Proof-of-concept implementation of human genomics variant calling pipeline using BioContainers and Galaxy

In the context of the Rare Disease Community, variant-calling workflows need to be validated, benchmarked and audited for the sake of quality and reproducibility. Software containers are suitable solutions to preserve reproducibility since they can run in isolation from other software on the same machine, in a different OS version (or flavor), preserving the same exact configuration, codebase and dependencies present in the initial image. In other words containers technology facilitate reproducibility of computations and reuse of software tools in workflows.

ELIXIR, BioContainers and Galaxy

ELIXIR is on the forefront of the utilization of containers and is collaborating across Europe in order to establish an infrastructure to allow end users to find, generate and benchmark software containers.

Part of this infrastructure are ELIXIR initiatives such as BioContainers, OpenEBench, Bio.tools, and many more.

Galaxy is a very popular platform used in Life Science, as in many other fields, to run analysis of large datasets typically in pipelines of tools or workflows.

BioContainers is "community-driven project that provides the infrastructure and basic guidelines to create, manage and distribute bioinformatics packages and containers. BioContainers is based on the popular frameworks Conda, Docker and Singularity."

Methods

We present a proof-of-concept workflow from Laurie et al 2016 [1], based on BioContainers and Galaxy. It includes one aligner, one BAM processor and one variant caller, see Fig. 1.

More specifically: Burrows Wheeler Aligner (BWA) is used to map NGS reads on the reference genome, the samtools and bcftools packages to process the BAM files and call variants, respectively. The tools, installed from the main Galaxy toolshed, see Fig. 2, were configured manually with Biocontainers images.

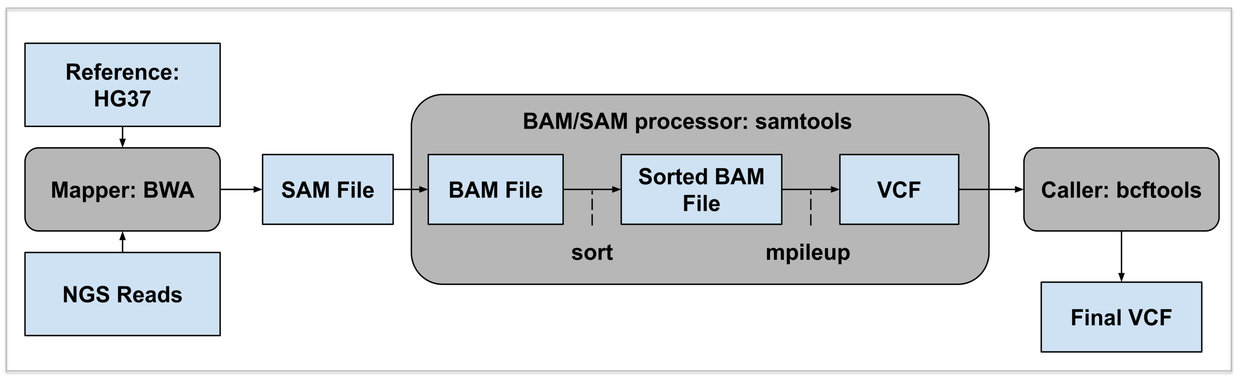

Fig. 1 Schematic representation of the workflow

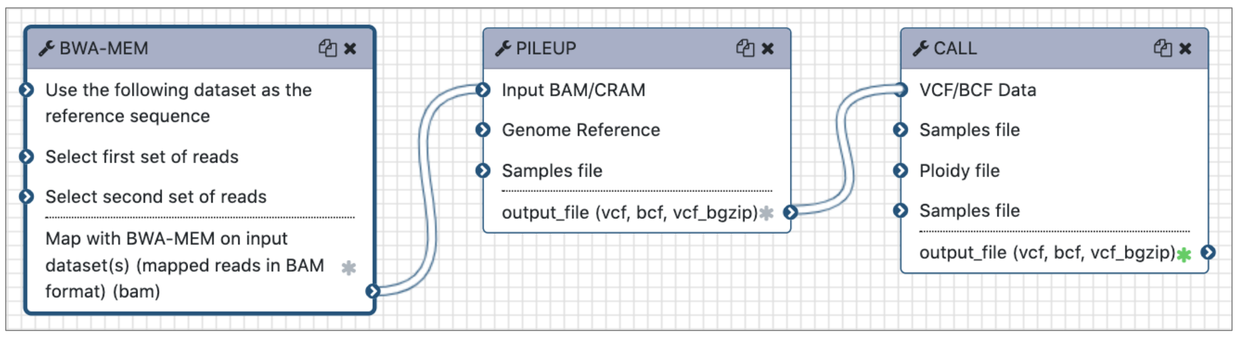

Fig. 2 Workflow representation in Galaxy

Results

To test the pipeline we used data from the NA12878 Platinum Genome FASTQ files and the hg37 human genome as reference [2]. Tools were used with standard parameters.

