DOI: 10.24425/academiaPAS.2023.147466

LEOWOLFERT/SHUTTERSTOCK.COM

Revolutionary Applications of DNA Sequencing

January Weiner, PhD
is a biologist by training and a bioinformatician by profession. For more than a decade he has been involved in the functional analysis of transcriptional responses in humans and model organisms, especially in the context of infectious diseases and vaccinations.
january.weiner@bih-charite.de

Few technologies have transformed the field of biology as profoundly as sequencing – the ability to decipher the sequence of base pairs in a fragment of DNA.

January Weiner

Berlin Institute of Health at Charité
– Universitätsmedizin Berlin

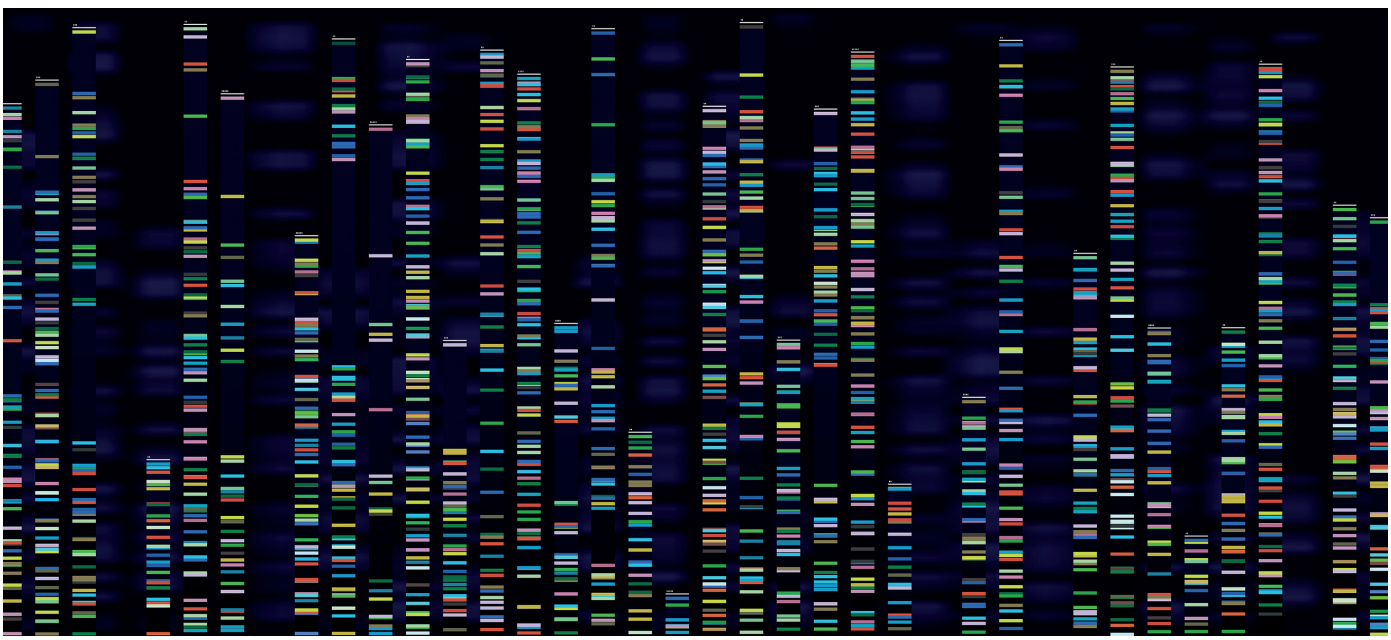
The first two human genomes were sequenced over 20 years ago, and since then, the genomes of hundreds of thousands of different individuals have been deciphered. The genes that make up our genomes exist in many variants, called *alleles*, and each of us has a unique set of such alleles. The point of sequencing more and more human genomes is not just about scientifically explaining the differences between people, the origins of our common traits, or identifying which genes encode, say, our height or intelligence – it's also about better understanding the genetic basis of many diseases, which may lead to new biomedical applications. Such knowledge can be useful both in diagnosis and in the search for new treatment methods. Genome sequencing also holds the promise of giving rise to *personalized medicine*, in which treatments are specifically tailored to individual patients, including their individual genomes.

