UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE DOCTORADO EN CIENCIAS BIOMÉDICAS

INSTITUTO DE ECOLOGÍA

Patrones macro y microevolutivos en calabazas (*Cucurbita*): ritmos de evolución genómica y patrones de genómica poblacional durante la domesticación

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
JOSUÉ BARRERA REDONDO

DIRECTOR DE TESIS
DR. LUIS ENRIQUE EGUIARTE FRUNS
INSTITUTO DE ECOLOGÍA, UNAM

COMITÉ TUTOR
DR. LUIS DAVID ALCARAZ PERAZA
FACULTAD DE CIENCIAS, UNAM
DR. ENRIQUE IBARRA LACLETTE
INSTITUTO DE ECOLOGÍA A.C. (INECOL)

CIUDAD UNIVERSITARIA, CDMX. AGOSTO 2020

El jurado de examen doctoral estuvo constituido por:

Dr. Diego Cortez Quezada	Presidente
Dr. Luis Enrique Eguiarte Fruns	Secretario
Dra. Eria Rebollar Caudillo	Vocal
Dra. Alicia Mastretta Yanes	Vocal
Dra. Araxi Urrutia Odabachian	Vocal

AGRADECIMIENTOS

Agradezco a la Universidad Nacional Autónoma de México (UNAM) por permitirme a mí y a miles de mexicanos realizar nuestros estudios superiores con la más alta calidad, de manera pública y gratuita. Gracias al Doctorado en Ciencias Biomédicas por permitirme realizar mis estudios de doctorado en el Instituto de Ecología de la UNAM y al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme los recursos económicos necesarios para mi manutención durante la realización de esta tesis (beca número 583146). Agradezco al Programa de Apoyo a los Estudios de Posgrado (PAEP) y al proyecto Ecos Nord M12-A03 “Genómica de poblaciones: estudios en el maíz silvestre, el teosinte (*Zea mays* ssp. *parviglumis* y *Zea mays* ssp. *mexicana*)” por el financiamiento otorgado para realizar estancias de investigación en México y en Francia, respectivamente. La investigación realizada en esta tesis fue financiada por la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) KE004 ‘‘Diversidad genética de las especies de *Cucurbita* en México e hibridación entre plantas genéticamente modificadas y especies silvestres de *Cucurbita*”, CONABIO PE001 ‘‘Diversidad genética de las especies de *Cucurbita* en México. Fase II. Genómica evolutiva y de poblaciones, recursos genéticos y domesticación’’, CONACyT Investigación Científica Básica 2011.167826 ‘‘Genómica de poblaciones: estudios en el maíz silvestre, el teosinte (*Zea mays* ssp. *parviglumis* y *Zea mays* ssp. *mexicana*)’’ y CONACyT Problemas Nacionales 247730. Durante el trascurso de esta tesis se realizaron análisis en clúster de cómputo de la CONABIO, el cual fue parcialmente financiado por la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) por medio del fondo ‘‘Contribución de la Biodiversidad para el Cambio Climático’’ otorgado a la CONABIO. Agradezco a Rodrigo García Herrera, jefe del Departamento de Cómputo Científico del Laboratorio Nacional de Ciencias de la Sostenibilidad (LANCIS), Instituto de Ecología, UNAM, por correr la infraestructura de computación de alto rendimiento del LANCIS con la que se realizaron parte de los análisis de esta tesis. Se reconoce el soporte técnico del MC. Emanuel Villafán y a los recursos de cómputo de alto rendimiento que el Instituto de Ecología A.C. (INECOL) hizo disponibles para realizar parte de los resultados descritos en esta tesis. Gracias a la Red de Estudios Moleculares Avanzados del INECOL por permitirme realizar diversas estancias en sus instalaciones y permitirme formar parte de su vida académica. Agradezco al Dr. Diego Cortez Quezada, la Dra. Eria Rebollar Caudillo, la Dra. Alicia Mastretta Yanes y la Dra. Araxi Urrutia Odabachian por sus valiosas observaciones y comentarios, los cuales mejoraron sustancialmente la calidad de esta tesis. Agradezco al Dr. Luis E. Eguiarte y a la Dra. Valeria Souza por permitirme realizar mis estudios de posgrado en su laboratorio, además de brindarme su apoyo académico y personal durante estos años. Agradezco al Dr. Rafael Lira por todas sus contribuciones y apoyo durante estos 5 años. Agradezco también al Dr. Daniel Piñero por todo el apoyo y la colaboración que hemos realizado estos años. Agradezco al Dr. Santiago Ramírez Barahona por haber sido mi tutor en la licenciatura y permitirme desarrollar mi potencial como investigador. Agradezco mucho a la Dra. Alejandra Vázquez Lobo por toda la ayuda y el apoyo que me brindó durante el inicio de mi doctorado, y por convencerme de realizar este proyecto. Agradezco a la Dra. Erika Aguirre Planter y a la Dra. Laura Espinosa Asuar por su ayuda y apoyo técnico. Agradezco a la Sra. Silvia Barrientos por su ayuda y por su amistad. Agradezco a los miembros del Laboratorio de Evolución Molecular y Experimental por formar parte de mi vida académica y personal durante estos 5 años. Le agradezco al Dr. Enrique Ibarra Laclette su apoyo y su amistad a lo largo de estos años, la cual quiero que dure toda la vida. Agradezco mucho a mi familia por brindarme su cariño y apoyo durante toda la vida, sé que cuento con ustedes y quiero que sepan que ustedes cuentan conmigo. Le agradezco a Jazmín muchas cosas: te agradezco por formar parte de mi vida, por soportarme y ayudarme en los momentos difíciles, por mejorar los momentos felices estando juntos, por las pláticas triviales y profundas, por darme tu cariño, por permitirme darte el mío y por hacerme crecer como persona.

INDICE

RESUMEN

El género *Cucurbita* (las calabazas) es un grupo de 13 especies de plantas que son nativas del continente americano. Son plantas herbáceas con una historia evolutiva que incluye una duplicación completa del genoma y al menos seis eventos independientes de domesticación. Estudiar los patrones de evolución molecular derivados de ambos procesos es útil para entender los procesos evolutivos a distintas escalas. Los resultados que se presentan en esta tesis se enfocan en la especie *Cucurbita argyrosperma* (la calabaza pipiana), un cultivo tradicional mexicano del cual se consume principalmente su semilla. La tesis presenta una revisión de los métodos modernos para estudiar el proceso de domesticación en plantas y animales. Se secuenció, ensambló y anotó el genoma de *C. argyrosperma* para compararlo con los genomas disponibles de otras especies del género *Cucurbita* y de la familia Cucurbitaceae. Esto con el objeto de comprender los cambios en el ritmo y modo de evolución molecular en las calabazas después del evento de duplicación completa del genoma. Usando datos de representación reducida del genoma de múltiples poblaciones, se identificaron patrones demográficos y los barridos selectivos que moldearon el genoma de esta especie durante su domesticación. Finalmente, se discute la importancia de los resultados obtenidos para entender la dinámica evolutiva de los genomas en las calabazas, sus aportaciones al entendimiento de la evolución molecular y la importancia teórica y práctica del estudio de la domesticación.

ABSTRACT

The *Cucurbita* genus (Pumpkins, squashes and gourds) is a group of 13 plant species that are native to the American continent. They are herbaceous plants with an evolutionary history that includes a whole-genome duplication and at least six independent domestication events. Studying the patterns of molecular evolution derived from both processes is useful to understand the evolutionary processes at different timescales. The results presented in this thesis focus on the species *Cucurbita argyrosperma* (the silver-seed gourd), a Mexican crop that is cultivated mainly for seed consumption. The thesis first summarizes the modern methods used to study the domestication process of plants and animals. We sequenced, assembled and annotated the genome of *C. argyrosperma* and compared it with the available genomes of other *Cucurbita* and Cucurbitaceae species to understand the changes in the tempo and mode of molecular evolution of the *Cucurbita* genus after its whole-genome duplication. Additional reduced-representation sequencing data from multiple populations allowed us to reconstruct the demographic patterns and identify selective sweeps that shaped the genome of this species throughout its domestication process. Finally, we discuss the importance of our results to understand the evolutionary dynamics of *Cucurbita* genomes, their contribution to our understanding of molecular evolution and both the theoretical and practical importance of domestication studies.

INTRODUCCIÓN

Evolución molecular y genómica

Desde la publicación de “El origen de las especies” en 1859 hasta mediados del siglo pasado, la evolución había sido estudiada con base en caracteres morfológicos (Hedrick, 2011). Con los avances teóricos y tecnológicos realizados en biología molecular desde la década de los 60’s hasta el día de hoy, se han utilizado distintos marcadores moleculares para comprender los procesos evolutivos que moldean a los organismos a lo largo del tiempo (Schlötterer, 2004). Emile Zuckerkandl y Linus Pauling argumentaron en 1965 que las proteínas, el ADN y el ARN pueden ser vistos como documentos de la historia evolutiva de las especies (Zuckerkandl y Pauling, 1965). Posteriormente Richard Lewontin y John Hubby realizaron en 1966 los primeros estudios con variantes alélicas de enzimas (aloenzimas) en poblaciones naturales de *Drosophila pseudoobscura*, abriendo las puertas a una nueva forma de estudiar la genética de poblaciones (Lewontin y Hubby, 1966). Más adelante, Motoo Kimura desarrolla en 1968 la teoría neutral de la evolución, argumentando que el costo de selección para mantener la cantidad observada de variación era demasiado alto para que cualquier población mamífera pudiera soportarla, por lo que la mayor parte de las diferencias genéticas entre organismos debía ser moldeada principalmente por la mutación y la deriva génica (Kimura, 1968). Todos estos eventos cimentaron el desarrollo de una nueva disciplina: la evolución molecular (Schlötterer, 2004). La evolución molecular estudia los procesos y eventos históricos responsables tanto del origen de la variación genética como de los cambios en el material genético a lo largo del tiempo (Futuyma, 2005).

Pronto aparecieron métodos que permitían analizar directamente la variación en el ADN, como los polimorfismos de longitud en fragmentos de restricción (RFLPs por sus siglas en inglés), los microsatélites o la secuenciación de ADN por el método de Sanger (Schlötterer, 2004). Recientemente se han desarrollado tecnologías de secuenciación masiva, las cuales permiten conocer y analizar los genomas completos de los organismos (Eguiarte *et al.*, 2013). Dado que los costos de secuenciación son cada vez menores, los estudios genómicos ya no están limitados a organismos modelo, como *Arabidopsis thaliana* o *Zea mays*, sino que pueden aplicarse a cualquier organismo de interés. El uso de estas nuevas tecnologías nos permite comprender el comportamiento en la totalidad del genoma de las poblaciones y de las especies, permitiéndonos estudiar procesos microevolutivos mediante la genómica de poblaciones, hasta procesos macroevolutivos mediante genómica comparada (Eguiarte *et al*., 2013).

Patrones micro y macroevolutivos de evolución molecular

La *microevolución* se suele definir como los patrones y cambios en la variación biológica a nivel poblacional, mientras que la *macroevolución* se suele atribuir a procesos observados a jerarquías taxonómicas superiores al de especie (Reznick y Ricklefs, 2009). El término *macroevolución* fue acuñado hace casi un siglo al pensarse que los cambios en los planes corporales no podían ser explicados por los cambios a escala poblacional propuestos originalmente por Charles Darwin (Philiptschenko, 1927). Hoy en día, la distinción entre la macroevolución y la microevolución suele ser ambigua e incluso controversial (Hautmann, 2020), ya que su distinción comúnmente se atribuye a la escala de tiempo en la que se observan los procesos microevolutivos que moldean a los organismos, mientras algunos autores sugieren que los patrones macroevolutivos requieren de procesos adicionales a los descritos por los estudios microevolutivos para poder ser explicados (Erwin, 2001), tales como los modelos de autoorganización (Kauffman, 1993) o las restricciones estructurales en los organismos (Webster y Goodwin, 1984). Independientemente de estas controversias, los procesos microevolutivos y macroevolutivos requieren de diferentes aproximaciones para poder ser analizadas a nivel molecular.

Los procesos microevolutivos pueden ser analizados bajo el marco de la genética de poblaciones, una herramienta útil que nos permite comprender cómo las distintas fuerzas evolutivas han moldeado la variación genética observada en las poblaciones a una escala ecológica. La genómica de poblaciones consiste en la aplicación de la teoría de genética poblacional sobre datos de genomas completos, lo cual nos permite discernir la demografía histórica de las poblaciones, identificar aquellas regiones que han sido sometidas a presiones de selección, conocer los tiempos de divergencia entre los organismos, detectar los patrones de introgresión y flujo genético entre poblaciones y encontrar asociaciones entre genotipos y fenotipos (Eguiarte *et al.*, 2013). Esto resulta particularmente útil para entender los procesos de adaptación local, o en el caso de las plantas cultivadas, los procesos de domesticación.

Por otro lado, los procesos macroevolutivos suelen analizarse bajo el marco de la biología comparada con un enfoque filogenético, ya sea mediante el estudio de fósiles, filogenias, el desarrollo embrionario o genomas completos (Reznick y Ricklefs, 2009). En este sentido, la genómica comparada, o sea el análisis comparativo de genomas completos, nos permite analizar procesos macroevolutivos de evolución molecular como el origen, el cambio, la ganancia y la pérdida de los genes y otros elementos funcionales en los genomas, los eventos de rearreglos cromosómicos entre distintos taxa, la dinámica evolutiva de los transposones en los genomas eucariontes, entender los cambios en el orden relativo de los genes en los cromosomas (*i.e.*, la sintenia), entender las convergencias evolutivas a nivel molecular, o la relación entre la aparición de ciertos genes e innovaciones morfológicas y fisiológicas en los organismos (Xia, 2013).

Los procesos de domesticación en plantas

La domesticación puede ser definida como un proceso de evolución mutualista en la cual un organismo asume control sobre la reproducción y cuidado de otro organismo, de manera que pueda aprovechar algún recurso de interés (Zeder 2015). Las presiones selectivas ejercidas sobre los taxa domesticados suele favorecer la fijación de caracteres que son de interés para el domesticador, lo cual conlleva a la divergencia morfológica, fisiológica y molecular entre el taxón domesticado y su pariente silvestre (Purugganan y Fuller 2009; Zizumbo-Villarreal y Colunga-GarcíaMarín 2010).

La agricultura surgió a consecuencia de la domesticación de las plantas (Diamond, 2002). La agricultura es considerada una de las innovaciones tecnológicas más importantes de la humanidad, con la cual los grupos nómadas pasaron a formar grupos sedentarios hace 10,500 años, gracias a la producción confiable y regular de alimento (Diamond, 2002). En este sentido, la domesticación podría considerarse una relación mutualista altamente exitosa, dado que en la actualidad las plantas cultivadas se han propagado a lo largo del planeta, ocupando grandes extensiones de tierra, mientras que las civilizaciones humanas han logrado prosperar y expandirse (Purugganan y Fuller, 2009).

Los organismos domesticados son sistemas útiles para comprender procesos evolutivos como la adaptación, la especiación, la coevolución y la evolución de los genomas (Hancock, 2005; Gepts, 2014); además de revelarnos información importante sobre caracteres de importancia agronómica en los taxa estudiados (Gustafson *et al*., 2008; Hufford *et al*., 2012). Estudios enfocados al análisis de la variación genética a nivel genómico han revelado loci candidatos a haber cambiado debido a las fuerzas de selección artificial durante la domesticación (Meyer y Purugganan, 2013; Gepts, 2014). Estos cambios genéticos explican las diferencias morfológicas y bioquímicas entre taxa domesticados y sus parientes silvestres (Qi *et al*., 2013; Shang *et al*., 2014). El grupo de interés en esta tesis son las calabazas, las cuales han pasado por múltiples eventos de domesticación (Castellanos-Morales *et al*., 2018).

El género *Cucurbita*

El género *Cucurbita*, comúnmente conocidos como calabazas en México o zapallos en Sudamérica, son un grupo de plantas herbáceas nativas del continente americano que presentan frutos pepónides (Lira *et al.*, 2009) y que consta de 13 especies (Paris, 2016). *Cucurbita* es un sistema de estudio interesante, dado que presenta seis eventos independientes de domesticación, que en su mayoría se propone que sucedieron en México (Nee, 1990; Zheng *et al*., 2013; Castellanos-Morales *et al*., 2018). Además, *Cucurbita* es uno de los géneros con mayor variedad morfológica en las angiospermas (Bisognin, 2002). La mayor parte de los taxa se distribuyen en México, con un origen relativamente reciente (16±7 Ma) desde su divergencia con su género hermano, *Peponopsis* (Schaefer *et al*., 2009). Estas plantas son de importancia alimenticia a escala nacional e internacional, dado que se consumen los frutos, las semillas, las flores y las partes "tiernas" de los tallos (Lira *et al.*, 2009). Además, poseen compuestos químicos de importancia económica, como las saponinas, que se usan como detergente y "jabón de calabaza" (Nee, 1990) y la cucurbitacina, un compuesto secundario amargo característico de la familia Cucurbitaceae (Bisognin, 2002) que se concentra en los frutos de las calabazas y que posee propiedades antitumorales y antifúngicas (Tallamy *et al.*, 2005; Lee *et al.*, 2010).

Uno de los clados del género *Cucurbita* [grupo Argyrosperma sensu Lira *et al*. (2009)] contiene a los taxa domesticados *C. argyrosperma* subsp. *argyrosperma* (la calabaza pipiana) y *C. moschata* (la calabaza de castilla), y al taxón silvestre *C. argyrosperma* subsp. *sororia* (Nee, 1990; Zheng *et al*., 2013), cuyas poblaciones se distribuyen por toda la costa del Pacífico en México y Centroamérica, desde Sonora hasta llegar a Nicaragua, con unas poblaciones aisladas en Veracruz (Lira *et al*., 2009). En particular, el taxón domesticado *C. argyrosperma* subsp. *argyrosperma* es interesante para el estudio de la domesticación, ya que se conoce su descendencia a partir de una población ancestral del taxón silvestre *C. argyrosperma* subsp. *sororia* (Sanjur *et al*., 2002; Sánchez-de la Vega *et al.,* 2018). Adicionalmente, ambos taxa están cercanamente emparentados con el taxón domesticado *C. moschata* (Sanjur *et al*., 2002; Castellanos-Morales *et al*., 2018), del cual se desconocen los ancestros silvestres de los cuales pudo originarse (Nee, 1990).

Los tres taxa son principalmente polinizados por abejas de los géneros *Peponapis* y *Xenoglossa*, siendo frecuente el entrecruzamiento ente el taxón silvestre y los dos taxa domesticados (Lira *et al*., 2009). Los registros arqueológicos más antiguos de *C. argyrosperma* subsp. *argyrosperma* datan de hace 8,700 años y fueron encontrados en el Valle Central del río Balsas en Guerrero (Piperno *et al.*, 2009), mientras que los registros más antiguos de *C. moschata* se han encontrado en el Valle de Tehuacán (Puebla) con 6900-5500 años de antigüedad (Whitaker, 1981). Sin embargo, el registro inequívoco más antiguo de calabazas cultivadas pertenece a *Cucurbita pepo* en la cueva de Guilá Naquitz (Oaxaca), con ~10,000 años de antigüedad (Smith, 1997).

El cultivo de *C. argyrosperma* se ha enfocado al uso de sus semillas, con menor énfasis en el uso de sus flores, tallos y frutos (Paris, 2016), desencadenando múltiples diferencias morfológicas que la distinguen de su pariente silvestre (Fig. 1). Esto lo vuelve nuevamente un buen modelo de estudio para comprender los procesos de domesticación en calabazas, ya que se piensa que los primeros caracteres en ser seleccionados durante la domesticación de estas especies fueron las semillas (Lira *et al.*, 2016). Por otro lado, *C. moschata* es aprovechada de manera más amplia, consumiéndose sus frutos maduros e inmaduros, semillas, tallos y flores como fuente de alimento, e incluso sus saponinas para la producción de jabón de calabaza (Paris, 2016). Esto sugiere que ambos taxa han pasado por presiones de selección similares en algunos rubros (*e.g.*, tamaño de la semilla) y presiones distintas en otros (*e.g.,* suculencia y sabor de la pulpa en el fruto).

Cucurbita argyrosperma y *C. moschata* presentan, como casi todas las especies del género, 20 pares de cromosomas (Whitaker, 1933) y el tamaño estimado de sus genomas por citometría de flujo son de ~366 Mbp y ~416 Mbp, respectivamente (Bennett y Smith, 1976; Šiško *et al*., 2003). Se han realizado estudios de mapeo cromosómico en *C. pepo* (Zraidi *et al*., 2007; Esteras *et al*., 2012), *C. moschata* (Gustafson *et al*., 2008) y *C. maxima* (Zhang *et al*., 2015), los cuales han revelado un alto grado de sintenia entre especies, sugiriendo que el orden relativo de los genes en los cromosomas esta conservado dentro del género *Cucurbita* (Gong *et al*., 2008; Gustafson *et al*., 2008).

Recientemente han avanzado los estudios genómicos enfocados en el género *Cucurbita*, y hasta el momento se han publicado genomas de referencia de alta calidad para las especies *C. pepo* (Montero-Pau *et al.*, 2018), *C. moschata* (Sun *et al*., 2017) y *C. maxima* (Sun *et al*., 2017). También se han secuenciado los genomas de otras especies de la familia Cucurbitaceae con distinto número cromo-

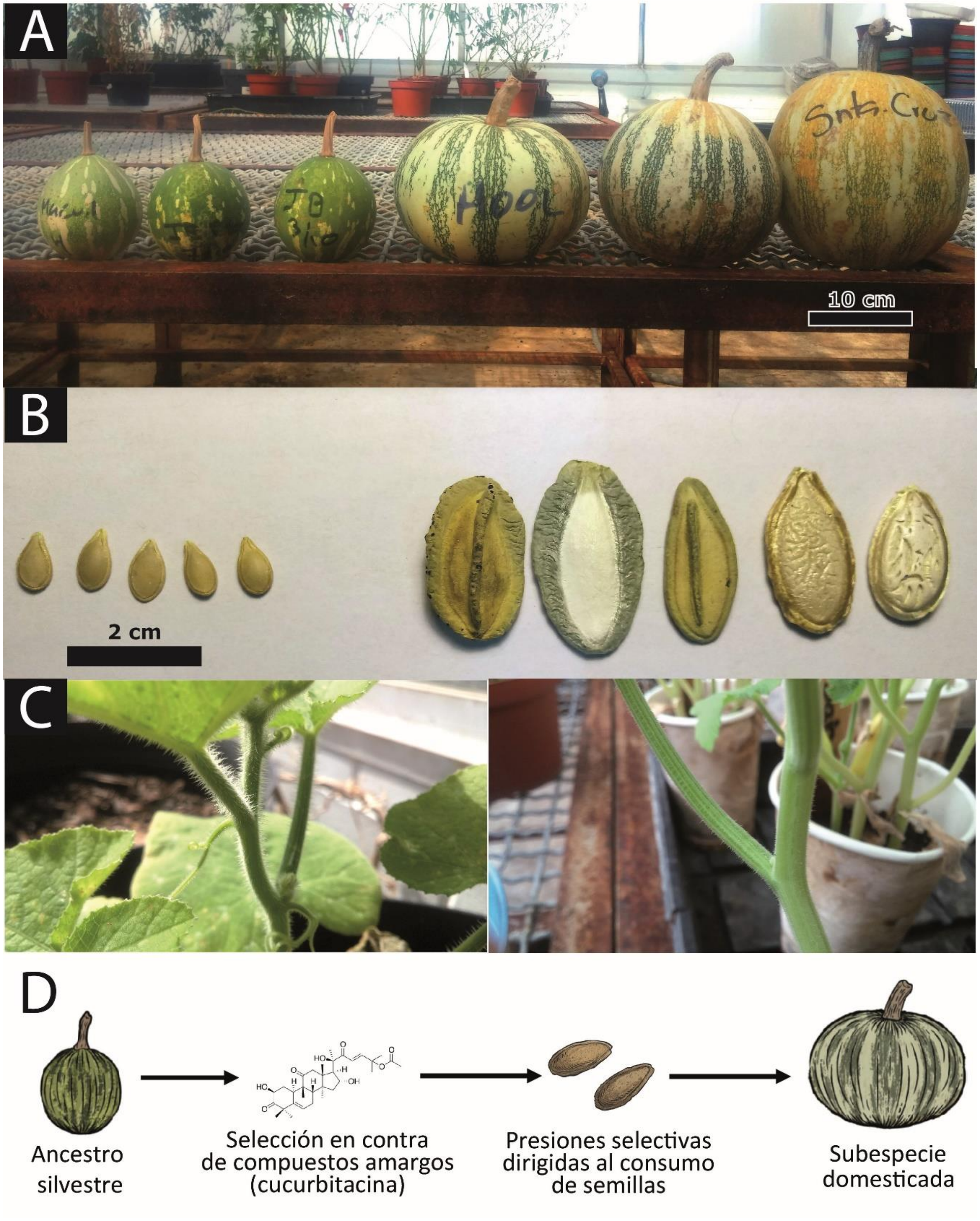

A
HOOL
Snta Cruz
10 cm
B
2 cm
C
D
Ancestro silvestre
Selección en contra de compuestos amargos (cucurbitacina)
Presiones selectivas dirigidas al consumo de semillas
Subespecie domesticada

Figura 1. Diferencias morfológicas entre *Cucurbita argyrosperma* subsp. *argyrosperma* (domesticada; derecha) y *Cucurbita argyrosperma* subsp. *sororia* (silvestre; izquierda). Se observan diferencias en **(A)** el tamaño y forma del fruto, **(B)** el tamaño y forma de la semilla, **(C)** y la presencia de tricomas urticantes. **(D)** Escenario de domesticación propuesto para *C. argyrosperma*.

-sómico cercanas a *Cucurbita* como el pepino (*Cucumis sativus*; 2n=14) (Huang *et al*., 2009), el melón (*Cucumis melo*; 2n=24) (Garcia-Mas *et al*., 2012), la sandía (*Citrullus lanatus*; 2n=22) (Guo *et al*., 2012), el guaje (*Lagenaria siceraria*; 2n=24) (Wu *et al*., 2017) y el melón amargo *(Momordica charantia*; 2n=22) (Urasaki *et al*., 2017).

Duplicación completa del genoma en *Cucurbita*

Desde hace décadas se ha sospechado de un origen alotetraploide (*i,e.*, un organismo que adquiere poliploidía a partir de un evento de hibridización) del género *Cucurbita* (Sun *et al*., 2017). La sospecha inicial surgió a raíz de un estudio citológico entre especies del mismo género (Weiling, 1959) y por el hecho de que las especies del género *Cucurbita* presentan un número cromosómico mayor al de otros miembros de la familia Cucurbitaceae (Xie *et al.*, 2019). Estudios citogenéticos y análisis de expresión genética con isoenzimas parecían apoyar la hipótesis de que el ancestro común del género *Cucurbita* había pasado por un evento de duplicación completa del genoma (Weeden, 1984; Singh, 1990). Un estudio más reciente utilizando un mapa genético de *C. pepo* basado en SNPs (*i.e.*, polimorfismos de un solo nucleótido) encontró bloques sinténicos que se sobrelapaban entre distintos cromosomas de esta especie, sugiriendo nuevamente un grado de duplicación dentro de los genomas de *Cucurbita* (Esteras *et al*., 2012).

No fue hasta que se generaron los primeros genomas de referencia para este género que se confirmó un evento de duplicación completa del genoma en *Cucurbita*, el cual se piensa que ocurrió hace 30± 4 millones de años (Sun *et al*., 2017; Montero-Pau *et al.*, 2018). La confirmación de esta duplicación genómica se basa en tres resultados clave: A) la observación de macrosintenia entre cromosomas de *Cucurbita* y otras cucurbitáceas, observando relaciones de sintenia 2:1; B) la observación de un exceso de genes parálogos por familia génica en *Cucurbita* a comparación de otras cucurbitáceas C) y el hallazgo de una divergencia sincrónica entre dichos genes parálogos dentro de los genomas de calabazas mediante un análisis de sitios sinónimos degenerados (Sun *et al*., 2017; Montero-Pau *et al*., 2018). Después de un evento de duplicación, los genomas suelen pasar por un proceso denominado fraccionamiento, el cual consiste en la perdida diferencial de genes y otros elementos no codificantes por parte de ambas copias del genoma, lo cual lleva a una diferenciación entre las dos copias y una diploidización (*i.e.*, pasar de un estado poliploide a un estado diploide) del genoma (Fig. 2; Cheng *et al*., 2018). Durante la diploidización del genoma duplicado, el contenido de genes dentro del genoma resultante suele ser similar al de sus parientes no duplicados, aunque la arquitectura de dicho genoma cambie radicalmente (Cheng *et al*., 2018). A diferencia de otros eventos de duplicación derivados de una alotetraploidización, los genomas del género *Cucurbita* pasaron por un proceso de fraccionamiento equilibrado, es decir, que ambas copias del genoma perdieron genes de manera equitativa durante el proceso de fraccionamiento y diploidización del genoma (Sun *et al*., 2017), a diferencia de otros alotetraploides donde un genoma suele dominar sobre el otro (Cheng *et al*., 2018).

A pesar de haber pasado por un fraccionamiento equilibrado, no se había estudiado el efecto que pudiera tener esta duplicación completa del genoma sobre la tasa de evolución de los genomas de calabazas, particularmente en familias de genes codificantes y no codificantes (Ponting *et al.*, 2009; Magadum *et al.*, 2013; Nelson y Shippen, 2015). Se ha especulado mucho sobre las consecuencias macroevolutivas que han tenido las duplicaciones completas del genoma sobre la historia evolutiva de las plantas, pero se han realizado pocos estudios empíricos que pongan a prueba dichas hipótesis (Clark y Donoghue, 2018).

Dadas las complicaciones en el estudio evolutivo de elementos no codificantes en los genomas, la mayor parte de los estudios enfocados en duplicaciones completas del genoma se enfocan en los genes codificantes (Ulitsky, 2016; Nelson *et al*., 2017). No obstante, se ha observado que los elementos no codificantes, tales como los ARNs largos intergénicos no codificantes (lincRNAs) aceleran su tasa de recambio después de un evento de “disturbio genómico”, tales como rearreglos cromosómicos o duplicaciones genómicas (Nelson y Shippen, 2015).

Por lo tanto, el estudio del género *Cucurbita* bajo un enfoque de genómica comparada nos permite poner a prueba los posibles efectos que puedan tener una duplicación completa del genoma sobre la dinámica y origen de familias de genes codificantes y no codificantes.

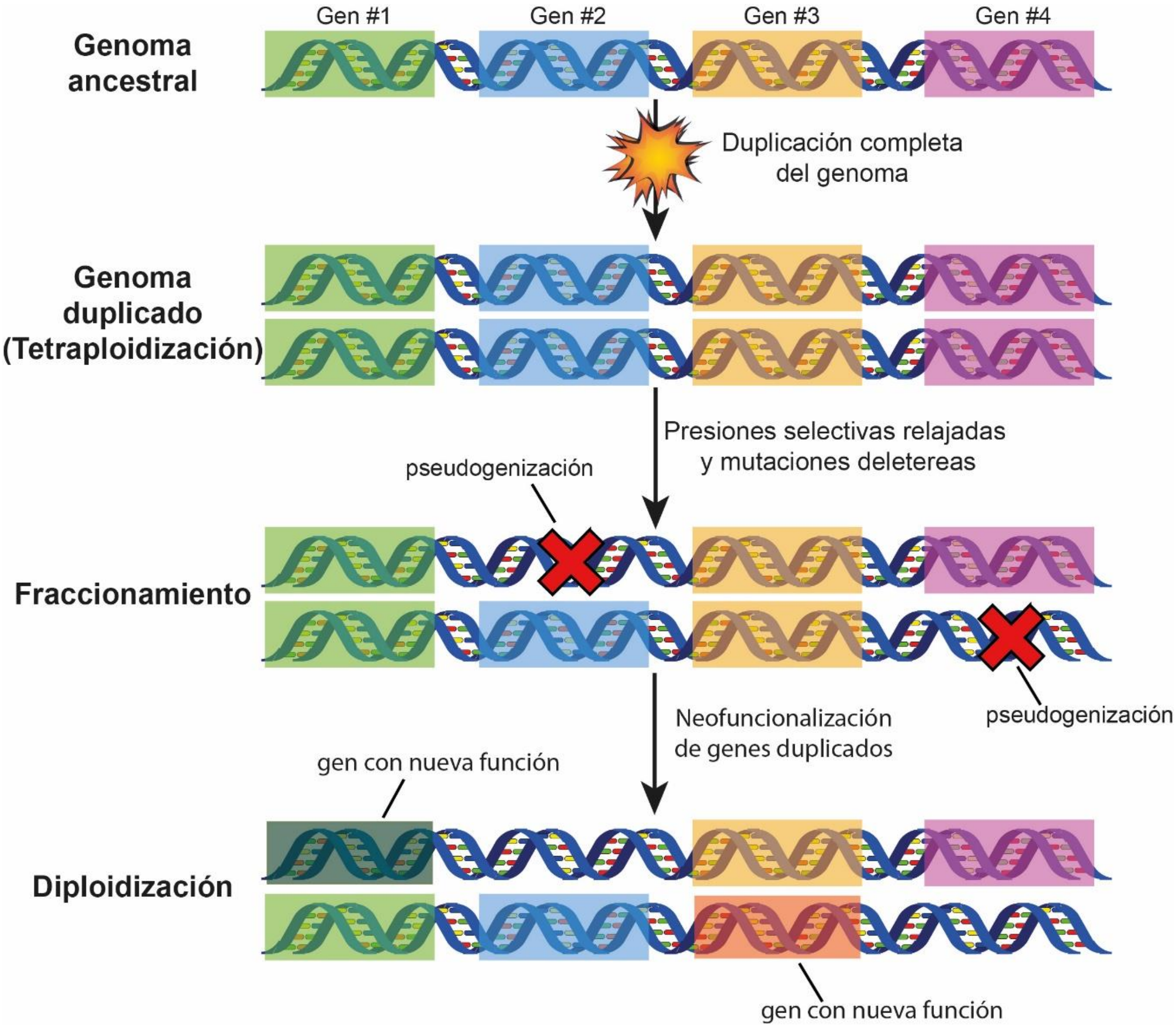

Gen #1
Gen #2
Gen #3
Gen #4
Genoma ancestral
Duplicación completa del genoma
Genoma duplicado (Tetraploidización)
Presiones selectivas relajadas y mutaciones deletereas
pseudogenización
Fraccionamiento
pseudogenización
Neofuncionalización de genes duplicados
gen con nueva función
Diploidización
gen con nueva función

Figura 2. Esquema de los procesos de fraccionamiento y diploidización posteriores a una duplicación completa del genoma.

Domesticación en el género *Cucurbita*

La disponibilidad de recursos genómicos en calabazas nos permite investigar los procesos microevolutivos que han moldeado los genomas de estas plantas durante su domesticación. Se han realizado estudios genómicos enfocados a la domesticación en cultivos americanos como el maíz (Hufford *et al*., 2012; Jiao *et al*., 2012; Romay *et al*., 2013), el frijol (Schmutz *et al*., 2014), el jitomate (Koenig *et al*., 2013; Lin *et al*., 2014) y el chile (Qin *et al*., 2014). También se han estudiado genómicamente aspectos de la domesticación en otros cultivos de importancia mundial como la soya (Lam *et al*., 2010; Li *et al*., 2013), el arroz (Xu *et al*., 2011) y el pepino (Qi *et al*., 2013; Shang *et al*., 2014).

Las calabazas, a pesar de ser cultivos de alta importancia a nivel mundial (Lira *et al*., 2009), carecen hasta el momento de estudios a nivel genómico donde se analice su domesticación. Sin embargo, los estudios genómicos y fisiológicos realizados en otras cucurbitáceas sugieren que se han seleccionado caracteres similares en las cucurbitáceas domesticadas, como el aumento en el tamaño del fruto y una pérdida sistemática de la amargura en los frutos (Chomicki *et al*., 2019). La pérdida de la amargura en otras cucurbitáceas está ligada a presiones selectivas convergentes sobre el factor transcripcional *Bt*, el cual regula la actividad transcripcional de la cucurbitadienol sintasa, la primera enzima en la vía de biosíntesis de la cucurbitacina (Shang *et al*., 2014; Zhou *et al*., 2016). Sin embargo, existen pocos casos donde se hayan seleccionado caracteres asociados a la semilla en cucurbitáceas (Chomicki *et al*., 2019), volviendo a las calabazas, y particularmente a *C. argyrosperma*, un buen modelo para entender la domesticación asociada al consumo de semilla.

Se han realizado estudios previos que elucidan la historia demográfica de las calabazas durante su domesticación, basándose en el uso de microsatélites y secuencias de cloroplasto (Sánchez-de la Vega *et al.*, 2018; Castellanos-Morales *et al.*, 2019). En ambos estudios, se encontró una diferenciación genética entre las subespecies domesticadas y sus parientes silvestres (Sánchez-de la Vega *et al.*, 2018; Castellanos-Morales *et al.*, 2019). En el caso de *C. argyrosperma*, se ha observado una variación genética similar en las poblaciones domesticadas y silvestres, además de evidencia de flujo genético entre ambos taxa (Sánchez-de la Vega *et al.*, 2018). Adicionalmente, se ha propuesto con base en la distancia genética entre las poblaciones silvestres y domesticadas, que la región del Balsas y Jalisco son posibles centros de domesticación para esta especie (Sánchez-de la Vega *et al.*, 2018). A pesar de los esfuerzos realizados para comprender la domesticación de *C. argyrosperma*, los microsatélites solo pueden revelar la historia demográfica de la especie, por lo que se desconocen las regiones del genoma que se encuentran bajo selección durante su domesticación.

El presente estudio consistió en generar un genoma de referencia de alta calidad para el taxón domesticado *C. argyrosperma.* subsp. *argyrosperma* y realizar análisis evolutivos a nivel genómico, comparando con su taxón hermano silvestre *C. argyrosperma.* subsp. *sororia* para dilucidar la historia demográfica y los patrones de selección entre ambas subespecies debido al proceso de domesticación. También se planteó realizar análisis comparativos con otros genomas del género *Cucurbita*, al igual que otros genomas de la familia Cucurbitaceae para describir el efecto que tuvo la duplicación completa del genoma (Sun *et al*., 2017; Montero-Pau *et al*., 2018) sobre la tasa de evolución de familias génicas codificantes y no codificantes en *Cucurbita*.

OBJETIVOS

Los objetivos generales de esta tesis corresponden a los capítulos que la componen, en el siguiente orden:

1. Describir los métodos modernos con los cuales se puede analizar el proceso de domesticación, de manera que se pueda plantear un proyecto de investigación enfocado en la domesticación a nivel genómico.
2. Ensamblar y anotar un genoma de referencia de *C. argyrosperma* para analizarlo con un enfoque de genómica comparada y evaluar el efecto que tuvo la duplicación completa del genoma sobre la dinámica evolutiva de los genes codificantes y no codificantes en el género *Cucurbita*.
3. Realizar análisis de genómica poblacional entre la subespecie domesticada (*C. argyrosperma* subsp. *argyrosperma*) y la subespecie silvestre (*C. argyrosperma* subsp. *sororia*) para dilucidar los eventos demográficos y las presiones de selección que ocurrieron durante el proceso de domesticación de *C. argyrosperma*.

CAPÍTULO 1: ENFOQUES MODERNOS AL ANÁLISIS DE LA DOMESTICACIÓN

Artículo de revisión: Genomic, transcriptomic and epigenomic tools to study the domestication of plants and animals: a field guide for beginners

Este capítulo es una guía teórica y práctica para el estudio de la domesticación mediante los enfoques más recientes de la genómica, transcriptómica, epigenética y la edición de genomas. Dicha guía es útil para estudiantes e investigadores poco familiarizados con estos enfoques, ya que describe las bases teóricas en cada sección, explica tanto las aplicaciones como su relevancia al estudio de la domesticación y expone ejemplos de cómo se han usado estas herramientas de manera exitosa en otros artículos. El capítulo describe los pasos necesarios para obtener un genoma de referencia y explica las distintas técnicas de secuenciación disponibles para realizar genómica de poblaciones, enfocándose en las ventajas y desventajas de cada una. Se exploran los modelos teóricos actuales y las herramientas necesarias para entender la historia demográfica y los procesos de selección a los fueron sometidos los organismos domesticados. Se expone el análisis de pan-genomas en plantas y animales como un modelo útil para conocer las variantes estructurales responsables de los procesos de domesticación. Se aborda el análisis de ADN antiguo extraído de material arqueológico para resolver preguntas que no son posibles de abordar analizando a las poblaciones de la actualidad. También se describe el diseño experimental necesario para estudiar los cambios en los patrones de expresión transcriptómica derivados de los procesos de domesticación. Se abordan las distintas herramientas disponibles para comprender el papel de los mecanismos de regulación epigenética en los procesos de domesticación. Finalmente, se describe la utilidad de las herramientas de edición genética para validar experimentalmente los genes candidatos de la domesticación, además del potencial uso de estas herramientas para el mejoramiento genético y realizar domesticación *de novo*. Este capítulo fue publicado en la revista *Frontiers in Genetics* en el mes de julio del 2020.

frontiers in Genetics

REVIEW
published: 15 July 2020
doi: 10.3389/fgene.2020.00742

Genomic, Transcriptomic and Epigenomic Tools to Study the Domestication of Plants and Animals: A Field Guide for Beginners

*Josué Barrera-Redondo, Daniel Piñero and Luis E. Eguiarte**

Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico

OPEN ACCESS

Edited by:
TingFung Chan,
The Chinese University of Hong Kong,
China

Reviewed by:
Eric Von Wettberg,
The University of Vermont,
United States
Martin Mascher,
Leibniz Institute for Plant Genetics
and Cultural Plant Research (IPK),
Germany
David Irwin,
University of Toronto, Canada

****Correspondence:***
Luis E. Eguiarte
fruns@unam.mx

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *25 April 2020*
Accepted: *22 June 2020*
Published: *15 July 2020*

Citation:
Barrera-Redondo J, Piñero D and Eguiarte LE (2020) Genomic, Transcriptomic and Epigenomic Tools to Study the Domestication of Plants and Animals: A Field Guide for Beginners. Front. Genet. 11:742. doi: 10.3389/fgene.2020.00742

In the last decade, genomics and the related fields of transcriptomics and epigenomics have revolutionized the study of the domestication process in plants and animals, leading to new discoveries and new unresolved questions. Given that some domesticated taxa have been more studied than others, the extent of genomic data can range from vast to nonexistent, depending on the domesticated taxon of interest. This review is meant as a rough guide for students and academics that want to start a domestication research project using modern genomic tools, as well as for researchers already conducting domestication studies that are interested in following a genomic approach and looking for alternate strategies (cheaper or more efficient) and future directions. We summarize the theoretical and technical background needed to carry out domestication genomics, starting from the acquisition of a reference genome and genome assembly, to the sampling design for population genomics, paleogenomics, transcriptomics, epigenomics and experimental validation of domestication-related genes. We also describe some examples of the aforementioned approaches and the relevant discoveries they made to understand the domestication of the studied taxa.

Keywords: population genomics, pangenomics, ancient DNA, differential expression analysis, epialleles, genome editing

INTRODUCTION

The modern study of domestication of plants and animals is multidisciplinary, and relevant contributions come from botany, zoology, archeology, genetics, ethnobiology, biogeography, and linguistics (Larson et al., 2014). Modern domestication studies seek to understand the dates of domestication, the places where domestication started and number of times that domestication took place, as well as the details of the evolutionary and ecological forces that led to the divergence between the domesticated taxa and their wild relatives and ancestors (Zeder, 2006; Larson et al., 2014).

Given that domestication is an evolutionary process, genetics emerged as a powerful tool to understand the domestication of plants and animals, revealing the demographic history of the domesticated taxa and the genetic variants that underlie their domesticated phenotypes (Zeder et al., 2006; Gepts, 2014). The advent of high-throughput sequencing technologies sparked the use of genomic studies to understand the domestication of crops and animals in a much deeper level than previously imagined, as researchers can now pinpoint the genetic changes that allowed domestication to happen (Ross-Ibarra et al., 2007; Gepts, 2014).

WHY AND HOW TO USE A GENOMIC APPROACH IN DOMESTICATION STUDIES? TOP-DOWN AND BOTTOM-UP APPROACHES FOR THE STUDY OF DOMESTICATION

In genetics, we refer to *top* or *up* when referring to a specific phenotype, while we refer to *bottom* or *down* when referring to the underlying genotype responsible for that trait. Thus, top-down approaches start by studying a particular phenotype and searching for its genetic basis. Huge advances in the genetic study of domestication traits have been made using classic top-down approaches (*e.g.*, Sax, 1923; Paterson et al., 1988; Doebley and Stec, 1991; Doebley et al., 1995), which are performed by analyzing the phenotypic traits of interest between wild and domesticated taxa, and then finding the genetic variant or variants that correlate with the phenotypic traits through the mapping of quantitative trait loci and linkage disequilibrium (Ross-Ibarra et al., 2007; Kantar et al., 2017). These top-down approaches are precise in finding causal variants involved in the evolution of specific traits, but usually they are very labor-intensive and are biased towards *a priori* selected phenotypes to be compared between wild and domesticated taxa (Ross-Ibarra et al., 2007; Kantar et al., 2017).

In contrast to top-down approaches, bottom-up approaches start by analyzing the genetic variation within genomes in order to detect potential signals of selection related to the domestication process and finally associate such evolutionary signals to important loci and domestication phenotypes (Ross-Ibarra et al., 2007; Kantar et al., 2017). In the last decade, high-throughput sequencing technologies allowed us to analyze entire genomes of one or several individuals of domesticated taxa, and to compare them to different varieties or to their wild relatives (*e.g.*, Hufford et al., 2012; Yang et al., 2012; Li et al., 2013; Wang et al., 2019; Zeng et al., 2019).

Bottom-up approaches do not need an *a priori* phenotypic target, enabling a genome-wide search of domestication-related loci without previous background of possible candidates, revealing important traits that can hardly be studied using a top-down approach (Ross-Ibarra et al., 2007; Kantar et al., 2017). Nevertheless, the results of bottom-up approaches can be limited by the sampling scheme, the density of genetic markers, and the detection of false positives (Tiffin and Ross-Ibarra, 2014), so these genomic approaches have to be properly and carefully designed in order to obtain satisfying results (De Mita et al., 2013).

Genomic data facilitated the widespread and reliable use of bottom-up approaches to study plant and animal domestication, but top-down strategies were also aided by genomics, allowing a more efficient search of genotype-phenotype correlations through genome-wide association studies (GWAS; Wang G.-D. et al., 2014), which can be defined as experimental designs that are used to detect the association between genetic variation in a population and phenotypical traits of interest (Visscher et al., 2017).

Genome-wide genetic markers allows to differentiate between global and local evolutionary signals occurring throughout the genome (Diao and Chen, 2012), discerning the signals of selection during domestication (Vitti et al., 2013) from other fine-scale signals of demographic events that occurred during the domestication process (Meyer and Purugganan, 2013; Guerra García and Piñero, 2017).

The use of modern genomic tools is not limited to population genetics, as other interesting approaches can reveal important aspects of the domestication process. For instance, one can analyze changes in the transcriptional activity of genes related to domestication (Hekman et al., 2015), demonstrate the phenotypic effects of certain alleles through the use of genomic editing tools (Zhou J. et al., 2019), search for epigenetic patterns that changed between domesticated and wild taxa (Janowitz Koch et al., 2016) or analyze the genetic makeup of archeological samples (Irving-Pease et al., 2019).

This review describes the necessary steps and data to start a genomic research project towards understanding domestication, the questions that can be approached using genomic data and the main results obtained from previous studies using these methods (**Figure 1**).

WHOLE-GENOME ASSEMBLY AND REFERENCE GENOMES

Whole-genome assembly is one of the first steps in modern domestication studies, since it generates a reference genome that is useful for downstream analyses. Whole-genome assembly projects require the use of high-throughput sequencing technologies such as Illumina (*e.g.,* Sun et al., 2017), PacBio (*e.g.,* Badouin et al., 2017; VanBuren et al., 2018), Oxford Nanopore (*e.g.,* Belser et al., 2018) or a combination of these (*e.g.,* Bickhart et al., 2017; Zhou Y. et al., 2019) to sequence the genome of interest of a single individual. Before starting a genome assembly project, a rough estimate of the haploid genome size must be known as well as the ploidy of the organism, since the assembly difficulty and sequencing cost are determined by both factors (Sims et al., 2014). In order to successfully assemble eukaryotic genomes, where repetitive elements usually comprise a significant portion of its content [ranging from 3% in tiny genomes such as *Utricularia gibba* (Ibarra-Laclette et al., 2013) up to 65.5% in huge genomes such as *Ambystoma mexicanum* (Nowoshilow et al., 2018)], it is necessary to generate sequencing libraries with large insert sizes – called mate-pair libraries – or use long-read sequencing technologies such as PacBio or Oxford Nanopore (Levy and Myers, 2016; Sohn and Nam, 2016). Additionally, the use of chromosome conformation capture (Mascher et al., 2017), optical mapping (Dong et al., 2013) or linkage maps obtained from crosses (Fierst, 2015) will help achieve chromosome-level assemblies that are highly desirable to adequately assess haplotypes, linkage disequilibrium, putative genomic rearrangements and the

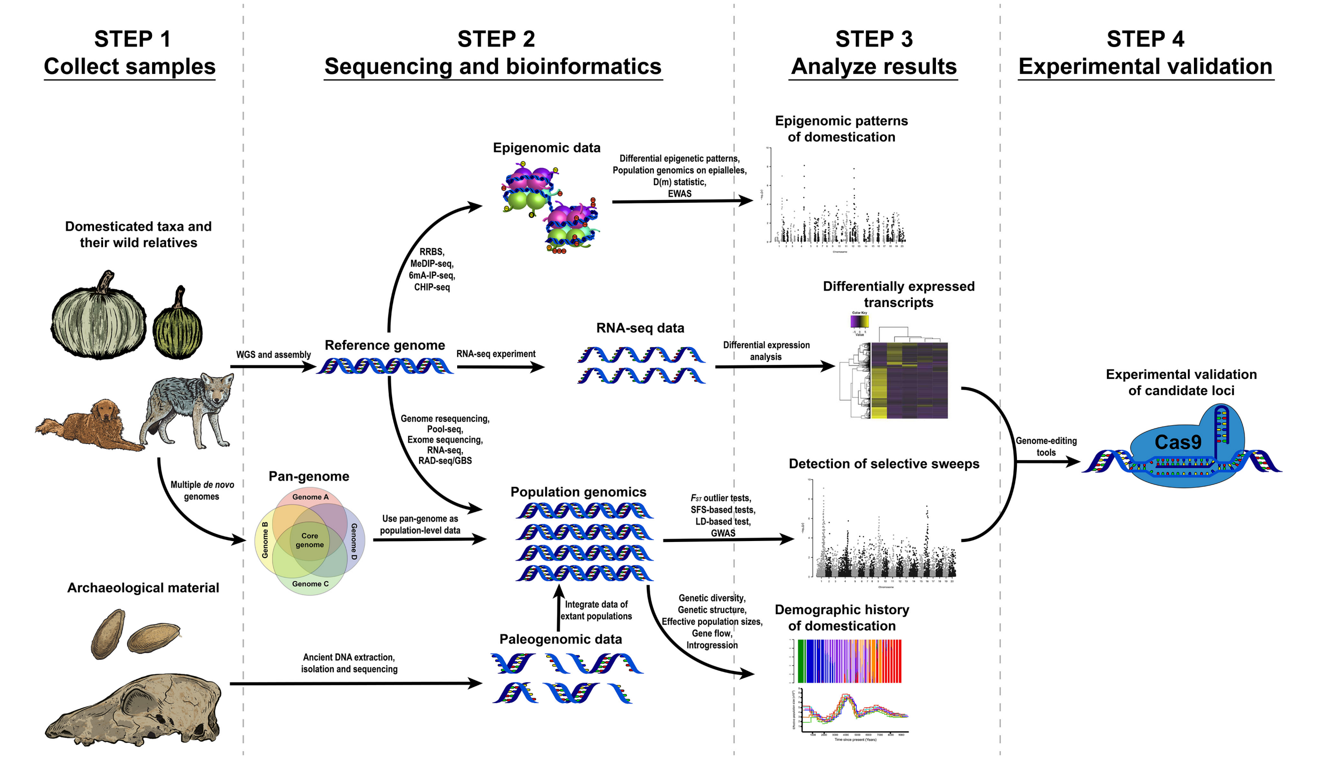

STEP 1
Collect samples
STEP 2
Sequencing and bioinformatics
STEP 3
Analyze results
STEP 4
Experimental validation
Domesticated taxa and their wild relatives
WGS and assembly
Reference genome
RNA-seq experiment
RNA-seq data
Differential expression analysis
Differentially expressed transcripts
RRBS, MeDIP-seq, 6mA-IP-seq, CHIP-seq
Epigenomic data
Differential epigenetic patterns, Population genomics on epialleles, D(m) statistic, EWAS
Epigenomic patterns of domestication
Genome resequencing, Pool-seq, Exome sequencing, RNA-seq, RAD-seq/GBS
Multiple de novo genomes
Pan-genome
Genome A
Genome B
Genome C
Genome D
Core genome
Use pan-genome as population-level data
Population genomics
F_ST outlier tests, SFS-based tests, LD-based test, GWAS
Detection of selective sweeps
Genome-editing tools
Experimental validation of candidate loci
Cas9
Archaeological material
Ancient DNA extraction, isolation and sequencing
Paleogenomic data
Integrate data of extant populations
Genetic diversity, Genetic structure, Effective population sizes, Gene flow, Introgression
Demographic history of domestication
Chromosome
Time since present (Years)
Effective population size

FIGURE 1 | Proposed workflows to study different problems related to the domestication of plants and animals through genomic, transcriptomic and epigenomic tools.

genomic location of candidate loci (Sohn and Nam, 2016; see **Table 1**).

After sequencing and assembling the genome of at least one individual, it must be properly annotated before it can be of any use. Since eukaryotic genes are structurally complex, genome assemblies require the additional sequencing of RNA data from the same species to be used as transcriptomic evidence, alongside homology evidence from other curated genomes and *ab initio* predictions based on the underlying structure of genes, in order to be successfully annotated (Yandell and Ence, 2012; see **Table 1**). Even though whole-genome assembly projects were previously restricted to large research groups (*e.g.,* Schnable et al., 2009; Tomato Genome Consortium, 2012), the sequencing cost per nucleotide is declining constantly in all the aforementioned technologies, making genome analyses accessible for a large part of the research community (Muir et al., 2016). The current bottleneck for small research groups is usually not the cost of sequencing itself, but rather the availability of computational resources capable of storing and analyzing huge amounts of data (Muir et al., 2016).

The main purpose of assembling a genome in a domestication study is to use it as a reference for high-quality population data to infer the selection, introgression and recombination processes, and to design posterior studies for experimental validation of candidate loci. Even though several population-level analyses based on reduced-representation genome sequencing can be performed in the absence of a reference genome (De Wit et al., 2012; Mastretta-Yanes et al., 2015), the use of a reference genome alongside population data enables the correct identification of otherwise anonymous loci into specific genes or regions within the genome and it makes possible the identification and the proper handling of linkage between loci (Fitz-Gibbon et al., 2017). Also, it can help to discriminate between orthologous and paralogous loci, which is critical given the large size of many genomes and the frequent genome duplication processes experienced during the evolution of plant and animal lineages (Clark and Donoghue, 2018; Zadesenets and Rubtsov, 2018).

Thus, the availability of a reference genome is desired for genomic analyses concerning domestication. Luckily, domesticated taxa are usually economically relevant, drawing the attention of several research groups worldwide and in some cases helping to fund the projects. Therefore, reference genomes are usually available for domesticated species, since such data is also relevant for other research areas, such as crop improvement and breeding programs (Ellegren, 2014). However, it should be noted that using a single reference genome can lead to reference bias, where sequenced individuals that are more distantly related to the reference will tend to have fewer predicted variants due to mismatches while mapping the reads (Günther and Nettelblad, 2019).

Besides its use as a reference genome for population-level data, the analysis of several whole-genome assemblies between

TABLE 1 | List of key publications with other reviews that are focused on specific topics, as well as some notable examples of research articles using some of the methods described in this review with reliable results.

Topic	Citation	Usefulness/importance
Genome assembly and reference genomes	Sohn and Nam, 2016	In-depth review on genome assembly. Includes compelling explanations behind the genome assembly algorithms and an extensive list of genome assembly strategies.
	Yandell and Ence, 2012	In-depth review on eukaryotic genome annotation, a description of the available tools to predict genes and best practices when predicting genes.
Sequencing strategies	Meirmans, 2015	Classic review concerning common pitfalls that should be avoided in a population genomic study. A compulsory review for any newcomer to population genomics.
	Dorant et al., 2019	Empirical study that compares the efficiency of Pool-seq, RADseq and Rapture to detect weak signals of genetic structure in lobsters.
	Inbar et al., 2020	Empirical study that compares the efficiency of whole-genome sequencing, Pool-seq and RADseq for GWAS in ants.
Pan-genomics	Golicz et al., 2016a	In-depth review about pan-genomics in plant species, its advantages over the use of reference genomes, a guide on how to generate pan-genomes and the importance of studying structural variants. The article is dedicated to plants, but the rationale and methods can also be applied to other eukaryotes.
	Khan et al., 2020	Opinion article detailing the relevance of pan-genomes as a necessary next step from reference genomes. The authors also highlight the importance of including wild taxa into pan-genomics and propose the idea of genus-level super-pan-genomes.
	Gao et al., 2019	Landmark study of the tomato pan-genome. The authors sequenced 725 accessions from the domesticated tomato and its wild relatives. They found 4,873 additional genes, including several well-characterized genes that were absent from the reference-genome. They also evaluated the presence-absence variants between the wild and domesticated tomatoes, which were enriched in disease-resistance genes.
Demographic analyses	Linck and Battey, 2019	Research study focused on the effects of minor allele-frequency filters to detect genetic structure in populations. Gives a good explanation on the rationale behind the clustering-based methods to detect structure.
	Mather et al., 2020	Review dedicated to the theoretical background and technical requirements of PSMS and MSMC to infer changes in effective population sizes and coalescent times.
	Gerbault et al., 2014	Excellent review on how to use Bayesian approaches to test different demographic models of domestication.
	Frantz et al., 2015	Landmark study on pig domestication. The authors make use of Approximate Bayesian computation to compare domestication scenarios, they use clustering-based methods to detect genetic structure and used a graph-based method to infer the genetic relationship between pig and wild boar populations.
Selection scans	Vitti et al., 2013	Good review focused on explaining the rationale behind many of the bottom-up tests to detect selection and the genomic signals they are sensitive to.
	De Mita et al., 2013; Lotterhos and Whitlock, 2015	Classic simulation-based studies that compare different scenarios to evaluate the best sampling strategies and the most powerful methods to detect selection throughout the genome, according to the reproductive nature of the organism under study.
	Gibson, 2018	A primer dedicated to understanding the principles behind GWAS and its ability to detect polygenic effects on quantitative traits.
	Hufford et al., 2012	A landmark paper that illustrates how to perform genome scans to detect domestication-related loci in domesticated taxa, and the importance of these loci for crop improvement. The paper studies the domestication of maize, but a similar study design can be applied to domesticated animals.
Paleogenomics	Irving-Pease et al., 2019	Exhaustive book chapter dedicated to the study of ancient DNA to understand domestication.
	Allaby et al., 2019	Research study that casts into doubt the long-lasting idea that domestication processes lead to strong population bottlenecks by re-analyzing data based on ancient DNA samples.
	Daly et al., 2018	Remarkable study that sequenced and analyzed 83 mitochondrial genomes and 51 nuclear genomes from ancient goat samples. The authors found signals of ancient introgression events, as well as ancient selective signals related to several traits that are shared with modern goats.
Transcriptomics	Fang and Cui, 2011	General guideline on how to adequately design an RNA-seq experiment to avoid technical mistakes and generate meaningful results.
	Yang and Kim, 2015	General guideline on how to analyze RNA-seq data to assess differential expression.
	Hekman et al., 2015	In-depth review dedicated to study the domestication process through transcriptomics, including methodological strategies and challenges.
	Hradilová et al., 2017	An excellent study that combines transcriptomic data with metabolomic data and morphological data between domesticated and wild peas. The analysis of multi-omic data allowed them to get a better understanding behind seed dormancy and pod dehiscence in domesticated peas.
Epigenomics	Guerrero-Bosagna, 2012; Heard and Martienssen, 2014; Burggren, 2016	Contrasting views on the role of transgenerational epigenetic inheritance in evolution. The topic is still debated and should be viewed critically.
	Jensen, 2015	In-depth review on the rationale and advances of epigenetic studies to understand domestication. The manuscript is focused on animal behavior, but many of the ideas can also be applied to domesticated plants.

(Continued)

TABLE 1 | Continued

Topic	Citation	Usefulness/importance
	Janowitz Koch et al., 2016	A landmark paper showing the importance of epigenetic marks on dog domestication and its association with behavioral traits. The study doesn't just compare the methylation marks between wolves and dogs, but also assess the heritability of the methylation marks and proposes a formal test to detect selection on epialleles.
Genome-editing tools	Boettcher and McManus, 2015).	Review on novel genome-editing techniques and RNA interference. Useful to compare and choose the best tool to validate candidate loci.
	Shan et al., 2020	A general guide on how to develop a CRISPR/Cas9 system on a non-model plant species.
	Soyk et al., 2017	Landmark paper that uses genome-editing to validate two candidate genes related to fruit size and reduced fruit dropping in tomato. The authors also detect the emergence of undesirable traits in domesticated tomatoes due to an epistatic effect between both domesticated loci and introduce wild alleles to generate new tomato phenotypes with reduced degrees of the undesirable traits.
Perspectives	Piperno, 2017	Review centered on the potential application of an extended synthesis framework to understand domestication. Centered around the concepts of niche construction, transgenerational epigenetic inheritance and developmental plasticity.

domesticated and wild taxa will help us reveal structural differences between the genome of a domesticated taxon and its closest wild relatives, such as duplications, chromosome rearrangements or presence/absence of entire genes and genomic regions (Yang et al., 2012; Wang W. et al., 2014; Xie et al., 2019). Since selection and bottlenecks during domestication often leads to the fixation of mutations that involve a loss of function (Renaut and Rieseberg, 2015; Moyers et al., 2018), comparative analyses using genome assemblies of wild ancestors may also reveal these changes in genes that could not be properly predicted within the domesticated genome (Moyers et al., 2018). In this sense, further efforts should be made to assemble high-quality genomes of wild relatives alongside the domesticated taxon of interest (Brozynska et al., 2016; Xie et al., 2019).

STRATEGIES TO GATHER ADEQUATE POPULATION GENOMICS DATA

Genome assemblies alone give us a limited view on domestication, unless several genomes of wild relatives (if known and available) and domesticated individuals are sequenced, because evolution is a population-level process, and in consequence population data is necessary to address most of the evolutionary questions in domestication (Wang G.-D. et al., 2014; Guerra García and Piñero, 2017). Population genomics examines the genetic variation within and between populations that is scattered across the entire genome to assess the demographic history, phylogenetic relations and selective pressures of a species (Jorde, 2001). Several types of genomic data can be evaluated at the population level, including single nucleotide polymorphisms (SNPs), indels and copy number variations; but SNPs are the most commonly analyzed of the three (Seal et al., 2014).

All population-level sequencing techniques share common pitfalls that should be known and avoided before investing any money on sequencing. Population sampling should be planned carefully, as the sampling scheme has a stronger impact over sequencing to obtain reliable results in any analysis (Meirmans, 2015). Also, different populations should be mixed, rather than being sequenced on separate libraries or sequencing lanes, as failing to do so will generate sequencing biases that can be confused with biological patterns (Meirmans, 2015; see **Table 1**).

Once adequate genomic population data is gathered, we need to analyze the demographic processes that shaped the genetic variation and the population structure of contemporary populations during the domestication process. This data is necessary to perform tests to detect natural and artificial selection, which are required to understand the genetic base of domestication syndromes (Ross-Ibarra et al., 2007). There are several approaches to obtain population data at a genomic scale, which differ in the fraction of the genome that is sequenced, therefore determining the sequencing cost of each sample (Schreiber et al., 2018).

Whole-Genome Sequencing of Populations

After assembling a reference genome, one of the next possible strategies to understand domestication is to sequence the complete genome of several individuals. This approach requires the alignment of the sequencing reads back to a reference genome, in order to infer the variable sites between individuals and know the genetic elements (e.g., genes, upstream regulators, repetitive elements, non-coding RNAs) associated to those sites. The main benefit of this approach is its potential to retrieve all the variant sites within an individual's genome that are structurally represented in the reference genome. Whole-genome sequencing can be used in almost any population-level test of interest (Schreiber et al., 2018). Common practices recommend a sequencing depth around 30× per individual, but empirical studies in pigs suggests that even 10x is enough to cover up to 99% of a genome with accurate detection of variant sites (Jiang L.G. et al., 2019). The main drawback of this approach is the sequencing cost of each sample, which is significantly higher compared with other approaches, especially for organisms with large genomes such as polyploid crops or mammals (Schreiber et al., 2018). This can lead researchers to evaluate a trade-off between sequencing depth and number of sampled individuals to optimize their resources. Simulation studies suggest that sequencing more individuals is more convenient to obtain reliable results, even at the expense of lower sequencing depths per individual (Fumagalli, 2013).

Alternatives to Whole-Genome Sequencing

Other approaches aim to reduce the sequencing cost per samples by pooling the DNA of several individuals into a single sequencing library (Futschik and Schlötterer, 2010) or by reducing the portion of the genome that is sequenced (often named as reduced-representation sequencing), either by sequencing arbitrary defined segments scattered across the genome, by targeting the desired portions of the genome or by sequencing the transcriptionally active portions of the genome (Schreiber et al., 2018). These techniques are especially helpful for organisms with very large genomes, and some of these methods can even be used in the absence of a reference genome (Mastretta-Yanes et al., 2015; Schreiber et al., 2018). Furthermore, the reduced representation of the genome means that those fewer regions that are targeted can have a high sequencing depth, leading to higher accuracy of the observed genetic variation and better heterozygosity estimations (Schreiber et al., 2018). Additionally, the reduced sequencing cost per sample allows for a large number of sequenced individuals and populations that, with a proper sampling strategy, can lead to robust results (De Mita et al., 2013; Lotterhos and Whitlock, 2015). Due to the fragmented nature of these sequencing techniques, reduced representation data alone may be insufficient to pinpoint all or even the most important possible causal genetic variants associated to the domestication syndromes (Lowry et al., 2017), but they are still useful to infer basic genetic statistics, infer demographic properties and past demographic scenarios, detect some signatures of selective sweeps across the genome and even perform GWAS for domestication traits of interest (Andrews et al., 2016; Schreiber et al., 2018).

Pool Sequencing

Pool sequencing (Pool-seq) is a promising alternative to whole-genome sequencing with a much lower cost (Futschik and Schlötterer, 2010). As the name suggests, Pool-seq consists of sequencing a large pool of individuals for a given population into a single high-throughput sequencing library, instead of sequencing each individual separately, allowing an accurate estimation of allele frequencies and other parameters of population genetics at the expense of losing individual-level information (Futschik and Schlötterer, 2010). This method requires to map the reads against a reference genome of the same species in order to work (Schlötterer et al., 2014). It is intended for sequencing large pools of individuals (>40 individuals per population is recommended, but >100 is optimal), otherwise the allele frequencies will not be estimated accurately (Schlötterer et al., 2014). The relative amount of pooled DNA of each individual in a Pool-seq study should be similar in order to avoid overrepresentation of individual alleles, a task that is often challenging (Schlötterer et al., 2014).

Pool-seq has several limitations that should be considered based on the objectives of the research project. It is difficult to discard a low-frequency allele from a sequencing error, but this problem is potentially fixed by either establishing a minor allele frequency threshold for SNP calling or by using pool replicates (Schlötterer et al., 2014). One important limitation is the inability of Pool-seq data to estimate linkage disequilibrium and haplotype phasing, which is particularly important to evaluate the non-independence of genetic signals in demographic studies and selective scans (Schlötterer et al., 2014). Finally, assessing genetic structure can be difficult and sometimes misleading when using Pool-seq, due to potential biases in individual allele representations within the pool (Dorant et al., 2019). This makes Pool-seq an adequate method for GWAS, selective sweeps and some methods based on allele frequencies when resources are limited (Luu et al., 2017; Inbar et al., 2020), but the loss of individual-level information makes many of the demographic inferences difficult, as populations need to be predefined before sequencing (Dorant et al., 2019), and the bioinformatic tools that handle Pool-seq data are scarce.

Exome Capture and Sequencing

Exome sequencing is another lower-cost alternative to whole-genome sequencing which targets the protein-coding regions of the genome (Warr et al., 2015; Kaur and Gaikwad, 2017). Protein-coding genes represent a small fraction of eukaryotic genomes, which is particularly useful for most population genomic studies, since it represents mostly functional elements within genomes (Kaur and Gaikwad, 2017). This technique is usually performed using hybridization probes, which requires previous knowledge of the genome content as well as a *priori* selection of regions of interest in order to design probes (Kaur and Gaikwad, 2017). Fortunately, hybridization probes are already available for several domesticated plants and animals (Warr et al., 2015; Kaur and Gaikwad, 2017).

Despite its advantages, exome sequencing can generate an uneven sequencing depth in certain genomic positions, unlike whole-genome sequencing that shows a uniform distribution of reads throughout the genome (Lelieveld et al., 2015). Another important limitation of exome sequencing is its bias towards the protein-coding portion of the genome, since increasing evidence shows that many of the genetic changes that have been directly associated to domestication traits are located within *cis*-regulatory elements, noncoding RNAs and other *trans*-regulatory elements, rather than within the open reading frame of the genes (Swinnen et al., 2016). Despite its limitations, demographic history and selective sweeps can still be detected using this sequencing method (Pankin et al., 2018).

RNA Sequencing of Populations

Transcriptome sequencing (also known as RNA-seq) is another useful approach to obtain population-level data from the transcriptionally active elements within genomes (De Wit et al., 2012). RNA-seq can be mapped against a reference genome to detect genetic variants and determine the genomic regions of interest, but it can also be analyzed in the absence of a reference genome (De Wit et al., 2012), since transcriptomes can be assembled *de novo* (Haas et al., 2013) and the functional annotation of the assembled transcripts is relatively easy (Bryant et al., 2017).

However, transcription profiles are dependent on the sequenced tissues and organs, the development stage of the organism, and the influence of external stimuli, capturing just the transcripts that are active at the moment of RNA extraction (Hekman et al., 2015). This complexity can generate important biases in the relative abundance of certain transcripts over others and overlook potential adaptative genes whose expression are context dependent (Hekman et al., 2015; Kaur and Gaikwad, 2017). Nonetheless, RNA-seq is still a good option for species with large genomes that are hard to assemble (De Wit et al., 2012). Similarly to exome-sequencing, RNA-seq data can be used to evaluate demographic history and selective sweeps, but the selective signals are restricted to the transcriptionally active part of the genome, and cannot be used to evaluate structural variants (Schreiber et al., 2018).

Restriction Site-Associated DNA Sequencing

Restriction site-associated DNA sequencing (RAD-seq), which may also be referred to as genotyping by sequencing (GBS), has been one of the most popular options for cost-affordable population genomics in the last decade (Davey and Blaxter, 2010). The technique consists in using restriction enzymes to digest the DNA and sequence the regions adjacent to the restriction sites that are scattered across the genome (Davey and Blaxter, 2010). It can also be combined with sequence capture techniques to target specific loci of interest (Ali et al., 2016). RADseq data can either be mapped against a reference genome or it can be assembled *de novo* (Catchen et al., 2013; Mastretta-Yanes et al., 2015), making it a versatile technique for species with scant genomic resources.

However, empirical studies show that using certain *de novo* approaches for RAD-seq data can lead to fewer predicted SNPs due to errors in the definition of loci and treatment of sequencing errors (Shafer et al., 2017), all which may subsequently alter downstream analyses, especially those based on the distribution of allele frequencies within the genome of a population, also known as the site frequency spectrum (SFS) (Shafer et al., 2017). For this reason, a reference-based approach is highly recommended as long as the reference genome is closely related to the population dataset (Shafer et al., 2017). Furthermore, RADseq data could involve errors when a polymorphism resides within a restriction site, which prevents the enzyme to cut in individuals carrying such polymorphism, leading to failures in sequencing that region in homozygous individuals (null alleles) and makes heterozygous individuals to look like homozygotes (allele dropout) (Andrews et al., 2016). Finally, the capacity of RADseq libraries to adequately perform selective scans has been casted into serious question (Lowry et al., 2017). Its potential capacity to detect selective sweeps is dependent on the genome size, the density of variants detected for a given genomic region and specially the length of the extent of linkage disequilibrium in the genome (Lowry et al., 2017). Thus, when a species genome has short regions in linkage disequilibrium (due to high recombination rates) and the SNP density is low (particularly in large genomes), odds are that the selective scans will likely miss a significant portion of selective sweeps associated to domestication (Lowry et al., 2017).

PAN-GENOME ANALYSES IN DOMESTICATED AND WILD TAXA

An increasing number of studies are revealing that structural variants (copy-number variation, presence/absence of genomic regions, inversions, transversions, translocations) are common within plant and animal populations (Khan et al., 2020). Thus, the use of a single reference genome hampers our ability to study the full repertoire of genetic variation within a species (Golicz et al., 2016a; Zhao et al., 2018). Structural variants such copy-number variation can contain functional genomic elements that are usually under relaxed selective pressures and can serve as the basis of adaptation given specific environments and selective regimes (Lye and Purugganan, 2019). Coincidentally, copy-number variation and other structural variants play an important role in the emergence of domestication traits, as well as diversification traits in landrace varieties (Lye and Purugganan, 2019). Some studies estimate that at least one third of the known domestication loci are structural variants, and up to one in seven genes can be hemizygous (*i.e.*, with one copy) in grapevine individuals (Zhou Y. et al., 2019). Despite its importance, structural variants cannot be properly analyzed using any of the aforementioned techniques. This led the research community to adopt the concept of the pan-genome, an idea that first appeared in microbiology (Tettelin et al., 2005), into the study of plant and animal genomes (Golicz et al., 2016a).

The concept of pan-genome rests on the idea that the genomes of individuals within a population or species share a core set of genes that unifies them (i.e., the core genome), but also contain a fraction of genes that are absent from one or more individuals (i.e., the accessory or dispensable genome), which altogether give rise to the pan-genome of such population or species (Tettelin et al., 2005).

There are three main methods to generate a pan-genome: the alignment and comparison of multiple *de novo* genome assemblies, the iterative assembly of several genomes from an initial reference or the use of *de Bruijn* graph assemblers to jointly assemble several genomes (Golicz et al., 2016a; see **Table 1**). Since domestication reduces the genetic diversity of a taxon, often eliminating portions of the dispensable genome that contain genes involved in local adaptation, the use of wild relatives is crucial to generate a representative pan-genome for a species (Khan et al., 2020). Once a pan-genome is generated, it can be used alongside whole-genome sequencing data to analyze the structural variants between and within populations, revealing novel loci involved in the development of domestication-related traits that would have stayed hidden when using a single reference genome (Li et al., 2014; Zhao et al., 2018). Besides, the use of a pan-genome alleviates the inherent reference biases of a single reference genome (Günther and Nettelblad, 2019).

Pan-genome studies have revealed additional selective sweeps and structural variants associated to the domestication process, which were not identified using sequencing data with a single reference genome (Li et al., 2014; Zhao et al., 2018). Pan-genomes are already available for several species (**Figure 2**) such as maize (Brohammer et al., 2018), wheat (Montenegro et al., 2017), *Brassica oleracea* (Golicz et al., 2016b) or *Brassica napus* (Hurgobin et al., 2018); and pan-genome analyses to study domestication have already been performed in soybean (Li et al., 2014), rice (Zhao et al., 2018), sunflower (Hübner et al., 2019) and tomato (Gao et al., 2019). While current eukaryote pan-genome analyses are focused on plant species (Golicz et al., 2016a, see **Table 1**) and goats (Li et al., 2019), other livestock researchers may soon venture into this field. As sequencing technologies become cheaper, multiple pan-genomes from different species of the same genus should eventually be combined to create a super-pan-genome that represents the entire genetic content available in a genus with one or more domesticated taxa, as it would include the diversity of all their wild relatives (Khan et al., 2020).

