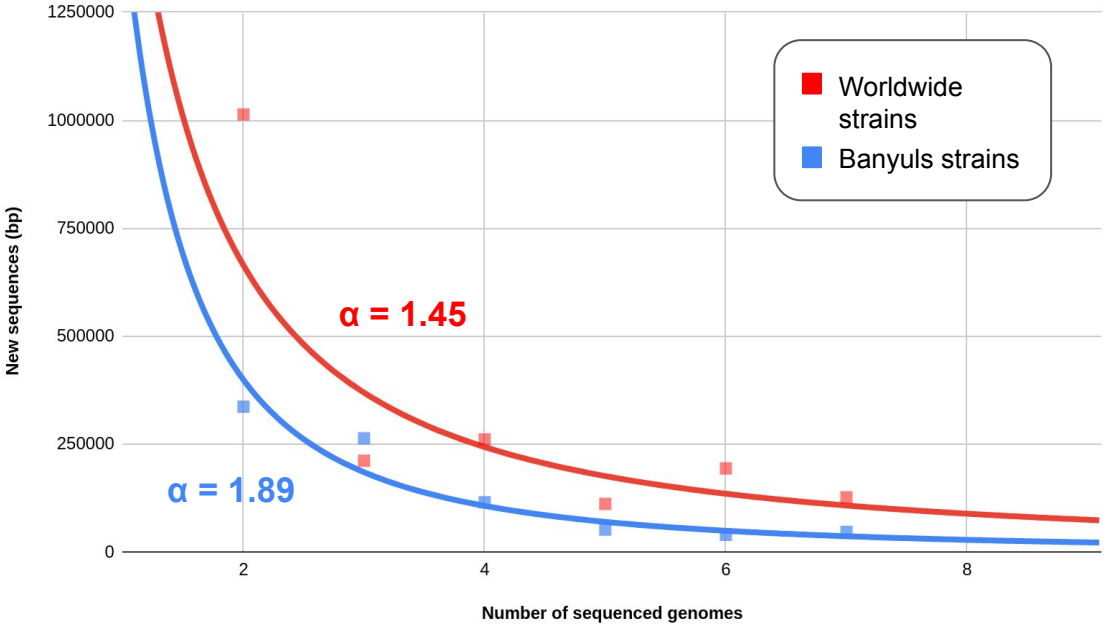
* Illumina unavailable

Structure of Bathycoccus Pangenome
Worldwide diversity and Local diversity

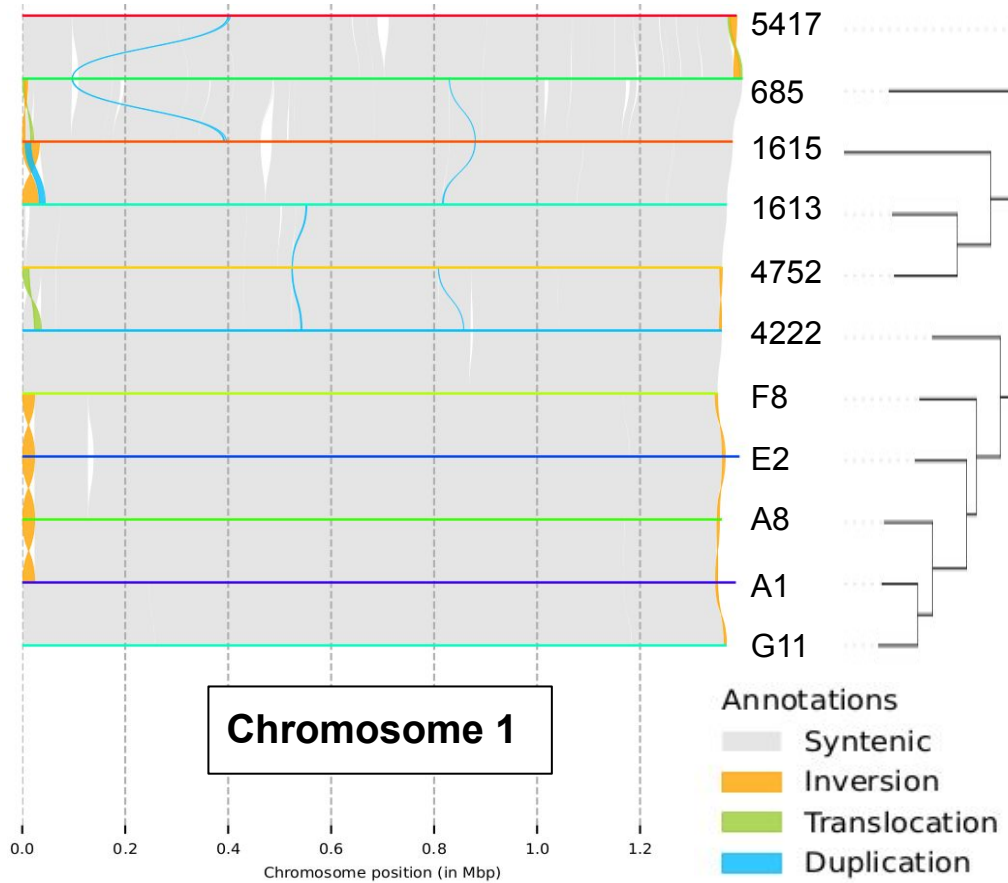
> 2Mb of new accessory sequences
(15% of genome size)

α > 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



high synteny and low number of structural variations

Observed genomic diversity is consistent with limited diversity predicted by pangenome structure

Pour finir… Objectifs remplis ?

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***

Pour finir… Objectifs remplis ?

