Capítulo 6
Filogenética de Plantas Basada en Secuenciación de Nueva Generación

Marcela Avendaño González
Departamento de Biología del Centro de Ciencias Básicas
Universidad Autónoma de Aguascalientes

Resumen

La filogenética permite establecer las relaciones ancestro – descendiente que unen a los organismos. En las plantas, como en todos los organismos, este campo de la sistemática ha representado y sigue representando un trabajo continuo que resulta en hipótesis filogenéticas cambiantes y propositivas. En la actual era de la sistemática molecular, la comparación de secuencias de ADN o proteínas, se emplean para generar filogenias donde las diferencias entre las secuencias indican divergencia genética como resultado de una evolución molecular a través del tiempo. Más recientemente, la identificación de miles de genes nucleares en bajo número de copias ha comenzado a transformar las investigaciones de biología evolutiva y sistemática. En este capítulo se ejemplifica un estudio de genética de poblaciones

empleando secuenciación de nueva generación. La suficiente variación genética obtenida a partir de la secuenciación de los múltiples loci empleados permitió conocer un poco más sobre la historia evolutiva de la especie *Bouteloua gracilis*, mientras que los resultados obtenidos, demuestran el potencial y versatilidad que tiene este método de secuenciación en estudios de genética ecológica y evolutiva.

Antecedentes

El estudio de las plantas, así como de cualquier organismo vivo, implica siempre comenzar situándolo en un contexto filogenético. La filogenética permite establecer las relaciones ancestro – descendiente que unen a los organismos vivos o extintos, dichas relaciones existen entre poblaciones y especies. A partir de los estudios filogenéticos es que otras relaciones, como las ecológicas, toman mayor relevancia pues se logra un mayor y mejor entendimiento de los procesos evolutivos que las han moldeado.

La determinación de las relaciones filogenéticas de las plantas ha representado y sigue representando un trabajo continuo pues se han empleado diversos enfoques que resultan en hipótesis filogenéticas cambiantes y propositivas. Es así como, a pesar de la variedad de sistemas de clasificación, el mejor será aquel que refleje de manera exacta las relaciones ancestro - descendiente entre los organismos clasificados. Como resultado de lo anterior tenemos que ahora esté bien establecido, por ejemplo, que evolutivamente ocurrió una única transición de las plantas al ambiente terrestre (Embriofitas) desde un ancestro acuático (Charophytas). Entre especies, las relaciones filogenéticas se infieren a partir de caracteres particulares, preferentemente aquellos originados a partir de cambios evolutivos que son heredados (sinapomorfias) o novedades evolutivas (autapomorfias) que permiten evidenciar estas relaciones.

En la actual era de la sistemática molecular, las filogenias moleculares se generan al comparar secuencias de ADN (ácido desoxirribonucleico) o proteínas, donde las diferencias entre las secuencias indican divergencia genética como resultado de una evolución molecular a través del tiempo. Los diferentes genes acumulan mutaciones a tasas de cambio diferentes por lo que no todos los genes o macromoléculas resultan ser marcadores moleculares aptos para brindar información filogenética. Tradicionalmente en la sistemática

molecular se han empleado, por su utilidad, marcadores de ADN ribosomal (rDNA) y de plástidos (cpDNA) para la reconstrucción de filogenias de plantas a nivel de género, pues en estos existen en un gran número de copias en el genoma de estos organismos. Los marcadores más frecuentemente empleados del rDNA han sido los espaciadores transcritos (ITS y ETS *siglas en inglés*) y la región codificante de la subunidad 26S del ARN. Entre los marcadores más empleados del cpDNA tenemos a los genes codificantes para las proteínas *rcbL* y *matk* y sus espaciadores no codificantes.

