Long-Read Sequencing in Clinical Bacterial Studies

Subjects: Agriculture, Dairy & Animal Science

Contributor: Mariem Ben khedher

The powerful combination of genome sequencing and bioinformatics analysis has played a crucial role in interpreting information encoded in bacterial genomes. High-throughput sequencing technologies have paved the way towards understanding an increasingly wide range of biological questions. This revolution has enabled advances in areas ranging from genome composition to how proteins interact with nucleic acids. This has created unprecedented opportunities through the integration of genomic data into clinics for the diagnosis of genetic traits associated with disease. Long-read sequencing has overcome previous limitations in terms of accuracy, thus expanding its applications in genomics, transcriptomics and metagenomics.

long-read sequencing bacterial genomes genomics

1. Long-Read Sequencing Developments

New sequencer machines appeared in 2011, proposing single-molecule sequencing technologies able to sequence over 10 kb of length. These long-read sequencings offer great advantages, including the ability to resolve repeats sequences [1].

Two technologies currently dominate the long-read sequencing space: ‘Pacific Biosciences’ (PacBio (Pacific Biosciences, Menlo Park, CA, USA)) single-molecule real-time (SMRT) sequences [2] and ‘Oxford Nanopore Technologies’ (ONT (Oxford Nanopore Technologies, Oxford, UK)) nanopore sequencing (Company history n.d.) which were commercially released in 2011 and 2014, respectively. The SMRT PacBio (Pacific Biosciences, Menlo Park, CA, USA) was the first long-read sequencer to be widely used. It is able to detect a single DNA molecule in real-time [3]. SMRT is based on DNA replication, utilizing the detection of released fluorophores as each nucleotide is added in the sequencing process. PacBio’s SMRT (Pacific Biosciences, Menlo Park, CA, USA) sequencing enables the real-time detection of nucleotide incorporation events during the elongation of the replicated strand from the non-amplified single-stranded template. The Nanopore from ONT (Oxford Nanopore Technologies, Oxford, UK) appeared later in 2014, and the MinIon (Oxford Nanopore Technologies, Oxford, UK) model was the first portable sequencer with a weight of only 100 g. The principle is based on a membrane including nanopores (transmembrane proteins), to which a low voltage is applied. The membrane detects the translocation signals, i.e., it acts as a nucleic acid counter by detecting the interruption to the current as they pass through the pore. Nanopore is less expensive than PacBio (Pacific Biosciences, Menlo Park, CA, USA). On the other hand, PacBio (Pacific Biosciences, Menlo Park, CA, USA) retains the advantage of better sequencing quality.

This third-generation sequencing has opened exciting avenues in genomics and has become suitable for an increasing number of applications. These capabilities have significantly improved accuracy and yield advances, making long-read sequencing key to a wide range of genomics applications for model and non-model organisms [4]. The advent of long-read technologies has the potential to transform clinical research and genomics analysis applications. An overview of the main advantages of long-read sequencing compared to short-read sequencing approaches are listed in **Table 1**.

Table 1. Summary of the main advantages of long-reads sequencing over short-read sequencing.

Short-Read Technologies	Long-Read Technologies
Fixed run time: - Increased time to results and inability to identify workflow errors before completed sequencing - Additional practical complexities associated with handling and storing large volumes of sequence data	Real-time data acquisition: - Achieve rapid turnaround with immediate access to results - Enrich single targets during sequencing, with no additional sample prep using adaptive sampling - Identify microbiome composition and resistance in real-time
Limited flexibility: - Sample batching often required for optimal efficiency - Potentially leads to long turnaround times - Benchtop devices confine sequencing to centralized locations	Scalable and flexible: - Scale to suit the throughput needs - Decentralize sequencing - No sample batching needed
Read length typically 50–300 bp	Unrestricted read length (>4 Mb achieved)
Limited genomic characterization: - Short reads do not span entire structural variants or important classes of genomic aberrations (repeat expansions and repeat-rich regions) - fragmented genome assemblies and ambiguous isoforms identification - Short sequencing reads may not span complex genomic regions such as genes duplications, transposons and prophage sequences - Potentially missing important genomic information	Comprehensive genomic characterization: - Identify mutations in complex and repetitive genomic regions - Accurately phase single nucleotide variants, structural variants, and base modifications - Can fully assemble genomes more easily - Simplify de novo assembly and correct microbial reference genomes - Possibility to completely assemble genomes and plasmids from metagenomic samples - Resolving complex genomic regions and similar species
Amplification required: - Amplification can introduce bias reducing uniformity of coverage and removes base modifications - Necessitating additional sample prep and sequencing runs	Amplification-free protocols: - Detect and phase base modifications as standard - No additional preparation required
Constrained to the lab: - Traditional sequencing technologies are typically expensive and require substantial site infrastructure	Sequence anywhere: - Sequence in your lab or in the field - Sequence at sample source and eliminate

