Facultad de Ciencias Experimentales

Grado en Biología

UNIVERSIDAD DE JAÉN
Facultad de Ciencias Experimentales

Trabajo Fin de Grado

Secuenciación de novo de genomas de plantas

Alumno: Julián Santiago Orta

Julio, 2021

UNIVERSIDAD DE JAÉN

Facultad de Ciencias Experimentales

GRADO EN BIOLOGÍA

Trabajo Fin de Grado

SECUENCIACIÓN DE NOVO DE GENOMAS DE PLANTAS

Julián Santiago Orta

Jaén, julio 2021

ÍNDICE

1. RESUMEN

La secuenciación del genoma se lleva a cabo mediante una serie de técnicas para determinar la secuencia genómica de un individuo. Durante los últimos 50 años, la tecnología ha evolucionado y perfeccionando estos métodos, haciendo que se obtengan lecturas de gran tamaño en muy poco tiempo, avanzando desde la secuenciación de Sanger hasta la más reciente Oxford Nanopore Technology, pasando por Illumina, la que posee mayor relevancia en la actualidad. También se profundiza en la importancia de los genomas vegetales. Debido a su naturaleza repetitiva y su poliploidía, se ha tenido que poner a prueba y mejorar cada tecnología, hasta el punto de haber sido secuenciados en su totalidad prácticamente. También cabe mencionar la importancia de las plantas modelo como *Arabidopsis thaliana*, la cual fue elegida gracias a su corto ciclo de vida y pequeño tamaño del genoma a la hora de ser secuenciada, y que sirvió de modelo para la secuenciación del resto de genomas de plantas que se conocen en la actualidad.

ABSTRACT

Genome sequencing is carried out using different techniques to determine the genomic sequence of an individual. During the last 50 years, technology has evolved and refined these methods, obtaining large readings in a very short time, advancing from Sanger sequencing to the most recent Oxford Nanopore Technology, passing through Illumina, the one with the biggest relevance today. The importance of plant genomes is also explored in this work. Due to its repetitive nature and polyploidy, it has been needed to test and improve that technology, making easier to practically complete a plant genome sequence. It is also worth mentioning the importance of model plants such as *Arabidopsis thaliana*, which was chosen due to its short life cycle and the small size of the genome at the time of being sequenced, and that served as a model for the sequencing of the rest of the plant genomes that are currently known.

2. ANTECEDENTES

La secuenciación del genoma consiste en una serie de técnicas bioquímicas para determinar la secuencia de ADN del genoma de un individuo, es decir, determinar el orden de los nucleótidos que conforman la molécula de ADN, la cual contiene la información genética del individuo en cuestión. La finalidad de esta serie de técnicas está en conocer toda la información genética presente en el genoma, lo que permite que, una vez analizada, se puedan identificar los genes que conforman dicho genoma para comparar los alelos entre diferentes individuos dentro de la misma especie y cotejar posteriormente especies o poblaciones para poder apreciar el efecto que producen en los fenotipos y comprender que función cumplen.

Gracias a estas técnicas se ha producido un gran aumento de nuestro conocimiento genético de las especies. En relativamente poco tiempo hemos avanzado desde conocer algunos genes aislados que provocaban algunos caracteres concretos a conocer el genoma completo de una gran diversidad de organismos. Además, al existir genes que poseen funciones similares en diferentes especies, se facilita su localización en el genoma además de poder extrapolarse la información a especies similares.

Estudiar los genomas en su totalidad permite un gran avance en la elaboración de una guía para las relaciones filogenéticas y una mayor comprensión de los fenómenos evolutivos. Las nuevas técnicas de secuenciación y edición de genes van a permitir la modificación genética de individuos para mejorar alguna característica concreta sin afectar al resto de características y de esta forma mejorar la productividad, eficiencia o corregir defectos o alteraciones negativas.

La comparación de diferentes secuencias de nucleótidos refleja las diferencias entre ellas. Entre las principales fuentes de variación en plantas se encuentran los transposones, que son secuencias de ADN que pueden replicarse e insertar una copia suya en un lugar concreto del genoma; reorganizaciones de cromosomas o inserciones y deleciones de nucleótidos. Cotejar las variaciones ayuda a conocer que zonas deberían analizarse en mayor profundidad. Además han de tenerse en cuenta otras condiciones como el factor epigenético, el cual se involucra en los cambios que se producen hereditariamente pero no se atribuyen a alteraciones de la secuencia de ADN.

2.1. Evolución de las técnicas de secuenciación

2.1.1. Primeros avances en secuenciación

En 1944, Oswald Avery, Colin McLeod y Maclyn McCarty demostraron que la información genética residía en el ADN y no en las proteínas como se pensaba anteriormente. A partir de entonces, se comenzó a llevar a cabo una carrera para descubrir su estructura. La estructura de doble hélice fue propuesta en 1953 por James Watson y Francis Crick, apoyados en los trabajos de Rosalind Franklin y Maurice Wilkins. Gracias al trabajo de Severo Ochoa, Marshall W. Nirenberg, Har Gobind Khorana y Robert W. Holley, se descubrió el código genético en 1961, por lo que se pudo determinar como la secuencia de nucleótidos de un determinado gen activo después de ser transcrita en moléculas de ARN, se traducían en proteínas. El estudio de las secuencias condujo al "Principio de colinealidad", en el que los tripletes o codones están relacionados con la síntesis de un aminoácido determinado. Desde entonces, muchos estudios se han centrado en conocer las secuencias de nucleótidos para comprender la información genética.

La secuenciación del ARN, se desarrolló previamente a la del ADN, ya que por razones técnicas en 1976 era más sencilla de llevar a cabo que la del ADN. El mayor hito en la secuenciación del ARN, fue la secuenciación del primer gen completo y del genoma del Bacteriófago MS2, lo cual fue publicado por Walter Fiers y sus colaboradores de la Universidad de Gante.

En 1977, se desarrollaron simultáneamente dos tecnologías de secuenciación. Por un lado, *"el método químico de Maxam y Gilbert, que consiste en fragmentar las cadenas de ADN, previamente marcadas mediante reacciones químicas específicas para cada base (purinas AG y pirimidinas CT) y después, separarlas por electroforesis en función de su tamaño en geles de alta resolución, típicamente de poliacrilamida"* (De Necochea *et al*., 2004). Por otro lado, Frederick Sanger desarrolló un método que lleva su nombre, también conocido como método de terminadores de cadena. Esto incluye la síntesis de una hebra complementaria a la hebra utilizada como molde mediante ADN polimerasa. En cada reacción, además de los cuatro nucleótidos, se agregan terminadores de cadena, que no tienen los grupos hidroxilo necesarios para elongar la copia de ADN. Cada vez que un dideoxinucleótido es incorporado, impide la continuación de la cadena y genera fragmentos de tamaños diferentes que son

separados por electroforesis. Este método se fue mejorando en los años posteriores para aumentar su eficacia y disminuir su precio, además de reducir el tiempo requerido para llevarse a cabo.

Algunos de los avances en la automatización del método Sanger fueron la técnica PCR desarrollada por Mullis en 1989, que permitía la amplificación de pequeñas muestras de ADN e implicaba el uso de Taq polimerasa, una enzima que permanece activa a altas temperaturas, es decir, termoestable. Otros de los avances fueron el uso de un sistema de separación molecular más eficiente, la electroforesis capilar, que permitía lecturas más largas, y el uso de terminadores marcados con fluorescencia, que permitían que las cuatro reacciones se pudieran llevar a cabo en un tubo en lugar de realizarse de forma individual como se hacía hasta ese momento.

"El primer secuenciador automático (ABI 370A) salió al mercado en 1986 a través de la empresa Applied Biosystems que permitía realizar lecturas de hasta 200 pb por hora y muestra, pudiendo cargar hasta 96 muestras" (Hutchison, 2007; Pareek *et al*., 2011).

Ya a finales del siglo XX, los investigadores y las empresas relacionadas con la biotecnología comenzaron a interesarse por las posibles aplicaciones de la secuenciación en medicina y se lanzó el proyecto de secuenciación del Genoma Humano. El PGH (Proyecto del Genoma Humano) se propuso en la década de 1990, y se esperaba que durara 15 años, con un coste aproximado de 3.000 millones de dólares estadounidenses. "*Distintos centros en varios países trabajaban con secuenciadores capilares, método automatizado de Sanger, que permitía una lectura más rápida y de fragmentos de mayor tamaño*" (Kchouk *et al*., 2017). De forma paralela*, "Craig Venter con su empresa Celera, comenzó a secuenciar el genoma humano mediante la técnica Whole Genome Shotgun sequencing, en la que el ADN es fragmentado, secuenciado y ensamblado mediante un programa informático a partir del alineamiento de contigs, que son secuencias de parte del genoma"* (Weissenbach, 2016). En 2001 fue publicado el primer borrador del genoma humano gracias a los datos proporcionados por Celera con una profundidad (número medio de veces que cada nucleótido ha sido secuenciado en distintas lecturas) de 5X además de los datos públicos, lo que le sumó 3X más.

En los años 2000 se secuenció prácticamente la totalidad del genoma humano gracias a la técnica de secuenciación "Shotgun", ya que se desarrolló un avance de esta

técnica empleando la secuenciación de extremos por pares, lo que creó la secuenciación Shotgun de doble cañón. Ésta empleaba fragmentos de distintos tamaños que orientados en direcciones opuestas eran superpuestos para recomponer la lectura de la secuencia.

2.1.2. Secuenciación de Nueva Generación (NGS)

Con el avance de la tecnología también surgieron nuevas técnicas de secuenciación, las cuales se denominaron NGS (Next Generation Sequencing), también conocidas como secuenciación masiva paralela, la cual permitía secuenciar un elevado número de fragmentos con un coste mínimo, aunque en su contra el tamaño de las lecturas que se obtenían era menor al que se conseguía con otras técnicas de secuenciación. "*Con la tecnología de NGS se consigue una mayor profundidad, lo que aumenta el grado de fiabilidad de las secuencias*" (Rodríguez-Santiago et al., 2012). A partir del año 2005 fueron desarrollados nuevos secuenciadores, los cuales se apoyaban en diferentes técnicas. "*El primer secuenciador de segunda generación que estuvo en el mercado fue desarrollado por 454 Life Sciences y se basaba en la pirosecuenciación, en la que se detecta la liberación de pirofosfato cuando se agrega un nuevo nucleótido*" (Hutchison, 2007; Kchouk et al., 2017).

La lectura de secuencias cortas se enfocó de manera que pudieran realizarse de dos formas distintas: secuenciación por ligación (SBL) y secuenciación por síntesis (SBS). Además de esto pueden ser clasificadas principalmente en tres métodos de secuenciación principales: Roche/454 lanzado en 2005, Illumina/Solexa en 2006 y en 2007 el ABI/SOLiD.

