Desarrollo de marcadores moleculares de alto rendimiento y análisis de datos genotípicos en plataformas de avanzada

María Guadalupe Quintos Cortes[1], Octavio Francisco Fernández Lozada[2], Dr. Cesar Daniel Petroli[3], Dr. Fernando Henrique Ribeiro Barrozo Toledo[4] , Dra. Teresa Romero Cortes[5], Dr. Jaime Alioscha Cuervo Parra[6], Dr. Martin Peralta-Gil[7]

Resumen: Actualmente el uso de marcadores genéticos moleculares ha sido de gran relevancia para el estudio de diversidad genética en plantas. En este sentido, el presente trabajo muestra los conocimientos adquiridos por estudiantes de la Escuela Superior de Apan, de la Universidad Autónoma del Estado de Hidalgo (UAEH) en una estrategia metodológica que involucra la aplicación de nuevas tecnologías generadoras de una gran cantidad de datos (High-throughput), el uso de análisis estadísticos y la implementación de herramientas bioinformáticas. La generación de datos High-throughput esta mediada por la tecnología DArTseq, la cual desarrolla miles de marcadores moleculares de tipo SNPs y Silico-DarT o PAVs (presencia/ausencia de variación). Esta novedosa tecnología se desarrolla en tres etapas: a) reducción de la complejidad del genoma con enzimas de restricción específicas; b) amplificación y secuenciación de los fragmentos generados mediante secuenciadores de nueva generación; c) identificación de polimorfismos o marcadores moleculares. El análisis estadístico, el uso de herramientas bioinformáticas y multiplataformas (CurlyWhirly y Flapjack) para la visualización de grandes cantidades de datos genotípicos representan factores importantes en la interpretación de la diversidad genética de las muestras evaluadas para el aprendizaje.
Palabras clave: DArTseq, Marcadores moleculares, Flapjack, CurlyWhirly

Introducción

México es considerado como el centro de origen y domesticación del maíz (*Zea mays* L.) y es uno de los países más reconocidos a nivel internacional por su gran diversidad de razas (Matsuoka *et al.*, 2002). Actualmente el conocimiento de la diversidad en maíces nativos de México es de fundamental importancia para el planteamiento de estrategias de caracterización y conservación de germoplasma, con potencial de uso en el mejoramiento genético. En este sentido, existen herramientas biotecnológicas que permiten diferenciar muestras de plantas a nivel genómico. Así, es posible la caracterización genotípica del germoplasma para llevar a cabo estudios de diversidad biológica entre las diferentes razas, así como la selección de caracteres de interés agronómico.

Los marcadores moleculares, determinados por los polimorfismos identificados en muestras de ADN, están representados por secuencias específicas de dicha molécula, con ubicación definida en un locus de un cromosoma, y cuya herencia genética puede ser observable o cuantificable. Existen diferentes tipos de métodos en el desarrollo de los marcadores moleculares, y algunos de ellos incluyen el Polimorfismo de Longitud de Fragmentos de Restricción (RFLP), Polimorfismos en la Longitud de Fragmentos Amplificados (AFLP), Amplificación Aleatoria de ADN Polimórfico (RAPD), Secuencias Simples Repetidas (SSR) y Polimorfismos de un Solo Nucleótido (SNP) (Parker *et al.*, 1998; Becerra y Paredes, 2000; Rentaría, 2007). Entre las muchas aplicaciones que tienen los marcadores moleculares se pueden mencionar: análisis filogenéticos, detección de relaciones entre diferente germoplasma en bancos de semillas y programas de mejoramiento (Reif *et al*., 2005), asociación fenotipo-genotipo, análisis de predicción por métodos como la selección genómica, así como la búsqueda de loci de características cuantitativas

[1] María Guadalupe Quintos Cortes, alumna de la Escuela Superior de Apan de la Universidad Autónoma del Estado de Hidalgo (ESAp-UAEH), Apan. Hgo. mari.quintos@yahoo.com (primer autor)
[2] Octavio Francisco Fernández Lozada, alumno de la ESAp-UAEH, Apan. Hgo. octaviolozada@live.com.mx
[3] Dr. Cesar Daniel Petroli, es Especialista en Genotipeo de Alto Rendimiento-SAGA, del Programa de Recursos Genéticos del CIMMYT, El Batán, Texcoco. c.petroli@cgiar.org
[4] Dr. Fernando Toledo, es Científico Asociado en Estadística Agrícola / Biometría, del Programa de Recursos Genéticos del CIMMYT, El Batán, Texcoco. f.toledo@cgiar.org
[5] Dra. Teresa Romero Cortes es miembro del Cuerpo Académico Biociencias Moleculares y profesor investigador en la ESAp-UAEH, Apan. Hgo. romero@uaeh.edu.mx
[6] Dr. Jaime Alioscha Cuervo Parra es miembro del Cuerpo Académico Biociencias Moleculares y profesor investigador en la ESAp-UAEH, Apan. Hgo. alioscha@uaeh.edu.mx
[7] Dr. Martin Peralta Gil es miembro del Cuerpo Académico Biociencias Moleculares y profesor investigador en la ESAp-UAEH, Apan. Hgo. martin_peralta10391@uaeh.edu.mx (**autor corresponsal**)

(QTLs) o inclusive de genes de interés que actúen directamente sobre la expresión de una característica especifica (Vigouroux *et al*., 2005; Riedelsheimer *et al.*, 2012; Dsechamps *et al.*, 2012). Para este efecto, la creciente información generada a partir de tecnologías de secuenciación de última generación (NGS) ha permitido impulsar el desarrollo de nuevos métodos con capacidad para generar "tsunamis" de datos genotípicos, los que muchas veces son difíciles de almacenar y emplear en la evaluación de los materiales. Sin embargo, actualmente estos datos masivos pueden estar ligados a herramientas bioinformáticas que permiten el procesamiento, almacenamiento, análisis y visualización del gran conjunto de datos.

En México, el Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) es una Institución que realiza investigación científica para el desarrollo de variedades mejoradas de maíz y trigo; por esta razón, utiliza diversas estrategias metodológicas que permitan generar dichas variedades con el objetivo de combatir la hambruna y pobreza a nivel global (https://www.cimmyt.org/). Así, CIMMYT ha decidido incluir a los marcadores moleculares dentro de sus estrategias para la identificación de materiales que puedan ser favorables para su propósito. El Programa de Recursos Genéticos dentro de CIMMYT, alberga el laboratorio Servicio de Análisis Genético para la Agricultura (SAGA), responsable del uso de una plataforma de genotipificación de alto rendimiento. Esta plataforma fue creada a partir de la iniciativa MasAgro (SADER), por el proyecto Descubrimiento de la Semilla (Seed of Discovery), administrado por el CIMMYT. Esta plataforma le aporta a México capacidad de genotipado y análisis genético de vanguardia (High-throughput Genotyping/Sequencing), proporcionando una visión única de la variación genética del trigo y el maíz a nivel de "secuencia" de ADN.

Los servicios de SAGA están disponibles para todos los científicos del CIMMYT, universidades, programas nacionales de investigación agrícola y empresas privadas. Además, como en este caso, está abierto a la participación de estudiantes que deseen conocer sobre estas tecnologías. A nivel mundial, pocas plataformas producen este tipo de datos y la mayoría son inaccesibles para los científicos que trabajan en instituciones financiadas con fondos públicos debido a sus dificultades económicas o logísticas. El presente trabajo sirve como ejemplo y guía para desarrollar la metodología e implementación de la tecnología DArTseq, mostrando las diferentes estrategias metodológicas aplicadas por los estudiantes, tanto experimentales, estadísticas e informáticas, utilizadas en el CIMMYT para la generación de marcadores moleculares, así como también la identificación y visualización de la diversidad genética en una raza de maíz.

Descripción del Método

El Programa de Recursos Genéticos del CIMMYT administra dos de los mayores bancos de germoplasma de maíz y trigo del mundo. Ambos bancos han sido caracterizados genéticamente por el método DArTseq, que es el método de genotipificación utilizado en esta capacitación. Durante el proceso para la generación de marcadores moleculares, se requiere de cuidados específicos, iniciando con la etapa de sembrado de las accesiones en invernadero donde se controlan las condiciones para dirigir la germinación de las semillas y obtener material foliar para las siguientes etapas. Posteriormente, las muestras vegetales pasan por subprocesos específicos y protocolarios que involucran la desinfección de muestras y captura de datos fisiológicos para realizar la extracción de ácidos nucleicos de alta calidad. El servicio proporcionado por SAGA utiliza una plataforma de alto rendimiento para la genotipificación de gran número de muestras de manera simultánea. En la plataforma se utiliza la tecnología DArTseq, desarrollada por la empresa Diversity Arrays Technology (DArT), la cual tiene la capacidad de identificar dos tipos de polimorfismos: SilicoDArT (PAVs) y SNPs DArTseq. Este es un método de genotipado por secuenciación eficiente, que hace posible el descubrimiento de decenas o centenas de miles de polimorfismos de alta calidad en el genoma completo mediante la reducción de la complejidad del genoma mediada por el uso de enzimas de restricción y la secuenciación de los fragmentos de restricción (Sansaloni *et al*., 2011; Edet *et al.*, 2018, Sansaloni et al., 2020).

La generación de una gran cantidad de datos obtenidos a través de la tecnología ofrecida en SAGA es de gran importancia, ya que ayuda a identificar la diversidad genética de un cultivo, así como comprender el control genético de los caracteres evaluados en una planta o cultivo de interés. Los pasos más destacados en un entrenamiento para el desarrollo de marcadores moleculares en SAGA son: el tratamiento de las muestras, la obtención de los marcadores moleculares, así como, el análisis estadístico y computacional de la información. Estos pasos son representados en un panorama general en la figura 1.

