

UF9

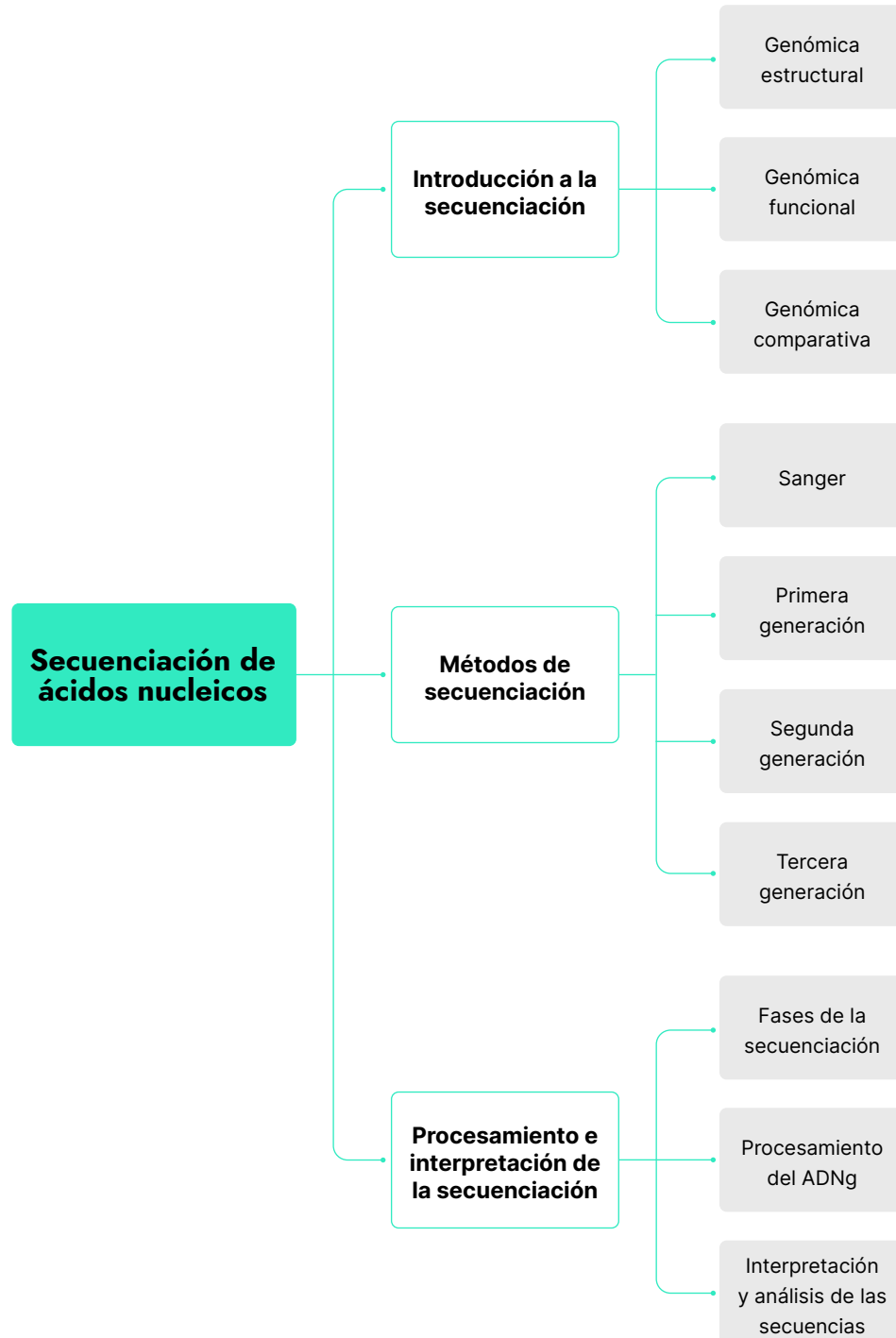
Secuenciación de ácidos nucleicos

Biología molecular y citogenética

ÍNDICE

Mapa conceptual	03
1. Introducción a la secuenciación	04
1.1. Genómica estructural.....	05
1.2. Genómica funcional y comparativa	06
2. Métodos de secuenciación.....	10
2.1. Método de Maxam y Gilbert.....	11
2.2. Método Sanger	12
A. Reacción de polimerización.....	12
B. Separación de los fragmentos	13
C. Detección y lectura de la secuencia	13
2.3. Primera generación.....	14
A. Reacciones de polimerización.....	14
B. Separación de fragmentos	15
C. Detección de la señal.....	15
D. Lectura de la secuencia.....	16
2.4. Segunda generación	18
2.5. Tercera generación.....	21
3. Procesamiento e interpretación de la secuenciación.....	23
3.1. Procesamiento del ADNg	24
3.2. Interpretación y análisis de las secuencias	25

Mapa conceptual



01 Introducción a la secuenciación

El término **genómica** lo definió Thomas Roderick en los años 80, para dar nombre al surgimiento de las técnicas que permitieron el descubrimiento de las secuencias nucleotídicas de multitud de organismos.

El **primer genoma** que se obtuvo fue el del bacteriófago ϕ **X174** (virus que infecta a la bacteria *E. coli*) con un tamaño de 5,4 Kb mediante la tecnología Sanger (véase el siguiente apartado) y más tarde se secuenció el primer genoma de un organismo vivo procariota, la bacteria *Haemophilus influenzae* (1,83 Mb). Desde entonces, esta rama de la genética no ha dejado de evolucionar exponencialmente. De hecho, en la actualidad se han **secuenciado multitud de genomas completos**, bacterias, levaduras, plantas, animales e incluso el genoma del ser humano.

La genómica actual comprende diversos **campos** de estudio:

- » **Genómica estructural.** Incluye la determinación de la secuencia de los genes, su localización en los cromosomas y la construcción de mapas genéticos.
- » **Genómica funcional.** Estudia las funciones de los genes, su regulación (conocida como «transcriptómica») y sus productos finales («proteómica» y «metabolómica»).
- » **Genómica comparativa.** Determina semejanzas o diferencias genéticas entre organismos que permitan detectar mutaciones relacionadas con el fenotipo o el origen evolutivo.

DEFINICIÓN

La **genómica** es la ciencia que permite cartografiar, secuenciar y analizar las funciones de genomas completos. También permite conocer la organización y evolución de la información genética de diferentes organismos.

1.1. Genómica estructural

Gracias al **Proyecto Genoma Humano** (HGP, por sus siglas inglesas) y junto con el avance de las tecnologías del ADN recombinante, de la clonación y de la **secuenciación de ADN**, se ha podido secuenciar todo el ADN humano, es decir, se han identificado todos los nucleótidos (3,1 Gb) que constituyen la información necesaria para la vida de nuestra especie.

El **HGP** se originó en 1990 de la mano de James Watson con el **objetivo de secuenciar y cartografiar cada uno de los 23 cromosomas** del ser humano. Cabe destacar que, hasta entonces, ya se habían secuenciado cientos de genomas de procariotas y eucariotas, pero con tamaños de genoma más reducidos. Después de diez años, en 2001 se presentó el primer borrador del genoma humano que contaba con el 96% de genoma secuenciado y, dos años más tarde, se publicó la secuencia completa del genoma con un coste total de 2.700 millones de dólares. Este genoma contenía multitud de cromosomas con huecos donde se desconocía las secuencias, eran zonas de **heterocromatina** con alta cantidad de repeticiones.

Este proyecto reveló las siguientes **características del genoma humano**:

- » Contiene más de 3.000 millones de nucleótidos, de los cuales solo el 2% contiene secuencias codificantes.
- » Contiene un número total de genes de entre 25.000 y 30.000, de los 100.000 genes predichos.
- » Del total de genes identificados, en el 40% de ellos no se había encontrado ninguna función molecular.
- » Los genes se distribuyen aleatoriamente entre los cromosomas, donde el cromosoma 19 es el de mayor densidad génica y los cromosomas 13 e Y son los de menor densidad.
- » Todos de los genes contenían secuencias intrónicas, sin contar con los genes de las histonas.

Actualmente, ya se conoce con más precisión el genoma humano. De hecho, de los miles de millones de nucleótidos secuenciados, se han identificado **57.771 genes**, 20.015 proteínas y 20.769 transcritos no codificantes, según muestra el reciente estudio publicado a **principios del 2022 por el Consorcio T2T**, por el cual se ha podido cartografiar el 100% del genoma humano, llamado **T2T-CHM13**.

Hay que añadir que el genoma **T2T-CHM13** se secuenció a partir de una célula haploide con el cromosoma X, lo que significa que faltaría finalizar de cartografiar el **cromosoma Y**.

N.º DE CROMOSOMA	TAMAÑO (MB)	N.º DE GENES	N.º DE CROMOSOMA	TAMAÑO (MB)	N.º DE GENES
1	248,4	6.865	13	113,6	1.990
2	242,7	5.049	14	101,2	2.528
3	201,1	3.934	15	99,7	2.508
4	193,6	3.206	16	96,3	2.468
5	182	3.466	17	84,3	3.048
6	172,1	4.006	18	80,5	1.253
7	160,6	3.735	19	61,7	2.992
8	146,3	2.952	20	66,2	1.661
9	150,7	3.054	21	45	1.207
10	134,7	2.890	22	51,3	1.613
11	135,1	3.509	X	154,2	2.587
12	133,3	3.069	Y	62,4	1.431
MT	0,016	37			

Tabla 1. Genoma T2T-CHM13 con longitud y número de genes de cada cromosoma. MT hace referencia al cromosoma mitocondrial.

1.2. Genómica funcional y comparativa

Tras secuenciar el genoma humano, el siguiente paso del HGP fue **determinar la función de los genes** que se habían identificado. Para algunos pocos de estos genes ya se conocía su función, pero poco se sabía del resto. Los investigadores del HGP utilizaron las siguientes herramientas de análisis para poder asignar la función a los genes cuya función era desconocida:

- » **Comparar la secuencia del ORF** con la de otro gen similar perteneciente a otro organismo.

ORF en inglés significa **Open Reading Frame**, que se puede traducir en español como “marco abierto de lectura”. Es un fragmento de ADN que no incluye codón de terminación, es decir, que no tiene ninguna señal que para que se detenga la síntesis de una proteína durante la traducción del ARN. El sitio de pausa de

terminación de la transcripción se localiza después del ORF, más allá del codón de parada de la traducción, ya que si la transcripción cesara antes del codón de terminación, se haría una proteína incompleta durante la traducción. Si se detecta que las secuencias de los dos ORF que se comparan tienen un alto grado de similitud, se puede pensar que estas secuencias han sido sujetas a presiones evolutivas similares y pueden desempeñar funciones biológicas importantes en ambos organismos.

- » **GenBank** es una base de datos pública que almacena secuencias de nucleótidos de ADN para alrededor de 400,000 especies. Estas secuencias provienen de laboratorios individuales y grandes proyectos de secuenciación. Está gestionada por el Centro Nacional de Información Biotecnológica (NCBI) y permite el intercambio de datos con otras bases de datos internacionales. En ella se pueden enviar secuencias o buscarlas y realizar búsquedas de similitud mediante la herramienta BLAST.

Buscar secuencias en el gen que **codifican dominios proteicos** ampliamente conocidos, como regiones reguladoras, canales de iones, etc.

IMPORTANTE

Todos los genes codificantes están formados por una **pauta de lectura abierta u ORF**. Son secuencias que empiezan con un triplete ATG y finalizan con una secuencia de TGA, TAA o TAG.

En la genómica funcional, el proceso de **anotación** consiste en identificar los posibles genes de una secuencia, clasificándolos en **genes codificantes** (generan proteína) y genes **no codificantes** (solo se transcriben, no se traducen, como el ARNt). También se identifican los **elementos móviles del genoma, los pseudogenes y el ADN satélite** (véase la Unidad 10).

Para esta anotación hizo falta comparar con otros genomas de **organismos modelos**, por lo que el HGP estimuló el estudio y la secuenciación de otros organismos eucarióticos como: *Saccharomyces cerevisiae* (12 Mb y 6.300 genes), *Caenorhabditis elegans* (97 Mb y 19.100 genes), *Drosophila melanogaster* (120 Mb y 13.600 genes), *Arabidopsis thaliana* (125 Mb y 25.500 genes) y *Mus musculus* (2,5 Gb y 30.000 genes).

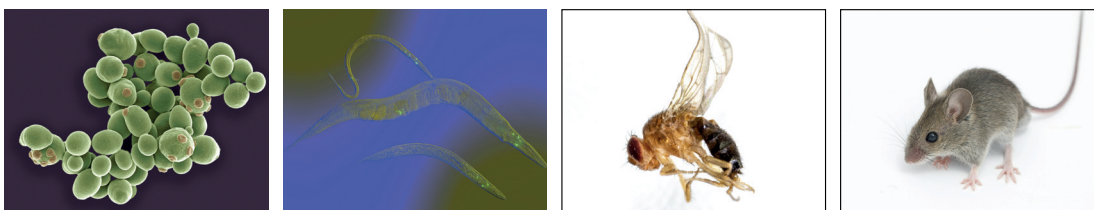


Fig. 1. Organismos modelos utilizados en estudios de genómica comparativa. De izquierda a derecha: Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster y Mus musculus.

La **genómica comparativa** ha permitido, además de ayudar en la identificación de la **función genética**, conocer los **procesos evolutivos** que han ocurrido entre especies filogenéticamente próximas, como el caso de los humanos con el resto de los simios. También ha permitido el **estudio de enfermedades genéticas** humanas modificando los genes homólogos de organismos modelo como el ratón, debido a la alta **sintenia** que existe entre este organismo y el humano.

DEFINICIÓN

La **sintenia** es la posición de los distintos genes dentro de los cromosomas y se conserva entre especies filogenéticamente relacionadas.

La **genómica** ha abierto la puerta a la exploración de enormes cantidades de datos y a la creación de nuevas bases de datos que aporten un instrumento tecnológico para el estudio de la función de los organismos a nivel genético y molecular. Este nivel de análisis requiere de un **equipamiento potente de secuenciación** y de **herramientas bioinformáticas** para el procesamiento de la gran cantidad de datos generados y que permitan predecir y catalogar tanto el número total de genes como las proteínas codificadas, conocer las regiones reguladoras y la estructura de los genes e interpretar la evolución los organismos a partir de sus genomas.

1

CASO PRÁCTICO



SINTENIA ENTRE HUMANOS Y CHIMPANCÉS

Debes buscar para el gen de la insulina (INS) si existe sintenia entre el *Homo sapiens* (humano) y el *Pan troglodytes* (chimpancé). Para ello, deberás hacer uso de la herramienta GenBank y comprobar la localización del gen en los diferentes cromosomas.

Pista: ID para INS (449570, 3630).

SOLUCIÓN

Sí que se observa que existe sintenia para el gen de la insulina (INS) entre ambos organismos. La localización para el humano sería *Chromosome 11: 2,159,779-2,161,221 reverse strand*, mientras que para el chimpancé sería *Chromosome 11: 2,161,977-2,175,544 reverse strand*.