La tecnología de Illumina significó un gran paso en el sector, ya que se pasó a realizar un alto número de lecturas de hasta 6000Gb. Las lecturas dejaron de ser tan cortas, ya que ahora cada segmento constaba con unos 250pb, y en adición a todo esto era capaz de llevar a cabo 20 mil millones de lecturas, lo cual disminuyó en gran medida el tiempo que se empleaba a la secuenciación.

Life Technologies sacó posteriormente el método IonTorrent, en el que se evitaba usar la fluorescencia y se sustituía por un análisis del cambio del pH que se produce cada vez que se añade un nuevo nucleótido. Estos métodos eran más eficaces que los anteriores, pero aun así tenían la principal limitación de que en organismos complejos no eran tan eficientes debido al tamaño de las lecturas que podían llevar a cabo, ya

que era del orden de unos cientos de bases. Con el tiempo, se fueron mejorando hasta que se llegó a las técnicas de tercera generación, también conocidas como "*Tecnología de secuenciación de molécula única*" (Pareek *et al*., 2011), las cuales reducían el tiempo empleado y no necesitaban amplificaciones, con el añadido de que permitían realizar la lectura de miles de bases. Entre estas técnicas destaca el secuenciador SMRT (single-molecule real-time) de Pacific Biosciences ya que son capaces de detectar la fluorescencia desprendida en el proceso a tiempo real.

"*Las técnicas NGS tienen un mayor porcentaje de error en comparación con el método de Sanger, pero esto se ve solventado ya que son capaces de realizar un número de lecturas muy elevado, en un menor tiempo empleado para ello (de días a horas), además de una mínima cantidad de muestra de ADN necesaria para el análisis*" (Kchouk *et al*., 2017; López de Heredia, 2016). A esto se suma "*la evolución de herramientas informáticas como el programa BLAST* (Basic Local Alignment Search Tool) *para comparar nuevas secuencias con las que se almacenan en bases de datos*" (Hutchison, 2007).

En la actualidad todas estas técnicas de secuenciación coexisten, además de sus variantes, por lo que es frecuente que en cada proyecto de secuenciación se utilice una combinación de algunas de ellas.

Para la secuenciación de grandes fragmentos de ADN pueden llevarse a cabo dos estrategias. El "Chromosome walking", el cual consiste en fragmentar ADN para introducir los fragmentos que se obtienen en un vector de clonación, crear un banco de moléculas y secuenciar cada fragmento que se solapa para conseguir la secuencia completa. Esta técnica fue empleada en el Proyecto Genoma Humano, pero es a su vez una estrategia muy costosa y laboriosa, ya que se necesita un número elevado de iniciadores ya que cada fragmento es único.

La otra estrategia es la conocida como "Shotgun", en ella se fragmenta el ADN aleatoriamente y es ensamblado mediante programas a partir de coincidencias en las secuencias, por lo que se va reconstruyendo el genoma. Sus ventajas frente a la técnica anterior radican su rapidez y es requerido un menor número de iniciadores. En el caso de esta técnica es necesario secuenciar varias veces el ADN para lograr una secuencia completa, por lo que se requiere un mayor esfuerzo bioinformático para el ensamblaje de los fragmentos, ya que son secuenciados sin seguir ningún orden.

"*Esta fue la técnica empleada por Venter para secuenciar el genoma humano de forma paralela al proyecto anunciado en los años 90*" (De Necochea et al., 2004).

"*El desarrollo de la genómica no hubiera sido posible sin el desarrollo paralelo de la bioinformática, imprescindible para almacenar, ensamblar y comparar la ingente información producida por los millones de lecturas actualmente conocidas y almacenadas en los bancos de información genética*" (Weissenbach, 2016). Dichas mejoras en las técnicas de secuenciación, así como el rápido desarrollo de la bioinformática permitieron el estudio de genomas completos.

Estos avances en las tecnologías de secuenciación han permitido pasar de la genética directa, es decir, buscar los genes que producen una característica determinada, a la genética inversa, la cual se basa en la secuencia genómica y busca identificar los genes presentes en ella e investigar su función.

2.1.3. Tecnología actual de secuenciación

A partir del año 2005, la estrategia shotgun comenzó a emplearse en combinación con técnicas más eficientes y actuales como son las tecnologías de Illumina, SMRT o nanopore.

Ya que de estas técnicas de secuenciación se obtienen grandes cantidades de datos, éstos son almacenados en grandes bases de datos con una gran potencia de cálculo y almacenamiento ya que, por ejemplo, en el caso de cada genoma humano existen unos tres mil millones de pares de bases. Todo esto no habría sido posible de no ser por los avances en el campo de la bioinformática, con la llegada de microprocesadores y ordenadores potentes, los cuales permiten almacenar semejante cantidad de información y disponer de ella de una forma rápida y sencilla.

En la actualidad una de las técnicas más efectivas para lograr una secuenciación de alto rendimiento para el genoma completo es la tecnología de nanoporos, la cual está patentada por la Universidad de Harvard junto con Oxford Nanopore Technologies para empresas del campo biotecnológico.

Otra de las técnicas que logran una secuenciación de alto rendimiento es mediante el uso de fluoróforos, los cuales son utilizados por las tecnologías de Illumina y SMRT de Pacific Biosciences.

Por último cabe mencionar la técnica de pirosecuenciación, pese a estar descatalogada desde hace varios años y actualmente considerarse obsoleta se basaba en el principio de síntesis. Ésta técnica fue desarrollada en 1996 por Pål Nyrén y su estudiante Mostafa Ronaghi en el Instituto Real de Tecnología en Estocolmo. En la actualidad se emplea la tecnología de Illumina, ya sea en solitario o en combinación con alguna de las técnicas anteriores (SMRT o nanopore)

2.1.4. Grandes hitos de la secuenciación del genoma

La base la vida es la información genética. Por lo tanto, el acceso a la secuencia genómica, y a cómo se codifican los genes dentro de ésta, se ha convertido en un recurso fundamental en biología. Aunque el ritmo de secuenciación del genoma en biología vegetal va por detrás del de los sistemas microbianos y de mamíferos, la aplicación de la genómica es muy importante entre las subdisciplinas de la ciencia vegetal, incluidas la agronomía, la bioquímica, la silvicultura, genética, horticultura, patología y sistemática.

Tabla 2.1 Lista de avances importantes en genómica vegetal

Año	Hito	Plataforma	Tamaño	Referencias
1993	Primer artículo EST (*Homo sapiens*)	Sanger	1813 secuencias	Adams *et al.* (1993)
1994	*Arabidopsis thaliana* ESTs	Sanger	1518 ESTs	Newman *et al*. (1994)
1995	Primer genoma bacteriano (*Haemophilus influenzae*)	Sanger	1.83 Mb	Fleischmann *et al.* (1995)
1999	Cromosoma humano 22	Sanger	33.4 Mb	Dunham *et al.* (1999)
1999	Cromosomas 2 y 4 de *A. thaliana* BAC-by-BAC	Sanger	19.6 y 17.4 Mb	Lin *et al*. (1999); Mayer *et al*. (1999)
2000	Genoma de *Drosophila melanogaster*; WGS	Sanger	114.8 Mb	Adams *et al*. (2000)
2000	Genoma completo de *A. thaliana* BAC-by-BAC	Sanger	115.4 Mb	Arabidopsis Genome Initiative (2000)
2001	Genoma humano completo	Sanger	2.91 Gb	Venter *et al*. (2001); Lander *et al*. (2001)
2002	Genomas de *Oryza sativa;* WGS	Sanger	390 Mb	Yu *et al*. (2002); Goff *et al*. (2002)
2003	Zea Mays; representación reducida	Sanger	132 Mb aprox.	Whitelaw *et al*. (2003)
2004	Transcriptoma de secuenciación masiva de lecturas de *A. thaliana*	Sanger	37 millones de lecturas	Meyers *et al.* (2004a,b)

2005	Genoma de *O. Sativa*; BAC-by-BAC	Sanger	370.7 Mb	International Rice Genome Sequencing Project (2005)
2007	Veinte genomas de *A. thaliana*; re-secuenciación	Sanger	20x119 Mb (2.4 Gb)	Clark *et al*. (2007)
2008	Tres genomas de *A. thaliana*; re-secuenciación	Illumina	3x119 Mb (357 Mb)	Ossowski *et al*. (2008)
2009	Genoma de *Z. mays*; BAC-by-BAC	Sanger	2.3 Gb	Schnable et al. (2009)
2009	Genoma de *Cucumis sativus*; WGS	Sanger e Illumina	243.5 Mb	Huang *et al*. (2009a)
2010	Transcriptoma de mRNA-seq de *A. thaliana*	Illumina	271M de lecturas	Filichkin *et al*. (2010)
2011	Genoma de *Fragaria vesca*; WGS	Illumina; SoLiD; 454	220 Mb	Shulaev *et al*. (2011)
2011	Ochenta genomas de *A. thaliana*; re-secuenciación	Illumina	80x119 Mb (9.5 Gb)	Cao *et al.* (2011)
2016	Genoma de *Olea europaea*	Illumina	1.31 Gb	Cruz *et al*. (2016)
2018	Genoma de *Rubus occidentalis*	SMRT	43 Mb	VanBuren *et al*. (2018)
2020	Genoma de *Solanum tuberosum*	Oxford nanopore	741 Mb	Pham *et al*. (2020)

Esta lista provee una perspectiva histórica sobre el avance de la genómica en la biología vegetal (Tabla 2.1) y destaca el uso de tecnologías de secuenciación de próxima generación y su rápida evolución. Los desafíos en genómica, especialmente para las plantas, son importantes para crear conciencia dentro de la comunidad científica en general sobre las limitaciones actuales en el campo.

2.1.5. El primer genoma vegetal secuenciado

Ya que constaba con unas características que la hacían idónea como especie modelo para la investigación del genoma vegetal, *A. thaliana* fue la primera especie vegetal destinada a la secuenciación del genoma de novo. Fue elegida para esto ya que es una especie con un ciclo de vida muy corto y un genoma muy pequeño, lo que lo hizo mucho más fácil de secuenciar y ensamblar.

Fue en el año 1996 con la formación de Arabidopsis Genome Initiative, cuando comenzó el proyecto que sería desarrollado por un consorcio de laboratorios internacionales y que utilizó un enfoque BAC-by-BAC con secuenciación de Sanger

para completar el genoma, lo que dio origen al mayor avance en este campo hasta la época, la Arabidopsis Genome Initiative 2000.

El genoma obtenido, con una extensión de 119 Mb se considera el estándar de oro para los genomas de plantas, ya que posee una muy alta calidad. Pese a ello cabe mencionar que el genoma de *A. thaliana* no es un genoma completo ya que existen brechas en la secuencia. "*La determinación estima un tamaño del genoma de 157 Mb*" (Bennett et al., 2003).

