JMBFS

Journal of Microbiology, Biotechnology and Food Sciences

International peer-reviewed scientific online journal

Published by Faculty of Biotechnology and Food Sciences

NEXT-GENERATION WHOLE-GENOME SEQUENCING PLATFORMS AND FACTORS TO CONSIDER FOR BACTERIAL APPLICATIONS

Yousef I. Hassan[1], *Dion Lepp*[1], *and Ting Zhou*[1*]

Address(es): Dr. Ting Zhou.
[1]Guelph Food Research Centre, Agriculture and Agri-Food Canada (AAFC), Guelph, Ontario N1G 5C9, Canada (Phone: 1 (226) 217-8084).

*Corresponding author: ting.zhou@agr.gc.ca

doi: 10.15414/jmbfs.2015.5.1.29-33

ARTICLE INFO

Received 26. 5. 2014
Revised 12. 1. 2015
Accepted 13. 5. 2015
Published 1. 8. 2015

Short communication

OPEN ACCESS

ABSTRACT

Next-generation sequencing (NGS) technologies are increasingly being used by microbiologists for characterizing bacterial isolates. A draft genome can now be obtained within a few days. With the recent technological advancements in NGS, the affordability and feasibility of this technique promises to redirect such tasks from research laboratories to the daily practices of clinical and environmental microbiology laboratories in the near future. This short technical note will introduce the average microbiologist, who may have only limited knowledge of NGS technologies, to the most commonly used platforms currently on the market. The advantages and disadvantages of each platform and the technical-terminology essential to navigate the planning phase of any bacterial genome assembly are also highlighted.

Keywords: Whole-genome assembly, next-generation sequencing, bacteria, applications

INTRODUCTION

Next-generation sequencing (NGS) technologies are increasingly being used by microbiologists for characterizing bacterial isolates, alongside more traditional methodologies. Advances in NGS, and the related bioinformatics tools, are allowing for higher quality *de novo* genomic assemblies and, at the same time, making the technology more feasible and affordable for small-laboratory settings **(Hernandez *et al.*, 2008; Hassan *et al.*, 2014)**. The objectives of this short technical note is to introduce the average microbiologist, who may have only limited knowledge of NGS technologies, to the most commonly used techniques currently found on the market, their advantages and disadvantages, and the technical-terminology essential during the planning phase of any *de novo* bacterial genome sequencing project.

With the recent technological advancements in NGS, a draft bacterial genome can now be obtained within a few days (figure 1). The affordability and feasibility of whole-genome sequencing, especially for smaller genomes (ie. 3-6 Mb), promises to redirect such tasks from research laboratories to the daily practices of clinical and environmental microbiology laboratories in the near future **(Edwards and Holt, 2013)**, particularly with the recent introduction of small-scale bench-top sequencing machines **(Perkins *et al.*, 2013)**. While there are currently a number of NGS platforms on the market, with development undergoing, this technical note will focus on two technologies, PacBio RSII and Illumina HiSeq/MiSeq (for sequencing of larger eukaryotic genomes, the readers are referred to more in-depth reviews **(Yandell and Ence, 2012)**). These two systems are arguably the most commonly used for bacterial genome sequencing at the time of writing this manuscript, and each has unique advantages for this task which distinguishes it from others **(Quail *et al.*, 2012)**.

a. The PacBio sequencing technology

PacBio sequencing chemistry depends on the action of DNA polymerase **(Hert *et al.*, 2008)**. The enzyme activity takes place within SMRT (single molecule real time sequencing) cells containing thousands of zero-mode waveguides (ZMWs), within each a single moiety of the enzyme is immobilized on the surface. During the highly efficient and accurate DNA replication process, the polymerase examines each individual base and then incorporates the matching nucleotide into the growing strand before moving to the next base. Labeled nucleotides with different fluorescent dyes diffuse into the ZMW chambers and emit a signal that identifies each base upon their incorporation into the DNA strand.

