Fundamentos de Secuenciación de Alto Rendimiento y Genómica Traslacional

Resumen

La asignatura aborda las tecnologías y metodologías actuales para la generación y análisis de datos multi-ómicos en biología y biomedicina, con un enfoque en técnicas de secuenciación masiva (NGS). Se estudian variantes genómicas puntuales y estructurales - como mutaciones, polimorfismos de nucleótido único (SNPs) y variantes en el número de copias génicas (CNVs) - que explican la variabilidad genética poblacional y su relación con enfermedades humanas.

A lo largo del curso, se aplican herramientas bioinformáticas para el análisis de datos reales, profundizando en la epidemiología molecular y los estudios de asociación genómica (GWAS) para explorar los factores de riesgo asociados a variantes genómicas en contextos clínicos y poblacionales. También se revisa el uso de datos genómicos en tratamientos personalizados, considerando su aplicación actual y potencial en la clínica.

Sandra Mingo Ramírez

UAM - 2024/25

4 de noviembre de 2024 19:03

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

Código en Github

Índice general

Capítulo I

Introducción a la genómica: caracterización del genoma mediante NGS

I.1. Introducción a la genómica traslacional

I.1.1. Definición e importancia de la genómica

La genómica, el estudio integral del ADN y de la estructura, función y dinámica de los genomas, representa un pilar fundamental en la biología moderna. Marcó un cambio de paradigma, pasando de un enfoque reduccionista en biología - donde se estudiaban componentes individuales y de manera aislada - a una perspectiva integradora que analiza las interacciones y relaciones entre los distintos elementos biológicos. Esta transición permitió evolucionar de la genética clásica, basada en hipótesis concretas, hacia la genómica, que integra análisis de datos masivos sin necesidad de preguntas iniciales específicas, aunque sí en constante búsqueda de respuestas biológicas complejas.

En el marco del dogma central de la biología, las "ómicas" representan tres niveles de estudio: la genómica (centrada en el ADN), la transcriptómica (ARN) y la proteómica (proteínas). Este curso se enfoca en la genómica, ya que la información genética determina las funciones bioquímicas y, por ende, los fenotipos de los organismos. Gracias a avances recientes, ahora es posible inferir la función bioquímica de las proteínas directamente a partir de la secuencia de ADN, sin necesidad de técnicas complejas como la cristalización. Además, herramientas de inteligencia artificial pueden predecir la estructura de las proteínas con precisión, acelerando la interpretación de funciones biológicas.

Las proteínas, incluyendo enzimas esenciales, son los elementos funcionales clave en la biología. La secuencia de aminoácidos en una cadena polipeptídica define sus propiedades funcionales, y, por tanto, conocer la secuencia genética subyacente (el ADN) facilita predecir la función de una proteína. Aunque determinar experimentalmente las propiedades de una proteína es complejo, la secuenciación genómica ha simplificado enormemente este proceso.

La mejora en tecnologías de secuenciación impulsó el **Proyecto Genoma Humano**, que logró identificar entre 20,000 y 25,000 genes y determinar la secuencia de los aproximadamente 3 mil millones de pares de bases del genoma humano. Este proyecto también fomentó la creación de bases de datos y herramientas para el análisis de datos genómicos, además de abrir el debate sobre los aspectos éticos, legales y sociales (conocidos como ELSI, por sus siglas en inglés), que siguen siendo temas vigentes y complejos en la actualidad.

Evolución de la bioinformática en la genómica La bioinformática ha crecido a la par de la genómica en múltiples niveles. Inicialmente, era una **disciplina** incipiente y se desarrollaba como apoyo experimental; sin embargo, ha evolucionado hasta convertirse en un campo esencial que impulsa la investigación. En cuanto a su **material, los datos,** la bioinformática ha tenido que adaptarse al fenómeno del big data, pasando de manejar cantidades limitadas de datos a enfrentar volúmenes masivos, propios de la genómica actual. Paralelamente, el **rol de los bioinformáticos** se transformó, pasando de ser técnicos a científicos de datos y académicos altamente reconocidos en la industria y en la investigación.

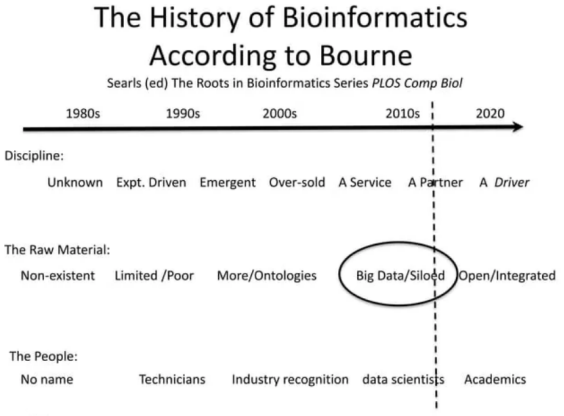

Figura I.1: *Breve historia de la bioinformática en tres niveles: como disciolina, como material que utiliza y como las personas que trabajan en ella. Evolución desde 1980 hasta 2020.*

I.1.2. Avances tecnológicos en secuenciación

Existen distintos tipos de tecnologías de secuenciación, comúnmente clasificadas en tres generaciones: la primera generación (first generation), la segunda o Next Generation Sequencing (NGS) y la tercera generación. Las dos primeras generaciones se enfocan en la secuenciación de fragmentos cortos de ADN, mientras que la tercera generación permite la lectura de fragmentos largos, facilitando el ensamblaje completo de genomas. Actualmente, uno de los mayores desafíos tecnológicos es detectar variantes de baja frecuencia y realizar secuenciaciones de ADN en células individuales (single-cell sequencing), lo cual tradicionalmente se hacía de forma masiva ("bulk").

A medida que el costo de la secuenciación ha disminuido y la capacidad de almacenamiento ha mejorado desde 1990, los datos generados también han crecido exponencialmente. En un experimento de secuenciación, los costos abarcan

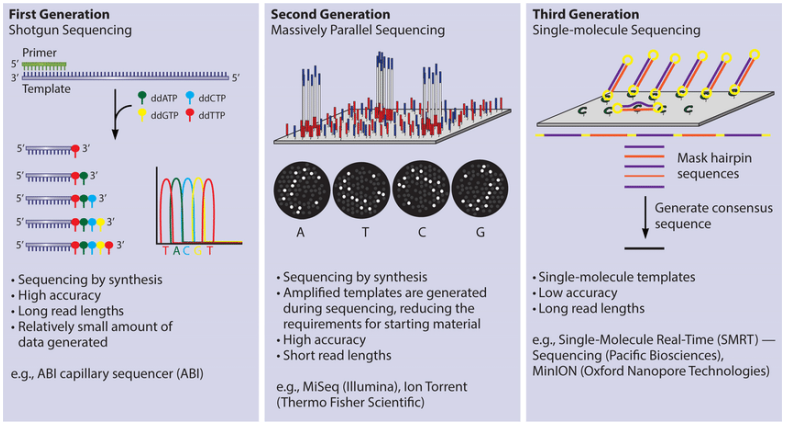

Figura I.2: *Las tres generaciones de secuenciación y su forma de actuar.*

tanto la secuenciación en sí como el procesamiento bioinformático, el reporte y el almacenamiento de los datos. La comunidad científica y muchos journals requieren que los datos de proyectos financiados públicamente estén disponibles en bases de datos accesibles, lo que asegura la transparencia y el acceso a esta información valiosa. Para obtener una cobertura de calidad, el ADN suele secuenciarse al menos 30 veces, lo que genera archivos de gran tamaño, como los archivos FastQ, que almacenan información de secuencia y calidad para cada base.

