Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
47(183):439-444, abril-junio de 2023. doi: https://doi.org/10.18257/raccefyn.1937

**Ensayo/Assay**

# Secuenciación de genomas utilizando lectura de cadenas largas

## Genome sequencing using long-read sequencing

El año pasado la secuenciación de cadenas largas "Long-read sequencing" (LRS), fue declarada como la metodología mas importante en su campo (**Marx,** 2023; **"Method of the Year 2022: long-read sequencing,"** 2023) Esta tecnología hace parte de la 3ra generación de secuenciación de ácidos nucleicos y ha sido de gran importancia en la generación del genoma humano completo de telómero a telómero (T2T) (**Gómez Gutiérrez,** 2022). La secuenciación de ácidos nucleicos comenzó el siglo pasado en la década de los 70, cuando Frederick Sanger desarrollo la técnica de secuenciación genética, y Walter Gilbert y Allan Maxam desarrollaron una técnica química (**Mukherjee,** 2016). La automatización de la secuenciación de cadenas terminales de Sanger permitió completar el proyecto del genoma humano, que tuvo un costo de 3 billones de dólares y tardó 10 años. A partir del 2005 emergen tecnologías de secuenciación de última generación (NGS), entre las cuales se desarrollaron el secuenciador 454 de Life Sciences® por pirosecuenciación y la que imperó en su momento con base en los secuenciadores Illumina®, que lograron reducir el costo de la secuenciación a 1000 dólares por genoma. La característica principal de las NGS es el alto rendimiento, al utilizar la amplificación clonal del ADN por medio de la reacción en cadena de la polimerasa (PCR), la cual permite obtener cada fragmento del genoma secuenciado más de 50 veces. Sin embargo, la desventaja de estas tecnologías conocidas como de segunda generación es que generaban lecturas cortas entre 250 y 800 pares de bases, las cuales demandan un alto procesamiento computacional para su ensamblaje y no logran obtener la secuencia contigua de cromosomas completos (**Figura 1**).

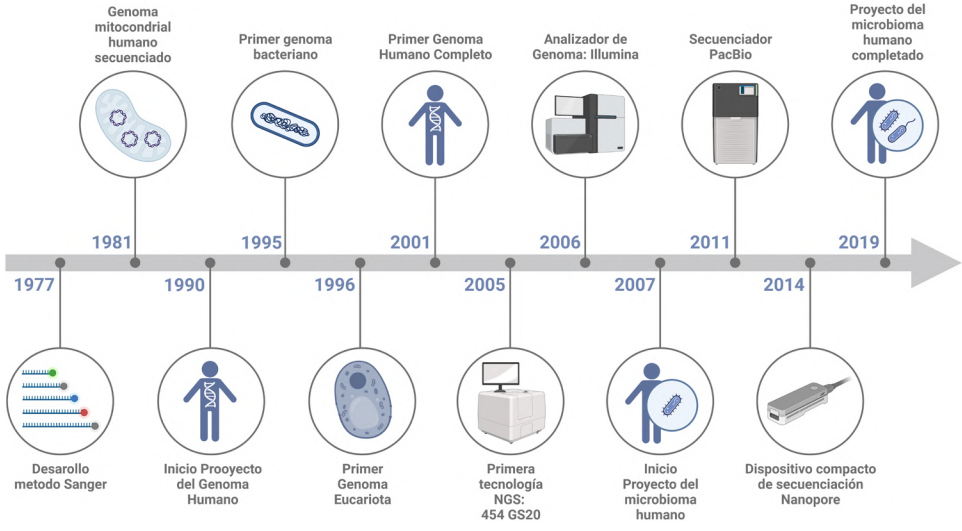


**Figura 1.** Historia de la secuenciación del ADN. Línea de tiempo con los principales eventos históricos acerca de las tecnologías de secuenciación. (Realizada por OMG usando https://app.biorender.com).

Los métodos de secuencia de tercera generación están enfocados en secuenciar cadenas largas tanto de ADN como de ARN; al momento hay 2 compañías: una es Pacific Biosciences (PacBio®), su tecnología se conoce como SMRT (Single Molecule Real-Time), en la que se emplea una plantilla de ADN circular de doble cadena con un adaptador de cadena sencilla en ambos extremos llamado "campana SMRT", luego se agrega ADN polimerasa y una vez se ensambla se colocan en una célula de lectura y a medida que la polimerasa genera la nueva cadena y se agregan los nucleótidos marcados con diferentes fluorocromos, se hace una detección del proceso por rayos láser y esta se graba. Este proceso se repite miles de veces determinando la secuencia con una precisión del 99,9%. El método SMRT es capaz de leer en promedio fragmentos de 2 a 20 Gb (gigabases). Dependiendo de la plataforma empleada se puede obtener un cubrimiento de 25 a 40 veces mayor (**Marx,** 2023), (**Figura 2**).

