BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 13 2014, pages 1928–1929
doi:10.1093/bioinformatics/btu135

Sequence analysis

Advance Access publication March 10, 2014

Orione, a web-based framework for NGS analysis in microbiology

Gianmauro Cuccuru*, Massimiliano Orsini, Andrea Pinna, Andrea Sbardellati, Nicola Soranzo, Antonella Travaglione, Paolo Uva, Gianluigi Zanetti and Giorgio Fotia

CRS4, Science and Technology Park Polaris, Piscina Manna, 09010 Pula (CA), Italy

Associate Editor: Michael Brudno

ABSTRACT

Summary: End-to-end next-generation sequencing microbiology data analysis requires a diversity of tools covering bacterial resequencing, *de novo* assembly, scaffolding, bacterial RNA-Seq, gene annotation and metagenomics. However, the construction of computational pipelines that use different software packages is difficult owing to a lack of interoperability, reproducibility and transparency. To overcome these limitations we present Orione, a Galaxy-based framework consisting of publicly available research software and specifically designed pipelines to build complex, reproducible workflows for next-generation sequencing microbiology data analysis. Enabling microbiology researchers to conduct their own custom analysis and data manipulation without software installation or programming, Orione provides new opportunities for data-intensive computational analyses in microbiology and metagenomics.

Availability and implementation: Orione is available online at http://orione.crs4.it.

Contact: gianmauro.cuccuru@crs4.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 27, 2013; revised on December 18, 2013; accepted on March 4, 2014

1 INTRODUCTION

Application of next-generation sequencing (NGS) in microbiology is becoming a common practice with a profound impact on research, diagnostic and clinical microbiology (Loman *et al*., 2012). Recent applications include genomic sequencing, differential transcription analysis, variant investigation, as well as metagenomics studies. Major challenges include draft assemblies finishing followed by reliable genome annotation or robust dissection of microbial communities including those associated with human health and disease. Furthermore, there is an increasing need to process and present data in a fashion that is transparent and reproducible and to provide analysis frameworks that are usable and cost-effective for biomedical researchers.

To address these challenges, we developed Orione, an online framework for integrative analysis of NGS microbiology data. Orione is based on Galaxy (Goecks *et al*., 2010), an open platform for reproducible data-intensive computational analysis used in many diverse biomedical research environment. Orione is the first freely available platform that supports the whole life cycle of microbiology research data from production and annotation to publication and sharing. Other commercial alternative exists (e.g. CLC Genomics Workbench by CLC Bio), but Orione is unique in transparently combining the most used open source bioinformatics tools for microbiology. Orione is currently applied to a variety of microbiological projects including bacteria resequencing, *de novo* assembling and microbiome investigations; see http://goo.gl/DwbgPD for a list. Furthermore, Orione is part of an ongoing project to integrate Galaxy with Hadoop-based tools to provide scalable computing (Leo *et al*., 2012); a specialized version of OMERO (Allan *et al*., 2012) to model biomedical data and the chain of actions that connect them; and iRODS (Rajasekar *et al*., 2010) to efficiently support inter-institutional data sharing. This infrastructure is already used in production at Center for Advanced Studies, Research and Development in Sardinia for the automated processing of sequencing data (Pireddu *et al*., 2013) and for quality control in gene therapy applications (Biffi *et al*., 2013).

2 FEATURES AND METHODS

Orione consists of 'best-of-breed' NGS bioinformatics tools covering end-to-end data analysis for bacterial resequencing, *de novo* assembly, scaffolding, bacterial RNA-Seq, gene annotation, metagenomics and metatranscriptomics. Publicly available research tools were integrated under the open source Galaxy framework with pipelines and workflows newly developed by our group for ready-to-go microbiological analysis. Although several of the tools for NGS microbiology data analysis were already available in Galaxy, a significant effort was required to expand the Galaxy functionalities with new features such as SSPACE (Boetzer *et al*., 2011), SSAKE (Warren *et al*., 2007), SOPRA (Dayarian *et al*., 2010), SEQuel (Ronen *et al*., 2012), EDGE-pro (Magoc *et al*., 2013), Gene Locator and Interpolated Markov ModelER (Delcher *et al*., 2007) and Prokka (http://goo.gl/aSuHb). We refer to the Supplementary information for a description of the complete set of Orione tools and workflows.

3 FUNCTIONALITIES

Orione complements the flexible Galaxy workflow environment, allowing microbiologists without any specific hardware or informatics skill to consistently access a set of NGS data analysis tools and conduct reproducible data-intensive computational analyses from quality control to microbial gene annotation. In the following paragraphs, we describe the main Orione functionalities.

*To whom correspondence should be addressed.

Preprocessing, quality control and trimming. The fundamental step before any NGS analysis is the quality control of reads and their trimming. To cope with long reads and paired-end technology, FastX (http://goo.gl/GxqyV) and FASTQC (http://goo.gl/6TUqD) were complemented with specifically developed tools (see also workflow #1 in the Supplementary information).