I.1.3. Procesos de llamada y priorización de variantes

Los datos de secuenciación se procesan en pipelines bioinformáticas que comienzan con archivos FastQ normalmente comprimidos y pasan por varias etapas: control de calidad, alineamiento y llamada de variantes (variant calling). Las variantes identificadas pueden incluir cambios de nucleótidos, variaciones en el número de copias de segmentos genómicos (copy number variation) o reordenamientos estructurales.

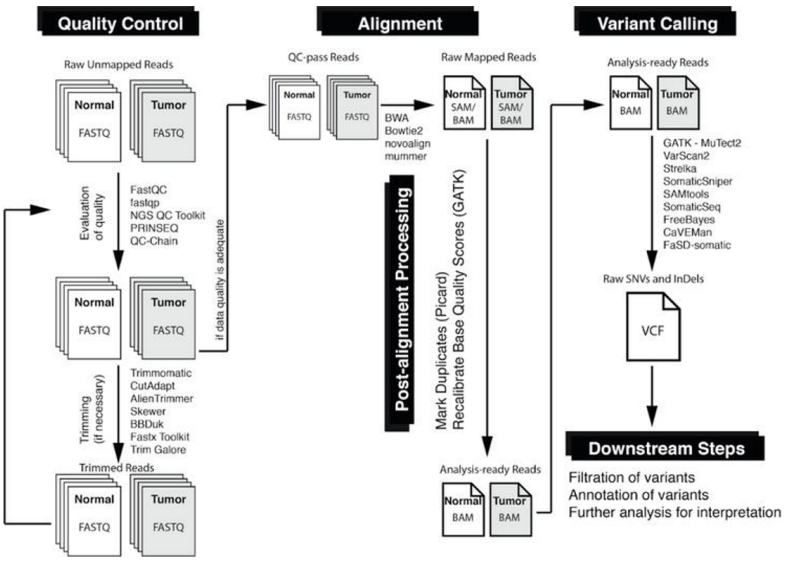

Figura I.3: *Esquema de la pipeline que se sigue en bioinformática para la llamada de variantes.*

La priorización de variantes se basa en factores como el impacto funcional, la frecuencia alélica en la población y la asociación con enfermedades. Sin embargo, muchas variantes requieren validación experimental, frecuentemente en modelos animales como ratones, para corroborar su relevancia funcional. El proceso de filtrado inicial se enfoca en variantes en exones de genes candidatos, analizando su frecuencia, patogenicidad y modelo de herencia; en caso de no hallarse variantes relevantes, se amplía el análisis a variantes oligogénicas o no codificantes.

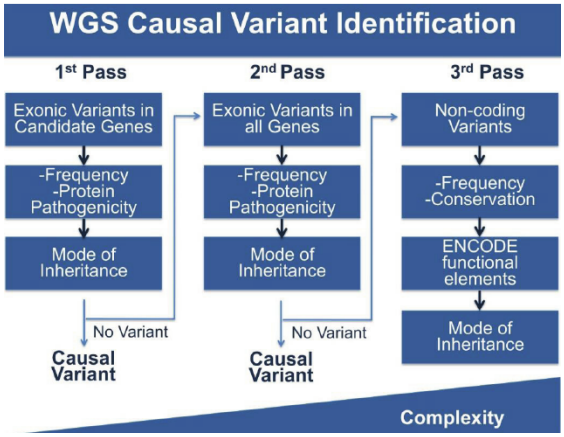

Figura I.4: *Ejemplo de la priorización de variantes.*

I.1.4. Genómica en medicina de precisión

La genómica ha transformado el enfoque de la medicina de precisión, permitiendo identificar enfermedades con bases genéticas, ambientales o una combinación de ambas. Algunas variantes genéticas confieren una predisposición a enfermedades sin ser causantes directas, lo cual es crucial para inferir relaciones causales y acelerar ensayos clínicos mediante la integración de grandes volúmenes de datos. Estas variantes pueden clasificarse en germinales (heredadas) o somáticas (adquiridas).

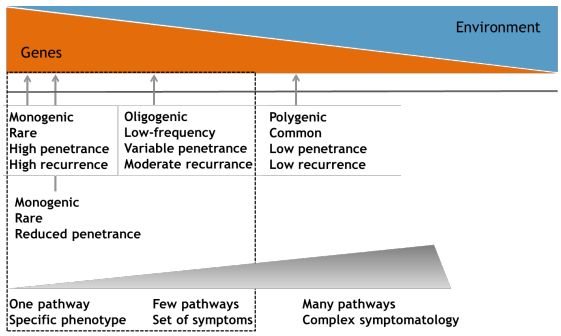

Figura I.5: *Representación gráfica de la relación entre enfermedades con base genética, ambientales o una mezcla de ambas.*

En medicina de precisión, la genómica es solo una capa de datos entre muchas. Para una comprensión holística de la salud y la enfermedad, es necesario combinarla con información de otras "ómicas" como la transcriptómica, epigenómica, proteómica, metabolómica, y datos de microbioma. Además, los datos clínicos y epidemiológicos

también forman parte del ecosistema de **Big Data Biomédico**, que actualmente se maneja mediante técnicas avanzadas de computación en clusters HPC, computación en la nube y algoritmos de GPU.

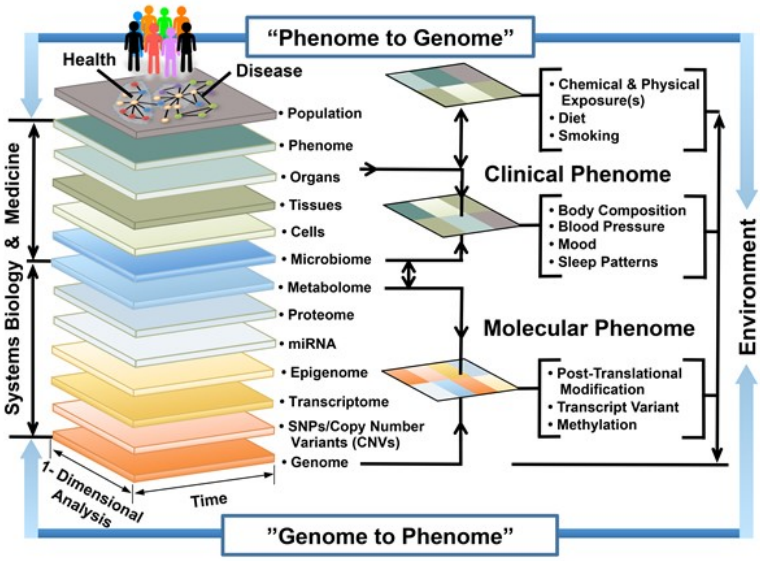

Figura I.6: *Esquema representando el dibujo general de la bioinformática.*

Varias bases de datos públicas permiten estudiar la transición entre salud y enfermedad. El estudio de Farmingham, por ejemplo, lleva más de 70 años recolectando datos de factores de riesgo cardiovascular en más de 15,000 participantes. En Reino Unido, el Biobank y, en Estados Unidos, la iniciativa All of Us, también representan recursos de gran envergadura. En España, el CNIC (Centro Nacional de Investigaciones Cardiovasculares) realiza el estudio PESA (Progression of Early Subclinical Atherosclerosis), que ha contribuido a identificar factores predictivos de aterosclerosis subclínica mediante el estudio multiómico, generando nuevos indicadores con un mayor poder predictivo de la formación de placas de colesterol.

Epigenética y la medición de la edad biológica El perfil de metilación del ADN es un factor epigenético que puede modificar la expresión genética y se ha utilizado para calcular la "edad biológica" o epigenética de una persona, lo que puede servir como predictor de esperanza de vida y salud. Al comparar estos perfiles con la edad cronológica, sexo y otros factores, se obtiene información sobre el envejecimiento y el riesgo de enfermedades, facilitando el desarrollo de estrategias de salud personalizadas.

