Beginner's guide to next-generation sequencing

Louise Aigrain
(Wellcome Sanger Institute, UK)

Since the publication of the first draft of the human genome 20 years ago, several novel sequencing technologies have emerged. Whilst some drive the cost of DNA sequencing down, others address the difficult parts of the genome which remained inaccessible so far. But the next-generation sequencing (NGS) landscape is a fast-changing environment and one can easily get lost between second- and third- generation sequencers, or the pros and cons of short- versus long-read technologies. In this beginner's guide to NGS, we will review the main NGS technologies available in 2021. We will compare sample preparation protocols and sequencing methods, highlighting the requirements and advantages of each technology.

40 years of sequencing

Over four decades ago, Frederick Sanger and his colleagues developed a method to decode the genetic information stored in DNA. Albeit labour intensive, the Sanger sequencing method still remains today's gold standard in terms of accuracy. Most importantly, it opened the door to a completely new world of possibilities. Researchers were suddenly able to access the book holding all the basic instructions for cellular life. The enthusiasm at the time was enough to spark one of the largest international projects in biology so far, the Human Genome Project. Launched in 1990, its cost estimates to $2.7bn over 13 years spanning between three continents and 20 collaborating sites.

Since the publication of the first draft of the human genome 20 years ago, the cost and turnaround time to sequence one human genome, or even several in parallel, dropped extraordinarily, surpassing all predictions. Such rapid development was initiated in the early 2000 s by virtue of two main factors: a new generation of high-throughput sequencing technologies referred to as second-generation sequencing and the simultaneous rapid development of computing power (Table 1). Second-generation sequencing benefited from a straightforward *in vitro* sample preparation and a high parallelization of the sequencing process. Among them Illumina established itself as the main actor on the sequencing market with its sequencing by synthesis (SBS) method and '$1,000 genome'.

From 2010 onward, a third generation of sequencers emerged led by PacBio and Oxford Nanopore Technology (ONT), both using single-molecule sequencing and allowing decoding of much longer stretches of DNA fragments, also called long reads. Initially more expensive and less accurate than previous sequencing technologies, the benefits of long reads lie in their name; as a puzzle with a few large pieces is much easier to solve than one with many small ones, assembling a genome is much easier with long-read data.

Words like 'genomics', 'variant' and 'sequencing' are routinely used in mainstream media reporting on exciting breakthroughs or worrying mutations, but technical aspects such as how the data is being produced are often overlooked. In this beginner's guide to next-generation sequencing (NGS), we will present and compare the main sequencing technologies available in 2021. We will not attempt to build an exhaustive image of all sequencers currently available, but rather focus on the most commonly used; nor will this article cover the data analysis aspect of NGS as it would require a beginner's guide to NGS bioinformatics of its own.

How it works

Sample preparation and considerations

Prior to its sequencing, DNA or RNA samples undergo a few laboratory steps called library preparation. Several protocols exist depending on the nature of the sample, its preservation, the amount of material available and of course the specific application for which the sequencing data is being generated. However, all sequencing technologies share in common the requirement to attach specific synthetized oligonucleotides, also called adapters, to both ends of the RNA or DNA fragments of interest (Figure 1). These adaptors will be recognized by the sequencer allowing the sequencing to take place.

Shearing. Prior to any adaptor ligation, the DNA sample often requires cutting, or shearing, to the appropriate insert length dictated ultimately by the sequencing technology used (short vs long reads,

Summary box

- Prior to sequencing, DNA or RNA samples are prepared following a library preparation protocol. With any sequencing platform, synthetic adaptors are required to be ligated to the DNA or RNA fragments of interest. These adaptors are recognized by the sequencing platform, allowing sequencing to start.
- Other optional steps such as barcoding, amplification or capture allow the user to focus on specific regions of interest in a genome and/or to lower sequencing costs.
- Short-read technologies can produce sequencing data from very low DNA amounts, in a high throughput manner and at the lowest cost available so far.
- Long-read technologies are more expensive but they can cover regions of a genome inaccessible to short-read technologies such as long repetitive regions, or sequence directly RNA or modified DNA.
- Whilst Illumina and PacBio technologies use fluorescence microscopy to detect the incorporation of fluorescently labelled nucleotides by a polymerase amplifying the DNA fragment of interest, Oxford Nanopore Technology measures the impedance across a membrane containing nanopore proteins through which DNA is translocating.