En las décadas pasadas, las miradas hacia la sistemática molecular destacaron la importancia de esta rama biológica, el desarrollo de las técnicas de secuenciación de ácidos nucleicos ultrarrápida y la generación de grandes cantidades de datos en muy poco tiempo, sumado al desarrollo de herramientas matemáticas y estadísticas, han permitido explicar los procesos evolutivos más fácilmente. En años más recientes, la identificación de miles de genes nucleares en bajo número de copias ha comenzado a transformar las investigaciones de biología evolutiva y sistemática molecular.

Secuenciación masiva de datos

La Secuenciación de Nueva Generación (SNG o Next Generation Sequencing: NGS) permite la secuenciación paralela masiva de ADN, y es capaz de incrementar el volumen y velocidad de generación de datos, siendo una de sus mayores ventajas que es aplicable a cualquier organismo. La SNG se han aplicado exitosamente en filogenética y filogeografía, además ha permitido abordar preguntas sobre la diversidad del genoma, naturaleza y frecuencia de duplicación del genoma entre linajes de plantas. Las metodologías de SNG desarrolladas hasta ahora incluyen la secuenciación completa del genoma, secuenciación del transcriptoma, enriquecimiento dirigido (targeted enrichment ó sequence capture), secuenciación RAD (RAD-Seq), secuenciación por genotipificación (Genotyping-by-Sequencing: GBS), y escaneo del genoma con o sin enriquecimiento dirigido (Hyb-Seq).

En este capítulo nos enfocaremos en el método de *sitios de restricción asociados al ADN* que por sus siglas en inglés es conocido como RAD-seq (restriction site-associated DNA sequencing), usado especialmente para estudios de genómica ecológica, evolutiva y de conservación. La secuenciación RAD se

ha convertido en el método más empleado para descubrir polimorfismos de nucleótido simple (single nucleotide polimorphism: SNP) y para genotipificar cualquier organismo.

Método RAD-seq

Explicado de manera sencilla, en este método, el ADN genómico es digerido por una enzima de restricción, seguido de un corte mecánico para reducir la longitud de los fragmentos para su secuenciación. El resultado es un muestreo de un gran número de datos adyacentes a un gran número de sitios de restricción en todas las áreas del genoma (codificantes y no codificantes) que posteriormente son secuenciados en las plataformas de Illumina. El análisis de las secuencias requiere de *software* especiales que permiten generar los alineamientos empleando una secuencia de referencia o, si esta no está disponible, los fragmentos RAD pueden analizarse de *novo*. Las lecturas idénticas se agrupan como posibles alelos, al agrupar todas las secuencias que tienen pocas diferencias entre ellas, se pueden ubicar los SNP e indeles entre los alelos de un mismo locus y se corrigen los errores al comparar base con base en cada uno de los sitios. Los alelos homocigotos o heterocigotos reales serán aquellos que tengan altos números de lecturas y pocos errores.

El método RAD-seq ha adoptado múltiples metodologías que varían por ejemplo en el número de enzimas empleadas o la selección directa o indirecta del tamaño de los fragmentos, manteniendo aspectos básicos de la metodología original (Fig. 6.1). Hasta la fecha, este método de secuenciación de nueva generación ha sido el que mayor impacto ha tenido en filogenética y filogeografía por el control que tiene sobre los fragmentos resultantes de la digestión, además por la versatilidad que presenta para resolver problemas de investigación al identificar múltiples marcadores para propósitos de genotipificación poblacional a gran escala y a un bajo costo.