I.1.5. Resumen

La genómica ha liderado una revolución científica en el siglo XX, evolucionando desde el estudio de componentes individuales hasta una perspectiva integral de sistemas biológicos y de investigación basada en datos masivos. La bioinformática se ha convertido en una disciplina central en el análisis genómico y predicción de estructuras proteicas, impulsada por el Proyecto Genoma Humano y el desarrollo de tecnologías de secuenciación. Los avances actuales buscan no solo la secuenciación del ADN, sino también la integración de estos datos con datos epidemiológicos y moleculares para

obtener una comprensión más profunda de la salud y la enfermedad. Así, el Proyecto Genoma Humano fue decisivo para sentar las bases de tecnologías de secuenciación, el desarrollo de la bioinformática en sí y el uso social e industrial de los datos ómicos.

La identificación de características genómicas relevantes causales de rasgos/enfermedades se basa en la anotación de variantes en bases de datos y en estudios poblacionales: hay margen de mejora y un gran éxito de la ciencia colaborativa. Hoy en día, los principales proyectos tratan no sólo de secuenciar el ADN, sino de integrar esta información con datos epidemiológicos y otros datos moleculares para comprender mejor la salud y la enfermedad. Las enfermedades, en función de su base genética, pueden clasificarse en monogénicas (mendelianas), oligogénicas (ej., cardiopatías familiares) y complejas (evaluadas mediante puntuaciones de riesgo poligénicas). Esta clasificación permite avanzar en la medicina de precisión, abordando enfermedades desde su origen genético para ofrecer intervenciones de salud más efectivas y personalizadas.

I.2. Métodos de secuenciación

La secuenciación permite pasar de la información contenida en el ADN a un dominio digital mediante una representación abstracta.

El primer método de secuenciación fue el **método Maxam-Gilbert**, que utilizaba un marcador en el extremo 5' del ADN. En este proceso, el ADN se trataba con diferentes compuestos químicos para provocar rupturas específicas en función de cada base nitrogenada. Los fragmentos resultantes se separaban en un gel de acrilamida mediante electroforesis, y se revelaban mediante autoradiografía de rayos X. La secuencia se deducía observando el patrón de bandas resultante.

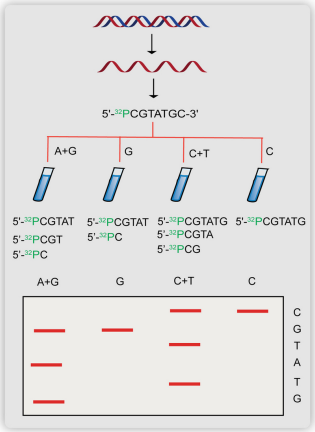

Figura I.7: ***Principio de la secuenciación Maxam-Gilbert****: Se llevaron a cabo cuatro reacciones separadas para la degradación de bases en un fragmento de ADN monocatenario: A+G, G, C+T y C. Se obtienen fragmentos de ADN de diferente longitud tras la degradación de las bases y la escisión del esqueleto de azúcar-fosfato. Los productos se cargan en cuatro pocillos separados de un gel de poliacrilamida. La secuencia se lee de abajo a arriba como GTATGC. Si se encuentra una G frente a un hueco en el gel, se confirma que se trata de 5-metilcitosina en la cadena molde.*

Otro método clave es el de **terminación de cadena, o método de Sanger**. Este utiliza deoxinucleótidos modificados, que tienen un átomo de hidrógeno en el grupo 2' de la pentosa, en lugar de un grupo hidroxilo (OH). Esto impide la unión del extremo 5' al 3', deteniendo así la extensión de la cadena de ADN. El resultado es una mezcla de fragmentos de distintos tamaños, los cuales se marcan con isótopos radioactivos o, en versiones más modernas, con fluoróforos. La secuencia se obtiene mediante detección de colores en una única reacción, simplificando el análisis. La clave de este método es el uso de dideoxinucleótidos, que interrumpen la actividad de la ADN polimerasa, permitiendo detener la cadena de manera controlada.

Pregunta examen

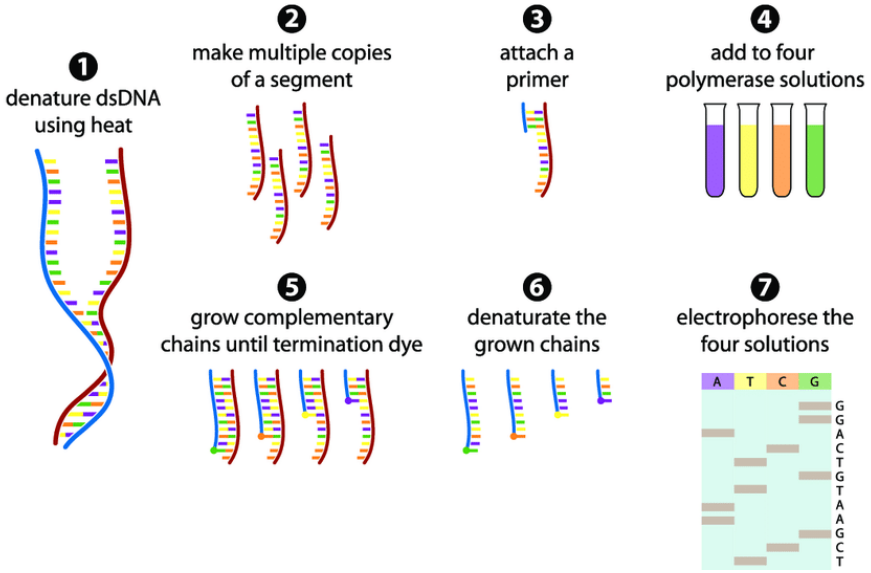

Figura I.8: ***El método de secuenciación Sanger en 7 pasos.*** *(1) El fragmento de dsADN se desnaturaliza en dos fragmentos de ssADN. (2) Un fragmento de ssADN se multiplica en millones de copias. (3) Se une un cebador que corresponde a un extremo del fragmento. (4) Los fragmentos se añaden a cuatro soluciones de polimerasa. Cada solución contiene los cuatro tipos de bases pero sólo un tipo de nucleótido de terminación. (5) La cadena crece hasta que se añade aleatoriamente un nucleótido de terminación. (6) Los fragmentos de dsADN resultantes se desnaturalizan para obtener una serie de ssADN de distintas longitudes. (7) Los fragmentos se separan por electroforesis y se lee la secuencia.*

El primer secuenciador automático fue el ABI370, capaz de secuenciar hasta 5000 bases al día. Sin embargo, se necesitarían aproximadamente 16,000 años para secuenciar todo el genoma humano usando esta tecnología. Este secuenciador innovador reemplazaba los geles por electroforesis capilar y un detector de fluorescencia. Durante el Proyecto Genoma Humano en los años 90 y 2000, desarrollado en colaboración entre el sector público y privado, se introdujeron mejoras significativas a los secuenciadores, como el modelo ABI377, que empleaba varios capilares para incrementar la eficiencia. Sin embargo, la secuenciación de regiones altamente repetitivas del genoma, como los telómeros y centrómeros, fue compleja, y la primera descripción completa del genoma humano fue publicada hace apenas un año.

