https://cibig-wave.github.io

Mentoring Project

October 21, 2024 – November 25, 2024

Theme:

Population structure within *Magnaporthe oryzea*

Attendees

- Attolou Raoul AGNIMONHAN
- **Pakyendou Estel NAME**

Tutors

- Aurore COMTE
- Sébastien RAVEL

PLAN

- ❑ BACKGROUND
- ❑ OBJECTIVES
- ❑ BIOINFORMATIC STRATEGY
- ❑ RESULTS & DISCUSSION
- ❑ CHALLENGES
- ❑ CONCLUSION
- ❑ PERSPECTIVES

BACKGROUND (1/2)

- *Magnaporte oryzae* (synonym of *Pyricularia oryzae*) is *an ascomycete* hemibiotrophic fungus that cause blast disease.
- Blast is one of the most serious and devastating diseases caused by fungal pathogen. As such, it is a major threat to global food security (Khush and Jena 2009).
- Rice Production : 30%–100% loss each year (Mutiga *et al.,* 2021)
- Blast disease is documented on more than 50 cultivated plant species

  rice (*Oryza sativa*), wheat (*Triticum aestivum*),

  barley (*Hordeum vulgare*),

  finger millet (*Eleusine coracana*) etc.

Field damage in the Camargue. Photos: JB Morel, INRAE

BACKGROUND (2/2)

- *M. oryzae*, first plant pathogenic fungus to be completely sequenced (Dean *et al.,* 2005)
- High genetic variation of *M. oryzae* was found in populations from

  South, East and Southeast Asia,

  which is more diverse than in other continents (Zeigler, 1998)
- Study based on microsatellites markers supported this observation

  Asia was the center of origin of *M. Oryzae* (Saleh *et al.*, 2014)
- Advances in high-throughput genomic sequencing have enhanced single nucleotide polymorphism (SNP) discovery.

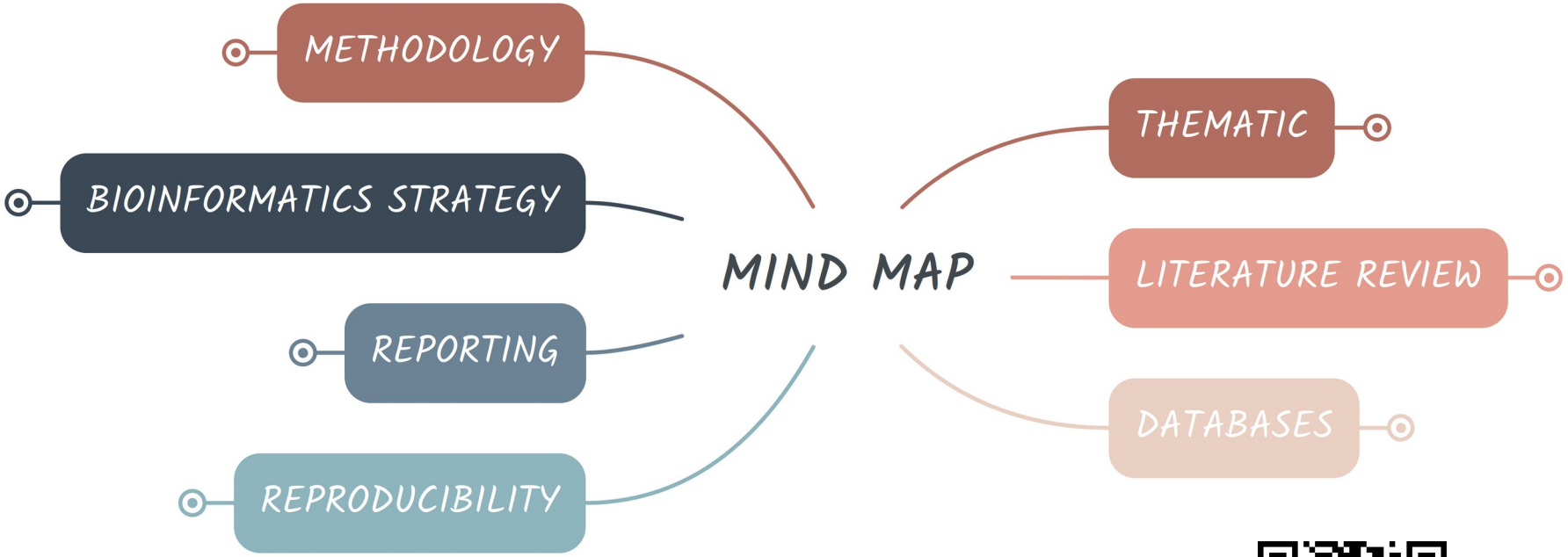
Rice blast symptoms
Diallo, 2021

OBJECTIVES

- We aim to understand the structure of *M. oryzae* populations
  - What is the link between host and population structure?
  - Are there cryptic species within *M.oryzae* ?

MIND MAP

Orgnanization for data analyzis : Mind Map



DATABASES & BIBLIOGRAPHY (1/2)

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa: Pyricularia oryzae (rice blast fungus) | Enter one or more taxonomic names

Filters

Download | Select columns | 392 Genomes | Rows per page 20 | 1-20 of 392

Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
MG8	GCA_000002495.2	GCF_000002495.2	Pyricularia oryzae 70-15	70-15 (strain)	NCBI RefSeq, Submitter	
Br48_v3	GCA_036493215.1		Pyricularia oryzae (rice blast fun...	Br48 (isolate)		
ASM1227299v1	GCA_012272995.1		Pyricularia oryzae (rice blast fun...	LpKY97 (strain)		
ASM434696v1	GCA_004346965.1		Pyricularia oryzae (rice blast fun...	MZ5-1-6 (isolate)	Submitter	
PoP131	GCA_000292605.2		Pyricularia oryzae P131	P131 (strain)		
ASM478572v2	GCA_004785725.2		Pyricularia oryzae (rice blast fun...	B71 (strain)		
ASM3071886v1	GCA_030718865.1		Pyricularia oryzae (rice blast fun...	T3 (strain)		

