MODESTUM
OPEN ACCESS

Review Article

Electron J Gen Med 2019;16(3):em136
ISSN:2516-3507
https://doi.org/10.29333/ejgm/108690

Relating next-generation sequencing and bioinformatics concepts to routine microbiological testing

Otávio Guilherme Gonçalves De Almeida[1], Elaine Cristina Pereira De Martinis[1]

ABSTRACT

Next-Generation Sequencing (NGS) is becoming a reality in the clinical microbiology laboratory because it can speed diagnosis when compared to traditional culture based-methods and moreover, to aid in unravelling key virulence traits of important pathogens. Nonetheless, there are many limitations for its wide application in routine testing, as the requirement of high performance hardware and software to support bioinformatics analysis, as well as the expertise in different programming languages to perform the analyses. In this context, this review was drawn to synthesize some basic concepts involved in NGS for Whole-Genome Sequencing (WGS), based on two international straightforward efforts to standardize WGS data acquisition and processing in the clinical routine, the PulseNet International and the ENGAGE project, allied with other tools available for WGS analysis, beginning from the available sequencing platforms to the main user-friendly pipelines dedicated for the pathogen identification, including the use of properly databases to search for virulence factors, resistance genes and software resources for molecular typing of isolates.

Keywords: high-throughput DNA sequencing, microbiology, pathogen, whole-genome sequencing

INTRODUCTION

The diagnosis of bacterial pathogens of clinical importance is still very dependent on phenotypic techniques, which include plate culture for isolation of colonies, differential staining, morphological analysis and biochemical tests (1). A second important step is the typing of clinical isolates, which can be based on serological tests to check for the presence/absence of selected antigens, as toxins or serotypes, and may be necessary to perform the genetic screening to search for antibiotic resistance genes for instance (2).

Nevertheless, culture-based diagnosis is laborious and time-consuming since it depends on multiple steps for isolation and identification of the target organism. There is a great deal of interest for alternative methods to replace or to complement the classic microbiological diagnostic tools, including tests based on polymerase chain reaction (PCR) and other technologies for detection of selected genes (3).

In this context, the variables and hypervariables regions of the gene coding for the 16S rRNA fraction are excellent phylogenetic markers for the identification of a bacterial isolate given the universal distribution of this gene in all prokaryotic organisms (4). The 16S rRNA gene is composed of approximately 1500 bp coding for the catalytic subunit of 30S rRNA (4,5). The structure of this taxonomic gene marker is characterized by nine variable and hypervariable regions flanked by conserved regions that evolve at different proportions making possible identify basal taxonomic levels (i.e. domains) and more derivate levels (i.e. species) as well (6). Due to the occurrence of differential evolution rates among these variable and hypervariable regions, this gene can be used as a "biological clock", for measurement of phylogenetic distances between groups and to generate hierarchical trees, evidencing the close relations among the isolates in the study (4).

Nevertheless, this marker presents a cumbersome limitation on the species level identification, since in 65% to 83% of the cases the taxonomical identification is achieved, which implies that around 1% to 14% of the isolates are kept unidentified after Sanger DNA sequencing (7). Even all variable and hypervariable regions alone are not capable of

[1] *Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Brazil.*

Correspondence: *Elaine Cristina Pereira De Martinis*
Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Brazil

E-mail: edemarti@usp.br

Received: 9 Aug 2018, Accepted: 22 Apr 2019

differentiate among all living bacteria, the V2, V3 and V6 regions are the three hypervariable regions characterized as the most heterogeneous, contributing for maximum differentiation of many similar bacterial groups (6). The major problem related to this limitation is the lower sufficient sequence evolution rate among bacterial species closely related that may not possess a significant evolutionary divergence in its 16S rRNA sequences to be properly distinguished (8).

Thus, for molecular taxonomic identification of a pure bacterial isolate derived from a pure culture, a DNA extraction method is employed (i.e. mechanical lysis, enzymatic or phenol-chloroform extraction) followed by 16S rRNA gene PCR amplification. The PCR products are analysed on agarose gel for purity and taken for DNA Sanger sequencing. The sequences can be analysed and compared to databases for taxonomic identification purposes (9).

However, the culture-based approach followed by the amplification of one specific piece of 16S rRNA (that may capture one or more variable and hypervariable regions for Sanger sequencing) may occur an underestimation of the microbial diversity in highly contaminated samples, such as faecal specimens, due to low resolution at species level (8). In addition, this method is useful if a particular feature of the pathogen is of interest, with amplification by PCR and sequencing of a target region of the genome (10). Because of these limitations and with the development and democratization of Next-Generation Sequencing (NGS) Technologies, some clinical laboratories are entering in the age of Whole-Genome Sequencing (WGS) (11). This technology allows the sequencing of multiple samples in a massive parallel way, resulting in a production of thousands of millions of DNA sequences at one single run, which after a bioinformatics downstream processing and analysis, may unravel the entire complete genome from various clinical isolates (11-13) without the need of prior knowledge of target conserved phylogenetic regions. At same time, it possesses high speed and throughput, while Sanger sequencing may took several years to close a single draft bacterial genome (14).

In the context of assessing serial bacterial genomes, the NGS has been a widely application to detect and to identify pathogens not only in clinical isolates, but also in food matrices as performed by Macori et al. (15). These researches revealed the complete genomes of four *Yersinia enterocolitica* bacterial pathogens isolated at wild ungulate carcasses using the WGS approach in combination with phenotypic biochemical tests to determinate their virulence profile (high virulence or nonvirulent). Another very interestingly work dedicated to the elucidation of whole-genome of a bacterial pathogen in cattle was published by Stevens, Stephan and Johler (16). According to the findings of these authors, the entire genome of *Staphylococcus aureus* strain 1608 mastitis causative agent was useful to establish further research novel features that may contribute to identification of factors related to toxic mastitis development in cattle (16).

The analysis of bacterial genomes in animals related to human consumption is crucial to an effective public health surveillance system, since animal pathogens can be transmitted to humans that maintain close relations with the contaminated animals, as occurs with mastitis causative agent (*S. aureus*), in such a way that the methicillin-resistant *S. aureus* strains (MRSA) (17) may infect the tits of cattle and, furthermore, can be transmitted to humans by feeding, what is a very important to the public health authorities (18).

Currently, the decreasing costs for WGS encourages its use in routine applications as well (11-13) but there are still many challenges related mainly to the need of technical expertise for choosing among the various NGS platforms available and for the analysis of the sequence data obtained (19). Therefore, in this review we will present relevant considerations on NGS, with emphasis on its use for identification of bacteria of clinical importance.

WHOLE-GENOME SEQUENCING

The authors of the present article define the WGS as an access of an individual DNA content present in a given organism through a chosen NGS technology that will generate a large amount of data which should be processed to unravel the entire genomic organisation of a targeted organism to achieve the taxonomy of the microbial isolate, as well as its knew and hidden functional properties (possibly).

For WGS it is necessary to obtain a pure culture and to perform DNA extraction. The isolate of interest is cultured on agar plate and next, genomic DNA must be extracted, similar to other genomic-based methods. For DNA extraction, either in-house protocols (classical DNA extraction) or commercial kits can be used. Quality and quantitative analysis of DNA extract have to be performed either by colorimetric or fluorimetric methods, prior to library preparation for NGS (20). Some steps of library preparation vary with the choice of the platform for sequencing but, regardless the platform, there is a random fragmentation of the genomic DNA, which can be accomplished by physical (i.e., sonication), enzymatic (i.e., non-specific endonucleases and tagmentation reactions by means of transposases) or chemical methods (20,21).

The next step involves the insertion of adapters to the DNA fragments, generating inserts. Adapters are oligonucleotide sequences complementary to sequences immobilized on commercial sequencing platforms and they

are needed for attachment of the inserts to a support where the DNA amplification and sequencing will be performed (22). Moreover, depending on the platform of choice, the adapters can be combined with indexes to speed up the library preparation process: by indexing the inserts, it is possible to process mixed samples and to keep track of each amplified fragment with regard to the original sample (23).

A clean-up process to separate only inserts with suitable size must be performed and, possible dimers of adapters must be removed to avoid consuming useful spots in the platform support to generate useless sequencing data. For this step, methodologies employing magnetic beads technology largely used (24).

Different companies offer commercially diverse sequencing platforms such as 454 (Roche), SOLiD (ThermoFisher), Illumina® (Illumina sequencers), PacBio (Pacific Biosciences) and MinION system (Oxford Nanopore Technologies) (25).

The platforms of Roche, SOLID and Illumina® companies are classified as second-generation technologies since they were launched after Sanger's sequencing technology and are characterized by different sequencing chemistries. In addition, second-generation sequencers are thus classified by their built-in nucleotide detection system, which is performed by optical sensors for light detection (26).

The basis of the Roche-454 sequencer is the detection of the pyrophosphate (PPi) released during the incorporation of the nitrogen base by DNA polymerase in the synthesis of the nascent DNA chain. The resulting PPi is converted to ATP by the ATP sulfurylase enzyme, which provides energy for the luciferase enzyme to oxidize luciferin, resulting in light emission. Visible light detected is proportional to the number of nucleotides incorporated and each DNA base has their own light colour, which allows the base calling identification (27). Nevertheless, Roche-454 technology is currently out of date (20).

The synthesis technology of the company Illumina® is the most used in the market due to the wide variety of sequencers available, including MiSeq, HiSeq, NextSeq and now NovaSeq. The generation of short-reads and the optical detection system for fluorophores-labeled dideoxinucleotides (dNTPs) are hallmarks of these platforms (20,28).

The Ion Torrent (Thermo Fisher) sequencer has a dilemma in its classification, as some authors classify it as a generation 2.5th equipment, due to its chemistry of sequencing based on the measurement of pH change caused by the release of a proton during the synthesis of the chain of DNA, by means of a chemical sensor (29).

According to Heather & Chain (30), third-generation sequencers are those capable of sequencing a single whole DNA molecule without the need for amplification of the fragments. The third generation sequencers are represented by PacBio (Pacific Biosciences) and minION (Oxford Nanopore) platforms (20,30).

In the PacBio platform, sequencing is performed on a chip that contains several zero-mode waveguide detectors (ZMW). Each detector has an added DNA polymerase and, when the enzyme adds the phosphor-linked dye-labeled nucleotides, the incorporation is detected in real-time by an imaging system (30).

On the other hand, MinION (Oxford Nanopore) technology is based on the measurement of changes in electrical conductivity caused by the passage of single strand of DNA through a biological pore. The DNA passes through the pore after the application of a voltage, which generates a current of ions. The change in the pattern or magnitude of the electric current in the nanopore is captured by a sensor "several thousand times per second" and the electrical current measures are passed to an application-specific integrated circuit (ASIC) microchip. The data are processed and analyzed in the MinKNOW software available on the MinION platform (31).

By having the ability to sequence single molecules without the amplification of DNA fragments, third-generation sequencing technology is characterized by the production of longer reads, which can include medium-sized reads of 10,000 bp and some reads larger than 100,000 bp, which confers advantages related to the resolution of gaps in the genome and allows the construction of large contiguous regions of DNA, facilitating the genome assembly and annotation (32).

In **Table 1** is presented a brief comparison among the sequencing commercial platforms available focusing on its sequencing chemistries and the main pros and cons.

Table 1: *Main features of NGS sequencing platforms and its pros and cons*

Sequencing Platform	Generation	Sequencing Chemistry	Read length (bp)	Supports PE reads?	Pros	Cons
Roche 454 GS Junior	1st	Pyrosequencing	500	Yes	Long-reads technology allows an acurated mapping to reference genomes and makes easier *De novo* genome assembly	High reagent costs Low-throughput High error rates in homopolymeric regions
NextSeq 550(Illumina®)	2nd	Flourophores detection by an optical sensor	150		High-throughput per running Lowest base-sequencing costs	Average error rate of 10^{-2} to 10^{-3} due to *in vitro* serial amplification steps
MiSeq (Illumina ®)			25-300[1]			
HiSeq 2500 (Illumina®)			50-250[2]			
NovaSeq 6000 (Illumina®)			50-150[3]			
Ion Torrent PGM (Thermo Fisher)	2nd/3rd	Proton detection by semiconductor technology	200	Yes	Fast running times (few hours)	High error rates in homopolymeric regions
PacBio (Pacific Biosciences)	3rd	Single Molecule Real Time technology	1500	No	Long-reads technology facilitates the conclusion of a genome assembly and elucidation of a draft genome	Lowest throughput among all the platforms High initial investment GC Bias
MinION (Oxford Nanopore)	3rd	Measurement of an ion current flow through a nanopore and the characterization of pattern or magnitude of the current generated	5.000	No	Portable sequencer (USB-like) makes easy the sampling in a field and the fast taxonomic identification. It is ideal for mobile labs No PCR biases (no need for DNA amplification) Lower initial investiment	Low numbers of generated reads High DNA input requirements (1 µg) GC Bias

Legend: [1],[2]Varies according to the run mode chosen and running kits employed. These values ranging are regarding to PE reads; [3]Depends on the flowcell type. Source from: van Dijik et al. (22); Goodwin et al. (28). Kircher & Kelso (33); Quail et al (34); Mikheyev & Tin (35) and Illumina® (36)

BIOINFORMATICS: DOWNSTREAM ANALYSIS

One advantage of the WGS over the classical bacterial identification approaches is that one single protocol serves for all kinds of bacteria (10), although specific knowledge is required to analyse the sequences generated. Computer programming skills are needed to comprehend scripts written in different programming languages in order to reconstruct and annotate the genomes. Therefore, bioinformatics is an integral part of the analysis process.

According to NIH Biomedical Information Science and Technology Initiative Consortium (BISTIC), bioinformatics can be defined as: "Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyse, or visualize such data" (37). Therefore, bioinformatics is an important computational tool to organize, visualize and integrate data from sequencing. In addition, knowledge about the available public databases is fundamental to succeed with taxonomic and functional annotations.

An interestingly report from the European Food Safety Authority (EFSA) showed that in the 2016 year all Europe reference laboratories (100%) employed WGS for pathogen identification, but there were discrepancies when compared the WGS employment among other specialized institutions as National Reference Laboratories (44%) and Official laboratories (7%). The main reasons for this reduced usage by non-reference laboratories can be summarized around two variables: the limitation in the budgets of the facilities and the absence of trained staff to conduct analysis of data using bioinformatics tools (38).

In the context of standardize the WGS application in clinical laboratories, two major consortia must be highlighted, the PulseNet International and the ENGAGE project. The PulseNet International is a network of 86 countries distributed in regions from Africa, Asia, Canada, Europe, Latin America and Caribbean, the Middle East and the United States. The concept of this consortium is to exchange information regarding genomic data among the countries laboratories and to implement standard operational procedures (SOPs) for genotyping (39).