POPULATION GENETICS AND DEMOGRAPHIC ANALYSES OF THE DOMESTICATION PROCESS

Demography and population size changes during the domestication process is tightly related to unraveling some of the most fundamental questions of the domestication process. These analyses can help answer questions such as possible centers of origin and diversification, patterns of migration and expansion throughout these centers, gene flow between domesticated and wild taxa, number of domestication events, the extent of genetic erosion in the domesticated taxon, levels of global genetic differentiation between wild and domesticated taxa, the patterns of adaptive and neutral introgression among them, and in some cases even the number of generations that have elapsed since domestication and other processes such as differentiation and local adaptation of domesticated taxa (Meyer and Purugganan, 2013; Guerra García and Piñero, 2017).

Genetic Diversity in Populations

A first necessary step for the SNP data is to extract and compare the summary statistics of population genetics within and between populations (Andrews et al., 2016). This information describes the genetic diversity in populations, including the estimate of allele frequencies (usually denoted as p or the frequency of the most abundant allele), observed heterozygosity (H_O), expected heterozygosity (H_E), nucleotide diversity (π), number of segregating sites (S) and number of private alleles (*i.e.*, alleles only found in one population). These summary statistics can reveal the level of genetic erosion in domesticated plants and animals when compared to the ancestral wild population, which is expected due to severe bottlenecks, selective sweeps and inbreeding (Groeneveld et al., 2010; Gepts, 2014). One should be aware that reference bias can influence the relative genetic variation observed between the wild and domesticated populations, which could be alleviated using more than one reference or using a pan-genome (Günther and Nettelblad, 2019).

Population Structure

It is also important to describe the population structure (*i.e.*, the genetic differentiation among populations) of domesticated taxa and of their wild relatives, as it can reveal the influence of historical events that shaped the genetic diversity of the organisms (Linck and Battey, 2019). The level of population structure between wild and domesticated taxa can be determined by several factors, such as the number of generations since domestication started, the intensity of the selective pressures imposed to the domesticated taxon, the intensity of the bottlenecks suffered though the domestication process, and the frequency of gene flow between the domesticated taxon and its wild relative (Meyer and Purugganan, 2013).

The F-statistics are classic estimates of population genetics that are based on the heterozygosity values within and among populations, which can reveal patterns of inbreeding, gene flow and differentiation between and within populations (Andrews et al., 2016). Of these, the F_{ST} statistic is of particular interest, since it can be used to detect population structure between wild and domesticated populations, or between different domesticated varieties (Andrews et al., 2016). These estimates are relatively simple to calculate, but they require *a priori* assignment of individuals to discrete populations, which may be wrongly assigned, may not reflect natural populations or may simply be unknown (Linck and Battey, 2019).

Methods based on population clustering have become popular for describing genetic structure, as they do not require *a priori* population assignment. These clustering methods can be classified into parametric and non-parametric methods (Linck and Battey, 2019). Parametric methods, also known as model-based methods, assign individuals into a predefined number of K populations based on their genotypes and the allele frequency of each locus (Pritchard et al., 2000). Several parametric methods have been described that successfully analyze genomic datasets to infer population structure (*e.g.*, Tang et al., 2006; Alexander et al., 2009; Raj et al., 2014), but one has to be careful when using them, as they assume linkage equilibrium and Hardy-Weinberg equilibrium in the dataset (Linck and Battey, 2019), so SNPs should be filtered accordingly before these methods can be confidently used (Wigginton et al., 2005; Mathew et al., 2018). Furthermore, parametric methods have been found to be susceptible to changes in the SFS generated by minor allele frequency thresholds that are commonly used to filter population genomics data because low-frequency polymorphisms are expected to contain information about recent events, which adds uncertainty to the assignation of individuals in populations that reflect ancient demographic events (Linck and Battey, 2019).

Non-parametric methods include principal component analyses, discriminant analyses of principal components and K-means clustering. These methods define populations and genetic structure by transforming the genetic data into uncorrelated variables – named eigenvectors or principal components – to identify groups within the dataset (Patterson et al., 2006; Jombart et al., 2010; Linck and Battey, 2019). Non-parametric methods were designed to work with large amounts

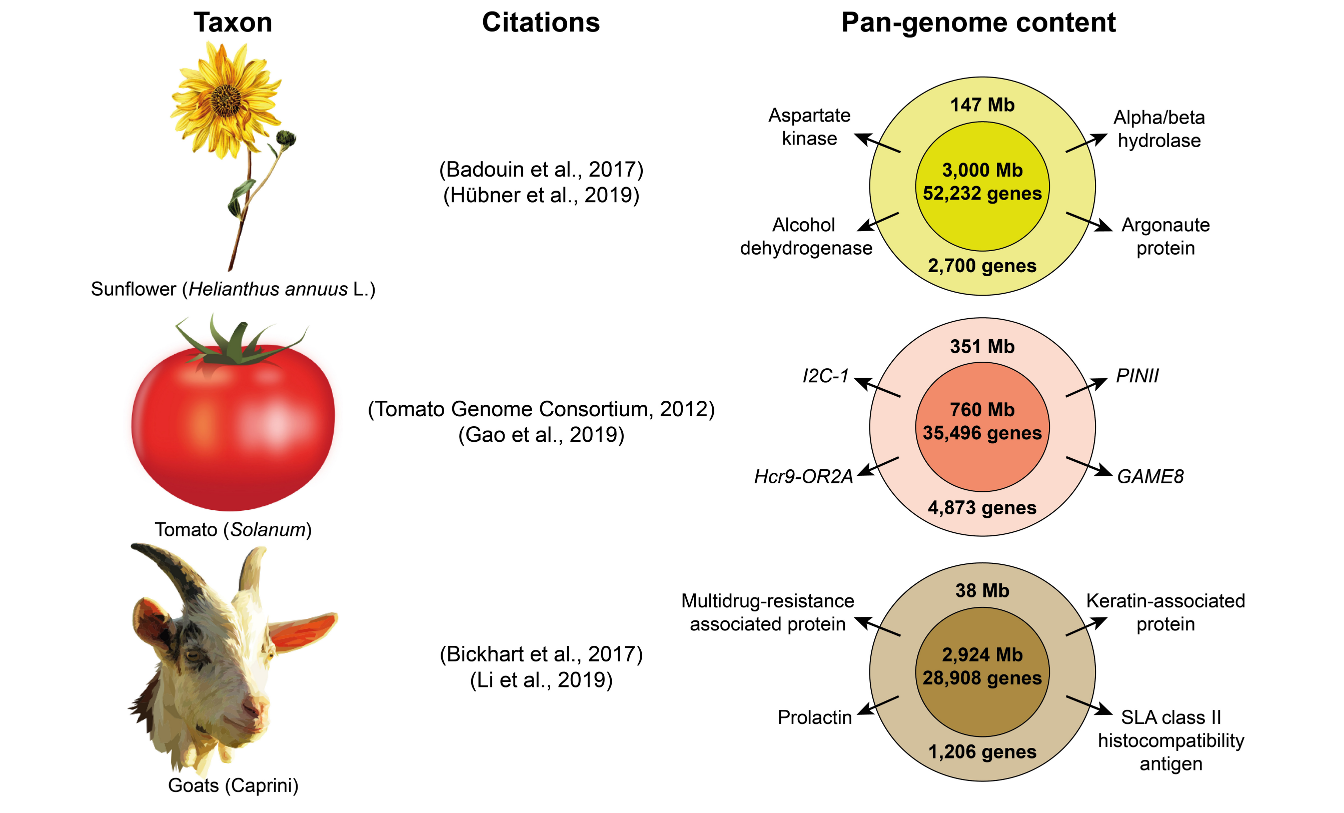


FIGURE 2 | Examples of three pan-genomes of domesticated taxa. The citations included correspond to the publications of the original reference genomes and the subsequent pan-genome assemblies. The inner circles on the right represent the content of the reference genome, while the outer circles represent the additional nonreference content retrieved with the pan-genome. Some examples of important genes are show for the nonreference part of each pan-genome. (images obtained from Openclipart).

of genomic data (Patterson et al., 2006; Jombart et al., 2010) and they are more robust to changes in the SFS than the parametric methods, so it is recommended to run both types of methods and compare their results before making further inferences (Linck and Battey, 2019).

Inferences in Changes of Population Sizes Throughout Time

An important aspect of the demographic history of domesticated taxa is the analysis of the change in the effective population size (N_e) in the populations throughout time (Chen J. et al., 2018). The concept of N_e reflects the estimated populations size in a Wright-Fisher model given an observed genetic variation, so these estimations hardly reflect the census population size of real populations (Charlesworth, 2009), and can also be affected by reference biases and allele dropouts. Changes in N_e can reveal or at least hint on the demographic history of taxa throughout the domestication process, such as expansions or bottlenecks. These changes can help to understand other evolutionary aspects of domestication concerning natural and artificial selection, such as the efficiency of selection and the accumulation of deleterious mutations in domesticated taxa (Chen J. et al., 2018; Allaby et al., 2019).

The domestication process is expected to include a bottleneck as a consequence of subsampling the genetic diversity in the wild ancestor, followed by a population expansion as domesticated taxa diversify (Meyer and Purugganan, 2013), although this idea has been recently challenged by paleogenomic studies (Allaby et al., 2019). Many methods exist to explore the changes in N_e throughout time, whose approach sometimes depends on the type of data available. It should be noted that all the methods to infer historical changes in N_e are susceptible to predicting false bottlenecks when populations are structured, so as indicated above, genetic structure should be evaluated and properly accounted for (Nielsen and Beaumont, 2009).

Studies with few individuals and high sequencing depth may use the Pairwise Sequentially Markovian Coalescent model (PSMC; Li and Durbin, 2011) or the Multiple Sequential Markovian Coalescent model (MSMC; Schiffels and Durbin, 2014) to analyze the demographic history of domesticated and wild taxa. The PSMC and MSMC models can infer changes in N_e throughout time (bottlenecks and expansions) by calculating the distribution of the time of coalescence between all the heterozygous loci in complete diploid genomes (Li and Durbin, 2011; Schiffels and Durbin, 2014). These models can also calculate the time of coalescence (*i.e.*, separation, and in some cases the domestication time) between two genomes given a specified mutation rate, recombination rate and generation time (Li and Durbin, 2011).

However, the genomes used in PSMC or MSMC must be of very good quality, having an average sequencing depth of the very least 18x, at least 10 reads per site, and less than 25% of missing data (Nadachowska-Brzyska et al., 2016). Besides, PSMC has several limitations compared to other estimators of N_{e} and is particularly susceptible to predicting false bottlenecks when populations are structured (Mazet et al., 2015). Nevertheless, this can be properly handled by comparing models of instantaneous N_{e} size change against models of classical symmetric islands using a maximum-likelihood approach (Mazet et al., 2015).

Multiple Sequential Markovian Coalescent can infer more recent changes in N_{e} compared to PSMC (Schiffels and Durbin, 2014), so it may be convenient to explore recent demographic expansions in diversified domesticated taxa (Allaby et al., 2019). For example, MSMC was used to infer population bottlenecks in East Asian and Western Eurasian dogs, as well as divergence times between wolves and dogs around 60,000–20,000 years ago (Frantz et al., 2016), while PSMC was used to determine a severe bottleneck in African rice around 15,000–13,000 years ago (Meyer et al., 2016).

Other methods rely on population data at a genomic scale from many (sometimes hundreds) individuals (as obtained from exome sequencing or RAD-seq), namely the extended Bayesian skyline plots (Heled and Drummond, 2008; Trucchi et al., 2014) and the stairway plots (Liu and Fu, 2015). Since N_{e} is a crucial concept in coalescent theory, extended Bayesian skyline plots and stairway plots rely on the SFS calculated from the population data to estimate N_{e} (Heled and Drummond, 2008; Liu and Fu, 2015). The inferences made from these two methods are comparable to those obtained from PSMC and MSMC, although they rely on different kinds of datasets (Liu and Fu, 2015). Furthermore, stairway plots are more efficient in inferring recent demographic history, whereas PSMC is more reliable for ancient demographic events (Liu and Fu, 2015).

Estimating Gene Flow and Introgression Between Populations

Ancient gene flow and local ancestry (*i.e.*, the genetic ancestry of an individual for an specific chromosomal position; Thornton and Bermejo, 2014) are also important aspects of plant and animal domestication that need to be addressed, since they can describe the genetic contribution of different ancestral populations in the genomic architecture of extant populations, such as wild and domesticated taxa (Price et al., 2009; Pickrell and Pritchard, 2012).

One approach to assess ancient gene flow are graph-based methods that incorporate the possibility of ancient gene flow between distantly related populations (Pickrell and Pritchard, 2012). This type of methods represents the relationships between populations as a bifurcating tree, where internal nodes can also be interconnected forming a graph that represents ancient gene flow that contributed to modern genetic variation (Pickrell and Pritchard, 2012). For example, graph-based analyses have revealed constant gene flow between sympatric populations of domesticated and wild pearl millet (Burgarella et al., 2018), constant gene flow between domesticated and wild pigs (Frantz et al., 2015) but lack of hybridization events between wild and domesticated populations of goats and sheep (Alberto et al., 2018).

Another popular test to infer ancient admixture is the ABBA-BABA test, also known as the *D*-statistic, which evaluates the allelic patterns of three taxa and compares them to an outgroup to identify genomic regions with an excess of shared derived variants that are not concordant to the species tree (i.e., ABBA-BABA patterns), which suggest introgression events (Durand et al., 2011). The $f_{\hat{d}}$ test, which is derived from the *D*-statistic, can help discriminate between introgression events and nonrandom mating in ancestral structured populations (Martin et al., 2015). The *D*-statistic is sensitive to both introgression and incomplete lineage sorting, so both signals can be separated by testing deviations in the symmetry of branch lengths between the gene trees and the species tree (Edelman et al., 2019). By the same logic, the D_3 test can also infer introgression events by analyzing the symmetry in branch lengths, without the need for an outgroup (Hahn and Hibbins, 2019). The *D*-statistic has been used to infer several introgression events between species of the *Bos* genus during domestication (Wu et al., 2018).

On the other hand, local ancestry methods can reveal which chromosomal segments in the genome were inherited from different ancestral source populations (Price et al., 2009). These methods use the data obtained from linkage disequilibrium between loci to assign ancestry in each portion of the genome in comparison to reference populations that depict ancestral source populations, requiring an *a priori* assignation of unadmixed reference populations in order to assign local ancestry to the populations of interest (Price et al., 2009). The analysis reveals chromosomic blocks that can be assigned to either a wild or a domesticated ancestry in hybrid populations, which may reveal historical processes of introgression and local adaptation in modern domesticated populations, as well as potential targets for selective breeding (Janzen et al., 2019).

Many methods exist that can infer local ancestry using genome-wide population data, and all of them require a high-quality reference genome (preferably assembled at a chromosome-level) in order to detect the ancestry of chromosomal segments (*e.g.*, Price et al., 2009; Baran et al., 2012; Maples et al., 2013; Dias-Alves et al., 2018). For example, a local ancestry analysis of East Asian domestic cattle revealed introgressed blocks inherited from ancient banteng and yak populations that contained genes enriched in sensory perception of smell, transmembrane transport and antigen processing (Chen N. et al., 2018).

Using Demographic Simulations to Infer Domestication Scenarios

The previous descriptive tools can help us explore possible evolutionary and demographic scenarios in the absence of *a priori* hypotheses (Liu and Fu, 2015). However, for domesticated taxa we usually have additional classic botanical, zoological, morphological, paleoclimatic, archeological, ethnobiological and biogeographical data that may suggest some likely scenarios (Gerbault et al., 2014). Thus, demographic modeling can be used to test explicit demographic scenarios by comparing simulations of SFS in such scenarios to the observed data (Gerbault et al., 2014; Liu and Fu, 2015). There are many methods available for demographic modeling, which can be more suitable depending on the type of scenarios that need to be tested (Anderson et al., 2005; Gutenkunst et al., 2009; Excoffier and Foll, 2011; Cornuet et al., 2014). All these methods rely on some basic tenets of coalescent theory (Liu and Fu, 2015), so they are also susceptible to possible biases in the observed genetic variation in the populations.

For example, the approximate Bayesian computation (ABC) method compares the summary statistics of several simulated scenarios against the observed data to accept or reject certain demographic hypotheses (Cornuet et al., 2014; Gerbault et al., 2014). This method can help us determine certain parameters of our models and can be used with genome-wide datasets (Cornuet et al., 2014).

Other methods based on diffusion approximation can help us infer the demographic history of multiple populations and their interaction through migration and admixture using biallelic SNP data (Gutenkunst et al., 2009). Demographic modeling has helped test the number of domestication events as well as intercontinental migratory events in cattle (Pitt et al., 2019). Coalescent simulations have supported a common origin for all the domesticated varieties of pearl millet (Burgarella et al., 2018), while the ABC method has revealed that the most likely scenario in the domestication of the scarlet runner bean consists of a single domestication event around 21,000 years ago with a mild bottleneck effect (Guerra-García et al., 2017).

IDENTIFYING GENES UNDER SELECTION DURING DOMESTICATION

Demographic processes are important to understand the general history that led to the domestication of plant and animal taxa, but many studies are specially interested in finding the selected genes that explain the phenotypic differences between domesticated taxa and their wild counterparts (Wang G.-D. et al., 2014; Kantar et al., 2017). Indeed, the detection of these genes under selection during domestication is critical to understand the genetic basis of domestication syndromes, especially for detecting genetic variation relevant for future improvement and selective breeding (Hufford et al., 2012).

When a genetic variant increases its frequency due to positive selection (*i.e.*, selection favoring the fixation of a new allele), the adjacent alleles (*i.e.*, physically connected in the same chromosomal region) also increase their frequency in a process known as hitchhiking (Smith and Haigh, 2007). Once the genetic variant under selection reaches a high frequency or fixation, the hitchhiking effect reduces or even eliminates the genetic variation around the selected locus, producing what is known as a selective sweep (Vitti et al., 2013; Pavlidis and Alachiotis, 2017). The size and intensity of a selective sweep depends on the rate of recombination in the genome, and on the intensity of the selective pressure (Smith and Haigh, 2007), which may be weaker in conscious selection compared to some cases of natural selection (Fugère and Hendry, 2018; Yang et al., 2019). Luckily, the signals of a selective sweep can be detected when the selection event occurred "recently" in an evolutionary timescale, as it is the case for domestication (Vitti et al., 2013).

Different bottom-up methods using population genomics data have been developed to detect the regions in the genome that were selected for during domestication, which we will refer to as candidate loci. We can mention methods for detecting regions with higher population differentiation compared to the rest of the genome, methods for detecting local changes in the SFS throughout the genome, and methods that detect extended regions with strong linkage disequilibrium compared to other haplotypes in the genome (see **Supplementary Table S1** for a summary of methods to detect selective sweeps). Alternatively, a GWAS can be performed to detect the association of a genetic variant to a specific phenotype of interest (Wang G.-D. et al., 2014).

F_{ST} Outlier Tests to Detect Candidate Genes

Besides the standard use of F_{ST} to detect global population structure, the F_{ST} statistic can also be used to detect signals of selective sweeps between populations, namely between wild and domesticated taxa (Gepts, 2014). While a global F_{ST} statistic (involving all the analyzed loci or SNPs) can reveal the overall genetic structure between populations, a local F_{ST} statistic calculated for each locus or SNP along the genome can evaluate whether particular regions of the genome are more differentiated from what is expected due to demographic processes, which can be interpreted as signals of a selective sweep (Nei and Maruyama, 1975). Many different methods exist that are based on the F_{ST} statistic, which are collectively known as F_{ST} outlier tests (Foll and Gaggiotti, 2008; Excoffier et al., 2009; Bonhomme et al., 2010; de Villemereuil and Gaggiotti, 2015; Lotterhos and Whitlock, 2015), that differ mainly on the underlying model used to calculate the null distribution of the F_{ST} values, and thus its ability to detect outliers (**Supplementary Table S1**).

F_{ST} outlier tests are able to detect selective pressures following a bottom-up approach, but their efficiency is determined by a multitude of factors that should be carefully accounted for before using them, such as the sampling scheme used to obtain the population data, the total size of the dataset (*i.e.*, number of populations, of individual per population and of SNPs analyzed), the intensity of the selective pressure, the selfing or allogamous nature of its sexual reproduction, and the migration patterns and genetic structure among populations (De Mita et al., 2013; Lotterhos and Whitlock, 2014, 2015).

Some successful examples in the use of F_{ST} outlier tests include the detection of domestication candidate genes in apple involved in fruit development, size, acidity and sugar metabolism (Khan et al., 2014), the finding of candidate domestication genes involved in metabolism and oil biosynthesis in sunflower

(Baute et al., 2015), the description of candidate diversification genes between pig breeds associated to the shape of the skull (Wilkinson et al., 2013), and the identification of candidate loci between wild and domesticated salmon strains involved in body weight, condition factor, male maturation and a brain related protein (Vasemägi et al., 2012).

Site Frequency Spectrum Based Tests to Detect Selective Sweeps

Selective sweeps alter the SFS that would be expected under neutral evolution processes because of the reduction in the genetic diversity around the loci under selection (Vitti et al., 2013). The genomic region under selection skews the SFS into an excess of high frequency derived alleles when the selective sweep was recent, since the alleles that were linked to the favored selected locus also reach high frequencies (Fay and Wu, 2000). However, after all the high-frequency alleles reached fixation, the genomic region under the selective sweep will have little to no variation, while mutations will slowly generate new allelic variants, skewing the SFS into an excess of low frequency variants (Zeng et al., 2006). Several tests have been developed to detect skews in the SFS, each of them capable of detecting changes in different parts of the SFS (**Supplementary Table S1**), making them complementary to one another (Zeng et al., 2006; Vitti et al., 2013).

Even though SFS based tests are powerful tools to detect selection, it is important to remember that the SFS at the global genomic scale is also altered by demographic events such as bottlenecks that produces an excess of low frequency variants, and expansions that generates an excess of intermediate frequency variants (Vitti et al., 2013). Thus, it is mandatory to have a previous prediction of the demographic history of the populations in order to properly adjust the null hypothesis in each test (Ross-Ibarra et al., 2007).

The well-known summary statistic called Tajima's *D* is sensitive to changes in low-frequency variants, making it particularly useful to detect selective sweeps before and after the selected locus reaches fixation, although low-frequency variants can also be observed in loci under purifying selection (Tajima, 1989; Zeng et al., 2006). Tajima's *D* is also sensitive to intermediate-frequency alleles, making it useful to detect balancing selection (Tajima, 1989) or even some forms of soft selective sweeps generated by standing genetic variation (Prezeworski et al., 2005).

Conversely, Fay and Wu's *H* is sensitive to changes in high-frequency variants, which are only altered by positive selection, making it very useful when used alongside Tajima's *D* (Fay and Wu, 2000). Unlike Tajima's *D*, Fay and Wu's *H* needs an outgroup species in order to differentiate ancestral alleles from derived alleles and thereby to know whether the derived alleles are at high or low frequencies (Fay and Wu, 2000).

Zeng et al. (2006)'s *E* is sensitive to both low and high frequency variants, making it particularly powerful to detect selective sweeps before or after the selected locus reached fixation, also needing an outgroup in order to differentiate derived alleles from ancestral alleles).

There are some tools available to implement SFS based tests using genome-wide data, that can perform all the above tests (*i.e.,* Korneliussen et al., 2013, 2014; Rozas et al., 2017). For example, Tajima's *D* test was used alongside other methods to detect selective sweeps associated to the domestication of yaks (Qiu et al., 2015), Zeng's *E* test helped discover 125 selective sweeps associated to the domestication of horses (Librado et al., 2016), and the complementary implementation of Tajima's *D*, Fay and Wu's *H* and Zeng's *E* revealed several candidate genes that share similar functions between peach and almond (Velasco et al., 2016).

The reduction of diversity (ROD) test is another popular SFS-based method to detect selective sweeps that has been particularly useful for the study of domestication (**Supplementary Table S1**). ROD compares local π values of domesticated taxa against the local π values of its wild relatives, using sliding windows alongside the genome (Guo et al., 2012; Huang et al., 2012; Qi et al., 2013; Schmutz et al., 2014). The ROD method has been used to successfully detect candidate domestication genes in rice (Huang et al., 2012), watermelon (Guo et al., 2012), cucumber (Qi et al., 2013), common bean (Schmutz et al., 2014), and chickpea (Varshney et al., 2019), to name a few.

Linkage Disequilibrium (LD) Based Methods to Detect Selection

Given that selective sweeps remove the variation in regions adjacent to the locus under selection, they can form haplotype blocks that extend in strong LD compared to other haplotypes in the same locus because they reached a medium-to-high frequency in the population swift enough so they are not yet disrupted by recombination (Sabeti et al., 2002; Vitti et al., 2013). This pattern has been exploited to develop several methods based on LD to detect selective sweeps of recent origin (Vitti et al., 2013). Interestingly, LD-based methods are sensitive enough to detect both strong and soft selective sweeps (Garud et al., 2015), as well as partial or incomplete selective sweeps (Vitti et al., 2013), making them excellent tools to study recent and ongoing selection events, such as those occurring during domestication and the subsequent diversification of landraces (**Supplementary Table S1**).

Since the above rationale relies on LD decay due to recombination, any method based on LD requires to control for local variation in recombination rates in order to reduce false positives (Sabeti et al., 2002). The extended haplotype homozygosity (EHH) is a widely used statistic in LD-based methods that is defined as the probability that two orthologous genomic regions carrying a "core" haplotype of interest (*i.e.*, the part of the haplotype that is shared by all the individuals carrying it, such as the allele under positive selection) in the population are identical by descent (*i.e.,* they were inherited by the same ancestor), as one looks to a specified distance farther away from the core region (Sabeti et al., 2002).

Among the LD based methods that uses the EHH, we can mention the long-range haplotype (LRH) test, sometimes named the relative EHH (rEHH) test, which controls for local recombination rates by comparing the EHH of several haplotypes localized within the same locus (Sabeti et al., 2002). Other EHH based methods include the whole-genome long-range haplotype (WGLRH) test that uses sliding windows to perform the LRH test (Zhang et al., 2006), the long-range haplotype similarity (LRHs) test (Hanchard et al., 2006), the integrated haplotype score (iHS) which is particularly sensitive to incomplete selective sweeps and soft sweeps (Voight et al., 2006) and the cross-population extended haplotype homozygosity (XP-EHH) statistic that is able to detect selective sweeps after the selected allele reached fixation (Sabeti et al., 2007). The iHS and the XP-EHH statistics can be regarded as complementary to each other, enabling the detection of incomplete and complete selective sweeps in the target population (Vitti et al., 2013).

All the LD-based tests that make use of the EHH statistic require the previous phasing of the chromosomes in order to work (*i.e.*, assignation of alleles in an individual to their corresponding maternal and paternal haplotypes), which may or may not be possible depending on the sequencing depth and type of data available for the analysis (Delaneau et al., 2013). For instance, a reference genome is usually needed in order to phase genotypes, since most methods rely on the information of proximity between alleles and their distribution within individuals in a population to assign haplotypes (Delaneau et al., 2013) although new methods are emerging that can phase genotypes without a reference genome (Money et al., 2017).

There are other LD-based methods that do not make use of the EHH statistic, such as the LD decay (LDD) test, which rely on individuals that are homozygous for any given SNPs to look for LD differences between alleles in a population (Wang et al., 2006) or the ω statistic that scans for high SNP correlation coefficients around a site under selection (Kim and Nielsen, 2004; Alachiotis et al., 2012). Another method that do not require chromosome phasing is the regression-based test, which relies on the reduction of heterozygosity as one approaches the locus under selection in a genome to infer selective sweeps (Wiener and Pong-Wong, 2011). Other LD-based methods exploit the estimation of identity-by-descent using genome-wide data to detect haplotypes that are shared between several unrelated individual (> 10 generations) to infer selective sweeps without previous knowledge of the pedigree of individual (Han and Abney, 2013), so they might prove useful to study recent domestication processes.

Some examples of LD-based methods used to explore the domestication process includes an analysis using LRH to detect signatures of selection associated to dairy and beef cattle breeds (Bomba et al., 2015), a study using the XP-EHH statistic to find signals of selective sweeps in Jinhua pigs (Li et al., 2016), and a paper focused on the diversification of goat landraces that calculated the iHS and the XP-EHH statistics alongside other tests to detect selective sweeps between goat breeds (Bertolini et al., 2018).

Other important tests include the XP-CLR test (Chen et al., 2010) and the μ statistic (Alachiotis and Pavlidis, 2018) which implement multiple signatures to detect selective sweeps (**Supplementary Table S1**) and have been used to detect candidate loci in maize and African rice, respectively (Hufford et al., 2012; Ndjiondjop et al., 2019).

Using GWAS to Detect Domestication-Associated Loci

Genome-wide association studies have been used extensively to uncover the genetic variants that underlie domestication traits (Shi and Lai, 2015). The domestication traits that can be analyzed through a GWAS can encompass any biological characteristic from simple morphological traits (Jiao et al., 2012) to the production of certain metabolites (Shang et al., 2014), tame behavior in animals (Ilska et al., 2017), resistance or susceptibility to certain diseases (Wang et al., 2012), or adaptation to certain environmental conditions (Song et al., 2018).

An important advantage of the GWAS over the bottom-up approaches is its ability to detect polygenic effects on single traits of interest, which is commonplace considering that genes interact between them and the environment to generate phenotypes (Gibson, 2018).

A prerequisite before preforming a GWAS is to have large sample sizes in both the number of sequenced genetic variants and the number of individuals included in the study, as they are necessary to obtain the statistical power to detect variants with small effects and to reduce the risk of false positives (Wang G.-D. et al., 2014).

Some recent examples include the use of a GWAS to identify candidate genes with unknown functions involved in several agronomic traits, including drought and heat tolerance in chickpea (Varshney et al., 2019); a GWAS that revealed loci associated to fruit size and quality in peach (Cao et al., 2019); and a GWAS that uncovered the genetic variants involved in the absence of anthocyanin in domesticated rice compared to its wild relative (Zheng et al., 2019).

ANCIENT DNA AND PALEOGENOMICS OF DOMESTICATED TAXA

Extant domesticated taxa lack the information of ancient genetic diversity that was lost through bottlenecks, selection and genetic drift (Ramos-Madrigal et al., 2016). However, the analysis of ancient DNA can allow the research community to overcome some of these limitations (Irving-Pease et al., 2019). Ancient DNA retrieved from archeological sites allows the study of the rate at which domestication happened, as well as revealing which genes were important at the beginning of this process (Vallebueno-Estrada et al., 2016; Irving-Pease et al., 2019). Thus, paleogenomics is becoming a novel research area for understanding the process of plant and animal domestication (Irving-Pease et al., 2019).

Extraction and Sequencing of Ancient DNA

An important limitation of paleogenomic analyses is the level of preservation of the ancient DNA itself, as well as the total yield of extracted DNA (Sawyer et al., 2012). The DNA molecules that are extracted from tissues that are not conserved on permafrost and are older than 100 years are usually shorter than 100 bp (Sawyer et al., 2012). The strand breaks of such fragments are also non-random, as purines are enriched before the strand breaks (Sawyer et al., 2012). Additionally, these fragments incorporate cytosine-to-uracil mutations on their ends, further hindering the analysis of the sequenced fragments (Sawyer et al., 2012). Even though these characteristics hamper the sequencing and analysis of ancient DNA, they are also useful to differentiate between real ancient DNA and extant DNA contamination (Sawyer et al., 2012). Furthermore, due to the scarce ancient material located throughout few archeological sites worldwide, sample sizes in paleogenomic studies are very small, usually one or few individuals per location and sometimes only one locality (*e.g.,* Wales et al., 2016; Ramos-Madrigal et al., 2016).

Given the above difficulties and the uniqueness of the biological material retrieved from archeological sites, it is crucial to extract and sequence as much ancient DNA as possible while avoiding DNA contamination (Gamba et al., 2016). Major efforts have been made to develop efficient protocols for ancient DNA extraction (Gamba et al., 2016) and single-strand library preparation for high-throughput sequencing (*e.g.,* Gansauge et al., 2017). Organelle genomes were usually the target for ancient DNA sequencing because multiple copies of these can be found within each plant and animal cell and can reveal several demographic processes (Wales et al., 2016; Irving-Pease et al., 2019). Nonetheless, more evolutionary information can be retrieved from nuclear DNA, which is the main target for modern paleogenomic studies (Wales et al., 2016; Irving-Pease et al., 2019).

Insights of Paleogenomic Data in Domestication

Paleogenomic studies are challenging some of our previous ideas of the domestication process, such as the occurrence of ancient domestication bottlenecks, which appear to be absent in several archeological plant genomes, suggesting that the reduced diversity in domesticated taxa may be a more gradual process from what was expected using DNA of extant populations (Allaby et al., 2019). For example, several archeological samples of *Sorghum bicolor* from different time periods (ranging from 1800 to 100 years ago) were compared to extant individuals of the species, revealing that this crop did not suffered an initial domestication bottleneck, but rather that the reduction in genetic diversity, and its associated mutational load, occurred gradually throughout time (Smith et al., 2019).

Paleogenomics is also revealing important aspects of plant and animal domestication, such as the first genetic steps towards domestication syndromes as well as the overall graduality of the process (Ramos-Madrigal et al., 2016; Vallebueno-Estrada et al., 2016; Daly et al., 2018). For example, archeological remains of goat populations have revealed multiple domestication processes in ancient wild goats, possible dispersal routes of ancient goat populations and signs of early selective pressures towards candidate genes involved in pigmentation, milk production, size, reproduction and changes in diet (Daly et al., 2018). Likewise, several archeological maize samples retrieved from the Tehuacán Valley in Mexico have revealed that early domesticates already presented signals of selective sweeps on important candidate genes, such as *teosinte branched1* and *brittle endosperm2*, but lacked selective sweep signals in other important candidate genes present on modern maize populations, even though these ancient maize populations were already endogamous and more closely related to modern maize than to wild teosinte, revealing that maize domestication was a gradual process ranging thousands of years (Ramos-Madrigal et al., 2016; Vallebueno-Estrada et al., 2016).

Other examples demonstrate the importance of paleogenomic studies in domesticated taxa, including grapevine (Wales et al., 2016), barley (Mascher et al., 2016), sunflower (Wales et al., 2019), horses (Schubert et al., 2014), dogs (Frantz et al., 2016) and cats (Ottoni et al., 2017).

RNA SEQUENCING TO DETECT DIFFERENTIALLY EXPRESSED GENES ASSOCIATED TO DOMESTICATION

Besides the use of RNA-seq to obtain population-level data, comparative transcriptomics is a good way to find or support the validity of candidate genes (Hekman et al., 2015). Transcriptomic analyses between domesticated and wild taxa can reveal important changes in gene expression associated to domestication (Koenig et al., 2013; Hekman et al., 2015; Hradilová et al., 2017). Likewise, the analysis of hybrids between domesticated and wild individuals can reveal important patterns of allele-specific regulation and the role of *cis*/*trans* regulatory elements in the emergence of domestication traits (Bell et al., 2013; Lemmon et al., 2014).

The Experimental Design of Differential Expression Analyses

Transcriptomic profiles are tissue-specific and time-dependent (Hekman et al., 2015). Thus, a good experimental design can reveal important loci involved in the phenotypic differences associated to domestication syndromes, such as suppression of secondary metabolites, changes in form, size, taste, absence of defense mechanisms, seed dormancy, docile behavior, among other traits (Hekman et al., 2015). This can be done by comparing the total RNA expression of the tissue or organ of interest (Koenig et al., 2013), as well as comparing RNA expression throughout the developmental stages of such tissue or organ (Hradilová et al., 2017).

Since transcriptomic analyses are experimental by nature, experimental designs require biological replicates for each treatment, condition or organ to assess the variability in the data; as well as controlled environmental conditions to reduce possible biases and sources of error (Fang and Cui, 2011; Schurch et al., 2016). Empirical studies recommend using at least six biological replicates for each condition in the experiment, even though the use of three replicates is common, but discouraged (Burden et al., 2014; Schurch et al., 2016). Additionally, it is important to avoid committing errors in the experimental design that can bias the results of the RNA-seq experiment, such as using different sequencing technologies for each sample, using different methods for library preparations throughout the samples, sequencing each treatment in a different sequencing flowcell or different lanes within a flowcell (Fang and Cui, 2011). Other technical biases associated to adapter ligation and within-lane variation can be properly assessed when using biological replicates (Fang and Cui, 2011; see **Table 1**).

RNA-seq data can also be complemented with metabolomic data to infer the association between the differential expression of genes and the presence/absence of metabolites between wild and domesticated taxa (Hradilová et al., 2017).

After obtaining high-quality data with an appropriate experimental design, RNA-seq analyses usually follow a similar workflow, which should culminate in the detection of differentially expressed genes between a wild plant and its domesticated counterpart (Yang and Kim, 2015; see **Table 1**). These differentially expressed genes are most likely candidates that may explain to some degree the changes associated to domestication (Koenig et al., 2013; Hradilová et al., 2017). Nonetheless, one must be careful while interpreting the results of these studies, as some differentially expressed genes between wild and domesticated taxa may be a consequence, rather than a cause, of the domestication traits under study (Albert et al., 2012).

Successful Examples of RNA-seq Experiments to Understand Domestication

RNA-seq analysis has been successfully employed to discover differentially expressed genes involved in the domestication of several plant species. For example, RNA-seq analyses between maize and teosinte found 600 differentially expressed genes and 1,100 genes with altered patterns of co-expression, mainly involved in biotic stress responses, and many of which were previously found as candidate genes using selective scans (Myers et al., 2012). Similar results have been found in tomato (Koenig et al., 2013), pea (Hradilová et al., 2017), common bean (Singh et al., 2018), and carrot (Machaj et al., 2018). This approach has also led to the discovery of differentially expressed genes between dogs and wolves associated to tameness (Li et al., 2013), as well as changes related to the immune system and aerobic capacity (Yang et al., 2018). Another study found differential isoform expression between wild and domesticated sorghum accessions, revealing that domestication can alter the patterns of alternative spicing (Ranwez et al., 2017). Hybrid studies have been performed between maize and teosinte, suggesting potential selection on *cis* regulatory elements associated with changes in ear tissue and previously reported candidate genes (Lemmon et al., 2014). Another hybrid study in *Capsicum annuum* using network analyses revealed that loss of function in *cis* regulatory sequences lead to transcriptional changes in *trans* elements that are associated with fruit morphology (Díaz-Valenzuela et al., 2020).

MODERN EPIGENOMICS AND METHODOLOGICAL STRATEGIES TO EXPLORE DOMESTICATION

Epigenetics is classically defined as the heritable mechanisms that regulate gene expression without direct modifications to the DNA sequence, namely DNA methylation, RNA methylation, covalent histone modifications and chromatin assembly states (Sakurada, 2010; Zhao et al., 2017). Epigenetic variants, sometimes called epialleles, are local differences in these epigenetic marks between individuals in a population, which can have similar dynamics to genetic variants (Weigel and Colot, 2012; Guo et al., 2015). Since epigenetic mechanisms underly the ability of organisms to respond to changing environmental conditions, some epigenetic marks associated to these responses are more susceptible to change due to environmental input, while other marks involved in cell differentiation, embryonic development and core cellular functions might be more stable (Turner, 2009).

Most of the domestication studies that explain phenotypic differences between wild and domesticated taxa focus on genetic variation. However, the study of epigenomics may explain some of the missing heritability in domestication traits (i.e., the gap between the heritability of a trait estimated by classic genetics and GWAS), the patterns of differentially expressed genes that do not have clear signs of selective sweeps, or even connect the causality between the genetic variation that was selected for during domestication and the resulting phenotypes (Schmitz et al., 2013; Trerotola et al., 2015; Janowitz Koch et al., 2016; Bélteky et al., 2018).

Epigenetic variation can be inherited from one generation to the next in a process known as trans-generational epigenetic inheritance, which has been documented in plants and animals (Heard and Martienssen, 2014), even though the overall importance of this trans-generational epigenetic inheritance in plant and animal evolution is still debated (see **Table 1**). Nevertheless, we consider that studying epigenetic patterns associated to transcriptional activity and phenotypic traits should help understand the emergence of domestication phenotypes (Bélteky et al., 2018). If epigenetic variants such as single methylation polymorphisms (SMPs) show complete transgenerational inheritance, they can even be analyzed using the theoretical tools of population genetics to detect selective sweeps (Schmitz et al., 2013; Janowitz Koch et al., 2016).

In a similar fashion to GWAS, the use of epigenome-wide association studies (EWAS) can also reveal the association of an epigenetic variant to a trait of interest in domesticated taxa (Feeney et al., 2014). The same precautions taken in transcriptomic data should also be taken for epigenomic data, since the patterns of epigenetic marks in organisms are tissue-specific, time-dependent and sensitive to environmental input, meaning that epigenomic data should be analyzed for specific organs or tissues of interest in a controlled environment (Jensen, 2015). This is particularly important for the epigenetic marks that respond to environmental input, since domesticated taxa and their wild relatives live under different environmental conditions. Growing both taxa under controlled conditions will alter the natural state of these marks, but will also help differentiate the heritable epialleles associated to domestication traits (Turner, 2009).

Obtaining Population Data From Epigenetic Marks

The most studied epigenetic mark is DNA 5-methylcytosine, which refers to the DNA methylation in cytosines which are usually associated to transcriptional gene silencing (He et al., 2011). Cytosine methylome data can be obtained using high-throughput sequencing technologies alongside bisulfite sequencing (Meissner, 2005). Bisulfite sequencing consists in the deamination of unmethylated cytosines through a bisulfite reaction, converting them into uracil, which are encoded as thymine by sequencing technologies (Frommer et al., 1992). The comparison of sequenced DNA that was treated with bisulfite alongside sequenced DNA without treatment can discriminate between methylated and unmethylated cytosines in an organ, tissue or cell-type of interest (Frommer et al., 1992).

Reduced representation bisulfite sequencing (RRBS) is a high-throughput technique with a similar rationale to RAD-seq that enriches the sequencing of CG rich regions of the genome after the digestion of restriction enzymes (Meissner, 2005). This makes the RRBS technique a cost-effective option to analyze cytosine methylation patterns in mammals, since its cytosine DNA methylation happens at CG sites (Meissner, 2005; He et al., 2011). Plant cytosine methylomes should instead be analyzed through MethylC-seq, which consists of whole-genome sequencing and bisulfite treatment (Urich et al., 2015), as cytosine methylation can also happen in CHG and CHH sites in plant genomes (He et al., 2011). Cytosine methylation can also be detected using methylated DNA immunoprecipitation sequencing (MeDIP-seq), which consists in shearing the genomic DNA into small pieces followed by the immunoprecipitation of the methylated cytosines using antibodies that recognizes 5-methylcytosine and finally sequencing the DNA sequences with the methylated sites using standard high-throughput sequencing technologies (Weber et al., 2005).

Besides cytosine methylation, adenine has also been shown to be methylated in both plants and animals (N6-methyldeoxyadenosine), which cannot be detected using bisulfite sequencing (Luo et al., 2015). However, genomic regions with methylated adenines can be detected using N6-methyldeoxyadenosine immunoprecipitation sequencing (6mA-IP-seq), which uses the same rationale as MeDIP-seq but requires antibodies that specifically targets N6-methyldeoxyadenosine (Fu et al., 2015). PacBio and Nanopore sequencing technologies are known to be sensitive to DNA methylation, regardless of it being on a cytosine or adenine, so they are currently being used as powerful, albeit expensive tools to evaluate DNA methylation patterns in genomes (Gouil and Keniry, 2019).

Histone modifications refers to either posttranslational covalent modifications in histones (methylations, acetylations, phosphorylations, ubiquitylations, ADP-ribosylations, sumoylations, crotonylations, malonylations, succinylations) or the substitution of canonical histones by histone variants with different amino acid composition (Bowman and Poirier, 2015). These histone modifications determine the functionality of local genomic regions by changing the state of the chromatin either through its direct effects on the chemical interactions between DNA and histones or through the recruitment of chromatin remodeling complexes (Bowman and Poirier, 2015).

Chromatin immunoprecipitation sequencing (ChIP-seq) can be used to assess the genome-wide association between DNA regions and specific histone modifications (Schmidt et al., 2009). ChIP-seq consists in the initial fixation of DNA-protein interactions using formaldehyde followed by DNA fragmentation and subsequent enrichment of the target histone modification using magnetic beads coupled to antibodies in order to sequence the genomic regions where the histone modification is present (Schmidt et al., 2009). ChIP-seq can also be used to assess the interaction between any DNA-binding protein such as transcriptional factors and specific genomic regions (Schmidt et al., 2009).

Epigenomic Studies Applied to Understand Domestication

The current epigenomic analyses regarding domestication have focused on DNA methylation patterns (Jensen, 2015; Ding and Chen, 2018), but some studies have also ventured into histone modification patterns (He et al., 2014). Recent efforts are trying to connect the discoveries of genomics and epigenetics to understand the evolution of tameness in domesticated animals (Jensen, 2015). A study using RRBS that compared the DNA methylation patterns between wolves and dogs revealed signals of natural selection acting on SMPs which are enriched in transposons and genes involved in the regulation of neurotransmitters, suggesting a dog-specific silencing of genes involved in behavior (Janowitz Koch et al., 2016). Similarly, a recent study using MeDIP-seq in red junglefowl populations that were bred to have either high or low fear to humans discovered genomic region that were differentially methylated in genes that were previously related to tameness (Bélteky et al., 2018).

Other studies focused on plant domestication have found differentially methylated sites associated to domestication syndromes (Song et al., 2017; Shen et al., 2018). A study using MethylC-seq found 519 differentially methylated genes between domesticated and wild cotton from which some of them are associated with the observed differences in flowering time and seed dormancy between the wild and domesticated taxa (Song et al., 2017). Another study using MethylC-seq found 4,248 differentially methylated regions between wild and domesticated soybean and 1,164 differentially methylated regions between domesticated and improved soybean (Shen et al., 2018). As expected, the differentially methylated regions in soybean had higher genetic diversity compared to the regions with evidence of selective sweeps that were previously found, and interestingly, 22.5% of the differentially methylated sites could be associated to a causal genetic variant (suggesting that these genetic variants were responsible for the observed epigenetic patterns), whereas the rest of the differentially methylated regions could be interpreted as genuine epialleles located within genes involved in carbohydrate metabolism (Shen et al., 2018).

EXPERIMENTAL VALIDATION OF CANDIDATE GENES

Once we have evidence of candidate genes involved in the domestication syndrome, the necessary next step to understand the genetic basis of domestication is to design *in vitro* systems, knock-out, knock-down or knock-in experiments that validate the involvement of such genes in the observed phenotypes (Zhang et al., 2017). This can be performed either by direct alteration of the genome in the organism of interest, by using RNA interference or by designing heterologous systems in a model organism (Boettcher and McManus, 2015). As an example, a knock-out experiment with backcrosses between domesticated and wild mice elucidated the role of some genes involved in behavioral changes associated to mouse domestication (Chalfin et al., 2014).

Previous knock-out and knock-in experiments were restricted to model organisms, but nowadays experimental validation of candidate genes can be supported via knock-out and knock-in experiments, using novel genome editing tools (*e.g.*, Shalem et al., 2014; Hahn et al., 2017; Ueta et al., 2017). Genome-editing tools are already available for a broad range of taxa, including dozens of crop species, but developing a working system in non-model organisms can still be a difficult task that can take several months or even years to accomplish (Shan et al., 2020), so doing collaborative studies alongside experimental researchers is recommended. In this moment, the leading toolset to perform genome editing is the Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR) system alongside the CRISPR associated protein 9 (Cas9), commonly known as CRISPR/Cas9, which can be used to eliminate, introduce or replace specific segments of DNA within a targeted site in a genome (Cong et al., 2013). Another useful tool for genome editing is the Transcription Activator-Like Effector Nuclease (TALEN) technology, which has its own advantages in comparison to CRISPR/CAS9 (Zhang et al., 2017). RNA interference can also help in validating the function of candidate genes, although it is limited to knock-down experiments (Boettcher and McManus, 2015). Heterologous expression in model organisms is a cost-effective alternative to validate candidate genes (e.g., Schweiger et al., 2010), although this method overlooks the interaction networks that exist *in vivo* which are accountable for the emergence of phenotypes (Rodríguez-Mega et al., 2015).

Regardless the genome-editing tool of choice (Boettcher and McManus, 2015; Zhang et al., 2017), genome edition is proving its usefulness to validate the effect of candidate genes involved in domestication through the introduction of domesticated alleles on wild relatives and vice-versa (Zhou J. et al., 2019), which can prove that the gene is indeed involved in the appearance of the domesticated phenotype (Zhou J. et al., 2019). This can be performed in the same way as a usual knock-out or knock-in experiment, where the edited locus must be validated through PCR and Sanger sequencing, a PCR-RFLP analysis or using Western-blot in case of a protein knock-out (*e.g.,* Ueta et al., 2017). The expected result of these type of studies is to find a modified phenotype after editing a candidate locus, either a wild individual with a domesticated-like phenotypic trait or a domesticated individual with a wild-like phenotypic trait (Zhou J. et al., 2019).

Of course, the above studies will hardly reproduce a complete domesticated or wild phenotype, since genetic elements interact in complex regulatory networks, including other elements within the genome as well as epigenetic and environmental components (Rodríguez-Mega et al., 2015), but nonetheless will be useful to understand the role of those genes in the emergence of domesticated phenotypes.

Once the candidate genes are validated, genome-editing tools can also become useful to introduce desirable traits from wild relatives to its domesticated counterparts, a goal of great interest for crop improvement (Zhou J. et al., 2019) and currently used to accelerate plant breeding and to fine-tune desirable traits (Wolter et al., 2019). Furthermore, recent efforts are trying to domesticate plant crops *de novo* by inserting the desired domestication alleles into their wild relatives, generating crops with the desired domestication phenotypes but without the problems of low genetic variation and accumulation of deleterious mutations that are an inevitable consequence of regular domestication processes (Fernie and Yan, 2019).

CONCLUSION AND PERSPECTIVES

Plant and animal domestication can be studied using genomic, transcriptomic and epigenomic strategies, revealing the action of evolutionary, ecological and anthropogenic processes (Kantar et al., 2017). These tools can lead us beyond the description of the possible historical scenarios that shaped the domesticated species, since we can explore the effects of domestication on the transcriptomic activity of a species (Hekman et al., 2015), test the validity of candidate genes associated to domestication phenotypes (Zhou J. et al., 2019) and analyze epigenetic patterns associated to domestication traits (Jensen, 2015). Many domesticated taxa remain genetically unexplored, and as sequencing technologies become cheaper and more efficient, domestication genomics will soon be available for polyploids and species with huge genomes (*e.g.,* Edger et al., 2019).

Nonetheless, the modern study of domestication of plants and animals should still be multidisciplinary, since genetics only tells us part of the story (Larson et al., 2014). An extended synthesis framework should also be considered to understand domestication, as these new studies are helping us understand niche construction and the emergence of domesticated phenotypes (Piperno, 2017). Other potential lines of work remain to be addressed in domestication studies, such as the changes in the chromatin architecture (e.g., Concia et al., 2020), the use of comparative proteomic atlases (e.g., Jiang Y. et al., 2019) and the analysis of cell-type divergences during development using single-cell RNA-seq data (Arendt et al., 2016). The use of this multi-omic approaches will help us create and compare developmental atlases (e.g., Walley et al., 2016) between wild and domesticated taxa to understand how morphology diverged during domestication.

AUTHOR CONTRIBUTIONS

JB-R, DP, and LE wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) KE004 "Diversidad genética de las especies de *Cucurbita* en México e hibridación entre plantas genéticamente modificadas y especies silvestres de Cucurbita" and CONABIO PE001 "Diversidad genética de las especies de Cucurbita en México. Fase II. Genómica evolutiva y de poblaciones, recursos genéticos y domesticación, both awarded to Rafael Lira-Saade and LE. DP was funded by Instituto de Ecología, UNAM". JB-R is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México and received fellowship 583146 from Consejo Nacional de Ciencia y Tecnología (CONACyT).

ACKNOWLEDGMENTS

We acknowledge the Doctorado en Ciencias Biomédicas for the academic support provided to JB-R during the development of this project. We thank Alejandra Moreno Letelier for her insights, which greatly improved the quality of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00742/full#supplementary-material

REFERENCES

Alachiotis, N., and Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Commun. Biol.* 1, 1–11. doi: 10.1038/s42003-018-0085-8

Alachiotis, N., Stamatakis, A., and Pavlidis, P. (2012). OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics* 28, 2274–2275. doi: 10.1093/bioinformatics/bts419

Albert, F. W., Somel, M., Carneiro, M., Aximu-Petri, A., Halbwax, M., Thalmann, O., et al. (2012). A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet.* 8:e1002962. doi: 10.1371/journal.pgen.1002962

Alberto, F. J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., de Villemereuil, P., et al. (2018). Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.* 9:813. doi: 10.1038/s41467-018-03206-y

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., et al. (2016). RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics* 202, 389–400. doi: 10.1534/genetics.115.183665

Allaby, R. G., Ware, R. L., and Kistler, L. (2019). A re−evaluation of the domestication bottleneck from archaeogenomic evidence. *Evol. Appl.* 12, 29–37. doi: 10.1111/eva.12680

Anderson, C. N. K., Ramakrishnan, U., Chan, Y. L., and Hadly, E. A. (2005). Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics* 21, 1733–1734. doi: 10.1093/bioinformatics/bti154

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92. doi: 10.1038/nrg.2015.28

Arendt, D., Musser, J. M., Baker, C. V., Bergman, A., Cepko, C., Erwin, D. H., et al. (2016). The origin and evolution of cell types. *Nat. Rev. Genet.* 17, 744–757. doi: 10.1038/nrg.2016.127

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 546, 148–152. doi: 10.1038/nature22380

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 1359–1367. doi: 10.1093/bioinformatics/bts144

Baute, G. J., Kane, N. C., Grassa, C. J., Lai, Z., and Rieseberg, L. H. (2015). Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytol.* 206, 830–838. doi: 10.1111/nph.13255

Bell, G. D., Kane, N. C., Rieseberg, L. H., and Adams, K. L. (2013). RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biol. Evol.* 5, 1309–1323. doi: 10.1093/gbe/evt072

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F. C., Falentin, C., et al. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* 4, 879–887. doi: 10.1038/s41477-018-0289-4

Bélteky, J., Agnvall, B., Bektic, L., Höglund, A., Jensen, P., and Guerrero-Bosagna, C. (2018). Epigenetics and early domestication: differences in hypothalamic DNA methylation between red junglefowl divergently selected for high or low fear of humans. *Genet. Sel. Evol.* 50:13. doi: 10.1186/s12711-018-0384-z

Bertolini, F., Servin, B., Talenti, A., Rochat, E., Kim, E. S., Oget, C., et al. (2018). Signatures of selection and environmental adaptation across the goat genome post-domestication. *Genet. Sel. Evol.* 50:57. doi: 10.1186/s12711-018-0421-y

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., et al. (2017). Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nat. Genet.* 49, 643–650. doi: 10.1038/ng.3802

Boettcher, M., and McManus, M. T. (2015). Choosing the right tool for the job: RNAi, TALEN, or CRISPR. *Mol. Cell* 58, 575–585. doi: 10.1016/j.molcel.2015.04.028

Bomba, L., Nicolazzi, E. L., Milanesi, M., Negrini, R., Mancini, G., Biscarini, F., et al. (2015). Relative extended haplotype homozygosity signals across breeds reveal dairy and beef specific signatures of selection. *Genet. Sel. Evol.* 47:25. doi: 10.1186/s12711-015-0113-9

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et al. (2010). Detecting selection in population trees: the lewontin and krakauer test extended. *Genetics* 186, 241–262. doi: 10.1534/genetics.104.117275

Bowman, G. D., and Poirier, M. G. (2015). Post-translational modifications of histones that influence nucleosome dynamics. *Chem. Rev.* 115, 2274–2295. doi: 10.1021/cr500350x

Brohammer, A. B., Kono, T. J. Y., and Hirsch, C. N. (2018). "The maize pan-genome," in *The Maize Genome. Compendium of Plant Genomes*, eds J. Bennetzen, S. Flint-Garcia, C. Hirsch, and R. Tuberosa (Cham: Springer), 13–29. doi: 10.1007/978-3-319-97427-9_2

Brozynska, M., Furtado, A., and Henry, R. J. (2016). Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol. J.* 14, 1070–1085. doi: 10.1111/pbi.12454

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., et al. (2017). A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Rep.* 18, 762–776. doi: 10.1016/j.celrep.2016.12.063

Burden, C. J., Qureshi, S. E., and Wilson, S. R. (2014). Error estimates for the analysis of differential expression from RNA-seq count data. *PeerJ* 2:e576. doi: 10.7717/peerj.576

Burgarella, C., Cubry, P., Kane, N. A., Varshney, R. K., Mariac, C., Liu, X., et al. (2018). A western Sahara centre of domestication inferred from pearl millet genomes. *Nat. Ecol. Evol.* 2, 1377–1380. doi: 10.1038/s41559-018-0643-y

Burggren, W. (2016). Epigenetic inheritance and its role in evolutionary biology: re-evaluation and new perspectives. *Biology (Basel).* 5:24. doi: 10.3390/biology5020024

Cao, K., Li, Y., Deng, C. H., Gardiner, S. E., Zhu, G., Fang, W., et al. (2019). Comparative population genomics identified genomic regions and candidate genes associated with fruit domestication traits in peach. *Plant Biotechnol. J.* 17, 1954–1970. doi: 10.1111/pbi.13112

Catchen, J., Hohenlohe, P., Bassham, S., Amores, A., and Cresko, W. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1016/j.biotechadv.2011.08.021.Secreted

Chalfin, L., Dayan, M., Levy, D. R., Austad, S. N., Miller, R. A., Iraqi, F. A., et al. (2014). Mapping ecologically relevant social behaviours by gene knockout in wild mice. *Nat. Commun.* 5:4569. doi: 10.1038/ncomms5569

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* 10, 195–205. doi: 10.1038/nrg2526

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402. doi: 10.1101/gr.100545.109

Chen, J., Ni, P., Li, X., Han, J., Jakovliæ, I., Zhang, C., et al. (2018). Population size may shape the accumulation of functional mutations following domestication. *BMC Evol. Biol.* 18:4. doi: 10.1186/s12862-018-1120-6

Chen, N., Cai, Y., Chen, Q., Li, R., Wang, K., Huang, Y., et al. (2018). Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat. Commun.* 9:2337. doi: 10.1038/s41467-018-04737-0

Clark, J. W., and Donoghue, P. C. J. (2018). Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* 23, 933–945. doi: 10.1016/j.tplants.2018.07.006

Concia, L., Veluchamy, A., Ramirez-Prado, J. S., Martin-Ramirez, A., Huang, Y., Perez, M., et al. (2020). Wheat chromatin architecture is organized in genome territories and transcription factories. *Genome Biol.* 21, 1–20. doi: 10.1186/s13059-020-01998-1

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823. doi: 10.1126/science.1231143

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., et al. (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* 30, 1187–1189. doi: 10.1093/bioinformatics/btt763

Daly, K. G., Maisano Delser, P., Mullin, V. E., Scheu, A., Mattiangeli, V., Teasdale, M. D., et al. (2018). Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. *Science* 361, 85–88. doi: 10.1126/science.aas9411

Davey, J. W., and Blaxter, M. L. (2010). RADSeq?: next-generation population genetics. *Brief. Funct. Genomics* 9, 416–423. doi: 10.1093/bfgp/elq031

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399. doi: 10.1111/mec.12182

de Villemereuil, P., and Gaggiotti, O. E. (2015). A new F ST -based method to uncover local adaptation using environmental variables. *Methods Ecol. Evol.* 6, 1248–1258. doi: 10.1111/2041-210X.12418

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., et al. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Mol. Ecol. Resour.* 12, 1058–1067. doi: 10.1111/1755-0998.12003

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., and Marchini, J. (2013). Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* 93, 687–696. doi: 10.1016/j.ajhg.2013.09.002

Diao, L., and Chen, K. C. (2012). Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. *Genetics* 192, 1503–1511. doi: 10.1534/genetics.112.144790

Dias-Alves, T., Mairal, J., and Blum, M. G. B. (2018). Loter: a software package to infer local ancestry for a wide range of species. *Mol. Biol. Evol.* 35, 2318–2326. doi: 10.1093/molbev/msy126

Díaz-Valenzuela, E., Sawers, R. H., and Cibrián-Jaramillo, A. (2020). *Cis*-and trans-regulatory variations in the domestication of the chili pepper fruit. *Mol. Biol. Evol.* 37, 1593–1603. doi: 10.1093/molbev/msaa027

Ding, M., and Chen, Z. J. (2018). Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. *Curr. Opin. Plant Biol.* 42, 37–48. doi: 10.1016/j.pbi.2018.02.003

Doebley, J., and Stec, A. (1991). Genetic analysis of the morphological differences between maize and teosinte. *Genetics* 129, 285–295.

Doebley, J., Stec, A., and Gustus, C. (1995). teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141, 333–346.

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., et al. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* 31, 135–141. doi: 10.1038/nbt.2478

Dorant, Y., Benestan, L., Rougemont, Q., Normandeau, E., Boyle, B., Rochette, R., et al. (2019). Comparing Pool—seq, Rapture, and GBS genotyping for inferring weak population structure: the American lobster (*Homarus americanus*) as a case study. *Ecol. Evol.* 9, 6606–6623. doi: 10.1002/ece3.5240

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28, 2239–2252. doi: 10.1093/molbev/msr048

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., et al. (2019). Genomic architecture and introgression shape a butterfly radiation. *Science* 366, 594–599. doi: 10.1126/science.aaw2090

Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., et al. (2019). Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* 51, 541–547. doi: 10.1038/s41588-019-0356-4

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29, 51–63. doi: 10.1016/j.tree.2013.09.008

Excoffier, L., and Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27, 1332–1334. doi: 10.1093/bioinformatics/btr124

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb).* 103:285. doi: 10.1038/hdy.2009.74

Fang, Z., and Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Brief. Bioinform.* 12, 280–287. doi: 10.1093/bib/bbr004

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.

Feeney, A., Nilsson, E., and Skinner, M. K. (2014). Epigenetics and transgenerational inheritance in domesticated farm animals. *J. Anim. Sci. Biotechnol.* 5:48. doi: 10.1186/2049-1891-5-48

Fernie, A. R., and Yan, J. (2019). De novo domestication: an alternative route toward new crops for the future. *Mol. Plant* 12, 615–631. doi: 10.1016/j.molp.2019.03.016

Fierst, J. L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front. Genet.* 6:220. doi: 10.3389/fgene.2015.00220

Fitz-Gibbon, S., Hipp, A. L., Pham, K. K., Manos, P. S., and Sork, V. L. (2017). Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (Quercus section Quercus). *Genome* 60, 743–755. doi: 10.1139/gen-2016-0202

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Frantz, L. A. F., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., et al. (2016). Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science* 352, 1228–1231. doi: 10.1126/science.aaf3161

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H.-J., Cagan, A., Bosse, M., et al. (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* 47, 1141–1148. doi: 10.1038/ng.3394

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., et al. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.* 89, 1827–1831. doi: 10.1073/pnas.89.5.1827

Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., et al. (2015). N6-methyldeoxyadenosine marks active transcription start sites in chlamydomonas. *Cell* 161, 879–892. doi: 10.1016/j.cell.2015.04.010

Fugère, V., and Hendry, A. P. (2018). Human influences on the strength of phenotypic selection. *Proc. Natl. Acad. Sci. U.S.A.* 115, 10070–10075. doi: 10.1073/pnas.1806013115

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One* 8:e79667. doi: 10.1371/journal.pone.0079667

Futschik, A., and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186, 207–218. doi: 10.1534/genetics.110.114397

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., et al. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Mol. Ecol. Resour.* 16, 459–469. doi: 10.1111/1755-0998.12470

Gansauge, M.-T., Gerber, T., Glocke, I., Korleviæ, P., Lippik, L., Nagel, S., et al. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:gkx033. doi: 10.1093/nar/gkx033

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 51, 1044–1051. doi: 10.1038/s41588-019-0410-2

Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet.* 11:e1005004. doi: 10.1371/journal.pgen.1005004

Gepts, P. (2014). The contribution of genetic and genomic approaches to plant domestication studies. *Curr. Opin. Plant Biol.* 18, 51–59. doi: 10.1016/j.pbi.2014.02.001

Gerbault, P., Allaby, R. G., Boivin, N., Rudzinski, A., Grimaldi, I. M., Pires, J. C., et al. (2014). Storytelling and story testing in domestication. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6159–6164. doi: 10.1073/pnas.1400425111

Gibson, G. (2018). Population genetics and GWAS: a primer. *PLoS Biol.* 16:e2005485. doi: 10.1371/journal.pbio.2005485

Golicz, A. A., Batley, J., and Edwards, D. (2016a). Towards plant pangenomics. *Plant Biotechnol. J.* 14, 1099–1105. doi: 10.1111/pbi.12499

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., et al. (2016b). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* 7, 1–8. doi: 10.1038/ncomms13390

Gouil, Q., and Keniry, A. (2019). Latest techniques to study DNA methylation. *Essays Biochem.* 63, 639–648. doi: 10.1042/EBC20190027

Groeneveld, L. F., Lenstra, J. A., Eding, H., Toro, M. A., Scherf, B., Pilling, D., et al. (2010). Genetic diversity in farm animals – a review. *Anim. Genet.* 41, 6–31. doi: 10.1111/j.1365-2052.2010.02038.x

Guerra García, A., and Piñero, D. (2017). Current approaches and methods in plant domestication studies. *Bot. Sci.* 95:345. doi: 10.17129/botsci.1209

Guerra-García, A., Suárez-Atilano, M., Mastretta-Yanes, A., Delgado-Salinas, A., and Piñero, D. (2017). Domestication genomics of the open-pollinated scarlet runner bean (*Phaseolus coccineus* L.). *Front. Plant Sci.* 8:1891. doi: 10.3389/fpls.2017.01891

Guerrero-Bosagna, C. (2012). Finalism in darwinian and lamarckian evolution: lessons from epigenetics and developmental biology. *Evol. Biol.* 39, 283–300. doi: 10.1007/s11692-012-9163-x

Günther, T., and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* 15:e1008302. doi: 10.1371/journal.pgen.1008302

Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., et al. (2012). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* 45, 51–58. doi: 10.1038/ng.2470

Guo, Z., Song, G., Liu, Z., Qu, X., Chen, R., Jiang, D., et al. (2015). Global epigenomic analysis indicates that Epialleles contribute to Allele-specific expression via Allele-specific histone modifications in hybrid rice. *BMC Genomics* 16:232. doi: 10.1186/s12864-015-1454-z

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695. doi: 10.1371/journal.pgen.1000695

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hahn, F., Eisenhut, M., Mantegazza, O., and Weber, A. (2017). Generation of targeted knockout mutants in *Arabidopsis thaliana* using CRISPR/Cas9. *Bio Protoc.* 7, 1–20. doi: 10.21769/BioProtoc.2384

Hahn, M. W., and Hibbins, M. S. (2019). A three-sample test for introgression. *Mol. Biol. Evol.* 36, 2878–2882. doi: 10.1093/molbev/msz178

Han, L., and Abney, M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. *Eur. J. Hum. Genet.* 21, 205–211. doi: 10.1038/ejhg.2012.148

Hanchard, N. A., Rockett, K. A., Spencer, C., Coop, G., Pinder, M., Jallow, M., et al. (2006). Screening for recently selected alleles by analysis of human haplotype similarity. *Am. J. Hum. Genet.* 78, 153–159. doi: 10.1086/499252

He, S., Yan, S., Wang, P., Zhu, W., Wang, X., Shen, Y., et al. (2014). Comparative analysis of genome-wide chromosomal histone modification patterns in maize cultivars and their wild relatives. *PLoS One* 9:e97364. doi: 10.1371/journal.pone.0097364

He, X.-J., Chen, T., and Zhu, J.-K. (2011). Regulation and function of DNA methylation in plants and animals. *Cell Res.* 21, 442–465. doi: 10.1038/cr.2011.23

Heard, E., and Martienssen, R. A. (2014). Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* 157, 95–109. doi: 10.1016/j.cell.2014.02.045

Hekman, J. P., Johnson, J. L., and Kukekova, A. V. (2015). Transcriptome analysis in domesticated species: challenges and strategies. *Bioinform. Biol. Insights* 9(Suppl. 4), 21–31. doi: 10.4137/BBI.S29334

Heled, J., and Drummond, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* 8:289. doi: 10.1186/1471-2148-8-289

Hradilová, I., Trnìnı, O., Válková, M., Cechová, M., Janská, A., Prokešová, L., et al. (2017). A combined comparative transcriptomic, metabolomic, and anatomical analyses of two key domestication traits: pod dehiscence and seed dormancy in pea (*Pisum* sp.). *Front. Plant Sci.* 8:542. doi: 10.3389/fpls.2017.00542

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497–501. doi: 10.1038/nature11532

Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., et al. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants* 5, 54–62. doi: 10.1038/s41477-018-0329-0

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44, 808–811. doi: 10.1038/ng.2309

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C. K. K., Tirnaz, S., Dolatabadian, A., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol. J.* 16, 1265–1274. doi: 10.1111/pbi.12867

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T. H., et al. (2013). Architecture and evolution of a minute plant genome. *Nature* 498, 94–98. doi: 10.1038/nature12132

Ilska, J., Haskell, M. J., Blott, S. C., Sánchez-Molano, E., Polgar, Z., Lofgren, S. E., et al. (2017). Genetic characterization of dog personality traits. *Genetics* 206, 1101–1111. doi: 10.1534/genetics.116.192674

Inbar, S., Cohen, P., Yahav, T., and Privman, E. (2020). Comparative study of population genomic approaches for mapping colony-level traits. *PLoS Comput. Biol.* 16:e1007653. doi: 10.1371/journal.pcbi.1007653

Irving-Pease, E. K., Ryan, H., Jamieson, A., Dimopoulos, E. A., Larson, G., and Frantz, L. A. F. (2019). "Paleogenomics of animal domestication," in *Paleogenomics: Genome-Scale Analysis of Ancient DNA*, eds C. Lindqvist and O. P. Rajora (Cham: Springer International Publishing), 225–272. doi: 10.1007/13836_2018_55

Janowitz Koch, I., Clark, M. M., Thompson, M. J., Deere-Machemer, K. A., Wang, J., Duarte, L., et al. (2016). The concerted impact of domestication and transposon insertions on methylation patterns between dogs and grey wolves. *Mol. Ecol.* 25, 1838–1855. doi: 10.1111/mec.13480

Janzen, G. M., Wang, L., and Hufford, M. B. (2019). The extent of adaptive wild introgression in crops. *New Phytol.* 221, 1279–1288. doi: 10.1111/nph.15457

Jensen, P. (2015). Adding "epi-" to behaviour genetics: implications for animal domestication. *J. Exp. Biol.* 218, 32–40. doi: 10.1242/jeb.106799

Jiang, L. G., Li, B., Liu, S. X., Wang, H. W., Li, C. P., Song, S. H., et al. (2019). Characterization of proteome variation during modern maize breeding. *Mol. Cell. Proteomics* 18, 263–276. doi: 10.1074/mcp.RA118.001021

Jiang, Y., Jiang, Y., Wang, S., Zhang, Q., and Ding, X. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinform.* 20:556. doi: 10.1186/s12859-019-3164-z

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., et al. (2012). Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* 44, 812–815. doi: 10.1038/ng.2312

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11:94. doi: 10.1186/1471-2156-11-94

Jorde, L. B. (2001). Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.* 10, 2199–2207. doi: 10.1093/hmg/10.20.2199

Kantar, M. B., Nashoba, A. R., Anderson, J. E., Blackman, B. K., and Rieseberg, L. H. (2017). The genetics and genomics of plant domestication. *Bioscience* 67, 971–982. doi: 10.1093/biosci/bix114

Kaur, P., and Gaikwad, K. (2017). From genomes to GENE-omes: exome sequencing concept and applications in crop improvement. *Front. Plant Sci.* 8:2164. doi: 10.3389/fpls.2017.02164

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., and Varshney, R. K. (2020). Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends. Plant Sci.* 25, 148–158. doi: 10.1016/j.tplants.2019.10.012

Khan, M. A., Olsen, K. M., Sovero, V., Kushad, M. M., and Korban, S. S. (2014). Fruit quality traits have played critical roles in domestication of the apple. *Plant Genome* 7, 1–18. doi: 10.3835/plantgenome2014.04.0018

Kim, Y., and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167, 1513–1524. doi: 10.1534/genetics.103.025387

Koenig, D., Jimenez-Gomez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., et al. (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc. Natl. Acad. Sci. U.S.A.* 110, E2655–E2662. doi: 10.1073/pnas.1309606110

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinform.* 15:356. doi: 10.1186/s12859-014-0356-4

Korneliussen, T. S., Moltke, I., Albrechtsen, A., and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinform.* 14:289. doi: 10.1186/1471-2105-14-289

Larson, G., Piperno, D. R., Allaby, R. G., Purugganan, M. D., Andersson, L., Arroyo-Kalin, M., et al. (2014). Current perspectives and the future of domestication studies. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6139–6146. doi: 10.1073/pnas.1323964111

Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A., and Gilissen, C. (2015). Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Hum. Mutat.* 36, 815–822. doi: 10.1002/humu.22813

Lemmon, Z. H., Bukowski, R., Sun, Q., and Doebley, J. F. (2014). The role of *Cis* regulatory evolution in maize domestication. *PLoS Genet.* 10:e1004745. doi: 10.1371/journal.pgen.1004745

Levy, S. E., and Myers, R. M. (2016). Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* 17, 95–115. doi: 10.1146/annurev-genom-083115-022413

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi: 10.1038/nature10231

Li, R., Fu, W., Su, R., Tian, X., Du, D., Zhao, Y., et al. (2019). Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. *Front. Genet.* 10:1169. doi: 10.3389/fgene.2019.01169

Li, Y., Von Holdt, B. M., Reynolds, A., Boyko, A. R., Wayne, R. K., Wu, D. D., et al. (2013). Artificial selection on brain-expressed genes during the domestication of dog. *Mol. Biol. Evol.* 30, 1867–1876. doi: 10.1093/molbev/mst088

Li, Y., Zhou, G., Ma, J., Jiang, W., Jin, L., Zhang, Z., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32, 1045–1052. doi: 10.1038/nbt.2979

Li, Z., Chen, J., Wang, Z., Pan, Y., Wang, Q., Xu, N., et al. (2016). Detection of selection signatures of population-specific genomic regions selected during domestication process in Jinhua pigs. *Anim. Genet.* 47, 672–681. doi: 10.1111/age.12475

Librado, P., Fages, A., Gaunitz, C., Leonardi, M., Wagner, S., Khan, N., et al. (2016). The evolutionary origin and genetic makeup of domestic horses. *Genetics* 204, 423–434. doi: 10.1534/genetics.116.194860

Linck, E., and Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* 19, 639–647. doi: 10.1111/1755-0998.12995

Liu, X., and Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nat. Genet.* 47, 555–559. doi: 10.1038/ng.3254

Lotterhos, K. E., and Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Mol. Ecol.* 23, 2178–2192. doi: 10.1111/mec.12725

Lotterhos, K. E., and Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* 24, 1031–1046. doi: 10.1111/mec.13100

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., et al. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* 17, 142–152. doi: 10.1111/1755-0998.12635

Luo, G.-Z., Blanco, M. A., Greer, E. L., He, C., and Shi, Y. (2015). DNA N6-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol.* 16, 705–710. doi: 10.1038/nrm4076

Luu, K., Bazin, E., and Blum, M. G. B. (2017). pcadapt?: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* 17, 67–77. doi: 10.1111/1755-0998.12592

Lye, Z. N., and Purugganan, M. D. (2019). Copy number variation in domestication. *Trends Plant Sci.* 24, 352–365. doi: 10.1016/j.tplants.2019.01.003

Machaj, G., Bostan, H., Macko-Podgórni, A., Iorizzo, M., and Grzebelus, D. (2018). Comparative transcriptomics of root development in wild and cultivated carrots. *Genes (Basel).* 9:431. doi: 10.3390/genes9090431

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288. doi: 10.1016/j.ajhg.2013.06.020

Martin, S. H., Davey, J. W., and Jiggins, C. D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* 32, 244–257. doi: 10.1093/molbev/msu269

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Mascher, M., Schuenemann, V. J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., et al. (2016). Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nat. Genet.* 48, 1089–1093. doi: 10.1038/ng.3611

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., and Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* 15, 28–41. doi: 10.1111/1755-0998.12291

Mather, N., Traves, S. M., and Ho, S. Y. (2020). A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecol. Evol.* 10, 579–589. doi: 10.1002/ece3.5888

Mathew, B., Léon, J., and Sillanpää, M. J. (2018). A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. *Heredity (Edinb).* 120, 356–368. doi: 10.1038/s41437-017-0023-4

Mazet, O., Rodríguez, W., and Chikhi, L. (2015). Demographic inference using genetic data from a single individual: separating population size variation from population structure. *Theor. Popul. Biol.* 104, 46–58. doi: 10.1016/j.tpb.2015.06.003

Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid them. *Mol. Ecol.* 24, 3223–3231. doi: 10.1111/mec.13243

Meissner, A. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 33, 5868–5877. doi: 10.1093/nar/gki901

Meyer, R. S., Choi, J. Y., Sanches, M., Plessis, A., Flowers, J. M., Amas, J., et al. (2016). Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* 48, 1083–1088. doi: 10.1038/ng.3633

Meyer, R. S., and Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* 14, 840–852. doi: 10.1038/nrg3605

Money, D., Migicovsky, Z., Gardner, K., and Myles, S. (2017). LinkImputeR: user-guided genotype calling and imputation for non-model organisms. *BMC Genomics* 18:523. doi: 10.1186/s12864-017-3873-5

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K. K., et al. (2017). The pangenome of hexaploid bread wheat. *Plant Journal* 90, 1007–1013. doi: 10.1111/tpj.13515

Moyers, B. T., Morrell, P. L., and McKay, J. K. (2018). Genetic costs of domestication and improvement. *J. Hered.* 109, 103–116. doi: 10.1093/jhered/esx069

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* 17:53. doi: 10.1186/s13059-016-0917-0

Myers, C. L., Springer, N. M., Schaefer, R., Ross-Ibarra, J., Swanson-Wagner, R., Tiffin, P., et al. (2012). Reshaping of the maize transcriptome by domestication. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11878–11883. doi: 10.1073/pnas.1201961109

Nadachowska-Brzyska, K., Burri, R., Smeds, L., and Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Mol. Ecol.* 25, 1058–1072. doi: 10.1111/mec.13540

Ndjiondjop, M. N., Alachiotis, N., Pavlidis, P., Goungoulou, A., Kpeki, S. B., Zhao, D., et al. (2019). Comparisons of molecular diversity indices, selective sweeps and population structure of African rice with its wild progenitor and Asian rice. *Theor. Appl. Genet.* 132, 1145–1158. doi: 10.1007/s00122-018-3268-2

Nei, M., and Maruyama, T. (1975). Lewontin-Krakauer test for neutral genes. *Genetics* 80:395.