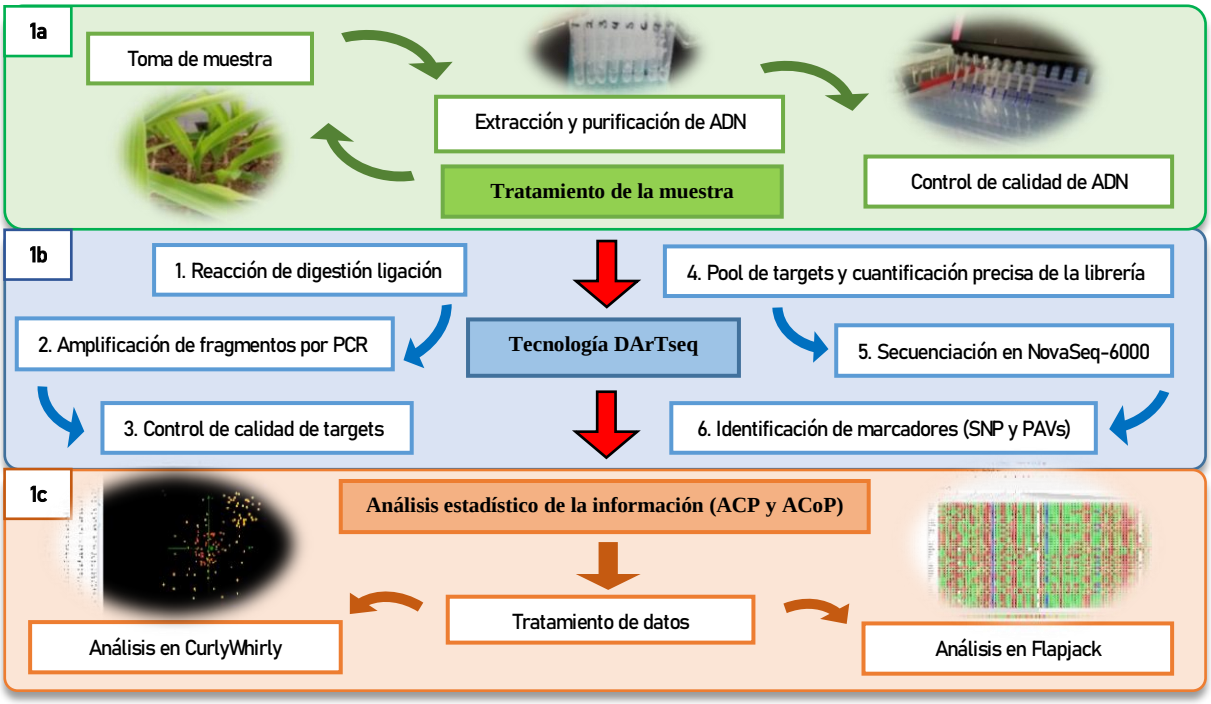

Figura 1. Metodología empleada en el proceso de capacitación de los estudiantes en la generación y análisis de marcadores moleculares mediante la implementación de la tecnología DArTseq. **1a.** Toma de la muestra, seguido por la extracción y purificación de ácidos nucleicos y finalmente el control de calidad del ADN mediante electroforesis. **1b.** Tecnología DArTseq, representada por una secuencia de 6 pasos. **1c.** Análisis estadístico de la información, representada por la asociación de datos con ACP y ACoP, y por último el uso de multiplataformas como CurlyWhirly y Flapjack para visualización e interpretación de la información.

Tratamiento de la muestra

Toma de muestra: El origen de las muestras puede ser del tejido foliar o de la semilla: esto depende de la cantidad de material que se disponga para los procesos. Generalmente se utiliza el tejido foliar fresco cosechado en invernadero o campo para la extracción de ADN. Las hojas recolectadas se depositan en sobres de papel, malla o plástico debidamente etiquetados para su identificación. El material colectado se congela por 24 horas, luego se traspasa a un liofilizador y las muestras se secan durante 72 horas. Posteriormente se cortan las hojas liofilizadas en pequeños círculos y se depositan en contenedores de plástico, los cuales contienen balines metálicos encargados de triturar las hojas hasta convertirlas en un polvo fino (Figura 1a).

Extracción y purificación de ADN: Los protocolos tradicionales consisten en cinco etapas principales: homogeneización del tejido, lisis celular, separación de proteínas y lípidos, precipitación y redisolución del ADN. En este caso se utiliza el método de Saghai-Maroof propuesto en 1984 con algunas modificaciones (CIMMYT Laboratory Protocols). Resumidamente, para realizar esta metodología se requiere agregar a un tubo de polipropileno una cantidad considerable del tejido molido, posteriormente se añade la solución amortiguadora CTAB para extracción. Se agitan los tubos y posteriormente las muestras se tratan con una serie de soluciones de cloroformo/octanol, isopropanol y etanol, para la purificación del ADN (Figura 1a).

Control de calidad de ADN: La cuantificación del ADN, extraído y purificado se realiza con geles de agarosa al 0.8%. Los geles se corren en buffer TAE, utilizando ADN Lambda sin cortar (λ) como marcador de peso molecular. Se utiliza un transiluminador de UV para capturar una fotografía del gel, la cual sirve para cuantificar la cantidad de ADN purificado, utilizando como referencia al marcador de peso molecular (Figura 1a).

Tecnología DArTseq

La tecnología DArTseq ha implementado la secuenciación de las representaciones genómicas en plataformas de secuenciación de próxima generación (NGS). La transición a una plataforma de secuenciación permite el aumento en fragmentos genómicos analizados, así como el número de marcadores detectados de manera más precisa y la

posibilidad de obtención de un marcador codominante. El método de reducción de complejidad genómica utilizado por DArTseq facilita la obtención de fracciones del genoma correspondientes a genes activos. Esto se debe a las características de las enzimas de restricción empleadas y combinadas, que separan las secuencias con baja copia, las cuales resultan ser más informativas para descubrir y tipificar marcadores (Diversity Arrays Technology, 2019).

