A quantitative assessment of the Hadoop framework for analysing massively parallel DNA sequencing data

Cover Letter

Dear editors,

We hereby submit our manuscript entitled *'A quantitative assessment of the Hadoop framework for analysing massively parallel DNA sequencing data'*. Hadoop is becoming increasingly popular for parallelising data analysis in a variety of disciplines. While some challenges and results on Hadoop for analysing sequencing data have been reported [1, 2], a more comprehensible assessment of this new framework is not available. In our manuscript we present the first quantitative comparison between Hadoop and traditional methods with high-performance computing resources for DNA sequence analysis in the form of short read alignment followed by variant calling. In order to make a fair comparison, we needed to define metrics and implement comparable protocols and pipelines, but also investigate user interfaces required to deliver these technologies as services for the computational biology community.

The authors have a long experience in providing high-performance e-infrastructure for biological sequence analysis, some of which was recently reported [3, 4]. Founded on the challenges we see in our current e-infrastructures we were motivated to investigate the benefits and drawbacks of the Hadoop framework and, more specifically, deduce at what problem sizes it becomes advantageous with respect to the current e-infrastructure best practices. As high-performance computing nowadays is used on wide scale in the computational biology community, we think that our paper would be of wide interest to the readers of PLoS Computational Biology as it can greatly affect how future large scale studies may be analyzed.

Sincerely,

Ola Spjuth and co-authors

Referenser

[1] Marx V. Biology: The big challenges of big data. *Nature*, 498(7453):255–60, 2013.

[2] McKenna A, Hanna M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–303, 2010.

[3] Lampa S, Dahlö M, et al. Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. *GigaScience*, 2(1):9, 2013. ISSN 2047-217X.

[4] Pireddu L, Leo S, et al. SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics*, 27(15):2159–60, 2011.