Nielsen, R., and Beaumont, M. A. (2009). Statistical inferences in phylogeography. *Mol. Ecol.* 18, 1034–1047. doi: 10.1111/j.1365-294X.2008.04059.x

Nowoshilow, S., Schloissnig, S., Fei, J. F., Dahl, A., Pang, A. W., Pippel, M., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature* 554, 50–55. doi: 10.1038/nature25458

Ottoni, C., Van Neer, W., De Cupere, B., Daligault, J., Guimaraes, S., Peters, J., et al. (2017). The palaeogenetics of cat dispersal in the ancient world. *Nat. Ecol. Evol.* 1:0139. doi: 10.1038/s41559-017-0139

Pankin, A., Altmüller, J., Becker, C., and von Korff, M. (2018). Targeted resequencing reveals genomic signatures of barley domestication. *New Phytol.* 218, 1247–1259. doi: 10.1111/nph.15077

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335, 721–726. doi: 10.1038/335721a0

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190

Pavlidis, P., and Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *J. Biol. Res.* 24:7. doi: 10.1186/s40709-017-0064-0

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8:e1002967. doi: 10.1371/journal.pgen.1002967

Piperno, D. R. (2017). Assessing elements of an extended evolutionary synthesis for plant domestication and agricultural origin research. *Proc. Natl. Acad. Sci. U.S.A.* 114, 6429–6437. doi: 10.1073/pnas.1703658114

Pitt, D., Sevane, N., Nicolazzi, E. L., MacHugh, D. E., Park, S. D. E., Colli, L., et al. (2019). Domestication of cattle: two or three events? *Evol. Appl.* 12, 123–136. doi: 10.1111/eva.12674

Prezeworski, M., Coop, G., and Wall, J. D. (2005). The signature of positive selection on standing genetic variation. *Evolution* 59, 2312–2323. doi: 10.1111/j.0014-3820.2005.tb00941.x

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., et al. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 5:e1000519. doi: 10.1371/journal.pgen.1000519

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1111/j.1471-8286.2007.01758.x

Qi, J., Liu, X., Shen, D., Miao, H., Xie, B., Li, X., et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* 45, 1510–1515. doi: 10.1038/ng.2801

Qiu, Q., Wang, L., Wang, K., Yang, Y., Ma, T., Wang, Z., et al. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nat. Commun.* 6:10283. doi: 10.1038/ncomms10283

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Ramos-Madrigal, J., Smith, B. D., Moreno-Mayar, J. V., Gopalakrishnan, S., Ross-Ibarra, J., Gilbert, M. T. P., et al. (2016). Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Curr. Biol.* 26, 3195–3201. doi: 10.1016/j.cub.2016.09.036

Ranwez, V., Serra, A., Pot, D., and Chantret, N. (2017). Domestication reduces alternative splicing expression variations in sorghum. *PLoS One* 12:e0183454. doi: 10.1371/journal.pone.0183454

Renaut, S., and Rieseberg, L. H. (2015). The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Mol. Biol. Evol.* 32, 2273–2283. doi: 10.1093/molbev/msv106

Rodríguez-Mega, E., Piñeyro-Nelson, A., Gutierrez, C., García-Ponce, B., Sánchez, M. D. L. P., Zluhan-Martínez, E., et al. (2015). Role of transcriptional regulation in the evolution of plant phenotype: a dynamic systems approach. *Dev. Dyn.* 244, 1074–1095. doi: 10.1002/dvdy.24268

Ross-Ibarra, J., Morrell, P. L., and Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8641–8648. doi: 10.1073/pnas.0700643104

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi: 10.1038/nature01140

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250

Sakurada, K. (2010). Environmental epigenetic modifications and reprogramming-recalcitrant genes. *Stem Cell Res.* 4, 157–164. doi: 10.1016/j.scr.2010.01.001

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* 7:e34131. doi: 10.1371/journal.pone.0034131

Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8, 552–560.

Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* 46, 919–925. doi: 10.1038/ng.3015

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* 15, 749–763. doi: 10.1038/nrg3803

Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., and Odom, D. T. (2009). ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions. *Methods* 48, 240–248. doi: 10.1016/j.ymeth.2009.03.001

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., et al. (2013). Patterns of population epigenomic diversity. *Nature* 495, 193–198. doi: 10.1038/nature11968

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713. doi: 10.1038/ng.3008

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Schreiber, M., Stein, N., and Mascher, M. (2018). Genomic approaches for studying crop evolution. *Genome Biol.* 19:140. doi: 10.1186/s13059-018-1528-8

Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., et al. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5661–E5669. doi: 10.1073/pnas.1416991111

Schurch, N. J., Schofield, P., Gierliñski, M., Cole, C., Sherstnev, A., Singh, V., et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 22, 839–851. doi: 10.1261/rna.053959.115

Schweiger, W., Boddu, J., Shin, S., Poppenberger, B., Berthiller, F., Lemmens, M., et al. (2010). Validation of a candidate deoxynivalenol-inactivating UDP-glucosyltransferase from barley by heterologous expression in yeast. *Mol. Plant Microbe Interact.* 23, 977–986. doi: 10.1094/MPMI-23-7-0977

Seal, A., Gupta, A., Mahalaxmi, M., Aykkal, R., Singh, T. R., and Arunachalam, V. (2014). Tools, resources and databases for SNPs and indels in sequences: a review. *Int. J. Bioinform. Res. Appl.* 10:264. doi: 10.1504/IJBRA.2014.060762

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., et al. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods Ecol. Evol.* 8, 907–917. doi: 10.1111/2041-210X.12700

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 343, 84–87. doi: 10.1126/science.1247005

Shan, S., Soltis, P. S., Soltis, D. E., and Yang, B. (2020). Considerations in adapting CRISPR/Cas9 in nongenetic model plant systems. *Appl. Plant Sci.* 8:e11314. doi: 10.1002/aps3.11314

Shang, Y., Ma, Y., Zhou, Y., Zhang, H., Duan, L., Chen, H., et al. (2014). Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* 346, 1084–1088. doi: 10.1126/science.1259215

Shen, Y., Zhang, J., Liu, Y., Liu, S., Liu, Z., Duan, Z., et al. (2018). DNA methylation footprints during soybean domestication and improvement. *Genome Biol.* 19:128. doi: 10.1186/s13059-018-1516-z

Shi, J., and Lai, J. (2015). Patterns of genomic changes with crop domestication and breeding. *Curr. Opin. Plant Biol.* 24C, 47–53. doi: 10.1016/j.pbi.2015.01.008

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132. doi: 10.1038/nrg3642

Singh, J., Zhao, J., and Vallejos, C. E. (2018). Differential transcriptome patterns associated with early seedling development in a wild and a domesticated common bean (*Phaseolus vulgaris* L.) accession. *Plant Sci.* 274, 153–162. doi: 10.1016/j.plantsci.2018.05.024

Smith, J. M., and Haigh, J. (2007). The hitch-hiking effect of a favourable gene. *Genet. Res. (Camb)* 89, 391–403. doi: 10.1017/S0016672308009579

Smith, O., Nicholson, W. V., Kistler, L., Mace, E., Clapham, A., Rose, P., et al. (2019). A domestication history of dynamic adaptation and genomic deterioration in Sorghum. *Nat. Plants* 5, 369–379. doi: 10.1038/s41477-019-0397-9

Sohn, J., and Nam, J.-W. (2016). The present and future of de novo whole-genome assembly. *Brief. Bioinform.* 19:bbw096. doi: 10.1093/bib/bbw096

Song, J., Li, J., Sun, J., Hu, T., Wu, A., Liu, S., et al. (2018). Genome-wide association mapping for cold tolerance in a core collection of rice (*Oryza sativa* L.) landraces by using high-density single nucleotide polymorphism markers from specific-locus amplified fragment sequencing. *Front. Plant Sci.* 9:875. doi: 10.3389/fpls.2018.00875

Song, Q., Zhang, T., Stelly, D. M., and Chen, Z. J. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biol.* 18:99. doi: 10.1186/s13059-017-1229-8

Soyk, S., Lemmon, Z. H., Oved, M., Fisher, J., Liberatore, K. L., Park, S. J., et al. (2017). Bypassing negative epistasis on yield in tomato imposed by a domestication gene. *Cell* 169, 1142–1155. doi: 10.1016/j.cell.2017.04.032

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., et al. (2017). Karyotype stability and unbiased fractionation in the paleo-allotetraploid cucurbita genomes. *Mol. Plant* 10, 1293–1306. doi: 10.1016/j.molp.2017.09.003

Swinnen, G., Goossens, A., and Pauwels, L. (2016). Lessons from domestication: targeting Cis -regulatory elements for crop improvement. *Trends Plant Sci.* 21, 506–515. doi: 10.1016/j.tplants.2016.01.014

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* 79, 1–12. doi: 10.1086/504302

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome.". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Thornton, T. A., and Bermejo, J. L. (2014). Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genet. Epidemiol.* 38, S5–S12. doi: 10.1002/gepi.21819

Tiffin, P., and Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* 29, 673–680. doi: 10.1016/j.tree.2014.10.004

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635. doi: 10.1038/nature11119

Trerotola, M., Relli, V., Simeone, P., and Alberti, S. (2015). Epigenetic inheritance and the missing heritability. *Hum. Genomics* 9:17. doi: 10.1186/s40246-015-0041-3

Trucchi, E., Gratton, P., Whittington, J. D., Cristofari, R., Le Maho, Y., Stenseth, N. C., et al. (2014). King penguin demography since the last glaciation inferred from genome-wide data. *Proc. R. Soc. B Biol. Sci.* 281:20140528. doi: 10.1098/rspb.2014.0528

Turner, B. M. (2009). Epigenetic responses to environmental change and their evolutionary implications. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 3403–3418. doi: 10.1098/rstb.2009.0125

Ueta, R., Abe, C., Watanabe, T., Sugano, S. S., Ishihara, R., Ezura, H., et al. (2017). Rapid breeding of parthenocarpic tomato plants using CRISPR/Cas9. *Sci. Rep.* 7:507. doi: 10.1038/s41598-017-00501-4

Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J., and Ecker, J. R. (2015). MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* 10:475. doi: 10.1038/nprot.2014.114

Vallebueno-Estrada, M., Rodríguez-Arévalo, I., Rougon-Cardoso, A., Martínez González, J., García Cook, A., Montiel, R., et al. (2016). The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proc. Natl. Acad. Sci. U.S.A.* 113, 14151–14156. doi: 10.1073/pnas.1609701113

VanBuren, R., Wai, C. M., Colle, M., Wang, J., Sullivan, S., Bushakra, J. M., et al. (2018). A near complete, chromosome-scale assembly of the black raspberry (*Rubus occidentalis*) genome. *Gigascience* 7, 1–9. doi: 10.1093/gigascience/giy094

Varshney, R. K., Thudi, M., Roorkiwal, M., He, W., Upadhyaya, H. D., Yang, W., et al. (2019). Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat. Genet.* 51, 857–864. doi: 10.1038/s41588-019-0401-3

Vasemägi, A., Nilsson, J., McGinnity, P., Cross, T., O'Reilly, P., Glebe, B., et al. (2012). Screen for footprints of selection during domestication/captive breeding of atlantic salmon. *Comp. Funct. Genomics* 2012, 1–14. doi: 10.1155/2012/628204

Velasco, D., Hough, J., Aradhya, M., and Ross-Ibarra, J. (2016). Evolutionary genomics of peach and almond domestication. *G3* 6, 3985–3993. doi: 10.1534/g3.116.032672

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47, 97–120. doi: 10.1146/annurev-genet-111212-133526

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72. doi: 10.1371/journal.pbio.0040072

Wales, N., Akman, M., Watson, R. H. B., Sánchez Barreiro, F., Smith, B. D., Gremillion, K. J., et al. (2019). Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3,000 years of sunflower domestication and improvement. *Evol. Appl.* 12, 38–53. doi: 10.1111/eva.12594

Wales, N., Ramos Madrigal, J., Cappellini, E., Carmona Baez, A., Samaniego Castruita, J. A., Romero-Navarro, J. A., et al. (2016). The limits and potential of paleogenomic techniques for reconstructing grapevine domestication. *J. Archaeol. Sci.* 72, 57–70. doi: 10.1016/j.jas.2016.05.014

Walley, J. W., Sartor, R. C., Shen, Z., Schmitz, R. J., Wu, K. J., Urich, M. A., et al. (2016). Integration of omic networks in a developmental atlas of maize. *Science* 353, 814–818. doi: 10.1126/science.aag1125

Wang, E. T., Kodama, G., Baldi, P., and Moyzis, R. K. (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc. Natl. Acad. Sci. U.S.A.* 103, 135–140. doi: 10.1073/pnas.0509691102

Wang, G.-D., Xie, H.-B., Peng, M.-S., Irwin, D., and Zhang, Y.-P. (2014). Domestication genomics: evidence from animals. *Annu. Rev. Anim. Biosci.* 2, 65–84. doi: 10.1146/annurev-animal-022513-114129

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y., et al. (2012). Genome-wide association study (GWAS) of resistance to head smut in maize. *Plant Sci.* 196, 125–131. doi: 10.1016/j.plantsci.2012.08.004

Wang, W., Feng, B., Xiao, J., Xia, Z., Zhou, X., Li, P., et al. (2014). Cassava genome from a wild ancestor to cultivated varieties. *Nat. Commun.* 5:5110. doi: 10.1038/ncomms6110

Wang, W., Zhang, X., Zhou, X., Zhang, Y., La, Y., Zhang, Y., et al. (2019). Deep genome resequencing reveals artificial and natural selection for visual deterioration, plateau adaptability and high prolificacy in chinese domestic sheep. *Front. Genet.* 10:300. doi: 10.3389/fgene.2019.00300

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., and Watson, M. (2015). Exome sequencing: current and future perspectives. *G3* 5, 1543–1550. doi: 10.1534/g3.115.018564

Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., et al. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* 37, 853–862. doi: 10.1038/ng1598

Weigel, D., and Colot, V. (2012). Epialleles in plant evolution. *Genome Biol.* 13, 249. doi: 10.1186/gb-2012-13-10-249

Wiener, P., and Pong-Wong, R. (2011). A regression-based approach to selection mapping. *J. Hered.* 102, 294–305. doi: 10.1093/jhered/esr014

Wigginton, J. E., Cutler, D. J., and Abecasis, G. R. (2005). A note on exact tests of hardy-weinberg equilibrium. *Am. J. Hum. Genet.* 76, 887–893. doi: 10.1086/429864

Wilkinson, S., Lu, Z. H., Megens, H.-J., Archibald, A. L., Haley, C., Jackson, I. J., et al. (2013). Signatures of diversifying selection in european pig breeds. *PLoS Genet.* 9:e1003453. doi: 10.1371/journal.pgen.1003453

Wolter, F., Schindele, P., and Puchta, H. (2019). Plant breeding at the speed of light: the power of CRISPR/Cas to generate directed genetic diversity at multiple sites. *BMC Plant Biol.* 19:176. doi: 10.1186/s12870-019-1775-1

Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J. M., Zhang, Y., Tokarska, M., et al. (2018). Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat. Ecol. Evol.* 2, 1139–1145. doi: 10.1038/s41559-018-0562-y

Xie, M., Chung, C. Y.-L., Li, M., Wong, F.-L., Wang, X., Liu, A., et al. (2019). A reference-grade wild soybean genome. *Nat. Commun.* 10:1216. doi: 10.1038/s41467-019-09142-9

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342. doi: 10.1038/nrg3174

Yang, C. J., Samayoa, L. F., Bradbury, P. J., Olukolu, B. A., Xue, W., York, A. M., et al. (2019). The genetic architecture of teosinte catalyzed and constrained maize domestication. *Proc. Natl. Acad. Sci. U.S.A.* 116, 5643–5652. doi: 10.1073/pnas.1820997116

Yang, I. S., and Kim, S. (2015). Analysis of whole transcriptome sequencing data: workflow and software. *Genomics Inform.* 13:119. doi: 10.5808/GI.2015.13.4.119

Yang, L., Koo, D. H., Li, Y., Zhang, X., Luan, F., Havey, M. J., et al. (2012). Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. *Plant J.* 71, 895–906. doi: 10.1111/j.1365-313X.2012.05017.x

Yang, X., Zhang, H., Shang, J., Liu, G., Xia, T., Zhao, C., et al. (2018). Comparative analysis of the blood transcriptomes between wolves and dogs. *Anim. Genet.* 49, 291–302. doi: 10.1111/age.12675

Zadesenets, K. S., and Rubtsov, N. B. (2018). Genome duplication in animal evolution. *Russ. J. Genet.* 54, 1125–1136. doi: 10.1134/S1022795418090168

Zeder, M. A. (2006). Central questions in the domestication of plants and animals. *Evol. Anthropol. Issues News Rev.* 15, 105–117. doi: 10.1002/evan.20101

Zeder, M. A., Emshwiller, E., Smith, B. D., and Bradley, D. G. (2006). Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.* 22, 139–155. doi: 10.1016/j.tig.2006.01.007

Zeng, K., Fu, Y.-X., Shi, S., and Wu, C.-I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174, 1431–1439. doi: 10.1534/genetics.106.061432

Zeng, L., Tu, X.-L., Dai, H., Han, F.-M., Lu, B.-S., Wang, M.-S., et al. (2019). Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* 20:79. doi: 10.1186/s13059-019-1686-3

Zhang, C., Bailey, D. K., Awad, T., Liu, G., Xing, G., Cao, M., et al. (2006). A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. *Bioinformatics* 22, 2122–2128. doi: 10.1093/bioinformatics/btl365

Zhang, H., Zhang, J., Lang, Z., Botella, J. R., and Zhu, J.-K. (2017). Genome editing—principles and applications for functional genomics research and crop improvement. *CRC. Crit. Rev. Plant Sci.* 36, 291–309. doi: 10.1080/07352689.2017.1402989

Zhao, B. S., Roundtree, I. A., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell. Biol.* 18, 31–42. doi: 10.1038/nrm.2016.132

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50, 278–284. doi: 10.1038/s41588-018-0041-z

Zheng, J., Wu, H., Zhu, H., Huang, C., Liu, C., Chang, Y., et al. (2019). Determining factors, regulation system, and domestication of anthocyanin biosynthesis in rice leaves. *New Phytol.* 223, 705–721. doi: 10.1111/nph.15807

Zhou, J., Li, D., Wang, G., Wang, F., Kunjal, M., Joldersma, D., et al. (2019). Application and future perspective of CRISPR/Cas9 genome editing in fruit crops. *J. Integr. Plant Biol.* 62, 269–286. doi: 10.1111/jipb.12793

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., et al. (2019). The population genetics of structural variants in grapevine domestication. *Nat. Plants* 5, 965–979. doi: 10.1038/s41477-019-0507-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CAPÍTULO 2: ENSAMBLE DEL GENOMA DE REFERENCIA DE *Cucurbita argyrosperma* Y DINÁMICA EVOLUTIVA DE LAS FAMILIAS GÉNICAS CODIFICANTES Y NO CODIFICANTES EN EL GÉNERO *Cucurbita*

Artículo de investigación: The genome of *Cucurbita argyrosperma* (silver-seed gourd) reveals faster rates of protein-coding gene and long noncoding RNA turnover and neofunctionalization within *Cucurbita*

Este capítulo trata sobre el ensamble del genoma de referencia de la calabaza domesticada *Cucurbita argyrosperma* subsp. *argyrosperma* y su uso para entender el ritmo y modo en que evolucionan los genes codificantes, los ARNs largos intergénicos no codificantes y los pseudogenes (*i.e.*, genes codificantes que perdieron su marco abierto de lectura, pero siguen siendo transcripcionalmente activos) después de una duplicación completa del genoma que había sido previamente descrita en las calabazas. El genoma de *C. argyrosperma* se generó usando las tecnologías de secuenciación de Illumina y PacBio, obteniendo un ensamble de alta calidad. También se secuenció ARN de cinco órganos distintos para poder predecir los genes codificantes y no codificantes en el genoma de *C. argyrosperma*. Mediante un análisis de genómica comparada, se encontró que la tasa de recambio de genes codificantes y ARNs largos no codificantes es más rápida dentro de los genomas de calabazas que en los demás genomas previamente reportados para la familia Cucurbitaceae, posiblemente como resultado de la redundancia funcional de la duplicación completa del genoma que facilitó la co-opción y reemplazo de genes. Finalmente, se encontró que algunos de los genes que perdieron su marco abierto de lectura se mantienen conservados estructural y filogenéticamente en los genomas de calabazas, además de que siguen siendo transcripcionalmente activos, sugiriendo que estos elementos pudieran ser funcionales, posiblemente como elementos regulatorios. Este capítulo fue publicado en la revista *Molecular Plant* en el número del mes de abril del 2019.

CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

The Genome of *Cucurbita argyrosperma* (Silver-Seed Gourd) Reveals Faster Rates of Protein-Coding Gene and Long Noncoding RNA Turnover and Neofunctionalization within *Cucurbita*

Josué Barrera-Redondo[1], Enrique Ibarra-Laclette[2], Alejandra Vázquez-Lobo[3], Yocelyn T. Gutiérrez-Guerrero[1], Guillermo Sánchez de la Vega[1], Daniel Piñero[1], Salvador Montes-Hernández[4], Rafael Lira-Saade[5,*] and Luis E. Eguiarte[1,*]

[1]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Circuito Exterior s/n Anexo al Jardín Botánico, 04510 Ciudad de México, Mexico

[2]Departamento de Estudios Moleculares Avanzados, Instituto de Ecología A.C., Carretera Antigua a Coatepec No. 351, Col. El Haya. C.P., Xalapa, Veracruz 91070, Mexico

[3]Centro de Investigaciones en Biodiversidad y Conservación, Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Col. Chamilpa, Cuernavaca, Morelos 62209, Mexico

[4]Campo Experimental Bajío, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP), Km 6.5 Carretera Celaya-San Miguel de Allende, Celaya, Guanajuato 38110, Mexico

[5]UBIPRO, Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Av. de los Barrios #1, Col. Los Reyes Iztacala, Tlanepantla, Edo. de Mex 54090, Mexico

***Correspondence: Rafael Lira-Saade (rlira@unam.mx), Luis E. Eguiarte (fruns@unam.mx)**

https://doi.org/10.1016/j.molp.2018.12.023

ABSTRACT

Whole-genome duplications are an important source of evolutionary novelties that change the mode and tempo at which genetic elements evolve within a genome. The *Cucurbita* genus experienced a whole-genome duplication around 30 million years ago, although the evolutionary dynamics of the coding and noncoding genes in this genus have not yet been scrutinized. Here, we analyzed the genomes of four *Cucurbita* species, including a newly assembled genome of *Cucurbita argyrosperma*, and compared the gene contents of these species with those of five other members of the Cucurbitaceae family to assess the evolutionary dynamics of protein-coding and long intergenic noncoding RNA (lincRNA) genes after the genome duplication. We report that *Cucurbita* genomes have a higher protein-coding gene birth–death rate compared with the genomes of the other members of the Cucurbitaceae family. *C. argyrosperma* gene families associated with pollination and transmembrane transport had significantly faster evolutionary rates. lincRNA families showed high levels of gene turnover throughout the phylogeny, and 67.7% of the lincRNA families in *Cucurbita* showed evidence of birth from the neofunctionalization of previously existing protein-coding genes. Collectively, our results suggest that the whole-genome duplication in *Cucurbita* resulted in faster rates of gene family evolution through the neofunctionalization of duplicated genes.

Key words: *Cucurbita argyrosperma*, comparative genomics, molecular evolution, neofunctionalization, long non-coding RNA, whole-genome duplication

Barrera-Redondo J., Ibarra-Laclette E., Vázquez-Lobo A., Gutiérrez-Guerrero Y.T., Sánchez de la Vega G., Piñero D., Montes-Hernández S., Lira-Saade R., and Eguiarte L.E. (2019). The Genome of *Cucurbita argyrosperma* (Silver-Seed Gourd) Reveals Faster Rates of Protein-Coding Gene and Long Noncoding RNA Turnover and Neofunctionalization within *Cucurbita*. Mol. Plant. **12**, 506–520.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

INTRODUCTION

Cucurbita is a genus with global agronomic relevance (Lira et al., 2016; Paris, 2016) and is one of the angiosperm genera with the highest number of independent domestication events (Nee, 1990; Zheng et al., 2013; Castellanos-Morales et al., 2018). Recent advances in the study of *Cucurbita* spp. genomes revealed a recent whole-genome duplication around 30 million years ago (Mya) in the common ancestor of the genus (Montero-Pau et al., 2017; Sun et al., 2017).

Genome duplications are important sources of evolutionary novelties in plants, since redundant elements in a genome can develop novel functions in a process called neofunctionalization (Ganfornina and Sánchez, 1999; Magadum et al., 2013). It is expected that a lineage that experienced a recent whole-genome duplication would have different gene evolution dynamics compared with other closely related species with non-duplicated genomes (Ponting et al., 2009; Magadum et al., 2013).

Even though the genomic footprints of a whole-genome duplication are strong in *Cucurbita*, the numbers of predicted protein-coding genes in *Cucurbita* genomes are roughly similar to those in other genomes of the Cucurbitaceae family (Huang et al., 2009; Garcia-Mas et al., 2012; Guo et al., 2012; Montero-Pau et al., 2017; Sun et al., 2017; Urasaki et al., 2017; Wu et al., 2017). This apparent lack of duplicated coding genes could be the result of ‘‘pseudogenization’’ processes, implying a loss of redundant genes throughout the evolution of the *Cucurbita* genomes, either by the accumulation of mutations resulting in loss of function or fractionation due to intrachromosomal recombination (Sun et al., 2017). However, duplicated coding genes can also evolve to perform novel functions through positive selection (Wang et al., 2015). These novel functions are not necessarily limited to the emergence of new protein-coding elements, since protein-coding genes can also evolve into regulatory elements as noncoding RNAs (Chen and Rajewsky, 2007). Most of the transcriptional activity in eukaryotes generates noncoding RNAs (Smith and Mattick, 2017), whose abundance correlates positively with organismal complexity, whereas the abundance of protein-coding genes does not scale with complexity (Liu et al., 2013). A particular category of noncoding transcripts called long noncoding RNAs (lncRNAs) seem to play critical roles in eukaryotic development and differentiation (Smith and Mattick, 2017).

lncRNAs are a heterogeneous group of noncoding RNAs larger than 200 nucleotides that lack coding potential (Mercer et al., 2009; Ulitsky, 2016). lncRNAs act as master regulatory genes, mainly through the recruitment of chromatin modifiers in the nucleus, such as DNA methyltransferases and histone posttranslational modifiers (Fatica and Bozzoni, 2014; Smith and Mattick, 2017), although lncRNAs can also act in the cytoplasm through sequence complementarity to other RNA molecules and the modulation of mRNA stability (Fatica and Bozzoni, 2014). In plants, lncRNAs are involved in several biological functions, including organ development, flowering and vernalization, phosphate homeostasis, photomorphogenesis, response to biotic and abiotic stress conditions such as heat stress and response to phytopathogens, alternative splicing of protein-coding genes, nodule formation, and cell-wall synthesis (Chekanova, 2015; Liu et al., 2015).

Studies regarding the evolutionary dynamics of lncRNAs are limited despite their evident importance in plant biology, due to a traditional focus on protein-coding genes in genome-wide studies (Ulitsky, 2016; Nelson et al., 2017). Furthermore, evolutionary analyses of these genes have been limited due to a lack of conservation at both the sequence level and the secondary structure level (Ulitsky, 2016), and research on their origin and evolution is still scarce (see Necsulea et al., 2014; Nelson et al., 2016; Zhao et al., 2018). Several hypotheses have been proposed to explain the emergence of new lncRNAs, such as the neofunctionalization of duplicated protein-coding genes, co-option of transposable elements in the genome, duplication followed by neofunctionalization from other lncRNAs, and *de novo* emergence (Kapusta et al., 2013). Previous studies have suggested that whole-genome duplications can lead to faster rates of evolution in lncRNA families (Ponting et al., 2009; Nelson and Shippen, 2015).

This study focuses on the effects of the *Cucurbita*-wide genome duplication on the evolutionary dynamics of both protein-coding and long intergenic noncoding RNA (lincRNA) genes in the *Cucurbita* genus. We propose that the *Cucurbita* genomes have faster gene evolutionary dynamics, including higher rates of gene birth and death, than the genomes of other members of the Cucurbitaceae family due to this whole-genome duplication. We expected that species belonging to the *Cucurbita* genus might have experienced several recent lincRNA birth events due to the whole-genome duplication (Ponting et al., 2009; Nelson and Shippen, 2015). We also analyzed the possibility that the duplication of protein-coding genes and posterior neofunctionalization may be a source of new lincRNA genes in *Cucurbita* (Kapusta et al., 2013). We explored these hypotheses by analyzing four *Cucurbita* genomes, including our novel genome assembly of *Cucurbita argyrosperma* ssp. *argyrosperma*, commonly known as cushaw or silver-seed gourd in English and ‘‘calabaza pipiana’’ in Spanish (Lira et al., 2016), by comparing the coding and noncoding genes in these genomes with those in other genome assemblies reported for the Cucurbitaceae family.

RESULTS

Cucurbita argyrosperma Genome and Transcriptome

We sequenced the genome of *C. argyrosperma* ssp. *argyrosperma* using three sequencing platforms: Illumina HiSeq2000, Illumina MiSeq, and PacBio RS II (see Supplemental Methods for detailed information on the genome and transcriptome sequencing). We obtained 38.4 Gb of data from HiSeq2000, 13.1 Gb from MiSeq, and 11.4 Gb from PacBio RS II. After applying quality filters to the data (see Supplemental Methods for detailed parameters) and filtering organelle reads, we obtained ~120× high-quality sequence coverage with the Illumina reads and ~31× coverage with the PacBio reads. We estimated the genome size of *C. argyrosperma* to be ca. 238 Mb using KmerGenie (Chikhi and Medvedev, 2014).

The chloroplast genome was assembled into a single circular contig of 157 623 bp and had the typical structures of a

Assembly size	228 814 150 bp
No. of scaffolds	920
Longest scaffold	2 746 581 bp
N_{50} of scaffolds	620 880 bp
L_{50} of scaffolds	103
No. of scaffolds >1 kbp	920 (100.0%)
No. of scaffolds >10 kbp	903 (98.2%)
No. of scaffolds >100 kbp	455 (49.5%)
No. of contigs	1481
Longest contig	2 172 140 bp
N_{50} of contigs	463 388 bp
L_{50} of contigs	132
No. of contigs >1 kbp	1481 (100.0%)
No. of contigs >10 kbp	1417 (95.7%)
No. of contigs >100 kbp	493 (33.3%)
CG content	36.22%
Illumina read coverage	120×
PacBio read coverage	31×
No. of protein-coding genes	28 298
Protein-coding gene average size	3457 bp
Protein-coding gene median size	2627 bp
No. of tRNAs	4387
No. of long noncoding intergenic RNAs	6124

Table 1. *Cucurbita argyrosperma* ssp. *argyrosperma* Genome Assembly Statistics.

chloroplast genome: a large single-copy region, a small single-copy region, and two inverted repeats (Daniell et al., 2016). The mitochondrial genome was assembled into 17 scaffolds composed of 1 062 053 bp and showed several instances of chloroplast sequence insertions, as previously described for the mitochondrial genome of *Cucurbita pepo* (Alverson et al., 2010).

We assembled the *C. argyrosperma* nuclear genome into 920 scaffolds (1481 contigs), with an N_{50} of 620 880 bp (Table 1). The total length of the assembled scaffolds was ~229 Mbp, around 96% of the estimated size of the genome and similar to the assembly size of previously reported *Cucurbita* genomes (Montero-Pau et al., 2017; Sun et al., 2017). Genome completeness was assessed by finding single-copy orthologous genes conserved in embryophytes (1440) using BUSCO (Simão et al., 2015). We found complete sequences for 93.2% (1342) of the BUSCO genes and fragmented sequences for 0.9% (13) of the BUSCO genes within the *C. argyrosperma* genome assembly, suggesting a high level of assembly completeness. We also found that 80.5% of the Illumina reads and 100% of the PacBio reads used for genome assembly mapped against the assembled scaffolds, indicating that most of the sequenced genome is present in the nuclear assembly.

We sequenced the *C. argyrosperma* transcriptome using Illumina HiSeq2000, obtaining 51 Gb of RNA sequencing (RNA-seq) data. We mapped 90.69% of the transcriptome reads back to either the nuclear or the organelle assemblies, indicating a high level of genome completeness. The transcriptome was assembled both *de novo* (Grabherr et al., 2011) and using a genome-guided assembly (Pertea et al., 2015) to aid in the prediction of gene models. We predicted 4387 tRNAs and 28 298 protein-coding genes within the genome assembly, numbers similar to those reported in *C. pepo*, *Cucumis sativus*, and *Cucumis melo* (Huang et al., 2009; Garcia-Mas et al., 2012; Montero-Pau et al., 2017) (Table 1). 78.9% of the protein-coding genes were functionally annotated using InterProScan (Jones et al., 2014; Supplemental Data 4), where 51.7% of the genes could be assigned with at least one gene ontology (GO) term (Ashburner et al., 2000; The Gene Ontology Consortium, 2017).

We predicted ~78 Mbp of transposable elements (TEs) within the *C. argyrosperma* genome, corresponding to 34.1% of the genome assembly. This proportion of TEs is similar to those found within the genomes of *C. pepo* (93 Mbp, 37.8% of the genome assembly), *Cucurbita maxima* (107 Mbp, 40.3%), and *Cucurbita moschata* (106 Mbp, 40.6%) (Montero-Pau et al., 2017; Sun et al., 2017). Of the 78 Mbp of TEs, 93.6% correspond to RNA transposons, with most being LTR retrotransposons (49.09%) and LARD retrotransposons (29.36%). Just 1.95% of the observed TEs correspond to DNA transposons, and 4.44% correspond to unidentifiable TEs (Supplemental Table 1). The dominance of LTR retrotransposons within the genome of *C. argyrosperma* is similar to that in *C. pepo* (50.7%) (Montero-Pau et al., 2017), *C. moschata* (62.9%), and *C. maxima* (69.9%) (Sun et al., 2017), revealing that TE families remained relatively stable within the *Cucurbita* genus.

Phylogeny and Evolution of Protein-Coding Gene Families

We compared the protein-coding genes of *C. argyrosperma* with those of *C. pepo* (Montero-Pau et al., 2017), *C. moschata*, and *C. maxima* (Sun et al., 2017); as well as other genera in the Cucurbitaceae family, *C. sativus* (Huang et al., 2009), *C. melo* (Garcia-Mas et al., 2012), *Citrullus lanatus* (Levi et al., 2011), *Lagenaria siceraria* (Wu et al., 2017), and *Momordica charantia* (Urasaki et al., 2017), to assess protein-coding gene family expansions and contractions within the Cucurbitaceae family, as well as within the *Cucurbita* genus. We used *Fragaria vesca* (Edger et al., 2018) and *Juglans regia* (Martínez-García et al., 2016) as outgroups.

We retrieved 23 247 protein-coding gene families, of which only 11 961 families included at least one homolog conserved in two or more different species; the remaining families were exclusive to a single species (Supplemental Figure 1). We found 698 gene families that remained multicopy within *Cucurbita* after the whole-genome duplication, and these families were functionally enriched ($p < 0.01$) in intracellular protein transport. We also found 858 gene families that remained a constant size throughout *Cucurbita* and the other Cucurbitaceae species, although we found no functional enrichment within these families. We identified 369 gene families as single-copy orthologs conserved in all Cucurbitaceae and outgroup species, which were used to obtain a time-calibrated phylogeny needed for gene family evolution analyses. The resulting species tree had approximate likelihood-ratio test (Ansimova and Gascuel, 2006) support

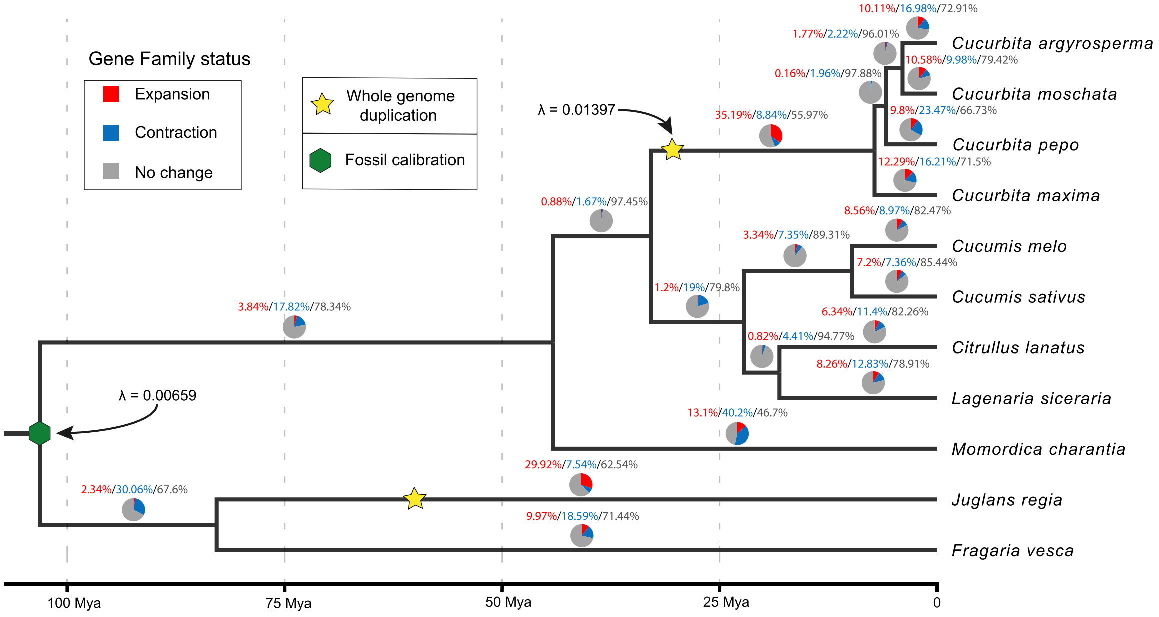


Figure 1. Dated Phylogeny of the Cucurbitaceae Family with Protein-Coding Gene Family Expansions and Contractions per Branch.
The phylogeny was generated with 369 single-copy orthologous genes. Fossil evidence was used to calibrate the basal node of the tree (green hexagon). The pie charts and the percentages at every branch of the tree indicate whether a gene family expanded (red), contracted (blue), or remained the same size (gray). The yellow stars indicate the estimated ages of the whole-genome duplication events in the *Cucurbita* genus (Montero-Pau et al., 2017) and in *Juglans regia* (Luo et al., 2015). The black arrows indicate the change from a basal gene birth/death rate (λ) in the most recent common ancestor of the phylogeny to a faster gene birth/death rate after the whole-genome duplication in *Cucurbita*. Every node in the phylogeny has an approximate likelihood-ratio test (aLRT) support value of 100%.

values of 100% at every node. The dated phylogeny with hishest posteior densities (HPS ± 95% confidence interval) obtained using mcmctree (Yang, 2007) supports a divergence time between *C. argyrosperma* and its sister species *C. moschata* of around 3.98 ± 1.7 Mya, while the divergence between *Cucurbita* and Benincaseae (*C. sativus* + *C. melo* + *C. lanatus* + *L. siceraria*; Schaefer et al., 2009) happened around 32.9 ± 11 Mya (Figure 1), concordant with the expected age of the whole-genome duplication event in *Cucurbita*, approximately 30 ± 4 Mya (Montero-Pau et al., 2017). The crown node of the included Cucurbitaceae species was dated at 44.1 ± 14 Mya.

We performed likelihood-ratio tests to compare the likelihood score of a global gene birth–death rate parameter (λ) across the tree against multiple λ values throughout the phylogeny. A model with a change in λ within *Cucurbita* had a significantly higher log-likelihood (−193 746.504) than a single λ (−198 328.178) throughout the tree (Figure 1 and Supplemental Figure 2). After accounting for genome assembly and annotation error rates, we found that the gene birth–death rate was twice as high in *Cucurbita* (λ = 0.01397) than in the rest of the phylogeny (λ = 0.00659).

We detected phylogenetic inconsistencies in gene content within *Cucurbita*, with some genomes containing ~32 000 protein-coding genes and others containing ~28 000 genes (Supplemental Table 2). To account for possible errors in gene prediction, we repeated the gene family analysis using only high-quality protein-coding gene predictions with annotation edit distances (eAED) lower than 0.5 (Yandell and Ence, 2012; Campbell et al., 2014) to eliminate low-quality gene models. We found a similar number of high-quality gene models in all *Cucurbita* genomes, which is much closer to the total number of predicted genes in *C. pepo* and *C. argyrosperma* (Supplemental Table 2). Even after discarding low-quality gene models, λ was still twice as high in *Cucurbita* (0.01188) than in the rest of the phylogeny (0.00566) (Supplemental Figures 3 and 4).

We found significantly rapid changes in gene family sizes ($p < 0.01$) throughout most of the branches within the phylogeny (Figure 1). We found that the branch leading to the crown node of the *Cucurbita* genus and the terminal branch of *C. argyrosperma* had unusually high rates of gene family evolution (Figure 1). Even though just a small number of gene families showed significantly rapid ($p < 0.01$) levels of change in the branch leading to the crown node of *Cucurbita* (six gene families), this branch had the second highest number of gene family changes within the entire phylogeny (Figure 1). Surprisingly, the terminal branch of *M. charantia* showed the highest number of gene family changes in the whole phylogeny (Figure 1), although there were only a few gene families with significantly rapid changes (27 gene families). Furthermore, the proportion of gene families that either expanded or contracted in the terminal branches of *Cucurbita* was higher compared with the proportion of gene families that either expanded or contracted in the terminal branches of Benincaseae (Figure 1).

The terminal branch of *C. argyrosperma* had an unusually high number of gene families with significantly rapid expansions/contractions (327 families). However, most of the rapidly evolving gene families within *C. argyrosperma* underwent contractions (78.3%), rather than expansions (21.7%). After performing

GO ID	GO term	FDR *p* value
Significantly expanded protein-coding gene families		
GO:0007018	Microtubule-based movement	<1.729E−27
GO:0006270	DNA replication initiation	7.8E−16
GO:0006855	Drug transmembrane transport	4.6E−05
GO:0007010	Cytoskeleton organization	0.00346
Significantly contracted protein-coding gene families		
GO:0042545	Cell-wall modification	<1.729E−27
GO:0006979	Response to oxidative stress	<1.729E−27
GO:0055114	Oxidation–reduction process	<1.729E−27
GO:0009733	Response to auxin	<1.729E−27
GO:0030244	Cellulose biosynthetic process	2.2E−22
GO:0006508	Proteolysis	1.2E−17
GO:0006887	Exocytosis	3.9E−06
GO:0006855	Drug transmembrane transport	4.7E−05
GO:0003333	Amino acid transmembrane transport	0.00048
GO:0005992	Trehalose biosynthetic process	0.00064
GO:0048544	Recognition of pollen	0.00064
GO:0005975	Carbohydrate metabolic process	0.00453
GO:0071577	Zinc II ion transmembrane transport	0.00939

Table 2. Enriched Biological Functions of Rapidly Evolving Protein-Coding Gene Families in *C. argyrosperma*.

a GO enrichment analysis, we found four overrepresented biological functions associated with the significantly expanded families in *C. argyrosperma* (Table 2), including microtubule-based movement in families mainly composed of proteins with kinesin motor domains, and drug transmembrane transport in families mainly composed of villin/gelsolin proteins. We also found 13 overrepresented biological functions associated with the significantly contracted families in *C. argyrosperma* (Table 2), including cell-wall modification in families mainly composed of pectinesterases, response to oxidative stress, oxidation–reduction processes, recognition of pollen, exocytosis, and several processes associated with transmembrane transport. Curiously, drug transmembrane transport was enriched in both significantly expanded families and significantly contracted families, which were composed mainly of multi-antimicrobial extrusion proteins.

lincRNA Prediction and Analysis

We used Evolinc-I (Nelson et al., 2017) to predict lincRNAs within the genome assembly of *C. argyrosperma*, as well as the genomes of *C. maxima*, *C. moschata*, *C. pepo*, *C. melo*, *C. sativus*, *C. lanatus*, and *L. siceraria*. The predicted lincRNAs were compared against the protein-coding gene transcripts of each genome to determine the percentage of lincRNAs produced from the neofunctionalization of duplicated protein-coding sequences. We also compared the predicted lincRNAs against the RepBase (Bao et al., 2015) sequences from eudicots to determine the percentage of lincRNAs produced from the neofunctionalization of TEs within each genome.

Since most of the species transcriptomes used to predict lincRNAs had differences in the organs sequenced, as well as differences in sequencing depth and library construction (see Supplemental Table 3), each species had a different number of predicted lincRNAs (Supplemental Table 2), and these numbers could not be directly compared. However, we expect that the proportion of lincRNAs derived from protein-coding genes and TEs relative to the total number of predicted lincRNAs in a genome remains relatively constant, despite differences in the RNA-seq strategy. Hence, we compared this proportion between *Cucurbita* species and the other species within Cucurbitaceae. Despite the differences in the number of predicted lincRNAs, the percentage of protein-coding-derived lincRNAs was roughly similar between *Cucurbita* species, while there was more variance in this proportion between the other cucurbits (Figure 2). We found a higher percentage of protein-coding-derived lincRNAs in *Cucurbita* species than in other cucurbits (Figure 2; $p = 0.041$), which fits the coding-to-noncoding neofunctionalization hypothesis (Kapusta et al., 2013). However, the proportion of TE-derived lincRNAs was lower in the *Cucurbita* genus compared with the other taxa of the Cucurbitaceae family (Figure 2; $p = 0.013$).

We analyzed the evolution of lincRNA families across the Cucurbitaceae family by using the *C. argyrosperma* predicted lincRNAs as queries to search for homologs within all the analyzed genomes. We compared the lincRNA homologs of each lincRNA family against the protein-coding genes and the Evolinc-I predicted lincRNAs for each species to assess the relationship between lincRNAs and protein-coding genes, as well as to assess the transcriptional potential of the lincRNA homologs. Since the evolutionary proximity of *C. argyrosperma* to the other *Cucurbita* species can lead to erroneous inferences about lincRNA family expansion in this genus, we also used the predicted lincRNAs of *C. lanatus* as sequence queries in the

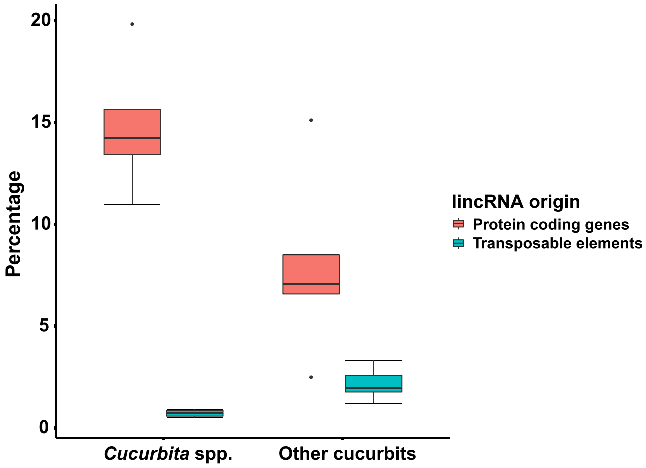


Figure 2. Differences in the Proportion of lincRNAs Associated with Protein-Coding Transcripts (Red) and Transposable Elements (Blue) between Four *Cucurbita* Genomes (*C. argyrosperma*, *C. moschata*, *C. maxima*, and *C. pepo*) and Five Genomes of Other Cucurbit Species (*C. sativus*, *C. melo*, *C. lanatus*, *L. siceraria*, and *M. charantia*).

Cucurbita spp. show a higher proportion of protein-coding gene-derived lincRNAs compared with other members of the same family (p = 0.041), whereas the proportion of transposable element (TE)-derived lincRNAs is lower in *Cucurbita* compared with the other cucurbit species (p = 0.013).

lincRNA family analysis to assess whether the patterns observed within *Cucurbita* were determined by the effect of lincRNA turnover throughout the phylogeny (Ulitsky, 2016).

We retrieved 5466 *C. argyrosperma* lincRNA families, of which 67.7% showed evidence of protein-coding gene neofunctionalization throughout their phylogenies, while only 32.3% were exclusively composed of noncoding elements. In contrast to the *C. argyrosperma* lincRNA gene families, we found that 34.3% of the 5231 *C. lanatus* lincRNA families had evidence of protein-coding neofunctionalization, while 65.7% were exclusively composed of noncoding elements. To assess the level of lincRNA conservation across the Cucurbitaceae family, we only used the subset of lincRNAs whose families were solely composed of noncoding elements, since protein-coding-derived lincRNAs can be traced to homologous protein-coding genes in distantly related taxa, therefore leading to overestimation of the conservation of lincRNAs throughout the phylogeny. This analysis showed that the conservation of lincRNAs between *C. argyrosperma* and the rest of the analyzed species steadily declines with phylogenetic distance, with an average of 2.27% of lincRNA homologs lost per million years (Figure 3A).

75.9%–92.4% of the lincRNAs found in *C. argyrosperma* had at least one homolog within the genomes of the other *Cucurbita* species, with an average lincRNA loss rate of about 2.4% per million years. Just 4.3%–8.8% of the lincRNA families had homologs within the genomes of species belonging to Benincaseae, whereas 6.5% of the lincRNA families had homologs within the genome of *M. charantia*, which is a higher percentage than that observed in some species belonging to Benincaseae. Just 1.2% of the lincRNA families in *J. regia* and 0.3% of those in in *F. vesca* were retained, and only five lincRNA homologs were conserved in both species. Despite the general trend between phylogenetic distance and lincRNA loss, the average rate of lincRNA loss between *C. argyrosperma* and Benincaseae increased to 2.8% per million years and then declined to around 2.1% per million years in *M. charantia* and to 0.96% per million years in the outgroup species. The decline in lincRNA conservation with respect to phylogenetic distance could also be observed for *C. lanatus* lincRNAs, with an average loss rate of 2.54% per million years in the whole phylogeny (Figure 3B), although the percentage of retained lincRNAs in *J. regia* and *F. vesca* was higher (2.3% and 0.7% respectively). The average rate of lincRNA loss between *C. lanatus* and *Cucurbita* was also around 2.8% per million years. However, the lincRNA loss rate within Benincaseae was approximately 3.4% per million years, the highest rate observed in the study. The loss rate also declined in *M. Charantia* to around 2% per million years and to 0.95% per million years in the outgroup species.

Some *C. argyrosperma* lincRNA families showed a high degree of conservation within Cucurbitaceae, that is, every species had at least one representative gene within the family (1016 families). While most of these conserved lincRNA families had at least one protein-coding gene within its phylogeny (95.17%), a small percentage of the families showing a high degree of conservation were composed of putatively noncoding elements (4.82%). We found that seven of these putatively noncoding families were present as single-copy orthologs within Cucurbitaceae, and the members of these families were mostly predicted independently as lincRNAs with Evolinc-I, that is, using transcriptional evidence (Figure 4A). The five lincRNAs that were shared between *C. argyrosperma* and both outgroup species showed complex evolutionary histories, with several instances of duplications and losses, and none of them were conserved in all the analyzed species (e.g., Supplemental Figure 5).

Our analyses show a high rate of lincRNA family birth within the *Cucurbita* genus (Figure 4B). 55.94% of the lincRNA families were exclusively found in the *Cucurbita* genus. Of these gene families, 78.87% were present in all four *Cucurbita* species and just 1.7% were exclusive to *C. argyrosperma*. We found a similar pattern for the *C. lanatus* lincRNA families, where 64.61% were exclusively found in Benincaseae. However, 47.24% of these lincRNA families were exclusive to *C. lanatus*, and only 10.47% were present in all four species.

Many of the *C. argyrosperma* lincRNA families (61.2%) showed signals of gene duplication within the *Cucurbita* genus. Many of these families showed symmetric expansion within *Cucurbita* (1948 families), that is, expansion where the number of lincRNA genes remained constant within *Cucurbita* but at least two-fold higher with respect to the rest of the Cucurbitaceae species (Figure 4C). This pattern is not a product of the phylogenetic distance between *Cucurbita* species, as it could also be seen within the *Cucurbita* clade when analyzing the *C. lanatus* lincRNA families (53 families, Supplemental Figure 6). Most of the lincRNA families with symmetric expansion within *Cucurbita* also showed evidence of protein-coding neofunctionalization (62.26%). The phylogenies of some of these families showed a duplication event corresponding to the whole-genome duplication in *Cucurbita*, where a protein-coding family gave rise to a lincRNA clade

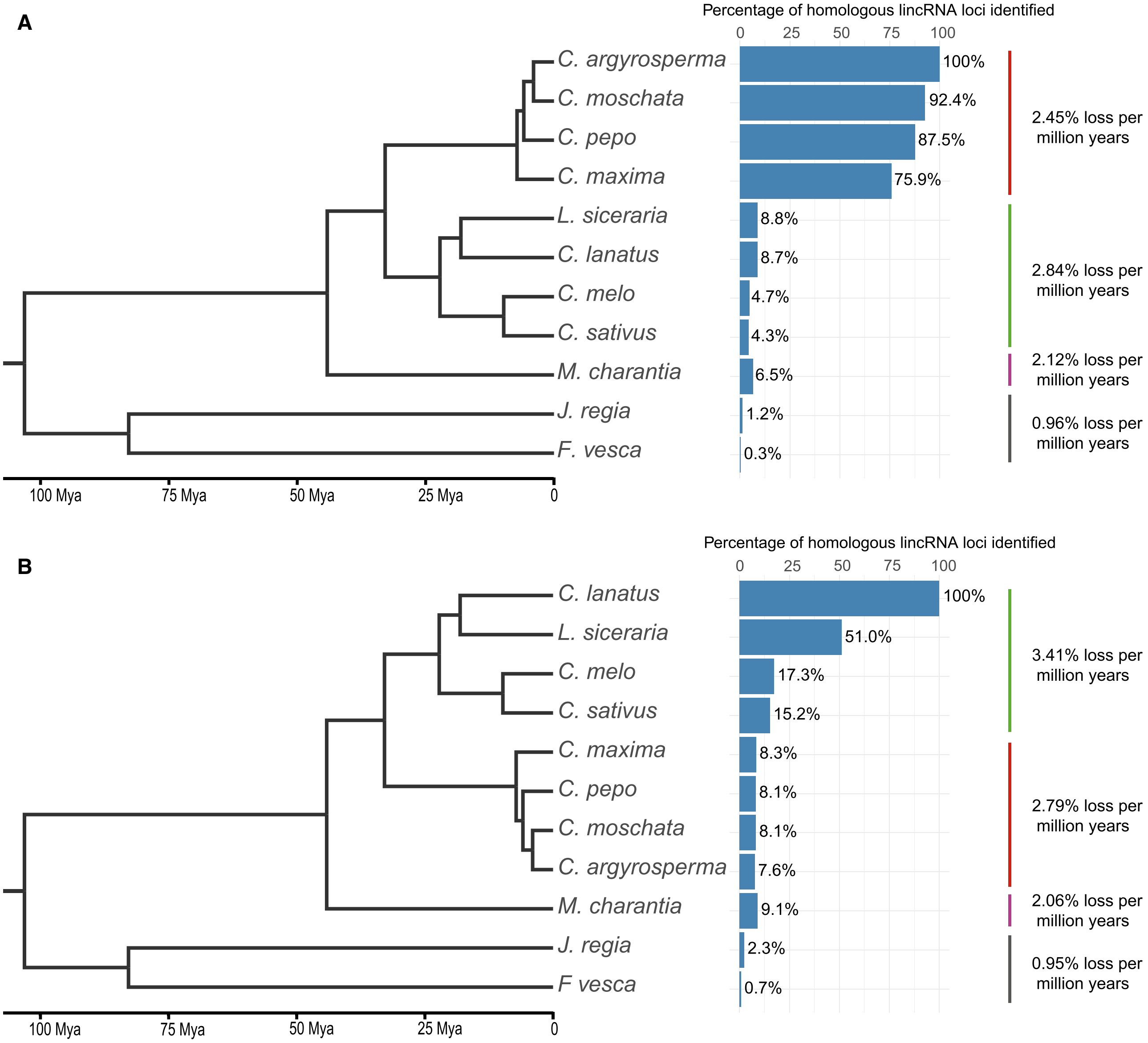


Figure 3. lincRNA Conservation across the Cucurbitaceae Family.
Each column alongside the phylogeny represents the percentage of homologs found within each cucurbit genome for the subset of lincRNAs without homology to the predicted protein-coding genes in the genomes of **(A)** *Cucurbita argyrosperma* and **(B)** *Citrullus lanatus*. The percentage of conserved lincRNAs between species declines drastically as the phylogenetic distance becomes larger, due to a high rate of lincRNA turnover (gene birth/death rate). Only a small fraction of lincRNAs are conserved between Cucurbitaceae and the outgroup species. The average rate of lincRNA loss per million years (right) is shown for the following clades: *Cucurbita* (red), Benincaseae (green), *Momordica charantia* (purple), and the outgroup species (gray).

(Figure 4D). After carefully inspecting one of these lincRNA families (Carg_TCONS _00015392), we found high levels of sequence conservation between the *Cucurbita* lincRNAs, although no conservation of the length, the ORF, or the codon structure of its protein-coding homolog was observed (Figure 5A). We also found a thermodynamically stable secondary structure in the lincRNA (Figure 5B and 5D), whereas the homologous protein-coding transcript showed lower structure stability, as well as signs of many equally stable structures throughout the transcript (Supplemental Figure 7), despite both the lincRNA and protein-coding transcript having similar dinucleotide frequencies (Supplemental Table 4). The structural stability of this lincRNA is even higher than that of one of the highly conserved lincRNAs found in single copy throughout the Cucurbitaceae (Carg_TCONS_00063022; Figure 5C and 5E).

DISCUSSION

The protein-coding gene content within the *Cucurbita* species has remained relatively constant, despite the whole-genome duplication that happened around 30 ± 4 Mya (Montero-Pau et al., 2017). However, our results indicate that this genome duplication event had a profound effect on the gene evolutionary dynamics within *Cucurbita*, namely, higher rates of protein-coding gene family evolution and higher rates of

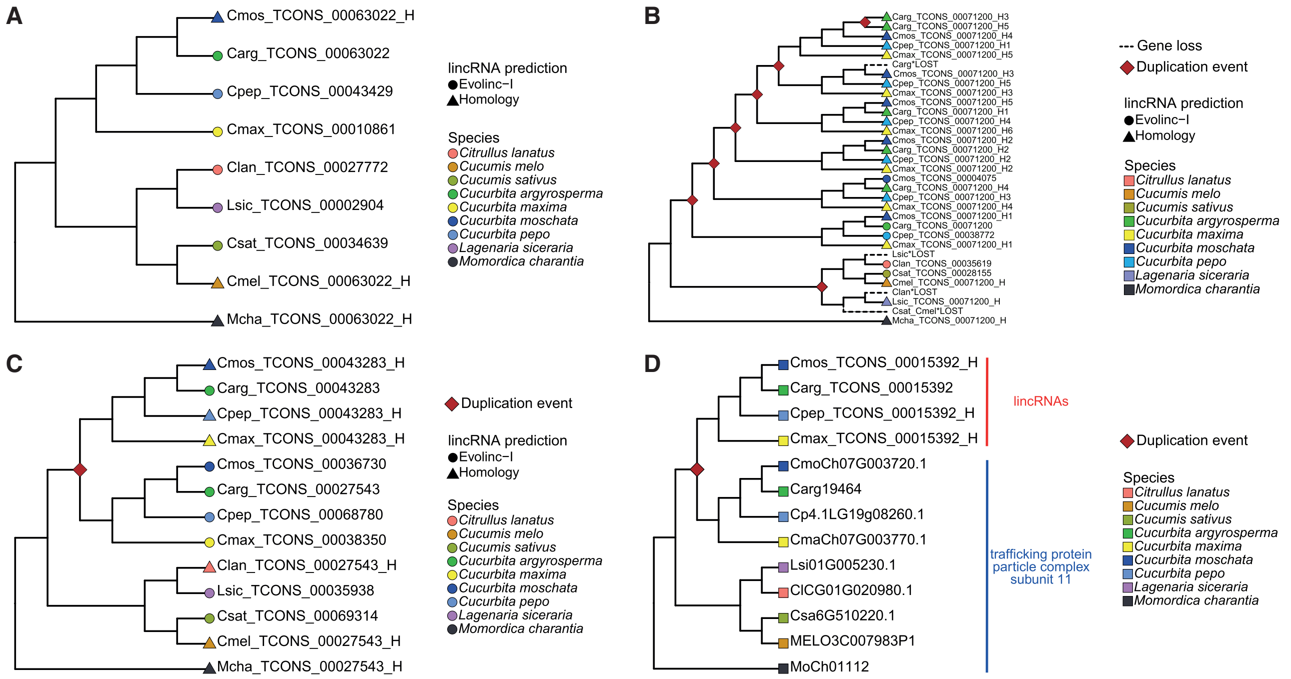


Figure 4. Patterns of lincRNA Family Evolution throughout the *Cucurbita* Genus.

(A) Presence of single-copy orthologous lincRNAs with a high degree of conservation throughout the Cucurbitaceae family suggests these lincRNAs have an essential biological function.

(B) Sudden duplication bursts (red diamonds) within the *Cucurbita* genus, as well as multiple gene losses (dotted branches), reveal a high rate of lincRNA turnover.

(C) Duplication of lincRNA families associated with the whole-genome duplication in *Cucurbita* (red diamond).

(D) Neofunctionalization of protein-coding genes (blue bar) into the novel Carg_TCONS_00015392 lincRNAs (red bar) after the whole-genome duplication in the *Cucurbita* genus (red diamond). Some lincRNAs were independently predicted based on homology using Evolinc-II (triangles in terminal nodes) and based on transcriptomic evidence using Evolinc-I (circles in terminal nodes), further supporting the transcription of these genes. The colors at the terminal nodes indicate the species to which each gene belongs.

coding-to-noncoding gene neofunctionalization. This idea is further supported by the concordance between the estimated age of the whole-genome duplication and the divergence time between *Cucurbita* and Benincaseae, as observed in our dated phylogeny. The divergence times estimated in this study mostly fall within the 95highest posterior density of previous phylogenetic studies, although our estimated ages within the *Cucurbita* genus are slightly older (Schaefer et al., 2009; Castellanos-Morales et al., 2018).

The whole-genome duplication within *Cucurbita* seems to be responsible for the observed acceleration in the rate of protein-coding gene family evolution, as almost half of the gene families experienced either expansions or contractions in the branch leading to the crown node of the *Cucurbita* genus, with a higher proportion of expansions than contractions. Furthermore, the terminal branches in the *Cucurbita* genus also showed larger proportions of gene family expansions and contractions compared with most of the other cucurbit taxa. Finally, the rate of gene family birth/death was two times higher in the *Cucurbita* clade compared with the rest of the phylogeny. All this evidence points toward a higher rate of gene family evolution in *Cucurbita* after the whole-genome duplication event that happened around 30 Mya (Montero-Pau et al., 2017). These patterns were observed despite performing statistical corrections for genome assembly and annotation errors during our gene family analyses, and were also observed after filtering low-quality gene models, suggesting that our results are robust to possible errors in gene model predictions. Furthermore, the number of gene models obtained after filtering low-quality gene models suggests that the real number of protein-coding genes in *Cucurbita* may be closer to 28 000 (Montero-Pau et al., 2017) than 32 000 genes (Sun et al., 2017).

The terminal branch of *J. regia* showed a high proportion of gene family expansion. These gene families have been previously shown to be involved in the biosynthesis of nonstructural polyphenols (Martínez-García et al., 2016). The genome of *J. regia* also shows evidence of a whole-genome duplication that happened around 60 Mya (Luo et al., 2015; Martínez-García et al., 2016), further suggesting that whole-genome duplications are correlated with higher rates of gene family evolution.

The results of GO enrichment analysis of the rapidly evolving families suggest that several biological functions changed in *C. argyrosperma* in relation to other *Cucurbita* species. For instance, we found a significant contraction in gene families associated with the recognition of pollen as well as pectinesterases, which are involved in pollen tube growth during pollination (Tian et al., 2006). Both kinesin motor protein and villin/gelsolin protein families, which expanded in *C. argyrosperma*, are also involved in pollen tube growth during pollination (Su et al., 2007; Li et al., 2012). Contraction and expansion of these families could be

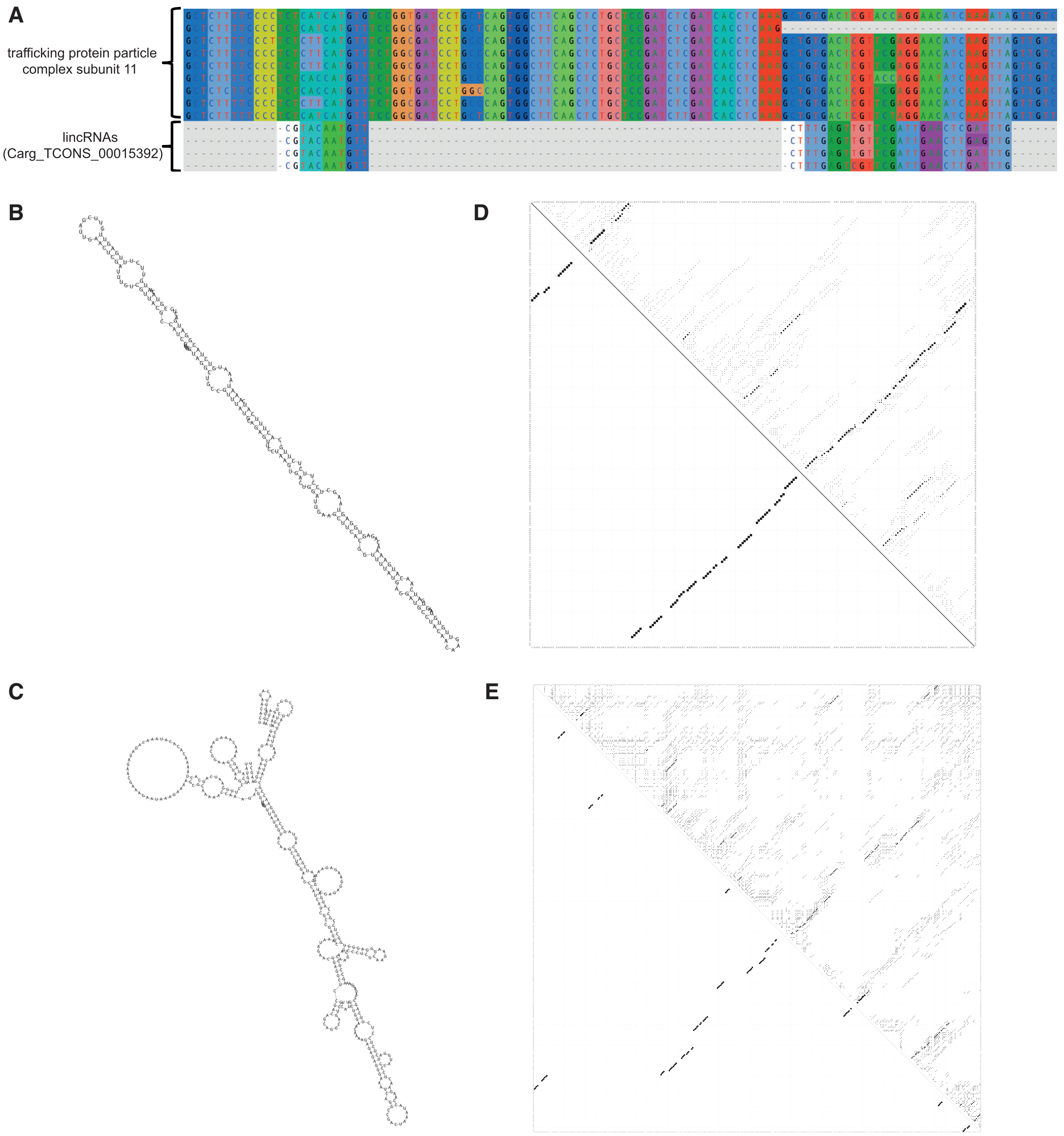

A
trafficking protein particle complex subunit 11
lincRNAs (Carg_TCONS_00015392)
B
C
D
E

Figure 5. Manual Assessment of Primary and Secondary Structure in Some lincRNAs Found within the Genome of *Cucurbita argyrosperma*.

(A) Part of a multiple sequence alignment between a protein-coding gene-derived lincRNA (Carg_TCONS_00015392; down) and the homologous protein-coding gene transcripts (trafficking protein particle complex subunit 11; up). Each codon in the alignment is shown in a different color. Carg_TCONS_00015392 lincRNAs do not contain an open reading frame or show codon conservation, but orthologous lincRNA sequences are conserved between species.

(B–C) Minimum free energy (MFE) structural predictions of **(B)** the protein-coding-derived lincRNA Carg_TCONS_00015392 and **(C)** the highly conserved lincRNA Carg_TCONS_00063022.

(D–E) RNA base-pairing probability matrices showing the MFE structural prediction (below the diagonal) and all possible suboptimal pairings (above the diagonal) for **(D)** Carg_TCONS_00015392 and **(E)** Carg_TCONS_00063022. Higher probabilities are represented as larger dots within the matrices.

associated with changes in reproductive isolation, since reproductive barriers that prevent hybridization are more stringent in *C. argyrosperma* than in its sister species, *C. moschata* (Hurd et al., 1971). Pectinesterases are also involved in cell-wall modification during fruit ripening, and changes in the pectinesterase family could explain the differences in the smoothness of the fruit flesh between *C. argyrosperma* and *C. moschata*. The reduction in the number of pectinesterases within *C. argyrosperma* could have had an impact in its domestication, since fewer genes were available for artificial selection to act upon, possibly restricting its use to seed consumption, unlike the rest of the domesticated *Cucurbita* species whose ripened fruit flesh is commonly consumed (Lira et al., 2016). A future comparison between the genomes of *C.*

argyrosperma ssp. *argyrosperma* and *C*. *argyrosperma* ssp. *sororia* will reveal whether this reduction in gene family size happened before or after the domestication of this species.

Several contracted families were functionally enriched in exocytosis and transmembrane transport functions, which are usually involved in the release of secondary metabolites, hormones, and numerous other compounds (Hedrich and Marten, 2006). Interestingly, different multi-antimicrobial extrusion protein families either expanded or contracted within the genome of *C*. *argyrosperma*. This suggests an adaptive transition between different families of multi-antimicrobial extrusion proteins, perhaps related to changes in the geographic distribution of *C*. *argyrosperma* from its ancestors (Castellanos-Morales et al., 2018). Since multi-antimicrobial proteins are usually involved in the removal of cytotoxic compounds (Eckardt, 2001), the levels of these compounds could change alongside the geographic distribution of the species, acting as selective pressures in these gene families.

Some of the observed evolutionary dynamics within the lincRNA families in Cucurbitaceae can be explained under a high-turnover model of lincRNA evolution, such as the decline in lincRNA conservation as a function of phylogenetic distance and sudden duplication bursts. These dynamics were observed in both the *C*. *argyrosperma* and *C*. *lanatus* lincRNA families, suggesting that gene duplication is a common mechanism of lincRNA birth within the Cucurbitaceae family.

The average rate of lincRNA loss observed in the Cucurbitaceae family is similar to that observed in the Brassicaceae family, around 2.47% per million years (Nelson et al., 2016). However, considerable variation can be observed within each clade in both phylogenies. In the case of Cucurbitaceae, the loss rate ranged from 2.8% per million years between *Cucurbita* and Benincaseae to 2.1% per million years between *Cucurbita* and *M. charantia* and 3.4% per million years within Benincaseae. The differences in loss rate between Brassicaceae species are even more drastic, ranging from 4.3% per million years between *Arabidopsis* and *Capsella* to 2% per million years between *Arabidopsis* and *Brassica* (Nelson et al., 2016). Even though the total number of shared lincRNAs decreases with phylogenetic distance, the loss rate seems to decrease between distantly related taxa. In the case of Brassicaceae, the loss rate declined to 1.5% per million years between *Arabidopsis* and Cleomaceae, which diverged 64 Mya (Nelson et al., 2016). In the case of Cucurbitaceae, we found that the loss rate declined to 0.96% per million years between Cucurbitaceae and the outgroup species. This decline could be explained by a survivor bias, whereby the most biologically important genes are conserved throughout distantly related taxa, thus slowing the rate of lincRNA loss per million years. The rate of lincRNA loss within Tetrapoda seems to be slower than in plants, as 3% of the lincRNAs in humans are also present in chickens, which diverged 300 Mya (Necsulea et al., 2014). The loss rate between *Cucurbita* and Benincaseae was higher than the rate within *Cucurbita*, as expected by the effect of the whole-genome duplication (Nelson and Shippen, 2015). However, the high loss rate within Benincaseae was unexpected, since it was higher than that observed between *Cucurbita* and Benincaseae. We propose that the acceleration in the rate of lincRNA turnover within Benincaseae was caused by the multiple changes in karyotype number throughout Benincaseae (Huang et al., 2009; Levi et al., 2011; Garcia-Mas et al., 2012; Wu et al., 2017), which are considered genomic disturbances that can accelerate this rate (Nelson and Shippen, 2015). This hypothesis is supported by the larger proportion of conserved lincRNAs between *C*. *argyrosperma* and *M*. *charantia* than between *C*. *argyrosperma* and *Cucumis*, which can be explained by additional genomic disturbances within the Benincaseae family.