Reacción de digestión/ligación: Es el primer paso y consiste en la reducción de la complejidad genómica. Este proceso se realiza en una placa de reacción con 96 pozos. Cada pozo suele corresponder a un genotipo específico el cual no debe mezclarse con los otros, para evitar tener muestras contaminadas, es decir ADN mezclado. Este primer paso, inicia con una mezcla de digestión/ ligación, que contiene ADN genómico, enzimas de restricción, buffer de restricción, adaptadores comunes, ligasa, ATP y "barcodes" (código de barras). Primero se utiliza una combinación de enzimas de restricción sensibles a la metilación, una de corte raro y otra de corte frecuente (por ejemplo: PstI y NspI), las cuales cortan el ADN genómico, estas enzimas permiten la obtención de aquellas regiones en el ADN que son más probables de contener información de genes activos, separándolas de las regiones repetidas del ADN dentro del genoma. Al combinar una enzima de corte frecuente con una de corte raro se regula la longitud de los fragmentos de ADN. Posteriormente se ligan los fragmentos digeridos mediante sus extremos cohesivos a adaptadores y un barcode único para ese genotipo (Figuras 1b y 2**)**, de tal manera que la construcción sea extremo cohesivo de corte de PstI-adaptador-barcode, mientras que en el otro extremo solo se unirá el adaptador. Los barcodes son secuencias de ADN combinadas de manera aleatoria, de modo que la secuencia resultante sea única para representar cada uno de los genotipos. En la Figura 2, se muestra la construcción completa, ligando también primers u oligonucleótidos *forward* (FW) y *reverse* (RV), los cuales servirán como punto de partida para realizar la amplificación de los fragmentos de ADN.

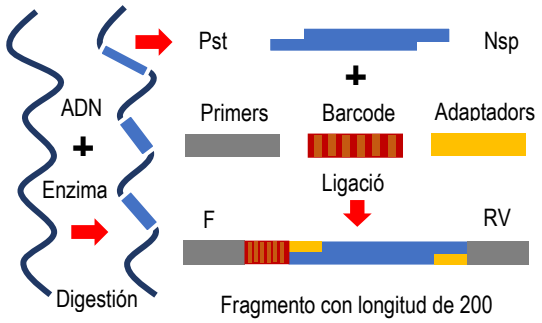

Figura 2. Representación gráfica de la reacción de digestión ligación. Esta reacción busca la reducción de la complejidad genómica, representando el primer paso particular de la tecnología DArTseq, así como el inicio del segundo paso, la amplificación.

Amplificación de fragmentos por PCR: El segundo paso corresponde a la amplificación de los fragmentos de ADN mediante la técnica de reacción en cadena de la polimerasa (PCR). En este paso se utiliza el ADN molde y los siguientes reactivos: ddH_2O, Buffer, oligonucleótidos, desoxirribonucleotidos trifosfato (dNTPs), ADN polimerasa (Taq) y cofactores como Mg_2Cl. La mezcla de reactivos para el PCR se combina con el ADN templado y se procesa en el termociclador, este equipo es programado con 40 ciclos de amplificación. Pasado el tiempo de programación del termociclador se extraen las placas de PCR de 96 pozos con los fragmentos de ADN amplificados.

Análisis de la calidad de los targets: Durante el tercer paso se efectúa el control de calidad de los targets. Los targets están representados por los fragmentos amplificados, los cuales contienen los adaptadores integrados, las secuencias de los primers y el barcode. Los targets amplificados se corren en geles de agarosa al 1.2% y buffer TAE. Una vez que ha corrido el gel en la cámara de electroforesis, se procede a tomar una fotografía del gel en el fotodocumentador y posteriormente se califican los targets con base a la amplificación de los fragmentos.

Pool de targets y cuantificación precisa de la librería: El cuarto paso consiste en realizar un pool de los targets que han sido calificados como de buena calidad, para ello se toma una muestra representativa de cada target, proveniente de la placa de 96 pozos, y se mezclan en un solo tubo Eppendorf para realizar una librería. En el caso de que se procese más de una placa se realiza un super pool, en este se toma también una muestra representativa de cada pool que se ha generado por cada placa y se mezclan en un nuevo tubo, recordando que cada muestra es identificada posteriormente por el barcode que posee. Una vez que la librería está lista se realiza su cuantificación precisa, para

esto se toma una muestra del super pool y esta es mezclada con un fluoróforo, se toman varias lecturas y se saca un promedio para saber la concentración de ADN presente en la librería.

Secuenciación en NovaSeq-6000: El quinto paso corresponde la secuenciación automática en el equipo NovaSeq 6000, este sistema provee una variedad de combinaciones de longitud de lectura y tipos de celda de flujo, ofreciendo así, flexibilidad en el rendimiento y duración del experimento. El equipo utiliza la secuenciación por síntesis, este método se basa en terminadores reversibles, en cada ciclo se incorpora una base fluorescente "marcada" y esta es identificada por excitación láser. Además, permite la secuenciación paralela a gran escala de miles de millones de fragmentos de ADN y permite detectar bases individuales a medida que se incorporan a las cadenas de ADN en crecimiento.

