sciendo

DOI: 10.2478/fv-2020-0019

FOLIA VETERINARIA, 64, 2: 66—73, 2020

HISTORY OF DNA SEQUENCING

Tyagi, P.[1], Bhide, M.[1, 2]

[1]Laboratory of Biomedical Microbiology and Immunology
University of Veterinary Medicine and Pharmacy in Košice, Komenskeho 73, Košice
[2]Institute of Neuroimmunology, Slovak Academy of Sciences, Dubravska cesta 9, Bratislava
Slovakia

bhidemangesh@gmail.com

ABSTRACT

The nucleotides are the building blocks of nucleic acids and determining their sequential arrangement had always been an integral part of biological research. Since the past seven decades, researchers from multidisciplinary fields has been working together to innovate the best sequencing methods. Various methods had been proposed, from some oligonucleotides to the whole genome sequencing, and the growth had gone through adolescence to the mature phase where it is now capable of sequencing the whole genome at a low cost and within a short time frame. DNA sequencing has become a key technology in every discipline of biology and medicine. This review aims to highlight the evolution of DNA sequencing techniques and the machines used, including their principles and key achievements.

Key words: DNA; NGS; RNA; sequencing; Sequencing machines

INTRODUCTION

We are witnessing the emergence of the cutting edge techniques in biological research that have gained a significant place in reducing the time and cost to obtain biological knowledge. Decades ago, the sequencing technology was time-consuming, labour-intensive, and relied on analytical chemistry. In 1976, the first single-stranded genome of the bacteriophage ØX174 was sequenced using a plus-minus sequencing method. Since then, several genomes (including the human genome) were drafted using the Sanger sequencing method which requires an enormous investment of time and cost [17, 19]. The most popular and highly used technique for DNA sequencing was established by Fred Sanger and it was referred to as the chain termination or dideoxy method. The Sanger method of sequencing has been characterized as the first-generation of sequencing. This phase of development can be considered as an adolescence period, when the human genome project was completed in 2003, whereas from 2007 onwards, the maturing phase was beginning.

There were continuous improvements in the sequencing methods which further led to high throughput sequencing which was collectively called as the next generation of sequencing. During this developmental journey, commercial platforms (such as ABI, Solexa, Ion Torrent, Illumina, Oxford Nanopore, etc.) with different sequencing strategies and concepts were developed, while the common sequencing steps remained conserved such as: the template preparation, clonal amplification and a cyclic round of massively parallel sequencing [17]. These sequencing platforms were capable of producing a huge amount of biological data in less time and money. These developments have opened new perspectives in the area of genomics, transcriptomics and metagenomics.

First-generation sequencing

The conceptual base for the replication and protein encoding by the nucleic acids was supported by the groundbreaking discovery of the three-dimensional structure of DNA by Watson and Crick [25] using photograph 51, produced by the Rosalind Franklin and Maurice Wilkins [26]. Still, the order of four nucleotides was unapproachable as the DNA molecule is longer and composed of only four nucleotide bases that made it difficult to sequence [6]. An initial study in the field of rapid sequencing was carried out in 1970 by Ray Wu. Then in early 1975, the first complete genome was sequenced at the RNA level; this involved the RNA bacteriophage MS2 [5, 19]. Primarily, the focus was on the pure species of RNA such as: transfer RNA, ribosomal RNA and the genome of single-stranded RNA bacteriophages; this was because those are abundant in cell culture, they are shorter, and not complicated with the complementary strands.

In the process of identification of small hypothetical DNA sequence, the first step was polymerization and elongation of DNA sequence using an already known short nucleotide (decamer). The radiolabeled complementary strands of unknown DNA template of various lengths were formed by incorporating four deoxynucleotide triphosphates (dNTPs), in which one was radiolabelled (32P). The second step was the removal of excess triphosphate using an agarose column and this mixture was further used for the minus and plus method [20].

