Bioinformatics, 31(17), 2015, 2870–2873
doi: 10.1093/bioinformatics/btv283
Advance Access Publication Date: 6 May 2015
Applications Note

Genome analysis

IonGAP: integrative bacterial genome analysis for Ion Torrent sequence data

Adrian Baez-Ortega[1,†], Fabian Lorenzo-Diaz[2,3,†], Mariano Hernandez[2], Carlos Ignacio Gonzalez-Vila[1], Jose Luis Roda-Garcia[4], Marcos Colebrook[4,*] and Carlos Flores[2,3,5,*]

[1]Information Technology Department, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain, [2]Applied Genomics Group (G2A), Instituto Universitario de Enfermedades Tropicales y Salud Pública de Canarias (CIBICAN), Universidad de La Laguna, Santa Cruz de Tenerife, Spain, [3]Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain, [4]Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Santa Cruz de Tenerife, Spain and [5]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on December 4, 2014; revised on March 23, 2015; accepted on April 29, 2015

Abstract

Summary: We introduce IonGAP, a publicly available Web platform designed for the analysis of whole bacterial genomes using Ion Torrent sequence data. Besides assembly, it integrates a variety of comparative genomics, annotation and bacterial classification routines, based on the widely used FASTQ, BAM and SRA file formats. Benchmarking with different datasets evidenced that IonGAP is a fast, powerful and simple-to-use bioinformatics tool. By releasing this platform, we aim to translate low-cost bacterial genome analysis for microbiological prevention and control in healthcare, agroalimentary and pharmaceutical industry applications.
Availability and implementation: IonGAP is hosted by the ITER's Teide-HPC supercomputer and is freely available on the Web for non-commercial use at http://iongap.hpc.iter.es.
Contact: mcolesan@ull.edu.es or cflores@ull.edu.es
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The application of Next-Generation DNA Sequencing (NGS) to bacterial genomic studies has a great impact in clinical microbiology diagnosis and epidemiology, as well as in the management of infectious disease outbreaks. Given the abundance of information produced by the NGS platforms, affordable and fast access to whole-genome sequence by means of intuitive bioinformatics tools demanding minimal training is mandatory to translate this technology into the clinical settings routine (Frike and Rasko, 2014).

Ion Torrent semiconductor sequencing technology (Thermo Fisher Scientific, Inc.) benefits from high sequencing speed at low cost, constituting one of the preferred choices in settings demanding fast turnaround NGS routine analysis (Frey *et al.*, 2014). Nevertheless, a comprehensive bacterial genomic data analysis demands a number of software packages, which often have cumbersome installation and use, and lack a graphical interface. Existing Web platforms such as Orione (Cuccuru *et al.*, 2014) aim to solve this issue *albeit* the user still must configure many genome assembly and analysis options. Alternatives for local execution of analyses, often requiring expert installation, are also available for genome assembly/annotation (Kislyuk *et al.*, 2010) or variant calling (Qi *et al.*, 2010). However, they were designed for Illumina and Roche/454 sequencing data. Therefore, there is no specific package providing an integrated collection of microbial genomic applications, specially

Table 1. Applications included in IonGAP to perform comparative genomics, annotation and bacterial classification routines.

Application	Process	Result
Comparative Genomics Module		
Mauve[a]	Genome alignment, contig reordering	Reordered contigs; alignment summary; information on indels and missing regions.
Cortex[b]	Variant calling	Variant calls in VCF file
TRAMS[c]	Annotation	Functional annotation of SNPs
MUMmer[d], Circos[e], Circoletto[f], genoPlotR[g]	Genome visualization	Linear and circular alignment graphs
Bacterial Classification & Annotation Module		
BLAST[h], NCBI 16S rRNA DB[i]	Taxonomic classification	16S rRNA sequence alignments for each contig
Torsten Seemann's MLST[J]	Multilocus sequence typing	Identified allele numbers and Sequence Type; allele sequences
Prokka[k]	Genome annotation	Annotated contigs (several formats) and protein sequences
BLAST[h], NCBI Plasmids DB[l]	Identification of plasmids	Plasmid sequence alignments for each contig
BLAST[h], MvirDB[m]	Identification of pathogenic factors	Antibiotic resistance/virulence genes, and pathogenicity islands alignments for each contig

