Principios básicos de la genómica y sus aplicaciones

Fundamentals of genomics and its applications

Jessica Pérez Alquicira*
Universidad de Guadalajara, Laboratorio Nacional de Identificación y Caracterización Vegetal (LaniVeg), Centro Universitario de Ciencias Biológicas y Agropecuarias, Zapopan, Jalisco, Mexico.
CONACyT, Ciudad de México, México.
Flor Rodríguez-Gómez
Universidad de Guadalajara, Laboratorio de Análisis de la Biodiversidad y Genómica. Departamento de Bioingeniería Traslacional. Centro Universitario de Ciencias Exactas e Ingenierías.
Ofelia Vargas-Ponce
Universidad de Guadalajara, Laboratorio Nacional de Identificación y Caracterización Vegetal (LaniVeg), Centro Universitario de Ciencias Biológicas y Agropecuarias, Zapopan, Jalisco, Mexico.
Zayra Arery Guadalupe Muñoz-González
Universidad de Guadalajara, Laboratorio de Análisis de la Biodiversidad y Genómica. Departamento de Bioingeniería Traslacional. Centro Universitario de Ciencias Exactas e Ingenierías.
Universidad de Guadalajara, Doctorado en Biosistemática, Ecología y Manejo de Recursos Naturales y Agrícolas. Centro Universitario de Ciencias Biológicas y Agropecuarias, Zapopan, Jalisco, México.
Gabriela Alcalá-Gómez
Universidad de Guadalajara, Laboratorio Nacional de Identificación y Caracterización Vegetal (LaniVeg), Centro Universitario de Ciencias Biológicas y Agropecuarias, Zapopan, Jalisco, Mexico.
Universidad de Guadalajara, Doctorado en Biosistemática, Ecología y Manejo de Recursos Naturales y Agrícolas. Centro Universitario de Ciencias Biológicas y Agropecuarias, Zapopan, Jalisco, México.
Pilar Zamora-Tavares
Universidad de Guadalajara, Laboratorio Nacional de Identificación y Caracterización Vegetal (LaniVeg), Centro Universitario de Ciencias Biológicas y Agropecuarias, Zapopan, Jalisco, Mexico.
*Autor para correspondencia: perezalquicira@gmail.com

Resumen

La genómica es una disciplina que estudia la estructura, función y evolución de los genomas y aborda los procedimientos metodológicos utilizados para secuenciar y ensamblar el genoma. El análisis de los datos de secuenciación requiere de recursos computacionales de alta gama y de algoritmos matemáticos y software especializado, que se engloba en lo que se conoce como Bioinformática. La historia de la secuenciación de genomas puede dividirse en tres etapas: 1) primera generación que se basa en la secuenciación de un solo fragmento por medio de electroforesis capilar 2) la segunda generación se caracteriza por la paralelización masiva de las reacciones de secuenciación, lo que resultó en un incremento en la cantidad de fragmentos de ADN secuenciados con una longitud de 50 a 300 pb, 3) la tercera generación incluye también la secuenciación masiva de fragmentos, pero de una longitud mucho mayor (> 10000 pb) lo que facilita el ensamblaje de genomas. El avance en la secuenciación masiva ha permitido obtener una gran cantidad de genomas, que a su vez, tienen una amplia aplicación en la medicina, en el mejoramiento de especies animales y vegetales de importancia económica, así como en estudios comparativos, filogenéticos, entre muchos otros. Entre las principales ramas de la genómica, se encuentra la metagenómica, que ha tenido gran relevancia en el conocimiento de la composición y diversidad de microorganismos en muestras ambientales de agua, suelo, aire, entre otras, facilitando el reconocimiento de nuevos taxa. Otra de las ramas importantes de la genómica es la filogenómica, que es utilizada en el estudio de las relaciones evolutivas. Por tanto, los avances en las plataformas de secuenciación así como los avances en el área de la bioinformática han resultado en una revolución del conocimiento de la complejidad de los genomas.

Palabras clave: Bioinformática, ensamblaje de genomas, librerías genómicas, metagenómica, plataformas de secuenciación, SNPs (single nucleotide polymorphism).