Assembly statistics

	RefSeq	GenBank
Genome size	41 Mb	41 Mb
Total ungapped length	40.9 Mb	40.9 Mb
Number of chromosomes	7	7
Number of scaffolds	53	53
Scaffold N50	6.6 Mb	6.6 Mb
Scaffold L50	3	3
Number of contigs	216	216
Contig N50	823.6 kb	823.6 kb
Contig L50	13	13
GC percent	51.5	51.5
Assembly level	Chromosome	Chromosome
View sequences	view RefSeq sequences	view GenBank sequences

Accession:

Pyricularia oryzae 70-15

WGS sequencing, assembly, and annotation

The International Rice Blast Genome Consortium, led by the Broad Institute and the Fungal Genomics Laboratory (FGL) at North Carolina State University (NCSU), sequenced the *Magnaporthe oryzae* strain 70-15 genome at >7X coverage using whole genome shotgun (WGS) sequencing. The genome assembly, which corresponds to release 5.0 from the Broad Institute, consists of 739 nuclear genome contigs in 197 scaffolds, and five mitochondrial genome contigs in three scaffolds. It represents the completion of the first stage of finishing. The genome assembly has been annotated using automated gene prediction tools. **Less...**

DATABASES & BIBLIOGRAPHY (2/2)

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Coexistence of Multiple Endemic and Pandemic Lineages of the Rice Blast Pathogen

Pierre Gladieux,[a] Sébastien Ravel,[a] Adrien Rieux,[b] Sandrine Cros-Arteil,[a] Henri Adreit,[a] Joëlle Milazzo,[a] Maud Thierry,[a] Elisabeth Fournier,[a] Ryohei Terauchi,[c] Didie

[a]UMR BGPI, Univ Montpellier, INRA, CIRAD, Montpellier SupAgro, Mc
[b]CIRAD, UMR PVBMT, St. Pierre de la Reunion, France
[c]Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Gene Flow between Divergent Cereal- and Grass-Specific Lineages of the Rice Blast Fungus *Magnaporthe oryzae*

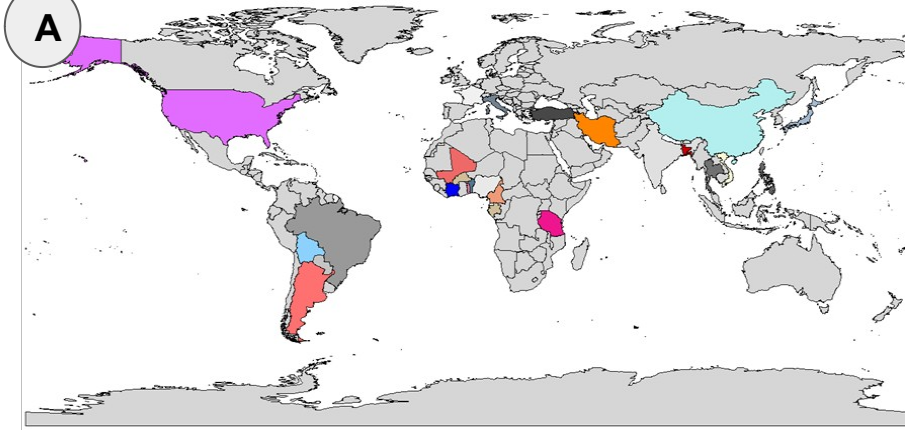
Pierre Gladieux,[a] Bradford Condon,[b] Sebastien Ravel,[a] Darren Soanes,[c] Joao Leodato Nunes Maciel,[d] Antonio Nhani, Jr,[e] Li Chen,[b] Ryohei Terauchi,[f] Marc-Henri Lebrun,[g] Didier Tharreau,[a] Thomas Mitchell,[h] Kerry F. Pedley,[i] Barbara Valent,[j] Nicholas J. Talbot,[c] Mark Farman,[b] Elisabeth Fournier[a]

BIOINFORMATIC STRATEGY (1/3)

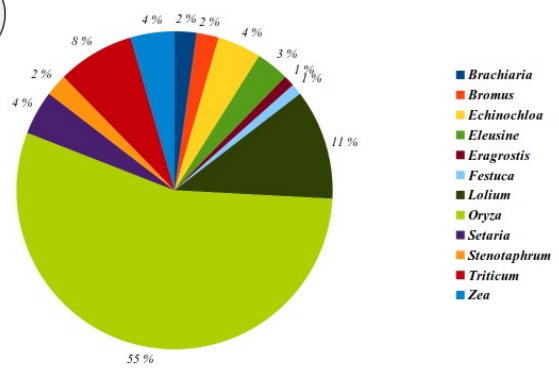
❑ **Datasets**

- Isolates : 89
- Host genus : 12
- Countries : 27
- Year: 1958- 2017
- Sequencing technology: Illumina Miseq, Illumina HiSeq, Solexa & 454.
- Type of sequencing : paired-end

