Medicine Papers, 2018, 4 (3): 33–38

Escherichia coli and *Salmonella enterica* genomes sequencing by NGS: a worldwide trial

STEFANO REALE*, FLORIANA BONANNO, IGNAZIO SAMMARCO, FABIOLA VAGLICA, PIETRO CICALA & FABRIZIO VITALE

Istituto Zooprofilattico Sperimentale della Sicilia "A. Mirri", Via G. Marinuzzi 3, 90129 Palermo, Italy
*Corresponding author: stefano.reale@izssicilia.it

Summary

In this study, we report the result of a Inter-laboratory proficiency-test which Innovative Technology Lab of Istituto Zooprofilattico Sperimentale attended in 2017. The Performance test, steered by the US FDA and Center of Genomic Epidemiology of Technical University of Denmark, is provided to facilitate harmonization and standardization in whole genome sequencing and data analysis, with the aim to produce comparable data for the GMI (Global Microbial Identifier) initiative. Our laboratory took part in phase 1a (DNA preparation and sequencing procedures) and phase 1b (laboratory's sequencing output) of the test called "wet lab". Two *Salmonella enterica* strains and two *Escherichia coli* strains, and four pre-prepared DNA extracted from those samples was delivered to the lab. DNA extraction, purification, library-preparation, and whole-genome-sequencing of the eight samples was performed by means of an Illumina MySeq NGS platform, with 150 bp paired end reads. Assembling of all the genomes was carried out with a bioinformatic pipeline using Illumina embedded Casava ver. 1.8 for sequence quality filter and Spades ver 3.11 for reads assembling. The final result of NGS workflow confirmed the identity of all bacterial isolates and applying a bioinformatic pipeline to reassemble the short reads produced in the first phase in a complete genome. Statistics on assembly quality were generated with Quast tool evaluating genome assemblies metrics like N50 and L50.

Key Words

Escherichia coli, Salmonella enterica; genome; DNA fragments; NGS.

Received 26.06.2018; accepted 05.08.2018; printed 30.08.2018

Introduction

Sequencing a genome means identifying the nucleotide order of a nucleic acid, either DNA or RNA. In general, the process consists of three basic steps common to all sequencing methods: sample preparation, physical sequencing and reassembly. In order to analyze the sample, it is necessary to fragment the genome and eventually to amplify the short sequences through molecular method such as PCR. The physical sequencing allows to identify the dNTP order by the synthesis of small sequences, called "reads", and using the DNA fragments as template to create the so-called "sequencing library". The overlapping readings are then reassembled through bioinformatics softwares that execute reads alignment with the reference genome (Global Microbial Identifier, Heather & Chain, 2016, Shendure & Ji, 2008).

The length of the reads depends on the sequencing method. First-generation sequencers, such as Sanger technology, are able to output reads of 800-1000 base pairs, but they present high costs although they are still used in diagnosis and research. The advent of Next-generation Sequencing techniques has allowed not only the reduction of costs and time, but also the possibility of sequencing a greater amount of DNA in parallel, producing millions of reads per single run of the instrument, even if giving rise to shorter readings (about 50-400 nt) (Schadt et al., 2011; Jarvie, 2005).

The "Global Microbial Identifier" (GMI) is currently an informal global, visionary taskforce of scientists

and other stakeholders who share the aim of making novel genomic technologies and informatics tools available for improved global patient diagnostics, surveillance, research and public health response.

The "Global Microbial Identifier" (GMI) initiative aims to build a database of whole microbial genome sequencing data linked to relevant metadata, which can be used to identify microorganisms, their communities and the diseases they cause. It would be a platform for storing "Whole-Genome Sequencing" (WGS) data of microorganisms, for the identification of relevant genes and for the comparison of genomes to detect outbreaks and emerging pathogens.

The main objective is to promote internationally the production of reliable and good quality results in microbial "Whole-Genome Sequencing" (WGS). In particular, with the Proficiency Tests, GMI aims to assess the suitability of members in DNA extraction, preparation of genomic libraries and sequencing, using their usual protocols, software and platforms (Global Microbial Identifier, Proficiency Test 2017). In this way it will be possible to obtain the standardization of the WGS procedures, including the last analysis, in order to have comparable data for the realization of the GMI initiative. To date, the PT project is still active and in 2016, 46 laboratories circa in 22 different countries, including Italy, partecipated. Among the participants there was also the Molecular Biology department of the "Istituto Zooprofilattico Sperimentale della Sicilia", based in Palermo and directed by Dr. S. Reale. This Institute also joined the PT set up in 2017, that was focused on the sequencing of *Salmonella enterica* and *Escherichia coli* genomes.