However, sequencing has far more applications than just reading out an individual's DNA. Apart from the alleles we inherit from our parents and any new mutations, which are present in all our cells, it turns out that new DNA variants arise within our bodies all the time. Firstly, the adaptive immune response that occurs in all vertebrates involves the creation of new antibodies through recombination and selection. As a result, the genomes of cells producing antibodies differ from the genomes of our other cells. Sequencing helps us better understand the mechanisms involved in developing immunity to infections, as well as in autoimmune diseases. Secondly, new, evolutionarily unintended mutations continually occur in the process of cell division in our bodies. These mutations are the cause of cancer and sequencing the genome of cancer cells makes it possible to identify key mutations and select effective therapies.

Gene expression

Another application of sequencing is transcriptomics. Each of our genes undergoes a process of expression, meaning that it is transcribed from DNA to RNA, and if it encodes a protein sequence, it is then translated into a protein. Our cells respond to their environment and communicate with each other largely through the regulation of gene expression: "switching on" or "switching off" the transcription of genes into RNA depending on the context. For example, if the organism has become infected by a virus, a cell may receive a signal stimulating it to produce proteins that inhibit viral infections. Sequencing the transcriptome allows us to detect and characterize the immune response. In recent years, methods that allow us to study the transcriptomes even of individual cells (single-cell RNA-seq, or sc-RNA-seq) have become widespread.

DNA sequencing is one of the fundamental technologies for studying organisms and cells at the molecular level. Unlike other methods, such as those based on antibodies or mass spectrometry, it makes it possible to analyze a vast number of samples in an incredibly short time. Numerous new DNA sequencing methods emerged at the beginning of the twenty-first



TARTILA/SHUTTERSTOCK.COM

century – these are sometimes referred to as *second-generation sequencing*, or more commonly as *high-throughput sequencing*. Short-read sequencing methods – in which short sequences of some 50 to 150 base pairs are sequenced – have gained particular popularity. For many of the aforementioned applications, short-read sequencing has enormous advantages: on the one hand, the sequenced segments are long enough to unambiguously identify, for instance, most messenger RNA (which encodes our proteins); on the other hand, the method allows the simultaneous sequencing of hundreds of millions of molecules within just a few hours.

However, two new families of sequencing-related technologies are worthy of attention: long-read sequencing and spatial transcriptomics.

Sequencing of long fragments

At the end of the twentieth century, automated sequencers were able to read a few hundred base pairs in a small number of fragments of DNA simultaneously. It was using such technology that the first nearly complete sequences of the human genome were published in the early twentieth century. Although almost all genes and over 90% of the entire human DNA were sequenced, many regions of our genome still resisted the sequencing techniques of that time and even newer technologies. The trouble was with regions of repeated DNA sequences, where the same sequence of base pairs is duplicated numerous times. Some regions of such sequences may encompass even millions of base pairs. For example, fragments found at both ends of all our chromosomes – telomeres – consist of hundreds or thousands of repetitions of short sequences.

> The point of sequencing human genomes is not just about scientifically explaining the differences between people, the origins of our common traits, but also better understanding the genetic basis of many diseases, leading to new biomedical applications.

And herein lies a fundamental problem: if the sequence read is shorter than the repetition region, it is often impossible to say unequivocally which exact position the sequence comes from. Telomeres therefore cannot be properly sequenced with the help of short-read sequencing, or even technologies that allow several hundred base pairs to be read. Unfortunately, the same applies to the structural variants already mentioned, i.e. large-scale mutations in our genome, which of course makes them difficult to detect.

Over the past few years, tremendous progress has been made in long-read sequencing. Two technologies, in particular, deserve attention: single molecule real-time sequencing (SMRT) and nanopore sequencing. Both allow very long DNA fragments to be sequenced – orders of magnitude longer than all other methods.