A



B



C

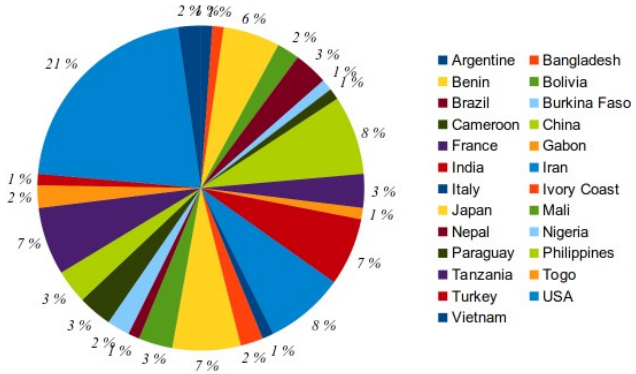
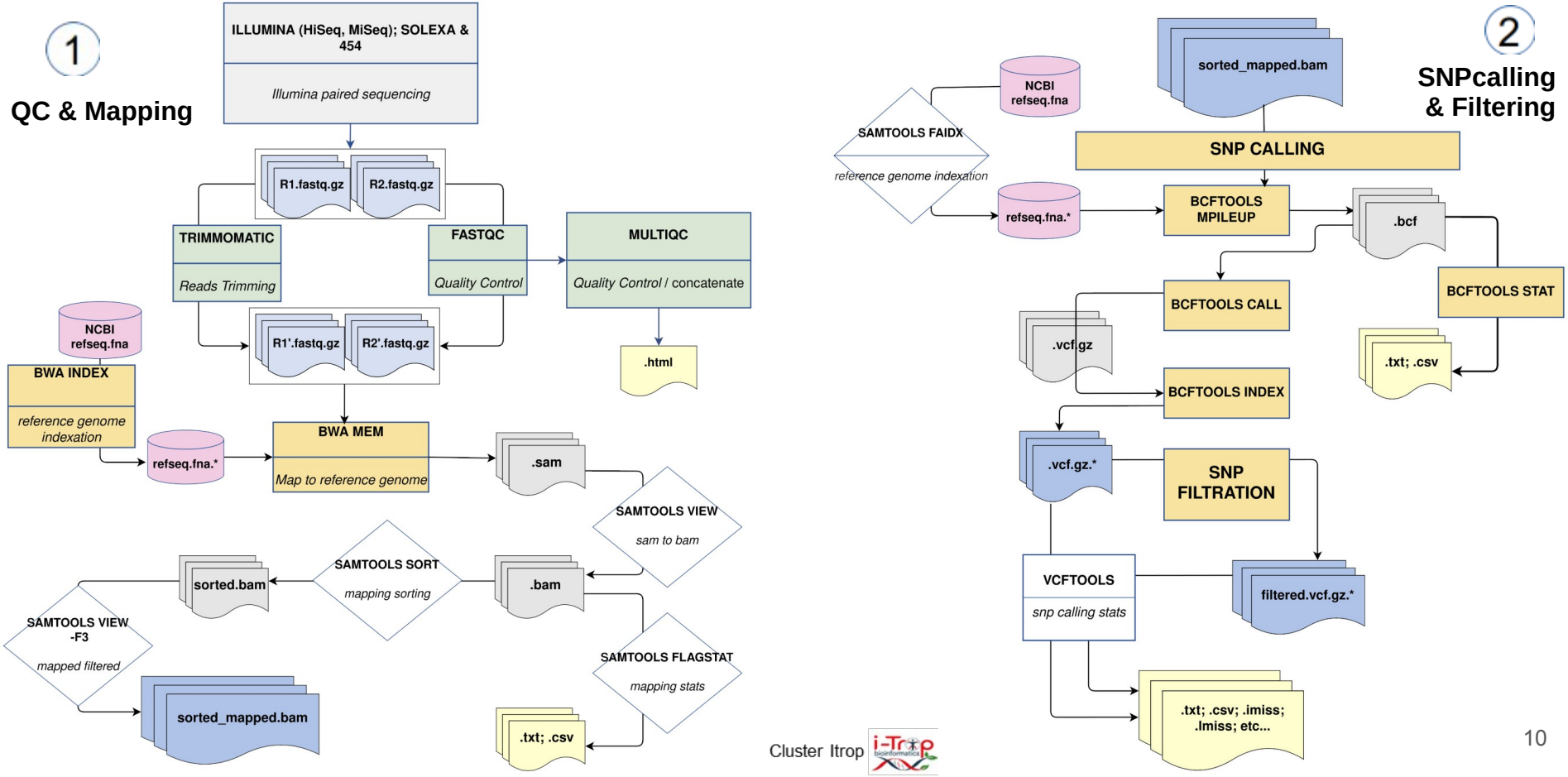


Figure 1 : (A)Map indicating 27 countries where the *M.Oryzae* were collected . Abbreviations (Codes ISO 3166-1) (B)Pie chart showing the number of *M.Oryzae i*solates reported from each country, (C) host plant of the study

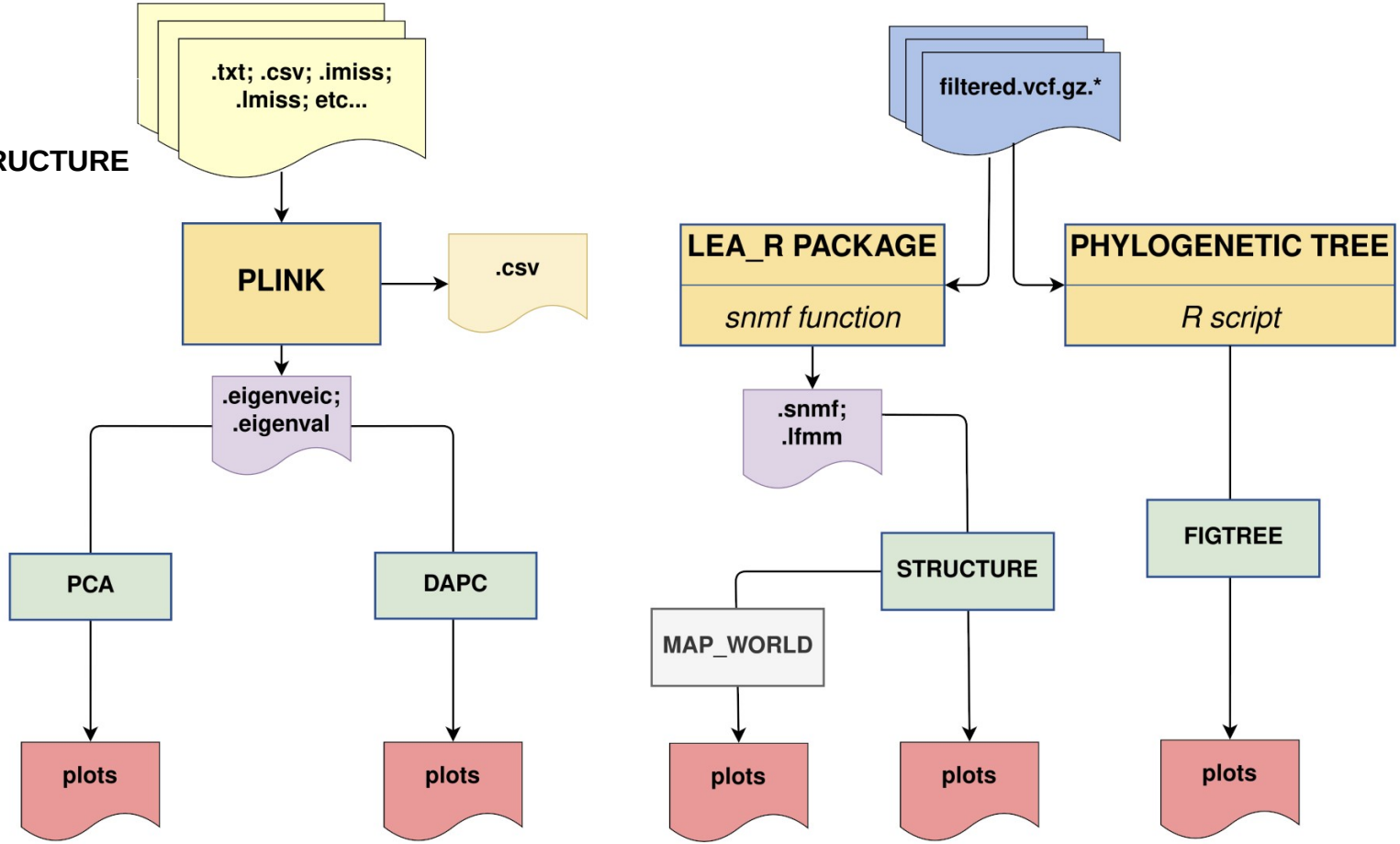
BIOINFORMATIC STRATEGY (2/3)