Figure 2). Illumina's SBS technology relies on the incorporation and imaging of one fluorescent base at a time. But the level of noise in the fluorescent signal accumulates with each SBS cycle, limiting Illumina's read length to a maximum of around 600 bp. Other short-read sequencing technologies also recommend fragment sizes typically under 1 kbp. With long-read technologies, genomic DNA might still be sheared but to much larger insert sizes to ensure optimal yield whilst preserving the longest reads possible. The optimal balance between overall data yield and maximal read length is generally obtained with inserts ranging from 10 to 20 kbp for PacBio instruments or around 30–50 kbp for ONT. However, both PacBio and ONT also offer the option of sequencing intact genomic DNA, and ONT users have reported reads longer than 2.3 Mbp.

DNA repair and adaptor ligation. Once DNA fragments are of the required size, adaptors are ligated to their ends (Figure 2). To improve the yield of this notoriously inefficient step, DNA fragments are first repaired into blunt-ended and 5′-phosphorylated fragments. Many protocols also recommend the addition of a single dATP to the 3′-end of the DNA fragments (A-tailing) to complement T-tailed adaptors and improve the ligation efficiency. During the last decade, rapid library preparation protocols using a transposase-based approach have been developed. These transposases, when added to the original DNA sample, are able to cut and ligate pre-loaded adaptors in a single step.

PCR-free vs amplification. Our adaptor-ligated DNA fragments are now ready for sequencing. Since they haven't been subjected to any amplification, this type of library is called PCR-free library and will produce high-quality sequencing data with minimal amplification bias. But in many applications such as single-cell studies or rare samples, the available DNA input is very low, much lower than 50 ng, and PCR-free library yields fail to reach the concentration required to load a sequencer. In these cases, an amplification step is required. The number of PCR cycles can be adjusted depending on the input material in order to minimize intrinsic amplification errors.

Barcoding and targeted sequencing. Depending on the output of the sequencing kit used and the size of the

Table 1. Different generations of DNA sequencing technologies developed over the past 50 years

	1977	1990s	2000s	2010s	2020s
Sequencing technology generation	*First-generation sequencing*		*Next-generation sequencing (NGS)*		*Future generation*
			Second-generation sequencing	Third-generation sequencing	
Sequencing technologies	► Sanger sequencing (manual)	► Sanger sequencing (automated)	► Illumina ► Roche ► Ion Torrent	► PacBio ► ONT ► …	► Genapsys ► MGI ► …
Technology breakthrough	Gel based	Capillary based	High throughput	Long reads	

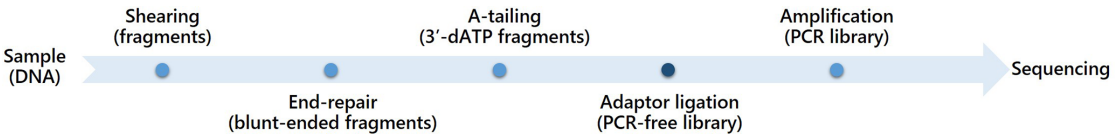

Figure 1. Protocol steps for preparing DNA library. Whilst protocols may vary, all of them include an adaptor ligation step. Adaptors are short synthetic DNA fragments which are recognized by the sequencers and allow the sequencing process to start.

genome, transcriptome or targeted region of interest, several libraries can be pooled and sequenced together to minimize cost. To ensure that each sequencing read can be linked to its original sample during data analysis, specific barcodes are added to each library. These barcodes are known synthetic sequences of DNA which can be incorporated either within the adaptors used during the ligation step or within the primers used during PCR amplification.

Another way to reduce sequencing cost is to focus on the exome, the coding regions of the genome. The human exome constitutes only 1% of the whole genome but is estimated to contain 85% of disease-causing mutations. Using synthetic nucleic acid probes complementary to exon regions, the DNA library fragments containing the exon sequence can be 'fished out' enabling a targeted approach.

