## Exploring Genetic Diversity: Big Data and Reproducibility Challenges in **Population Genomics**

## Louis OLLIVIER<sup>1,2</sup>, Sarah COHEN-BOULAKIA<sup>1</sup>, Gilles FISCHER<sup>2</sup>, Fanny POUYET<sup>1</sup>

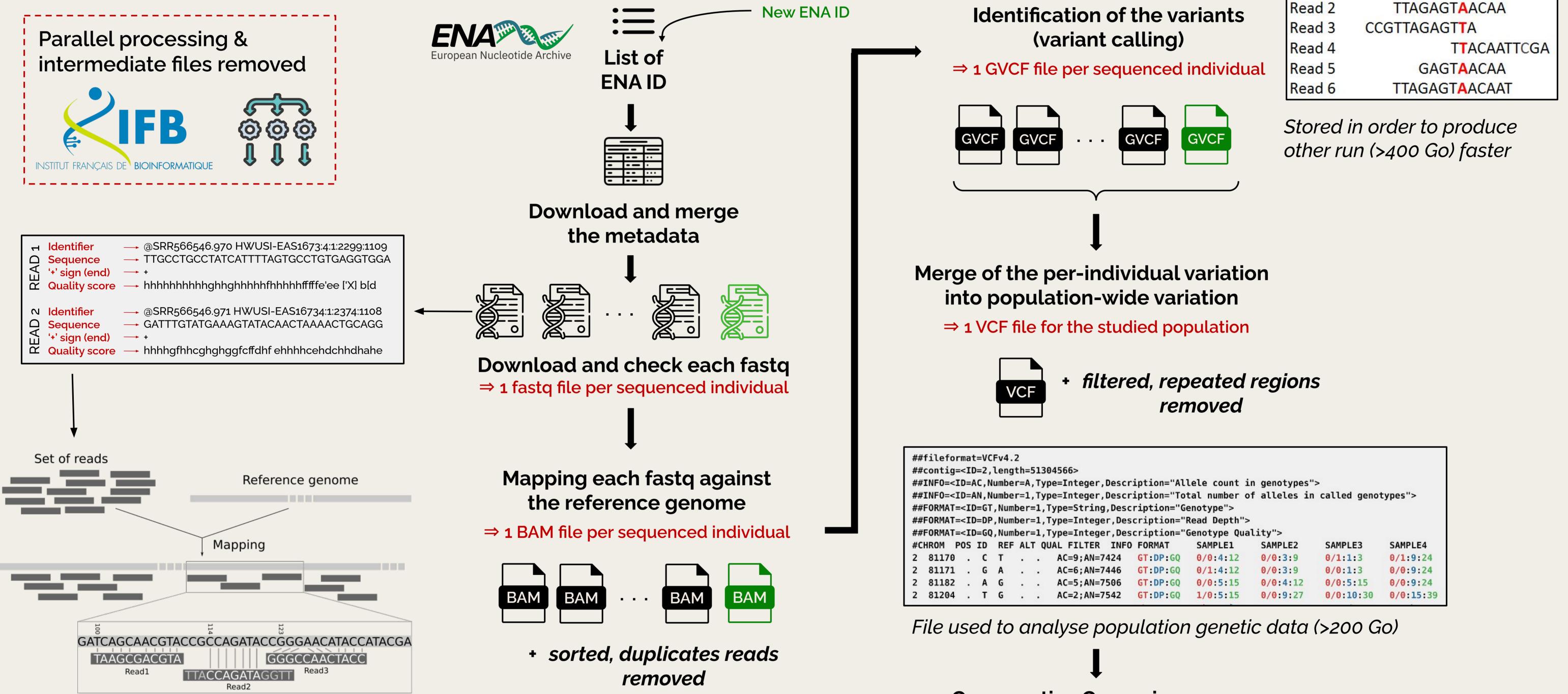
<sup>1</sup> Université Paris-Saclay, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), UMR 9015, Équipe BioInfo <sup>2</sup> Sorbonne Université, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), UMR 7238, Équipe Biologie des Génomes

## **1** INTRODUCTION

**DNA**, the **genetic** blueprint of every organism, varies between individuals, influencing traits and disease susceptibility. High-throughput technologies like Illumina [1] sequencing generate extensive datasets of short DNA reads, necessitating efficient assembly and alignment for variant analysis. Managing this **big data** is crucial, especially when studying **populations** rather than individual genomes. Robust variant calling pipelines are essential for extracting meaningful insights from large-scale genomic data, advancing our understanding of genetic diversity and disease genetics. Here, I present a variant calling pipeline initially made for the analysis of the yeast Saccharomyces cerevisae population.