I.2.1. Métodos de secuenciación empleados en el Proyecto Genoma Humano

Los métodos que se utilizaron en el proyecto fueron los siguientes:

- **Hierarchical Shotgun:** En este método, el ADN se clona en fragmentos más pequeños usando enzimas de restricción. Estos fragmentos se solapan y forman contigs, los cuales se ensamblan progresivamente para reconstruir la secuencia original.
- **Whole-genome Shotgun:** Similar al método anterior, pero se realiza directamente sobre el genoma completo en lugar de partir de cromosomas bacterianos. El ADN se clona en bacterias, se fragmenta y se ensamblan los contigs mediante solapamiento.

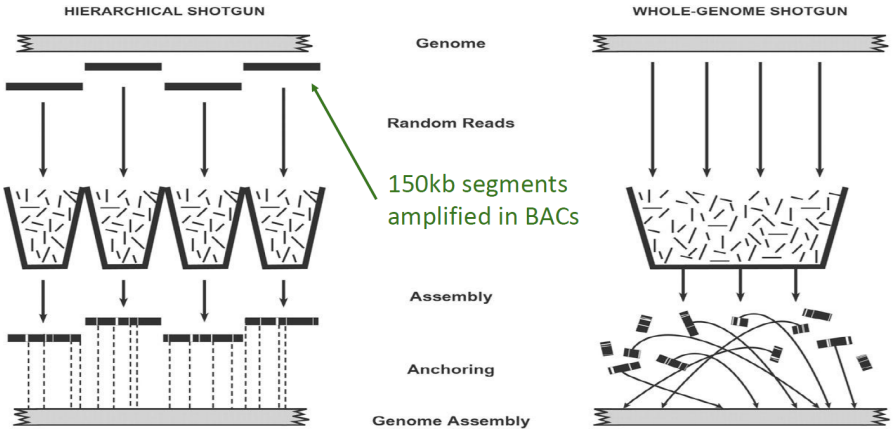

Figura I.9: ***Estrategias de secuenciación en el Proyecto Genoma Humano.*** *(Izquierda) La estrategia de hierarchical shotgun (HS) consiste en descomponer el genoma en un camino de mosaico de clones BAC (bacterial artificial chromosome) superpuestos, realizar la secuenciación en cada BAC y volver a ensamblarlo, y luego fusionar las secuencias de clones adyacentes. El método tiene la ventaja de que todos los contigs de secuencias y scaffolds derivados de un BAC pertenecen a un único compartimento con respecto al anclaje al genoma. (Derecha) La estrategia WGS (Whole-genome shotgun) consiste en secuenciar todo el genoma e intentar reensamblar toda la colección. Con el método WGS, cada contig y scaffold es un componente independiente que debe anclarse al genoma. En general, muchos scaffolds no pueden anclarse sin esfuerzos dirigidos. (Los contigs son bloques contiguos de secuencia; los scaffolds son conjuntos de contigs unidos por lecturas emparejadas de ambos extremos de un inserto plasmídico).*

La electroforesis capilar, usada en ambos métodos, permite separar fragmentos de ADN de diferentes tamaños a través de un capilar con un detector de fluorescencia, logrando una lectura precisa de aproximadamente 500 pares de bases por fragmento. Con el tiempo, los costos de secuenciación disminuyeron gracias a avances en técnicas posteriores al método de Sanger.

I.2.2. NGS: la siguiente generación de tecnología de secuenciación del ADN

La secuenciación de segunda generación o Next-Generation Sequencing (NGS) permite una secuenciación paralela y masiva, también conocida como **high-throughput sequencing**. Los principales métodos NGS incluyen 454 Roche, Solexa Illumina, ABI/SOLiD, Complete Genomics, Pacific Biosciences, Ion Torrent y Oxford Nanopore.

I.2.2.1. Preparación de librerías de NGS

Las librerías de secuenciación se preparan fragmentando el ADN y generando secuencias que luego se amplifican y procesan en el secuenciador, obteniendo las lecturas o reads. Estas librerías se amplifican clonalmente mediante tres métodos:

- **Beads:** pequeñas bolitas recubiertas de primers, donde el ADN se adhiere y se amplifica.
- **Fase sólida:** el ADN se adhiere a una superficie de cristal donde se amplifica.
- **Nanobolas:** se produce un ovillo de ADN amplificado en forma circular, que se adhiere a una placa metálica funcionalizada (con grupos funcionales) para secuenciación.

La secuenciación NGS utiliza un gran número de moléculas idénticas, permitiendo una secuenciación paralela de alta eficiencia y alto rendimiento o high-throughput. La característica de la segunda generación es que utiliza la molécula de ADN original y, sobre ella, la amplifica, es decir, la utiliza como molde para generar muchas moléculas iguales.

I.2.2.2. Clasificación de NGS: secuenciación por síntesis y por ligación

Los métodos de secuenciación de segunda generación se pueden clasificar en secuenciación por síntesis (con la enzima polimerasa) o secuenciación por ligación (con la enzima ligasa).

- **Secuenciación por síntesis (SBS)**
  - **Ciclo de terminación reversible (CRT):** una evolución del método Sanger. Se utiliza ADN unido a beads o cristales y se añaden dNTPs modificados con el grupo 3' OH bloqueado, limitando así la duplicación de la polimerasa. Cada ciclo implica la incorporación de un nucleótido, seguido de una señal fluorescente específica del nucleótido unido. Posteriormente, el grupo OH se desbloquea con un químico de lavado para que el proceso continúe. La señal que se detecta no es de un único nucleótido, si no del conjunto de nucleótidos del cluster, que debido a la amplificación clonal, debería ser la misma señal amplificada. Esto se realiza por el límite de detección de fluorescencia de los microscopios. Además, la placa con los

moldes tiene en los límites unos marcadores que permiten que el microscopio se enfoque a la altura a la que debe.

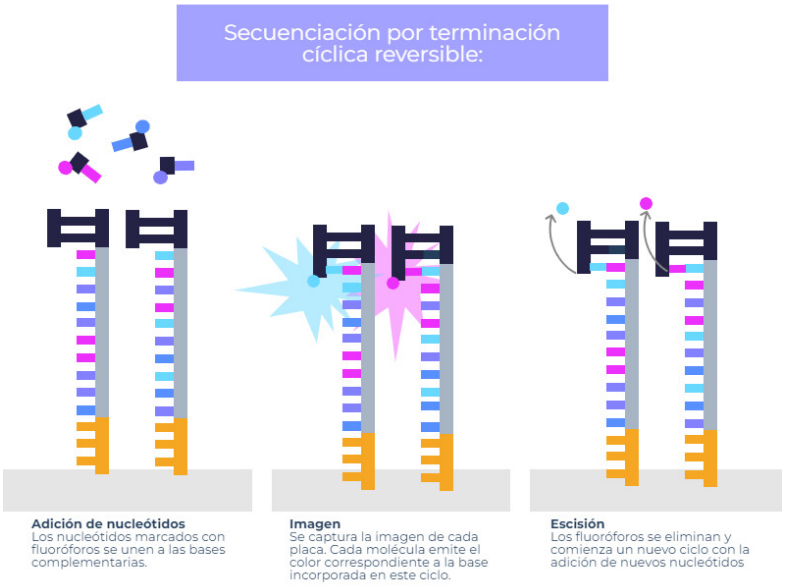

Figura I.10: ***Secuenciación por terminación cíclica reversible:*** *Esta metodología se basa en la utilización de nucleótidos marcados con fluoróforos en una reacción de síntesis de ADN. Cada vez que uno de estos nucleótidos se incorpora a la cadena, el sistema toma una captura y registra de qué tipo de nucleótido se trata. Una vez tomada la captura, se eliminan los fluoróforos de los nucleótidos que se han incorporado y se continúa la síntesis de la cadena con nuevos nucleótidos marcados.*

Una vez terminada la secuenciación, se utiliza como primer para secuenciar la cadena contraria. Esto se debe a que el microscopio va enfocando peor y se pierde calidad. Cada señal emitida por el fluoróforo se conoce como call o llamada. Cada call tiene una confident score de Q, que se calcula mediante la fórmula $Q = -10 \cdot log_{10} P$. Por tanto, si Q es 30, P sería 10^{-3}, representando P la probabilidad de error. La información que se obtiene en el archivo es la secuencia obtenida con un valor Q asociado codificado en ASCII.

