WDL-based pipelines for whole genome and exome sequencing analysis

Adrián Muñoz-Barrera[1], Luis A. Rubio-Rodríguez[1], José M. Lorenzo-Salazar[1], Carlos Flores[1,2,3], Marcos Colebrook[4], Carlos J. Pérez-González[5], José L. Roda-García[4]

[1]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain; [2]Research Unit, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain; [3]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain; [4]Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [5]Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Santa Cruz de Tenerife, Spain

Introduction

Next Generation Sequencing data analysis comprises a series of computational tasks frequently based on the use of command line tools. These analyses are defined in workflows[1] that group all the necessary tasks, improving data processing performance and results interpretation. Some Domain Specific Languages (DSLs), such as WDL and Nextflow[2], have been recently created to define and program complex pipelines, as well as to improve the parallelization, the scalability[3] and the reusability. We have developed complete pipelines programmed in WDL via scripting and Rabix Composer based on the Broad Institute's best practices and the Genome Analysis Toolkit (GATK4)[4] to analyze whole-genome (WGS) and whole-exome (WES) data.

For benchmarking, we are following the guidelines of the Truth and Consistency precisionFDA challenges using Genome In A Bottle Consortium released genomes data. A full pipeline is currently running on TeideHPC to analyze WGS and WES germline data produced by an Illumina HiSeq4000 sequencing platform for research purposes.

Materials and methods

We have developed two workflows based in GATK4 using WDL and Cromwell technologies, and both of them could run in local mode, over a HPC infrastructure or in a dockerized cluster[5] (**Figure 1**). These pipelines allows the user to abstract from the implementation details and focus on the data analysis.

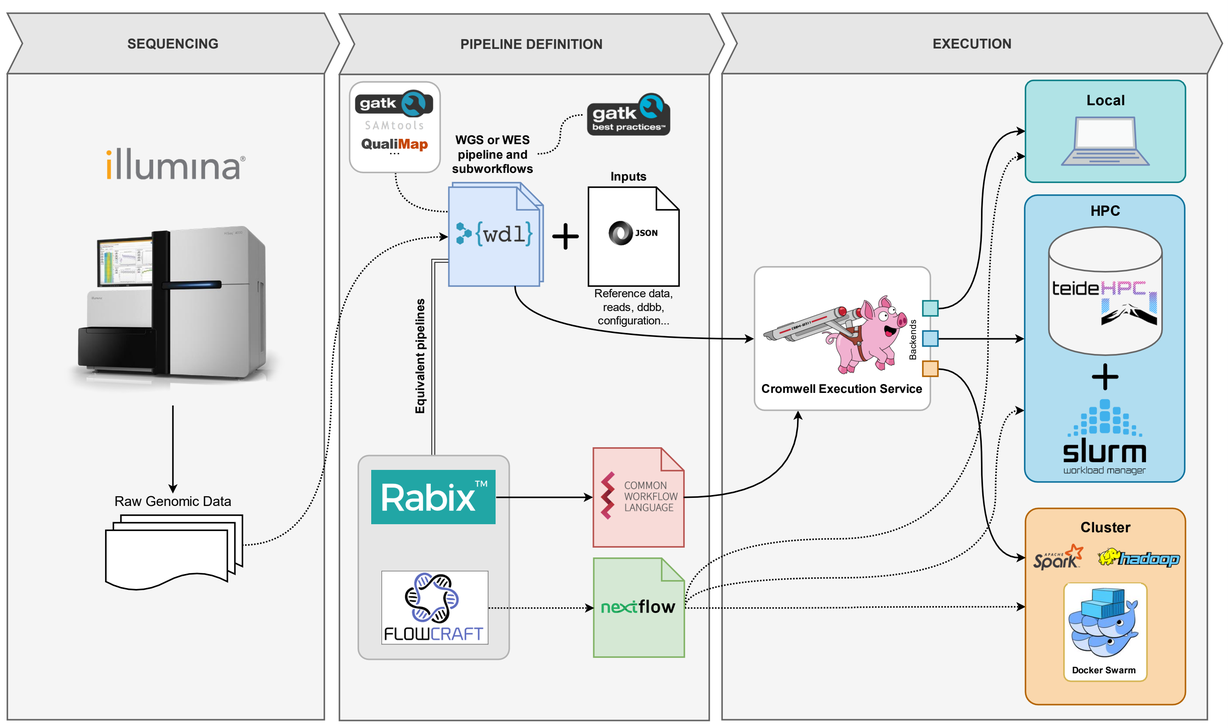

Figure 1. From sequencing to execution. See **Poster 22** for information about a dockerized cluster configuration for bioinformatics.