The availability of over 3,000 fully sequenced Saccahrmyces cerevisae genomes (12Mbp each, ~10 To of data in total) using Illumina's short-read (~ 150 bp) technology presents a significant optimization challenge. Our robust pipeline efficiently processes this large-scale dataset, ensuring scalability to incorporate new available data efficiently while accurately identifying genetic variants.

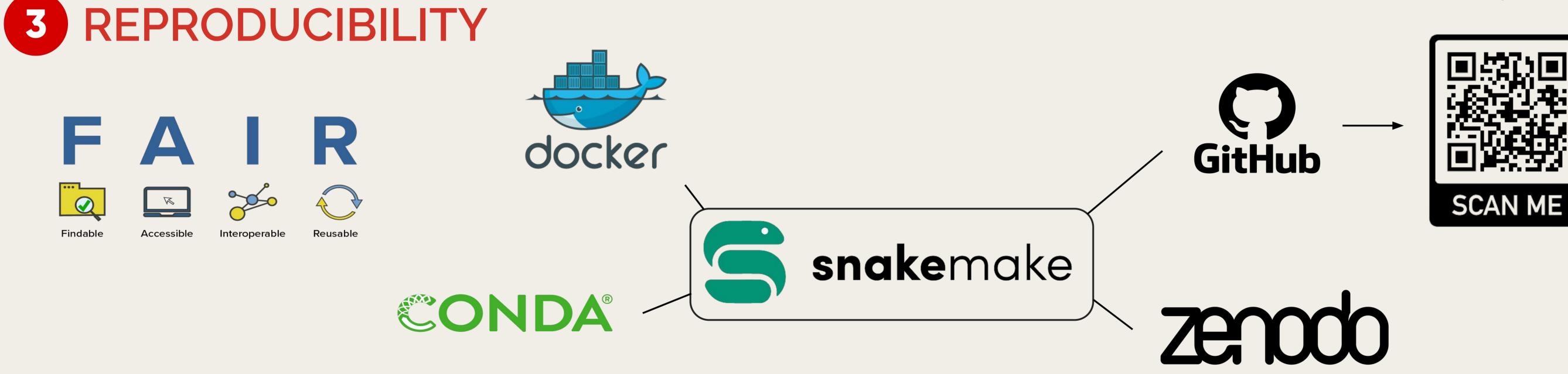


BAM	BAM		BAM	BAM

##	##INFO= <id=ac,number=a,type=integer,description="allele count="" genotypes"="" in=""></id=ac,number=a,type=integer,description="allele>												
##INF0= <id=an,number=1,type=integer,description="total alleles="" called="" genotypes"="" in="" number="" of=""></id=an,number=1,type=integer,description="total>													
##FORMAT= <id=gt,number=1,type=string,description="genotype"></id=gt,number=1,type=string,description="genotype">													
##FORMAT= <id=dp,number=1,type=integer,description="read depth"=""></id=dp,number=1,type=integer,description="read>													
##FORMAT= <id=gq,number=1,type=integer,description="genotype quality"=""></id=gq,number=1,type=integer,description="genotype>													
#C	HROM P	05	ID	REF	AL	T QUA	L FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2	SAMPLE3	SAMPLE4
2	81170		С	Т			AC=9; AN=7	424	GT:DP:GQ	0/0:4:12	0/0:3:9	0/1:1:3	0/1:9:24
2	81171		G	Α		. 1	AC=6; AN=7	446	GT:DP:GQ	0/1:4:12	0/0:3:9	0/0:1:3	0/0:9:24
2	81182		Α	G			AC=5;AN=7	506	GT:DP:GQ	0/0:5:15	0/0:4:12	0/0:5:15	0/0:9:24
2	81204		т	G			AC=2 ; AN=7	542	GT:DP:GQ	1/0:5:15	0/0:9:27	0/0:10:30	0/0:15:39

Reference CCGTTAGAGTTACAATTCGA

- Comparative Genomics
- Variant Annotation and Interpretation (diseases)
- Population Genetics and Evolutionary Studies





The integration of high-throughput sequencing data from over 3,000 genomes presents substantial challenges in data processing and optimization. Our robust and scalable pipeline addresses these issues by efficiently handling large datasets, ensuring accurate variant identification. By prioritizing reproducibility and resource efficiency, we can reliably analyze genetic diversity and accommodate the continuous influx of new genomic data. Additionally, automating the search for new ENA IDs will further increase the population size, advancing our understanding of population genomics.

## REFERENCES

[1] https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html IFB: https://www.france-bioinformatique.fr/cluster-ifb-core/ ENA: https://www.ebi.ac.uk/ena/browser/home Github: <u>https://github.com/ & https://github.com/Louis-XIV-bis/varcall\_snakemake</u> Conda: https://conda.io/projects/conda/en/latest/index.html Docker: <u>https://www.docker.com/</u>



This work has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF

