Avenços en les tecnologies de seqüenciació del DNA

Berta Fusté,[1] Elena Vila[1] i Mònica Bayés[1, 2]

[1] Centre Nacional d'Anàlisi Genòmica - Centre de Regulació Genòmica (CNAG-CRG), Barcelona Institute of Science and Technology (BIST)
[2] Universitat Pompeu Fabra (UPF)

Correspondència: Berta Fusté. Centre Nacional d'Anàlisi Genòmica. C. de Baldiri Reixac, 4. 08028 Barcelona. Tel.: +34 934 037 289. Adreça electrònica: *berta.fuste@cnag.crg.eu.*

DOI: 10.2436/20.1501.02.213
ISSN (ed. impresa): 0212-3037
ISSN (ed. digital): 2013-9802
http://revistes.iec.cat/index.php/TSCB
Rebut: 24/01/2022
Acceptat: 16/03/2022

Resum

L'any 1977 Frederick Sanger va desenvolupar un mètode per a determinar l'ordre de les bases dels fragments de DNA. Aquesta tecnologia encara s'utilitza actualment i ha sigut clau per a aconseguir fites tan importants com la primera seqüenciació completa del genoma humà. L'aparició d'una nova generació de tecnologies de seqüenciació del DNA, tecnologies d'NGS (de l'anglès *next generation sequencing*), més la gran explosió d'eines computacionals per a analitzar-lo, ha permès seqüenciar de manera ràpida, econòmica i amb una elevada precisió genomes de microbis, plantes i animals. Durant l'última dècada, hi ha hagut una gran expansió de plataformes d'NGS: primer van sorgir les tecnologies d'NGS de cadena curta i, més endavant, les de cadena llarga. Tot i que les tecnologies d'NGS de lectures llargues prometien grans avenços per a resoldre genomes complexos, les freqüències d'error d'aquestes tecnologies són elevades. Això ha fet que, en els últims anys, hagin aparegut una sèrie de mètodes de seqüenciació complementaris per a resoldre les deficiències de les tecnologies d'NGS de lectures curtes i llargues.

Paraules clau: genoma, seqüenciació de nova generació (NGS), seqüenciació per síntesi (SBS), seqüenciació de molècules úniques en temps real (SMRT-seq), seqüenciació per nanoporus (ONT).

Advances in DNA sequencing technology

Abstract

In 1977, Frederick Sanger developed a method for determining the order of the bases of DNA fragments. This technology still works today and has been crucial in achieving such important milestones as the first complete sequencing of the human genome. The emergence of the new generation of DNA sequencing technologies (NGS) plus the great explosion of computer tools for their analysis has become a matter of routine and allows the sequencing of the genomes of microbes, plants and animals in a way that is quick, relatively cheap and highly precise. There has been a great expansion of NGS sequencing platforms over the last decade, first involving short-read and later long-read NGS sequencing technologies. Although long-read NGS sequencing promised great advances in solving complex genomes, the error rates of these technologies are high. This has led to the appearance in recent years of a number of complementary sequencing methods to address the shortcomings of NGS sequencing of short and long readings.

Keywords: genome, next-generation sequencing (NGS), sequencing by synthesis (SBS), single-molecule real-time sequencing (SMRT-seq), nanopore sequencing (ONT).

Introducció a la genòmica

El genoma és el conjunt del material genètic de les cèl·lules d'un organisme que s'emmagatzema en forma d'àcid desoxiribonucleic o DNA i que conté tota la informació per al seu desenvolupament i funcionament correctes. Anomenem *gens* els segments de DNA que confereixen instruccions específiques a la cèl·lula, sovint mitjançant la síntesi de proteïnes a partir de molècules intermèdies anomenades *ARN missatgers*. Les proteïnes són les que formen els òrgans i teixits del cos, i controlen les reaccions químiques i la comunicació entre les cèl·lules.

Tal com van descriure James Watson i Francis Crick l'any 1953 gràcies a les observacions prèvies de Rosalind Franklin, la molècula de DNA està formada per dues cadenes d'unes unitats químiques que anomenem *nucleòtids*, enrotllades al voltant d'un eix comú formant una doble hèlix (Watson i Crick, 1953). Hi ha quatre tipus de nucleòtids, que identifiquem amb les lletres A, T, G i C (adenina, timina, guanina i citosina, respectivament). Les dues cadenes queden unides per ponts d'hidrogen formats entre la base d'una cadena i la de l'altra cadena, amb la qual queda enfrontada. Els aparellaments sempre són entre A-T i G-C. El DNA dels individus d'una mateixa espècie varia en un petit percentatge (<1 %); hi trobem canvis d'un únic nucleòtid (*single nucleotide variants* o SNV), però també reordenaments genòmics o variants estructurals (*structural variants* o SV) que afecten centenars o milers de bases, com ara inversions, delecions o duplicacions. Aquestes variacions en el DNA són responsables de les diferències entre les persones, i en alguns casos també poden donar lloc a malalties diverses.

El genoma humà té uns tres mil milions de nucleòtids i al voltant de vint-i-cinc mil gens, cada un dels quals dona lloc a tres proteïnes diferents de mitjana. La primera seqüència del genoma humà, és a dir, l'ordre de pràcticament tots els nucleòtids en la cadena del DNA, es va completar l'any 2003, en el marc d'una col·laboració internacional, el Projecte Genoma Humà (*http://www.genome.gov/10001772*). Es tracta d'un dels assoliments més importants de la biologia; el projecte es va dur a terme en tretze anys i es calcula que va tenir un cost d'uns tres mil milions de dòlars (NHGRI, 2020).

L'any 1977, un investigador de la Universitat de Cambridge, Frederick Sanger, va desenvolupar un mètode per a determinar l'ordre dels nucleòtids i obtenir-ne la seqüència de DNA (Sanger *et al.*, 1977). Es basa en l'ús d'un enzim, la DNA polimerasa, per a generar noves cadenes a partir de la cadena que es vol seqüenciar. En aquesta síntesi, es generen fragments que acaben en les quatre possibles bases del DNA, cadascuna marcada amb una molècula fluorescent diferent. Aquests fragments se separen després segons la seva mida en una matriu porosa i es detecten mitjançant el senyal fluorescent que emeten. L'any 1980 Sanger va ser guardonat amb el Premi Nobel de Química per aquest descobriment. Actualment, el mètode de Sanger continua sent el

més adequat per a seqüenciar petites regions de DNA en moltes mostres o, fins i tot, per a validar resultats obtinguts amb tecnologies més noves (Hert *et al.*, 2008).

