Introducing nf-core/ phaseimpute from idea to release

Louis LE NÉZET[*], Anabella TRIGILLA[#], Pascale QUIGNON[*], Catherine ANDRÉ[*]
[*]Université de Rennes, CNRS, Institut de Génétique & Développement de Rennes, UMR 6290 CNRS - Rennes, France
[#] ZS Discovery, Argentina

INTRODUCTION

Genome imputation is a statistical technique that enhances the resolution of genotyping arrays and low-pass sequencing (<1x) by filling missing data with information from reference panels. While existing pipelines primarily focus on the imputation step and in the human species, crucial steps such as panel preparation, phasing, and imputation assessment are often overlooked.

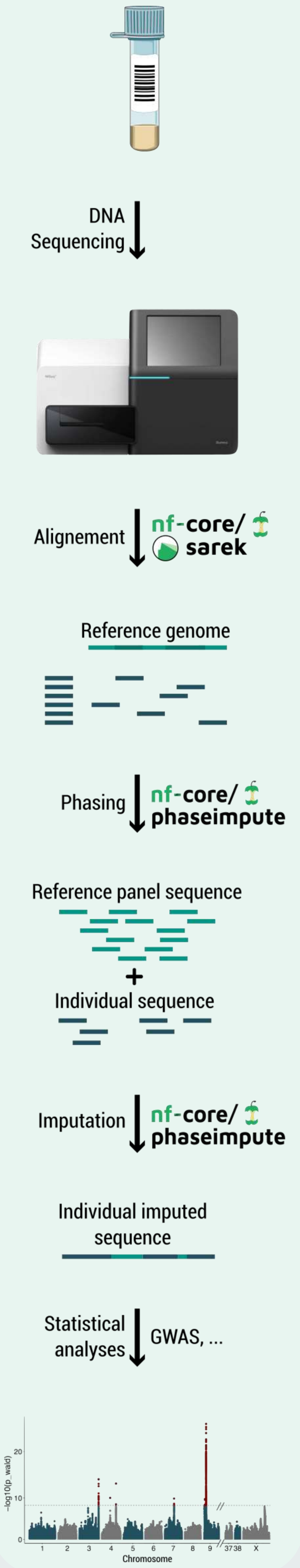
To address this gap, we introduce nf-core/phaseimpute, a comprehensive pipeline performing panel preparation, genomic data simulation, imputation, and tool assessment. Each step is designed for independent execution, enabling users to save outputs and computational time for subsequent analysis. It offers flexibility by allowing execution with or without reference panels, making it invaluable for non-model species where phased haplotypes may not always be available.

The journey from the initial idea to the first release of nf-core/phaseimpute has been an extensive one. We benefited from advancements made in Nextflow plugins, such as nf-validation and nf-test, to enforce schema validation and to ensure that each update maintains the pipeline's accuracy and stability.

OBJECTIVES

- Easy pre-processing of data: panel normalisation, phasing, filtering
- Imputation of different data with different tools: low-pass (GLIMPSE 1 & 2, Stitch, Quilt), SNP chip array (Beagle 5, Impute 5)
- Simulation of data and validation process

