Genome analysis

IonGAP: integrative bacterial genome analysis for Ion Torrent sequence data

Adrian Baez-Ortega1,†, Fabian Lorenzo-Diaz2,3,†, Mariano Hernandez2, Carlos Ignacio Gonzalez-Vila1, Jose Luis Roda-Garcia4, Marcos Colebrook4,* and Carlos Flores2,3,5,*

1Information Technology Department, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain, 2Applied Genomics Group (G2A), Instituto Universitario de Enfermedades Tropicales y Salud Pública de Canarias (CIBICAN), Universidad de La Laguna, Santa Cruz de Tenerife, Spain, 3Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain, 4Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Santa Cruz de Tenerife, Spain and 5CIBER 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 (Caccuru 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, especially...
configured for rapid handling of data generated by Ion Torrent sequencing. For this reason, we have developed IonGAP, a web-based genome analysis platform that straightforwardly performs 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.

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

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

<sup>a</sup>Darling et al. (2014).
<sup>b</sup>Iqbal et al. (2012).
<sup>c</sup>Reumerman et al. (2013).
<sup>d</sup>Kurtz et al. (2004).
<sup>e</sup>Altschul et al. (1990).
<sup>g</sup>https://github.com/Victorian-Bioinformatics-Consortium/mlst.
<sup>i</sup>Seemann (2014).
<sup>j</sup>Darling et al. (2007).

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 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.
were consistent with those previously reported for the reference genome. For instance, from the total SNP calls retrieved for \textit{E.coli} genomics, comparison, annotation and classification bacterial genomes (Table 1). In this regard, the results generated by analyzing the mentioned datasets were similar to those generated by IonGAP, a N50 length of 4 589 812 bp, with a N50 length of 118 424 bp. In order to compare 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, the variant calls also revealed differences between the two datasets. For \textit{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 \textit{M.ulcerans} dataset (>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 \textit{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 \textit{M.ulcerans} dataset (>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 \textit{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 \textit{M.ulcerans} dataset (>1 kb) in high-copy number (Doig et al., 2012).

### 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