RNA sample and nucleic modifications. RNA samples, such as transcriptomes or viral RNA, can of course be sequenced too. Most sequencing technologies require the RNA samples to be transcribed into DNA

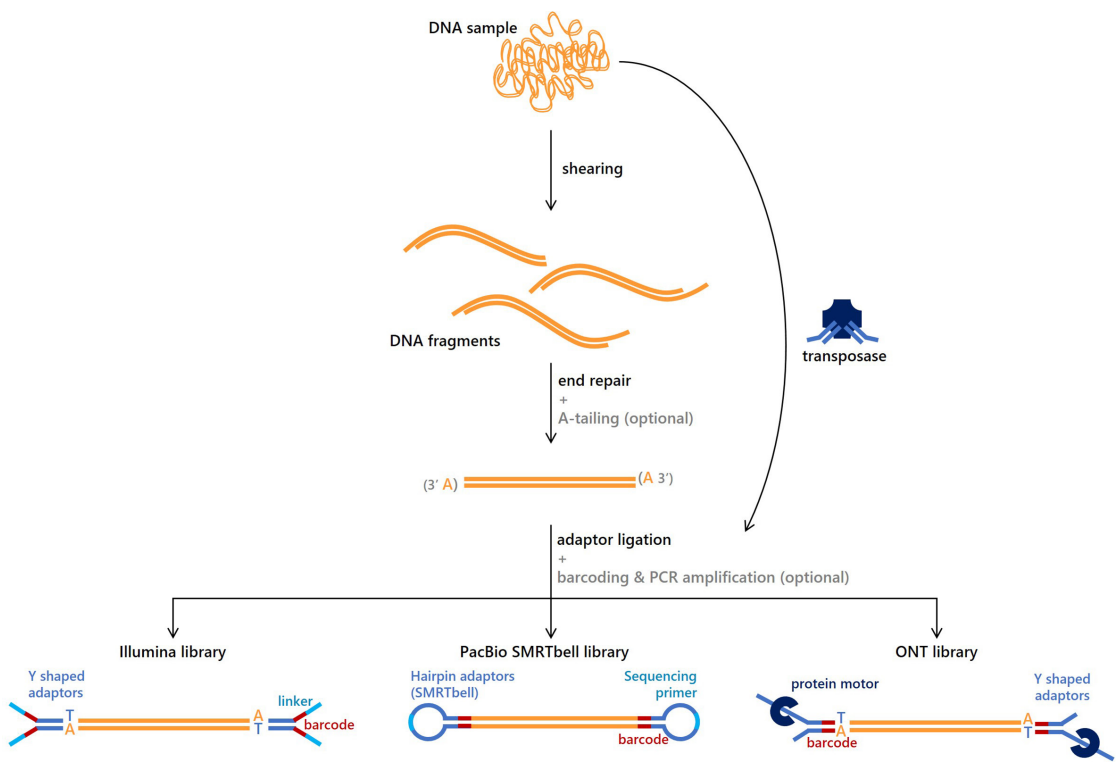

Figure 2. DNA library preparation for sequencing. The DNA sample is sheared into fragments and repaired. An optional A-tailing step corresponding to the addition of a unique dATP to the fragment's 3'-end can increase adaptor ligation efficiency. Alternatively shearing and adaptor ligation can be combined within a single step using a transposase-based method. Fragment bearing adaptors (blue) at their ends can be called libraries and are ready for sequencing. Adaptors or PCR primers can contain a barcode (red) allowing several samples to be sequenced together. Illumina's adaptors contain the linker (cyan) able to hybridize onto Illumina's sequencing flow cells. PacBio SMRTbell adaptors complement the primer sequence (cyan) allowing the polymerase to bind prior PacBio's flow cell loading. ONT adaptors are bound to a helicase motor protein (dark blue) which will regulate the speed at which DNA translocates through the pores.

prior to library preparation, the rest of the process being similar to the DNA library protocols described earlier. Enrichment steps can significantly improve the RNA library yield by capturing, e.g., messenger RNA molecules through their poly-A tails. ONT platforms also offer the ability to directly sequence RNA molecules, and both ONT and PacBio platforms can detect in their raw sequencing signal methylations and other nucleic modifications which play a critical role in gene expression and regulation.