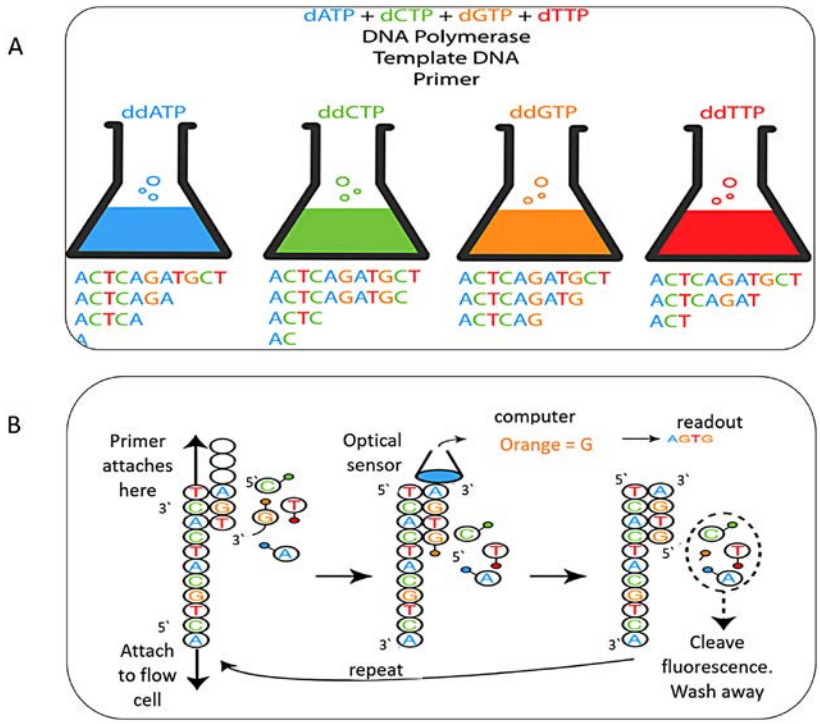

Fig. 1. Pictures briefing the chemistry and techniques used in the evolution of first and second generation DNA sequencing
(A)—The Sanger chain termination methods using types of ddNTPs with DNA polymerase in four separate reactions to infer the DNA sequence (Image Source- Snipcadmy.com); (B)—Illumina sequencing by synthesis techniques, each freely available nucleotide added recognized by the optical sensor and connected to a computer to readout the nucleotide pattern (Image source—Medium.com)

In the minus method, the random mixture of radiolabeled complementary strands was incubated with polymerase I in the presence of three deoxyribose-triphosphates (dNTPs), whenever in the synthesis of DNA, a triphosphate missed, chain terminated at 3' end before that specific residue. The four incubation mixture was synthesized by missing one triphosphate among the four each time and further denatured and subjected to gel electrophoresis for molecular size-based separation. In the case of the plus method, the above obtained random mixture was incubated with only a single type of deoxynucleotide triphosphates and T4 DNA polymerase, because of T4 polymerase exonuclease activity, all extension ended with that triphosphate present. Further, the radioautograph produced from the plus and minus method was used to infer the positions of the nucleotides in the hypothetical DNA sequence [20]. However, this technique was limited to approximately 50—100 nucleotides that consisted of small stretches of DNA and involved lots of analytical chemistry and fractionation steps [7].

After the development of the plus and minus methods, the first rapid sequencing was developed by Gilbert and Sanger with chemical cleavages and chain termination, respectively [11, 19]. In the chain termination method, the DNA monomeric unit (deoxyribonucleic acid mimicked by the chemical analogue di-deoxyribonucleotide (ddNTPs) that lacked a hydroxyl group at 3' prime) which was required for the extension of the DNA chain, resulted in the hindrance of the bond formed between the 5' phosphate of the next dNTP [19]. Whereas, an alternative method (the Maxam and Gilbert chemical cleavage method) involved complex chemistry, in which the double-stranded or single-stranded DNA was first digested with the restriction enzymes and then the end-labelled with 32P phosphate subjected to the random cleavage at adenine (A), cytosine (C), guanine (G) or thymine (T) positions using specific chemical agents [11]. The products of these four reactions were then separated using polyacrylamide gel electrophoresis to inferred the DNA sequence. Due to the benefits of the less toxic chemicals and less complex procedure, the Sanger method was further adopted and modified by the replacement of phospho-radiolabelling with the fluorometric based detection and capillary-based electrophoresis technique to increase the capability.

