Summary of the DTL Focus meeting on NGS production pipeline

Utrecht, the Netherlands, 2014-04-15

(Leon Mei, LUMC)

Due to the increasing number of NGS datasets and projects, there is a need for genomics core facility to build reliable, scalable, and reproducible NGS data analysis production pipelines. Furthermore, introducing a service model on top of such pipelines requires clear specification of the file system, backup policy, responsibilities, and SOPs. In this DTL focus meeting, we discussed current practice at different groups: GNU makefile used at LUMC, Molgenis-compute used at UMCG, GATK queue, bash scripts used at UMCU, Snakemake used at UvA and NIOO. Some frameworks have a clear limitation, for example the steep learning curve and readability issue with GNU makefile framework. Therefore the meeting participants at this DTL focus meeting aimed at comparing these different approaches and selecting the most promising framework(s) for implementing their future NGS production pipelines.

To assist the comparison, the following list of requirements of NGS production pipeline framework are agreed upon by the participants at the start of the meeting:

- **Sustainable** good support and reliable community
- **Robust** can rerun part of the pipeline if certain step failed
- **Scalable** utilize multiple cores (in a cluster)
- **Modular and no boiler plate code** swap in/out similar components (e.g. switching aligners), modules can be written in different languages
- **Portable** can easily run on a different server/cluster.
- **Transparent control** directly manage script, file location and change parameters
- **Readability** code should be relatively easy to understand
- **Provenance** explicit tracking of all scripts and options used executed steps for report generation or monitoring using a webpage

The meeting is continued with two presentations on existing frameworks used in LUMC and UMCG. Wibowo Arindrarto (LUMC) presented the modular makefile based pipeline framework developed at LUMC where the main advantages are the sustainability of GNU makefile and the native support of SGE cluster. In the discussion, Peter van 't Hof (LUMC) and Ies Nijmans (UMCU) also shared their experience with GATK Queue as an alternative framework to implement pipelines. Freerk van Dijk (UMCG) presented the Molgenis-compute system developed and used at GCC in UMCG. One of the advantages is that the pipelines based on Molgenis-compute can run on either a local cluster or the LSG Grid transparently.

The second part of the meeting consists a presentation from Johannes Köster (TU) and a presentation from Pjotr Prins (UMCU) where most interesting discussions in this meeting taken place.

Johannes explained the rationale and features of Snakemake which is specifically designed to implement data analysis pipelines and inspired by makefiles. Snakemake files are Python language based, hence it has a much cleaner syntax and offers the possibility to embed complicate logic within a snakemake file itself. Mattias de Hollander (NIOO) and Mateusz Kuzak (UvA) have been already using Snakemake for a while and both are sharing a very positive view on it. One of the current problems is that the support of cluster is a bit limiting at the moment, but Johannes confirmed that adding support of DRMAA library is the next immediate feature he will work on Snakemake.

Snakemake is also used by Johannes' group now to perform regular NGS data analysis, so there is a strong support from the group to sustain the development and support of Snakemake in future.

Pjotr motivated his presentation with a reference to his recent “small tools manifesto” initiative (https://github.com/pjotrp/bioinformatics). Pjotr has been an active contributor in the bioinformatics open source software community for many years. So one highlight of his presentation is also the link to various available open source programs (including several developed by himself) that can help bioinformaticians to simplify their pipelines, e.g., pfff, once-only, GUIX, etc. Pjotr also explained some of his short term plans on extending an existing language framework and concepts (e.g., JavaScript and the future concept) to a more bioinformatics workflow friendly language, where he also see a strong need for open communication and collaboration within the bioinformatics community.

The last presentation is from Mahdi Jaghouri (UvA) who gave a presentation from the perspective of scientific workflow community. One of the goals is to support real end users using Grid or Cluster infrastructure. Thus they have developed a portal based solution to ease the accessibilities. The meeting is ended with a short round discussion from each participant on which framework he likes the most and is planning to test back to their group. Out of 31 participants, more than 20 are interested in following up with Snakemake thanks to its clean code and makefiles like feature.