UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

BIOLOGÍA EVOLUTIVA

GENÓMICA Y DOMESTICACIÓN DEL FRIJOL AYOCOTE

(*Phaseolus coccineus* L.)

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTORA EN CIENCIAS

PRESENTA:

AZALEA GUERRA GARCÍA

TUTOR PRINCIPAL DE TESIS: DR. DANIEL IGNACIO PIÑERO DALMAU
INSTITUTO DE ECOLOGÍA, UNAM

COMITÉ TUTOR: DR. ALFONSO OCTAVIO DELGADO SALINAS
INSTITUTO DE BIOLOGÍA, UNAM
DR. ALEJANDRO CASAS FERNÁNDEZ
INSTITUTO DE INVESTIGACIONES EN ECOSISTEMAS Y SUSTENTABILIDAD, UNAM

CD. MX. ENERO, 2019.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

BIOLOGÍA EVOLUTIVA

GENÓMICA Y DOMESTICACIÓN DEL FRIJOL AYOCOTE

(*Phaseolus coccineus* L.)

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTORA EN CIENCIAS

PRESENTA:

AZALEA GUERRA GARCÍA

TUTOR PRINCIPAL DE TESIS: DR. DANIEL IGNACIO PIÑERO DALMAU
INSTITUTO DE ECOLOGÍA, UNAM

COMITÉ TUTOR: DR. ALFONSO OCTAVIO DELGADO SALINAS
INSTITUTO DE BIOLOGÍA, UNAM
DR. ALEJANDRO CASAS FERNÁNDEZ
INSTITUTO DE INVESTIGACIONES EN ECOSISTEMAS Y SUSTENTABILIDAD, UNAM

MÉXICO, CD. MX. ENERO, 2019.

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

OFICIO CPCB/086/2019

Asunto: Oficio de Jurado para Examen de Grado.

M en C. Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted, que el Subcomité de Biología Evolutiva, en su sesión ordinaria del día 22 de octubre de 2018, aprobó, el jurado para la presentación del examen para obtener el grado de **DOCTORA EN CIENCIAS**, a la alumna **GUERRA GARCÍA AZALEA** con número de cuenta 304189256, con la tesis titulada: **"GENÓMICA Y DOMESTICACIÓN DEL FRIJOL AYOCOTE (*Phaseolus coccineus* L.)"**, bajo la dirección del **DR. DANIEL IGNACIO PIÑERO DALMAU:**

Presidente:	DR. LUIS ENRIQUE EGUIARTE FRUNS
Vocal:	DR. LUIS DAVID ALCARAZ PERAZA
Secretario:	DR. ALFONSO OCTAVIO DELGADO SALINAS
Suplente:	DRA. ALMA PIÑEYRO NELSON
Suplente	DR. ALEJANDRO CASAS FERNÁNDEZ

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPIRITU"
Cd. Universitaria, Cd. Mx., a, 18 de enero de 2019

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

DR. ADOLFO GERARDO NAVARRO SIGÜENZA
COORDINADOR DEL PROGRAMA

Agradecimientos institucionales

Al Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México (UNAM), por la formación académica que me ha brindado y por el apoyo dado a lo largo de mis estudios de doctorado.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca de manutención otorgada durante mis estudios de posgrado (440709), así como al Programa de Apoyo a los Estudios de Posgrado (PAEP) por el apoyo dado para la realización de una estancia de investigación.

Al proyecto “La variación genética de las plantas cultivadas en México: Estrategias para enfrentar el cambio climático” (247730) del CONACyT, el cual fue el sustento financiero de mi proyecto.

Al Dr. Daniel Piñero, quien fue mi tutor principal y el director de esta tesis.

A los miembros de mi comité tutor: el Dr. Alfonso Delgado Salinas y el Dr. Alejandro Casas, quienes constantemente aportaron su valioso conocimiento para el desarrollo de este trabajo.

Agradecimientos personales

El doctorado ha sido un proceso lleno de grandes aprendizajes académicos y personales. No sería justo recibir todo el mérito por un trabajo que ha resultado de la suma de muchas personas que han contribuido directa e indirectamente. Aquí trataré, aunque sé no será suficiente, de reconocer a aquellos que han estado presentes en mi vida académica y no académica.

Me encuentro profundamente agradecida con el Dr. Daniel Piñero, quien no solo me ha apoyado académicamente desde el inicio de este proyecto, sino que además se ha convertido en una fuente de admiración e inspiración por sus conocimientos, experiencia y especialmente por su calidez humana.

También quiero dar mi reconocimiento a los miembros de mi comité. El Dr. Alfonso Delgado ha compartido conmigo sus conocimientos, y gracias a él aprendí a observar, y no solo mirar, a las plantas y sus flores, y la historia que nos cuentan. El Dr. Alejandro Casas me ha ayudado a enfocar mi proyecto, especialmente me ha dado una visión amplia de la domesticación, haciéndome ver más allá de la genética.

Agradezco a los sinodales quienes amablemente accedieron a revisar este trabajo y aportaron valiosas críticas. Gracias Dr. Luis Eguiarte, Dr. Luis David Alcaraz y Dra. Alma Piñeyro.

Toda mi admiración al Dr. Jeffrey Ross-Ibarra, quien me permitió visitar su laboratorio, donde realicé análisis, desarrollé ideas y, pese al corto periodo que estuve ahí, me mostró cómo hacer ciencia de forma innovadora y colaborativa. También agradezco al Dr. Roberto Papa y a los integrantes de su laboratorio, con quienes también realicé una estancia de investigación.