Gradual refinement contributed to the development of the first-generation automated sequencing machine, however, these machines were only capable of producing reads of less than 1 kilobase length [6]. To address this limitation and to sequence longer fragments of DNA, researchers came up with the shotgun strategy, in which the overlapping regions in the genome were fragmented, cloned using a vector and then sequenced separately and reassembled using computational tools [16]. The overall advances in the sequencing technology led to the enablement to draft the first human genome sequence on 14 April 2003 using the Sanger chain termination method (Fig. 1). The automated DNA sequencer ABI prism 3700 with 96 capillaries was used for this genome sequencing [16, 24].

Second-generation sequencing

The first-generation sequencing (mainly the Sanger sequencing technique) continued to dominate the sequencing market for approximately two decades; however, the researchers were in search of a better alternative technique that would have lower cost, higher throughput and capable of massively parallel sequencing (Table 1). Rather than a chain termination method, a new sequencing method had evolved based on the production of light, whenever a nucleotide was incorporated the release of a pyrophosphate occurred, hence it was called pyrosequencing. During the synthesis of DNA, nucleotides were incorporated by the polymerase enzyme and each incorporation released a pyrophosphate. This pyrophosphate was then, in the presence of ATP sulfurylase and adenylyl sulfate, converted to ATP and then this ATP was used as a substrate for luciferase to produce light that was proportional to the amount of pyrophosphate. In this method at a time, one type of dNTPs was added and if complementary nucleotide found on the unknown DNA template, the light emits. This process further continued with washing and adding different dNTPs to inferred the DNA sequence in real-time [16]. The pyrosequencing technique came into existence in 1993 and was commercialized in 1997 by a company (Pyrosequencing AB) owned by Pål Nyrén and colleagues [15]. This method was not dependable upon electrophoresis and fragment separation, hence it was more rapid than the chain termination. The second-generation sequencing can also be termed as a short-read sequencing approach and can be broadly divided into Sequencing-by-ligation, Sequencing-by-synthesis and Ion semiconductor sequencing [8].

In 1998, Pål Nyrén and colleagues used one more en-

Table 1. Showing the DNA sequencing companies with the details of their platforms and their pros and cons

Company	Sequencing Principle	System platform	Read length and accuracy	Pros	Cons
Illumina	Reversible terminator sequencing by synthesis	HiSeq 2500/1500	36/50/100 SE and > 99 %	Very high throughput, cost-effective, steadily improving read length	Long run time, short read length
		MiSeq	35/50/75/100 SE > 99 %	Short run time, cost-effective, high coverage	Short read length
Roche	Pyrosequencing	454 GS FLX+	1 Million, 99.97 %	Longer reads, high throughput, high coverage	High reagent cost, the higher error rate in homopolymers region
Helicos Biosciences	Single-molecule sequencing	HeliScope	25—55 (average—32) 99.99 %	Non-bias representation of a template for genome	Expensive instrument, very short read length
ABI Life technologies	Ligation	5500 SOLID	75+35 99.99 %	Low reagent cost and high throughput	Long run time and very short reads
	Proton detection	Ion Personal Genome Machine(PGM)	35/200/400 > 99 %	Short run time, low cost/sample	High reagent cost, the high error rate in homopolymers
Pacific Biosciences	The real-time single-molecule DNA sequencing	PacBioRS	Average 3000 84—85 %	Short runtime, very long read length, low reagent cost	No paired reads, the high error rate
Oxford nanopore	Nanopore exonuclease sequencing	gridION	Tens of Kilobytes 96 %	Extremely long reads, no fluorescent labelling and no optics	4 % error rate, difficult to fabricate a device with multiple parallel pores
	Nanopore sequencing	MinION	Up to 1 Megabyte, 99 %	Longest read length, portable, affordable	High cost/Megabytes, No protocol yet