Los microscopios se clasifican en microscopios de 4 canales y de 2 canales. Los microscopios de 4 canales tienen una mayor calidad al poder distinguir cada uno de los nucleótidos, mientras que los de 2 canales utilizan la combinación de dos fluoróforos: se detecta verde, rojo, la combinación entre verde y rojo, y la ausencia de fluorescencia. Esto último es algo arriesgado, ya que algunos nucleótidos podrían perder el fluoróforo y se consideraría ausencia de fluorescencia. No obstante, estos microscopios de 2 canales, pese a tener una peor calidad, son más rápidos y baratos. Respecto al secuenciador, hay varios tipos, por lo que al elegir uno se tendrá que tener cuenta el caso de uso y el dinero disponible (la página de Illumina tiene tablas comparativas para elegir el mejor secuenciador para cada caso).

Las ventajas de la secuenciación CRT es que es la que produce la mayor cantidad de secuencias secuenciadas a la vez (mayor throughput). La

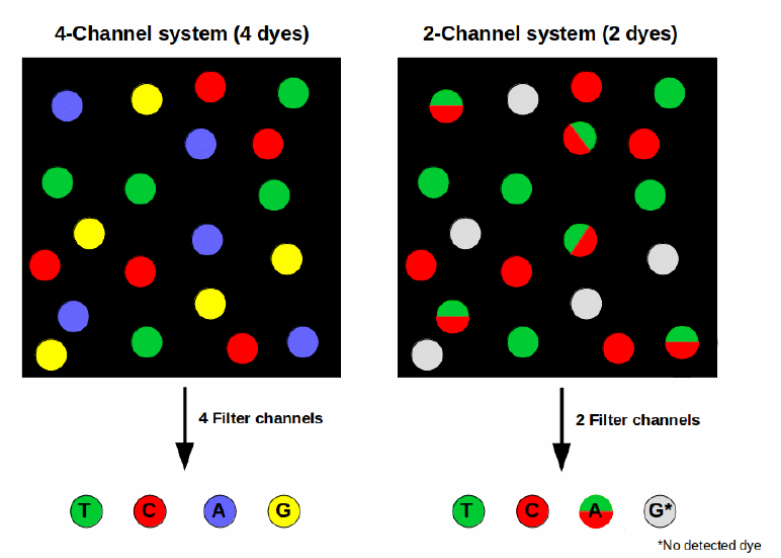

Figura I.11: *Comparación entre los microscopios de 4 y de 2 canales.*

desventaja es el límite que puede secuenciar, que es en torno a 150 bases por cada extremo.

- **Adición de nucleótidos simple (SNA):** en cada ciclo se añade un solo tipo de nucleótido, detectando su incorporación. Este método es sensible a los homopolímeros (repeticiones del mismo nucleótido), lo que puede generar problemas de fase si la señal no es proporcional al número de nucleótidos añadidos.
  - **Pirosecuenciación:** emplea pirofosfato liberado en la síntesis de ADN. Debido a su enlace de alta energía, la acción de la pirofosfatasa acoplada a la luciferasa produce que se emita una señal de luz proporcional al número de nucleótidos añadidos. Este método es rápido, económico y preciso, aunque presenta limitaciones con secuencias largas debido al cambio de fase en el momento en el que se produzca un error. La calidad de la secuenciación es Q45 (99,997 %).
  - **Ion Torrent proton detection:** mide el cambio de pH (cambio de potencial) que ocurre al liberar un protón durante la polimerización del ADN. Al final de cada ciclo es necesario lavar para evitar la señal cruzada. La técnica es económica y ampliamente utilizada en hospitales, pero presenta desafíos con secuencias largas debido a la falta de proporcionalidad en la señal en secuencias con regiones muy repetitivas (si se unen dos nucleótidos en lugar de uno, la señal es proporcional a los dos, pero cuando se unen 50 nucleótidos, el cambio de potencial no es proporcional a los 50).

- **Secuenciación por ligación (SBL)**
  - **Secuenciación por SOLiD:** Este método emplea sondas de ligación con dos bases complementarias a la base que se secuencia. En cada ciclo, una ligasa une una sonda marcada con un fluoróforo y luego se elimina la fluorescencia para repetir el ciclo, generando datos precisos, aunque menos comunes en la práctica. Se van mapeando dos nucleótidos a la

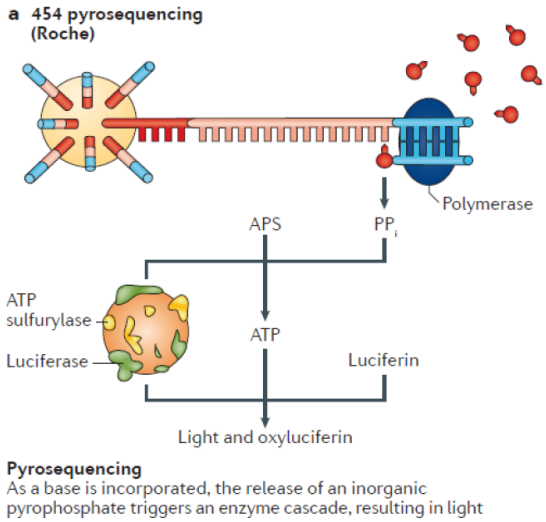

Figura I.12: *Esquema de la pirosecuenciación, tecnología que permite determinar el orden de una secuencia de ADN mediante luminiscencia.*

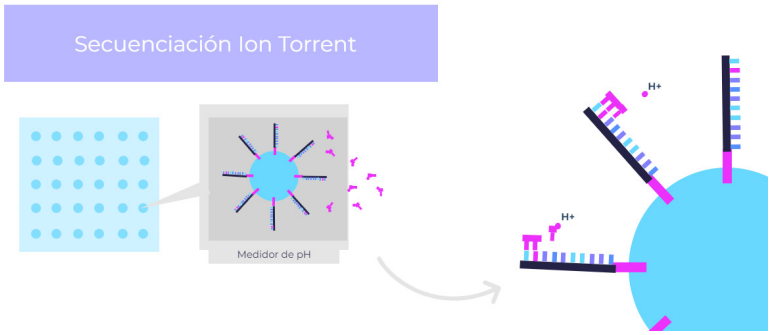

Figura I.13: ***Secuenciación por Ion Conductor (Ion Torrent Sequencing)**: Se trata de una estrategia que se basa en la detección de las modificaciones en el pH que se producen en la síntesis de ADN. Para ello, se van incorporando nucleótidos a una cadena de ADN, provocando que se libere un protón (H+) en la reacción y, por tanto, que se vea modificado el pH. Para poder diferenciar cuál de los cuatro tipos de nucleótidos se ha introducido en cada posición de la secuencia, se repiten varios ciclos, cada uno de ellos, con la adición de un único tipo de nucleótido.*

vez, dejando un espacio de 3 nucleótidos. Por ello, se repite cinco veces añadiendo espaciadores que permitan el solapamiento de las lecturas.