WORKFLOW



DATA INPUTS & OUTPUTS

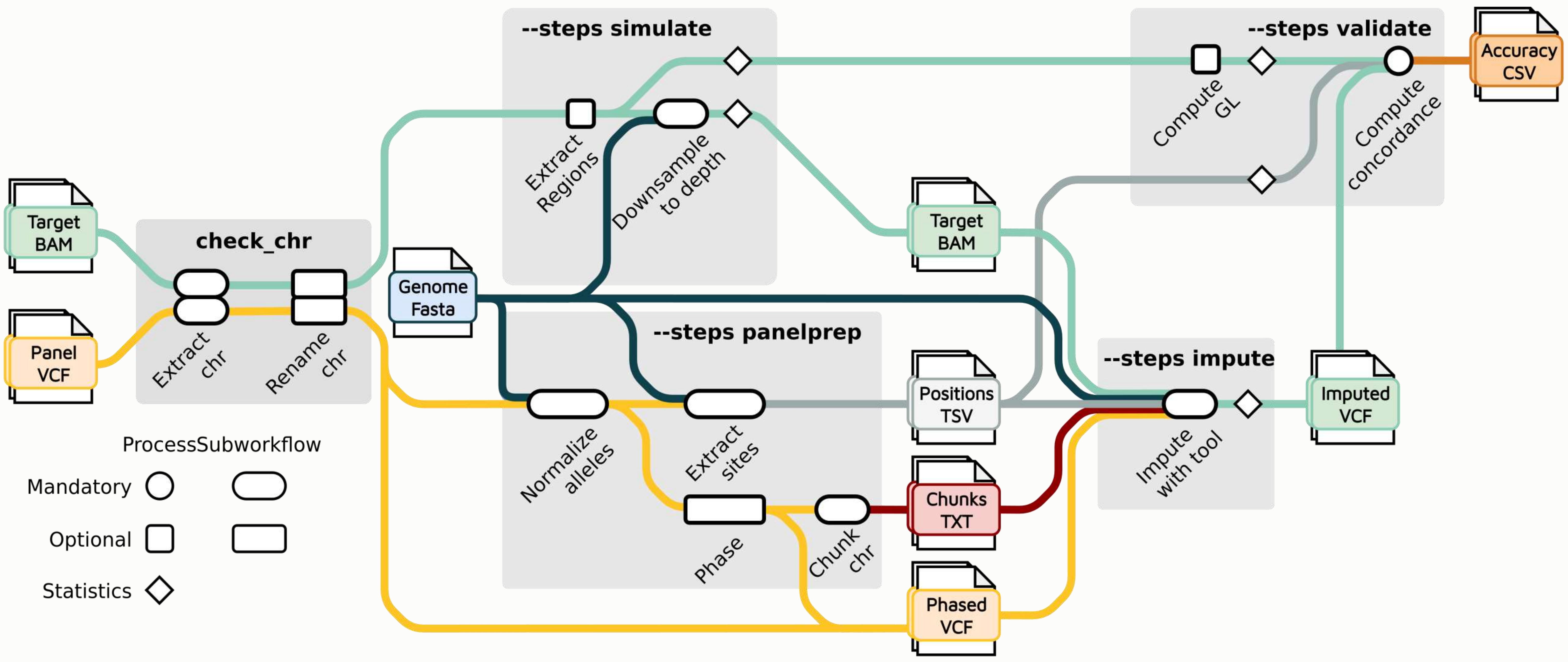
To use the « phaseimpute » pipeline you need:

- `--input` : Target individual's BAM files (id, bam, bai)
- `--genome` or `--fasta` : Reference genome fasta
- `--reference` : Phased reference panel VCF (id, chr, vcf, index)
- Other parameters specified in `.json` files

STEPS

The pipeline consists of 4 steps that can be run independently or together :

- `--steps panelprep` : normalize, extract and phase reference variants
- `--steps impute` : impute target bam files with ≠ tools (`--tools <glimpse1,glimpse2,quilt,stitch>`)
- `--steps simulate` : downsample bam files given in `--input` to the `--depth` specified
- `--steps validate`: compute imputation accuracy between imputed files and `--input_truth` (if present, if not use `--input` files)