Evolució de la nova generació de tecnologies de seqüenciació

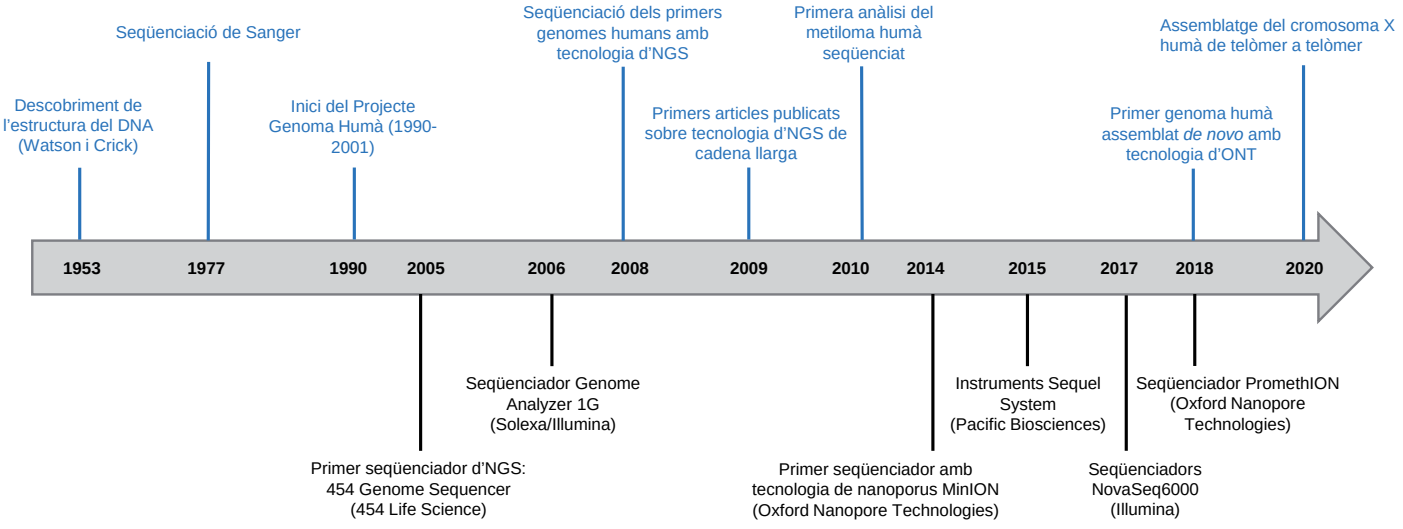
L'any 2004 hi va haver un canvi de paradigma en el camp de la seqüenciació del DNA i de la genòmica, amb l'aparició d'una nova generació de tecnologies de seqüenciació (tecnologies d'NGS, de l'anglès *next negeneration sequencing*) o seqüenciació massiva. Les tecnologies d'NGS combinen l'ús de tècniques d'enginyeria genètica, la nanotecnologia i la generació de milions de dades basades en la imatge. A diferència de la seqüenciació pel mètode de Sanger, que es basa en l'anàlisi individual de fragments de DNA, els seqüenciadors d'NGS són capaços d'analitzar milions de fragments de DNA en paral·lel i, en conseqüència, de seqüenciar un genoma humà sencer en pocs dies. Generen una quantitat de dades genòmiques impensable fa vint anys, de qualsevol organisme, en coneguem o no el genoma prèviament, i de manera sòlida. Les tecnologies d'NGS, juntament amb la gran explosió d'eines computacionals a finals de la primera dècada del segle XXI, com ara els programes informàtics per a alinear milions de lectures curtes en genomes de referència o per a detectar variants genètiques, han provocat un increment sense precedents de les dades de seqüenciació, a una velocitat superior a la llei de Moore, segons la qual la capacitat dels ordinadors es dobla cada any.

Amb l'ús d'aquestes noves tecnologies, el 2005 es van seqüenciar dos genomes bacterians sencers (Shendure *et al.*, 2005; Margulies *et al.*, 2005), i el mateix any l'empresa biotecnològica 454 Life Sciences va comercialitzar el primer equip de seqüenciació massiva, el 454 Genome Sequencer, capaç de produir lectures d'uns 110 parells de bases (pb) i un total de 20 megabases (Mb) en cada carrera (unitat de funcionament de cada seqüenciador). Roche Diagnostics va adquirir 454 Life Sciences i va comercialitzar-ne els instruments de seqüenciació durant uns anys, però actualment aquesta plataforma ja no està en funcionament.

L'any 2006, l'empresa Solexa, una empresa derivada de la Universitat de Cambridge, va llançar al mercat el seqüenciador Genome Analyzer 1G, que permetia seqüenciar 1 gigabase (Gb) de seqüència en una única carrera. La multinacional Illumina va adquirir Solexa poc després i des d'aleshores s'han seqüenciat una gran quantitat de genomes de microbis, plantes i animals amb aquesta tecnologia, que és avui en dia la principal tecnologia d'NGS de cadena curta. El primer genoma humà seqüenciat per NGS va ser el de James Watson el 2008, que ho va fer utilitzant els equips Genome Analyzer 1G, amb un cost d'un milió de dòlars i en un període de temps de dos mesos. Des del seu inici, el 2005, fins avui, Illumina ha dedicat tots els seus esforços a incrementar la producció dels seus seqüenciadors al mateix temps que es reduïa el cost per gigabase de dades generades. El NovaSeq6000 és, ara mateix, el seqüenciador més potent del mercat, capaç de produir 6 terabases (Tb) de seqüència i 20 bilions de lectures en menys de 48 h, l'equivalent a 48 genomes humans a una cobertura 30 ×.

El 2009 van sorgir dues tècniques d'NGS de cadena llarga que utilitzen aproximacions molt diferents: la seqüenciació de molècules úniques en temps real (SMRT), de Pacific Biosciences (PacBio), i la seqüenciació basada en nanoporus, d'Oxford Nanopore Technologies (ONT). PacBio va comercialitzar el primer instrument el 2015. En l'actualitat, PacBIO opera amb tres seqüenciadors de tecnologia d'NGS de lectures llargues, els Sequel Systems. Els seqüenciadors més grans són els Sequel II i Sequel IIe, amb una capacitat de seqüenciar 8 milions de molècules alhora. Pel que fa a ONT, el 2014 va comercialitzar el primer seqüenciador, el MinION, capaç de produir entre 5 i 10 Gb de seqüència per carrera. Gràcies a les millores continuades en els mètodes de preparació de biblioteques i en la química del MinION, el 2018 es va seqüenciar el primer genoma humà amb lectures ultrallargues (>100 kb). Avui dia, l'empresa comercialitza dos seqüenciadors més, que són més potents: el GridION, que pot generar entre 50 i 250 Gb de seqüència per carrera, i el més potent dels instruments, el PromethION, capaç de generar entre 1.000 i 2.000 Gb de seqüència en una sola carrera.

