SparkSeq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision

Marek S. Wiewiórka 1,*; Antonio Messina 2, Alicja Pacholewska 3 Sergio Maffioletti 2 Piotr Gawrysiak 1 and Michał J. Okoniewski 4

1Institute of Computer Science, Warsaw University of Technology, Poland; 2Grid Computing Competence Center - GC3, University of Zurich, Switzerland; 3Swiss Institute of Equine Medicine, University of Bern and ALP-Haras, and Institute of Genetics, Vetsuisse Faculty, University of Bern, Switzerland 4Functional Genomics Center Zurich, Switzerland.

ABSTRACT

Many time-consuming analyses of next generation sequencing data can be addressed with modern cloud computing. The Apache Hadoop-based solutions have become popular in genomics due to their scalability in a cloud infrastructure. So far, most of these tools have been used for batch data processing rather than interactive data querying.

The SparkSeq software has been created to take advantage of a new MapReduce framework, Apache Spark, for next generation sequencing data. SparkSeq is a general-purpose, flexible and easily extendable library for genomic cloud computing. It can be used to build genomic analysis pipelines in Scala and run them in an interactive way. SparkSeq opens up the possibility of customised ad hoc secondary analyses and iterative machine learning algorithms. This paper demonstrates its scalability and overall very fast performance by running the analyses of sequencing datasets. Tests of SparkSeq also prove that the use of cache and HDFS block size can be tuned for the optimal performance on multiple worker nodes.

Availability and Implementation: Available under open source Apache 2.0 license: https://bitbucket.org/mwiewiorka/sparkseq/

Supplementary information Supplementary information available as a file attached to the manuscript.

Contact: marek.wiewiorka@gmail.com

In recent years next generation sequencing (NGS) has become the essential technology that is producing precise insight into the genomes and transcriptomes of living organisms. A standard sequencer run produces hundreds gigabase pairs of short reads. The experimental results (fastq or alignment files) can be regarded as big data, in particular with the high number of biological samples. This can be less overwhelming when lab techniques of pre-selection are applied or when the data are aggregated. Computationally, the reduction of the datasets is done by counting the reads in genes or finding genome variants. The resulting count tables or variant lists are then much smaller than the original BAM files and in most cases can be processed in memory. However, aggregation and summarization always result in the loss of information on such phenomena as novel expression regions, alternative splicing or low coverage areas. Sequencing can pinpoint various properties of genomes and transcriptomes (Frazee et al., 2014; Leśniewska and Okoniewski, 2011) with a nucleotide precision but often the size of datasets is prohibitive to study such details. The nucleotide-level analysis can be a driving force for the many new and additional applications of sequencing such as linc-RNA discovery, verification of gene and UTR boundaries in the species with imprecise annotation or studies of alternative splicing.

The developments in computer science, especially in the area of distributed and cloud computing, are trying to keep pace with the rapidly growing amount of experimental data. There are several applications already available, e.g. (Taylor, 2010; Langmead et al., 2010; Schumacher et al., 2014). Currently there are many initiatives within the IT and bioinformatics community that can be utilized for creating a useful and scalable system for sequencing data analysis. One of the most frequently used parallel and distributed programming models is MapReduce which has its open-source implementation – Apache Hadoop (Borthakur, 2007). Since initially it was not possible to analyse NGS data stored in HDFS (Hadoop Distributed File System) without conversion to file formats supported by Apache Hadoop, a library Hadoop-BAM was developed for direct access and manipulation of formats commonly used in bioinformatics: BAM, FASTQ and VCF (Niemennmaa et al., 2012).

Many analytical tools that process data within the Hadoop ecosystem have been developed recently in the open-source community: new high-level languages, database engines and application frameworks. They use the same Apache Hadoop execution model, primarily designed and optimized for one-pass batch processing of on-disk data but not for interactive and in-memory ad-hoc data exploration. A good example of the successful combination of Hadoop-BAM and the Apache Pig language is SeqPig, an extension for processing of sequencing data (Schumacher et al., 2014) similar to standard samtools (Li et al., 2011).
Fig. 1. Speedup (A) and scalability (B) of SparkSeq and SeqPig on 8 BAM files RNA-sequencing dataset (total 9 GB) and a different number of worker nodes. The benchmark consists of the filtering out duplicates and reads with low quality or gaps in CIGAR string, then calculating the coverage/depth for all the samples. samtools used as a baseline.

et al., 2009). However, using high-level dataflow languages such as Apache Pig has one drawback: in many cases, data processing with them is far from optimal, since it is harder to do fine-grained code optimizations.