[a]Darling *et al.* (2014).
[b]Iqbal *et al.* (2012).
[c]Reumerman *et al.* (2013).
[d]Kurtz *et al.* (2004).
[e]Krzywinski *et al.* (2009).
[f]Darzentas (2010).
[g]Guy *et al.* (2010).
[h]Altschul *et al.* (1990).
[i]ftp://ftp.ncbi.nih.gov/blast/db.
[j]https://github.com/Victorian-Bioinformatics-Consortium/mlst.
[k]Seemann (2014).
[l]ftp://ftp.ncbi.nlm.nih.gov/genomes/Plasmid.
[m]Zhou *et al.* (2007).

configured for rapid handling of data generated by Ion Torrent sequencing. For this reason, we have developed IonGAP, a web-based genome analysis platform to straightforwardly perform both the assembly process and subsequent comparative analysis, variant calling, functional annotation and bacterial typing routines in a seamless, user-friendly way. Therefore, IonGAP offers a first-line solution, making data analysis accessible to non-specialists but compatible with downstream advanced applications.

2 Features and Functionalities

The IonGAP processing pipeline is structured in three independent modules, which can be optionally disabled in order to allow a user-customized workflow. Once the project has finished, the user is notified by an email from which the compressed results folder can be downloaded. This includes a summary HTML file for the user to facilitate browsing the results.

2.1 Genome assembly

The first module is composed by the MIRA assembler (Chevreux *et al.*, 1999) and the FastQC quality control software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The assembler configuration has been greatly simplified by means of a Web interface, which allows to start a *de novo* assembly project simply by submitting a FASTQ, BAM or SRA file. The platform admits compressed files in *.zip* or *.tar.bz2* format, which can be obtained directly using the FileExporter plugin of the Torrent Suite. The input file may be provided by direct upload or, alternatively, through a FTP or cloud storage URL. At this moment, the pipeline limits the input file size for this step to approximately 3 million reads or 1 GB (assuming an uncompressed FASTQ file). The user is allowed to configure a variety of relevant assembly parameters, in case the default assembly is not satisfactory, as well as choosing between 11 assembly output formats. The assembler conducts read pre-processing and filtering. The assembly stage may also be omitted, as the service allows supplying a file of assembled contigs (in FASTA format) as an input for the rest of the pipeline, making it practical also for pre-processed data from other NGS technologies (e.g. Illumina, Roche/454, Pacific Biosciences).

2.2 Comparative genomics

To execute this module, a unique reference genome sequence (FASTA format) or its NCBI accession/Sequence identification number must be provided by the user. Comparative analysis is conducted by Mauve (Darling *et al.*, 2014), Cortex (Iqbal *et al.*, 2013) and TRAMS (Reumerman *et al.*, 2013), and involves contig alignment and reordering based on the reference genome, as well as subsequent identification of genetic variants, including single nucleotide polymorphisms (SNPs), indels, complex polymorphisms, structural variants and missing regions (Table 1). As part of this, IonGAP outputs a VCF file for downstream epidemiological and evolutionary analyses, and a table of SNPs with their functional annotation. Furthermore, publication-ready reports are generated by different comparative visualization tools (Table 1 and Supplementary data). Integrating Mauve in this module offers diverse capabilities and versatility (Edwards *et al.*, 2013), and provides assembly quality metrics. Cortex allows discovering and genotyping variants with high specificity and sensitivity, and provides genotype confidence scores to derive a high quality variant call set (Iqbal *et al.*, 2012). This memory-efficient suite works perfectly with haploid genomes (Iqbal *et al.*, 2013), and has been extensively used for bacterial epidemiological analyses.

Table 2. Summary of data generated by IonGAP on distinct control (*E.coli*) and case study (*S.aureus, M.ulcerans*) datasets.

Dataset	*E.coli* (100/400)[a]	*S.aureus*	*M.ulcerans*
#SNPs against reference[b,c]	0/8	19 671	218
Synonymous	0/3	6 416	21
Non-synonymous	0/1	1 822	61
#Annotated genes	4 225/4 453	2 627	5 694
Virulence[d]	111/94	48	0
Antibiotic resistance[d]	25/20	6	2
#Plasmids alignments[e]	16/19	16	14
Total assembly size (Mb)	4.48/4.59	2.83	5.37
Reference genome (Mb)[c]	4.69	2.82	5.63

[a]Summary data of the two control datasets (separated by a slash) is provided.