The decline in lincRNA conservation throughout the Cucurbitaceae phylogeny could be explained by high levels of gene birth and death, making the search for homology between distantly related species futile, as the vast majority of these genes either arose before the divergence of both taxa or became extinct in one of the lineages (Ulitsky, 2016). It is also possible that lincRNAs have high rates of nucleotide substitution due to positive selection (Smith and Mattick, 2017), which hinders the search for homologous sequences as species become more distantly related. Given that lncRNAs are involved in epigenetic regulation through several mechanisms (Mercer et al., 2009), the possibility of positive selection acting on such dynamic genes may be an important factor in adaptive radiation (Smith and Mattick, 2017).

The observation of highly conserved lincRNAs, as well as the evidence of their transcription suggests they have an important biological function within the Cucurbitaceae family. However, pinpointing the specific biological functions of lincRNAs is still difficult without experimental data. The high rate of turnover in lincRNA families limits the inference of biological functions from distantly related plants such as *Arabidopsis thaliana*, in which experimental validation of gene functions is more common (Ulitsky, 2016). Future experimental studies should focus on the functional characterization of these lincRNAs.

The proportion of Evolinc-II lincRNA families with evidence of protein-coding neofunctionalization was higher than initially expected based on our direct comparison between lincRNAs and protein-coding genes. Such events are not exclusive of the *Cucurbita* crown node, as they are rather common throughout the Cucurbitaceae family, although they are particularly frequent within the *Cucurbita* clade. This suggests that the neofunctionalization of protein-coding genes into novel lincRNAs is more common than initially suspected (Kapusta et al., 2013), and may be a recurrent source of noncoding genes in the Cucurbitaceae family (Chen and Rajewsky, 2007) alongside other sources of lincRNA genes, such as neofunctionalization from TEs (Kapusta et al., 2013).

The proportion of lincRNA families with evidence of protein-coding neofunctionalization predicted from comparisons with the *C*. *argyrosperma* lincRNAs was higher compared with those predicted from comparisons with the *C*. *lanatus* lincRNAs. Furthermore, the proportion of protein-coding gene-derived lincRNAs calculated from the direct comparison between coding transcripts and lincRNAs was significantly higher in *Cucurbita* compared with the proportion in other cucurbit species, whereas the proportion of TE-derived lincRNAs was lower in *Cucurbita* with respect to the rest of the Cucurbitaceae family. These results suggest that the whole-genome duplication in *Cucurbita* acted as a genomic disturbance that altered lincRNA family birth dynamics, with more lincRNAs being derived from protein-coding genes than from TEs (Kapusta et al., 2013; Nelson and Shippen, 2015). This is consistent with the apparent conservation of TE proportions between *Cucurbita* species, suggesting that TEs did not play an important role in lincRNA family evolution within *Cucurbita* after the whole-genome duplication, unlike in other taxa such as vertebrate species, where TEs play an important role in lincRNA birth (Kapusta et al., 2013). Our results also differ from those obtained in cotton, where a larger proportion of lincRNAs were homologous to TEs than to protein-coding loci (Zhao et al., 2018). This suggests that the evolutionary dynamics of lincRNAs can change drastically between different plant taxa. After the whole-genome duplication within *Cucurbita*, many duplicated protein-coding genes that were functionally redundant were coopted into novel lincRNAs (Ponting et al., 2009; Kapusta et al., 2013).

The observed differences in length between Carg_TCONS_00015392 and its protein-coding homologs, as well as the disruption of the ORF and the lack of codon structure, suggest this lincRNA is not a cryptic protein-coding transcript but a true noncoding element in *Cucurbita*. Furthermore, the level of sequence conservation between species, as well as the thermodynamic stability of its predicted secondary structure, suggests that this lincRNA may be functional (Smith and Mattick, 2017). Even though secondary structure alone is insufficient evidence to support the hypothesis that a noncoding RNA is functional, especially considering that some lincRNAs have more than one functional structure (Smith and Mattick, 2017), secondary structures in functional noncoding RNAs are expected to be more stable than those in other sequences with similar compositions (Clote et al., 2005). This can be observed when comparing the stability between Carg_TCONS_00015392 and its protein-coding homolog, both of which have similar dinucleotide frequencies but different structural stabilities. This structural stability is also present in the highly conserved lincRNA Carg_TCONS_00063022. The stability of the predicted secondary structures in both lincRNAs suggests that they are functional (Clote et al., 2005), unlike the secondary structure in the transcript of the protein-coding gene homologous to Carg_TCONS_00015392, which appears to be random. Whether all predicted lincRNAs behave similarly or whether they are indeed functional remains to be experimentally validated.

We propose that the whole-genome duplication within the *Cucurbita* genus allowed for faster rates of gene family evolution, since functional redundancy within the genome facilitated the cooption of complex genetic elements, such as previously existing genes, into new functions. During the fractionation process after the whole-genome duplication (Sun et al., 2017), a substantial fraction of the duplicated protein-coding genes with redundant functions either diverged (acquiring novel functions as protein-coding genes, thereby increasing the rate of gene family birth/death), or neofunctionalized into noncoding elements, such as lincRNAs.

METHODS

Biological Samples and DNA/RNA Extraction

We obtained seeds from a cultivated individual of *C. argyrosperma* ssp. *argyrosperma* collected in the region of Tepec, Jalisco (see Supplemental Data 1 for detailed methods and data on fruit selection). One of the seeds was germinated in a greenhouse, and plants were grown to maturity, when flower buds started to develop. We selected one of the germinated plants and extracted total DNA from fresh leaves (Doyle and Doyle, 1987) for genome sequencing.

For transcriptome sequencing, we extracted total RNA from leaves, stems, roots, male flower buds, and tendrils using the RNeasy Plant Mini Kit (Qiagen) according to the manufacturer's protocol. Each RNA sample was precipitated in salty ethanol (260 mM lithium chloride and 66% EtOH).

The plant used for whole-genome sequencing and transcriptome sequencing was deposited in the National Herbarium of Mexico (MEXU) under accession number SMH-JMG-627. The details of DNA and RNA sequencing are available in Supplemental Methods.

Genome and Transcriptome Assembly

The chloroplast genome of *C. argyrosperma* was assembled with NOVOPlasty (Dierckxsens et al., 2016), and the mitochondrion genome was assembled using the Organelle-PBA pipeline (Soorni et al., 2017). Both organelles were scaffolded using SSPACE-longread (Boetzer and Pirovano, 2014). Gap-filling and base corrections were performed with Pilon (Walker et al., 2014). The nuclear genome was assembled with a hybrid approach, using Platanus (Kajitani et al., 2014) and DBG2OLC (Ye et al., 2016). We used Minimap and Racon (Vaser et al., 2017) to obtain a consensus sequence assembly, then base corrections were made using Pilon. Scaffolding was done using BESST (Sahlin et al., 2014) and SSPACE-longread. Gap closing was performed with GapFiller (Boetzer and Pirovano, 2012) and a final base correction was done with Pilon. See Supplemental Methods for a detailed description of the organelle and nuclear genome assemblies and a description of the transcriptome assembly.

The Illumina and PacBio sequence reads were mapped against the genome using BWA *mem* (Li, 2013) and BlasR (Chaisson and Tesler, 2012), respectively, to assess the completeness of the genome assembly. We mapped the transcriptome reads against the nuclear and organelle genomes using Hisat2 (Kim et al., 2015) to assess assembly completeness. The percentage of reads that mapped to the assembly was calculated using *flagstat* within SAMtools (Li et al., 2009).

Prediction of Transposable Elements and Protein-Coding Gene Models

We used the REPET package (Flutre et al., 2011) to predict *de novo* the TEs within the *C. argyrosperma* genome assembly, generating a library of non-redundant consensus sequences. These consensus sequences were classified according to Wicker's classification system (Wicker et al., 2007) using PASTEC (Hoede et al., 2014) within the REPET pipeline (repeat library available in Supplemental Data 2). The repeat library was used to annotate and mask the TEs within the genome assembly with RepeatMasker (Smit et al., 2013).

MAKER3 (Cantarel et al., 2008) was used to predict protein-coding gene models in the *C. argyrosperma* genome assembly. We incorporated AUGUSTUS (Stanke et al., 2006) GeneMark-ES (Lomsadze et al., 2005) and SNAP (Korf, 2004) as *ab initio* gene predictors within MAKER3. We also used EvidenceModeler (Haas et al., 2008) to obtain additional gene models within MAKER3. We incorporated tRNAscan-SE (Lowe and Eddy, 1997) within the MAKER3 pipeline to predict tRNA genes. See Supplemental Methods for a detailed description of the prediction of protein-coding gene models. The protein-coding genes predicted within *C. argyrosperma* were functionally annotated with InterProScan (Jones et al., 2014). The annotation table is available in Supplemental Data 4.

Phylogenetic and Protein-Coding Gene Family Analyses

The details of the phylogenetic analyses can be found within Supplemental Methods. We performed an all-VS-all BLASTp (Camacho et al., 2009) analysis to identify protein-coding gene families with MCL (Enright et al., 2002), using an inflation parameter of 3. All gene families were aligned with MAFFT (Katoh et al., 2002).

We generated the species phylogeny with PhyML (Guindon et al., 2010), using SMS (Lefort et al., 2017) to determine the best amino acid substitution model for our sequence alignment. To obtain a dated phylogeny, we used a Bayesian Markov chain Monte Carlo approach with approximate likelihood calculation, as implemented in mcmctree (Yang, 2007). The mcmctree trace files are available in Supplemental Data 3.

We assessed changes in protein-coding gene family sizes across the dated phylogeny using CAFE v4.0.2 (De Bie et al., 2006). We initially estimated the gene birth–death parameter λ to assess gene family evolution using a subset of gene clusters that had <100 differences in gene content between any pair of species and used it to calculate significant changes (p value <0.01) in gene family size at every branch in the dated phylogeny for every gene family.

We compared three different λ schemes: (a) a change in λ within Cucurbitaceae, (b) a change in λ within *Cucurbita*, and (c) two changes in λ, one within Cucurbitaceae and another within *Cucurbita*. After finding the best scheme of λ parameters within the phylogeny, we estimated error models for genome assembly and annotation errors (Han et al., 2013), and used those models to analyze significant gene family expansions and contractions throughout the tree. We performed the clustering, molecular clock, and gene family analyses using the same methodology as described above using a subset of high-quality protein-coding gene models obtained after filtering MAKER predictions with eAED values lower than 0.5 (input files, CAFE scripts, and final outputs are available in Supplemental Data 5). We performed GO enrichment analyses with topGO using the *weight01* method (Alexa et al., 2006). Significantly enriched terms were assessed after performing Fisher's exact tests and performing false discovery rate (FDR) adjustments of the p values (Benjamini and Hochberg, 1995).

Prediction and Analysis of Long Noncoding RNAs

We used Evolinc-I (Nelson et al., 2017) to predict lincRNAs within the genome-guided transcriptome assembly of *C. argyrosperma*. In brief, Evolinc-I predicted as lincRNAs all transcripts longer than 200 bp that did not overlap with any of the predicted protein-coding genes within the genome and did not contain an ORF longer than 100 amino acids (Nelson et al., 2017). lincRNAs were also predicted for the genomes of *C. maxima*, *C. moschata*, *C. pepo*, *C. melo*, *C. sativus*, *C. lanatus*, and *L. siceraria* using the same methodology as described above, using RNA-seq data available in the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra). SRA accessions used for each species are available in Supplemental Table 3. The lincRNAs of *M. charantia* were extracted from the gff3 file available under the NCBI RefSeq genome accession PRJNA433137 (Urasaki et al., 2017). We used BLASTn (Camacho et al., 2009) with a cutoff of 50% coverage and 30% identity to define similarity due to sequence homology between lincRNAs, protein-coding transcripts, and TEs. We performed Student's *t*-tests with the *stats* package in R (R Core Team, 2016) to assess statistical differences in the proportion of protein-coding gene-derived lincRNAs and TE-derived lincRNAs between *Cucurbita* and the other cucurbit species.

We used Evolinc-II (Nelson et al., 2017) to assess the evolution of lincRNA families across the Cucurbitaceae family. Homologous sequences for each query lincRNA were filtered using an e-value of $<1e^{-20}$ after a reiterative reciprocal BLAST search against each genome. We used MAFFT (Katoh et al., 2002) to align each lincRNA family and RAxML (Stamatakis, 2014) to generate gene family trees. Each gene family tree was compared against the species tree to detect duplication or loss of lincRNAs across the phylogeny using Notung (Chen et al., 2000). Since there is a possibility that the predicted lincRNAs are protein-coding genes that could not be predicted with MAKER3, we wanted to exclude as many false positives as possible to obtain a conservative estimation of the number of lincRNA families that evolved from protein-coding genes. Thus, we eliminated 648 possible spurious lincRNAs (that is, protein-coding genes that were mistakenly predicted as lincRNAs) where the only noncoding elements within the gene family belonged to the query species, and the other elements were protein-coding genes predicted from the other species. Likewise, we defined lincRNA families with evidence of protein-coding gene neofunctionalization as those with noncoding elements from two or more different species within the phylogeny, as the rate of parallel misidentification of lincRNA genes in several species should be low.

Finally, we defined putatively noncoding families as those that lacked any protein-coding gene within the phylogeny, that is, they were composed exclusively of lincRNA genes. All lincRNA family alignments and phylogenies are available in Supplemental Data 6 and 7. Rates of lincRNA loss were calculated as the percentage of lincRNAs in the query species without a homolog in another species divided by the mean divergence time between the two taxa. The secondary structure predictions and the base-pairing probability matrix of the lincRNAs were calculated with RNAfold from the ViennaRNA package (Hofacker et al., 1994; Lorenz et al., 2011).

ACCESSION NUMBERS

The raw sequence reads of all the genomic and transcriptomic data are available in the NCBI SRA database (accession SRP157098). The nuclear and organelle genome assemblies; as well as the protein-coding gene, tRNA, and lincRNA predictions of *C. argyrosperma* are available in the CoGe database (IDs 53608, 52005, 52006) and in the Figshare database (https://doi.org/10.6084/m9.figshare.7728608.v1). The predicted lincRNAs of the other species are available in the CoGe database (IDs 52078–52084). The dated species tree and the single-copy ortholog alignment used for the phylogenetic analyses are available in TreeBase (submission ID S23151).

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This study was funded by Comisión Natural para el Conocimiento y Uso de la Biodiversidad (CONABIO) KE004 "Diversidad genética de las especies de *Cucurbita* en México e hibridación entre plantas genéticamente modificadas y especies silvestres de *Cucurbita*," CONABIO PE001 "Diversidad genética de las especies de *Cucurbita* en México. Fase II. Genómica evolutiva y de poblaciones, recursos genéticos y domesticación" (both awarded to R.L.-S. and L.E.E.), Consejo Nacional de Ciencia y Tecnología (CONACyT) Investigación Científica Básica 2011.167826 "Genómica de poblaciones: estudios en el maíz silvestre, el teosinte (*Zea mays* ssp. *parviglumis* y *Zea mays* ssp. *mexicana*)" (awarded to L.E.E.), and CONACyT Problemas Nacionales through grant number 247730 (awarded to D.P.). J.B.-R. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 583146 from CONACyT. The sabbatical leave of L.E.E. at Dr. Peter Tiffin's laboratory, Department of Plant and Microbial Biology, University of Minnesota, was supported by the program PASPA-DGAPA, UNAM. The Evolinc-II analyses were carried out using CONABIO's computing cluster, which was partially funded by Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) through the grant "Contribución de la Biodiversidad para el Cambio Climático" to CONABIO.

AUTHOR CONTRIBUTIONS

J.B.-R., E.I.-L., A.V.-L., S.M.-H., R.L.-S., and L.E.E. jointly conceived and designed the study. G.S.d.l.V. and S.M.-H. collected the biological samples and carried out morphological analyses to select the parental individual of *C. argyrosperma* used for whole-genome sequencing and assembly. J.B.-R., G.S.d.l.V., and A.V.-L. carried out the molecular laboratory work. J.B.-R. performed the genome assembly and gene predictions, and carried out most of the analyses. E.I.-L. trained the AUGUSTUS model for gene prediction and ran REPET for TE prediction. Y.T.G.-G. did the functional annotation of the protein-coding gene families. E.I.-L. and Y.T.G.-G. helped in several bioinformatics analyses. J.B.-R. drafted the manuscript. D.P., R.L.-S., and L.E.E. obtained the funding for the study. All authors revised the final version of the manuscript.

ACKNOWLEDGMENTS

This manuscript is presented in partial fulfillment of the requirements to obtain a PhD degree by J.B.-R. in the Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM). We acknowledge the Doctorado en Ciencias Biomédicas for the support provided during the development of this project. Special thanks to Valeria Souza for supporting this research, and the technical support of Laura Espinosa-Asuar and Erika Aguirre-Planter. We thank Xitlali Aguirre-Dugua for depositing the plant accession in the herbarium, and all the Laboratorio de Evolución Molecular y Experimental at Instituto de Ecología, UNAM. We also acknowledge the support of the Facultad de Estudios Superiores Iztacala, UNAM, and of the Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Campo Experimental Bajío. The authors thank Rodrigo García Herrera, head of the Scientific Computing Department at LANCIS, Instituto de Ecología, UNAM, for running the HTC infrastructure we used for part of the analyses. The authors acknowledge the technical support of MSc Emanuel Villafán and the resources for high-performance computing that the Institute of Ecology (INECOL) made available for conducting the genome assembly and annotation reported in this paper. L.E.E. acknowledges the support of the program PASPA-DGAPA, UNAM to conduct his sabbatical year. No conflict of interest declared.

Received: August 27, 2018
Revised: December 12, 2018
Accepted: December 28, 2018
Published: January 7, 2019

REFERENCES

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics **22**:1600–1607.

Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K., and Palmer, J.D. (2010). Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol. Biol. Evol. **27**:1436–1448.

Ansimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. **55**:539–552.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**:25–29.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA **6**:11.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **57**:289–300.

Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. Genome Biol. **13**:R56.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics **15**:211.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST plus: architecture and applications. BMC Bioinformatics **10**:421.

Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinform. **48**:4.11.1–4.11.39.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. **18**:188–196.

Castellanos-Morales, G., Paredes-Torres, L.M., Gámez, N., Hernández-Rosales, H.S., Sánchez-de la Vega, G., Barrera-Redondo, J., Aguirre-Planter, E., Vázquez-Lobo, A., Montes-Hernández, S., Lira-Saade, R., et al. (2018). Historical biogeography and phylogeny of *Cucurbita*: insights from ancestral area reconstruction and niche evolution. Mol. Phylogenet. Evol. **128**:38–54.

Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics **13**:238.

Chekanova, J.A. (2015). Long non-coding RNAs and their functions in plants. Curr. Opin. Plant Biol. **27**:207–216.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet. **8**:93–103.

Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. J. Comput. Biol. **7**:429–447.

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics **30**:31–37.

Clote, P., Ferré, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA **11**:578–591.

Daniell, H., Lin, C.-S., Yu, M., and Chang, W.-J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. **17**:134.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics **22**:1269–1271.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. **45**:gkw955.

Doyle, J.J., and Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**:11–15.

Edger, P.P., VanBuren, R., Colle, M., Poorten, T.J., Wai, C.M., Niederhuth, C.E., Alger, E.I., Ou, S., Acharya, C.B., Wang, J., et al. (2018). Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. Gigascience **7**:1–7.

Eckardt, N.A. (2001). Move it on out with MATEs. Plant Cell **13**:1477–1480.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30**:1575–1584.

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. **15**:7–21.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in *de novo* annotation approaches. PLoS One **6**:e16526.

Ganfornina, M.D., and Sánchez, D. (1999). Generation of evolutionary novelty by functional shift. BioEssays **21**:432–439.

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., Gonzalez, V.M., Henaff, E., Camara, F., Cozzuto, L., Lowy, E., et al. (2012). The genome of melon (*Cucumis melo* L.). Proc. Natl. Acad. Sci. U S A **109**:11872–11877.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. **29**:644–652.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**:307–321.

Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W.J., Zhang, H., Zheng, Y., Mao, L., Ren, Y., Wang, Z., et al. (2012). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat. Genet. **45**:51–58.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. **9**:R7.

Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. **30**:1987–1997.

Hedrich, R., and Marten, I. (2006). 30-year progress of membrane transport in plants. Planta **224**:725–739.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. PLoS One **9**:1–6.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatsh. F. Chem. **125**:167–188.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. Nat. Genet. **41**:1275–1281.

Hurd, P.D., Jr., Linsley, E.G., and Whitaker, T.W. (1971). Squash and gourd bees (*Peponapis*, *Xenoglossa*) and the origin of the cultivated *Cucurbita*. Evolution **25**:218–234.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30**:1236–1240.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. **24**:1384–1395.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. **9**:e1003470.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**:3059–3066.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**:357–360.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics **5**:59.

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. Mol. Biol. Evol. **34**:2422–2424.

Levi, A., Hernandez, A., Thimmapuram, J., Donthu, R., Wright, C., Ali, C., Wechter, W.P., Reddy, U., and Mikel, M. (2011). Sequencing the genome of the heirloom watermelon cultivar charleston gray. Plant and Animal Genome Conference. P047.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics **25**:2078–2079.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv 1303.3997.

Li, J., Xu, Y., and Chong, K. (2012). The novel functions of kinesin motor proteins in plants. Protoplasma **249**:S95–S100.

Lira, R., Eguiarte, L., Montes, S., Zizumbo-Villarreal, D., Colunga-GarcíaMarín, P., and Quesada, M. (2016). *Homo sapiens-Cucurbita* interaction in Mesoamerica: domestication, dissemination and diversification. In Ethnobotany of Mexico, R. Lira, A. Casas, and J. Blancas, eds. (New York: Springer-Verlag), pp. 389–402.

Liu, G., Mattick, J.S., and Taft, R.J. (2013). A meta-analysis of the genomic and transcriptomic composition of complex life. Cell Cycle **12**:2061–2072.

Liu, X., Hao, L., Li, D., Zhu, L., and Hu, S. (2015). Long non-coding RNAs and their biological roles in plants. Genomics Proteomics Bioinformatics **13**:137–147.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. **33**:6494–6506.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA package 2.0. Algorithms Mol. Biol. **6**:26.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.

Luo, M.-C., You, F.M., Li, P., Wang, J.-R., Zhu, T., Dandekar, A.M., Leslie, C.A., Aradhya, M., McGuire, P.E., and Dvorak, J. (2015). Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. BMC Genomics **16**:707.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. J. Genet. **92**:155–161.

Martínez-García, P.J., Crepeau, M.W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J., Stevens, K.A., Paul, R., Butterfield, T.S., Britton, M.T., Reagan, R.L., et al. (2016). The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. Plant J. **87**:507–532.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat. Rev. Genet. **10**:155–159.

Montero-Pau, J., Blanca, J., Bombarely, A., Ziarsolo, P., Esteras, C., Martí-Gómez, C., Ferriol, M., Gómez, P., Jamilena, M., Mueller, L., et al. (2017). *De novo* assembly of the zucchini genome reveals a whole genome duplication associated with the origin of the *Cucurbita* genus. Plant Biotechnol. J. **12**:3218–3221.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature **505**:635–640.

Nee, M. (1990). The domestication of *Cucurbita* (Cucurbitaceae). Econ. Bot. **44**:56–68.

Nelson, A.D.L., and Shippen, D.E. (2015). Evolution of TERT-interacting lncRNAs: expanding the regulatory landscape of telomerase. Front. Genet. **6**:1–6.

Nelson, A.D.L., Devisetty, U.K., Palos, K., Haug-Baltzell, A.K., Lyons, E., and Beilstein, M.A. (2017). Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs. Front. Genet. **8**:1–12.

Nelson, A.D.L., Forsythe, E.S., Devisetty, U.K., Clausen, D.S., Haug-Batzell, A.K., Meldrum, A.M., Frank, M.R., Lyons, E., and Beilstein, M.A. (2016). A genomic analysis of factors driving lincRNA diversification: lessons from plants. G3 (Bethesda) **6**:2881–2891.

Paris, H.S. (2016). Genetic resources of pumpkins and squash, *Cucurbita* spp. In Genetics and Genomics of Cucurbitaceae, R. Grumet, N. Katzir, and J. Garcia-Mas, eds. (Cham, Switzerland: Springer), pp. 111–154.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. **33**:290–295.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell **136**:629–641.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., and Arvestad, L. (2014). BESST—efficient scaffolding of large fragmented assemblies. BMC Bioinformatics **15**:281.

Schaefer, H., Heibl, C., and Renner, S.S. (2009). Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. Proc. R. Soc. B Biol. Sci. **276**:843–851.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker. *Open4.0*. http://www.repeatmasker.org.

Smith, M.A., and Mattick, J.S. (2017). Structural and functional annotation of long noncoding RNAs. In Bioinformatics. Methods in Molecular Biology, J.M. Keith, ed. (New York: Humana Press), pp. 65–85.

Soorni, A., Haak, D., Zaitlin, D., and Bombarely, A. (2017). Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics **18**:49.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313.

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics **7**:62.

Su, H., Wang, T., Dong, H., and Ren, H. (2007). The villin/gelsolin/fragmin superfamily proteins in plants. J. Integr. Plant Biol. **49**:1183–1191.

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., Zhang, J., Zhang, H., Gong, G., Jia, Z., et al. (2017). Karyotype stability and unbiased fractionation in the Paleo-Allotetraploid *Cucurbita* genomes. Mol. Plant **10**:1293–1306.

The Gene Ontology Consorsium. (2017). Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res. **45**:D331–D338.

Tian, G.W., Chen, M.H., Zaltsman, A., and Citovsky, V. (2006). Pollen-specific pectin methylesterase involved in pollen tube growth. Dev. Biol. **294**:83–91.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat. Rev. Genet. **17**:601–614.

Urasaki, N., Takagi, H., Natsume, S., Uemura, A., Taniai, N., Miyagi, N., Fukushima, M., Suzuki, S., Tarora, K., Tamaki, M., et al. (2017). Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. DNA Res. **24**:51–58.

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. **27**:737–746.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963.

Wang, J., Chu, S., Zhu, Y., Cheng, H., and Yu, D. (2015). Positive selection drives neofunctionalization of the UbiA prenyltransferase gene family. Plant Mol. Biol. **87**:383–394.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**:973–982.

Wu, S., Shamimuzzaman, M., Sun, H., Salse, J., Sui, X., Wilder, A., Wu, Z., Levi, A., Xu, Y., Ling, K.-S., et al. (2017). The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a Papaya ring-spot virus resistance locus. Plant J. **92**:963–975.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. **13**:329–342.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591.

Ye, C., Hill, C.M., Wu, S., Ruan, J., and Ma, Z.S. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci. Rep. **6**:31900.

Zhao, T., Tao, X., Feng, S., Wang, L., Hong, H., Ma, W., Shang, G., Guo, S., He, Y., Zhou, B., et al. (2018). LncRNAs in polyploid cotton interspecific hybrids are derived from transposon neofunctionalization. Genome Biol. **19**:195.

Zheng, Y.H., Alverson, A.J., Wang, Q.F., and Palmer, J.D. (2013). Chloroplast phylogeny of *Cucurbita*: evolution of the domesticated and wild species. J. Syst. Evol. **51**:326–334.

CAPÍTULO 3: ANÁLISIS GENÓMICO DE LA DOMESTICACIÓN DE *Cucurbita argyrosperma*

Artículo de investigación: The domestication patterns of *Cucurbita argyrosperma* are consistent with early migration events in Mesoamerica and general domestication syndromes in squashes.

Este capítulo trata sobre el estudio de la domesticación de *Cucurbita argyrosperma* utilizando genómica comparada y genómica de poblaciones. Para ello, se secuenció y ensamblo el genoma de la subespecie silvestre *C. argyrosperma* subsp. *sororia* con la misma metodología utilizada en el capítulo 1 para el genoma de *C. argyrosperma* subsp. *argyrosperma*. Ambos genomas se ensamblaron a nivel de cromosoma y se compararon, encontrando múltiples variantes estructurales asociadas al proceso de domesticación de esta especie. Se colectaron individuos de poblaciones silvestres y domesticadas de *C. argyrosperma* a lo largo de su distribución en México y se secuenciaron con la técnica de GBS para realizar análisis de demografía histórica y pruebas que permitieran detectar señales de selección. Los resultados sugieren que *C. argyrosperma* comenzó a ser domesticada en Jalisco hace aproximadamente 13,800 años, poco después de la llegada de los humanos a Mesoamérica. Dicho evento está asociado a una reducción en el tamaño efectivo de las poblaciones domesticadas, aunque el flujo genético con las poblaciones silvestres parece haber mitigado los efectos del cuello de botella. Las pruebas de selección detectaron genes candidatos asociados a características de la subespecie domesticada, tales como la pérdida de tricomas urticantes, el aumento en el tamaño de la planta, la pérdida de latencia en semillas, el aumento en el tamaño de las semillas y la concentración de carotenoides en los frutos. Finalmente, se encontraron señales de selección asociadas a un pseudogen producido por la duplicación completa del genoma en *Cucurbita*. Este capítulo fue sometido para su publicación y se encuentra en proceso de revisión.

The domestication patterns of *Cucurbita argyrosperma* are consistent with early migration events in Mesoamerica and general domestication syndromes in squashes.

Josué Barrera-Redondo°, Guillermo Sánchez-de la Vega°, Jonás A. Aguirre-Liguori, Gabriela Castellanos-Morales, Xitlali Aguirre-Dugua, Salvador Montes-Hernández, Yocelyn T. Gutiérrez-Guerrero, Erika Aguirre-Planter, Maud Tenaillon, Rafael Lira-Saadde*, Luis E. Eguiarte*

° J.B-R. and G.S-V. contributed equally to this work.

* To whom correspondence may be addressed.

Email: fruns@unam.mx and rlira@unam.mx

Abstract

Cucurbita argyrosperma is an important Mexican crop whose consumption has been focused to its seeds. It was an important crop for early Mesoamerican cultures and is a useful model to understand *Cucurbita* domestication, as seeds were probably the first trait to be selected in domesticated *Cucurbita*. We investigated the domestication history of this species through genome-wide analyses of genetic variants between the wild and domesticated populations ranging its known distribution in Mexico. We sequenced 192 individuals using tGBS libraries and assembled a novel wild reference genome to a chromosome level. Our results indicate that the domesticated subspecies of *C. argyrosperma* descends from an ancient wild population in Jalisco ˜13,800 years ago under constant gene flow. The demographic patterns of the domesticated populations coincide with the archaeological records and the migration patterns of humans throughout Mesoamerica. We detected several selective sweeps associated with the domestication of *C. argyrosperma*. We found candidate genes involved in the synthesis and regulation of growth hormones, plant defense mechanisms and phospholipid transportation, possibly associated with its domestication. We found selective signals on several uncharacterized proteins and noncoding transcripts, some of which arose during the whole-genome duplication that happened in the origin of the genus *Cucurbita*.

Significance Statement

Cucurbita argyrosperma was an important Mesoamerican crop that is still harvested today in Mexico. Its domestication process helps us understand the early years of plant domestication in America and the importance of this species as an ancient nutritional source of lipids. The analysis of population-level genomic data reveals that the domestication of *C. argyrosperma* happened soon after humans arrived at Mesoamerica, 13,800 years ago. The domestication process of this species took place in Jalisco, likely alongside maize. We also found signals of selective pressures within the genome of *C. argyrosperma* that can be attributed to this domestication process. These selective signals are associated with changes that could likely influence fruit size, seed size, seed dormancy and lipid content in seeds.

Introduction

Domestication is an evolutionary process where an organism (usually humans) selects, modifies and eventually assumes control over the reproduction of another organism, leading in many cases to a mutualistic relationship where the humans can exploit a particular resource of interest while the domesticated organism increases its fitness and its geographical range (1; 2). This is particularly true for the species of the *Cucurbita* genus (pumpkins, squashes, and gourds), whose domestication process led to their ecological success after its natural dispersers went extinct during the Pleistocene (3). This domestication process has led to the morphological, physiological and molecular differentiation between the domesticated *Cucurbita* and their wild relatives (4; 5), the latter who now have geographically restricted distributions throughout the American continent (6). Today, the domesticated *Cucurbita* taxa are successful crops that are consumed worldwide, having a global annual production of around 24 million tons with an estimated value of four billion dollars (7).

The *Cucurbita* genus is composed of ca. 14 species which have experienced at least six independent domestication events (8; 6). Despite being independent domestications, most of the *Cucurbita* crops share several traits, including the loss of bitter compounds known as cucurbitacins, the loss of defense mechanisms, loss of seed dormancy, gigantism, larger fruits, larger seeds and a high diversity of fruit morphology (7; 5). However, each *Cucurbita* crop also experienced unique selective regimes. Some good examples are the zucchini variety of *Cucurbita pepo* whose fruit is consumed at an early developmental stage, while the fruit flesh of *C. moschata* is consumed at a mature stage in most regions of Mexico but also as an immature fruit in Yucatan (4; 7). These selective regimes were predominantly defined by the nutritional and cultural necessities of early human populations in Mesoamerica (9; 10).

The initial steps of *Cucurbita* crop domestication were probably directed towards seed consumption (11; 12) since they are rich in both carbohydrates and fatty acids (13), and cucurbitacins can be removed through boiling and washing, as is still seen in the consumption of wild *Cucurbita* seeds in southern Mexico (14). While maize (*Zea mays*) acted as a main source of carbohydrates and the common bean (*Phaseolus vulgaris*) was well suited as a source of proteins, *Cucurbita* seeds were a combined source of carbohydrates and lipids for prehispanic civilizations (10; 13). In this sense, the Mexican crop *Cucurbita argyrosperma* is a good model to study the early steps of *Cucurbita* domestication, since its cultivation is directed mainly towards seed production, whereas the fruit flesh is rarely consumed, due to its poor quality compared to other *Cucurbita* crops, aside from some local specialized varieties (4).

C. argyrosperma is composed of two subspecies corresponding to the domesticated taxon (*Cucurbita argyrosperma* subsp. *argyrosperma*) and its wild relative (*Cucurbita argyrosperma* subsp. *sororia*) (4). Both subspecies are sympatrically distributed throughout the Pacific Coast of Mexico and Central America, with a few populations scattered in the coast of the Gulf of Mexico (4; 15). The domesticated taxon is also distributed in the Yucatan Peninsula, where its wild counterpart is absent (15).

The domestication of *Cucurbita argyrosperma* is presumed to have happened around the Jalisco-Balsas Basin, as hinted by archaeological and genetic evidence (8; 16; 15). The earliest archaeological record of *C. argyrosperma* was found in the Central Balsas Valley (Guerrero) alongside maize with an estimated age of 8,700 years (16), while the oldest archaeological record of a *Cucurbita* crop is that of *C. pepo*, which was found in Guil´a Naquitz (Oaxaca) with an estimated age of 10,000 years (17). Since crop domestication in Mesoamerica is linked to migration patterns and cultural development of early human populations in America (18; 10), it is expected that Mesoamerican crop species share historical demographical patterns with humans.

We studied different populations of *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia* using genome-wide data alongside a novel genome assembly of *C. argyrosperma* subsp. *sororia* and the previously reported reference genome of *C. argyrosperma* subsp. *argyrosperma* (19), both of which were assembled to a chromosome level. We investigated the demographic history of the wild and domesticated populations to infer their evolutionary relationships and propose a domestication scenario. We also performed scans to detect selection throughout the genome of *C. argyrosperma* to detect candidate regions associated with the domestication of this species and the possible link between the candidate genes and the domestication syndromes observed in *C. argyrosperma* subsp. *argyrosperma*.

Results

Genome assembly of *C. argyrosperma* subsp. *sororia*

We sequenced and assembled the genome of the wild subspecies *C. argyrosperma* subsp. *sororia* from an individual located in Puerto Escondido (Oaxaca), to find structural differences when compared to the genome of *C. argyrosperma* subsp. *argyrosperma* (19). We assembled the nuclear genome in 828 contigs with an N50 contig size of 1.3 Mbp and an L50 of 58 contigs (Table S1). A BUSCO analysis (20) against the *embryophyta odb9* database revealed 92.8% of complete BUSCOs, 1.2% fragmented BUSCOs and 6.0% missing BUSCOs within the genome assembly, similarly to other *Cucurbita* genome assemblies (21; 22; 19). We predicted 31,452 protein-coding genes within the genome assembly using BRAKER2 (23; 24). This reference genome is 9.23% larger than the genome of *C. argyrosperma* subsp. *argyrosperma* (19).

Anchoring the reference genomes into pseudomolecules

We anchored 99.97% of the *C. argyrosperma* subsp. *argyrosperma* genome assembly (19) and 98.8% of the *C. argyrosperma* subsp. *sororia* genome assembly into 20 pseudomolecules using RaGOO (25), which corresponds to the haploid chromosome number of *Cucurbita* genomes (26). Both chromosome assemblies of *C. argyrosperma* show high synteny conservation across the *Cucurbita* genus (Fig. S1), revealing a previously reported inversion in chromosome four that is shared with *Cucurbita moschata* but not with the other *Cucurbita* species (22). Most of the *C. argyrosperma* chromosomes also show chromosome-wide homoeologous pairs within the genome assembly (Fig. S1), which have been previously attributed to a whole-genome duplication event in the *Cucurbita* genus (22; 21).

We found several putative rearrangements in the genome of *C.argyrosperma* subsp. *argyrosperma* when compared to *C. argyrosperma* subsp. *sororia* (Fig. 1) and using *C. moschata* as an outgroup (Figs. S2 S21). Some of the centromeres in the wild genome were larger than in the domesticated, possibly due to a better assembly of the repetitive regions. We found several copy-number variants (CNVs), inversions, translocations and unalignable regions between the wild and the domesticated genomes (Fig. 1). The wild genome is 9.23% larger than the domesticated one (19), which could be explained by these structural variations. The genes found within the CNV losses in the domesticated genome were enriched in pectinesterases and microtubule-based processes (*p*-value < 0.05). We also found sucrose-6^F-phosphate phosphohydrolase within a CNV loss in the domesticated genome. Even though some regions were unalignable between both genomes, they share some common genes such as microtubule-associated proteins and genes related to tryptophan biosynthesis. However, such unaligned regions contain more genes in the wild genome than in the domesticated, including some proteolytic enzymes and sucrose biosynthetic genes that are not contained in the domesticated genome.

Population data and SNP genotyping

We collected plant samples across the known distribution of *C. argyrosperma* in both its domesticated populations (*C. argyrosperma* subsp. *argyrosperma*) and wild populations (*C. argyrosperma* subsp. *sororia*) throughout Mexico (Table S2). The samples were sequenced using the tunable Genotype by Sequencing (tGBS) method (27) to obtain genome-wide genetic information of the *C. argyrosperma* populations. The reads were quality-filtered and mapped against the chromosome-level genome assembly of *C. argyrosperma* ssp. *argyrosperma* to predict single nucleotide polymorphisms (SNPs) throughout the genome of each individual. We obtained an initial dataset consisting of 12,813 biallelic SNPs with a mean read depth of 50 reads per SNP and a minor allele frequency (MAF) of at least 1% corresponding to 108 individuals of *C. argyrosperma* subsp. *argyrosperma*, 44 individuals of *C. argyrosperma* subsp. *sororia*, 12 individuals of *C. argyrosperma* that were previously reported to have a semi-wild phenotype and a cultivated genotype (15) and 5 individuals of *C. moschata* that were used as outgroups (13k dataset, Dataset S1).

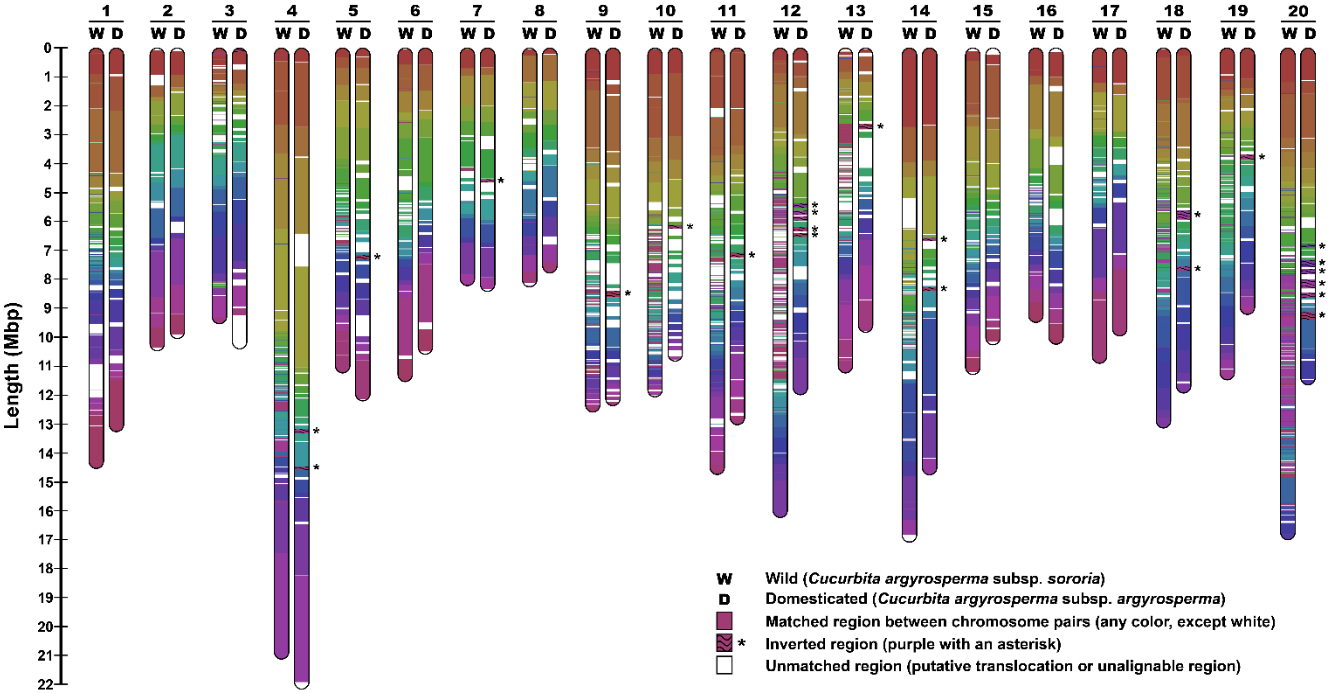

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
W D
Length (Mbp)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
*
W Wild (*Cucurbita argyrosperma* subsp. *sororia*)
D Domesticated (*Cucurbita argyrosperma* subsp. *argyrosperma*)
Matched region between chromosome pairs (any color, except white)
* Inverted region (purple with an asterisk)
Unmatched region (putative translocation or unalignable region)

Figure 1. Chromosome map representing the matching regions and putative structural variants between the genome assemblies of *C. argyrosperma* subsp. *sororia* (wild) and *C. argyrosperma* subsp. *argyrosperma* (domesticated). Matching colors represent the aligned homologous regions between both genomes, while white segments represent regions that could not be aligned to the other genome. Inverted regions are highlighted with an asterisk.

Demographic history of *C. argyrosperma* during its domestication

For this set of demographic analyses, the 13K dataset was pruned to eliminate SNPs that deviated from Hardy-Weinberg equilibrium ($p < 0.01$) as well as adjacent SNPs under linkage disequilibrium ($r^2 > 0.25$ in 100 kbp sliding windows) to obtain 2,861 SNPs that could be used to infer the demographic history of *C. argyrosperma* during its domestication (marker density of 12 SNPs per Mb).

We found a slightly higher average genetic variation in *C. argyrosperma* subsp. *sororia* compared to *C. argyrosperma* subsp. *argyrosperma* (Table 1), suggesting that the possible effects of a domestication bottleneck were alleviated by constant gene flow. At a population scale, the wild population in Jalisco had the highest genetic diversity within *C. argyrosperma* subsp. *sororia*, whie the highest diversity in *C. argyrosperma* subsp. *argyrosperma* was found in the domesticated populations distributed alongside the Pacific Coast (Table S3). The domesticated and wild populations of *C. argyrosperma* showed genetic differentiation within the species (F_{ST} = 0.0646; 95% confidence interval from 0.0565 to 0.0751), while the feral populations are more closely related to the domesticated populations (F_{ST} = 0.0479) than to the wild populations (F_{ST} = 0.1006).

We used SNPhylo (28) and ADMIXTURE (29) to evaluate the genealogical relationships and genetic structure among the wild and domesticated populations of *C. argyrosperma* (Fig. 2), finding genetic differentiation between wild and domesticated populations, as previously detected by the F_{ST} analyses. The crown node of the domesticated populations was highly supported in the Maximum Likelihood (ML) tree (bootstrap = 83), indicating a single domestication event (Fig. 2A). The wild subspecies showed additional genetic differentiation between the populations in southeastern Mexico and the populations in Jalisco, in both the ADMIXTURE assignations (Fig. 2B) and their positions in the ML tree (Fig. 2A). The wild populations of Jalisco are genetically closer to the domesticated taxon, according to the admixed assignation in the ADMIXTURE test (Fig. 2B) and by their paraphyletic position in the ML tree (Fig. 2A). The domesticated populations in Guerrero and Jalisco appear as the basal branches of the *C. argyrosperma* subsp. *argyrosperma* clade (Fig. 2A), all showing instances of genetic similarity to the wild populations in Jalisco in the K = 4 ADMIXTURE assignation (Fig. 2B).

Table 1. Average genetic diversity of the wild, domesticated and feral populations of *Cucurbita argyrosperma* using 2,861 unlinked SNPs ($r^2 < 0.25$, MAF > 1%). (N_{ind} = number of individuals, N_{pop} = number of populations, H_O = observed heterozygosity, H_E = expected heterozygosity, π = nucleotide diversity, F_{IS} = inbreeding coefficient, Var = variance)

Taxon	N_{ind}	N_{pop}	H_O (Var)	H_E (Var)	π (Var)	F_{IS} (Var)
Cucurbita argyrosperma subsp. *sororia*	44	4	0.1 (0.02)	0.1 (0.01)	0.1 (0.02)	0.01 (0.03)
Cucurbita argyrosperma subsp. *argyrosperma*	109	19	0.09 (0.01)	0.09 (0.01)	0.09 (0.01)	0.03 (0.03)
feral populations	14	3	0.1 (0.03)	0.09 (0.02)	0.1 (0.02)	-0.02 (0.02)
Cucurbita moschata (outgroup)	5	1	0.08 (0.04)	0.06 (0.02)	0.07 (0.03)	-0.02 (0.02)

The ADMIXTURE results (Fig. 2B) also show admixture events between some wild and domesticated populations, particularly between the populations of Oaxaca and the populations of Sinaloa (Fig. 2C), suggesting gene flow between sympatric populations. Despite the evidence of gene flow in Sinaloa, the feral populations are consistently grouped alongside their sympatric domesticated populations within the domesticated subspecies (Fig. 2A-B), indicating that these populations diverged recently from nearby domesticated populations. The domesticated subspecies show genetic differentiation between the western populations alongside the Pacific Coast in Mexico and the eastern populations distributed around the South Coast of Mexico, the Gulf of Mexico and the Yucatán Peninsula, with a possible recent anthropogenic migration event of eastern populations into Onavas, Sonora (Fig. 2B-C).

We used Fastsimcoal 2 (30; 31) to explore whether *C. argyrosperma* was domesticated from a wild population in southeast Mexico or from a wild population in Jalisco (Fig. 3A) and infer demographic parameters such as effective population sizes, gene flow and the time of the split between the domesticated and wild populations. Given that interbreeding has been previously observed between wild and domesticated *Cucurbita* taxa (15; 32; 33), we compared three different gene flow models (continuous gene flow, secondary contact or no gene flow) for each scenario (Fig. 3A). The 20 replicates of each model converged to similar likelihoods, indicating that the simulations performed well.

Based on the likelihood distribution of the models, the AIC values and the difference between simulated and expected SFS values, we found that the Jalisco domestication scenario with continuous gene flow had the highest likelihood (Tukey's range test *p-value* < 0.01; Fig. 3B-C). This model indicates that domestication occurred around 13,829 generations ago with an interval of 12,160 to 22,216 generations within a 1.5x interquartile range (IQR). As this species is strictly annual (4), these generations correspond to years. Our model indicates that in general migration rates were low but occurred between all the genetic groups (ranging between 1E-4 and 2E-4 migrants per generation). However, we observed higher gene flow from the wild Southern populations to the wild populations in Jalisco (m=0.033) and from the wild Southern populations to the domesticated populations (m=0.004). The simulations indicate that *C. argyrosperma* subsp. *argyrosperma* had a higher effective population size (N_e=1,238) than *C. argyrosperma* subsp. *sororia* (N_e southern=110, N_e Jalisco=105).

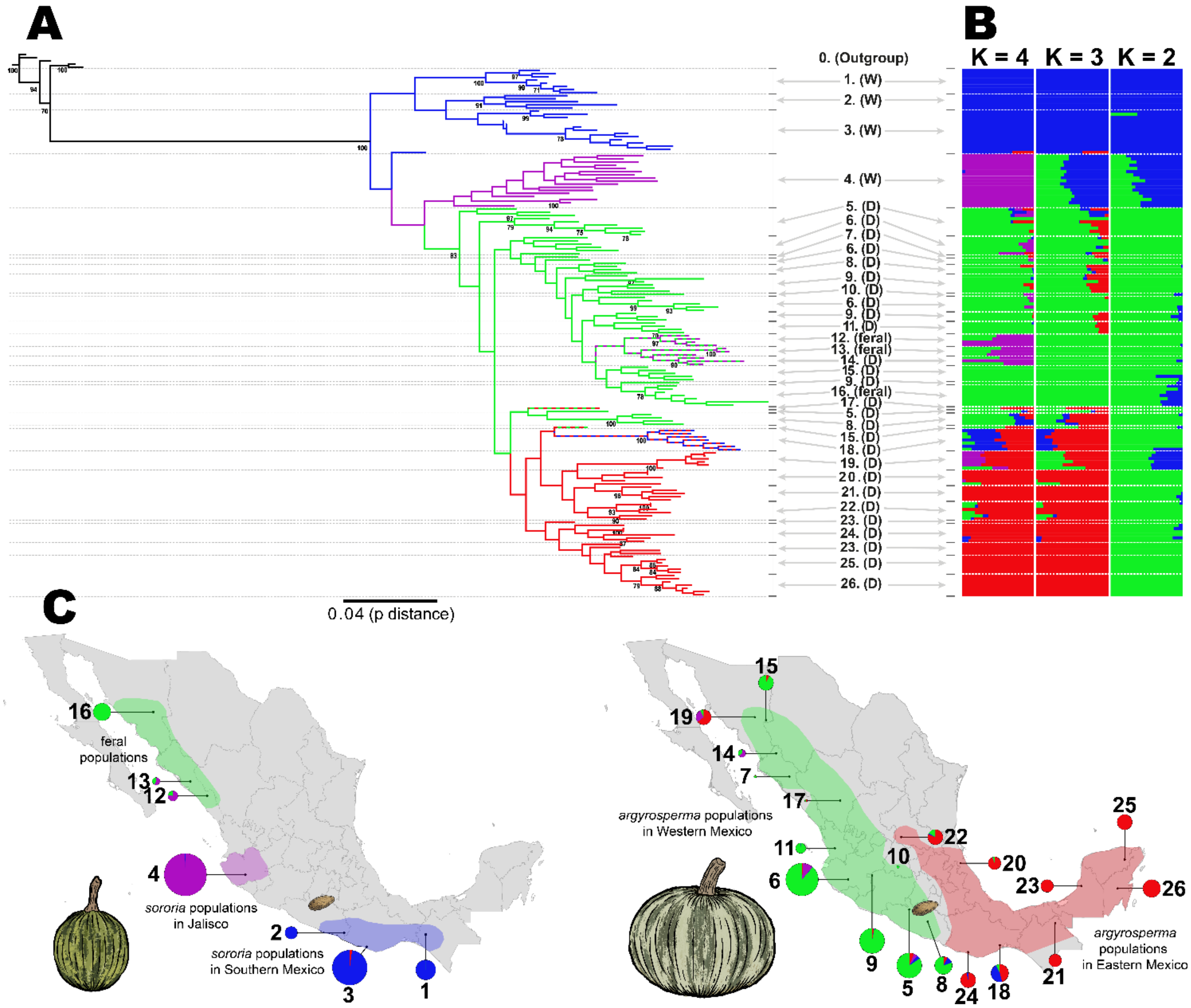


Figure 2. Genetic structure and phylogenetic relationships between wild and domesticated populations of *Cucurbita argyrosperma* based on 2,861 SNPs. (A) Maximum Likelihood tree with 100 bootstraps (only bootstrap values > 70 shown) (B) ADMIXTURE analysis using K values ranging from 2 to 4. (C) Geographic distribution of wild (down left) and domesticated (down right) populations, with pie chart colors representing the ADMIXTURE assignation of the individuals in 4K ancestral populations (size of pie charts proportional to sample size). The seed in the maps represent the earliest archaeological record of argyrosperma from Xihuatoxtla, Guerrero (dated 8,700 years BP) (16).

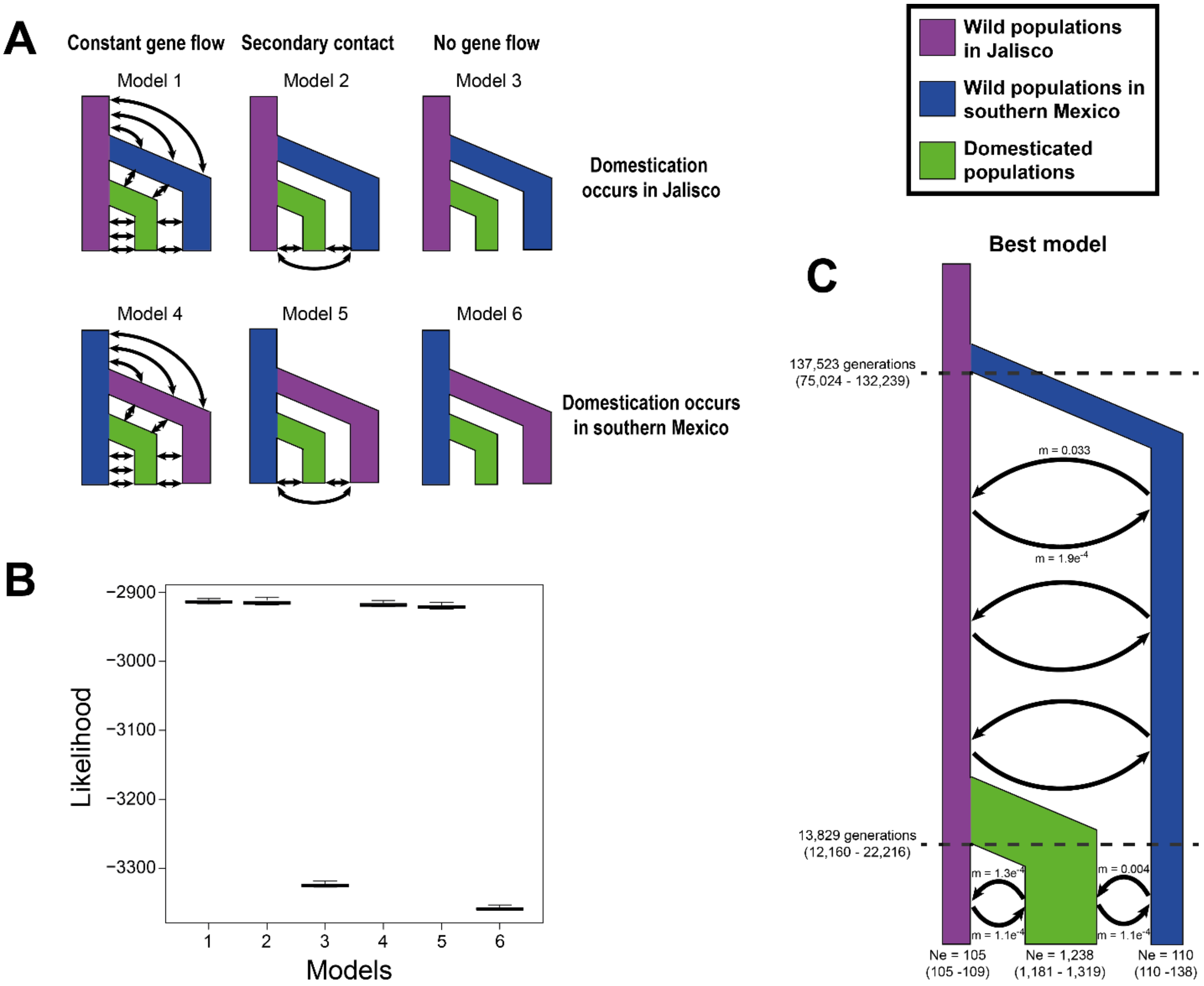


Figure 3. Coalescent simulations and most likely domestication scenario of *Cucurbita argyrosperma*. (A) Six different models were simulated to assess their compatibility with the unfolded multidimensional Site Frequency Spectrum of our data. (B) Comparison of the likelihood of all the models. All models were significantly different based on a Tukey test (p-value < 0.01). (C) Representation of the model that best fits the data, including the estimated parameters of divergence time, migration rates and effective population sizes.

Domestication sweeps in *C. argyrosperma*

In order to perform the tests to detect selective sweeps associated with the domestication of *C. argyrosperma*, we removed from the 13k dataset the *C. moschata* individuals as well as the feral individuals of *C. argyrosperma*. We used the same 1% MAF threshold for this subset, obtaining a 10,617 SNP dataset suitable to detect selective sweeps, with a marker density of 44 SNPs per Mb and a considerably fast LD decay throughout the genome (Fig. S22). We performed two F_{ST} based tests as implemented in BayeScEnv (34) and PCAdapt (35) to detect selective sweeps between the domesticated and the wild populations of *C. argyrosperma* (Fig. 4A).

We found 334 outlier SNPs by either BayeScEnv or PCAdapt , of which 19 SNPs were predicted as outliers by both tests (Fig. 4A). By assuming a moderate LD within the selective sweeps, we compared the genomic regions 5kp upstream and downstream these 19 candidate SNPs between the wild and the domesticated reference genomes, finding several structural differences that could be attributed to the selective signals (Table S4). We found the arabinogalactan protein AGP18 close to one of these SNPs, which showed two insertions and one deletion in the first exon of the domesticated genome, disrupting its open reading frame (Fig. 4B). We also found several indels between a 17.8 kDa class I heat shock protein (HSP17.8) and an uncharacterized protein, one of whose size was 422 nt long, possibly disrupting an unknown regulatory sequence (Fig. 4C). We found two serine/threonine-protein kinases, PBL10 and PBL23, where the former shows several deletions upstream of the gene (Table S4) and the later shows a 915 nt insertion downstream of the gene, possibly disrupting its transcriptional activity (Fig. 4D). In addition, we found a strong candidate SNP within an intron of a Ribonucleases P/MRP protein subunit (POP1) homolog with a 229 nt insertion upstream of the gene (Fig. 4E).

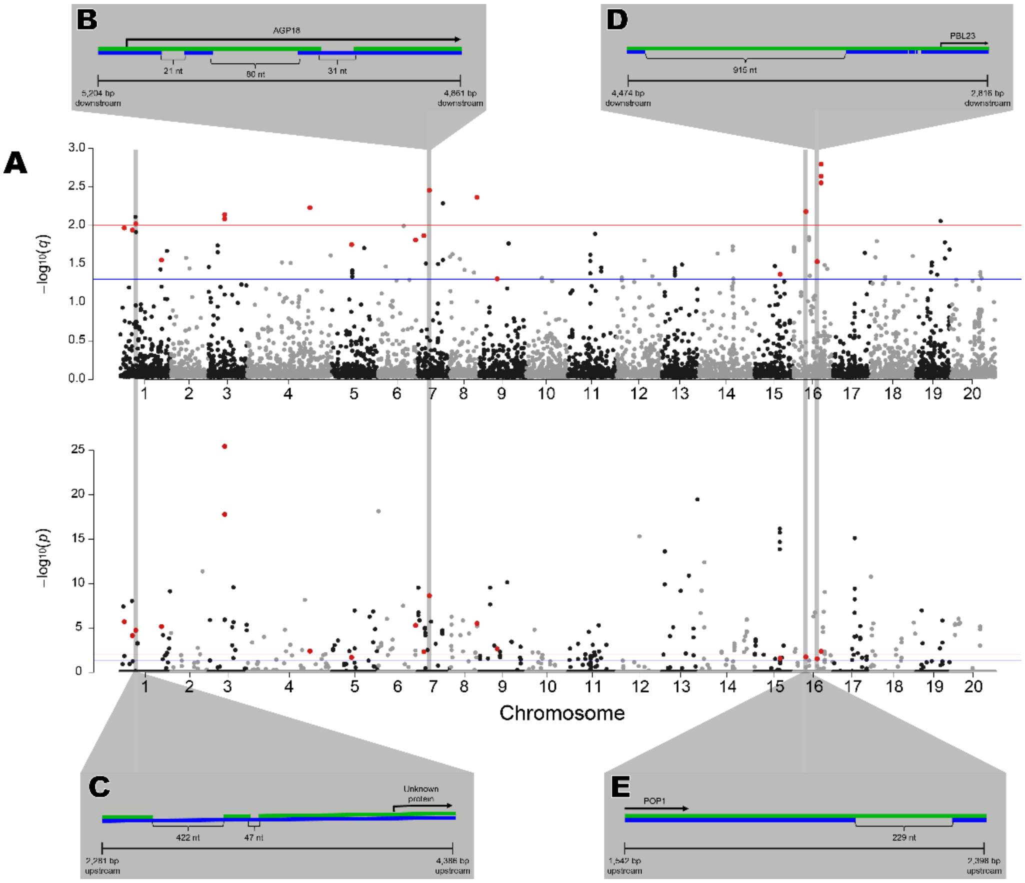


Figure 4. Footprints of selection associated with the domestication of *Cucurbita argyrosperma*. (A) Manhattan plots representing the BayeScEnv (up) and PCAdapt (down) tests in each chromosome of the genome. The blue line indicates a p-value < 0.05 and the red line indicates a p-value < 0.01 for each of the tests. The 19 SNPs that were considered as outliers by both tests are represented as red dots. (B-E) These 19 outlier SNPs were further analyzed by aligning the domesticated reference genome (green lines) against the wild reference genome (blue line) 5,000 bp in the vicinity of the candidate SNP (see Table S4). Some examples include: (B) the disruption of the AGP18 gene by two insertions and one deletion in the domesticated genome, (C) two large deletions in the domesticated genome between the HSP17.8 gene and an uncharacterized protein, (D) a huge insertion downstream of the PBL23 gene, (E) and a large insertion upstream of a POP1 homolog.

We found 125 genes including candidate SNPs by either test within their structure (*i.e.*, introns, exons, UTRs; Table S5). After performing a Gene Ontology enrichment analysis, we found seven significantly enriched biological processes in these 125 candidate genes, including abscisic acid (ABA) biosynthesis and brassinosteroid-mediated signaling (Table S6). Among these candidate genes, we found the homolog of Auxin response factor 1 (ARF1), WRKY transcription factor 2 (WRKY2), ABC transporter C family member 2 and ABC transporter E family member 2. We also found that the zeaxanthin epoxidase activity was significantly enriched in the 125 candidate genes (Table S6), including one of the two copies of zeaxanthin epoxidase within the genome of *C. argyrosperma*. We detected CLAVATA3/ESR-related protein 10 (CLE10), Tubulin alpha chain (TUBA), Cellulose synthase interactive 1 (CSI1), seven serine/threonine-protein kinases and a phosphatidylinositol/phosphatidylcholine transfer protein (SFH9) among our candidate genes, the last which was also a candidate gene by both BayeScEnv and PCAdapt (Table S4).

We looked for the orthologous sequences of the cucurbitacin pathway reported in *Cucumis sativus* (36) within the *C. argyrosperma* subsp. *argyrosperma* genome through a bidirectional BLAST analysis (Table S7). However, we found no evidence of selective pressures on any cucurbitacin-related gene.

We found 13 uncharacterized proteins among the 125 candidate genes (Table S5) and three uncharacterized proteins among the convergent selective signals (Table S4). All these uncharacterized proteins were multiexonic and showed Annotation Edit Distances < 0.5 (37), suggesting they are most likely real genes and not annotation artifacts. Eleven of these uncharacterized proteins were found to have homologs among the NR database, revealing their phylogenetic conservation. However, one of these genes (Carg11153) was only found within genomes of the Cucurbitaceae species, finding single orthologs in several cucurbits but in double copy within the *Cucurbita* genomes (22; 21). Finally, we found an uncharacterized protein (Carg25109) from which we could not retrieve any homolog sequences when comparing it to the NR database, making it a potential *de novo* gene in *C. argyrosperma*.

We found that eight of the PCAdapt candidate SNPs reside within seven previously described long noncoding RNAs in the genome of *C. argyrosperma* subsp. *argyrosperma* (19), as well as two long noncoding RNAs in close proximity to the convergent selective signals (Table S4). Two of these long noncoding RNAs reside within chromosome 3 and are homologous to the RMD5 gene found in chromosome 7 (Fig. S23). Coincidentally, these long noncoding RNAs are found in the same chromosome as another previously described long noncoding RNA with possible signals of purifying selection that is homologous to a trafficking protein particle found in chromosome 7 (19)(Fig. S23). The other five long noncoding RNAs show sequence similarity to either mitochondrial-like or plastid-like genomic sequences but reside within four different chromosomes of the nuclear genome. Most of these transcripts are multi-exonic, suggesting they are not artifacts produced by "transcriptional noise" (38; 39). Furthermore, a BLASTn search of these organelle-like transcripts revealed they were also found in the same nuclear chromosomes of the *C. argyrosperma* subsp. *sororia* genome assembly, suggesting they are not a product of genome misassemblies, but rather organellar introgressions into the nuclear genome.

Discussion

Effects of domestication on genetic diversity

The genome assembly of the domesticated subspecies was smaller than that of wild subspecies, which is possibly caused by the loss of structural variants during its domestication. Many of these unaligned regions contain entire genes, making this wild taxon a reservoir of potentially adaptive presence/absence variants. The genetic diversity of the domesticated populations (*C. argyrosperma* subsp. *argyrosperma*) is just slightly lower than in the wild populations (*C. argyrosperma* subsp. *sororia*), suggesting that the effects of a domestication bottleneck were not as dramatic as in many other crop species (1). The effects of the domestication bottleneck were probably alleviated by the constant genetic flow between both subspecies, as observed in our data and in previous studies (33). This constant gene flow may be related to the sympatric distribution of the wild and domesticated populations of *C. argyrosperma* throughout the pacific coast of Mexico (15) and by their coevolved pollinators *Peponapsis* and *Xenoglossa* in both subspecies (40). In this sense, the phenotype of the feral populations could be attributed to a combination of wild genetic introgression and phenotypic plasticity, rather than fast adaptive changes to the wild conditions.

The concept of a drastic domestication bottleneck has been cast into doubt by recent archaeogenetic studies, as this phenomenon may be accentuated latter throughout a long period of time as well as during crop improvement (41). Since *C. argyrosperma* subsp. *argyrosperma* is a traditional crop composed mostly of landraces (4), the effects of a genetic bottleneck may not be as harsh as in other crop species. The genetic diversity in *C. argyrosperma* subsp. *argyrosperma* is maintained by the traditional agricultural practices and selection criteria performed by Mexican farmers, which promote the conservation of local landrace varieties throughout the country (42). Several studies have shown that traditional agricultural practices are a fundamental force that maintains the diversity of crop species (43; 44). This is also true for *Cucurbita* species, where traditional agriculture and diverse criteria of human selection promotes the diversity of landraces (45; 46).

Furthermore, the wild *Cucurbita* populations have also been subjected to demographic bottlenecks after the extinction of the Pleistocene megafauna which acted as their natural dispersals (3), previously reducing their genetic diversity in a similar fashion to a domestication bottleneck. This is supported through chloroplast and microsatellite data in *Cucurbita pepo*, whose wild populations show a lower genetic diversity than their domesticated counterparts due to its small geographic distribution (47).

Domestication of *C. argyrosperma* in Jalisco

Our results suggest that the initial population of *C. argyrosperma* subsp. *sororia* from which *C. argyrosperma* subsp. *argyrosperma* originated was distributed within the state of Jalisco, as revealed by its genetic closeness to *C. argyrosperma* subsp. *argyrosperma*. The genetic closeness between the wild populations in Jalisco and the domesticated taxon was also observed with mitochondrial and microsatellite markers (8; 15). Our coalescent simulations also support this scenario, suggesting that *C. argyrosperma* started its domestication process around 13,800 generations ago, which can be roughly extrapolated to years considering that mesophilic *Cucurbita* species display an annual life cycle (4). This coalescent event predated the oldest archaeological record of a domesticated *C. argyrosperma*, around 8,700 years ago (16; 48) and is in rough agreement with the earliest archaeological records of human presence in Mesoamerica, around 13,000 years ago (49), although humans inhabited the continent as early as 18,500 years ago (50). This suggests that *C. argyrosperma* may have undergone management practices during early human arrival to Mesoamerica (17), leading to a genetic differentiation prior to the fixation of clear domestication traits that were subsequently found on the archaeological records (51; 16; 48).

Our data show that the domesticated populations found within Guerrero and Jalisco are the most closely related to the wild population in Jalisco. This means that even though the wild ancestors of *C. argyrosperma* subsp. *argyrosperma* came from Jalisco, the domestication process may have occurred throughout the JaliscoBalsas region, bringing these ancestral populations to other states such as Guerrero, where their oldest archaeological remains were found (16). The idea of a domestication center in the Jalisco-Balsas basin has also been previously suggested using mitochondrial and microsatellite markers (8; 15), as well as archaeological data (16). Similarly, ancient human migration events have been documented from Jalisco to the rest of the Balsas Basin between

10,000 to 7,000 years ago, probably bringing with them several crop species such as *C. argyrosperma*, *Zea mays*, and *Phaseolus vulgaris* as food sources (16; 18; 52). These migration patterns may explain the genetic cohesiveness between the domesticated populations of *C. argyrosperma* found throughout the western Pacific Coast in Mexico, representing the first fully domesticated lineages of the species (18). Previous studies based on 8,700 years old phytoliths found in Xihuatoxtla, Guerrero, suggest a parallel domestication of *Zea mays* and *C. argyrosperma* in the Balsas-Jalisco region within this time period (16; 48) that may have been disseminated along these human migration events (18). Overall, the genetic patterns of *C. argyrosperma* structure are coherent with the archaeological evidence of early human migration throughout Mesoamerica (18; 49) (Fig. 5).

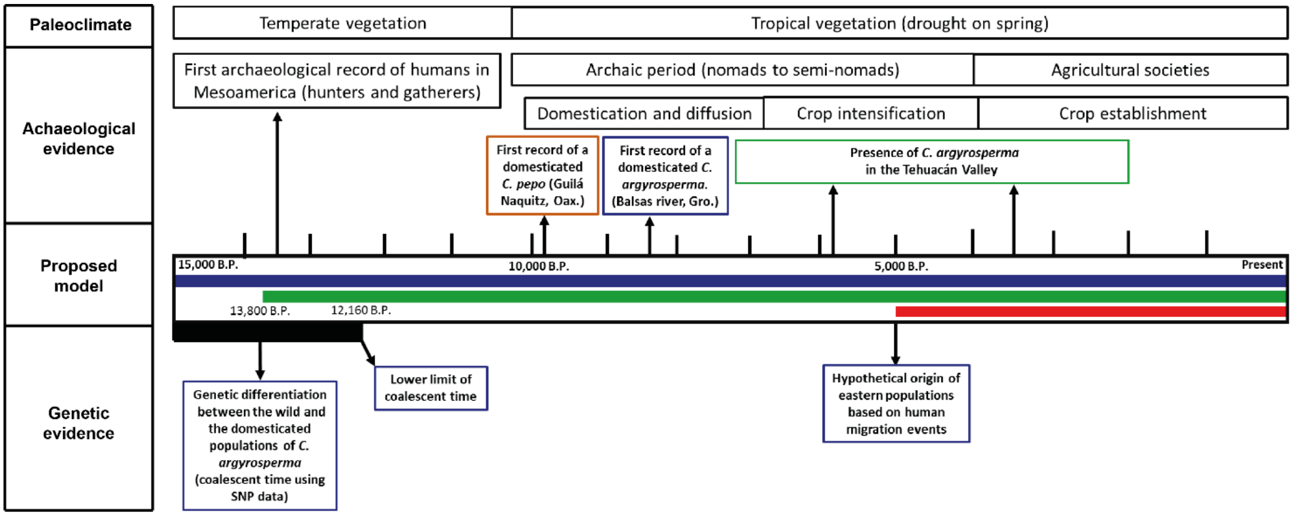

Paleoclimate
Temperate vegetation
Tropical vegetation (drought on spring)
Achaeological evidence
First archaeological record of humans in Mesoamerica (hunters and gatherers)
Archaic period (nomads to semi-nomads)
Agricultural societies
Domestication and diffusion
Crop intensification
Crop establishment
First record of a domesticated *C. pepo* (Guilá Naquitz, Oax.)
First record of a domesticated *C. argyrosperma.* (Balsas river, Gro.)
Presence of *C. argyrosperma* in the Tehuacán Valley
Proposed model
15,000 B.P.
10,000 B.P.
5,000 B.P.
Present
13,800 B.P.
12,160 B.P.
Genetic evidence
Genetic differentiation between the wild and the domesticated populations of *C. argyrosperma* (coalescent time using SNP data)
Lower limit of coalescent time
Hypothetical origin of eastern populations based on human migration events

Figure 5. Model of *Cucurbita argyrosperma* domestication based on archaeological and genetic data. The colored lines represent the presence of each *C. argyrosperma* population in Mesoamerica (blue = wild populations, green = domesticated populations in western Mexico, red = domesticated populations in eastern Mexico) Archaeological and paleoclimatic data obtained from (8; 48; 18).

The selective sweeps in *C. argyrosperma* can be traced back to some domestication traits

Even though tGBS sequencing has a limited capacity to detect selective sweeps across the genome (53), we found several signals of outlier SNPs between the domesticated and wild populations of *C. argyrosperma*. The SNP density for our selection tests was of 44 SNPs per Mb, which is one order of magnitude denser than other studies using reduced representation genome sequencing to detect selective sweeps (54). This is a consequence of the relatively small genome size of *C. argyrosperma*, around 229 Mb (19). Nonetheless, the LD in *C. argyrosperma* decays at a shorter length than our SNP density, so our genome scans should be interpreted as a partial representation of the selective sweeps associated with the domestication process (53). Fortunately, the use of a wild reference genome helped us investigate the vicinity of the strong candidate SNPs, revealing large insertions or deletions either upstream or downstream of the genes that were close to the region under selection. These structural variants could potentially be linked with the disruption of the transcriptional activity of their nearby genes, as the mutations associated with domestication traits are usually found within cis-regulatory regions, rather than within the gene itself (55). Future studies should include large indels within the selective scans to confirm its role in the appearance of domestication traits.

Some SNPs that were retrieved as outlier SNPs in both tests were found near genes involved in biotic and abiotic plant defense responses. For example, HSP17.8 has been proven to activate in response to heat (56), oxidative stress (57), and salt stress (58). Likewise, PBL10 and PBL23 have been suggested to be involved in plant defense pathways due to its similarity to other serine/threonine-protein kinases (59). This is concordant with previous studies showing that induced defense mechanisms are usually selected against during plant domestication, as the products of these responses are usually unpleasant or harmful to humans when the plant is consumed (60).