The practical PacBio sequencing process include the following steps: the preparation of 8–12 kb libraries, sequencing on the PacBio RS II (100X raw data is recommended), filtering and error correction of long reads through alignment of shorter reads, *de novo* assembly of the error-corrected reads, and finally the assembly through re-alignment of continuous long reads (CLR).

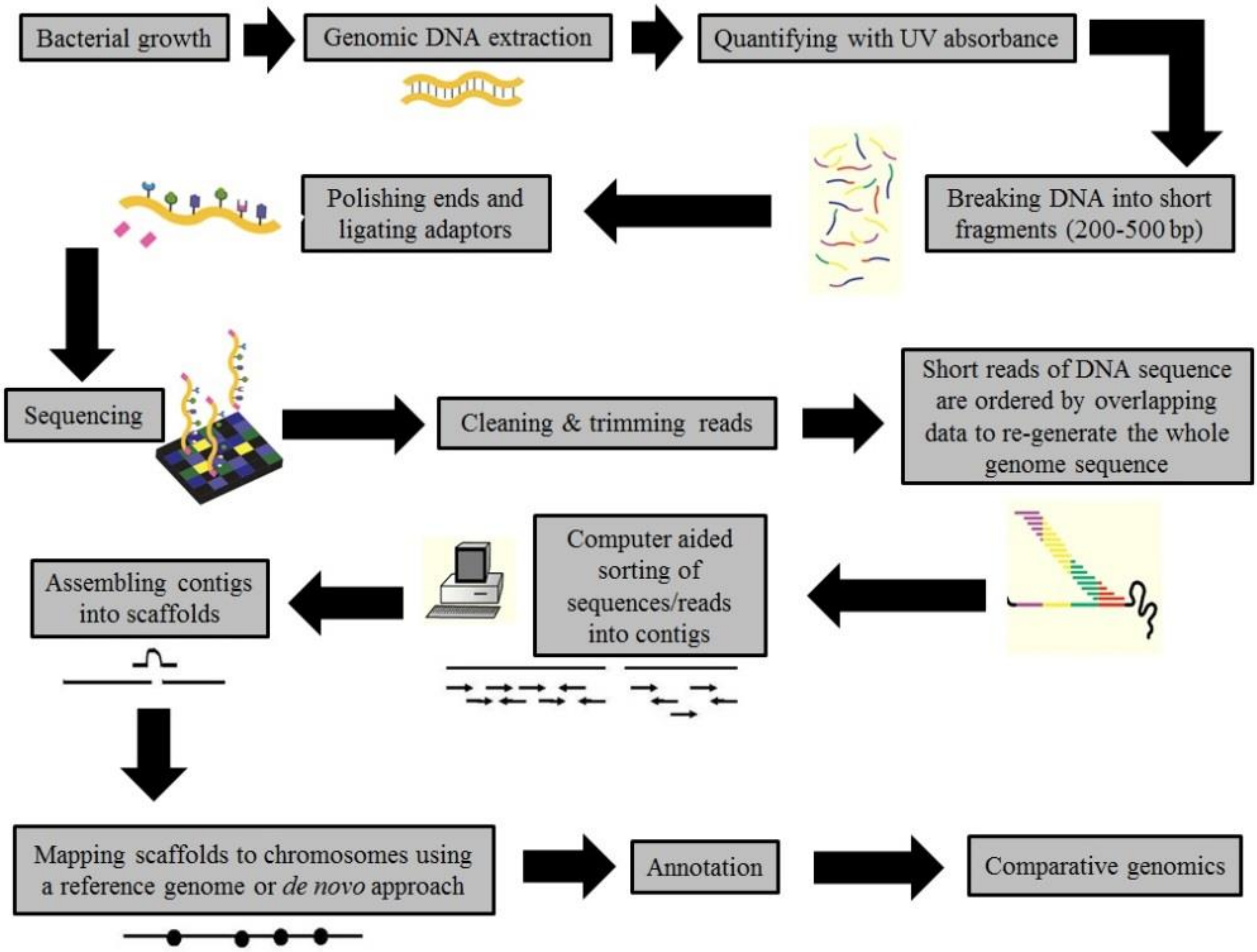
Bacterial growth
Genomic DNA extraction
Quantifying with UV absorbance
Breaking DNA into short fragments (200-500 bp)
Polishing ends and ligating adaptors
Sequencing
Cleaning & trimming reads
Short reads of DNA sequence are ordered by overlapping data to re-generate the whole genome sequence
Computer aided sorting of sequences/reads into contigs
Assembling contigs into scaffolds
Mapping scaffolds to chromosomes using a reference genome or de novo approach
Annotation
Comparative genomics