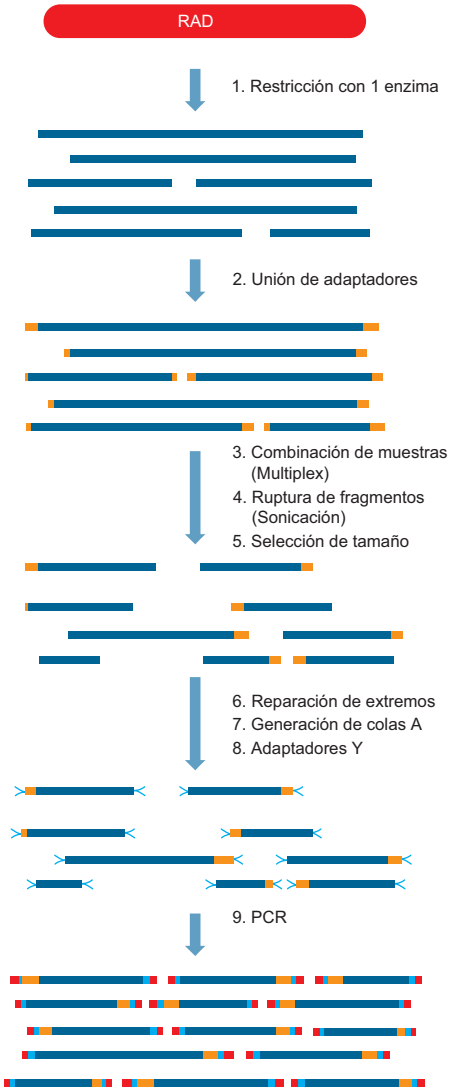

Figura 6.1. Ilustración paso a paso de la elaboración de la biblioteca genómica para secuenciación RAD (Elaborada a partir de Andrews *et al.* 2016).

Filogenética de poblaciones de *Bouteloua gracilis* (Chloridoideae: Poaceae)

En los pastizales de Norteamérica se distribuye el pasto navajita azul *(Bouteloua gracilis* (Kunth) Lag. ex Griffiths), un pasto perenne nativo con metabolismo C4. Su distribución va desde el sur de Canadá, centro y oeste de Estados Unidos hasta México. *Bouteloua gracilis* ha sido muy estudiada debido a su amplia distribución, su importancia en el pastoreo histórico y su rol clave en la ecología de las comunidades de pastizales de Norteamérica. A lo largo de su distribución ha sido reportado como ecológica, morfológica y genéticamente variable. Lo anterior se relaciona a que se puede encontrar en climas áridos y semiáridos, es tolerante al frío y a la sequía, así como a los suelos alcalinos. A pesar de la gran variación reportada para la especie, solo un estudio se había enfocado en investigar la variación genética poblacional de la especie *B. gracilis*, reportando que, para cuatro poblaciones del centro de México, zona que representa la distribución sureña de la especie, la variación genética dentro las poblaciones era mayor que la interpoblacional. Sin embargo, a partir de los estudios filogenéticos previos, lo que se podía inferir sobre la variación genética de esta especie era muy poco dado que los estudios realizados emplearon marcadores (ITS, rpL32-trnL and rps16-trnK) que resultaron invariantes a nivel infraespecífico. Ante esta problemática, resultaba necesario abordar el estudio de la filogenética y la genética de poblaciones de *B. gracilis* a partir de datos generados con una metodología más potente de SNG como lo es la secuenciación RAD.

Muestreo y análisis de datos

A partir del muestreo de varios individuos de 42 poblaciones diferentes a lo largo del área de distribución de la especie se pudieron muestrear 33807 loci del genoma nuclear, que contenían 164045 SNP. A partir de estos datos, la inferencia filogenética generada con el método de Máxima Verosimilitud mostró una topología bien resuelta, en la que se observa que las muestras de México y EUA comparten ancestría (Fig. 6.2). El clado que incluye a todas las muestras de México (A, 99% BS) se divide en dos clados altamente soportados (A1& A2, 93% cada uno); el clado A2 contiene muestras del centro y norte de

México incluyendo a la población de Sonora como la población más tempranamente divergente. El clado hermano A1 incluye muestras de Chihuahua y Texas. En la otra rama, el clado de EUA (B, 100% BS) se separa en dos grupos que coinciden con su origen geográfico: la gran planicie (B1) y la región montañosa del oeste (B2, 100% BS).