La segunda compañía es Oxford Nanopore Technologies (ONT®), cuya tecnología de secuenciación emplea moléculas de ADN lineal en vez de las circulares usadas por PacBio. El sistema de ONT inicia la secuenciación uniendo la doble cadena de ADN a un adaptador que es precargado con una proteína motora; esta mezcla se carga en una célula de flujo que tiene miles de nanoporos insertados en una membrana sintética. La proteína motora desenrolla la doble cadena de ADN y junto con una corriente eléctrica empuja el ADN cargado negativamente a través del poro a un paso controlado. A medida que el ADN pasa por el poro causa una disrupción en la corriente que es analizada en tiempo real para determinar la secuencia de las bases. ONT lee fragmentos de 1 a 6 Gb con una precisión del 99%. Para lecturas ultra-largas puede llegar hasta 1 a 2 Mb (**Marx,** 2023) (**Figura 2**).

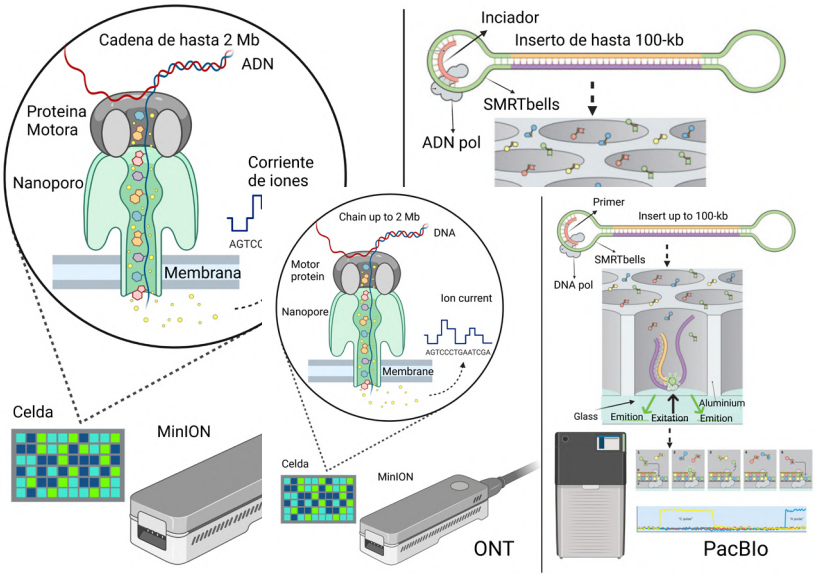


**Figura 2.** Tecnologías de secuenciación genómica basadas en lecturas de cadena larga. En el panel izquierdo, se muestra el esquema de la tecnología de nanoporos utilizando el MinION de ONT como ejemplo. Esta técnica utiliza una plantilla de ADN lineal cargada en una proteína motora y un nanoporo molecular. A medida que el ADN se desenrolla y es conducido a través del nanoporo, los cambios característicos en la corriente de iones permiten determinar la secuencia de las bases residentes en la cadena de ADN utilizando un algoritmo entrenado en la identificación de las bases. En el panel derecho se muestra el esquema secuenciación SMRT, utilizando la plataforma PacBio como ejemplo. Esta técnica utiliza plantillas en forma de campanas SMRT, que se secuencian como moléculas individuales dentro de una matriz que cada uno contiene una ADN polimerasa con su iniciador. Un láser excita a cada nucleótido marcado con un fluorocromo específico durante su incorporación por la ADN polimerasa y cada señal de fluorescencia es captada y transformada en secuencia por un algoritmo para cada nucleótido. (Realizada por OMG usando https://app.biorender.com)

Uno de los factores más importantes en la obtención de LRS, es la preparación de los ácidos nucleicos, que deben ser de alto peso molecular. También, la secuenciación es basada directamente en la cadena de ADN nativo obtenido del organismo sin requerir amplificación por PCR, lo que puede eliminar la tasa de error de la polimerasa, en el cálculo de la incertidumbre del proceso de secuenciación.