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

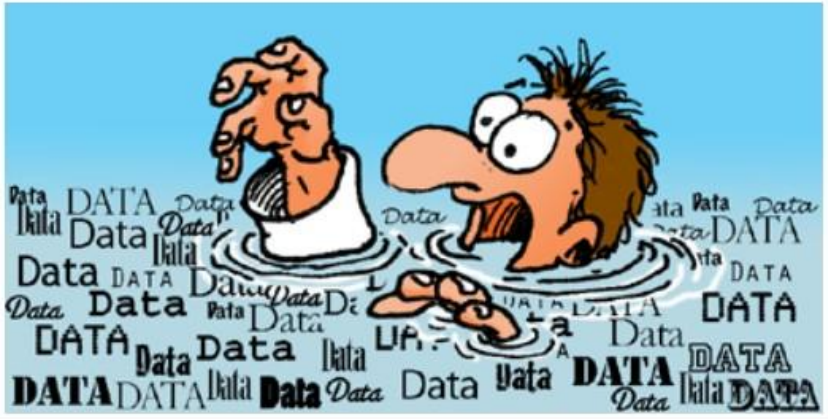
Take home message

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!



Citations

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”

Projets

- 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 !*

Last but not least !

Louis Dennu *pour les données du "practice Day 4"*

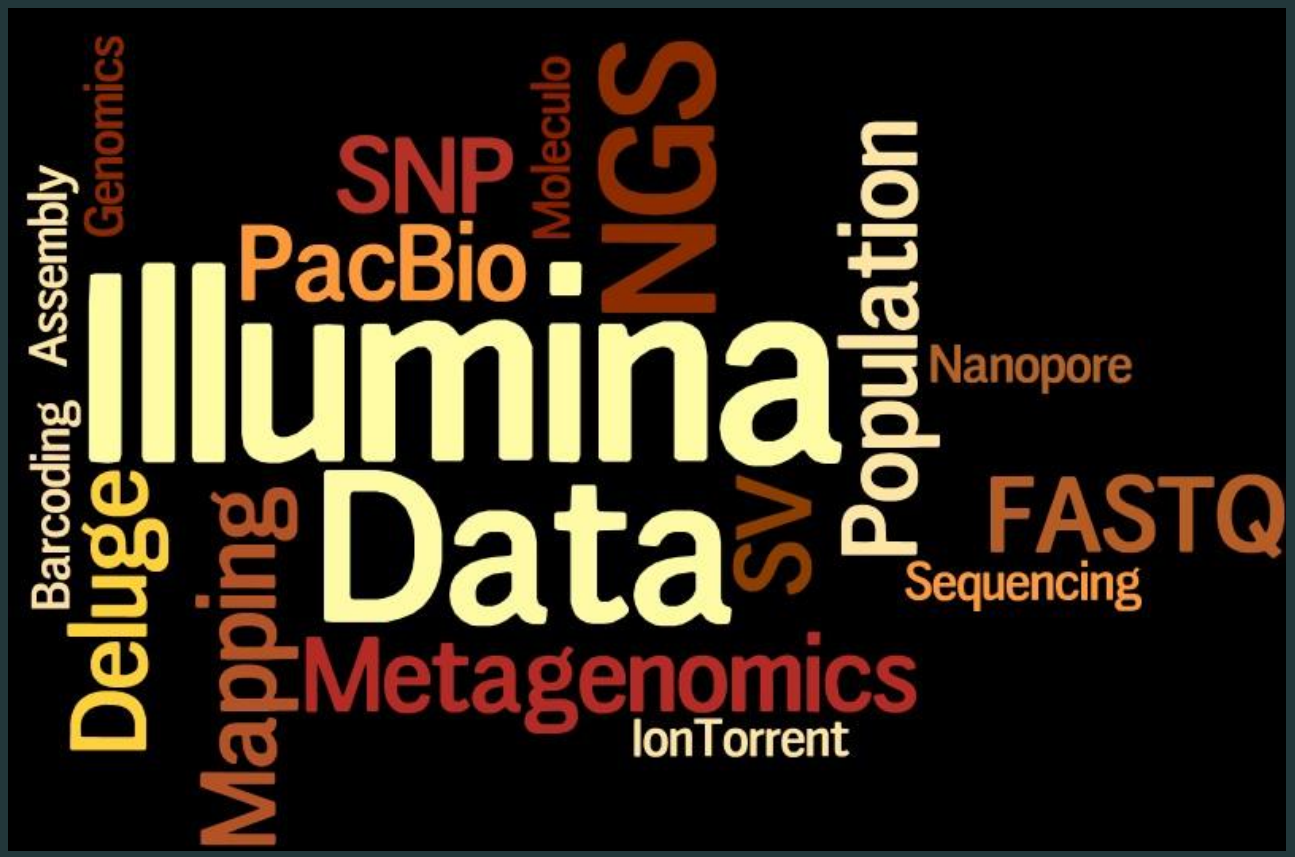
Service formation IRD et CIRAD

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