zyme called apyrase to remove the nucleotides that were not incorporated by the DNA polymerase, hence they established the automated setup for Pyrosequencing. Using the principal of pyrosequencing, 454 pyrosequencing method attached the DNA that was to be sequenced to the solid phase fibre-optic slides which consisted of millions of wells and each well was capable for the separate enzymatic reactions; this achievement boosted the rapid growth in parallel sequencing techniques [18, 21]. One of which was Solexa/Illumina sequencing platform that includes: DNA fragmentation, adapter ligation (library preparation), fixation at the flow cell, clustering (PCR- bridge amplification), adding four types (A/T/G/C) of fluorescently labelled reversible terminating nucleotides and the clusters were exited using laser and signals were detected using coupled-charge diode (CCD) [17]. The advantage of this sequencing method was that it was capable to perform paired-end sequencing that increases the accuracy of the information and helps in mapping reads to the reference genome, later this technique was acquired by Illumina.

The growing field of DNA sequencing witnessed another technique called Ion Torrent (Thermo Fisher Scientific) that utilized a semiconductor sequencing technology. In which the hydrogen ions were released whenever the nucleotide was incorporated in a single strand of DNA shifts the pH of the surrounding solution during the polymerization of DNA and these changes were detected by the sensor on the bottom of each well [22]. Each nucleotide was added with the washing cycle and according to the change in voltage, the sequence of the nucleotide was recorded. Another approach that was used commercially in ABI/SOLiD (Supported oligonucleotides ligation and detection) was the sequencing-by-ligation, not sequencing-by-synthesis, that consists of the attaching an adapter to the DNA fragment, one fragment-one bead complex formation and cloned by PCR emulsion, further processed with purification and