The ENGAGE (Establishing Next Generation sequencing Ability for Genomic Analysis in Europe – www.engage-europe.eu) project, composed by eight European institutions, published SOPs for genomic data collection from a pure culture isolate, WGS DNA sequencing and data analysis suggesting dedicated tools for quality control, genome assembly and annotation viewing to standardize pathogen identification in different laboratories across Europe. The project also makes available on-line learning material to update WGS and bioinformatics knowledge (40).

So, due to the highly quality and the hot topics pointed out by both projects, this paper will discuss briefly some tools proposed in these projects guidelines, specifically those related to pathogen molecular typing, and the basic bioinformatics concepts dispersed in the current literature starting by the first steps required to genomic data analysis.

The basic *in silico* steps for processing NGS data prior to taxonomic and functional analyses include: (i) the quality control of the sequences, (ii) the removal of the adapters inserted in the stage of preparation of the library, (iii) the assembly of the genomes guided by reference or the *De novo* assembly (20).

Regarding the quality control of sequences, the base calling quality measure reflects the probability of a given base being randomly or incorrectly mismatched, and is commonly calculated on the basis of the Phred score (also entitled Q score) (20-21, 41). The Q score value is a logarithmic representation of the error probability and its calculation varies according to the sequencing platform used. The data generated in the sequencing are commonly addressed in FASTQ files, an ASCII-format file format, in which the DNA sequences present the ASCII coded quality score (Q score) for each identified nitrogen base (20). The use of ASCII coding is not only justified by adjusting the logarithmic-scale probabilities to their nearest integer value, but also by saving space in the file and optimizing computational resources (42). The quality control of the base calls is fundamental to avoid misinterpretations during the assembly of the genomes (20).

The Centers for Disease Control and Prevention (USA), one of the countries that integrates the PulseNet International (38), establishes standard protocols for WGS-based diagnosis by addressing the stage of preparation of libraries for the MiSeq platform of the company Illumina® and standard operating procedures for the quality control of generated sequences in sequencing using User-friendly tools like FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (43). The same tool is recommended for QC of reads by the ENGAGE project (40).

However, if the analyst possesses bioinformatics background it is possible to opt by other tools for QC of sequences. These tools are not user-friendly and require programming expertise. One of the most used pipelines to processing reads is the NGS QC Toolkit (available in http://www.nipgr.res.in/ngsqctoolkit.html). This pipeline is divided in two major workflows: one dedicated to guarantee the reads quality control for illumina® sequencing platforms (IlluQC) and other specific for Roche-454 sequencing platform (454QC). This tool includes trimming of adapters, quality control of sequences, trimming of homopolymeric regions, statistics for average quality of each read by FASTA file and FASTQ to FASTA format conversion (44). Nevertheless, there is no mentation to the use of this software in any SOP from PulseNet International or the ENGAGE project.

The ENGAGE project, recommends a specifically developed tool for illumina® sequencing platforms, the Trimmomatic software (available in http://www.usadellab.org/cms/?page=trimmomatic) may be employed to QC of paired-end and single-end sequence reads that are filtered based on their Phred score values (+30 or +64 depending on Illumina® pipeline chosen) and the pipeline also supports adapter trimming. The software was written in java language and uses FASTQ files as an input (45).

Alternatively, if the analyst opts for other tool to adapter trimming it is possible to employ the Cutadapt (available in https://cutadapt.readthedocs.io/en/stable/), a standalone software projected for adapter removal that supports FASTQ, FASTA and SOLID .csfasta/.qual as inputs. The rational of this tool is looking for multiple adapters in a one single run of the program based on previous alignment among the reads and adapters sequences. The algorithm removes the sequences with the best matching and, optionally, may search for an adapter and remove it multiple times, what is useful when a problem related to library preparation results in an insertion of the same adapter multiple times. Although this software is designed for command line interface, it offers an easy-to-use command line interface (46). However, this famous algorithm is not in the list of recommended tools for adapter trimming in clinical diagnosis from any SOPs of the PulseNet International and the ENGAGE project.

Several softwares available from the commercial platforms are available for sequence preprocessing and downstream analysis, such as BaseSpace (Illumina®) (47), NextGENe™ (Ion Torrent) (48) and SMRT Analysis Software (PacBio) for instance (49), but they are limited to a default established by the companies, which may limit the data analysis if the analyst is searching for specific features in a given pathogen.

The QC of sequences is an important step for all WGS downstream analysis because low quality base calls dispersed in the dataset could add useless and misinterpreted information due to the presence of random reads. It becomes clearer during the genome assembly step, in which these low-quality reads may generate false k-mers (substrings of sequences with a determined size), that may increase the complexity of genome assembly phase (50).

According to Miller, Koren and Sutton (51), a genome assembly is a hierarchical data structure that maps the data sequence into a putative reconstruction of the target. In other words, it is the grouping of reads into contigs (contiguous DNA sequences) and contigs into scaffolds. The contigs provide for multiple sequence alignment and also the generation of a consensus sequence, whereas the scaffolds define the order of the contigs and their orientation (51).