Core Cloud

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

Thanks you all, and hoping the teaching was helpful



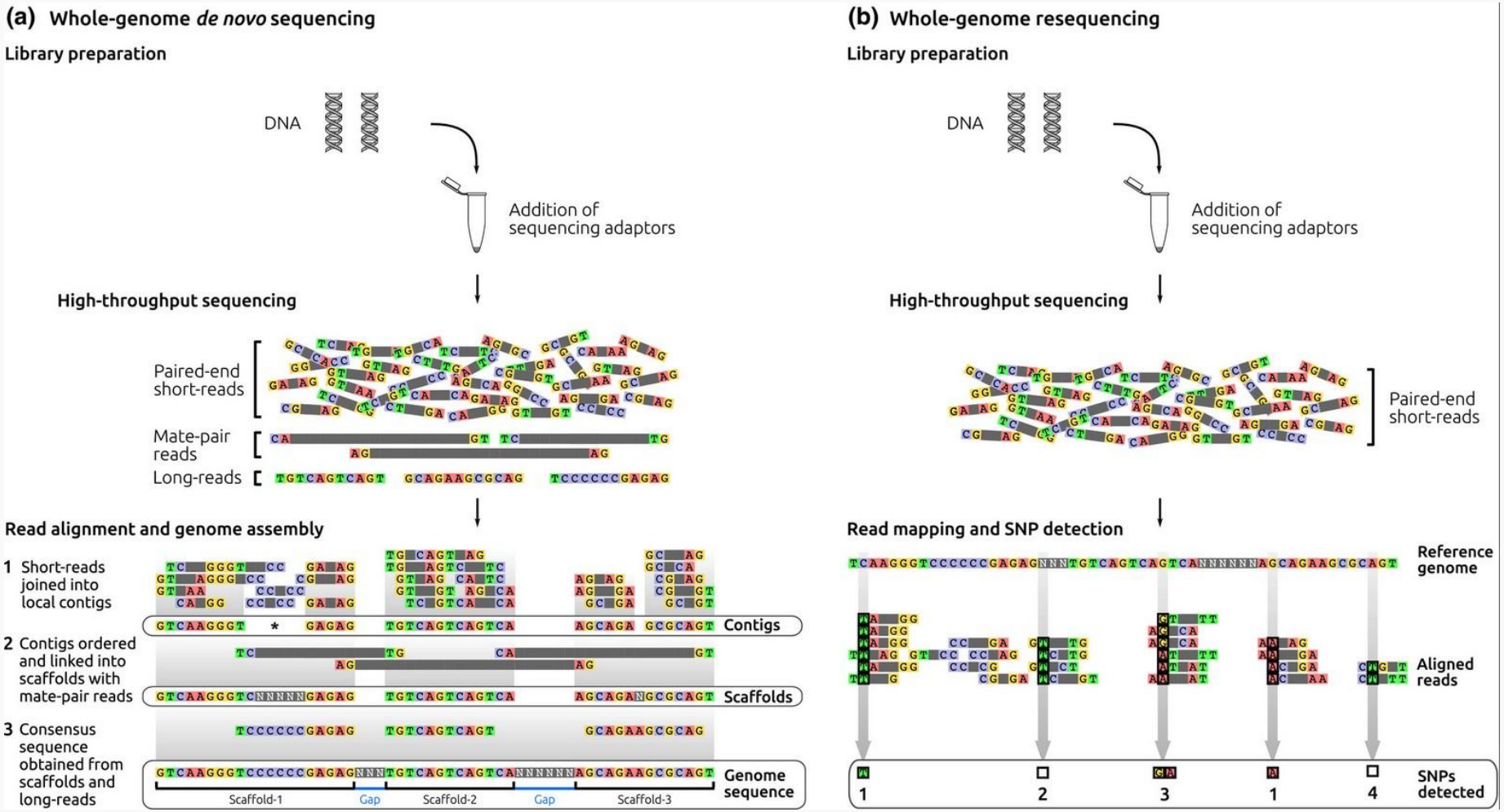
Formateurs

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