02 Métodos de secuenciación

La secuenciación de ácidos nucleicos proceso ha revolucionado nuestra comprensión de la genética, la evolución, la biomedicina y muchas otras áreas de la ciencia. Desde el descubrimiento de la estructura del ADN por Watson y Crick, los métodos de secuenciación han avanzado notablemente, permitiendo a los científicos descifrar genomas completos de organismos, estudiar la diversidad genética, identificar mutaciones asociadas a enfermedades y desarrollar nuevas terapias y diagnósticos.

La primera secuenciación completa de un genoma, el del bacteriófago ϕ X174, se logró en 1977, marcando el inicio de una nueva era en la investigación genética. Desde entonces, se han desarrollado múltiples técnicas de secuenciación, cada una con sus propias ventajas y limitaciones. Estos métodos van desde los primeros enfoques manuales hasta las tecnologías de secuenciación de nueva generación (NGS), que han incrementado la velocidad y la eficiencia del proceso, reduciendo significativamente los costos y ampliando el alcance de los estudios genómicos.

DEFINICIÓN

La **secuenciación** es el proceso que permite conocer la posición de cada nucleótido que compone una cadena de ADN o ARN. Se describe nombrando en mayúscula la base nitrogenada de cada nucleótido.

El desarrollo de nuevas herramientas de secuenciación como el **método Sanger** o método de terminación de cadena permitió el avance del Proyecto Genoma Humano.

- » Después, y gracias a nuevas técnicas de biología molecular como la PCR y la electroforesis, se pudo automatizar el proceso de secuenciación, lo que dio paso a la **primera generación** de secuenciadores automáticos.
- » En los últimos años han surgido nuevos secuenciadores que, gracias al perfeccionamiento de las técnicas enzimáticas y al análisis de millones de secuencias en una misma reacción, ha dado lugar a la secuenciación de **segunda generación**.
- » Actualmente, existen procesos que permiten secuenciar, de manera continua y a un bajo coste, millones de moléculas con tamaños muy largos (Kb) de ADN o ARN. Estos métodos se conocen como los de **tercera generación**.

2.1. Método de Maxam y Gilbert

La **técnica de secuenciación de Maxam y Gilbert**, desarrollada por Allan Maxam y Walter Gilbert en 1977, fue uno de los primeros métodos de secuenciación de ADN y se basa en la modificación química del ADN y la posterior escisión selectiva de las bases purinas y pirimidinas.



WEB

En este enlace puedes consultar el artículo que publicaron sus desarrolladores en el que describen la técnica:

<https://bit.ly/3L5IPdx>

Son necesarias varias etapas para obtener la secuencia de un fragmento de ADN:

1. Marcado del ADN: el primer paso consiste en marcar uno de los extremos del fragmento de ADN con un **isótopo radiactivo**.
2. Modificación química: el ADN marcado se divide en cuatro tubos diferentes, cada uno de los cuales se trata con reactivos distintos que modifican específicamente una o dos de las cuatro bases del ADN (adenina, guanina, citosina y timina). Estos reactivos son DMS, ácido fórmico, hidracina e hidracina con NaCl.
3. Escisión selectiva: después de la modificación, cada tubo se trata con **piperidina** que consigue romper la cadena de ADN en las posiciones de las bases modificadas anteriormente.
4. Electroforesis en gel: los fragmentos resultantes se separan por tamaño mediante electroforesis en gel de poliacrilamida al 10-20%. Como el ADN está marcado radiactivamente, los fragmentos se pueden visualizar al final mediante una autorradiografía.
5. Lectura de la secuencia: Al comparar la posición de las bandas en los diferentes carriles del gel, se puede determinar la secuencia de nucleótidos del ADN original. Hay que tener en cuenta que en algunos carriles la cadena puede terminar en una o en otra base, y para deducirlo debemos mirar los otros carriles.

Esta técnica fue revolucionaria en su momento porque permitió secuenciar el ADN de una manera relativamente precisa. Sin embargo, el método de Maxam y Gilbert ha sido

reemplazado por la secuenciación de Sanger y las tecnologías de secuenciación de nueva generación debido a su complejidad, el uso de productos químicos tóxicos y la dificultad para automatizar el proceso.

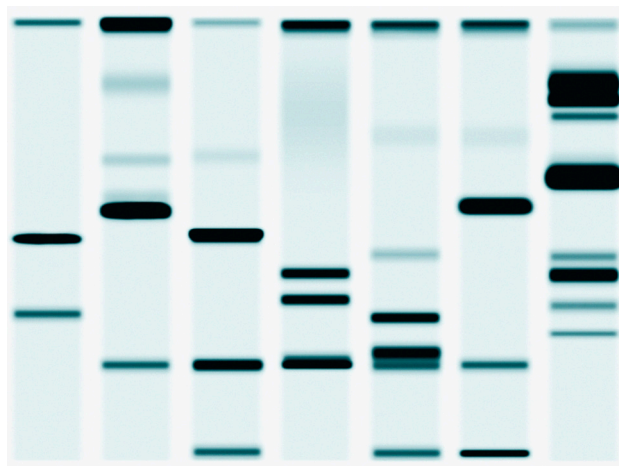


Fig. 2. Gel de secuenciación de ADN.

2.2. Método Sanger

DEFINICIÓN

El método fue desarrollado por Frederick Sanger en 1977 y se conoce también como secuenciación por terminación de cadena debido al uso de un tipo de nucleótidos llamados **ddNTPs**. El procedimiento se desarrolla en varias etapas: *reacción de polimerización, separación de los fragmentos y detección y lectura de la secuencia*.

A. Reacción de polimerización

Es una reacción de polimerización que se lleva a cabo en cuatro tubos distintos, cada uno de los cuales lleva una parte de la muestra de ADN pero algunos de los reactivos que contiene son diferentes.

El ADN problema se mezcla con un **primer** o **cebador** específico complementario a una región inicial del ADN que se quiere secuenciar.

Otros de los reactivos que se añaden son los **nucleótidos**, de dos tipos: dNTPs y ddNTPs. Los dNTPs son los desoxinucleótidos trifosfato normales que se usan para crear la cadena de ADN mientras que los ddNTPs son un tipo de nucleótido especial sintético que carece de grupo -OH en el extremo 3' lo que impide la adición de más nucleótidos una vez que se han incorporado y se produce la terminación de la cadena.

Existen cuatro tipos de nucleótidos terminadores: ddATP, ddTTP, ddCTP, y ddGTP. Estos ddNTPs están marcados con isótopos radiactivos o fluorocromos, lo que permitirá identificar cada fragmento de ADN. En la mezcla de reacción se añaden nucleótidos de los dos tipos, normales y terminadores, pero en cada tubo de reacción solamente se añade un tipo de ddNTP.

En cada tubo se añade también **ADN pol I**, una polimerasa de ADN, que sintetizará las nuevas cadenas copiando las hebras originales, añadiendo los nucleótidos complementarios. Cada vez que se incorpora un ddNTP en lugar de un dNTP, la síntesis de la cadena se interrumpe, y la ADN polimerasa se desprende, quedando un fragmento de longitud variable en función de la posición en la que se haya incorporado el ddNTP.

B. Separación de los fragmentos

El producto resultante de la reacción de secuenciación se somete a PAGE (electroforesis en gel de poliacrilamida) vertical, cargando cada tubo en un carril distinto. Así se consiguen diferenciar los fragmentos de ADN. Antes de cargar las muestras en el gel, es necesario una desnaturalización previa para que las cadenas puedan migrar adecuadamente.