La secuenciación de especies distintas así como la resecuenciación de individuos de la misma especie, permiten una mejor comprensión de la función de los genes, además de una comparación de su estructura y expresión, ya que las diferentes especies comparten muchos genes además de funciones específicas como se ilustra a continuación (Figura 2.1). En la imagen podemos apreciar que 7914 genes de los 26819 encontrados en *A. Thaliana* en 2007, son comunes en otras tres especies vegetales secuenciadas (*Vitis vinifera*, *Oryza sativa* y *Populus trichocarpa*).

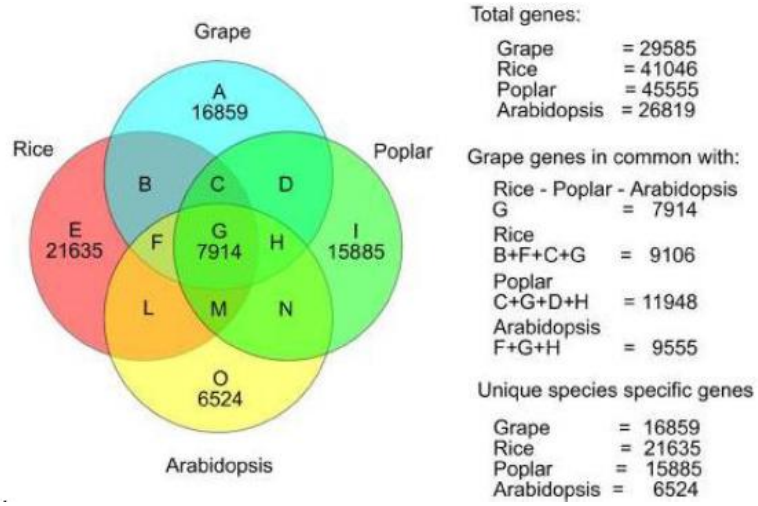

Figura 2.1 Comparación de genes de cuatro genomas de plantas con una longitud parecida de más del 60% de similitud. (Fuente: Velasco *et al*., 2007)

Basándose en el enfoque adoptado por la Arabidopsis Genome Initiative, el Proyecto Internacional de Secuenciación del Genoma del Arroz siguió sus pasos y secuenció el genoma del arroz (*Oryza sativa*) utilizando un enfoque BAC-by-BAC, produciendo la única otra secuencia del genoma vegetal con una gran calidad final. Al igual que con el genoma de *A. thaliana*, el genoma de *O. sativa* está casi completo, con espacios en la mayoría de los centrómeros y un número limitado de espacios en los brazos eucromáticos.

3. OBJETIVOS

Los objetivos que se persiguen con esta revisión son los de elaborar un modelo actualizado de secuenciación de novo de genomas de plantas, hablando de las plantas modelo como *A. thaliana*, de los problemas que se encuentran a la hora de llevar a cabo la secuenciación de una especie vegetal, de la importancia de ésta para entender el genoma, además de remarcar los principales avances en este campo.

En adición a esto, también se busca recoger la información de la secuenciación genómica en general. Comenzando sobre las diferentes generaciones de métodos y de tecnologías, recoger los avances en este campo y los posibles avances que podrían darse en un futuro.

Con estos conocimientos se pretende elaborar una revisión que sirva para entender más fácilmente esta tecnología, la importancia que tiene para los avances de la genética y las aplicaciones que se pueden dar especialmente en lo referido a las plantas.

4. TECNOLOGÍA IMPLICADA EN LA SECUENCIACIÓN DEL GENOMA

4.1. Métodos utilizados en el ensamblaje de novo

Cada método de ensamblaje se basa en la suposición de que fragmentos de ADN muy parecidos se originan de la misma posición dentro del genoma. De forma que la gran similitud entre secuencias de ADN es usada para conectar fragmentos individuales en secuencias más largas que se denominan contigs, las cuales se obtienen consensuadamente a partir del ensamblaje de las lecturas o “reads”.

Una vez tenemos estos fragmentos, se emplea el método de scaffolding, el cual consiste en orientar una serie no contigua de contigs y ordenarlos, sabiendo que hay presentes ciertos espacios entre ellos (gaps) de una longitud conocida. Las secuencias que se emplean son secuencias contiguas que corresponden a superposiciones de lectura. Todo este proceso se da de forma muy simplificada en la figura representada a continuación (Figura 4.1).

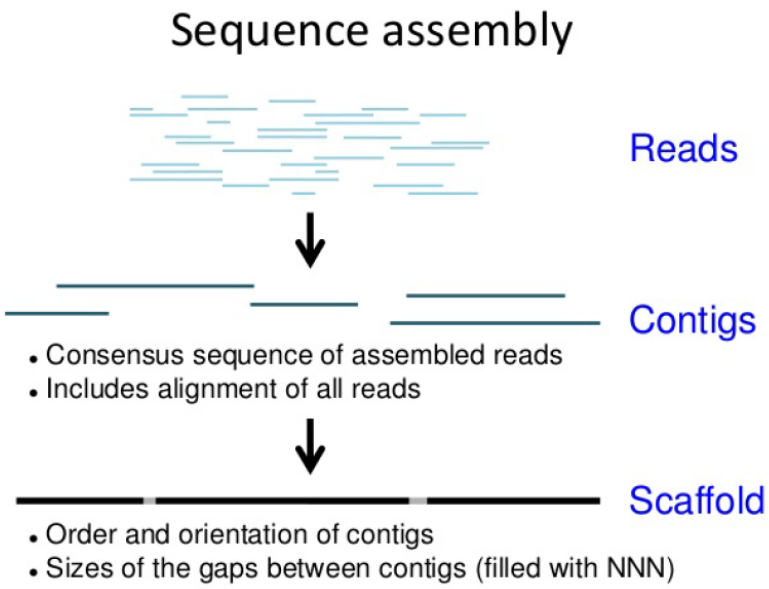

Figura 4.1 Ensamblaje de secuencias genómicas para dar origen a los Scaffolds

La tecnología NGS ha supuesto un gran avance para la biología actual, incluyendo el ensamblaje de genomas. En comparación con el tradicional método de Sanger, obtenemos un significativamente mayor número de datos así como unos costos menores, con lo que el rendimiento obtenido es mayor. Sin embargo, desde el punto de vista computacional representan un nuevo reto para el ensamblaje de novo, ya que la longitud de los fragmentos secuenciados es muy corta.

"*Los reads de corta longitud constituyen un problema cuando hay repeticiones en el genoma*" (Treangen & Salzberg, 2011). Las regiones repetitivas son fragmentos de ADN que aparecen más de una vez en el genoma. Cuando un read proviene de una región repetitiva y su longitud es menor a esta, no se sabe qué copia de la repetición se obtuvo. Debido a esto, durante el ensamblaje se pueden producir falsas uniones en el genoma en estas regiones repetitivas, lo cual hace que resulte difícil ensamblar todos los fragmentos para que esto se dé en un solo evento.

Para solucionar estos problemas debidos a los cortos fragmentos obtenidos por las tecnologías de NGS de segunda generación, "*se han desarrollado nuevos secuenciadores que incrementan la longitud de los mismos mientras mantienen el rendimiento*" (Schadt, Turner & Kasarskis, 2010).

Computacionalmente, también es posible extender los reads de NGS mediante la técnica de paired-ends. Para ello es necesario conocer la secuencia de ambos

extremos de un segmento de ADN, así como la distancia entre ellos. Cuando el tamaño del fragmento es menor que dos veces el del read, ambos extremos se solapan, por lo que se pueden unir para formar una secuencia de mayor tamaño.

Los programas ensambladores de secuencias utilizan tres algoritmos principales para llevar a cabo su trabajo: Greedy, Overlap-Layout-Consensus y gráficos de Bruijn.

El algoritmo de Greedy es el más sencillo e intuitivo. Siempre se conectan los reads con mayor afinidad iterativamente, siempre que no contradigan el ensamblaje previamente elaborado. Pese a esto, no es el método más utilizado, ya que se enfoca más en un ensamblaje local, sin emplear información global, por lo que no resuelve eficientemente regiones repetitivas de gran longitud. Usualmente es usado para el ensamblaje de datos obtenidos mediante secuenciación Sanger.

En el caso de Overlap-Layout-Consensus (OLC) funciona de una forma distinta, ya que primero localiza los pares de reads que se solapan y organiza dicha información en un gráfico en el cual hay un nodo y un conector por cada solapamiento. Este gráfico permite tener en cuenta la relación global entre los reads. Este método dominaba el mundo del ensamblaje hasta que aparecieron las tecnologías NGS.

El último algoritmo es el gráfico de Bruijn, el cual modela la relación entre subcadenas con una longitud de k (un número entero positivo) exactamente igual dentro de los reads. Este método guarda similitud con el anterior, ya que posee nodos que representan k-mers y conectores que se solapan con estos, relacionando ambas longitudes. En adición a esto, incorporan métodos de corrección de errores que mejoran la calidad del ensamblaje.

4.2. Las estrategias de secuenciación de Primera Generación

“Los métodos de Sanger y de Maxam-Gilbert fueron consideradas como las tecnologías que conformarían las estrategias de secuenciación de Primera Generación” (Liu *et al*., 2012), y que sentarían las bases de la secuenciación del ADN con su publicación en 1977.

Estos métodos se emplearían para la investigación biomédica, de genomas demás de presentar también aplicaciones clínicas.

4.2.1. Método de secuenciación de Sanger

Este método se conoce como método de terminación de cadena o secuenciación por método de síntesis. Su proceso se basa en utilizar una hebra de ADN bicatenario como plantilla para ser secuenciado. Esto se produce a partir de dNTP, los cuales son marcados para cada base de ADN por ddA, ddG, ddC y ddT. Estos son utilizados para la elongación de los nucleótidos, y una vez que son incorporados a la hebra de ADN evitan que este siga elongándose (Figura 4.2).

Los primeros genomas que se secuenciaron con el método de Sanger fueron ΦX174, con un tamaño de 5374 pb y el del bacteriófago lambda, con un tamaño de 48501 pb. En 1995 Applied Biosystems construyó una máquina de secuenciación automática llamada ABI Prism 370, la cual aprovechaba este método en combinación con la electroforesis por capilaridad, lo cual se convirtió en toda una revolución en cuanto a la velocidad de secuenciación.

La secuenciación de Sanger fue ampliamente utilizada durante tres décadas, e incluso hoy en día lo es para realizar la secuenciación de ADN simple o de bajo rendimiento, ya que era difícil mejorar aún más la velocidad de análisis que permitía la secuenciación de genomas complejos, como los genomas de especies vegetales.

