Capítulo 9

Herramientas informáticas para el estudio de los recursos genéticos

Lorena Jacqueline Gómez-Godínez[1*], *Francisco Fabián Calvillo Aguilar*[1], *Arturo Vera Ponce de León*[2], *Colin K. Khoury3, María Victoria Díaz*[3]
1 Centro Nacional de Recursos Genéticos, INIFAP. 2 Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 3 International Center for Tropical Agriculture (CIAT)
gomez.lorena@inifap.gob.mx

Introducción.

En el CNRG además de la conservación de los recursos genéticos, se realizan estudios de nivel informático que incluye la descripción y modelación de la distribución geográfica de las especies conservadas, así como el análisis del genoma y diversidad. A través de herramientas informáticas, es posible determinar el contenido genómico y como este cambia o responde a diversas condiciones ambientales. Por otro lado, con el apoyo de la modelación espacial y los sistemas de información geográfica se ha realizado la planeación y evaluación en las colectas, para determinar la distribución actual y potencial de las especies conservadas. Además, sirven para conocer las condiciones ambientales del suelo, clima y relieve, en las que dichas colectas se desarrollan.

El continúo avance y desarrollo de tecnologías informáticas ha permitido realizar estudios cada vez más complejos. La ciencia de datos (*data science*), ha permitido el modelado y caracterización de los recursos genéticos. Así la bioinformática y los sistemas de información geográfica, son herramientas que contribuyen al conocimiento y conservación de la biodiversidad.

Surgimiento de las técnicas de secuenciación masiva: del capilar al nanopore.

La secuenciación del primer genoma (*Haemophilus influenza* 1.8 Mb) se realizó utilizando tecnología de secuenciación capilar desarrollada por Frederick Sanger en la decada de los 70 (Fleischmann *et al.,* 1995). Al inicio del siglo XXI, se planteó el proyecto de secuenciar organismos eucariontes comenzando por la mosca de la fruta en el año 2000, el primer bosquejo del genoma humano en el 2001 y el ratón en el 2002 (Goodwin *et al*., 2016). Usando este tipo de tecnología capilar (de aquí en adelante secuenciación Sanger), producir un millón de bases de DNA costaba aproximadamente $1,500 USD, y se necesitaba de un día entero para la corrida del experimento (Escobar-Zepeda *et al.,* 2015). Es por lo que el primer bosquejo ("*draft*") del genoma humano (3 billones de bases o 3 Gb), se calcula que tuvo un costo de billones de dólares y se necesitaron años de horas máquina para refinarlo. En 2005 apareció formalmente el primer equipo de secuenciación de nueva generación, o "NGS" por sus siglas en inglés *next generation sequencing* (Escobar-Zepeda *et al.,* 2015).

La secuenciación por síntesis de Roche 454, implementó una química de secuenciación basada en la detección y medición de luz emitida por una luciferasa al momento de que se hidroliza ATP, por la incorporación de un nucleótido en las hebras recién sintetizadas de DNA. El uso de ATP en la reacción confiere el nombre de pirosecuenciación a la técnica (Buermans y Den Dunnen 2014). Esta tecnología hizo posible obtener un genoma humano, con una cobertura 7.4x, en tan solo dos meses y con un costo 10 veces menor en comparación con la secuenciación Sanger (Escobar-Zepeda *et al.,* 2015). Aunque esta tecnología revolucionó la forma de secuenciar, en el año 2016 la empresa Roche anunció la descontinuación de los equipos 454 (Buermans y Den Dunnen 2014).

La segunda NGS en comercializarse fue utilizando la tecnología Illumina, esta difiere de 454 ya que adoptó la secuenciación por síntesis usando nucleótidos fluorescentes removibles usados en la polimerización del DNA (Buermans y Den Dunnen 2014). Actualmente, y con el precipitado avance en las nuevas tecnologías de secuenciación, se tienen plataformas con la capacidad de generar entre 120-1500 Gb por corrida y una media del tamaño de secuencia de 150 bp. En estos se encuentra además una gama compacta para laboratorios llamada MiSeq, la cual es pequeña en tamaño,

y se pueden lograr de 0.3 Gb a 15 Gb de datos en un tiempo que puede ser de 4 horas (Buermans y Den Dunnen 2014).

Además, se ha desarrollado una nueva forma de secuenciar DNA sin requerir un paso previo de amplificación por PCR. Este tipo de tecnología se denomina secuenciación a partir de molécula única (*single-molecule sequencing*). Este tipo de secuenciación implementa una polimerasa directamente embebida en una celda de vidrio o un poro dependiendo el modelo de las plataformas. Usando estas tecnologías se puede obtener hasta 100 Gb de información (aprox. 33 veces el genoma humano) a un costo aproximado de $ 1,000 dólares (Ameur *et al*., 2019). El uso de las NGS ha revolucionado el estudio de diversos organismos y su información genética (genomas y metagenomas), así como la respuesta de cambio en los genes a diversos estímulos (transcriptomas y metatranscriptomas).

Herramientas bioinformáticas para el estudio de la diversidad de los recursos genéticos.

Mediante las NGS podemos responder algunas preguntas, como ¿que se encuentra en determinado ambiente? y ¿qué función biológica está realizando?, información básica en la ecología microbiana, etc. Podemos resolver estas cuestiones con dos enfoques diferentes: 1) secuenciación de genes marcadores (*metabarcoding*) o 2) secuenciación de todo el contenido genético (*whole-genome sequencing* WGS). La secuenciación de uno o varios genes blanco, comunes en los organismos de la comunidad son utilizados para la estrategia del *metabarcoding*. Para esto se utilizan iniciadores, dirigidos al gen de interés, generalmente se utiliza el gen 16S rRNA para bacterias y arqueas y la secuencias intergénicas del gen 18S rRNA (ITS) para hongos. Estos métodos son rápidos y rentables lo que permite obtener una visión global de los organismos presentes en una comunidad microbiana en poco tiempo y a bajo costo (Knight *et al.,* 2018; Calle, 2019).

El segundo enfoque, es un método que permite capturar todos los genomas presentes en cierta comunidad, para esto se secuencia el DNA total de la muestra y bioinformáticamente se separan los genomas de cada individuo presente en la comunidad (*binning*). Al poder analizar no sólo un marcador sino todos los genes presentes en los distintos organismos de la comunidad (bacterias, arqueas y hongos), se obtiene no

sólo una visión taxonómica si no funcional de los miembros del ecosistema secuenciado (Knight *et al*., 2018; Calle, 2019).

Análisis taxonómico.