C. Detección y lectura de la secuencia

La detección de los fragmentos depende del método de marcaje de los ddNTPs, de forma que si se han usado isótopos radiactivos el gel de poliacrilamida debe ponerse en contacto con una película radiográfica que quedará impresionada en las zonas donde haya fragmentos con ddNTPs que emitan radiación, y si se han marcado con fluorocromos el gel se expondrá a la luz ultravioleta y se registrará mediante una fotografía.

Los fragmentos aparecerán en orden escalonado en cada carril, lo que nos permitirá deducir la secuencia de nuestro ADN problema (Figura 3).

A la hora de leer la secuencia e interpretar los resultados, es muy importante tener en cuenta que en la película o en la fotografía lo que queda registrado es la secuencia complementaria de la que queríamos secuenciar, ya que en una de las fases del procedimiento ha tenido lugar una reacción de polimerización mediante la ADN pol I, que ha copiado la hebra original.

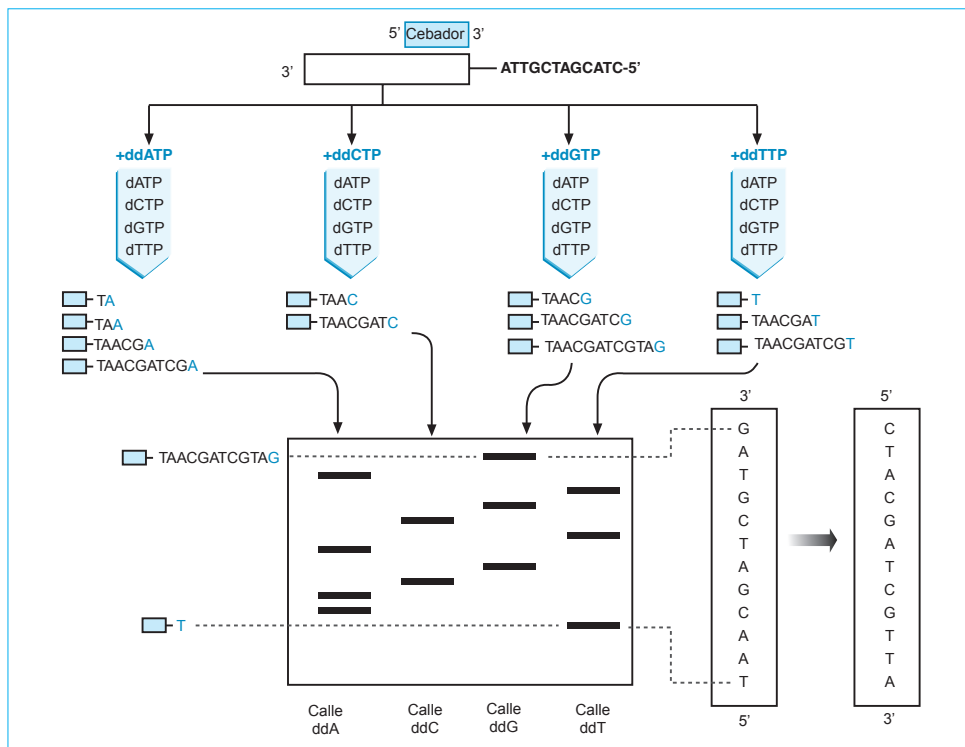


Figura 3. Método Sanger de secuenciación por terminación de cadena. La ordenación de las calles, de manera decreciente, de cada reacción donde aparece un ddNTP diferente permite conocer la secuencia del fragmento de ADN.

Tomado de <https://bit.ly/45oZMbO>.

2.3. Primera generación

La **secuenciación de Sanger automatizada** o **método de primera generación** es una evolución del método original desarrollado por Frederick Sanger. Este método es considerado de primera generación y ha sido fundamental en el desarrollo de la genómica. Las principales innovaciones consisten en que una parte del proceso está automatizada y se usa una polimerasa termoestable, lo que acelera la obtención de los resultados.

Tiene lugar en varias **fases**: *reacciones de polimerización, separación de los fragmentos, detección de la señal y lectura de la secuencia.*

A. Reacciones de polimerización

Esta primera fase debe realizarse manualmente y tiene lugar fuera del secuenciador. Durante ella tienen lugar las reacciones de polimerización, gracias al uso de:

- » cebador específico se une al principio del ADN molde que se quiere secuenciar.
- » ADN polimerasa termoestable amplifica la cadena mediante la incorporación de nucleótidos extendiendo el cebador.
- » dNTPs normales.

- » ddNTPs: estos nucleótidos carecen del grupo 3'-OH, lo que impide la adición de más nucleótidos y, por lo tanto, termina la síntesis de la cadena.

Se deben hacer dos reacciones en cadena de la polimerasa (PCR) de forma análoga a la *nested PCR* (PCR anidada), es decir, dos reacciones de amplificación consecutivas siendo la primera una **PCR de amplificación convencional** en la que se usan dos cebadores, y luego sobre el producto obtenido se emplea un cebador que permite la **PCR de secuenciación**.

Esta segunda reacción de amplificación, la PCR de secuenciación se puede realizar de dos formas, usando un **primer fluorescente** o **terminadores fluorescentes**.

- » Si se usa un **primer fluorescente** se deben usar cuatro tubos por cada muestra de forma que en cada tubo se añade el mismo primer pero marcado con un fluorocromo distinto en cada caso. Cuando termina la PCR, todos los fragmentos que están en el mismo tubo están marcados con el mismo color, pero existen distintos tamaños de secuencias. Al final se mezclan todos los tubos en uno que se carga en el secuenciador.
- » Si se emplean **terminadores fluorescentes** solamente será necesario un tubo de reacción, pero en este caso cada ddNTP lleva un marcador fluorescente diferente (un color distinto). Cuando termina la PCR el tubo contiene fragmentos de ADN que están marcados con un color distinto dependiendo del ddNTP que se haya incorporado en el extremo 3', es decir, el que finalice la cadena. Luego se carga el contenido del tubo directamente en el secuenciador. Este sistema es más utilizado actualmente porque requiere menos tiempo y menos manipulaciones.

B. Separación de fragmentos

Los productos de la etapa de polimerización se cargan en un secuenciador. Esta fase está automatizada, y consiste en una electroforesis **capilar de sílice fundida** o en **gel de poliacrilamida** en la que los fragmentos de ADN se separan por tamaño debido a que se mueven a través del sílice o del gel sobre los que se aplica un campo eléctrico de manera que los fragmentos más cortos se mueven más rápidamente y llegan primero al detector, mientras que los más grandes lo hacen al final.

C. Detección de la señal

En la zona final del capilar o del gel hay un detector de fluorescencia, y a medida que los fragmentos pasan por delante de él un láser excita los fluorocromos que emitirán luz. Cada color de luz se corresponde con uno de los cuatro ddNTPs al final del fragmento que pasa en ese momento por el detector. Un software registra la fluorescencia emitida y genera un **electroferograma**.

El **electroferograma** es un gráfico que refleja los picos de fluorescencia de los ddNTP detectados por el láser. Los **fluoróforos** o **fluorocromos** son moléculas que emiten fluorescencia tras ser excitadas con una longitud de onda concreta y permiten la detección de moléculas biológicas marcadas con estos.

D. Lectura de la secuencia

El software traduce los picos del electroferograma en una secuencia de nucleótidos y el resultado final es una secuencia de ADN, representada como una serie de letras: A, T, C y G.