Durant l'última dècada hem vist una gran expansió de plataformes de seqüenciació, i seqüenciar el genoma d'un vertebrat ja és una cosa rutinària. Tot i així, l'assemblatge de quasi tots els genomes diploides continua sent incomplet i altament fragmentat. En els últims anys, juntament amb les tecnologies d'NGS, han aparegut mètodes de seqüenciació complementaris, com el Hi-C (Bonev i Cavalli, 2016) o el mapatge òptic (Giani *et al.*, 2019) per a resol-



↑ Figura 1. Fites clau en les tecnologies de seqüenciació del DNA. Elaboració pròpia.

dre les deficiències de l'NGS de lectures curtes i llargues. De fet, s'ha acabat demostrant que la combinació d'aquestes tècniques pot solucionar les limitacions de cada una per separat. El 2020 es va publicar per primera vegada l'assemblatge d'un cromosoma humà, concretament el cromosoma X, de telòmer a telòmer (T2T) sense haver-hi cap buit per resoldre (Miga *et al*., 2020) (vegeu la figura 1).

Principis bàsics de les tecnologies d'NGS

Seqüenciació per síntesi (SBS), d'Illumina

Avui dia, quan parlem de tecnologies d'NGS de lectures curtes o NGS de cadena curta, parlem de seqüenciació de DNA amb instruments d'Illumina (*www.Illumina.com*). Més del 90 % de les dades de seqüenciació al món es generen mitjançant aquesta plataforma. Es parla de *seqüenciació de cadena curta* quan els fragments que se seqüencien van entre 50 i 300 pb.

La tecnologia d'Illumina sintetitza la nova cadena de DNA usant el mètode de seqüenciació per síntesi (SBS), que consisteix en la fabricació d'una cadena de DNA complementària a una cadena motlle mitjançant la DNA polimerasa. Tot i que la tecnologia d'Illumina està basada en l'aproximació de Sanger, difereix en la longitud de les lectures, que són més curtes, però, sobretot, en la capacitat d'analitzar milions de fragments de DNA de manera massiva i en paral·lel. La taxa d'error amb aquest mètode és més alta que amb el de Sanger, però es veu compensada per la gran capacitat de generar moltíssimes dades de seqüència alhora i, per tant, incrementar la cobertura de cada base seqüenciada. Per aquesta raó, la química d'Illumina és considerada seqüenciació d'alta resolució i precisió (precisió del 99,9 %). El 80 % de les bases seqüenciades per Illumina presenten un valor de qualitat *Phred quality score* igual o més alt de Q30, és a dir, que la probabilitat d'identificar erròniament una base és d'una entre un miler. Aquest nivell de fiabilitat és ideal per a abordar tot el ventall d'aplicacions d'NGS, incloent-hi les de diagnòstic clínic.

En general, l'NGS implica quatre passos bàsics, que es divideixen en: 1) la preparació de la biblioteca, que consisteix en la lligació de seqüències curtes conegudes, anomenades *adaptadors*; 2) la generació de milers de molècules de DNA idèntiques, o sigui la immobilització i clonació de les molècules de DNA que es vol seqüenciar; 3) la seqüenciació, i 4) l'anàlisi de les dades.

El primer pas en la seqüenciació d'Illumina consisteix a trencar el DNA en fragments més manejables d'entre 200 i 600 pb. Als fragments de DNA se'ls uneixen unes seqüències curtes conegudes, anomenades *adaptadors*. Aquests adaptadors tenen tres funcions, que són la clau en els mètodes d'NGS de cadena curta. En primer lloc, serveixen per a immobilitzar les seqüències de DNA en una superfície sòlida (*flowcells* o FC) que contenen nanopous, on s'amplifica i se seqüencia el DNA. En segon lloc, s'empren per a replicar les seqüències ancorades i produir milers de molècules de DNA idèntiques. Aquest procés es coneix amb el nom de *bridge PCR amplification* i és necessari per a després emetre un senyal prou fort per a ser detectat per una càmera. I en tercer lloc, aquests adaptadors són la seqüència complementària a l'encebador que, juntament amb la DNA polimerasa, elongaran la cadena i, per tant, generaran la seqüència que després llegirem (Goodwin *et al*., 2016; Barton *et al*., 2018).

Igual que el mètode de Sanger, Illumina utilitza la incorporació de nucleòtids modificats (dNTP, de l'anglès *deoxynucleotide triphosphates*) i reversibles durant diferents cicles consecutius, de manera que, una vegada incorporats, impedeixen l'elongació de la cadena de DNA. Els nucleòtids modificats són marcats amb un fluorocrom diferent que en ser excitat per un làser dona diferents longituds d'ona. L'emissió del senyal generat és capturada per una càmera i emmagatzemada en un ordinador. El procés de seqüenciació d'Illumina és un procés cíclic. A cada cicle de seqüenciació s'incorpora un únic dNTP a la molècula de DNA. La incorporació d'aquest dNTP, juntament amb el senyal emès pel fluoròfor, queda enregistrat per la càmera i, per tant, guardat a l'ordinador. Al final de cada cicle hi ha un trencament del grup que bloqueja a 3' del nucleòtid i de l'etiqueta fluorescent que permetrà la incorporació del següent nucleòtid en el cicle posterior, i així successivament. El nombre de cicles es repeteix *n* vegades i és equivalent a la longitud de lectures seqüenciades, lectures de *n* bases de longitud. La seqüència de DNA s'analitza base a base durant la seqüenciació d'Illumina, per la qual cosa és un mètode molt precís. Una vegada acabat el procés de seqüenciació, la seqüència generada es pot alinear amb una seqüència de referència per a buscar coincidències o canvis en el DNA seqüenciat.