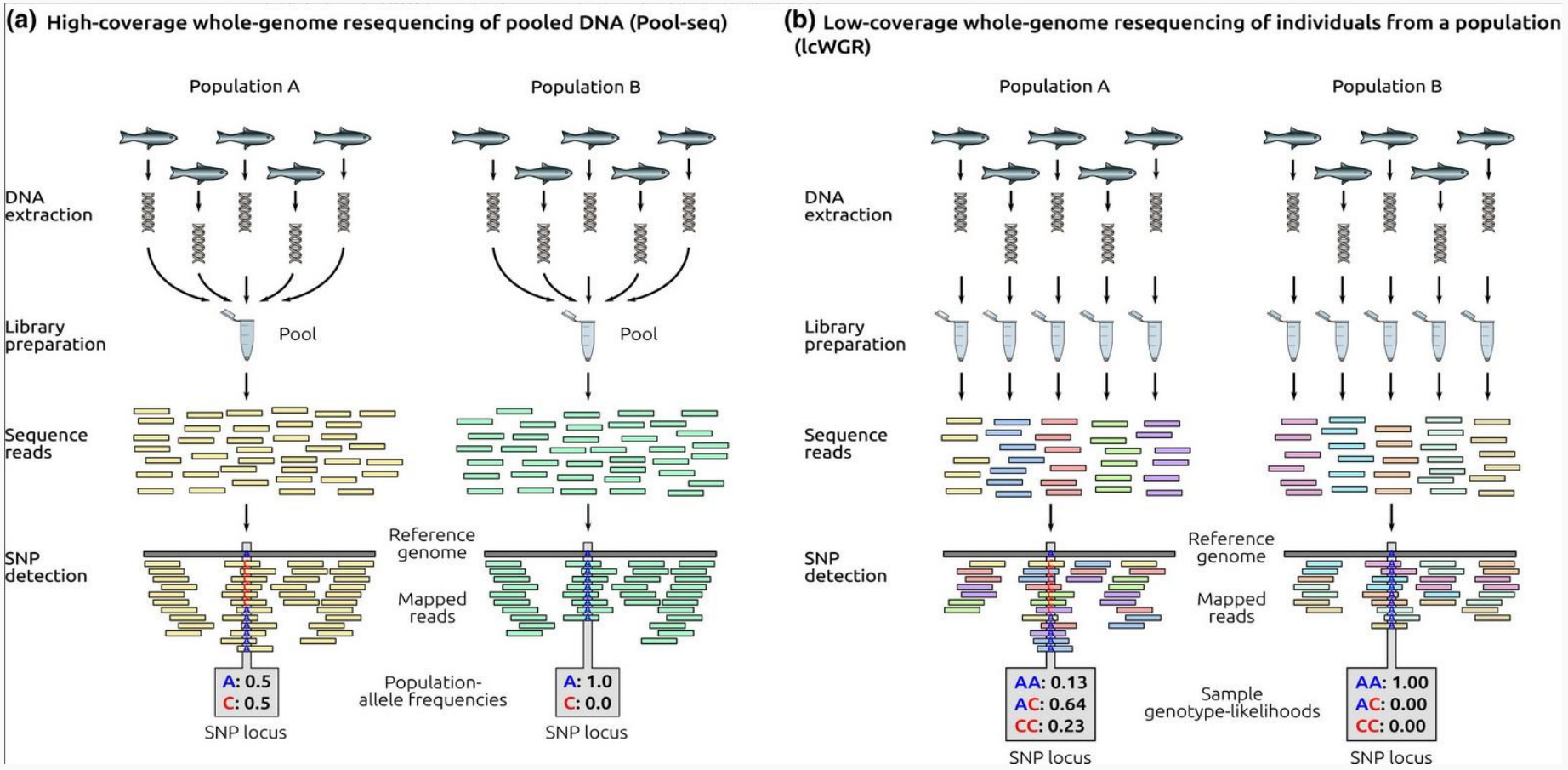

Variations around genotypes and SNP anthem

Standard approaches



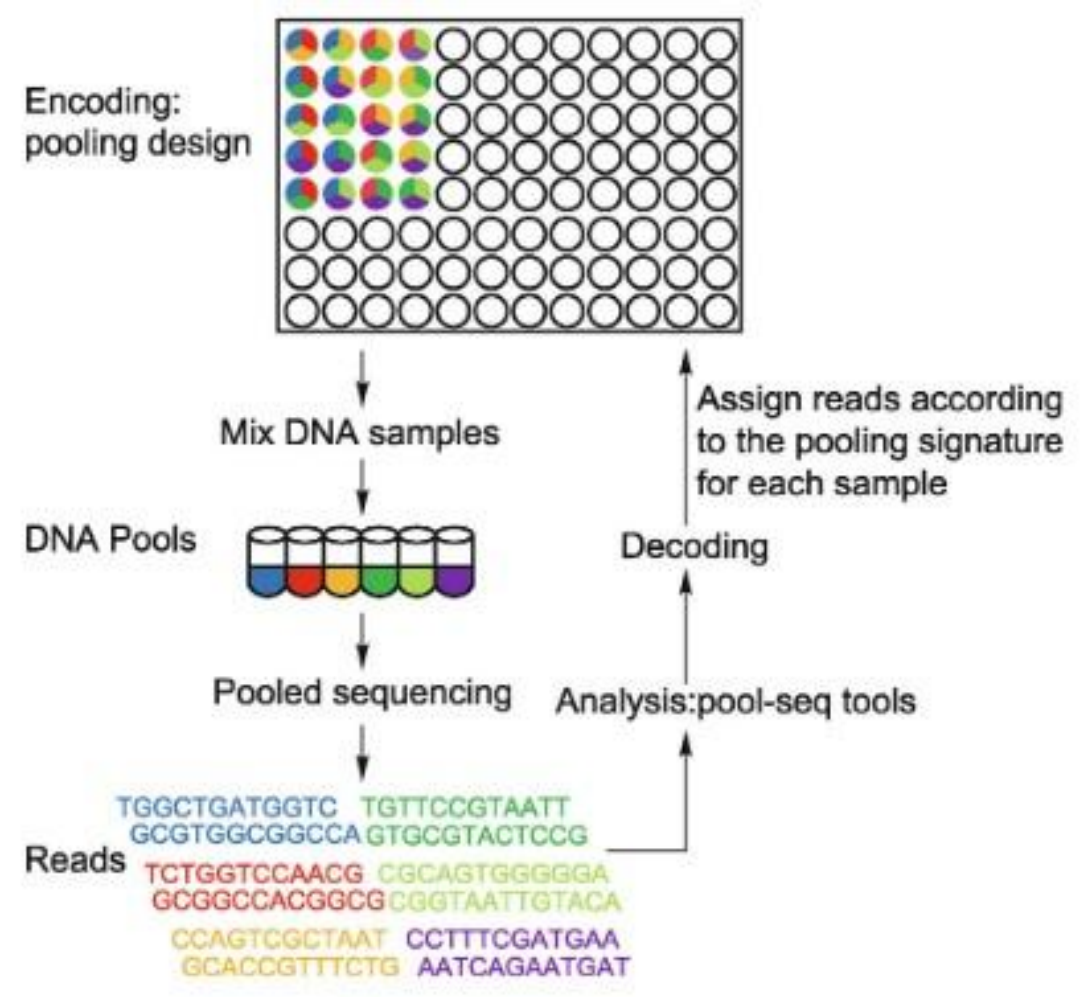
Fuentes-Pardo and Ruzzante, 2017

PoolSeq



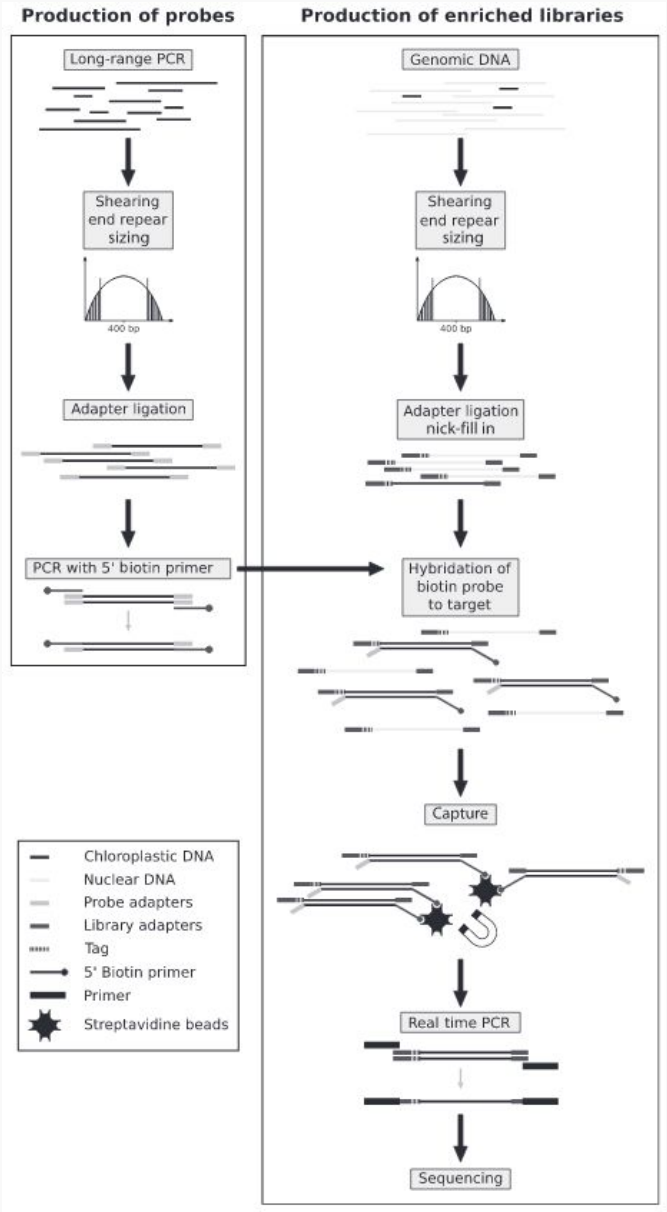
Fuentes-Pardo and Ruzzante, 2017