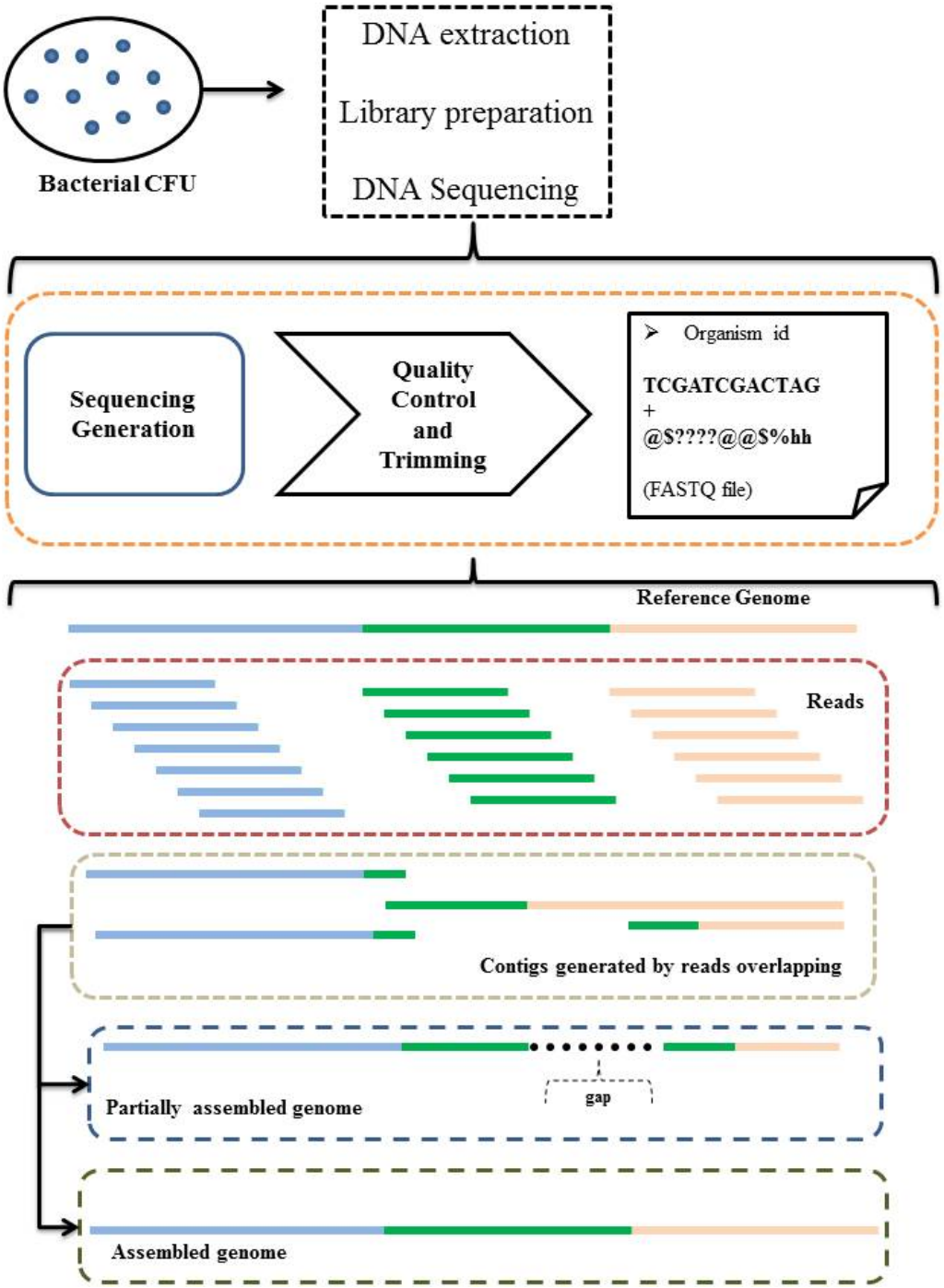

Figure 1: *General overview regarding NGS approach for Whole-Genome Sequencing. After a bacterial colony selection, follows the genomic DNA extraction and the library preparation steps to prepare the sample for NGS sequencing. The analysis of the sequenced genome is performed with bioinformatics pipelines that must ensure the quality control of reads and adapters trimming. If the researcher or analyst is using short-reads sequencing technologies, such as those based on the second generation technologies, the total genome assembly may be hard, but possible. By the other hand, if the analyst chooses to employ a third-generation sequencer, like SMRT technology, the genome assembly step can be easier due to long-reads generation. At the end, the analyst should compare the genomes against a reference to identify and to characterize the isolate (not shown in the figure).*
Source: Elaborated by the authors

The assembly of genomes can be based on two approaches: guided by reference or *De novo* assembly. The assembly guided by a reference depends on the construction of overlapping contigs and the alignment of these sequences against sequences belonging to reference genomes available in the databases, which corroborates the assembly of the scaffolds and the resolution of the draft genome (**Figure 1** shows a brief overview on reference-guided genome assembly for

WGS). On the other hand, *De novo* assembly refers to the reconstruction of contiguous DNA sequences without using any reference sequences (52), i.e. sequences of sequenced microorganisms deposited in databases are not available to guide this step (for further study on the art of genome assembly, see references: 41, 42, 51 and 52).

One of the adequate tools recommended by the ENGAGE project for genome assembly procedure is SPAdes (http://bioinf.spbau.ru/spades), an open source software that uses the mathematical concept of de Bruijn graph variations (multisized de Bruijn graph and paired assembly graph) for a genome assembly which makes possible identify insert size variation, chimeric reads (reads formed by the union of distant substrings of the genome) and sequencing errors (53). This assembler is recommended because it achieved, during a standardized benchmarking exercise among the EU laboratories, the longer contigs generation and high accuracy to predict correctly (100%) Multi-locus Sequencing Type (MLST) of *Salmonella* genomes in comparison with the Velvet assembler (40).

Nevertheless, the bacterial genomes structure is not only marked by the presence of a single double-stranded DNA (dsDNA) molecule arranged in a nucleoid configuration at the central part of the bacterial cell (54), but many of them may also possess an extra genetic material characterized by a circular dsDNA that confers fitness advantages in a specific environment. The acquisition of antibiotic resistance genes (ARGs) is an example of the plasticity of bacterial cells in a stressful condition (55).

Many ARGs are carried by plasmids. Notwithstanding, some of them may be intrinsic to the bacterial chromosome structure. So, search for ARGs using WGS from bacterial isolates offer an interestingly approach because it makes possible to trace the localization of the resistance gene if in plasmid or in chromosome and allows to analyse the genetic context of the acquired gene, via bacterial transformation or conjugation for example (56). Although with the WGS technology it is now a reality to sequence entire plasmids from bacterial cultures, it is still a hardworking task, since the data generated by short-reads sequencing platforms may limits the analysis of plasmids DNA sequences in a sample due to the presence of repetitive sequences that may be shared by plasmid or the chromosome (56) and the high presence of several elements as transposons that could be a result of multiple insertion processes. One advantage of search for plasmid sequences in a WGS data is to detect plasmids with new functions not annotated in the databases (55). Thus, the development of softwares with the ability to unravel the plasmids sequences dispersed in whole bacterial genomes is extremely necessary.

To fill up this gap and allow capture the whole plasmid sequences in a given genome, the tool plasmidSPAdes was developed by Antipov et al. (56). This software uses the genome assembly method performed by SPAdes tool and computes the median coverage of the assembly graph that will be useful to identify short dead-end and long chromosomal edges in the assembly graph for cut them off during the analysis. So, after computing the median is possible to find a subgraph in an assembly graph that is referred as plasmid graph. Further, an exSPAnder function is recruited to search for repetitive regions in the plasmid graph, which results in the generation of plasmidic contigs. With this rational, is feasible to recover plasmid sequences for WGS data (57).

PIPELINES FOR PATHOGENS IDENTIFICATION

Although the aim of this manuscript is to address the application of WGS for molecular identification of bacterial isolates, we must also point out that through the NGS technology it is possible to perform analyses of clinical samples by independent culture methods such as metagenomics and metataxonomics, where the first is based on the sequencing of the genomes of all the microorganisms present in the sample, returning a set of annotated theoretical genomes, while the second one refers to the sequencing of specific regions of the 16S rRNA gene present in the microbial genomes of a given sample (58).

Such analyses are culture-independent and allow the study of complex samples, with high microbial diversity. However, metagenomics requires extensively high computational resources (processing, memory and storage), and the user's need to have programming skills. In this paper we also discuss on the use of WGS for lab analysts and clinicians with a focus on metagenomics.

For the processing of NGS data files suitable pipelines should be employed. A pipeline is a series of transformations by which an input file is processed (59). In other words, a pipeline is a set of scripts, which can be written in various programming languages and its function is to contain tasks that modify the input file. There are pipelines dedicated to pathogen detection and identification in WGS data that will be summarized below.

In this context, SUPRI (sequence-based ultrarapid pathogen identification- http://chiulab.ucsf.edu/surpi/) is a complete pipeline composed of a series of scritps encoded in the languages, shell, Python and Perl coded on the Linux operating system, including also open-source tools such as SNAP and RAPSearch aligners. The pipeline recognizes

FASTQ files as inputs and removes (trimming) low quality reads and adapter sequences. Subsequently, the user can choose to run the software in fast or comprehensive modes, which are characterized by extensive classification of reads in virus and bacterial databases in the prior parameter and by comparison against the entire NCBI nucleotide sequence database in the second one. These analyses are carried out by means of external software fixed to the pipeline, such as those related to the trimming of adapters and low-quality reads, for example. The output of the pipeline is characterized by a list of all taxonomically ranked reads and a set of coverage maps for detected microbial genomes, which reflect the highest percentage of coverage value generated by the reads alignment against the most likely reference genomes (60).