A completely new MapReduce paradigm implementation, Apache Spark ( Zaharia et al., 2012), addresses both of those problems. It introduces a new API, suitable for both high-level data processing workflow definition as well as low-level code tuning, using the concise Scala language. Apache Spark includes also a new storage primitive: resilient distributed datasets (RDDs). It lets the user cache precomputed data in memory and/or disk across queries. Apache Spark has also much lower launching overheads which is very important for running ad-hoc queries. Another advantage of Apache Spark is its REPL (read-eval-print-loop) environment: Spark-shell.

To study the utility of Apache Spark in the genomic context, we created SparkSeq. SparkSeq is a general-purpose tool for RNA and DNA sequencing analyses, tuned for processing in the cloud big alignment data with nucleotide precision. It combines Picard SAM-JDK via Hadoop-BAM library and Apache Spark to introduce versatile sequencing analyses in MapReduce environment. It implements currently operations on many alignment files at a time, falling into three categories: filtering of reads, summarising genomic features and statistical analyses. It can be run directly with Scala, or in R using the RSparkSeq package. The scalability study done with the prototype has proven that by using big enough computational infrastructure it is possible to handle the constantly growing number of alignment files with base-pair resolution in a manageable time. Much shorter processing time due to scalability can lead to more interactive analysis with a deeper biological insight for the genomics researchers. The computing infrastructure for the tests included: cluster nodes equipped with 8-core CPU and 16GB RAM running Linux Ubuntu 12.04 virtual machines with Apache Hadoop 1.2.1, Apache Spark 0.8.0 and Scala 2.9.3 installed. Scalability and performance has been evaluated by increasing the number of worker nodes from 1 to 10 (8 to 80 cores in total) and comparing this to samtools and SeqPig on the coverage/depth function of a multi-sample dataset. In separate tests, such parameters as HDFS block size and the use of caching have been examined to find the optimal processing time. Implemented analysis pipelines include the calculation of read coverage for all the nucleotides in the reference genome and the counts of reads within all exons, that can be processed then by the standard statistical methods for RNA sequencing (Anders et al., 2013). Testing of SparkSeq has been done mainly using two datasets: one 30GB BAM file from multiple-amplicon sequencing experiment and 32 BAM files (approx. 40GB) from a whole-transcript RNA-sequencing experiment.

The results presented in Figure 1 show that SparkSeq outperforms SeqPig in terms of speed (8.4 to 9.15 times) and that finding the coverage for all the nucleotides scales up very well. The cache experiments described in detail in the Supplementary File show that using a fast data serializer (like KryoSerializer) instead of the Java built-in one together with LZF or Snappy compression can speed up multi-pass data querying up to 110 times and reduce memory consumption by a factor of 13. HDFS block size adjustment can also result in a performance boost.

The tests done on SparkSeq clearly prove the utility of Apache Spark/Hadoop-BAM based solution for high-performance analysis of sequencing alignment files. The efficiency is comparable to standards (samtools) on a standalone machine and it offers the added value of horizontal scalability. Similar implementations can be done for other types of analyses and other type of genomic alignment files. The benefit for biology researchers will be the speedup of the long analyses and getting even the nucleotide precision of multisample results in non-prohibitive time. This will enable unsupervised searches for novel genomics phenomena in many samples in parallel, as well as optimizing standard analyses by running them many times, tuning their parameters in an interactive way. Acknowledgements: We would like to thank Christian Parise, Riccardo Murri, Tzanko Aleksiev and Martin Ryan for the valuable discussions. Funding: The work was supported by the Sciex.ch grant no. 12.289 and in part by the Swiss National Science Foundation grant (310000-116502).

REFERENCES