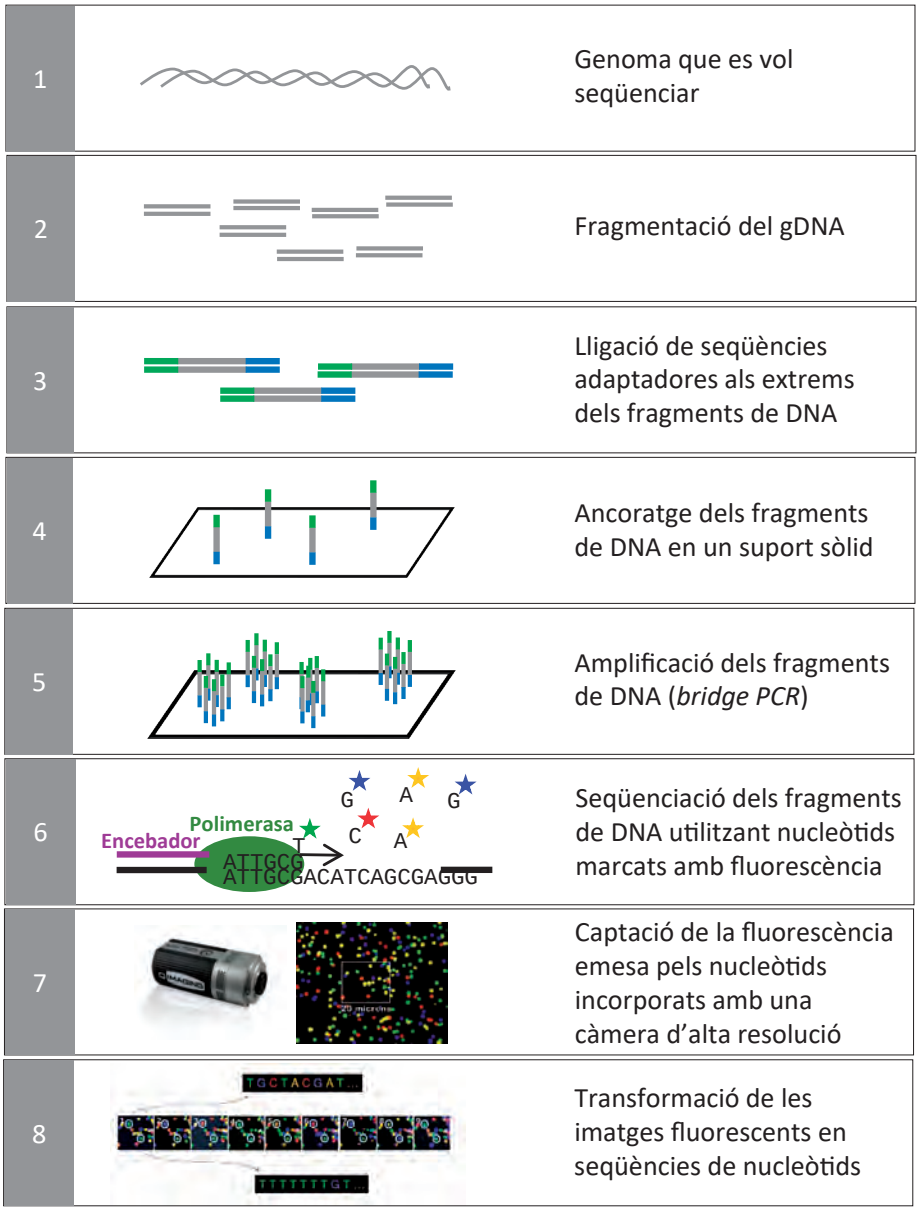
Seqüenciació de molècules úniques en temps real (SMRT), de PacBio

La seqüenciació de molècules úniques en temps real (SMRT, de l'anglès *single-molecule real-time*) és una tècnica d'NGS de lectures llargues que permet seqüenciar milions de molècules llargues de DNA al mateix temps utilitzant el procés natural de la replicació del DNA.

La tecnologia SMRT genera seqüències molt llargues que poden ser de 30-50 quilobases (kb) o inclús més llargues. A més, la tecnologia SMRT, tal com n'indica el nom, sintetitza una única molècula de DNA i ho fa en temps real. A diferència de la tècnica d'NGS de cadena curta, SMRT no cal que amplifiqui fragments de DNA ni de cicles químics repetitius per a elongar la cadena, fet que afavoreix l'eliminació de tots els errors sistemàtics induïts per la mateixa amplificació del DNA.

La tecnologia SMRT parteix de DNA circular de cadena simple. Al DNA se li uneixen uns adaptadors de cadena doble a cada extrem que en permeten la circularització i també l'ancoratge de les molècules a la DNA polimerasa. La reacció de seqüenciació es produeix sobre un suport sòlid (*flowcell*, SMRT FC) que conté milions de nanopous (pous ZMW, *zero-mode waveguide*). El pou ZMW és una cavitat de desenes de nanòmetres de diàmetre que es fabrica en una pel·lícula metàl·lica de 100 nm dipositada sobre un substrat de vidre i on la llum no pot entrar. A la superfície de vidre inferior de cada pou ZMW hi ha ancorada una DNA polimerasa. En el procés de seqüenciació s'afegeix a cada pou ZMW una única molècula de DNA circular, més els quatre nucleòtids marcats amb fluorescència. A mesura que la DNA polimerasa incorpora nucleòtids a la cadena, s'alliberen els fluoròfors que, juntament amb l'excitació per làser, emeten diferents longituds d'ona. Cada longitud d'ona s'identifica amb una de les bases. El procés es fa en temps real i és enregistrat pel mateix instrument en format de vídeo (Eid *et al*., 2009; Ansorge *et al*., 2017).

Un dels avantatges de la tecnologia d'NGS de cadena llarga SMRT és la capacitat de sintetitzar seqüències de diverses kilobases (10-25 kb) amb la resolució necessària per a resoldre de manera senzilla zones del genoma amb un alt nombre d'elements repetitius o identificar les diferents isoformes d'un mateix gen. Per contra, un dels principals desavantatges que té és la taxa d'error, que és força més alta que el de les tecnologies d'NGS de cadena curta. En aquest sentit, PacBio ha desenvolupat dues es-



↑ Figura 2. Flux de seqüenciació per síntesi (SBS), d'Illumina. Elaboració pròpia.

tratègies per a resoldre aquest problema: una d'enfocada a obtenir lectures molt llargues (CLR, de l'anglès *continous long reads*) i un nou mètode anomenat d'*alta fidelitat* (HiFi), en què la mateixa molècula de DNA és seqüenciada diverses vegades, de manera que la seqüència final és un consens d'alta qualitat autocorregit (CCS, de l'anglès *circular consensus sequencing*) que aconsegueix lectures molt precises (99,8 %) (Taishan *et al.*, 2021).