BIOINFORMATIC STRATEGY (3/3)

3

PCA
DAPC
GENETIC STRUCTURE



RESULTS & DISCUSSION (1/9)

❑ QUALITY CONTROL & TRIMMING

Use Trimmomatic to trim reads with a Phred score > 30 to retain only the best quality reads for further analysis (74% reads treated)

Figure 2 : Quality control of reads a) MultiQC before trimming , b) MultiQC after the trimming

RESULTS & DISCUSSION (2/9)

❑ **MAPPING**

Mapped reads = 77% to 93% and

Unmapped reads= 7% to 23 %

(successfully alignment, good quality and relevance of the sequencing.)

Sequence	total	mapped	%_mapped	unmapped	%_unmapped
AG0004	14592207	12599588	86,3	1992619	13,7
BN0123	9937668	8707660	87,6	1230008	12,4
CH0461	16620893	15201485	91,5	1419408	8,5
G22	6372205	5390226	84,6	981979	15,4
IE1K	8127130	7251751	89,2	875379	10,8
IR0015	17648916	14743456	83,5	2905460	16,5
ML33	4123811	3592671	87,1	531140	12,9
PH42	2138588	1819713	85,1	318875	14,9
TN0057	15170565	13331011	87,9	1839554	12,1
Arcadia	2395926	2073528	86,5	322398	13,5
BN0202	10871986	9615772	88,4	1256214	11,6
CH0533	11543000	10459604	90,6	1083396	9,4
GFSI1-7-2	47816125	36818519	77,0	10997606	23,0
IN0017	7200346	6212309	86,3	988037	13,7
IR0083	14051512	11985467	85,3	2066045	14,7
NG0012	9486090	8263271	87,1	1222819	12,9
PL2-1	6689378	5727658	85,6	961720	14,4
TN0065	16045920	13915802	86,7	2130118	13,3
B2	464467	401904	86,5	62563	13,5
BN0252	15067680	13213012	87,7	1854668	12,3
CH1103	10036214	9113996	90,8	922218	9,2
GG11	1187056	1003768	84,6	183288	15,4
IN0054	15949676	14099140	88,4	1850536	11,6
IR0084	20356013	16769886	82,4	3586127	17,6
NG0054	10590503	9318184	88,0	1272319	12,0
SSFL02	3387358	2835242	83,7	552116	16,3
TN0090	13977922	12330440	88,2	1647482	11,8
B71	19392697	16849194	86,9	2543503	13,1
Br7	6545075	5636467	86,1	908608	13,9
CH1164	8323560	7741468	93,0	582092	7,0
GN0001	15763678	13224279	83,9	2539399	16,1

RESULTS & DISCUSSION (3/9)

❑ SNP CALLING & FILTERING

```
[3]key  [4]value
number of samples:      89
number of records:      1613204
number of no-ALTs:      0
number of SNPs: 1535749
number of MNPs: 0
number of indels:       77455
number of others:       0
number of multiallelic sites:   1413
number of multiallelic SNP sites:       347
```

Raw SNP calling

Filter 1

Less stringent mode SNP filtering
MAF = 0.05
MISS = 0.95
QUAL= 5000

Filter parameter
Allele frequency
Mean depth per individual
Mean depth per site
Site quality
Proportion of missing data per individu
Heterozygosity and inbreeding coefficient per individual

R

```
[3]key  [4]value
number of samples:      89
number of records:      4914
number of no-ALTs:      0
number of SNPs: 4914
number of MNPs: 0
number of indels:       0
number of others:       0
number of multiallelic sites:   264
number of multiallelic SNP sites:       264
```