Historical overview

Maxam & Gilbert (1977) published the first sequencing of a DNA fragment long about 24bp. This method consists in the chemical denaturation by DMSO (dimethylsulfoxide) of the nucleic acid, that is radioactively labeled at one end. The sequence order is determined thanks to the arrangement of the fragments along the bands on an electrophoretic run.

This chemical method (Maxam & Gilbert, 1977) was soon replaced by an enzymatic one, developed by Sanger et al. (1977). The Sanger method uses 2', 3'-didehoxy and arabinonucleosides, labeled and analogues of the normal deoxynucleoside triphosphates, which act as specific DNA-polymerase inhibitors. This method is called "*chain termination sequencing*" because the ddNTPs prevent the formation of phosphodiesteric bonds with the nucleotides in the mix (Sanger et al., 1977).

The Sanger method revolutionized the molecular biology field, and it was a means through which one of the most important projects of the last decades has been realized, the "Human Genome Project". The project was carried out by two different consortia: in fact, Greg Verter, detached from the National Institute of Health, has found a private research body, the Celera Genomics, whose contribution was fundamental for the application of a new sequencing approach, the "shotgun" sequencing. This approach also characterizes The Next-generation Sequencing technologies, introduced in 2005 with the purpose of speeding up the sequencing process, lowering costs and increasing their use in the research areas. The NGS methods make it possible to sequence a bacterial genome in just a few days and rapidly compare genetic sequences among multiple genome. With its ultra-high throughput, NGS enables researchers to perform a wide variety of applications and study biological systems at a level never before possible. The term "high-throughput" refers to the possibility to carry out a large number of measurements simultaneously, and to analyze multiple samples in a single session.

Many technologies are available, and in most, the DNA sample is broken into a library of small fragments and then attached to oligonucleotide adapters. The construct is placed on a slide or in a flow-cell, and the strings of nucleotide bases that make up the fragments are then sequenced in hundreds of millions of parallel reactions.

NGS platforms

Over the years, many companies have invested in the production of advanced equipment contributing to the rapid evolution of the sequencing methods. There are several sequencers that have been placed on the market and each platform uses different biochemical processes. Making a brief overview of the main technologies, we have to enunciate the first sequencer on the market (since 2005), the Roche/454. The 454 technology is based on two fundamental processes: the emulsion PCR amplification and the pyrosequencing. In the emulsion PCR method, the amplification takes place into water-oil emulsion droplets that contain particles, on whose surface are attached oligonucleotide probes that are complementary to the adaptors of the single-strand DNA templates. The pyrosequencing method is instead based on the use of enzymes which, during the synthesis of the new filaments, allow to associate the addition of each complementary nucleoside with a light signal.

Another important sequencer is SOLiD, produced by Applied Biosystems. It differs from the other sequencers because of its different way of sequencing, that is not by synthesis but through ligation. The sequencing takes place by the action of a DNA ligase able to create covalent bonds between marked DNA fragments that will give rise to a single filament complementary to the one to be sequenced. As the pairing occurs, the markers are excited so that the light emission - associated with the first two base pairs - is detected by the instrumentation to reconstruct the nucleotide sequence during the cycles.

Lastly, the Illumina Genome Analyzer, introduced in 2006: its technology is based on SBS (sequencing-by-synthesis) process. The DNA cluster is formed through a so-call “bridge PCR”, with the DNA immobilization by hybridization on the plate.