Pathosphere (https://www.pathosphere.org/) is another free open source software with the ability to process NGS data generated on the 454, Ion Torrent and Illumina platforms. Its analytical capabilities include taxonomic analysis of sequence files, sequence assembly, and pathogen identification using a variety of databases, including NCBI. The software allows automating the work, being only necessary to upload the FASTQ files (or other formats) in the user interface, then select the identification pipeline and, finally, wait for the report generation (61).

One critical comment is that while both pipelines described above are designed for application of data files from metagenomic sequences, the two pipelines employ assembles dedicated to single genomes. In SUPRI, the software ABySS and Minimus are used to assemble the metagenomes, but as they are appropriate also to assembly unique genomes and, to use them for WGS analysis is also feasible. As in Pathosphere the assemblers used are Velvet and GS Newbler (Roche) also applied to unique genomes, the ECBC pipeline would be useful for WGS data analysis, whereas for the metagenomic analysis the USAMRIID-WRAIR pipeline would be more appropriate since it employs a specific metagenomic assembler, the Ray Meta.

Moreover, in the case of diagnostic metagenomics applications, another very useful user-friendly software designed for the clinical routine is the PathoScope (http://sourceforge.net/projects/pathoscope/), characterized as an accurate pipeline for identification of virus and bacteria in clinical samples. The algorithm is able to discriminate among multiple pathogens in the sample, as well as to differentiate new strains and also those with high mutation rates. The concept of the software is to employ a Bayesian method of statistical inference to process sequence alignment and return the probability of the profiles of the organisms present, based on the NCBI database, although custom databases can be incorporated as user needs. Since PathoScope relies on sequence alignment without prior assembly stage, which would delay the analysis, its results are generated in a shorter time interval (62).

A complete pipeline for diverse purposes is supplied by the Galaxy software (http://galaxy.project.org) that processes the sequences of the user by its own defaults, but it results in hiding the computational details from the analyst. The software is available as a web-based service, in which is possible to access several pipelines for genomic and functional analysis, or it can be downloaded as a package in a laboratory informatics dedicated server (63).

The WGS approach in clinical microbiology leads to a deep analysis for variant discovery in bacterial isolates which is important for the management of antibiotic resistance genes due to the possibility of evaluate a single nucleotide mutation (SNP) and/or indel variants in a given bacterial genome to identify know/unknown ARGs or pathogenic islands by comparison of the input sequences with sequences deposited in public databases (12).

For this purpose, a GATK pipeline (64) was drawn for variant discovery analysis in High-Throughput sequencing (HTS) data. This software is open source and is available in https://software.broadinstitute.org/gatk/download/. The pipeline consists in three steps for data processing: (i) Pre-processing of raw sequence data to produce a BAM file (a file derivate from the sequence alignment against a reference database); (ii) production of a VCF file (variant calling format file) which presents the variant genomic information and; (iii) additional steps for genomic annotation and downstream analysis. This pipeline supports data from WGS, exomes, gene panels and RNA-seq (https://software.broadinstitute.org/gatk/best-practices/).

The Galaxy cloud-based service and GATK pipeline are recommended by the ENGAGE project. The former is useful when there is no infrastructure for bioinformatics server implementation (40).

DATABASES FOR MAIN APPLICATIONS

Virulence Factors

Several pipelines may be customized by the user due to the possibility of change defaults and to choose a specific database to improve the analysis. In the clinical microbiology field it is crucial to search for virulence factors in a bacterial isolate since, bacteria use many strategies to cause infections and diseases. This is due to the fact that these microorganisms present structures and metabolic networks related to their virulence, with a focus on adhesion and host

invasion. Many structures and metabolites may be considered as virulence factors, for example, the presence of polysaccharide capsule, cell wall, toxin production and expression of microbial adhesins, as well as the expression of enzymes and proteins that contribute to antimicrobial resistance mechanisms (65).

The virulence factor (VF) database (VFDB- http://www.mgc.ac.cn/VFs/), released in 2005, presents information on the principal VFs related to bacterial pathogens, emphasizing aspects of structural and functional biology. The database is user-friendly and the search for VFs can be performed simply by text query or by selecting the function category under consideration. The software also supports BLAST, which compares its sequences with the entire VFs database (66).

In 2012, this database was updated to an advanced user interface and also with new contents related to the comparative analysis of VFs between intergenera that are related to host adhesion and invasion, bacterial secretion systems and its effectors, secretion of toxins and ion capture systems (67).

Antibiotic Resistance

Three major databases stand out in the search for resistance genes in bacterial genomes derived from WGS: ResFinder (https://cge.cbs.dtu.dk/services/ResFinder/), ARG-ANNOT (http://www.mediterranee-infection.com/article.php?laref=282&titer=arg-annot) and CARD (https://card.mcmaster.ca/) for being constantly updated and revised.

The ResFinder server came from an effort by The Center for Genomic Epidemiology (Denmark) to provide researchers with poor knowledge in bioinformatics a user-friendly interface, which allows analysis of WGS data for outbreak investigation, laboratory diagnosis and also epidemiological surveillance. The construction of this database was carried out on the basis of information deposited with other databases, as well as information available in review articles. Sequences related to resistance genes were taken from the NCBI database and used for the assembly of ResFinder (68).

The user has different possibilities with ResFinder, such as insert as input unassembled sequences files, that can be assembled by the server's own algorithm or even complete or partially assembled genome sequences. The server supports sequences files from the 454, Illumina, Ion Torrent and SOLiD platforms. By means of the BLAST algorithm, the server aligns the user sequences with the resistance gene sequences available in the database, returning as output the gene sequences with the highest match (68).

The ARG-ANNOT server is another valuable tool available for the analysis of the presence/absence of antibiotic resistance genes (AR). This server was designed for the analysis of genomic and metagenomic data, to serve as a tool for quick identification and prediction of the presence of existing, putative or new AR genes as well as point mutations in chromosomes that contain the sequences of interest. This database was created based on the classification of databases for resistance genes and previous publications available in PubMed (69). The highlight of this tool is that it is able to identify AR genes not only in complete gene sequences, but also in partially assembled sequences and in those with low levels of similarity to the sequences of the databases. According Gupta et al. (69), when compared to ResFinder, ARG-ANNOT demonstrated greater ability to detect resistance genes because ResFinder only detects genes with high similarity (≥50%) and high sequencing coverage, predicting only known resistance genes, which limits the application from this server for the prediction of unknown AR genes (69).

In response to the findings of Gupta et al. (69), Zankari (70) reported that the first version of the ResFinder was compared by the former author. Currently, ResFinder allows the user to choose to reduce the identity to 20% and coverage size to 30%, also allowing the detection of AR genes with low similarities. However, with that choice, specificity for hits with resistance genes may decrease and consequently, may leads to increase the number of false-positive results for AR genes. The ResFinder, according to Zankari (70), has 99.74% agreement between AR prediction data and the results of *in vitro* antimicrobial susceptibility assays due to its high specificity. Therefore, to choose between these tools it is necessary to compare novel WGS bioinformatics pipelines with the conventional methods.

Finally, CARD is a user-friendly database of gene sequences data from antimicrobial resistance (AMR) genotypes obtained from genomic data. The deposited data includes information on mechanisms of intrinsic resistance, resistance genes and acquisition of resistance by mutations at target sites of antibiotics and associated elements. The CARD server core consists of a new database titled Antibiotic Resistance Ontology (ARO), which describes the targets of antimicrobial molecules, resistance mechanisms, genes and mutations and their interactions. In addition, CARD includes a software called Resistance Gene Identifier (RGI), capable of predicting resistance genes from sequenced genomes and contigs. The advantage of this server is that it allows carrying out integrated epidemiological surveillance with data from health institutions, agricultural regions and the environment to track an outbreak (71).