Quiero agradecer a personas que me aportaron su ayuda de forma técnica o teórica. Gracias a Marco Suárez, quien me apoyó con diversos análisis, y a Alicia Mastretta, que dentro de sus muchas aportaciones está el haberme introducido y guiado en el (intimidante) campo de la bioinformática. También doy gracias a las técnicas del laboratorio Susette Castañeda y Tania Garrido, de quienes siempre recibí ayuda.

Un aspecto fundamental del desarrollo de este proyecto fue salir a campo a buscar frijoles, tarea más difícil de lo que suena, pero llena de experiencias enriquecedoras. Quiero dar mi agradecimiento a las personas que a lo largo de estos cuatro años me acompañaron a campo: Idalia Rojas (mi asidua compañera de campo), Myriam Campos, Nancy Gálvez, Alfredo Villarruel, Verónica González, Laura Figueroa.

Ahora quiero reconocer a aquellos que han estado presentes más allá de la vida profesional. Inicio agradeciendo a mis padres. Mamá, Doña Rosita, te quiero y te estoy eternamente agradecida porque de ti aprendí a seguir caminando pese a lo atemorizante que pueda parecer el camino. También doy gracias a mi padre, una mitad mía viene de ti, y has contribuido enormemente a definirme como persona. Reconozco y llevo siempre conmigo a mi familia: mi hermana Azucena, mis sobrinos Fer y Alonsito, mis tíos, mis abuelos y los que ya no están.

Agradezco a mis viejos y entrañables amigos: Stevens, Joana, Adán y Nadia. Los quiero harto. Me siento muy afortunada de haber pasado estos años en el Laboratorio de Genética y Ecología, a lado de grandes personas con quienes la cotidianidad ha sido amena y divertida: Alejandro, Juan Carlos, Raquel, Laura, Myriam, Vero, Verito y todos con los que he compartido una taza de café o una comida en la terraza. Quiero hacer un agradecimiento especial a Idalia, a quien le tomé un profundo cariño y admiración por tantos viajes, momentos, discusiones académicas y pláticas que compartimos.

Finalmente quiero dar gracias a Chris, con quien he aprendido a amar bonito. Gracias por el apoyo y la confianza, por tomar mi mano dándome libertad, gracias por el cariño, las sonrisas, los abrazos y por compartir este momento de nuestras vidas.

Mi profunda admiración y agradecimiento a quienes,
a través de su cuidadoso y continuo trabajo,
nos han dado la diversidad de colores, formas
y sabores de lo que ponemos en la mesa.

ÍNDICE

Resumen …… 1

Abstract …… 2

Capítulo 1: Introducción ……... 3

La domesticación como proceso evolutivo …….. 3

Mesoamérica, lugar de origen de cultivos …… 4

México y sus (sabrosos) frijoles ……..... 6

Distribución e historia evolutiva de los frijoles……..…… 8

Frijol común (*Phaseolus vulgaris*) 8

Frijol lima (*Phaseolus lunatus*) 9

Frijol tepari (*Phaseolus acutifolius*) y piloy (*Phaseolus dumosus*) 10

Phaseolus coccineus: Ayocote, botil, xut chenek, xutito, pocohuinic, 11

tacahuaquetl, tecomarí, patol, cimatl ……

Diversidad genética del ayocote …….. 12

Métodos y análisis en genómica de poblaciones ……..……. 14

Explorar la estructura poblacional …… 15

Inferir la historia de las poblaciones con métodos de coalescencia…… 16

Detección de firmas de selección …….. 16

Capítulo 2:

Current approaches and methods in plant domestication studies ……... 18

Capítulo 3:

Domestication genomics of the open-pollinated Scarlet Runner Bean

(*Phaseolus coccineus* L.) 37

Capítulo 4: Discusión ……..…….. 53

Estructura poblacional de *Phaseolus coccineus* y el origen de sus cultivos …….. 54

Flujo genético e introgresión …….. 55

Diversidad genética y el cuello de botella de la domesticación …… 55

Huellas de selección en el genoma de *Phaseolus coccineus* …….. 56

Conclusiones y perspectivas …….. 59

Referencias ……..……. 61

Material suplementario 69

Resumen

La domesticación es un proceso complejo en el que interactúan fuerzas evolutivas y la selección ejercida por los humanos, y que produce cambios en los fenotipos y en los genomas. En las últimas décadas los marcadores moleculares y el desarrollo de tecnologías de secuenciación masiva han permitido rastrear las huellas genómicas de la domesticación y encontrar patrones en especies domesticadas. Resultado de lo anterior, las especies cultivadas se han convertido en modelos de estudio. Sin embargo, los esfuerzos se han enfocado en los cultivos de mayor importancia económica y se han dejado de lado especies de importancia regional y a los parientes silvestres.