Una vez obtenidos los datos por NGS se realiza el análisis bioinformático y spon diversos protocolos los que se han desarrollado para conocer la composición taxonómica de los miembros de una comunidad microbiana. Sin embargo, todos siguen algunos lineamientos generales para el procesamiento y análisis de los datos. El primer paso para el análisis es la remoción de secuencias de baja calidad y artefactos de secuenciación (adaptadores). Una vez obtenidas secuencias de alta calidad, por lo regular se procede a la agrupación o *clustering* de secuencias y su asignación a unidades operacionales taxonómicas (OTUs) o variantes de secuencias ribosomales (ASVs) (Callahan *et. al*., 2017). Esto puede hacerse directamente de los datos obtenidos por el método de *metabarcoding* o bien extrayendo las secuencias de los genes ribosomales (u otro gene taxonómicamente informativo) de los datos producidos por *WGS* (Rausch *et al.,* 2019). Finalmente, cada una de estas OTUs son asignadas a un nivel taxonómico determinado utilizando búsquedas en bases de datos por ejemplo Silva, RPD o GreenGenes y emparentándolas con su homólogo en dicha base (Balvočiūtė *et al.,* 2017). Todo esto finaliza en la obtención de una matriz de relación OTU, taxonomía y muestra, comúnmente conocida como matriz de OTUs (Callahan *et al.,* 2017). Programas informáticos como Mothur (Schloss *et al.,* 2009) y Qiime 1 y 2 (Quantitative Insights Into Microbial Ecology) (Caporaso *et al.,* 2010), y Phyloseq (McMurdie y Holmes, 2013) son softwares que siguen protocolos comunes para el análisis de datos. Estos protocolos consisten en (1) limpiar y filtrar las secuencias obtenidas, (2) asignar secuencias a OTUs o ASVs y (3) describir la diversidad (α y β), composición y diferencias o similitudes entre las comunidades. Con todo esto, las diferencias en la abundancia relativa de los miembros presentes en las comunidades microbianas pueden ser calculadas por medio de análisis multivariados de agrupación y minería de datos (ejemplo análisis de componentes principales o PERMANOVA). Estos análisis pueden ser después visualizados por graficas de ordenación tipo análisis de componentes principales (*PCA)* o *NMDS* (Rausch *et al.,* 2019) (Figura1).

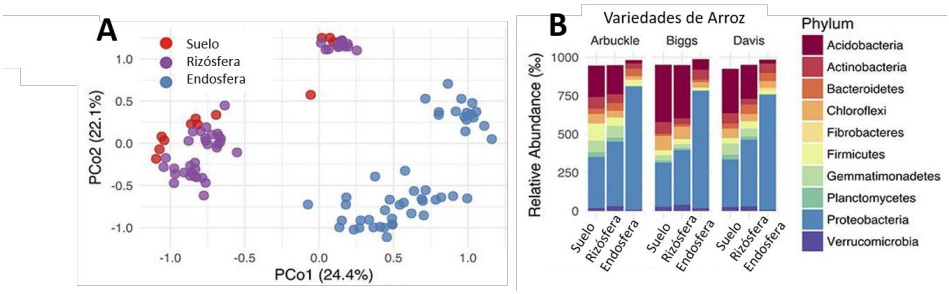

Figura 1. Representaciones graficas de información obtenida de un análisis de diversidad (metabarcoding). A) Análisis de componentes principales (PCA), mostrando diversos grupos de comunidades microbianas presentes en diferentes ambientes del suelo. B) Histogramas de abundancia de phylum en cada compartimento (tomado y modificado de Edwars *et al.,* 2015).

Análisis taxonómico: ¿Que se encuentra en determinado ambiente, y qué función biológica está realizando?

El hecho de obtener la secuencia del genoma completo (o cuasi completo) de la mayoría de organismo presentes en una muestra vía WGS permite conocer la taxonomía y posible nicho que ocupan los miembros de la comunidad. Los primeros pasos de la mayoría de los protocolos descritos para analizar datos tipo WSG son similares a los utilizados en el *metabarcoding*. Esto es el filtrado por calidad de secuencias y remoción de adaptadores. Programas como *fast*QC, software desarrollado para proporcionar una visión general de las secuencias obtenidas, ayuda a visualizar de manera gráfica la calidad y presencia de artefactos en las secuencias (Andrews, 2010). El filtrado por remoción de lecturas (secuencias tipo NGS) con baja calidad puede llevarse a cabo utilizando paqueterías bioinformáticas tipo trimmomatic (Bolger *et al.*, 2014). Este tipo de software permite remover utilizado un corte en valor *Phred* de calidad además de la remoción de artefactos como quimeras y adaptadores presentes en las lecturas.

Para poder recuperar la información genética de los organismos presentes en las muestras secuenciadas por WGS, es necesario ensamblar los genomas. El ensamble consiste en recuperar secuencias de mayor longitud, denominadas “contigs” en los cuales se podrá recuperar las regiones codificantes de los genes. Programas como SPAdes (Bankevich *et al*., 2012), IDBA (Peng *et al.,* 2010) o MEGAHIT (Li *et al.,* 2014), permiten la recuperación de ensambles genómicos. Todas estas herramientas trabajan bajo el mismo principio, el uso de gráficas de Brujin y

fragmentación de lecturas en *k-mer* para la extensión de secuencias cortas (Bankevich *et al.,* 2012). Con este tipo de estrategias es posible recuperar genomas casi completos o completos a partir de muestras metagenómicas.

Al recuperar secuencias con una longitud mayor a 1,000 nucleótidos (tamaño promedio de un gen bacteriano) podemos empezar a predecir unidades codificantes (genes) presentes en el ambiente. Genes taxonómicamente informativos como secuencias ribosomales, genes mitocondriales o genes de copia única pueden ser utilizados para la asignación taxonómica de los individuos presentes en la comunidad (Wu *et al*., 2008; Darling *et al.,* 2014). MetaPhlAn1 y 2 (Truong *et al.,* 2015), Kraken (Wood and Salzberg, 2014), AMPHORA (Wu *et al*., 2008) y PhiloSift (Darling *et al.,* 2014), son ejemplos de herramientas que proveen información y clasificación taxonómica a partir de búsqueda de genes de copia única en los genomas. Esta información puede ser representada en forma de árboles filogenéticos o matrices de comparación.

Una vez obtenida la clasificación taxonómica de los miembros presentes en la comunidad, es posible separar cada uno de los genomas utilizando una técnica denominada *binning*. El *binning* consiste en clasificar y separar por similitud aquellas secuencias comunes pertenecientes al genoma de un organismo en particular. En donde un *bin* representara un genoma (Wu *et al.,* 2016). Rasgos como cobertura de secuencias, porcentaje y frecuencia de *k*-meros, frecuencia de tetranucleótidos, así como motivos particulares en las secuencias (Kislyuk *et al.,* 2009, Strous *et al.,* 2012), son utilizados para agrupar aquellos *contigs* similares y separarlos en genomas individuales. Programas informáticos como MaxBin (Wu *et al.,* 2016), MetaBat (Kang *et al.,* 2015) y CONCOCT (Alneberg *et al.,* 2014) emplean este tipo de estrategias para aislar genomas presentes en comunidades metagenómicas.

Con los genomas individuales de cada uno de los organismos presentes en el ambiente podemos empezar a clasificar el contenido de genes presentes en ellos, proceso denominado anotación genómica. Programas como Prokka (Seemann, 2014), KOALA (Kanehisa *et al.,* 2016), y el Prokaryotic Genome Annotation Pipeline del National Center for Biotechnology Information NCBI, (Tatusova *et al.,* 2016) son capaces de predecir las secuencias codificantes y anotar los genes presentes en ellas

a partir de búsquedas de homología con bases de datos (ejemplo Uniprot (Apweiler *et al.,* 2004), Pfam (Bateman *et al.,* 2002), KEGG (Kanehisa y Goto, 2000).

Una vez ensamblados los *contigs* se puede utilizar MetaPhlAn2, que se encarga de la asignación taxonómica, utilizando regiones genómicas y marcadores moleculares de copia única (Truong *et al.,* 2015), para esto también es posible utilizar Kraken, el cual asigna etiquetas taxonómicas a secuencias de DNA, mucho más rápido y eficiente que MetaPhlAn2 (Wood y Salzberg, 2014). Una vez realizada la clasificación taxonómica se pueden utilizar otras herramientas para predecir y anotar genes en vías metabólicas, prokka (Seemann, 2014) e InterProScan (Mitchell *et al.,* 2019) son herramientas que se utiliza con estos fines (Figura 2).

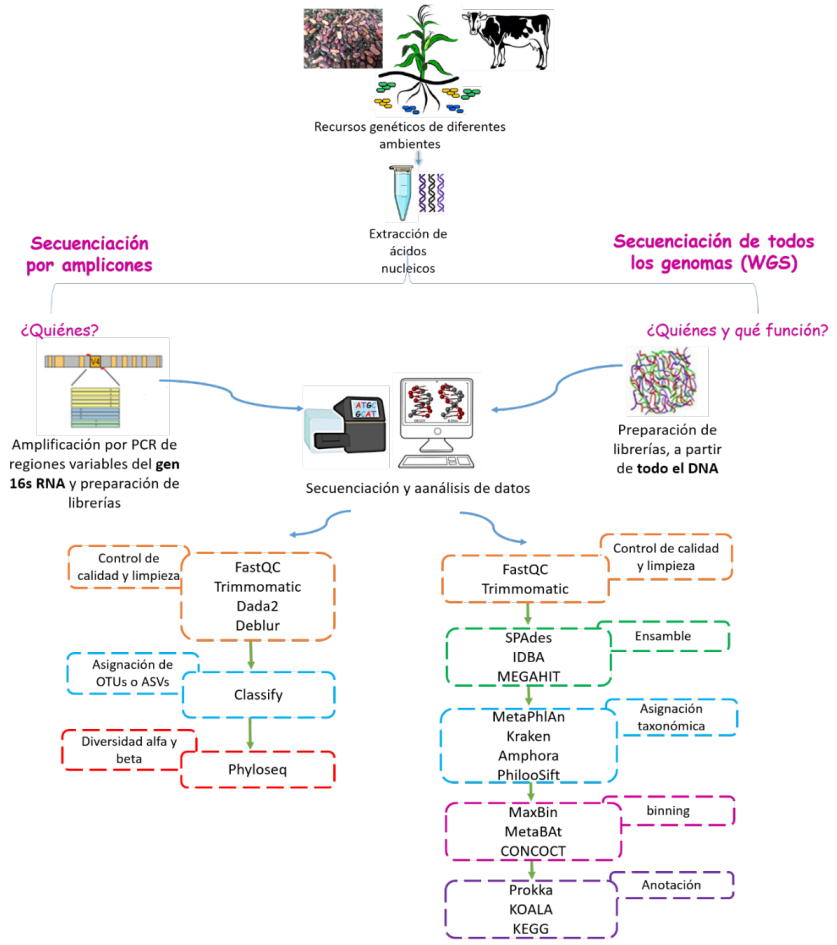

Figura 2. Flujo de trabajo que representa las diferencias de secuenciación y análisis bioinformático entre metabarcoding (amplicones) y WGS.

Herramientas bioinformáticas para el estudio de la expresión diferencial de genes en los recursos genéticos.

Toda la información de un organismo se encuentra contenida en el genoma, sin embargo, no toda se expresa al mismo tiempo. Por ejemplo, una planta sometida a estrés por sequía, expresará ciertos genes para contrarrestar los efectos del estrés, que no expresaría bajo condiciones normales. A la expresión de ciertos genes o transcritos (moléculas de RNA) en determinadas condiciones como estrés, enfermedades o tipos celulares se le conoce como transcriptoma. Se considera que el

transcriptoma es dinámico debido a que este depende de las condiciones de crecimiento del organismo.

Existen dos técnicas para el análisis de los transcritos, por un lado están los microarrays/microarreglos, que cuantifican un conjunto de secuencias predeterminadas o conocidas, y la secuenciación de RNA (RNA-Seq), que utiliza NGS para capturar todas las secuencias (Figura 3).

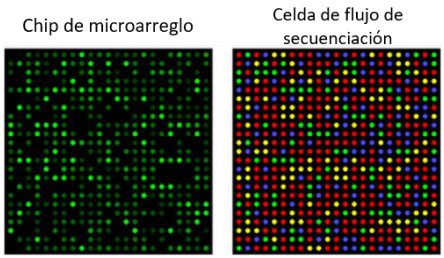

Figura 3. Microarreglos y RNA-Seq se basan en el análisis de imágenes de diferentes maneras. En un chip de microarreglos, cada punto en un chip es una sonda de oligonucleótidos definida, y la intensidad de fluorescencia detecta la abundancia de una secuencia hibridada y específica. En la secuenciación de alto rendimiento, se secuencia un nucleótido a la vez, el color en cada ronda indica el nucleótido añadido.

Existen diversas herramientas informáticas para el análisis de transcriptomas, este capítulo se centrará en el análisis de RNA-*seq*, debido a que es la técnica más novedosa y actualmente más usada, para la identificación de los niveles de expresión, es los diferentes recursos genéticos. El análisis de secuencias de RNA-*seq*, se puede dividir en cuatro etapas: control de calidad, alineación, cuantificación y expresión diferencial. Para el control de calidad, generalmente se usa FastQC (Andrews, 2010), el cual permite visualizar la calidad de las secuencias, este software hace un llamado de base por base y evalúa su calidad, asignándole un puntaje según la escala, permite ver el contenido de GC, y la cantidad de secuencias obtenidas. Posterior a la visualización de calidad de las secuencias se realiza una limpieza de datos, la cual consiste en quitar aquellas bases de mala calidad, secuencias sobrerrepresentadas, quimeras, y los adaptadores. Las herramientas más utilizadas para estos fines son Trimmomatic (Bolger *et al.,* 2014), Trimgalore (Krueger, 2015) y FastX (FastX, 2015). Una vez limpias las secuencias se procede al alineamiento, elk cual consiste en empalmar las secuencias obtenidas por NGS contra un genoma de referencia o *de novo*. Dentro de los programas utilizados para el alineamiento se encuentran Burrow-Wheeler Aligner (BWA) (Li y Durbin 2009), Bowtie (Langmead *et al.,* 2009), TopHat (Trapnell *et al*., 2009), y GSNAP (Wu y Nacu 2010), la elección de estos dependerá del

genoma de referencia utilizado ya que TopHat y GSNAP aumenta la probabilidad de identificar nuevas transcripciones generadas por *splicing* alternativos (Yang y Kim, 2015).