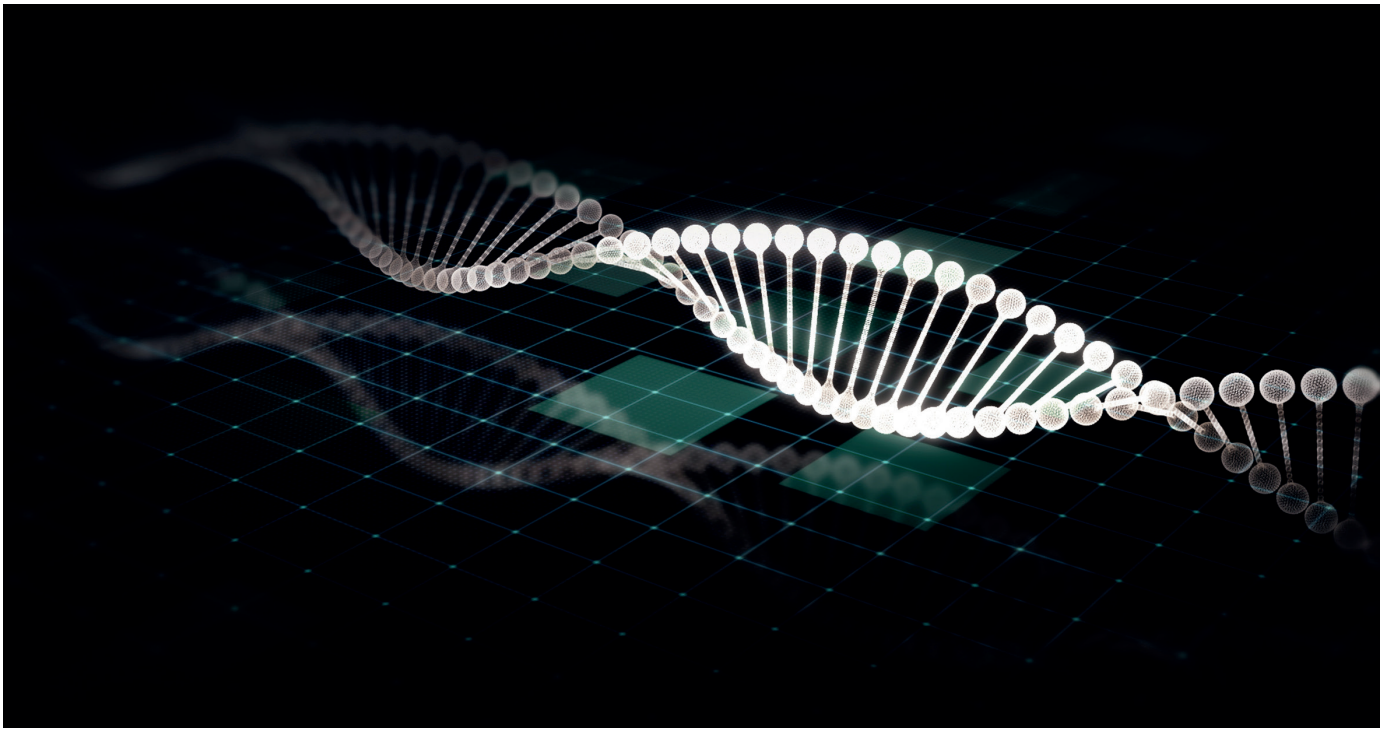
SMRT is very similar to the classical Sanger sequencing method in that the basic mechanism of sequencing involves the synthesis of complementary DNA strands. Fluorescently labeled nucleotides also involved, making it possible to detect nucleotides incorporated into the synthesized DNA strand. The crucial difference, however, is that individual molecules are observed in real time. As a result, the DNA polymerase works continuously and can synthesize even fragments tens of thousands of base pairs in length.