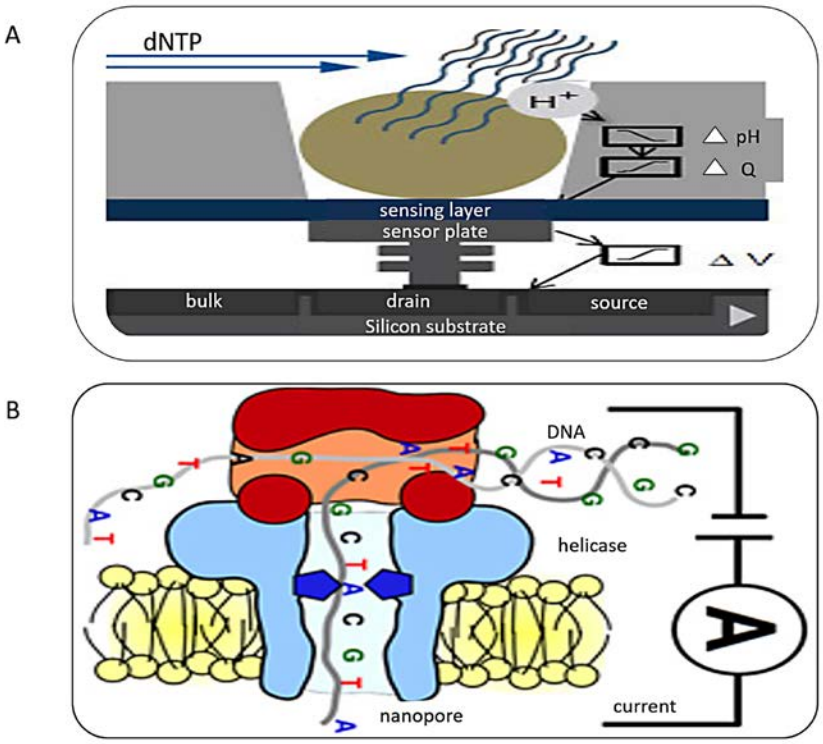

Fig. 2. Pictures briefing the chemistry and techniques used in the evolution of third or next generation DNA sequencing
(A)—Ion sequencing protons were released when growing DNA strands were incorporated by dNTP and change in pH in the well detected by the sensor and recorded as a nucleotide (Image source—en.genomics.cn); (B)—Oxford nanopore technique, nanopore incorporated into phospholipids bilayer and electric potential opposite side help DNA in translocation through the nanopore because of negative charge on it and membrane and ionic current is partially blocked to differentiate four nucleotides (Image source—author Steinbock, L. J., & Radenovic, A. (2015)

immobilization of the beads on a glass slide [8, 10] (Figure 2). The shortcoming of this method was the data analysis as the read length and depth was not the same as Illumina and created a problem in the assembly preparation [2].

Third-generation sequencing

The second-generation sequencing approach was incapable of handling: repetitive regions, to produce long reads, to recognize thousands of novel isoforms and gene fusion, therefore, more advanced and improved sequencing techniques were required [9] (Fig. 3). The third-generation techniques mainly focused on the single-molecule sequencing (SMS) first developed in the lab of Stephen Quake and the single-molecule real-time sequencing (SMRT) approach. Contrasting the SMS technology worked the same as the Illumina technique but without bridge amplification and this technique was relatively slow and expensive [14]. The basis of SMRT was the recognition of signals discharged using an array detection charge-coupled diode (CDD) in real-time, when they were incorporated, although, these two main techniques used DNA-polymerase and the terminal-phosphate-labelled nucleotide that allowed for sequencing long read length and short runtimes [1, 13].

The third-generation techniques were successful in the meantime, as they were capable of producing a huge amount of data at low cost and with less time as compared to the first generation sequencing [12]. During this phase, the setup of the sequencing machine reduced from giant size sequencing machine to a small cell phone size MinION sequencer (3rd Generation) and SmidION even smaller than a MinION (Figure 4).

The small size sequencer was designed in such a way that it can be connected into the laptop using the USB port and can be controlled by a Smartphone [23]. Among the seven types of nanopore sequencer, MinION and GridION worked with a biological nanopore in which the negatively charged DNA translocates through the nanopore placed into the phospholipids bilayer and when the positive electric potential introduced to the opposite side of the membrane translocation occurred and to allow the detection of four different nucleotides, the ionic current was partially blocked, leading to a reduction in the current, hence, DNA