Short-Read Technologies	Long-Read Technologies
- Usually limited its usage to well-resourced environments - Delay in transmitting the results	sample shipping delays - Scale-up with high-throughput

omes with finishing steps [5]. Loman et al. showed the feasibility of assembling a complete bacterial genome (*Escherichia coli* K-12 MG1655) in good quality using only long-reads produced by a MinION sequencer (Oxford Nanopore Technologies, Oxford, UK) [5] since long-read technology is now mainly used to obtain complete genomes.

Long-read technology also has other advantages. It improves the identification of transcription isoforms [6], the detection of structural variants [7], enables the direct detection of haplotypes and even whole chromosome phasing [8][9]. Finally, it makes it possible to sequence single molecules in real-time, avoiding DNA amplification which could be a bias inherent to second generation sequencing [10]. The ease of use of the Nanopore MinIon (Oxford Nanopore Technologies, Oxford, UK) has allowed sequencing to be performed with limited resource environments and in situ natural environments [11]. The machine also presents the opportunity to decentralize sequencing with fast run times, accurate performance and the ability to simply drop a sample onto the sequencer without any preparation. The consequences of this evolution towards long-read sequencing has given rise to numerous studies [12][13][14][15].

The affordability and usability of long-read single-molecule sequencing instruments has facilitated new real-time applications of disease outbreaks [16]. As shown by Joshua Quick and Nicholas Loman in 2015, they attempted to eradicate and stamp out the West Africa epidemic in Guinea and succeeded in the sequencing of Ebola viruses two days after sample collection [16][17]. Furthermore, Nanopore sequencing has already been applied for the rapid identification of microorganisms [18] and could be used for the detection of antibiotic-resistant pathogens such as *Salmonella* [19].

However, there are still some limitations to long-read technologies. They produce a higher rate of sequencing errors (5–20%) compared to other NGS data (<1%) [20], which are mostly randomly distributed. Nevertheless, long-read technologies are continuously improving, and the error rate is steadily decreasing with new machines. Moreover, bioinformatics algorithms have also evolved and now allow us to generate satisfactory read correction when the sequencing depth is high enough, reaching in some cases an accuracy over 99.9%. Aware of these limitations, the Oxford Nanopore company has refined resolution and throughput sequencing. For this purpose, several Oxford Nanopore products have been developed, including the GridION X5 (Oxford Nanopore Technologies, Oxford, UK) commercialized since March 2017 that can generate up to 100 GB of data per cycle. The PromethION (Oxford Nanopore Technologies, Oxford, UK), a high-throughput desktop device, contains channels for 144,000 nanopores (compared to 512 for the MinIon (Oxford Nanopore Technologies, Oxford, UK). Other platforms are in development, such as the SmidgION (Oxford Nanopore Technologies, Oxford, UK), a sequencer that can be connected to a smartphone and aims to make outdoor sequencing even more accessible.

2. Disruption of Clinical Studies on Prokaryotes

The democratization of high-throughput sequencing has made these techniques accessible to many clinical microbiology and public health laboratories. Due to the cost decrease, these structures are equipped with

genomics and sequencing platforms or collaborate with external providers. These new resources have changed the way by which hospitals or public health laboratories determine the agents involved in infectious diseases, in addition to the epidemiology and evolution of various infectious pathogens. The following sections describe the main clinical applications of NGS in clinical microbiology and their evolution.

2.1. Molecular Detection and Identification of Pathogens

Molecular markers or signatures are small nucleic acid fragments that are specific motifs to the genome of an organism. These signatures make it possible to determine the taxon to which the organism belongs, to predict a restriction profile, to find specific PCR primers or hybridization probes and to develop DNA arrays. The full sequencing of genomes has made it possible to move from a small choice of target sequences such as ribosomal subunits 16S, 23S or housekeeping genes (i.e., rpoB) to a wider choice of sequences, more specific to each biological question. For example, C.R Laing et al. analyzed the 4939 genome sequences of *Salmonella enterica* and identified 404 new subsp. markers in *S. enterica* subsp. [21]. They also identified 1610 universal markers along 10 serovars of *S. enterica* (Typhi, Typhimurium, Enteritidis, Heidelberg, Paratyphi, Kentucky, Agona, Weltevreden, Bareilly and Newport). These new signatures are intended to refine and improve the identification and diagnosis of *S. enterica* strains.