Nanopore sequencing, in turn, is based on a completely different mechanism. Unlike most practically proven sequencing technologies, the detection of successive nucleotides is not achieved by synthesizing a new DNA strand. Rather, in this technology, an electric field is used to move a DNA strand through a microscopic opening (nanopore) in a membrane. This gives rise to changes in voltage that depend on which nucleotide is passing through the opening at a given moment. By monitoring these voltage changes, we can read off a sequence whose length is theoretically limited only by the length of the DNA strand being sequenced. In practice, although the longest sequences so obtained have been over a million base pairs long, more typical read lengths range from around 50 to over 20,000 base pairs.

Both long-read methods have recently been used to sequence the complete human genome for the first time in history – including telomeres, centromeres, and all other sequences that had not been read before. There is no doubt that this marks the beginning of a new era in both basic science and biomedicine. Currently, both long-read methods are quite expensive, but sequencing costs are decreasing rapidly. Already now, the cost of sequencing a single human genome has dropped to just a few hundred dollars – for comparison, just 15 years ago sequencing a single human genome cost millions of dollars.

Monitoring changes

Among the various uses to which genetic sequencing can be put, transcriptomics holds a prominent place. It is undoubtedly the simplest and cheapest method



COLOR4260/SHUTTERSTOCK.COM

that can be used to monitor changes in the processes occurring in living cells, and so it has been routinely used for well over two decades. It has found a great number of applications in almost all fields of biological and medical science. It allows for precise tracking of changes in gene expression in entire organisms, tissues, selected cell types, and even individual cells. We are now observing another breakthrough in this technology, known as *spatial transcriptomics*.

Microscopic methods have long made it possible to study the processes underway inside our tissues. By visualizing individual genes and proteins, we can determine where in the tissue or cell the expression of a specific gene occurs. However, current technological advancements have enabled transcriptomics (high-throughput analysis of the expression of many genes simultaneously) to be combined with microscopic visualization of tissue organization. This makes it possible to determine not only which cells regulate gene expression and how they do so, but also where in the examined tissue they are located and how gene expression changes at the boundary between healthy and diseased tissue.

There are roughly two approaches to spatial transcriptomics. The first type of methods is based on the preparation of special glass slides, where each fragment of the surface is specially marked with short DNA sequences (oligonucleotides). A tissue preparation is applied to the slide. When RNA from the cells of the preparation is synthesized into cDNA, oligonucleotides attach to it. Then, cDNA from the entire preparation is harvested and sequenced in the usual way. Labelling makes it possible to later recognize from which place on the slide a particular sequence came and to reconstruct the spatial structure of the preparation, which can then be overlaid onto microscopic images. The advantage of these methods is a relatively large "depth" – the number of different RNA molecules that can be identified through them. The drawback is relatively low resolution.

The second type of method is based on *in situ* detection, similar to what has been done for many decades, only on a larger scale. The preparation is labeled with fluorescent probes corresponding to different mRNA sequences. This often involves marking short, characteristic DNA sequences with fluorescent markers. Such probes bind specifically to mRNAs, which can then be directly visualized on the preparation. There are many variations in this family of spatial transcriptomic methods, but most of them only allow for the detection of a limited repertoire of genes defined *a priori*.

Progress in developing new methods in biology and medicine takes years, even decades. Methods that were primarily experimental and rarely appeared in top-tier journals ten years ago are now being refined and routinely used in laboratories worldwide, and their costs, while not trivial, are no longer prohibitive. The methods described in this article – long-read sequencing and spatial transcriptomics – are now entering this second stage. Their practical importance for both fundamental science and biomedicine can be expected to increase significantly in the coming years. ■

Further reading:

Deamer D., Akeson M., Branton D., Three Decades of Nanopore Sequencing, *Nature Biotechnology* 2016.

Lightbody G., Haberland V., Browne F., Taggart L., Zheng H., Parkes E., Blayney J.K., Review of Applications of High-Throughput Sequencing in Personalized Medicine: Barriers and Facilitators of Future Progress in Research and Clinical Application, *Briefings in Bioinformatics* 2019.

Logsdon G.A., Vollger M.R., Eichler E.E., Long-Read Human Genome Sequencing and Its Applications, *Nature Reviews Genetics* 2020.