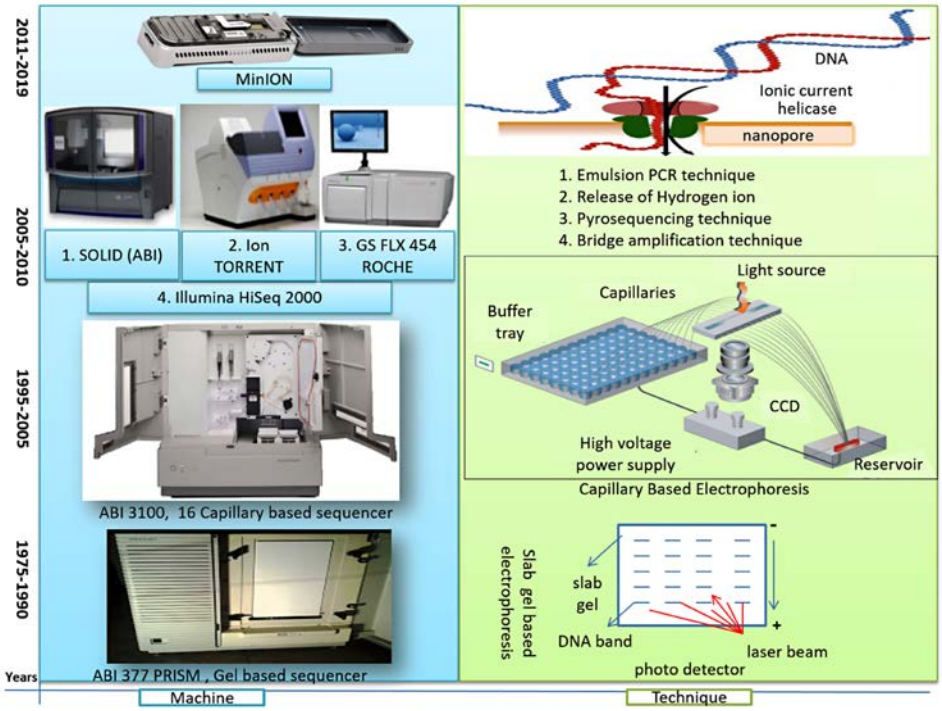

Fig. 3. Scheme depicting the evolution of sequencing machines and techniques

The development from bottom to top demonstrates the advancement in DNA sequencing and from giant size machine to small size sequencer. The Sanger chain termination methods using types of ddNTPs in 1975—2005; later, Illumina sequencing by synthesis techniques, Ion sequencing, pyrosequencing, sequencing by ligation, and the latest Oxford nanopore technique. The sources of images were from google images search and their respective official website

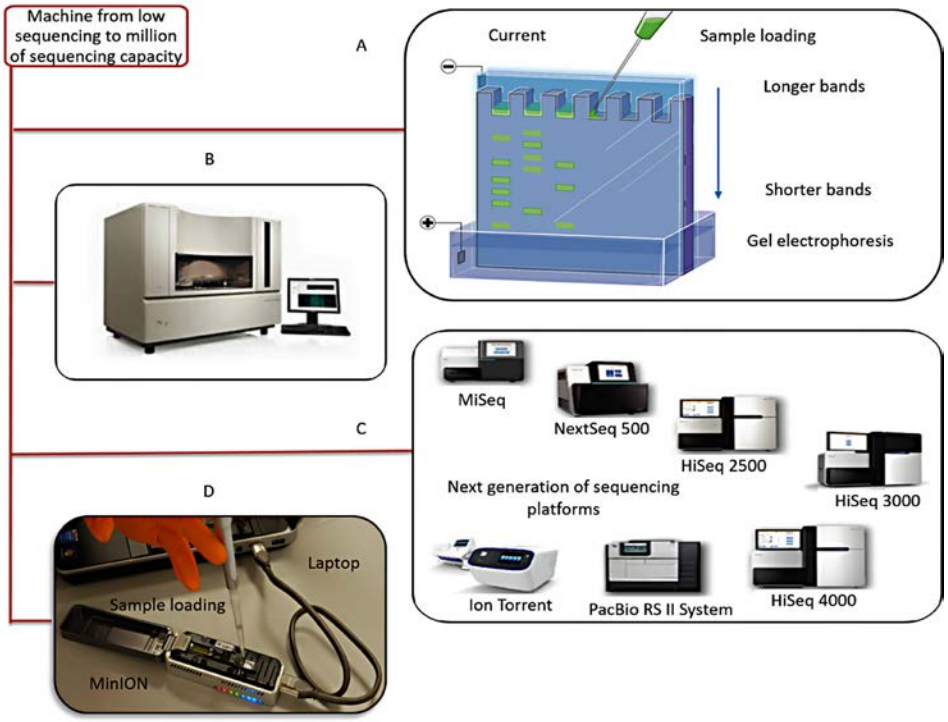

Fig. 4. Pictures explaining the evolution of DNA sequencing capacity from A low to D high