La suficiente variación genética obtenida de los múltiples loci empleados permitió conocer un poco más sobre la historia evolutiva de la especie pues al observar dentro del clado de México (A) las relaciones revelan que las muestras del centro de México son las ultimas en divergir (Actipan de Morelos, Puebla in A5) y se originan de ancestros que debieron estar en el Norte de México (clados A1, A3. A4 y el clado tempranamente divergente A5), indicando que el origen de las poblaciones mexicanas ancestrales fue en el desierto chihuahuense. De hecho, algunos autores han concluido que el centro de origen y diversificación de las especies del género Bouteloua Lag. ocurrió en las áreas abiertas del norte de México, además existe evidencia de que algunas especies incluyendo a *B. gracilis*, migraron hacia el norte desde México.

En este estudio, la secuenciación RAD permitió obtener datos suficientes para establecer: a) las frecuencias alélicas de cada individuo y agruparlos de acuerdo a las mismas (Fig. 6.3, *izquierda*), b) calcular un índice de flujo genético y el modelo de cómo éste está ocurriendo, c) los haplotipos y su diversidad así como su diversificación, permitiendo esto evidenciar el movimiento de dispersión de la especie (Fig. 6.3, *centro* y *derecha*), d) la diversidad alélica dentro de las poblaciones, dando indicios de que algunas poblaciones se encuentran aisladas, e) los modelos de expansión demográfica para determinar cuáles son las poblaciones que tienen un crecimiento estable y cuáles han sido las ultimas en expandirse. Los resultados enlistados anteriormente demuestran el potencial y versatilidad que tiene este método de secuenciación en estudios de genética ecológica y evolutiva.