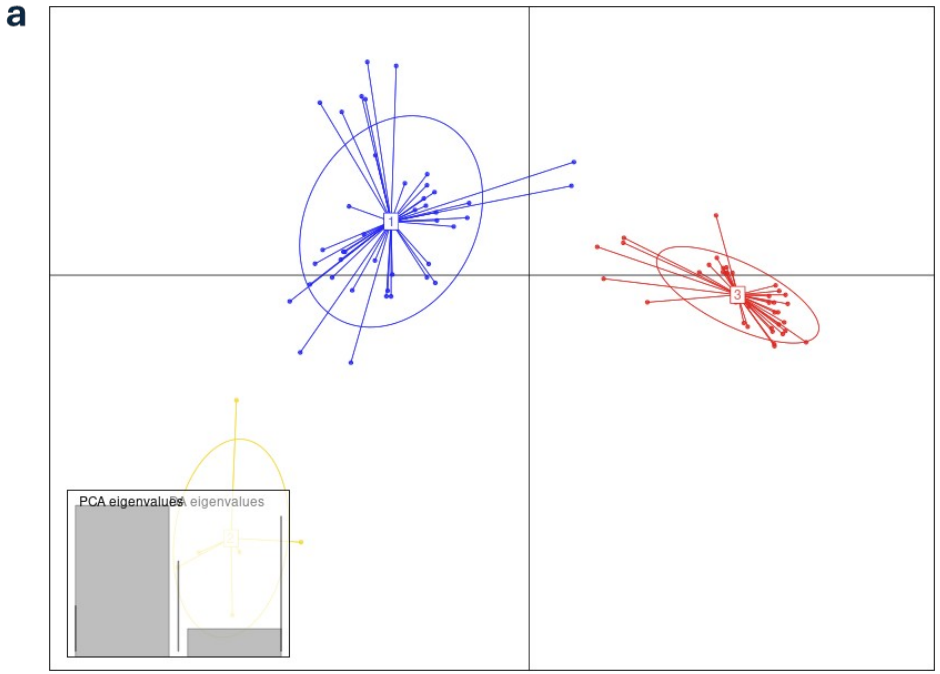
Filter 2

String mode SNP filtering
MAF = 0.1
MISS = 0.9
QUAL= 19000

```
[3]key  [4]value
number of samples:      89
number of records:      795
number of no-ALTs:      0
number of SNPs: 795
number of MNPs: 0
number of indels:       0
number of others:       0
number of multiallelic sites:   66
number of multiallelic SNP sites:       66
```

RESULTS & DISCUSSION (4/9)

- ❑ **DAPC & PCA using Total SNPs**



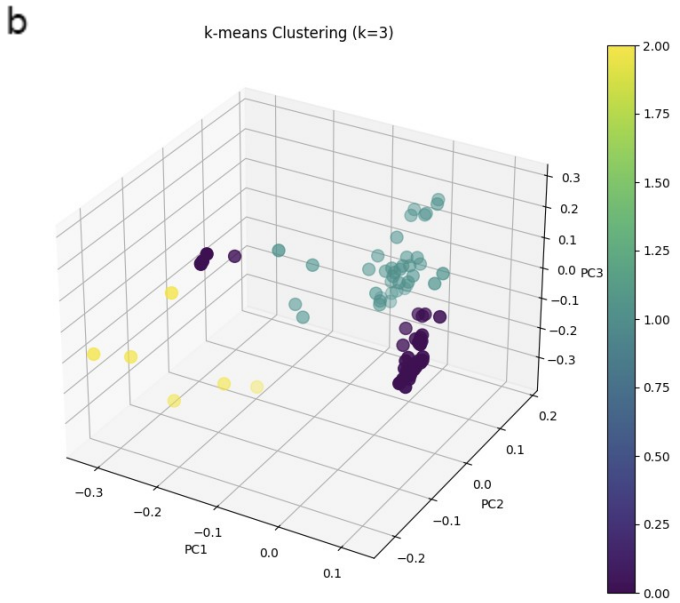


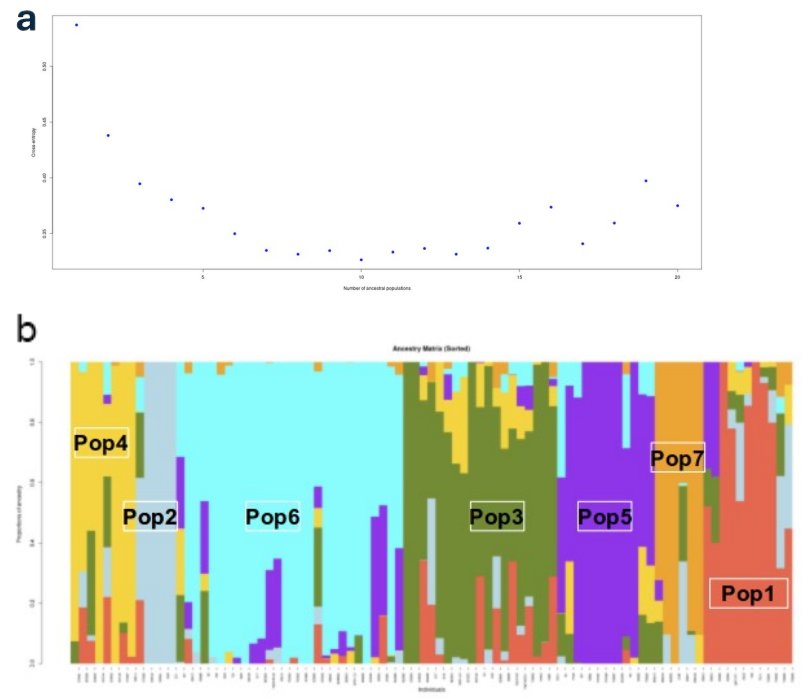
Figure 3 : (a) A DAPC plot showing 3 clusters; (b) PCA plot with k-means clustering (k=3)

RESULTS & DISCUSSION (5/9)

- **GENETIC STRUCTURE using R LEA package (snmf function)**

String mode SNP filtering

- The optimal *K* = 7 (Evanno's method)



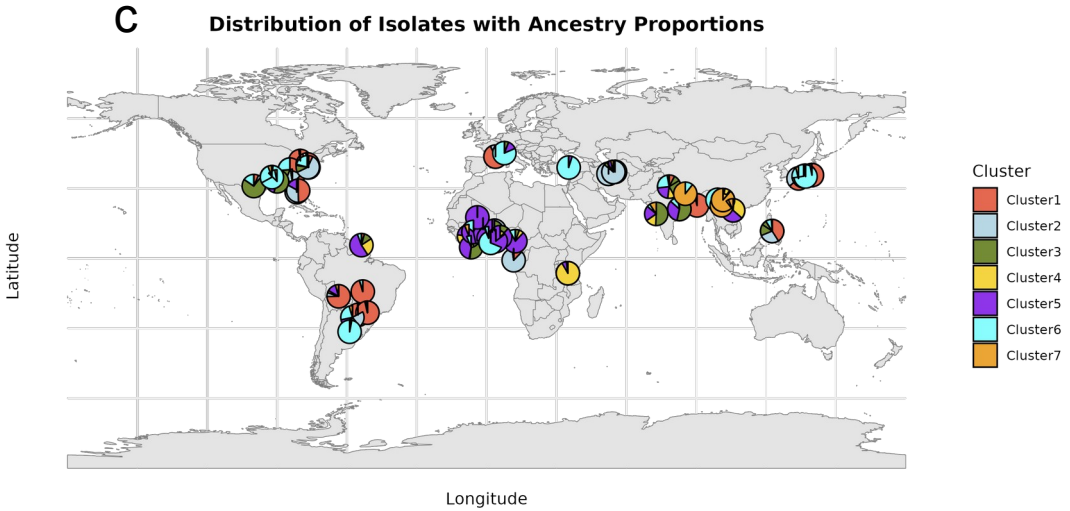


Figure 4 : (a) A plot of cross-entropy versus the number of ancestral populations K,, where K= 1-20 (b) graphical display of the population structure of *M.Oryzae* with stringent filtering (SNP_filter :795)); (c) Map showing distribution of isolates with Ancestry proportions