Figure 1 A draft bacterial genome can now be assembled within a few days. After preparing and quantifying the genomic DNA of the bacterial isolate, a fragmentation and end-polishing step with known adaptors is carried out. The adapters incorporate sites for sequencing primers and immobilized flow cell primers. The reads assembly is carried out using different algorithms and usually refined by mapping to a matching reference genome (if available).

If successful, the process above will yield approximately 25-fold genome coverage with quality-filtered and error-corrected reads. The PacBio RS II sequencing technology has the following advantages:

(a) It combines large insert size libraries (8 to 12 kb or longer, depending on DNA quality) and long sequencing reads (with an average > 4000-4500 bp). This is the biggest advantage of this technology, as longer reads are more likely to cover repetitive regions in the genome. Repetitive regions lead typically to more gaps and sequence errors during the assembly process.

(b) The sequencing provides uniform coverage (even for AT or GC rich regions) so it can span repeats, sequence palindrome (see Tab 1), and microsatellite regions.

(c) The pattern of errors along the sequence has a random distribution, and it's not dependent on the base composition of a defined region in the genome.

(d) Has the highest N50 value (Tab 1) and lowest number of contigs. While it is favorable to have a smaller number of contigs and higher N50 value, these metrics must also be considered in terms of overall assembly quality, since the incorrect joining of contigs may artificially alter these values.

(e) This technique provides information about any structural variations including deletions, duplications and rearrangements of DNA sequences (this is more important in eukaryotic and human genomes than bacterial ones).

Among the drawbacks of the PacBio RS II sequencing technology are: (a) relatively high error rates (11-15% on average) in single reads (figure 2); (b) the need for more sequencing cycles/repeats/SMRT cells in order to compensate for the high percentage of read-errors which **(c)** increases the overall costs. Finally, since the PacBio sequencing does not depend on any prior amplification steps (even though it can be done in some cases), (d) DNA quantity and quality is a major factor that influences the final assembly quality and should be considered during the planning phase. Steps should be taken to minimize freeze/thaw cycles or exposure to environmental factors (such as reagents/light/buffers) that may affect DNA stability. If obtaining 10 µg (200 ng/µl concentration) of good quality genomic DNA is not possible, then the Illumina platform may be more appropriate. Despite these disadvantages and the high single-read error rates, the PacBio technology has been used to successfully generate several single contig genome assemblies **(Chin *et al.*, 2013; Rehvathy *et al.*, 2013; Schmuki *et al.*, 2012; Zhang *et al.*, 2012)**.

Table 1 Definition of terms.

Term	Definition
Genome	The genetic material of an organism encoded either in DNA (genomic/plasmid/ribosomal) or RNA (in case of viruses).
Read	A continuous DNA sequence that is acquired by the sequencing machine. Multiple reads are assembled to create a contig (see contig).
Sequencing depth (Coverage)	The number of times a nucleotide is read during the sequencing process. Empirically, it represents the average number of reads representing a given nucleotide in the reconstructed genome.
Open reading frame (ORF)	Sequence of nucleotides in the DNA that contain no stop/termination codons so can potentially translate as a polypeptide chain.
de novo assembly	A method for creating draft genomes without the aid of a reference genome.
Draft genome	A genome that has been sequenced around 4-5 times, providing a template for DNA fragment assembly and potentially allowing up to 96-97% of genes to be identified/annotated. Draft genomes may contain gaps and the true order and orientation may not be completely known/correct.