Este trabajo se centra en *Phaseolus coccineus*, una de las cinco especies del género que fueron domesticadas, pero que a diferencia del frijol común (*P. vulgaris*), ha recibido poca atención. *Phaseolus coccineus*, junto con otras tres especies de frijoles, fue domesticada en lo que hoy es el territorio mexicano. Es una planta perenne pero que es cultivada como anual por pequeños agricultores para el consumo de sus semillas y vainas inmaduras. En este trabajo se estudiaron los patrones de diversidad y diferenciación en las poblaciones silvestres y domesticadas para tratar de elucidar parte de su historia de domesticación e identificar loci relacionados con distintos procesos selectivos (adaptación, domesticación y diversificación de cultivos). Se colectaron individuos silvestres de 10 localidades, 11 cultivares de México, una línea mejorada (Blanco Tlaxcala) y un cultivo de España, además de tres poblaciones de ferales. Se generaron datos genómicos basados en la técnica de *Genotyping by Sequencing* (GBS), la cual hace un muestreo del genoma, resultando en la obtención de 42,548 marcadores genéticos. Se identificaron ocho grupos genéticos, la mitad de éstos corresponden a poblaciones silvestres y la otra mitad a cultivos. Contrario a lo propuesto por otros autores, quienes sugieren dos eventos de domesticación, las plantas cultivadas integran un clado monofilético, lo que sugiere un evento de domesticación, el cual probablemente ocurrió a partir de poblaciones silvestres de la Faja Volcánica Transmexicana.

Uno de los patrones reconocidos en las especies domesticadas es la reducción de diversidad genética, resultado del cuello de botella en las fases iniciales de la domesticación. En el caso de *P. coccineus*, los valores de heterocigosis más bajos fueron encontrados en los cultivos de Oaxaca (H_E=0.148) y de España (H_E=0.134). Sin embargo, no hay un patrón claro entre los niveles de variación y la condición de silvestre o domesticada. Por ejemplo, poblaciones silvestres de la Sierra Madre Occidental presentan valores de heterocigosis tan bajos como algunos cultivos, y las poblaciones domesticadas de la Faja Volcánica Transmexicana muestran niveles superiores a algunas poblaciones silvestres. Lo anterior sugiere que el cuello de botella no fue tan severo. Un factor que pudo haber contribuido al mantenimiento de la variación genética es el sistema de polinización cruzada que presenta esta especie. En el caso del cultivo de España, la baja diversidad puede deberse al subsecuente cuello de botella que ocurrió durante la introducción a Europa. Finalmente, usando dos métodos de detección de *outliers*, se identificaron 24 SNPs candidatos relacionados a domesticación, 13 a diversificación de cultivos y ocho a selección natural. Dentro de los SNPs relacionados a domesticación, cuatro se han reportado altamente expresados en flores y vainas, estructuras que han sido modificadas durante la domesticación. Sin embargo, la mayoría de estos marcadores candidatos no se encontraron en regiones anotadas. Este trabajo constituye un primer acercamiento a la diversidad genómica de las poblaciones silvestres y cultivadas de *P. coccineus* en México, donde fue domesticada esta especie. A través de los datos generados, fue posible hacer inferencias acerca de la historia de domesticación y de las regiones que han sido afectadas en este proceso. Ampliar el muestreo, incluyendo poblaciones de Centroamérica y cultivos de diferentes países de Europa, contar con un genoma de referencia de la especie e incrementar la densidad de marcadores permitiría profundizar en la historia evolutiva de la especie e identificar con mayor precisión regiones que han estado sujetas a selección.

Abstract

Domestication is a complex process where both evolutionary forces and human selection interact, affecting the phenotypes and the genomes of the domesticated species. In the last few decades, the development of molecular markers and more recently, of massive sequencing tools, have enabled the identification of genomic signatures of domestication and the discovery of common patterns across domesticated species. Crops have become model species for evolutionary and genetic studies, but most works have focused in economically important crops, while domesticated species important at regional levels and wild relatives have received little attention.

This work focuses on the scarlet runner bean (*Phaseolus coccineus*), which is one of the five domesticated species in the genus *Phaseolus*. The domestication of runner bean and other three species of the genus occurred in Mexico. The scarlet runner bean is a perennial species, but it is usually cultivated as an annual by smallholder farmers as a self-sufficiency crop for its dry seeds and immature pods. The aims of this study were to provide information about the genetic diversity and differentiation patterns of wild and domesticated populations to elucidate its domestication history and to detect signatures of selective processes (adaptation, domestication and cultivar diversification). Individuals from 11 wild populations, 11 Mexican cultivars, one breeding line and one cultivar from Spain were collected. Genomic data were generated using a *Genotyping by Sequencing* (GBS) approach, which samples the genome, obtaining 42,548 genetic markers. Eight genetic groups were identified, of which half correspond to wild populations and the rest to cultivars. Previous works suggested that two domestication events took place in *P. coccineus*, but our data show that cultivars integrate a monophyletic group, suggesting a single domestication event that probably occurred from wild populations in the Trans-Mexican Volcanic Belt.

A recognized pattern across crop species is the loss of diversity due to the bottleneck associated with the initial phase of domestication. In the samples included in this study, the lowest heterozygosity values were estimated in cultivars from Oaxaca (H_E=0.148) and from Spain (H_E=0.134). Nonetheless, no clear pattern was observed with regards of the diversity levels when comparing wild versus cultivated populations. For example, wild samples from Sierra Madre Occidental presented lower heterozygosity than cultivars from the same geographic region, and cultivars from the Trans-Mexican Volcanic Belt have higher diversity than two wild populations. Therefore, our data suggest that the genetic bottleneck related to domestication was not severe. An important factor that probably contributed to maintain diversity is the open-pollinated system observed in runner bean. In contrast, the low heterozygosity estimated in the cultivar from Spain could be due to the second bottleneck that occurred during its introduction into Europe. To identify signatures of selection and obtain candidate genetic markers, two methods based on outlier detection were used. Twenty four SNPs putatively related to domestication, 13 to cultivar diversification and eight to natural selection were detected. While four of SNPs related with domestication have been reported as highly expressed in flowers and pods, most of the candidate SNPs fell within no annotated regions of the genome. This is the first work that focuses on analyzing the genomic diversity of wild and cultivated populations of *P. coccineus* in Mexico, its center of domestication. We provide information about the domestication history of scarlet runner bean and the genomic regions that were affected during the domestication process.

