*Bioinformatics*, 34(19), 2018, 3405–3407
doi: 10.1093/bioinformatics/bty363
Advance Access Publication Date: 3 May 2018
Applications Note

OXFORD

Gene expression

# NGScloud: RNA-seq analysis of non-model species using cloud computing

**Fernando Mora-Márquez[1], José Luis Vázquez-Poletti[2] and Unai López de Heredia[1,*]**

[1]GI Genética, Fisiología e Historia Forestal, Dpto. Sistemas y Recursos Naturales, ETSI Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid and [2]GI Arquitectura de Sistemas Distribuidos, Dpto. Arquitectura de Computadores y Automática, Facultad de Informática, Universidad Complutense de Madrid, Ciudad Universitaria, 28040 Madrid, Spain

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on September 21, 2017; revised on April 23, 2018; editorial decision on April 30, 2018; accepted on May 2, 2018

## Abstract

**Summary:** RNA-seq analysis usually requires large computing infrastructures. NGScloud is a bioinformatic system developed to analyze RNA-seq data using the cloud computing services of Amazon that permit the access to *ad hoc* computing infrastructure scaled according to the complexity of the experiment, so its costs and times can be optimized. The application provides a user-friendly front-end to operate Amazon's hardware resources, and to control a workflow of RNA-seq analysis oriented to non-model species, incorporating the cluster concept, which allows parallel runs of common RNA-seq analysis programs in several virtual machines for faster analysis.
**Availability and implementation:** NGScloud is freely available at https://github.com/GGFHF/NGScloud/. A manual detailing installation and how-to-use instructions is available with the distribution.
**Contact:** unai.lopezdeheredia@upm.es

## 1 Introduction

RNA-seq experiments often yield huge amount of data, especially when several NGS libraries are involved. The algorithms used in the bioinformatic analyses are very complex, particularly those referred to the assembly of reads (Miller *et al.*, 2010). Thus, the hardware requirements to run RNA-seq analysis are very high in terms of CPUs and GiBs of RAM memory, and computing infrastructure to fulfill such requirements is not always available in small research centers. In such cases, cloud computing is a solution that provides resizable computing capacity, and therefore, allows to fit the hardware to the nature of the experimental data. One of the main cloud computing solutions is the Elastic Compute Cloud (EC2), a service of the Amazon Web Services (AWS). The EC2 has a wide range of scalable instances that allow the optimization of the experiment costs, because the user only pays for the time of use of the resources. Also, the EC2 provides immediacy, since a virtual machine can be booted in only a few minutes.

We present NGScloud, a bioinformatic system developed to analyze RNA-seq data using the cloud computing offered by EC2. NGScloud is oriented to non-model species whose reference genomes are not available, and it implements parallel runs in several virtual machines for faster analysis. The application aims to ease the researcher the use of EC2 resources and the performance of RNA-seq analysis.

## 2 Materials and methods

### 2.1 Software

NGScloud was programmed in Python3, and it runs in any computer with an OS that allows for Python3: Linux, Microsoft Windows, Mac OS X and other platforms. To work properly, NGScloud has the following dependencies for the local computer of the user: (i) StarCluster, an open source cluster-computing toolkit for EC2 (http://star.mit.edu/cluster/); (ii) Boto3, the AWS SDK for

Python (https://boto3.readthedocs.io/); (iii) Paramiko, an implementation of the SSHv2 protocol in Python (http://www.paramiko.org/) and (iv) AWS CLI (https://aws.amazon.com/cli/).

NGScloud offers a user-friendly front-end to operate the EC2 resources, to control the implement RNA-seq workflow and to handle the data. NGScloud runs in graphical mode using the graphical user interface (GUI) by default, but it can also be run in console mode on server machines without GUI installed.

In addition, several free bioinformatic applications of common use in RNA-seq workflows are easily set up from the front-end (Point 2.3).

### 2.2 Cloud computing

NGScloud philosophy is based on the cluster concept. A cluster is a set of virtual machines of an AWS instance type. Each instance type has its hardware features (machine type, CPU number, memory amount, etc).

Data volumes allow to save data and keep them even if there is not any cluster created. NGScloud uses Amazon's EBS volumes to hold applications, read files, references, databases and results of analysis.

Through the NGScloud front-end, the user can easily to: (i) create and terminate clusters; (ii) show the cluster composition; (iii) add and remove nodes dynamically to a cluster; (iv) create and remove volumes; mount and unmount volumes; (v) submit and kill jobs to the RNA-seq workflow; (vi) show the status of the batch job; (vii) view the log of every batch job to inspect correct program operability and (viii) upload, download, compress, decompress and remove datasets.

When a cluster is created, it has only a virtual machine named master node. After the master node creation, subsidiary nodes can be added if necessary, to run some processes in parallel. In this case, the new job will run in the node determined according to the workload.

### 2.3 RNA-seq workflow

The RNA-seq workflow implemented has the standard steps of a RNA-seq analysis in non-model species (López de Heredia and Vázquez-Poletti, 2016), including: (i) pre-processing: read quality assessment with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc.), trimming with Trimmomatic (Bolger, 2014) and the *insilico-read-normalization* procedure of Trinity; (ii) *de novo* assembly with SOAPdenovo-Trans (Xie, 2014), Trans-Abyss (Robertson, 2010) and Trinity (Grabherr, 2011; Haas, 2013), and reference-based assembly with STAR (Dobin, 2013); (iii) assessment of the assembly quality and transcript quantification with Transrate (Smith-Unna, 2016), GMAP (Wu and Watanabe, 2005), BUSCO (Waterhouse, 2017), QUAST (Gurevich, 2013), rnaQUAST (Bushmanova, 2016) and RSEM-EVAL included in DETONATE package (Li, 2014); (iv) post-filtering with CD-HIT-EST (Li and Godzik, 2006) and *transcript-filter* included in NGShelper package, which also has some other tools to perform the RNA-seq analysis workflow and (v) annotation with *transcriptome-blast* that encapsulates blastx runs in several nodes, and is included in NGShelper. The results may be downloaded for downstream analysis (Fig. 1).

The workflow steps are run separately by the user, with each step requiring the setup of the cloud resources. NGScloud is configured to read the output generated in each step as the required input file(s) of subsequent steps, or to download the output to a local machine. For instance, to perform an assembly, reads are uploaded, the assembly program runs in the cluster and the output is downloaded,

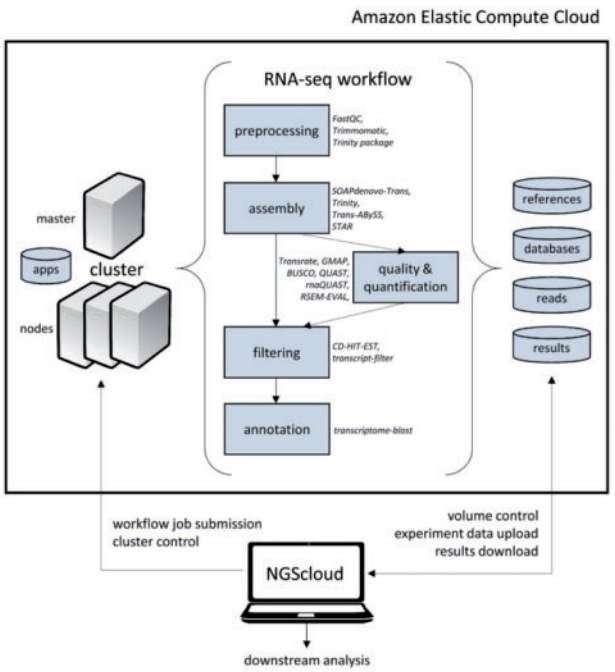


**Fig. 1.** NGScloud architecture. NGScloud operates EC2 resources, submits workflow and manages datasets from RNA-seq experiments

or submitted to the next step of the RNA-seq workflow. Running of bioinformatic applications is easy since the researcher is guided in the choice of the input files and the parameters to be used, encapsulating the complexity of the command line. Multiple runs of applications can be run in parallel creating nodes. In addition, the annotation step supports parallelization, so it can use several nodes to increase the run speed.

## 3 Conclusions

NGScloud provides a user-friendly front-end to operate the EC2 resources, and to control the workflow of non-model species oriented RNA-seq experiment in a modular way. The application allows to optimize the cost-efficiency ratio of RNA-seq experiments when appropriate computational facilities are not available.

## Funding

This work has been supported by the projects SPIP2014-01093 (Spanish National Parks Agency, Ministry of Agriculture and AGL2015-67495-C2-2-R and FedCloudNet) (MINECO TIN2015-65469-P) (Spanish Ministry of Economy and Competitiveness) and by an Amazon Research Grant.

*Conflict of Interest*: none declared.

## References

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bushmanova,E. *et al.* (2016) rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*, **32**, 2210–2212.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Gurevich,A. *et al*. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Haas,B.J. *et al*. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–1512.

Li,B. *et al*. (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.*, **15**, 553.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

López de Heredia,U. and Vázquez-Poletti,J.L. (2016) RNA-seq analysis in forest tree species: bioinformatic problems and solutions. *Tree Genet. Genomes*, **12**, 30.

Miller,J.R. *et al*. (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–327.

Robertson,G. *et al*. (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.

Smith-Unna,R. *et al*. (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.*, **26**, 1134–1144.

Waterhouse,R.M. *et al*. (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, **35**, 1–6.

Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Xie,Y. *et al*. (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–1666.