Finished/Reference genome	The genome has been sequenced at least 10 times to reduce ambiguities in the draft sequence, close gaps, and allow for only a single error every 10,000 bp.
Contig	A contiguous DNA sequence generated by assembling smaller overlapping reads/consensus sequences.
N50	A parameter used to evaluate the quality of the final assembly. It represents the size of the smallest of all large contigs covering 50% of the genome.
Paired-end library	Refers to the two ends of the same DNA fragment where the sequence is obtained from the upper end (with the help of forward adaptor) and from the lower end (with the help of reverse adaptor) simultaneously. Paired-end sequencing facilitates detection of repetitive sequence elements and genomic rearrangements, as well as gene fusions and novel transcripts.
Microsatellite regions	DNA repeat units typically 1-5 bp, with repeat length rarely exceeding hundreds of repetitions in order. Microsatellites are very common in the genome, highly polymorphic, and are very often used as genetic markers.
Repetitive regions	Non-coding DNA stretches that come with high copy numbers. If copies of the sequence motif lie adjacent to each other in a block or an array, they referred to as tandem repeats.
Sequence palindrome	DNA strand/sequence that reads the same way in the 5′ → 3′ direction as the complementary strand reads in the 5′ → 3′ direction.

b. The Illumina-based sequencing systems

The Illumina-based systems utilize an entirely different chemistry, beginning with a solid phase amplification step that generates thousands of copies for each library fragment to be sequenced. A modified polymerase then begins incorporating dye-labeled and terminated nucleotides, followed by a detection step, and finally the removal of terminator/label to start the sequencing cycle once again. In practice, the Illumina sequencing process depends on fragmenting the whole genome to a small library (~250-1500 bp fragments), to which adapters are ligated onto the ends. One of the adapters serves to hybridize the fragment to the surface of a flow cell, upon which a localized Polymerase Chain Reaction (PCR) is performed to generate clusters. A mixture of modified nucleotides is added, each carrying a base-specific fluorescent label and a modified 3'-OH group to ensure the incorporation of only one nucleotide at a time.

Illumina technology has the following advantages:

(a) Since an amplification step is involved, the required starting amount of DNA is much less with as little as 2 µg of purified DNA (with a concentration ranging from 20 ng/µl to 500 ng/µl) required for some applications.

(b) Pair-end sequencing can be performed, in which both ends of the DNA fragment are sequenced in a single run, producing two reads per fragment. This not only doubles the total number of reads obtained, but since the distance between the paired reads is known, this spatial information can be used to assist in the assembly process and span repetitive regions.

(c) It is the highest throughput technique on the market and since many samples may be multiplexed in a single run (up to 24 samples), this substantially lowers the costs for the end-user.

(d) Has one of the lowest error rates among all platforms, with 85% of the reads having < 0.1% error rate **(Glenn, 2011)**.