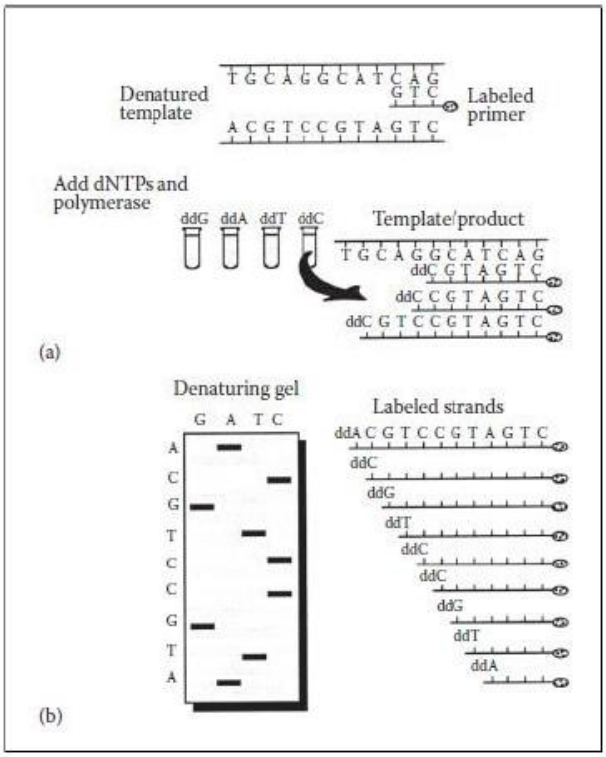

Figura 4.2 Método de secuenciación de Sanger

4.2.2. Método de secuenciación de Maxam-Gilbert

Es otro método de secuenciación perteneciente a la Primera Generación, y es conocido como el método de degradación químico. Se basa en la escisión de nucleótidos mediante químicos y es muy eficaz con polímeros de nucleótidos de pequeño tamaño. El tratamiento químico produce roturas en pequeña proporción de uno o dos nucleótidos de cada par de bases en cada reacción (G, C+T, C, G+A). Esto conduce a *"una serie de fragmentos marcados que pueden ser fácilmente separados según su tamaño mediante electroforesis*" (Maxam & Gilbert, 1977).

En el caso de este método, la secuenciación se da sin necesidad de realizar clonación. Este método sin embargo sería declarado obsoleto por culpa del desarrollo del método de Sanger, ya que era mucho más rápido y eficiente.

4.3. Las estrategias de secuenciación de Segunda Generación

Las estrategias que pertenecían a la Primera Generación dominaron la escena durante tres décadas, en especial el método de Sanger, sin embargo se percibía que su costo y tiempo empleados eran aún muy elevados, lo que llevaría a desarrollar nuevas técnicas. Desde 2005 se ha destacado la aparición de una nueva generación de técnicas de secuenciación que podían superar las limitaciones de los pertenecientes a la Primera Generación.

Las principales características de estas nuevas técnicas son la generación de millones de lecturas cortas en paralelo, el aumento de velocidad del proceso en comparación con la anterior generación, el bajo coste del proceso y la facilidad de detección del producto final sin necesidad de una electroforesis.

"La secuenciación de lecturas cortas se separó en dos enfoques principales: la secuenciación por ligación (SBL) y la secuenciación por síntesis (SBS)" (Goodwin *et al*., 2016). Además de esto, principalmente se clasificaron en tres plataformas de secuenciación: Roche/454 lanzada en 2005, Illumina/Solexa en 2006 y ABI/SOLiD en 2007.

4.3.1. Tecnología de secuenciación Roche/454

Esta tecnología comenzó a comercializarse en el 2005, y empleaba la técnica de pirosecuenciación, la cual se basaba en la detección de fotones, producidos por la enzima luciferasa al utilizar pirofosfatos, que se liberan tras cada incorporación de un nucleótido en la secuencia de ADN.

En este método las muestras de ADN se fragmentan aleatoriamente, entonces cada fragmento se une a un molde que lleva primers con oligonucleótidos y que es complementario a los fragmentos. Tras eso se aíslan y amplifican mediante una emulsión por PCR y se lleva a una placa denominada placa de picotiter (PTP), donde se lleva a cabo la técnica de pirosecuenciación (Figura 4.3).

"*Gracias a esta placa se pueden producir miles de reacciones en paralelo, lo cual aumenta en gran medida e rendimiento del método*" (Vezzi *et al*., 2012).

Entre las ventajas que ofrecía Roche/454 se encontraba su habilidad para generar reads de mediana longitud y la posibilidad de leer secuencias separadas por distancias conocidas (como por ejemplo 20 kb), lo cual es de gran ayuda a la hora de mapear un genoma de referencia. Los errores (inserciones y deleciones) eran detectados mediante la presencia de regiones homopolímeras, las cuales se identificaban gracias a la intensidad de la luz emitida por la pirosecuenciación.

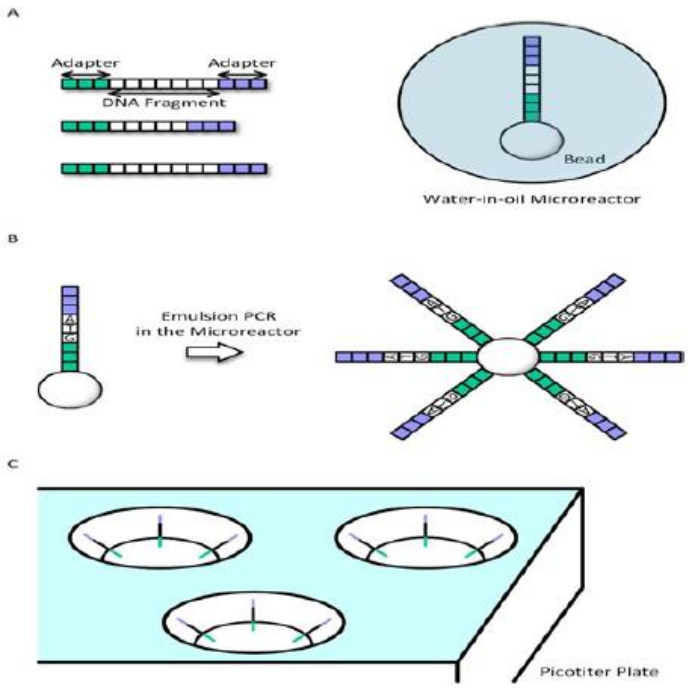

Figura 4.3 Método de secuenciación Roche/454

En 2013 este método dejó de ser competitivo, con lo cual la compañía que lo comercializaba dejó de hacerlo.

4.3.2. Tecnología de secuenciación Ion torrent

Esta tecnología empezó a ser comercializada por Life Technologies en 2010. Era similar a la pirosecuenciación de 454 pero no empleaba nucleótidos marcados con fluorescencia como otros métodos de segunda generación, ni la emisión de fotones, sino que "*se basaba en la detección del ion hidrógeno (protón) que era liberado durante la secuenciación*" (Rotheberg *et al*., 2011) y que producía un cambio de potencial eléctrico transitorio.

Ion torrent emplea un chip que contiene una serie de micropocillos cada uno con fragmentos idénticos. Cada vez que se incorpora un nucleótido se libera un ion hidrógeno, lo que cambia el pH de la solución. Esto es detectado por un sensor situado en el fondo del micropocillo, que lo convierte en una señal eléctrica proporcional al número de nucleótidos que se incorporan (Figura 4.4).

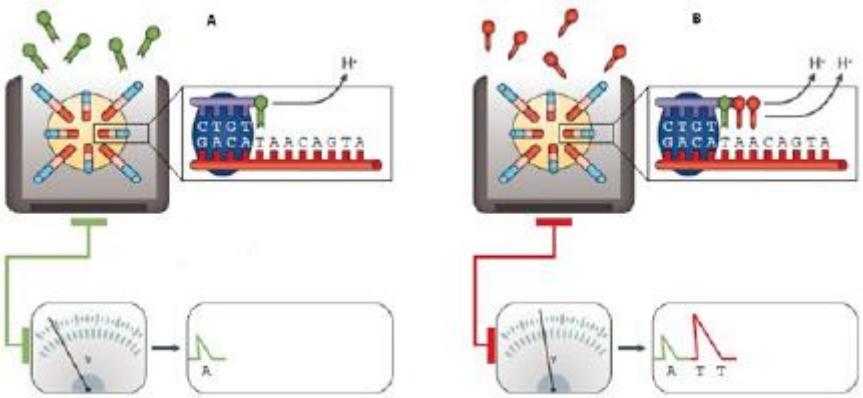

Figura 4.4 Método de secuenciación Ion torrent

Los secuenciadores de Ion torrent son capaces de crear reads que van desde 200 pb hasta 600 pb y que pueden alcanzar un rendimiento de 10 Gb para secuenciadores de iones protones. La principal ventaja de este método es su tiempo de secuenciación, situado entre 2 y 8 horas. La mayor desventaja del mismo es la dificultad para interpretar secuencias homopolímeras de más de 6 pb.

4.3.3. Tecnología de secuenciación Illumina/Solexa

La compañía Solexa, que más tarde sería comprada por Illumina, desarrolló un nuevo método de secuenciación que sería denominado como secuenciador Illumina/Solexa Genome Analyzer (GA). Esta tecnología se enfoca en la secuenciación por síntesis y es la técnica de Secuenciación de Nueva Generación más empleada actualmente.

En la primera fase el ADN es fragmentado en secuencias aleatoriamente y los adaptadores se unen a los extremos de cada secuencia. Tras eso se fijan a los adaptadores complementarios y se sitúan en una placa sólida. En la segunda fase se realiza la PCR sobre dichas secuencias. En la fase final se determinan los nucleótidos en cada secuencia mediante el enfoque de síntesis, el cual emplea terminadores reversibles en la que los cuatro nucleótidos modificados, los primers de secuenciación y la ADN polimerasa se añaden como un mix y los cebadores se hibridan con las secuencias. Cada tipo de nucleótido es marcado con un fluoróforo específico para ser identificado fácilmente. Tras eso los cluster se excitan mediante laser y emiten una señal de luz para cada nucleótido, la cual es detectada por una cámara y transmitida al ordenador que traducirá esto en una secuencia de nucleótidos (Figura 4.5). "*Para terminar el proceso se elimina el terminador y comienza un nuevo ciclo con nuevas incorporaciones*" (Reuter *et al*., 2015).

