An update on the Seal Hadoop-based sequence processing toolbox

Luca Pireddu, Simone Leo, Gianluigi Zanetti
CRS4, Pula, Italy
Email: luca.pireddu@crs4.it

Web site URL:	`http://biodoop-seal.sourceforge.net/`
Code URL:	`git://git.code.sf.net/p/biodoop-seal/code`
License:	GPLv3

Regular advances in high-throughput DNA and RNA sequencing technologies are continuously pushing the limits of typical bioinformatics data processing techniques. Medium-sized sequencing laboratories can generate Terabytes of data per week; large laboratories can produce even more. Unfortunately, most software tools available for sequence processing are not designed to scale easily to such high data rates, nor are the typical bioinformatics workflow designs. Data scalability issues such as these have already been faced by the "big data revolution" in data-based activities resulting in novel computational paradigms such as MapReduce and computing frameworks such as Hadoop.

Seal is a suite of tools that harnesses the Hadoop framework to process sequencing data. It is currently used in the production pipeline at the CRS4 Sequencing and Genotyping Platform, which houses 3 Illumina HiSeq 2000 sequencers for a total capacity of about 5000 Gbases/month. While in its first release Seal only included tools to perform sequence alignment on Hadoop (with an embedded version of BWA [1]), it has since grown, gradually removing processing bottlenecks with new scalable tools. The current (stable) `master` branch includes the following Hadoop-based distributed processing tools:

Demux: demultiplex reads from a multiplexed sequencing run;

Prq: reformat reads in `qseq` or `fastq` format in the `prq` format for alignment with Seqal;

Seqal: BWA-based distributed read mapping and duplicate identification;

ReadSort: distributed read sorting based on read id or alignment position;

RecabTable: extract empirical base quality statistics (for recalibration).

Shortly Seal will also acquire Hadoop-based tools to convert Illumina BCL files (produced by the sequencer) to `fastq` and to recalibrate base qualities; the former has already been written and will be included in the next release.

The Seal tools can be chained consecutively to implement most of the typical variant calling pipeline. At CRS4 work has been done to integrate Seal with Galaxy in order to manage the workflows through the popular web application, while the toolbox has also been independently integrated into other high-level workflow tools such as Clougene [3]. In addition, Seal can also be used as a library, borrowing its functionality for new custom and complementary applications—e.g., SeqPig (`http://seqpig.sf.net/`).

Seal tools have been shown to scale well in the amount of input data and the amount of computational nodes available [2]; therefore, with Seal one can increase processing throughput by simply adding more computing nodes. Moreover, thanks to the robust platform provided by Hadoop, the effort required by operators to run the analyses on a large cluster is generally reduced, since Hadoop transparently handles most hardware and transient network problems, and provides a friendly web interface to monitor job progress and logs. Finally, the Hadoop Distributed File System (HDFS) provides a scalable storage system that scales its total throughput and volume hand in hand with the number of processing nodes. Thus, it avoids creating a bottleneck at the shared storage volume and the need for an expensive high-performance parallel storage device.

References

[1] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25(14):1754—1760, 2009.

[2] Luca Pireddu, Simone Leo, and Gianluigi Zanetti. Seal: a distributed short read mapping and duplicate removal tool. *Bioinformatics*, 27(15):2159–2160, 2011.

[3] Sebastian Schonherr, Lukas Forer, Hansi WeiSZensteiner, Florian Kronenberg, Gunther Specht, and Anita Kloss-Brandstatter. Cloudgene: A graphical execution platform for mapreduce programs on private and public clouds. *BMC Bioinformatics*, 13(1):200, 2012.