Capítulo 1: Introducción

La domesticación como proceso evolutivo

La domesticación es un proceso que ha sido ampliamente estudiado, comenzando por el mismo Darwin, quien documentó cómo la selección ejercida por el ser humano es capaz de modificar atributos de las especies cultivadas, y usó este caso como analogía de la selección natural. La domesticación es un proceso dinámico y continuo que comienza con la explotación, manejo y/o cultivo deliberado de poblaciones silvestres de una especie (Pickersgill, 2007), e involucra factores biológicos, ecológicos y culturales (Larson et al., 2014). Con el tiempo, las poblaciones manejadas divergen de sus ancestros silvestres y se adaptan a las nuevas condiciones agroecológicas, las cuales en muchos casos son dependientes del ser humano (Meyer y Purugganan, 2013). Estas adaptaciones resultan en cambios fenotípicos morfológicos y fisiológicos que en conjunto son conocidos como “síndrome de domesticación”. Frecuentemente las diferencias entre los atributos de los cultivos y parientes silvestres no son discretas, sino que integran un continuo morfológico (Abbo et al., 2014).

Conforme las especies son cultivadas en jardines, traspatios o campos de cultivo y son dispersadas a nuevas áreas geográficas, se establecen poblaciones a partir de un número relativamente pequeño de individuos fundadores (Ladizinsky, 1985). La introducción a nuevos agrosistemas conduce a nuevas adaptaciones, así como a cambios genómicos que reflejan las historias demográficas, las presiones selectivas y el flujo genético. Por tanto, al caracterizar y conocer los patrones de diversidad, estructura y organización de los genomas de las especies domesticadas, así como identificando cambios causales en loci específicos, es posible entender los mecanismos moleculares, las respuestas genéticas y fenotípicas a la selección artificial, y tratar de extrapolar a procesos de evolución natural (Kantar et al., 2017). Por otro lado, debido a que los patrones de diversidad muestran las historias de ancestría, migración y cambios en el tamaño de las poblaciones, los estudios de genética de poblaciones contribuyen a nuestro entendimiento de dónde ocurrió la domesticación y cómo aprovechar los recursos genéticos de los cultivos y sus parientes silvestres (Larson et al., 2014; Kantar et al., 2017).

Los principales cuestionamientos evolutivos en especies domesticadas pueden clasificarse en seis grandes grupos (Gerbault et al., 2014 ; Gaut et al., 2018): 1) ¿Dónde, cuándo y cuántas veces fue domesticada una especie?; 2) ¿Cuáles fueron las rutas de distribución?; 3) ¿Ha ocurrido flujo genético entre los parientes silvestres y las poblaciones domesticadas?: 4) ¿En cuántas generaciones se fijaron los atributos fenotípicos que diferencían

a las formas domesticadas?; 5) ¿Cómo se adaptan los cultivos a nuevos ambientes antropogénicos?; 6) ¿Cuál es la historia demográfica de las poblaciones cultivadas? Muchos de estos cuestionamientos pueden ser contestados, al menos de forma parcial, a través de la genética de poblaciones, ya que la variación a lo largo del genoma es moldeada por la historia demográfica, mientras que la diversidad en genes relacionados con atributos clave está dada por selección. Sin embargo, diferenciar entre estos dos procesos no es trivial y puede ser metodológicamente complejo (Ross-Ibarra et al., 2007; Gerbault et al., 2014).

La investigación y el cúmulo de conocimiento de la domesticación, así como del origen espacio temporal de las especies domesticadas, se ha acelerado en los últimos años, resultado de estudios arqueológicos, el avance en las técnicas de secuenciación del material genético y el desarrollo de métodos para identificar cambios claves en plantas y animales domesticados (Meyer y Purugganan 2013; Larson et al., 2014). La domesticación ofrece la oportunidad de dilucidar preguntas evolutivas, ya que se trata de un proceso reciente en términos de la escala geológica, y porque las presiones selectivas que el humano ha ejercido en ocasiones son conocidas (Fuller et al., 2014). Así, los principales cultivos se han convertido en modelos de estudio en el área de la genética y la evolución, y han permitido estudiar la interacción de los humanos con especies bajo manejo (Kantar et al., 2017).

Actualmente una cantidad importante del material conservado en bancos de germoplasma ha sido secuenciado por completo o genotipado. Por ejemplo, 539 líneas mejoradas de maíz que se encuentran en el Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) han sido genotipados (Wu et al., 2016). Se espera que en los próximos años aumente la cantidad de información genética disponible, tanto de cultivos como de parientes silvestres. Pese a que el objetivo de estos esfuerzos se centra en mejoramiento, estos recursos son también útiles para estudios de genética de poblaciones, evolución de los genomas y de diversidad (Kantar et al., 2017).