Both pipelines satisfy a series of requirements and features needed to cover all the demands that we established during the development. Some of these are listed below:

- Possibility to run on a HPC infrastructure connecting the Cromwell engine and the Slurm scheduler.
- Demultiplexing of samples pooled across the flowcell.
- Data processing both on a per-lane and a per-sample basis.
- Possibility to handle hg19 and hg38[6] reference genomes.
- Programmed to restart from every step in case of fail.

We are following the guidelines of the precisionFDA challenges to validate both workflows, using NA12878 training dataset for WGS pipeline and Garvan_NA12878_HG001_HiSeq_Exome data for WES pipeline. All data is available at the Genome in a Bottle Consortium (GIAB) FTP and was sequenced on an Illumina HiSeq2500.

We are currently working to adapt both pipelines to the Nextflow specification aiming to test all different pipelines versions for benchmarking purposes.

Results

WGS and WES workflows run at three execution levels (per-sample per-lane, per-sample, and multi-sample) and include preprocessing, BQSR, variant calling, joint genotyping, VQSR, as well as several quality control steps (**Figure 2**).

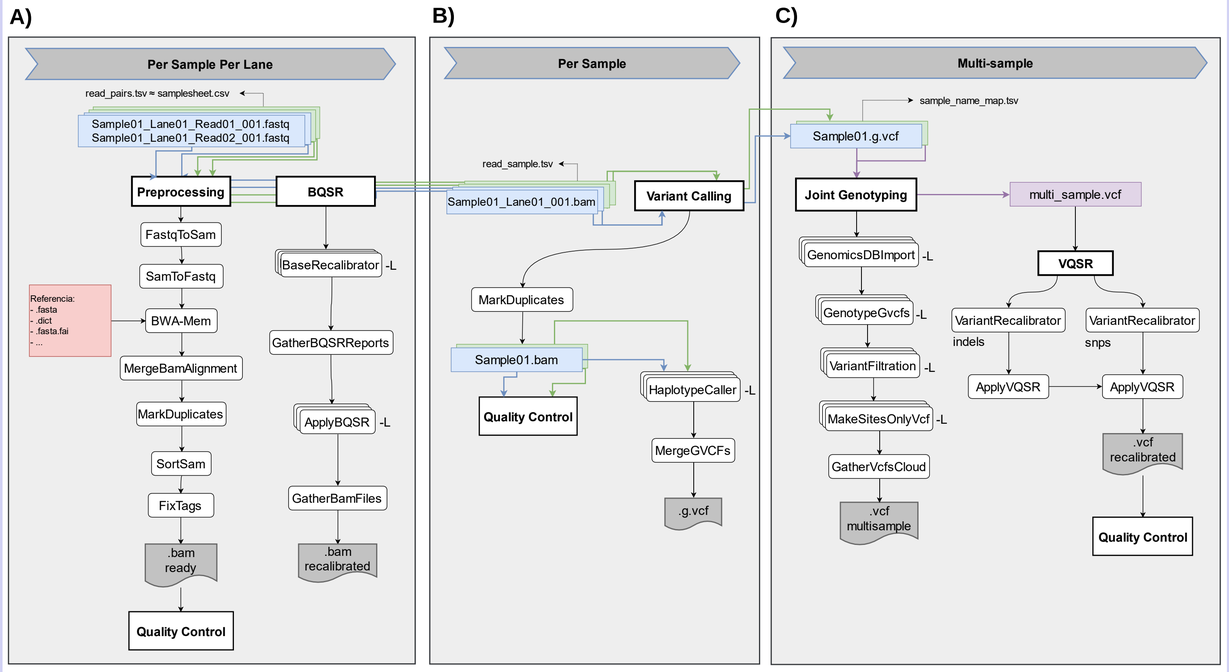

Figure 2. Pipeline stages for WGS and WES data analysis: a) per-sample per-lane; b) per-sample; and c) multi-sample.