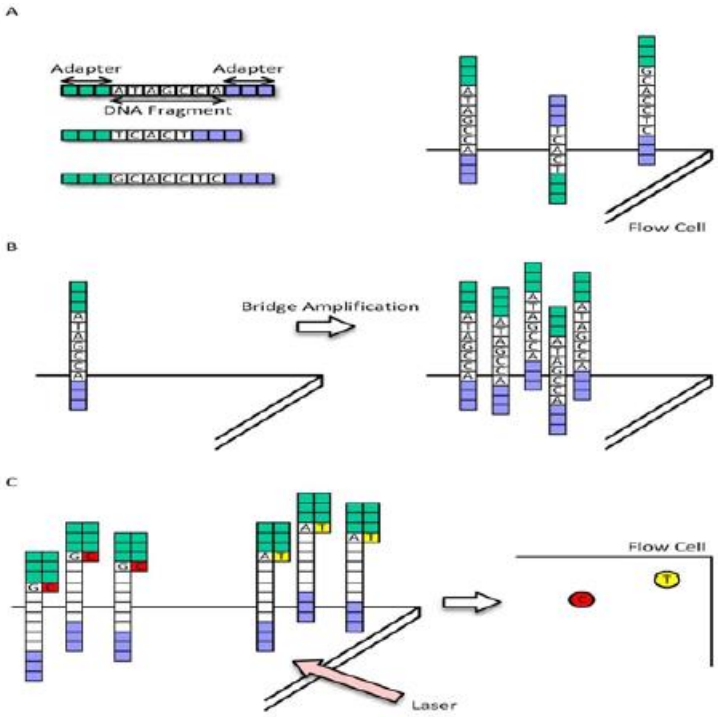

Figura 4.5 Método de secuenciación Illumina/Solexa

Una de las principales desventajas de la tecnología de Illumina, como sucede con la mayoría de estas tecnologías, es que necesita un alto control de carga, ya que una sobrecarga puede generar cluster superpuestos, lo cual se traduce en una calidad de secuenciación deficiente. No obstante, ha resultado ser la tecnología más eficiente y económica, por lo que se ha convertido en la más utilizada con gran diferencia con respecto a las otras tecnologías disponibles. Pese a esto la tasa de error de este método de secuenciación ronda el 1%, y se debe a la errónea interpretación de la incorporación de un nucleótido.

4.3.4. Tecnología de secuenciación ABI/SOLiD

La ligación y detección de oligonucleótidos compatibles (SOLiD) fue comercializada por Life Technologies. "*En 2007 Applied Biosystems (ABI) adquirió esta empresa y desarrollaron este método que emplea el enfoque de ligadura*" (Shendure *et al*., 2008).

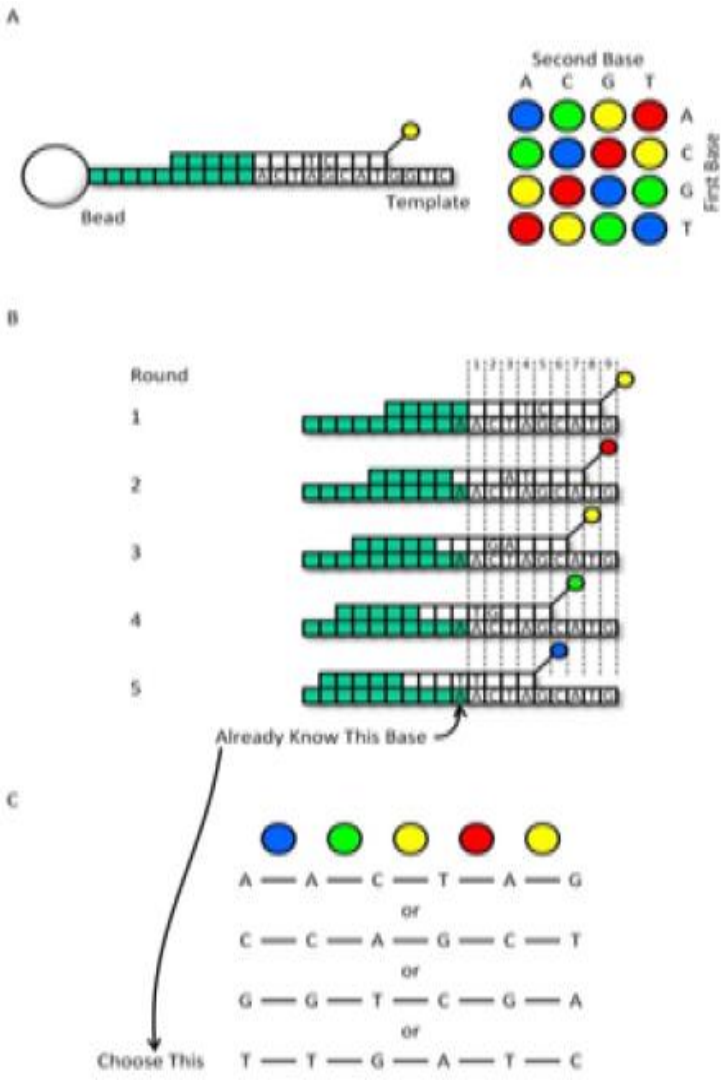

Figura 4.6 Método de secuenciación ABI/SOLiD

El funcionamiento de este método se basa en múltiples rondas de secuenciación. Comienza uniendo adaptadores a fragmentos de ADN y los clona mediante PCR. El formato de salida es el espacio de color que aparece al secuenciarse. Pueden obtenerse hasta 16 posibles combinaciones de dos bases. El ciclo es repetido hasta que se secuencia dos veces y los datos obtenidos se traducen en las secuencias del fragmento de ADN (Figura 4.6).

La mayor ventaja de este método es que cada base se lee dos veces, lo cual le confiere una alta precisión, mientras que su mayor inconveniente es que las lecturas son relativamente cortas y su tiempo de ejecución es largo. Los errores que pueden obtenerse con esta tecnología se deben a errores de sustitución de bases. Su gran fiabilidad le hizo inicialmente ideal para su uso en clínica, pero actualmente es una tecnología obsoleta y los equipos descatalogados ya que Illumina contrarresta la mayor fiabilidad de lectura de SOLiD con una enorme capacidad de secuenciación a muy bajo coste de Illumina. Al leerse muchas veces la misma base es fácil determinar cuál es la base correcta y eliminar las lecturas erróneas.

4.4. Las estrategias de secuenciación de Tercera Generación

Las tecnologías de la Segunda Generación revolucionaron el análisis de ADN y fueron más usadas en comparación con las pertenecientes a la Primera Generación. Sin embargo, las tecnologías de Segunda Generación requieren PCR, lo que limitaba mucho el tamaño de los reads obtenidos, ya que la PCR solo es capaz de amplificar de forma eficiente pequeños fragmentos de ADN. En adición a eso, los genomas muy complejos con áreas repetitivas eran incapaces de ser resueltos por estas tecnologías, lo cual llevó a los científicos a desarrollar nuevas técnicas denominadas secuenciación de Tercera Generación. Estas técnicas presentan un bajo costo de secuenciación y fácil preparación de la muestra sin necesidad de PCR, lo que se traduce en que se puedan obtener lecturas de gran tamaño, además de facilitar el problema del ensamblaje y las regiones repetitivas del genoma.

Las estrategias de Tercera Generación se pueden dividir en dos: la tecnología SMRT y la tecnología de nanoporos.

4.4.1. Tecnología de secuenciación SMRT de Pacific Biosciences

Pacific Biosciences desarrolló el primer secuenciador genómico que empleaba SMRT, la cual es la tecnología de Tercera Generación más utilizada actualmente.

Este método emplea los mismos métodos de fluorescencia que las tecnologías anteriores, pero en lugar de ejecutar ciclos de amplificación, detecta las señales en tiempo real una vez se producen las incorporaciones (Figura 4.7). Este método emplea estructuras compuestas por muchas nanoestructuras llamadas zeromode waveguides (ZMWs), las cuales emplean las propiedades de la luz cuando pasa a través de aberturas con un diámetro menor que su longitud de onda, por lo que ésta no puede propagarse. Cada ZMW posee ADN polimerasa y un fragmento de ADN diana para su secuenciación. Siempre que se incorpora un nucleótido se emite una señal luminosa que es registrada por los sensores y permite determinar la secuencia de ADN.

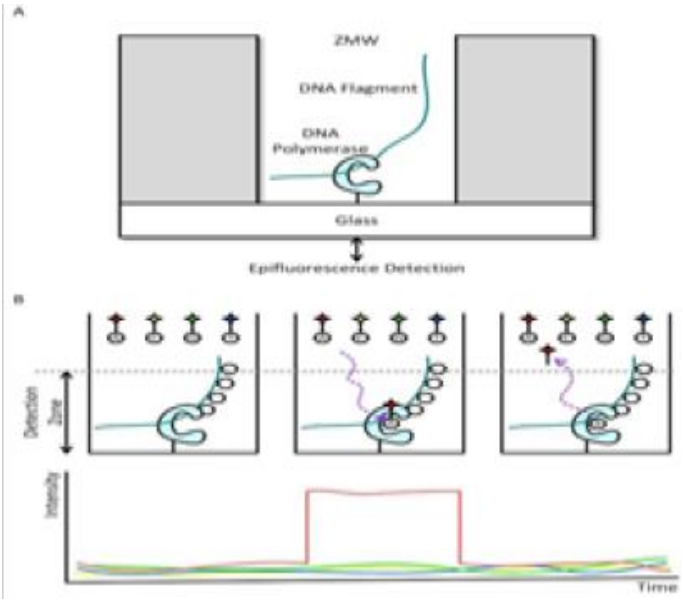

Figura 4.7 Tecnología de secuenciación Pacific Biosciences SMRT

Las principales ventajas de este método radican en parte en su rapidez de preparación de muestra, la cual solo lleva 4 o 6 horas en lugar de días, pero sobre todo, aumenta enormemente las longitudes de las lecturas, las cuales pueden llegar a los 60 kb, mucho más largo que cualquier tecnología perteneciente a la Segunda Generación.

4.4.2. Tecnología de secuenciación Oxford Nanopore

Esta tecnología fue desarrollada para determinar el orden de los nucleótidos en una secuencia de ADN. En 2014, Oxford Nanopore Technologies desarrolló el dispositivo MinION, el cual podía generar lecturas más largas y que presentaba la resolución de variantes genómicas estructurales y contenido repetitivo.

En este método, una hebra de ADN se une a una proteína que la dirige hacia un poro que tiene una diferencia de potencial. Al pasar la hebra por el poro, cada base produce una alteración diferente del potencial del poro y eso es lo que se usa para obtener la secuencia de bases (Figura 4.8). Además, esta tecnología permite que a continuación pase la otra hebra de ADN por el poro, lo cual sirve para confirmar la secuencia obtenida y detectar posibles errores en la secuenciación. Esto es registrado en un modelo gráfico para posteriormente ser interpretado e identificar la secuencia.

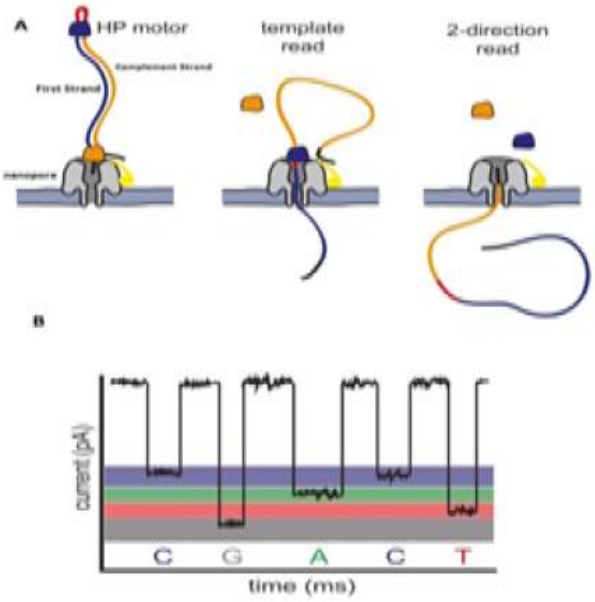

Figura 4.8 Tecnología de secuenciación MinION de Oxford Nanopore

Entre las ventajas que presenta este método se encuentran su pequeño tamaño y su reducido costo, además de proporcionar largas lecturas similares a las producidas por la tecnología de PacBio, lo cual ayuda a mejorar el ensamblaje de novo y la obtención de scaffolds de gran tamaño.