Our Gene Ontology enrichment analysis found several genes enriched in the ABA biosynthesis and brassinosteroidmediated signaling. This suggests that the alteration of growth hormones may play an important role in *C. argyrosperma* domestication. ABA is involved in a myriad of functions such as the regulation of plant growth, plant development, seed dormancy and response to biotic/abiotic stress (61). In this sense, the lack of dormancy in seeds and gigantism are both common domestication changes (62) that are present in domesticated cucurbits that may be caused by changes in the regulation of ABA and brassinosteroids (63; 5). Particularly, phytohormone regulation may be involved in *Cucurbita* fruit size alongside microtubule-related genes (5). We found WRKY2 among our candidate genes, which is involved in seed germination and postgermination development in *A. thaliana* (64) and may explain the lack of seed dormancy in *C. argyrosperma* subsp. *argyrosperma*. We also found CLE10 among our candidate loci, which is involved in the regulation of cell fate and proliferation in the apical meristems that is homologous to CLAVATA3 (65). CLAVATA3 and its pathway have been linked to fruit size in tomato (66), as meristem proliferation is correlated to fruit size (67). We also detected two genes involved in microtubule formation, namely TUBA and CSI1 (68; 69), which may play a role in fruit size or other morphological differences between the wild and domesticated subspecies.

Among specific candidate genes, we found two ABC transporters under selection in *C. argyrosperma*. ABC transporters have been known to be involved in the transmembrane transport of ABA-GE, an ABA conjugate that is usually attributed to the plant response against water stress (70). However, studies made in *Hordeum vulgare* suggest that the transport of ABA-GE may play a role in seed development alongside *de novo* ABA synthesis within the developing seed (71), suggesting a role of ABC transporters in the seed development of *C. argyrosperma*. Previous studies have also found an association between variants in ABC transporter proteins and seed size in *Cucurbita maxima* (72) and *Linum usitatissimum* (73), further suggesting that ABA may be deregulated via selective pressures on ABC transporters to enhance seed size in *C. argyrosperma* during its domestication. Additionally, variants in a serine/threonine-protein kinase as the ones we found in our selective scans have also been associated with seed size in *Cucurbita maxima* (72).

We found SFH9 as a candidate gene by both BayeScEnv and PCAdapt, which belongs to a gene family involved in the transportation of phosphatidylinositol and phosphatidylcholine throughout the plant cells (74). Phosphatidylcholine is a major component of the phospholipid content in *Cucurbita* seeds (75), suggesting phospholipid content could be targeted through artificial selection for its nutritional value (10).

We also found a significant enrichment of zeaxanthin epoxidase activity, and a zeaxanthin epoxidase homolog, among our candidate genes. Zeaxanthin epoxidase is linked to the degradation of carotenoids in plant seeds (76), which in turn are important precursors of ABA biosynthesis, regulating processes such as germination and maturation (77). Carotenoid degradation may also be deregulated in fruit tissue, as the domesticated *Cucurbita* shows a larger concentration of carotenoids in their fruits compared to their wild relatives, giving them their characteristic orange fruit flesh (78).

Bitterness is a monogenic trait that differentiates the wild and cultivated subspecies of *C. argyrosperma* and the rest of the domesticated *Cucurbita* (51). Even though we expected strong selective pressures against bitterness in *C. argyrosperma* subsp. *argyrosperma*, none of the genes associated with the selective sweeps could be traced back to an ortholog of the cucurbitacin pathway reported in *Cucumis sativus* (36). This may probably be a limitation of our tGBS sequencing data, limiting our ability to detect this selective sweep (54). Thus, future studies should aim to use whole-genome resequencing to obtain a better understanding of the selective pressures in *Cucurbita* species during domestication.

Selective sweeps on novel genomic elements

We found 13 uncharacterized genes associated with the domestication sweeps of *C. argyrosperma*. Most of these proteins are conserved across multiple plant species, suggesting they are important in a broad biological context. However, we also found two uncharacterized genes with a restricted taxonomic presence. *De novo* genes have been previously shown to be involved in agronomically relevant domestication traits (79), which should be further analyzed directly on the species of interest.

Interestingly, we found several long noncoding RNAs under selection, which suggests a few trans-regulatory elements may play a part in *C. argyrosperma* domestication, as opposed to the usual selection of cis-regulatory elements throughout domesticated plants (55). Selective pressures on long noncoding RNAs have been previously reported on other domesticated crops such as *Phaseolus vulgaris* (80), indicating the importance of analyzing noncoding transcripts on domestication studies. Two of the long noncoding RNAs with evidence of selective sweeps were found in chromosome 3 and were homologous to the RMD5 gene found in chromosome 7 of *C. argyrosperma*. Interestingly, chromosomes 3 and 7 share chromosome-wide homoeologous blocks, strongly suggesting that these two overlapping long noncoding RNAs originated from a single pseudogenization process of one of the duplicated copies of RMD5 after the whole-genome duplication event in *Cucurbita* (22; 21; 19). Despite becoming pseudogenes, these noncoding transcripts show signals of selective sweeps during the domestication process of *C. argyrosperma*, suggesting they might still be functional. A previously reported long noncoding RNA in *Cucurbita argyrosperma* (Carg TCONS 00015392) also showed homology to the trafficking protein particle complex subunit 11 (TRAPPC11) and presented both evidence of purifying selection and a thermodynamically stable secondary structure (19). We found that Carg TCONS 00015392 and TRAPPC11 are also found in chromosomes 3 and 7, respectively. This strongly suggests that the fractionation process between chromosomes 3 and 7 led to the pseudogenization of the RMD5 and TRAPPC11 copies in chromosome 3, who are still transcriptionally active. Given the evidence of selection acting on these pseudogenes, they may now play regulatory roles as long noncoding RNAs within the genome of *C. argyrosperma* (19).

The five organelle-like long noncoding transcripts with evidence of selective sweeps reside within different chromosomes of the nuclear genome of *C. argyrosperma* and most of them are multi-exonic, suggesting they are real noncoding RNAs rather than annotation artifacts. The presence of organellar sequences within the nuclear genome is not surprising, as *Cucurbita* genomes are known to suffer from multiple duplication and transfer events between the different genomic compartments of the cell, especially within the mitochondrial genome (81; 82). This phenomenon, coupled with the already high rates of nuclear genome evolution in *Cucurbita* (19), indicates that organellar transfers into the nucleus may be a source of novel regulatory elements in *Cucurbita* genomes. The function of all these candidate long noncoding RNAs or their role in the domestication of *C. argyrosperma* remains elusive, but our results indicate that these noncoding transcripts are good candidates for future functional characterization.

Concluding remarks

This work represents the first study of domestication in a *Cucurbita* species using genomic data, unraveling its underlying demographic patterns and selective pressures. We consider that the sympatric distribution and constant gene flow between *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, in addition to the maintenance of genetic diversity through traditional agricultural practices, has aided in the conservation of genetic variation in this species.

Our analyses support the notion that *C. argyrosperma* was domesticated in Jalisco and the adjacent Balsas region, as suggested by previous studies (8; 15), alongside other important crop species such as *Zea mays* (16) and *Phaseolus vulgaris* (80). The demographic patterns of *C. argyrosperma* we observe in our genomic data coincide with the earliest presence of humans in Mesoamerica (49) and the emerging patterns of human migration throughout the biological and cultural corridors in Mesoamerica (18).

We found several signals of selection associated with domestication traits in *C. argyrosperma* subsp. *argyrosperma,* as well as selective pressures on uncharacterized proteins and long noncoding RNAs that should be further studied. The list of candidate loci will be useful for future studies where such genes could be validated, as candidate domestication genes are usually of agronomic value for crop improvement (83) or *de novo* crop domestication (84; 85).

The availability of a wild *Cucurbita* genome is a helpful resource for future studies that may want to avoid using reference genomes of domesticated taxa in a comparative fashion. It will also serve as a resource to find genetic variation in wild *Cucurbita* populations that may be useful for improvement programs (86).

Methods

Genome assembly and annotation of *Cucurbita argyrosperma* subsp. *sororia*

We sequenced the genome of a *C. argyrosperma* subsp. *sororia* individual collected in Oaxaca, Mexico. Total DNA was extracted from leaf tissue and sequenced using PacBio Sequel and Illumina HiSeq4000. We filtered the Illumina sequences using the qualityControl.py script (https://github.com/Czh3/NGSTools) to retain the reads with a PHRED quality ≥ 30 in 85% of the sequence and an average PHRED quality ≥ 25. The Illumina adapters were removed using SeqPrep (https://github.com/jstjohn/SeqPrep) and the paired reads that showed overlap were merged. The chloroplast genome was assembled using NOVOplasty (87) and the organellar reads were filtered using Hisat2 (88) against the chloroplast genome of *C. argyrosperma* subsp. *argyrosperma* (19) and the mitochondrial genome of *C. pepo* (81). We assembled the nuclear genome into small contigs using the Illumina reads and the Platanus assembler (89). The Platanus contigs were assembled into larger contigs using the PacBio Sequel reads and DBG2OLC (90). We performed two iterations of minimap2 and racon (91) to obtain a consensus genome assembly by mapping the PacBio reads and the Platanus contigs against the DBG2OLC backbone. We performed three additional polishing steps using PILON (92) by mapping the Illumina reads against the consensus genome assembly with BWA mem (93).

The genome annotation processes were performed using the GenSAS v6.0 online platform (94). The transposable elements within the genome were predicted and masked using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/). We downloaded five RNA-seq libraries of *Cucurbita argyrosperma* available on the Sequence Read Archive (accessions SRR7685400, SRR7685404 - SRR7685407), preformed the same quality filters described above for the Illumina genomic sequences, and aligned the high-quality reads against the masked genome of *C. argyrosperma* subsp. *sororia* using STAR v2.7 (95). We used filterBAM from the Augustus repository (96) to filter low-quality alignments and used the remaining alignments as RNA-seq evidence for BRAKER2 (96; 97; 23; 24).

Anchoring the reference genomes into pseudomolecules

We obtained the Pacbio corrected reads from the PacBio RSII reads of *C. argyrosperma* subsp. *argyrosperma* (NCBI SRA accession SRR7685401) and the PacBio Sequel reads of *C. argyrosperma* subsp. *sororia* using CANU (98). We anchored the genome assemblies of *Cucurbita argyrosperma* subsp. *argyrosperma* (19) and *C.*

argyrosperma subsp. *sororia* into pseudomolecules using RaGOO (25) alongside the PacBio corrected reads of each taxon to detect and correct misassemblies, using a gap size of 2600 bp for chromosome padding (average gap length of *C. argyrosperma* subsp. *argyrosperma* genome assembly) and using the genome assembly of *Cucurbita moschata* (22) as reference. We evaluated the synteny and rearrangements between *Cucurbita* genomes using Synmap2 (99), prommer (100) and Smash (101) using a minimum block size of 100,000 nt, a threshold of 1.9 and a context of 28.

Sample collection, DNA extraction and sequencing

We collected seeds from 26 populations of *C. argyrosperma* subsp. *argyrosperma* (domesticated populations), 15 populations of *C. argyrosperma* subsp. *sororia* (wild populations), and three feral populations, covering most of the reported distribution of this species throughout Mexico (6). The seeds were germinated in a greenhouse until the seedlings started to grow leaves. We extracted total DNA from fresh leaves using a DNeasy Plant Kit (Qiagen) of 192 individuals across the collected populations (Table S2), including five individuals of *Cucurbita moschata* to be used as outgroup, all which were sequenced by Data2Bio LLC using the tGBS method (27) with an Ion Proton instrument and two restriction enzymes (Sau3AI/BfuCI and NspI).

Data filtering and SNP genotyping

The raw reads were trimmed using LUCY2 (102), removing bases with PHRED quality scores < 15 using overlapping sliding windows of 10bp. Trimmed reads shorter than 30 bp were discarded. The trimmed reads were mapped against the chromosome-level genome assembly of *Cucurbita argyrosperma* subsp. *argyrosperma* using segemehl (103), since empirical studies suggest this program outperforms other read-mapping software while using Ion Torrent reads (104). We only retained the reads that mapped uniquely to one site of the reference genome for subsequent analyses. We used BCFtools (105; 106) for an initial variant calling step, retaining variants with at least 6 mapped reads per individual per site where the reads had a minimum PHRED quality score of 20 in the called base and a minimum mapping quality score of 20 (107). We used plink (108) to perform additional filters, such as retaining only biallelic SNPs, retaining SNPs with no more than 50% of missing data, individuals with no more than 50% of missing data and sites with a MAF of at least 1% (13K dataset).

In order to obtain an adequate SNP dataset to infer the demographic history of *C. argyrosperma*, we performed additional filters to the 13k dataset with plink (108), including the elimination of all the SNPs that diverged significantly ($p < 0.01$) from the Hardy-Weinberg equilibrium exact test (109), and the elimination of adjacent SNPs in the genome with a squared correlation coefficient (R^2) larger than 0.25 within 100 kbp sliding windows with a step size of 100 bp.

In order to obtain an adequate SNP dataset to detect selective sweeps associated with the domestication of *C. argyrosperma*, we eliminated all the feral individuals of *C. argyrosperma*, which could not be assigned to either a wild or a domesticated population, as well as the five individuals of *C. moschata*. We also eliminated the SNP sites with more than 50% missing data and performed a MAF filter of 1% after reducing the number of individuals in the 13K dataset. The SNP density in this dataset was calculated using VCFtools (110) and the LD decay was calculated using plink (108) with a minimum R^2 threshold of 0.001.

Population structure

We used diveRsity (111) to calculate the pairwise F_{ST} statistics, using 100 bootstraps to calculate the 95% confidence intervals. We used STACKS (112) to calculate the genetic variation in the wild, domesticated and feral populations. We used ADMIXTURE (29) to evaluate the genetic structure among the wild and domesticated populations of *C. argyrosperma*, evaluating the assignation of individuals into one (CV error = 0.26205), two (CV error = 0.25587), three (CV error = 0.25806) and four (CV error = 0.26658) *K* populations. We used SNPhylo (28) to reconstruct a neighbor-joining maximum-likelihood tree with genetic distances between all the individuals in the dataset. We performed 100 bootstraps to assess the reliability of the tree topology.

Coalescent simulations

We used coalescent simulations to test two different possible scenarios of divergence between the genetic groups observed in our ADMIXTURE and SNPhylo results to determine the most likely region where domestication occurred. In the first scenario, the wild populations in Jalisco represent the oldest lineage and the wild populations in southern Mexico and the domesticated populations coalesce with it. In the second scenario, the wild group in southern Mexico is the oldest lineage and the wild populations in Jalisco and domesticated populations coalesce with it. We also tested three scenarios of divergence. The first scenario assumes that gene flow occurred continuously during the divergence, the second scenario assumes that divergence occurred without gene flow and there was a posterior secondary contact, where gene flow occurred. The third scenario assumes that divergence occurred without gene flow and divergent populations never came into contact. For each model, we estimated the time of divergence (T) between genetic groups and their effective population size (Ne). For models that included migration between groups, we also estimated the migration rate (m). We used Fastsimcoal 2 (30; 31) to determine the parameters that maximize the composite likelihood of each model given the unfolded multidimensional Site Frequency Spectrum (SFS).

The unfolded multidimensional SFS was obtained with DADI (113), using 17 wild individuals of Jalisco, 27 wild individuals of southern populations, 123 domesticated individuals and 5 individuals of *C. moschata* used as an outgroup to unfold the SFS. We ran 100,000 simulations with 20 replicates for each model (two divergence scenarios and three gene flow scenarios) using the following settings: -M 0.001, -C, -L 40 and –N 200,000. We also selected log-uniform priors for parameter estimations, setting times of divergence between 1000 and 200,000 generations, effective population sizes between 100 and 60,000 individuals and migration rates between 0.0001 and 0.5. We also constringed the times of divergence in all scenarios, forcing the domesticated taxa to diverge after the wild relatives. After corroborating that all replicates converged to similar likelihoods, we combined all replicates and retained all outputs that were above the 95% of the likelihood distribution. We selected the best model based on a Tukey test (p-value < 0.01) between the top 5% distribution of the composite likelihoods of the six models, the Akaike information criteria of the highest likelihood of each model and the differential between the highest observed and estimated composite likelihoods.

Tests to detect selective sweeps

We used BayeScEnv (34) to detect candidate SNPs that were differentiated between the wild and domesticated populations of *C. argyrosperma*. For the "environmental" values used by BayeScEnv, we assigned each population as either wild (0) or domesticated (1). The SNPs with *q-values* < 0.05 were regarded as candidate loci under selection.

The Mahalanobis distances implemented in PCAdapt (35) were used to detect candidate SNPs after controlling for the first two principal components in our dataset, which correspond to the subspecies and geographical differentiation observed during the population structure analysis. We performed Bonferroni corrections to adjust the *p-values* and the SNPs with *p-values* <0.05 were regarded as candidate loci under selection.

We used snpEff (114) to associate the candidate loci found in both tests with the genome annotation of *C. argyrosperma* subsp. *argyrosperma* (19). The genes that could be unambiguously assigned to a candidate SNP were screened for a Gene Ontology enrichment analysis using topGO and the *weight01* algorithm (115). We determined the significantly enriched biological functions by performing Fisher's exact test.

The convergent SNPs under selection were compared between the *C. argyrosperma* subsp. *argyrosperma* and the *C. argyrosperma* subsp. *sororia* genome assemblies by extracting 5 kb upstream and downstream of the convergent SNPs under selection and detecting the orthologous region in both assemblies through a BLAST search. The orthologous regions were aligned using MUSCLE (116) and manually inspected using AliView (117). The synteny map between chromosomes 3 and 7 was generated using SynMap2 (99).

Data Availability

The chromosome-level genome assemblies of *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia* are available in the Cucurbit Genomics Database (118) and in the NCBI RefSeq database (accessions XXXXXXXXX). The raw sequencing reads of the *C. argyrosperma* subsp. *sororia* genome are available in the NCBI Sequence Read Archive (accessions SRPXXXXXX- SRPXXXXXX). The raw sequencing reads of each individual sequenced by tGBS are available in the NCBI Sequence Read Archive (accessions SRPXXXXXX- SRPXXXXXX; table S2). The 13k SNP dataset is available in Dataset S1.

Acknowledgments

This work was funded by CONABIO KE004 "Diversidad genetica de las especies de *Cucurbita* en Mexico e hibridacion entre plantas geneticamente modificadas y especies silvestres de *Cucurbita*" and CONABIO PE001 "Diversidad genetica de las especies de Cucurbita en Mexico. Fase II. Genomica evolutiva y de poblaciones, recursos geneticos y domesticacion" (both awarded to R.L-S., L.E.E. and S.M-H.). J.B-R. is a doctoral student from Programa de Doctorado en Ciencias Biomedicas, Universidad Nacional Autonoma de Mexico, and received fellowship 583146 from CONACyT. We acknowledge the Doctorado en Ciencias Biomedicas for the support provided during the development of this project. We thank Jodi Lynn Humann for her technical support while using the GenSAS web server. Special thanks to Laura Espinosa-Asuar for her technical support. We thank Rodrigo Garcia Herrera, head of the Scientific Computing Department at LANCIS, Instituto de Ecologia UNAM, for running the HTC infrastructure we used for the genome assembly and chromosome anchoring. The population genomic analyses were carried out using CONABIO's computing cluster, which was partially funded by SEMARNAT through the grant "Contribucion de la Biodiversidad para el Cambio Climatico" to CONABIO.

References

1. R. S. Meyer, M. D. Purugganan, Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet* **14**, 840–52 (2013).
2. M. A. Zeder, Core questions in domestication research. *Proc Natl Acad Sci U S A* **112**, 3191–8 (2015).
3. L. Kistler, *et al.*, Gourds and squashes (*Cucurbita* spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proc Natl Acad Sci U S A* **112**, 15107–12 (2015).
4. R. Lira, *et al.*, "*Homo sapiens–Cucurbita* interaction in Mesoamerica: Domestication Dissemination, and Diversification" in *Ethnobotany of Mexico*, (Springer New York, 2016), pp. 389–401.
5. G. Chomicki, H. Schaefer, S. S. Renner, Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. *New Phytol* (2019).

6. G. Castellanos-Morales, *et al.*, Historical biogeography and phylogeny of *Cucurbita*: Insights from ancestral area reconstruction and niche evolution. *Mol Phylogenet Evol* **128**, 38–54 (2018).

7. H. S. Paris, "Genetic Resources of Pumpkins and Squash *Cucurbita* spp." in *Genetics and Genomics of Cucurbitaceae*, (Springer International Publishing, 2016), pp. 111–154.

8. O. I. Sanjur, D. R. Piperno, T. C. Andres, L. Wessel-Beaver, Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: Implications for crop plant evolution and areas of origin. *Proc Natl Acad Sci U S A* **99**, 535–40 (2002).

9. Z. Kerem, S. Lev-Yadun, A. Gopher, P. Weinberg, S. Abbo, Chickpea domestication in the NeolithicLevant through the nutritional perspective. *Journal of Archaeological Science* **34**, 1289–1293 (2007).

10. D. Zizumbo-Villarreal, A. Flores-Silva, P. C.-G. Marín, The Archaic Diet in Mesoamerica: Incentive for Milpa Development and Species Domestication. *Economic Botany* **66**, 328–343 (2012).

11. T. W. Whitaker, H. C. Cutler, Cucurbits and cultures in the Americas. *Economic Botany* **19**, 344–349 (1965).

12. M. Nee, The domestication of *Cucurbita* (Cucurbitaceae). *Economic Botany* **44**, 56–68 (1990).

13. R. L. Jarret, I. J. Levy, T. L. Potter, S. C. Cermak, L. C. Merrick, Seed oil content and fatty acid composition in a genebank collection of *Cucurbita moschata* Duchesne and *C. argyrosperma* C. Huber. *Plant Genetic Resources* **11**, 149–157 (2013).

14. R. Lira, J. Caballero, Ethnobotany of the wild Mexican Cucurbitaceae. *Economic Botany* **56** (2002).

15. G. Sánchez-de la Vega, *et al.*, Genetic Resources in the Calabaza Pipiana Squash (*Cucurbita argyrosperma*) in Mexico: Genetic Diversity, Genetic Differentiation and Distribution Models. *Front Plant Sci* **9**, 400 (2018).

16. D. R. Piperno, A. J. Ranere, I. Holst, J. Iriarte, R. Dickau, Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proc Natl Acad Sci U S A* **106**, 5019–24 (2009).

17. B. D. Smith, The Initial Domestication of *Cucurbita pepo* in the Americas 10,000 Years Ago. *Science* **276**, 932–934 (1997).

18. D. Zizumbo-Villarreal, P. Colunga-GarcíaMarín, Origin of agriculture and plant domestication in West Mesoamerica. *Genetic Resources and Crop Evolution* **57**, 813–825 (2010).

19. J. Barrera-Redondo, *et al.*, The Genome of *Cucurbita argyrosperma* (Silver-Seed Gourd) Reveals Faster Rates of Protein-Coding Gene and Long Noncoding RNA Turnover and Neofunctionalization within *Cucurbita*. *Mol Plant* **12**, 506–520 (2019).

20. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–2 (2015).

21. J. Montero-Pau, *et al.*, *De novo* assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the *Cucurbita* genus. *Plant Biotechnol J* **16**, 1161–1171 (2018).

22. H. Sun, *et al.*, Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid *Cucurbita* Genomes. *Mol Plant* **10**, 1293–1306 (2017).

23. K. J. Hoff, S. Lange, A. Lomsadze, M. Borodovsky, M. Stanke, BRAKER1: Unsupervised RNA-Seq Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–9 (2016).

24. K. J. Hoff, A. Lomsadze, M. Borodovsky, M. Stanke, Whole-Genome Annotation with BRAKER. *Methods Mol Biol* **1962**, 65–95 (2019).

25. M. Alonge, *et al.*, RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**, 224 (2019).

26. T. W. Whitaker, W. P. Bemis, Origin and evolution of the cultivated *Cucurbita*. *Bulletin of the Torrey Botanical Club* **102** (1975).

27. A. Ott, *et al.*, tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. *Nucleic Acids Res* **45**, e178 (2017).

28. T. H. Lee, H. Guo, X. Wang, C. Kim, A. H. Paterson, SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 162 (2014).

29. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–64 (2009).

30. L. Excoffier, M. Foll, fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–4 (2011).

31. L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, M. Foll, Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**, e1003905 (2013).

32. R. Cruz-Reyes, G. Avila-Sakar, G. Sánchez-Montoya, M. Quesada, Experimental assessment of gene flow between transgenic squash and a wild relative in the center of origin of cucurbits. *Ecosphere* **6**, art248 (2015).

33. S. Montes-Hernandez, L. E. Eguiarte, Genetic structure and indirect estimates of gene flow in three taxa of *Cucurbita* (Cucurbitaceae) in western Mexico. *Am J Bot* **89**, 1156–63 (2002).

34. P. de Villemereuil, O. E. Gaggiotti, A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution* **6**, 1248–1258 (2015).

35. K. Luu, E. Bazin, M. G. Blum, pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol Ecol Resour* **17**, 67–77 (2017).

36. Y. Shang, *et al.*, Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084–8 (2014).

37. M. S. Campbell, C. Holt, B. Moore, M. Yandell, Genome Annotation and Curation Using MAKER and MAKER-P. *Curr Protoc Bioinformatics* **48**, 4.11.1–39 (2014).

38. T. Derrien, *et al.*, The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775–89 (2012).

39. A. D. L. Nelson, *et al.*, Evolinc: A Tool for the Identification and Evolutionary Comparison of Long Intergenic Non-coding RNAs. *Front Genet* **8**, 52 (2017).

40. H. D. Wilson, Gene Flow in Squash Species. *BioScience* **40**, 449–455 (1990).

41. R. G. Allaby, R. L. Ware, L. Kistler, A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evol Appl* **12**, 29–37 (2019).

42. R. Lira-Saade, *Estudios taxonomicos ecogeograficos de las Cucurbitaceae Latinoamericanas de importancia economica* (International Plant Genetic Resources Institute, Rome, Italy, 1995).

43. D. Jarvis, T. Hodgkin, "Farmer decision making and genetic diversity" in *Genes in the Field*, (CRC Press, 1999) https:/doi.org/10.1201/9781420049824.ch11.

44. D. I. Jarvis, *et al.*, A global perspective of the richness and evenness of traditional crop-variety diversity maintained by farming communities. *Proc Natl Acad Sci U S A* **105**, 5326–31 (2008).

45. S. Montes-Hernández, L. C. Merrick, L. E. Eguiarte, Maintenance of Squash (*Cucurbita* spp.) Landrace Diversity by Farmers' Activities in Mexico. *Genetic Resources and Crop Evolution* **52**, 697–707 (2005).

46. J. Barrera-Redondo, *et al.*, Variedades locales y criterios de selección de especies domesticadas del género *Cucurbita* (Cucurbitaceae) en los Andes Centrales del Perú: Tomayquichua, Huánuco. *Botanical Sciences* **98**, 101–116 (2020).

47. G. Castellanos-Morales, *et al.*, Tracing back the origin of pumpkins (*Cucurbita pepo* ssp. *pepo* L.) in Mexico. *Proc Biol Sci* **286**, 20191440 (2019).

48. A. J. Ranere, D. R. Piperno, I. Holst, R. Dickau, J. Iriarte, The cultural and chronological context of early Holocene maize and squash domestication in the Central Balsas River Valley, Mexico. *Proc Natl Acad Sci U S A* **106**, 5014–8 (2009).

49. W. Stinnesbeck, *et al.*, The earliest settlers of Mesoamerica date back to the late Pleistocene. *PLoS One* **12**, e0183345 (2017).

50. T. D. Dillehay, *et al.*, New Archaeological Evidence for an Early Human Presence at Monte Verde, Chile. *PLoS One* **10**, e0141923 (2015).

51. D. R. Piperno, I. Holst, L. Wessel-Beaver, T. C. Andres, Evidence for the control of phytolith formation in *Cucurbita* fruits by the hard rind (Hr) genetic locus: Archaeological and ecological implications. *Proc Natl Acad Sci U S A* **99**, 10923–8 (2002).

52. D. R. Piperno, The Origins of Plant Cultivation and Domestication in the New World Tropics. *Current Anthropology* **52**, S453–S470 (2011).

53. D. B. Lowry, *et al.*, Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Mol Ecol Resour* **17**, 366–369 (2017).

54. D. B. Lowry, *et al.*, Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol Ecol Resour* **17**, 142–152 (2017).

55. G. Swinnen, A. Goossens, L. Pauwels, Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement. *Trends Plant Sci* **21**, 506–515 (2016).

56. D. H. Kim, *et al.*, Small heat shock protein Hsp17.8 functions as an AKR2A cofactor in the targeting of chloroplast outer membrane proteins in *Arabidopsis*. *Plant Physiol* **157**, 132–46 (2011).

57. K. H. Stanley, *et al.*, Transcriptional divergence of the duplicated oxidative stress-responsive genes in the *Arabidopsis* genome. *Plant J* **41**, 212–20 (2005).

58. P. Gaudet, M. S. Livstone, S. E. Lewis, P. D. Thomas, Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform* **12**, 449–62 (2011).

59. J. Zhang, *et al.*, Receptor-like cytoplasmic kinases integrate signaling from multiple plant immune receptors and are targeted by a *Pseudomonas syringae* effector. *Cell Host Microbe* **7**, 290–301 (2010).

60. X. Moreira, L. Abdala-Roberts, R. Gols, M. Francisco, Plant domestication decreases both constitutive and induced chemical defenses by direct selection against defensive traits. *Sci Rep* **8**, 12678 (2018).

61. K. Chen, *et al.*, Abscisic acid dynamics, signaling, and functions in plants. *J Integr Plant Biol* **62**, 25–54 (2020).

62. J. S. (P. Heslop-Harrison, T. Schwarzacher, "Genetics and genomics of crop domestication" in *Plant Biotechnology and Agriculture*, (Elsevier, 2012), pp. 3–18.

63. A. B. Martínez, *et al.*, Differences in seed dormancy associated with the domestication of *Cucurbita maxima*: elucidation of some mechanisms behind this response. *Seed Science Research* **28**, 1–7 (2017).

64. W. Jiang, D. Yu, Arabidopsis WRKY2 transcription factor mediates seed germination and post germination arrest of development by abscisic acid. *BMC Plant Biol* **9**, 96 (2009).

65. T. J. Strabala, *et al.*, Gain-of-function phenotypes of many CLAVATA3/ESR genes, including four new family members, correlate with tandem variations in the conserved CLAVATA3/ESR domain. *Plant Physiol* **140**, 1331–44 (2006).

66. G. R. Rodríguez, *et al.*, Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity. *Plant Physiol* **156**, 275–85 (2011).

67. C. Xu, *et al.*, A cascade of arabinosyltransferases controls shoot meristem size in tomato. *Nat Genet* **47**, 784–92 (2015).

68. J. L. Carpenter, S. E. Ploense, D. P. Snustad, C. D. Silflow, Preferential expression of an alpha-tubulin gene of *Arabidopsis* in pollen. *Plant Cell* **4**, 557–71 (1992).

69. Y. Mei, H. B. Gao, M. Yuan, H. W. Xue, The Arabidopsis ARCP protein, CSI1, which is required for microtubule stability, is necessary for root and anther development. *Plant Cell* **24**, 1066–80 (2012).

70. B. Burla, *et al.*, Vacuolar transport of abscisic acid glucosyl ester is mediated by ATP-binding cassette and proton-antiport mechanisms in *Arabidopsis*. *Plant Physiol* **163**, 1446–58 (2013).

71. C. Seiler, *et al.*, ABA biosynthesis and degradation contributing to ABA homeostasis during barley seed development under control and terminal drought-stress conditions. *J Exp Bot* **62**, 2615–32 (2011).

72. Y. Wang, *et al.*, Construction of a High-Density Genetic Map and Analysis of Seed-Related Traits Using Specific Length Amplified Fragment Sequencing for *Cucurbita maxima*. *Front Plant Sci* **10**, 1782 (2019).

73. D. Guo, *et al.*, Resequencing 200 Flax Cultivated Accessions Identifies Candidate Genes Related to Seed Size and Weight and Reveals Signatures of Artificial Selection. *Front Plant Sci* **10**, 1682 (2019).

74. J. Huang, R. Ghosh, V. A. Bankaitis, Sec14-like phosphatidylinositol transfer proteins and the biological landscape of phosphoinositide signaling in plants. *Biochim Biophys Acta* **1861**, 1352–1364 (2016).

75. T. J. Raharjo, L. Nurliana, S. Mastjeh, Phospholipids from pumpkin (*Cucurbita moschata* (Duch.) Poir) seed kernel oil and their fatty acid composition. *Indonesian Journal of Chemistry* **11**, 48–52 (2011).

76. S. Gonzalez-Jorge, *et al.*, ZEAXANTHIN EPOXIDASE Activity Potentiates Carotenoid Degradation in Maturing Seed. *Plant Physiol* **171**, 1837–51 (2016).

77. A. Frey, J. P. Boutin, B. Sotta, R. Mercier, A. Marion-Poll, Regulation of carotenoid and ABA accumulation during the development and germination of *Nicotiana plumbaginifolia* seeds. *Planta* **224**, 622–32 (2006).

78. M. Murkovic, U. Mülleder, H. Neunteufl, Carotenoid Content in Different Varieties of Pumpkins. *Journal of Food Composition and Analysis* **15**, 633–638 (2002).

79. G. Li, *et al.*, Orphan genes are involved in drought adaptations and ecoclimatic-oriented selections in domesticated cowpea. *J Exp Bot* **70**, 3101–3110 (2019).

80. M. Rendón-Anaya, *et al.*, Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biol* **18**, 60 (2017).

81. A. J. Alverson, *et al.*, Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol Biol Evol* **27**, 1436–48 (2010).

82. X. Aguirre-Dugua, *et al.*, Evolutionary Dynamics of Transferred Sequences Between Organellar Genomes in *Cucurbita*. *J Mol Evol* **87**, 327–342 (2019).

83. M. B. Hufford, *et al.*, Comparative population genomics of maize domestication and improvement. *Nat Genet* **44**, 808–11 (2012).

84. A. Zsögön, *et al.*, *De novo* domestication of wild tomato using genome editing. *Nat Biotechnol* (2018).

85. A. R. Fernie, J. Yan, *De Novo* Domestication: An Alternative Route toward New Crops for the Future. *Mol Plant* **12**, 615–631 (2019).

86. M. Xie, *et al.*, A reference-grade wild soybean genome. *Nat Commun* **10**, 1216 (2019).

87. N. Dierckxsens, P. Mardulyn, G. Smits, NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res* **45**, e18 (2017).

88. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).

89. R. Kajitani, *et al.*, Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* **24**, 1384–95 (2014).

90. C. Ye, C. M. Hill, S. Wu, J. Ruan, Z. S. Ma, DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Sci Rep* **6**, 31900 (2016).

91. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

92. B. J. Walker, *et al.*, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

93. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–95 (2010).

94. J. L. Humann, T. Lee, S. Ficklin, D. Main, Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. *Methods Mol Biol* **1962**, 29–51 (2019).

95. A. Dobin, *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

96. M. Stanke, O. Sch¨offmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).

97. A. Lomsadze, P. D. Burns, M. Borodovsky, Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119 (2014).

98. S. Koren, *et al.*, Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).

99. A. Haug-Baltzell, S. A. Stephens, S. Davey, C. E. Scheidegger, E. Lyons, SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* **33**, 2197–2198 (2017).

100. S. Kurtz, *et al.*, Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

101. D. Pratas, R. M. Silva, A. J. Pinho, P. J. Ferreira, An alignment-free method to find and visualize rearrangements between pairs of DNA sequences. *Sci Rep* **5**, 10203 (2015).

102. S. Li, H. H. Chou, LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics* **20**, 2865–6 (2004).

103. S. Hoffmann, *et al.*, Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol* **5**, e1000502 (2009).

104. S. Caboche, C. Audebert, Y. Lemoine, D. Hot, Comparison of mapping algorithms used in high throughput sequencing: application to Ion Torrent data. *BMC Genomics* **15**, 264 (2014).

105. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

106. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–93 (2011).

107. H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851–8 (2008).

108. S. Purcell, *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–75 (2007).

109. J. E. Wigginton, D. J. Cutler, G. R. Abecasis, A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 887–93 (2005).

110. P. Danecek, *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).

111. K. Keenan, P. McGinnity, T. F. Cross, W. W. Crozier, P. A. Prodöhl, diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution* **4**, 782–788 (2013).

112. J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, W. A. Cresko, Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**, 3124–40 (2013).

113. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695 (2009).

114. P. Cingolani, *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

115. A. Alexa, J. Rahnenführer, T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–7 (2006).

116. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–7 (2004).

117. A. Larsson, AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–8 (2014).

118. Y. Zheng, *et al.*, Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Res* **47**, D1128–D1136 (2019).

DISCUSIÓN

El estudio genómico de las calabazas nos permitió estudiar con detalle eventos evolutivos que ocurrieron en distintas escalas temporales, una a una escala macroevolutiva, como los efectos de una duplicación completa del genoma que ocurrió en el género *Cucurbita* hace ~30 millones de años (Barrera-Redondo *et al*., 2019) y otra a una escala microevolutiva, el proceso de domesticación de *C. argyrosperma* que comenzó hace tan solo ~13,800 años (Barrera-Redondo *et al*., 2020a). Los dos eventos estudiados en esta tesis están relacionados en términos de contingencia histórica (Blount *et al.*, 2018). La duplicación completa del genoma derivó en el cambio y origen de elementos nuevos que pudieron ser o no seleccionados durante el proceso de domesticación. Asimismo, muchos de los resultados de la domesticación fueron deterministas en vez de contingentes (Blount *et al.*, 2018), ya que las calabazas domesticadas muestran muchos de los síndromes de domesticación que se han observado en otras cucurbitáceas, tales como la pérdida de la cucurbitacina, el crecimiento del tamaño de la planta, el crecimiento del fruto, el crecimiento de la semilla y la pérdida de genes relacionados al estrés biótico y abiótico (Chomicki *et al*., 2019).

Los resultados obtenidos en esta tesis nos abordan diversas cuestiones de relevancia: ya sea conocimiento básico de los procesos de evolución molecular, o el entendimiento del papel que juegan estos procesos evolutivos en la aparición de uno de los organismos domesticados que forman la base de nuestra alimentación. Tomando como ejemplo los procesos de duplicación completa del genoma, se puede discutir la posible existencia de una distinción relevante entre los procesos considerados como microevolutivos y macroevolutivos. La observación de nuevos genes codificantes y ARNs largos no codificantes en *Cucurbita* nos permite abordar otra cuestión crucial en evolución, que es el origen de las funciones a escala molecular, es decir: a qué nos referimos con una función en biología, de qué manera se pueden inferir funciones en los distintos elementos que componen a un genoma y cómo pueden surgir funciones nuevas en los genomas. Los análisis poblacionales realizados en *C. argyrosperma* nos revelan cómo la genética de poblaciones, a pesar de sus limitantes, puede integrar la información genómica con el contexto ecológico e histórico de esta planta para elucidar un escenario concreto de su domesticación en Mesoamérica. Finalmente, el entendimiento de estos procesos de domesticación nos guía a plantear posibles usos de estos conocimientos para el beneficio de la sociedad, ya sea por medio de la conservación de los recursos fitogenéticos o mediante mejoramiento genético de los cultivos.

La duplicación completa del genoma como evento macroevolutivo: ¿Una cuestión de escala de tiempo o de propiedades emergentes?

Se ha argumentado que la distinción entre la macro y microevolución se suele abarcar en tres contextos distintos: 1) cuando aparecen planes corporales nuevos que puedan catalogarse como el origen de un jerarquía taxonómica supra-específica; 2) cuando se habla de procesos de especiación y extinción, donde la unidad de evolución son las especies; 3) y cuando se habla de la acumulación de procesos evolutivos en largas escalas de tiempo (Hautmann, 2020).

El contexto del origen de nuevos planes corporales dio pie a la distinción original entre macro y microevolución (Philiptschenko, 1927), y ha sido un recurso teórico frecuentemente utilizado por los defensores del saltacionismo (Goldschmidt, 1933), quienes más recientemente recurren a la biología evolutiva del desarrollo (Evo-Devo) para tratar de encontrar cambios en los patrones de expresión genética durante el desarrollo que puedan generar cambios importantes en los planes corporales (Theißen, 2009). En la actualidad existen algunos ejemplos que se ajustan al modelo saltacionista, aunque parecen tratarse de eventos excepcionales, al contrario de los cambios graduales que parecen ser la norma en la evolución biológica (Chouard, 2010).

A pesar de la plausibilidad de los eventos saltacionistas (Chouard, 2010), este tipo de eventos aparentemente no son cualitativamente distintos de los procesos microevolutivos, ya que la fijación de este tipo de caracteres “saltacionistas” u otros fenómenos observados por la Evo-Devo pueden analizarse desde una perspectiva poblacional (Nunes *et al*., 2013). Bajo esta lógica, solo los procesos macroevolutivos donde ocurre selección a nivel de especie (Jablonski, 2008) podrían considerarse como fenómenos cualitativamente distintos que no pueden extrapolarse a la acumulación de procesos microevolutivos (Hautmann, 2020). Sin embargo, dicha definición tiene sus propias complicaciones, como lo es la delimitación, y por lo tanto definición, de una especie (Mallet, 1995).

La duplicación completa de un genoma, como la que distingue a las calabazas de otros miembros de la familia Cucurbitaceae resulta ser un evento puntual de alopoliploidía (Sun *et al*., 2017; Montero-Pau *et al*., 2018). Éste podría interpretarse con un evento semi-saltacionista a nivel genómico, ya que los procesos de fraccionamiento y diploidización en plantas suelen involucrar la escisión repentina de grandes segmentos del genoma, más que una pérdida gradual gen por gen (Wang *et al*., 2011). Sin embargo, nuestros resultados con el genoma de *C. argyrosperma* indican la prevalencia de un gran número de pseudogenes, los cuales suelen encontrarse con mayor frecuencia en uno de los dos pares de cromosomas homeólogos — como encontramos en el análisis de sintenia entre los cromosomas 3 y 7 — pero separados por millones de pares de bases que contienen genes codificantes conservados (Barrera-Redondo *et al*., 2020a). Lo anterior apoya la idea de un proceso más gradual de pérdida gen por gen en los genomas de *Cucurbita*, ya que su diploidización fue mayoritariamente el resultado de eventos independientes de pseudogenización, en lugar de eventos de escisión de los genes duplicados (Barrera-Redondo *et al*., 2019; Barrera-Redondo *et al*., 2020a). Adicionalmente, el análisis de expansión y contracción de familias génicas dentro de *Cucurbita* muestra que sigue habiendo un ritmo alto de recambio de genes dentro del género. Esto indica que el proceso de fraccionamiento continúa ocurriendo de manera paralela y gradual más de 30 millones de años después del evento de duplicación de genoma en las distintas especies de *Cucurbita*, en conjunto con la tasa basal de recambio de genes de la familia Cucurbitaceae (Barrera-Redondo *et al*., 2019).

Por lo tanto, la evolución de familias génicas y de la estructura del genoma, más que un cambio de naturaleza saltacionista puede considerarse como una extensión a largo plazo de la microevolución. Los elementos de presencia/ausencia, las variaciones en número de copia y los cambios estructurales en los cromosomas se han observado y analizado en otros genomas a nivel poblacional desde hace más de una década (*e.g*., Wong *et al*., 2007; Tan *et al*., 2012; Hu *et al*., 2018), siendo estos catalogados como variación intraespecífica (Hautmann, 2020). Inclusive, los procesos de duplicación completa del genoma y fraccionamiento se han visto asociados a procesos de estructuración poblacional y eventual especiación entre poblaciones (Williams *et al*., 2016), convirtiéndolos nuevamente en una extensión de procesos microevolutivos, pero con repercusiones a gran escala en la evolución de los linajes (Williams *et al*., 2016; Hautmann, 2020).

Cuando se habla de eventos de tipo saltacionista se suele hacer referencia a cambios en el fenotipo, más que a cambios dramáticos en la arquitectura del genoma (Goldschmidt, 1933; Theißen, 2009). El estudio de estos eventos se suele enfocar a cambios en las redes de regulación transcripcional (Theißen, 2009; Chouard, 2010). Sin embargo, los posibles efectos fenotípicos de las duplicaciones completas del genoma han sido menos estudiados, salvo por algunos ejemplos en Brassicales y Eudicotiledonias (Edger *et al*., 2015; Chanderbali *et al*., 2017; Clark y Donoghue, 2018). Recientemente, se realizó un estudio filogenómico en Cucurbitaceae donde se encontraron cuatro eventos independientes de duplicaciones genómicas en la familia (Guo *et al*., 2020). Estas incluyen una duplicación previamente descrita asociada al origen de la familia Cucurbitaceae (Wang *et al.*, 2018) y a duplicación encontrada en el género *Cucurbita* (Sun *et al*., 2017). La duplicación completa del genoma asociada al origen de Cucurbitaceae fue determinante en la aparición de zarcillos, y posiblemente también del fruto pepónide, en dicha familia. Además, se encontró una asociación cercana entre dicha duplicación genómica y dos eventos de diversificación en Cucurbitaceae (Guo *et al*., 2020). Esto sugiere que las duplicaciones completas del genoma son una importante fuente de innovación evolutiva en las plantas, ya que la redundancia funcional resultante de estas duplicaciones las hace más evolucionables, promoviendo la aparición de novedades fenotípicas y la aparición de nuevos linajes (Barrera-Redondo *et al*., 2020b).

La repercusión de la duplicación completa del genoma sobre la aparición de diferencias morfológicas concretas que distingan a *Cucurbita* como un género no ha sido abordado de manera puntual. Los análisis de enriquecimiento funcional en genes duplicados de *Cucurbita* sugieren que éstos están asociados al desarrollo de estrategias adaptativas de supervivencia y al crecimiento y maduración del fruto (Sun *et al*., 2017). Sin embargo, los análisis filogenómicos de Guo *et al*. (2020) sugieren que esta duplicación genómica es compartida por todas las especies de la tribu Cucurbiteae, que incluye a *Cucurbita* y a otros 12 géneros. Por lo tanto, es necesario realizar más estudios a lo largo de los distintos géneros de la tribu Cucurbiteae para entender mejor qué caracteres morfológicos o fisiológicos coinciden cronológicamente con el evento de la duplicación (Barrera-Redondo *et al*., 2020b).

Estudios enfocados en duplicaciones del genoma en diversos linajes de plantas han propuesto que los genes retenidos durante una duplicación completa del genoma suelen estar enriquecidos en factores transcripcionales y proteínas que forman complejos multi-proteicos que son sensibles a los cambios de dosis. Durante los eventos de poliploidización, ocurre un balanceo de dosis que permite la retención de este tipo de genes, a diferencia de las duplicaciones en tandem (Clark y Donoghue, 2018; Rendón-Anaya *et al*., 2019). Esto sugiere que las duplicaciones en tándem suelen favorecer la expansión de genes involucrados en adaptaciones locales recientes, mientras que las duplicaciones completas del genoma favorecen la expansión de genes involucrados en procesos basales de fisiología y desarrollo, implicando un posible efecto sobre los planes corporales (Clark y Donoghue, 2018; Rendón-Anaya *et al*., 2019).

En el caso de las calabazas, los estudios de expresión diferencial muestran cambios sustanciales en la expresión genética entre genes homeólogos de los dos subgenomas del género *Cucurbita* (*i.e.*, las dos copias diferenciadas de los cromosomas generadas durante el evento de alopoliploidización). Esto indica que las copias de los genes duplicados han cambiado en cuanto a patrones relativos de expresión debido a rearreglos en las redes de regulación transcripcional (Sun *et al.*, 2017). Esto abre la posibilidad de que la aparición de elementos reguladores novedosos, como los RNAs largos no codificantes, pudieron jugar un papel importante en los cambios de las redes de regulación transcripcional. Un estudio de expresión diferencial durante las etapas del desarrollo de los órganos podría revelar el papel de los genes duplicados en Cucurbiteae con rasgos morfológicos particulares de la tribu.

Un rasgo particular en calabazas es su alta diversidad morfológica en frutos, considerada una de las más altas en el reino Plantae (Bisognin, 2002). Esta alta variación morfológica podría estar directa o indirectamente relacionada a la duplicación del genoma (Mattenberger *et al*., 2017). Consistente con esta hipótesis, cuatro de los genes duplicados en *Cucurbita* pertenecen a la familia OVATE, implicados en la regulación de la forma del fruto (Sun *et al*., 2017). Dicha variación morfológica puede estar dada por una alta evolucionabilidad de nuevos fenotipos en Cucurbiteae, capacidad dada por la duplicación de estos genes (Sun *et al*., 2017; Barrera-Redondo *et al*., 2020b). También es posible que las calabazas presenten una mayor plasticidad fenotípica que pueda haber surgido a partir de la duplicación completa del genoma, permitiendo una amplia variedad de formas en los frutos (Mattenberger *et al*., 2017). Esta hipótesis se puede poner a prueba, realizando un análisis con métodos filogenéticos comparativos para evaluar la correlación de la duplicación completa del genoma contra un incremento en la diversidad morfológica en los distintos géneros que componen a Cucurbiteae, comparando con otros taxa cercanos sin eventos de duplicación (*e.g.*, Clark y Donoghue, 2018).

El proceso de fraccionamiento y retención de genes duplicados en calabazas no parece ser azaroso, dado que se encuentran enriquecidos en ciertas funciones, a pesar de haber ocurrido un proceso de fraccionamiento equilibrado (Sun *et al*., 2017; Clark y Donoghue, 2018). La retención de genes duplicados puede ser entendida por el efecto de la selección positiva sobre las copias que lograron neofuncionalizarse, mediante la acumulación de mutaciones con cambio de función, que a su vez fue posible por la relajación en las presiones de selección purificadora dada una redundancia funcional (Clark y Donoghue, 2018). Otra forma de ver este proceso es mediante la aparición de nuevos elementos que interactúan en las redes de regulación transcripcional, que mediante la pérdida de ciertos elementos desembocan en la formación de nuevas configuraciones estables llamadas atractores, las cuales dan pie a los diferentes linajes celulares de las plantas (Crombach y Hogeweg, 2008). La evolución de estas redes, y por lo tanto la evolución de los fenotipos que generan, se ve ligada a la tasa de recambio de los elementos que componen dicha red (Crombach y Hogeweg, 2008). En este sentido, la alta tasa de nacimiento y muerte de genes que observamos en el género *Cucurbita* pudo haber cambiado las conexiones en las redes de regulación transcripcional (Barrera-Redondo *et al*., 2019).

Por otro lado, la tasa de recambio de lincRNAs se suele acelerar cuando ocurren "perturbaciones" a grande escala en los genomas de las plantas, tales como rearreglos cromosómicos, cambios de cariotipo y duplicaciones genómicas (Nelson y Shippen, 2015). Este aumento en la tasa de recambio de lincRNAs podría explicarse mediante un cambio en el paisaje epigenético de un organismo después de una perturbación genómica, lo cual conlleva a la pérdida y ganancia de transcritos reguladores, hasta llegar a una nueva configuración estable de regulación transcripcional (Crombach y Hogeweg, 2008). La aceleración en la tasa de recambio de lincRNAs en *Cucurbita* se ajustaría a esta hipótesis (Barrera-Redondo *et al*., 2019).

La limitante de los análisis de redes radica en que no analiza explícitamente la aparición y cambio de los elementos en las redes, sino que solo modela los cambios en las interacciones de los elementos y su potencial fenotipo (Crombach y Hogeweg, 2008). Para ello, pareciera ineludible invocar a procesos de selección para entender los procesos de neofuncionalización de los genes y demás transcritos redundantes que se generaron durante la duplicación completa del genoma en las calabazas (Clark y Donoghue, 2018). De tal manera que estudiar las causas últimas del origen de la variación solo puede realizarse mediante análisis de evolución molecular, más que estudios de la dinámica de las redes, los cuales se enfocan en la aparición de los fenotipos dada una red de elementos preexistentes o en los cambios de la red dada una tasa de recambio de elementos (Crombach y Hogeweg, 2008).

Innovación evolutiva a escala molecular

El estudio de genómica comparada que se realizó en el género *Cucurbita* permitió elucidar el origen evolutivo de nuevos genes y lincRNAs en el género, de los cuales se podría discutir su funcionalidad en el genoma (Barrera-Redondo *et al*., 2019). El origen de nuevos elementos funcionales en los genomas de los organismos es una pregunta fundamental en la biología evolutiva (Lynch y Walsh, 2007; Van Oss y Carvunis, 2019). Sin embargo, definir qué es un elemento funcional en un genoma es complicado en sí mismo (Kellis *et al.*, 2014).

Existen diversas líneas de evidencia que no necesariamente se sobrelapan para determinar que un elemento en el genoma es funcional, tales como la observación de consecuencias fenotípicas cuando son perturbados, su conservación a nivel evolutivo o la presencia de su actividad bioquímica en las células (Kellis *et al.*, 2014). Cada una de estas evidencias tiene sus limitaciones, en particular de aquellos elementos con tan sólo actividad bioquímica, de los cuales se puede sospechar, pero no corroborar función (Graur *et al.*, 2015). Además, los elementos en los genomas no actúan de manera aislada, sino que sólo operan de acuerdo con una red de interacciones, por lo que su relación con un fenotipo es en ocasiones insuficiente para definir que son funcionales (Crombach y Hogeweg, 2008; Pigliucci, 2010). Tal es el caso de algunos microRNAs, los cuales no están directamente involucrados en la emergencia de un fenotipo, sino que ayudan a que el desarrollo embrionario avance de manera adecuada cuando las condiciones abióticas no son ideales (Bartel, 2009). Finalmente, se tiene evidencia de ARNs largos no codificantes que carecen de conservación a nivel evolutivo, pero que han sido validados como funcionales (Kapusta y Feschotte, 2014).

Inclusive el concepto de función en biología es problemático en un contexto evolutivo, considerando que los elementos que forman a un organismo no aparecen con un fin preexistente, sino que son la culminación de fenómenos de causalidad (Amundson, y Lauder, 1994), ya sean procesos de autoorganización (Kauffman, 1993) o dinámicas históricas de mutación y selección (Long *et al*., 2013; Vakirlis *et al*., 2020). Por ello utilizaré del concepto "etiológico" de función, mejor conocido como el concepto de función como efecto seleccionado, donde un elemento se considera funcional de acuerdo con el efecto positivo que tiene en la adecuación del organismo que lo posee, independientemente de su origen causal (*i.e.*, incluye adaptaciones y exaptaciones) (Wright, 1973; Amundson y Lauder, 1994; Graur *et al.*, 2015). El concepto de función como efecto seleccionado nos permite enmarcar a las funciones biológicas en un contexto histórico, a diferencia de otras definiciones utilizadas en biología, como o es el concepto de función de rol causal. Este segundo concepto define a la función como la capacidad de un elemento (*e.g.*, la capacidad de contracción de un músculo) de realizar una acción (*e.g.*, movimiento) dentro de un sistema (*e.g.*, una extremidad) (Cummins, 1975; Amundson y Lauder, 1994), dejando de lado el contexto histórico que dio origen a dicho elemento.

Adicionalmente, el uso del concepto de función como efecto seleccionado nos permite clasificar a los elementos del genoma en cuatro categorías distintas. El "ADN literal" se puede definir como una secuencia funcional cuyo orden de nucleótidos se encuentra bajo selección (*e.g.,* el marco de lectura de un gen codificante). El "ADN indiferente" son secuencias funcionales cuyo orden o identidad de nucleótidos no es importante, pero cuya presencia se encuentra bajo selección (*e.g.*, la tercera posición de un codón de leucina). El ADN selectivamente neutral es aquel cuya presencia o ausencia no afecta la adecuación del organismo (*e.g.*, algunas secuencias intergénicas), pero que tiene el potencial de adquirir una función nueva mediante la acumulación de mutaciones. Finalmente, el ADN deletéreo representa aquellas secuencias cuya existencia perjudica la adecuación del organismo que las posee (*e.g*., una inserción que altere un marco abierto de lectura) (Graur *et al*., 2015).

El problema de definir elementos funcionales en los genomas es particularmente importante en el estudio de los lincRNAs, ya que a pesar de existir múltiples ejemplos de actividad regulatoria en estos transcritos (Chekanova, 2015), su repercusión en el fenotipo es desconocida para la mayor parte de los lincRNAs descritos y tienen una baja tasa de conservación a nivel evolutivo (Ulitsky, 2016; Nelson *et al.*, 2017). Algunos de los lincRNAs que han sido validados experimentalmente pueden ser clasificados como "ADN literal", en el caso de aquellos transcritos que realizan su función reguladora a partir de su secuencia primaria o de su estructura secundaria, la cual se suele encontrar evolutivamente conservada (Chekanova, 2015; Graur *et al.*, 2015). Pero también hay lincRNAs que pueden ser clasificados como "ADN indiferente", que son aquellos transcritos cuya secuencia es irrelevante para su función, pero cuya actividad transcripcional actúa como regulador trascripcional en *cis* de uno o más genes en los alrededores por medio de interferencia transcripcional, ya sea porque comparten o se sobelapan sus secuencias promotoras (Shearwin *et al.*, 2005; Chekanova, 2015). Es por ello por lo que los análisis de conservación evolutiva suelen ser insuficientes para determinar si un lincRNA es funcional o no (Kapusta y Feschotte, 2014).

Nosotros encontramos ejemplos de lincRNAs con señales de selección purificadora y selección positiva en el genoma de *C. argyrosperma* (Barrera-Redondo *et al*., 2019; Barrera-Redondo *et al*., 2020a), por lo cual pueden considerarse funcionales como efecto seleccionado (Wright, 1973; Graur *et al.*, 2015). Sin embargo, sigue siendo un misterio su repercusión en el fenotipo, dado que no presentan homología contra ningún transcrito validado experimentalmente en *A. thaliana* (Barrera-Redondo *et al*., 2019).

A pesar de encontrar señales de selección en algunos de los lincRNAs de *C. argyrosperma*, el comportamiento de muchos de estos se asemeja al de ADN selectivamente neutral, ya que observamos una relación directa entre el nivel de conservación de los lincRNAs en cucurbitáceas y la distancia evolutiva entre los taxa (Barrera-Redondo *et al*., 2019), aunque esto no descarta la posible función de estos transcritos (Kapusta y Feschotte, 2014).

Otros elementos genómicos que presentan estos mismos problemas son los genes de origen evolutivo reciente, conocidos como genes huérfanos o genes *de novo* (Keeling *et al.*, 2019; Van Oss y Carvunis, 2019). El nacimiento de genes *de novo* se refiere al origen de un gen nuevo y funcional a partir de ADN intergénico y selectivamente neutral (Van Oss y Carvunis, 2019).

A pesar de representar una de las interrogantes más básicas en la biología evolutiva (Van Oss y Carvunis, 2019), el análisis funcional de los genes con aparición *de novo* ha sido complicado no solo por su origen reciente, sino también por la omisión de utilizar un concepto concreto de función al analizar estos elementos (Keeling *et al.*, 2019). Nosotros encontramos señales de selección positiva en al menos dos genes con un origen potencialmente *de novo* exclusivos de calabazas y de *C. argyrosperma* (Barrera-Redondo *et al*., 2020a), lo cual nuevamente se ajusta al concepto de función como efecto seleccionado (Wright, 1973; Graur *et al.*, 2015). Sin embargo, la tasa de nacimiento *de novo* puede llegar a sobreestimarse con métodos tradicionales para buscar homología, por lo que los análisis de sintenia con otros genomas pueden ayudar a verificar la aparición *de novo* de un gen (Casola, 2018).

Se requieren de validaciones experimentales que vayan más allá de los organismos modelo para entender la función de aquellos elementos con conservación evolutiva baja como los genes *de novo* y los lincRNAs, lo que nos ayudará a comprender la importancia de estas novedades evolutivas y sus repercusiones en el fenotipo (Casola, 2018). Esto se puede realizar por medio de ensayos "*knockout*" de estos elementos, utilizando tecnologías de edición genética que se encuentran disponibles en diversos taxa de plantas (Zhang *et al*., 2017). Otra forma eficiente de inferir la función de estos elementos sería mediante experimentos de expresión diferencial, para asociar su actividad transcripcional a cierto órgano, etapa del desarrollo o respuesta a estrés (Nelson *et al*., 2017).

Los eventos de duplicación completa del genoma pueden promover la aparición de genes *de novo* a partir de ADN no codificante, como se ha observado en genomas de hongos microesporidios (Williams *et al.*, 2016). Esto parece haber ocurrido en las calabazas, ya que se encontraron algunos genes codificantes en el genoma de *C. argyrosperma* sin homología a otras proteínas reportadas en GenBank (Barrera-Redondo *et al*., 2020a).

En cuanto a la tasa de nacimiento de lincRNAs, la duplicación completa del genoma resultó en la aparición de una gran cantidad de transcritos exclusivos de calabazas, varios de los cuales surgieron por medio de la duplicación y pseudogenización de genes codificantes (Barrera-Redondo *et al*., 2019). El haber encontrado señales de selección en estos elementos sugiere que los pseudogenes pueden neofuncionalizarse como elementos regulatorios (Barrera-Redondo *et al*., 2019; Barrera-Redondo *et al*., 2020a).

También se ha reportado que algunos genes codificantes surgen *de novo* a partir de ARNs no codificantes que adquieren marcos abiertos de lectura espontáneamente por mutaciones (Ruiz-Orera *et al*., 2014; Casola, 2018; Vakirlis *et al*., 2020). Esto indica que la neofuncionalización de elementos transcripcionalmente activos puede ser una fuente esencial de innovación evolutiva, generando así nuevos elementos funcionales, ya sea de genes codificantes a lincRNAs (Kapusta *et al*., 2013; Barrera-Redondo *et al*., 2019) o viceversa (Ruiz-Orera *et al*., 2014; Vakirlis *et al*., 2020).

Los resultados comparativos con los genomas de calabazas y cucurbitáceas muestran que, en comparación con los genes codificantes, los lincRNAs tienen una alta tasa de recambio, incluyendo una alta tasa de duplicación y extinción (Barrera-Redondo *et al*., 2019). Esta alta tasa de recambio no es exclusiva de calabazas, ya que también se ha observado en otras plantas (Nelson *et al*., 2016) y en menor medida en tetrápodos (Necsulea *et al.*, 2014), por lo que los lincRNAs pueden ser considerados una fuente importante de innovación evolutiva en los organismos (Ruiz-Orera *et al*., 2014). Podemos concluir que la importancia evolutiva de los lincRNAs no está atada únicamente a su función como efecto seleccionado (Graur *et al.*, 2015), sino también a su potencial evolutivo de generar nuevos elementos funcionales en el genoma (Kapusta y Feschotte, 2014).

Además de poseer una alta tasa de recambio, los lincRNAs también tienen una alta tasa de sustitución (Ulitsky, 2016) debido a que la función de muchos de estos transcritos depende únicamente de su estructura secundaria o de su posición relativa a la de otros genes (Kapusta y Feschotte, 2014). De esta manera, los lincRNAs pueden perder rápidamente las señales de homología con otros transcritos con los que compartan ancestría común. La alta tasa de sustituciones de los lincRNAs puede llevarnos a inferir erróneamente que estos genes tienen un origen *de novo* (Kapusta y Feschotte, 2014) cuando en realidad pudieron surgir por eventos antiguos duplicación (Casola, 2018). Si además consideramos que los genes codificantes sin homología a otras proteínas pueden tener un origen en regiones de transcritos no codificantes (Casola, 2018; Vakirlis *et al*., 2020), entonces debemos replantear la posibilidad de que mucha de la evolución aparentemente "*de novo*" que observamos en los genomas, sean en realidad procesos de neofuncionalización a partir de elementos preexistentes con homología profunda, difícilmente detectable por métodos de similitud (Ruiz-Orera *et al*., 2014; Casola, 2018).

Lo anterior apoya la idea de que las fuerzas evolutivas operan primordialmente modificando los recursos biológicos disponibles en sistemas imperfectos, pero con el potencial de llegar a ser altamente eficientes (Gould, 1994). Para comprender mejor estos procesos, es necesario realizar futuros estudios basados en sintenia y validación experimental que analicen a profundidad la evolución de los ARNs no codificantes con homología detectable a genes codificantes, en conjunto con los genes codificantes potencialmente *de novo* con aparición evolutiva reciente.

El estudio de la domesticación a la luz de la genómica de poblaciones

La genómica de poblaciones es útil para comprender el comportamiento de las innovaciones evolutivas bajo un contexto de interacción con el entorno de los organismos que las poseen (Jorde, 2001; Ross-Ibarra *et al*., 2007). Los procesos de domesticación a la luz de la genómica de poblaciones son muy útiles para entender la dinámica evolutiva tanto de los elementos funcionales como de los elementos neutrales en los genomas (Ross-Ibarra *et al*., 2007), ya que estos procesos demográficos y selectivos han ocurrido recientemente en relación con otros procesos evolutivos (Purugganan y Fuller, 2009). Adicionalmente, el estudio del registro arqueológico y las prácticas de agricultura tradicional nos permiten entender el contexto ecológico en que se desencadenaron los procesos de domesticación (Casas *et al*., 1996; Fuller *et al.*, 2014; Milla *et al*., 2015). Por lo tanto, el estudio de la domesticación de *C. argyrosperma* nos permitió estudiar a detalle la interacción entre su genoma y el ambiente, revelando tanto la historia demográfica como los patrones de selección en esta especie (Kantar *et al.*, 2017; Barrera-Redondo *et al.*, 2020a).

Dado que las huellas históricas de la domesticación son recientes, diversas diciplinas como la arqueología, la ecología, la etnobotánica y la lingüística pueden aportar información valiosa al entendimiento de este proceso (Larson *et al*., 2014; Milla *et al*., 2015). Nuestros resultados de la demografía histórica de *C. argyrosperma* fueron congruentes y complementarios a la evidencia arqueológica que se tiene para esta especie, permitiéndonos elucidar un escenario claro de su domesticación (Barrera-Redondo *et al.*, 2020a).

Los análisis demográficos revelan que la domesticación de *C. argyrosperma* está íntimamente ligada a los patrones de migración y a los hábitos alimenticios de los humanos en Mesoamérica (Barrera-Redondo *et al*., 2020a). Dichos resultados fueron obtenidos mediante el análisis de la variación no funcional del genoma (Barrera-Redondo *et al.*, 2020a), por lo que el estudio del ADN neutral es valioso para tener una comprensión global de los procesos evolutivos. A su vez, el entendimiento de las prácticas agrícolas actuales (Montes-Hernández *et al*., 2005) y las necesidades nutricionales de las primeras poblaciones humanas en Mesoamérica (Zizumbo-Villarreal *et al*., 2012) nos permitió contextualizar las señales de selección positiva que observamos a lo largo del genoma de *C. argyrosperma* (Barrera-Redondo *et al.*, 2020a).

Nuestros resultados sugieren que las prácticas agrícolas tradicionales con las que se cosecha a *C. argyrosperma* promueven la diversidad genética y reducen el efecto de cuello de botella que se esperaría en un evento de domesticación (Montes-Hernández *et al*., 2005; Barrera-Redondo *et al*., 2020a). La congruencia entre los resultados obtenidos con marcadores genómicos y los resultados de otras diciplinas como la antropología (Zizumbo-Villarreal y Colunga-GarcíaMarín, 2010; Zizumbo-Villarreal *et al*., 2012) y la etnobotánica (Montes-Hernández *et al*., 2005) es evidencia del poder explicativo de la genética de poblaciones en el estudio de la evolución (Baedke *et al.,* 2020).

Algunas de las limitantes de la genómica de poblaciones en el entendimiento de la domesticación están basadas en su enfoque exclusivo hacia la variación genética (Baedke *et al.,* 2020). Sin embargo, también existen limitantes tecnológicas y económicas que impiden un análisis más profundo de la domesticación por la genómica de poblaciones, como es el uso alternativas a la secuenciación completa de genomas que capturan un menor número de variantes genéticas a favor de una reducción en los costos de secuenciación (Schreiber *et al*., 2018).

A pesar de que la reducción en los costos de secuenciación masiva ha revolucionado el estudio de la evolución (Eguiarte *et al*., 2013), muchas veces se tienen que recurrir a métodos de representación reducida del genoma para realizar estudios genómicos costeables a una escala poblacional. Como su nombre lo indica, estas técnicas de secuenciación reducen a porción total del genoma que es secuenciado para abaratar costos, ya sea mediante la secuenciación dirigida hacia la porción codificante del genoma (*e.g.*, exomas y RNAseq) o mediante a secuenciación de regiones adyacentes a sitios de restricción (*e.g.*, GBS y RADseq) (Schreiber *et al*., 2018). El uso de tGBS en *C. argyrosperma* eficientizó el costo de secuenciación por individuo, permitiendo el llamado de decenas de miles de variantes a lo largo del genoma de cientos de individuos con una alta confianza (Ott *et al*., 2017), con los cuales estimamos de manera precisa la historia demográfica de *C. argyrosperma* durante su domesticación (Barrera-Redondo *et al*., 2020a). Por otro lado, la búsqueda de barridos selectivos en *C. argyrosperma* fue fructífera pero limitada, dado que las librerías de tGBS no son aconsejables para realizar este tipo de pruebas (Lowry *et al*., 2017). Adicionalmente, se encontró un bajo desequilibrio de ligamiento entre los SNPs de *C. argyrosperma*, volviendo problemática la búsqueda de barridos electivos (Barrera-Redondo *et al*., 2020a). Esta problemática se debe a que la densidad de SNPs que se pueden recobrar utilizando tGBS no es suficiente para abarcar la totalidad de haplotipos que componen al genoma. Por lo tanto, la búsqueda de barridos selectivos se limita a un subconjunto de todas las posibles señales de selección que existen en el genoma, dado que pueden existir haplotipos bajo selección que pudieran no haber sido secuenciados con tGBS. Además, esto limita la asociación entre una variable detectada con señales de selección y su posible papel como variante causal del barrido detectado, dado que múltiples loci en un mismo haplotipo pueden presentar la misma señal de selección por efectos de *hitchhiking* (Lowry *et al*., 2017).

La secuenciación completa de genomas siempre será una alternativa superior al uso de representaciones reducidas del genoma para la búsqueda de señales de selección en el genoma, ya que se pueden secuenciar muchas más variantes a lo largo de cada cromosoma (Schreiber *et al*., 2018). Esto es particularmente útil cuando el desequilibrio de ligamiento se pierde rápidamente en los haplotipos y se pretenden encontrar las variantes causales de los barridos selectivos (Lowry *et al*., 2017). Sin embargo, los análisis de secuenciación completa de genomas a nivel poblacional también tienen sus limitaciones, como utilizar un solo genoma de referencia para realizar la búsqueda de variantes, además de los costos restrictivos que limitan el número de individuos y poblaciones secuenciadas (Golicz *et al*., 2016; Schreiber *et al*., 2018).

Considero que el siguiente paso en los estudios de domesticación en calabazas con enfoques de genómica poblacional deberá basarse en el análisis de pan-genomas, donde cada individuo sea secuenciado y ensamblado *de novo* (Golicz *et al*., 2016). Los estudios pan-genómicos permitirán el análisis de las variantes estructurales y de presencia/ausencia que estén involucradas en los procesos de domesticación (Golicz *et al*., 2016; Zhao *et al*., 2018). El elevado costo de esta aproximación lo vuelve una promesa en un futuro no muy cercano, cuando la secuenciación de tercera generación sea más barata y, con suerte, con una menor tasa de error (Golicz *et al*., 2016).

Eventualmente será necesario llevar a cabo estudios enfocados a la domesticación de las calabazas desde otras líneas de investigación que no solo nos permitan estudiar la evolución del genoma y su correlación con los fenotipos, sino también entender los mecanismos de causalidad que dan origen a los fenotipos (Baedke *et al.,* 2020). Un primer paso para ello serán los estudios de transcriptómica y metabolómica comparada entre calabazas domesticadas y sus parientes silvestres, con los cuales se puedan encontrar los patrones de expresión diferencial que liguen a los genotipos con los fenotipos (*e.g.*, Hradilová *et al*., 2017).

Los estudios de domesticación como herramienta biotecnológica

El estudio de la domesticación no solo es relevante para comprender las bases de la evolución o la conexión entre los cultivos y la humanidad (Ross-Ibarra *et al*., 2007; Zizumbo-Villarreal y Colunga-GarcíaMarín, 2010). También es un área que genera conocimiento agronómicamente relevante y de una aplicación directa y relativamente sencilla (Bellón *et al*., 2009; Hufford *et al*., 2012; Fernie y Yan, 2019). En concreto, el estudio de la domesticación y diversidad genética de los cultivos es de particular relevancia para la agricultura, ya que nos ayuda a reconocer y entender los recursos fitogenéticos. Con esto nos referimos a describir la variación genética de los cultivos con valor agronómico para la seguridad alimenticia, tanto de las plantas domesticadas como de sus parientes silvestres (Hajjar y Hodgkin, 2007; Bellón *et al*., 2009).

Un aspecto fundamental que influye en la conservación de los recursos fitogenéticos son las prácticas agrícolas bajo las que se siembran los cultivos (Jarvis *et al*., 2008). Se ha documentado en diversos cultivos del planeta que las prácticas agrícolas intensivas llevan a la extinción de variedades locales (Jarvis *et al*., 2008), lo cual erosiona la diversidad genética de las plantas domesticadas (Jarvis *et al*., 2008). Nuestras observaciones sugieren que las prácticas agrícolas tradicionales realizadas en *C. argyrosperma* (Montes-Hernández *et al*., 2005) promueven la diversidad genética en sus poblaciones domesticadas (Barrera-Redondo *et al*., 2020a), por lo que es importante salvaguardar estas prácticas tradicionales como reservas *in situ* de los recursos fitogenéticos (Jarvis *et al*., 2008).

Por lo general, las plantas domesticadas pasan por cuellos de botella y barridos selectivos que reducen dramáticamente la diversidad genética de los cultivos, reduciendo su capacidad adaptativa y fijando mutaciones deletéreas (Moyers *et al*., 2018). Este proceso se ve acentuado durante los procesos de mejoramiento, ya que se forman cuellos de botella adicionales que agotan la diversidad genética de los cultivos (Moyers *et al*., 2018).

Es importante señalar que las poblaciones domesticadas de *C. argyrosperma* no mostraron señales drásticas de pérdida de diversidad genética como resultado del cuello de botella (Sánchez-de la Vega *et al.*, 2018; Barrera-Redondo *et al*., 2020a). Consideramos que esto se debe en parte a que esta especie no ha pasado aún por procesos de mejoramiento genético moderno (Lira *et al.*, 2016), a diferencia de otros cultivos importantes como el maíz o el frijol (Moyers *et al*., 2018). Sin embargo, la atenuación del cuello de botella también puede estar dada por flujo genético constante con sus parientes silvestres (Barrera-Redondo *et al*., 2020a).

Los parientes silvestres representan una diversidad genética más amplia a la que presentan las poblaciones domesticadas. Estas poblaciones silvestres preservan elementos funcionales que se perdieron durante los cuellos de botella que sufrieron los taxa domesticados (Hajjar y Hodgkin, 2007). Por lo tanto, no solo es relevante conservar la diversidad genética de las calabazas domesticadas, sino también realizar esfuerzos de conservación en los taxa silvestres, quienes fungen como un reservorio invaluable de diversidad genética aprovechable en los cultivos domesticados (Hajjar y Hodgkin, 2007).

La búsqueda de genes candidatos asociados a la domesticación también resulta útil en el proceso del mejoramiento genético (Hufford *et al*., 2012). Se ha observado que los genes que fueron seleccionados durante el proceso de domesticación también son blanco de los esfuerzos por generar variedades mejoradas en los cultivos, por lo que identificarlos resulta de interés agronómico (Hufford *et al*., 2012).

Las señales de selección que encontramos en el genoma de *C. argyrosperma* revelaron genes candidatos que podrían estar asociados a caracteres de importancia agronómica como el tamaño del fruto, el tamaño de la semilla y el contenido de fosfolípidos en la semilla (Barrera-Redondo *et al*., 2020a). Estos resultados podrían guiar el eventual mejoramiento de los cultivos de *C. argyrosperma* por medio de selección asistida por marcadores o por la introgresión de alelos deseados con técnicas de edición genómica (Jiang, 2013; Zhou *et al*., 2019).