Abstract

Genomics is the discipline that studies the structure, function, and evolution of genomes and addresses the methodological processes used to sequence and assemble the genome. Analyzing sequencing data requires state-of-the-art computational resources and specialized mathematical algorithms and software, which are together known as bioinformatics. The history of genome sequencing can be divided into three stages: 1) first-generation, which is based on the sequencing of a single fragment using capillary electrophoresis; 2) second generation, characterized by the mass parallelization of sequencing reactions, resulting in an increase in the amount of DNA fragments sequenced with a length of 50 to 300 base pairs; and 3) third generation, which also includes the mass sequencing, but of much longer fragments (> 10000 base pairs), which facilitates genome assembly. Advances in massive sequencing have allowed for the sequencing of a large number of genomes, which has had broad applications in medicine, the improvement of economically important plant and animal species, and phylogenetic studies, among many others. One of the main branches of genomics is metagenomics, which has been highly important in generating knowledge of the composition and diversity of microorganisms in environmental samples of water, air, and other materials, facilitating the recognition of new taxa. Another branch of genomics is phylogenomics, which is used to infer the evolutionary relationships among species. Therefore, the advances in sequencing platforms as well as advances in the area of bioinformatics have resulted in a revolution of knowledge of genome complexity.

Keywords: Bioinformatics, genome assembly, genomic libraries, metagenomics, sequencing platforms, SNPs (single nucleotide polymorphism).

La genómica es una disciplina que se encarga del estudio de la estructura, función y evolución de los genomas o una parte de ellos. El término genómica fue propuesto por Thomas Roderick en 1986 (Kuska 1998). Esta disciplina está estrechamente relacionada con los procesos metodológicos utilizados para secuenciar y ensamblar el genoma; la gran cantidad de datos que se generan requieren de recursos computacionales de alta gama para ser analizados y almacenados y de métodos matemáticos conocidos comúnmente como bioinformática. La genómica tiene amplias y distintas aplicaciones en la medicina, biología, agronomía, taxonomía, ecología y muchas otras áreas.

Historia de la secuenciación de los genomas y plataformas de secuenciación

Por más de 50 años, los investigadores han aplicado diferentes técnicas para secuenciar los ácidos nucleicos como son el ADN y el ARN. Secuenciar el ADN por ejemplo, es conocer la composición y el orden de sus nucleótidos (i.e. adenina A, guanina G, citosina C y timina T), así como la cantidad de cada uno de los nucleótidos o pares de bases que conforman un fragmento de ADN. La historia de la secuenciación se puede dividir en tres etapas, de acuerdo con el tamaño y número de fragmentos que se pueden secuenciar. La primera de ellas es la secuenciación de la primera generación, la cual se caracteriza por la secuenciación de un solo fragmento de ADN menor a 900 pares de bases (pb). La técnica de Maxam-Gilbert (1977) y la de Sanger y colaboradores (1977) son consideradas de primera generación. La técnica de Sanger, ha sido utilizada con mayor frecuencia y hasta la fecha sigue siendo utilizada. Esta técnica consiste en la secuenciación del ADN a través de electroforesis capilar y la incorporación de dideoxinucleótidos marcados con fluorescencia (los dideoxinucleótidos carecen del grupo 3´hidroxilo, que es importante en la extensión de la cadena de ADN y no puede formar un enlace con el fosfato 5´ del siguiente nucleótido). También se adicionan nucleótidos sin marcaje que, a diferencia de los anteriores, permiten la adición de nucleótidos y el alargamiento de la cadena de ADN. De tal manera que al final quedarán fragmentos de diferente tamaño, siempre con los dideoxinucleótidos en el extremo terminal de cada cadena de ADN. A través de la electroforesis se podrá determinar el tamaño de cada fragmento y por tanto la posición del nucleótido marcado con fluorescencia, la cual se encuentra al final de cada fragmento. Es así como se puede reconstruir el orden de los nucleótidos para posteriormente identificar los aminoácidos y las proteínas correspondientes a dichos fragmentos.
La técnica de secuenciación de Sanger o de primera generación es utilizada para secuenciar distintos genomas, de tamaño pequeño o fragmentos de genomas cortos. El primer genoma secuenciado fue el del virus PhuX174 en 1977, posteriormente en 1995 se completó la secuencia genómica de la bacteria *Haemophilus influenza*, ambos organismos poseen genomas pequeños. Después se fueron incrementando las secuencias de genomas o fragmentos de genomas de organismos eucariotas.
El método de secuenciación propuesto por Sanger o de primera generación (automatizado en la actualidad), tiene más de cuatro décadas y sigue siendo utilizado; sin embargo, la secuenciación de alto rendimiento o de nueva generación o Next Generation Sequencing (NGS) por sus siglas en inglés, ha ido ganando terreno en las ciencias biológicas, incluida la medicina.
El proyecto del genoma humano se inició en 1990, la idea era secuenciar alrededor de 3 mil millones de nucleótidos. Once años después, se anunció la versión preliminar de la secuencia completa del genoma humano que fue publicada en 2001, y en 2003 se completó el proyecto del genoma humano y se confirmó que los humanos tenían entre 20 a 25 mil genes. Con este acontecimiento molecular comenzó la bien llamada “revolución genómica”. Aunque en un principio secuenciar un genoma humano fue muy caro (i.e. 100 millones de dólares), pronto el costo y la facilidad para obtener secuencias de ADN mejoró drásticamente. La revolución genómica trajo nuevas plataformas y equipos de secuenciación, lo que ha conducido a una manera distinta de analizar e interpretar los bloques que constituyen el árbol de la vida.
Las primeras plataformas masivas de segunda generación, sentaron el cambio de paradigma, debido a que permitieron la paralelización masiva de las reacciones de secuenciación, lo que resultó

en un incremento de la cantidad de ADN que podía ser secuenciado en un tiempo corto y cada vez a un menor costo. La primera plataforma de secuenciación masiva que se fabricó corresponde al secuenciador 454 de Life Sciences en 2005, después la compañía ROCHE adquirió la plataforma y liberó el secuenciador al mercado en 2007. Esta técnica no requería de la fluorescencia de los nucleótidos sino del método de luminiscencia para medir la síntesis de pirofosfato. Esta técnica fue propuesta por Pål Nyrén y colaboradores, y tiene ventajas con respecto a otras técnicas debido a que no requiere nucleótidos modificados y tampoco electroforesis. Sin embargo, la capacidad de generar y almacenar nucleótidos secuenciados, no fue suficiente para esta plataforma comparado con nuevas técnicas que comenzaron a surgir, las cuales producían más datos y a menor costo que el secuenciador de ROCHE. Una gran cantidad de técnicas de secuenciación vinieron después del éxito del 454, la cual desapareció prácticamente del mercado en el 2016.

Al año siguiente de que se liberara el 454 al mercado, la compañía Illumina (http://www.illumina.com) lanza una nueva gama de secuenciadores. Hasta la fecha Illumina es el sistema de secuenciación más exitoso, principalmente con sus dos secuenciadores el HiSeq y el MiSeq. El secuenciador Illumina se diferencía del secuenciador Roche 454 por la tecnología de secuenciación por síntesis (ver Illumina Sequencing by Synthesis), utilizando nucleótidos de terminación de cadena extraíbles marcados con fluorescencia que pueden producir una mayor cantidad de lecturas a un menor costo (un genoma humano por menos de 1000 dólares).

Illumina proporciona varios equipos de secuenciación, por ejemplo el secuenciador MiSeq, aunque tiene un tamaño pequeño, el número de datos oscila entre 0.3 a 15 Gb, y tasas de rotación rápidas adecuadas para la secuenciación dirigida, para aplicaciones clínicas y de laboratorio pequeño; genera los resultados en 1 o 2 días a un costo mucho menor. El NOVASeq 6000 es el secuenciador que produce el mayor número de lecturas y puede generar hasta 10 billones de lecturas, “paired-end” de 100 o 250 pb, en un periodo de 36 horas y esto lo puede hacer dos veces, es decir puede hacer dos flujos, generando el doble de datos.

En la secuenciación de tercera generación, se han desarrollado tecnologías como la de Pacific Biosciences o PacBio SMRT (Eid *et al*., 2009) y Oxford Nanopore (Mikheyev y Tin, 2014), que producían lecturas mayores a 10 mil pb pero con una precisión de 75-90%. Como resultado de esto, los datos eran raramente utilizados para detectar variantes estructurales o indels (deleciones-inserciones de nucleótidos). Sin embargo, en los últimos años, PacBio HiFi (Wenger *et al*., 2019), se ha utilizado con gran éxito ya que permite la secuenciación de fragmentos más largos y con menor error. La tecnología utilizada es una optimización de secuenciación circular consenso (CCS por sus siglas en inglés, circular consensus sequencing), para incrementar la precisión de la secuenciación en tiempo real de una sola molécula (SMRT por sus siglas en inglés, single-molecule real-time) y generar lecturas con una precisión muy alta (99.8%) y con una longitud promedio de 13500 pb. No requiere una amplificación o PCR previa del ADN y permite obtener la secuencia de moléculas de cadena simple en tiempo real. Con esta tecnología se ha caracterizado el genoma humano HG002/NA24385; el proyecto es patrocinado por “The Genome in a Bottle Consortium” y tiene como finalidad generar materiales de referencia y datos de secuenciación del genoma humano con fines médicos. El ensamblaje de este genoma con PacBio tiene una precisión de 99.91% para las variantes genéticas y el 95.98% para las inserciones y deleciones < 50 pb y 95.99% para las variantes estructurales (cambio en la disposición de los genes o fragmentos cromosómicos).