Recientemente también "*se ha desarrollado un nuevo instrumento denominado PromethION"* (Karow *et al*., 2014), el cual es un secuenciador de mesa de trabajo autónomo que posee 48 celdas de flujo individuales con 3000 poros cada una, lo que

equivaldría a unos 48 MinION, que funcionan a 500pb por segundo, lo cual es capaz de alcanzar un rendimiento enorme.

5. ¿QUÉ HACE ESPECIAL LA SECUENCIACIÓN DEL GENOMA VEGETAL?

Para secuenciar un genoma debemos tener en cuenta una serie de factores como su tamaño, las duplicaciones y el contenido que se repite a lo largo de éste. Los genomas vegetales suelen aparecer como grupos aislados de información entre una gran extensión de secuencias repetitivas de hasta el 80% del genoma total. Esta condición puede llegar a obstaculizar el ensamblaje de las secuencias, limitando severamente la productividad de las investigaciones genómicas.

5.1. Muestreo

A la hora de secuenciar genomas vegetales, el mayor problema se encuentra en la dificultad de obtener una gran cantidad de ADN de alta calidad a partir de material vegetal. Esto supone un gran inconveniente a la hora de elaborar bibliotecas genómicas. En el caso de la secuenciación de transcriptomas es aun más complicado, ya que para cualquier proyecto de secuenciación de un genoma es suficiente con muestras de una sola planta, pero en este caso "*es necesario emplear diferentes variedades de tejidos, lo que lleva a unigenes muy similares representando al mismo gen*" (Gonzalez-Ibeas *et al*., 2007). Cuando se elabora un transcriptoma heterogéneo empleando lecturas de longitud elevada, la presencia de alelos múltiples no obstaculiza el ensamblaje, pero cuando se emplean técnicas de secuenciación de nueva generación, los alelos y parálogos realmente perjudican el ensamblaje.

5.2. Tamaño y complejidad del genoma

Las plantas tienen una serie de necesidades que no podrían sustentarse de no ser por el surgimiento de nuevos genes basados en duplicaciones, superposición de genes, ploidías o cualquier estrategia similar, lo que hace que los genomas vegetales sean grandes y complejos. El tamaño general de plantas terrestres se encuentran alrededor de unas 6 Gb. Las tecnologías actuales de secuenciación son capaces de hacerse cargo de genomas complejos como los pinos, cuya longitud oscila entre las

22 y las 33 Gb; o el trigo, con una longitud de 16 Gb. Esto deja en evidencia que el tamaño de un genoma ya no es un problema en la actualidad.

Actualmente el problema radica en la complejidad de un genoma, ya que el tamaño del genoma no varía en la misma medida que el número de genes. La longitud de las regiones con una sola copia varía ampliamente entre las distintas especies de plantas.

El tamaño del genoma parece estar relacionado con el tipo de intercalación: especies vegetales con genomas pequeños (como *Arabidopsis*) tienen largas intercalaciones y por consecuencia una mayor longitud de secuencias. Por otro lado, "*las especies de plantas con genomas mayores como el trigo, el centeno o el maíz tienen intercalaciones de menor tamaño y debido a esto, secuencias no repetitivas de menor longitud*" (Lapitan, 1992). Esto confirma la razón de que los genomas pequeños son mas sencillos de ensamblar que los de mayor tamaño.

5.3. Elementos transponibles

A lo largo de la evolución, se ha producido una serie de grandes cambios en el tamaño, estructura y función del genoma entre especies, en buena medida dichos cambios se deben a los transposones.

Esto podría explicar misterios como por qué los cromosomas 1 y 2 de *A. thaliana* son una fusión de los cromosomas 1 y 2, y 3 y 4 respectivamente de *A. lyrata.* Las secuencias repetitivas más representadas en genomas vegetales están compuestas en su mayoría por elementos transponibles, debido a la naturaleza replicativa del proceso de retrotransposición, los transposones de Clase I pueden representar hasta el 90% de los transposones totales, mientras que los pertenecientes a la Clase II abundan menos. En plantas con genomas pequeños, los transposones son escasos, mientras que en plantas con genomas de mayor tamaño se encuentran alcanzando hasta el 85% del total, como en el caso del maíz. "*Debido a su abundancia y naturaleza repetitiva, los elementos transponibles son el principal dolor de cabeza a la hora de ensamblar el genoma ya que complican enormemente este proceso, particularmente cuando se utilizan tecnologías de lectura corta"* (Schatz et al., 2010).

5.4. Heterocigosidad

Excluyendo los vegetales que han sido domesticados en laboratorios, la mayoría de plantas alógamas son altamente heterocigotas. Dado que se presenta una especie de redundancia que siempre es un factor en contra en los ensamblajes, solo se pueden ensamblar regiones eucromáticas de los genomas, y un alto porcentaje de las lecturas proporcionadas por NGS permanecen sin ensamblar. Como consecuencia de esto, algunos proyectos de secuenciación vegetal tienden a centrarse en homocigotos obtenidos experimentalmente como por ejemplo los dobles haploides. Éstos se crean a partir de una planta haploide mediante el cultivo de anteras. Dicho haploide se trata con colchicina para inducir una endorreduplicación de los cromosomas para finalmente obtener un diploide homocigoto para todos sus genes. Estos estudios se llevan a cabo incluso en el caso de que no sean comerciales o importantes para el campo de la agricultura. Otro problema que surge debido a la heterocigosidad es la creación de segmentos duplicados falsos en el proceso de ensamblaje, que "*aparecen cuando las secuencias heterocigotas de dos haplotipos son ensamblados en contigs separados y scaffolds adyacentes entre sí en lugar de juntarse entre ellos*" (Kelley & Salzberg, 2010). Solo el uso de lecturas de mayor tamaño mejoraría la capacidad de ensamblar haplotipos separados dentro de un genoma.

5.5. Poliploidía

Es el resultado de la fusión de dos o más genomas en un mismo núcleo. Su origen es debido a la duplicación de un genoma completo (autopoliploidía) o a hibridaciones interespecíficas o intergenéricas seguido de duplicación de cromosomas (alopoliploidía).

La duplicación trae una serie de ventajas: "*permite la aparición de nuevos genes con funciones y fenotipos nuevos* y s*on capaces de adaptarse mejor debido a la plasticidad adquirida por el genoma*" (Comai, 2005). Ganan reproducción asexual y permiten un aumento del nivel de heterocigosidad pese a perder su autoincompatibilidad, lo cual explica la generalización de las poliploidías en el reino vegetal.

La poliploidía es sin duda alguna uno de los principales motores de evolución de las plantas, ya que ha proporcionado una extensa variedad de diversidad y ha permitido la aparición de nuevas especies. El resultado de la secuenciación de un genoma

vegetal depende de si la especie es autopoliploide, alopoliploide o de la antigüedad del evento de ploidización. La secuenciación de un poliploide reciente será más compleja según la divergencia de los genes duplicados. Por otra parte, la redundancia producida por la presencia de dos o más conjuntos de genes en un mismo núcleo puede afectar a la precisión del ensamblaje.

Para resolver el revés de la ploidía surgieron varios métodos diferentes que ayudarían a simplificar el problema. En el caso de la fresa de cultivo, que es alo-octaploide, se secuenció la especie diploide *Fragaria vesca.* Para el trigo hexaploide se elaboró una estrategia de separación por citometría de flujo y se construyó un mapa físico de clones BAC en mosaico.

5.6. Contenido de genes y familias

La información genética albergada en los genes de plantas puede ser muy compleja, lo cual es demostrado por la existencia de grandes familias de genes y "*una gran variedad de pseudogenes originados a partir de eventos de duplicación recientes y a la actividad de los transposones*" (Schnable *et al*., 2009). Más de la mitad de los genes vegetales pertenecen a las mismas familias, teniendo el 45% de ellos la misma función pero distintos patrones de expresión.

Un hallazgo interesante es la existencia de genes específicos de linaje en prácticamente todos los genomas eucariotas secuenciados. Estos genes podrían emplearse para realizar estudios funcionales, ya que pueden distinguir taxones relacionados, pero estos genes específicos de linaje podrían ser solo el resultado de ensamblajes incorrectos, por lo que deberían realizarse estudios al respecto.

Los estudios de movimiento de genes encontraron que muchas categorías de genes en Arabidopsis, papaya y uva transpusieron a una frecuencia basal del 5%. Lo más impactante frente este hecho es que algunas familias de genes presentaron frecuencias de movimiento de hasta el 90%. Este hecho no debería presentar ningún problema para los prodecimientos de ensamblaje, ya que ocurrieron hace mucho tiempo, pero el principal inconveniente es que la región que rodea al gen transpuesto presenta una alta composición de transposones reales, defectuosos y pseudogenes, lo que causa problemas a la hora de analizar las secuencias repetitivas.

5.7. ARNs no codificantes

"Este tipo de ARN (ncARN) se describió por primera vez en 1993 y presentó un cambio a la hora de entender la regulación genética en animales y vegetales" (Claros *et al*., 2012). Con la aparición de las NGS surgieron una gran cantidad de ncARN, como los de pequeña longitud (inferior a 30pb) y los de gran longitud (de 200pb o superiores). Estas secuencias a menudo participan en una serie de transcripciones codificantes y no codificantes y en sentido y antisentido, lo cual aumenta aún más la diversidad. Cuando empleas NGS con lecturas cortas, la complejidad que presentan no permite el ensamblaje debido a estas secuencias de ARN no codificantes y su carácter repetitivo. Solo las estrategias que empleen lecturas largas y puedan cubrir la longitud de estas secuencias de ARN no codificantes pueden llevar a cabo esta tarea.

5.8. Secuencias repetitivas ampliamente extendidas

"Las secuencias repetitivas siempre son un problema para realizar los ensamblajes de los genomas" (Gore *et al.,* 2009). Las principales fuentes de repetición son:

Las repeticiones entre cromosomas, estas duplicaciones se producen tanto dentro de los cromosomas (se encuentran duplicaciones dentro de un cromosoma concreto de una especie) como entre cromosomas (se encuentran regiones comunes entre dos cromosomas distintos de una misma especie).

Las unidades de ADNr (ADN ribosómico). Estas ADNr contienen los genes de los ARNr (ARN ribosómicos) y están compuestas por cientos de copias, las cuales pueden llegar a representar hasta el 10% del genoma y que no han sido bien resueltas aún por ninguna tecnología de secuenciación.