NGS Applications

NGS techniques immediately represented the possibility of implementing the development of different research areas, such as personalized medicines - thanks to the possibility to make a gene expression profiling, the chromosomal count or the identification of an individual's epigenetic mutations. Due to their high sensitivity and specificity, these systems also allow the monitoring over time of genetic variations in viruses and pathogenic bacteria. The NGS technology can gain critical genetic insight into bacteria and viruses with microbial genome sequencing, whether you are performing metagenomics studies, or monitoring disease outbreaks

Whole-genome sequencing

Whole-genome sequencing (WGS) is a comprehensive method for analyzing entire genomes. Genomic information has been instrumental in identifying inherited disorders, characterizing the mutations that drive cancer progression, and tracking disease outbreaks. Rapidly dropping sequencing costs and the ability to produce large volumes of data with today’s sequencers make whole-genome sequencing a powerful tool for genomics research.

Despite this method is commonly associated with sequencing human genomes, the scalable, flexible nature of next-generation sequencing (NGS) technology makes it equally useful for sequencing any species, such as agriculturally important livestock, plants, or disease-related microbes. Small genome sequencing (≤ 5 Mb) involves sequencing the entire genome of a bacterium, virus, or other microbe, and then comparing the sequence to a known reference. Sequencing small microbial genomes can be useful for food testing in public health, infectious disease surveillance, molecular epidemiology studies, and environmental metagenomics.

Material and Methods

DNA extraction

The lyophilized material shipped from DTU Denmark was inoculate into liquid soil with subsequent distribution in plate with added solid soil. From each plate, the colonies were collected and the DNA was extracted by boiling.

In the end, the genomic material was purified by extracting on silica gel columns with the EZNA kit, after adding lysis buffer, proteinase K and ethanol.

The DNA quality was evaluated using the Nanodrop ND-J000 spectrophotometer (Nanodrop Technology) by reading the absorbance at 260nm, in order to test the DNA purity and the absence of solvents that could alter following steps.

DNA extracted was quantified through Qubit® fluorimetric assay.

NGS Library preparation and sequencing

Illumina NGS workflows include four basic steps:

1. Library Preparation: the sequencing library is prepared by random fragmentation of the DNA or cDNA samples, followed by 5′ and 3′ adapter ligation. An alternative is the “tagmentation” process, which uses transposons to break DNA and to bind adapters in a single pass, in order to speed up the process. The fragments at this point are amplified by PCR and purified.

2. Cluster Generation: the library is loaded into the plate (flow-cell) where the fragments can be anchored because captured by oligonucleotides complementary to the library adapters. At this point, thanks to the "bridge" amplification (bridge PCR), each fragment is cloned to form numerous copies of DNA template.

3. Sequencing: SBS technologies use fluorochrome-labeled reversible terminators. At each cycle, a DNA polymerase incorporates a labeled base to the clusters, the laser light causes the emission of their fluorescence and the first image of the plate is captured. The software records the successive images, recognizes the bright spots associated with each individual fragment and the colors that indicate the base: in this way it will process the reads.

4. Data Analysis: the output reads are aligned (alignment) and the identified sequences are compared with the reference DNA.

Genomic libraries were obtained using Nextera XT DNA Library Prep kit.

Protocol for library preparation

The protocol used for the preparation of genomic libraries was the Nextera XT DNA Library Prep. It consists of:

1. Genomic DNA Tagmentation: after normalization of all the samples bringing them to the concentration of 0.2 ng / µl, Tagment DNA buffer and Amplicon tagment mix were added on each well of the PCR plate, which is centrifuged and placed in the thermal cycler. The frammentation is then stopped through a neutralized tagment buffer;
2. DNA framments amplification: the primers used in this phase are the Nextera XT Index 1-2 Primers (Illumina), whose different combination allowed to index the different samples;
3. DNA purification: carried out with AMPure XP beads;
4. Library normalization and pooling;
5. Library Dilution and denaturing.

To execute a run on MiSeq, a reagent kit that includes the flow-cell, a bottle of PR2 and the reagent cartridge must be used. The MiSeq flow cell is a single-use glass substrate on which clusters are generated and the sequencing reaction is made. The libraries are loaded into the reagent cartridge before starting the run and then automatically transferred to the flow-cell. The MiSeq reagent cartridge is a disposable material consisting of small sealed tanks pre-filled with sufficient reagents for cluster generation and sequencing a flow-cell.

At this point, it is possible to start a sequencing run which can be monitored from the Sequencing screen or using “Sequencing Analysis Viewer” (SAV). Run duration depends on the number of cycles performed. During cluster generation, individual DNA molecules bound to the surface of the flow-cell are amplified through bridge PCR to generate the cluster.