PoolSeq, barcoded

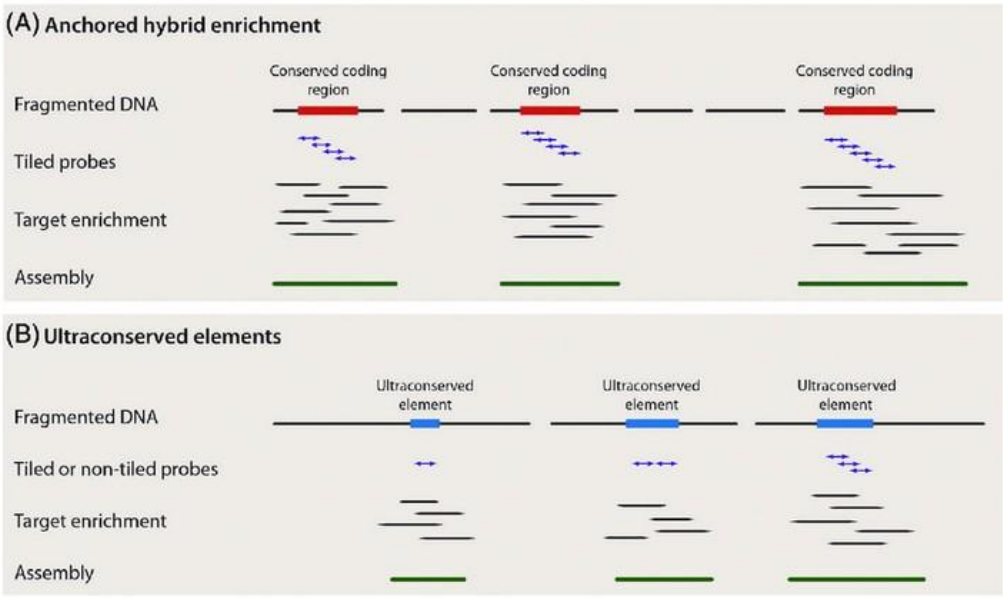


Cao-Sun et al, 2016

Phylogenomics using probes

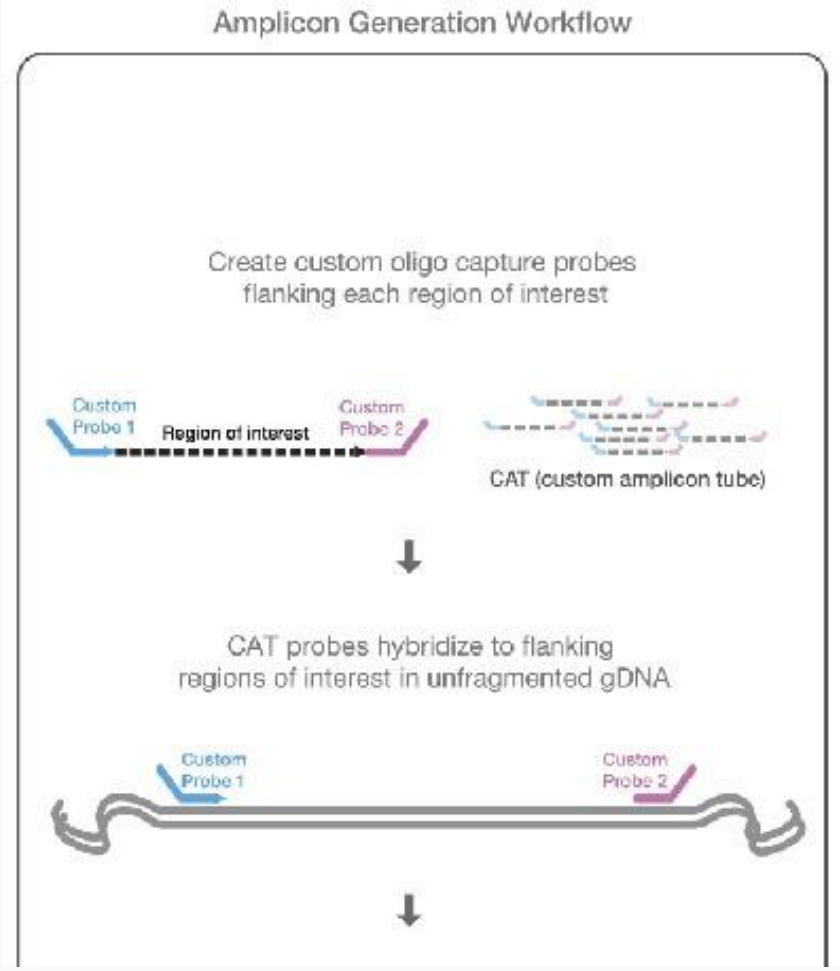


Mariac et al, 2014

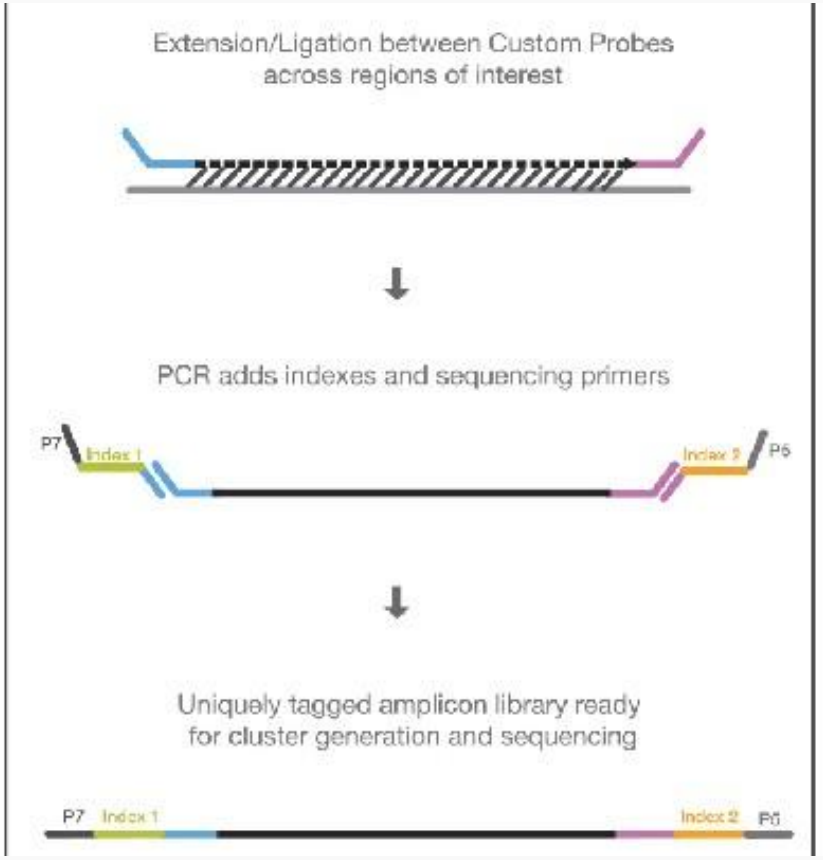


Young and Gillung, 2019