Pipelines are easily configurable and reproducible through a UI editor as Rabix Composer which allows to build and visualize workflows using drag-and-drop actions (**Figure 3**).

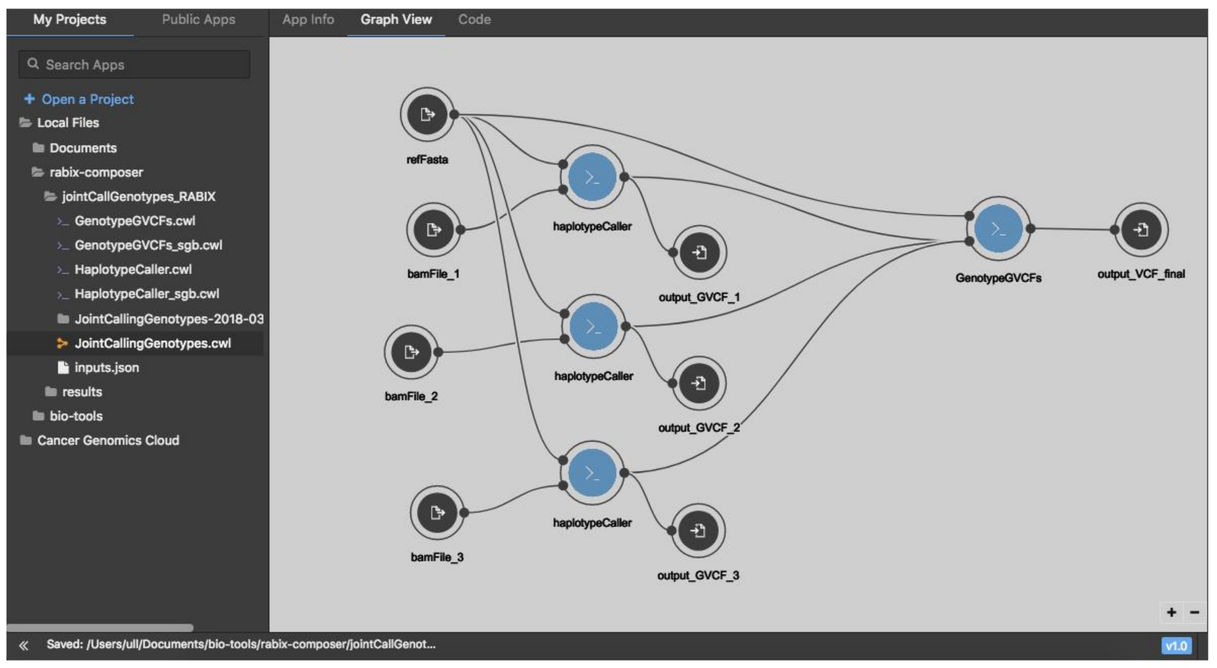

Figure 3. Variant Calling stage based on GATK's HaplotypeCaller in WGS and WES pipelines using Rabix Composer.

Availability and Contact data

Source code and documentation:

https://github.com/genomicsITER-developers/wdl

Contact data and poster download:

Funding and Acknowledgements

Funded by Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE). This work has been supported by the CEDeI program (Centro de Excelencia de Desarrollo e Innovación, Cabildo de Tenerife). The authors also thankfully acknowledge the computer resources and the technical support provided by TARO Research Group of the University of La Laguna.

References

1. J. Leipzig, Briefings in Bioinformatics, pp. bbw020, 2016
2. P. Di Tommaso et al. Nature Biotechnology, vol. 35, pp. 316-319, Apr 1,. 2017
3. B. Fjukstad et al. Data Sci. Eng, vol. 2, pp. 245-251, Sep. 2017.
4. Van der Auwera GA et al. Current Protocols in Bioinformatics 43:11.10.1-11.10.33, 2013
5. W.L. Schulz et al. Journal of Pathology Informatics, vol. 7, pp. 53, 2016
6. Y. Guo et al. Genomics, vol. 109, pp. 83-90, Mar. 2017.