Adicionalmente, se pueden realizar prácticas de domesticación *de novo*, en las cuales a una planta silvestre o semi-domesticada se le introducen los alelos responsables de los fenotipos de domesticación con métodos de edición genómica (Fernie y Yan, 2019). Esta es una manera creativa en que se puede conseguir un cultivo con caracteres deseables sin el problema de acarrear los costos de la domesticación como la pérdida de resistencia a plagas u otros caracteres deletéreos e indeseables (Fernie y Yan, 2019). Para que un proyecto de domesticación *de novo* sea exitoso, es esencial conocer las variantes causales de los caracteres fenotípicos deseables en una planta cultivada (Fernie y Yan, 2019).

Por ello, es importante que los estudios de genes candidatos no solo estén basados en la asociación entre una variante y un fenotipo, sino que se realicen validaciones experimentales que demuestren una relación de causalidad (Zhang *et al*., 2017). Sin embargo, no hay que dejar que las expectativas tecnológicas de la ingeniería genética y de las domesticaciones *de novo* le resten importancia a la conservación *in situ* de los recursos fitogenéticos, ya que estos sientan las bases biológicas y culturales de los procesos de domesticación y mejoramiento (Bellón *et al*., 2009).

CONCLUSIONES

1. Se secuenciaron, ensamblaron y anotaron dos genomas de referencia a nivel de cromosoma de *C. argyrosperma*, uno para la subespecie domesticada y otro para la subespecie silvestre. Ambos sientan las bases para estudios evolutivos y agronómicos tanto en *C. argyrosperma* como en el resto del género *Cucurbita*.

2. Proponemos que la duplicación completa del genoma en el género *Cucurbita* produjo una aceleración en la tasa de evolución de familias de genes codificantes y ARNs largos no codificantes, ya que la redundancia funcional dentro del genoma facilitó la exaptación de elementos genéticos previamente existentes hacia nuevas funciones.

3. Durante el proceso de fraccionamiento y diploidización en calabazas, una proporción de los genes codificantes duplicados pasaron por procesos de pseudogenización y se propone que posteriormente se neofuncionalizaron como ARNs largos intergénicos no codificantes.

4. Los análisis de genómica poblacional revelaron que la domesticación de *C. argyrosperma* comenzó poco después de la llegada de los humanos a Mesoamérica, hace aproximadamente 13,800 años en Jalisco. Esto ocurrió bajo un flujo genético constante con su pariente silvestre, lo cual atenuó los efectos de cuello de botella durante la domesticación de la especie.

5. Detectamos señales de selección atribuibles al proceso de domesticación de *C. argyrosperma*, los cuales están asociados a genes que podrían explicar algunos aspectos del fenotipo de la subespecie domesticada como la pérdida de mecanismos de defensa, el crecimiento del fruto, el crecimiento de la semilla, la pérdida de la dormancia en la semilla y el contenido de lípidos en la semilla.

REFERENCIAS

Amundson, R., Lauder, G. V. (1994). Function without purpose. *Biology and philosophy* **9:** 443-469.

Baedke, J., Fábregas-Tejeda, A., Vergara-Silva, F. (2020). Does the extended evolutionary synthesis entail extended explanatory power? *Biology & Philosophy* **35:** 20.

Barrera-Redondo, J., Ibarra-Laclette, E., Vázquez-Lobo, A., Gutiérrez-Guerrero, Y. T., de la Vega, G. S., Piñero, D., Montes-Hernández, S., Lira-Saade, R., Eguiarte, L. E. (2019). The genome of *Cucurbita argyrosperma* (silver-seed gourd) reveals faster rates of protein-coding gene and long noncoding RNA turnover and neofunctionalization within *Cucurbita*. *Molecular Plant* **12:** 506-520.

Barrera-Redondo, J., Sánchez-de la Vega, G., Aguirre-Liguori, J. A., Castellanos-Morales, G., Aguirre-Dugua, X., Aguirre-Planter, E., Gutiérrez-Guerrero, Y. T., Tenaillon, M., Montes-Hernández, S., Lira-Saade, R., Eguiarte, L. E. (2020a). The domestication patterns of *Cucurbita argyrosperma* are consistent with early migration events in Mesoamerica and general domestication syndromes in squashes. *Under review.*

Barrera-Redondo, J., Lira-Saade, R., & Eguiarte, L. E. (2020b). Gourds and tendrils of Cucurbitaceae: how their shape diversity, molecular and morphological novelties evolved via whole-genome duplications. *Molecular Plant* **13:** 1108-1110.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* **136:** 215-233.

Bellón, M. R., Barrientos-Priego, A. F., Colunga-GarcíaMarín, P., Perales, H., Reyes Agüero, J. A., Rosales-Serna, R., Zizumbo-Villarreal, D. (2009). Diversidad y conservación de recursos genéticos en plantas cultivadas. *Capital natural de México* **2:** 355-382.

Bennett, M. D., Smith, J. B. (1976). Nuclear DNA Amounts in Angiosperms. *Philos. Trans. R. Soc. B Biol. Sci.* **274:** 227–274.

Bisognin, D. A. (2002). Origin and evolution of cultivated cucurbits. *Ciência Rural* **32:** 715–723.

Blount, Z. D., Lenski, R. E., Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life's tape. *Science* **362:** eaam5979.

Brigandt, I., Love, A. C. (2012). Conceptualizing evolutionary novelty: moving beyond definitional debates. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **318:** 417-427.

Casas, A., del Carmen Vázquez, M., Viveros, J. L., Caballero, J. (1996). Plant management among the Nahua and the Mixtec in the Balsas River Basin, Mexico: an ethnobotanical approach to the study of plant domestication. *Human Ecology* **24:** 455-478.

Casola, C. (2018). From *de novo* to "*de nono*": the majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. *Genome biology and evolution* **10:** 2906-2918.

Castellanos-Morales, G., Paredes-Torres, L. M., Gámez, N., Hernández-Rosales, H. S., Sánchez-de la Vega, G., Barrera-Redondo, J., ... Eguiarte, L. E. (2018). Historical biogeography and phylogeny of *Cucurbita*: insights from ancestral area reconstruction and niche evolution. *Molecular Phylogenetics and Evolution* **128:** 38-54.

Chanderbali, A. S., Berger, B. A., Howarth, D. G., Soltis, D. E., Soltis, P. S. (2017). Evolution of floral diversity: genomics, genes and gamma. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372:** 20150509.

Chekanova, J. A. (2015). Long non-coding RNAs and their functions in plants. *Current opinion in plant biology* **27:** 207-216.

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants* **4:** 258-268.

Chomicki, G., Schaefer, H., Renner, S. S. (2019). Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. *New Phytologist* **226:** 1240-1255.

Chouard, T. (2010). Evolution: revenge of the hopeful monster. *Nature* **463:** 864–867.

Clark J. W., Donoghue C. J. (2018). Whole-Genome Duplication and Plant Macroevolution. *Trends in Plant Science* **23:** 933-945.

Crombach, A., Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. *PLoS computational biology*, **4:** e1000112.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy* **72:** 741–765.

Diamond J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature* **418:** 700–707.

Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., ..., Hofberger, J. A. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences* **112:** 8362-8366.

Eguiarte, L. E., Aguirre-Liguori, J. A., Jardón-Barbolla, L., Aguirre-Planter, E., Souza, V. (2013). Genómica de poblaciones: nada en evolución va a tener sentido si no es a la luz de la genómica, y nada en genómica tendrá sentido si no es a la luz de la evolución. *TIP Revista Especializada En Ciencias Químico-Biológicas*, **16(1):** 42-56.

Esteras, C., Gómez, P., Monforte, A. J., Blanca, J., Vicente-Dólera, N., Roig, C., ... Picó, B. (2012). High-throughput SNP genotyping in *Cucurbita pepo* for map construction and quantitative trait loci mapping. *BMC Genomics* **13:** 80.

Fernie, A. R., Yan, J. (2019). *De novo* domestication: an alternative route toward new crops for the future. *Molecular Plant* **12:** 615-631.

Fuller, D. Q., Denham, T., Arroyo-Kalin, M., Lucas, L., Stevens, C. J., Qin, L., ... Purugganan, M. D. (2014). Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. *Proceedings of the National Academy of Sciences* **111:** 6147-6152.

Futuyma, D. J. (2005). Evolution. Sinauer Associates, Massachusetts. 11p.

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., González, V. M., ... Alioto, T. (2012). The genome of melon (*Cucumis melo* L.). *Proceedings of the National Academy of Sciences* **109:** 11872-11877.

Gepts, P. (2014). The contribution of genetic and genomic approaches to plant domestication studies. *Curr. Opin. Plant Biol.* **18:** 51–59.

Goldschmidt, R. (1933). Some aspects of evolution. *Science* **78:** 539–547.

Golicz, A. A., Batley, J., Edwards, D. (2016). Towards plant pangenomics. *Plant Biotechnology Journal* **14:** 1099-1105.

Gong, L., Stift, G., Kofler, R., Pachner, M., Lelley, T. (2008). Microsatellites for the genus *Cucurbita* and an SSR-based genetic linkage map of *Cucurbita pepo* L. *Theor. Appl. Genet.* **117:** 37–48.

Gould, S. J. (1994). El pulgar del panda: reflexiones sobre historia natural y evolución (No. 575.8 GOU).

Graur, D., Zheng, Y., Azevedo, R. B. (2015). An evolutionary classification of genomic function. *Genome biology and evolution* **7:** 642-645.

Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., ... Min, J. (2013). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nature genetics* **45:** 51-58.

Guo, J., Xu, W., Hu, Y., Huang, J., Zhao, Y., Zhang, L., Huang, C.-H., Ma, H. (2020). Phylotranscriptomics in Cucurbitaceae Reveal Multiple Whole Genome

Duplications and Key Morphological and Molecular Innovations. *Molecular Plant* **13:** 1117–1133.

Gustafson, P., Gong, L., Pachner, M., Kalai, K., Lelley, T. (2008). SSR-based genetic linkage map of *Cucurbita moschata* and its synteny with *Cucurbita pepo*. *Genome* **51:** 878–887.

Hajjar, R., Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* **156:** 1-13.

Hancock, J. F. (2005). Contributions of Domesticated Plant Studies to our Understanding of Plant Evolution. *Ann. Bot.* **96:** 953–963.

Hautmann, M. (2020). What is macroevolution*? Frontiers in Palaeontology* **63:** 1-11.

Hedrick, P. W. (2011). *Genetics of populations*. Jones & Bartlett Learning.

Hu, Z., Wang, W., Wu, Z., Sun, C., Li, M., Lu, J., Fu, B., Shi, J., Xu, J., Ruan, J., Wei, C. (2018). Novel sequences, structural variations and gene presence variations of Asian cultivated rice. *Scientific Data* **5:** 180079.

Hradilova, I., Trněný, O., Valkova, M., Cechova, M., Janska, A., Prokešová, L., ... Varshney, R. K. (2017). A combined comparative transcriptomic, metabolomic, and anatomical analyses of two key domestication traits: pod dehiscence and seed dormancy in pea (*Pisum* sp.). *Frontiers in Plant Science* **8:** 542.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., ... Ren, Y. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nature Genetics* **41:** 1275-1281.

Hufford, M. B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J. M., Cartwright, R. A., ... Lai, J. (2012). Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44:** 808-811.

Jablonski, D. (2008). Species selection: theory and data. *Annual review of ecology, evolution, and systematics* **39:** 501-524.

Jarvis, D. I., Brown, A. H., Cuong, P. H., Collado-Panduro, L., Latournerie-Moreno, L., Gyawali, S., ... Hue, N. T. N. (2008). A global perspective of the richness and evenness of traditional crop-variety diversity maintained by farming communities. *Proceedings of the National Academy of Sciences* **105:** 5326-5331.

Jiang, G. L. (2013). Molecular markers and marker-assisted breeding in plants. *Plant breeding from laboratories to fields*, 45-83.

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., ... Zhang, M. (2012). Genome-wide genetic changes during modern breeding of maize. *Nature Genetics* **44:** 812-815.

Jorde, L. B. (2001). Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics* **10:** 2199–2207.

Kantar, M. B., Nashoba, A. R., Anderson, J. E., Blackman, B. K., Rieseberg, L. H. (2017). The genetics and genomics of plant domestication. *BioScience* **67:** 971-982.

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS genetics* **9:** e1003470.

Kapusta, A., Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics* **30:** 439-452.

Kauffman, S. A. (1993). The origins of order: Self-organization and selection in evolution. OUP USA.

Keeling, D. M., Garza, P., Nartey, C. M., Carvunis, A. R. (2019). The meanings of 'function' in biology and the problematic case of *de novo* gene emergence. *eLife* **8:** e47014.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., ... Dunham, I. (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences* **111:** 6131-6138.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217:** 624-626.

Koenig, D., Jiménez-Gómez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., ... Tohge, T. (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proceedings of the National Academy of Sciences* **110:** E2655-E2662.

Lam, H. M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F. L., ... Li, J. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* **42:** 1053-1059.

Larson, G., Piperno, D. R., Allaby, R. G., Purugganan, M. D., Andersson, L., Arroyo-Kalin, M., Barton, L., Climer Vigueira, C., Denham, T., Dobney, K., *et al*. (2014). Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences* **111:** 6139–6146.

Lee, D.H., Iwanski, G.B., Thoennissen, N.H. (2010). Cucurbitacin: ancient compound shedding new light on cancer treatment. *Scientific World Journal* **10:** 413-418.

Lewontin, R. C. Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54:** 595-609.

Li, Y. H., Zhao, S. C., Ma, J. X., Li, D., Yan, L., Li, J., ... Chang, R. Z. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14:** 1-12.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., ... Huang, Z. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics* **46:** 1220-1226.

Lira, R., Eguiarte, L. E., Montes-Hernández, S. (2009). Proyecto Recopilación y análisis de la información existente de las especies de los géneros *Cucurbita* y *Sechium* que crecen y / o se cultivan en México. CONABIO, México, D.F. 107p.

Lira, R., Eguiarte, L. E., Montes, S., Zizumbo-Villarreal, D., Colunga-GarcíaMarín, P., Quesada, M. (2016). *Homo sapiens-Cucurbita* interaction in Mesoamerica: Domestication, Dissemination and Diversification. *In*: Lira, R., Casas, A., Blancas, J. (eds.). Ethnobotany of Mexico. Springer-Verlag, New York, 389–402. ISBN: 978-1-4614-6669-7

Long, M., VanKuren, N. W., Chen, S., Vibranovski, M. D. (2013). New gene evolution: little did we know. *Annual review of genetics* **47:** 307-333.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., Storfer, A. (2017). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular ecology resources* **17:** 142-152.

Lynch, M., Walsh, B. (2007). The origins of genome architecture (Vol. 98). Sunderland, MA: Sinauer Associates.

Mallet, J. (1995). A species definition for the modern synthesis. *Trends in Ecology & Evolution* **10:** 294-299.

Mattenberger, F., Sabater-Muñoz, B., Toft, C., Fares, M. A. (2017). The phenotypic plasticity of duplicated genes in *Saccharomyces cerevisiae* and the origin of adaptations. *G3: Genes, Genomes, Genetics* **7:** 63-75.

Meyer, R. S., Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nature Review Genetics* **14:** 840–52.

Milla, R., Osborne, C. P., Turcotte, M. M., Violle, C. (2015). Plant domestication through an ecological lens. *Trends in ecology & evolution* **30:** 463-469.

Montero-Pau, J., Blanca, J., Bombarely, A., Ziarsolo, P., Esteras, C., Martí-Gómez, C., Ferriol, M., Gómez, P., Jamilena, M., Mueller, L., *et al.* (2018). *De novo* assembly of the zucchini genome reveals a whole genome duplication associated with the origin of the *Cucurbita* genus. *Plant Biotechnol. J.* **12:** 3218–3221.

Montes-Hernández, S., Merrick, L. C., Eguiarte, L. E. (2005). Maintenance of squash (*Cucurbita* spp.) landrace diversity by farmers' activities in Mexico. *Genetic Resources and Crop Evolution* **52:** 697-707.

Moyers, B. T., Morrell, P. L., McKay, J. K. (2018). Genetic costs of domestication and improvement. *Journal of Heredity* **109:** 103-116.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505:** 635-640.

Nee, M. (1990). The domestication of *Cucurbita* (Cucurbitaceae). *Economic Botany* **44:** 56–68.

Nelson, A. D. L., Shippen, D. E. (2015). Evolution of TERT-interacting lncRNAs: expanding the regulatory landscape of telomerase. *Frontiers in Genetics* **6:** 1–6.

Nelson, A. D. L., Forsythe, E. S., Devisetty, U. K., Clausen, D. S., Haug-Batzell, A. K., Meldrum, A. M., ... Beilstein, M. A. (2016). A genomic analysis of factors driving lincRNA diversification: lessons from plants. *G3: Genes, Genomes, Genetics* **6:** 2881-2891.

Nelson, A. D. L., Devisetty, U. K., Palos, K., Haug-Baltzell, A. K., Lyons, E., Beilstein, M. A. (2017). Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs. *Frontiers in Genetics* **8:** 1–12.

Nunes, M. D., Arif, S., Schlötterer, C., McGregor, A. P. (2013). A perspective on micro-evo-devo: progress and potential. *Genetics* **195:** 625-634.

Ott, A., Liu, S., Schnable, J. C., Yeh, C. T. E., Wang, K. S., Schnable, P. S. (2017). tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. *Nucleic acids research* **45:** e178.

Paris, H. S. (2016). Genetic Resources of Pumpkins and Squash, *Cucurbita* spp. Pp. 111–154. In: Grumet, R., Katzir, N., Garcia-Mas, J. (eds.). Genetics and Genomics of Cucurbitaceae. Springer, Gewerbestrasse 11, 6330 Cham, Switzerland.

Pigliucci, M. (2010). Genotype–phenotype mapping and the end of the 'genes as blueprint' metaphor. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365:** 557-566.

Piperno, D. R., Ranere, A. J., Holst, I., Iriarte, J., Dickau, R. (2009). Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, Mexico. *Proceedings of the National Academy of Sciences* **106:** 5019-5024.

Purugganan, M. D., Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature* **457:** 843–848.

Qi, J., Liu, X., Shen, D., Miao, H., Xie, B., Li, X., ... Du, Y. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nature genetics* **45:** 1510.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., ... Yang, Y. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proceedings of the National Academy of Sciences* **111:** 5135-5140.

Rendón-Anaya, M., Ibarra-Laclette, E., Méndez-Bravo, A., Lan, T., Zheng, C., Carretero-Paulet, L., …, Farr, K. M. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the National Academy of Sciences* **116:** 17081-17089.

Reznick, D. N., Ricklefs, R. E. (2009). Darwin's bridge between microevolution and macroevolution. *Nature* **457:** 837-842.

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., ... McMullen, M. D. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome biology* **14:** R55.

Ross-Ibarra, J., Morrell, P. L., Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences* **104:** 8641-8648.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *elife*, **3:** e03523.

Sanjur, O. I., Piperno, D. R., Andres, T. C., Wessel-Beaver, L. (2002). Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: Implications for crop plant evolution and areas of origin. *Proceedings of the National Academy of Sciences* **99:** 535-540.

Schaefer, H., Heibl, C., Renner, S. S. (2009). Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proceedings of the Royal Society B: Biological Sciences* **276:** 843-851.

Schlötterer, C. (2004). The evolution of molecular markers—just a matter of fashion? *Nature reviews genetics* ***5*****(1)**: 63.

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., ... Torres-Torres, M. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics* **46:** 707-713.

Schreiber, M., Stein, N., Mascher, M. (2018). Genomic approaches for studying crop evolution. *Genome biology* **19:** 140.

Shang, Y., Ma, Y., Zhou, Y., Zhang, H., Duan, L., Chen, H., ... Liu, M. (2014). Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346:** 1084-1088.

Shearwin, K. E., Callen, B. P., Egan, J. B. (2005). Transcriptional interference–a crash course. *Trends in Genetics* **21:** 339-345.

Singh, A.K. (1990). Cytogenetics and evolution in the Cucurbitaceae. *In*: Bates, D., Robinson, R., Jeffrey, C. (Eds.). Biology and Utilization of the Cucurbitaceae. Cornell University Press, Ithaca, New York. pp. 10-28.

Šiško, M., Ivančič, A., Bohanec, B. (2003). Genome size analysis in the genus *Cucurbita* and its use for determination of interspecific hybrids obtained using the embryo rescue technique. *Plant Sci.* **165:** 663–669.

Smith, B. D. (1997). The Initial Domestication of *Cucurbita pepo* in the Americas 10,000 Years Ago. *Science* **276:** 932–934.

Tallamy, D. W., Hibbard, B. E., Clark, T. L., Gillespie, J. J. (2005). Western corn rootworm: ecology and management. *In:* Vidal, S., Kuhlmann, U., Edwards, C.R. (Eds.). Western corn rootworm: Ecology and management. CAB international, UK. pp. 67-94.

Tan, S., Zhong, Y., Hou, H., Yang, S., Tian, D. (2012). Variation of presence/absence genes among *Arabidopsis* populations. *BMC evolutionary biology* **12:** 86.

Theißen, G. (2009). Saltational evolution: hopeful monsters are here to stay. *Theory in Biosciences* **128:** 43–51.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nature Reviews Genetics* **17:** 601.

Vakirlis, N., Acar, O., Hsu, B., Coelho, N. C., Van Oss, S. B., Wacholder, A., … Parikh, S. B. (2020). *De novo* emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature Communications* **11:** 1-18.

Van Oss, S. B., & Carvunis, A. R. (2019). *De novo* gene birth. *PLoS genetics* **15:** e1008160.

Wang, B., Zheng, C., Sankoff, D. (2011). Fractionation statistics. In: *BMC bioinformatics* (Vol. 12, No. S9, p. S5). BioMed Central.

Wang, J., Sun, P., Li, Y., Liu, Y., Yang, N., Yu, J., ... Ge, D. (2018). An overlooked paleotetraploidization in Cucurbitaceae. *Molecular Biology and Evolution* **35:** 16-26.

Webster, G., Goodwin, B. (1984). A structuralist approach to morphology. *Rivista di Biologia* **77:** 503-510.

Weeden, N. (1984). Isozyme studies indicate that the genus *Cucurbita* is an ancient tetraploid. *Cucurbit Genet. Coop. Rep.* **7:** 84-85

Weiling F. (1959). Genomanalytische Untersuchungen bei Kürbis (*Cucurbita* L.). *Der Züchter*, ***29*****(4):** 161–179.

Whitaker, T. W. (1981). Archeological cucurbits. *Econ. Bot.* **35:** 460–466.

Whitaker, T. W. (1933). Cytological and Phylogenetic Studies in the Cucurbitaceae. *Bot. Gaz.* **94:** 780–790.

Williams, T. A., Nakjang, S., Campbell, S. E., Freeman, M. A., Eydal, M., Moore, K., Hirt, R. P., Embley, T. M., Williams, B. A. (2016). A Recent Whole-Genome Duplication Divides Populations of a Globally Distributed Microsporidian. *Molecular biology and evolution* **33:** 2002–2015.

Wong, K. K., deLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., MacAulay, C., Ng, R. T., Brown, C. J., Eichler, E. E., Lam, W. L. (2007). A comprehensive analysis of common copy-number variations in the human genome. *The American Journal of Human Genetics* **80:** 91-104.

Wright, L. (1973). Function. *Philosophical Review* **82:** 139-168.

Wu, S., Shamimuzzaman, M., Sun, H., Salse, J., Sui, X., Wilder, A., Wu, Z., Levi, A., Xu, Y., Ling, K-S., *et al.* (2017). The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a Papaya ring-spot virus resistance locus. *Plant J.* **92:** 963–975.

Xia, X. (2013). What is Comparative Genomics? *In*: Comparative Genomics (pp. 1-20). Springer, Berlin, Heidelberg.

Xie, D., Xu, Y., Wang, J., Liu, W., Zhou, Q., Luo, S., Huang, W., He, X., Li, Q., Peng, Q., Yang, X., Yuan, J., Yu, J., Wang, X., Lucas, W. J., Huang, S., Jiang, B., Zhang, Z. (2019). The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nature Communications* **10(1):** 5158.

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., ... Li, J. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature biotechnology* **30:** 105-111.

Zeder, M.A. (2015). Core questions in domestication research. *Proc. Natl. Acad. Sci.* **112:** 3191–3198.

Zhang, G., Ren, Y., Sun, H., Guo, S., Zhang, F., Zhang, J., Zhang, H., Jia, Z., Fei, Z., Xu, Y., Li, H. (2015). A high-density genetic map for anchoring genome sequences and identifying QTLs associated with dwarf vine in pumpkin (*Cucurbita maxima* Duch.). *BMC Genomics* **16:** 1101. BMC Genomics.

Zhang, H., Zhang, J., Lang, Z., Botella, J. R., Zhu, J. K. (2017). Genome editing — principles and applications for functional genomics research and crop improvement. *Critical Reviews in Plant Sciences* **36:** 291-309.

Zheng, Y. H., Alverson, A. J., Wang, Q. F., Palmer, J. D. (2013). Chloroplast phylogeny of *Cucurbita*: evolution of the domesticated and wild species. *Journal of Systematics and Evolution* **51:** 326-334.

Zhou, Y., Ma, Y., Zeng, J., Duan, L., Xue, X., Wang, H., ... & Zhang, S. (2016). Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nature plants* **2:** 16183.

Zhou, J., Li, D., Wang, G., Wang, F., Kunjal, M., Joldersma, D., Liu, Z. (2019). Application and future perspective of CRISPR/Cas9 genome editing in fruit crops. *Journal of integrative plant biology* **62:** 269-286.

Zizumbo-Villarreal, D., Colunga-GarcíaMarín, P. (2010). Origin of agriculture and plant domestication in West Mesoamerica. *Genetic Resources and Crop Evolution* **57:** 813-825.

Zizumbo-Villarreal, D., Flores-Silva, A., Marín, P. C. G. (2012). The archaic diet in Mesoamerica: incentive for milpa development and species domestication. *Economic Botany* **66:** 328-343.

Zraidi, A., Stift, G., Pachner, M., Shojaeiyan, A., Gong, L., Lelley, T. (2007). A consensus map for *Cucurbita pepo*. *Mol. Breed.* **20:** 375–388.

Zuckerkandl, E., Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology* **8:** 357-366.

ANEXO 1: MATERIAL SUPLEMENTARIO DEL CAPÍTULO 1

Supplementary Material

1 Supplementary Figures and Tables

TABLE S1 | Some examples of bottom-up tests designed to detect selective sweeps in populations using genomic data and their underlying models (modified from Vitti et al., 2013). See the main text to find the references cited in this table.

Underlying model	Test name	Rationale
F_{ST} outlier tests	BayeScan	This method estimates the correlation between allele frequencies (*i.e.*, the genetic structure caused by demographic events) by calculating a multinomial-Dirichlet distribution using a Markov chain Monte Carlo Bayesian approach to subsequently detect outlier loci. This method assumes an island model (*i.e.*, gene flow is equally frequent between each pair of populations) which can lead to false positives when used in populations with limited gene flow and strong population bottlenecks, as expected for many domesticated taxa (Foll and Gaggiotti, 2008).
	Excoffier, Hofer and Foll (EHF) test	Uses coalescent simulations to calculate the distribution of genetic diversity within and between populations to obtain a null distribution of FST values and detect outliers. This method is optimized for hierarchically structured populations (*i.e.*, genetic flow is more frequent within populations and between populations belonging to the same group such as domesticated or wild taxa) by using a hierarchical island model in which populations are assigned to groups *a priori* (Excoffier et al., 2009).
	$T_{F\text{-}LK}$ statistic	This method calculates a kinship matrix between populations by generating a neighbor-joining population tree in order to obtain a null distribution to detect F_{ST} outliers. The statistic seems to be robust to complex demographic scenarios, reducing the emergence of false candidate loci (Bonhomme et al., 2010).
	BayeScEnv	Uses the same approximation as BayeScan to detect F_{ST} outliers, but it also incorporates environmental cues to detect adaptation to local environment (de Villemereuil and Gaggiotti, 2015).
	PCAdapt	This method calculates the underlying population structure using a principal component analysis and then detects candidate loci under selection using Mahalanobis distance. This method is particularly powerful when handling admixed individuals and hierarchical structure in the studied populations (Luu et al., 2017).

SFS based methods	Tajima's *D* statistic	Calculated from the comparison of the nucleotide diversity (π) against Watterson's Theta (θ_W). When $\pi < \theta_W$, the region has an excess of low-frequency variants, suggesting purifying or positive selection, whereas when $\pi > \theta_W$, the region has an excess of middle-frequency variants, suggesting balancing selection or a soft selective sweep (Tajima, 1989).
	Fai and Wu's *H* statistic	Calculated from the comparison of π against θ_H or θ_L, which are estimators of theta weighted by the homozygosity of derived variants. When $\pi < \theta_H$, the region has an excess of high-frequency variants, a characteristic signature of selective sweeps (Fay and Wu, 2000).
	Zeng *et al.*'s *E* statistic	Calculated from the comparison of θ_L against θ_W, rendering it sensible to changes in high and low-frequency variants, which are signals of selective sweeps before and after the fixation of the locus under selection (Zeng et al., 2006).
	Reduction of diversity (ROD) test	The selective sweeps associated to domestication will form a pattern where the domesticated taxon will have a significantly lower diversity in that region compared to the overall diversity in its genome, while the wild taxon will not show reduced genetic diversity in that locus ($\pi_{wild} > \pi_{domesticate}$) (Guo et al., 2012).
LD based methods	Long-range haplotype (LRH) test	This test uses the EHH statistic to detect whether an haplotype is inherited throughout the population without its disruption by recombination, suggesting that such haplotype is under positive selection (Sabeti et al., 2002).
	Whole-genome long-range haplotype (WGLRH) test	The WGLRH test performs the LRH test throughout the entire genome using sliding windows to detect EHH outliers (Zhang et al., 2006).
	Long-range haplotype similarity (LRHs) test	Calculates the similarity between homologous haplotypes by calculating an haplosimilarity score throughout the genome using sliding windows in order to detect haplotypes that contain alleles with low frequencies that are similar between each other, suggesting large haplotypes under a recent selective pressure that hasn't been disrupted by recombination (Hanchard et al., 2006).
	Integrated haplotype score (iHS)	The iHS compares the area under the curve defined by the EHH, which allows the identification of incomplete selective sweeps and soft sweeps throughout the genome (Voight et al., 2006).
	Cross-population extended haplotype homozygosity (XP-EHH) statistic	This test compares the EHH in a locus between a population with a fixed haplotype against other populations where such locus remains polymorphic (*e.g.*, domesticated and wild populations), allowing it to detect selective sweeps after the selected allele reached fixation (Sabeti et al., 2007).

	LD decay (LDD) test	The LDD test sorts individuals according to their homozygosity for each of the alleles found in any given SNP in the genome, and then calculates the fraction of heterozygous SNPs that are adjacent to each of the allelic variants in the SNP that is being evaluated. This way the LDD test can determine whether or not the LD decays significantly slower in one of the alleles of the evaluated SNP, suggesting a recent selective sweep. Since the test calculates the decay in LD only for the individuals that are homozygous in the SNP being evaluated, there is no necessity to obtain phased haplotypes (Wang et al., 2006).
	Regression-based test	Calculates the reduction of heterozygosity as one approaches the locus under selection in a genome to infer selective sweeps (Wiener and Pong-Wong, 2011).
	OmegaPlus	Implements the ω statistic by detecting regions with high SNP correlation coefficient across the genome to find regions under selection (Alachiotis et al., 2012).
	GIBDLD	Calculates the identity-by-descent (IBD) between all pairs of individuals for each locus in the dataset to detect segments of IBD (*i.e.*, haplotypes) that are shared between several unrelated pairs of individuals, suggesting the action of selective pressures (Han and Abney, 2013).
Composite tests	Cross-population composite likelihood ratio (XP-CLR) test	The XP-CLR test searches for genomic regions with extended allele differentiation between a population under selective pressures (*i.e.*, domesticated taxa) and a "control" population (*i.e.*, a close wild relative) in order to detect the haplotypes where differentiation happened quicker than expected under neutrality (Chen et al., 2010).
	RAiSD	This program uses the μ statistic to score genomic regions as candidate loci based on the joint analysis of LD, changes in the SFS and a reduction in the genetic diversity within sliding windows (Alachiotis and Pavlidis, 2018).

ANEXO 2: MATERIAL SUPLEMENTARIO DEL CAPÍTULO 2

Supplemental information

The genome of *Cucurbita argyrosperma* (silver-seed gourd) reveals faster rates of protein-coding gene and long noncoding RNA turnover and neofunctionalization within *Cucurbita*

Content

Supplemental Methods

Library construction and DNA/RNA sequencing

Illumina library construction and sequencing were done at the Vincent J. Coates Genomics Sequencing Laboratory in UC Berkeley (NIH S10 Instrumentation Grants S10RR029668 and S10RR027303). *C. argyrosperma* genomic DNA was sheared using a Bioruptor Pico sonication device (Diagenode). Illumina libraries of 350 bp and 1,000 bp were constructed and sequenced on HiSeq2000 and MiSeq systems, respectively, using paired-end mode.

PacBio RS II (P6/C4 chemistry) sequencing was done at the University of Washington PacBio Sequencing Services. Libraries were prepared and size-selected with fragments >8 kbp using the BluePippin system (Sage Science). Sequencing was done on 8 SMRT cells, using MagBeads to prioritize the sequencing of longer DNA fragments.

For the *C. argyrosperma* transcriptome, Illumina ribo-depleted libraries (350 bp) were prepared for each RNA sample after cDNA transformation and shearing. All libraries were multiplexed and sequenced in a single HiSeq2000 lane using paired-end mode.

Sequence read quality filters

Raw Illumina reads were quality-filtered using the qualityControl.py script (https://github.com/Czh3/NGSTools), retaining only those read pairs that showed a PHRED quality ≥ 30 in at least 85% of its sequence and an average PHRED quality ≥ 25. We used SeqPrep (https://github.com/jstjohn/SeqPrep) to trim Illumina adapter sequences and merge read pairs that showed at least 20 bp of overlap. PacBio subreads were obtained after filtering reads with a minimum length of 50 bp and a minimum quality score of 75.

Organelle and nuclear genome assemblies

The chloroplast genome of *C. argyrosperma* was assembled using the Illumina quality trimmed reads with the NOVOPlasty assembler (Dierckxsens *et al.*, 2016). We assembled the chloroplast genome in two contigs using the chloro2 algorithm, a k-mer size of 39 and a chloroplast genome previously reported by Kistler *et al.* (2015) as a sequence seed (GenBank accession number KT898803.1). The chloroplast contigs were merged in a single scaffold using the PacBio reads and the scaffolding software SSPACE-longread (Boetzer and Pirovano, 2014). A final gap-filling and base correction step was done with Pilon (Walker *et al.*, 2014) by mapping the Illumina quality trimmed reads using BWA *mem* (Li, 2013) to the chloroplast scaffold.

We filtered the reads belonging to the organelle genomes before nuclear genome assembly. Illumina quality trimmed reads were filtered by mapping them to the assembled chloroplast genome and the previously assembled *C. pepo* mitochondrial genome (Alverson *et al.* 2010, GenBank accession number GQ856148.1) using Hisat2 (Kim *et al.*, 2015). PacBio data was filtered by mapping the reads to the organelle genomes using BlasR (Chaisson and Tesler, 2012).

The PacBio reads that mapped to the *C. pepo* mitochondrion were used to assemble the *C. argyrosperma* mitochondrion genome using the Organelle-PBA pipeline (Soorni *et al.*, 2017). The mitochondrial contigs were merged into scaffolds using the mitochondrial PacBio reads and the SSPACE-longread script. A final gap-filling and base correction step was done to the mitochondrial scaffolds using the Illumina quality trimmed reads by doing 5 iterations of Pilon and BWA *mem*.

The Illumina nuclear reads were assembled into contigs with the Platanus assembler (Kajitani *et al.*, 2014), using an initial k-mer size of 32, a step size of 10 for k-mer extension and a bubble crush parameter of 0.1. The assembled contigs were used alongside the PacBio nuclear reads to construct an assembly backbone according to the DBG2OLC pipeline (Ye *et al.*, 2016), considering a k-mer size of 17, a k-mer matching threshold of 5, an adaptive k-mer threshold of 0.01 and a minimum matching k-mer number of 30 between two PacBio reads.

We used Minimap and Racon (Vaser *et al.*, 2017) twice to map the *Platanus* contigs and the nuclear PacBio reads to the assembly backbone and obtain a consensus sequence assembly. We did three iterations of base correction by mapping the Illumina nuclear reads to the consensus sequence assembly using BWA *mem* and Pilon. Scaffolding was done with the paired-end information of the Illumina reads using BESST (Sahlin *et al.*, 2014) and with the PacBio reads using SSPACE-longread. We did 30 iterations of gap closing with GapFiller (Boetzer and Pirovano, 2012), using Bowtie (Langmead *et al.*, 2009) to map the Illumina reads to the assembled genome. Three final base corrections were done with BWA *mem* and Pilon.

De novo **and genome-guided transcriptome assemblies**

The transcriptome was assembled *de novo* using Trinity (Grabherr *et al.*, 2011). DeconSeq (Schmieder and Edwards, 2011) was used to remove contigs with at least 90% identity to a database of contaminant sequences (bacteria, virus, fungi and arthropods). Poly A tails were trimmed from the remaining contigs with SeqClean (Tae *et al.*, 2012). The remaining contigs were reassembled using CAP3 (Huang and Madan, 1999) to merge fragmented transcripts.

The transcriptome reads were also used to generate a genome-guided transcriptome assembly. We mapped the transcriptome reads of each organ to the genome assembly using Hisat2 (Kim *et al.*, 2015). The mapped reads of each organ were assembled into transcripts using StringTie (Pertea *et al.*, 2015) using a minimum coverage of 1, a minimum intron length of 20 bp and a maximum intron length of 500,000 bp. All transcripts were merged into a single genome-guided transcriptome assembly using Cuffmerge (Trapnell *et al.*, 2012).

Prediction of protein-coding gene models

We searched for ORFs within the *de novo* assembled transcriptome to identify coding transcripts using AlignWise (Evans and Loose, 2015). The proteins of the predicted ORFs were compared to the proteomes of *Citrullus lanatus* (Guo *et al.*, 2012), *Cucumis sativus* (Huang *et al.*, 2009) and *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) with BLASTp (Camacho *et al.*, 2009). We considered as complete or nearly complete proteins those that had an alignment coverage of at least 70% compared to its best BLAST hit (6,987 proteins) and used them to train AUGUSTUS (Stanke *et al.*, 2006) and obtain a hidden Markov model (HMM) file for gene prediction.

MAKER3 (Cantarel *et al.*, 2008) was used to obtain protein-coding gene models in the *C. argyrosperma* genome assembly. We used the repeat library obtained from REPET to mask TEs with RepeatMasker, as well as RepeatRunner

(Smith *et al.*, 2007) to mask transposable element proteins. GeneMark-ES (Lomsadze *et al.*, 2005) was used to obtain a self-trained HMM file from the masked genome. We used AUGUSTUS and GeneMark-ES HMM files for *ab initio* gene prediction. Both the *de novo* and the genome-guided transcriptome assemblies were used as transcriptomic evidence. The SwissProt database (downloaded on July 13, 2017) was used as protein homology evidence. We used EvidenceModeler (Haas *et al.*, 2008) within MAKER3 to obtain additional gene models from the weighted evidence of *ab initio* predictions, transcript evidence and protein homology evidence. The gene models obtained from MAKER3 were used to train SNAP (Korf, 2004) and obtain a HMM file for the second round of gene model prediction.

We used MAKER3 for a second round of gene model prediction, incorporating the SNAP HMM training file, the previous MAKER3 models as an annotation pass-through and using the proteomes of *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000), *Cucumis sativus* (Huang *et al.*, 2009), *Cucumis melo* (Garcia-Mas *et al.*, 2012), *Theobroma cacao* (Argout *et al.*, 2011), *Vitis vinifera* (Jaillon *et al.*, 2007), *Fragaria vesca* (Edger *et al.*, 2018) and *Citrus sinsensis* (Xu *et al.*, 2012) as additional protein homology evidence. We incorporated tRNAscan-SE (Lowe and Eddy, 1997) within the MAKER3 pipeline to predict tRNA genes.

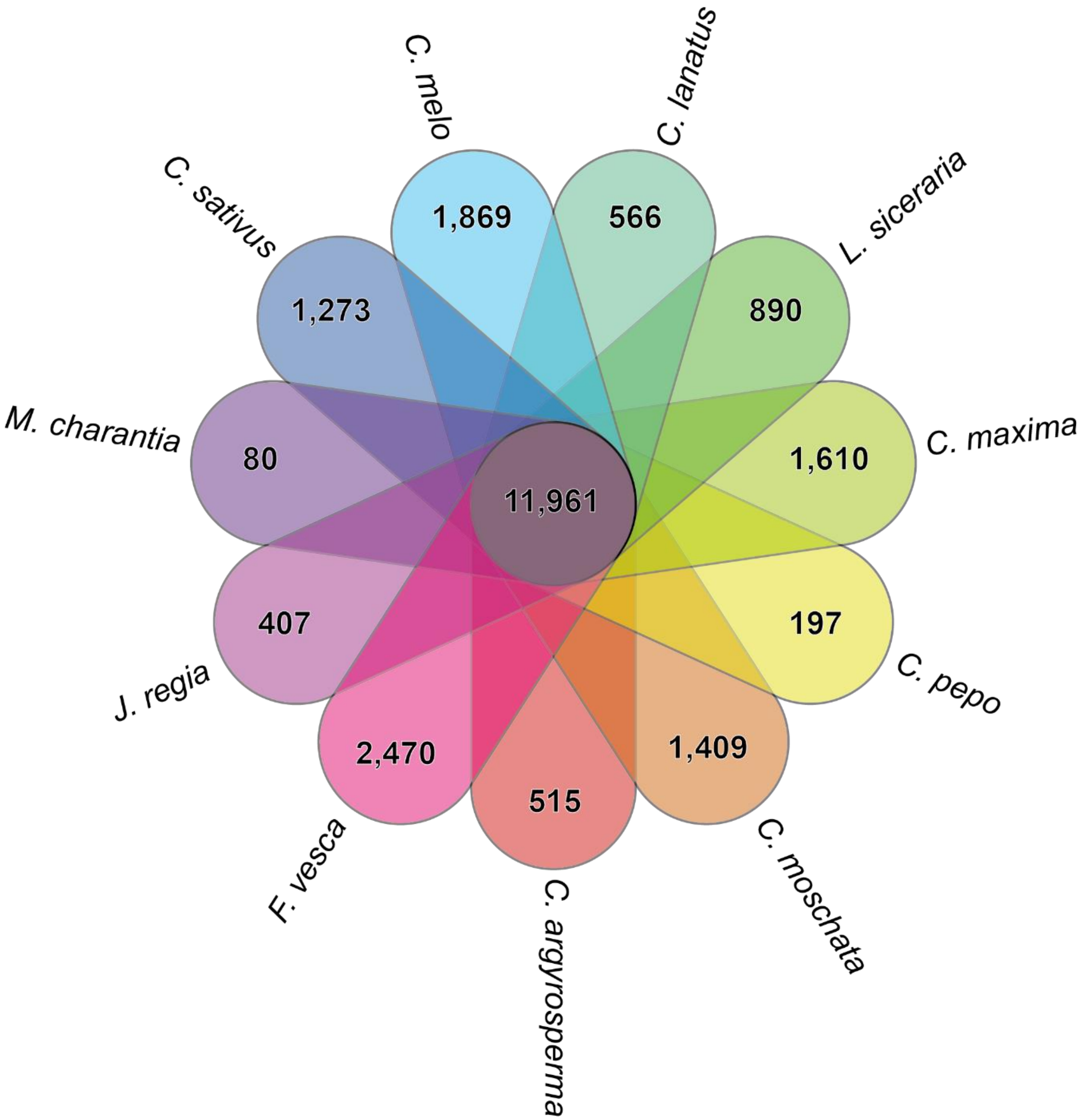
Phylogenetic analyses

We detected single-copy ortholog genes that were conserved in all the analyzed species, that is, gene families where every species had just one representative gene. The aligned sequences of the single-copy orthologs were concatenated and used to obtain a species phylogeny with PhyML (Guindon *et al.*, 2010). We used the Akaike

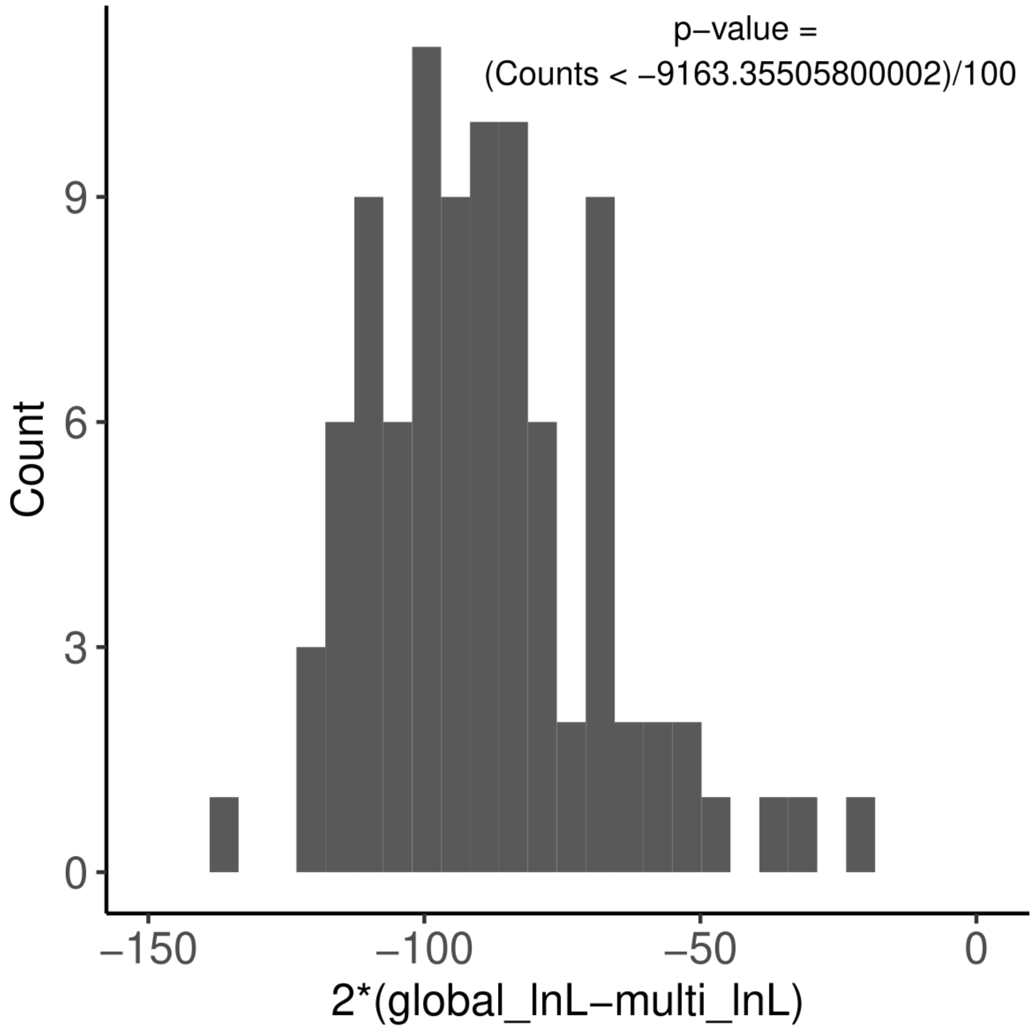
Information Criterion within SMS (Lefort *et al.*, 2017) to determine the best amino acid substitution model for our sequence alignment: JTT matrix (Jones *et al.*, 1992) with a gamma shape parameter (G = 0.906) of four categories and a proportion of invariant sites (I = 0.081). We used BioNJ for the initial tree with an SPR search for tree topology and performed aLRT to test branch support at each node (Ansimova and Gascuel, 2006). To obtain a dated phylogeny, we used a Bayesian Markov Chain Monte Carlo (MCMC) approach with approximate likelihood calculation, as implemented in mcmctree (Yang, 2007).

We incorporated the ages of *Protofagacea* (Herendeen *et al.*, 1995) and *Antiquacupula* (Sims *et al.*, 1998) as a minimum date of divergence for the root of the tree (84 Mya), since they can be unambiguously assigned to the Fagales Order, thereby functioning as a minimum date of divergence between Cucurbitales and Fagales (Wikstrom *et al.*, 2001). We also used the estimated age of the origin of core eudicots (115 Mya) as a maximum time of divergence for the root of the tree (Chang *et al.*, 2004). We ran two independent MCMC analyses for 11,000,000 generations, sampling every 500 generations and specifying an initial burn-in of 1,000,000 generations. We confirmed convergence of the two chains using Tracer v1.6 (Rambaut *et al.*, 2014). Both mcmctree trace files are available in Supplemental Data 3.

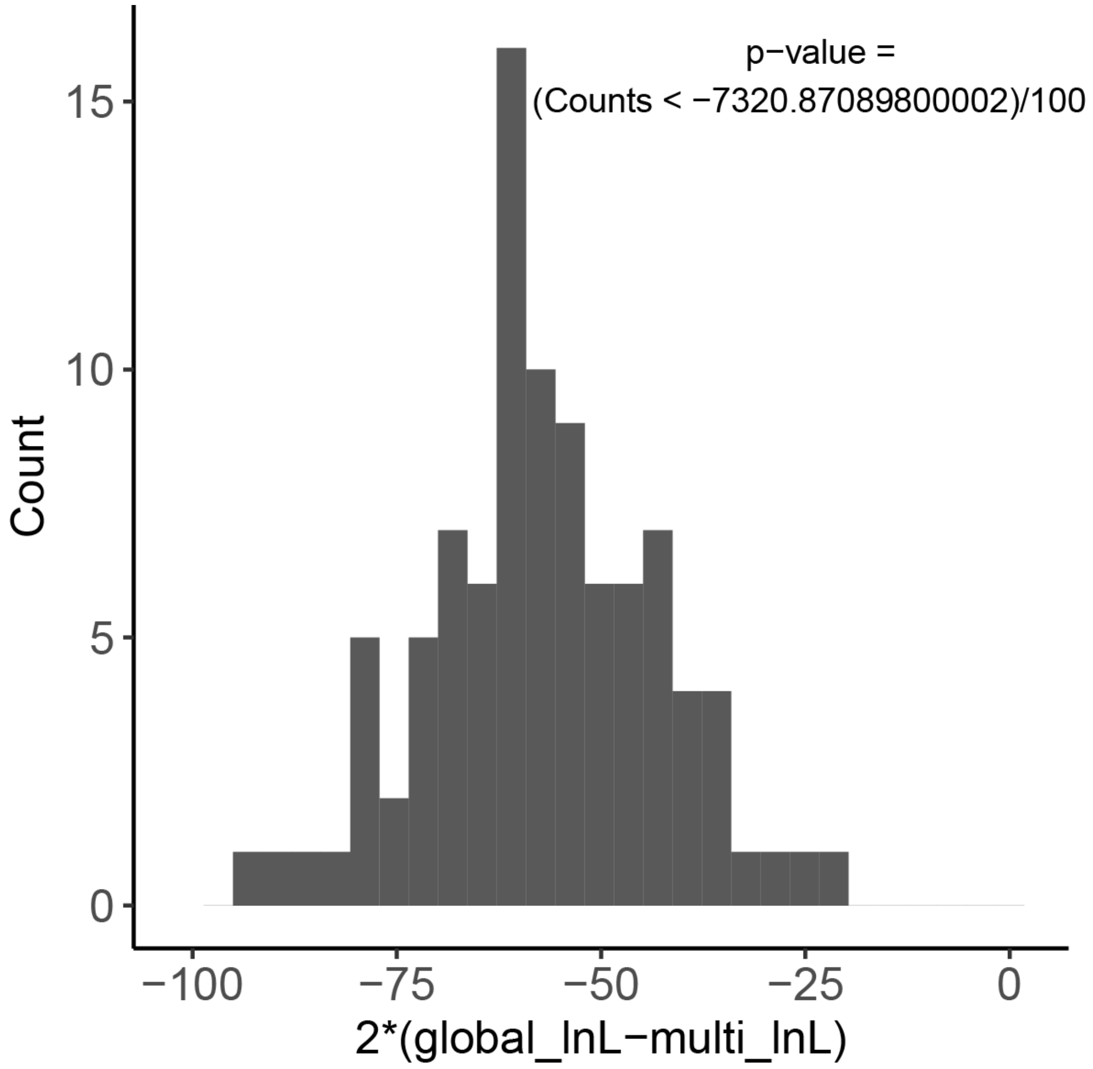
Supplemental Figures



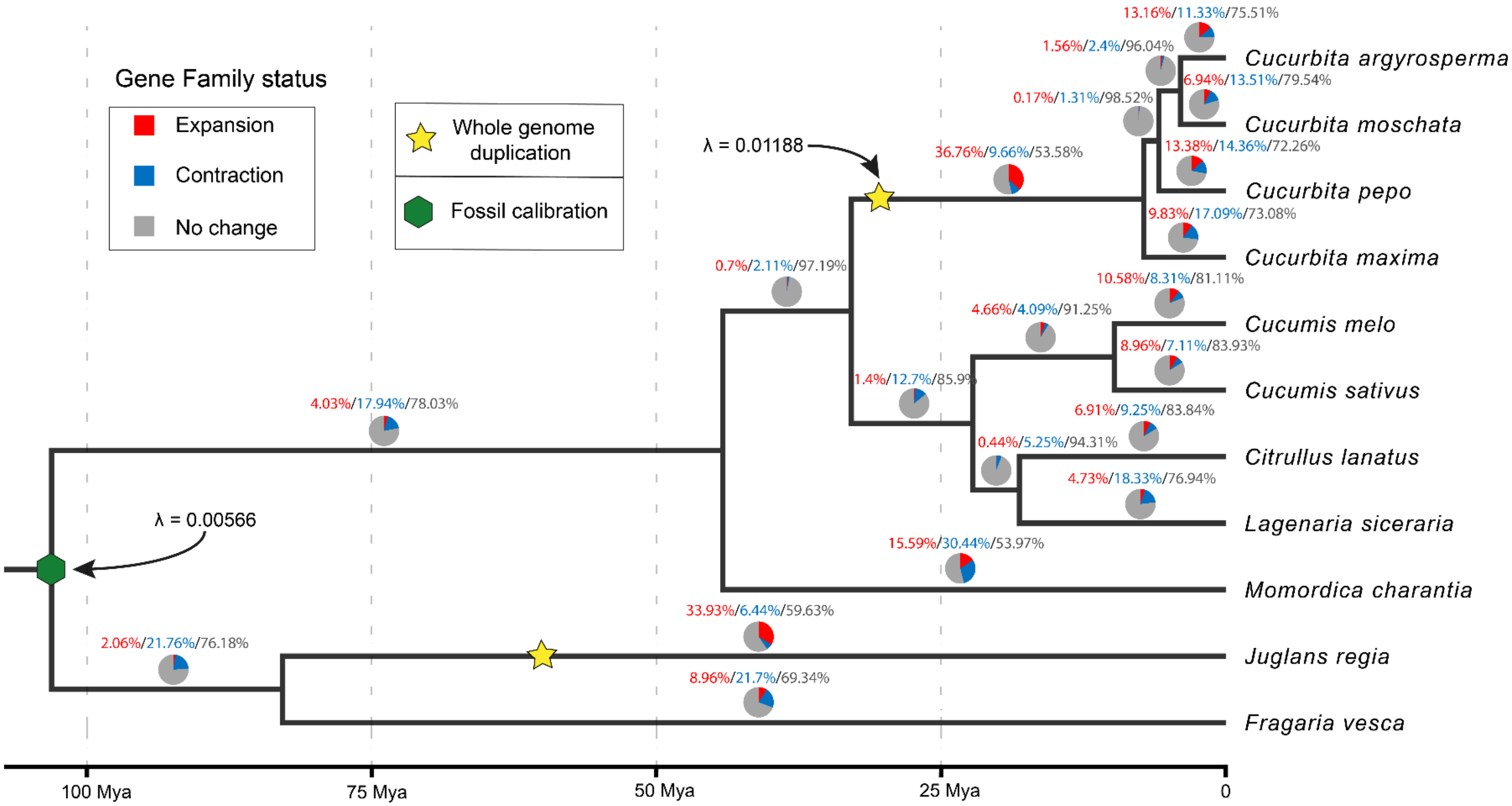
Supplemental Figure 1. Species-exclusive protein-coding gene families from all the analyzed taxa. The species-exclusive gene families were discarded from the gene family expansion/contraction analyses. The remaining 11,961 gene families (grey circle) were shared between at least two species and were used to calculate gene birth-death rates.



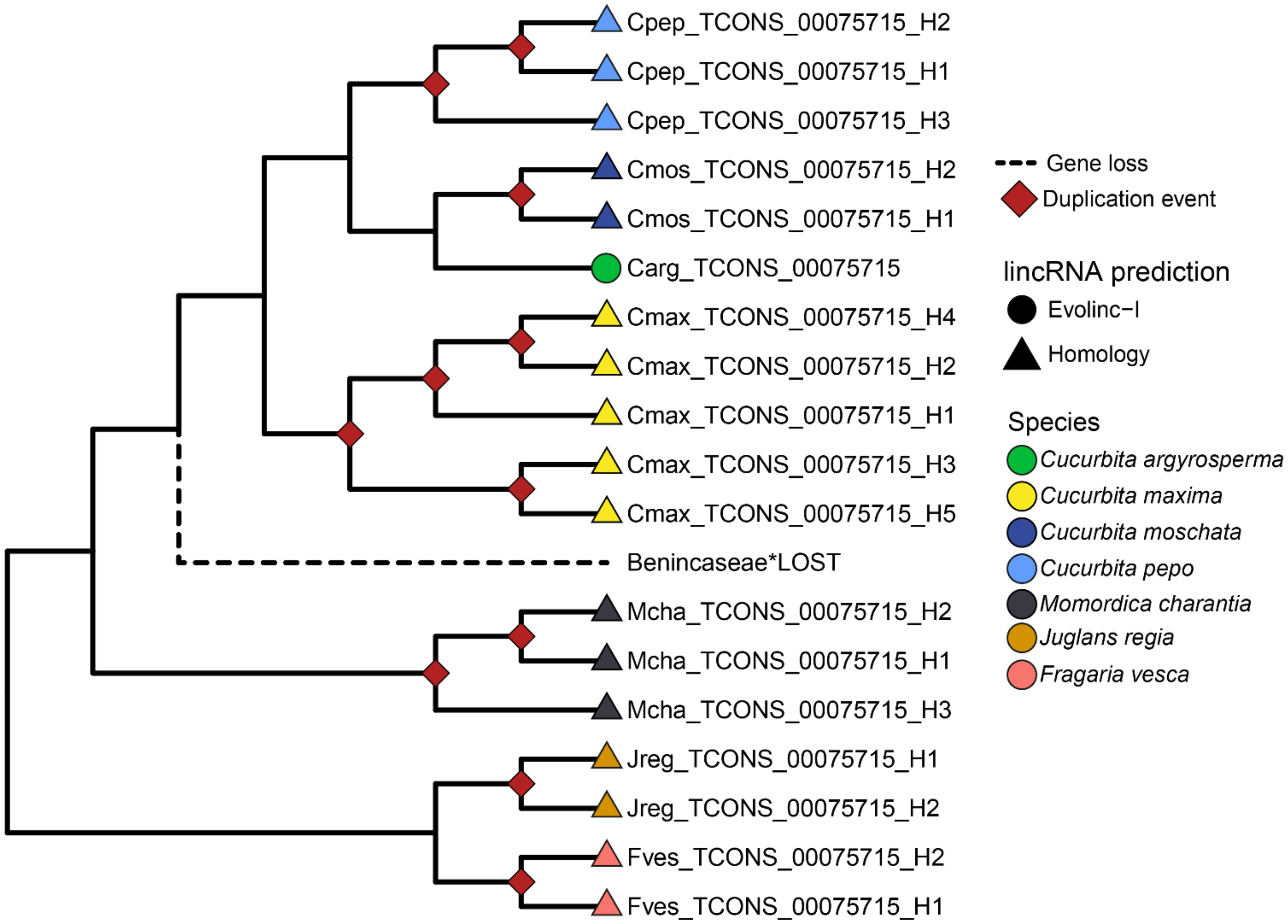
Supplemental Figure 2. Likelihood ratio test between a single gene birth-death rate (lambda) throughout the tree and a change in lambda within the *Cucurbita* genus. After calculating the observed likelihood ratio between a single lambda and two lambdas [2*((-198328.178461)-(-193746.504245))], 100 simulations were made to obtain a null distribution and asses if the observed likelihood ratio could be explained by chance. None of the simulated likelihood ratios were lower than the observed likelihood ratio (-9163.35), yielding a significantly low p-value ($p < 0.01$).



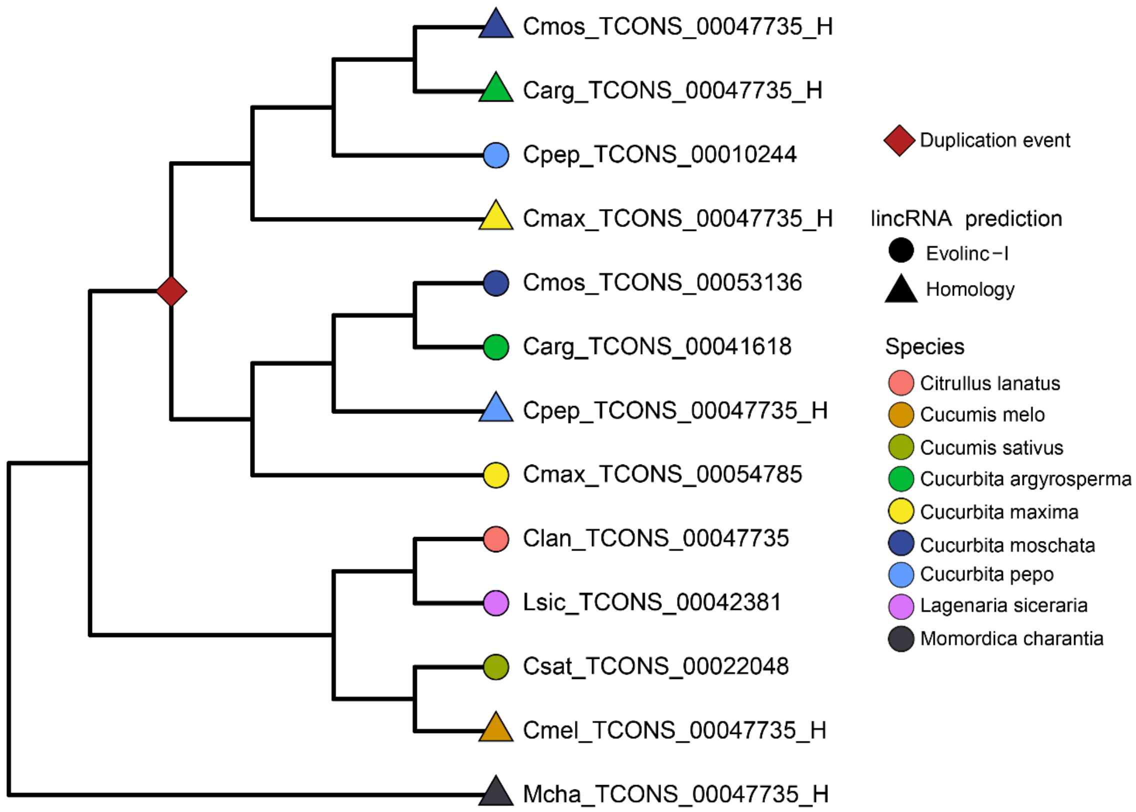
Supplemental Figure 3. Likelihood ratio test between a single gene birth-death rate (lambda) throughout the tree and a change in lambda within the *Cucurbita* genus using high-confidence protein-coding gene models. After calculating the observed likelihood ratio between a single lambda and two lambdas [2*((-169502.088320)-(-165841.652871))], 100 simulations were made to obtain a null distribution and asses if the observed likelihood ratio could be explained by chance. None of the simulated likelihood ratios were lower than the observed likelihood ratio (-7320.87), yielding a significantly low p-value (p < 0.01).



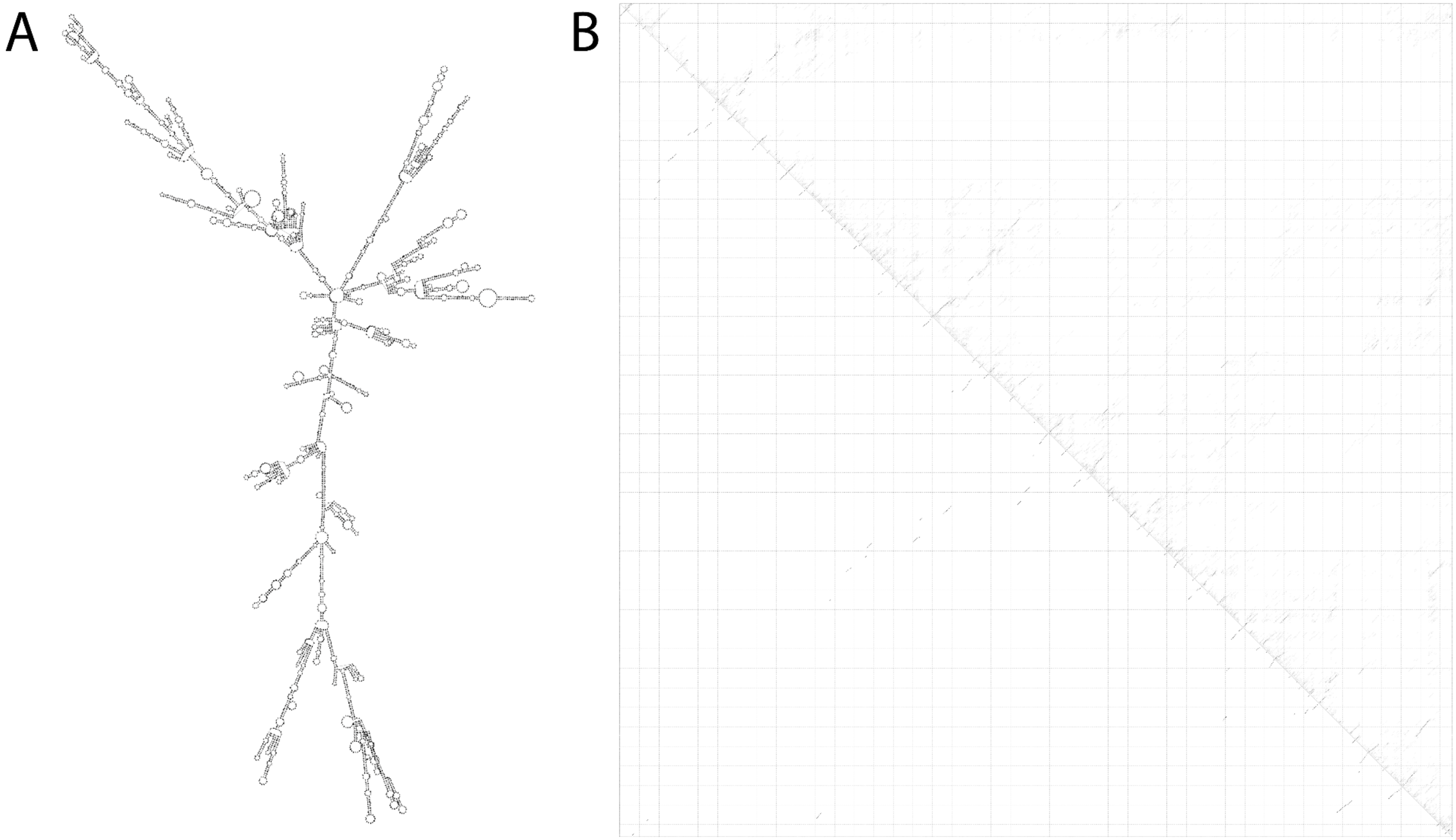
Supplemental Figure 4. Dated phylogeny of the Cucurbitaceae family with protein-coding gene family expansions and contractions per branch after discarding low-quality protein-coding gene models. Fossil evidence was used to calibrate the basal node of the tree (green hexagon). The pie charts and the percentages at every branch of the tree indicate whether a gene family expanded (red), contracted (blue) or remained the same size (gray). The yellow stars indicate the estimated ages of whole-genome duplication events within the phylogeny. The black arrows indicate the change in the basal gene birth/death rate of the phylogeny after the whole-genome duplication in *Cucurbita*.



Supplemental Figure 5. Pattern of lincRNA conservation in the outgroup species, *Juglans regia* and *Fragaria vesca*. Only a few lincRNA loci were conserved between *Cucurbita argyrosperma* and the outgroup species. All of these lincRNAs show complex evolutionary patterns and none of them are conserved in all the analyzed species.



Supplemental Figure 6. Duplication of lincRNA family associated to the whole-genome duplication in *Cucurbita* (red diamond). The lincRNA family was predicted using *C. lanatus* “Clan_TCONS_00047735” gene as query with Evolinc-II (Nelson *et al.*, 2017). Some lincRNAs within the family were independently predicted by homology using Evolinc-II (triangles in terminal nodes) and with transcriptomic evidence using Evolinc-I (circles in terminal nodes), further supporting the transcriptional activity of such genes. The colors at the terminal nodes indicate which gene belongs to each species.



Supplemental Figure 7. Secondary structure of the protein-coding transcript Carg19464 (trafficking protein particle complex subunit 11). A) Minimum free energy (MFE) structural prediction. B) RNA base-pairing probability matrix showing the MFE structural prediction (below the diagonal) and all the possible suboptimal pairings (above the diagonal). Higher probabilities are represented as larger dots within the matrix.

Supplemental Tables

Supplemental Table 1. Transposable elements within the genome of *Cucurbita argyrosperma*, classified according to Wicker's classification system (Wicker et al., 2007).

Class	Subclass	Order	Number of elements	Number of bp	percentage (genome assembly)
DNA transposons	Subclass-I (autonomous)	TIR	787	856472	0.3743
		Crypton	405	205905	0.0899
	Subclass-II (autonomous)	Helitron	326	336888	0.1472
		Maverick	65	107728	0.0471
	Non-autonomous DNA transposons	MITE	32	17023	0.0074
Retrotransposons	Autonomous retrotransposons	LINE	1507	974219	0.4257
		LTR	32222	38316833	16.7458
		SINE	34	7161	0.0031
		DIRS	5683	7746666	3.3855
	Non-autonomous retrotransposons	LARD	42230	22917378	10.0157
		TRIM	3830	2868582	1.2536
	Unclassifiable retrotransposon	NA	273	231598	0.1012
Unclassifiable transposon	NA	NA	248	106628	0.0466
Unknown repetitive element	NA	NA	2572	3364279	1.4703
TOTAL	NA	NA	90214	78057360	34.1138

Supplemental Table 2. Assembly size and gene content of all the analyzed genomes.

Group	Species	Assembly size	Protein-coding genes	High-quality gene models*	Predicted lincRNAs
Cucurbita spp.	*Cucurbita argyrosperma*	229 Mbp	28,298	26,603	6,124
	Cucurbita moschata	270 Mbp	32,205	26,425	998
	Cucurbita maxima	271 Mbp	32,076	26,843	1,054
	Cucurbita pepo	263 Mbp	27,867	27,017	3,272
Other cucurbits (Cucurbitaceae)	*Cucumis sativus*	243 Mbp	26,682	NA	3,488
	Cucumis melo	375 Mbp	27,427	NA	153
	Citrullus lanatus	353 Mbp	23,440	21,244	5,549
	Lagenaria siceraria	313 Mbp	22,472	18,903	2,446
	Momordica charantia	285 Mbp	23,514	NA	1,062
Outgroups	*Juglans regia*	667 Mbp	32,496	30,994	NA
	Fragaria vesca	220 Mbp	28,588	23,700	NA

* Gene predictions with Annotation Edit Distances lower than 0.5.

Supplemental Table 3. SRA accessions and sequenced organs of the RNAseq data used to predict lincRNA genes for each species using Evolinc-I (Nelson et al., 2017).

Species	leaves	roots	flowers	stems	fruits	tendrils	multiple pooled organs
Cucurbita argyrosperma	SRR7685407	SRR7685405	SRR7685400	SRR7685406	NA	SRR7685404	NA
Cucurbita maxima	SRR5504358 SRR5504359 SRR5504360	SRR5504353 SRR5504354 SRR5504355	NA	SRR5504356 SRR5504357	SRR5504361 SRR5504362 SRR5504363	NA	NA
Cucurbita moschata	SRR5504348 SRR5504349	SRR5504344 SRR5504343 SRR5504342	NA	SRR5504345 SRR5504346 SRR5504347	SRR5504350 SRR5504351 SRR5504352	NA	NA
Cucurbita pepo	NA	NA	NA	NA	NA	NA	SRR091276 SRR091277 SRR1182476 SRR1182477
Cucumis melo	NA	NA	NA	NA	NA	NA	SRR411100 SRR411102 SRR411103 SRR411104 SRR411105 SRR411106
Cucumis sativus	SRR351906	SRR351499	SRR351908 SRR351912	SRR351905	SRR351476 SRR351489 SRR351495	SRR351910 SRR351911	NA
Citrullus lanatus	SRR3156549	SRR3706796	SRR2033940 SRR2033941	SRR494474 SRR494479	SRR4046409 SRR4046416 SRR4046425	NA	NA
Lagenaria siceraria	SRR5590272	SRR5590273	SRR5590270	SRR5590274	SRR5590271	NA	NA

Supplemental Table 4. Dinucleotide content within a lincRNA and its protein-coding homolog.

	lincRNA (Carg_TCONS_00015392)		protein-coding homolog (Carg19464)	
Dinucleotide	Count	Frequency	Count	Frequency
AA	16	0.0740741	365	0.0856405
AC	12	0.0555556	204	0.0478649
AG	13	0.0601852	300	0.0703895
AT	14	0.0648148	331	0.0776631
CA	12	0.0555556	290	0.0680432
CC	6	0.0277778	226	0.0530267
CG	8	0.037037	86	0.0201783
CT	15	0.0694444	316	0.0741436
GA	17	0.0787037	323	0.075786
GC	8	0.037037	187	0.0438761
GG	8	0.037037	180	0.0422337
GT	18	0.0833333	208	0.0488034
TA	10	0.0462963	222	0.0520882
TC	14	0.0648148	301	0.0706241
TG	23	0.1064815	332	0.0778977
TT	22	0.1018519	391	0.091741

Supplemental References

Alverson, A. J., Wei, X., Rice, D. W., Stern, D. B., Barry, K., and Palmer, J. D. (2010). Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol. Biol. Evol.* **27**:1436–1448.

Ansimova, M., **and Gascuel, O.** (2006). Approximate Likelihood-Ratio Test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* **55:**539–552.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S. N., et al. (2011). The genome of *Theobroma cacao*. *Nat. Genet.* **43**:101–8.

Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biol.* **13**:R56.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**:211.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST plus : architecture and applications. *BMC Bioinformatics* **10**.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., and Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**:188–196.

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**:238.

Chang, C.-C., Chen, H.-L., Li, W.-H., and Chaw, S.-M. (2004). Dating the Monocot-Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes. *J. Mol. Evol.* **58**:424–441.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**:gkw955.

Edger, P. P., VanBuren, R., Colle, M., Poorten, T. J., Wai, C. M., Niederhuth, C. E., Alger, E. I., Ou, S., Acharya, C. B., Wang, J., et al. (2018). Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *Gigascience* **7**:1–7.

Evans, T., and Loose, M. (2015). AlignWise: a tool for identifying protein-coding sequence and correcting frame-shifts. *BMC Bioinformatics* **16**:376.

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., Gonzalez, V. M., Henaff, E., Camara, F., Cozzuto, L., Lowy, E., et al. (2012). The genome of melon (*Cucumis melo* L.). *Proc. Natl. Acad. Sci.* **109**:11872–11877.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**:644–652.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate Maximum-

Likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59:**307-321.

Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., Zheng, Y., Mao, L., Ren, Y., Wang, Z., et al. (2012). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**:51–58.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**:R7.

Herendeen, P. S., Crane, Peter, R., and Drinnan, A. N. (1995). Fagaceous flowers, fruits, and cupules from the Campanian (late Cretaceous) of central Georgia, USA. *Int. J. Plant Sci.* **156**:93–116.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W. J., Wang, X., Xie, B., Ni, P., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**:1275–1281.