In recent years, the determination of new molecular markers has been facilitated by the massive use of WGS. This provided epidemiologists with a great tool to understand and predict the spread of bacterial species or to study the diversity of bacterial clones and their relationships. A wide genomic study of samples from various locations of a hospital revealed a reservoir of bacterial plasmids conferring carbapenem resistance [22]. The study is part of a large bacterial sequencing project at the Sanger Institute that widely use SMRT Pacific Biosciences (Pacific Biosciences, Menlo Park, CA, USA) technology, leading to sequencing and assembly of over 3000 complete bacterial genomes (from PHE's National Collection of Type Cultures (NCTC) https://www.phe-culturecollections.org.uk/collections/nctc-3000-project.aspx, accessed on 8 December 2021).

2.2. SNPs Genotyping

Genotyping is another strategy for molecular identification. Genotyping is the discipline that aims to determine the identity of a genetic variation for a given organism, at some specific positions, on the whole or only a part of its genome. Current methods of genotyping include restriction fragment length polymorphism identification (RFLPI) of genomic DNA, random amplified polymorphic detection (RAPD) of genomic DNA, amplified fragment length polymorphism detection (AFLPD), polymerase chain reaction (PCR), allele-specific oligonucleotide (ASO) probes, hybridization to DNA microarrays and more recently, DNA sequencing using NGS. The availability of complete genomes due to NGS has made new genotyping methods such as Microsatellites SSR (simple sequence repeats), SNP (Single Nucleotide Polymorphisms) or ISBP (Insertion Site-Based Polymorphisms) possible.

Genotyping by microsatellites SSR is now commonly used to classify isolates from one another. It consists of using tandem repeats in the genomes, called VNTRs (variable number tandem repeats). These repeats are amplified,

and the different sizes of the fragments obtained make it possible to determine to which strains they belong.

2.3. Phenotype Prediction to Track Virulence Factors and Antimicrobial Resistance

The current availability of a large number of genomes enables us to achieve a “genome wide association study” (GWAS). GWAS aims to identify significant associations between genetic traits and phenotypes. Regarding microbes, these GWAS studies generally focus on associations between nucleotide polymorphisms (SNPs) and phenotypes. Genome-based phenotypic prediction can relate to the detection of virulence factors. Herein then speak about “pathogenomics”. Understanding the genetic variations and mechanisms of infectious disease emergence and adaptation holds promise to improve disease prevention, intervention and to develop more targeted therapies [23].

The presence of a virulence factor does not necessarily imply that the bacterium will be pathogenic, and some bacteria may have one or more virulence genes in their genome without providing a pathogenic phenotype. This is illustrated by the study carried out by Armougom et al., which shows that the bacteria *Citrobacter Koseri*, despite possessing the *Pla* gene identical to that of *Yersinia pestis*, does not provide any particular pathogenicity [24]. The prediction of pathogenicity must take into account the whole genome, integrating the possible associations between virulence factors, the presence of other genes that may repress the virulence factors or the structure of the genome itself [25]. Phenotypic prediction can also be used to detect antimicrobial resistance (AMR). Therefore, predicting these resistances from the genomes can be an efficient tool to anticipate and propose treatments. Thus, the complete sequencing of genomes offers the possibility of accurately predicting the potential resistance of various strains [26].

2.4. Comparative Genomics to Understand bacterial Strains Evolution

The discovery of genetic variants underlying bacterial phenotypes and the prediction of phenotypic traits are fundamental tasks of bacterial genomics [27][28][29][30][31]. Thus, comparative genomics can be used for the prediction of specific microbial phenotypes for various clinical applications such as characterization of outbreaks, performing phylogeography allowing tracing and monitoring pathogen evolution or analysis of genomic diversity of strains. Comparative genomics corresponds to the comparison of biological information derived from whole-genome sequences and genome reconstructions. Comparative genomics therefore began in 1995, when the first two whole organism genomes, *Haemophilus influenzae* and *Mycoplasma genitalium*, were published [32][33].