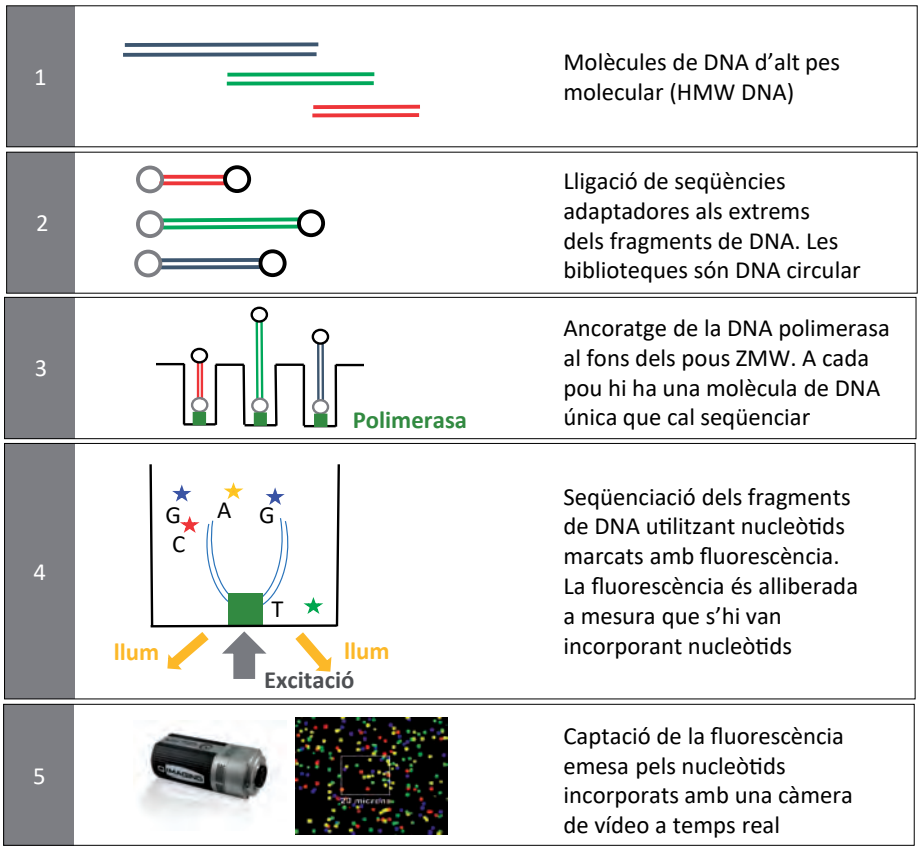
Seqüenciació per nanoporus, d'Oxford Nanopore Technologies (ONT)

Oxford Nanopore Technologies (ONT, *www.nanoporetech.com*) ha desenvolupat una tecnologia d'NGS de cadena llarga que permet seqüenciar molècules de DNA i RNA en temps real. De la mateixa manera que la tecnologia SMRT, la seqüenciació per nanoporus (ONT) també pot seqüenciar molècules úniques de DNA i RNA sense passos d'amplificació. La seqüenciació d'ONT és el mètode d'NGS que aconsegueix les lectures de seqüència més llargues: pot generar seqüències de més d'1 Mb de longitud i arribar a les 2 Mb per fragment amb ajuda computacional.

La seqüenciació d'ONT es basa a passar una molècula de DNA o RNA per uns petits forats, anomenats *nanoporus*, que es troben incrustats en una membrana sintètica. Tots els dispositius de seqüenciació d'ONT utilitzen uns suports sòlids (FC) que contenen centenars d'aquests nanoporus, l'un al costat de l'altre. Cada nanoporus conté el seu propi elèctrode connectat a un sensor que mesura el corrent elèctric que flueix a través del nanoporus. Quan una molècula passa a través d'un nanoporus, el corrent canvia i produeix un senyal elèctric característic per cada nucleòtid o grup de nucleòtids. Aquest senyal elèctric es descodifica utilitzant algoritmes bioinformàtics i determina l'ordre de les bases seqüenciades (Jain *et al.*, 2018; Lin *et al.*, 2021).

La seqüenciació d'ONT analitza tota la cadena de DNA i RNA que passa pel nanoporus; és a dir, la longitud de les seqüències és equivalent a la longitud de la mostra inicial que volem processar. La longitud de les lectures, doncs, és condicionada pel protocol que s'utilitza per a extreure el DNA o l'RNA, però també pel protocol utilitzat en la preparació de les biblioteques. En treballar amb molècules de DNA o RNA sense manipular, el procés de preparació de biblioteques és molt simple i ràpid. A les molècules de DNA o RNA se'ls afegeix un adaptador de cadena senzilla i una proteïna motora. Aquesta cadena senzilla permet l'entrada d'una de les dues cadenes de la molècula de DNA dins del porus amb l'ajut de la proteïna motora que comença el procés d'elongació de la seqüència.

La seqüenciació d'ONT es produeix en temps real, ja que, a mesura que se seqüencia, es pot anar llegint la seqüència generada en un ordinador. Aquest pas és un avantatge respecte d'altres sistemes de seqüenciació perquè la qualitat de la seqüència pot ser validada en el moment precís i no cal esperar que l'experiment finalitzi per a veure'n els resultats. Un altre avantatge és la mida reduïda dels instruments utilitzats. Un MinION és més petit que un telèfon mòbil, la qual cosa fa que sigui totalment transportable i que pugui connectar-se directament a qualsevol ordinador per mitjà d'un port USB. Ara bé, igual que la tecnologia SMRT, la seqüenciació d'ONT, tot i produir seqüències molt llargues, presenta valors de qualitat de seqüència inferiors a les plataformes d'NGS de cadena curta. Això es pot resoldre mitjançant estratègies de combinació de tecnologies d'NGS de lectures llargues i curtes, tot i que recentment ONT està desenvolupant reactius d'alta fiabilitat per a la seva tecnologia, Q20+, que incrementa la fiabilitat de la seqüenciació i, en conseqüència, la precisió de les lectures s'aproxima a nivells similars als de l'NGS de cadena curta (Taishan *et al.*, 2021).



↑ Figura 3. Flux de seqüenciació de molècules úniques en temps real (SMRT), de PacBio. Elaboració pròpia.