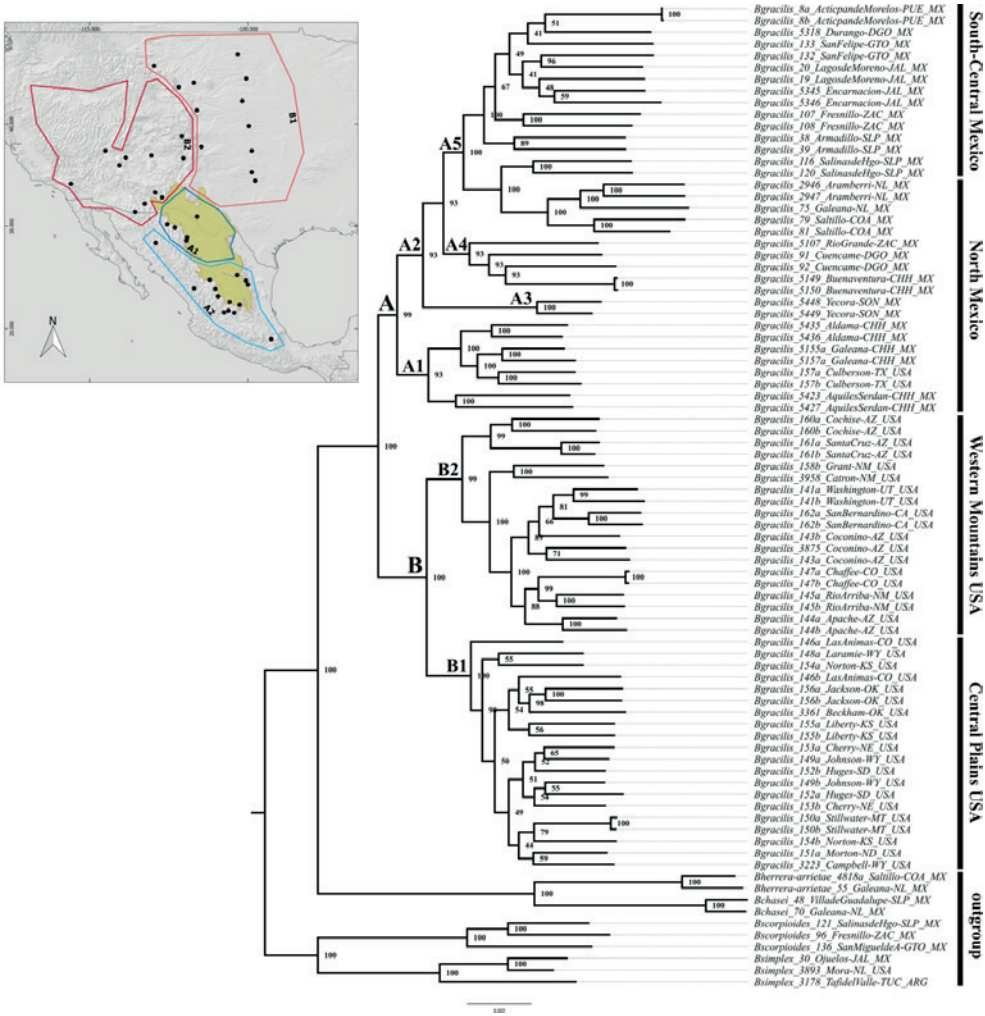

Figura 6.2. Inferencia filogenética molecular con el método de máxima verosimilitud empleando 33807 loci y 164045 SNPs de *B. gracilis*. Generado empleando IQTREE con 1000 réplicas de bootstrap ultrarrápido, los valores de soporte se muestran en los nodos; Log-likelihood: -7270163.539, longitud total del árbol: 0.296. Letras en los nodos indican el nombre asignado al clado, las etiquetas de la derecha señalan la ubicación geográfica de las muestras. Mapa de la ubicación geográfica de los clados A1 y A2 (México), B1 y B2 (EUA), Desierto chihuahuense en amarillo.

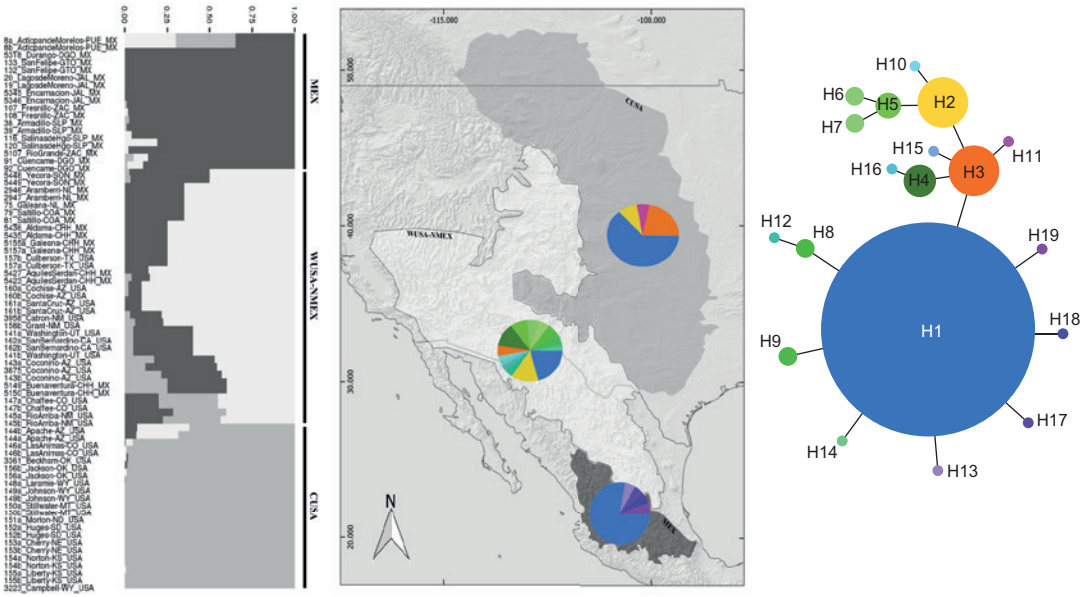

Figura 6.3. *Izquierda*: gráfico del análisis Bayesiano de agrupamiento con Structure, se muestra la probabilidad (barras) de cada muestra de ser asignada a uno de los tres grupos genéticos (K=3) de acuerdo con sus frecuencias alélicas. *Centro*: localización geográfica de los grupos: MEX- gris oscuro, CUSA- gris, WUSA-NMEX- gris claro. Los límites de los grupos se establecieron con base en las regiones ecoterrestres (L2 y L3); además se muestran los gráficos de pastel con la distribución de haplotipos para cada grupo genético, los colores corresponden a los observados en la red de haplotipos. *Derecha*: Red de haplotipos donde el tamaño del círculo corresponde a el número de muestras con el haplotipo.