Por otra parte, Oxford Nanopore desarrollada por Oxford Nanopore Systems en 2014 (https://nanoporetech.com/) es una tecnología basada en nanoporos y se basan en la identificación de las distintas bases de la cadena de ADN gracias a una señal óptica o por la variación que se produce en una corriente eléctrica de iones al pasar la cadena a través de un nanoporo anclado a una membrana formada de dos proteínas. Tiene grandes ventajas, al igual que PacBio, no necesita de una amplificación o PCR previa, sus corridas son en un tiempo corto y sus lecturas son muy largas, además

de que presenta un secuenciador portátil (MinION, https://nanoporetech.com/products/minion) que puede ser llevado en el bolsillo.

La importancia de tener secuencias más largas radica en una mayor eficacia y facilidad en la reconstrucción de la secuencia del genoma, es decir, es más sencillo el ensamblaje de las lecturas. En el caso de las lecturas cortas, éstas son de utilidad en el llamado de variantes genéticas (SNVs, *single-nucleotide variants*) y de pequeños indels, pero son poco útiles y más difíciles de utilizar para los ensamblados de novo y variantes estructurales.

Preparación de bibliotecas genómicas

En general, en la secuenciación masiva se incluyen tres pasos básicos: la preparación de bibliotecas, la secuenciación y el análisis de datos (i.e. bioinformática). Una biblioteca genómica es la representación parcial o total de todo el conjunto de ADN y se encuentra fragmentado en segmentos de diferente tamaño. Dado que secuenciar un genoma completo implica un gran esfuerzo de secuenciación, así como un intenso trabajo bioinformático, en muchos proyectos genómicos se busca tener una representación más reducida o parcial del genoma. Es por esta razón que algunas técnicas utilizan enzimas de restricción que cortan sitios específicos del ADN. Por ejemplo, la enzima EcoR1, tiene dentro del genoma un sitio de reconocimiento de 5 ´G^AATTC y 3´CTTAA^G y corta el ADN en estas regiones. El tipo de enzimas de restricción que se utilizan en la elaboración de librerías genómicas es muy variable, en algunos casos se incluye una, dos o más enzimas. La elección de las enzimas es muy importante porque puede haber enzimas que no corten lo suficiente el ADN y por tanto se tendría una representación del genoma muy reducida. Por esta razón, antes de construir una biblioteca genómica se requiere hacer pruebas con diferentes enzimas de restricción. La digestión del ADN con las diferentes enzimas, se verifica a través de un gel de agarosa.

Existen diferentes tipos de bibliotecas genómicas, algunas técnicas o métodos realizan la digestión del ADN mediante la sonicación (frecuencia de sonidos), a través de aparatos como el Bioruptor, el cual corta el ADN en fragmentos de tamaños específicos dependiendo el número de ciclos de sonicación que se programen. También se utilizan enzimas de restricción para la digestión del ADN, por ejemplo en la técnica de RADseq (Restriction-site Associated DNA sequencing, Baird *et al*., 2008) y GBS (Genotyping by Sequencing, Elshire *et al*., 2011) se usa una enzima. Además, existen otras como MIG-seq (Multiplexed ISSR genotyping by sequencing, Suyama y Matsuki, 2015) que no utilizan enzimas; esta técnica emplea cebadores ISSR para amplificar numerosas regiones de ADN y a partir de ellas construir las bibliotecas genómicas.

El método de RADseq ha sido uno de los más utilizados, y existen diferentes variantes como ddRAD (Double Digested RADseq, Peterson *et al*., 2012) en el cual se utilizan dos enzimas de restricción, ezRAD que usa una o varias enzimas (Toonen *et al*., 2013), hyRAD (Suchan *et al*., 2016) y nextRAD (Fu *et al*., 2017). A continuación se describen los pasos básicos en la construcción de las bibliotecas con el método ddRAD. En el primer paso, se realiza la fragmentación del ADN (Figura 1a), para lo cual se utilizan dos enzimas de restricción, que cortan el ADN en sitios específicos. En el caso de la química de Illumina, una vez que se ha fragmentado el ADN, se utilizan adaptadores diseñados para esta plataforma que son ligados a cada uno de los fragmentos (Figura 1b). Este adaptador es una secuencia de ADN que incluye 1) una secuencia complementaria que se localiza en la plataforma del secuenciador, 2) una etiqueta o código de barras que identifica a cada muestra y 3) una última región que se va a ligar al ADN de la muestra. Para obtener información más detallada sobre la química de Illumina se recomienda visitar la siguiente página https://www.youtube.com/watch?v=fCd6B5HRaZ8. Una vez que las muestras están etiquetadas individualmente se colocan todas en un mismo tubo para continuar con la selección de tamaño de los fragmentos a ser secuenciados . Los fragmentos seleccionados van desde los 200 - 600 pb (Figura 1c), para realizar este proceso, se usan geles de agarosa diseñados para correrse en equipos como: Pippin Prep y E-Gel Size-Select System, o un analizador de fragmentos. Finalmente a través de PCR se incrementa la concentración de los fragmentos (Figura 1d), se continúa con la cuantificación de la biblioteca genómica y se procede