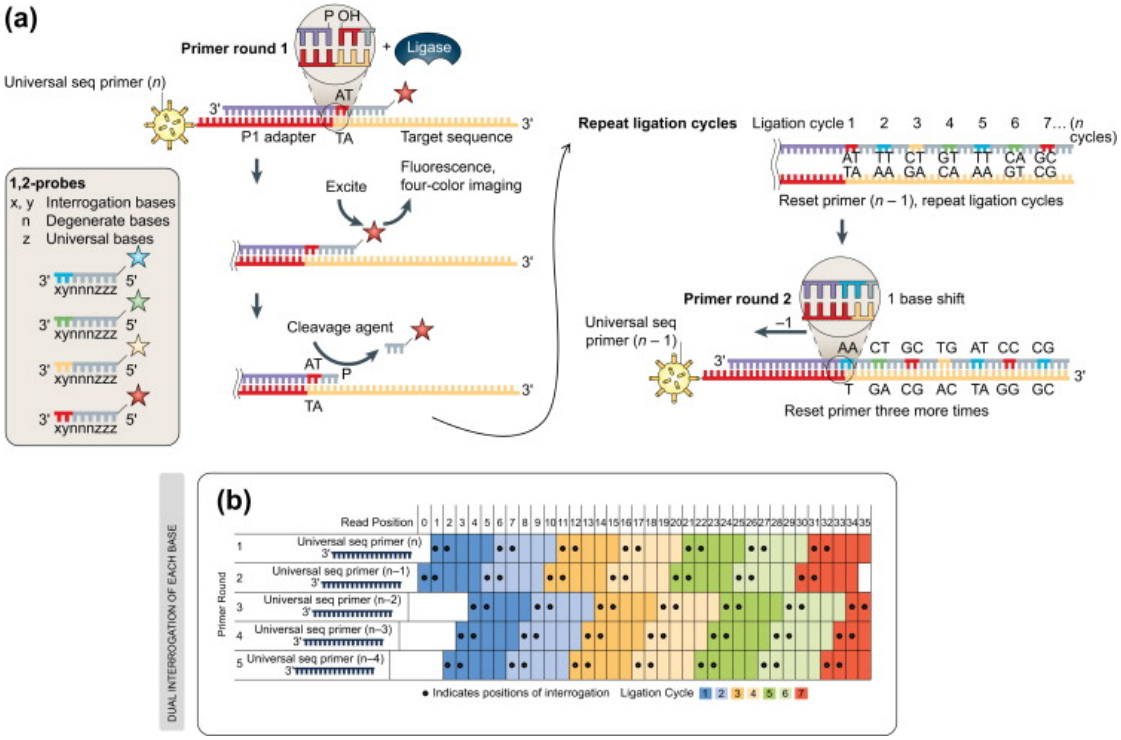

Figura I.14: ***Ilustración del método de secuenciación por ligación utilizando la plataforma SOLiD.*** *(a) Esquema de los diferentes pasos seguidos por el método SOLiD de ligadura por cuatricromía: hibridación de primers, ligadura selectiva de las sondas, obtención de imágenes por cuatricromía y escisión de la sonda. El ciclo SOLiD se repite nueve veces más. El producto de extensión se elimina y la plantilla se reajusta con un cebador complementario a la posición n -1 para una segunda ronda de ciclos de ligación. (b) Se realizan cinco rondas de reajuste del cebador para cada etiqueta de secuencia. Mediante el procedimiento de reajuste de cebadores, prácticamente todas las bases se consultan en dos reacciones de ligación independientes con dos cebadores diferentes.*

- **Complete Genomics (Nanoballs, BGI):** Amplifica ADN mediante rolling circle para formar ovillos, que luego se hibridan en una placa. Las sondas fluorescentes hibridan de forma iterativa para secuenciar el ADN. Este método permite una alta densidad de secuencias pequeñas, teniendo así una precisión muy alta. No obstante, permite secuenciar lecturas pequeñas y, al final, se utilizan distintos adaptadores a los que se unen las sondas, permitiendo aumentar el throughput. Hay partes de la tecnología que no se conocen y está principalmente disponible en China.

I.2.2.3. Limitaciones y desafíos en NGS

La NGS de segunda generación presenta una tasa de error de Q25-Q35 y ciertas limitaciones:

- **Lecturas cortas** generan dificultades en el ensamblaje de genomas completos y en la identificación de variantes estructurales.
- **Errores de secuenciación** especialmente en regiones complejas y repetitivas, como secuencias AT/GC (SBS) y homopolímeros (SNA).
- **Sesgo de amplificación** algunas regiones se amplifican mejor que otras, afectando la uniformidad en las lecturas.
- **Alto coste de los equipos**
- **Fenómeno de la cadena retrasada** cuando no se incorpora un nucleótido, produciendo un descabalgamiento del ciclo de lectura real y el ciclo de lectura en el que creemos que estamos.
- **Persistencia de errores en los cluster** al producirse un error en el cluster, el error se queda a lo largo de la secuenciación.
- **Cambios epigenéticos**

I.2.3. Resumen

La secuenciación ha cambiado la forma de hacer y entender la biología. La secuenciación de segunda generación o NGS permite secuenciar millones de moldes de ADN al mismo tiempo. Generalmente, el molde de ADN es amplificado clonalmente, y las llamadas se hacen mediante el consenso de los moldes clonales. Hay dos tipos de secuenciación NGS: por síntesis con la polimerasa o por ligación (SOLiD y Nanoballs). Hay dos tipos de secuenciación por síntesis. La adición simple de nucleótidos (pirosecuenciación 454 y Ion Torrent) añade un dNTP distinto en cada ciclo, pero tiene problemas con moldes homopoliméricos. La terminación cíclica reversible (Illumina) añade todos los dNTP en cada ciclo y secuencia la misma posición en el molde, pero puede sufrir de desfase.

I.2.4. Quizz

1. Which of the following NGS platforms offers the highest accuracy?
    - Ion Torrent
    - 454
    - Illumina
    - SOLiD

    Answer: Illumina

2. What is the reversible chain termination method (CRT)?
    - A method that uses modified nucleotides to stop DNA synthesis
    - A method that involves the use of anchors and fluorescent probes

- A sequencing process based on detecting pH changes
- A real-time PCR technique

Answer: A method that uses modified nucleotides to stop DNA synthesis

3. Which sequencing method employs PCR amplification and ddNTP?

   - Maxam-Gilbert
   - Ion Torrent
   - Sanger
   - Nanoballs

   Answer: Sanger

4. What is a common problem in second generation sequencing methods?

   - Low cycling efficiency
   - Difficulty detecting homopolymers
   - Low precision in GC regions
   - Very long execution times

   Answer: Difficulty detecting homopolymers

5. What NGS technology allows real-time sequencing?

   - 454
   - SOLiD
   - Illumina
   - PacBio

   Answer: PacBio

6. What achievement was reached with the Human Genome Project (HGP)?

   - Sequencing of the complete human genome
   - Sequencing of the mouse genome
   - The first automated sequencing
   - Creation of the nanoball sequencing method

   Answer: Sequencing of the complete human genome

7. What technology uses circle displacement amplification to generate nanoballs?

   - PacBio
   - SOLiD
   - Illumina
   - BGI

Answer: BGI

8. What was the first NGS instrument developed?

   - Illumina
   - Ion Torrent
   - 454
   - SOLiD

   Answer: 454

9. What error is common in sequencing based on the nucleotide addition method (SNA)?

   - Errors from long reads
   - Errors in low complexity regions
   - Difficulty detecting single nucleotide polymorphisms (SNPs)
   - Problems with homopolymers

   Answer: Problems with homopolymers

10. What is the basis of the Maxam-Gilbert method for DNA sequencing?

    - Amplification of fragments on a solid surface
    - Use of chemicals to break the DNA molecule
    - Adding nucleotides iteratively
    - Electronic detection of pH changes