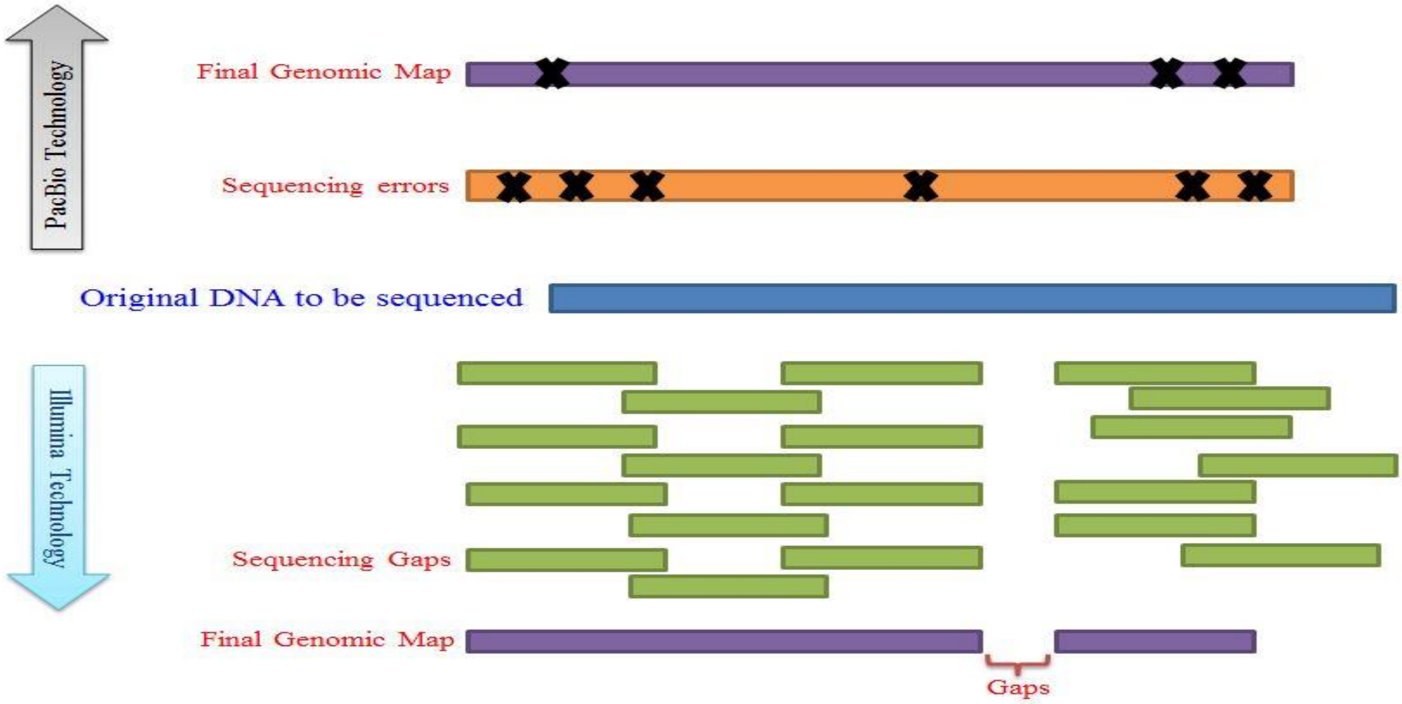

Figure 2 A comparison of PacBio RS II and Illumina next-generation sequencing technologies. Both technologies are predominant in *de novo* sequencing of bacterial genomes, yet with inherent disadvantages that can be reduced by hybrid approaches and clear understanding of the expected experimental outcomes. For example, if the purpose is to detect single nucleotide polymorphisms (SNP's)/mutations within the bacterial genome then the Illumina systems will be considered superior over here. While identifying novel genes and open reading frames (ORF) that possibly code for functional enzymes might be easier with complete/closed genomic maps assembled using the PacBio RS II technology.

The biggest disadvantages of Illumina systems are (a) relatively short generated reads (35-300 bp) (figure 2), which makes it more difficult to accurately assemble bacterial genomes as (b) gaps are more likely to exist with this approach. Furthermore, this approach leads to (c) larger number of contigs within the final map. Also with this approach, the (d) amplification efficiency is more likely to be jeopardized by difficult DNA templates such as high GC or AT content regions or that contain repetitive elements.