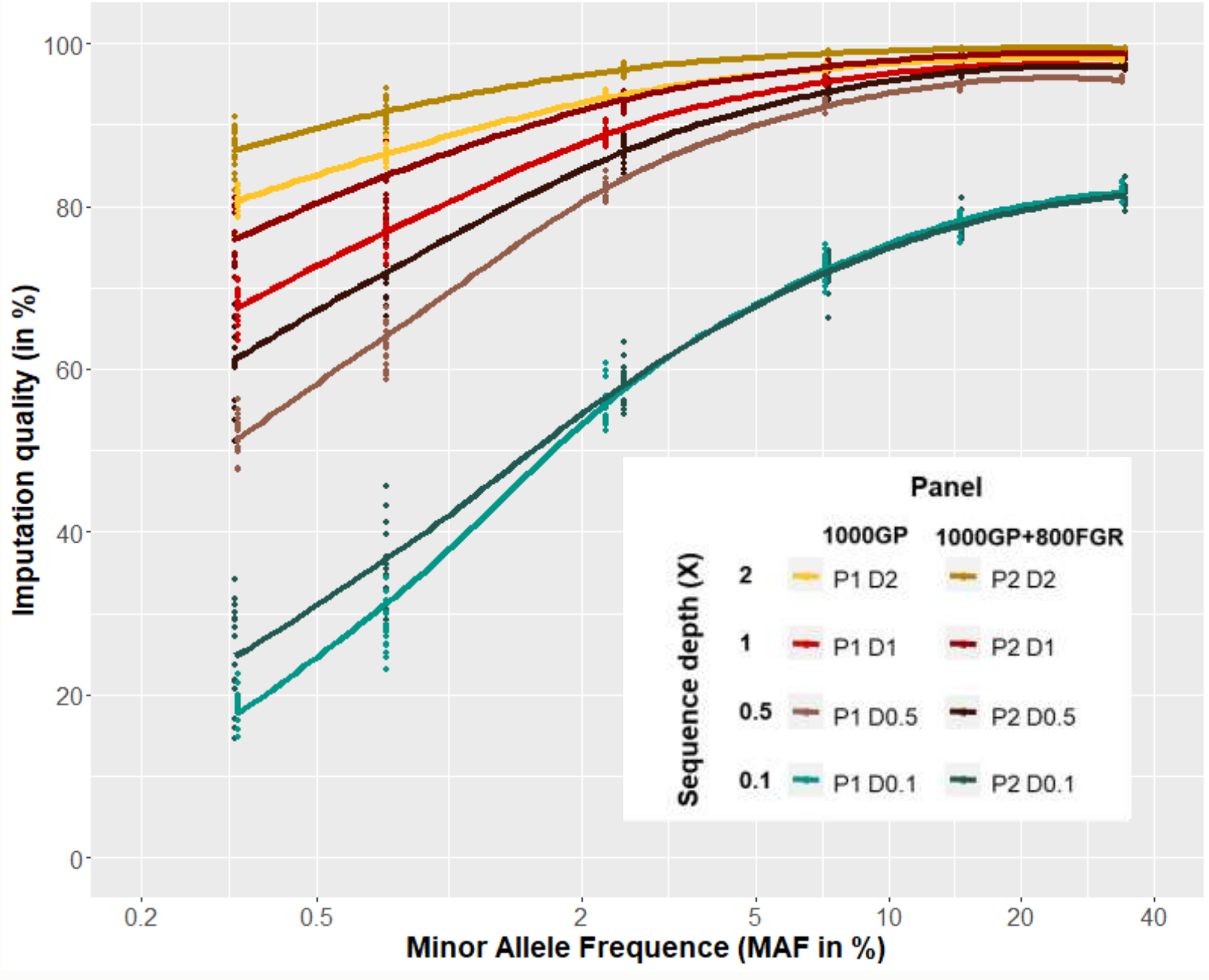


DEVELOPMENT

- Back and forth with NF-Core modules repository
- NF-test every modules and subworkflows
- 65 issues (53 closed)
- 55 Pull Request (PR)
- CI/CD through github actions
  - For each PR
  - Mega-tests with

PRELIMINARY RESULTS

- Different simulated depths (0.1 to 2X)
- 16 French individuals
- Panels: 1k Genome Project [1] (GP) or 1kGP + 800 France Gen Ref [2] (FGR)
- Imputation quality increases with:
  - Minor Allele Frequency
  - Sequencing depth
  - Reference panel with shared genetic background



OTHER INFORMATIONS

- GLIMPSE2_validate added to MultiQC.
- Before main pipeline launch check the contigs name in all files to ensure smooth running and add / remove `chr` prefix if necessary `--rename_chr`.

PERSPECTIVES

- Add simulation and imputation for SNP chips data using new tools (Beagle5, impute5, minimap2).
- Allow imputation in batch.
- Provide cost and environment impact of each run with nf-CO2footprint.

[1] Sudmant et al. 2015. "*An integrated map of structural variation in 2,504 human genomes.*" Nature 526, 75–81.

[2] Herzig et al. 2022. "Can imputation in a European country be improved by local reference panels? The example of France." BioRxiv

With the financial support of the ACGAO and the Visio Fundation: