NGScloud2: optimized bioinformatic analysis using Amazon Web Services

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

[1] GI Sistemas Naturales e Historia Forestal, Dpto. Sistemas y Recursos Naturales, ETSI Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Madrid, Spain

[2] GI Arquitectura de Sistemas Distribuidos, Dpto. de Arquitectura de Ordenadores y Automática, Facultad de Informática, Universidad Complutense de Madrid, Madrid, Spain

Corresponding Author: Unai López de Heredia
Email address: unai.lopezdeheredia@upm.es

Background. NGScloud was a bioinformatic system developed to perform de novo RNAseq analysis of non-model species by exploiting the cloud computing capabilities of Amazon Web Services. The rapid changes undergone in the way this cloud computing service operates, along with the continuous release of novel bioinformatic applications to analyze next generation sequencing data, have made the software obsolete. NGScloud2 is an enhanced and expanded version of NGScloud that permits the access to ad hoc cloud computing infrastructure, scaled according to the complexity of each experiment.

Methods. NGScloud2 presents major technical improvements, such as the possibility of running spot instances and the most updated AWS instances types, that can lead to significant cost savings. As compared to its initial implementation, this improved version updates and includes common applications for de novo RNAseq analysis, and incorporates tools to operate workflows of bioinformatic analysis of reference-based RNAseq, RADseq and functional annotation. NGScloud2 optimizes the access to Amazon's large computing infrastructures to easily run popular bioinformatic software applications, otherwise inaccessible to non-specialized users lacking suitable hardware infrastructures.

Results. The correct performance of the pipelines for de novo RNAseq, reference-based RNAseq, RADseq and functional annotation was tested with real experimental data. NGScloud2 code, instructions for software installation and use are available at https://github.com/GGFHF/NGScloud2. NGScloud2 includes a companion package, NGShelper that contains python utilities to post-process the output of the pipelines for downstream analysis at https://github.com/GGFHF/NGShelper.

NGScloud2: optimized bioinformatic analysis using Amazon Web Services

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

[1] GI Sistemas Naturales e Historia Forestal, Dpto. Sistemas y Recursos Naturales, ETSI Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Madrid, Spain.
[2] GI Arquitectura de Sistemas Distribuidos, Dpto. de Arquitectura de Ordenadores y Automática, Facultad de Informática, Universidad Complutense de Madrid, Madrid, Spain.

Corresponding Author:
Unai López de Heredia[1]
C/ José Antonio Nováis 10, Madrid, 28040, Spain
Email address: unai.lopezdeheredia@upm.es

Abstract

Background. NGScloud was a bioinformatic system developed to perform de novo RNAseq analysis of non-model species by exploiting the cloud computing capabilities of Amazon Web Services. The rapid changes undergone in the way this cloud computing service operates, along with the continuous release of novel bioinformatic applications to analyze next generation sequencing data, have made the software obsolete. NGScloud2 is an enhanced and expanded version of NGScloud that permits the access to ad hoc cloud computing infrastructure, scaled according to the complexity of each experiment.

Methods. NGScloud2 presents major technical improvements, such as the possibility of running spot instances and the most updated AWS instances types, that can lead to significant cost savings. As compared to its initial implementation, this improved version updates and includes common applications for de novo RNAseq analysis, and incorporates tools to operate workflows of bioinformatic analysis of reference-based RNAseq, RADseq and functional annotation. NGScloud2 optimizes the access to Amazon's large computing infrastructures to easily run popular bioinformatic software applications, otherwise inaccessible to non-specialized users lacking suitable hardware infrastructures.

Results. The correct performance of the pipelines for de novo RNAseq, reference-based RNAseq, RADseq and functional annotation was tested with real experimental data. NGScloud2 code, instructions for software installation and use are available at https://github.com/GGFHF/NGScloud2. NGScloud2 includes a companion package, NGShelper that contains python utilities to post-process the output of the pipelines for downstream analysis at https://github.com/GGFHF/NGShelper.

Introduction

The large output size of Next Generation Sequencing (NGS) technologies and the algorithms and applications employed in their analysis, present processing limitations typical of big data, such as RAM size, CPU capacity, storage and data accessibility (Yang et al., 2017). Therefore, research labs have to allocate a significant part of their budget to provisioning, managing and maintaining their computational infrastructure (Kwon et al., 2015). A cost-efficient alternative for NGS analysis that presents several advantages over local or HPC hardware infrastructure resides in cloud computing (Langmead & Nellore, 2018). Cloud computing is flexible and scalable, allowing various configurations of OS, RAM size, CPU number and almost unlimited storage to fit the hardware resources for a specific bioinformatic workflow. Once the workflow computing requirements are provisioned, hardware resources are readily available, and the workflow performance and data can be securely accessed and monitored at any time from any local computer with internet access. Moreover, for public cloud services, the user only pays for the effectively used resources, reducing experiment times and costs.

Here we present NGScloud2, a new version of the NGScloud software (Mora-Márquez, Vázquez-Poletti & López de Heredia U, 2018). NGScloud was developed as a bioinformatic system to perform *de novo* RNAseq analysis of non-model species. This was accomplished using the cloud computing infrastructure from Amazon Web Services (AWS), the Elastic Compute Cloud (EC2), and its high-performance block storage service, the Amazon Elastic Block Store (EBS). NGScloud allowed to create one or more EC2 instances (virtual machines) of M3, C3 or R3 instance types forming clusters where analytic processes were run using StarCluster, an open source cluster-computing toolkit for EC2 (http://star.mit.edu/cluster/). However, NGScloud did not support the new instance types that AWS has made available since the original application release. Below we describe the major new features of NGScloud2 that significantly expand NGScloud2 functionality with respect to the original version.

Materials & Methods

NGScloud2 is a free and open source program written in Python3. Source code and a complete manual with installation instructions and tutorials to exploit all the potential of NGScloud2 are available from the GitHub repository (https://github.com/GGFHF/NGScloud2). NGScloud2 presents remarkable differences with respect to NGScloud both in the way AWS resources are managed to better exploit all the potential of EC2 and EBS, but also by incorporating the possibility of running a more complete set of bioinformatic applications and pipelines for *de novo* RNAseq, reference-based RNAseq, Restriction site Associated DNA sequencing (RADseq) and functional annotation (see Results and Discussion section). In addition, a toolkit of Python programs useful to post-process the output of RNAseq and RADseq experiments is available in NGShelper (https://github.com/GGFHF/NGShelper).

The correct operability of the pipelines for *de novo* RNAseq, reference-based RNAseq, RADseq and funcional annotation was tested with data generated by our research group. Test data for RNAseq and RADseq workflows consisted of two sets of Illumina^{TM} reads: (1) Pcan, a paired-ended RNA library of xylem regeneration tissue of the conifer tree *Pinus canariensis* (Mora-Márquez et al. 2020a). (2) Suberintro, a set of 16 paired ended Illumina^{TM} libraries of *Quercus suber*, *Quercus ilex* and their hybrids obtained from leaf tissue; eight libraries correspond to genotyping-by-sequencing with MslI and other eight libraries correspond to ddRADseq with PstI-MspI (see details in Guillardín-Calvo et al., 2019). Read data are available at NCBI: SRX5228139 -SRX5228161 for Pcan, and SRX5019123-SRX5019138 for Suberintro. The functional annotation workflow was tested with a small subset of transcripts corresponding to the monolignol biosynthesis gene family in Arabidopsis (Raes et al., 2003).

Results & Discussion

Technical improvements

NGScloud2 introduces a more efficient architecture of instances and volumes than the original version (Figure1). While NGScloud used one volume for each type of existing datasets (applications, databases, references, reads and results), NGScloud2 offers the possibility of holding all dataset types in a unique volume, thus reducing the complexity in volume management. NGSCloud2 philosophy is based on the "cluster" concept. A cluster is a set of 1 to n virtual machines with the same instance type. Each instance type has its hardware features: processor type, CPU number, memory amount, etc. (https://aws.amazon.com/ec2/instance-types/).

NGScloud2 includes two cluster modes, StarCluster and native. The StarCluster mode uses StarCluster (http://star.mit.edu/cluster/), an open source cluster-computing toolkit for EC2, which implements clusters of up to 20 virtual machines, enabling faster analysis. The last version of Starcluster (0.95.6) dates from 2013 and can only use AWS's previous generation instance types, i.e. m3, c3 or r3. In NGScloud2, we provide a patch to enable using m4, c4 and r4 instance types.

To reduce the dependency of NGScloud from StarCluster, which only allows to create clusters of previous generation instances, NGScloud2 has incorporated a "native" instance creation mode that sets a single virtual machine with any of the currently available on-demand EC2 instance types (m4, c4, r4, m5, m5a, c5, c5a, r5 and r5a). The new generation instance types are slightly cheaper and their hardware improves over equivalent hardware from previous generations. Moreover, the new version enables launching "spot instances" that derive from unused EC2 capacity in the AWS cloud (https://aws.amazon.com/ec2/spot/). Spot instances have the advantage of being up to 50-80% cheaper than on-demand instances at the cost of suffering unpredictable interruption out of control of the user (Supplemental Table 1). Therefore, using spot instances is highly recommended for data transfer and for certain bioinformatic processes that run fast, process small volume input or include the possibility to be re-launched from the process interruption point.

NGScloud2 includes a user-friendly graphical front-end to operate the hardware resources, submit processes, and manage the data. The front-end includes a drop-down menu to configure AWS resources (clusters, nodes and volumes) and to install available bioinformatic software.

Data transfer between the cloud and the local computer is operated through another drop-down menu. Additional drop-down menus are available to run de novo RNAseq, reference-based RNAseq, RADseq and functional annotation workflows, respectively. Log files of each executed process can be consulted in the "Logs" menu.

New methods and applications available

The other major improvements of NGScloud2 over NGScloud are related to the implementation of new bioinformatic pipelines and application tools (Table 1) that are automatically installed using Bioconda (Grüning et al., 2018), thus giving access to updated versions of the software without worrying about dependencies and software requirements. While the original purpose of NGScloud was to help in *de novo* RNAseq analysis, NGScloud2 includes pipelines and applications to perform reference based RNAseq, RADseq and functional annotation. A summary of the AWS instances employed and the total elapsed times for the pipelines run on the test data is available in Supplemental Table 2, Table3, Table 4 and Table 5.

De novo RNAseq

The original software was mainly focused on *de novo* assembly of RNAseq libraries using either Trinity, and included pre-processing of reads with FASTQC (Andrews, 2010), Trimmomatic (Bolger, Lohse & Usadel, 2014) and three *de novo* RNAseq assemblers: Trinity (Haas et al., 2013), SoapDeNovo-Trans (Xie et al., 2014) and Transabbys (Robertson et al., 2010). NGScloud2 *de novo* RNAseq workflow has been improved (Figure 2) by including cutadapt (Martin, 2011) to perform read pre-processing, a new read alignment step with Bowtie2 (Langmead & Salzberg, 2012) to map back the reads to the assembled transcriptome and software to quantify total counts of transcripts for further differential expression analysis: eXpress (Roberts & Pachter, 2013) and Kallisto (Bray el al., 2016). Intensive processes, such as Trinity and SOAPdenovo-Trans transcriptome assemblers can now be re-launched from the point where the process interruption occurred, thus preventing unexpected malfunctioning of the cloud system or software bugs (Mora-Márquez et al. 2020a). A variant calling step is also included to find SNPs or indels using SAMtools (Li et al. 2009), BEDtools (Quinlan & Hall, 2010) and BCFtools (Danecek & McCarthy, 2017).

Reference-based RNAseq

In the last years, an increasing number of genomic and transcriptomic resources are available for many plant and animal species. Therefore, reference-based RNAseq is expected to become a usual practice not only for model species. NGScloud2 includes a workflow to accomplish read pre-processing, read alignment, reference-guided assembly, quantitation, differential expression and variant calling (Figure 3). Read pre-processing is done with the same tools as for *de novo* RNAseq (Trimmomatic and cutadapt). Read alignment to a reference genome assembly can be performed with Bowtie2, or with popular splice-aware aligners: Hisat2 (Kim et al., 2019), TopHat2 (Kim et al., 2013), STAR (Dobin et al., 2013) or GSNAP (Wu et al., 2016). Moreover, read alignment can also be run against a reference transcriptome using GMAP (Wu et al., 2016). After read alignment, a transcriptome can be assembled using Cufflinks-Cuffmerge (Trapnell et al., 2012). Reference-guided *de novo* assembly can also be performed with Trinity's genome guided version (Haas et al., 2013). Transcript or isoform abundance can be quantified with Cuffquant (Trapnell et al., 2012) or HT-seq-count (Anders, Pyl & Huber, 2015), and differential expression analysis can be run with Cuffdiff and Cuffnorm (Trapnell et al., 2012). A variant calling step that operates in a similar way than for de novo RNA-seq is also included.

RADseq

Another major novelty in NGScloud2 is the possibility of running RAD-seq bioinformatic workflows. This reduced genome representation methodology and its derivates (e.g. ddRADseq) are used to find out polymorphism in specific genomic regions nearby restriction enzyme cut sites in populations of multiple individuals, and has revealed powerful in phylogenetics, population genetics, and association mapping studies, among others (Andrews et al., 2016). In NGScloud2, we have included ddRADseqTools (Mora-Márquez et al., 2017) and RADdesigner (Guillardín-Calvo et al., 2019) to assess the optimal experimental design of a RADseq experiment, i.e. to choose the enzyme combinations, simulate the effect of allele dropout and PCR duplicates on coverage, quantify genotyping errors, optimize polymorphism detection parameters or determine sequencing depth coverage.

The workflow of RADseq data in NGScloud2 allows to analyze the data using two strategies (Figure 4). RADseq libraries can be mapped with Bowtie2, GSNAP or HISAT2 to an available genome or psedudogenome assembly. The pseudogenome can be assembled using the same (or complementary) reads with SOAPdenovo2 genomic assembler (Luo et al., 2012), or with the Starcode sequence clusterizer (Zorita, Cuscó & Filion, 2015). After read mapping, variant calling is performed in a similar way than for *de novo* RNA-seq. The alternative is to perform read clusterization, filtering and variant calling in a single step with the robust iPyrad pipeline (Eaton & Overcast, 2020).

Functional annotation

As a last improvement over the original version, NGScloud2 encapsulates our standalone application TOA (Mora-Márquez et al., 2020b), so it can run in EC2. This application automates the extraction of functional information from genomic databases, both plant specific (PLAZA) and general-purpose genomic databases (NCBI's RefSeq and NR/NT), and the annotation of sequences (Figure 5). TOA can be a good complement for both RNAseq and ddRADseq workflows in non-model plant species that has shown optimal performance in AWS's EC2 cloud. TOA aims to establish workflows geared towards woody plant species that automate the extraction of information from genomic databases and the annotation of sequences. TOA uses the following databases: Dicots PLAZA 4.0, Monocots PLAZA 4.0, Gymno PLAZA 1.0, NCBI RefSeq Plant and NCBI Nucleotide Database (NT) and NCBI Non-Redundant Protein Sequence Database (NR). Although TOA was primarily designed to work with woody plant species, it can also be used in the analysis of experiments on any type of plant organism. Additionally, NCBI Gene, InterPro and Gene Ontology databases are also used to complete the information.

NGShelper

Besides the cloud infrastructure deployed in NGScloud2, we have included a companion package, NGShelper that contains python utilities to post-process the output of NGScloud2 pipelines. The package contains some Bash (Linux) and Bat (Windows) scripts to facilitate running the Python3 programs.

NGShelper facilitates format conversion of output files, filtering and subsetting of results, VCF and FASTA files statistics extraction, among others. Utilities list and their usage and parameters can be consulted at https://github.com/GGFHF/NGShelper/blob/master/Package/help.txt.

Conclusions

NGScloud2 has significantly expanded the types of bioinformatic workflows to run using Amazon Web Services since its previous version. This new version has incorporated major technical improvements that optimize the use of popular software applications otherwise inaccessible to non-specialized users lacking suitable hardware infrastructures. Moreover, these technical improvements are oriented to significantly reduce costs by simplifying data access and taking advantage of EC2 spot instances that may produce savings of up to 50-80% in many steps of the analysis.

Acknowledgements

We'd like to thank M. Hurtado and Dr. V.M. Chano for beta-testing of NGScloud2 operability.

References

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–169. DOI: 10.1093/bioinformatics/btu638.

Andrews,S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17:81–92. DOI: 10.1038/nrg.2015.28.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34:525–527. DOI: 10.1038/nbt.3519.

Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. 2016. rnaQUAST: a quality assessment tool for de novo transcriptome assemblies: Table 1. *Bioinformatics* 32:2210–2212. DOI: 10.1093/bioinformatics/btw218.

Danecek P, McCarthy SA. 2017. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* 33:2037–2039. DOI: 10.1093/bioinformatics/btx100.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21. DOI: 10.1093/bioinformatics/bts635.

Eaton DAR, Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36:2592–2594. DOI: 10.1093/bioinformatics/btz966.

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods* 15:475–476. DOI: 10.1038/s41592-018-0046-7.

Guillardín-Calvo L, Mora-Márquez F, Soto Á, López de Heredia U. 2019. RADdesigner: a workflow to select the optimal sequencing methodology in genotyping experiments on woody plant species. *Tree Genetics & Genomes* 15:64. DOI: 10.1007/s11295-019-1372-3.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. DOI: 10.1093/bioinformatics/btt086.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 37:907–915. DOI: 10.1038/s41587-019-0201-4.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14:R36. DOI: 10.1186/gb-2013-14-4-r36.

Kwon T, Yoo WG, Lee W-J, Kim W, Kim D-W. 2015. Next-generation sequencing data analysis on cloud computing. *Genes & Genomics* 37:489–501. DOI: 10.1007/s13258-015-0280-7.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8:1494–1512. DOI: 10.1038/nprot.2013.084.

Langmead B, Nellore A. 2018. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics* 19:208–219. DOI: 10.1038/nrg.2017.113.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359. DOI: 10.1038/nmeth.1923.

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN. 2014. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology* 15:553. DOI: 10.1186/s13059-014-0553-5.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27:718–719. DOI: 10.1093/bioinformatics/btq671.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659. DOI: 10.1093/bioinformatics/btl158.

López de Heredia U, Mora-Márquez F, Goicoechea PG, Guillardín-Calvo L, Simeone MC, Soto Á. 2020. ddRAD Sequencing-Based Identification of Genomic Boundaries and Permeability in Quercus ilex and Q. suber Hybrids. *Frontiers in Plant Science* 11:1–16. DOI: 10.3389/fpls.2020.564414.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18. DOI: 10.1186/2047-217X-1-18.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10–12. DOI: 10.14806/ej.17.1.200.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nature Reviews Genetics* 12:671–682. DOI: 10.1038/nrg3068.

Mora-Márquez F, Chano V, Vázquez-Poletti JL, López de Heredia U. 2020b. TOA: a software package for automated functional annotation in non-model plant species. *Molecular Ecology Resources* (on-line). DOI: 10.1111/1755-0998.13285.

Mora-Márquez F, García-Olivares V, Emerson BC, López de Heredia U. 2017. ddRADseqTools : a software package for in silico simulation and testing of double-digest RADseq experiments. *Molecular Ecology Resources* 17:230–246. DOI: 10.1111/1755-0998.12550.

Mora-Márquez F, Vázquez-Poletti JL, Chano V, Collada C, Soto Á, de Heredia UL. 2020a. Hardware performance evaluation of de novo transcriptome assembly software in Amazon Elastic Compute Cloud. Current Bioinformatics 15:420–430. DOI: 10.2174/1574893615666191219095817.

Mora-Márquez F, Vázquez-Poletti JL, López de Heredia U. 2018. NGScloud: RNA-seq analysis of non-model species using cloud computing. *Bioinformatics* 34:3405–3407. DOI: 10.1093/bioinformatics/bty363.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842. DOI: 10.1093/bioinformatics/btq033.

Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W. 2003. Genome-wide characterization of the lignification toolbox in *Arabidopsis*. *Plant Physiology* 133(3):1051-1071. DOI: 10.1104/pp.103.026484

Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods* 10:71–73. DOI: 10.1038/nmeth.2251.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A-L, Tam A, Zhao Y, Moore RA, Hirst M, Marra

MA, Jones SJM, Hoodless PA, Birol I. 2010. De novo assembly and analysis of RNA-seq data. *Nature Methods* 7:909–912. DOI: 10.1038/nmeth.1517.

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research* 26:1134–1144. DOI: 10.1101/gr.196469.115.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7:562–578. DOI: 10.1038/nprot.2012.016.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V., Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* 35:543–548. DOI: 10.1093/molbev/msx319.

Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In: Mathé E, Davis S eds. *Statistical Genomics: Methods and Protocols*. Springer Science+Business Media New York, 283–334. DOI: 10.1007/978-1-4939-3578-9_15.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam T-W, Li Y, Xu X, Wong GK-S, Wang J. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30:1660–1666. DOI: 10.1093/bioinformatics/btu077.

Yang A, Troup M, Ho JWK. 2017. Scalability and Validation of Big Data Bioinformatics Software. *Computational and Structural Biotechnology Journal* 15:379–386. DOI: 10.1016/j.csbj.2017.07.002.

Zorita E, Cuscó P, Filion GJ. 2015. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* 31:1913–1919. DOI: 10.1093/bioinformatics/btv053.

Table 1(on next page)

Software available for de novo RNA-seq, reference-based RNAseq, RADseq and functional annotation in NGScloud2.

Recommendations for use spot or on demand instances are provided to optimize costs at every step of the workflows. (*) For time consuming processes that can be re-launched from the point of interruption, both spot or on demand instances may produce optimal performance, depending on the user's needs.

Workflow	Step	Software	Spot/On demand	Reference
de novo RNA-seq	Read pre-processing	FastQC	spot	Andrews, 2010
		cutadapt	spot	Martin, 2011
		Trimmomatic	spot	Bolger et al., 2014
		insilico_read_normalization (*)	spot	Haas et al., 2013
	Assembly	SOAPdenovo-Trans (*)	spot/on demand	Xie et al., 2014
		Trinity (*)	spot/on demand	Haas et al., 2013
		Trans-Abyss	on demand	Robertson et al., 2010
	Read alignmen	Bowtie2	on demand	Langmead & Salzberg, 2012
	Transcriptome quality assessment	BUSCO	spot	Waterhouse et al., 2018
		QUAST	spot	Gurevich et al., 2013
		rnaQUAST	spot	Bushmanova et al., 2016
		RSEM-EVAL	on demand	Li et al., 2014
		Transrate	spot/on demand	Smith-Unna et al., 2016
	Transcritpomefiltering	CD-HIT-EST	spot/on demand	Li & Godzik, 2006
		transcript-filtering	spot	https://github.com/GGFHF/NGShelper
	Quantitation	eXpress	spot	Roberts & Pachter2013
		Kallisto	spot	Bray et al., 2016
	Annotation	transcriptome-blast	on demand	https://github.com/GGFHF/NGShelper
	Variant calling	SAMtools BEDtools BCFtools Tabix (*)	spot	Li et al. 2009 Quinlan & Hall, 2010 Danecek & McCarthy, 2017 Li, 2011
Reference-based RNA-seq	Read pre-processing	FastQC	spot	Andrews, 2010
		cutadapt	spot	Martin, 2011
		Trimmomatic	spot	Bolger et al. 2014
	Read alignment	Bowtie2	on demand	Langmead & Salzberg, 2012
		GSNAP	on demand	Wu et al., 2016
		HISAT2	on demand	Kim et al., 2019
		STAR	on demand	Dobin et al., 2013
		TopHat	on demand	Kim et al., 2013
	Assembly	Cufflinks-Cuffmerge	spot	Trapnell et al., 2012
		Genome-guided Trinity (*)	spot/on demand	Haas et al., 2013
	Transcriptome alignment	GMAP	on demand	Wu et al., 2016
	Quantitation	Cuffquant	spot	Trapnell et al., 2012
		ht-seq-count	spot	Anders et al., 2015
	Differential expression	Cuffdiff	spot	Trapnell et al., 2012
		Cuffnorm	spot	Trapnell et al., 2012
	Variant calling	SAMtools BEDtools BCFtools Tabix (*)	spot	Li et al. 2009 Quinlan & Hall, 2010 Danecek & McCarthy, 2017 Li, 2011
RAD-seq	Design	rsitesearch	spot	Mora-Márquez et al., 2017
		ddRADseq simulation (*)	spot	Mora-Márquez et al., 2017
		RADdesigner (*)	spot	Guillardín-Calvo et al.,

				2019
	Read pre-processing	FastQC	spot	Andrews, 2010
		cutadapt	spot	Martin, 2011
		Trimmomatic	spot	Bolger et al. 2014
	Pseudo assembly	SOAPdenovo2 (*)	spot/on demand	Luo et al., 2012
	Read alignment	Bowtie2	on demand	Langmead & Salzberg, 2012
		GSNAP	on demand	Wu *et al.*, 2016
		HISAT2	on demand	Kim *et al.*, 2019
	Variant calling	SAMtools BEDtools BCFtools Tabix (*)	spot	Li *et al.* 2009 Quinlan & Hall, 2010 Danecek & McCarthy, 2017 Li, 2011
	Pipelines	ipyrad	on demand	Eaton & Overcast, 2020
Functional annotation	TOA annotation processes	TOA (*)	spot/on demand	Mora-Márquez et al., 2020b

Figure 1

Technical improvements of NGScloud2.

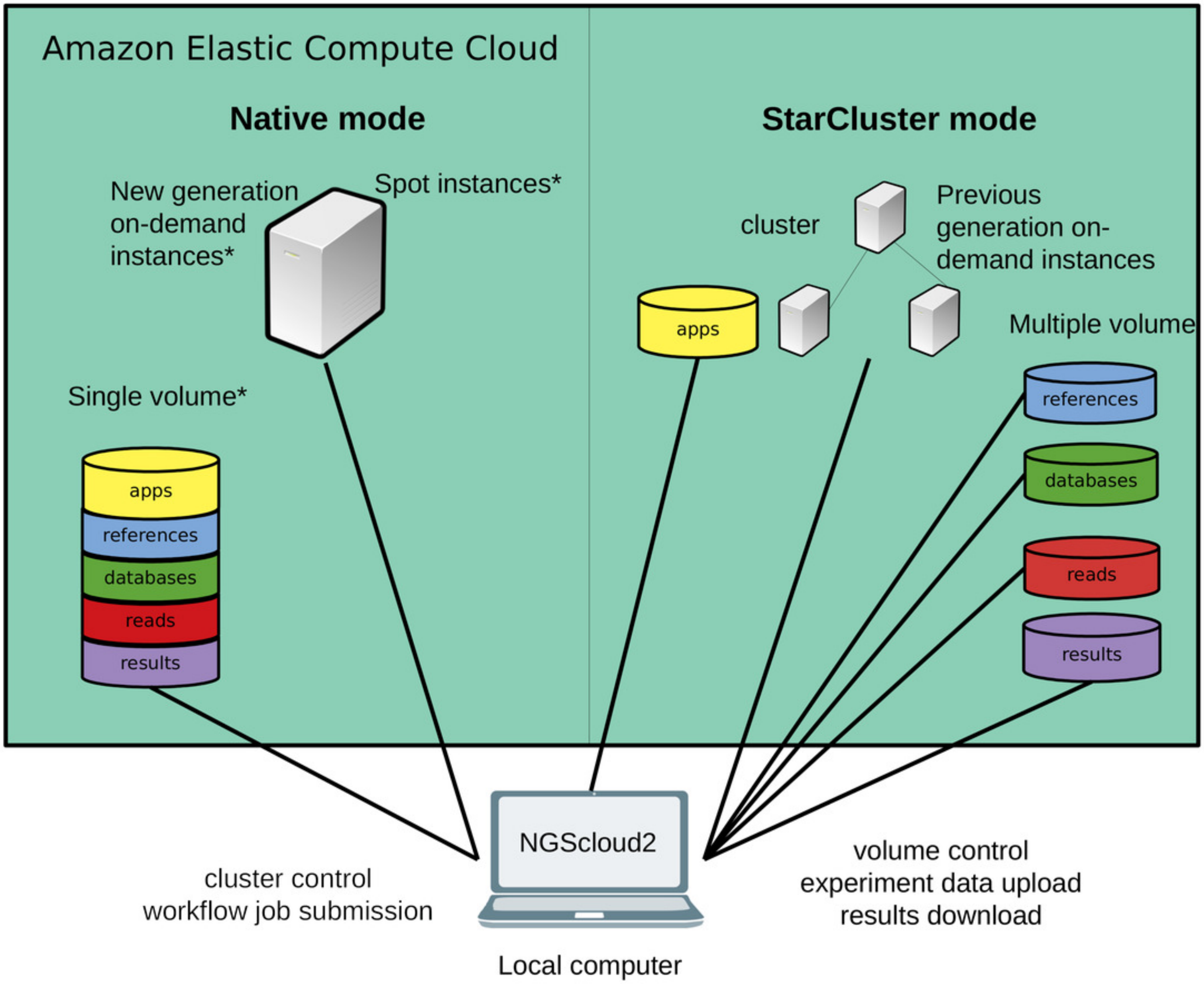


Figure 2

De novo RNAseq workflow in NGScloud2.

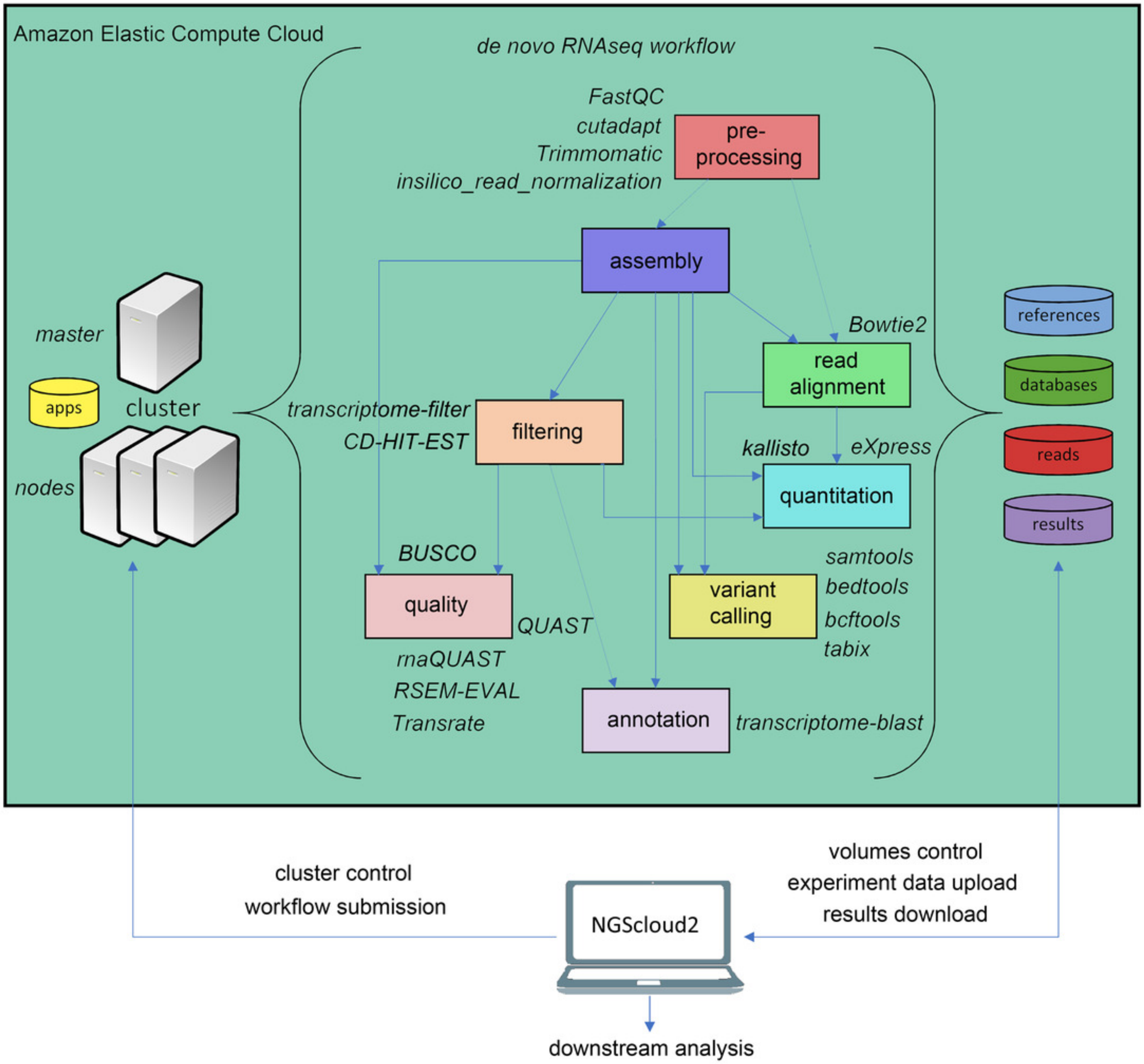


Figure 3

Reference-based RNAseq workflow in NGScloud2.

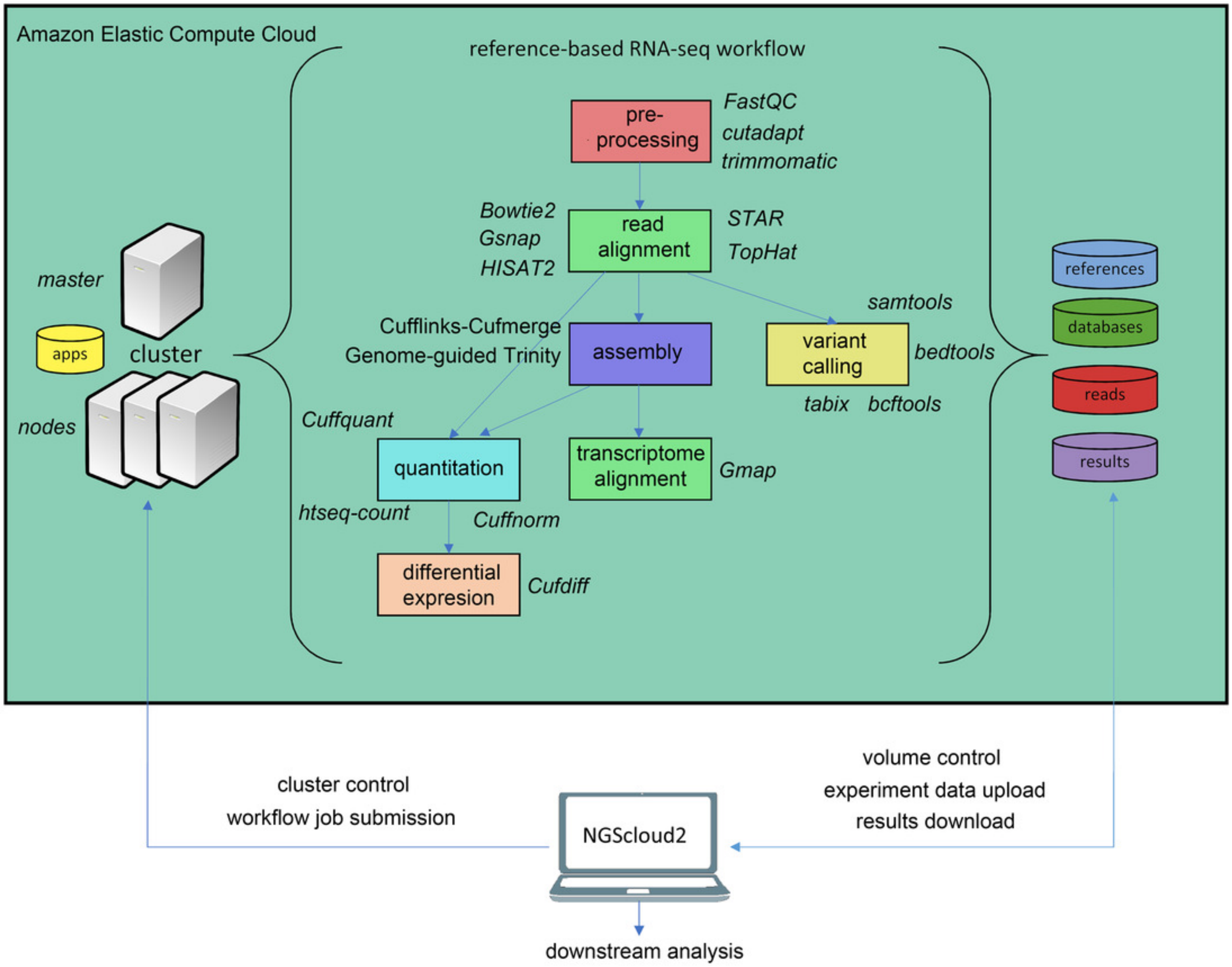


Figure 4

Reference-based RADseq workflow in NGScloud2.

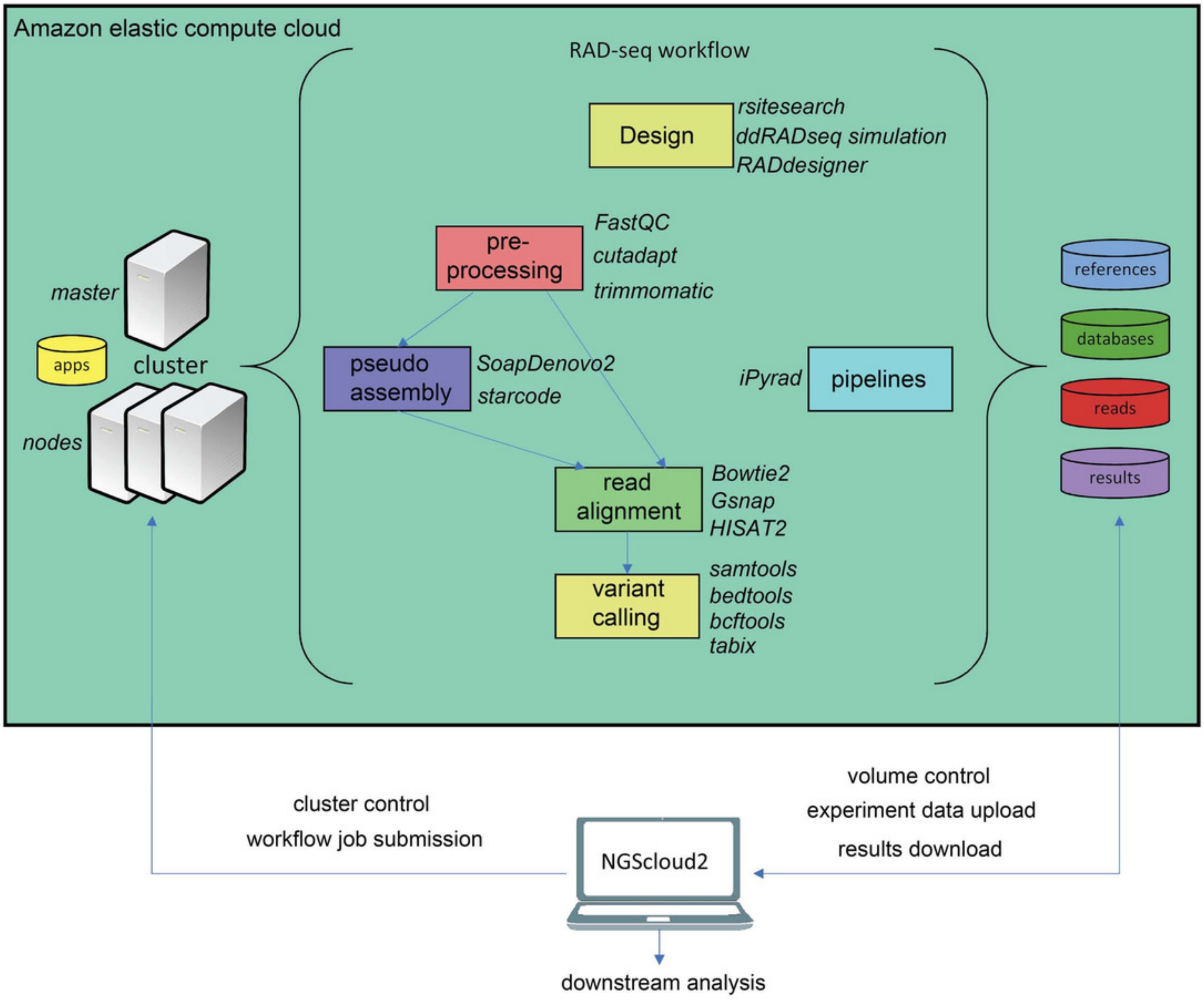
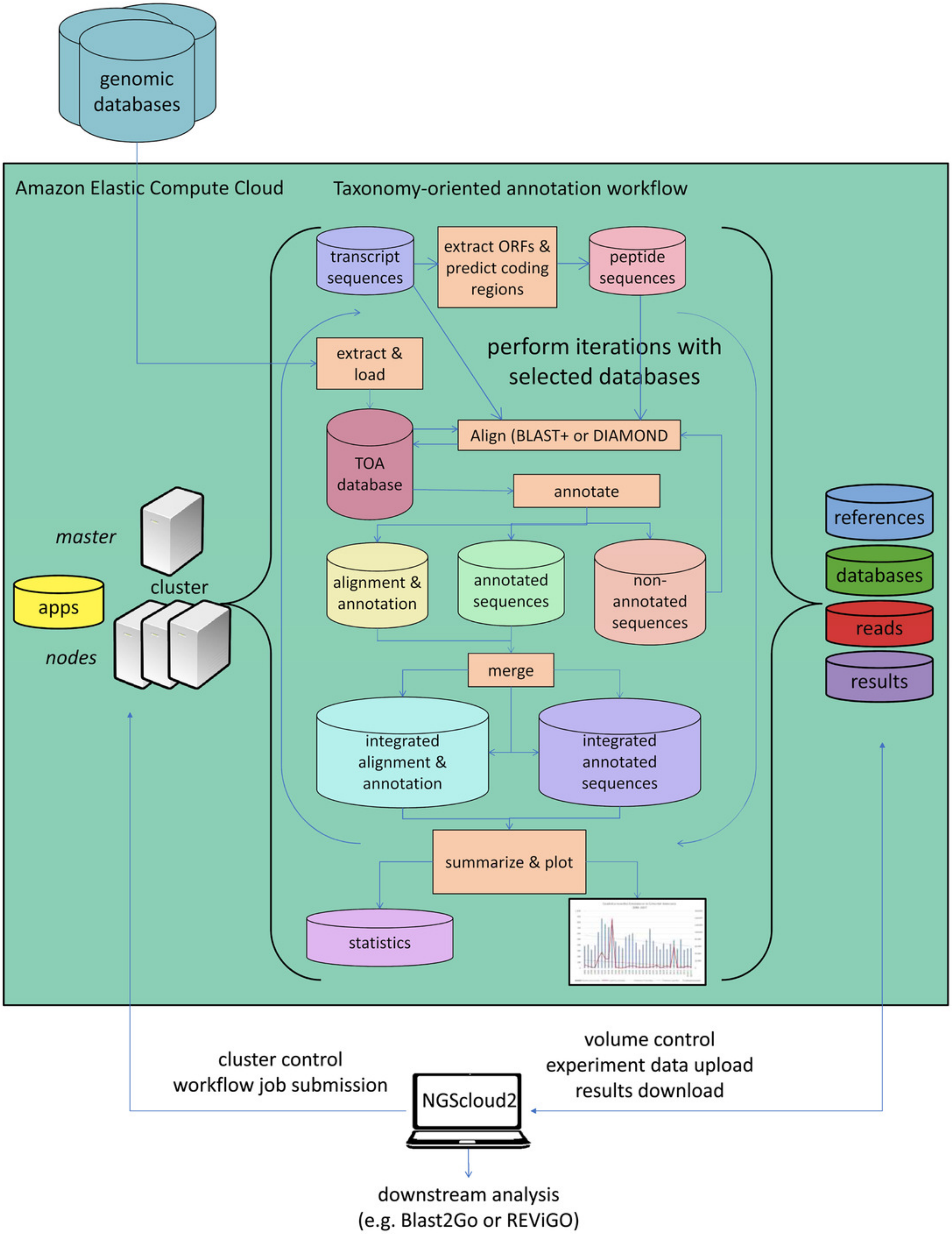


Figure 5

Functional annotation workflow in NGScloud2.


genomic databases
Amazon Elastic Compute Cloud
Taxonomy-oriented annotation workflow
transcript sequences
extract ORFs & predict coding regions
peptide sequences
extract & load
perform iterations with selected databases
TOA database
Align (BLAST+ or DIAMOND
annotate
master
cluster
apps
nodes
alignment & annotation
annotated sequences
non-annotated sequences
merge
integrated alignment & annotation
integrated annotated sequences
summarize & plot
statistics
references
databases
reads
results
cluster control
workflow job submission
NGScloud2
volume control
experiment data upload
results download
downstream analysis
(e.g. Blast2Go or REViGO)