Una vez alineadas las secuencias al genoma de referencia, se realiza la cuantificación de secuencias alineadas por *contig*/gen, para esto es posible utilizar programas como RSEM (Li y Dewey, 2011), que proporciona estimaciones a nivel de genes e isoformas como salida primaria al calcular estimaciones de abundancia de máxima verosimilitud basadas en el algoritmo de Expectación-Maximización (EM) después de leer el mapeo. RSEM puede dar estas cuantificaciones normalizadas por millón (TPM), así como también admite la visualización de la alineación y la profundidad de lectura mediante un navegador genómico como el navegador genómico Santa Cruz (UCSC) de la Universidad de California. Cufflinks es el el programa informático más utilizado y estima la abundancia, mediante la abundancia de probabilidad máxima, basada en la cobertura de la transcripción. Las abundancias se informan por kilobase por millón de fragmentos mapeados (FPKM o RPKM). Finalmente, para el análisis de expresión diferencial, se han desarrollado varias paqueterías de software que incluyen EdgeR (Robinson *et al*. 2010), DESeq (Anders y Huber, 2010), que ocupan modelos binomiales negativos y NOIseq (Tarazona *et al*. 2015), que son no paramétricos. Los programas anteriores adoptan uno o más de los varios métodos de normalización disponibles (recuento total, cuartil superior, mediana, normalización DESeq, media recortada de valores M, normalización de cuantil y RPKM) para corregir los sesgos que pueden aparecer entre las muestras (profundidad de secuenciación) o dentro de la muestra (longitud del gen y contenido de GC) (Yang y Kim, 2015). Estos programas permiten encontrar aquellos genes involucrados, con alguna respuesta a cierta condición, estrés o ambiente al cual se encuentren sometidos, los diferentes recursos genéticos.

Sistemas de información geográfica.
Datos pasaporte, una primera etapa a los mapas de distribución.
Actualmente se ha difundido en gran medida el uso de tecnologías como drones, principalmente para fotografía artística y técnica (Figura 4), servidores de mapas en teléfonos móviles para ubicar una dirección, incluso el cine 3D o la llamada realidad aumentada; todo lo mencionado es tecnología cuyos orígenes se remontan a los años 50's con el desarrollo

de los hoy llamados Sistemas de Información Geográfica (SIG), impulsados con fines científicos y militares para el estudio del territorio de las naciones (Olaya, 2014).

Figura 4. Vuelo de dron y georreferenciación de puntos de control con estación total para un estudio de hidrología.

Los SIG son una herramienta informática que consiste en software capaz de almacenar, capturar, verificar, gestionar, analizar, transformar, mostrar y transferir datos especialmente referidos a la tierra (georreferenciados), con la finalidad de realizar diversos análisis de carácter territorial (SGM, 2019).

Los SIG se conforman de cinco componentes (FAO, 2006):

1. Hardware: Equipo de cómputo.
2. Software: Programas que permiten el manejo y visualización de las bases de datos.
3. Datos: Conjunto de información con carácter espacial.
4. Personas: Especialistas y técnicos que diseñan, operan y mantienen el SIG.
5. Métodos: Modelos y prácticas realizadas en el análisis y mantenimiento de los datos.

La función principal de un SIG es servir como una herramienta para la toma de decisiones, ya que permite mediante la visualización de mapas, responder a preguntas sobre la localización, condición, cambio histórico, modelación y simulación de un fenómeno (INEGI, 2014). Estos mapas se utilizan entre otras cosas para (ESRI, 2013):

1. Conocer y compartir información.
2. Compilar y mantener datos.

3. Organizar y visualizar.
4. Mediante geoprocesos, generar nueva información.

Debido a los avances en el desarrollo de tecnologías de la información y cómputo, la aplicación de los SIG es cada vez más diversa, por ejemplo, en ámbitos como el productivo, científico, cultural y de gobernanza (McCall, 2003; Siabato, 2018). Una de las áreas de interés en la actividad científica, es el uso de los SIG en la conservación de los recursos genéticos, generando mapas de distribución geográfica de las especies con fines de resguardo.

Un mapa de distribución geográfica conocida, muestra los sitios o regiones donde se tiene registro de la presencia de una especie (Figura 5), planta o animal, y puede ser utilizado para la planeación de acciones de conservación *in situ*, así como para la distribución de los sitios de colecta en la conservación *ex situ* (Maxted *et al.,* 2013). Realizar un mapa de distribución geográfica conocida, requiere que los registros tengan las coordenadas geográficas donde se determinó la presencia de la especie (avistamiento/colecta); es precisamente cuando los datos pasaporte tienen especial relevancia, ya que en ellos se registra un lote de información respecto al origen de las colectas de germoplasma, incluidas las coordenadas, localidad, municipio o referencias geográficas.

En el CNRG, el proceso de ingreso de accesiones de germoplasma solicita que se incluyan los datos pasaporte de cada accesión; estos datos cuentan con una estructura con base en acuerdos internacionales, como el Tratado Internacional Sobre los Recursos Filogenéticos para la Alimentación y la Agricultura, ITPGRFA, por sus siglas en inglés, de la Organización Mundial para la Agricultura y la Alimentación (FAO) Biodiversidad Internacional (Alercia *et al.,* 2015).

Figura 5. Mapa de distribución geográfica conocida de *Echinocactus playacanthus*, especie en *peligro de extinción* según la NOM-059-SEMARNAT 2010. Fuente: (adaptado de CONABIO, 2014).

Ecogeografía, una caracterización multidisciplinaria.

El primer resultado de realizar un mapa de distribución geográfica conocida, consiste en agrupar y visualizar los sitios y regiones donde está presente una especie. Además, es posible conocer las características ambientales de estas regiones, bióticas, abióticas y antrópicas; como altitud, temperatura, precipitación, suelo, vegetación dominante, uso del suelo y topografía, entre muchas otras. En ecología, al conjunto de estas características ambientales que determinan las regiones dónde puede o no estar presente una especie se le denomina hábitat, y puede modelarse tanto para las condiciones actuales como para las históricas y futuras, por ejemplo, para condiciones de cambio climático (Figura 6).

