Reproducible Data Analysis in Precision Medicine

Bjørn Fjukstad[1], Vanessa Dumeaux[2], Michael Hallett[2], and Lars Ailo Bongo[1]

1) Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway
2) Department of Biology, Concordia University, Montreal, Canada

PRECISION MEDICINE IN PRACTICE

We have analyzed a breast cancer patient's primary tumor and adjacent normal tissue, including subsequent metastatic lesions, to provide insights and recommendations for the cancer treatment. From the analyses we:

- Found germline mutations and deletion events associated with breast cancer.
- Discovered mutations in a specific gene which could explain ineffective drug treatment.
- Discovered somatic events across the whole genome.
- Identified several mutations and copy number changes in key driver genes.

BIOINFORMATICS CHALLENGES IN THE CLINIC

Implementing a precision medicine approach in the clinic has several challenges:

- Analysis pipelines must be tailored to fit each patient and biological sample.
- Large computational resources required, both with regards to storage and also computational power.
- Not feasible to use modern cloud infrastrucutres to analyze the data.
- Record all data processing steps, tool versions, optional arguments and flags, as well as output and intermediate data.
- Standardized systems make it difficul to experiment with novel tools and input parameters.

A SOLUTION

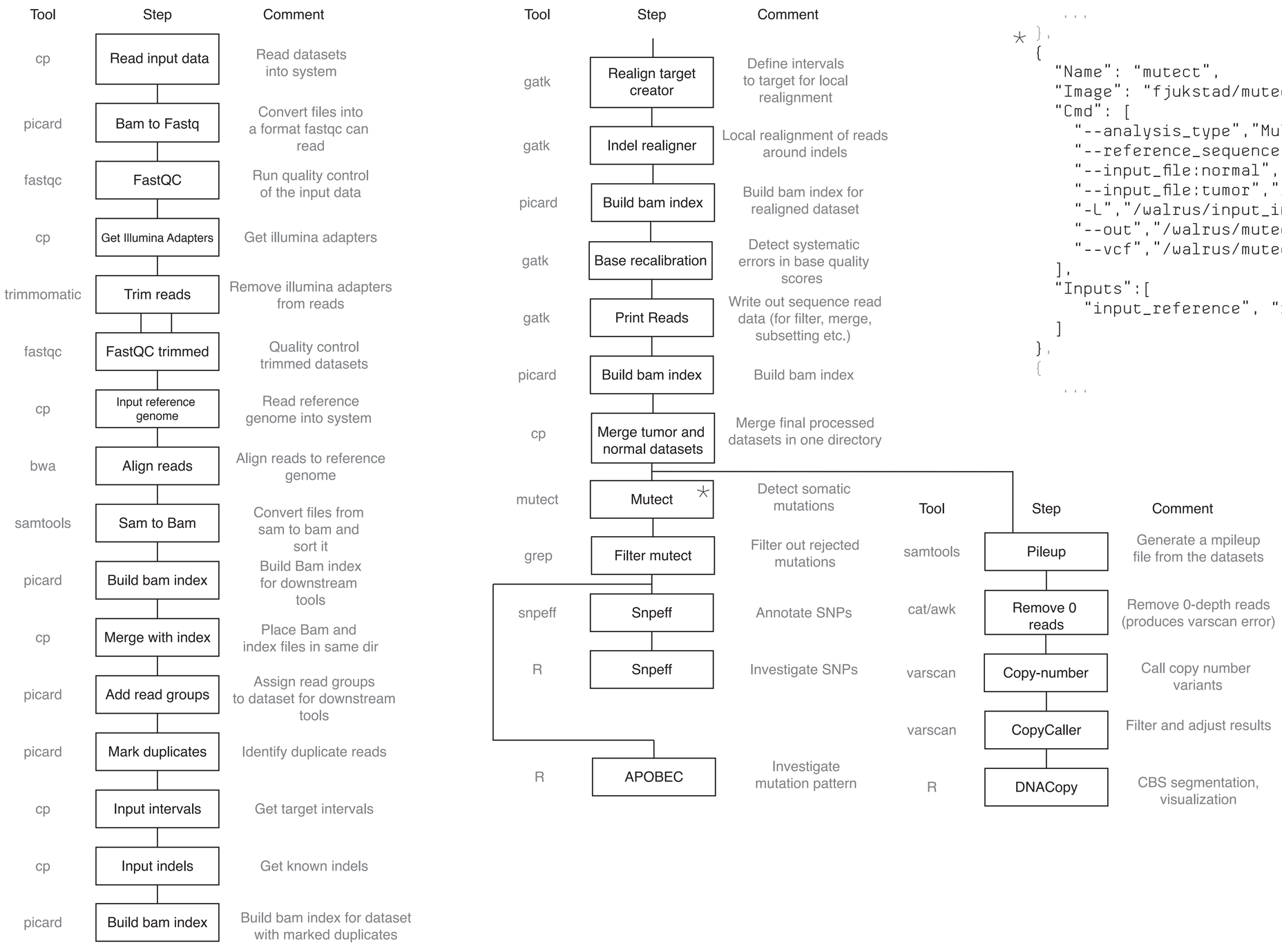
Our solution **walrus** provides:

- A flexible tool for interactively developing reproducible analysis pipelines in precision medicine.
- Full provenance management of the analysis pipeline including tool versions, tool input parameters, and data.
- Reproducible execution environments that simplifies reusing and sharing data analysis tools and pipelines.
- The functionality to use analysis tools written in any programming language.
- A simple set up process suitable for compute infrastructures in clinical settings where researchers have little or no opportunity to install complex tools.
- The functionality to run on a wide range of compute infrastructures.

walrus executes analysis pipelines within Docker containers and integrates with git using git-lfs. It uses pipeline descriptions written in either JSON or YAML, to generate an execution plan and orchestrate the execution. walrus will automatically parallelize analysis pipelines, making it possible to analyze multiple datasets simultaneously.

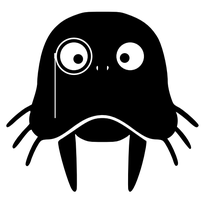
It tracks tool versions, parameters, and any intermediate data including logs, to provide full provenance of the analysis. By capturing all this data we enable researchers to reproduce the analyses that went into generating diagnostic reports for a patient.

EXAMPLE




```
    ...
  },
  {
    "Name": "mutect",
    "Image": "fjukstad/mutect",
    "Cmd": [
      "--analysis_type","MuTect",
      "--reference_sequence","/walrus/input_reference/reference.fasta",
      "--input_file:normal","/walrus/merge/normal.bam",
      "--input_file:tumor","/walrus/merge/tumor.bam",
      "-L","/walrus/input_intervals/targets.bed",
      "--out","/walrus/mutect/mutect-stats-txt",
      "--vcf","/walrus/mutect/mutect.vcf"
    ],
    "Inputs":[
       "input_reference", "input_intervals", "merge"
    ]
  },
  {
    ...
```

An example pipeline for investigating SNPs, copy-number variants, and a specific mutational pattern. Each box represents a pipeline stage and is run within a Docker container. In the top right we show an excerpt of the pipeline description for the mutect step which uses MuTect to call somatic points mutations.



github.com/uit-bdps/walrus
bjorn.fjukstad@uit.no