RESULTS & DISCUSSION (6/9)

String mode SNP filtering

❑ DAPC & PCA using filtering SNPs

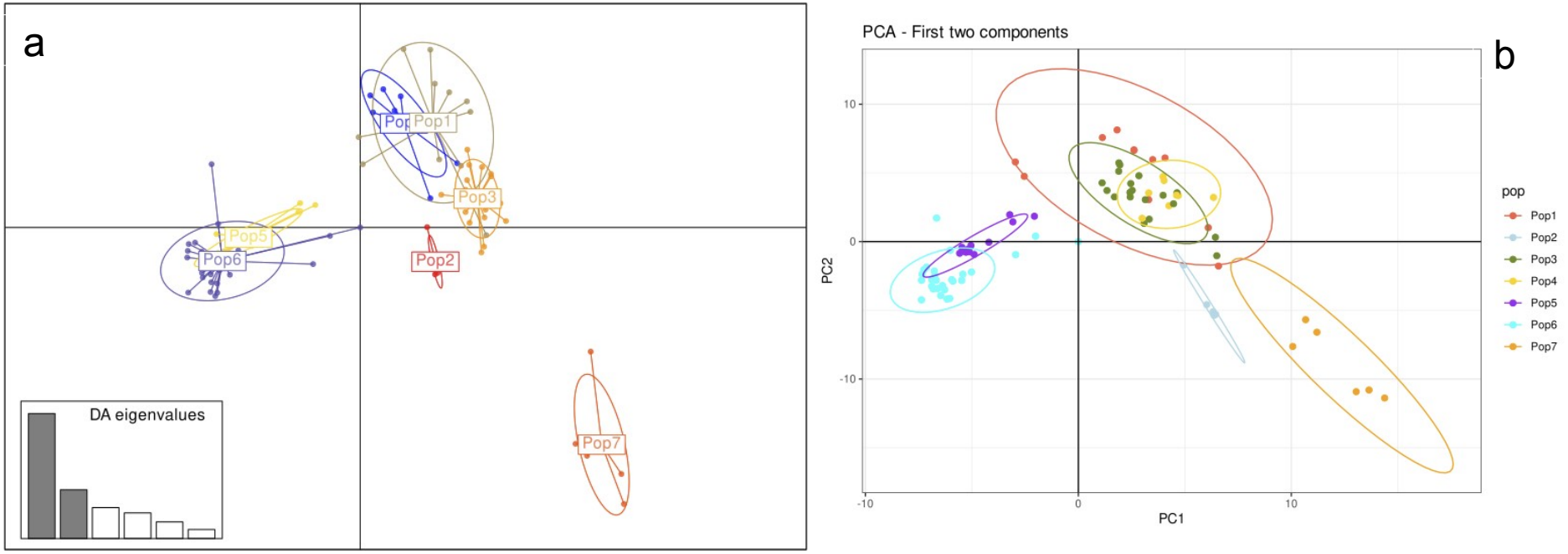


Figure 5 : DAPC (a) and PCA (b) to examine the number of clusters or groups. Individuals with the same color belong to the same group. DAPC & PCA generated by package R adegenet and custom python script

RESULTS & DISCUSSION (7/9)

String mode SNP filtering

1	2	3	4	5	6	7
BF0072	IB49	BN0119	AG0004	Bm88324	Arcadia	CH0043
BN0019	TN0001	BN0123	BN0202	GFSI1-7-2	B2	CH0072
BN0252	TN0002	CD0065	CH0533	GN0001	B71	CH0461
CM0028	TN0065	CD0142	IA1	IR00102	Bd8401	CH1103
GY0040	TN0090	CH0452	IN0116	IR0013	BdBar	CH1164
ML0060		IB33	INA168	IR0015	Br7	NP0058
ML33		IC17	IT0010	IR0083	Br80	
NG0012		IE1K	TR0025	IR0084	CHRF	
NG0054		IN0017		IR0088	CHW	
VT0027		IN0054		IR0095	FR1067	
VT0030		IN0059		JP0091	FR1069	
		IN0114		US0064	G17	
		IN0115			G22	
		ML0062			GG11	
		TG0004			HO	
		TG0032			LpKY-97-1	
		TN0050			P28	
		TN0057			P29	
		US0041			P3	
					Pg1213-22	
					PgKY4OV2-1	
					PgPA18C-02	
					PH42	
					PL2-1	
					SSFL02	
					SSFL14-3	
					T25	
					Z2-1	