(A)—Slab gel-based DNA sequencing platform to separate labelled nucleotide based on their size (Image source—Smart.servier.com); (B)—First-generation automated Sanger sequencing machine with capillary method ABI 3730 sequencer; (C)—Advancement and second-generation Illumina sequencers with different in optic power and output capacity (Image source—base-asia.com); (D)—Third generation portable DNA sequencing instrument of Oxford Nanopore, MinION (Image source—Science-practise.com). Source: An original drawing with images sources from their respective origins

sequence inferred [22]. Soon, there was also a possibility to use non-biological, solid-state technology to design hybrid nanopores that might be capable of sequencing double-stranded DNA molecules [6].

We were aware that every technology had advantages and disadvantages; the potential drawbacks as compared to the second generation sequencing was the error rate that was over 90 % when the analysis was done using MinION on the lambda phage genome and amplicon of snake venom gland transcriptome [4]. Although, the advantage of compact size and compatibility to carry anywhere, Joshua Quick and Nicholas Loman successfully sequenced the Ebola virus genome in just 48 hours after the sample collection [3]. Nanopore technology was still in its initial phase of development in terms of accuracy and it will take time to be used in a wide range of application with high specificity.

CONCLUSIONS

It was desirable to achieve the highest possible accuracy in the field of sequencing because multiple factors can impact the final biological results. Therefore, there were scientists from the multidisciplinary fields worked together to refine these techniques. Over the last five-decades sequencing had progressed through sequencing the limited reads of pure RNA species to the whole set of eukaryotic genomes which was supported by the advancements in molecular biology, analytical chemistry and laser-induced fluoresce detection methods. Using DNA sequencing technology we are now able to understand the fundamental level properties that help to differentiate life. The overall evolution of sequencing techniques from using radioactive isotopes to changes in ionic current for the detection of nucleotide pattern and sequence opens up the possibility to sequence highly complex genomes with low cost, time and effort. The strength of DNA sequencing was that it can be applied to various omics and molecular diagnostic studies. It was reasonable that the future challenges will be aimed at achieving connectivity between data generated from the massively parallel sequencing and making repositories that can further help the researcher to get deeper biological insight. We believe that understanding the rich history of sequencing will establish a foundation for new sequencing techniques, as learning from previous factors leads to further progress.

ACKNOWLEDGEMENTS

Punit Tyagi was supported by The European Union's Horizon 2020 research and innovation programme H2020-MSCA-ITN-2017-EJD: Marie Skłodowska-Curie Innovative Training Networks (European Joint Doctorate)—Grant agreement No. 765423—Molecular Animal Nutrition (MANNA) and Mangesh Bhide was supported by APVV-18-0259.

REFERENCES

1. **Ambardar, S., Gupta, R., Trakroo, D., et al., 2016:** High throughput sequencing: An overview of sequencing chemistry. *Indian J. Microbiol.*, 56, 4, 394—404. DOI: 10.1007/s12088-016-0606-4.
2. **Buermans, H. P., den Dunnen, J. T., 2014:** Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta*, 1842, 10, 1932—1941. DOI: 10.1016/j.bbadis.2014.06.015.
3. **Check Hayden, E., 2015:** Pint-sized DNA sequencer impresses first users. *Nature*, 521, 7550, 15—16. DOI: 10.1038/521015a.
4. **Feng, Y., Zhang, Y., Ying, C., et al., 2015:** Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics and Bioinformatics*, 13, 1, 4—16. DOI: 10.1016/j.gpb.2015.01.009.
5. **Fiers, W., Contreras, R., Duerinck, F., et al., 1976:** Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260, 5551, 500—507. DOI: 10.1038/260500a0.
6. **Heather, J. M., Chain, B., 2016:** The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107, 1, 1—8. DOI: 10.1016/j.ygeno.2015.11.003.
7. **Hutchison, C. A., 3rd, 2007:** DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.*, 35, 18, 6227—6237. DOI: 10. 1093/nar/gkm688.
8. **Kchouk, M., Gibrat, J. F., Elloumi, M., 2017:** Generations of sequencing technologies : From first to next generation. B*iol. Med. (Aligarh)*, 9, 3, DOI: 10.4172/0974-8369.1000395.
9. **Lee, H., Gurtowski, J., Yoo, S., et al., 2016: T**hird-generation sequencing and the future of genomics. *BioRxiv.* DOI: 10.1101/048603.
10. **Liu, L., Li, Y., Li, S., et al., 2012:** Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, 2012, 251364. DOI: 10.1155/2012/251364.