Mesoamérica, lugar de origen de cultivos

Un cuerpo rico y diverso de datos arqueológicos y moleculares sugieren que la agricultura surgió de forma independiente en los distintos continentes y en diferentes culturas (Fuller et al., 2014; Larson et al., 2014). Los primeros estudios en plantas domesticadas impulsaron los conceptos de los *Centros de Origen* o *Centros de Domesticación*. Esta hipótesis, que fue originalmente propuesta por Alphonse de Candolle en 1886, y posteriormente refinada y popularizada por Nikolai Vavilov (1951), proponía que la domesticación ocurrió en unas pocas y

discretas regiones geográficas (“centros”), a partir de los cuales las especies domesticadas se expandieron a través de las migraciones humanas y el comercio. Lo que Vavilov observó es que la variación de los cultivos se concentra en ciertas regiones geográficas, y estos centros de diversidad frecuentemente coinciden entre varias especies domesticadas.

Actualmente, datos genéticos y arqueológicos sugieren que estos sitios discretos o centros de domesticación pueden estar simplificando la historia de las especies cultivadas (Harlan, 1971). Sin embargo, el concepto de Centro de Origen es útil para tratar de responder preguntas relacionadas con encontrar variación importante en los cultivos, estudiar el cambio en la diversidad dentro y fuera de su rango natural de distribución, y examinar la evolución fenotípica y genotípica, especialmente cuando los cultivos son introducidos a nuevos sitios (Kantar et al., 2017).

Mesoamérica es una de las áreas de América donde surgió la agricultura, lo que llevó al origen de distintos cultivos en esta región (Harlan 1971). Esto puede estar relacionado con la alta diversidad de plantas y culturas presentes (Casas et al., 2007; Pickersgill, 2007). Por ejemplo, sólo en el territorio mexicano, Villaseñor (2003) estima la presencia de más de 22,300 especies de angiospermas, de las cuales el 56% son endémicas de México. Por otro lado, de acuerdo con Toledo (2001) en el territorio mexicano habitan 56 grupos étnicos indígenas, cuyos ancestros han ocupado el área por 12,000-14,000 años. Así, los grupos humanos en Mesoamérica se han establecido y diversificado en una amplia gama de ecosistemas, explorando y modificando a través de selección un gran número de especies de plantas (Delgado-Salinas et al., 2004).

Se calcula que en Mesoamérica se utilizan entre 5,000 y 7,000 especies de plantas, y más de 200 especies nativas fueron domesticadas en esta región, lo que implica que coexisten con sus parientes silvestres (Casas et al., 1994). Dentro de estos cientos de especies domesticadas en Mesoamérica se incluyen algunas de gran importancia económica a nivel mundial, como el maíz (*Zea mays*), el chile (*Capsicum annum*), distintas especies de calabaza (*Cucurbita* spp.), el algodón (*Gossypium hirsutum*) y los frijoles (*Phaseolus* spp.). Pero también hay especies económicamente importantes a nivel regional con niveles de domesticación que van desde incipiente hasta avanzada (Casas et al., 2007). Ejemplos de esto incluyen especies de los géneros *Agave*, *Opuntia*, *Leucaena*, *Chenopodium* y *Amaranthus*; así como al tomate (*Physalis* sp.), el aguacate (*Persea americana*), el chayote (*Sechium edule*), el zapote blanco (*Casimiroa edulis*), el zapote negro (*Diospyros digyna*), la jícama (*Pachyrhizus erosus*) y la vainilla (*Vanilla planifolia*), entre otras (Delgado-Salinas et al., 2004; Casas et al., 2007)

Existe un amplio espectro de formas en las que el humano interactúa con las especies manejadas. Se pueden distinguir dos tipos principales (Casas et al., 1996):

- *In situ*: interacciones que se llevan a cabo en los espacios naturalmente ocupados por plantas silvestres. El humano puede o no alterar la estructura fenotípica y/o genética de las poblaciones. Incluye la recolección, tolerancia, fomento o inducción, y la protección.
- *Ex situ*: interacciones que se llevan a cabo fuera de las poblaciones naturales, en hábitats creados y controlados por el hombre.

México y sus (sabrosos) frijoles

El género *Phaseolus* (Fabaceae) contiene alrededor de 70 especies, la mayoría de ellas distribuidas en Mesoamérica, donde se diversificó el género hace 4-6 millones de años (Delgado-Salinas et al., 2006). *Phaseolus* (2n = 2X = 22) representa un caso destacable de domesticación ya que ocurrieron varios eventos, resultando en cinco especies domesticadas: *Phaseolus vulgaris*, *P. coccineus*, *P. acutifolius*, *P. dumosus*, y *P. lunatus*. Cuatro de estas especies fueron domesticadas en lo que hoy es el territorio de México (Blair et al., 2012; Schmutz et al., 2014; Chacón-Sánchez y Martínez-Castillo, 2017; Guerra-García et al., 2017; Chacón-Sánchez, 2018; Fig. 1). Únicamente *P. dumosus* fue domesticada fuera del territorio mexicano (Chacón-Sánchez, 2018), y dos especies (*P. vulgaris* y *P. lunatus*) se domesticaron de forma independiente en Mesoamérica y en los Andes (Andueza-Noh et al., 2012; Schmutz et al., 2014; Andueza-Noh et al., 2015; Chacón-Sánchez y Martínez-Castillo, 2017; Fig. 1).