Identificación de marcadores moleculares (SNP y silico-DArT o PAVs): El sexto paso es la identificación de los marcadores moleculares, en el cual la intensidad de fluorescencia es traducida en *reads* (lecturas) de secuenciación. Durante el proceso se hace un filtrado de la calidad de los datos, eliminando los errores de secuenciación para clasificar los *barcodes*. Los *reads* de secuenciación son alineados a una biblioteca genómica de referencia interna. Al comparar las muestras con las representaciones del cultivo evaluado se obtienen los marcadores moleculares SNPs, así como los marcadores de tipo PAVs a través de la presencia o ausencia de cada fragmento. Por último, se lleva a cabo un filtrado de calidad de los datos de manera específica, utilizando varios parámetros de calidad para cada experimento, en el que se obtienes decenas o centenas de miles de polimorfismos por cada tipo de marcador (PAV y SNP) y por cada muestra analizada.

Multiplataformas utilizados para el análisis y visualización de datos

CurlyWhirly: La plataforma de CurlyWhirly (CW) es una herramienta que desarrollada en un esfuerzo conjunto entre el proyecto Seed of Discovery (https://seedsofdiscovery.orgr) y el Instituto James Hutton de Escocia (https://www.hutton.ac.uk) que permite representar una gran cantidad de datos en un gráfico tridimensional. Esta plataforma es adecuada para interpretar los Análisis de Coordenadas Principales (PCoA), Análisis de Componentes Principales (PCA) y Escalado Multidimensional (MDS) (James Hutton Institute, 2019). El programa se encuentra de forma libre y gratuita en la página del James Hutton Institute, una vez instalado se carga la matriz de frecuencias alélicas en dicha plataforma con las gráficas como resultado. CurlyWhirly ayuda a tener una visualización relativamente sencilla de la diversidad o similitud genética entre los diferentes genotipos analizados. Dentro de sus características se encuentra la jerarquización de la información genéticay filtrar puntos específicos que pueden ser seleccionables, o deseleccionar categorías al visualizarlo coloreado o sin color respectivamente. También contiene opciones que permiten manipular el gráfico, tales como: restaurar imagen, girar el gráfico, captura de la pantalla, captura de la película del gráfico mientras está en rotación y entre otras opciones.

Flapjack: La multiplataforma denominada Flapjack también ha sido creada por el mismo proyecto, y permite la observación gráfica de genotipos y haplotipos utilizando los datos de los SNPs). Estos datos con diferentes densidades y tres formatos (mapeados, no mapeados o combinados) permite también analizar y visualizar una gran cantidad de datos a nivel cromosómico, facilitando una navegación fluida para el investigador. Permitiendo realizar diversas comparaciones entre cromosomas, marcadores moleculares y diversas líneas, así como identificar posiciones para un Locus de Carácter Cuantitativo (QTL, por sus siglas en inglés). Flapjack también consiente el cálculo de matrices de similitud y contribuye a los análisis de diversidad genética junto con CurlyWhirly, entre otras características "mapeables" (Milne *et al*., 2010; Seeds of Discovery, 2017).

Análisis y Resultados

Una vez realizada la capacitación en la obtención de las muestras de ADN y el desarrollo de los marcadores moleculares con la tecnología DArTseq, el entrenamiento se continua con el proceso de análisis de los datos obtenidos a partir de las muestras evaluadas. Aquí son expuestos algunos ejemplos sobre estos procesos y su aplicación final para un grupo de accesiones pertenecientes al banco internacional de germoplasma de maíz del CIMMYT.

Análisis estadístico de la información

Los resultados generados por el secuenciador automático están representados por un archivo de Excel delimitado por comas con extensión.csv, el cual consiste en una matriz de frecuencias alélicas. En este archivo las columnas contienen los nombres de los individuos (genotipos), y en las filas se encuentran los marcadores que se identificaron en el genoma de esas muestras. Para el tratamiento de los datos, y con el objetivo de mejorar la calidad de los resultados, se suele implementar parámetros como el de Menores Frecuencias Alélicas (MAF) a la matriz. De esta

forma, únicamente se conservan los marcadores localizados dentro del rango de frecuencia >0.05 y <0.95; mientras que los marcadores con frecuencias extremas, o fuera del rango antes mencionado, se descartan del estudio. También son descartados los marcadores con datos monomórficos para la población estudiada, los que poseen altos valores de datos perdidos, y aquellos con bajos valores de reproducibilidad entre la muestra original y su réplica. Una vez que se han filtrado los marcadores, el archivo en Excel está listo para realizar análisis estadísticos.

Por lo tanto, los datos de alto rendimiento obtenidos y tratados pueden ser empleados en diversas técnicas estadísticas para su estudio. El Análisis de Componentes Principales (ACP) y el Análisis de Coordenadas Principales (ACoP) son dos herramientas estadísticas que le permiten al usuario manejar los datos e interpretarlos como información valiosa (Figura 1c). El ACP reduce la dimensión de un gran número de variables correlacionadas, genotípicas o fenotípicas, que representen debidamente la variabilidad de los datos. Esto se realiza con la finalidad de convertirlas en un nuevo conjunto de variables, no correlacionadas entre sí, según el orden de importancia de la variabilidad total que representan en la muestra. El análisis mediante ACoP ordena los datos en función de un conjunto de variables y los transforma en medidas de distancia; la principal diferencia con ACP es que ACoP puede utilizar tanto variables cuantitativas como cualitativas. Estas herramientas de agrupamiento de datos se aplican en estudios de diversidad genética, análisis de comunidades, variaciones morfológicas entre especies y estructura espacial de las comunidades.