Sequencing

Our libraries are now ready to be loaded on a sequencing platform (Figure 3). The technology behind each type

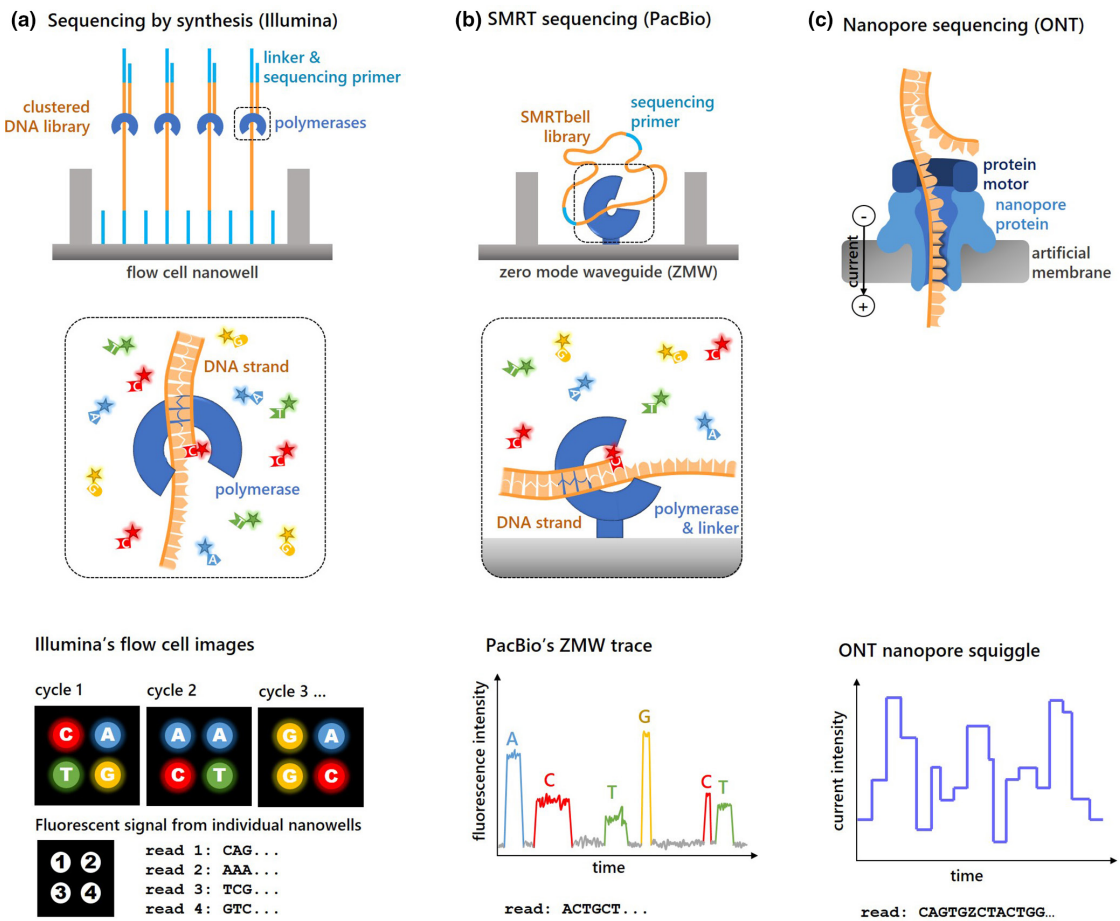

Figure 3. DNA library sequencing on different NGS technologies. (a) A single Illumina DNA library fragment hybridizes via the linkers (cyan) at the bottom of each nanowell and is amplified into a cluster of identical DNA library fragments. Sequencing primers allow the polymerase (blue) to start amplifying each DNA fragment in a synchronized parallel manner using fluorescent nucleotides. DNA synthesis is paused upon each nucleotide incorporation allowing the instrument to image the entire flow cell under laser illumination at each SBS cycle. A base-calling algorithm translates the colour of each cluster at a specific SBS cycle into the corresponding base sequence (one read per cluster). (b) PacBio SMRTbell libraries are annealed to a sequencing primer (cyan) and bound to the polymerase (blue) prior to flow cell loading. A DNA–polymerase complex is connected to the bottom of each nanowell (ZMW) via a linker. The incorporation of each fluorescent nucleotide leads to a burst of light captured in the raw video data. A base-calling algorithm translates the fluorescent intensity signal into its original DNA sequence. (c) Protein nanopores are embedded into an artificial membrane along the flow cell. A current is applied across the membrane forcing the DNA library strands to translocate through the nanopores. The motor protein loaded on the library adaptor slows down and guides the single-stranded DNA. The obstruction of a nanopore by a single-stranded DNA fragment leads to a change in the current which is measured continuously by an electronics chip integrated within the flow cell. The current intensity traces from each nanopore (squiggles) form ONT's raw sequencing data and are analysed in real time by ONT's base-calling algorithm which translates the current variation back into the original DNA sequence.

of sequencer will come with some advantages and constraints, which we will discuss below.

Illumina's SBS. Illumina's sequencing technology combines fluorescent microscopes with microfluidics devices allowing the sample and reagents to flow through and hybridize at the surface of a glass flow cell. The flow cells contain one or more channels through which the libraries flow. Their glass surface is etched with millions of nanowells grafted with synthetic DNA linkers complementary to the libraries' adaptors. Upon flow cell loading, one library fragment hybridizes at the surface of each nanowell (Figure 3a). In order to increase the nanowell signal from a single molecule to several thousands, the newly hybridized libraries are amplified using a surface PCR chemistry, also called isothermal or bridge amplification. After the formation of these mono-template clusters, SBS can start.