Reads mapping. Mapping is a key step in many NGS applications from bacteria resequencing to variant calling. The most widely used aligners are integrated in Orione, including BWA (Li and Durbin, 2009), Bowtie1 (Langmead *et al*., 2009), Bowtie2 (Langmead and Salzberg, 2012), SOAP (Li *et al*., 2008) and MOSAIK (http://git.io/QrYWXg). We further added BLAT (Kent, 2002), SHRiMP (David *et al*., 2011), LASTZ (Harris, 2007) and BFAST (Homer *et al*., 2009) for use with long reads from 454 Roche.

***De novo* assembly.** *De novo* assembly produces contigs without the aid of a reference genome. Different methods, either based on a de Bruijn graph [Velvet (Zerbino and Birney, 2008), ABySS (Simpson *et al*., 2009) and SPAdes (Bankevich *et al*., 2012)] or on a greedy approach [SSAKE, Edena (Hernandez *et al*., 2008)], are available in Orione.

Scaffolding. After mapping, contigs are ordered and oriented to produce even longer sequences called scaffolds, exploiting the mate-pair/paired-end information. Orione includes the most established scaffolders such as SSAKE, SSPACE, SEQuel and SOPRA.

Post assembly, contigs statistics, (multi) aligning and variant calling. This section of Orione includes tools we have developed covering task such as genome-scale alignment, high-quality contigs extraction, statistics over contigs or draft genomes (N50/NG50 values, contigs length distribution, high/low quality regions/gaps in draft genomes).

Annotation. Annotation is the process of identifying meaningful biological information from sequences. Glimmer and tRNAscan-SE (Lowe and Eddy, 1997) were wrapped into Orione together with the Prokka pipeline, enabling easy Genbank/DDJB/ENA submission.

RNA-Seq. We integrated EDGE-pro tool for bacterial RNA-Seq analysis. As EDGE-pro requires genome annotation files, we developed an accessory tool ('Get EDGE-pro files') that retrieves them directly from the NCBI RefSeq repository.

Metagenomics and other tools. We added to the standard Galaxy metagenomics pipeline MetaPhlAn (Segata *et al*., 2012) and MetaVelvet (Namiki *et al*., 2012). The MetaGeneMark (Zhu *et al*., 2010) annotation tool has been added for gene prediction in metagenomic sequences and a workflow has been developed for (bacterial) metatranscriptome analysis. We complete this section with instruments for data filtering, conversion and taxonomy abundance displaying into the Krona visualizer (Ondov *et al*., 2011).

ACKNOWLEDGEMENTS

The authors would like to thank Dr Cesare Cammà (Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise) and Prof. Sergio Uzzau (Università di Sassari and Porto Conte Ricerche) for providing us with the data we used for the set up of Orione.

Funding: This work was partially supported by the Sardinian Regional Authorities and the Wellcome Trust (095931).

Conflict of Interest: none declared.

REFERENCES

Allan,C. *et al.* (2012) OMERO: flexible, model-driven data management for experimental biology. *Nat. Methods*, **9**, 245–253.

Bankevich,A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Biffi,A. *et al.* (2013) Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science*, **341**, 1233158.

Boetzer,M. *et al.* (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.

David,M. *et al.* (2011) Shrimp2: sensitive yet practical short read mapping. *Bioinformatics*, **27**, 1011–1012.

Dayarian,A. *et al.* (2010) SOPRA: scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics*, **11**, 345.

Delcher,A.L. *et al.* (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Harris,R.S. (2007) Improved pairwise alignment of genomic DNA. PhD Thesis, Pennsylvania State University, PA.

Hernandez,D. *et al.* (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.*, **18**, 802–809.

Homer,N. *et al.* (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, **4**, e7767.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Leo,S. *et al.* (2012) SNP genotype calling with MapReduce. In: *Proceedings of The Third International Workshop on MapReduce and its Applications*. MapReduce'12. ACM, pp. 49–56. New York, NY.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,R. *et al.* (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Loman,N.J. *et al.* (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.*, **10**, 599–606.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Magoc,T. *et al.* (2013) EDGE-pro: estimated degree of gene expression in prokaryotic genomes. *Evol. Bioinform. Online*, **9**, 127–136.

Namiki,T. *et al.* (2012) MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.*, **40**, e155.

Ondov,B.D. *et al.* (2011) Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, **12**, 385.

Pireddu,L. *et al.* (2013) Automated and traceable processing for large-scale high-throughput sequencing facilities. *EMBnet. J.*, **19**, 23–24.

Rajasekar,A. *et al.* (2010) iRODS primer: integrated rule-oriented data system. In: *Synthesis Lectures on Information Concepts, Retrieval, and Services*. Vol. 2, Morgan&Claypool, San Rafael, CA, pp. 1–143.

Ronen,R. *et al.* (2012) SEQuel: improving the accuracy of genome assemblies. *Bioinformatics*, **28**, i188–i196.

Segata,N. *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–814.

Simpson,J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.

Warren,R.L. *et al.* (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23**, 500–501.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.

Zhu,W. *et al.* (2010) *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.*, **38**, e132.