Ejemplo de aplicación en CurlyWhirly

Para visualizar los datos genotípicos obtenidos a partir de la tecnología DArTseq en la plataforma CurlyWhirly, se necesita tener el filtrado de estos en un archivo delimitado por comas. Al cargarlo desde una ubicación conocida a la plataforma, los datos serán interpretados por el software tomando la matriz de frecuencias alélicas de los marcadores moleculares pertenecientes a cada genotipo estudiado. La versión 1.19.09.04 de CurlyWhirly permite importar los datos seleccionando “Importar un archivo a CurlyWhirly” desde la pantalla de bienvenida de la plataforma, el botón “Abrir datos” de la barra de herramientas, o arrastrando y soltando un archivo de datos con formato legible en la ventana cargada de CurlyWhirly.

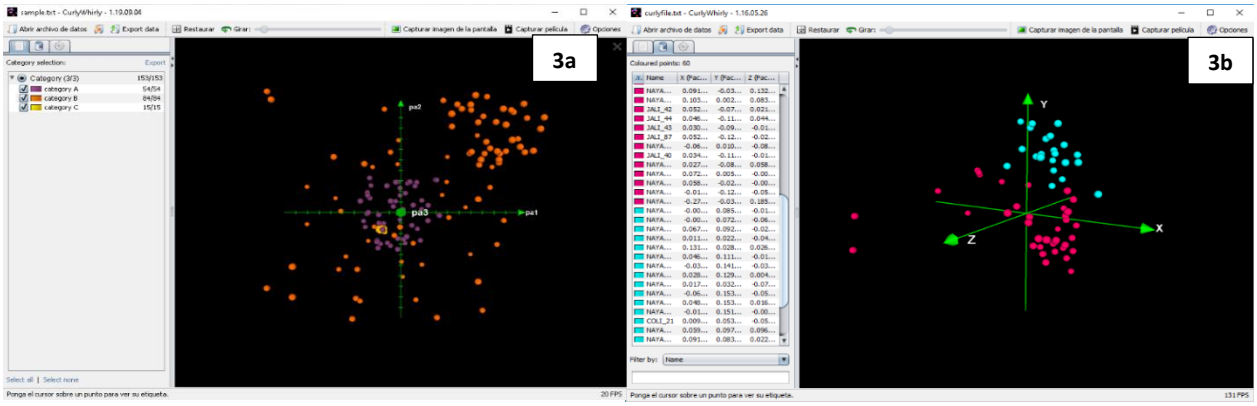

Figura 3. Visualización de la pantalla de trabajo de CW con datos importados. **3a**. Distribución de los datos de muestra por deafult del programa, representa un total de 153 accesiones o genotipos, agrupados en tres categorías A, B y C (morado, naranja y amarillo) con 54, 84 y 15 datos respectivamente (James Hutton Institute, 2019) y **3b.** Distribución de genotipos de maíz de raza Jala con un total de 60 accesiones, agrupados en dos categorías con base a su distanciamiento genético. Se representa la distribución espacial de la diversidad genética encontrada en los genotipos analizados mediante la tecnología DArTseq.

En la figura 3, el software trabajará con los datos importados y los mostrará en la pantalla de trabajo principal, reemplazando la pantalla de bienvenida. Una gráfica 3D básica de los puntos de datos del conjunto de datos, con los ejes X, Y, Z se mostrarán como en el inciso a y b de la figura 3 que representan un archivo muestra y un archivo de datos trabajados de maíz de raza Jala, respectivamente. El panel izquierdo expondrá las categorías (y los valores asociados a esas categorías) que se encontraron en el conjunto de datos (Figuras 3a y 3b).

Ejemplo de aplicación en Flapjack

Flapjack está diseñado para la visualización y exploración de datos de genotipos de plantas, en conjuntos de datos que contienen sólo unos pocos a muchos miles de líneas y marcadores. La versión 1.20.10.07 permite importar conjuntos de datos desde la barra de menú, en el botón archivo y seleccionando la opción “Importar datos”, es aquí donde se puede elegir qué tipo de datos se pueden cargar, tales como; genotipos, fenotipos, mapas genéticos, QTL, etc.

Cada conjunto de datos debe contener cierto tipo de información, en el caso de los mapas estos deben contener información sobre los marcadores, el cromosoma en el que están y su posición dentro de ese cromosoma. Flapjack los agrupará y los ordenará por cromosoma y distancia una vez cargados. En el caso de los genotipos, estos deben contener una lista de líneas de variedad, con datos de alelo por marcador para esa línea. Para ambos casos, los archivos deben estar en formato de texto sin formato delimitado por tabulaciones o comas.

La pantalla principal de Flapjack referente a la visualización de genotipos se divide en varias áreas separadas, cada una de ellas muestra un componente diferente relacionado con el genotipo. Flapjack asignará colores a cada tipo de alelo como se muestra en la Figura 4. Se pueden llevar a cabo una gran cantidad de acciones dentro del software, el tipo de estudio o análisis que se requiera definirá qué acciones son clave para llevarlo a su realización.