El registro arqueológico sugiere que la domesticación de las especies de este género ocurrió posterior a la de otros integrantes de la milpa. Por ejemplo, los restos arqueológicos más antiguos de calabaza, los cuales se encontraron en la cueva de Guilá Naquitz, en Oaxaca, pertenecen a la especie *Cucurbita pepo* y datan de hace alrededor de 10,000 años (Smith, 1997). Por su parte, los restos más antiguos de maíz son granos de almidón localizados en la cueva Xihuatoxtla, en la Cuencas del Balsas, con una edad estimada de 8,700 años (Piperno et al., 2009). Esto concuerda con datos genéticos que ubican temporalmente a la domesticación del maíz hace 9,000 años (Matsuoka et al., 2002).

El registro arqueológico en Mesoamérica de las especies domesticadas del género *Phaseolus* no es tan antiguo como en el caso de la calabaza y el maíz. Los restos de mayor antigüedad de frijol común (*P. vulgaris*) y teparis (*P. acutifolius*) en Mesoamérica fueron encontrados en el Valle de Tehuacán y tienen una edad de 2,500 años. Los siguientes más antiguos provienen del Valle de Oaxaca y datan de hace 2,100 años. De ayocote, en el Valle de

Oaxaca y en Río Zape, Durango, se han encontrado restos de semillas que datan de hace 1,100 años (Kaplan & Lynch, 1998). En la Región Andina de Perú se han descubierto restos de frijoles domesticados de hace 4,400 años, mientras que en los valles costeros de ese mismo país se han hallado restos de *P. lunatus* de 5,600 años (Kaplan & Lynch, 1998). Sin embargo, datos genéticos y genómicos sitúan a la domesticación de estas leguminosas miles de años atrás. Por ejemplo, Schmutz et al. (2014) estimaron que la domesticación del frijol común ocurrió hace alrededor de 8,000 años de forma independiente en Mesoamérica y en la Región Andina.