We compared our results with the gold-standard VCF generated by the U.S. National Institute of Standards and Technology (NIST) and the Genome in a Bottle Consortium [3] obtaining results on SNPs comparable with the golden standard:

True Positive Rate (=Recall) = 99.7%

Positive Predictive Value (=Precision) = 99.8%.

Discussion

Containerized workflows provides a prominent advantage: the same fine tuned steps of the computation can be available to each site that requires them and can be replicated even if the execution environments differ.

There are, however, few issues still to be addressed: potential version mismatch caused by manual tool configuration, and the potential incompatibility between the command line specified in the Galaxy tool and the container.

Both issues are being worked on by providing automatic container creation via **Conda**, **Mulled** and **Involucro** .

In this way the exact same version and container will be used transparently (still beta at the moment of writing).

Future directions

This proof-of-concept workflow is a stepping stone towards an effort in benchmarking different workflow managers in executing a well established variant calling pipeline, see Fig. 3. The same workflow will be implemented in NextFLow, CWL and Galaxy and the tools included in the workflow will run in containers [4].

The results will be available in OpenEBench.

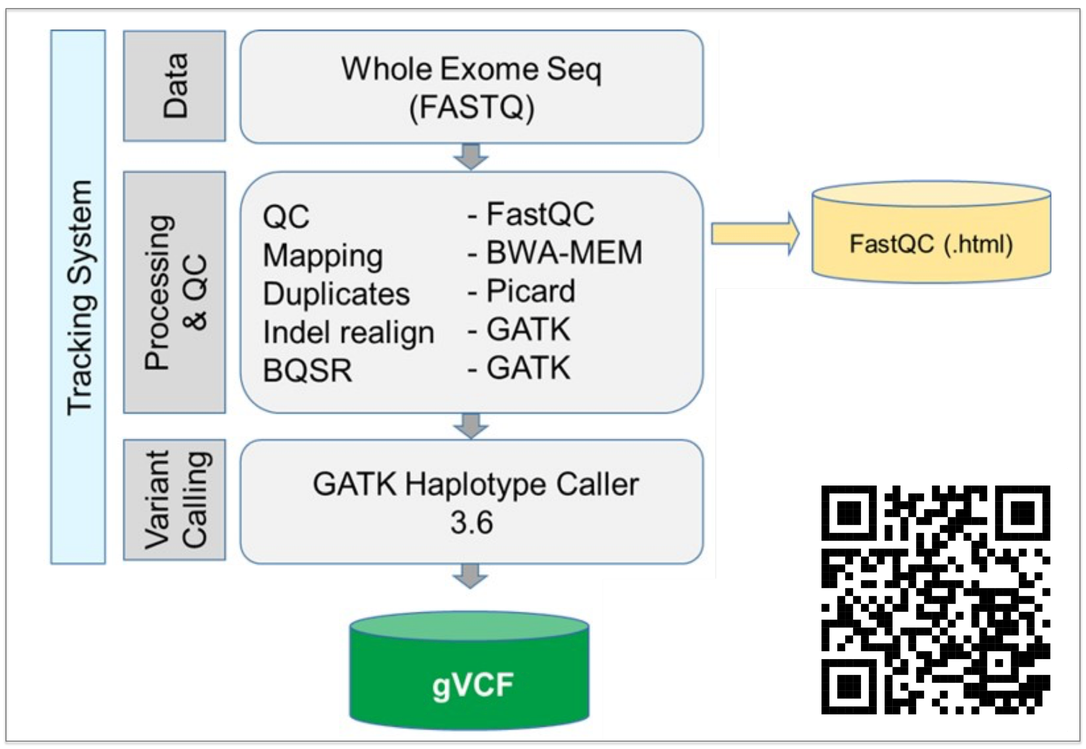

Fig. 3 RD-Connect variant calling pipeline
R. Tonda, S. Laurie, S.Berltran, et al. from [4]

References

[1] Laurie, S et al. (2016) From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. Hum Mutat. 37(12):1263-1271.

[2] https://www.illumina.com/platinumgenomes.html

[3] Zook, JM *et al.* (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 32(3):246-51.

[4] https://github.com/inab/Wetlab2Variations

Contact

Giacomo Tartari g.tartari@ibiom.cnr.it
Salvador Capella Gutierrez salvador.capella@bsc.es
Castrense Savojardo castrense.savojardo2@unibo.it
Pier Luigi Martelli pierluigi.martelli@unibo.it
Rita Casadio rita.casadio@unibo.it