Es importante destacar que la secuencia obtenida en el electroferograma requiere de una **normalización de las señales fluorescentes** para corregir posibles alteraciones en la movilidad de las bandas en el gel o errores en la intensidad de emisión. El resultado final es una secuencia informatizada en formato FASTA o FASTQ.



WEB

Método Sanger automatizado:

<https://bit.ly/3BSWS17>

Vídeo Secuencia de Sanger

<https://bit.ly/3RNN01a>

Los secuenciadores de primera generación basados en sistemas electroforéticos permiten secuenciar fragmentos de hasta 1.000 bases en una sola reacción. La **secuenciación a gran escala** persigue secuenciar fragmentos mucho mayores a 1 kb e incluso genomas completos. El obstáculo principal de la secuenciación de un ácido nucleico no es la secuenciación en sí misma, sino la ordenación de los miles de millones de fragmentos obtenidos en bloques contiguos (llamados contigs) y asignarles sus posiciones dentro del genoma.

Las dos estrategias más usadas son el paseo cromosómico y shotgun.

Paseo cromosómico: es una técnica para definir la posición de los genes en fragmentos de ADN largos. El paseo cromosómico consta, en líneas generales, de los siguientes pasos:

1. Obtención de fragmentos superponibles mediante tratamiento con enzimas de restricción.
2. Clonación de fragmentos en plásmidos o cósmidos.
3. Creación de una biblioteca genómica con los fragmentos como insertos.
4. "Paseo cromosómico" (*chromosome walk*) con sondas de hibridación partiendo de un punto inicial. Primero, se selecciona el fragmento de ADN de uno de los clones y se usa como sonda de hibridación contra la biblioteca completa. Se emplea uno de los extremos de este fragmento para crear una sonda que detecte un clon con una secuencia que se solape con la del fragmento original. Este procedimiento se repite en serie hasta secuenciar toda la biblioteca.



WEB

En el siguiente enlace puedes consultar un gráfico sobre el paseo cromosómico:

<https://bit.ly/3VQqe9>

Autora: Samanthi Udayangani, doctora en Microbiología aplicada, (24 de marzo de 2017).

Método "Shotgun": la estrategia por "Shotgun" o "fuerza bruta" se realiza a partir de fragmentos al azar a partir de los cuales un software ordena las secuencias que se solapan entre las secuencias individuales quedando así organizados los fragmentos. Esta técnica es rápida ya que requiere la síntesis de muchos menos cebadores que el paseo cromosómico. Para conseguir ensamblar las secuencias necesitamos tener previamente una gran cantidad de secuencias equivalentes, y que así se pueda deducir la secuencia de la molécula original.



WEB

National Human Genome Research Institute: Secuenciación Shotgun:

<https://bit.ly/3L4zxhQ>



¿SABÍAS QUE...?

El **formato FASTA** es el archivo de texto estándar utilizado en **bioinformática** para representar secuencias de nucleótidos obtenidas por procesos de secuenciación. Este formato permite una fácil lectura y análisis mediante lenguajes informáticos como **Python o R**. Se caracteriza por presentar en su primera línea el símbolo > seguido de la procedencia de la secuencia o identificador. En la segunda línea y sucesivas se detalla la secuencia de nucleótidos.

El archivo FASTQ permite almacenar en el mismo archivo la secuencia nucleotídica y su correspondiente valor de calidad (Q-score). Este archivo es muy utilizado en la secuenciación de segunda generación. (<https://doi.org/10.1093/nar/gkp1137>)

2.4. Segunda generación

La secuenciación de segunda generación surgió como resultado de la búsqueda de procesos que permitieran **disminuir los precios** (primera generación: 1 \$/nucleótido) y fueran **más rápidos** que los de la primera secuenciación. Otras formas de llamar a la segunda generación son tecnología **NGS** (del inglés *Next Generation Sequencing*) o **HTS** (*High Throughput Sequencing*), ya que se caracteriza por secuenciar **enormes cantidades de lecturas** (entre diez y 500 millones de secuencias al mismo tiempo) con **precios muy reducidos** (100-500 \$/3 Gb), pero su **principal inconveniente es el tamaño de sus lecturas**, pues solo permite secuenciar fragmentos de 50-400 pb, lo que complica su ensamblaje posterior.

Existen diferentes estrategias comercializadas para la secuenciación NGS, pero todas ellas necesitan un **paso previo de amplificación** debido a que no son capaces de secuenciar moléculas individuales y requieren de múltiples copias de un mismo fragmento para poder obtener una lectura. Dependiendo de la plataforma utilizada, se realiza un tipo específico de ampliación:

- » **PCR en emulsión.** La reacción ocurre en una emulsión de aceite-agua en la que dentro de las gotas de agua se encuentran unas perlas con **adaptadores** para que el fragmento de ADN que secuenciar se quede unido a estas perlas y aislado del resto de fragmentos. La PCR se realiza dentro de cada perla, lo que da como resultado millones de fragmentos idénticos inmovilizados y unidos a las perlas, a las que se conoce como **colonias** (agrupación de fragmentos clonales obtenidos por PCR). Se utiliza en las tecnologías **SOLiD y Ion Torrent**.

- » **PCR puente.** Los fragmentos de ADN que pirosecuenciar tienen incorporados unos adaptadores en los extremos, lo que les permite unirse a otros adaptadores complementarios que a su vez están unidos a una lámina de vidrio. Cuando cada fragmento caiga en lugar diferente y quede unido a la lámina, se da el proceso de PCR puente. La cadena amplificada se desnaturaliza y queda unida en su extremo, en forma de puente, a otro adaptador. Si se repite varias veces, se obtienen miles de fragmentos iguales en regiones concretas de la placa, que se conocen como **clusters**. Este proceso lo utiliza la tecnología **Illumina**.

En el enlace <https://bit.ly/3OsaCrr> puedes visualizar los métodos de ampliación para segunda generación: PCR en emulsión (a) y PCR puente (b).



¿SABÍAS QUE...?

Para poder llevar a cabo la secuenciación de segunda generación hace falta realizar tres pasos previos. Se debe **extraer y purificar** el ADNg, **fragmentarlo** por enzimas de restricción o sonicación (para obtener tamaños de fragmentos deseados) y **generar las genotecas** (flanquear cada fragmento con un **adaptador** de secuencia conocida).

Después de la ampliación se procede a la fase de secuenciación que, dependiendo de la plataforma utilizada, se puede realizar por **síntesis o ligación**. Actualmente, el método más utilizado es por síntesis.

La **secuenciación por síntesis**, al igual que en el método Sanger, se basa en detectar cada nucleótido que se incorpora durante la síntesis de la cadena complementaria al fragmento que se desea secuenciar. Dentro de este tipo de tecnología, **existen tres variantes**, atendiendo a la forma de detección de los nucleótidos añadidos:

- » Por **luminiscencia**. Se basa en la tecnología de **pirosecuenciación** que permite detectar emisiones de luz producidas durante la elongación de la cadena complementaria mediante la producción de pirofosfato (es la molécula que al reaccionar con otros reactivos produce la bioluminiscencia). Permite secuenciar

en tiempo real, puesto que la generación de luz se produce por cuatro reacciones químicas acopladas que se detectan en forma de pico tras añadir un nucleótido nuevo a la reacción. Su principal limitación son las lecturas de homopolímeros (secuencias repetidas de un mismo nucleótido). Usada por la empresa Roche. Está en desuso.