Las secuencias satélites, que presentan un sinfín de copias idénticas o casi idénticas de una sola unidad y que abundan en los centrómeros y la heterocromatina.

Los microsatélites o SSR, los cuales son series de pequeñas repeticiones en tándem, que en algunos casos pueden encontrarse repetidos hasta cientos de veces, aunque normalmente presentan repeticiones no muy elevadas pero muy variables entre diferentes individuos.

Las secuencias teloméricas que están formadas por breves repeticiones de secuencias similares a TTTAGGG a lo largo de cientos de unidades en los extremos de cada brazo cromosómico y cuyo tamaño es variable, por lo que no se considera relevante determinar su tamaño.

6. GENOMAS VEGETALES DE ESPECIAL RELEVANCIA

Con la publicación de la primera secuencia genómica vegetal perteneciente a *A. thaliana* la genética comenzó a profundizar en este campo, allanando el camino para la secuenciación de otros genomas vegetales. Con el avance en otros campos también han cambiado los métodos y herramientas para la investigación de plantas y la mejora de cultivos como Arabidopsis, y más tarde *Oryza sativa* (arroz), *Carica papaya* (papaya) y *Zea mays* (maíz) se secuenciaron utilizando el método clásico de Sanger. La llegada de las tecnologías de secuenciación de próxima generación (NGS) ha permitido el rápido avance de recursos genómicos para especies de plantas no modelo o huérfanas. Sin embargo, solo Arabidopsis y el arroz han sido secuenciados totalmente a día de hoy, siendo el resto borradores en mayor o menor grado de finalización.

Desafortunadamente, incluso los genomas completos contienen lagunas en sus secuencias, las cuales pertenecen a secuencias altamente repetitivas, que complican los métodos de secuenciación y ensamblaje. En la siguiente tabla se puede apreciar una lista con las principales especies vegetales cuyos genomas han sido secuenciados a día de hoy (Tabla 6.1).

Tabla 6.1 Lista de especies vegetales con mayor relevancia cuyo genoma ha sido secuenciado

Especie vegetal	Familia	Enfoque	Referencias
Arabidopsis thaliana	Brassicaceae	BAC-by-BAC	Arabidopsis Genome Initiative (2000)
Oryza sativa (indica)	Poaceae	WGS	Yu *et al*. (2002)
*Oryza sativa (*japonica)	Poaceae	WGS y BAC-by-BAC	Goff *et al.* (2002); International Rice Genome Sequencing Project (2005)
Populus trichocarpa	Salicaceae	WGS	Tuskan *et al.* (2006)
Vitis vinifera	Vitaceae	WGS	Jaillon *et al.* (2007)

Carica papaya	Caricaceae	WGS	Ming *et al.* (2008)
Physcomitrella patens	Funariaceae	WGS	Rensing *et al.* (2008)
Sorghum bicolor	Poaceae	WGS	Paterson *et al.* (2009)
Zea mays	Poaceae	BAC-by-BAC	Schnable *et al.* (2009)
Bracypodium distachyon	Poaceae	WGS	International Brachypodium Initiative (2010)
Glycine max	Fabaceae	WGS	Schmutz *et al.* (2010)
Glycine soja	Fabaceae	WGS	Kim *et al.* (2010)
Malus × domestica	Rosaceae	WGS	Velasco *et al.* (2010)
Ricinus communis	Euphorbiaceae	WGS	Chan *et al.* (2010)
Arabidopsis lyrata	Brassicaceae	WGS	Hu *et al.* (2011)
Brassica rapa	Brassicaceae	WGS	*Brassica rapa* Genome Sequencing Project Consortium, (2011)
Cajanus cajan	Fabaceae	WGS	Varshney *et al.* (2011)
Cucumis sativus	Cucurbitaceae	WGS	Huang *et al.* (2009a,b) and Woycicki *et al.* (2011)
Fragaria vesca	Rosaceae	WGS	Shulaev *et al.* (2011)
Medicago truncatula	Fabaceae	WGS y BAC-by-BAC	Young *et al.* (2011)
Phoenix dactylifera	Arecaceae	WGS	Al-Dous *et al.* (2011)
Selaginella moellendorffii	Selaginellaceae	WGS	Banks *et al.* (2011)
Solanum tuberosum	Solanaceae	WGS	Potato Genome Sequencing Consortium, (2011)
Thellungiella parvula	Brassicaceae	WGS	Dassanayake *et al.* (2011)
Theobroma cacao	Malvaceae	WGS	Argout *et al.* (2011)
Olea europaea	Oleaceae	WGS	Cruz et al. (2016)
Rubus occidentalis	Rosaceae	WGS	VanBuren *et al*. (2018)
Solanum tuberosum	Solanaceae	WGS	Pham *et al*. (2020)

6.1. *Arabidopsis thaliana*

Esta planta herbácea perteneciente a la familia de las brasicáceas es una especie típica de Europa, Asia y el noroeste de África, además es empleada como un sistema modelo para identificar genes y determinar sus funciones. Las regiones secuenciadas abarcan 115,4 megabases del genoma de las 125 megabases totales y además se extiende a regiones centroméricas. En el interior de este genoma se encuentran 25498 genes que codifican proteínas de 11000 familias, similar a la diversidad funcional de

otras especies como Drosophila, otro modelo multicelular secuenciado en eucariotas. Arabidopsis presenta muchas familias de proteínas nuevas, pero también carece de otras tantas familias de proteínas comunes.

Esta especie vegetal fue la primera en ser secuenciada completamente y sentó las bases para una comparación en mayor profundidad de todos los eucariotas, identificando una amplia gama de funciones génicas específicas de las plantas y estableciendo rápidas formas para identificar genes para poder llevar a cabo la mejora de cultivos.

6.2. *Triticum aestivum* (trigo)

Esta variedad de cereal pertenece al género Triticum, y es la más extendida a lo largo del mundo, abarcando hasta un 95% de la producción. Una de las características que hacen especial a esta planta es que es un género hexaploide que posee 42 copias de cromosomas, es decir cada uno posee 6 copias.

Con un tamaño de más de quince mil millones de bases, posee uno de los genomas más complejos. Su secuenciación se intentó repetidamente, pero los resultados que se obtenían tenían un rendimiento por debajo del esperado. Mediante el uso de las tecnologías de Illumina y PacBio se logró secuenciar su genoma prácticamente en su totalidad, con una longitud de 15 Gb. "*El ensamblaje final contiene 15344693583 bases y tiene un tamaño de contig promedio ponderado (N50) de 232 659 bases*" (Zimin *et al*., 2017).

Con la secuenciación de este genoma, se pueden obtener aplicaciones clave para avanzar en futuros estudios relacionados con el campo de la alimentación o la agricultura, lo cual resulta altamente beneficioso.

6.3. *Oryza sativa* (arroz)

Esta especie pertenece a la familia de las Poáceas y posee una semilla comestible y muy ampliamente extendida en la dieta de la población mundial.

La investigación de esta especie comenzó en el año 2005 con el International Rice Genome Sequencing Project (IRGSP). Gracias a este proyecto, se obtuvo fácilmente acceso a la secuencia del genoma de referencia de alta calidad, lo que permitió un

desarrollo en el campo de la investigación. *"El ensamblaje del genoma de Nipponbare se actualizó revisando y validando el minimal tilling path de clones con el mapa óptico para el arroz*" (Kawahara et al., 2013).

Las especies domesticadas se han dejado influir por las poblaciones de los ancestros predomesticados, así como por el sistema de cría y la complejidad de las prácticas de cría ejercidas por los humanos. Dentro de este género, existe una divergencia antigua y bien establecida entre las dos principales subespecies, indica y japónica, pero la historia de su cultivo sugiere niveles más finos de estructura genética.

6.4. *Solanum lycopersicum* (tomate)

El tomate es una variedad perteneciente a la familia Solanaceae, cuyo origen se sitúa en Centroamérica, y el norte de Sudamérica. Además de su importancia como cultivo, también se emplea como sistema modelo para el desarrollo de frutos.

"El genoma de la variedad de tomate Heinz 1706 fue secuenciado y ensamblado empleando una mezcla del método de Sanger y NGS. El genoma que se obtuvo tenía un tamaño de 900 Mb, de las que 760 Mb fueron ensambladas en 91 scaffolds, y la mayoría de los gaps pertenecían a la región pericentromérica". ("The tomato genome sequence provides insights into fleshy fruit evolution", 2021)

Los frutos carnosos son una estrategia importante en el ámbito de la reproducción vegetal, ya que al atraer a los frugívoros vertebrados para la dispersión de semillas, ésta aumenta su probabilidad de dispersión y gracias a ello de éxito. Mediante investigaciones, se ha demostrado que las triplicaciones presentes en el genoma del tomate añadieron nuevos miembros a familias de genes, los cuales llevan a cabo importantes funciones específicas de los frutos. "*Entre estos grupos destacan factores de transcripción y enzimas necesarias para la biosíntesis del etileno, así como fotorreceptores de luz roja que influyen en la calidad de la fruta*". ("The tomato genome sequence provides insights into fleshy fruit evolution", 2021)

6.5. *Olea europaea* (olivo)

El olivo es un árbol de hoja perenne muy longevo y que puede llegar a alcanzar hasta los 10 metros de altura. Pertenece a la familia Oleaceae y es uno de los principales motores económicos de la región gracias al aceite de oliva, el cual goza de unas

propiedades nutricionales tan únicas como beneficiosas. Su domesticación en España data de la época del neolítico hace unos 7500 años.

Mediante el estudio y análisis de la variedad “Picual”, se encontró que su genoma está compuesto por 79667 genes modelo, de los cuales 78079 codifican para proteínas y el resto son ARNt y otros ARN de pequeño tamaño. “*Además se encontró que la gran variabilidad genética presente en el olivo se debe a los elementos transponibles, los cuales se encuentran implicados en rasgos como la reproducción, la fotosíntesis o la producción de aceite”* (Jiménez-Ruiz *et al*., 2020).

El genoma de la variedad “Picual” posee un tamaño mayor al genoma de las variedades salvajes. La domesticación produjo un efecto conocido como cuello de botella, el cual se traduce en una escasa variabilidad genética debido a un drástico descenso de población en el pasado, pero pese a este hecho, el olivo ha recuperado una amplia diversidad genética. “*Los elementos transponibles han tenido un papel fundamental en la variabilidad genética actual, ya que durante el proceso de domesticación, muchos grupos de elementos transponibles se expandieron”* (Jiménez-Ruiz *et al*., 2020). Todos estos hechos hacen al olivo una especie única, surgida de los cambios que se han ido produciendo en su genoma a lo largo del proceso de domesticación hasta la actualidad.

7. CONCLUSIONES

Desde que surgiera hace aproximadamente medio siglo el primer método de secuenciación, la tecnología ha continuado desarrollándose y avanzando, especialmente tras el surgimiento de la tecnología de Secuenciación de Nueva Generación (NGS), la cual apareció en el año 2005. Gracias a su elevado rendimiento y a su velocidad de ejecución, permitían producir millones de lecturas a un coste mínimo, lo cual permitió un gran avance en el campo de la investigación y el análisis de las secuencias genómicas marcando un punto de inflexión.