Protocolo

Colecta en campo: se visitaron 42 localidades en donde se colectaron de uno a cinco individuos por población, de los cuales se obtuvieron muestras de hoja en fresco que se colocaron en bolsas con sílica gel para su deshidratación y conservación.

Extracción de ADN: se pesaron 0.3 g de tejido de hoja deshidratada. En un tubo de 1.5 ml el tejido se homogenizó empleando el Tissue Lyser II- QIAGEN con perlas de acero inoxidable. Se aplicó un ciclo (o dos) de disrupción

a 30/s por 2 a 3 min. Una vez homogeneizado el tejido, el ADN se extrajo siguiendo el protocolo CTAB 2X (Doyle, 1987). El ADN de cada muestra fue cuantificado usando el fluorómetro Qbit (kit dsDNA Assay) de Invitrogen y todas las muestras se ajustaron a la misma concentración molar (20 ng/µL).

Biblioteca genómica y secuenciación: la biblioteca genómica se preparó siguiendo el protocolo de Etter *et al.* (2011) empleando la enzima de alta fidelidad *SbfI* (New England Biolabs) para digerir 300 ng de ADN de cada muestra. Para este estudio se procesaron 94 muestras simultáneamente para las cuales los códigos de identificación de sus fragmentos fueron de 6 a 10 pb de longitud difiriendo por lo menos en dos bases. Después de restringir el ADN de cada muestra, se juntaron todas para formar la biblioteca genómica la cual se sometió a sonicación en un Covaris S220 para producir fragmentos de aproximadamente 400 pb, posteriormente se seleccionaron los fragmentos de entre 360 y 600 pb empleando cartuchos de agarosa al 1.5% (Pippin Prep; Sage Science, Beverly, MA). La amplificación final para el enriquecimiento de la biblioteca genómica se dividió en reacciones múltiples de 25µl, realizando 18 ciclos para cada amplificación. El tamaño, calidad y cantidad de la biblioteca genómica se evaluó con el Agilent 2100 Bioanalyzer (Agilent Technologies) empleando el DNA 1000 Kit. La secuenciación de la biblioteca se obtuvo mediante lecturas sencillas de 150 pb de longitud en una línea de la plataforma Illumina NextSeq500 (IIGB Genomic Core de la Universidad de California, Riverside).

Filtrado de calidad y recuperación de Polimorfismos de Nucleótido Simple (SNP): las secuencias crudas se procesaron empleando el *software* ipyrad v.0.7.17 (Eaton, 2014). El alineamiento se realizó de *novo* (sin secuencia de referencia). A cada lectura de secuenciación se le cortaron 6 pb al inicio para eliminar el remanente del sitio de restricción de la enzima y se estableció que las lecturas quedaran a una longitud máxima de 145 pb. Los parámetros de alineamiento empleados fueron los de fábrica estableciendo un umbral de alineamiento de 90%. El alineamiento final se realizó solo con 74 muestras de poblaciones de *B. gracilis* y algunas muestras de las especies hermanas para formar el grupo externo (*B. chasei*, *B. herrera-arrietae*, *B. scorpioides* y *B. simplex*).

Inferencia filogenética: La inferencia de máxima verosimilitud (ML) se realizó en el programa IQTREE (Nguyen *et al.*, 2015) especially for maximum-likelihood (ML. Se realizaron 1000 réplicas de bootstrap ultrarápido. Para las inferencias se empleó el modelo de evolución molecular GTR+gamma empleado anteriormente en estudios con RAD-seq (Eaton *et al.*, 2017). Los

árboles resultantes se enraizaron y visualizaron con el programa FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Glosario

Filogenia: estudio de las relaciones evolutivas entre organismos empleando diagramas representativos parecidos a árboles.

