Scripting for large-scale sequencing based on Hadoop

André Schumacher[1,2,3], Luca Pireddu[4], Aleksi Kallio[5], Matti Niemenmaa[6], Eija Korpelainen[5], Gianluigi Zanetti[4], Keijo Heljanko[2,3]✉

[1]ICSI, Berkeley, USA
[2]Helsinki Institute for Information Technology HIIT, Helsinki, Finland
[3]Aalto University, Espoo, Finland
[4]CRS4, Pula, Italy
[5]CSC-IT Center for Science, Helsinki, Finland
[6]Aalto University, Espoo, Finland

Motivation and Objectives

The large volumes of data generated by modern sequencing experiments present significant challenges in their manipulation and analysis. Traditional approaches, such as scripting and relational database queries, are often found to be inadequate, frustratingly slow, or complicated to scale. These problems have already been faced by the "big data revolution" in data-based activities resulting in novel computational paradigms such as MapReduce and scalable tools such as Hadoop and Pig.

We describe our ongoing work on SeqPig, a tool that facilitates the use of the Pig Latin scripting language to manipulate, analyze and query sequencing data. SeqPig provides access to popular data formats and implements a number of high level functions. Most importantly, it grants users access to the proven to be scalable platform that is Hadoop from a high level scripting language, whether the cluster is run locally or *in the cloud*.

Methods

SeqPig operates on top of *Hadoop* and *Pig* and augments them to facilitate their use to process sequencing data. Hadoop is a distributed computing framework that implements the MapReduce programming model, which expresses computations as sequences of *side-effect* free Map and Reduce functions. Hadoop was initially developed at Yahoo!, but has since been widely adopted, e.g. by Facebook, Twitter and LinkedIn. Pig is a set-based scripting language whose instructions are compiled to a sequence of MapReduce jobs, which are then executed on a Hadoop cluster. It effectively simplifies the use of a Hadoop cluster through its concise SQL-like logic. Both Hadoop and Pig are projects supported by the Apache Software Foundation (http://hadoop.apache.org, http://pig.apache.org).

SeqPig

SeqPig extends Pig with a number of features and functionalities conceived for processing sequencing data. Specifically, it provides: 1) data input and output components, 2) specialized functions to extract fields and to transform data and 3) a collection of scripts for frequent tasks (e.g., pileup, QC statistics).

SeqPig provides import and export functions for file formats commonly used for sequencing data: Fastq, Qseq, SAM and BAM. SeqPig supports ad hoc – scripted or even interactive – distributed manipulation and analysis of large sequencing datasets. Unlike traditional methods, the scalable nature of Pig allows the speed of its operations to scale with the computing resources available. SeqPig includes functions to access SAM flags, split reads by base (for computing base-level statistics), reverse-complement reads, calculate read reference positions in a mapping (for pile-ups, extracting SNP positions), and more. The authors are currently working on expanding the library of functions, and SeqPig is an open source project that welcomes and encourages contributions from the community.

Using cloud-based resources

SeqPig has been tested on Amazon's Elastic MapReduce service. Users may rent computing time on the cloud to run their SeqPig scripts, and even share their S3 storage buckets with other cloud-enabled software.

Dependencies

SeqPig builds on Hadoop-BAM (Niemenmaa *et al.,* 2012), Seal (Pireddu *et al.,* 2011), and Picard (http://picard.sourceforge.net). Hadoop-BAM implements a number of file formats for Hadoop, while Seal and Picard implement some of the sequence analysis functiona-lity that SeqPig exposes at a higher level.

Results and Discussion

SeqPig enables the manipulation and analysis of sequencing data on the Hadoop big-data computational platform. At CRS4 SeqPig is already used routinely for some steps in the production workflow; in addition, SeqPig scripts have been used for ad hoc investigations into data quality issues, comparison of alignments tools, and reformatting or packaging data. In the future we plan to expand its function library and thoroughly test its scalability and performance characteristics.

Acknowledgements

This work was supported by the Cloud Software and D2I Programs of the Finnish Strategic Centre for Science, Technology and Innovation TIVIT and by the Sardinian (Italy) Regional Grant L7-2010/COBIK.

References

Niemenmaa M, Kallio A, Schumacher A, Klemelä P, Korpelainen E, and Heljanko K. (2012) Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. *Bioinformatics* **28**(6):876-877. doi:10.1093/bioinformatics/bts054

Pireddu L, Leo S, and Zanetti G. (2011) SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics* **27**(15):2159-2160. doi:10.1093/bioinformatics/btr325