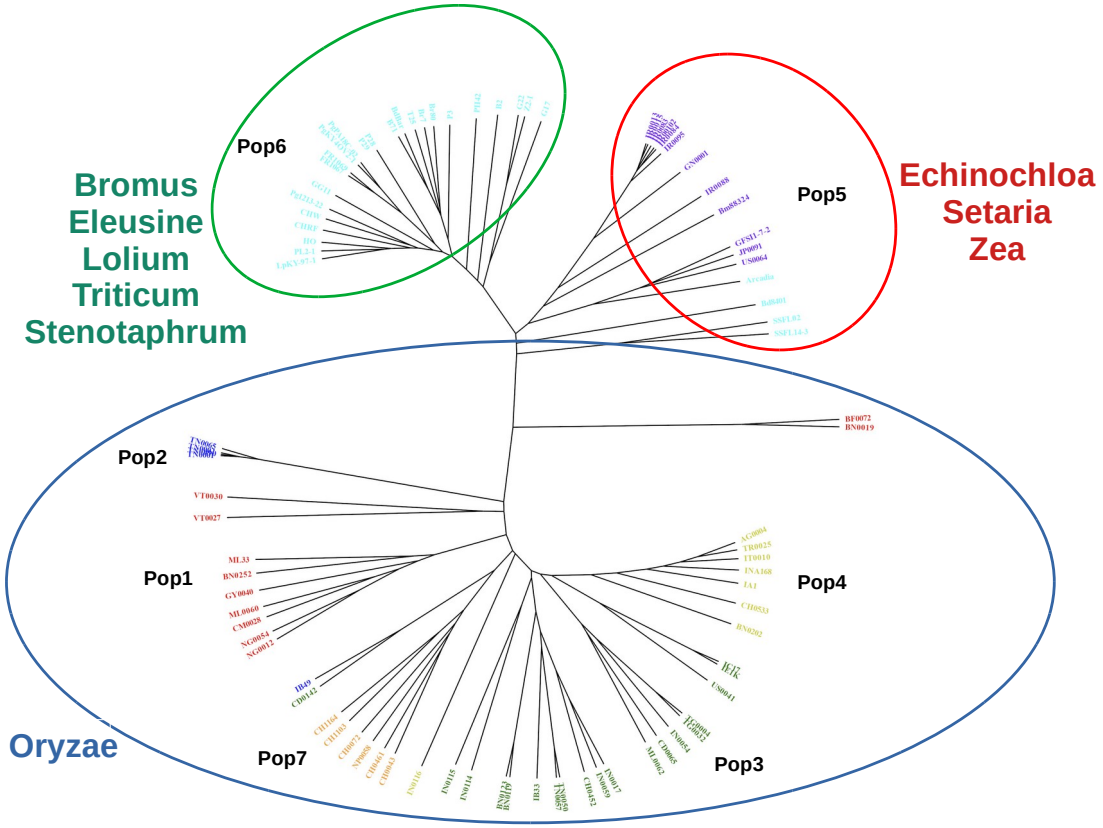


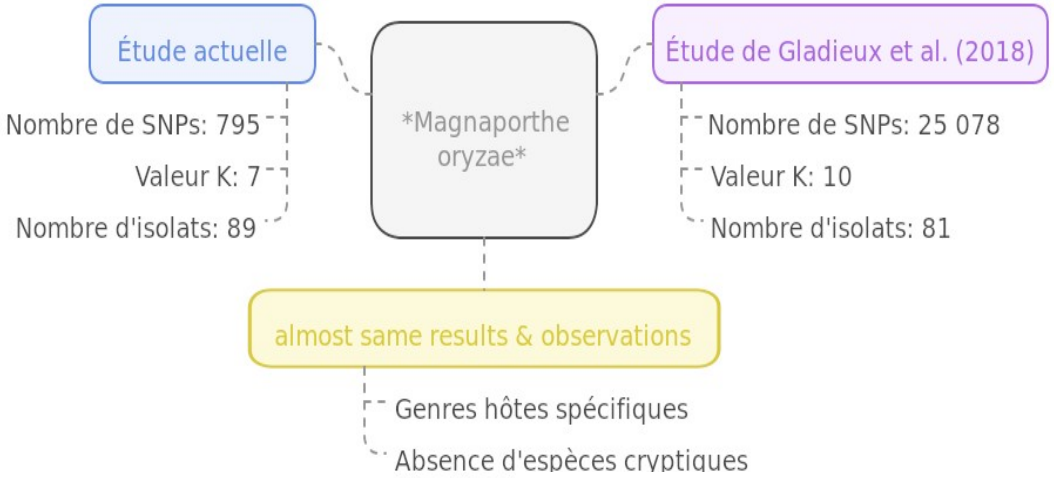
Figure 6 : Phylogenetic tree

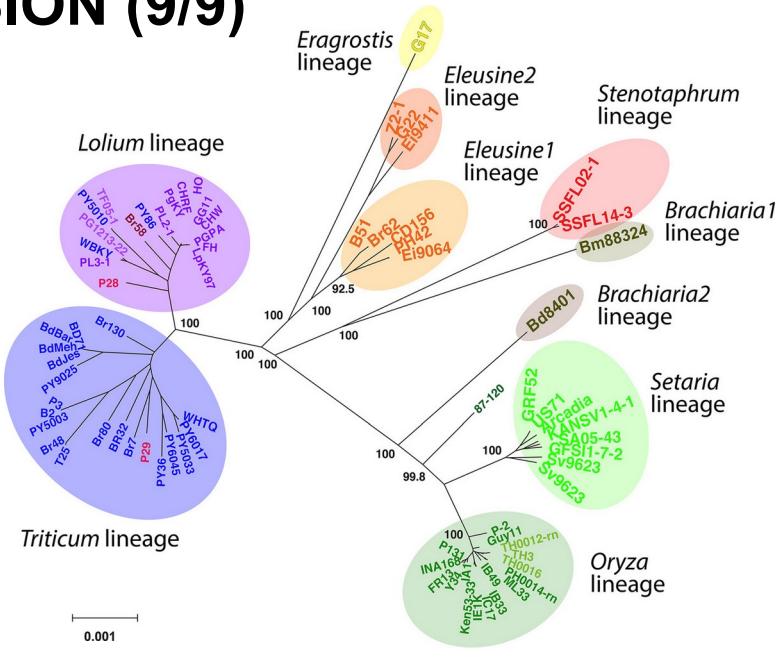
RESULTS & DISCUSSION (8/9)

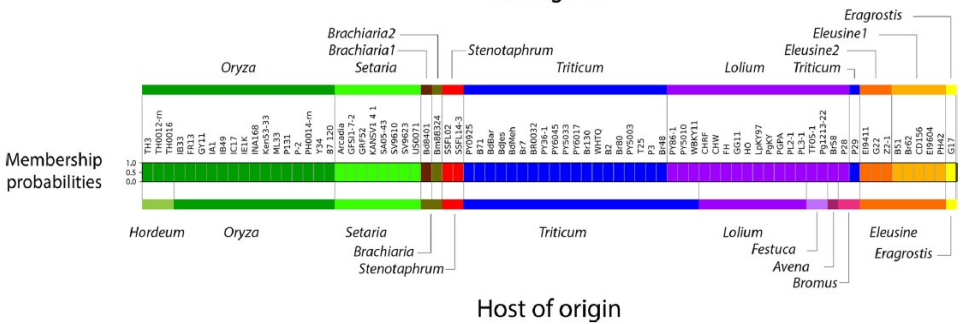
String mode SNP filtering

- Genetic diversity reduced among *M. oryzae* isolates
- Correlation observed between pathogen and host, although relatively small
    - *relatively small genetic differences between isolates remain*
    - *regardless of country of origin*
- Strong genetic proximity between rice isolates
- Isolates from other hosts also tend to form distinct groups.
    - Not validating the existence of well-defined cryptic species within M. oryzae
- These observations suggest that the host plays a key role in structuring *M. oryzae* populations,
    - *but with no marked separation between countries*