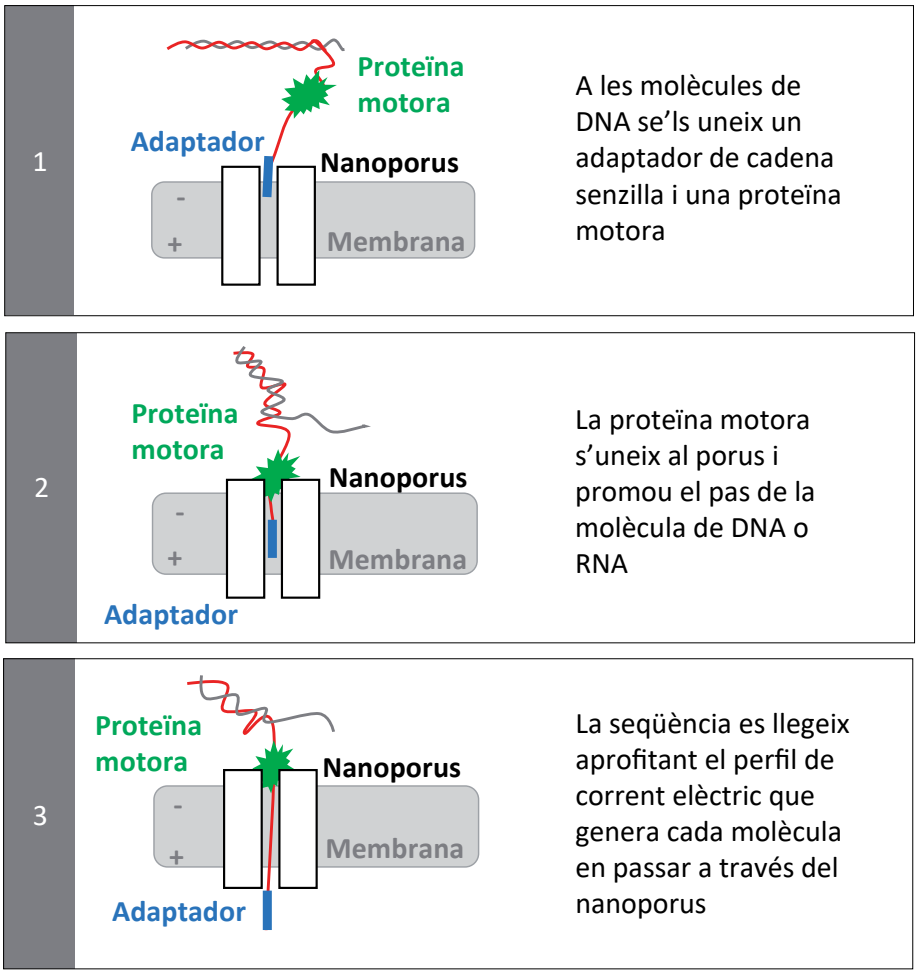
Principals aplicacions de les tecnologies d'NGS

Les tecnologies d'NGS possibiliten l'estudi del material genètic a un nivell de resolució sense precedents i amb protocols molt diversos segons l'objectiu de l'estudi (Buermans *et al.*, 2014). Actualment, i gràcies a la gran disminució del seu cost, la seqüenciació de genomes sencers (WGS, de l'anglès *whole genome sequencing*) és una de les aplicacions d'NGS més utilitzades. Permet obtenir una visió completa de tot el genoma, comparar-la amb el genoma de referència de l'espècie i identificar-ne les variants de seqüència (SNV, SV i CNV), que en determinen les característiques fenotípiques, incloent-hi les responsables d'algunes malalties. La seqüenciació del genoma d'espècies d'interès s'utilitza també en agrigenòmica per a accelerar els processos de millora genètica en plantes i animals.

En l'entorn clínic, la seqüenciació d'exomes sencers (WES, de l'anglès *whole exome sequencing*) o de panells de gens (panel-seq) d'interès per a una malaltia són les aplicacions d'NGS més comunes (Manolio *et al.*, 2019). S'utilitzen mètodes basats en la hibridació de fragments de DNA per a capturar regions genòmiques específiques, com ara el conjunt de regions codificants de tot el genoma o l'exoma, usant sondes de DNA que són complementàries a les regions dianes. Aquests fragments de DNA enriquits en les regions d'interès després se seqüencien amb tècniques d'NGS. En comparació amb la seqüenciació de genomes sencers, la seqüenciació d'exomes o de panells de gens genera una quantitat de dades més manejable i permet obtenir més lectures de les regions codificants, que són les que concentren la gran majoria de mutacions que causen malalties, i tot a un preu més reduït.

Les tecnologies d'NGS permeten també estudiar els mecanismes de regulació del genoma. L'ús de *microarrays* per a l'estudi de l'expressió gènica ha donat pas a la seqüenciació de transcriptomes sencers (RNA-seq) per a quantificar de manera precisa l'expressió dels gens en una mostra i identificar-ne els fenòmens de *splicing* alternatiu i preservar-ne la informació sobre la cadena que es transcriu. Hi ha mètodes específics per a seqüenciar l'RNA total (*coding RNA* i *long-noncoding RNA* [lncRNA]), l'RNA missatger (mRNA) i els RNA de mida petita (*small RNA*) (Stark *et al.*, 2019). La caracterització de modificacions epigenètiques, com la detecció de la metilació de les citosines, també es fa actualment per NGS. El mètode més utilitzat es basa en un tractament amb bisulfit sòdic que converteix les citosines no metilades en uracils, però deixa les citosines metilades intactes. Quan se seqüencia el DNA tractat d'aquesta manera (WGBS, de l'anglès *whole-genome bisulfite sequencing*), s'obtenen dades precises sobre la freqüència de la metilació de totes les citosines del genoma. També podem identificar les interaccions entre determinades proteïnes i el DNA utilitzant tècniques d'immunoprecipitació seguides d'NGS (ChIP-seq), o alternativament estudiar les regions no protegides pels nucleosomes mitjançant l'ús de transposases o d'altres estratègies que permeten enriquir els fragments de DNA accessibles de la cromatina (ATAC-seq, FAIRE-seq). Finalment, s'han desenvolupat protocols específics per a explorar l'organització tridimensional de la cromatina al nucli, com ara el Hi-C. En la seqüenciació Hi-C s'estabilitzen els complexos DNA-proteïna amb formaldehid, s'enriqueixen aquells fragments que es troben en un espai proper en el nucli de la cèl·lula i se seqüencien, de manera que s'obté un mapa precís de les interaccions de la cromatina.