con la secuenciación.

Análisis de datos

Una vez obtenidos los datos de las secuencias o lecturas (también llamadas reads en inglés) éstos requieren procesarse. La cantidad de datos puede ir desde pocos millones de secuencias hasta cientos de millones. El procesamiento de los datos debe hacerse generalmente a través de un cluster, el cual es un sistema que incluye varias computadoras unidas entre sí, generalmente por una red de alta velocidad. Cuando la cantidad de datos es menor, es posible hacer los análisis a través de una computadora personal. La mayor parte de los programas para el análisis de datos están en un sistema operativo de Linux o Unix, y en ocasiones en Windows. También existen programas que se pueden utilizar en línea, por ejemplo en la plataforma Galaxy (https://usegalaxy.org/). Esta plataforma está disponible en línea y es de acceso gratuito para el procesamiento de datos masivos genómicos y transcriptomas (RNA mensajero). Galaxy incluye numerosos programas para el análisis de datos, evitando la necesidad de instalarlos en una computadora, no requiere del uso de código sino que se pueden utilizar a través de ventanas.

La información de las secuencias comúnmente viene en un formato llamado Fastq, el cual es un formato basado en texto para almacenar una secuencia de ADN y la respectiva calidad de cada nucleótido. La calidad se mide a través del Phred score o Q, el cual se define como una propiedad que está relacionada logarítmicamente con las probabilidades de error de las bases (P) y los valores se encuentran en código ASCII. Por ejemplo, un Phred score de 10 indica que las probabilidades de error es 1 en 10, es decir el 90% de precisión. El valor de calidad mínimo aceptado es de 20, es decir 1 en 100 la probabilidad de error o bien el 99% de precisión.

A continuación, se mencionan los pasos básicos en el análisis de variantes genéticas (Figura 2). Primero, las muestras se separan de acuerdo al código de barras (demultiplex), es decir, las secuencias con el mismo código de identificación (que corresponden a una especie o muestra particular) son integradas en un sólo archivo.

Enseguida se realiza la visualización de la calidad de las secuencias, a través de programas como FastQC (Andrews, 2010). Los resultados se obtienen en un archivo html, el cual incluye gráficas para la visualización de diferentes parámetros, por ejemplo calidad de las secuencias por nucleótido, presencia de adaptadores, tamaño de las secuencias, entre otros. El segundo paso consiste en la edición de las secuencias, se pueden utilizar programas como Trimmomatic (Bolger *et al*., 2014) y Cutadapt (Martin, 2011), los cuales recortan las lecturas y quitan adaptadores. Entre las principales ediciones que se hacen, incluye la eliminación de nucleótidos de mala calidad, de adaptadores y de fragmentos más pequeños, por ejemplo menores a 50 bp. Una vez editadas las secuencias, el siguiente paso es 1) alinear con un genoma de referencia o 2) realizar los ensamblados de novo. En el primer caso, se requiere tener un genoma de referencia de la misma especie de la muestra analizada o bien de una especie cercanamente relacionada. En el segundo caso, se realiza la unión de fragmentos contiguos de ADN para reconstruir la secuencia original del ADN. Existen múltiples programas para hacer estos ensamblajes, entre los más utilizados son aquellos que utilizan las gráficas de Bruijn. En estas gráficas, los nodos son las secuencias y las ligas de los nodos describen secuencias que se traslapan. Entre los programas o ensambladores que utilizan estos algoritmos, se encuentran Velvet (Zerbino y Birney, 2008) y SPADES (Bankevich *et al*., 2012), entre otros. SPADES es uno de los programas más utilizados y el que genera mejores ensamblados. La calidad de un ensamblado se puede evaluar con diferentes parámetros, uno de los más comunes es el tamaño del contig. Un contig es la unión de las lecturas contiguas (mientras que a la unión de contigs se le conoce como scaffolds). Entre más grandes sean los contigs, el ensamblado es de mejor calidad, dado que se logrará una reconstrucción más completa del genoma. Existen programas que evalúan la calidad de los ensamblados, uno de ellos es Quast (http://cab.cc.spbu.ru/quast/). El tercer paso es el llamado de las variantes genéticas, es decir, se extraen aquellos sitios que son variables entre las muestras. Este análisis se puede realizar con el software mpileup de samtools (Li, 2011), el cual