LRS ha encontrado muchas aplicaciones en todas las áreas de la genómica, entre ellas la posibilidad de leer genomas de eucariotas de telómero a telómero (T2T) (**Gómez Gutiérrez,** 2022), lo que permite el análisis de muchas regiones de estos genómicas antes desconocidas, debido a su contenido en secuencias altamente repetitivas, localizadas en regiones centroméricas y teloméricas. Además, permite el estudio preciso y detallado de variantes cromosómicas producidas por rearreglos, por variaciones en el número de copias (CNVs) de genes o por aneuploidías **(Nurk *et al.,*** 2022; **Warburton & Sebra,** 2023).

La aplicación mas novedosa corresponde al estudio de la regulación de la expresión genética, debido a la forma de lectura de ONT y SMRT permite detectar modificaciones en el ADN y ARN, tales como la metilación 5 metilcitosina (5mC) y N6-metiladenosina entro otros. Como la secuenciación es del ADN nativo obtenido del organismo, y no de un clon obtenido por PCR, este proceso ha permitido obtener patrones de metilación de genomas (epigenómica) sin ningún tratamiento adicional del ADN. Por ejemplo, las plataformas detectan diferencias entre la citosina y 5-metilcitosina, a través de un cambio en la corriente en ONT o cambio en la intensidad de la señal en SMRT, luego con algoritmos desarrollados para la búsqueda de estos patrones de modificación en secuencias se logran obtener epigenomas. Esto permitirá un amplio estudio de la epigenética y epigenotranscriptómica, para entender mejor las bases de la regulación genética (**Warburton & Sebra,** 2023).

Finalmente, esta tecnología se está aplicando al estudio de la metagenómica, y el estudio de los microbiomas, al igual de la generación de una metacodificación de los organismos existentes que permita la realización de un árbol de la vida más completo y confiable (**Hoang *et al.,*** 2021). La gran versatilidad de la LRS y sus múltiples aplicaciones, claramente nos permiten entender por qué esta metodología fue seleccionada como la tecnología del año 2022 (**“Method of the Year 2022: long-read sequencing,”** 2023).

Last year, Long-read sequencing (LRS) was declared the most important methodology in its field. (**Marx**, 2023; **“Method of the Year 2022: long-read sequencing,”** 2023). This technology is part of the 3rd generation of nucleic acid sequencing and has been of great importance in the generation of the complete human genome from telomere to telomere (T2T) (**Gómez Gutiérrez**, 2022). Nucleic acid sequencing began last century in the 1970s, when Frederick Sanger developed the genetic sequencing technique and Walter Gilbert and Allan Maxam developed a chemical technique for the sequencing of nucleic acids. (**Mukherjee**, 2016). The automation of Sanger end chain-termination sequencing enabled the completion of the human genome project, which cost $3 billion and took 10 years. From in 2005 onwards, next generation sequencing technologies (NGS) emerged, among which the Life Sciences 454 sequencer® by pyrosequencing; and the one that prevailed at the time based on Illumina sequencer®, which managed to reduce the cost of sequencing to $1000 per genome, were developed. The main characteristic of the NGS is its high yield, as it uses clonal amplification of the DNA by polymerase chain reaction (PCR), which makes it possible to obtain each fragment of the sequenced genome more than 50 times. However, the disadvantage of these technologies known as second generation is that they generate short reads between 250 and 800 base pairs, which require high computational processing for their assembly and do not obtain the contiguous sequence of complete chromosomes. **(Figure 1)**.