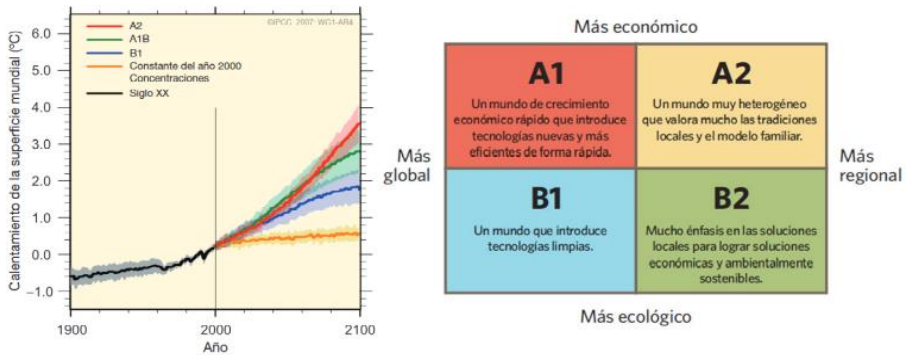

Figura 6. Escenarios de cambio climático propuestos por el Panel Intergubernamental del Cambio Climático (IPCC). Fuente: (adaptado de IPCC, 2007).

De este modo, al estudio de los factores que determinan el hábitat y el ambiente al cual un individuo, población o especie se ha adaptado, se le conoce como análisis ecogeográfico, y es importante dado que estas adaptaciones tienen su representación en la información genética de cada individuo. El análisis ecogeográfico requiere de la recopilación y síntesis de información geográfica, taxonómica y genética; sus resultados son de carácter predictivo, pueden utilizarse para formular y priorizar proyectos de recolección y conservación de especies (Castañeda *et al.,* 2011).

La caracterización ecogeográfica de una región o país se realiza mediante el compilado y análisis de información espacial edáfica, bioclimática, geofísica, biótica y antrópica, a través de un SIG; y se representa con mapas, que al combinarse reflejan los diferentes escenarios de adaptación ambiental. Además, con la cartografía y datos espaciales, se pueden realizar análisis de distribución potencial, riesgo e índices de biodiversidad (Figura 7).

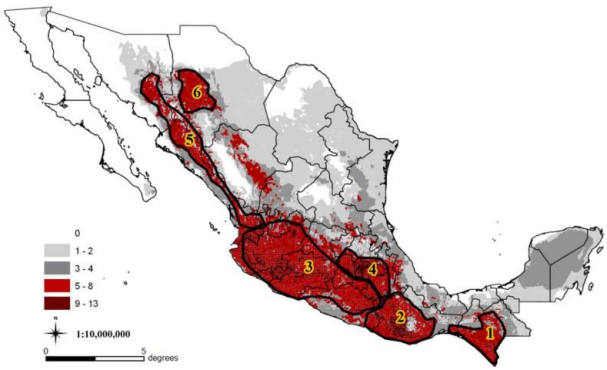

Figura 7. Riqueza de razas de maíz en México (número de razas que ocurren dentro de cada celda ó pixel). Fuente: (Perales y Golicher, 2014).

Según Maxted *et al.,* (2013), el uso de los SIG en el análisis ecogeográfico permite:

1. Caracterización ambiental de los sitios de colecta.
2. Optimización de las colectas de germoplasma orientada a una mayor representatividad de la diversidad genética.
3. Interpretación de patrones geográficos, ecológicos y taxonómicos.

4. Representatividad y sesgo ecogeográfico en colectas existentes (análisis de vacíos).
5. Determinación de sitios para establecer reservas genéticas.
6. Impacto del cambio climático en las poblaciones naturales.

La caracterización ecogeográfica de los recursos genéticos, es una herramienta que permite determinar el rango adaptativo de las especies, y con ello determinar los factores ambientales más determinantes. El valor genético de estos rasgos puede utilizarse en el mejoramiento genético de especies de interés agrícola, forestal y pecuario. Por otra parte, en el caso de cultivos, la regeneración del germoplasma puede realizarse en los sitios más acordes a las condiciones ecogeográficas nativas, para garantizar un mayor éxito de la regeneración y reducir la erosión genética (Parra *et al.,* 2012).

La planeación: el análisis de vacíos y la caracterización predictiva.
Anteriormente se mencionó que derivado del análisis ecogeográfico y los mapas de distribución geográfica, la información permite determinar la calidad o sesgo de las colectas de germoplasma con base en la diversidad ecogeográfica no representada, a esta determinación se le conoce como análisis de vacíos, y puede aplicarse en la planeación de las jornadas de colecta para mejorar la representatividad ecogeográfica de la conservación *ex situ* (Parra *et al.,* 2012), incluso para determinar zonas de interés por su diversidad en recursos genéticos. Ejemplo de ello, es el análisis de vacíos realizado por Contreras *et al.,* (2019) para parientes silvestres de cultivos en México, donde de forma complementaria propuso áreas para el establecimiento de reservas genéticas (Figura 8).

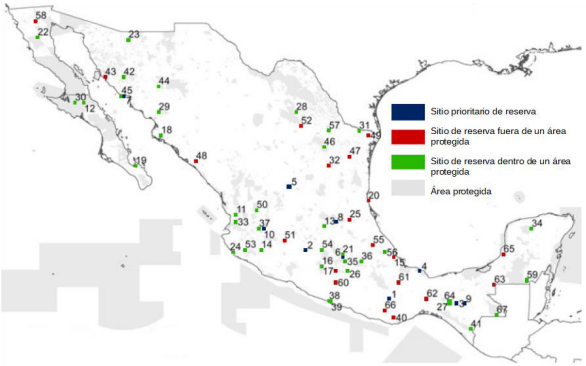

Figura 8. Sitios propuestos para el establecimiento de reservas genéticas de parientes silvestres de cultivos en México. Fuente: (adaptado de Contreras *et al.,* 2019).

Determinar los sitios de colecta con el objetivo de obtener rasgos adaptativos para realizar mejoramiento genético, es una herramienta de selección denominada *caracterización predictiva*, y ha generado metodologías y protocolos internacionales como el Focused Identification of Germoplasm Strategy (FIGS). La metodología FIGS consiste en la búsqueda de rasgos específicos, con base en la relación fenotipo-genotipo, utilizando herramientas geográficas para determinar la presión del medio ambiente sobre los individuos; FIGS asume la probabilidad de que el germoplasma refleje las adaptaciones de las muestras colectadas (Bari *et al.,* 2012). Esta herramienta se utiliza en la búsqueda de rasgos como la tolerancia a sequía, plagas y enfermedades (Maxted *et al.,* 2013).

Caso de estudio: Análisis de vacíos geográficos y ecológicos de conservación para plantas silvestres.

Para abordar las necesidades persistentes de indicadores para la conservación de la biodiversidad y recursos genéticos, particularmente con respecto a la evaluación eficiente de la conservación de la diversidad genética, dentro y entre los taxones, se han desarrollado análisis de vacíos de conservación.