- » Por **fluorescencia**. Es una secuenciación **terminal cíclica reversible (CTR)** muy parecida a la tecnología Sanger. Usa **nucleótidos marcados con fluoróforos y bloqueados en sus extremos 3'**. Este bloqueo es reversible y se puede eliminar tras la lectura de la fluorescencia, lo que permite continuar con las síntesis de la cadena complementaria. Se mide la fluorescencia en cada punto de la placa (donde se encuentra un fragmento aislado y amplificado), por lo que es capaz de detectar millones de destellos fluorescentes a la vez, lo cual genera **lecturas de 150 pb** de cada fragmento. Con una pequeña adaptación, esta tecnología permite secuenciar en ambos sentidos (cadena original y cadena complementaria) cada fragmento, lo que aumenta la cobertura de lectura y facilita la secuenciación de homopolímeros o regiones con deleciones e inserciones. Se conoce como secuenciación por *paired-ends*. Es la más utilizada de todas las NGS y la distribuye la empresa **Illumina**.
- » Por **liberación de H⁺**. Al igual que la tecnología de pirosecuenciación, no utiliza dNTP modificados, lo cual la hace más rentable que Illumina. Se basa en la tecnología de **semiconductores** que permite detectar la liberación de protones durante la síntesis de la cadena complementaria. Esto genera un cambio de pH que es proporcional a la cantidad de dNTP incorporados durante cada ciclo. Esta técnica se conoce como Ion Torrent y es la más barata de todas.

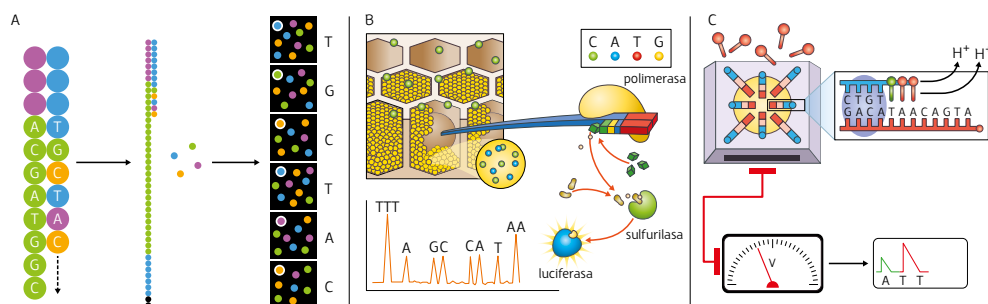


Fig. 4. Métodos de Segunda generación: A) CTR, B) Pirosecuenciación y C) Liberación de H⁺.

Respecto a la **secuenciación por ligación**, hoy en día solo se utiliza la tecnología **SOLiD (soporte de detección por ligación de oligonucleótidos)**, que permite, mediante hibridaciones de 16 cebadores diferentes y marcados con cuatro fluoróforos distintos, la lectura de fragmentos de **50 pb** midiendo la fluorescencia emitida por los cebadores. Debido al pequeño tamaño de sus lecturas, no se utiliza mucho.

Para facilitar la comprensión de este apartado, se muestra un **resumen** (Tabla 2) con las características principales de los métodos de segunda generación.

Tipo de secuenciación	Variantes de secuenciación	Tipo de PCR utilizada en la amplificación	Tamaño de fragmentos secuenciados	Nombre la empresa
Síntesis	Luminiscencia o pirosecuenciación	PCR emulsión	400 pb	Roche
	Fluorescencia ddNTP o CTR	PCR puente	150 pb + paired-ends	Illumina
	Liberación de H ⁺ o Ion Torrent	PCR emulsión	200 pb	Thermo Fisher
Ligación	SOLiD	PCR emulsión	50 pb	Applied Biosystems

Tabla 2. Principales características de la secuenciación de segunda generación.

2.5. Tercera generación

Todas las técnicas de segunda generación presentan una **limitación** importante debido al **pequeño tamaño de lecturas que generan**. Esto provoca que resulte muy difícil ensamblar lecturas y cartografiar cromosomas cuando se analizan genomas grandes como el humano. Además, necesitan de un paso previo de amplificación, lo que aumenta los costes y las posibles introducciones de errores en la secuencia final.

Los métodos de secuenciación de tercera generación **utilizan moléculas individuales**, es decir, no necesitan de ampliaciones previas, y secuencian **tamaños de lecturas Kb o Mb**, lo que permite un mejor ensamblaje con homopolímeros o regiones de cromatina compactada (heterocromatina).

Actualmente, existen dos **técnicas** ampliamente usadas que utilizan este tipo de secuenciación mediada por nanoporos:

PacBio. Es una secuenciación por síntesis de molécula única y basada en la tecnología **SMRT** (del inglés *Single Molecule, Real-Time*). Utiliza moléculas de ADN pol inmovilizadas en unos nanoporos, así como nucleótidos marcados con fluoróforos, lo que permite que sean excitados justo antes de añadirse a la cadena en síntesis.

Oxford Nanopore. Es una secuenciación por **diferencia de voltaje**. Esta tecnología se basa en **una membrana con nanoporos** que mide la diferencia de corriente entre los dos extremos del poro. En cada nanoporo se encuentra fijada una **helicasa** que permite el paso del ADNss y, cuando este atraviesa el poro, dependiendo del nucleótido que está pasando en ese momento, se produce un cambio de voltaje. Cada uno de los cuatro nucleótidos

interrumpe la señal eléctrica de una forma distinta, lo que permite registrarla mediante un ordenador que asigna el paso de un determinado nucleótido.

IMPORTANTE

Puedes ver métodos de secuenciación por tercera generación en el artículo:

<https://go.nature.com/3VLhQsx>

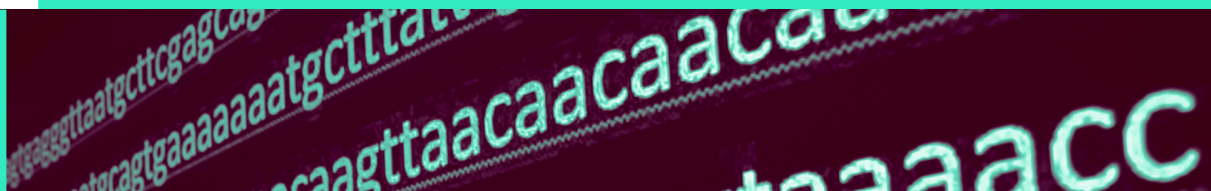


VÍDEO LAB

Para consultar el siguiente video sobre secuenciación del SARS-CoV-2: Validación de variantes mediante Sanger automatizado, escanea el código QR o [pulsa aquí](#).

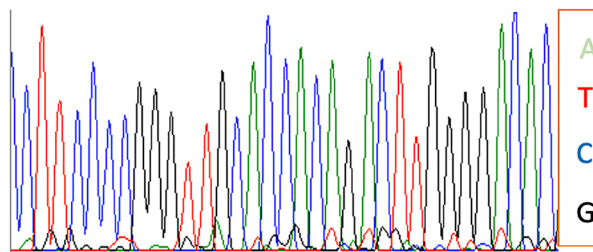
2

CASO PRÁCTICO



ANÁLISIS DEL ELECTROFEROGRAMA

Tras realizar una secuenciación, se te pide que interpretes los resultados obtenidos mediante la visualización del siguiente electroferograma.



1. Indica la técnica de secuenciación utilizada.
2. Obtén la secuencia final mostrada en el electroferograma.

SOLUCIÓN

1. La técnica utilizada es el método Sanger automatizado.
2. La secuencia final sería: 5' CCTTCCCGGTTGCACCACAGACTTGGGGACAC-3'

03 Procesamiento e interpretación de la secuenciación

A menudo, al análisis de genomas completos se lo conoce como «secuenciación del genoma», pero el proceso de secuenciar no es el primero ni el último. Por ello, se requieren una serie de **pasos necesarios** antes y después de la fase de secuenciación.