[b]Only SNP calls with genotype confidence ≥10 were annotated.

[c]Strains utilised as references were *E.coli* DH10B, *S.aureus* NCTC8325, and *M.ulcerans* Agy99 (see Supplementary data for details).

[d]Virulence and antibiotic resistance gene alignments ≥500 bp are indicated.

[e]Only plasmid alignment lengths ≥3 000 bp were considered.

2.3 Bacterial classification and annotation

The last module of the platform is mainly dedicated to classification, gene annotation and identification of mobile elements and pathogenic factors. The underlying applications and their output results are summarized in Table 1. The classification based on 16S rRNA gene sequence might be useful as a checkpoint for sample DNA contaminations and/or phenotypic misidentification. Multilocus sequence typing (MLST) analysis runs only if the bacterial species is selected by the user, and might help in a rational determination of the closest reference genome(s) for further comparisons in case a MLST scheme is available for the species. Projects running the whole IonGAP routines employ the reordered contigs (based on a given reference genome). Nevertheless, the original (unordered) contigs can be used if the *Comparative genomics* module is disabled.

3 Results and Discussion

3.1 Testing using control datasets

The IonGAP pipeline has been tested using two public *Escherichia coli* DH10B genome datasets, generated by the Ion PGM System with 100 or 400 bp sequencing chemistries (refer to Supplementary data). As expected, the best assembly results were obtained using 400 bp reads; IonGAP assembled 155 contigs containing a total of 4 589 812 bp, with a N50 length of 118 424 bp. In order to compare with a reference pipeline, the same datasets were analyzed in the Orione software by running the 'Pre-processing' and 'Bacterial *de novo* assembly' workflows (Cuccuru *et al.*, 2014). The whole process required 2 h and 38 min (single-end 400 bp dataset), and although total assembled bases and annotated protein coding genes were similar to those generated by IonGAP, a N50 length of 57 757 bp was obtained (see Supplementary data).

Besides assembly, IonGAP integrates two modules devoted to compare, annotate and classify bacterial genomes (Table 1). In this regard, the results generated by analyzing the mentioned datasets were consistent with those previously reported for the reference genome, corresponding to the same strain (Durfee *et al.*, 2008). A closer inspection of the annotated SNPs list generated by the *Comparative genomics* module evidenced the high performance of the variant calling. For instance, from the total SNP calls retrieved for *E.coli* control datasets against the reference, as few as eight of them supported a nucleotide substitution in the worst case scenario (400 bp dataset) when a minimally stringent threshold for confidence was imposed (see Table 2 and Supplementary data). Based on the 16S rRNA gene classification and MLST, sequencing reads were unequivocally assigned to *E.coli* ST1060. Sequence annotation revealed up to 4 453 protein-coding genes, as well as 19 (>3 kb in length) and 25 (>500 bp) contig alignments to previously described plasmid and antibiotic resistance gene sequences, respectively (Table 2). Furthermore, different virulence proteins coding genes ($n = 111$) and pathogenicity islands ($n = 49$) were identified. Noteworthy, the whole IonGAP process only required up to 1 h and 44 min.

3.2 Validation using case study datasets

To further evaluate its performance, IonGAP was tested with two different case study datasets derived with 200 bp chemistry corresponding to bacteria with low or high percentage of GC genome content (see Supplementary data). One of the datasets corresponds to *Staphylococcus aureus* (strain M34-B-1_11), a Gram-positive pathogen with 32.9% GC content that can be found in the upper respiratory tract and skin of healthy individuals, representing one of the most common causes of hospital acquired infections worldwide. The other dataset corresponds to *Mycobacterium ulcerans* (strain Mu_F74), a Gram-positive bacterium with 65.5% GC content that causes ulcerative skin diseases mainly confined to sub-Saharan African human populations.