Fluorescently labelled nucleotides containing a blocker group are added onto the flow cell and incorporated to the clustered fragments one base at a time. A well-optimized chemistry pauses the library synthesis whilst a fluorescence image of the flow cell is recorded. After imaging, the fluorophore attached to the freshly incorporated nucleotide is cleaved, allowing a new cycle of synthesis and imaging to take place. An algorithm then performs the base calling by translating each image taken at each SBS cycle into a sequence.

Illumina's SBS technology has established its dominance over the past 15 years. The robustness and quality of the data produced coupled with the ease of high parallelization have allowed Illumina to offer a $1,000 dollar (human) genome 6 years ago and prices continued their decrease since. However, short-reads sequences leave some significant gaps for long-read technologies to fill in.

PacBio's SMRT sequencing. PacBio instruments are based on a fairly similar technology to Illumina platforms since they too use fluorescence microscopy. The main differences lie in the facts that PacBio sequencers are single-molecule fluorescence microscopes and that the DNA synthesis is continuous rather than paused at each sequencing cycle, a principle named single-molecule real-time (SMRT) DNA sequencing. A PacBio flow cell is also composed of millions of nanowells, called zero mode waveguides (ZMW), at the bottom of which a single polymerase is immobilized (Figure 3b). Before being loaded onto the sequencer, the SMRTbell libraries are annealed to complementary sequencing primers and bound to a polymerase molecule, which is then linked to a ZMW surface.

In each ZMW, library fragments are continuously synthesized using fluorescently labelled nucleotides emitting a strong colourful signal upon incorporation. To reduce the background noise, a specific illumination pattern restricts the fluorescence signal to the bottom of the ZMW. This feature allows an accurate detection of the burst of fluorescence originating from the incorporation of each nucleotide. The video of the fluorescence intensity in each ZMW is then converted into a sequence by the base-calling algorithm.