- » La **primera fase** tiene como objetivo preparar el ADN para poder secuenciarlo mediante cualquiera de las tres generaciones de secuenciación descritas en el apartado anterior. En esta fase se necesita **aislar y purificar el ADN o ARN** problema. Después se **fragmenta** por métodos de sonicación o usando enzimas de restricción para que estos fragmentos sean compatibles con el tipo de tecnología elegida para secuenciarlos. Por último, se añaden adaptadores en ambos extremos de los fragmentos que permiten identificar a los fragmentos (permite secuenciar diferentes muestras en un único proceso) y poder unirlos a la celdas o placas de secuenciación. A este proceso se lo conoce como **construcción de genotecas** (librerías genéticas).
- » La **segunda fase** es el proceso de secuenciación que se ha visto en el apartado anterior.
- » La **tercera fase** consta de las **herramientas bioinformáticas** que permiten **evaluar la calidad** de las lecturas obtenidas, alinear todas las lecturas mediante programas de **ensamblaje computacional**, y **anotar y analizar** los datos obtenidos.

Fase de procesamiento o preparación del ADN	1. Aislamiento del ADN 2. Fragmentación del ADNg 3. Construcción de librerías genéticas
Fase de secuenciación	4. Secuenciación por Sanger, NGS o nanoporos
Fase de interpretación y análisis de las secuencias	5. Ensamblaje computacional 6. Anotación 7. Análisis funcional

Esquema 1. Flujo de trabajo de la secuenciación.

3.1. Procesamiento del ADNg

Todos los proyectos de secuenciación siguen los pasos descritos en la Tabla 9.3, pero no de la misma forma. El flujo de trabajo en los proyectos de secuenciación va a depender del objetivo del proyecto. Es decir, los genomas completos pueden secuenciarse por primera vez (lo que se conoce como secuenciación **de novo**) o pueden ser proyectos donde ya existen un **genoma de referencia**. Esto hace que la preparación de librerías genéticas y los programas bioinformáticos utilizados sean diferentes.

En el **aislamiento del ADN** es importante saber si se quiere secuenciar un genoma completo o algunas zonas concretas (como exomas) y zonas del genoma con características especiales. Para **genomas completos** se usa el ADNg extraído de cualquier fuente como fósiles, tejidos o células individuales. Para **exomas** se suelen usar paneles comerciales que capturen estas regiones del genoma o, directamente, secuenciar el ARNm. Para **regiones especiales** como metilaciones, se usa ADN pretratado con bisulfito. También es posible secuenciar el ARN pasando por un paso previo de generación de ADNc por retrotranscripción (en NGS) o secuenciarlo directamente (usando nanoporos). Esta tecnología se conoce como RNA-seq.

DEFINICIÓN

El **exoma** es la parte del genoma que está compuesta únicamente por las secuencias que codifican proteínas.

En los protocolos de **construcción de librerías** se suele incluir el proceso de fragmentación. Para **fragmentar** es necesario saber qué tipo de método de secuenciación se utilizará, ya que, por ejemplo, para NGS se necesitan fragmentos de 500 pb como máximo, por lo que se suelen utilizar procesos de sonicación o de nucleasas de restricción. Una vez se cortan en fragmentos, estos se seleccionan según el tamaño deseado por métodos de electroforéticos. Después, **se les añade un adaptador** (secuencias conocidas de oligonucleótidos) en cada extremo del fragmento. Estos contienen diferentes regiones, como en los adaptadores usados en la tecnología de **Illumina**, donde la **región más externa** permite el anclaje al soporte de la secuenciación, las **regiones intermedias** contienen los índices utilizados en los procesos de identificación del fragmento o de ensamblaje informático y **las regiones internas** permiten la unión de los *primers* para la secuenciación por síntesis.



Fig. 5. Adaptadores de Illumina. La región externa está marcada en rojo y verde; los índices, en negro; y las regiones internas, en azul y naranja.

3.2. Interpretación y análisis de las secuencias

Una vez obtenidas las secuencias en formato **FASTA** o **FASTQ**, y antes de proceder a realizar el ensamblaje de las lecturas, se debe evaluar la calidad de los datos. Para este control de calidad se utiliza el programa **FastaQC** (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>), que permite observar el número total de las lecturas de nuestro archivo FASTQ, así como conocer el tamaño medio de las lecturas y la calidad de estas.

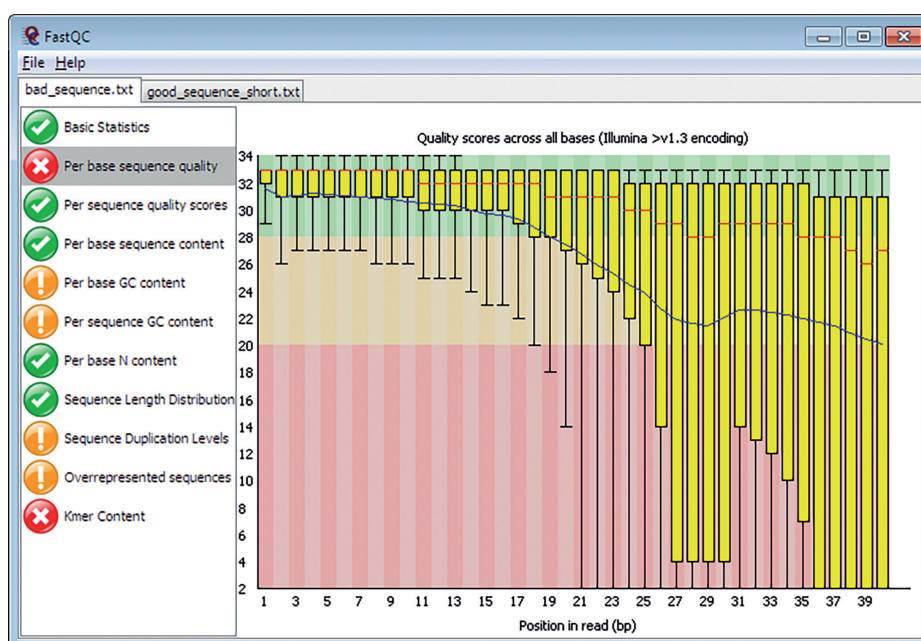


Fig. 6. Programa FastaQC que evalúa la calidad de la secuenciación. Tomado de <https://bit.ly/42cBPBp>.

Tras comprobar que existe buena calidad en las lecturas, se puede pasar al **ensamblaje de los fragmentos** obtenidos. Este proceso se lleva a cabo mediante **programas informáticos** que **buscan regiones** en los fragmentos que sean **solapantes** de otros fragmentos (fragmentos con mismas secuencias) y van uniéndolos hasta conseguir una única secuencia formada por la unión de miles de fragmentos, llamados **contig**. Normalmente, son necesarios varios **contigs** para cubrir un cromosoma, y es posible que existan zonas del cromosoma donde los programas de ensamblaje no hayan podido conocer la secuencia que se sitúa en esa región, por lo que dejan huecos entre **contigs** (conocidos como **gaps**). Los **gaps** suelen rellenarse fácilmente si se utiliza secuenciación de tercera generación para estas regiones problema.