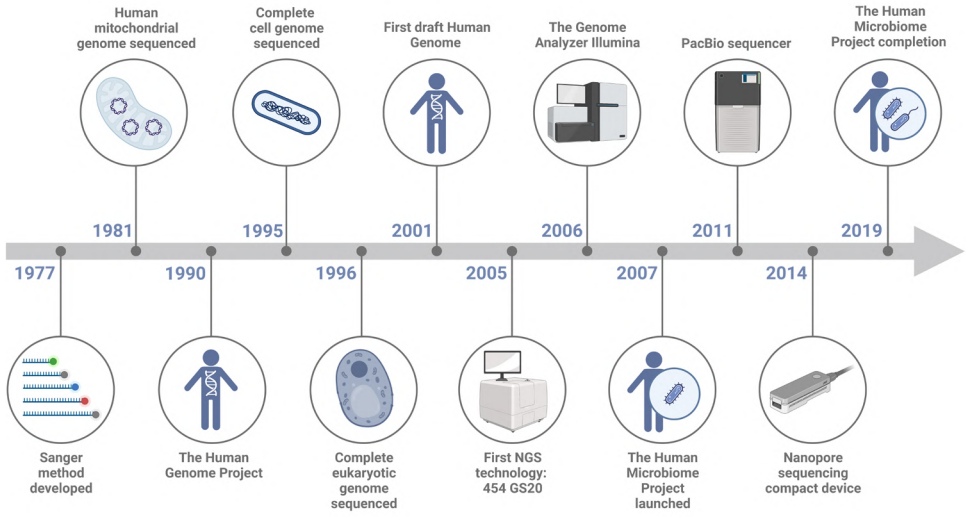


**Figure 1.** History of DNA sequencing. Timeline with major historical events about sequencing technologies. (Made by OMG using https://app.biorender.com).

Third generation sequencing methods are focused on sequencing long strands of both DNA and RNA; at the moment there are 2 companies; one is Pacific Biosciences (PacBio®), their technology is known as SMRT (Single Molecule Real-Time), in which a double-stranded circular DNA template is used with a single-stranded adapter at both ends called an SMRT bell, then DNA polymerase is added and once it is assembled, it is placed in a reading cell and as the polymerase generates the new strand, the nucleotides labeled with different fluorochromes are added, the process is detected by a laser beams and recorded. This process is repeated thousands of times determining the sequence with an accuracy of 99.9%. The SMRT method is capable of reading on average fragments of 2 to 20 Gb (gigabases). Depending on the platform used, a coverage of 25 to 40 times higher can be obtained (**Marx**, 2023)2023, **(Figure 2)**.

The second company is Oxford Nanopore Technologies (ONT®), whose sequencing technology employs linear DNA molecules instead of the circular ones used by PacBio. ONT's system initiates sequencing by attaching double-stranded DNA to an adapter that is preloaded with a motor protein, this mixture is loaded into a flow cell that has thousands of nanopores embedded in a synthetic membrane. The motor protein unwinds the double-stranded DNA and together with an electrical current pushes the negatively charged DNA through the pore at a controlled rate. As the DNA passes through the pore it causes a disruption in the current that is analyzed in real time to determine the sequence of the bases. ONT reads fragments from 1 to 6 Gb with 99% accuracy. For ultra-long reads it can go up to 1 to 2 Mb (**Marx**, 2023) **(Figure 2)**.

One of the most important factors in obtaining LRS is the preparation of nucleic acids, which must be of high molecular weight. Also, sequencing is based directly on the native DNA strand obtained from the organism without requiring PCR amplification, which can eliminate the polymerase error rate in calculating the uncertainty of the sequencing process.

LRS has found many applications in all areas of genomics, among them the possibility of reading eukaryotic genomes from telomere to telomere (T2T) (**Gómez Gutiérrez**, 2022). This allows the analysis of many previously unknown regions of these genomes, due to their content in highly repetitive sequence, located in centromeric and telomeric regions. In addition, it allows the precise and detailed study of chromosomal variants produced by rearrangements and variations in the number of copies (CNVs) of genes or aneuploidies (**Nurk *et al.***, 2022; **Warburton & Sebra**, 2023).