Hi ha hagut avenços en els darrers deu anys que permeten seqüenciar el DNA, el transcriptoma o l'epigenoma de cèl·lules individuals (scDNA-seq, de l'anglès *single cell DNA sequencing*; scRNA-seq, de l'anglès *single cell RNA sequencing*, i scWGBS, de l'anglès *single cell whole genome bisulfite sequencing*), i que han revelat una elevada heterogeneïtat de tipus cel·lulars en molts teixits i noves poblacions de cèl·lules, la qual cosa afegeix una nova capa de complexitat en molts processos biològics i malalties (Anaparthy *et al.*, 2019). El primer pas consisteix a aïllar les cèl·lules individuals utilitzant mètodes de separació de les cèl·lules per citometria de flux (FACS, de l'anglès *fluorescence activated cell sorting*), micromanipulació, microdissecció per captura làser (LCM, de l'anglès *laser capture microdissection*) o sistemes de microfluids. Hi ha equips, com ara els desenvolupats per l'empresa 10X Genomics, que permeten analitzar desenes de milers de cèl·lules individuals en paral·lel, que queden encapsulades en petites gotes en les quals tenen lloc les primeres reaccions: lisi cel·lular, transcripció inversa en el cas de l'RNA i



↑ Figura 4. Flux de seqüenciació per nanoporus, d'Oxford Nanopore Technologies (ONT). Elaboració pròpia.

el marcatge de tots els fragments de material genètic que provenen d'una mateixa cèl·lula utilitzant oligonucleòtids específics. A partir d'aquí, se n'extreuen els àcids nucleics, s'amplifica tot el genoma, les regions d'interès o el transcriptoma, es preparen les biblioteques i se seqüencien per NGS. Es tracta d'un camp en evolució contínua, en el qual cada mes es publiquen protocols nous que permeten analitzar moltes més cèl·lules i que minimitzen els biaixos causats per l'ínfim material de partida.

Més del 90 % de les dades de seqüenciació es generen amb la tecnologia i els instruments de l'empresa Illumina. La seva elevada capacitat de producció, precisió i el baix cost que té per base fan que sigui la tecnologia més adequada per a dur a terme la major part de les aplicacions descrites anteriorment. Les lectures relativament curtes dels instruments d'Illumina (50-150 pb) són, però, insuficients per a resoldre zones complexes del genoma. Per aquest motiu, cal recórrer a les plataformes de lectures llargues (10-40 kb de mitjana) quan no coneixem el genoma de referència d'una espècie (*de novo* WGS), o per a identificar variants estructurals o les diverses isoformes d'un mateix gen. Tot i així, per a obtenir el genoma de referència d'una espècie diploide normalment es combinen la seqüenciació genòmica amb lectures curtes i llargues juntament amb altres mètodes complementaris com el Hi-C o el mapatge òptic (Giani *et al.*, 2019; Graham *et al.*, 2020).

Darrers desenvolupaments en tecnologies de seqüenciació

El gran potencial de les tecnologies d'NGS i l'impacte que han generat en tots els camps de la biologia han desencadenat un gran interès en l'ampliació de noves tecnologies i aplicacions enfocades a augmentar el rendiment, la velocitat i la precisió de la seqüenciació. En aquest sentit, avui dia hi ha un gran ventall de noves tecnologies o modificacions de les existents en desenvolupament.

Un exemple en són els avenços que s'estan fent al voltant de la seqüenciació per nanoporus. La tecnologia de nanoporus, actualment, utilitza nanoporus basats en proteïnes. Amb la finalitat d'augmentar la resolució i la velocitat de seqüenciació, s'estan investigant nous tipus de nanoporus fabricats a partir de l'ús de materials sintètics com són el grafè o el carboni (Wang *et al.*, 2015). També, altres millores van encaminades a fer els instruments cada vegada més petits per tal de fer-los encara més portables; un exemple en seria el nou instrument SmidgION (Oxford Nanopore Technologies), el funcionament del qual aniria lligat a un telèfon mòbil (Kumar *et al.*, 2019).

Durant els últims anys, els progressos en la resolució de la microscòpia òptica, juntament amb l'aparició dels mètodes de seqüenciació de RNA de cèl·lula única, han unit forces i han desencadenat l'aparició de tot un ventall de tècniques que tenen l'objectiu d'estudiar l'expressió gènica en el seu context en l'espai, seqüenciació *in situ* (ISS, de l'anglès *in situ sequencing*). El marcatge de l'expressió gènica es fa directament sobre els teixits i s'utilitza microscòpia òptica per a detectar-la. El marcatge de l'expressió gènica pot ser via hibridació *in situ* (seqFISH), que implica l'ús de múltiples oligonucleòtids amb etiquetes fluorescents que s'uneixen a les molècules d'RNA, o via ISS, on l'RNAm se seqüencia directament en una secció del teixit (Marx *et al.*, 2021).

Tot i així, els darrers avenços en seqüenciació i en microscòpia no només van encaminats a entendre millor el paper de les cèl·lules i la seva ubicació en els processos biològics, sinó que també ens han proporcionat eines per a sondejar l'estructura del mateix genoma. Les tecnologies actuals de microscòpia òptica d'alta resolució estan limitades a observar un grapat de gens o, en el millor dels casos, un 1 % del genoma. Tanmateix, recentment s'ha publicat una nova tecnologia, OligoFISSEQ, que combina la microscòpia d'alta resolució, l'NGS i la hibridació amb oligonucleòtids marcats amb fluorescència, que permetrà visualitzar amb una resolució molecular el genoma sencer (Huy *et al.*, 2020). Continuant en aquesta línia, el Centre Nacional d'Anàlisi Genòmica - Centre de Regulació Genòmica (CNAG-CRG) formarà part del Center for Genome Imaging, una nova infraestructura amb seu a la Universitat de Harvard, amb l'objectiu de desenvolupar tecnologies que permetin la visualització, l'anàlisi i la modelització de tot el genoma humà en 3D i a un nivell extremament d'alta resolució.

Les tecnologies de seqüenciació de DNA a Catalunya: el Centre Nacional d'Anàlisi Genòmica

Catalunya té una reconeguda tradició en l'àmbit de la genòmica i la bioinformàtica. Els principals hospitals, centres de recerca en ciències de la vida i universitats del territori tenen unitats de genòmica amb tecnologies de seqüenciació massiva, generalment de petita o mitjana escala.