Huang, X., and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Res.* **9**:868–877.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**:463–7.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8:**275-282.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**:1384–1395.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**:357–360.

Kistler, L., Newsom, L. A., Ryan, T. M., Clarke, A. C., Smith, B. D., and Perry, G. H. (2015). Gourds and squashes (*Cucurbita* spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proc. Natl. Acad. Sci.* **112**:15107–15112.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* **5**:59.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**:R25.

Lefort, V., Longueville, J-E., and Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, **34:**2422-2424.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv* **00**:3.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**:6494–6506.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**:955–964.

Nelson, A. D. L., Devisetty, U. K., Palos, K., Haug-Baltzell, A. K., Lyons, E., and Beilstein, M. A. (2017). Evolinc: A Tool for the Identification and Evolutionary Comparison of Long Intergenic Non-coding RNAs. *Front. Genet.* **8**:1–12.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**:290–295.

Rambaut, A., Suchard, M., Xie, D., and Drummond, A. (2014). Tracer 1.6, Available at http://tree.bio.ed.ac.uk/software/tracer/.

Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., and Arvestad, L. (2014). BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15**:281.

Schmieder, R., and Edwards, R. (2011). Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS One* **6**:e17288.

Sims, H. J., Herendeen, P. S., and Crane, Peter, R. (1998). New genus of fossil Fagaceae from the Santonian (Late Cretaceous) of central Georgia, U.S.A. *Int. J. Plant Sci.* **159**:391–404.

Smith, C. D., Edgar, R. C., Yandell, M. D., Smith, D. R., Celniker, S. E., Myers, E. W., and Karpen, G. H. (2007). Improved repeat identification and masking in Dipterans. *Gene* **389**:1–9.

Soorni, A., Haak, D., Zaitlin, D., and Bombarely, A. (2017). Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics* **18**:49.

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**:62.

Tae, H., Ryu, D., Sureshchandra, S., and Choi, J.-H. (2012). ESTclean: a cleaning tool for next-gen transcriptome shotgun sequencing. *BMC Bioinformatics* **13**:247.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**:796–815.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**:562–578.

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* Advance Access published January 18, 2017, doi:10.1101/gr.214270.116.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9**:e112963.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**:973–982.

Wikstrom, N., Savolainen, V., and Chase, M. W. (2001). Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. B Biol. Sci.* **268**:2211–

2220.

Xu, Q., Chen, L.-L., Ruan, X., Chen, D., Zhu, A., Chen, C., Bertrand, D., Jiao, W.-B., Hao, B.-H., Lyon, M. P., et al. (2012). The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**:59–66.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**:1586–1591.

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. (Sam) (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Sci. Rep.* **6**:31900.

ANEXO 3: MATERIAL SUPLEMENTARIO DEL CAPÍTULO 3

Supplementary information for:

The domestication patterns of *Cucurbita argyrosperma* are consistent with early migration events in Mesoamerica and general domestication syndromes in squashes.

Josué Barrera-Redondo, Guillermo Sánchez-de la Vega, Jonás A. Aguirre-Liguori, Gabriela Castellanos-Morales, Xitlali Aguirre-Dugua, Erika Aguirre-Planter, Yocelyn T. Gutiérrez-Guerrero, Maud Tenaillon, Salvador Montes-Hernández, Rafael Lira-Saade, Luis E. Eguiarte

Contact: fruns@unam.mx and rlira@unam.mx

This PDF file includes:

Supplemental Tables

S1: Assembly metrics of *C. argyrosperma* subsp. *sororia* genome.

S2: Information of 192 individuals sequenced with tGBS.

S3: Genetic diversity of each *C. argyrosperma* population.

S4: Strong candidate SNPs.

S5: Candidate genes under selection.

S6: GO enrichment results

S7: Cucurbitacin-related genes in *C. argyrosperma*.

Supplemental Figures

S1: Gene synteny dot plots between *Cucurbita* genomes.

S2-S21: Nucleotide synteny dot plots for each chromosome.

S22: LD decay in *C. argyrosperma*.

S23: Synteny dot plot between proteins and pseudogenes.

Supplemental Tables

Table S1. Assembly metrics of the reference genome of *Cucurbita argyrosperma* subsp. *sororia*.

Assembly size	253,587,946 bp
No. of contigs	828
Longest contig	4,976,248 bp
N50	1,323,288 bp
L50	58 contigs
No. of contigs > 1 kbp	828 (100.0%)
No. of contigs > 10 kbp	810 (97.8%)
No. of contigs > 100 kbp	292 (35.3%)
CG content	36.56%
Illumina read coverage	213x
PacBio read coverage	75.4x
No. of genes	31,452
Average gene size	3,269 bp
Complete BUSCOs	92.80%
Fragmented BUSCOs	1.20%
Missing BUSCOs	6.00%

Table S2. Information of 192 individuals sequenced using tGBS libraries, including population name, geographical coordinates and SRA accession. (within population names: W = wild, D = domesticated)

Individual ID	Population name	Population number	Taxon	Latitude	Longitude	SRA accession
129	Alamos, Sonora (Outgroup)	0	*Cucurbita moschata* (Outgroup)	27.02694	-108.93659	XXXXXX
131	Alamos, Sonora (Outgroup)	0	*Cucurbita moschata* (Outgroup)	27.02694	-108.93659	XXXXXX
M1_145	Alamos, Sonora (Outgroup)	0	*Cucurbita moschata* (Outgroup)	27.02694	-108.93659	XXXXXX
M2_146	Alamos, Sonora (Outgroup)	0	*Cucurbita moschata* (Outgroup)	27.02694	-108.93659	XXXXXX
M3_147	Alamos, Sonora (Outgroup)	0	*Cucurbita moschata* (Outgroup)	27.02694	-108.93659	XXXXXX
M4_148	Alamos, Sonora (Outgroup)	0	*Cucurbita moschata* (Outgroup)	27.02694	-108.93659	XXXXXX
S_CHIS14	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_CHIS18	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_CHIS19	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_CHIS22	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_CHIS8	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_CHISA	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_CHISB	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_CHISC	Jiquipilas, Chiapas (W)	1	*Cucurbita argyrosperma* subsp. *sororia*	16.597486	-93.626878	XXXXXX
S_GRO40	Ometepec, Guerrero (W)	2	*Cucurbita argyrosperma* subsp. *sororia*	16.688139	-98.406319	XXXXXX
S_GRO43	Ometepec, Guerrero (W)	2	*Cucurbita argyrosperma* subsp. *sororia*	16.688139	-98.406319	XXXXXX
S_GRO46	Ometepec, Guerrero (W)	2	*Cucurbita argyrosperma* subsp. *sororia*	16.688139	-98.406319	XXXXXX
S_GRO49	Ometepec, Guerrero (W)	2	*Cucurbita argyrosperma* subsp. *sororia*	16.688139	-98.406319	XXXXXX
S_GRO51	Ometepec, Guerrero (W)	2	*Cucurbita argyrosperma* subsp. *sororia*	16.688139	-98.406319	XXXXXX
S_OAX10	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX11	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX1	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX2	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX

S_OAX3	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX4	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX5	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX7	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX8	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_OAX9	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.918225	-97.076308	XXXXXX
S_MAC2	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.9528333	-97.0772778	XXXXXX
S_MAC3	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.9528333	-97.0772778	XXXXXX
S_MAC4	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.9528333	-97.0772778	XXXXXX
S_MAC5	Puerto Escondido, Oaxaca (W)	3	*Cucurbita argyrosperma* subsp. *sororia*	15.9528333	-97.0772778	XXXXXX
22	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.682	-104.333278	XXXXXX
23	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.682	-104.333278	XXXXXX
90	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.7019583	-104.204314	XXXXXX
53	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9005833	-104.160222	XXXXXX
54	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9005833	-104.160222	XXXXXX
65	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9005833	-104.160222	XXXXXX
40	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9005833	-104.160222	XXXXXX
26	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.6997222	-104.203056	XXXXXX
27	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.6997222	-104.203056	XXXXXX
28	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.6997222	-104.203056	XXXXXX
88	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.8753889	-104.072333	XXXXXX
45	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9123056	-104.116833	XXXXXX
49	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9123056	-104.116833	XXXXXX

51	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9123056	-104.116833	XXXXXX
55	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9123056	-104.116833	XXXXXX
59	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9600833	-104.037472	XXXXXX
44	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.9574722	-103.988389	XXXXXX
87	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.8753889	-104.072333	XXXXXX
85	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.8356111	-104.081417	XXXXXX
86	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.8356111	-104.081417	XXXXXX
80	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.6909722	-104.360833	XXXXXX
83	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.6909722	-104.360833	XXXXXX
84	Jalisco (W)	4	*Cucurbita argyrosperma* subsp. *sororia*	19.6909722	-104.360833	XXXXXX
97	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
98	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
99	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
100	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
101	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
102	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
103	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
104	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
105	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
106	Tlapehuala, Guerrero (D)	5	*Cucurbita argyrosperma* subsp. *argyrosperma*	18.2416667	-100.534722	XXXXXX
77	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.7019583	-104.204314	XXXXXX
79	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.7019583	-104.204314	XXXXXX
32	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.6997222	-104.203056	XXXXXX
34	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.6997222	-104.203056	XXXXXX
15	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8753889	-104.072333	XXXXXX
17	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8753889	-104.072333	XXXXXX
56	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.9600833	-104.037472	XXXXXX
70	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8714722	-104.217333	XXXXXX
73	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8714722	-104.217333	XXXXXX
39	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.9574722	-103.988389	XXXXXX

1	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8356111	-104.081417	XXXXXX
2	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8356111	-104.081417	XXXXXX
3	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8356111	-104.081417	XXXXXX
4	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8356111	-104.081417	XXXXXX
6	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8356111	-104.081417	XXXXXX
7	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8356111	-104.081417	XXXXXX
19	Jalisco (D)	6	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.8356111	-104.081417	XXXXXX
144	Badiraguato, Sinaloa (D)	7	*Cucurbita argyrosperma* subsp. *argyrosperma*	25.359173	-107.558408	XXXXXX
MTP14	Matlalapa, Guerrero (D)	8	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5971639	-99.4579917	XXXXXX
MTP1	Matlalapa, Guerrero (D)	8	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5971639	-99.4579917	XXXXXX
MTP2	Matlalapa, Guerrero (D)	8	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5971639	-99.4579917	XXXXXX
MTP3	Matlalapa, Guerrero (D)	8	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5971639	-99.4579917	XXXXXX
MTP4	Matlalapa, Guerrero (D)	8	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5971639	-99.4579917	XXXXXX
MTP5	Matlalapa, Guerrero (D)	8	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5971639	-99.4579917	XXXXXX
MTP9	Matlalapa, Guerrero (D)	8	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5971639	-99.4579917	XXXXXX
92	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH14	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH1	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH2	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH3	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH4_B01	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH4_D07	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH5	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH6	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
SAH7	Sahuayo, Michoacán (D)	9	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.0587194	-102.716233	XXXXXX
91	Salamanca, Guanajuato (D)	10	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.5205778	-101.190992	XXXXXX
SJI4_2	San José Iturbide, Guanajuato (D)	10	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.9988889	-100.385	XXXXXX
NAY2	Tepic, Nayarit (D)	11	*Cucurbita argyrosperma* subsp. *argyrosperma*	21.519956	-104.893423	XXXXXX
NAY3	Tepic, Nayarit (D)	11	*Cucurbita argyrosperma* subsp. *argyrosperma*	21.519956	-104.893423	XXXXXX

NAY4	Tepic, Nayarit (D)	11	*Cucurbita argyrosperma* subsp. *argyrosperma*	21.519956	-104.893423	XXXXXX
NAY5	Tepic, Nayarit (D)	11	*Cucurbita argyrosperma* subsp. *argyrosperma*	21.519956	-104.893423	XXXXXX
109	El Platanar, Sinaloa (feral)	12	feral individual	24.0303667	-106.432561	XXXXXX
110	El Platanar, Sinaloa (feral)	12	feral individual	24.0303667	-106.432561	XXXXXX
112	El Platanar, Sinaloa (feral)	12	feral individual	24.0303667	-106.432561	XXXXXX
113	El Platanar, Sinaloa (feral)	12	feral individual	24.0303667	-106.432561	XXXXXX
114	El Platanar, Sinaloa (feral)	12	feral individual	24.0303667	-106.432561	XXXXXX
93	El Platanar, Sinaloa (feral)	12	feral individual	24.0303667	-106.432561	XXXXXX
94	El Platanar, Sinaloa (feral)	12	feral individual	24.0303667	-106.432561	XXXXXX
95	Culiacán, Sinaloa (feral)	13	feral individual	24.817335	-107.416667	XXXXXX
96	Culiacán, Sinaloa (feral)	13	feral individual	24.817335	-107.416667	XXXXXX
107	Culiacán, Sinaloa (feral)	13	feral individual	24.817335	-107.416667	XXXXXX
108	Culiacán, Sinaloa (feral)	13	feral individual	24.817335	-107.416667	XXXXXX
121	Culiacán, Sinaloa (feral)	13	feral individual	24.817335	-107.416667	XXXXXX
115	Choix, Sinaloa (D)	14	*Cucurbita argyrosperma* subsp. *argyrosperma*	26.5967833	-108.335581	XXXXXX
116	Choix, Sinaloa (D)	14	*Cucurbita argyrosperma* subsp. *argyrosperma*	26.5967833	-108.335581	XXXXXX
117	Choix, Sinaloa (D)	14	*Cucurbita argyrosperma* subsp. *argyrosperma*	26.5967833	-108.335581	XXXXXX
118	Choix, Sinaloa (D)	14	*Cucurbita argyrosperma* subsp. *argyrosperma*	26.5967833	-108.335581	XXXXXX
119	Choix, Sinaloa (D)	14	*Cucurbita argyrosperma* subsp. *argyrosperma*	26.5967833	-108.335581	XXXXXX
138	Yecora, Sonora (D)	15	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.3720417	-108.926986	XXXXXX
139	Yecora, Sonora (D)	15	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.3720417	-108.926986	XXXXXX
140	Yecora, Sonora (D)	15	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.3720417	-108.926986	XXXXXX
141	Yecora, Sonora (D)	15	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.3720417	-108.926986	XXXXXX
142	Yecora, Sonora (D)	15	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.3720417	-108.926986	XXXXXX
143	Yecora, Sonora (D)	15	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.3720417	-108.926986	XXXXXX
128	Onavas, Sonora (feral)	16	feral individual	28.533333	-109.583333	XXXXXX
130	Onavas, Sonora (feral)	16	feral individual	28.533333	-109.583333	XXXXXX
132	Onavas, Sonora (feral)	16	feral individual	28.533333	-109.583333	XXXXXX
133	Onavas, Sonora (feral)	16	feral individual	28.533333	-109.583333	XXXXXX
134	Onavas, Sonora (feral)	16	feral individual	28.533333	-109.583333	XXXXXX

136	Onavas, Sonora (feral)	16	feral individual	28.533333	-109.583333	XXXXXX
137	Onavas, Sonora (feral)	16	feral individual	28.533333	-109.583333	XXXXXX
DGO1	Durango, Durango (D)	17	*Cucurbita argyrosperma* subsp. *argyrosperma*	24.066667	-104.583333	XXXXXX
TEH16	Tehuantepec, Oaxaca (D)	18	*Cucurbita argyrosperma* subsp. *argyrosperma*	16.3328306	-95.2330361	XXXXXX
TEH1	Tehuantepec, Oaxaca (D)	18	*Cucurbita argyrosperma* subsp. *argyrosperma*	16.3328306	-95.2330361	XXXXXX
TEH21	Tehuantepec, Oaxaca (D)	18	*Cucurbita argyrosperma* subsp. *argyrosperma*	16.3328306	-95.2330361	XXXXXX
TEH2	Tehuantepec, Oaxaca (D)	18	*Cucurbita argyrosperma* subsp. *argyrosperma*	16.3328306	-95.2330361	XXXXXX
TEH3_B07	Tehuantepec, Oaxaca (D)	18	*Cucurbita argyrosperma* subsp. *argyrosperma*	16.3328306	-95.2330361	XXXXXX
TEH3_E04	Tehuantepec, Oaxaca (D)	18	*Cucurbita argyrosperma* subsp. *argyrosperma*	16.3328306	-95.2330361	XXXXXX
TEH4	Tehuantepec, Oaxaca (D)	18	*Cucurbita argyrosperma* subsp. *argyrosperma*	16.3328306	-95.2330361	XXXXXX
122	Onavas, Sonora (D)	19	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.533333	-109.583333	XXXXXX
123	Onavas, Sonora (D)	19	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.533333	-109.583333	XXXXXX
124	Onavas, Sonora (D)	19	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.533333	-109.583333	XXXXXX
125	Onavas, Sonora (D)	19	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.533333	-109.583333	XXXXXX
126	Onavas, Sonora (D)	19	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.533333	-109.583333	XXXXXX
127	Onavas, Sonora (D)	19	*Cucurbita argyrosperma* subsp. *argyrosperma*	28.533333	-109.583333	XXXXXX
VER1	Tihuatlán, Veracruz (D)	20	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.7200667	-97.5395028	XXXXXX
VER2	Tihuatlán, Veracruz (D)	20	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.7200667	-97.5395028	XXXXXX
VER3	Tihuatlán, Veracruz (D)	20	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.7200667	-97.5395028	XXXXXX
VER4	Tihuatlán, Veracruz (D)	20	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.7200667	-97.5395028	XXXXXX
120	Tihuatlán, Veracruz (D)	20	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.7200667	-97.5395028	XXXXXX
PAL1	Palenque, Chiapas (D)	21	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5128139	-91.9877611	XXXXXX
PAL13	Palenque, Chiapas (D)	21	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5128139	-91.9877611	XXXXXX
PAL20	Palenque, Chiapas (D)	21	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5128139	-91.9877611	XXXXXX
PAL21	Palenque, Chiapas (D)	21	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5128139	-91.9877611	XXXXXX
PAL3	Palenque, Chiapas (D)	21	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5128139	-91.9877611	XXXXXX
PAL4	Palenque, Chiapas (D)	21	*Cucurbita argyrosperma* subsp. *argyrosperma*	17.5128139	-91.9877611	XXXXXX
SLP1	Tanquián, San Luis Potosí (D)	22	*Cucurbita argyrosperma* subsp. *argyrosperma*	22.1154861	-101.009331	XXXXXX
SLP2	Tanquián, San Luis Potosí (D)	22	*Cucurbita argyrosperma* subsp. *argyrosperma*	22.1154861	-101.009331	XXXXXX

SLP3	Tanquián, San Luis Potosí (D)	22	*Cucurbita argyrosperma* subsp. *argyrosperma*	22.1154861	-101.009331	XXXXXX
SLP4	Tanquián, San Luis Potosí (D)	22	*Cucurbita argyrosperma* subsp. *argyrosperma*	22.1154861	-101.009331	XXXXXX
SLP5	Tanquián, San Luis Potosí (D)	22	*Cucurbita argyrosperma* subsp. *argyrosperma*	22.1154861	-101.009331	XXXXXX
SLP6	Tanquián, San Luis Potosí (D)	22	*Cucurbita argyrosperma* subsp. *argyrosperma*	22.1154861	-101.009331	XXXXXX
CHAMP1	Champotón, Campeche (D)	23	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.5011556	-90.4613778	XXXXXX
CHAMP2	Champotón, Campeche (D)	23	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.5011556	-90.4613778	XXXXXX
CHAMP3	Champotón, Campeche (D)	23	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.5011556	-90.4613778	XXXXXX
CHAMP4	Champotón, Campeche (D)	23	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.5011556	-90.4613778	XXXXXX
CHAMP5	Champotón, Campeche (D)	23	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.5011556	-90.4613778	XXXXXX
MIXT1	Mixtepec, Oaxaca (D)	24	*Cucurbita argyrosperma* subsp. *argyrosperma*	15.95875	-97.0849167	XXXXXX
MIXT2	Mixtepec, Oaxaca (D)	24	*Cucurbita argyrosperma* subsp. *argyrosperma*	15.95875	-97.0849167	XXXXXX
MIXT3	Mixtepec, Oaxaca (D)	24	*Cucurbita argyrosperma* subsp. *argyrosperma*	15.95875	-97.0849167	XXXXXX
MIXT4	Mixtepec, Oaxaca (D)	24	*Cucurbita argyrosperma* subsp. *argyrosperma*	15.95875	-97.0849167	XXXXXX
MIXT5	Mixtepec, Oaxaca (D)	24	*Cucurbita argyrosperma* subsp. *argyrosperma*	15.95875	-97.0849167	XXXXXX
MIXT6	Mixtepec, Oaxaca (D)	24	*Cucurbita argyrosperma* subsp. *argyrosperma*	15.95875	-97.0849167	XXXXXX
EK1	Ek Balam, Yucatan (D)	25	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.9166667	-87.9166667	XXXXXX
EK2	Ek Balam, Yucatan (D)	25	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.9166667	-87.9166667	XXXXXX
EK3	Ek Balam, Yucatan (D)	25	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.9166667	-87.9166667	XXXXXX
EK4	Ek Balam, Yucatan (D)	25	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.9166667	-87.9166667	XXXXXX
EK5	Ek Balam, Yucatan (D)	25	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.9166667	-87.9166667	XXXXXX
EK6	Ek Balam, Yucatan (D)	25	*Cucurbita argyrosperma* subsp. *argyrosperma*	20.9166667	-87.9166667	XXXXXX
CHAN1	Chan Santa Cruz, Quintana Roo (D)	26	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.3670639	-88.3328917	XXXXXX
CHAN2	Chan Santa Cruz, Quintana Roo (D)	26	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.3670639	-88.3328917	XXXXXX
CHAN33	Chan Santa Cruz, Quintana Roo (D)	26	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.3670639	-88.3328917	XXXXXX

CHAN36	Chan Santa Cruz, Quintana Roo (D)	26	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.3670639	-88.3328917	XXXXXX
CHAN3	Chan Santa Cruz, Quintana Roo (D)	26	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.3670639	-88.3328917	XXXXXX
CHANT4	Chan Santa Cruz, Quintana Roo (D)	26	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.3670639	-88.3328917	XXXXXX
CHANT5	Chan Santa Cruz, Quintana Roo (D)	26	*Cucurbita argyrosperma* subsp. *argyrosperma*	19.3670639	-88.3328917	XXXXXX

Table S3. Genetic diversity of each wild, domesticated and feral population of *Cucurbita argyrosperma* using 2,861 unlinked SNPs (r2 < 0.25, MAF > 1%). (*N* = sample size, H_O = observed heterozygosity, H_E = expected heterozygosity, π = nucleotide diversity, F_{IS} = inbreeding coefficient, Var = variance) (within population names: W = wild, D = domesticated)

Population name	Population number	*N*	H_O (Var)	H_E (Var)	π (Var)	F_{IS} (Var)
Cucurbita moschata (Outgroup)	0	5	0.077 (0.042)	0.058 (0.019)	0.068 (0.029)	-0.018 (0.017)
Jiquipilas, Chiapas (W)	1	8	0.091 (0.039)	0.073 (0.020)	0.082 (0.027)	-0.020 (0.017)
Ometepec, Guerrero (W)	2	5	0.089 (0.038)	0.076 (0.021)	0.088 (0.030)	0.001 (0.027)
Puerto Escondido, Oaxaca (W)	3	14	0.094 (0.031)	0.079 (0.018)	0.085 (0.022)	-0.021 (0.017)
Jalisco (W)	4	17	0.115 (0.033)	0.099 (0.020)	0.106 (0.024)	-0.020 (0.020)
Tlapehuala, Guerrero (D)	5	10	0.098 (0.029)	0.086 (0.018)	0.093 (0.022)	-0.010 (0.023)
Jalisco (D)	6	13	0.099 (0.025)	0.090 (0.017)	0.096 (0.020)	-0.006 (0.022)
Badiraguato, Sinaloa (D)	7	1	0.095 (0.086)	0.047 (0.021)	0.095 (0.086)	0.000 (0.000)
Matlalapa, Guerrero (D)	8	7	0.092 (0.029)	0.081 (0.018)	0.090 (0.023)	-0.006 (0.017)
Sahuayo, Michoacán (D)	9	10	0.101 (0.027)	0.091 (0.018)	0.097 (0.021)	-0.008 (0.022)
Salamanca, Guanajuato (D)	10	1	0.087 (0.079)	0.043 (0.020)	0.087 (0.079)	0.000 (0.000)
Tepic, Nayarit (D)	11	4	0.089 (0.044)	0.067 (0.020)	0.083 (0.034)	-0.012 (0.012)
El Platanar, Sinaloa (feral)	12	4	0.107 (0.062)	0.074 (0.024)	0.097 (0.046)	-0.019 (0.018)
Culiacán, Sinaloa (feral)	13	3	0.099 (0.064)	0.066 (0.023)	0.094 (0.053)	-0.010 (0.014)
Choix, Sinaloa (D)	14	3	0.101 (0.060)	0.069 (0.023)	0.096 (0.050)	-0.008 (0.014)
Yecora, Sonora (D)	15	6	0.104 (0.040)	0.081 (0.020)	0.092 (0.027)	-0.025 (0.013)
Onavas, Sonora (feral)	16	7	0.101 (0.044)	0.077 (0.022)	0.088 (0.029)	-0.026 (0.015)
Durango, Durango (D)	17	1	0.076 (0.071)	0.038 (0.017)	0.076 (0.071)	0.000 (0.000)
Tehuantepec, Oaxaca (D)	18	7	0.085 (0.032)	0.071 (0.018)	0.079 (0.023)	-0.014 (0.014)
Onavas, Sonora (D)	19	6	0.089 (0.040)	0.067 (0.018)	0.078 (0.026)	-0.025 (0.013)
Tihuatlán, Veracruz (D)	20	5	0.088 (0.041)	0.069 (0.020)	0.082 (0.030)	-0.011 (0.018)
Palenque, Chiapas (D)	21	5	0.089 (0.035)	0.072 (0.019)	0.083 (0.026)	-0.013 (0.015)
Tanquián, San Luis Potosí (D)	22	6	0.090 (0.033)	0.074 (0.019)	0.084 (0.025)	-0.013 (0.013)
Champotón, Campeche (D)	23	5	0.096 (0.039)	0.076 (0.020)	0.090 (0.030)	-0.011 (0.016)
Mixtepec, Oaxaca (D)	24	6	0.101 (0.038)	0.081 (0.020)	0.092 (0.027)	-0.020 (0.016)
Ek Balam, Yucatan (D)	25	6	0.098 (0.041)	0.075 (0.020)	0.086 (0.027)	-0.026 (0.014)
Chan Santa Cruz, Quintana Roo (D)	26	7	0.089 (0.032)	0.073 (0.018)	0.081 (0.022)	-0.018 (0.016)

Table S4. Structural analysis performed 5,000 bp upstream and downstream of the 19 SNPs that were predicted as outliers by both BayeScEnv and PCAdapt. The structural analysis was performed by comparing the wild and domesticated reference genomes.

CHR	Candidate SNP position	Candidate gene (domesticated)	Candidate gene (wild)	Gene annotation	Description
1	1131161	Carg07327	Csor.00g196260	Uncharacterized protein	We found some small indels on the intergenic DNA and the introns, most probably neutral. We also found three non-synonymous mutations within the uncharacterized protein.
1	3254988	Carg12374	Csor.00g192590	serine/threonine-protein kinase PBL10	We found several large deletions upstream of PBL10 (2000 bp upstream of the candidate SNP), possibly disrupting its transcriptional activity. The gene structure of PBL10 is well conserved between subspecies, with a few mutations within the introns, most likely neutral. Four substitutions were also found within the exons, two of which are non-synonymous.
1	4158310	Carg06982	Csor.00g203760	HSP17.8 17.8 kDa class I heat shock protein	We detected two large indels downstream of HSP17.8 (4,700 bp from the candidate SNP) and a 3 bp insertion in the 5'UTR of HSP17.8. We also found several indels between HSP17.8 and an uncharacterized protein, one of whose size was 422 nt long, possibly disrupting an unknown regulatory sequence.
1	10918103	Carg_TCONS_00084066	NA	lincRNA	Several insertions/deletions are disrupting the lincRNAs around the candidate SNP.
3	4256305, 4256319	Carg19982	Csor.00g199760	UPF0503 protein At3g09070 (OCTOPUS)	There are two nonsynonymous substitutions and a 12 nt insertion within the open reading frame of the OCTOPUS gene in *C. argyrosperma* subsp. *sororia*, (~2 kb away from the candidate SNPs). However, this insertion may be a derived state in *C. argyrosperma* subsp. *sororia*, as the genomes of *C. argyrosperma* subsp. *argyrosperma*, *C. moschata*, and *C. pepo* subsp. *pepo* all lack that same 12 nt insertion (Sun 2017, Montero-Pau 2018, Barrera-Redondo 2019).
5	5237485	TE	TE	NA	The candidate SNP was localized within a transposable element.
4	16492565	Intergenic DNA	Intergenic DNA	NA	We detected a 711 nt deletion within the domesticated genome, in the intergenic DNA between Rab7 and a long-noncoding RNA
6	9854562	Carg_TCONS_00051014	NA	lincRNA	We found 4 mutations between the wild and the domesticated genome in a predicted long noncoding RNA downstream of the candidate SNP (<100 bp away from the candidate SNP), suggesting it is the possible target of the selective sweep.

7	1584624	Carg07889	Csor.00g113710	Phosphatidylinositol/ phosphatidylcholine transfer protein SFH9	Candidate SNP located within an intron. Other variants are seen within the intron, but they don't disrupt the structure of the gene. A 29 nt deletion was detected within one of the introns (~1,800 bp upstream of candidate SNP). A 1 nt indel was found on the 5'UTR of SFH9. A 20 nt insertion was also found between SFH9 and KINB2 (1,000 bp downstream of SNP).
7	3024258	Carg18571	Csor.00g110800	AGP18 Lysine-rich arabinogalactan protein 18	There are two insertions and one deletion in the first exon of AGP18 in the domesticated genome, disrupting its open reading frame.
8	7179125	Carg15929	Csor.00g237260	uncharacterized protein At5g19025	No relevant differences between reference genomes.
9	4725402	Intergenic DNA	Intergenic DNA	NA	We observed a 1,545 nt insertion within the domesticated genome, in an intergenic spacer between a long noncoding RNA and CPK9.
15	6673629	Carg26784	Csor.00g054230	Uncharacterized protein	There is a 48 nt insertion in the domesticated genome within an intron of the uncharacterized protein. We also found a 63 nt deletion in the domesticated genome upstream of a small heat shock protein.
16	3211595	Carg22232	Csor.00g242160	Ribonucleases P/MRP protein subunit POP1	We found a 229 nt insertion in the domesticated genome upstream of POP1.
16	6190329	Carg24812	NA	serine/threonine-protein kinase PBL23	There is a 915 nt insertion downstream of PBL23 in the domesticated genome. Both genomes become unalignable after 2500 nt upstream of the candidate SNP.
16	7224386, 7224386, 7224444	TE	TE	Possible uncharacterized protein	Three candidate SNPs found in a row (within 100 bp). Many large and small rearrangements throughout the candidate region. No gene was found on either reference genome, but the site containing the candidate SNPs closely resembles an uncharacterized protein predicted in *C. maxima* (93.16% identity, 91% query cover), possibly a transposon protein.

Table S5. *Cucurbita argyrosperma* genes containing a candidate SNP (predicted by either BayeScEnv or PCAdapt) within their inner structure (introns, exons, UTRs). (AED = Annotation Edit Distance; GO ID = Gene Ontology ID)

Gene ID (domesticated genome)	Gene ID (wild genome)	Chromosome location	Functional annotation against SwissProt	AED	GO ID
Carg00678	Csor.00g059490	Chr03	Similar to CSI1 Protein CELLULOSE SYNTHASE INTERACTIVE 1 (*Arabidopsis thaliana*)	0.06	GO:0005515
Carg00749	Csor.00g058770	Chr03	Similar to WDR5A COMPASS-like H3K4 histone methylase component WDR5A (*Arabidopsis thaliana*)	0.01	GO:0005515
Carg00754	Csor.00g058710	Chr03	Similar to TUBA Tubulin alpha chain (*Prunus dulcis*)	0.07	GO:0003924
Carg00755	NA	Chr03	Protein of unknown function	0.16	
Carg00942	Csor.00g292650	Chr06	Similar to CTN Cactin (*Arabidopsis thaliana*)	0.14	GO:0005515
Carg01177	Csor.00g247900	Chr06	Protein of unknown function	0.13	GO:0071816
Carg01302	Csor.00g249070	Chr06	Similar to SSL10 Protein STRICTOSIDINE SYNTHASE-LIKE 10 (*Arabidopsis thaliana*)	0.13	GO:0009058, GO:0016844
Carg01309	Csor.00g249150	Chr06	Similar to ARF1 Auxin response factor 1 (*Arabidopsis thaliana*)	0.15	
Carg01823	Csor.00g164870	Chr04	Similar to PP2AA2 Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform (*Arabidopsis thaliana*)	0.19	GO:0005515
Carg02429	Csor.00g032070	Chr15	Protein of unknown function	0.11	
Carg02490	Csor.00g220720	Chr02	Similar to FZR1 Protein FIZZY-RELATED 1 (*Arabidopsis thaliana*)	0.17	
Carg02612	Csor.00g221980	Chr02	Similar to TIC100 Protein TIC 100 (*Arabidopsis thaliana*)	0.21	
Carg02896	Csor.00g176320	Chr09	Similar to AKT1 Potassium channel AKT1 (*Arabidopsis thaliana*)	0.14	GO:0005216, GO:0006811, GO:0016020, GO:0055085
Carg02996	NA	Chr09	Similar to At5g49980 Transport inhibitor response 1-like protein (*Arabidopsis thaliana*)	0.02	GO:0005515

Carg03224	Csor.00g101370	Chr20	Similar to DDB_G0284757 OTU domain-containing protein DDB_G0284757 (*Dictyostelium discoideum*)	0.31	
Carg03669	NA	Chr14	Similar to At3g10130 Heme-binding-like protein At3g10130, chloroplastic (*Arabidopsis thaliana*)	0.2	
Carg03798	Csor.00g009600	Chr08	Similar to At1g09760 U2 small nuclear ribonucleoprotein A' (*Arabidopsis thaliana*)	0.06	
Carg04098	Csor.00g278800	Chr19	Similar to GNTI Alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase (*Arabidopsis thaliana*)	0.23	GO:0006486, GO:0008375
Carg04189	Csor.00g196640	Chr01	Similar to RAD23B Ubiquitin receptor RAD23b (*Arabidopsis thaliana*)	0.16	GO:0003684, GO:0005515, GO:0005634, GO:0006289, GO:0043161
Carg04403	Csor.00g121120	Chr04	Similar to P4H7 Probable prolyl 4-hydroxylase 7 (*Arabidopsis thaliana*)	0.15	GO:0016491, GO:0055114
Carg04520	Csor.00g122340	Chr04	Similar to RBCMT Ribulose-1,5 bisphosphate carboxylase/oxygenase large subunit N-methyltransferase, chloroplastic (*Nicotiana tabacum*)	0.23	GO:0005515
Carg04587	Csor.00g123000	Chr04	Similar to At3g53190 Probable pectate lyase 12 (*Arabidopsis thaliana*)	0.06	
Carg04908	Csor.00g009170	Chr08	Similar to RPS15AE 40S ribosomal protein S15a-5 (*Arabidopsis thaliana*)	0.11	GO:0003735, GO:0005840, GO:0006412
Carg04919	Csor.00g009070	Chr08	Similar to U2AF65B Splicing factor U2af large subunit B (*Nicotiana plumbaginifolia*)	0.3	GO:0003676, GO:0003723, GO:0005634, GO:0006397
Carg04921	Csor.00g009020	Chr08	Similar to ATX2 Histone-lysine N-methyltransferase ATX2 (*Arabidopsis thaliana*)	0.23	GO:0005515, GO:0005634

Carg04961	Csor.00g008680	Chr08	Similar to CRK3 CDPK-related kinase 3 (*Arabidopsis thaliana*)	0.18	GO:0004672, GO:0005524, GO:0006468
Carg04969	Csor.00g008610	Chr08	Similar to CAT9 Cationic amino acid transporter 9, chloroplastic (*Arabidopsis thaliana*)	0.15	GO:0016020, GO:0022857, GO:0055085
Carg05198	Csor.00g110710	Chr07	Similar to MYB44 Transcription factor MYB44 (*Arabidopsis thaliana*)	0.01	GO:0003677
Carg05727	Csor.00g172100	Chr16	Similar to KIN7G Kinesin-like protein KIN-7G (*Arabidopsis thaliana*)	0.11	GO:0003777, GO:0005524, GO:0007018, GO:0008017
Carg05757	Csor.00g172420	Chr16	Similar to BGAL3 Beta-galactosidase 3 (*Arabidopsis thaliana*)	0.17	GO:0030246
Carg06107	Csor.00g118480	Chr04	Similar to OPR1 12-oxophytodienoate reductase 1 (*Arabidopsis thaliana*)	0.32	GO:0010181, GO:0016491, GO:0055114
Carg06628	Csor.00g046050	Chr18	Similar to NMT1 Phosphoethanolamine N-methyltransferase 1 (*Arabidopsis thaliana*)	0.2	GO:0008168
Carg06661	Csor.00g045750	Chr18	Protein of unknown function	0.15	
Carg06696	Csor.00g045440	Chr18	Similar to At5g11010 Polynucleotide 5'-hydroxyl-kinase NOL9 (*Arabidopsis thaliana*)	0.19	
Carg06792	Csor.00g044470	Chr18	Similar to MSP1 Protein MSP1 (*Saccharomyces cerevisiae* (strain ATCC 204508 / S288c))	0.22	GO:0005515, GO:0005524
Carg06849	Csor.00g043850	Chr18	Similar to SS3 Soluble starch synthase 3, chloroplastic/amyloplastic (*Solanum tuberosum*)	0.11	GO:2001070
Carg06968	Csor.00g203640	Chr01	Similar to ORRM6 Organelle RRM domain-containing protein 6, chloroplastic (*Arabidopsis thaliana*)	0.18	GO:0003676
Carg06997	Csor.00g080170	Chr01	Similar to IREH1 Probable serine/threonine protein kinase IREH1 (*Arabidopsis thaliana*)	0.08	GO:0004672, GO:0005524, GO:0006468
Carg07232	Csor.00g265760	Chr17	Similar to HULK3 Protein HUA2-LIKE 3 (*Arabidopsis thaliana*)	0.22	

Carg07327	Csor.00g196260	Chr01	Similar to dusA tRNA-dihydrouridine(20/20a) synthase (*Vibrio vulnificus* (strain CMCP6))	0.27	GO:0008033, GO:0017150, GO:0050660, GO:0055114
Carg07674	Csor.00g064770	Chr13	Similar to apaG Protein ApaG (*Magnetospirillum magneticum* (strain AMB-1 / ATCC 700264))	0.21	GO:0005515
Carg07889	Csor.00g113710	Chr07	Similar to SFH9 Phosphatidylinositol/phosphatidylcholine transfer protein SFH9 (*Arabidopsis thaliana*)	0.14	
Carg07954	Csor.00g113070	Chr07	Similar to BHLH121 Transcription factor bHLH121 (*Arabidopsis thaliana*)	0.03	GO:0046983
Carg08549	Csor.00g041110	Chr02	Similar to At4g10930 Uncharacterized protein At4g10930 (*Arabidopsis thaliana*)	0.19	
Carg09452	Csor.00g084200	Chr17	Similar to ALA4 Probable phospholipid-transporting ATPase 4 (*Arabidopsis thaliana*)	0.09	GO:0000166, GO:0000287, GO:0005524, GO:0015914, GO:0016021, GO:0140326
Carg09511	Csor.00g083590	Chr17	Similar to IAA27 Auxin-responsive protein IAA27 (Arabidopsis thaliana)	0.14	
Carg10718	Csor.00g080500	Chr01	Similar to FBL15 F-box/LRR-repeat protein 15 (*Arabidopsis thaliana*)	0.16	GO:0005515
Carg10909	Csor.00g085670	Chr05	Protein of unknown function	0.04	
Carg10970	Csor.00g236380	Chr08	Similar to trc Serine/threonine-protein kinase tricorner (*Drosophila pseudoobscura pseudoobscura*)	0.14	GO:0004672, GO:0004674, GO:0005524, GO:0006468
Carg11153	Csor.00g150050	Chr14	Protein of unknown function	0.07	
Carg11621	Csor.00g072250	Chr02	Similar to Zeaxanthin epoxidase, chloroplastic (*Prunus armeniaca*)	0.12	GO:0005515, GO:0009507, GO:0009540, GO:0009688, GO:0016020,

					GO:0055114, GO:0071949
Carg11936	Csor.00g166690	Chr05	Similar to PUB6 U-box domain-containing protein 6 (*Arabidopsis thaliana*)	0.1	GO:0004842, GO:0016567
Carg12108	Csor.00g006010	Chr05	Similar to ATAD1 ATPase family AAA domain-containing protein 1 (*Bos taurus*)	0.17	GO:0005524
Carg12374	Csor.00g192590	Chr01	Similar to PBL10 Probable serine/threonine-protein kinase PBL10 (*Arabidopsis thaliana*)	0.19	GO:0004672, GO:0006468
Carg12525	Csor.00g094660	Chr14	Similar to EMB1444 Transcription factor EMB1444 (*Arabidopsis thaliana*)	0.24	GO:0046983
Carg12589	Csor.00g095250	Chr14	Similar to AP2 Floral homeotic protein APETALA 2 (*Arabidopsis thaliana*)	0.07	GO:0003677, GO:0003700, GO:0006355
Carg12845	Csor.00g037030	Chr09	Similar to GDPDL4 Glycerophosphodiester phosphodiesterase GDPDL4 (*Arabidopsis thaliana*)	0.1	GO:0006629, GO:0008081
Carg13010	Csor.00g214890	Chr10	Similar to At4g29530 Thiamine phosphate phosphatase-like protein (*Arabidopsis thaliana*)	0.11	GO:0016791
Carg13344	Csor.00g042780	Chr07	Similar to CSTF77 Cleavage stimulation factor subunit 77 (*Arabidopsis thaliana*)	0.23	GO:0005515, GO:0005634, GO:0006397
Carg13413	Csor.00g005190	Chr08	Similar to Sacs Sacsin (*Mus musculus*)	0.06	
Carg13432	Csor.00g005000	Chr08	Similar to efr3b Protein EFR3 homolog B (*Danio rerio*)	0.16	
Carg13651	Csor.00g132700	Chr02	Similar to Unc45a Protein unc-45 homolog A (*Mus musculus*)	0.12	GO:0005515
Carg14212	Csor.00g024080	Chr04	Similar to SCAI Protein SCAI (*Homo sapiens*)	0.08	GO:0003714, GO:0006351
Carg14376	Csor.00g056580	Chr04	Similar to CLE10 CLAVATA3/ESR (CLE)-related protein 10 (*Arabidopsis thaliana*)	0.22	
Carg14512	Csor.00g157750	Chr18	Similar to MKP1 Protein-tyrosine-phosphatase MKP1 (*Arabidopsis thaliana*)	0.01	GO:0008138, GO:0016311
Carg14536	Csor.00g207860	Chr06	Similar to At5g03900 Uncharacterized protein At5g03900, chloroplastic (*Arabidopsis thaliana*)	0.18	

Carg14932	Csor.00g116960	Chr07	Similar to VPS54 Vacuolar protein sorting-associated protein 54, chloroplastic (*Arabidopsis thaliana*)	0.26	GO:0005515, GO:0008080, GO:0042147
Carg14976	Csor.00g116510	Chr07	Similar to ABCC2 ABC transporter C family member 2 (*Arabidopsis thaliana*)	0.88	GO:0005524, GO:0016021, GO:0016887, GO:0042626, GO:0055085
Carg15060	Csor.00g282910	Chr16	Similar to IQD1 Protein IQ-DOMAIN 1 (*Arabidopsis thaliana*)	0.09	GO:0005515
Carg15210	Csor.00g217960	Chr10	Protein of unknown function	0.33	
Carg15512	Csor.00g251220	Chr06	Similar to SEC23 Protein transport protein SEC23 (*Ustilago maydis* (strain 521 / FGSC 9021))	0.14	GO:0006886, GO:0006888, GO:0008270, GO:0030127
Carg15691	Csor.00g281360	Chr19	Similar to Os03g0733400 Zinc finger BED domain-containing protein RICESLEEPER 2 (*Oryza sativa* subsp. *japonica*)	0.11	GO:0003677, GO:0046983
Carg15904	Csor.00g236940	Chr08	Similar to SGS3 Protein SUPPRESSOR OF GENE SILENCING 3 homolog (*Oryza sativa* subsp. *indica*)	0.13	GO:0031047
Carg15929	Csor.00g237260	Chr08	Similar to At5g19025 Uncharacterized protein At5g19025 (*Arabidopsis thaliana*)	0.28	
Carg16055	Csor.00g112640	Chr07	Similar to AUL1 Auxilin-like protein 1 (*Arabidopsis thaliana*)	0.04	
Carg16433	NA	Chr12	Similar to Gtp-bp Signal recognition particle receptor subunit alpha homolog (*Drosophila melanogaster*)	0.08	GO:0003924, GO:0005047, GO:0005525, GO:0005785, GO:0006614, GO:0006886
Carg17189	Csor.00g139890	Chr01	Similar to FPP4 Filament-like plant protein 4 (*Arabidopsis thaliana*)	0.11	

Carg17222	Csor.00g304480	Chr18	Similar to CUT1 3-ketoacyl-CoA synthase 6 (*Arabidopsis thaliana*)	0.02	GO:0003824, GO:0006633, GO:0016020, GO:0016747
Carg17814	Csor.00g228870	Chr10	Protein of unknown function	0.21	
Carg17827	Csor.00g228730	Chr10	Similar to BSL2 Serine/threonine-protein phosphatase BSL2 (*Arabidopsis thaliana*)	0.1	GO:0004721, GO:0005515, GO:0009742, GO:0016787
Carg18146	Csor.00g156040	Chr11	Protein of unknown function	0.13	
Carg18171	Csor.00g211980	Chr11	Similar to AHA11 ATPase 11, plasma membrane-type (*Arabidopsis thaliana*)	0.11	GO:0008553, GO:0016021, GO:0120029
Carg18484	Csor.00g111760	Chr07	Similar to At1g04910 Uncharacterized protein At1g04910 (*Arabidopsis thaliana*)	0.25	
Carg18727	Csor.00g105010	Chr17	Similar to serinc Probable serine incorporator (*Nematostella vectensis*)	0.07	GO:0016020
Carg18786	Csor.00g231070	Chr08	Similar to CURT1C Protein CURVATURE THYLAKOID 1C, chloroplastic (*Arabidopsis thaliana*)	0.26	
Carg18944	Csor.00g084780	Chr17	Similar to ATL3 RING-H2 finger protein ATL3 (*Arabidopsis thaliana*)	0.01	
Carg19818	Csor.00g017820	Chr05	Similar to SPCC1672.07 U3 small nucleolar RNA-associated protein 21 homolog (*Schizosaccharomyces pombe* (strain 972 / ATCC 24843))	0.18	GO:0005515, GO:0006364, GO:0032040
Carg20078	Csor.00g227200	Chr09	Similar to ABCE2 ABC transporter E family member 2 (*Arabidopsis thaliana*)	0.12	GO:0005524, GO:0016887
Carg20623	Csor.00g016670	Chr13	Protein of unknown function	0.03	GO:0005515
Carg20889	Csor.00g273290	Chr06	Protein of unknown function	0.18	
Carg21397	Csor.00g161930	Chr13	Similar to AGD12 ADP-ribosylation factor GTPase-activating protein AGD12 (*Arabidopsis thaliana*)	0.16	GO:0005096

Carg21521	Csor.00g160720	Chr13	Similar to SNI1 Negative regulator of systemic acquired resistance SNI1 (*Arabidopsis thaliana*)	0.19	
Carg21706	Csor.00g203020	Chr01	Similar to DGK1 Diacylglycerol kinase 1 (*Arabidopsis thaliana*)	0.06	GO:0004143, GO:0007205, GO:0016301, GO:0035556
Carg22034	Csor.00g277150	Chr19	Similar to PDV2 Plastid division protein PDV2 (*Arabidopsis thaliana*)	0.12	
Carg22092	Csor.00g304880	Chr18	Similar to Arfrp1 ADP-ribosylation factor-related protein 1 (*Rattus norvegicus*)	0.09	GO:0005525
Carg22232	Csor.00g242160	Chr16	Similar to POP1 Ribonucleases P/MRP protein subunit POP1 (*Homo sapiens*)	0.1	
Carg22875	Csor.00g267800	Chr15	Similar to SPBC3E7.09 Uncharacterized protein slp1 (*Schizosaccharomyces pombe* (strain 972 / ATCC 24843))	0.11	
Carg22878	Csor.00g267780	Chr15	Similar to PTD Protein PARTING DANCERS (*Arabidopsis thaliana*)	0.32	
Carg22996	Csor.00g086420	Chr05	Similar to FLK Flowering locus K homology domain (*Arabidopsis thaliana*)	0.11	GO:0003723
Carg23167	Csor.00g003720	Chr08	Similar to CES101 G-type lectin S-receptor-like serine/threonine-protein kinase CES101 (*Arabidopsis thaliana*)	0.04	GO:0004672, GO:0004674, GO:0005524, GO:0006468
Carg23235	Csor.00g188450	Chr16	Similar to Lon protease homolog 2, peroxisomal (*Spinacia oleracea*)	0.12	GO:0004176, GO:0004252, GO:0005524, GO:0006508
Carg23389	Csor.00g206240	Chr06	Similar to nop12 Nucleolar protein 12 (*Schizosaccharomyces pombe* (strain 972 / ATCC 24843))	0.08	GO:0003676
Carg23772	Csor.00g220140	Chr04	Similar to SAC3A SAC3 family protein A (*Arabidopsis thaliana*)	0.14	

Carg23802	NA	Chr14	Similar to At1g06840 Probable LRR receptor-like serine/threonine-protein kinase At1g06840 (*Arabidopsis thaliana*)	0.19	GO:0005515
Carg24347	Csor.00g079020	Chr07	Similar to ITN1 Ankyrin repeat-containing protein ITN1 (*Arabidopsis thaliana*)	0.03	
Carg24693	Csor.00g076650	Chr01	Similar to At1g17220 Translation initiation factor IF-2, chloroplastic (*Arabidopsis thaliana*)	0.08	GO:0003743, GO:0003924, GO:0005525, GO:0006413
Carg24812	NA	Chr16	Similar to PBL23 Probable serine/threonine-protein kinase PBL23 (*Arabidopsis thaliana*)	0.07	GO:0004672, GO:0005524, GO:0006468
Carg24979	Csor.00g267730	Chr15	Similar to Cag_1601 UPF0301 protein Cag_1601 (*Chlorobium chlorochromatii* (strain CaD3))	0.29	
Carg25109	NA	Chr13	Protein of unknown function	0.23	
Carg25229	Csor.00g030470	Chr16	Protein of unknown function	0.05	
Carg25230	Csor.00g030480	Chr16	Similar to CLC-F Chloride channel protein CLC-f (*Arabidopsis thaliana*)	0.05	GO:0005247, GO:0006821, GO:0016020, GO:0055085
Carg25231	Csor.00g030490	Chr16	Similar to WRKY2 Probable WRKY transcription factor 2 (*Arabidopsis thaliana*)	0.03	GO:0003700, GO:0006355, GO:0043565
Carg25337	Csor.00g070030	Chr19	Similar to Glycerol-3-phosphate acyltransferase, chloroplastic (*Cucumis sativus*)	0.17	GO:0004366, GO:0006650, GO:0016746
Carg25546	Csor.00g002770	Chr17	Similar to TMKL1 Putative kinase-like protein TMKL1 (*Arabidopsis thaliana*)	0.3	GO:0004672, GO:0005515, GO:0006468
Carg25626	Csor.00g028980	Chr01	Similar to ASP3 Aspartate aminotransferase 3, chloroplastic (*Arabidopsis thaliana*)	0.14	GO:0009058, GO:0030170
Carg25639	Csor.00g029150	Chr01	Similar to OVA7 Serine--tRNA ligase, chloroplastic/mitochondrial (*Arabidopsis thaliana*)	0.12	GO:0000166, GO:0004812, GO:0004828,

					GO:0005524, GO:0006418, GO:0006434
Carg26216	Csor.00g010650	Chr15	Similar to FTSH11 ATP-dependent zinc metalloprotease FTSH 11, chloroplastic/mitochondrial (*Arabidopsis thaliana*)	0.27	GO:0004222, GO:0005524, GO:0006508, GO:0016020
Carg26378	Csor.00g260430	Chr11	Similar to DLO1 Protein DMR6-LIKE OXYGENASE 1 (*Arabidopsis thaliana*)	0.16	GO:0016491, GO:0055114
Carg26784	NA	Chr15	Protein of unknown function	0.02	
Carg26826	Csor.00g030720	Chr16	Similar to PA200 Proteasome activator subunit 4 (*Arabidopsis thaliana*)	0.14	
Carg26857	Csor.00g260990	Chr11	Similar to DDB_G0292320 Protein unc-50 homolog (*Dictyostelium discoideum*)	0.24	
Carg26907	Csor.00g220120	Chr04	Similar to maea Macrophage erythroblast attacher (*Danio rerio*)	0.14	
Carg27113	Csor.00g014920	Chr19	Similar to VPS13C Vacuolar protein sorting-associated protein 13C (*Homo sapiens*)	0.11	
Carg27299	Csor.00g267650	Chr15	Similar to SAC1 Phosphoinositide phosphatase SAC1 (*Arabidopsis thaliana*)	0.18	GO:0042578
Carg27622	Csor.00g103900	Chr20	Similar to SUMO2 Small ubiquitin-related modifier 2 (*Arabidopsis thaliana*)	0.32	

Table S6. Significantly enriched (p < 0.05) Gene Ontology terms for the 125 genes with candidate SNPs within their structure (introns, exons, UTRs).

	Biological Process	
GO ID	**Term**	***p*-value**
GO:0009688	Abscisic acid biosynthetic process	0.0055
GO:0006434	Seryl-tRNA aminoacylation	0.0111
GO:0071816	Tail-anchored membrane protein insertion into ER membrane	0.0111
GO:0009742	Brassinosteroid mediated signaling pathway	0.0274
GO:0042147	Retrograde transport, endosome to Golgi	0.0328
GO:0043161	Proteasome-mediated ubiquitin-dependent protein catabolic process	0.0328
GO:0007205	Protein kinase C-activating G protein-coupled receptor signaling pathway	0.0382
	Molecular Function	
GO ID	**Term**	***p*-value**
GO:0009540	Zeaxanthin epoxidase [overall] activity	0.0059
GO:0004828	Serine-tRNA ligase activity	0.0117
GO:0004366	Glycerol-3-phosphate O-acyltransferase activity	0.0117
GO:0004176	ATP-dependent peptidase activity	0.0117
GO:0005515	Protein binding	0.0148
GO:0003714	Transcription corepressor activity	0.0233
GO:0005047	Signal recognition particle binding	0.0291
GO:0004674	Protein serine/threonine kinase activity	0.0324
GO:0017150	tRNA dihydrouridine synthase activity	0.0348
GO:0004143	Diacylglycerol kinase activity	0.0405
GO:0016844	Strictosidine synthase activity	0.0405
GO:2001070	Starch binding	0.0405
	Cellular Component	
GO ID	**Term**	***p*-value**
GO:0005785	Signal recognition particle receptor complex	0.011
GO:0032040	Small-subunit processome	0.043

Table S7. Cucurbitacin-related gene orthologs of *C. argyrosperma* identified by a bidirectional BLAST against the genes reported by Shang *et al.* (2014) on *Cucumis sativus*. (AED = Annotation Edit Distance; GO ID = Gene Ontology ID)

Gene ID (domesticated genome)	**Gene ID (wild genome)**	**Chromosome location**	**Functional annotation against SwissProt**	**AED**	**GO ID**
Carg02215	Csor.00g312480	Chr15	Similar to 5MAT Malonyl-CoA:anthocyanidin 5-O-glucoside-6''-O-malonyltransferase (*Arabidopsis thaliana*)	0.11	GO:0016747
Carg02488	Csor.00g220700	Chr02	Similar to rpoB DNA-directed RNA polymerase subunit beta (*Cucumis sativus*)	0.08	GO:0003677, GO:0003899, GO:0006351, GO:0046983
Carg03795	NA	Chr08	Similar to CYP81E1 Isoflavone 2'-hydroxylase (*Glycyrrhiza echinata*)	0.01	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg03796	Csor.00g009580	Chr08	Similar to BAHD1 BAHD acyltransferase At5g47980 (*Arabidopsis thaliana*)	0.02	GO:0016747
Carg03797	Csor.00g009590	Chr08	Similar to CYP87A3 Cytochrome P450 87A3 (*Oryza sativa* subsp. *japonica*)	0.14	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg06672	Csor.00g045650	Chr18	Similar to CYP87A3 Cytochrome P450 87A3 (*Oryza sativa* subsp. *japonica*)	0.1	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg07313	Csor.00g196390	Chr01	Similar to CYP705A5 Cytochrome P450 705A5 (*Arabidopsis thaliana*)	0.12	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg08824	Csor.00g156780	Chr02	Similar to CYP88D6 Beta-amyrin 11-oxidase (*Glycyrrhiza uralensis*)	0.03	GO:0005506, GO:0016705,

					GO:0020037, GO:0055114
Carg11550	Csor.00g001510	Chr17	Similar to CYP89A9 Cytochrome P450 89A9 (*Arabidopsis thaliana*)	0	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg11551	NA	Chr17	Similar to CYP81E8 Cytochrome P450 81E8 (*Medicago truncatula*)	0.1	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg11552	Csor.00g001500	Chr17	Similar to CPQ Cucurbitadienol synthase (*Cucurbita pepo*)	0.16	GO:0016866
Carg14872	Csor.00g181690	Chr14	Similar to KAO1 Ent-kaurenoic acid oxidase 1 (*Arabidopsis thaliana*)	0.07	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg15241	Csor.00g217650	Chr10	Similar to CPR NADPH--cytochrome P450 reductase (*Catharanthus roseus*)	0.18	GO:0003958, GO:0010181, GO:0016491, GO:0055114
Carg15423	Csor.00g250380	Chr06	Similar to CYP51G1 Sterol 14-demethylase (*Arabidopsis thaliana*)	0.03	GO:0005506, GO:0016705, GO:0020037, GO:0055114
Carg16672	Csor.00g308950	Chr15	Similar to BHLH120 Transcription factor bHLH120 (*Arabidopsis thaliana*)	0.14	GO:0046983

Supplemental Figures

Figure S1. Synteny dot plots between the chromosome-level genome assemblies of *Cucurbita argyrosperma* subsp. *argyrosperma* and *Cucurbita argyrosperma* subsp. *sororia* against the reference genomes of *Cucurbita moschata* and *Cucurbita maxima* (Sun *et al*., 2017).

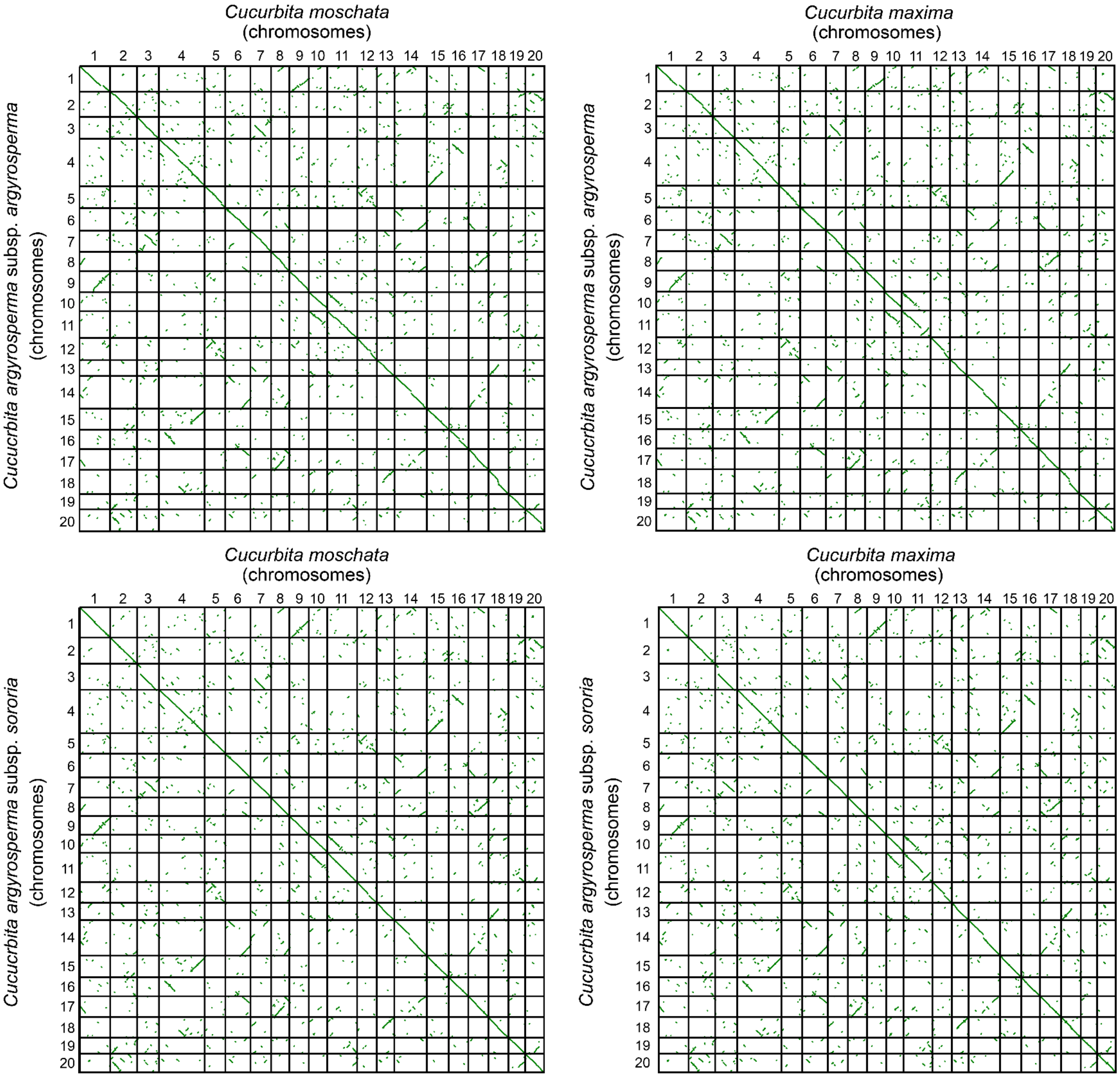


Figure S2. Synteny dot plots for chromosome 1 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*, (D) *C. moschata* and *C. maxima*. We used *C. moschata* and *C. maxima* as outgroups to identify the evolutionary orientation of the observed differences. We found a putative ~1,000,000 nt deletion in chromosome 1 of *C. argyrosperma* subsp. *argyrosperma* that occurred after diverging from *C. argyrosperma* subsp. *sororia*. We also found a putative inversion in chromosome 1 of *C. moschata* that occurred after diverging from *C. argyrosperma*.

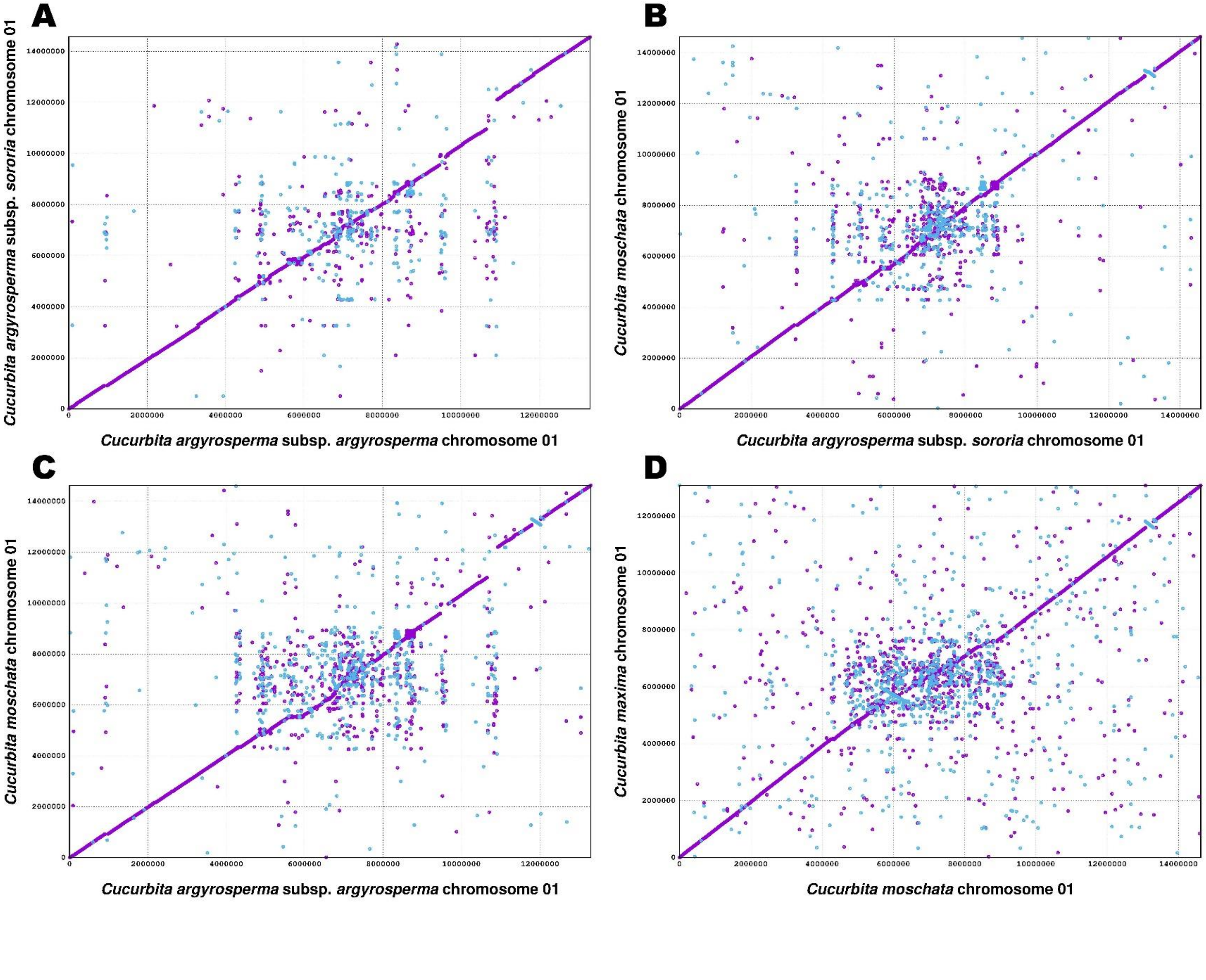


Figure S3. Synteny dot plots for chromosome 2 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found two putative deletions and one insertion in chromosome 2 of *C. argyrosperma* subsp. *argyrosperma* that occurred after diverging from *C. argyrosperma* subsp. *sororia*.

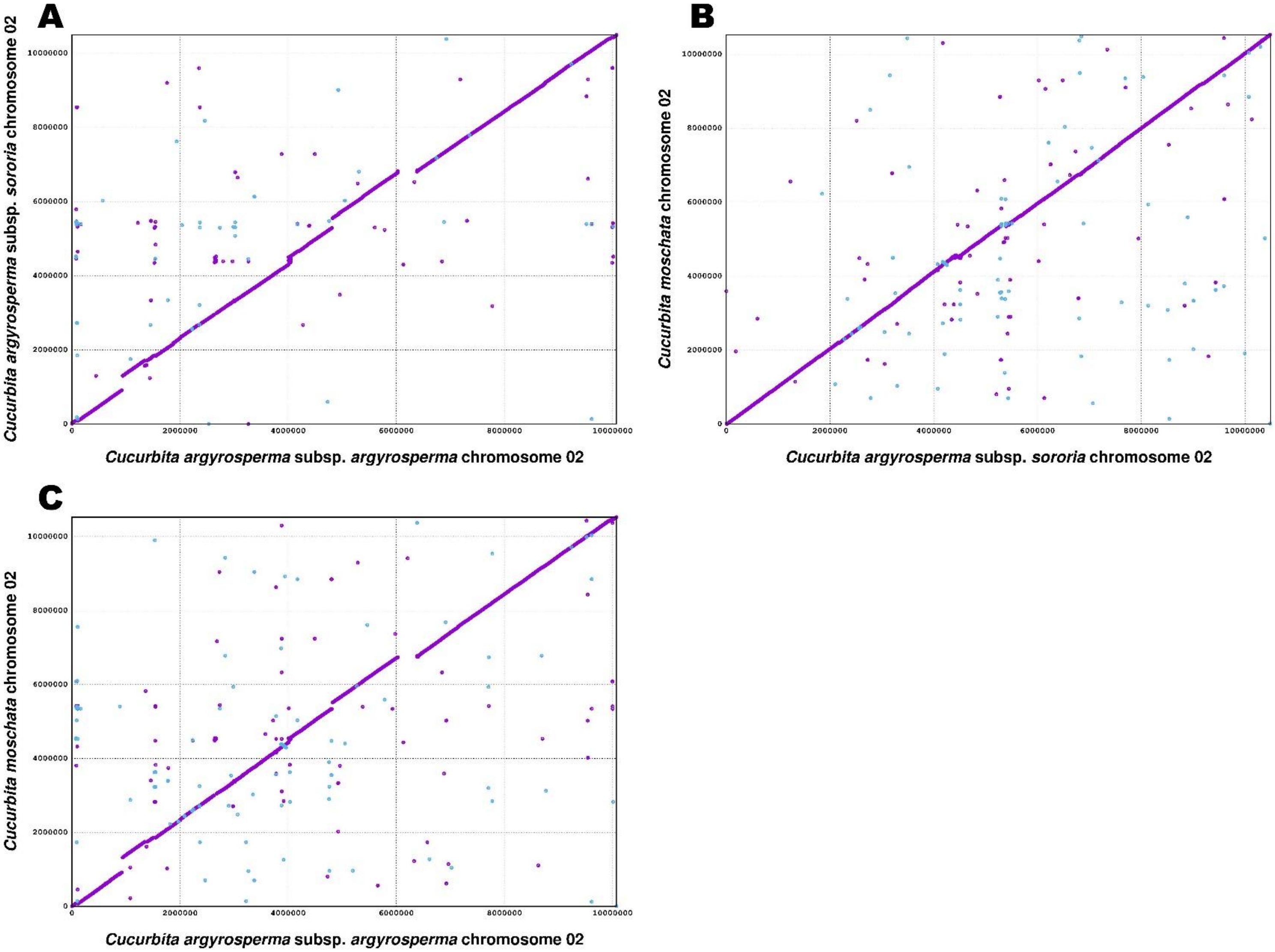


Figure S4. Synteny dot plots for chromosome 3 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a huge deletion (>1,000,000 nt) at the end of chromosome 3 in *C. argyrosperma* subsp. *sororia*, possibly due to an artifact during chromosome anchoring, as this sequence was found alongside the centromere of chromosome 15.

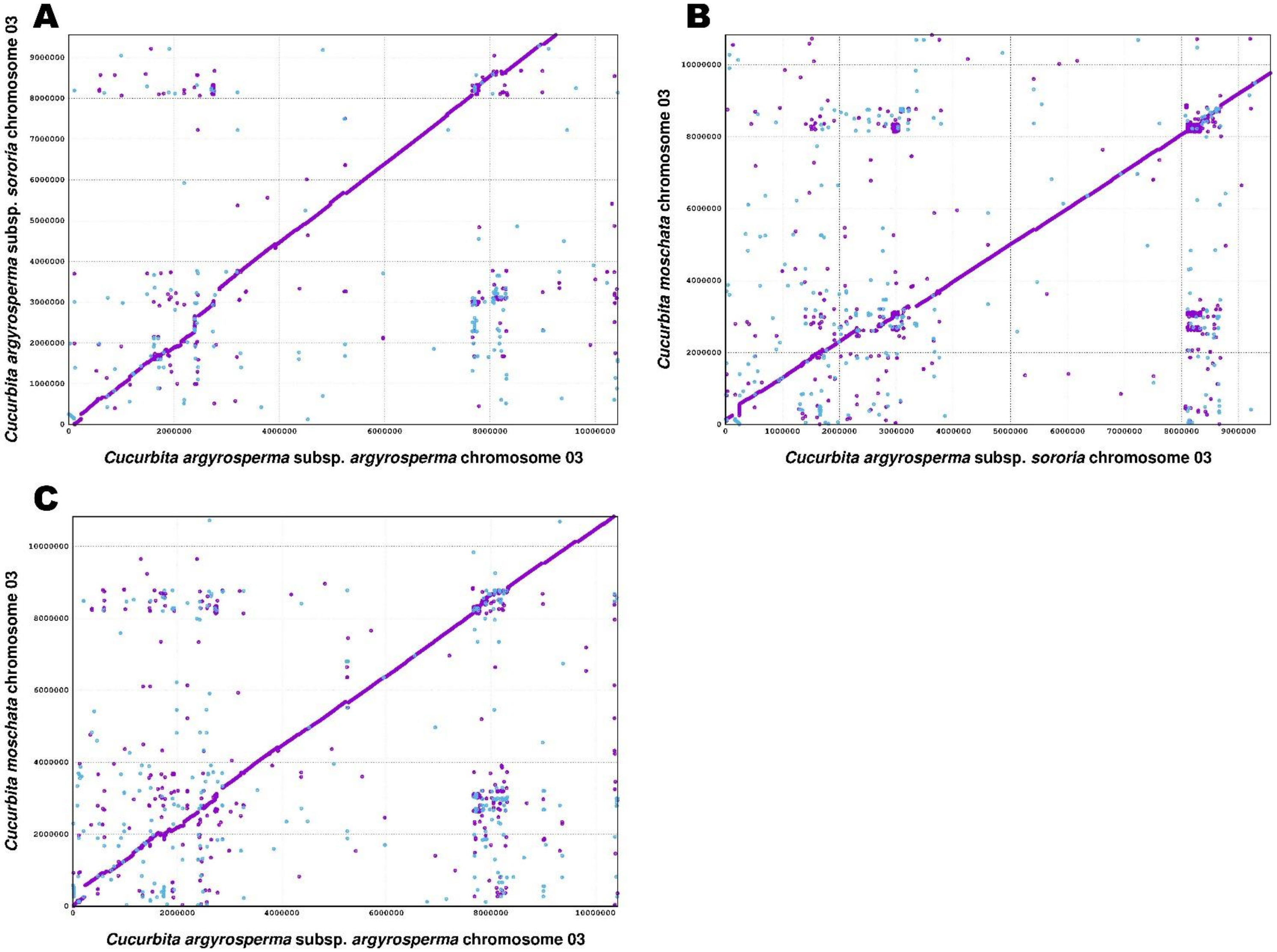


Figure S5. Synteny dot plots for chromosome 4 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a putative insertion (> 1,000,000 nt) in chromosome 4 of *C. argyrosperma* subsp. *argyrosperma* that occurred after diverging from *C. argyrosperma* subsp. *sororia*.