Khoury *et al*. (2019b) ofreció una metodología de análisis de brechas aplicada a los sistemas de conservación *ex situ* e *in situ* para plantas silvestres. El método proporcionó una aproximación de la distribución de la diversidad genética de una especie de planta silvestre, utilizando el

alcance de la variación ecogeográfica (es decir, geográfica y ecológica) en su rango nativo, predicho como un *proxy*, que se ha demostrado que es un sustituto efectivo (Hanson *et al.,* 2017, Hoban *et al*., 2018), facilitando la planificación de la conservación a pesar de las brechas persistentes en los datos genéticos a nivel de población (Balmford *et al*., 2005, Hanson *et al*., 2017; Hoban *et al*., 2020).

La variación ecogeográfica evidente a partir de un análisis de la ubicación de los sitios de recolección de muestras, salvaguardadas en repositorios de conservación (es decir, bancos de genes, semillas, y jardines botánicos) (*ex situ*), y en el rango de distribución de las especies dentro de áreas naturales protegidas (*in situ*), se midió contra la variación ecogeográfica encontrada dentro del rango nativo general predicho de la especie. El proceso identificó brechas geográficas y ecológicas en la protección actual, que pueden representar puntos focales para acciones futuras. Posteriormente, se priorizó los taxa para realizar más esfuerzos de conservación, y se combinaron los puntajes de múltiples taxones para proporcionar indicadores a diferentes escalas, locales, nacionales, regionales y globales (Khoury *et al*., 2019a). La metodología se basó en datos y herramientas de acceso abierto (Khoury *et al*., 2019a) y fue reportada en formatos de resumen fácilmente comprensibles, al tiempo que proporcionaron información específica por taxón útil para acciones de conservación. Además, cuando se aplica repetidamente a lo largo del tiempo, los resultados podrían usarse para visualizar el progreso hacia el objetivo de la conservación integral, incluida la determinación de cuándo se ha alcanzado ese objetivo.

El análisis de vacíos de conservación se basó en métodos desarrollados durante la última década, primero para medir el estado de conservación de los taxones en repositorios y para ayudar a guiar los esfuerzos de recolección adicionales, destinados a construir colecciones *ex situ* más diversas (Ramirez-Villegas *et al*., 2010, Castañeda-Álvarez *et al*., 2016). Recientemente, el enfoque se adaptó para medir la representación dentro de las áreas naturales protegidas (Khoury *et al*., 2019b,c,d, Lebeda *et al*., 2019; Mezghani *et al*., 2019). Tales estudios se han llevado a cabo con mayor frecuencia en una variedad de especies dentro de un género, aunque también se han aplicado a nivel nacional (Norton *et al*., 2017; Khoury *et al*., 2020) y global para grupos específicos de plantas (Castañeda-Álvarez *et al*., 2016; Khoury *et al*., 2019b).

En México, un resultado interesante de este análisis de vacíos para parientes silvestres, es el de *Cucurbita argyrosperma* C. Huber subsp. *sororia* (LH. Bailey) L. Merrick & D. M. Bates, el progenitor de C. *argyrosperma* C. Huber subsp. *argyrosperma* (calabaza pipiana), que fue domesticada en el sur de México unos 7000 años pb. (Antes del presente, *Before Present*) (Smith, 2006). Este taxón anual mesofítico se distribuye a lo largo de las costas tropicales del Pacífico y del Golfo de México, desde el estado de Sonora en México hasta el sur de Nicaragua, y ha sido reconocido como una fuente de resistencia a varios virus de importancia económica en el cultivo (Khoury *et al*., 2019d). Usando el método de análisis de vacíos, se descubrió que las 59 ocurrencias de germoplasma estaban relativamente bien distribuidas en el rango geográfico y ecológico de los taxones, aunque tal vez faltan representaciones en las partes más al norte y más al sur de su rango (Fig. 9A). La comparación de su distribución prevista con las áreas protegidas oficialmente reconocidas, encontró que las áreas principales de su distribución geográfica no están representadas en áreas protegidas, mientras que la mayoría de su variación ecológica está posiblemente representada (Figura 9). A la especie se le asignó un valor de acuerdo al grado de conservación, en un rango de 0 a 100, 46.8 y 31.8, para la conservación *ex situ* e *in situ*, respectivamente; interpretándose 100 como una conservación integral y 0 limitada (Khoury *et al*. 2019d).

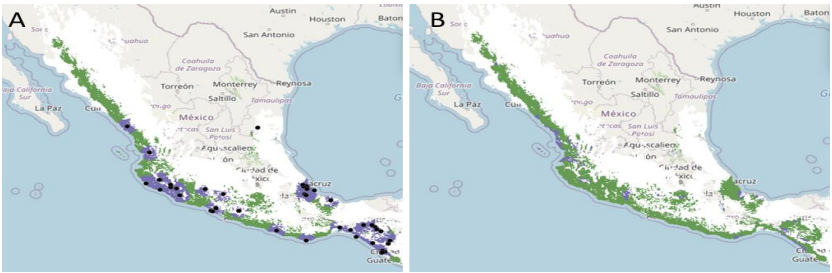

Figura 9. Análisis de vacíos de conservación para *Cucurbita argyrosperma* C. Huber subsp. *sororia* (L. H. Bailey) L. Merrick y D. M. Bates, que muestra la representación geográfica del pariente silvestre en la conservación *ex situ* (A) y en las áreas protegidas *in situ* (B). El verde representa el rango predicho del taxón basado en información de ocurrencia, datos climáticos y topográficos. El púrpura representa áreas consideradas como conservadas, con base en colecciones *ex situ* anteriores (A) y en áreas protegidas existentes (B). Datos y mapas derivado de Khoury *et al*., (2019d).

Análisis de vacíos geográficos y ecológicos de conservación *ex situ* para plantas cultivadas

El grado de representación de las variedades tradicionales de los agricultores (variedades locales) en la conservación *ex situ* es poco conocido, en parte debido a la falta de métodos que puedan identificar los determinantes antropogénicos y ambientales de sus distribuciones geográficas. Ramírez-Villegas *et al*., (2020) desarrollaron un nuevo marco de modelado espacial y de análisis de vacíos de conservación *ex situ* para variedades locales de cultivos, utilizando frijol (*Phaseolus vulgaris* L.) como estudio de caso.

El modelado de cada una de las variedades locales incluyó cinco pasos principales: (1) determinar grupos relevantes de variedades locales utilizando literatura, de manera que, al probar modelos estadísticos de clasificación, fuera posible encontrar que existe una diferencia significativa entre ellos según las características ambientales y socioeconómicas de su distribución geográfica; (2) modelar la distribución geográfica potencial de estos grupos utilizando datos de presencia (local), contemplando predictores ambientales y socioeconómicos; (3) calcular puntajes de vacíos geográficos y ambientales para las colecciones actuales en bancos de germoplasma; (4) mapear los vacíos de conservación *ex situ*; y (5) compilar aportes de expertos (Ramírez-Villegas *et al*., 2020).