A lo largo de este trabajo se ha presentado una descripción de los métodos y las tecnologías de secuenciación así como sus procedimientos y evolución. Además se han remarcado los principales avances y los resultados más importantes,

profundizando en el campo vegetal, el cual ha sido de vital importancia para su desarrollo, ya que sin los modelos simplificados o la complejidad de dichos genomas no habría sido posible darse cuenta de las limitaciones de cada tecnología y a partir de ahí haberla mejorado o haber cambiado el enfoque para marcar un cambio.

De cara al futuro, todavía hay aspectos a mejorar en estas tecnologías de NGS, entre los que destacan la dificultad de almacenamiento y de análisis de los datos que se obtienen a partir de estos métodos. En los próximos años surgirán nuevas tecnologías que aumenten aún más el rendimiento o disminuyan el coste de estos procedimientos, además de producir lecturas de mayor longitud.

8. BIBLIOGRAFÍA

ALTSCHUL, S.F, MADDEN, T.L., SCHÄFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W., & LIPMAN D.J. (1997). “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. Nucleic Acids Res.

BAUSHER, M.G., SINGH, N.D., MOZORU, J., LEE, S-B., JANSEN, R.K., & DANIELL, H. (2006). “The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var. ‘Ridge Pineapple’: organization and phylogenetic relationships to other angiosperms”. BMC Plant Biol. 6: 21.

CLAROS, M., BAUTISTA, R., GUERRERO-FERNÁNDEZ, D., BENZERKI, H., SEOANE, P., & FERNÁNDEZ-POZO, N. (2012). “Why Assembling Plant Genome Sequences Is So Challenging”. Biology, 1(2), 439-459.

DE NECOCHEA R & CANUL JC (2004). “Métodos fisicoquímicos en biotecnología: Secuenciación de ácidos nucleicos”. Instituto de biotecnología-UNAM.

DO, H.D.K., KIM, J.S., & KIM, J.H. (2013). “Comparative genomics of four liliales families inferred from the complete chloroplast genome sequence of Veratrum patulum O. Loes. (Melanthiaceae)”.

FELSENSTEIN, J. (1985). “Confidence limits on phylogenies: An approach using the bootstrap”. Evolution 39: 783-791.

FUNK, H., BERG, S., KRUPINSKA, K., MAIER, U., & KRAUSE, K. (2007). “Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, Cuscuta reflexa and Cuscuta gronovii”.

GOODWIN S, MCPHERSON JD, RICHARD MCCOMBIE W (2016) “Coming of age: Ten years of next-generation sequencing technologies”. Nature Reviews Genetics 17: 333–351.

GRAY, M.W. (1999). “Evolution of organellar genomes”. Curr. Opin. Genet. Dev. 9: 678-687

GUO, X.Y., RUAN, S.L., HU, W.M., CA, D.G., & FAN, L.J. (2008). “Chloroplast DNA insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved”.

HOWE, C.J., BARBROOK, A.C., KOUMANDOU, V.L., NISBET, R.E.R., SYMINGTON, H.A., & WIGHTMAN, T.F. (2003). “Evolution of the chloroplast genome”.

HUANG, Y.Y., MATZKE, A.J., & MATZKE, M. (2013). “Complete sequence and comparative analysis of the chloroplast genome of coconut palm” (Cocos nucifera).

HUTCHISON, C. (2007). “DNA sequencing: bench to bedside and beyond. Nucleic Acids Research”, 35(18), 6227-6237.

JIMÉNEZ-RUIZ, J., RAMÍREZ-TEJERO, J., FERNÁNDEZ-POZO, N., LEYVA-PÉREZ, M., YAN, H., & ROSA, R. et al. (2020). “Transposon activation is a major driver in the genome evolution of cultivated olive trees (Olea europaea L.)”. The Plant Genome, 13(1).

JO, M.H., HAM, I.K., MOE, K., KWON, S.-W., LU, F.-H., PARK, Y.-J., KIM, W.S., KIM, M.K., KIM, T., & LEE, E.M. (2012). “Classification of genetic variation in garlic (Allium sativum L.) using SSR markers”. Australian J. Crop Sci. 6: 625-631.

KAROW J (2014) “Oxford Nanopore presents details on new highthroughput sequencer, improvements to MinIon”.

KAWAHARA, Y., DE LA BASTIDE, M., HAMILTON, J., KANAMORI, H., MCCOMBIE, W., & OUYANG, S. et al. (2021). “Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data”.

KCHOUK, M., GIBRAT, J., & ELLOUMI, M. (2017). “Generations of Sequencing Technologies: From First to Next Generation”. Biology And Medicine, 09(03).

KIM, S., KIM, M.S., KIM, Y.M., YEOM, S.I., CHEONG, K., KIM, K.T., JEON, J., KIM, S., KIM, D.S., SOHN, S.H., LEE, Y.H., & CHOI, D. (2015a). “Integrative structural annotation of de novo RNA-Seq provides an accurate reference gene set of the enormous genome of the onion (Allium cepa L.)”. DNA Res. 22: 19-27.

KOONIN, E.V. (2001). “Computational genomics”. Curr. Biol. 11: R155-158.

LANGMEAD, B., TRAPNELL, C., POP, M., & SALZBERG, S.L. (2009). “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. Genome Biol.

LANGMEAD, B., & SALZBERG, S.L. (2012). “Fast gapped-read alignment with Bowtie 2”. Nat. Methods

LASLETT, D., y CANBACK, B. (2004). “ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences”. Nucl. Acids Res. 32: 11-16.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R., & 1000 GENOME PROJECT DATA PROCESSING SUBGROUP (2009). “The Sequence Alignment/Map format and SAMtools”. Bioinformatics 25: 2078-2079.

LI, S., WANG, J., DONG, R., ZHU, H., LAN, L., & ZHANG, Y. et al. (2020). “Chromosome-level genome assembly, annotation and evolutionary analysis of the ornamental plant Asparagus setaceus”. Horticulture Research, 7(1).

LIU, L. (2012). Table 1 | “Comparison of Next-Generation Sequencing Systems”. https://www.hindawi.com/journals/bmri/2012/251364/tab1/

LÓPEZ DE HEREDIA, U. (2016). “Las técnicas de secuenciación masiva en el estudio de la diversidad biológica”. Munibe Ciencias Naturales, 64.

MAIER, R.M., NECKERMANN, K., IGLOI, G.L., & KÖSSEL, H. (1995). “Complete sequence of maize chloroplast genome: Gene content, hotspots of divergence and the tuning of genetic information by transcript editing”. J. Mol. Biol. 251: 614-628.

MARGULIS, L. (1975). “Symbiotic theory of the origin of eukaryotic organelles”

MAXAM, A., & GILBERT, W. (1977). “A new method for sequencing DNA”. Proceedings Of The National Academy Of Sciences, 74(2), 560-564.

MICHAEL, T., & VANBUREN, R. (2020). “Building near-complete plant genomes. Current Opinion In Plant Biology”, 54, 26-33.

NATURE, (2000). “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. 408(6814), pp.796-815.

NATURE, (2012). “The tomato genome sequence provides insights into fleshy fruit evolution”. 485(7400), pp.635-641.

NOTREDAME, C., HIGGINS, D.G., & HERINGA, J. (2000). “T-Coffee: A novel method for fast and accurate multiple sequence alignment”. J. Mol. Biol. 302: 205-17.

PALMER, J.D. (1990). “Contrasting modes and tempos of genome evolution in land plant organelles”. Trends Genet. 6: 115-120.

PAREEK, C., SMOCZYNSKI, R., & TRETYN, A. (2011). “Sequencing technologies and genome sequencing”. Journal Of Applied Genetics, 52(4), 413-435.

PIRES, J.C., MAUREIRA, I.J., GIVNISH, T.J., SYTSMA, K.J., SEBERG, O., PETERSEN, G., DAVIS, J.I., STEVENSON, D.W., RUDALL, P.J., FAY, M.F., & CHASE, M.W. (2006). “Phylogeny, genome size, and chromosome evolution of the Asparagales”. Aliso 22: 287-304.

REUTER JA, SPACEK DV, SNYDER MP (2015) “High-throughput sequencing technologies”. Molecular Cell 58: 586-597.

RODRÍGUEZ-SANTIAGO, B., & ARMENGOL, L. (2012). “Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal”. Diagnóstico Prenatal, 23(2), 56-66.

ROTHEBERG JM, HINZ W, REARRICK TM, SCHULTZ J, MILESKI W, et al. (2011) “An integrated semiconductor device enabling non-optical genome sequencing”. Nature 475: 348-52.

ROUZE, P., PAVY, N., & ROMBAUTS, S. (1999). “Genome annotation: which tools do we have for it?” Curr. Opin. Plant Biol. 2: 90-95.

SCHADT, E., TURNER, S., & KASARSKIS, A. (2010). “A window into third generation sequencing. Human Molecular Genetics”, 20(4), 853-853.

SHENDURE J, JI H (2008) “Next-generation DNA sequencing”. Nature Biotechnology 26: 135–1145.

SHEPPARD, A.E., & TIMMIS, J.N. (2009) “Instability of plastid DNA in the nuclear genome”. PloS Genet. 55: e1000323.

SUN, X., ZHOU, S., MENG, F., & LIU, S. (2012). “De novo assembly and characterization of the garlic (Allium sativum) bud transcriptome by Illumina sequencing”. Plant Cell Reports 31: 1823-1828.

TANKSLEY, S.D., & PICHERSKY, E. (1988). “Organization and evolution of sequences in the plant nuclear genome”. Gottlieb L.D., Jain S.K., editors. Plant Evolutionary Biology. Chapman and Hall. p: 55-83.

TATUSOVA, T.A., & MADDEN, T.L. (1999). “BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences”. FEMS Microbiol. Lett. 174: 247-250.

TREANGEN, T., & SALZBERG, S. (2011). “Repetitive DNA and next-generation sequencing: computational challenges and solutions”. Nature Reviews Genetics, 13(1), 36-46.

VEZZI F (2012) “Next generation sequencing revolution challenges: Search, assemble, and validate genomes”. Ph.D, Universita degli Studi di Udine, Italy.

WALKOWIAK, S., GAO, L., MONAT, C., HABERER, G., KASSA, M., & BRINTON, J. et al. (2021). “Multiple wheat genomes reveal global variation in modern breeding”.

WEISSENBACH, J. (2016). “The rise of genomics”. Comptes Rendus Biologies, 339(7-8), 231-239.

ZERBINO, D.R., & BIRNEY, E. (2008). “Velvet: algorithms for de novo short read assembly using de Bruijn graphs”. Genome Res. 18: 821-829.

ZIMIN, A., PUIU, D., HALL, R., KINGAN, S., CLAVIJO, B. & SALZBERG, S., (2021). “The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum”.