también se puede encontrar en la plataforma Galaxy. El archivo que incluye los sitios variantes, puede convertirse a un formato vcf (variant call format, por sus siglas en inglés). Este, es un archivo de texto que contiene metadatos de los sitios variantes o SNP (single nucleotide polymorphism), así como información detallada de la calidad del genotipo, frecuencia alélica y profundidad del SNP (es decir, cuantas repeticiones por nucleótido existen para esa variante genética), así como otras características. Los archivos vcf, pueden utilizarse para estimar distintos parámetros de diversidad genética, flujo génico así como para construir árboles filogenéticos, en análisis de ordenamiento (por ejemplo PCAs) y estudios filogeográficos (donde se analiza la importancia de factores históricos como geológicos y ambientales en relación a los datos de las variantes genéticas).

A su vez, existen softwares que procesan los datos crudos de las muestras y realizan todos los pasos en conjunto (no requiere utilizar los programas por separado), a partir de la edición de las secuencias, ensamblado y llamado de variantes. Entre los programas más utilizados se encuentran ipyrad (Eaton y Overcast, 2020) y stacks (Catchen *et al*., 2011, Catchen *et al*., 2013).

Aplicaciones y ramas de la genómica

La genómica nos ha permitido conocer la estructura de genomas completos, detectar diferencias o variaciones y a la par identificar un gran número de genes presentes en diferentes grupos de seres vivos. Entre las ramas de la genómica se encuentra la genómica comparativa, la cual permite realizar estudios comparativos entre especies cercanas y lejanamente emparentadas, aportando información sobre similitudes y diferencias del genoma, su historia evolutiva y los procesos involucrados (como mutaciones, duplicaciones de genes, rearreglos cromosomales, entre otros). Por ejemplo, determina la funcionalidad del genoma y ayuda al reconocimiento de determinantes de virulencia de organismos patógenos como algunos virus y bacterias (Primrose y Twyman, 2003).

El estudio de la funcionalidad de los genes, además de la información evolutiva que proporcionan, tiene múltiples aplicaciones desde agronómicas (como el mejoramiento genético de cultivos), hasta médicas (como el desarrollo de vacunas y conocer las causas de enfermedades, por ejemplo el cáncer y la hipertensión). La farmacogenómica estudia la asociación entre la respuesta a fármacos respecto a los datos genómicos y proteómicos. Ayuda a desarrollar fármacos óptimos con base en la variación genómica individual, ya que esta variación es la causante de altas tasas de fracaso de nuevos fármacos candidatos durante la etapa de ensayos clínicos (Primrose y Twyman, 2003).

Otra rama de la genómica es la filogenómica que es utilizada en el estudio de las relaciones evolutivas entre especies cercanas y en niveles taxonómicos superiores. Por ejemplo, el genoma completo de cloroplasto es utilizado para reconstruir historias evolutivas empleando métodos filogenéticos. Por su parte, la genómica de poblaciones se enfoca en el análisis de secuencias o variantes (SNPs), obtenidas mediante la secuenciación masiva, para responder preguntas particulares de la diversidad genética y adaptativa de poblaciones de especies silvestres y cultivadas; cuya información puede ser útil en la conservación biológica. La genómica también se aplica en el estudio de comunidades completas de microorganismos, mediante la metagenómica.

El término metagenoma consiste en dos palabras: genoma, que es la secuencia de ADN, y meta, que implica que se están analizando para muchos organismos al mismo tiempo (Segre, 2022). Este enfoque o método se utiliza cuando estudiamos comunidades de microorganismos (i.e. bacterias, virus, eucariotas) donde no podemos separarlos unos de otros de manera eficiente. La metagenómica permite el estudio de todos los microorganismos, independientemente de que sean o no cultivables, aportando conocimiento de las especies presentes y de las funciones que realizan en su hábitat natural (Coughlan *et al*., 2015). Las secuencias de fragmentos del genoma o genes se obtienen directamente de una muestra ambiental como agua, aire o sedimentos y sustratos, así como los intestinos humanos o de otra especie. Las muestras se procesan en instrumentos con tecnología de secuenciación masiva de segunda o tercera generación, ya sea dirigida a un gen o a muchos.