To capitalize on the advantages of the PacBio and Illumina systems, sequence data from both platforms may also be combined together or else combined with Sanger sequencing reads for genomes with particularly abundant or complex repeats. Such a hybrid assembly method effectively overcomes some of inherent disadvantages of any single method. In particular, it may close gaps that frequently remain in assemblies from Illumina data, while reducing sequence errors inherent to the PacBio system. This approach is particularly suited to sequencing uncharacterized genomes where no reference sequence is available, or known genomes where significant structural variation is expected **(Koren *et al.*, 2012; Wang *et al.*, 2012)**.

Due to the competition between different platforms and service providers, the actual costs involved in generating the raw reads of genome-assembly projects are dropping substantially. It's possible to accomplish a whole genome sequencing for a 4-5 Mbp bacterial genome with >100X coverage at an estimated cost between $2,000-3,000 (this figure was correct when this contribution was in preparation). We expect that these prices will further fall due to the increasing numbers of service providers in academic and commercial centers.

Despite the potential applications of NGS, this approach comes with some technical, computing, and storage challenges for the generated sequencing data (**Lampa *et al*., 2013**). The average-user computers are not suitable to process, assemble, and annotated medium to large size genomes (even some commercial software packages claim so). The demands on memory and storage space dictate certain levels of hardware capacity. Furthermore, the free available academic algorithms designed to tackle such assemblies need certain levels of comfort and proficiency in dealing with command-line bioinformatics. The alternative is to use the not so cheap (yet powerful and easy to use) commercial packages such as CLC Genomics Workbench or DNASTAR (Tab 2). The annotation of the generated genomes is another tedious task. Luckily the availability of government supported servers, such as the Rapid Annotation using Subsystem Technology (RAST) and NCBI Prokaryotic Genome Annotation Pipeline (PGAP), made this mission more feasible. Table (2) provides the reader with some of the leading technologies and helpful resources available on the web to consider before embracing such a project.

CONCLUSIONS

Genome sequence analysis is a key step for all branches of biological sciences. In microbiology, the falling costs of whole-genome sequencing are promising to take this field to another unmatched level. Sequencing of bacterial genomes is now a pivotal part of deciphering pathogen virulence and the phylogenetic relationship among related strains. It is also opening the door for gene/enzyme discoveries with agricultural and industrial applications. The three areas in clinical microbiology that can particularly benefit the most from routine whole-genome sequencings are: **(a)** detection and Identification, **(b)** drug susceptibility testing, and finally **(c)** epidemiological typing. Whole genome sequencings can accurately be used to detect nonculturabl or difficult-to-culture microorganisms, including fastidious bacteria and anaerobes. For example, this can be very important during the investigation of meta-genomic changes associated with intestinal microbiota alterations preceding bloodstream invasion by specific pathogen which might provide novel opportunities for treatment intervention **(Bertelli and Greug, 2013)**. NGS techniques can also permit the detection of resistance genes in the sequenced genomes and any related variants (such as point mutations or small insertions/deletions). This is currently considered very useful especially when obtaining the needed phenotypic data (such as antibiotic susceptibility) is time consuming, as the case of *M. tuberculosis* which requires weeks of growth and testing (**Koser *et al*., 2012**). Finally, whole genome sequencing can support outbreak investigations and make tracking the transmission pathways of pathogens easier due to the embedded high resolution nature of this method compared to the current conventional genotyping methods that investigate only short regions of bacterial genomes (**Dunne *et al*., 2012**). Readers are referred to more in-depth sources for further details and field-specific examples **(Padmanabhan *et al*., 2013; Chen, 2014; Lecuit and Eloit, 2014; Livermore and Wain, 2013; Nikolaki and Tsiamis, 2013)**.

In our opinion, whole-genome sequencing will soon become routine practice within environmental and clinical microbiology laboratories as a quick and affordable technique that allows: the characterization and tracking of microorganisms at strain levels, the comparison of genomic features and structures among closely-related species, and the understanding of genetic pathways in the identification of virulence factors **(Perkins *et al*., 2013)**.