A més a més, des de 2009, a Catalunya hi trobem un dels centres europeus de referència en seqüenciació i anàlisi de dades genòmiques, el CNAG-CRG. Gràcies a un equip multidisciplinari d'investigadors, tècnics de laboratori, bioinformàtics i enginyers, el CNAG-CRG ofereix serveis de seqüenciació massiva amb tecnologies de segona i tercera generació i d'anàlisi bioinformàtica de les dades.

Actualment, la plataforma del CNAG-CRG té tres unitats dels seqüenciadors de lectures curtes més potents del mercat, cinc equips de seqüenciació massiva mitjançant nanoporus, diverses plataformes per a fer experiments de seqüenciació a partir de cèl·lules individuals, així com d'altres equips de laboratori complementaris. Aquest parc de seqüenciadors pot generar més de 10.000 Gb de seqüència cada 24 hores, o el que és el mateix, permet seqüenciar cada dia 100 genomes humans sencers amb una cobertura suficient per a identificar-ne de manera fiable les variants en el 95 % del genoma. Per tal de processar aquesta gran quantitat de dades, el centre té un superordinador amb més de 3.500 nodes de computació, 8 petabytes d'espai de disc per a emmagatzemament i una xarxa interna de 56 GB per segon. Tots els processos del centre es fan sota controls de qualitat estrictes, queden enregistrats en el LIMS (*laboratory management information system*) i disposen de la certificació ISO 9001:2015 i l'acreditació ISO 17025:2017. El 2020, el CNAG-CRG va seqüenciar més de 12.000 mostres de DNA o RNA en el marc de 521 projectes de 190 investigadors d'hospitals, universitats i centres de recerca.

Les activitats de recerca i suport a la recerca del CNAG-CRG s'articulen al voltant de sis àrees: la medicina personalitzada, les malalties rares, la genòmica del càncer, la genòmica de cèl·lules individuals, la genòmica funcional i la biodiversitat. Els investigadors del centre participen activament en algunes de les principals iniciatives nacionals i internacionals en aquestes àrees, com ara els projectes IMPaCT de medicina personalitzada de l'Instituto de Salud Carlos III (*www.isciii.es/QueHacemos/Financiacion/IMPaCT/Paginas/default.aspx*), l'International Rare Diseases Research Consortium (*irdirc.org*), el consorci del *Human Cell Atlas* (*www.humancellatlas.org*) i el *European Reference Genome Atlas* (*www.erga-biodiversity.eu*).

En resum, la seqüenciació de DNA i RNA ha esdevingut una eina imprescindible per a la recerca bàsica i aplicada, i la missió del CNAG-CRG és facilitar-ne la implementació al país, oferint equipaments de darrera generació, preus competitius i el saber fer d'experts en diverses disciplines genòmiques.

Bibliografia

Anaparthy, N. [*et al.*] (2019). «Single-cell applications of next-generation sequencing». *Cold Spring Harb. Perspect. Med*., 9 (10): a026898.

Ansorge, W. J. [*et al.*] (2017). «Prespectives for future DNA sequencing techniques and applications». A: *Molecular diagnostics*, cap. 8: 141-153.

Barton, E. [*et al.*] (2018). «Overview of next generation sequencing technologies». *Curr. Protoc. Mol. Biol.*, 122 (1): e59.

Bonev, B.; Cavalli, G. (2016). «Organization and function of the 3D genome». *Nat. Rev. Genet*., 17 (11): 661-678.

Buermans, H. P. J.; Dunnen, J. T. [*et al.*] (2014). «Next generation sequencing technology: Advances and applications». *Biochim. Biophys. Acta*, 1842 (10): 1932-1941.

Eid, J. [*et al.*] (2009). «Real-time DNA sequencing from single polymerase molecules». *Science*, 323 (5910): 133-138.

Giani, A. M. [*et al.*] (2019). «Long walk to genomics: History and current approaches to genome sequencing and assembly». *Comput. Struct. Biotechnol. J.*, 18: 9-19.

Goodwin, S. [*et al.*] (2016). «Coming of age: Ten years of next-generation sequencing technologies». *Nat. Rev. Genet.*, 17: 333-351.

Graham, J. E. [*et al.*] (2020). «Sequencing smart: De novo sequencing and assembly approaches for non-model mammal». *GigaScience*, 9: 1-14.

Hert, D. G. [*et al.*] (2008). «Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods». *Electrophoresis*, 23: 4618-4626.

Huy, Q. [*et al.*] (2020). «3D mapping and accelerated super-resolution imaging of the human genome using *in situ* sequencing». *Nat. Methods*, 17: 822-832.

Jain, M. [*et al.*]. (2018). «Nanopore sequencing and assembly of a human genome with ultra-long reads». *Nat. Biotechnol*., 36 (4): 338-345.

Kumar, K. R. [*et al.*] (2019). «Next-generation sequencing and emerging technologies». *Semin. Thromb. Hemost*., 45 (7): 661-673.

Lin, B. [*et al.*] (2021). «Nanopore technology and its applications in gene sequencing». *Biosensors*, 11: 214.

Manolio, T. A. [*et al.*] (2019). «Genomic medicine year in review: 2019». *Am. J. Hum. Genet.*, 105 (6): 1072-1075.

Margulies, M. [*et al.*] (2005). «Genome sequencing in microfabricated high-density picolitre reactors». *Nature*, 437: 376-380.

Marx, V. [*et al.*] (2021). «Method of the year: Spatially resolved transcriptomics». *Nat. Methods*, 18: 9-14.

Miga, K. H. [*et al.*] (2020). «Telomere-to-telomere assembly of a complete human X chromosome». *Nature*, 585: 79-84.

National Human Genome Research Institute (NHGRI) (2020). «Human Genome Project completion: Frequently asked questions» [en línia]. <www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project> [Consulta: 11 desembre 2022].

Sanger, F. [*et al.*] (1977). «DNA sequencing with chain-terminating inhibidors». *Proc. Natl. Acad. Sci. USA*, 74 (12): 5463-5467.

Shendure, J. [*et al.*] (2005). «Accurate multiplex polony sequencing of an evolved bacterial genome». *Science*, 309 (5741): 1728-1732.

Stark, R. [*et al.*] (2019). «RNA sequencing: The teenage years». *Nat. Rev. Genet*., 20 (11): 631-656.

Taishan, H. [*et al.*] (2021). «Next-generation sequencing tecnologies: An overview». *Hum. Immunol.*, 82: 801-811.

Wang, Y. [*et al.*] (2015). «The evolution of nanopore sequencing». *Front. Genet*., 5: 449.

Watson, J.; Crick, F. (1953). «Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid». *Nature*, 171: 737-738.