RESULTS & DISCUSSION (9/9)







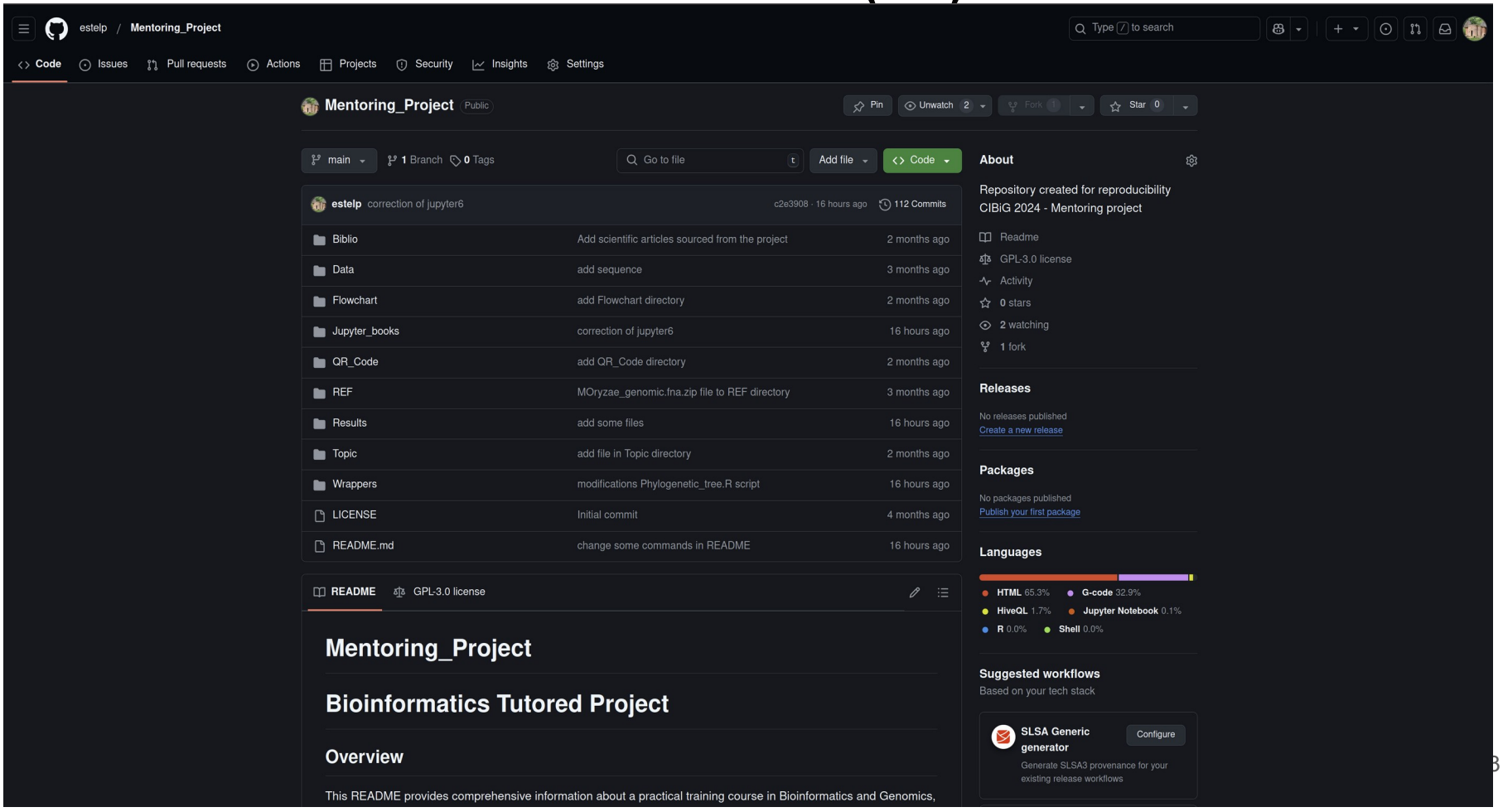
CHALLENGES

- ❑ Lack of time for optimal management of different steps
  - ❑ limits the depth of certain analyses and the quality
- ❑ New and complex topic to understand
  - ❑ considerable time for learning and experimentation
- ❑ Multiplicity of research axes with SNPs
  - ❑ risk of dispersing efforts, potentially compromising the depth of priority analyses.

CONCLUSION (1/2)

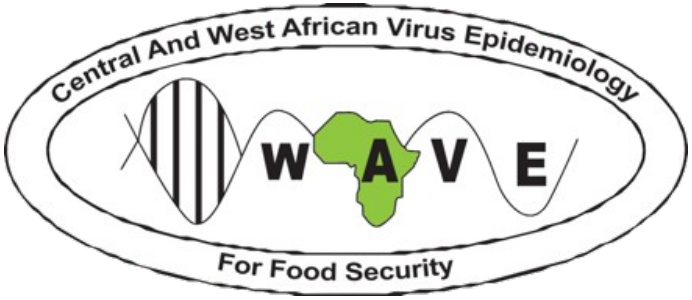
- The strategy adopted is different from that of the 2 publications, but the results are the same.
- It is very useful to perform different strategies to confirm a result
- The objective of this tutored project was achieved
- Production of a Github page
  - (https://github.com/estelp/Mentoring_Project/tree/main)
- Other aspects remain to be explored to better address objectives of this project

CONCLUSION (2/2)



PERSPECTIVES

- ❑ Explore other aspects of the study
- ❑ Perform mapping
- ❑ Optimize SNP filtering
- ❑ Organize and improve the GitHub repository
- ❑ Extend the project to produce a Snakefile



THANKS YOU FOR YOUR ATTENTION