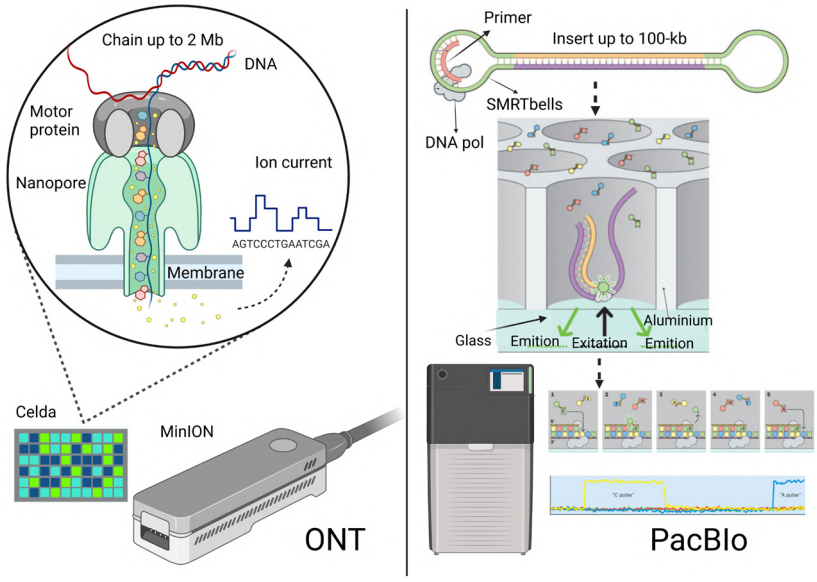


**Figure 2.** Genomic sequencing technologies based on long-chain reads. In the left panel, the schematic of nanopore technology is shown using ONT MinION as an example. This technique uses a linear DNA template loaded onto a motor protein and a molecular nanopore. As the DNA unwinds and is driven through the nanopore, characteristic changes in the ion current allow the sequence of resident bases on the DNA strand to be determined using a trained base caller algorithm. The right panel shows the SMRT sequencing scheme, using the PacBio platform as an example. This technique uses dumbbell-shaped templates "SMRTbells", which are sequenced as individual molecules within an array that each contain a DNA polymerase with its primer. A laser excites each nucleotide labeled with a specific fluorochrome during its incorporation by the DNA polymerase and each fluorescence signal is captured and transformed into sequence by an algorithm for each nucleotide. (Done by OMG using https://app.biorender.com)

The most ground-breaking application corresponds to the study of the regulation of gene expression, due to the way ONT and SMRT are read, which allows the detection of modifications in DNA and RNA, such as methylation of 5-methylcytosine (5mC) and N6-methyladenosine, among others. As the sequencing uses the native DNA obtained from the organism, and not from cloned fragments obtained by PCR, this process has made it possible to obtain genome methylation patterns (epigenomics) without any additional DNA treatment. For example, the platforms detect differences between cytosine and 5mC, through a change in current in ONT or change in signal intensity in SMRT, then with algorithms developed to search for these patterns of modification in sequences, epigenomes can be obtained. This will allow a broad study of epigenetics and epigeno transcriptomics to better understand the basis of gene regulation (**Warburton & Sebra**, 2023)2023.

Finally, this technology is being applied to the study of metagenomics, and the study of microbiomes, as well as the generation of a meta-coding of existing organisms that allows the construction of a more complete and reliable tree of life (**Hoang *et al.***, 2021). The great versatility of the LRS and its multiple applications clearly allow us to understand why this methodology was selected as the technology of the year 2022 (**"Method of the Year 2022: long-read sequencing,"** 2023).

**Juan Guillermo McEwen**[1, 2,*], **Oscar Mauricio Gómez**[1, 3]
[1] Corporación para Investigaciones Biológicas (CIB), Medellín, Colombia.
[2] Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia.
[3] Escuela de Microbiología, Universidad de Antioquia, Medellín, Colombia.

## Referencias

**Gómez-Gutiérrez, A.** (2022). El genoma humano: llenando los vacíos. *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, *46*(179), 577-579. https://doi.org/10.18257/raccefyn.1712

**Hoang, M.T.V., Irinyi, L., Hu, Y., Schwessinger, B., Meyer, W.** (2021). Long-Reads-Based Metagenomics in Clinical Diagnosis With a Special Focus on Fungal Infections. *Frontiers in Microbiology*, *12*, 708550. https://doi.org/10.3389/fmicb.2021.708550

**Marx, V.** (2023). Method of the year: long-read sequencing. *Nature Methods*, *20*(1), 6-11. https://doi.org/10.1038/s41592-022-01730-w

**Method of the Year 2022: long-read sequencing.** (2023). *Nature Methods*, *20*(1), 1. https://doi.org/10.1038/s41592-022-01759-x

**Mukherjee, S.** (2016). The Gene: An Intimate History. *Scribner.* https://books.google.com.co/books?id=S4nHjgEACAAJ

**Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S.Y., . . . Phillippy, A. M.** (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44-53. https://doi.org/10.1126/science.abj6987

**Warburton, P.E., Sebra, R.P.** (2023). Long-Read DNA Sequencing: Recent Advances and Remaining Challenges. *Annual Review of Genomics and Human Genetics*, *24*. https://doi.org/10.1146/annurev-genom-101722-103045