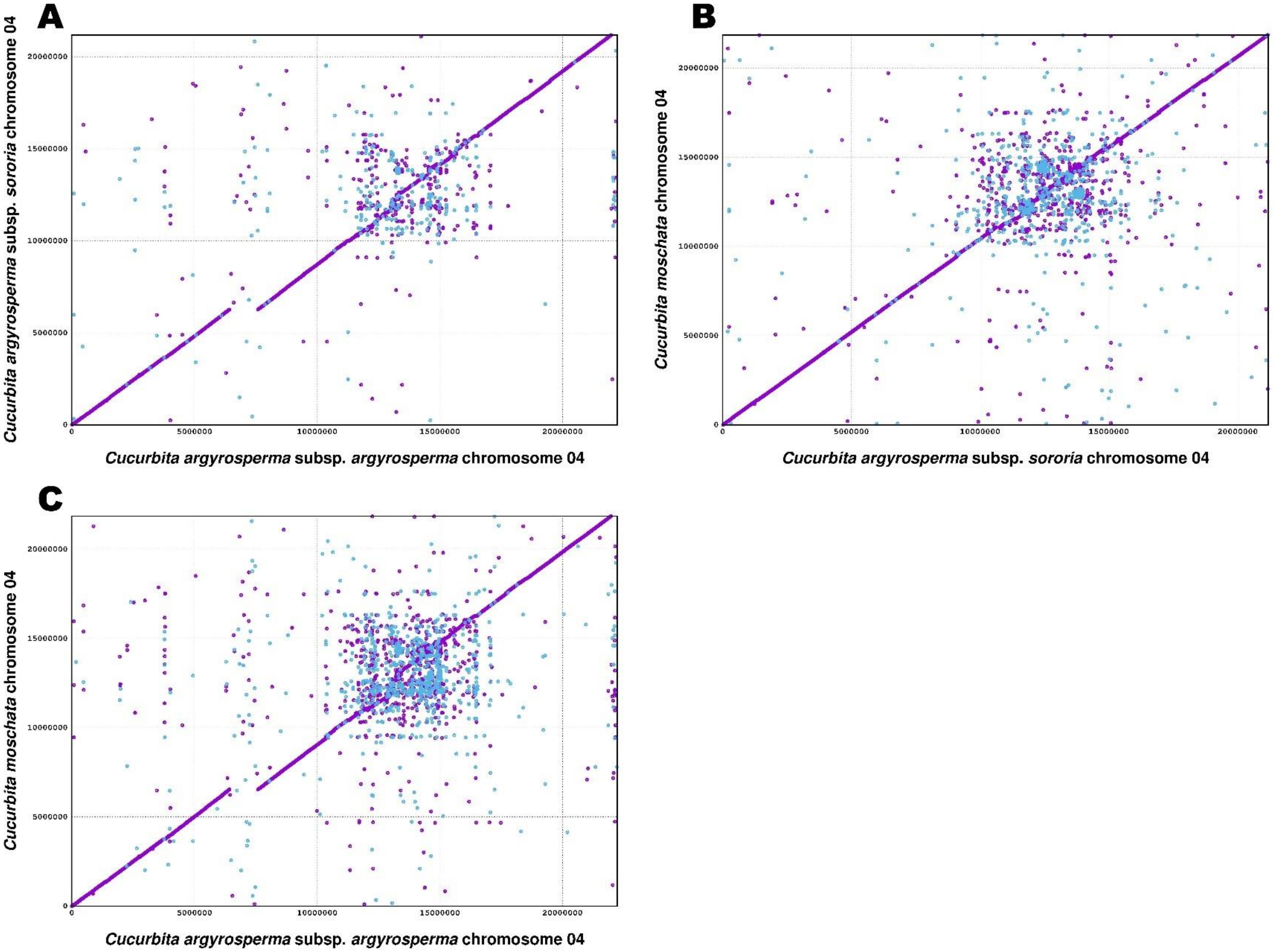


Figure S6. Synteny dot plots for chromosome 5 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found three insertions in chromosome 5 of *C. argyrosperma* subsp. *argyrosperma*, including one that is ~800,000 nt long, that occurred after diverging from *C. argyrosperma* subsp. *sororia*.

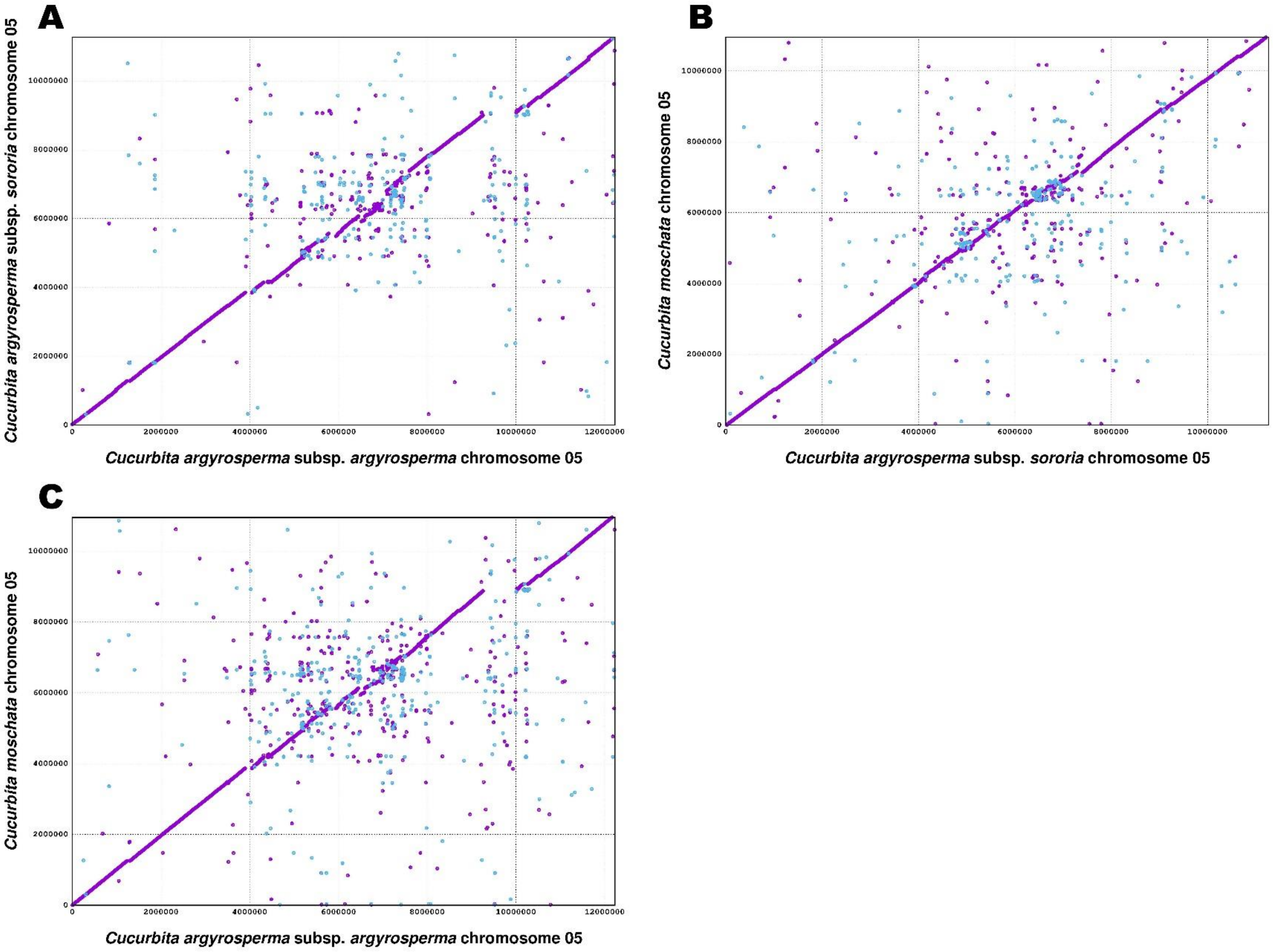


Figure S7. Synteny dot plots for chromosome 6 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a collapsed centromere assembly in *C. argyrosperma* subsp. *argyrosperma*, as well as a ~200,000 nt putative deletion that occurred after diverging from *C. argyrosperma* subsp. *sororia*.

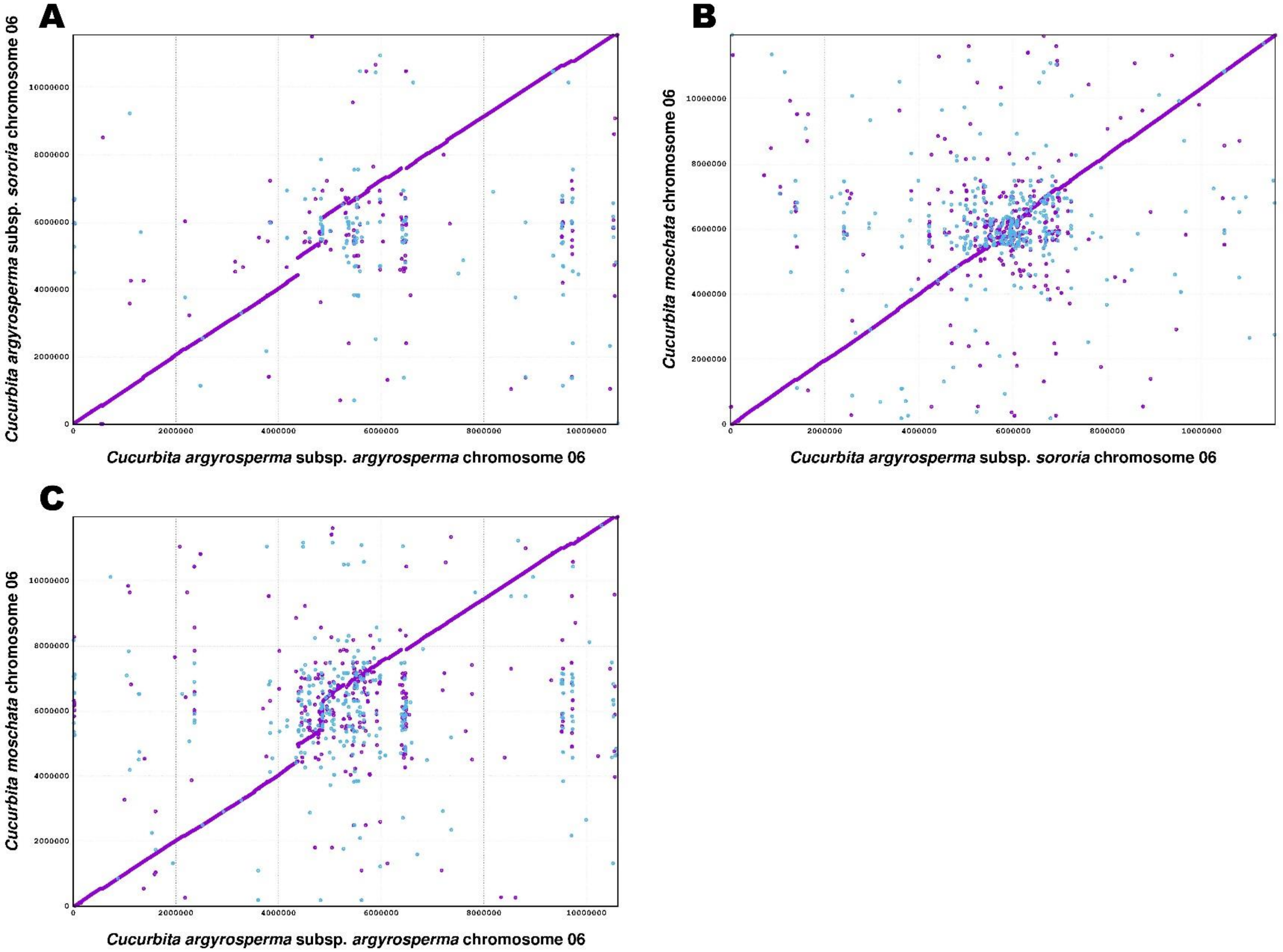


Figure S8. Synteny dot plots for chromosome 7 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We recovered a larger centromere in *C. argyrosperma* subsp. *sororia*, possibly due to a better assembly of the repetitive regions. We found a putative insertion (~400,000 nt) and a putative inversion (~200,000 nt) in chromosome 7 of *C. argyrosperma* subsp. *argyrosperma* that occurred after diverging from *C. argyrosperma* subsp. *sororia*. We found a candidate SNP within the ~200,000 nt inversion, making this putative rearrangement a possible candidate for selection during domestication.

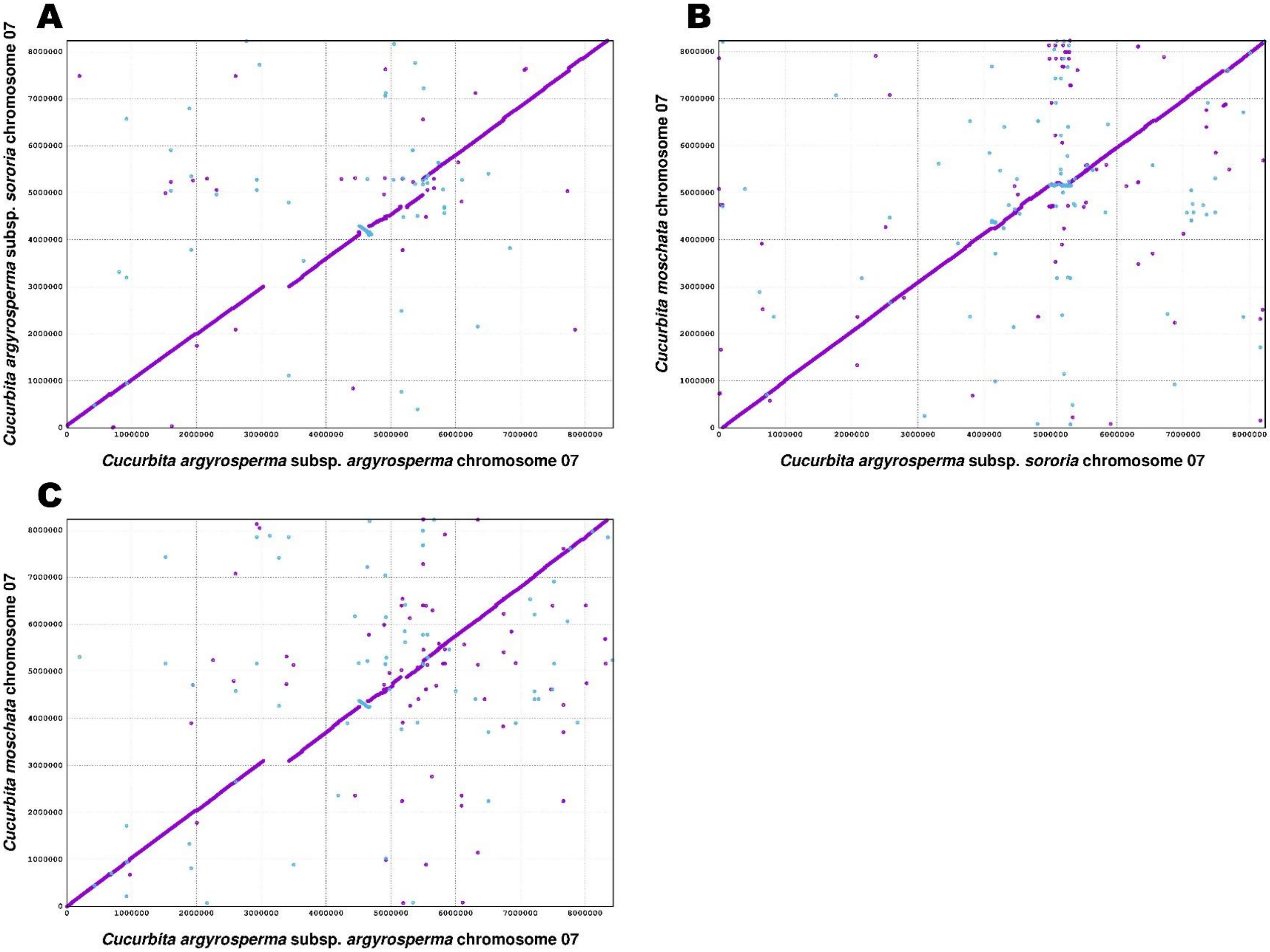


Figure S9. Synteny dot plots for chromosome 8 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a larger centromere in *C. argyrosperma* subsp. *sororia*, as well as other repetitive region upstream, possibly due to a better assembly of the repetitive regions. We also found one putative deletion and one putative insertion in chromosome 8 of *C. argyrosperma* subsp. *argyrosperma* that occurred after diverging from *C. argyrosperma* subsp. *sororia*.

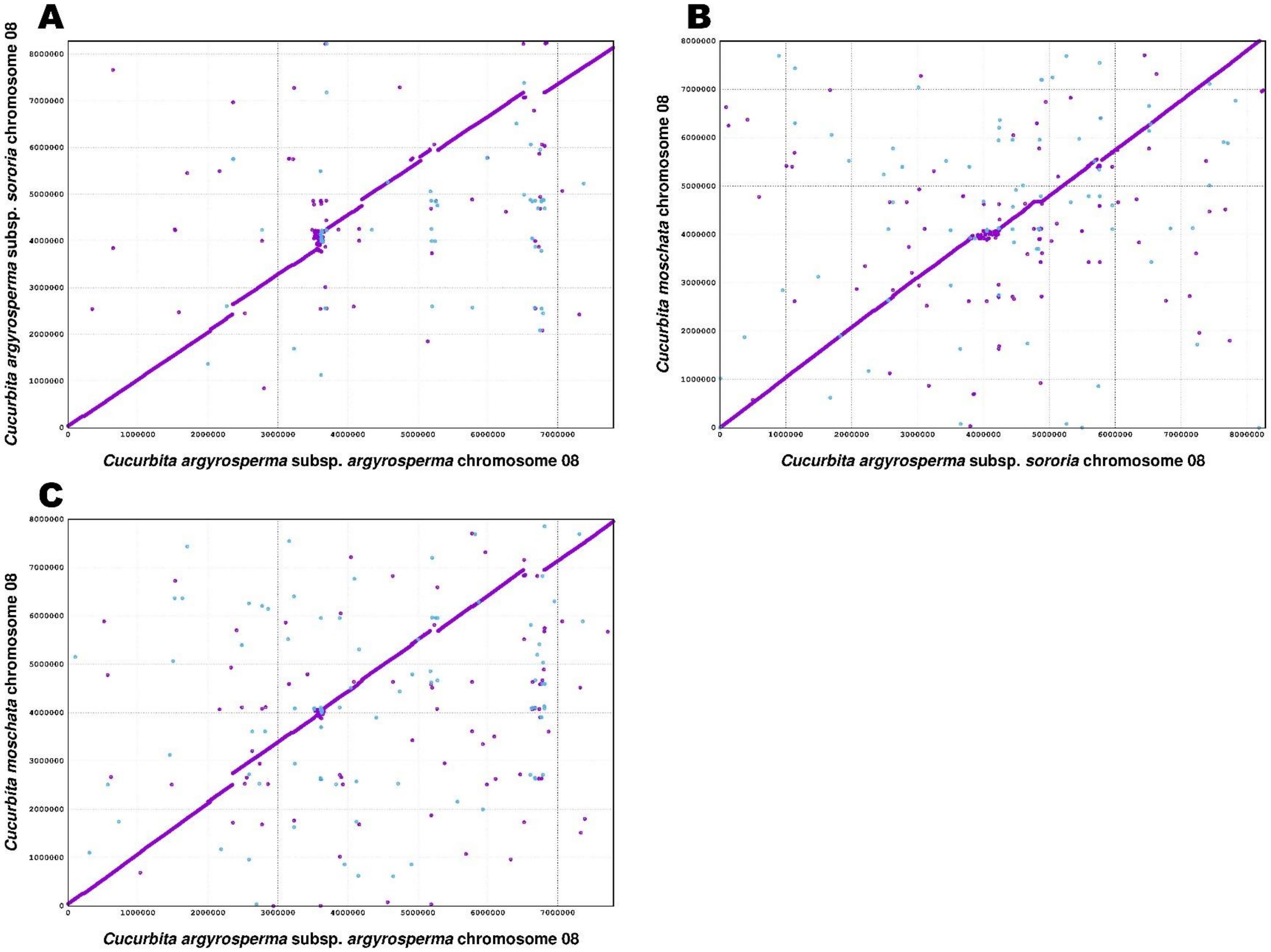


Figure S10. Synteny dot plots for chromosome 9 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found two putative insertions in the chromosome 9 of *C. argyrosperma* subsp. *argyrosperma* and one putative insertion in the chromosome 9 of *C. argyrosperma* subsp. *sororia*.

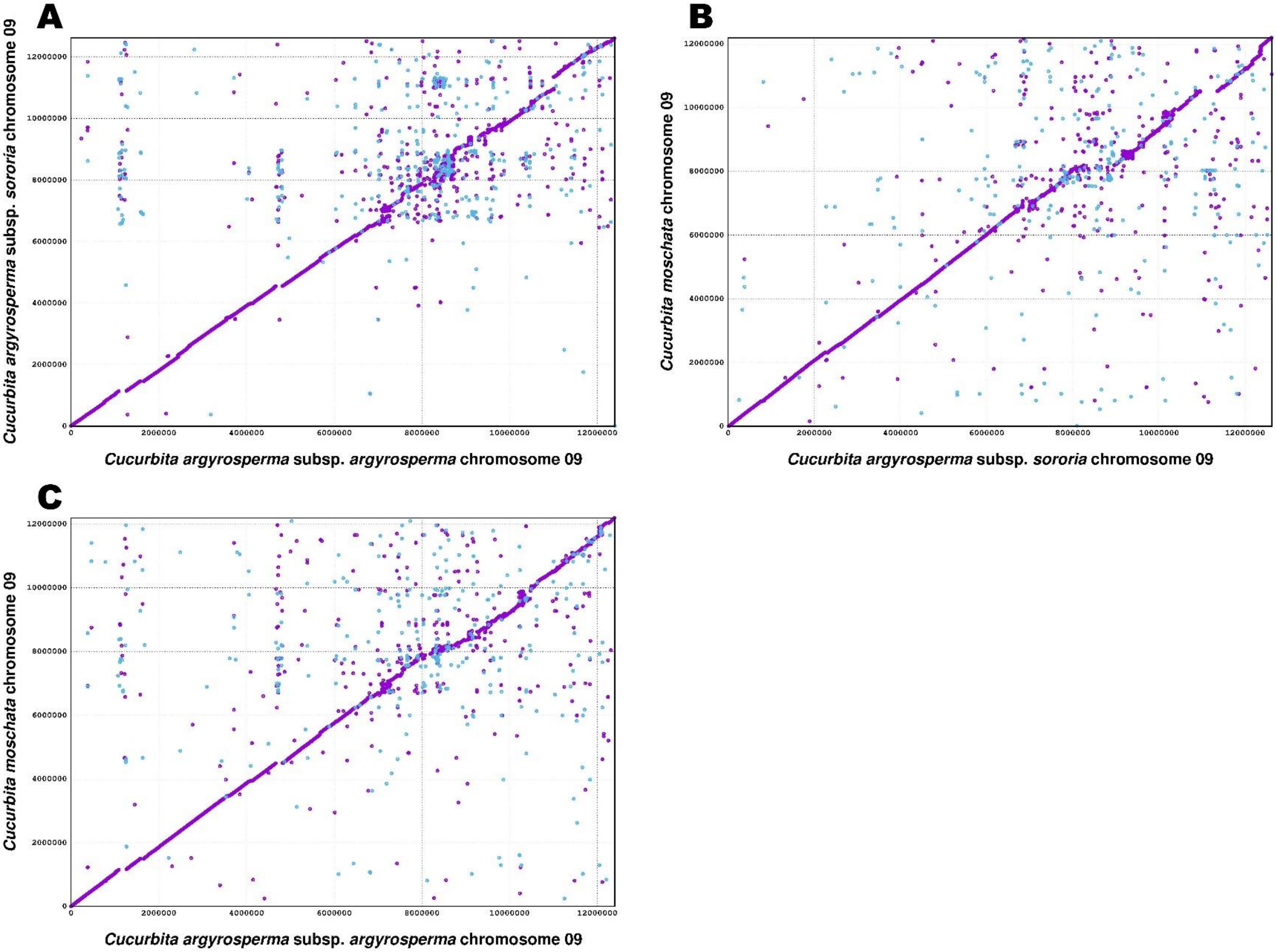


Figure S11. Synteny dot plots for chromosome 10 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a shorter centromere in the assembly of chromosome 10 of *C. argyrosperma* subsp. *argyrosperma*, possibly due to a better assembly of the repetitive regions in the other two genomes.

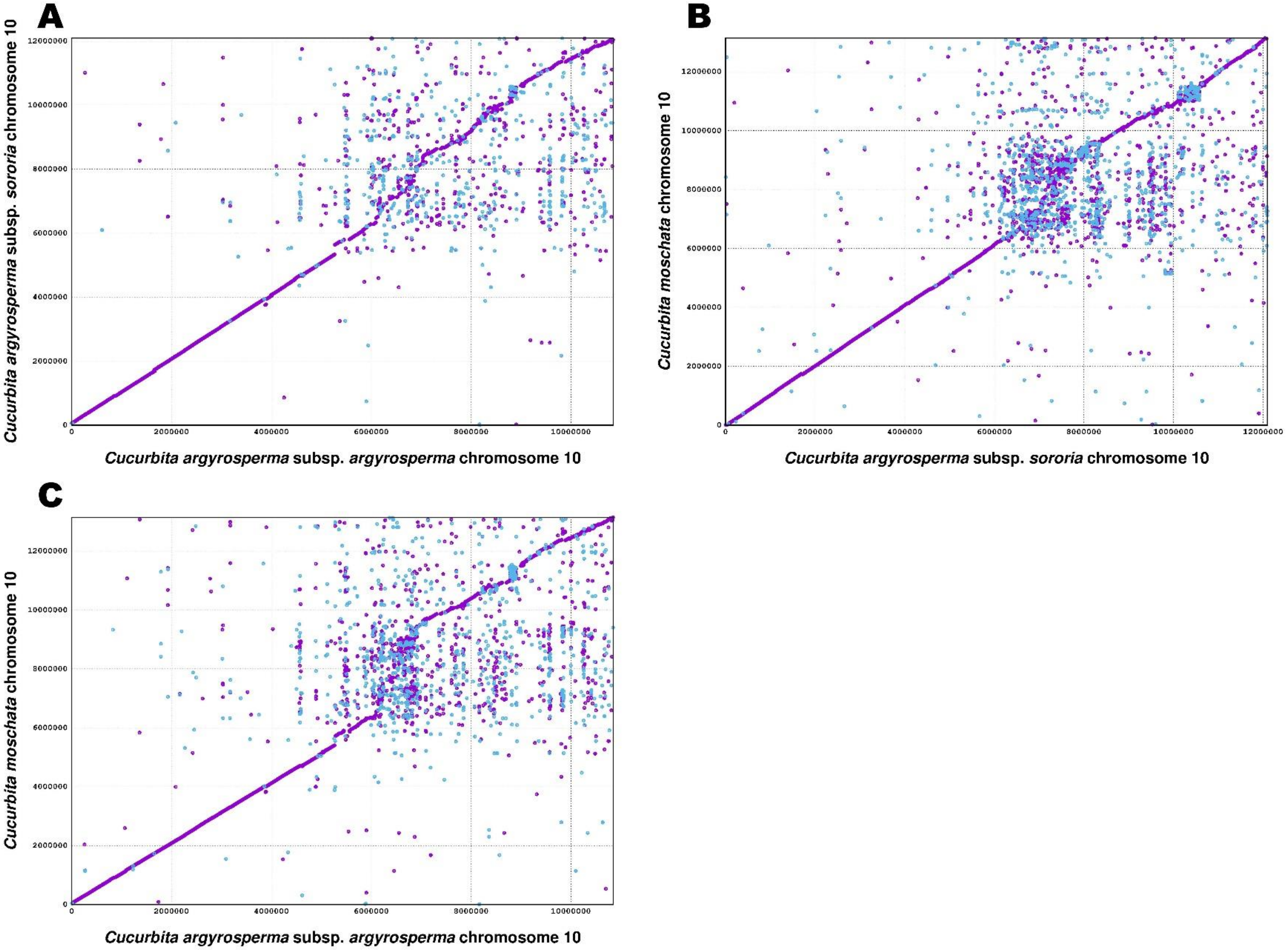


Figure S12. Synteny dot plots for chromosome 11 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We detected a larger centromere in *C. argyrosperma* subsp. *sororia*, possibly due to a better assembly of the repetitive regions. We also found three putative deletions in chromosome 11 of *C. argyrosperma* subsp. *argyrosperma* that occurred after diverging from *C. argyrosperma* subsp. *sororia*.

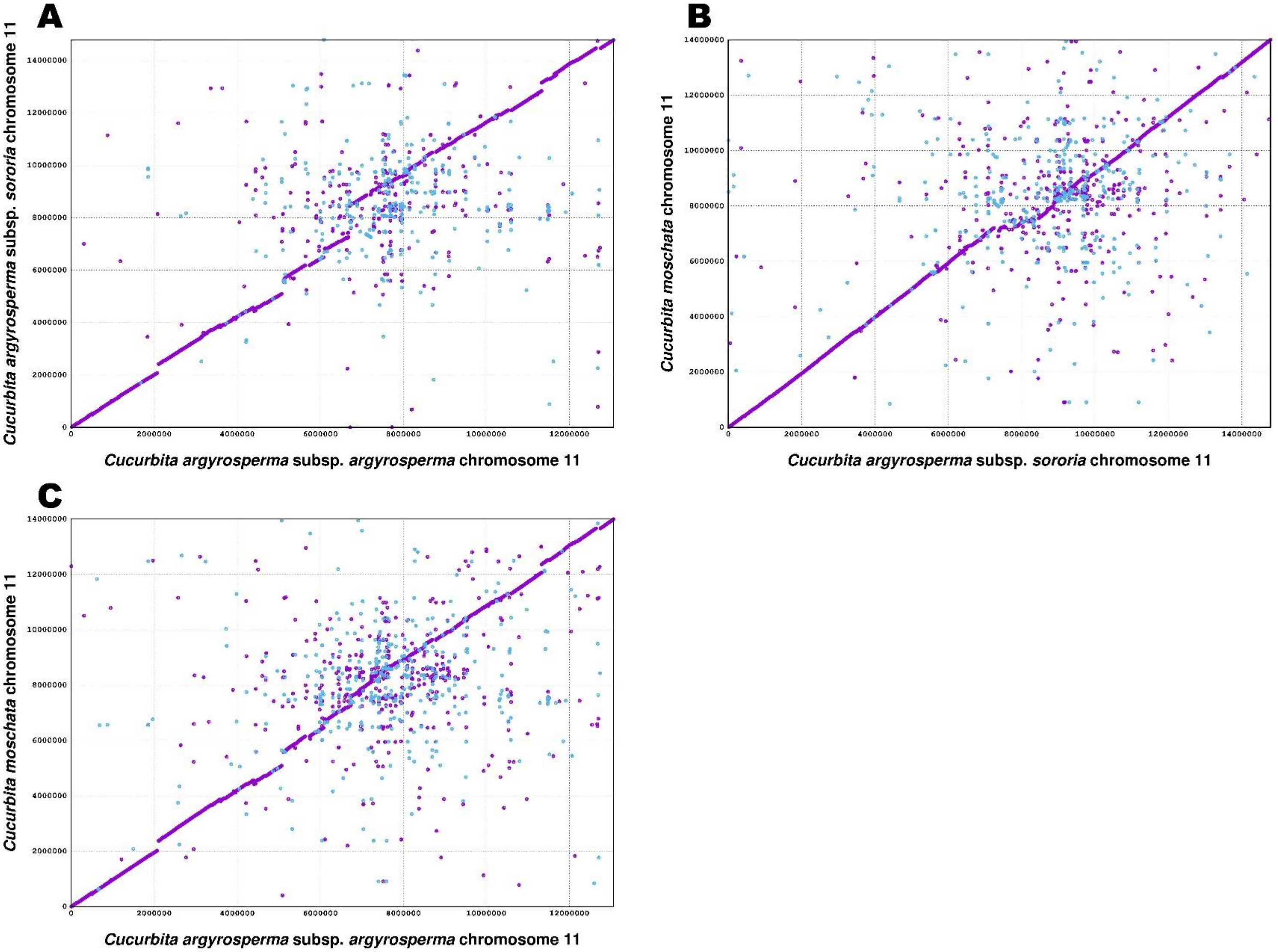


Figure S13. Synteny dot plots for chromosome 12 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We detected a larger centromere in *C. argyrosperma* subsp. *sororia*, possibly due to a better assembly of the repetitive regions. We also found five putative insertions in chromosome 12 of *C. argyrosperma* subsp. *argyrosperma* that occurred after diverging from *C. argyrosperma* subsp. *sororia*.

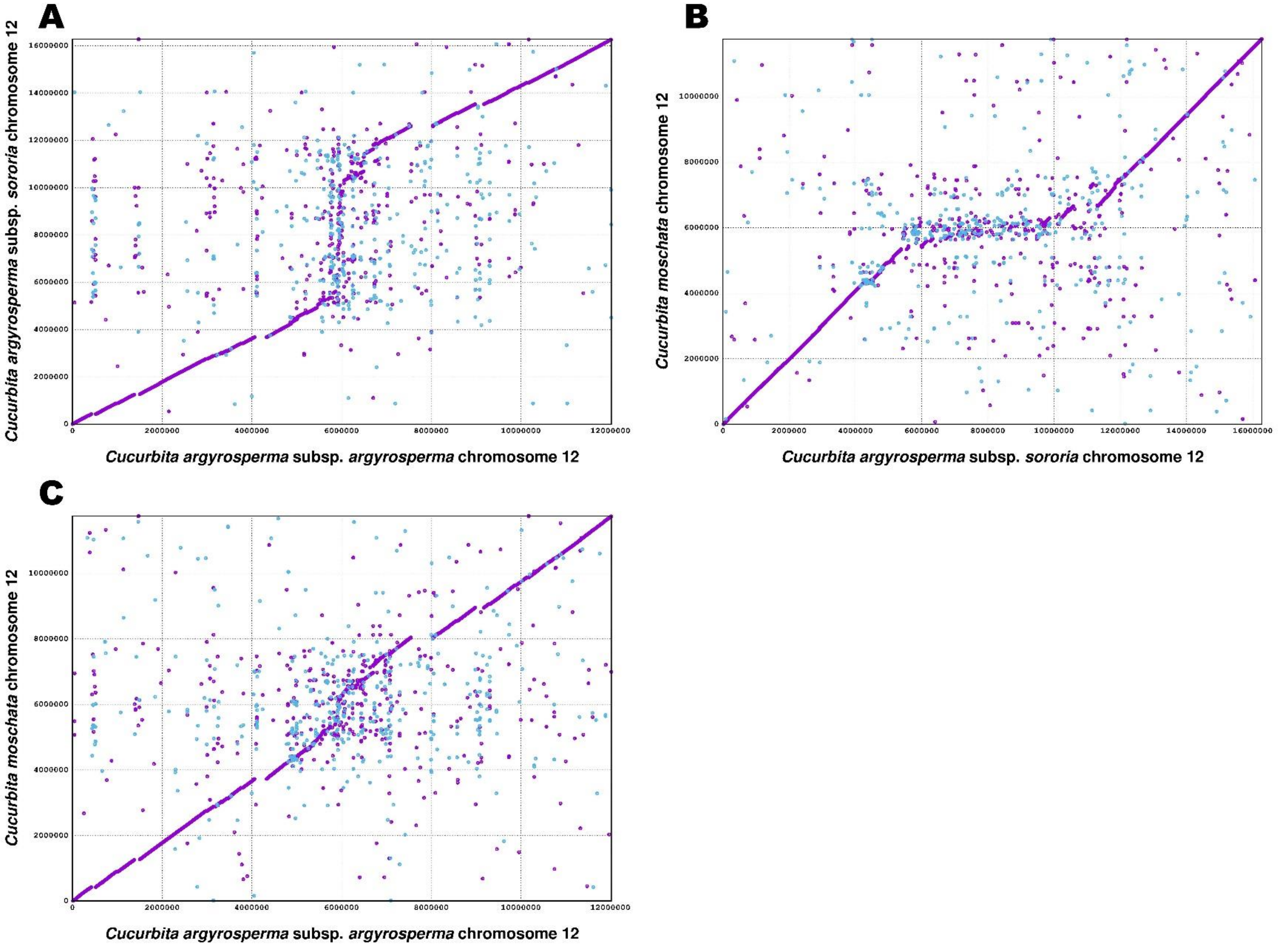


Figure S14. Synteny dot plots for chromosome 13 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We detected a larger centromere in *C. argyrosperma* subsp. *sororia*, possibly due to a better assembly of the repetitive regions.

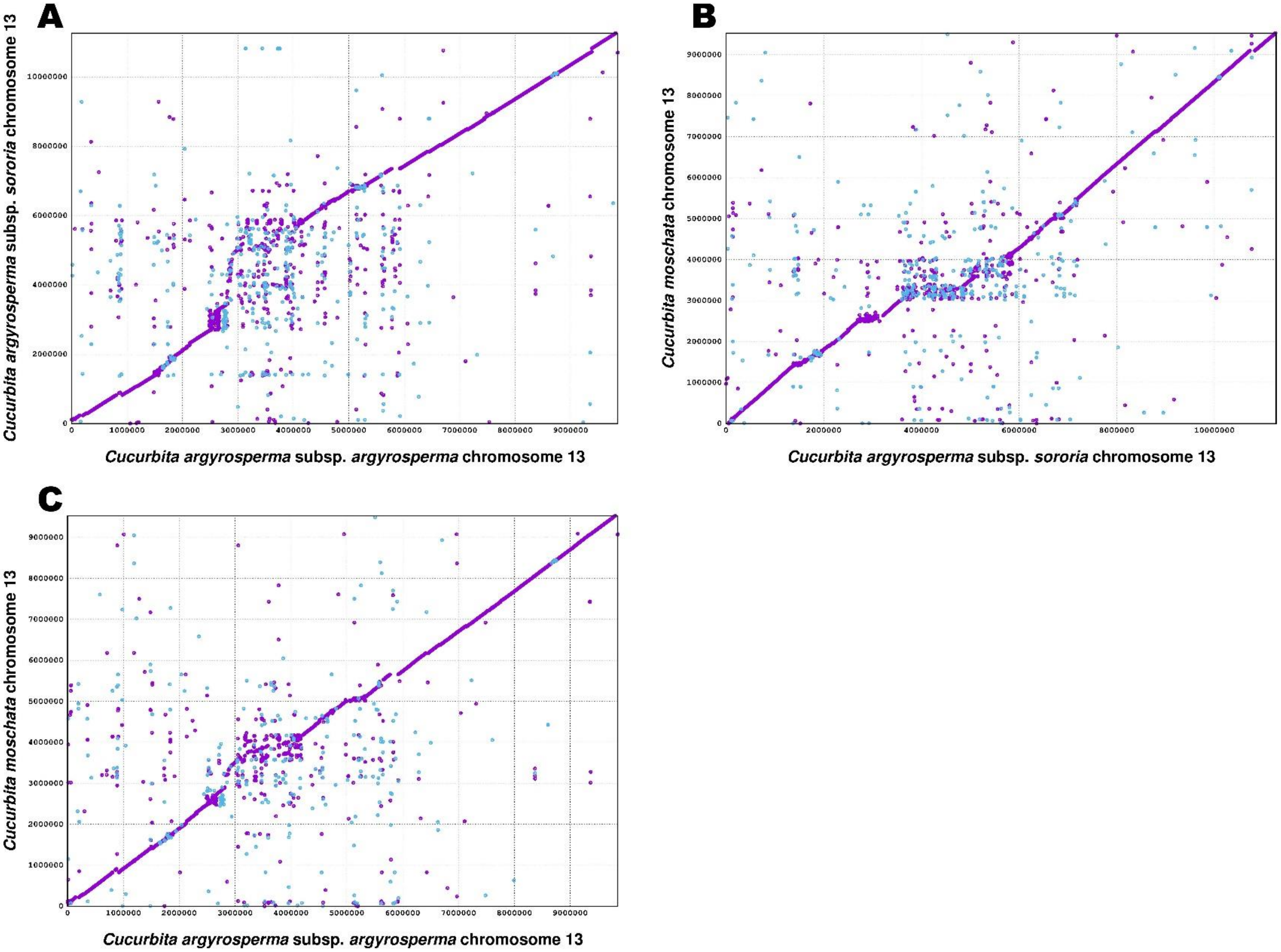
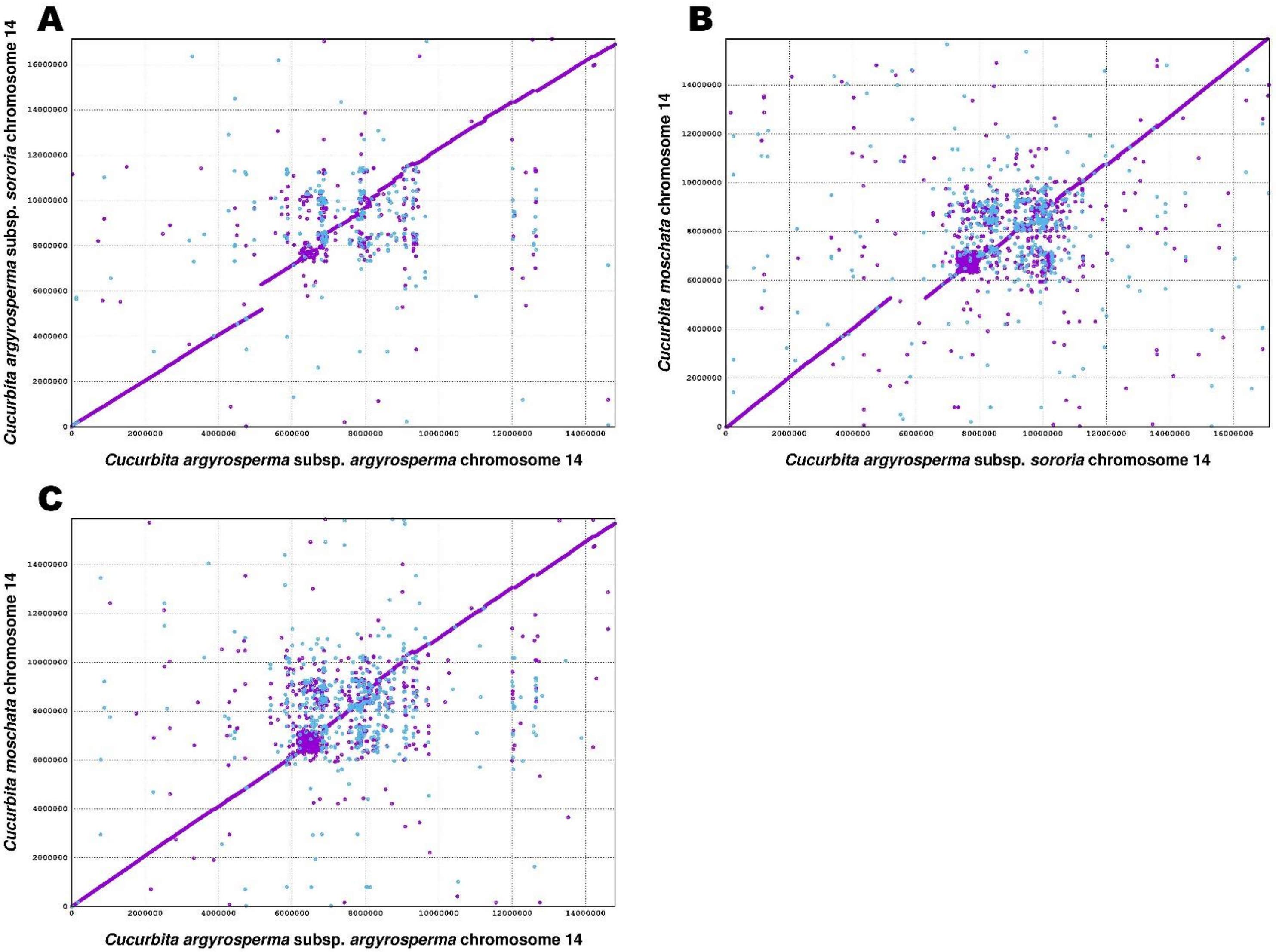


Figure S15. Synteny dot plots for chromosome 14 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a ~1,000,000 nt putative insertion in chromosome 14 of *C. argyrosperma* subsp. *sororia*.



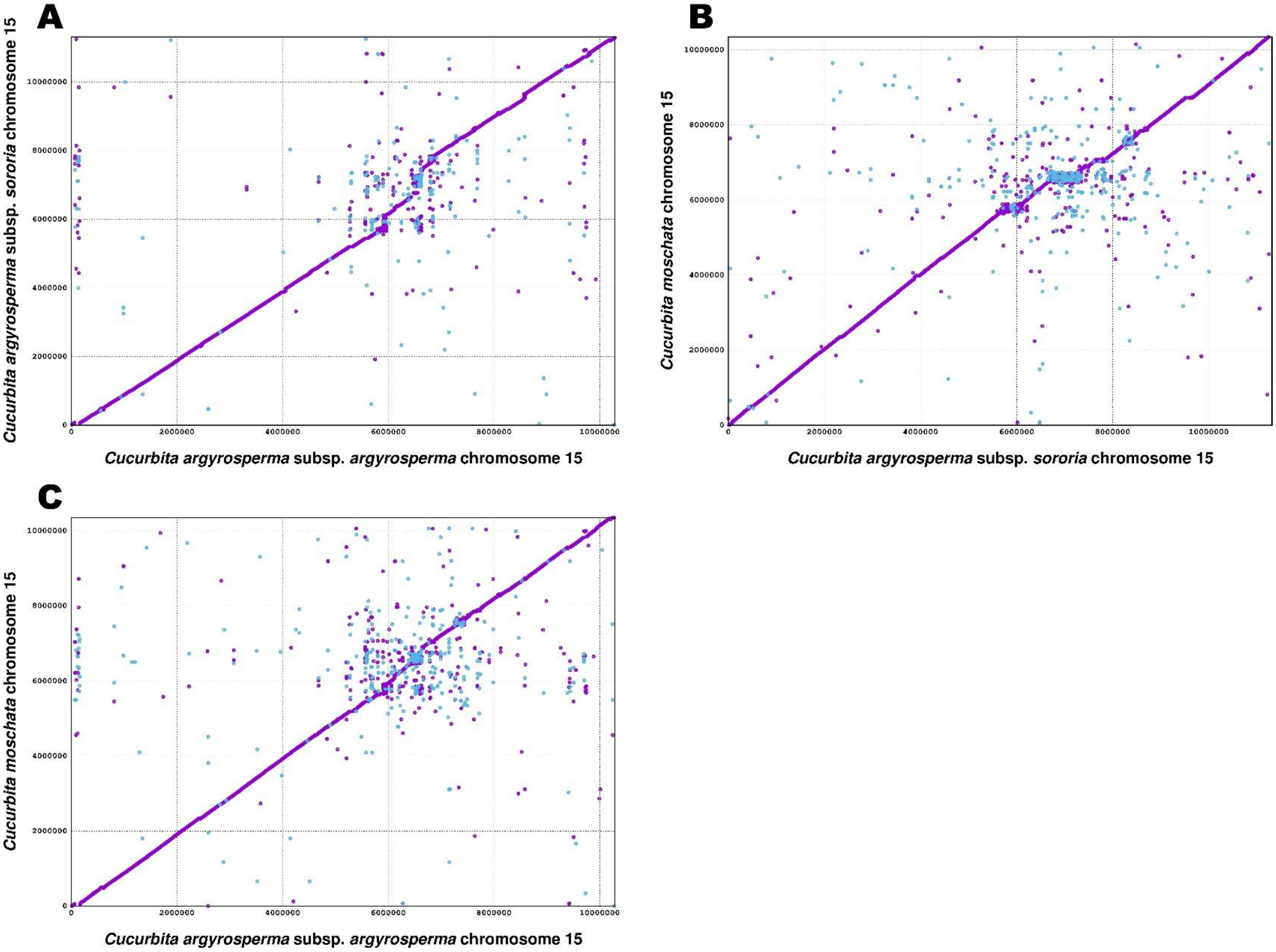


Figure S16. Synteny dot plots for chromosome 15 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We detected an insertion of a segment of chromosome 3 near the centromere of chromosome 15 in *C. argyrosperma* subsp. *sororia*, possibly due to an artifact during chromosome anchoring.

Figure S17. Synteny dot plots for chromosome 16 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found one large putative insertion (~700,000 nt) in *C. argyrosperma* subsp. *argyrosperma* between the bases 3,300,000 and 4,000,000 of chromosome 16. We found two smaller putative deletions in *C. argyrosperma* subsp. *argyrosperma* corresponding to the positions ~1,200,000 and ~7,500,000 of chromosome 16.

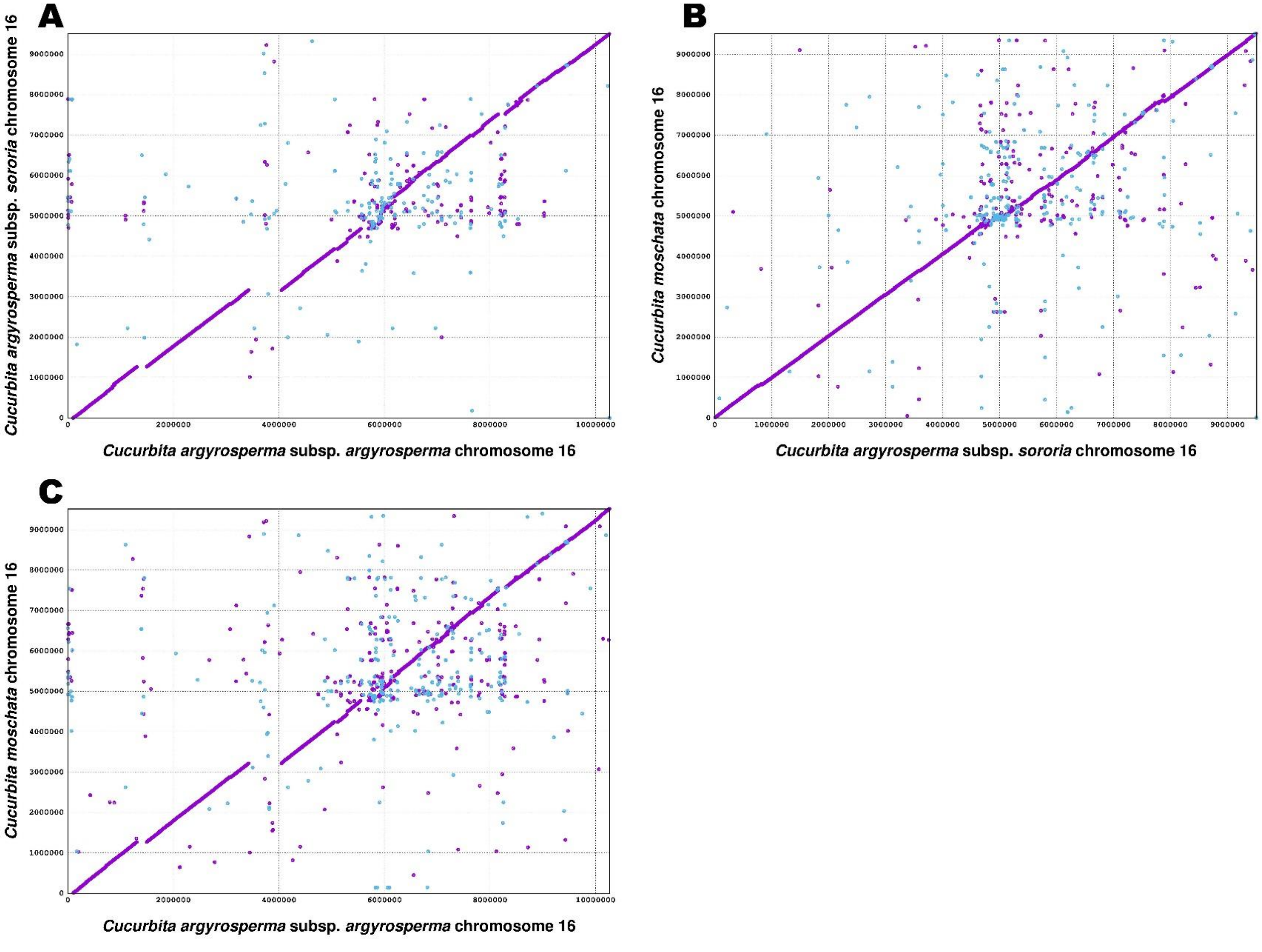


Figure S18. Synteny dot plots for chromosome 17 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a shorter centromere in the assembly of chromosome 17 of *C. argyrosperma* subsp. *argyrosperma*.

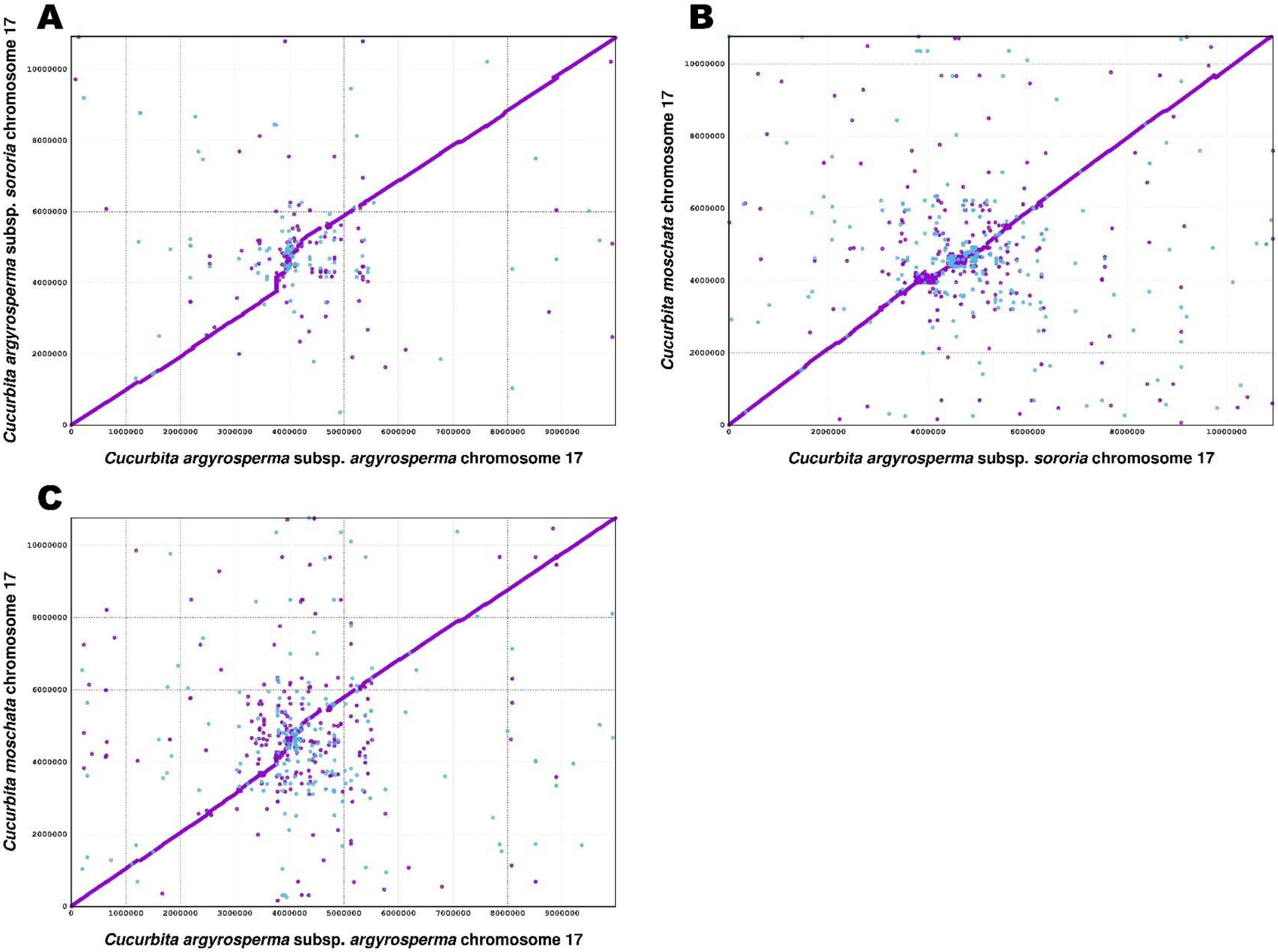


Figure S19. Synteny dot plots for chromosome 18 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We found a ~600,000 nt putative insertion near the centromere of *C. argyrosperma* subsp. *sororia*.

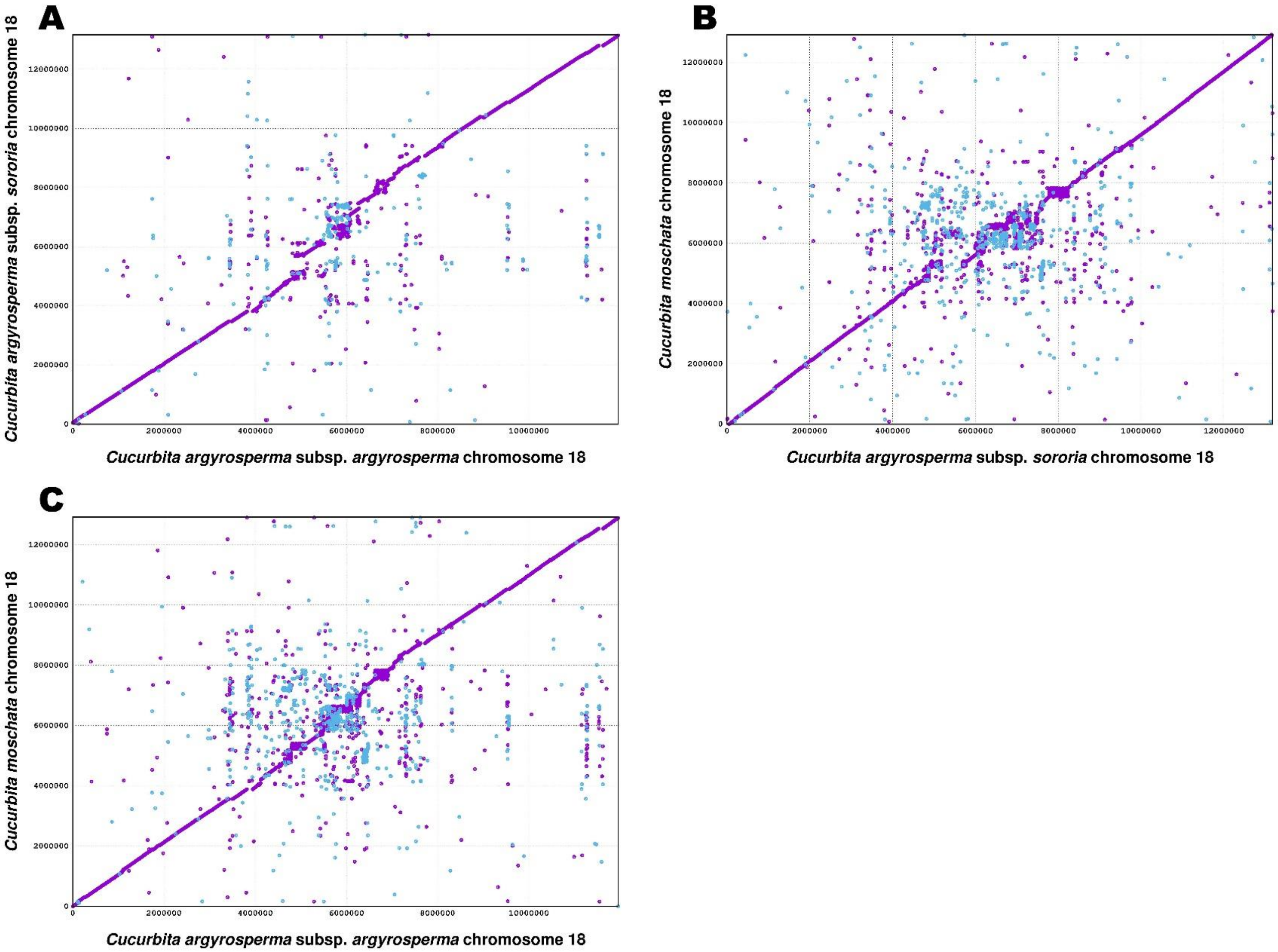


Figure S20. Synteny dot plots for chromosome 19 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We detected a larger centromere in *C. argyrosperma* subsp. *sororia*, possibly due to a better assembly of the repetitive regions.

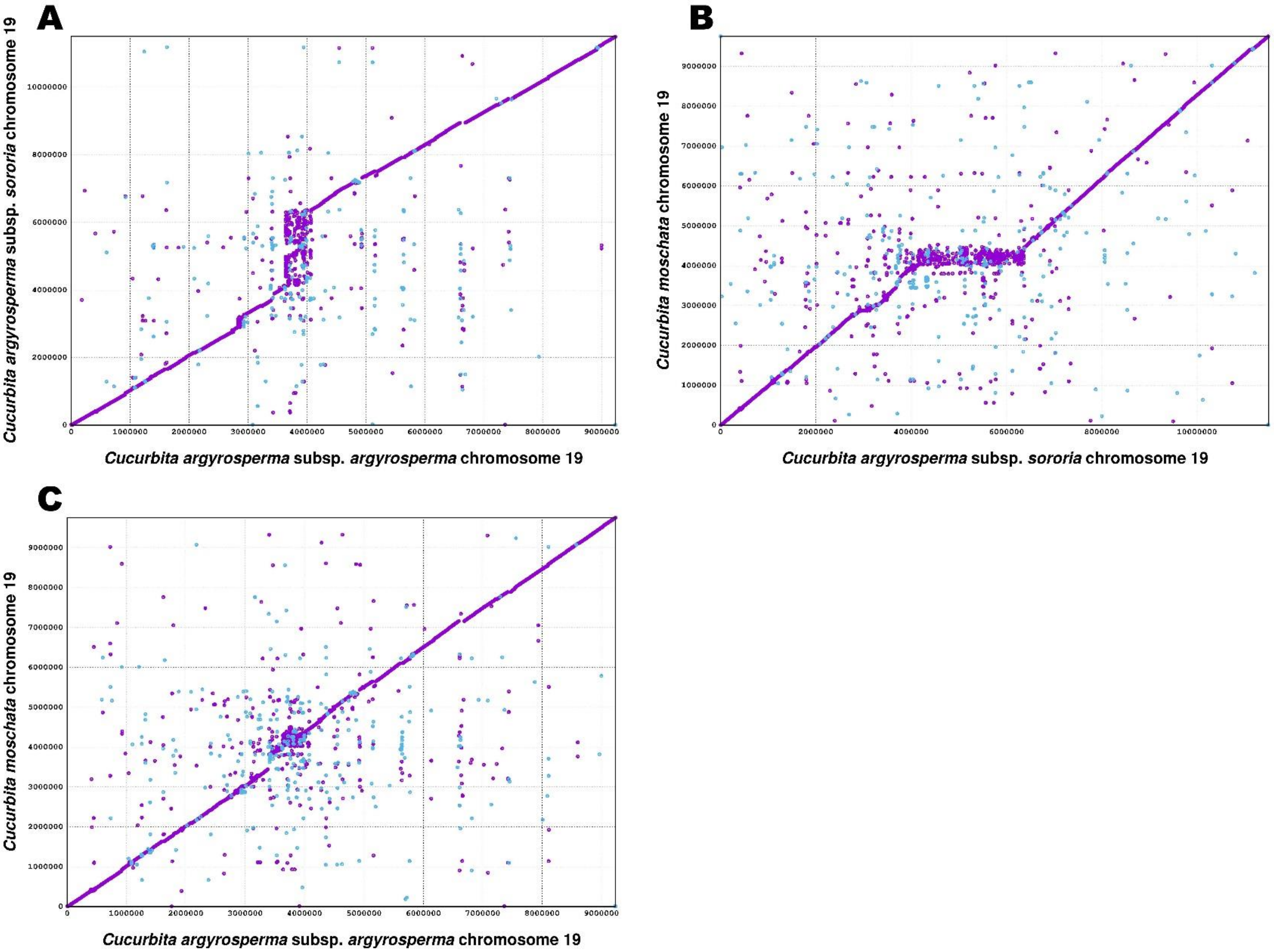


Figure S21. Synteny dot plots for chromosome 20 between (A) *C. argyrosperma* subsp. *argyrosperma* and *C. argyrosperma* subsp. *sororia*, (B) *C. argyrosperma* subsp. *sororia* and *C. moschata*, (C) *C. argyrosperma* subsp. *argyrosperma* and *C. moschata*. We used *C. moschata* as outgroup to identify the evolutionary orientation of the observed differences. We detected a larger centromere in *C. argyrosperma* subsp. *sororia*, possibly due to a better assembly of the repetitive regions.

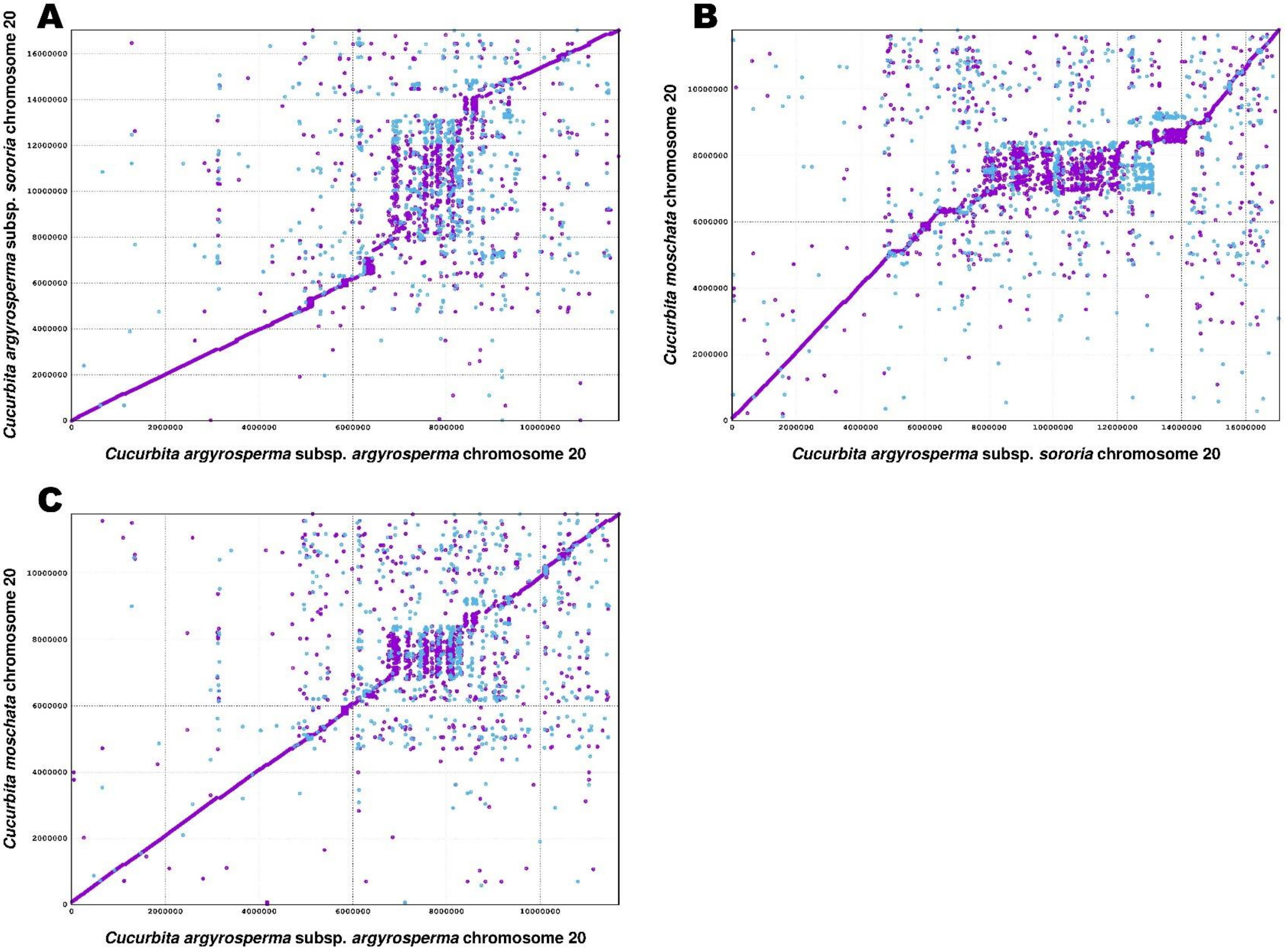


Figure S22. LD decay graph between the 10,617 SNPs used to perform the selective scans in *Cucurbita argyrosperma*.

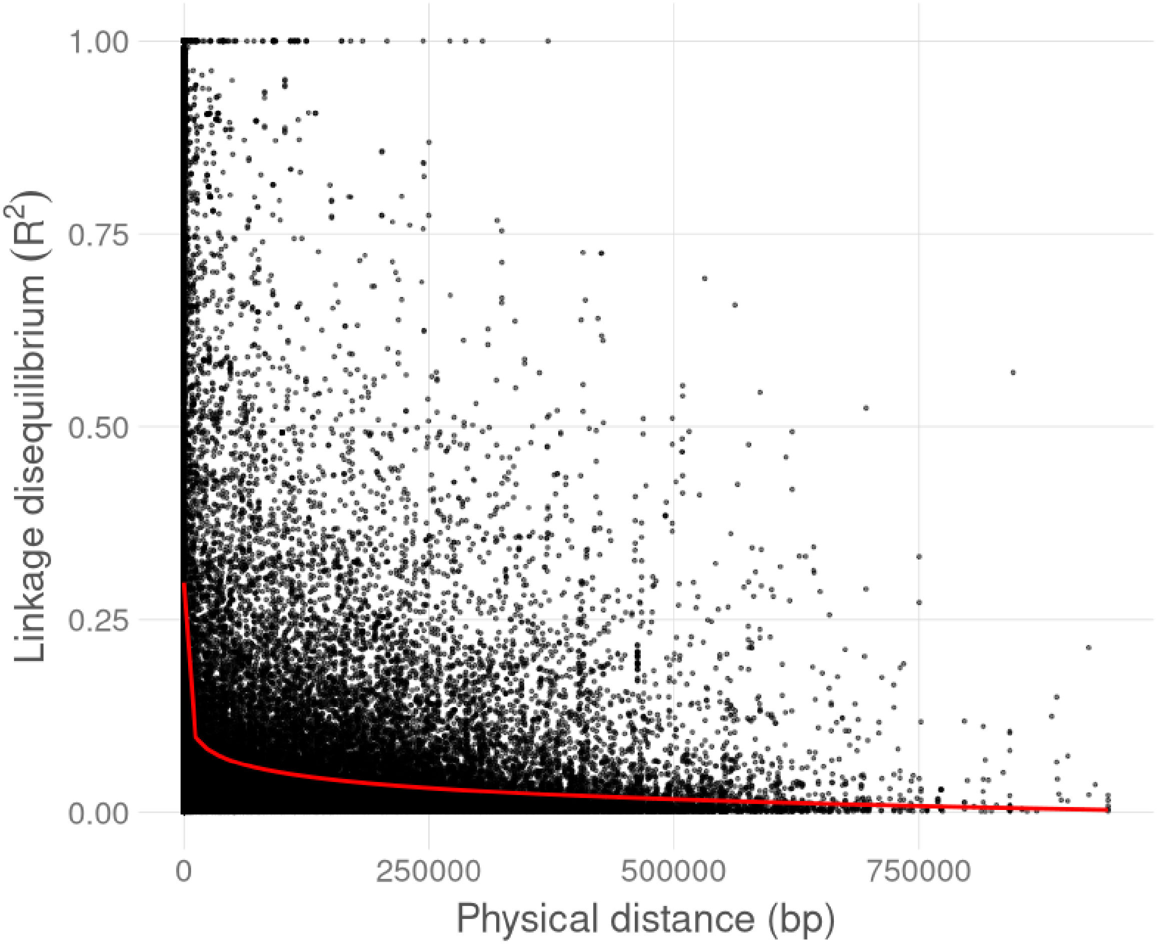
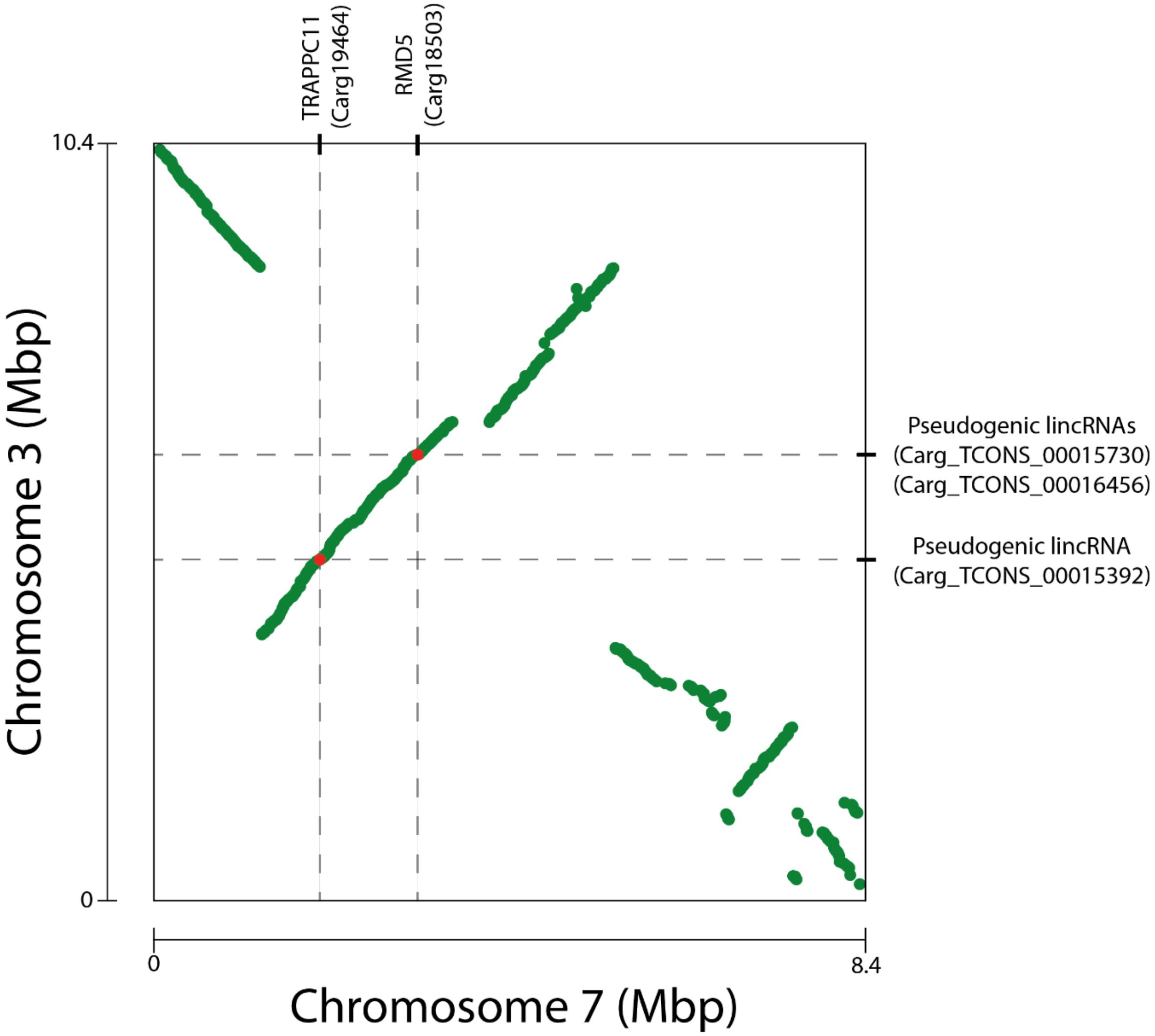


Figure S23. Synteny plot between Chromosomes 3 and 7 of *Cucurbita argyrosperma*. The syntenic location between the two long-noncoding RNAs under positive selection during domestication in chromosome 3 (Carg_TCONS00015730 and Carg_TCONS_00016456) and its protein-coding homolog in chromosome 7 (RMD5) is highlighted in red, as well as the syntenic location of a previously identified long-noncoding RNA that is potentially under purifying selection in chromosome 3 (Carg_TCONS_00015392) and its protein-coding homolog in chromosome 7 (TRAPPC11).


TRAPPC11
(Carg19464)
RMD5
(Carg18503)
Chromosome 3 (Mbp)
Pseudogenic lincRNAs
(Carg_TCONS_00015730)
(Carg_TCONS_00016456)
Pseudogenic lincRNA
(Carg_TCONS_00015392)
Chromosome 7 (Mbp)