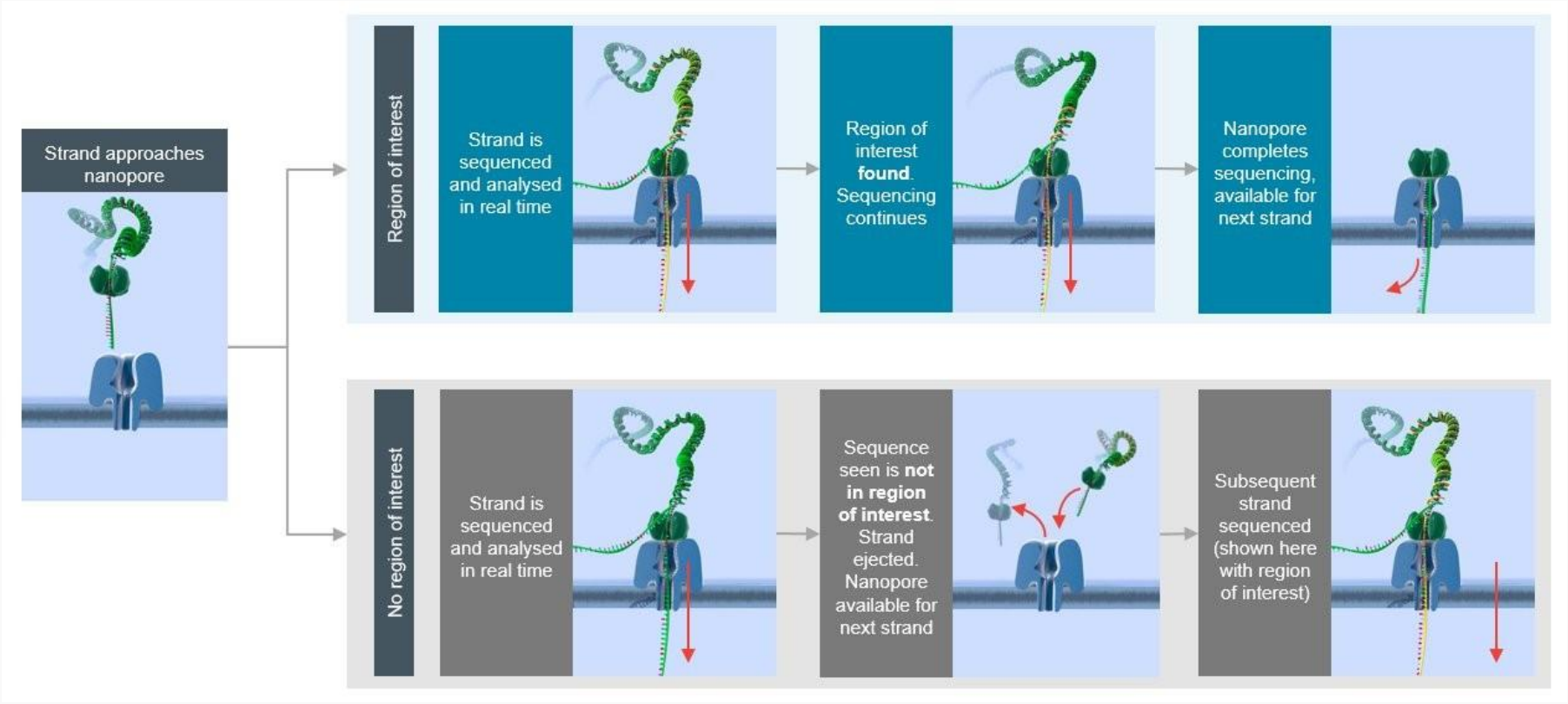
Phylogenomics using amplicons



From Illumina



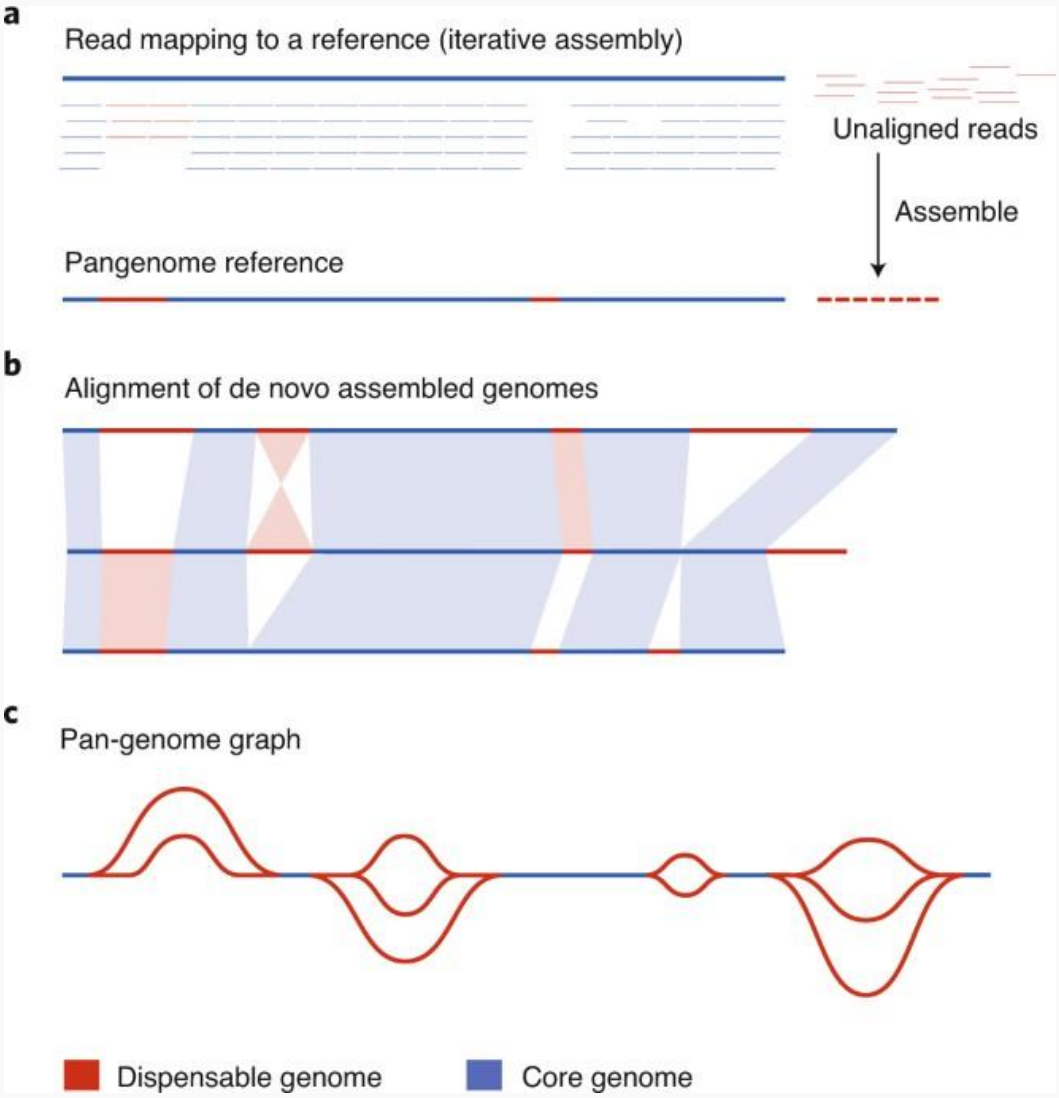
Adaptative sequencing



From Nanopore

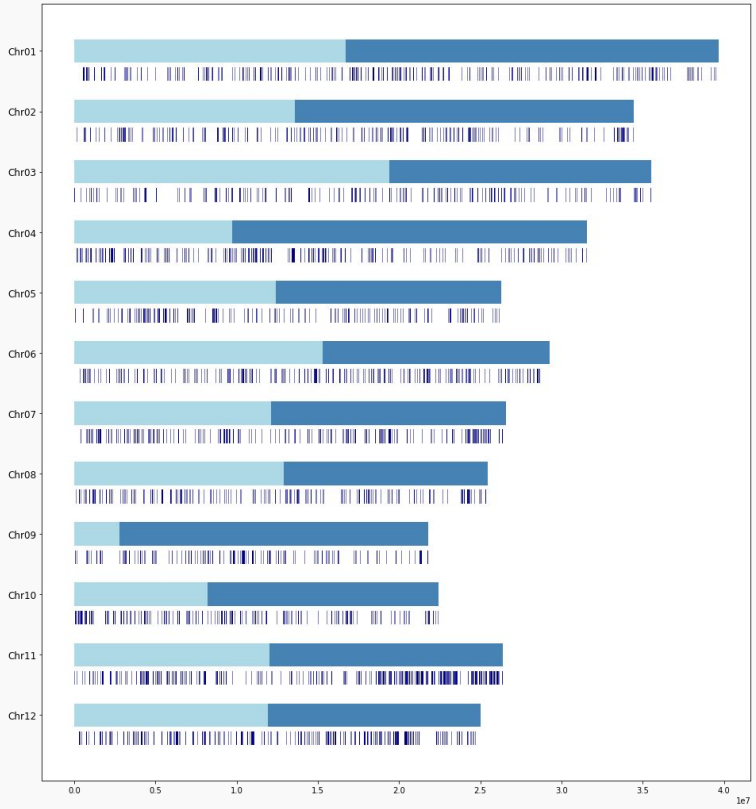
Other ways to analyse diversity...