Bioinformatics tools then appear that provide a way to compare the genome sequences themselves, RNAs, proteins, and gene annotations that can be derived from them. These tools are constantly evolving to deal with the exponential proliferation of sequenced genomes driven by advances in sequencing technology and to become more comprehensive and user-friendly. The use of comparative genomic approaches is reaching maturity. However, the use of short reads can limit the comparative genomics analysis for microbes. Genomes are rarely fully completed, and even if they are, some assembly uncertainties often remain, which leads to doubts about the final genome structure. This is particularly the case for large genomes, which often contain repeated regions (e.g.,

operons or repetitive mobile elements) that are difficult to assemble [34]. Furthermore, even if genomes are released as completed on public databases, the comparison of synteny rearrangements between closed species or comparisons of redundant regions are still problematic. Indeed, structural variations (SV) within the genomes play an important role and have to be assembled correctly. SV refers to chromosomal rearrangements typically classified as insertions, inversions, duplications, deletions and translocations describing resulting combinations of DNA losses or gains.

3. Conclusions

Today, it seems obvious that, whatever technology is imposed on the market, the future of sequencing will be turned towards long reads or even reads that can represent the entirety of a chromosome or a mobile element. In this case, it will no longer be necessary to facilitate assembly. Costs will also obviously continue to fall, making these new technologies more and more common. Sample preparation is simplified with each new generation, and already manufacturers such as Nanopore propose to simply place a sample on the sequencer chip. In addition, the automatisation of analysis methods is also developing rapidly. The biologist or clinician can quickly obtain an overview of the results in an intelligible way without needing bioinformatics skills. More advanced analyses requiring bioinformatics skills will still be necessary in some cases, especially for more fundamental projects or those requiring more investigation. However, routine clinical applications can often be satisfied with the results produced through in-line platforms to which the sequencers are connected. These cloud platforms integrate pipelines that automate data processing by software suites, and the results are graphically displayed and standardized.

Finally, similar to the first computers, sequencers have largely decreased in size and can, for some models, be transported directly to the field. Often associated with large computers such as computing clusters, it is now possible to perform routine analyses and real-time sequencing from a simple laptop computer equipped with a good video card. The quality and quantity of information produced by these machines will continue to increase, leading to a better understanding of the biological mechanisms governing the functioning of microorganisms.

References

1. Pollard, M.O.; Gurdasani, D.; Mentzer, A.J.; Porter, T.; Sandhu, M.S. Long Reads: Their Purpose and Place. Hum. Mol. Genet. 2018, 27, R234–R241.
2. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 2009, 323, 133–138.
3. Venkatesan, B.M.; Bashir, R. Nanopore sensors for nucleic acid analysis. Nat. Nanotechnol. 2011, 6, 615–624.

4. Karczewski, K.J.; Snyder, M.P. Integrative omics for health and disease. Nat. Rev. Genet. 2018, 19, 299–310.
5. Loman, N.J.; Quick, J.; Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods 2015, 12, 733–735.
6. Soneson, C.; Yao, Y.; Bratus-Neuenschwander, A.; Patrignani, A.; Robinson, M.D.; Hussain, S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat. Commun. 2019, 10, 1–14.
7. Begum, G.; Albanna, A.; Bankapur, A.; Nassir, N.; Tambi, R.; Berdiev, B.; Akter, H.; Karuvantevida, N.; Kellam, B.; Alhashmi, D.; et al. Long-Read Sequencing Improves the Detection of Structural Variations Impacting Complex Non-Coding Elements of the Genome. Int. J. Mol. Sci. 2021, 22, 2060.
8. Feng, Z.; Clemente, J.C.; Wong, B.; Schadt, E.E. Detecting and phasing minor single-nucleotide variants from long-read sequencing data. Nat. Commun. 2021, 12, 1–13.
9. Shafin, K.; Pesout, T.; Chang, P.-C.; Nattestad, M.; Kolesnikov, A.; Goel, S.; Baid, G.; Eizenga, J.M.; Miga, K.H.; Carnevali, P.; et al. Haplotype-Aware Variant Calling Enables High Accuracy in Nanopore Long-Reads Using Deep Neural Networks. bioRxiv 2021.
10. Rhee, M.; Burns, M.A. Nanopore sequencing technology: Research trends and applications. Trends Biotechnol. 2006, 24, 580–586.
11. Leggett, R.M.; Clark, M.D. A world of opportunities with nanopore sequencing. J. Exp. Bot. 2017, 68, 5419–5429.
12. Heather, J.M.; Chain, B. The sequence of sequencers: The history of sequencing DNA. Genomics 2015, 107, 1–8.
13. Ku, C.-S.; Roukos, D. From next-generation sequencing to nanopore sequencing technology: Paving the way to personalized genomic medicine. Expert Rev. Med. Devices 2013, 10, 1–6.
14. Loman, N.j.; Pallen, M.j. Twenty years of bacterial genome sequencing. Nat. Rev. Genet. 2015, 13, 787–794.
15. McGinn, S.; Gut, I.G. DNA sequencing—spanning the generations. New Biotechnol. 2013, 30, 366–372.
16. Quick, J.; Loman, N.J.; Duraffour, S.; Simpson, J.T.; Severi, E.; Cowley, L.; Bore, J.A.; Koundouno, R.; Dudas, G.; Mikhail, A.; et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 2016, 530, 228–232.
17. Hayden, E.C. Pint-sized DNA sequencer impresses first users. Nature 2015, 521, 15–16.

