e-Bioscience Solutions and Challenges for Next Generation Sequencing Experiments

Barbera DC van Schaik*, Mark Santcroos, Souley Madougou, Aldo Jongejan, Antoine HC van Kampen, and Silvia D Olabarriaga

Academic Medical Center,
Department of Clinical Epidemiology Biostatistics and Bioinformatics,
Bioinformatics Laboratory, Amsterdam, the Netherlands
{b.d.vanschaik,m.a.santcroos,s.madougou,a.jongejan,
a.h.vankampen,s.d.olabarriaga}@amc.uva.nl
http://www.bioinformaticslaboratory.nl/

Abstract

Next generation sequencing (NGS) produces large volumes of data. To keep the processing time within bounds, there is the need to optimize the bioinformatics analysis pipelines. We have been using scientific workflow technology for agile development of analysis pipelines and grid infrastructure to accelerate the processing. Although these methods were successfully applied to a diverse range of sequencing experiments we also encountered several bottlenecks during the analysis. The challenge is to pinpoint the bottlenecks and to remove them to make optimal use of e-infrastructures for NGS experiments.

The generic platform we have developed and work with, called e-BioInfra, is currently used for medical imaging, metabolomics and large scale DNA sequencing projects. It has been used for comparison of small genomes, analysis of RNAseq experiments[1, 2], virus discovery[3], and the analysis of exome sequencing experiments[4]. The platform has a layered architecture and is based on a national grid infrastructure. It consists of a workflow management system[5], several user interfaces for workflow submission[6], and a provenance store that gathers workflow execution information from the other system components[7]. Upon execution the user can choose a previously developed workflow, define which files to analyze, and which parameters to use. The workflow management system then takes the input and translates these into jobs that are automatically executed on a distributed infrastructure.

In an earlier study we showed that this approach can result in a 30x speed-up compared to serial execution on a local system[8]. In practice we see that some applications perform well indeed and that similar speed-up rates can be obtained. However, with some software applications we experience high job error rates which are not all automatically recovered by the platform. One of the popular workflows we experienced problems with (BWA alignment) was optimized to reduce the error rate and decrease total workflow runtime[9]. A subsequent

* Corresponding author

statistical analysis of the provenance store, which is originally meant to trace back all steps of an experiment, was used to identify main causes for failure[10].

The improvements to the BWA workflow resulted in an increase of the success rate from 10% to 70% and a reduction of processing time to a third. The analysis of the provenance store indicated that BWA uses more memory than some sites offer, which was not obvious from manual inspection of the log files. This problem could be easily solved by blacklisting certain sites for BWA alignments.

To summarize, the e-Bioscience platform hides the complexities of distributed computing infrastructures from end-users. It enables scientists to perform analysis faster and be flexible in tool integration in analysis workflows. It furthermore facilitates the sharing of data and methods via the e-infrastructure. By using the e-BioInfra platform, optimizing workflows, and thorough examination of provenance and logging information, the analysis of NGS data can be greatly accelerated.

Acknowledgments

The e-BioInfra uses resources provided by the Dutch e-science Infrastructure, BiG Grid project, financed by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research, NWO). This work is supported by the ER-Flow project (financed by the FP7 infrastructures-2012-1 call under contract n312579) and by the BiGGrid/e-Biogrid programme (www.biggrid.nl).

References

1. R. Huis In 't Veld et al., *Deep Sequencing Whole Transcriptome Exploration of the s Regulon in Neisseria meningitidis*, PLoS One, 2011, vol. 6(12), e29002.
2. N. Schopman et al., *Deep sequencing of virus-infected cells reveals HIV-encoded small RNAs*, Nucleic Acids Research, 2011, vol. 40(1), pp. 414-427.
3. M. de Vries et al., *A sensitive assay for virus discovery in respiratory clinical samples*, PLoS One, 2011, vol. 24 6(1), e16118.
4. J. Van Houdt et al., *Heterozygous missense mutations in SMARCA2 cause Nicolaides-Baraitser syndrome*, Nature Genetics, 2012, vol. 44(4), pp. 445-449
5. S. Olabarriaga et al., *A Virtual Laboratory for Medical Image Analysis*, IEEE Transactions on Information Technology In Biomedicine (TITB), 2010, vol. 14(4), 979-985.
6. S. Shahand et al., *Front-ends to Biomedical Data Analysis on Grids*, Proceedings of HealthGrid, June 2011, Bristol. UK.
7. S. Madougou et al., *Provenance for distributed biomedical workflow execution*, Proceedings of HealthGrid, June 2012, Amsterdam, NL.
8. A. Luyf et al., *Initial steps towards a production platform for DNA sequence analysis on the grid*, BMC Bioinformatics, 2010, vol. 11, 598.
9. B. van Schaik et al., *Challenges in DNA sequence analysis on a production grid*, Proceedings of Science (EGICF12-EMITC2) 039, March 2012, Munich, DE.
10. S. Madougou et al., *Characterizing workflow-based activity on a production e-Infrastructure using provenance data*, 2012, submitted.