Los programas utilizados para ensamblar genomas de *novo* son **MEGAHIT**, **MIRA** y **Velvet**, mientras que para ensamblar secuencias en las que se disponga de un genoma de referencia, como es el caso del genoma humano, el programa más usado es **BWA**. Se basa en el algoritmo de alineamiento de Burrows-Wheeler. Esta herramienta permite alinear millones

de secuencias de manera rápida y ágil a partir de un genoma de referencia, obtenido de un archivo en **formato SAM** (del inglés *Sequence Alignment Map*).

Antes de la anotación de las secuencias de genomas completos, se suele usar el programa **SAMtools** para eliminar secuencias duplicadas, filtrarlas e indexar los alineamientos obtenidos en el archivo SAM.

Para realizar la anotación, se usan programas como **snpEff** o **VEP**, que permiten anotar las secuencias obtenidas en genes codificantes o no codificantes, a partir de genomas completos. Si se quiere anotar o conocer la función de secuencias independientes extraídas de proyectos de secuenciación de genes concretos, se puede emplear **BLAST** (*Basic Local Alignment Search Tool*: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

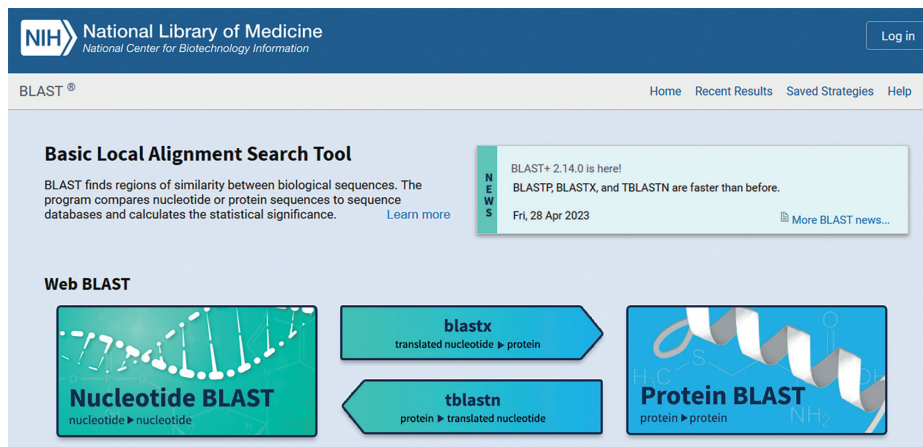


Fig. 7. Herramienta BLAST.

Por último, para **analizar la función** de los genes anotados se pueden utilizar herramientas bioinformáticas como **Ensembl** o **GenBank**. Estas bases de datos tienen información de todos los genes conocidos de multitud de especies, incluida la humana. Además, te permiten conocer de manera intuitiva datos sobre los transcritos de un determinado gen, su tamaño en pb, la secuencia de aminoácidos que genera y la posición del cromosoma donde se encuentra.

Gene: INS ENSG00000254647

Description: insulin [Source:HGNC Symbol;Acc:HGNC:60814]

Gene Synonyms: IDDM1, IDDM2

Location: Chromosome 11: 2,159,779-2,161,221 reverse strand. GRCh38: CM000673.2

About this gene: This gene has 5 transcripts (splice variants), 253 orthologues, 2 paralogues and is associated with 6 phenotypes.

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000381330.5	INS-202	465	110aa	Protein coding	CCDS7729.4	P01308-14	MM_000207.3	MANE Select Ensembl Canonical GENCODE basic APPRIS
ENST00000397262.5	INS-203	639	110aa	Protein coding	CCDS7729.4	P01308-14	-	GENCODE basic APPRIS P1 TSL1
ENST00000250971.1	INS-201	503	110aa	Protein coding	CCDS7729.4	P01308-14	-	GENCODE basic APPRIS P1 TSL1
ENST00000512521.1	INS-205	297	98aa	Protein coding	-	A6QL24	-	GENCODE basic TSL1
ENST00000421783.1	INS-204	464	92aa	Non stop decay	-	C9JNS4	-	TSL2 CDS 3' incomplete

Summary

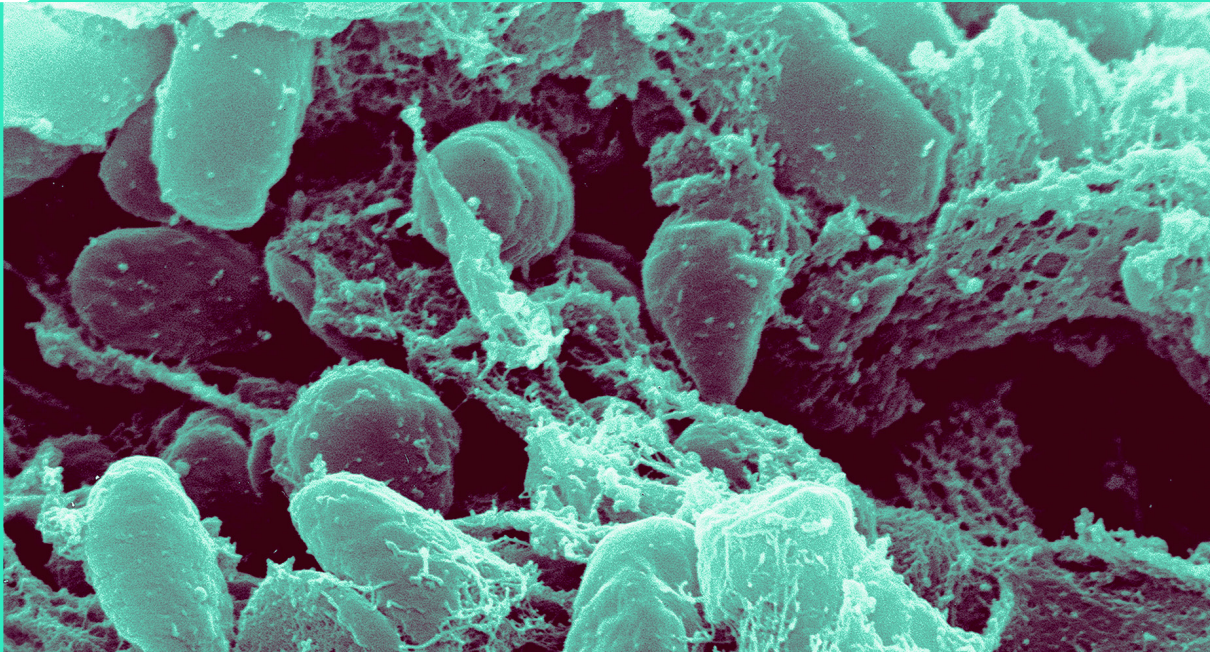
Name: INS (HGNC Symbol)

MANE: This gene contains MANE Select ENST00000381330 ENSP00000370731

Fig. 8. Programa Ensembl con la información funcional del gen de insulina humano.

3

CASO PRÁCTICO



ANOTACIÓN DE SECUENCIAS

Trabajas en el laboratorio de microbiología y te piden que analices por Sanger una secuencia problema de una posible bacteria que está dando negativo en todas las pruebas de PCR. Tras utilizar Sanger automatizado obtienes la siguiente secuencia en formato FASTA:

```
>SEQUENCE_1
```

```
ACTGCGCCAGCCCCAACCGTTCAGGATTATCGATAATCATCAAGCTGGCATTGGGCACATCAACTCCCAC
```

Debes utilizar una herramienta de anotación para identificar la bacteria bajo estudio.

Pista: usar BLAST.

SOLUCIÓN

Tras utilizar la herramienta BLAST, obtenemos que la bacteria problema se identifica como *Yersinia pestis* con un 100% de exactitud.

X T A R T
FP SALUD