Pangenomics



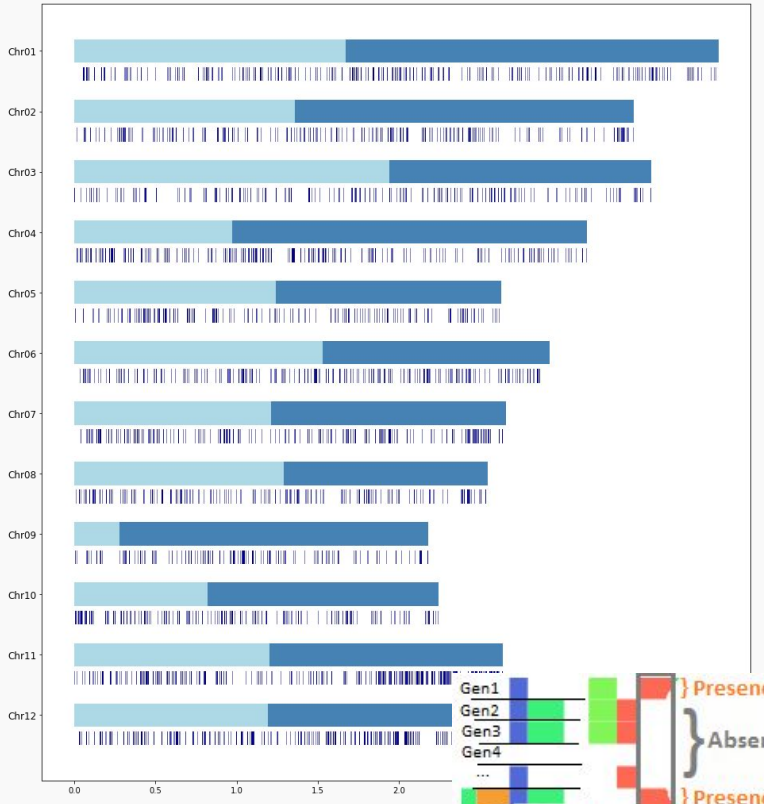
From Bayer et al, 2020

Pangenomics & visualisation

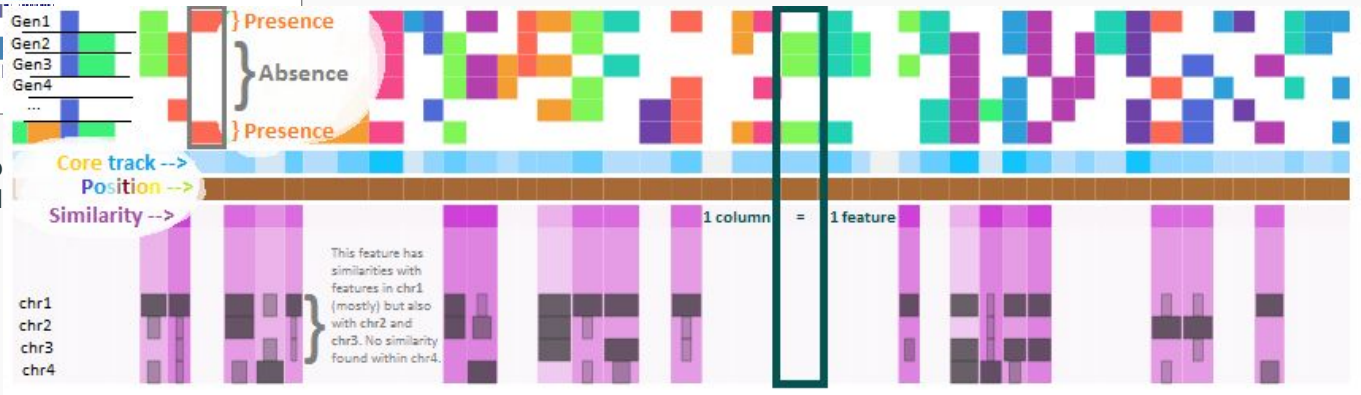


From Tranchant-Dubreuil et al, unpublished

Pangenomics & visualisation

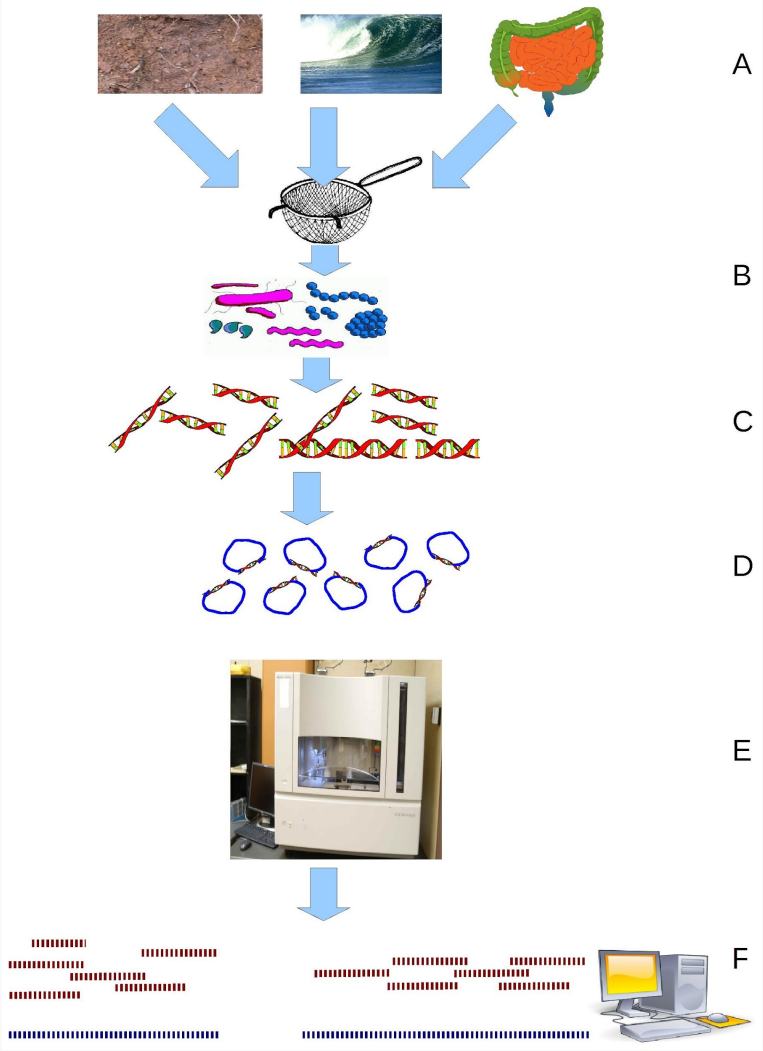


From Tranchant-Dubreuil et al, unpu b
d

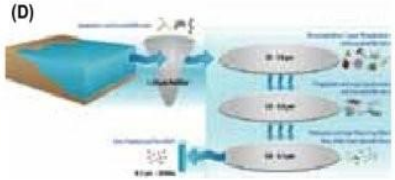


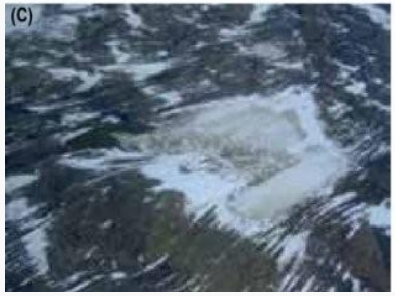
From Durant et al, unpublished

Metagenomics



From Wikipedia





From Yau and Cavicchioli, 2011