We first implemented Eoulsan for differential analysis of transcript expression. The workflow dedicated to this task includes 6 steps: configuration of the cloud computing infrastructure and work with computation. This software has been developed in order to automate high-throughput sequencing data by the use of distributed algorithms that are not common among bioinformatics software solutions available. Our tests with Amazon Web Services demonstrated that the computation cost is linear with the number of instances booked as is the running time with the increasing amounts of data.

Availability and implementation: Eoulsan is implemented in Java, supported on Linux systems and distributed under the LGPL License at: http://transcriptome.ens.fr/eoulsan/

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 19, 2011; revised on March 23, 2012; accepted on April 2, 2012

1 INTRODUCTION

High-throughput sequencing data analysis requires large computer clusters associated with expensive costs that are only profitable for large genomics centers. Outsourcing data analysis over cloud computing infrastructure is one alternative. Nevertheless, software that use distributed algorithm are not common among bioinformatics developers, and only few solutions have been made available for sequencing data (Langmead et al., 2010; McKenna et al., 2010). Here we present Eoulsan, an open source framework, to facilitate high-throughput sequencing data by the use of distributed computation. This software has been developed in order to automate the analysis of a large number of samples at once, simplify the configuration of the cloud computing infrastructure and work with various already available analysis solutions.

2 SOFTWARE IMPLEMENTATION

We first implemented Eoulsan for differential analysis of transcript expression. The workflow dedicated to this task includes 6 steps:

1. The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

ABSTRACT

Summary: We developed a modular and scalable framework called Eoulsan, based on the Hadoop implementation of the MapReduce algorithm dedicated to high-throughput sequencing data analysis. Eoulsan allows users to easily set up a cloud computing cluster and automate the analysis of several samples at once using various software solutions available. Our tests with Amazon Web Services demonstrated that the computation cost is linear with the number of instances booked as is the running time with the increasing amounts of data.

Availability and implementation: Eoulsan is implemented in Java, supported on Linux systems and distributed under the LGPL License at: http://transcriptome.ens.fr/eoulsan/

Contact: eoulsan@biologie.ens.fr

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 19, 2011; revised on March 23, 2012; accepted on April 2, 2012

1 INTRODUCTION

High-throughput sequencing data analysis requires large computer clusters associated with expensive costs that are only profitable for large genomics centers. Outsourcing data analysis over cloud computing infrastructure is one alternative. Nevertheless, software that use distributed algorithm are not common among bioinformatics developers, and only few solutions have been made available for sequencing data (Langmead et al., 2010; McKenna et al., 2010). Here we present Eoulsan, an open source framework, to facilitate high-throughput sequencing data by the use of distributed computation. This software has been developed in order to automate the analysis of a large number of samples at once, simplify the configuration of the cloud computing infrastructure and work with various already available analysis solutions.

2 SOFTWARE IMPLEMENTATION

We first implemented Eoulsan for differential analysis of transcript expression. The workflow dedicated to this task includes 6 steps:

1. The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

ABSTRACT

Summary: We developed a modular and scalable framework called Eoulsan, based on the Hadoop implementation of the MapReduce algorithm dedicated to high-throughput sequencing data analysis. Eoulsan allows users to easily set up a cloud computing cluster and automate the analysis of several samples at once using various software solutions available. Our tests with Amazon Web Services demonstrated that the computation cost is linear with the number of instances booked as is the running time with the increasing amounts of data.

Availability and implementation: Eoulsan is implemented in Java, supported on Linux systems and distributed under the LGPL License at: http://transcriptome.ens.fr/eoulsan/

Contact: eoulsan@biologie.ens.fr

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 19, 2011; revised on March 23, 2012; accepted on April 2, 2012

1 INTRODUCTION

High-throughput sequencing data analysis requires large computer clusters associated with expensive costs that are only profitable for large genomics centers. Outsourcing data analysis over cloud computing infrastructure is one alternative. Nevertheless, software that use distributed algorithm are not common among bioinformatics developers, and only few solutions have been made available for sequencing data (Langmead et al., 2010; McKenna et al., 2010). Here we present Eoulsan, an open source framework, to facilitate high-throughput sequencing data by the use of distributed computation. This software has been developed in order to automate the analysis of a large number of samples at once, simplify the configuration of the cloud computing infrastructure and work with various already available analysis solutions.

2 SOFTWARE IMPLEMENTATION

We first implemented Eoulsan for differential analysis of transcript expression. The workflow dedicated to this task includes 6 steps:

1. The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
We saw that the relationship between running time and number of available solution in Eoulsan.

The distributed calculation process we used is based on Hadoop and it can be installed on numerous cluster server configurations. We made our proof of concept using AWS solution as it includes EMR, an advanced service based on EC2 that easily allows Hadoop distributed computation solutions to be deployed. However, we are working to make our software independent of EMR to work directly on EC2 services. Such an evolution will allow Eoulsan to be run on any other cloud computing solutions such as Open Source initiatives like OpenNebula or OpenStack. This would open future possibilities by creating regional genomic computer infrastructures (like iPlant for example) to be shared among several local high-throughput sequencing users. With a dedicated high-speed network, this can speed up the time transfer process. In addition, this could also favor the standardization of analysis pipelines developed from the bioinformatics community, making high-throughput sequencing technologies really accessible for a wide audience.

4 CONCLUSION

With our framework, we aimed to facilitate high-throughput sequencing analysis on cloud computing services with an automatic, modular and efficient tool. This approach differs from other solutions already available for distributed calculation such as Myrna, CloudBurst or Crossbow (Langmead et al., 2010; Langmead et al., 2009a; Schatz, 2009), where the cloud computing implementation is made around a unique analysis solution. In addition, Eoulsan is also complementary to the customizable Galaxy server solution as it allows for batch analyses and it contains a full automation process able to handle external file locations and distributed file system. More generally, we have first implemented an automated RNA-Seq analysis pipeline but all the tools are already included within Eoulsan for other applications such as SNP calling or ChIPPeak analysis. The Java plug-in system we developed as well as the full documentation we provide, allow for user contribution by the integration of other available solution in Eoulsan.

and Supplementary Table S3). We observed that the whole time spent for the calculations strongly drop with low instance number and remained close to linear with more than five instances booked. However, one question remains. How many instances do we need to use for data analysis? This is critical as each hourly booked EC2 instance costs a fixed price. Our tests demonstrate that the cost per hour is linear over the number of instances used (Supplementary Fig. 4). This means that the number of instances can be increased in order to speed up the data analysis process without the risk to fall in a suboptimal configuration.

Finally, we assessed the impact of an increase in raw data on the computation time by running Eoulsan with 16 and 32 samples of 23.5 millions of reads each, respectively, 376 and 752 millions of total reads (Supplementary Fig. S5 and Supplementary Table S4). We saw that the relationship between running time and number of samples is also linear (Supplementary Fig. S6). This demonstrates that Eoulsan is able to handle the increase in raw data coming from future evolutions of Illumina sequencing devices.

ACKNOWLEDGEMENTS

We thank J. Le Men and P. Gilardi Hebenstreit for providing RNA-Seq samples and the Pasteur Institute Transcriptomics and Epigenomics platform for the Illumina runs. We express our gratitude to T. Portnoy for helpful discussions and to A. Kandil and J. Bansse for their analysis of the cloud computing market. Finally, we thank P. Surnn for the careful review of the English spelling.

Funding: This work was funded by the network Infrastructures en Biologie Santé et Agronomie (IBiSA) and by an Amazon Web Services research grant.

Conflict of Interest: none declared.

REFERENCES