Figura 4. La figura muestra la interfaz de Flapjack y el área de trabajo, representando el resultado de los datos de muestra cargados en la multiplataforma. Entre otros aspectos se observa: el panorama general de la distribución de los marcadores moleculares a lo largo del genoma.

Experiencia de los alumnos durante la estancia profesional en CIMMYT

Las prácticas profesionales se llevaron a cabo en el Programa de Recursos Genéticos del CIMMYT, en el laboratorio de SAGA y la Unidad de Bioestadística. Durante este tiempo los alumnos desarrollaron capacidades en la producción de marcadores moleculares de alto rendimiento a partir de la tecnología DArTseq implementada en dicho laboratorio. Los alumnos estuvieron involucrados desde el tratamiento de las muestras y dieron seguimiento al proceso hasta la secuenciación de fragmentos de ADN en el secuenciador NovaSeq-6000. Además, el conocimiento acerca del análisis estadístico de la información se adquirió durante cursos previos en el área de estadística y biometría impartidos por el Doctor Fernando Henrique Toledo y el Doctor César Daniel Petroli del programa de recursos genéticos, obteniendo así una visualización completa de la implementación de DArTseq, el análisis estadístico de la información genética y el uso de multiplataformas como CurlyWhirly y Flapjack. En este caso, fueron evaluadas accesiones del banco de germoplasma internacional de maíz perteneciente al CIMMYT.

Los alumnos también estuvieron involucrados en otras actividades de aprendizaje dentro de SAGA, algunas de ellas fueron: participar en las actividades de sembrado y colecta de muestras de poblaciones específicas de maíz dentro de los invernaderos, captura de datos de germinación de maíz para distintos proyectos del programa de recursos genéticos, preparación de muestras para su posterior análisis genético, seguimiento de los protocolos y apoyo en la preparación de soluciones para la extracción y purificación de ADN de alta calidad de maíces, asistencia y apoyo en la calificación de “targets” y “barcodes” implementados en los análisis genéticos de muestras específicas procesadas, así como, ayudar en la preparación de los reactivos exclusivos necesarios para el funcionamiento del secuenciador NovaSeq-6000 para procesar las librerías genómicas.

Conclusiones

El presente trabajo muestra los conocimientos adquiridos por los estudiantes María Guadalupe Quintos Cortes y Octavio Francisco Fernández Lozada de la Licenciatura en Ingeniería en Biociencias de la Escuela Superior de Apan, UAEH. Los alumnos realizaron prácticas profesionales en el Laboratorio de Servicio de Análisis Genético para la Agricultura (SAGA) en el CIMMYT, ubicado en El Batán, Texcoco, Estado de México. La experiencia obtenida durante sus prácticas reafirmó el conocimiento adquirido durante la Licenciatura, así como el desarrollo de nuevas competencias disciplinares y genéricas. El trabajo también reforzó las Líneas de Generación y Aplicación del Conocimiento (LGCA) de la Escuela Superior de Apan de la UAEH, así como el impulso a las Líneas de Investigación de los Investigadores participantes. Por este medio, se reconoce y agradece al Dr. Daniel Cesar Petroli, al Dr. Fernando Henrique Ribeiro Barrozo Toledo, a la Ingeniera Guadalupe Valdés, así como al CIMMYT en general, por el apoyo incondicional y las facilidades prestadas durante el desarrollo de las actividades académicas de los alumnos.

En la actual era postgenómica, podemos considerar a la tecnología DArTseq como un método efectivo de genotipificación y consecuente caracterización genética a gran escala, capaz de desarrollar simultáneamente el perfil genómico de miles de muestras de ADN para cualquier cultivo agronómicamente importante. La tecnología DArTseq ayuda a comprender la diversidad genética de un cultivo, así como el control genético de los caracteres evaluados en una planta. A partir de la producción de decenas o incluso cientos de miles de marcadores moleculares tipo silico DArT (PAV - Presencia/Ausencia de Variación) y SNPs (Single Nucleotide Polymorphism). A partir de la obtención de la información genotípica, las estrategias estadísticas de asociación (ACP y ACoP) son importantes para la identificación de la variabilidad genética, las cuales pueden ser complementadas por la visualización facilitada de los datos en multiplataformas como CurlyWhirly o Flapjack.

Limitaciones

Las limitaciones del proyecto se enfocan en aspectos teóricos y en la calidad del trabajo a nivel experimental: cabe destacar que estas se deben a la necesidad de trabajar con ADN de alta calidad. Es importante mencionar que debido a que DArTseq es una técnica novedosa, no se puede implementar en cualquier laboratorio y por lo tanto se requiere de equipo especializado. Por ejemplo, el sistema de secuenciación masiva, requiere del uso del equipo NovaSeq 6000, el cual cuenta con una tecnología de nueva generación. Dicha tecnología genera una gran cantidad de resultados y ofrece una alta productividad y flexibilidad, además de adaptarse prácticamente a cualquier genoma. Lamentablemente, en México, este equipo de secuenciación solamente se encuentra en el CIMMYT y se utiliza para desarrollar proyectos relacionados con el mejoramiento de las semillas de maíz y trigo.