    Answer: Use of chemicals to break the DNA molecule

11. What is one of the main advantages of massively parallel sequencing?

    - Generation of long and accurate reads
    - Ability to sequence multiple DNA templates at the same time
    - Capability to perform sequencing at low cost
    - Reduction of error rates in reads

    Answer: Ability to sequence multiple DNA templates at the same time

12. What technique uses clonal amplification of DNA on solid surfaces?

    - Ion Torrent
    - Pyrosequencing
    - Sequencing by synthesis
    - Nanoballs

    Answer: Sequencing by synthesis

13. Which sequencing platform is based on proton detection

- SOLiD
- Illumina
- Ion Torrent
- PacBio

Answer: Ion Torrent

14. What was the main technique used in the Human Genome Project?

    - Maxam-Gilbert
    - Pyrosequencing
    - Sanger sequencing
    - Second generation NGS

    Answer: Sanger sequencing

15. Which NGS platform is known for its low cost per Mb sequenced?

    - Illumina
    - SOLiD
    - 454
    - Ion Torrent

    Answer: Ion Torrent

16. What is one of the advantages of reversible terminator sequencing technology (CTR)?

    - Does not require clonal amplification
    - Long read length
    - High accuracy in called bases
    - Can handle RNA templates

    Answer: High accuracy in called bases

17. What NGS technology uses a system of up to two colors for detection?

    - Ion Torrent
    - Illumina
    - PacBio
    - SOLiD

    Answer: Illumina

18. What are the disadvantages of second generation sequencing systems?

    - Low precision in SNP detection
    - Short read lengths

- High costs per sequence
- Low coverage of repetitive regions

Answer: Short read lengths

19. What is the main limitation of pyrosequencing?

    - High error rate in low complexity regions
    - Problems with homopolymers
    - High error rate in short segments
    - Difficulty in detecting structural variants

    Answer: Problems with homopolymers

20. What is one of the main disadvantages of the SOLiD platform?

    - Problems with long reads
    - High error rate in homopolymers
    - Low precision in variation detection
    - Requires an additional cycle for each read

    Answer: High error rate in homopolymers

21. Which NGS platform is based on luminescence detection?

    - 454
    - Illumina
    - SOLiD
    - Ion Torrent

    Answer: 454

22. What is a main disadvantage of secong-generation sequencing systems?

    - Low precision in SNP detection
    - Short read lengths
    - High costs per sequence
    - Low coverage of repetitive regions

    Answer: Short read lengths

23. What key feature defines the Sanger chain termination method?

    - Use of specific enzymes to emit light
    - Use of ddNTPs to stop DNA replication
    - Probe and anchor-based sequencing
    - Use of a microchip with electronic sensors

Answer: Use of ddNTPs to stop DNA replication

24. What type of methods are grouped under the term NGS?

    - Massively parallel high-capacity sequencing methods
    - Manual low-precision sequencing methods
    - Methods based on RNA synthesis
    - Methods for detecting three-dimensional structures

    Answer: Massively parallel high-capacity sequencing methods

25. Which sequencing technique was the first to implement the concept of sequencing by synthesis?

    - Nanoballs
    - SOLiD
    - Sanger
    - 454

    Answer: Sanger

I.3. Aligners

I.3.1. Preparación de librería

Una librería es una colección de fragmentos de ADN de tamaño aleatorio obtenidos a partir de una muestra que se desea secuenciar. El proceso comienza con la extracción del material genético (ADN o ARN), seguido de su fragmentación en piezas pequeñas que posteriormente serán leídas. A continuación, se añaden adaptadores a los extremos de los fragmentos, lo que permite su hibridación con una fase sólida para realizar la amplificación. Finalmente, se purifican los fragmentos para obtener solo las moléculas del tamaño deseado, dependiendo del método de secuenciación que se vaya a utilizar.

I.3.1.1. Fragmentación del material genético

Existen distintas técnicas para fragmentar el ADN, cada una con sus ventajas y desventajas:

- **Aproximación por ligación:** En este método, los fragmentos de ADN se preparan añadiendo una adenina en los extremos, lo que permite la unión complementaria de los adaptadores. Esto produce fragmentos con un adaptador en cada extremo. Cada fragmento tiene así dos componentes: la secuencia del adaptador para la secuenciación y la secuencia molde de ADN. Además, cada fragmento puede llevar un identificador único (UMI, por sus siglas en inglés). Las técnicas de fragmentación por ligación incluyen:

- *Fragmentación física (sonicación)*: Mediante un sonicador, se aplican ondas sonoras que generan vibración por resonancia, dividiendo el ADN en fragmentos. La frecuencia de las ondas determina el tamaño de los fragmentos obtenidos.
- *Fragmentación química:* Se utilizan agentes químicos, como ácidos o bases fuertes, para romper los enlaces fosfodiéster del ADN. Este método es útil en casos donde la epigenética no es relevante, ya que los químicos fuertes pueden modificar las marcas epigenéticas mediante procesos de oxidación o reducción.
- *Fragmentación enzimática:* Se emplean endonucleasas, que son enzimas capaces de cortar las cadenas de ADN en puntos específicos, produciendo fragmentos con extremos cohesivos o romos. Este método permite una fragmentación precisa, pero puede introducir sesgos en la representatividad de la librería generada.

	Physical	Chemical	Enzymatic
Pro	Broad range Unbiased Less sample variation Even sized of fragments No interferences Easy to implement	Well for RNA Lower input of material	Standard lab equipment Highly scalable
Cons	Expensive equipment Loss of material Modification of bases	Cations interfere with some seq methods	Fragmentation bias Ratio material/enzymes Sample-to-sample variation

Tabla I.1: *Pros y contras de cada método de fragmentación por ligación*

- **Aproximación por tagmentación:** En este método, se utiliza la enzima tagmentasa, una transposasa que corta la secuencia de ADN e incorpora adaptadores de manera enzimática [1]. La tagmentación es rápida y eficiente, pues combina la fragmentación y la adición de adaptadores en un solo paso.

I.3.1.2. Reparación de extremos y ligación de adaptadores

Después de la fragmentación del ADN, es necesario reparar los extremos de los fragmentos. Como la fragmentación no suele producir cortes limpios, los fragmentos generados suelen presentar extremos sobresalientes (overhangs). Para corregir esto, se realiza un tratamiento enzimático con polimerasas, que además añade una adenina (A) en los extremos 3'. Estos extremos, con la adenina añadida, facilitan la ligación de los adaptadores, los cuales suelen tener un overhang de timina (T) para permitir una unión complementaria con los fragmentos de ADN.

Los adaptadores empleados en la secuenciación tienen distintas configuraciones. Por lo general, se colocan dos tipos de adaptadores: el adaptador P5 en un extremo y el adaptador P7 en el otro. Esta disposición permite identificar las direcciones de lectura durante el proceso de secuenciación. Además, algunos adaptadores incluyen identificadores moleculares únicos, conocidos como UMIs (Unique Molecular

[1] Los transposones son elementos móviles dentro del ADN.

Identifiers), que permiten rastrear de forma única cada molécula de ADN. Durante la amplificación, todas las moléculas con el mismo UMI corresponden a la misma molécula de ADN original. Esto tiene múltiples beneficios:

- **Eliminación de duplicados de PCR:** Permite distinguir duplicados generados por PCR de secuencias originales.
- **Disminuir el ratio de error:** La lectura de la misma molécula varias veces permite detectar posibles errores generados durante la construcción de la librería o la amplificación. Si diferentes secuencias presentan un UMI idéntico pero difieren en algún nucleótido, se puede deducir que ha habido un error, ya que las secuencias deberían ser idénticas. Esto ayuda a reducir el índice de error y a detectar variantes poco frecuentes.