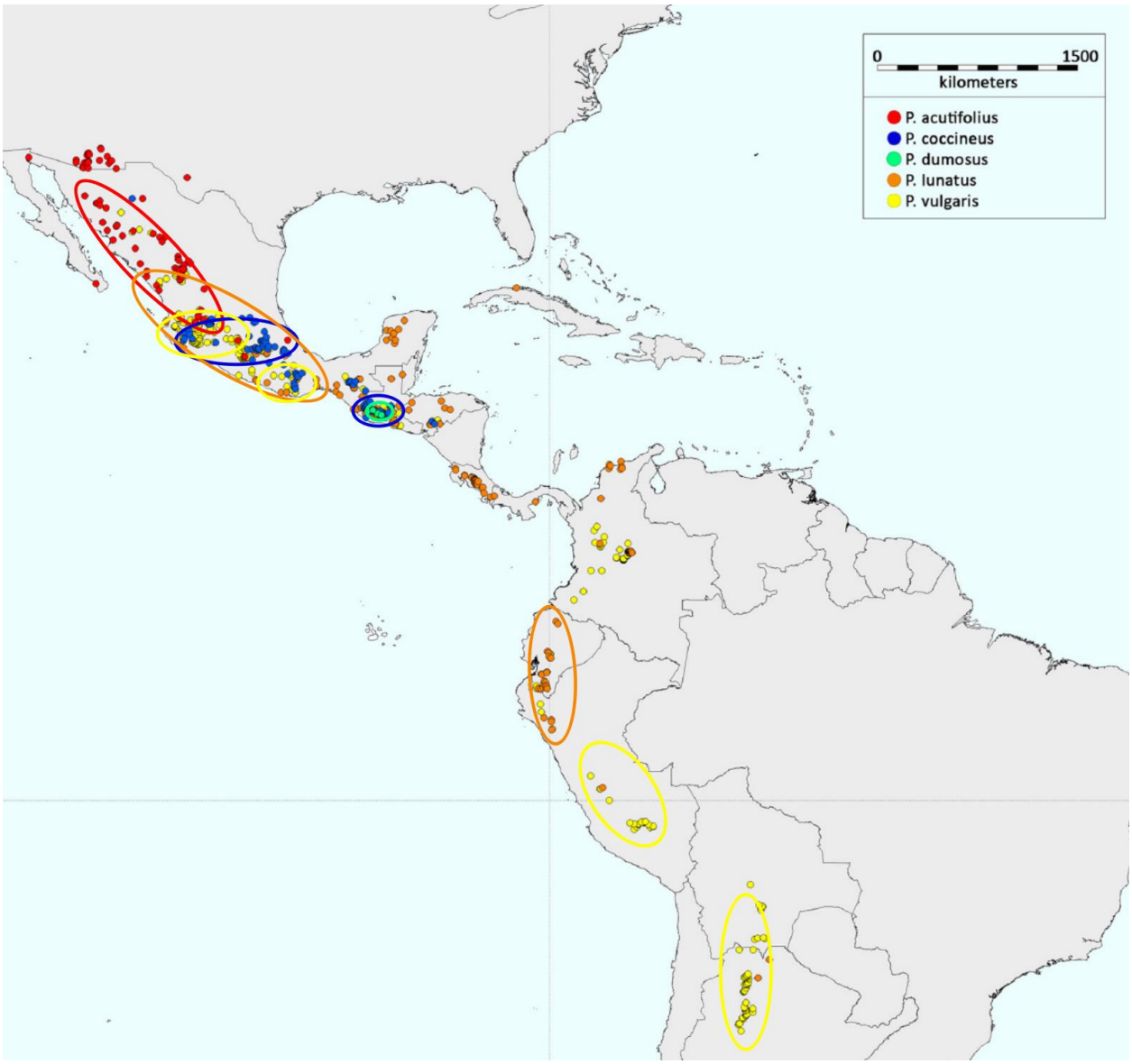


Fig. 1. Distribución de poblaciones silvestres de las cinco especies de frijoles domesticados. Las áreas delimitadas indican los centros de domesticación que han sido propuestos para cada especie. Tomado y modificado de Bitocchi et al. (2017).

Estas especies ofrecen la oportunidad de estudiar y evaluar las bases genéticas del proceso de domesticación al comparar a los frijoles domesticados, y poner a prueba paralelismos y convergencias dentro de una misma especie. Pese a esto, existe una gran diferencia en el cúmulo de información de las distintas especies, siendo el frijol común la especie más estudiada, seguido del frijol ibe o lima (*P. lunatus*; Bitocchi et al., 2017; Chacón-Sánchez & Martínez-Castillo 2017). En contraste, la cantidad de estudios acerca del origen, historia y patrones de diversidad genética en las otras tres especies es menor.

Distribución e historia evolutiva de los frijoles

*Frijol común (*Phaseolus vulgaris)

Las poblaciones silvestres de *P. vulgaris* se distribuyen desde el norte de México hasta el noroeste de Argentina (Fig. 1), y tres pozas genéticas eco-geográficas habían sido descritas. Dos de estas pozas, la mesoamericana y la andina, representan las mayores pozas de la especie y cada una incluye tanto formas silvestres como cultivadas (Bitocchi et al., 2013). La tercera poza génica está constituida por poblaciones silvestres del norte de Perú y Ecuador. Las poblaciones silvestres de México se encuentran subdivididas, y las relaciones filogenéticas entre los grupos muestra un cluster mesoamericano más cercanamente relacionado con la poza andina, y otro grupo mexicano más cercano a las poblaciones del norte Perú y Ecuador. Lo anterior llevó a la conclusión de que cada poza sudamericana se originó a partir de distintos eventos de migración provenientes de México (Bitocchi et al., 2013; Schmutz et al., 2014; Rendón-Anaya et al., 2017b) hace alrededor de 165,000 años (Schmutz et al., 2014). Sin embargo, recientemente Rendón-Anaya et al. (2017a), usando información de secuenciación de genomas completos, sugirieron que el grupo Perú y Ecuador son una especie hermana.

Dos regiones han sido sugeridas como áreas de domesticación en Mesoamérica: Kwak et al. (2009), usando microsatélites, proponen la cuenca del Río Lerma-Río Grande de Santiago, en en centro oriente de México; mientras que los trabajos de Bitocchi et al. (2013) y de Rodriguez et al. (2016) sitúan el inicio de la domesticación en el Valle de Oaxaca. En el caso de la domesticación en la Región Andina, también dos zonas han sido propuestas: centro-sur de Perú (Chacón-Sánchez et al., 2005) y norte de Argentina-sur de Bolivia (Beebe et al., 2001; Rodriguez et al., 2016).

La teoría de genética de poblaciones predice que durante la domesticación ocurre una reducción en la diversidad genética y un incremento en la divergencia entre poblaciones silvestres y cultivadas, lo cual se debe a procesos demográficos que afectan a todo el genoma y

a la selección artificial en ciertos loci. En el caso del frijol común, varios estudios han identificado el decremento de variación genética consecuencia de los cuellos de botella por los que han pasado las poblaciones domesticadas (Kwak et al., 2009; Mamidi et al., 2011; Bitocchi et al., 2013; Bellucci et al., 2014). Esta reducción en la diversidad genética es mayor en los cultivos mesoamericanos que en los andinos (Bitocchi et al., 2013). Además, Bellucci et al. (2014), con información de secuenciación de ARN, mostraron que además del decremento en la variación, los cultivos también presentan una reducción de la expresión génica.

Hasta el momento, *P. vulgaris* es la única especie del género que cuenta con un genoma secuenciado (Schmutz et al., 2014), el cual es el más pequeño de las cinco especies domesticadas de *Phaseolus*, con un tamaño de 587 Mb. Papa et al. (2007), usando 2,506 AFLPs para detectar firmas de selección, estimó que alrededor del 16% del genoma está bajo efectos selectivos. Más recientemente, Bellucci et al. (2014) con información de transcriptomas y simulando la dinámica demográfica, calculó que el 9% de los genes han estado bajo presiones de selección artificial. Finalmente, Schmutz et al. (2014), secuenciado genomas completos, estimaron que alrededor de 74 Mb del genoma de *P. vulgaris* fueron afectadas por la domesticación, pero menos del 10% de estas regiones son compartidas entre los cultivos andinos y mesoamericanos.