After images analysis, the software performs the identification of the bases, the filtering and the calculation of the qualitative scores.

Bioinformatic elaboration

The main objective of the Proficiency test was to quantify differences among laboratories in order to facilitate the development of reliable laboratory results of consistently good quality within the area of DNA preparation, sequencing, and analysis (for example phylogeny).

Result data from wet lab operation were elaborated using a simple bioinformatic pipeline whose purpose was to verify the final quality of resulting reads from Illumina MySeq NGS workflow to assemble a whole genome from the reads.

Another item was a variant detection analysis and phylogenetic/clustering analysis of assembled genomes, but because the Proficiency Test consisted of three parts, each of which were optional, and because we had not enough time nor expertise at the moment of the execution of the test, we chose not to execute the variant detection and phylogenetic/clustering analysis. On the contrary, we just generated FASTQ file from sequencing and assembled it in two complete genomes of *S. enterica* and *E. coli*.

FastQ file were generated from base call file directly with Illumina Base space cloud platform, including demultiplexing, quality filtering, trimming of low quality ends of reads.

Assembling was performed with software Spades ver 3.11 installed on Ubuntu server 10.2 virtual machine with 128 GB of RAM and 2.5 Tb HD. Assembly quality was assessed with Quast ver. 4.6.1 software.

Results

Final result of GMI Proficiency test was the submission of raw reads FASTQ file and two FASTA format files containing the assembled genomes of *S. enterica* and *E. coli* samples. Files have been uploaded by through the GMI Proficiency Test web platform and by HTTP client Filezilla, when web interface stopped the upload.

Figure 1 and figure 2 show reads quality score graph for sample 1 of *S. enterica* bacterial culture and pre-extracted DNA.

The result statistics of the assembled genome for *S. enterica* - sample 1, elaborated with Quality Assessment Tool for Genome Assemblies software ver. 4.6.3 are reported in Table 1. Estimated reference genome size parameter was 5.000.000 bp.

The statistics show an overall good quality of the assembled genome for all samples, only data for sample 1 (bacterial culture and pre extracted DNA) are shown.

NGS technology enables authors to make further elaboration with raw reads data. We are planning to perform the identification of variant sites (e.g. SNP) within whole genome sequence and to distinguish different clusters of samples based on those variant polymorphisms. This will be important for MLST typing of bacterial strain, Multidrug Resistance Identification and for phylogenetic studies of different population of micro-organisms.

Figure 1. Quality score graph for sample 1 of *Salmonella enterica* bacterial culture.
Figure 2. Quality score graph for sample 1 of *Salmonella enterica* pre-extracted DNA.

Statistics	Salmonella enterica Sample 1
# contigs	66
Largest contig	719,795
Total length	4,924,049
Total length (>= 1000 bp)	4,916,901
N50	173,791
# contigs (>= 1000 bp)	56
GC (%)	52.08
Estimated reference length	5,000,000

Table 1. Statistics for quality assessment of assembly for Sample 1 - *Salmonella enterica*. All statistics are based on contigs of size >= 500 bp.

References

Global Microbial Identifier. [Online] http://www.globalmicrobialidentifier.org/-/media/Sites/gmi/Work-groups/GMI_PT_report_2016_AllFigures_AllAppendices_ISBN.ashx?la=da&hash=482FCDF87390CFFD04847722298843D960557AB8

Global Microbial Identifier, Proficiency Test, 2017. [Online]: http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-2017

Heather J.M. & Chain B., 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107: 1–8. DOI: 10.1016/j.ygeno.2015.11.003

Jarvie T., 2005. Next generation sequencing technologies. *Drug Discovery Today Technologies*, 2: 255–260.

Maxam A.M. & Gilbert W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74: 560–564.

Sanger F., Nicklen S. & Coulson A.R., 1977. DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74: 5463–5467.

Schadt E.E., Turner S. & Kasarskis A., 2011. A window into third-generation sequencing. *Human Molecular Genetics*, 19: R227-240. DOI: 10. 1093/hmg/ddq416

Shendure J. & Ji H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26: 1135–1145. DOI: 10.1038/nbt1486.