Although PacBio's raw error rate is higher than Illumina's one (15% for PacBio vs 0.1% for Illumina at a quality score of Q30), it's random. Accuracy is greatly improved by reading through the same circular library fragment many times until the polymerase degrades. The consensus sequence of these multiple reads of the same library molecule is called HiFi (for high-fidelity) reads and can reach a base call accuracy similar to Illumina (99.9%). Despite its relatively high cost, the constant improvement of data quality coupled with reads of several tens of thousands base pairs has established PacBio as a sequencing method of choice to produce very high quality genomes, for the *de novo* sequencing of species for which no reference genome is available yet, or to cover long repetitive regions of a genome.

Oxford Nanopore Technology sequencing. ONT is based on a completely different principle. Whilst PacBio and Illumina engineered polymerases to slow down their nucleotide incorporation or attach them at the surface of a flow cell, ONT made the most out of another type of naturally occurring protein, the pore-forming protein α-haemolysin. These barrel-shaped proteins are typically found embedded within cell membranes and regulate which molecules can enter or leave the cell. α-Haemolysin happened to have an inner diameter of 1 nm, just large enough to allow a single strand of DNA (ssDNA) through (Figure 3c).

ONT flow cells are controlled by an application-specific integrated circuit (ASIC) which sits just under an artificial membrane containing hundreds or thousands of these protein nanopores. A voltage is set across the membrane attracting negatively charged DNA molecules through the nanopores which obstructs the current across the membrane. Because the four bases of DNA have different shapes and sizes, their translocation through a nanopore leads to different current variations which constitute ONT's raw sequencing signal (squiggle). A base-calling algorithm then converts the squiggles into a sequence, each intensity step corresponding to the multiple bases obstructing the pore at a certain time.

Despite a lower raw read accuracy (>98% at a quality score of Q20), ONT offers several unique advantages. First of all if a DNA sample is handled carefully and preserved throughout the library preparation, reads can reach several millions of base pairs. These ultra-long reads significantly decrease the number of puzzle pieces

(contigs) when assembling the sequencing data into a genome.

Depending on the model, ONT instrument's size ranges between a mobile phone and microwave. For comparison, Illumina and PacBio largest sequencers are similar to large fridge-freezer units. Of course large sequencers offer greater throughput, but a small portable device requiring nothing more than a powerful laptop can be key to perform experiments outside of the classical academic laboratories. ONT's smallest sequencer, the MinIon, has been used in low-income countries, notably during the Ebola pandemic in 2014, in Antarctica, or even on board the International Space Station.

Finally, ONT offers real-time analysis as the base-calling algorithm starts converting the squiggles into a sequence as soon as the run begins. This provides a very rapid turnaround time in comparison to other sequencers and allows the user to produce just the right amount of data for a specific application. We are now entering an era where data processing and storage represent as big a challenge as sample processing and sequencing itself. Minimizing data production whilst ensuring its quality is therefore a critical endeavour.

Other technologies. A few other very promising NGS technologies are now appearing on the market. MGI Technology uses a fluorescence SBS chemistry similar to Illumina's and could offer some competition to the current world leader. ONT itself is pursuing research on solid-sate nanopore sequencers using synthetic materials such as graphene containing nano holes through which the ssDNA could translocate. GenapSys is among the latest to have entered the market with its non-fluorescent SBS chemistry. The incorporation of each negatively charged nucleotide is detected via an electronic chip. These devices hold the promise of compact, robust and cheaper instruments.

What's next

The recent dramatic drop of sequencing cost meant that the technology could escape the academic niche and start invading the clinical world with hospitals introducing NGS as a routine diagnosis tool for rare genetic diseases and cancers. These programs started as national or international consortiums with presidents and prime ministers announcing endeavours such as DeCode (founded in 1996 in Iceland, 70% of the Icelandic population had been sequenced by 2019), the 100,000 Genomes Projects (launched in 2012 by NHS England and spun out as NHS Genomic Medicine Service) or the NIH All of Us research program (founded in 2015 and aiming to sequence 1 million human genomes), just to cite a few. These projects are building the first blocks of personalized medicine. Other sequencing databanks are drawing the landscape of a new population genomics era covering different continents and ethnicities (UK Biobank, H3Africa, Genome Asia 100 k).