Divergencia: referente a la evolución divergente en la cual dos poblaciones estrechamente relacionadas por exposición a presiones selectivas diferentes terminan por diferenciarse una de la otra.

Clado: grupo de organismos (linajes) que tienen un ancestro en común.

Grupo externo: taxón o grupo en un árbol filogenético conocido por haber divergido más tempranamente que el resto de los taxa en el árbol y se emplea para determinar la posición de la raíz del árbol.

Filogeografía: sub-rama de la biogeografía histórica que consiste en el análisis espacial de los linajes que se fundamenta en el estudio de los procesos geográficos históricos que podrían ser responsables de las distribuciones contemporáneas de individuos.

Genotipificación: determinar el genotipo

Restricción: cuando algunas enzimas reconocen una o varias secuencias donde cortan el ADN.

Polimórfico: que tiene o puede tener muchas formas, en el ADN los polimorfismos incluyen las diferencias en nucleótidos o en secuencias, al compararlos.

Indel: contracción de "inserción o deleción", en referencia a los dos tipos de mutaciones genéticas.

De novo: se refiere a la secuenciación de un genoma nuevo sin una secuencia de referencia disponible para su alineamiento.

Perenne: planta que vive más de dos años.

Haplotipo: conjunto de variaciones del ADN, o polimorfismos, que tienden a ser heredados juntos.

Bibliografía

Aguado-Santacruz, G. A., Leyva-López, N. E., Pérez-Márquez, K. I., García-Moya, E., Arredondo-Moreno, J. T., & Martínez-Soriano, J. P. (2004). Genetic variability of Bouteloua gracilis populations differing in forage production at the southernmost part of the North American Gramine-tum. Plant Ecology, 170(2), 287-299.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17(2), 81-92.

Avendaño-González, M., Morales-Domínguez, J. F., & Siqueiros-Delgado, M. E. (2019). Genetic structure, phylogeography, and migration routes of Bouteloua gracilis (Kunth) Lag. Ex Griffiths (Poaceae: Chloridoideae). Molecular phylogenetics and evolution, 134, 50-60.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE, 3(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. Briefings in Functional Genomics, 9(5-6), 416-423. https://doi.org/10.1093/bfgp/elq031

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem bull, 19, 11-15.

Eaton, D. A. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, btu121.

Egan, A. N., Schlueter, J., & Spooner, D. M. (2012). Applications of next-generation sequencing in plant biology. American Journal of Botany, 99(2), 175-185. https://doi.org/10.3732/ajb.1200020

Gould, F. W. (1980). The genus Bouteloua (Poaceae). Annals of the Missouri Botanical Garden, 348-416.

Karol, K. G., McCourt, R. M., Cimino, M. T., & Delwiche, C. F. (2001). The closest living relatives of land plants. Science (New York, N.Y.), 294(5550), 2351-2353. https://doi.org/10.1126/science.1065156

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating

Maximum-Likelihood Phylogenies. Molecular Biology and Evolution, 32(1), 268-274. https://doi.org/10.1093/molbev/msu300

Soltis, D. E., & Soltis, P. S. (1998). Choosing an approach and an appropriate gene for phylogenetic analysis. En Molecular systematics of plants II (pp. 1-42). Springer.

Zimmer, E. A., & Wen, J. (2013). Reprint of: Using nuclear gene data for plant phylogenetics: Progress and prospects. Molecular Phylogenetics and Evolution, 66(2), 539-550. https://doi.org/10.1016/j.ympev.2013.01.005

Zimmer, E. A., & Wen, J. (2015). Using nuclear gene data for plant phylogenetics: Progress and prospects II. Next-gen approaches. Journal of Systematics and Evolution, 53(5), 371-379. https://doi.org/10.1111/jse.12174