La metodología, logró distinguir las distribuciones (Figura 10A) y las brechas de conservación para los dos principales grupos genéticos de frijol (andino y mesoamericano), y los resultados se alinearon bien con la opinión de expertos. Se encontró que ambos grupos genéticos estaban relativamente bien conservados en bancos de germoplasma, respecto a sus distribuciones geográficas previstas, con colecciones *ex situ* que representaban el 78.5% del grupo andino y el 98.2% del mesoamericano. Las prioridades de recolección de variedades locales mesoamericanas se concentran en varias zonas de México, Belice y Guatemala (Figura 10B).

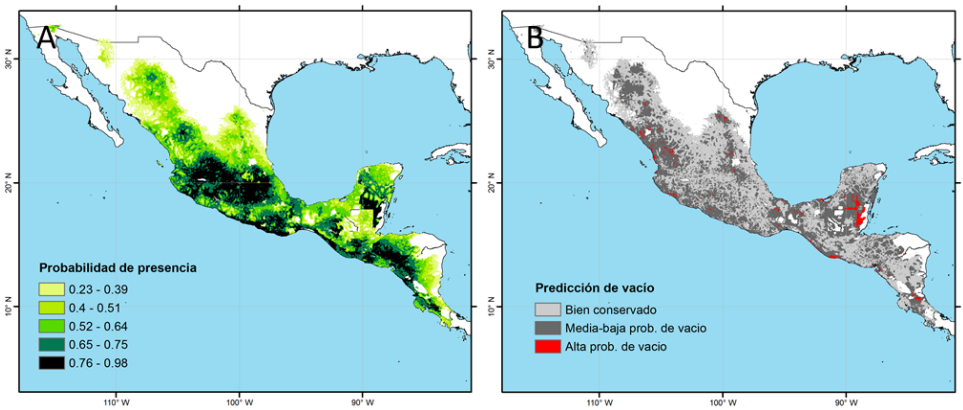

Figura 10. Distribución potencial (A) y análisis de vacíos geográficos de conservación *ex situ* (B) para variedades tradicionales del grupo genético mesoamericano de frijol (*Phaseolus vulgaris* L.). Datos y mapas derivado de Ramírez-Villegas *et al*., (2020).

Literatura consultada.

- Alercia A, Diulgheroff, S; Mackay, M. 2015. Descriptores de pasaporte para cultivos múltiples FAO/BIOVERSITY V.2.1 [MCPD V.2.1] - diciembre 2015. Organización de las Naciones Unidas para la Alimentación y la Agricultura y Bioversity International, 12 p. https://cgspace.cgiar.org/handle/10568/76136.Consultado 21 Abril 2020
- Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Quince C. Binning metagenomic contigs by coverage and composition. Nature Methods. 2014;11:1144-1146.
- Ameur A, Kloosterman WP, Hestand MS. Single-molecule sequencing: towards clinical applications. Trends in biotechnology. 2019;37:72-85.
- Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;11:R106.
- Andrews S. FastQC: A quality control tool for high throughput sequence data. Babraham Bioinforma. 2010;48-9843:00144-3.
- Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Martin MJ. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2004;32:D115-D119.
- Balmford A, Crane P, Dobson A, Green RE, Mace GM. The 2010 challenge: data availability, information needs and extraterrestrial

insights. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005;360:221–228.
- Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics. 2017;18:114.
- Bankevich, A. *et al*. SPAdes: a new genome assembly algorithm and itsapplications to single-cell sequencing. Journal Comput. Biol. 2012;19:455–477.
- Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In: Berger B. (eds) Research in Computational Molecular Biology. RECOMB 2010. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2010;6044.
- Bari A, Street K, Mackay M, Dag F, De-Pauw E, Amri A. Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables. Genetics Resourse Crop Evolution. 2012;59:1465–148.
- Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Research. 2002;30:276-280.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-2120.
- Buermans HPJ, Den Dunnen JT. Next generation sequencing technology: advances and applications. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. 2014;1842:1932-1941.
- Callahan B, McMurdie P, Holmes S. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME Journal. 2017;11:2639–2643.
- Calle M. Statistical Analysis of Metagenomics Data. Genomics Informayics. 2019;17:e6.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al*. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 2010;7:335–336.
- Castañeda NP, Vincent HA, Kell SP, Eastwood RJ, Maxted, N. Ecogeographic surveys. In Guarino L, Ramanatha Rao V, Goldberg E (editors). Collecting Plant Genetic Diversity: Technical Guidelines. Bioversity International, Rome. 2011. Disponible en: https://cropgenebank.sgrp.cgiar.org/index.phpoption=com_content&view=article&id=679.

- Castañeda-Álvarez NP, Khoury CK, Achicanoy HA, Bernau V, Dempewolf H, Eastwood RJ, Guarino L, Harker RH, Jarvis A, *et al*. Global conservation priorities for crop wild relatives. Nature Plants 2016;2:16022.
- CONABIO. Distribución conocida de biznaga tonel grande (*Echinocactus platyacanthus*). Distribución conocida. Catálogo de metadatos geográficos. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. México. 2014. Disponibl en: http://geoportal.conabio.gob.mx/metadatos/doc/html/echplat_gcagw.html.
- Contreras TA, Cortés CM, Costich D, Rico AM, Magos BJ, Maxted N. Diversity and conservation priorities of crop wild relatives in Mexico. Plant Genetic Resources: Characterization and Utilization. 2019;17: 140-150.
- Darling AE, Jospin G, Lowe E, Matsen IV FA, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243.
- Edwards J, Johnson C, Santos-Medellín C, Lurie E, Podishetty NK, Bhatnagar S, Sundaresan V. Structure, variation, and assembly of the root-associated microbiomes of rice. Proceedings of the National Academy of Sciences of US. 2015;112:E911-E920.
- Escobar-Zepeda A, Vera-Ponce de Leon A, Sanchez-Flores A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Frontiers in Genetics. 2015;6:348.
- ESRI. Introducción a SIG. ArcGIS Resources. ESRI Company. 2013.
- FAO. Sistemas de Información Geográfica (SIG) en Salud Animal. Componentes y Funciones de los SIG. Organización de las Naciones Unidas para la Agricultura y la Alimentación. 2006. Disponible en: http://www.fao.org/tempref/GI/Reserved/FTP_FaoRlc/old/prior/segalim/animal/sig/intro/compo.htm7
- FASTX-Toolkit. Cold Spring Harbor: Cold Spring Harbor Laboratory; 2015. Disponible en: http://hannonlab.cshl.edu/fastx_toolkit/
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Merrick JM. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;269:496-512.
- Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17:333.