Population genomics also applies to other species than human, particularly in times of epi- or pandemics such as Ebola in 2014 or SARS-Cov-2 today. Sequencing the viral genomes of positive cases in such outbreaks allows researchers to build up the phylogenetic tree of an epidemic and trace contamination cases to their origin, sometimes spotting clusters and super-spreaders before they become obvious to the medical surveillance organisms. When vaccines or treatments exist, monitoring the evolution of infectious genomes is a powerful tool to characterize new variants and overcome pathogen resistance. This strategy has already shown its potential in the fight against drug- or vaccine-resistant pathogens such as malaria in Southeast Asia or the current SARS-Cov2 pandemic with the work carried out by the CoG-UK consortium.

The scientific community continues to dream bigger and pushes NGS technologies further with the ability to sequence DNA and RNA extracted from a single cell, and to associate genomic data to its original tissue or cell location (spatial transcriptomic). A multi-omics approach where scientists build up multifaceted images of a single sample, gathering transcriptomic, genomic, long-range DNA interactions, DNA/RNA–protein interactions and/or cell localization information, constitutes one of the most exciting developments of the NGS field. The Human Cell Atlas project uses such an approach to build up a map of the human body made of 'a collection of cellular reference maps characterizing each of the thousands of cell types in the human body'. On the other hand, the development of long-read sequencing and bioinformatics technologies means that *de novo* sequencing is not confined to small genomes or model organisms anymore. The Earth Biogenome Project aims to sequence, catalogue and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of 10 years and is described by its founder as 'a *moonshot* for biology'. ■

Acknowledgements

I would like to thank Laura Olivares Boldu from the Wellcome Connecting Science team for our very helpful discussion and her advices when writing this article and preparing the figures.

Further Reading & Viewing

- https://www.yourgenome.org/ and the associated YouTube channel https://www.youtube.com/c/yourgenome/
- Shendure, J., Balasubramanian, S. Church, G.M. et al. (2017) DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353. DOI: https://doi.org/10.1038/nature24286
- Logsdon, G.A., Vollger, M.R. and Eichler, E.E. (2020) Long-read human genome sequencing and its applications. *Nat. Rev. Genet*. **21**, 597–614. DOI: https://doi.org/10.1038/s41576-020-0236-x
- The 1000 Genomes Project Consortium, Auton, A., Brooks, L.D. et al. (2015) A global reference for human genetic variation. *Nature* **526**, 68–74. DOI: https://doi.org/10.1038/nature15393
- van Dijk, E.L., Jaszcyszyn, Y., Naquin, D. et al. (2018) The Third Revolution in Sequencing Technology. *Trends Genet*. **34**, 666–681. DOI: https://doi.org/10.1016/j.tig.2018.05.008
- Office of the Press Secretary (2000) President Clinton announces the completion of the first survey of the entire human genome. https://web.ornl.gov/sci/techresources/Human_Genome/project/clinton1.shtml [Accessed 31 August 2021]
- Department of Health and Social Care (2018) 100,000 whole genomes sequenced in the NHS. https://www.gov.uk/government/news/100000-whole-genomes-sequenced-in-the-nhs [Accessed 31 August 2021]
- Giani, A.M., Gallo, G.R., Gianfranceschi, L. et al. (2019) Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J*. **18**, 9–19. DOI: https://doi.org/10.1016/j.csbj.2019.11.002
- Whibley, A., Kelley, J.L. and Narum, S.R. (2021) The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. *Mol. Ecol. Resour*. **21**, 641–652. DOI: https://doi.org/10.1111/1755-0998.13312

Louise Aigrain is Senior Staff Scientist in the R&D DNA pipelines team of the Wellcome Sanger Institute. After a PhD in biochemistry (CNRS – Université Paris-Sud 11, 2010), she pursued a post-doctoral position in the Physics Department of the University of Oxford using single-molecule fluorescence methods to study DNA polymerase dynamics and accuracy. In 2014, Louise joined the Wellcome Sanger Institute where her expertise in molecular biology and biophysics allows her to develop and optimize new NGS processes, from bespoke ancient DNA sequencing, to large sequencing projects such as UK Biobank and CoG-UK. email: louise.aigrain@sanger.ac.uk