The databases for AR gene detection differ according to the degree of update and the definitions of resistance adopted (69,71). Therefore, the choice of the database or more than one of them must be performed according to the purpose and need of the work. The ENGAGE project report presents a list of databases dedicated to pathogen identification and characterization, which is available from Hendriksen et al. (40).

GENOTYPING

One of the most widely used methods for identifying pathogenic strains from bacterial isolates is the Multilocus Sequence Typing (MLST), which is of crucial importance for outbreak tracking and public health decision-making to design control strategies (72). The classic MLST analyses were developed to be compatible with the Sanger sequencing platform, based on PCR amplification of nucleotide sequences ranging from 400-500 bp of seven housekeeping genes (72,73). The amplified gene regions are sequenced and those that are present only within a given species receive an allele number. Thus, each isolate is characterized according to its alleles at each of the seven loci, constituting its allelic profile or Sequence Type (ST) (72).

The MLST was developed to overcome the limitation of the precursor's molecular techniques to typing microorganisms and the lack of reproducibility around laboratories based on these methods (74). The first bacterial specie tested with the MLST approach was *Neisseria meningitides* in 1998. On that occasion, this approach showed a powerful discriminatory power to differentiate clonal lineages with distinct rates of recombination. The highly clonal species may be discriminated by the phylogenetic relationships between isolates in dendograms built from the pairwise distances between STs and independently from a consensus tree assembled from the gene sequences, while weakly clonal lineages the dendograms exhibits clusters of isolates with identical or very similar STs (75).

After the *N.* meningitides MLST scheme others authors proposed for bacteria and fungi novel schemes and the MLST configured as the "gold standard" for pathogen typing. However, the time-consuming and the costly procedures make the WGS technology attractive, since with the NGS technologies advancement and their cost-decreasing price of the DNA sequencing it was more feasible a clinical laboratory to apply this methodology in routine (74).

With WGS it is possible to access all the genes of a bacterial genus or species, including those housekeeping genes used for MLST. These data can be used to perform whole-genome MLST (wgMLST). Another approach for analysis of WGS is focused on the whole set of genes universally present in a particular genus or species, which is defined as core genome MLST (cgMLST). In both approaches, the genome sequences of the isolates are compared against an appropriate database containing sequences of all deposited allelic variants, which allows tracing the relationships among the isolates and the reference species (76).

For genotyping using WGS data, it is possible to use the BacWGSTdb (Bacterial Whole Genome Sequence Database - http://bacdb.org/BacWGSTdb), which allows the extraction of MLST information in sequenced bacterial genomes. What stands out in this database is the possibility of relating data from the clinical isolate under study with phylogenetically related isolates in the database (77). To use this database, it is necessary simply to upload the query genome, which will be aligned against a user-selected reference genome, generating a Variant Call Format (VCF) standard extension file. In this file there is the SNP data generated in the alignment. In the next step, the SNP data are compared with the repository in the database and the most closely related isolates are listed in a phylogenetic tree. This tool is very useful for outbreak tracking (77).

Considering that for epidemiological surveillance the analysis of generated wgMLST data by different laboratories should be standardized and comparable, the authors Liu, Chiou and Chen (78) developed a WEB-based computational tool to create a pan-genome allele database (PGAdb - http://wgmlstdb.imst.nsysu.edu.tw/), to enable the comparison of data among laboratories and to establish a repository containing information on the allelic variants of the populations of a given bacterial organism (77).

The software works through two modules: in the first, called "Build_PGAdb", the contigs of the file uploaded by the user are functionally annotated by using the software Prokka and later the module extracts the information from the identified alleles based on orthologous genes, producing a pan-genomic database. In the second module, entitled "Build_wgMLSTtree", the contigs of the uploaded genomes are compared to the PGAdb database generated through the BLASTN between the contigs and the PGAdb for the construction of phylogenetic trees (78).

CONCLUSIONS

Even with so many different applications of the WGS for diagnostic microbiology, barriers related to the interpretation of data by microbiologists, which rely on bioinformaticians for data analysis, creates a major limitation for NGS technology to be incorporated into routine laboratories. Therefore, the development of user-friendly software is very important to facilitate the work of the laboratory analyst unfamiliar with advanced bioinformatics tools. On the other hand, in reference laboratories, the WGS may be more plausible to use focusing on confirmation of suspicious pathogens and/or on molecular typing of bacterial isolates important for public health (25).

Although many refinements must be made for the insertion of NGS technology into the diagnostic routine in microbiology laboratories, WGS-based diagnosis is a potential tool that can be made a reality, at any time, in the clinical microbiology field, since some research groups are using this technology to solve outbreaks, as those caused by *Salmonella* Typhimurium *S*. Enteritidis (79,80) and to identify pathogens and their virulence factors, as Verotoxigenic *Escherichia coli* (81).

In addition, the prepare of human resources to handle with NGS technology can be achieved employing dedicated educational resources that are proposed using practical simulations of NGS data acquisition and bioinformatics analysis as suggested by Macori et al. (82). This approach may be useful to microbiology and life sciences students interested to learn NGS applications in different fields.

Allied with the educational resources, it is expected that standardized pipelines for the analysis of genomic data will be improved and/or developed and, in the case of reference laboratories carrying out epidemiological surveillance, these pipelines must allow the comparison of data among laboratories regarding to clinical isolates and must be evaluated by certified institutions around the world to ensure a high-quality diagnostic protocol.

ACKNOWLEDGEMENTS

This work was supported by the São Paulo Research Foundation - FAPESP, Brazil [grant number 2017/13759-6], the National Council for Scientific and Technological Development [grant number 306762/2006-4] and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

1. Weile J, Knabbe C. Current applications and future trends of molecular diagnostics in clinical bacteriology. Anal Bioanal Chem. 2009;394(3):731-42. https://doi.org/10.1007/s00216-009-2779-8 PMid:19377839
2. Pitt TL, Saunders NA. Molecular bacteriology: a diagnostic tool for the millennium. J Clin Pathol. 2000;53:71-5. https://doi.org/10.1136/jcp.53.1.71 PMCid:PMC1731063
3. Pierro A, Sambril V. Molecular methods: are their results of help or they make more confuse the clinical management of patients? Microbiologia Medica. 2016;31(4):97-8. https://doi.org/10.4081/mm.2016.6494
4. Srinivasan R, Karaoz U, Volegova M, et al. Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. PLoS ONE. 2015;10(2):e0117617. https://doi.org/10.1371/journal.pone.0117617
5. Clarridge III, JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004; 17(4):840-62. https://doi.org/10.1128/CMR.17.4.840-862.2004 PMid:15489351 PMCid:PMC523561
6. Yarza P, Yilmaz P, Pruesse E, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12(9):635-45. https://doi.org/10.1038/nrmicro3330 PMid:25118885
7. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007;45(9):2761-4. https://doi.org/10.1128/JCM.01228-07 PMid:17626177 PMCid:PMC2045242
8. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE. 2011;6(12):e27310. https://doi.org/10.1371/journal.pone.0027310
9. Patel JB. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. Mol Diagn. 2001;6(4):313-21. https://doi.org/10.1054/modi.2001.29158
10. Deurenberg RH, Bathoorn E, Chlebowicz MA, et al. Application of Next-Generation Sequencing in clinical microbiology and infection prevention. J Biotechnol. 2017;243:16-24. https://doi.org/10.1016/j.jbiotec.2016.12.022 PMid:28042011

11. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The Next-Generation Sequencing revolution and its impact on genomics. Cell. 2013;155(1):27-38. https://doi.org/10.1016/j.cell.2013.09.006 PMid:24074859 PMCid:PMC3969849
12. Mardis ER. The impact of Next-Generation Sequencing technology on genetics. Trends Genet. 2008;24(3):133-41. https://doi.org/10.1016/j.tig.2007.12.007 PMid:18262675 PMCid:PMC2680276
13. Buchan BW, Ledeboer NA. Emerging technologies for the clinical microbiology laboratory. Clin Microbiol Rev. 2014;27(4):783-822. https://doi.org/10.1128/CMR.00003-14 PMid:25278575 PMCid:PMC4187641
14. Ansorge WJ. Next-Generation DNA Sequencing techniques. N Biotechnol. 2009; 25(4):195-203. https://doi.org/10.1016/j.nbt.2008.12.009 PMid:19429539
15. Macori G, Romano A, Adriano D, et al. Draft genome sequences of four Yersinia enterocolitica strains, isolated from wild ungulate carcasses. Genome Announc. 2017; 5(15). pii: e00192-17. https://doi.org/10.1128/genomeA.00192-17
16. Stevens MJ, Stephan R, Johler S. Draft genome sequence of Staphylococcus aureus 1608, a strain that caused toxic mastitis in twin cows. Genome Announc. 2017;5(1). pii: e01438-16. https://doi.org/10.1128/genomeA.01438-16
17. Schmidt T, Kock MM, Ehlers MM. Molecular characterization of Staphylococcus aureus isolated from bovine mastitis and close human contacts in South African dairy herds: genetic diversity and inter-species host transmission. Front Microbiol. 2017; 8:511. https://doi.org/10.3389/fmicb.2017.00511 PMid:28428772 PMCid:PMC5382207
18. Kluytmans JA. Methicillin-resistant Staphylococcus aureus in food products: cause for concern or case for complacency? Clin Microbiol Infect. 2010; 16(1):11-5. https://doi.org/10.1111/j.1469-0691.2009.03110.x PMid:20002686
19. Tsogalis GJ, Chao E, Hagenkord JM, Hambuch T, Moore JH. Bioinformatics: what the clinical laboratorian needs to know and prepare for. Clin Chem. 2013; 59(9):1301-5. https://doi.org/10.1373/clinchem.2012.198226 PMid:23723312
20. Almeida OGG, De Martinis ECP. Bioinformatics tools to assess metagenomic data for applied microbiology. Appl Microbiol Biotechnol. 2019; 103(1):69-82. https://doi.org/10.1007/s00253-018-9464-9 PMid:30362076
21. El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: four stages of data processing and computational challenges. PLOS Comput Biol. 2013; 9(12):e1003345. https://doi.org/10.1371/journal.pcbi.1003345
22. van Djik EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of Next-Generation Sequencing technology. Trends Genet. 2014;30(9):418-26. https://doi.org/10.1016/j.tig.2014.07.001 PMid:25108476
23. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010(6):pdb.prot5448. https://doi.org/10.1101/pdb.prot5448
24. Head SR, Komori HK, LaMere SA, et al. Library construction for Next-Generation Sequencing: overviews and challenges. Biotechniques. 2014; 56(2):61-4, 66, 68, passim. https://doi.org/10.2144/000114133
25. Török ME, Peacock SJ. Rapid Whole-Genome Sequencing of bacterial pathogens in the clinical microbiology laboratory-pipe dream or reality?. J Antimicrob Chemother. 2012;67(10):2307-08. https://doi.org/10.1093/jac/dks247 PMid:22729921
26. Varshney RK, Nayak SN, May GD, Jackson SA. Next-Generation Sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 2009;27(9):522-30. https://doi.org/10.1016/j.tibtech.2009.05.006 PMid:19679362
27. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Res. 2001;11(1):3-11. https://doi.org/10.1101/gr.11.1.3 PMid:11156611
28. Goodwin S, McPherson JD, McCombie R. Coming of age: ten years of Next-Generation Sequencing technologies. Nat Rev Genet. 2016;17(6):333-51. https://doi.org/10.1038/nrg.2016.49 PMid:27184599
29. Schadt EE, Truner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227-40. https://doi.org/10.1093/hmg/ddq416
30. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics. 2016;107(1):1-8. https://doi.org/10.1016/j.ygeno.2015.11.003 PMid:26554401 PMCid:PMC4727787

31. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. Genomics Proteomics Bioinformatics. 2016;14(5):265-79. https://doi.org/10.1016/j.gpb.2016.05.004 PMid:27646134 PMCid:PMC5093776
32. Kulkarni P, Frommolt P. Challenges in the setup of large-scale Next-Generation Sequencing analysis workflows. Comput Struct Biotechnol J. 2017;15:471-7. https://doi.org/10.1016/j.csbj.2017.10.001 PMid:29158876 PMCid:PMC5683667
33. Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. Bioessays. 2010;32(6):524-36. https://doi.org/10.1002/bies.200900181 PMid:20486139
34. Quail MA, Smith M, Coupland P, et al. A tale of three Next Generation Sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:341. https://doi.org/10.1186/1471-2164-13-341 PMid:22827831 PMCid:PMC3431227
35. Mikheyev AS, Tin MM. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014;14(6):1097-102. https://doi.org/10.1111/1755-0998.12324 PMid:25187008
36. Illumina sequencing platforms. [online] [accessed 2019-01-13]. Available from: https://www.illumina.com/systems/sequencing-platforms.html
37. Bioinformatics Definition Committee. NIH working definition of bioinformatics and computational biology. 2000 [online] [accessed 2019-01-13]. Available from: http://2digitstechcom.ipage.com/uploads/2/9/0/1/2901227/compubiodef.pdf
38. Fierro RG, Thomas-Lopez D, Deserio D, Liebana E, Rizzi V, Guerra B. Outcome of EC/EFSA questionnaire (2016) on use of Whole Genome Sequencing (WGS) for food- and waterborne pathogens isolated from animals, food, feed and related environmental samples in EU/EFTA countries. EFSA Supporting Publications. 2018;15(6):1432E. https://doi.org/10.2903/sp.efsa.2018.EN-1432
39. Nadon C, Van Walle I, Gerner-Smidt P, et al. PulseNet International: vision for the implementation of Whole Genome Sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 2017;22(23). pii: 30544. https://doi.org/10.2807/1560-7917.ES.2017.22.23.30544
40. Hendriksen RS, Pedersen SK, Leekitcharoenphon P, et al. Final report of ENGAGE - Establishing Next Generation sequencing Ability for Genomic Analysis in Europe. EFSA Supporting Publications. 2018;15(6):1431E. https://doi.org/10.2903/sp.efsa.2018.EN-1431
41. Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics. 2015;31(21):3476-82. https://doi.org/10.1093/bioinformatics/btv401 PMid:26139637
42. Wajid B, Serpedin E. Do it yourself guide to genome assembly. Brief Funct Genomics. 2016;15(1):1-9. https://doi.org/10.1093/bfgp/elu042 PMid:25392234
43. Centers for Disease Control and Prevention (USA). WGS protocols. [online] [accessed 2019-01-13]. Available from: https://www.cdc.gov/pulsenet/pathogens/protocols.html
44. Patel RK, Jain M. NGS QC toolkit: A toolkit for quality control of Next Generation Sequencing data. PloS One. 2012;7(2):e30619. https://doi.org/10.1371/journal.pone.0030619
45. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. https://doi.org/10.1093/bioinformatics/btu170 PMid:24695404 PMCid:PMC4103590
46. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10-2. https://doi.org/10.14806/ej.17.1.200
47. Illumina®. BaseSpace sequence Hub. [online] [accessed 2019-01-13]. Available from: https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html
48. Thermo Fisher. NextGENeTM Software for Ion TorrentTM Academic/Network License. [online] [accessed 2019-01-13]. Available from: https://www.thermofisher.com/order/catalog/product/4467742
49. PACBIO®. SMRT Analysis Software. [online] [accessed 2019-01-13]. Available from: https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/
50. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS ONE. 2013;8(12):e85024. https://doi.org/10.1371/journal.pone.0085024
51. Miller JR, Koren S, Sutton G. Assembly algorithms for Next-Generation Sequencing data. Genomics. 2010; 95(6):315-27. https://doi.org/10.1016/j.ygeno.2010.03.001 PMid:20211242 PMCid:PMC2874646
52. Ekblom R, Wolf JB. A field guide to Whole-Genome Sequencing, assembly and annotation. Evol Appl. 2014;7(9):1026-42. https://doi.org/10.1111/eva.12178 PMid:25553065 PMCid:PMC4231593

53. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77. https://doi.org/10.1089/cmb.2012.0021 PMid:22506599 PMCid:PMC3342519
54. Wang W, Li GW, Chen C, Xie XS, Zhuang X. Chromosome organization by a nucleoid-associated protein in live bacteria. Science. 2011;333(6048):1445-9. https://doi.org/10.1126/science.1204697 PMid:21903814 PMCid:PMC3329943
55. Smalla K, Jechalke S, Top EM. Plasmid detection, characterization and ecology. Microbiol Spectr. 2015;3(1):PLAS-0038-2014. https://doi.org/10.1128/microbiolspec.PLAS-0038-2014
56. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genom. 2017;3(10):e000128. https://doi.org/10.1099/mgen.0.000128
57. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling plasmids from Whole Genome Sequencing data. Bioinformatics. 2016;32(22):3380-7. https://doi.org/10.1093/bioinformatics/btw493
58. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. Microbiome. 2015;3:31. https://doi.org/10.1186/s40168-015-0094-5 PMid:26229597 PMCid:PMC4520061
59. Leipzig J. A review of bioinformatic pipeline frameworks. Brief Bioinform. 2017;18(3):530-6.
60. Naccache SN, Federman S, Veeeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from Next-Generation Sequencing of clinical samples. Genome Res. 2014; 24(7):1180-92. https://doi.org/10.1101/gr.171934.113 PMid:24899342 PMCid:PMC4079973
61. Kilianski A, Carcel P, Yao S, et al. Pathosphere.org: pathogens detection and characterization through a web-based, open source informatics platform. BMC Bioinformatics. 2015;16:416. https://doi.org/10.1186/s12859-015-0840-5 PMid:26714571 PMCid:PMC4696252
62. Byrd AL, Perez-Rogers JF, Manimaran S, et al. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics. 2014;15(1):262. https://doi.org/10.1186/1471-2105-15-262 PMid:25091138 PMCid:PMC4131054
63. Blankenberg D, Kuster GV, Coraor N, et al. Galaxy, a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010;Chapter 19:Unit 19.10.1-21. https://doi.org/10.1002/0471142727.mb1910s89
64. De Summa, S Malerba, G Pinto, R Mori, A Mijatovic V, Tommasi S. GATK hard filtering: tunable parameters to improve variant calling for Next Generation Sequencing targeted gene panel data. BMC Bioinformatics. 2017;18(Suppl 5):119. https://doi.org/10.1186/s12859-017-1537-8
65. Wilson JW, Schurr MJ, LeBlanc CL, Ramamurthy R, Buchanan KL, Nickerson CA. Mechanisms of bacterial pathogenicity. Postgrad Med J. 2002;78(918):216-24. https://doi.org/10.1136/pmj.78.918.216 PMid:11930024 PMCid:PMC1742320
66. Chen L, Yang J, Yu J, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33(database issue):D325-8. https://doi.org/10.1093/nar/gki008
67. Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 2011;40(database issue):D641-5. https://doi.org/10.1093/nar/gkr989
68. Zankari E, Hasman H, Cosentino S, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67(11):2640-4. https://doi.org/10.1093/jac/dks261 PMid:22782487 PMCid:PMC3468078
69. Gupta SK, Padmanabhan BR, Diene SM, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58(1):212-20. https://doi.org/10.1128/AAC.01310-13 PMid:24145532 PMCid:PMC3910750
70. Zankari E. Comparison of the web tools AR-ANNOT and ResFinder for detection of resistance genes in bacteria. Antimicrob Agents Chemother. 2014;58(8):4986. https://doi.org/10.1128/AAC.02620-14 PMid:25028728 PMCid:PMC4136053
71. Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017; 45(D1):D566-D573. https://doi.org/10.1093/nar/gkw1004
72. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect Genet Evol. 2013;16:38-53. https://doi.org/10.1016/j.meegid.2013.01.009 PMid:23357583

73. Dekker JP, Frank KM. Next-generation epidemiology: using real-time core genome multilocus sequence typing to support infection control policy. J Clin Microbiol. 2016; 54(12):2850-3. https://doi.org/10.1128/JCM.01714-16 PMid:27629902 PMCid:PMC5121370
74. Larsen MV, Cosentino S, Rasmussen S, et al. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 2012;50(4):1355-61. https://doi.org/10.1128/JCM.06094-11 PMid:22238442 PMCid:PMC3318499
75. Maiden MC, Bygraves JA, Feil E, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95(6):3140-5. https://doi.org/10.1073/pnas.95.6.3140 PMid:9501229 PMCid:PMC19708
76. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-Generation Sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect. 2018;24(4):335-41. https://doi.org/10.1016/j.cmi.2017.10.013 PMid:29074157 PMCid:PMC5857210
77. Ruan Z, Feng Y. BacWGSTdb, a database for genotyping and source tracking bacterial pathogens. Nucleic Acids Res. 2016; 44(D1):D682-7. https://doi.org/10.1093/nar/gkv1004
78. Liu YY, Chiou CS, Chen CC. PGAdb-builder: A web service tool for creating pan-genome allele database for molecular fine typing. Sci Rep. 2016;6:36213. https://doi.org/10.1038/srep36213 PMid:27824078 PMCid:PMC5099940
79. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. Evaluation of Whole Genome Sequencing for outbreak detection of Salmonella enterica. PLoS One. 2014;9(2):e87991. https://doi.org/10.1371/journal.pone.0087991
80. Inns T, Ashton PM, Herrera-Leon S, et al. Prospective use of Whole Genome Sequencing (WGS) detected a multi-country outbreak of Salmonella Enteritidis. Epidemiol Infec. 2017;145(2):289-98. https://doi.org/10.1017/S0950268816001941 PMid:27780484
81. Joensen KG, Scheutz F, Lund O, et al. Real-time Whole-Genome Sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol. 2014;52(5):1501-10. https://doi.org/10.1128/JCM.03617-13 PMid:24574290 PMCid:PMC3993690
82. Macori G, Romano A, Decastelli L, Cotter, PD. Build the read: a hands-on activity for introducing microbiology students to Next-Generation DNA Sequencing and bioinformatics. J Microbiol Biol Educ. 2017;18(3). pii: 18.3.62. https://doi.org/10.1128/jmbe.v18i3.1363

◇◈◆◈◇

http://www.ejgm.co.uk