La metagenómica funcional se basa en identificar

genes y su expresión, lo que puede aplicarse al descubrimiento de nuevas proteínas de interés alimentario, farmacéutico o industrial codificadas por los genes de microorganismos previamente inaccesibles (Coughlan *et al*., 2015). Otra aproximación de la metagenómica es el metacódigo de barras o metabarcoding, el cual permite la identificación de múltiples taxones presentes en una misma comunidad. Esta aproximación es altamente relevante en estudios de estimación de la diversidad y facilita el reconocimiento de nuevos taxa (Cristescu, 2014).

Por otro lado, en la biología molecular, el término ómica se utiliza como sufijo para referirse al estudio de la totalidad o del conjunto de moléculas. Entre ellas, además de la genómica, se incluye a la transcriptómica (Figura 3) que estudia específicamente las moléculas de ARN mensajero, es decir las regiones codificantes, y es posible evaluar qué genes se expresan en condiciones específicas. Por ejemplo, en la interacción entre los organismos (especies) y su entorno (biótico, abiótico). Por otro lado, la proteómica analiza las proteínas producidas por una célula así como sus interacciones y perfiles de expresión. Y finalmente, la metabolómica, analiza el perfil metabólico de una muestra, de forma cuantitativa y cualitativa.

Literatura citada

- Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data Version 0.11. 2. [Publicación en línea], disponible en http://www.bioinormatics.babraham.ac.uk/projects/fastqc
- Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, W.A. Cresko y E.A. Johnson. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, 3(10), e3376. https://doi.org/10.1371/journal.pone.0003376
- Bankevich, A., S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev y P.A. Pevzner. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), 455-477. https://doi.org/10.1089/cmb.2012.0021
- Bolger, A.M., M. Lohse y B. Usadel. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, 30(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170
- Catchen, J., A. Amores, P. Hohenlohe, W. Cresko y J. Postlethwait. (2011). Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics, 1(3), 171-182. https://doi.org/10.1534/g3.111.000240
- Catchen, J., P. Hohenlohe, S. Bassham, A. Amores y W. Cresko. (2013). Stacks: an analysis tool set for population genomics. Molecular Ecology, 22(11), 3124-3140. https://doi.org/10.1111/mec.12354
- Coughlan, L., P. Cotter, C. Hill y A. Alvarez-Ordóñez. (2015). Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. Frontiers in Microbiology, 6, 672. https://doi.org/10.3389/fmicb.2015.00672
- Cristescu, M.E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol Evol, 29(10), 566-71. https://doi.org/10.1016/j.tree.2014.08.001
- Eaton, D.A.R. e I. Overcast. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics, 36(8), 2592-2594. https://doi.org/10.1093/bioinformatics/btz966
- Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach y S. Turner. (2009). Real-time DNA sequencing from single polymerase molecules. Science, 323(5910), 133-138. https://doi.org/10.1126/science.1162986.
- Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, y S.E. Mitchell. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, 6(5), e19379. https://doi.org/10.1371/journal.pone.0019379
- Fu, Z., B. Epstein, J.L. Kelley, Q. Zheng, A.O. Bergland, C.I. Castillo Carrillo, A.S. Jensen, J. Dahan, A.V. Karasev y W.E. Snyder (2017). Using NextRAD sequencing to infer movement of herbivores among host plants. PloS one, 12(5), e0177742. https://doi.org/10.1371/journal.pone.0177742
- Kuska, B. (1998). Beer, Bethesda, and biology: how 'genomics' came into being. Journal of the National Cancer Institute, 90(2), 93.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987-93. https://doi.org/10.1093/bioinformatics/btr509