En el método conocido como **Duplex Sequencing**, se emplean UMIs distintos en ambos extremos del fragmento de ADN. De este modo, durante el análisis de consenso, se pueden identificar las posiciones que muestran concordancia entre ambas hebras y descartar los nucleótidos mutados por error.

Aunque los métodos basados en UMIs son altamente eficaces para detectar variantes de muy baja frecuencia, su uso es limitado debido a su alto costo, ya que requieren múltiples lecturas de la misma secuencia para asegurar precisión.

I.3.1.3. Adaptadores para secuenciación de célula única (single cell)

En la secuenciación de célula única (Single Cell), se emplean adaptadores y primers que contienen UMIs específicos tanto para cada célula como para cada molécula de ADN. Esto permite diferenciar si las lecturas corresponden a una célula particular y, dentro de esa célula, a una molécula específica. En Single Cell, se utiliza un enfoque de consenso: las lecturas múltiples de la misma molécula permiten generar una secuencia de consenso para mejorar la precisión de los datos.

La construcción de librerías en Single Cell se realiza mediante chips que permiten el paso de flujo de células individuales junto con liposomas que contienen la mezcla de PCR. Cada mezcla de PCR tiene adaptadores con un barcode específico de célula, pero diferentes barcodes de molécula. Esto permite identificar y diferenciar las células individuales entre sí.

I.3.2. Formatos de datos

El formato del alineador es el FastQ, Fast5 o HDF5 (este último solo para single cell).

Cuando se trabaja con alineadores, normalmente se genera un perfil con picos a los que se asigna el nombre de la base. El alineador da un archivo final con las bases asignadas y la calidad de lectura de cada base en formato Phred33 codificado en caracteres ASCII.

Antiguamente, en la cabecera del ID se daba una serie de elementos como el nombre del instrumento, el flowcell, las coordenadas x e y del cluster, el número de la muestra y si el índice del par. El formato nuevo tiene una cabecera distinta.

Single-end secuencia desde un lado hacia delante. Pairend tiene una lectura en ambas direcciones, teniendo así información en ambos lados. Se puede hacer solapante (cuando se quiere tener más evidencia de lo que se secuencia) y no solapante (cuando se quiere mapear las estructuras).

Al realizar un experimento de secuenciación, se pasa por un programa llamado FastQC que da la calidad a lo largo de la lectura de las bases mediante el Q score. El tipo de gráfico visual es un diagrama de barras y cajas. Lo normal es que conforme se van produciendo los ciclos de lectura, la calidad va bajando, pero debería mantenerse en un rango de confianza.

FastQC devuelve distintas métricas:

- La calidad por la media de la secuencia
- Las proporciones de bases por posición
- El contenido de las bases por posición
- El contenido GC: al principio suele haber un desabarajuste debido a la secuencia de los alineadores, pero si el contenido GC se mantiene estable después, significa que la secuenciación ha salido bien y simplemente se corta el principio.

Preprocesado El preprocesado aumenta la calidad, mejora el mapeo, elimina los contaminantes, elimina los sesgos y elimina los segmentos no informativos.

La secuenciación se compara con genomas de referncia. Normalmente, estos genomas están en formato Fasta. Dentro de ellos se encuentran los nucleótidos representados en el código IUPAC. Los genomas de referencia se hacían mediante la información genética de un individuo. Actualmente, se hace a través del consenso de los genomas de muchos individuos, por lo que hay posiciones que pueden tener mucha confianza, pero son diferentes entre un individuo y otro. Los genomas de referencia se pueden encontrar en https://hgdownload.soe.ucsc.edu/downloads.html. Es imporante conocer el genoma de referencia y la anotación (hay diferencias entre los repositorios europeos y americanos a la hora de anotar los cromosomas).

Alineamiento Un alineamiento es una forma de comparar secuencias de ADN, ARN o proteínas e identificar regiones de similitud que puedan tener relaciones funcionales, estructurales o evolutivas entre especies. Los objetivos del alineamiento son:

- Determinar el grado de homología para inferir relaciones filogenéticas
- Identificar dominios funcionales
- Comparar el gen con sus productos
- Encontrar posiciones homólogas
- Identificar diferencias

Alineamiento vs mappeo En el alineamiento, cada secuencia se mide con la secuencia de referncia para determinar lo buena que es la secuencia en cada posición.

En el caso del mapeado, se busca los mejores loci en el que una secuencia se podría alinear.

Hay programas que permiten realizar el mapeo y se conocen como alineadores de corta lectura. Buscan tener una alta eficiencia de velocidad y memoria utilizada. Un mapeado se puede configurar de diferntes maneras: para que haya mapeados únicos, mapeados múltiples o que devuelva mapeados con calidades parciales.

Dependiendo del problema que se esté abordando, se utiliza un alineador distinto. Hay alineadores para ADN (Novoalign, Bowtie2, BWA) y alineadores para ARN (RSEM, SALMON, SLEUTH).

Los algoritmos de mapeo hacen una primera etapa de indexación del genoma. Este proceso es muy lento y consume mucha memoria, pero la ventaja es que solo se hace una vez por cada programa que se utilice y por cada genoma de referencia. Utilizando hashes o índices, se utilizan k-mers que van anotando las posiciones en las que se encuentran. Mediante ventana móvil se van anotando los distintos k-mers y las posiciones en las que se encuentran. Una vez con las lecturas, se busca en el diccionario de k-mers las posiciones en las que se puede encontrar (sin utilizar ventana móvil para las lecturas). Una vez teniendo el árbol de posibilidades, se va evaluando la compatibilidad. Si hay una mutación que no coincide con la referencia, en el índice no se va a encontrar el k-mer, por lo que se realiza un alineamiento con los posibles k-mers. El mismatch se puntúa con una calidad de mapeo más baja. Ocurre lo mismo cuando la evaluación de las posiciones no es compatible. Así, los match parciales reducen la puntuación de mapeo en base a la distancia de edición. El objetivo es minimizar los gaps.

La transformación de Burrows-Wheeler utliza un algoritmo de compresión que permite buscar de manera rápida con qué mapean las lecturas en el genoma de referencia.

Una vez que se ha mapeado la secuencia al genoma, se pueden tener distintos estados de las lecturas: lecturas que no han mapeado, lecturas que han mapeado a una posición y lecturas que mapean a varias posiciones. Normalmente se trabaja con lecturas que mapean a una sola posición. De los multimappers, se pueden subdividir en alineamientos primarios y secundarios en función a su puntuación. Se puede elegir reportar todos los alineamientos, los mejores alineamientos, alineamientos aleatorios de los mejores y todos aquellos que superen un threshold.

Mapping quality Utiliza la misma escala que las lecturas, es decir, Phred33. Los alineamientos están en formato SAM, BAM o CRAM. SAM es un formato de texto plano, BAM en binario. Se pueden interconvertir con la herramienta samtools, especificando el grado de compresión.

La cabecera de BAM empieza con y guarda la versión de SAM, si está ordenado, los contigs, la información del mapeo y la secuencia de la lectura. Todo lo que empieza por es metadato, y lo que no es lectura. El alineamiento tiene:

- Nombre de la lectura
- Flag: si está mapeado, no mapeado, etc.
- Chromosoma

- Posición de inicio en el cromosoma
- Mapping quality
- Cigar: indica los eventos de la lectura a la hora de alinearse con la referencia; eventos de match, mismatch, insert, delete.
- Información donde ha mapeado la lectura y su par en caso de que sea pairend

CRAM es como BAM, pero lo utiliza el EBI fundamentalmente.