Recomendaciones

De acuerdo al trabajo realizado en esta investigación, es recomendable utilizar la tecnología DArTseq para el análisis de perfiles genómicos de diferentes muestras de ADN, provenientes de cultivos agronómicamente importantes. DArTseq incluye una plataforma de secuenciación por síntesis (NovaSeq 6000) que permite identificar miles de marcadores moleculares y ofrece alta productividad y flexibilidad para cualquier genoma. El sistema es recomendable para los investigadores interesados en esta área y debido a que no es un procedimiento sencillo, es importante considerar los aspectos técnicos relacionados con el funcionamiento del proceso. Se recomienda a los usuarios de la tecnología DArTseq enfocarse en utilizar los servicios de extracción de ADN, ofrecidos por SAGA, debido a que el procesamiento de las muestras inicia desde el sembrado en invernadero, con todos los controles y cuidados necesarios para realizar una extracción de ADN de alta calidad. Con la finalidad de obtener información molecular para su uso en programas de mejoramiento genético, así como de estudios de diversidad genética y también para aplicar conservación de la biodiversidad.

Referencias

Becerra, V. y Paredes M. (2000). Uso de marcadores bioquímicos y moleculares en estudios de diversidad genética. Agricola Técnica, 60 (3).

CIMMYT. Laboratory protocols: CIMMYT applied molecular genetics laboratory,Third Edition. 2005.

Diversity Arrays Technology (2019). "DArTseq/LD/Met," Recuperado 09/07/2020, desde https://www.diversityarrays.com/technology-and-resources/dartseq/

Dsechamps, S., Llaca, V. and May, G. D. (2012). Genotyping-by-sequencing in plants. Biology 1: 460-83.

Edet, O. U., Gorafi, Y. S. A., Nasuda, S., & Tsujimoto, H. (2018). "DArTseq-based analysis of genomic relationships among species of tribe Triticeae," Scientific Reports, 8(1).

James Hutton Institute. (2019). "CurlyWhirly," Recuperado el 29/06/2020, desde https://ics.hutton.ac.uk/curlywhirly/

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez, J., Buckler, E., & Doebley, J. (2002). "A single domestication for maize shown by multilocus microsatellite genotyping," Proceedings of the National Academy of Sciences. 99(9), 6080-6084.

Milne I., Shaw P., Stephen G., Bayer M., Linda Cardle, Thomas W.T. B., Flavell A.J. & Marshall D. (2010). "Flapjack—graphical genotype visualization," Bioinformatics, 26 (24), 3133–3134.

Parker P.G., Snow, A.A., M.D. Schug, G.C., Booton & P.A. Fuerst. (1998). What molecules con tell us about population: choosing and using a molecular marker. Ecology. 79: 361-382.

Reif, J., Melchinger, A., & Frisch, M. (2005). "Genetical and Mathematical Properties of Similarity and Dissimilarity Coefficients Applied in Plant Breeding in Seed Bank Management". Crop Science Society of America, 1-7.

Rentaría Alcántara, M. (2007). Capítulo 18. Breve revisión de los marcadores moleculares. En L. E. Eguiarte, V. Souza, & X. Aguirre, Ecología Molecular, México. 541-562.

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L. & Melchinger, A. E. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc. Nat. Acad. Sci.109: 8872-77.

Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., … Pixley, K. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. Nature Communications, 11(1).

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., & Kilian, A. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proceedings, 5(Suppl 7), P54.

Seeds of Discovery, CIMMYT. (2017). "Descubriendo la diversidad genética de la semilla. Taller – SAGA – Servicio de Análisis Genético para la Agricultura," Recuperado el 10/07/ 2020, desde: https://seedsofdiscovery.org/es/catalogo/saga-servicio-de-analisis-genetico-para-la-agricultura/

Vigouroux, Y., Mitchell, S., Matsuoka, Y., Hamblin, M., Kresovich, S., Smith, J. S. C., … Doebley, J. (2005). An Analysis of Genetic Diversity Across the Maize Genome Using Microsatellites. Genetics, 169 (3), 1617–1630.

Notas Biográficas

Los alumnos María Guadalupe Quintos Cortes y Octavio Francisco Fernández Lozada son estudiantes egresados de la Licenciatura en Ingeniería en Biociencias de la Escuela superior de Apan (ESAp), de la Universidad Autónoma del Estado de Hidalgo. Son miembros honorarios de la Asociación Mexicana de Profesores de Bioquímica A.C. Han presentado trabajos académicos y proyectos de investigación en Congresos Nacionales e Internacionales. La alumna María Guadalupe Quintos Cortes ha sido galardonada, en dos ocasiones consecutivas, con la Medalla Rosalind Franklin al Mérito Académico, durante sus estudios de Licenciatura en Ingeniería en Biociencias de la ESAp-UAEH.

Los Doctores César Daniel Petroli y Fernando Henrique Ribeiro Barrozo Toledo son Científicos pertenecientes al Programa de Recursos genéticos del CIMMYT, El Batán, Texcoco.

Los Dres. Martin Peralta Gil, Jaime Alioscha Cuervo Parra y la Dra. Teresa Romero Cortes son profesores investigadores en la Licenciatura en Ingeniería en Biociencias de la Escuela Superior de Apan de la UAEH, y pertenecen al Cuerpo Académico Biociencias Moleculares.

Agradecimientos

A la Maestra en Narrativa Annett Marianne Peralta Arteaga por la revisión del manuscrito. La Maestra es egresada de la Escuela de Escritores de Madrid, España.