- Hanson JO, Rhodes JR, Riginos C, Fuller RA. Environmental and geographic variables are effective surrogates for genetic variation in conservation planning. Proceedings of the National Academy of Sciences of USA. 2017;114:12755–12760.
- Hoban S, Callicrate T, Clark J, Deans S, Dosmann M, Fant J, Gailing O, *et al*. Taxonomic similarity does not predict necessary sample size for ex situ conservation: a comparison among five genera. Proc. R. Soc. B. 2020;287:20200102.
- Hoban S, Kallow S, Trivedi C. Implementing a new approach to effective conservation of genetic diversity, with ash (Fraxinus excelsior) in the UK as a case study. Biological Conservation. 2018;225:10–21.
- INEGI, 2014. Sistema de Información Geográfica. Servicio Profesional de Carrera – Documentación. Instituto Nacional de Estadística, Geografía e Informática. México. Disponible en: https://www.inegi.org.mx/inegi/spc/doc/internet/sistemainformacionngeografica.pdf.
- IPCC, 2007: Summary for Policymakers. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 2007
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000;28:27-30.
- Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. Journal of Molecular Biology. 2016;4284:726-731.
- Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
- Khoury CK, Amariles D, Soto JS, Diaz MV, Sotelo S, Sosa CC, Ramírez-Villegas J, *et al*. Data for the calculation of an indicator of the comprehensiveness of conservation of useful wild plants. Data in Brief. 2019a;22:90-97.
- Khoury CK, Amariles D, Soto JS, Diaz MV, Sotelo S, Sosa CC, Ramírez-Villegas J, *et al*. Comprehensiveness of conservation of useful wild

plants: an operational indicator for biodiversity and sustainable development targets. Ecological Indicators. 2019b;98:420-429.
- Khoury CK, Carver D, Barchenger DW, Barboza G, van Zonneweld M, Jarret R, Bohs L, *et al*. Modeled distributions and conservation status of the wild relatives of chile peppers (*Capsicum* L). Diversity and Distributions. 2019c;26:209-225.
- Khoury, C. K., Carver, D., Greene, S. L., Williams, K. A., Achicanoy, H. A., León, B., Wiersema, J. H. and Frances, A. Crop wild relatives of the United States require urgent conservation action. Proceedings National Academic Science of USA. 2020;117:33351-33357.
- Khoury CK, Carver D, Kates HR, Achicanoy HA, van Zonneweld M, Thomas E, Heinitz C, *et al*. Distributions, conservation status, and abiotic stress tolerance potential of wild cucurbits (*Cucurbita L.*). Plants, People, Planet 2019d;2:269-283.
- Kislyuk A, Bhatnagar S, Dushoff J, Weitz JS. Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics. 2009;10:316.
- Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Melnik AV. Best practices for analysing microbiomes. Nature Reviews Microbiology. 2018;16:410-422.
- Krueger F. Trimgalore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. 2015;516:517.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;10:R25.
- Lebeda A, Křístková E, Kitner M, Majeský Ľ, Doležalová I, Khoury CK, Widrlechner MP, *et al*. Research gaps and challenges in the conservation and use of North American wild lettuce germplasm. Crop Science. 2019;59:2337-2356.
- Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–1760.
- Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics

assembly via succinct de Bruijn graph. Bioinformatics. 2014;31:1674–1676.

- Maxted N, Magos BJ. Kell S. Resource book for preparation of national conservation plans for crop wild relatives and landraces. University of Birmingham. United Kingdom. 2013. Disponible: http://www.fao.org/fileadmin/templates/agphome/documents/PGR/PubPGR/ResourceBook/TEXT_ALL_2511.pdf.
- McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 2013;8:e61217.
- Mezghani N, Khoury CK, Carver D, Achicanoy HA, Simon P, Martínez FF, Spooner D. Distributions and conservation status of carrot wild relatives in Tunisia: a case study in the Western Mediterranean Basin. Crop Science. 2019;59:1–12.
- McCall KM. Seeking good governance in participatory-GIS: a review of processes and governance dimensions in applying GIS to participatory spatial planning, Habitat International. 2003;27:4.
- Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Gough J. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research. 2019;47:D351-D360.
- NOM-059-SEMARNAT 2010. Protección ambiental-Especies nativas de México de flora y fauna silvestre-Categorías de riesgo y especificaciones para su inclusión, exclusión o cambio-Lista de especies en riesgo. Norma Oficial Mexicana. Diario Oficial de la Federación, Noviembre 30, 2010. México.
- Norton SL, Khoury CK, Sosa CC, Castañeda-Álvarez NP, Achicanoy HA, Sotelo S. Priorities for enhancing the *ex situ* conservation and use of Australian crop wild relatives. Australian Journal of Botany. 2017;65:638-645.
- Olaya V. Sistemas de información geográfica. Proyecto OSGeo. España. 2014. Repositporio libre: https://github.com/volaya/libro-sig
- Parra M, Iriondo JM, Torres E. Review. Applications of ecogeography and geographic information systems in conservation and utilization of plant genetic resources. Spanish Journal of Agricultural Research. 2012;10:419.
- Perales H, Golicher D. Mapping the Diversity of Maize Races in Mexico. PLOS One. 2014;9: e114657.

- Ramírez-Villegas J, Khoury CK, Achicanoy HA, Mendez AC, Diaz MV, Sosa CC, Debouck DG, Kehel Z, Guarino L. A gap analysis modeling framework to prioritize collecting for *ex situ* conservation of crop landraces. Diversity and Distributions. 2020;26(6):730-742.
- Ramírez-Villegas J, Khoury C, Jarvis A, Debouck DG, Guarino L. A gap analysis methodology for collecting crop genepools: a case study with *Phaseolus* beans. PLoS One. 2010;5:e13497.
- Rausch P, Rühlemann M, Hermes BM. *et al*. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome. 2019;7:133.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–140.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister E, *et al*. Introducing mothur: open-source. platform-independent, community-supported software for describing and comparing microbial communities. Applied Environmental Microbiol. 2009;75:7537–7541.
- Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068-2069.
- SGM. 2019. Sistemas de Información Geográfica. Museo Virtual. Servicio Geológico Mexicano. México. Disponible en: https://www.sgm.gob.mx/Web/MuseoVirtual/SIG/Introduccion-SIG.html
- Siabato Willington. Sobre la evolución de la información geográfica: las bodas de oro de los sig. Cuadernos de Geografía: Revista Colombiana de Geografía. 2018;27:1-9.
- Smith B. D. Eastern North America as an independent center of plant domestication. Proceedings of the National Academy of Sciences of USA. 2006;103:12223–12228.
- Strous M, Kraft B, Bisdorf R, Tegetmeyer H. The binning of metagenomic contigs for microbial physiology of mixed cultures. Frontiers in Microbiology. 2012;3:410.
- Tarazona S, Furio-Tari P, Turra D, Pietro AD, Nueda MJ, Ferrer A, *et al*. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43:e140.

- Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Ostell J. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Research. 2016;44:6614-6624.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–1111.
- Truong DT. *et al*. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 2015;12:902–903.
- Wood DE, Salzberg SL, Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15:R46
- Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–881.
- Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biology. 2008;9:R151.
- Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605-607.
- Yang IS, Kim S. Analysis of whole transcriptome sequencing data: workflow and software. Genomics & Informatics. 2015;13:119.