Analysis of the total processing time as well as assembly quality metrics such as the number of contigs, N50 length and total assembly size, evidenced that IonGAP outperformed Orione in both case study datasets (see Supplementary data). Conversely, although both datasets are comparable in terms of sequence reads number, average read length and total sequenced bases, the two datasets performed remarkably different in IonGAP, resulting in dissimilar assembly quality metrics and processing time (see Supplementary data). Ninety-four contigs with a N50 length of 218 499 bp were obtained for *S.aureus* in as few as 3 h and 32 min. However, a total of 1 352 contigs (N50 length of 7 674 bp) were generated for *M.ulcerans*. Furthermore, in this latter case, IonGAP required more than 20 h for completing the assembly process. Such discrepancies might be attributed to the *M.ulcerans* genome complexity, which harbours a megaplasmid as well as several insertion sequence elements (>1 kb) in high-copy number (Doig *et al.*, 2012).

Some of the most relevant results generated by IonGAP for the case study datasets are summarized in Table 2. Apart from the total assembly size, the number of annotated genes in each dataset fits with values reported for their respective reference genomes. As expected, the variant calls also revealed differences between the two datasets. For *S.aureus*, the comparison with a genetically distant reference genome (ST398 versus ST8) rendered a total of 19 671 confident SNP calls, from which 32.6 and 9.3% accounted for synonymous and non-synonymous SNPs, respectively. In contrast, only 278 SNP calls were retrieved from the *M.ulcerans* dataset (~1:3 synonymous/non-synonymous ratio). Worth mentioning, IonGAP only required 21 and 27 min to execute the *Comparative genomics* and *Bacterial classification and annotation* modules with the *S.aureus* and *M.ulcerans* datasets, respectively.

Acknowledgements

Thanks are due to members of the ETSII Computing Centre (Universidad de La Laguna, Tenerife, Spain) and the Information Technology Department of the ITER (Tenerife, Spain). We also thank the anonymous reviewers for their critical advice.

Funding

Research was funded by the Instituto de Salud Carlos III (grants PI14/00844 to C.F., and Sara Borrell CD13/00304 to F.L.D.), and grants MTM2013-43396-P (to M.C.) and INP-2011-0063-PCT-430000-ACT (INNPLANTA program) from the Spanish Ministry of Economy and Competitiveness, co-financed by the European Regional Development Funds 'A way of making Europe' from the European Union, and by the 7th Framework Programme (FP7-REGPOT-2012-CT2012-31637-IMBRAIN). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

References

Altschul,S. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Chevreux,B. *et al*. (1999) Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol*., **99**, 45–56.

Cuccuru,G. *et al*. (2014) Orione, a web-based framework for NGS analysis in microbiology. *Bioinformatics*, **30**, 1928–1929.

Darling,A.C. *et al*. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*., **14**, 1394–1403.

Darzentas,N. (2010) Circoletto: visualizing sequence similarity with Circos. *Bioinformatics*, **26**, 2620–2621.

Doig,K.D. *et al*. (2012) On the origin of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *BMC Genomics*, **13**, 258.

Durfee,T. *et al*. (2008) The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse. *J. Bacteriol*., **190**, 2597–2606.

Edwards,D.J. and Holt,K.E. (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb. Inform Exp*., **3**, 2.

Frey,K.G. *et al*. (2014) Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genomics*, **15**, 96.

Fricke,W.F. and Rasko,D.A. (2014) Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat. Rev. Genetics*, **15**, 49–55.

Guy,L. *et al*. (2010) genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, **26**, 2334–2335.

Iqbal,Z. *et al*. (2012) *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet*., **44**, 226–232.

Iqbal,Z. *et al*. (2013) High-throughput microbial population genomics using the Cortex variation assembler. *Bioinformatics*, **29**, 275–276.

Kurtz,S. *et al*. (2004) Versatile and open software for comparing large genomes. *Genome Biol*., **5**, R12.

Kislyuk,A.O. *et al*. (2010) A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics*, **26**, 1819–1826.

Krzywinski,M. *et al*. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*., **19**, 1639–1645.

Qi,J. *et al*. (2010) inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics*, **26**, 127–129.

Reumerman,R.A. *et al*. (2013) Tool for rapid annotation of microbial SNPs (TRAMS): a simple program for rapid annotation of genomic variation in prokaryotes. *Antonie van Leeuwenhoek*, **104**, 431–434.

Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

Zhou,C. *et al*. (2007) MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res*., **35**, D391–D394.