Out of the box cloud solution for Next-Generation Sequencing analysis

F. van Dijk[1*], H. Byelas[1*], L. Jensma[2] , P. Neerincx[1], D. van Enckevort[1], M. Swertz[1]

[1] Genomics Coordination Center, Department of Genetics, University Medical Center Groningen, The Netherlands; Emails: f.van.dijk02@umcg.nl m.a.swertz@rug.nl

[2] University of Groningen, The Netherlands

Project Website: http://www.molgenis.org/wiki/ComputeVM
Source Code: https://github.com/molgenis/; https://github.com/molgenis/molgenis-pipelines; https://github.com/molgenis/molgenis-compute
License: GNU LESSER GENERAL PUBLIC LICENSE

The Genomics Coordination Center (Groningen, the Netherlands) has gained broad experience in running complex workflows, in which large datasets are analyzed using heterogeneous computational resources from projects like the Genome of the Netherlands [1] and LifeLines. The primary goal was to analyze large data and deliver results as quick as possible.

Now, sharing analysis protocols between institutions and reproducing analysis results has become an important issue. Setting up an execution environment in a new cluster or grid computational site introduces certain latency to the time needed to obtain results. Furthermore, the executional settings environments can change in computational grids and clusters. Hence, we have considered using the cloud infrastructure to automate setting up and sharing analysis infrastructure.

In this work, we present the complete execution environment for NGS analysis, which can be used out of the box in computational resources based on OpenStack platform. We created a configurable VM with widely used in bioinformatics software, implemented into a single pipeline to analyze NGS data. It starts with BWA aligning the raw data (FASTq), followed by realignment around known indels, quality score recalibration and variant calling using Genome Analysis ToolKit (GATK). Several quality metrics are obtained using picard-tools during data processing, variant calls are annotated using SnpEff and SnpSift tools to aid variant interpretation, producing a tab-delimited variant file. We use the EasyBuild framework and wget tool to install software and download resources respectively to ensure reproducibility of environment set-up. This enables users to reproduce installation using pre-defined easyblocks, which are available in the software repository. Furthermore, the necessary resources are installed by executing a shell script after the VM is initiated. The pipeline is available in the github repository.

To summarize, we have shown how to ensure reproducibility and efficient sharing of NGS analyses with the prepared OpenStack VM, which contains all open-source software needed for analysis grouped in a single analysis pipeline. This OpenStack VM can also be combined with MOLGENIS database [2] management system via OpenStack API to track executions. Also, all NGS analysis jobs can be generated as scripts from the MOLGENIS-Compute command-line tool [3]. Scripts can be started manually or via the pre-installed SLURM scheduler.

[1] The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014; 46:818–825.
[2] Swertz M. *et. al*. The molgenis toolkit: rapid prototyping of biosoftware at the push of a button. *BMC bioinformatics* 2010; 11(Suppl 12)
[3] Byelas H. *et. al*. Scaling bio-analyses from computational clusters to grids. *in proc. of the 5th IWSG conference)*, CEUR-WS.org

[*] Contributed equally.