- López De Heredia, U. (2016). Las técnicas de secuenciación masiva en el estudio de la diversidad biológica. Munibe Ciencias Naturales, 64, 7-31. https://doi.org/10.21630/mcn.2016.64.07
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), 10-12. https://doi.org/10.14806/ej.17.1.200
- Maxam, A.M., y W. Gilbert. (1977). A new method for sequencing DNA. Proceedings of the National Academy of Sciences, 74(2), 560-564. https://doi.org/10.1073/pnas.74.2.560
- Mikheyev, A.S., y M.M. Tin. (2014). A first look at the Oxford Nanopore MinION sequencer. Molecular ecology resources, 14(6), 1097-1102. https://doi.org/10.1111/1755-0998.12324
- Peterson, B.K., J.N. Weber, E.H. Kay, H.S. Fisher y H.E. Hoekstra. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One, 7(5), e37135. https://doi.org/10.1371/journal.pone.0037135
- Primrose, S.B. y R.M. Twyman. (2003). Principles of genome analysis and genomics. Blackwell Publishing, Reino Unido. 263 pp.
- Sanger F., S. Nicklen y A.R. Coulson. (1977). DNA sequencing with chain-termination inhibitors. Proceedings of the National Academy of Sciences. 74,(12), 5463-5467.
- Segre, A.J. (26 de enero de 2022). Glossary of Genetic Terms. National Human Genome Research Institute. [Publicación en línea], disponible desde Internet en https://www.genome.gov/genetics-glossary/Metagenomics#:~:text=%22Metagenomics%22%20is%20the%20two%20words,separate%20one%20microbe%20from%20another.
- Suchan, T., C. Pitteloud, N.S. Gerasimova, A. Kostikova, S. Schmid, N. Arrigo, M. Pajkovic, M. Ronikier, y N. Alvarez. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. PLoS One, 11(3), e0151651. http://doi.org/10.1371/journal.pone.0151651
- Suyama, Y. e Y. Matsuki. (2015). MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. Scientific Reports, 5(1), 1-12. https://doi.org/10.1038/srep16963
- Toonen, R.J., J.B. Puritz, Z.H. Forsman, J.L. Whitney, I. Fernandez-Silva, K.R. Andrews y C.E. Bird. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ, 1, e203. https://doi.org/10.7717/peerj.203
- Wenger, A. M., P. Peluso, W.J. Rowell, P.C. Chang, R.J. Hall, G.T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N.D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C. Chin, A. Phillippy, M.C. Schatz, G. Myers, M.A. DePristo, J. Ruan, T. Marschall, F.J. Sedlazeck, J.M. Zook, H. Li, S. Koren, A. Carroll, D.R. Rank y M.W. Hunkapiller. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature biotechnology, 37(10), 1155-1162. https://doi.org/10.1038/s41587-019-0217-9
- Zerbino, D.R. y E. Birney. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18(5), 821-829. https://doi.org/10.1101/gr.074492.107

Anexos

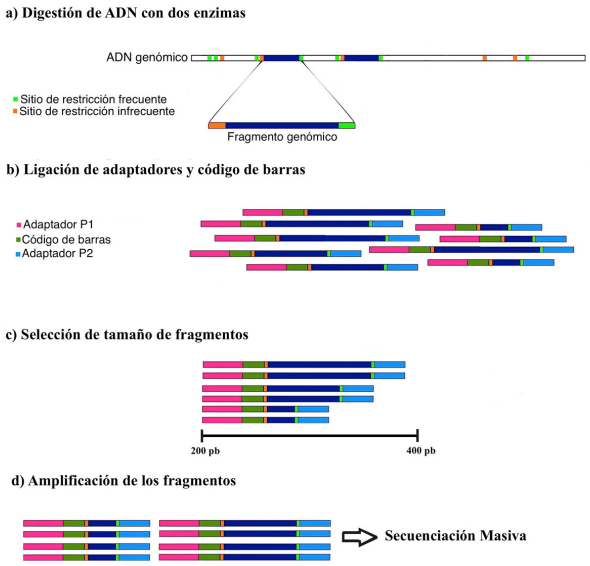

Figura 1. Flujo de trabajo para la preparación de bibliotecas genómicas con el método ddRAD. Modificado de Peterson et al., 2012 y López de Heredia, 2016.

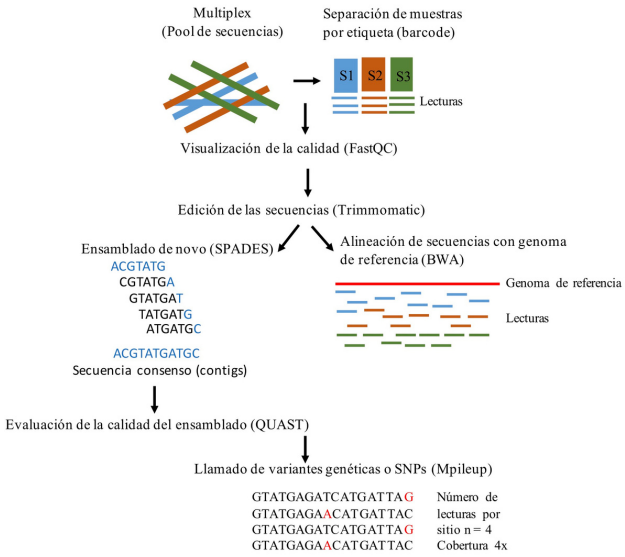

Figura 2. Flujo de trabajo para el llamado de variantes genéticas.

Figura 3. Se muestran 4 de las principales disciplinas “ómicas”