11. **Maxam, A. M., Gilbert, W., 1977:** A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, 74, 2, 560—564. DOI: 10.1073/pnas.74.2.560.
12. **Metzker, M. L., 2010:** Sequencing technologies—the next generation. *Nat. Rev. Genet.*, 11, 1, 31—46. DOI: 10.1038/nrg2626.
13. **Munroe, D. J., Harris, T. J., 2010:** Third-generation sequencing fireworks at Marco Island. *Nat. Biotechnol.*, 28, 5, 426—428. DOI: 10.1038/nbt0510-426.
14. **Niedringhaus, T. P., Milanova, D., Kerby, M. B., et al., 2011:** Landscape of next-generation sequencing technologies. *Anal. Chem.*, 83, 12, 4327—4341. DOI: 10.1021/ac2010857.
15. **Nyren, P., 2007:** The history of pyrosequencing. *Methods Mol. Biol.*, 373, 1—14. DOI: 10.1385/1-59745-377-3:1.
16. **Pareek, C. S., Smoczynski, R., Tretyn, A., 2011:** Sequencing technologies and genome sequencing. *J. Appl. Genet.*, 52, 4, 413—435. DOI: 10.1007/s13353-011-0057-x.
17. **Reuter, J. A., Spacek, D. V., Snyder, M. P., 2015:** High-throughput sequencing technologies. *Mol. Cell*, 58, 4, 586—597. DOI: 10.1016/j.molcel.2015.05.004.
18. **Rothberg, J. M., Leamon, J. H., 2008:** The development and impact of 454 sequencing. *Nat. Biotechnol.*, 26, 10, 1117—1124. DOI: 10.1038/nbt1485.
19. **Sanger, F., Air, G. M., Barrell, B. G., et al., 1977:** Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265, 5596, 687—695. DOI: 10.1038/265687a0.
20. **Sanger, F., Coulson, A. R., 1975:** A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94, 3, 441—448. DOI: 10.1016/0022-2836(75)90213-2.
21. **Siqueira, J. F., Jr., Fouad, A. F., Rocas, I. N., 2012:** Pyrosequencing as a tool for better understanding of human microbiomes. *J. Oral Microbiol.*, 2012, 4. DOI: 10.3402/jom.v4i0.10743.
22. **Steinbock, L. J., Radenovic, A., 2015:** The emergence of nanopores in next-generation sequencing. *Nanotechnology*, 26, 7, 074003. DOI: 10.1088/0957-4484/26/7/074003.
23. **van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., et al., 2018:** The third revolution in sequencing technology. *Trends in Genetics: TIG*, 34, 9, 666—681. DOI: 10.1016/j.tig.2018.05.008.
24. **Venter, J. C., Adams, M. D., Myers, E. W., et al., 2001:** The sequence of the human genome. *Science*, 291, 5507, 1304—1351. DOI: 10.1126/science.1058040.
25. **Watson, J. D., Crick, F. H., 1953:** Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171, 4356, 737—738. DOI: 10.1038/171737a0.
26. **Zallen, D. T., 2003:** Despite Franklin's work, Wilkins earned his Nobel. *Nature*, 425, 6953, 15. DOI: 10.1038/425015b.

Received May 5, 2020
Accepted June 6, 2020