18. Karlsson, E.; Lärkeryd, A.; Sjödin, A.; Forsman, M.; Stenberg, P. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci. Rep. 2015, 5, 11996.
19. Judge, K.; Harris, S.R.; Reuter, S.; Parkhill, J.; Peacock, S.J. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. J. Antimicrob. Chemother. 2015, 70, 2775–2778.
20. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. Nat. Rev. Genet. 2016, 17, 333–351.
21. Laing, C.R.; Whiteside, M.D.; Gannon, V.P.J. Pan-genome Analyses of the Species Salmonella enterica, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar. Front. Microbiol. 2017, 8, 1345.
22. Weingarten, R.A.; Johnson, R.; Conlan, S.; Ramsburg, A.M.; Dekker, J.P.; Lau, A.F.; Khil, P.; Odom, R.T.; Deming, C.; Park, M.; et al. Genomic Analysis of Hospital Plumbing Reveals Diverse Reservoir of Bacterial Plasmids Conferring Carbapenem Resistance. mBio 2018, 9, e02011-17.
23. Slobounov, S.; Cao, C.; Jaiswal, N.; Newell, K.M. Neural basis of postural instability identified by VTC and EEG. Exp. Brain Res. 2009, 199, 1–16.
24. Armougom, F.; Bitam, I.; Croce, O.; Merhej, V.; Barassi, L.; Nguyen, T.-T.; La Scola, B.; Raoult, D. Genomic Insights into a New Citrobacter koseri Strain Revealed Gene Exchanges with the Virulence-Associated Yersinia pestis pPCP1 Plasmid. Front. Microbiol. 2016, 7, 340.
25. Yang, T.; Zhong, J.; Zhang, J.; Li, C.; Yu, X.; Xiao, J.; Jia, X.; Ding, N.; Ma, G.; Wang, G.; et al. Pan-Genomic Study of Mycobacterium tuberculosis Reflecting the Primary/Secondary Genes, Generality/Individuality, and the Interconversion Through Copy Number Variations. Front. Microbiol. 2018, 9, 1886.
26. Codoñer, F.M.; Pou, C.; Thielen, A.; García, F.; Delgado, R.; Dalmau, D.; Alvarez-Tejado, M.; Ruiz, L.; Clotet, B.; Paredes, R. Added Value of Deep Sequencing Relative to Population Sequencing in Heavily Pre-Treated HIV-1-Infected Subjects. PLoS ONE 2011, 6, e194612011.
27. Freddolino, P.L.; Goodarzi, H.; Tavazoie, S. Revealing the Genetic Basis of Natural Bacterial Phenotypic Divergence. J. Bacteriol. 2014, 196, 825–839.
28. Brbic, M.; Piškorec, M.; Vidulin, V.; Kriško, A.; Šmuc, T.; Supek, F. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. 2016, 44, 10074–10090.
29. Lees, J.A.; Tien Mai, T.; Galardini, M.; Wheeler, N.E.; Horsfield, S.T.; Parkhill, J.; Corander, J.; Lees, C.J.; Jacques Ravel, E.; Wilson, D. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. ASM J. 2020, 4, e01344-20.
30. Goberna, M.; Verdú, M. Predicting microbial traits with phylogenies. ISME J. 2015, 10, 959–967.

31. Weimann, A.; Mooren, K.; Frank, J.; Pope, P.B.; Bremges, A.; McHardy, A.C. From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems 2016, 1, e00101-16.
32. Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.F.; Dougherty, B.A.; Merrick, J.M.; et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995, 269, 496–512.
33. Fraser, C.M.; Gocayne, J.D.; White, O.; Adams, M.D.; Clayton, R.A.; Fleischmann, R.D.; Bult, C.J.; Kerlavage, A.R.; Sutton, G.; Kelley, J.M.; et al. The Minimal Gene Complement of Mycoplasma genitalium. Science 1995, 270, 397–404.
34. Schmid, M.; Frei, D.; Patrignani, A.; Schlapbach, R.; Frey, J.E.; Remus-Emsermann, M.; Ahrens, C.H. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. Nucleic Acids Res. 2018, 46, 8953–8965.