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

Louis OLLIVIER[1,2], Sarah COHEN-BOULAKIA[1], Gilles FISCHER[2], Fanny POUYET[1]

[1] Université Paris-Saclay, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), UMR 9015, Équipe BioInfo

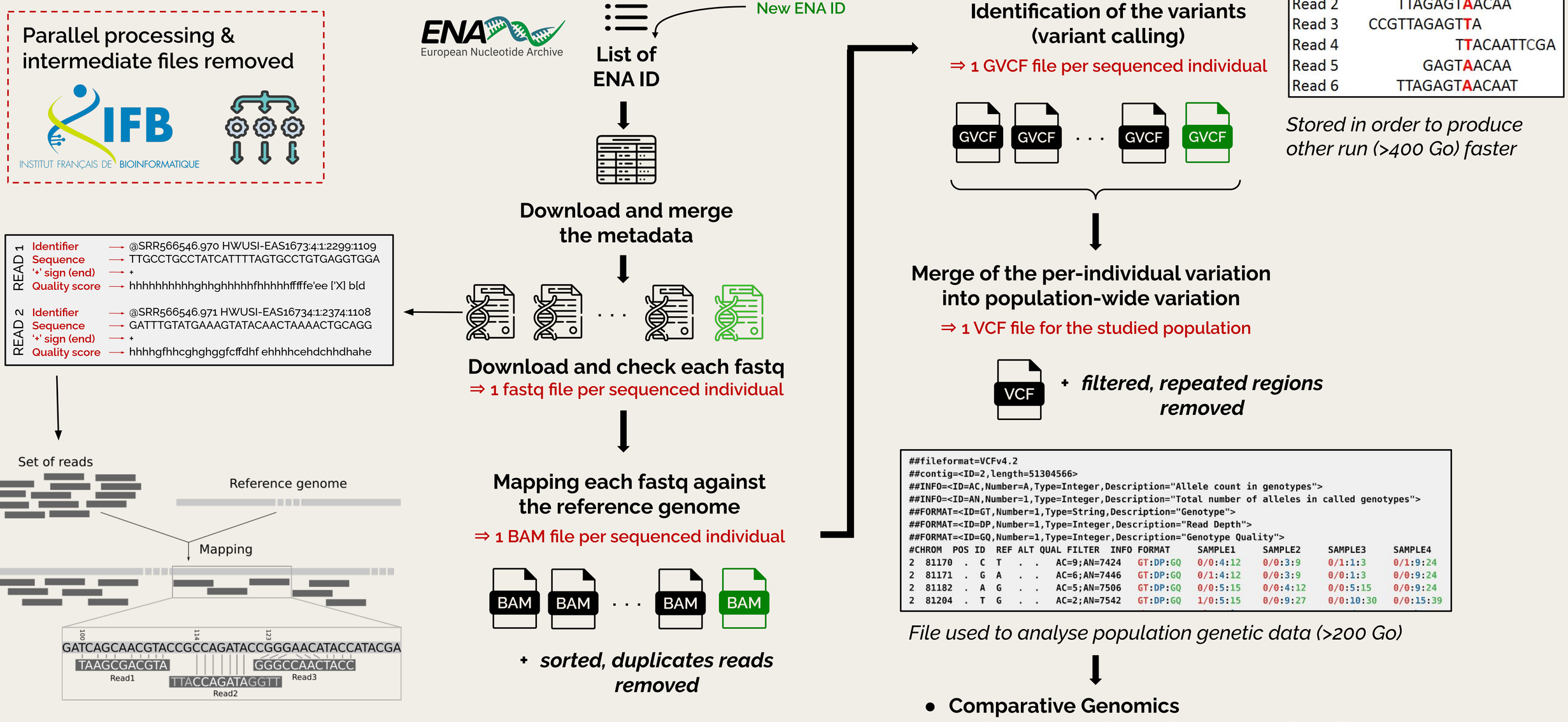
[2] Sorbonne Université, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), UMR 7238, Équipe Biologie des Génomes

1 INTRODUCTION

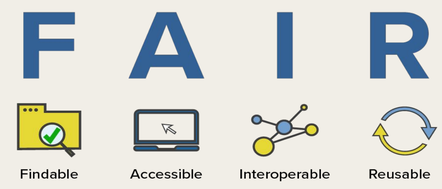
DNA, the **genetic** blueprint of every **organism**, v**aries between individual**s, influencing traits and **disease susceptibility**. **High-throughput** technologies like **Illumina** [1] sequencing generate **extensive datasets** of **short DNA reads**, necessitating efficient **assembly** and **alignment** for **variant analysis**. Managing this **big data** is crucial, especially when studying **populations** rather than i**ndividual genomes**. **Robust** variant calling **pipelines** are essential for **extracting** meaningful insights from **large-scale genomic dat**a, 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**.

2 PIPELINE FOR GENOMIC DATA EXPLORATION

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**.



3 REPRODUCIBILITY



4 CONCLUSION

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: https://github.com/ & https://github.com/Louis-XIV-bis/varcall_snakemake
Conda: https://conda.io/projects/conda/en/latest/index.html
Docker: https://www.docker.com/

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