EMBnet.journal
Bioinformatics in Action

Report on the Swiss-Colombian workshop: *"Assembly, annotation and comparison of bacterial genomes"*

Laurent Falquet[1], Sandra Patricia Calderon-Copete[2], Emiliano Barreto-Hernández[3], Jaime Enrique Moreno Castañeda[4]

[1]University of Fribourg and Swiss Institute of Bioinformatics, Biochemistry Unit, Fribourg, Switzerland; [2]LGTF-UniL, Génopode, Lausanne, Switzerland; [3]Centro de Bioinformática, Instituto de Biotecnología - Universidad Nacional de Colombia, Bogotá, Colombia; [4]Grupo de Microbiología-Investigación Instituto Nacional De Salud, Bogotá, Colombia.

Introduction

This workshop was organised as part of the Swiss-Colombian project, *A pilot integrative knowledgebase for the characterization and tracking of multi resistant Acinetobacter baumannii in Colombian Hospitals*, sponsored by the Leading House Cooperation and Development Centre[1] of the École Polytechnique Fédérale Lausanne (EPFL).

The aim of this project is to develop a prototype centralised knowledgebase. Initially, we selected complete genome sequences obtained from a collection of *Acinetobacter baumannii* strains collected from the Antimicrobial-Resistant Healthcare-Associated Infections Surveillance Program, during 2012–2015, by the Colombian National Health Institute (NHI) and the Biotechnology Institute of the National University of Colombia (IBUN-UNAL). In addition, complete *Acinetobacter baumannii* genome sequences were added from public databases. The prototype will consist of fully assembled and annotated genomes associated with geographical, temporal and clinical data, allowing tracking of a variety of infection outbreaks. The resulting knowledgebase will serve as a reference to help clinicians to track rapid dissemination of highly pathogenic and resistant strains.

The workshop was held at the Bioinformatics centre of the National University of Bogotá, 23-27 May 2016, and gathered 18 participants from diverse institutions in Colombia.

Programme

Each half-day was split into theoretical lectures (60-90 minutes each), followed by hands-on practicals (150 minutes each). The programme is shown in Table 1.

Table 1. Programme of the Swiss-Colombian workshop, 23-27 May 2016.

Day 1	Introduction to UNIX and computer clusters Introduction to sequencing techniques, QC and data cleaning (adapter removal, trimming, filtering, etc.)
Day 2	De novo assembly Assembly by re-mapping
Day 3	SNP and small indel calling: how to detect variants? Annotation and profiling of resistance and virulence factors
Day 4	Comparative genomics (core/pan genomes, structural variants, phylogeny distribution)
Day 5	Presentation of individual research projects of participants

Organisation of the work

Given the lack of access to a high-performance cluster, the participants were divided into nine groups of two, each being responsible for the analysis of a set of paired-end 100 bp reads from Illumina sequencing of a strain of *Acinetobacter baumannii* from the Sequence Read Archive (SRA). To distribute the workload, we divided the work across five computing nodes (16 cores, 64 Gb RAM). After a brief reminder of computing and UNIX operating system basics, participants had the opportunity to refresh their knowledge of the command line. The genome analysis comprised data quality control with FastQC[2], cleaning both the adapter content with CutAdapt (Martin, 2011) and low quality sequences with sickle[3]. The cleaned sequences were assembled with SOAPdenovo (Luo *et al.*, 2012), using various kmers, and SPAdes (Bankevich *et al.*, 2012). The draft genomes were compared using summary statistics, QUAST (Gurevich *et al.*, 2013) and MAUVE (Darling *et al.*, 2010). The best draft genome was annotated using Prokka (Seemann, 2014) and a set of HMMs built from the Virulence Factor database (Chen *et al.*, 2016) and downloaded from the ResFam database (Gibson *et al.*, 2015). The gff files of ten genomes (nine, plus reference) were compared, looking for core and pan genomes using Roary (Page *et al.*, 2015) and Phandango[4]. The reads were also re-mapped to the reference genome, and SNPs and indels called with BWA (Li and Durbin, 2010), SAMtools and BCFtools (Li, 2011). The SNP vcf files were annotated with snpEff, filtered with SnpSift (Cingolani *et al.*, 2012) and finally visualised with IGV (Thorvaldsdottir *et al.*, 2013). After conversion to multi-fasta format, a tree was constructed

Article history
Received: 22 August 2016
Published: 9 November 2016

[1] http://cooperation.epfl.ch/cms/lang/fr/pid/118892

[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] https://github.com/najoshi/sickle

[4] http://jameshadfield.github.io/phandango/

with FastTree (Price *et al.*, 2010) and visualised using the Newick viewer (Boc *et al.*, 2012).

Finally, participants had the opportunity to present their own current research work and to receive feedback from other course participants and trainers, promoting an enriching exchange of valuable research experiences in the area of genomics.

Figure 1. Participants and trainers.

Evaluation of the course

Participants from different research institutions expressed satisfaction with the high academic level of the course in general. They gave high value to the knowledge shown by trainers, and to the materials used in the lectures and practical exercises. Some respondents said that knowledge acquired during the course had allowed them to solve their own data-analysis problems.

They also made recommendations regarding the inclusion of additional practicals, and the possibility of additional access to the servers used for the hands-on sessions, in order to become more familiar with Linux. Course servers will be available to them for a few months more.

Figure 2. Workshop hand-on session.

Conclusions

According to the attendees' course evaluation and the organisers comments, this workshop was very useful both for biologists working on assembly and annotation of bacterial genomes, and researchers of the Colombian NHI, interested in tracking resistance and virulence factors in clinical isolates.

Acknowledgements

This work was supported by the "Leading House Cooperation and Development Centre". We thank Hermes Perez Cardona and Dra. Maria Teresa Reguero, for their help with local organisation of the workshop.

References

1. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. http://dx.doi.org/10.1007/978-3-642-37195-0_13
2. Boc A, Diallo AB, Makarenkov V (2012) T-REX: a Web server for inferring validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.* **40**, W573–W579. http://dx.doi.org/10.1093/nar/gks485
3. Chen L, Zheng D, Liu B, Yang J and Jin Q (2016) VFDB 2016: hierarchical and refined dataset for big data analysis - 10 years on. *Nucleic Acids Res* **44**, D694–D697. http://dx.doi.org/10.1093/nar/gkv1239
4. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2; iso-3. *Fly* (Austin) **6**, 80–92. http://dx.doi.org/10.4161/fly.19695
5. Darling AE, Mau B and Perna NT (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain Loss and Rearrangement. *PLoS ONE* **5**, e11147. http://dx.doi.org/10.1371/journal.pone.0011147
6. Gibson MK, Forsberg KJ, Dantas G (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* **9**, 207–216. 6. https://dx.doi.org/10.1038/ismej.2014.106
7. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. http://dx.doi.org/10.1093/bioinformatics/btt086
8. Li H (2011) A statistical framework for SNP calling mutation discovery association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993. http://dx.doi.org/10.1093/bioinformatics/btr509
9. Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. http://dx.doi.org/10.1093/bioinformatics/btp698
10. Luo R, Liu B, Xie Y, Li Z, Huang W *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18. http://dx.doi.org/10.1186/2047-217x-1-18
11. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12. http://dx.doi.org/10.14806/ej.17.1.200
12. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693. http://dx.doi.org/10.1093/bioinformatics/btv421
13. Price MN, Dehal PS and Arkin AP (2010) FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490. http://dx.doi.org/10.1371/journal.pone.0009490
14. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30**, 2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153
15. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192. http://dx.doi.org/10.1093/bib/bbs017