Table 2 Some websites that can serve as useful resources while starting up whole-genome sequencing projects.

Resource	**Website**
Illumina	http://www.illumina.com/
Pacific Biosciences	http://www.pacificbiosciences.com/
List of PacBio sequencing service providers	http://www.pacificbiosciences.com/support/sequencing_provider/
Beijing Genomics Institute	http://www.genomics.cn/en/index
GATC Biotech	http://www.gatc-biotech.com/en/index.html
Genome Quebec	http://www.genomequebec.com/
DNASTAR software package for NGS analysis	http://www.dnastar.com/
CLC Genomics Workbench, NGS sequencing software package	http://www.clcbio.com/
Rapid Annotation using Subsystem Technology (RAST)	http://rast.nmpdr.org/
Next-generation sequencing forum	http://seqanswers.com/
Transcriptome Sequencing Research & Industry News	http://www.rna-seqblog.com/

Acknowledgment: Authors would like to acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) and Agriculture and Agri-Food Canada (AAFC), Growing Forward 2 program for their financial support to the first author.

Disclosure: Authors do not have any conflict of interests to disclose nor they endorse the use of any product/technology/service over others. The purpose of this note is to acquaint the non-expert reader with terms used by the industry when they plan for whole-genome sequencing projects and it should serve as a guideline only.

REFERENCES

BERTELLI, C., GREUG, G. 2013. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin Microbiol Infect.*, 19, 803-13 (http://dx.doi.org/10.1111/1469-0691.12217).

CHEN, CY. 2014. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present, *Front Microbiol.*, 5, 305 (http://dx.doi.org/10.3389/fmicb.2014.00305).

CHIN, C.S., ALEXANDER, D.H., MARKS, P., KLAMMER, A.A., DRAKE, J., HEINER, C., CLUM, A., COPELAND, A., HUDDLESTON, J., EICHLER, E.E., TURNER, S.W., KORLACH, J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*, 10, 563-569 (http://dx.doi.org/10.1038/nmeth.2474).

DUNNE, W.M. Jr., WESTBLADE, L.F., FORD B. 2012. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory, *Eur J Clin Microbiol Infect Dis.*, 31, 1719-26 (http://dx.doi.org/10.1007/s10096-012-1641-7).

EDWARDS, D.J., HOLT, K.E. 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation*, 3: 2 (http://dx.doi.org/10.1186/2042-5783-3-2).

GLENN, T.C. 2011. Field guide to next-generation DNA sequencers. *Molecular ecology resources*, 11, 759-769 (http://dx.doi.org/10.1111/j.1755-0998.2011.03024.x).

HASSAN, Y.I., LEPP, D., HE, J., ZHOU, T. 2014. Draft Genome Sequences of *Devosia* sp. Strain 17-2-E-8 and *Devosia riboflavina* Strain IFO13584. *Genome Announc.*, 2, 5, pii: e00994-14 (http://dx.doi.org/10.1128/genomeA.00994-14).

HERNANDEZ, D., FRANCOIS, P., FARINELLI, L., OSTERAS, M., SCHRENZEL, J. 2008. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*, 18, 802-809 (http://dx.doi.org/10.1101/gr.072033.107).

HERT, D.G., FREDLAKE, C.P., BARRON, A.E. 2008. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, 29, 4618-4626 (http://dx.doi.org/10.1002/elps.200800456).

KOREN, S., SCHATZ, M.C., WALENZ, B.P., MARTIN, J., HOWARD, J.T., GANAPATHY, G., WANG, Z., RASKO, D.A., MCCOMBIE, W.R., JARVIS, E.D., ADAM, M.P. 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30, 693-700 (http://dx.doi.org/10.1038/nbt.2280).

KOSER, C.U., ELLINGTON, M.J., CARTWRIGHT, E.J., GILLESPIE, S.H., BROWN, N.M., FARRINGTON, M., HOLDEN, M.T., DOUGAN, G., BENTLEY, S.D., PARKHILL, J., PEACOCK, S.J. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology, *PLoS Pathog.*, 8, e1002824 (http://dx.doi.org/10.1371/journal.ppat.1002824).

LAMPA, S., DAHLO, M., OLASON, P.I., HAGBERG, J., SPJUTH, O. 2013. Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data, Gigascience, 2(1):9, (http://dx.doi.org/ 10.1186/2047-217X-2-9).

LECUIT, M., ELOIT, M. 2014. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening, *Front Cell Infect Microbiol*., 4, 25 (http://dx.doi.org/10.3389/fcimb.2014.00025).

LIVERMORE, D.M., WAIN, J. 2013. Revolutionising bacteriology to improve treatment outcomes and antibiotic stewardship, *Infect Chemother*., 45, 1-10 (http://dx.doi.org/10.3947/ic.2013.45.1.1).

NIKOLAKI, S., TSIAMIS, G. 2013. Microbial diversity in the era of omic technologies. *Biomed Res Int*., 2013, 958719 (http://dx.doi.org/10.1155/2013/958719).

PADMANABHAN, R., MISHRA, AK., RAOULT, D., FOURNIER, PE. 2013. Genomics and metagenomics in medical microbiology, *J Microbiol Methods*., 95, 415-24 (http://dx.doi.org/10.1016/j.mimet.2013.10.006).

PERKINS, T.T., TAY, C.Y., THIRRIOT, F., MARSHALL, B. 2013. Choosing a benchtop sequencing machine to characterise *Helicobacter pylori* genomes. *PLoS One*, 8, e67539 (http://dx.doi.org/10.1371/journal.pone.0067539).

QUAIL, M.A., SMITH, M., COUPLAND, P., OTTO, T.D., HARRIS, S.R., CONNOR, T.R., BERTONI, A., SWERDLOW, H.P., GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13, 341 (http://dx.doi.org/10.1186/1471-2164-13-341).

REHVATHY, V., TAN, M.H., GUNALETCHUMY, S.P., THE, X., WANG, S., BAYBAYAN, P., SINGH S., ASHBY, M., KAAKOUSH, N.O., MITCHELL, H.M., CROFT, L.J., GOH, K.L., LOKE, M.F., VADIVELU, J. 2013. Multiple Genome Sequences of *Helicobacter pylori* Strains of Diverse Disease and Antibiotic Resistance Backgrounds from Malaysia, *Genome Announcements*, 1 (http://dx.doi.org/10.1128/genomeA.00687-13).

SCHMUKI, M.M., ERNE, D., LOESSNER, M.J., KLUMPP, J. 2012. Bacteriophage P70: unique morphology and unrelatedness to other *Listeria* bacteriophages, *Journal of Virology*, 86, 13099-13102 (http://dx.doi.org/10.1128/JVI.02350-12).

WANG, Y., YU, Y., PAN, B., HAO, P., LI, Y., SHAO, Z., XU, X., LI, X. 2012. Optimizing hybrid assembly of next-generation sequence data from *Enterococcus faecium*: a microbe with highly divergent genome, *BMC Systems Biology*, 6, Suppl 3, S21 (http://dx.doi.org/10.1186/1752-0509-6-S3-S21).

YANDELL, M., ENCE, D. 2012. A beginner's guide to eukaryotic genome annotation, *Nature Reviews Genetics*, 13, 329-342 (http://dx.doi.org/10.1038/nrg3174).

ZHANG, X., DAVENPORT, K.W., GU, W., DALIGAULT, H.E., MUNK, A.C., TASHIMA, H., REITENGA, K., GREEN, L.D., HAN, C.S. 2012. Improving genome assemblies by sequencing PCR products with PacBio, *Biotechniques*, 53, 61-62 (http://dx.doi.org/ 10.2144/0000113891).