*Frijol lima (*Phaseolus lunatus*)*

La segunda especie del género de mayor importancia económica es *P. lunatus* o frijol lima, la cual se distribuye desde el centro de México hasta el norte de Argentina (Fig. 1). Los estudios sobre la historia evolutiva del frijol lima se han basado en espaciadores intergénicos de cloroplasto (*atpB-rbcL*, *trnL-trnF*; Serrano-Serrano et al., 2010), en la secuencia ribosomal 5.8S y en espaciadores internos transcritos (región ITS; Motta-Aldana et al., 2010; Serrano-Serrano et al., 2012), así como en 10 microsatélites (Martínez-Castillo et al., 2014). Los resultados de estos estudios presentan discordancias, especialmente al contrastar conclusiones obtenidas con marcadores nucleares y citoplásmicos.

Al igual que en el frijol común, el frijol lima se domesticó dos veces. Uno de estos eventos ocurrió en la Región Andina y dio origen a los cultivares conocidos como “Lima grande”, y el otro ocurrió en Mesoamérica, a partir del cual surgieron cultivares cuya semilla es de menor tamaño (Motta-Aldana et al., 2010; Serrano-Serrano et al., 2012). Recientemente, Chacón-Sánchez y Martínez-Castillo (2017) realizaron el primer análisis genómico de la especie. Sus resultados también apoyan dos eventos de domesticación.

Análisis de diversidad genética con distintos marcadores moleculares han detectado una severa reducción de variación al comparar poblaciones silvestres y cultivadas del frijol lima, tanto en la poza génica mesoamericana como en la andina (Motta-Aldana et al., 2010; Serrano-Serrano et al., 2010; Serrano-Serrano et al., 2012; Martínez-Castillo et al., 2014; Chacón-Sánchez y Martínez-Castillo, 2017). Este decremento de diversidad parece ser aún más pronunciado en el cloroplasto (Serrano-Serrano et al., 2010).

*Frijol tepari (*Phaseolus acutifolius*) y piloy (*Phaseolus dumosus*)*

Dada la importancia económica del frijol común, es la especie de *Phaseolus* más estudiada, seguida del frijol lima. De las otras tres especies de frijoles domesticados poco se conoce de su historia evolutiva y proceso de domesticación.

El frijol tepari silvestre (*P. acutifolius*) crece desde el centro de México hasta el suroeste de Estados Unidos (Fig. 1). Estudios de esta especie basados en faseolina (Schinkel y Gepts, 1988), isozimas (Garvin y Weeden, 1994), microsatélites (Blair et al., 2012) y 645 SNPs (Gujaria-Verma et al., 2016) indican un solo evento de domesticación y una reducción de diversidad en los cultivos. Además, el trabajo realizado por Gujaria-Verma et al. (2016) muestra que los cultivos de frijol tepari se encuentran divididos en dos grupos genéticos (centro de México y noroeste de México-Sur de Estados Unidos), lo que sugiere que el evento de domesticación fue seguido de la separación de los cultivos en dos regiones geográficas (Gujaria-Verma et al., 2016).

De las especies domesticadas de frijoles, *P. dumosus* presenta el área de distribución más reducida, ya que sus parientes silvestres únicamente crecen en la parte central de Guatemala, donde muy probablemente se originó la especie. Los datos por Schmit y Debouck (1991) de faseolina muestran la presencia de un sólo grupo genético de poblaciones silvestres, lo cual fue corroborado por Mina-Vargas et al. (2016) usando 600 marcadores moleculares. A partir de esta poza génica de Guatemala se domesticó el frijol piloy, y posteriormente se expandió a Chiapas, Oaxaca, Puebla y Veracruz (Schmit y Debouck, 1991).

***Phaseolus coccineus*: Ayocote, botil, xut chenek, xutito, pocohuinic, tacahuaquetl, tecomarí, patol, cimat**

Al igual que muchas especies domesticadas, las poblaciones domesticadas de *Phaseolus coccineus* tienen distintos nombres de acuerdo a la región y características del cultivo. Por ejemplo, el nombre común en el centro del México es ayocote, mientras que en la zona de Chiapas es conocido como botil o shbotil, y en la Sierra Tarahumara es llamado tecomarí. Otros nombres comunes son xut chenek, xutito, pocohuinic, tacahuaquetl, tachena, tukámuli, shashana, tangashtpu, chambotorte, patol, yeguas y cimatl.

Phaseolus coccineus se distribuye en zonas templadas (1,000-3,000 m.s.n.m.) desde Chihuahua en el norte de México hasta Costa Rica (Salinas, 1988). Freytag y Debouck (2002) con base a comparaciones morfológicas proponen dos subespecies: *P. coccineus* subsp. *coccineus* y *P. coccineus* subsp. *striatus*. De acuerdo con esta clasificación, la principal diferencia entre las subespecies es que la primera presenta flores de color rojo, mientras que la segunda posee flores rosas, lilas o moradas. Cabe mencionar que las formas domesticadas corresponden a *P. coccineus* subsp. *coccineus* var. *coccineus*, junto con otras 11 variedades más. Por su parte, dentro de P. *coccineus* subsp. *striatus* se incluye a 8 variedades (Tabla 1).

Tabla 1. Subespecies y variedades de *P. coccineus* descritas por Freytag y Debouck (2002).

Subespecie	Variedades	Características y distribución
P. coccineus subsp. *coccineus*	var *coccineus* var *parvibracteolatus* var *griseus* var *lineatibracteolatus* var *tridentatus* var *splendens* var *strigillosus* var *semperbracteolatus* var *condensatus* var *pubescens* var *argentus* var *zongolicensis*	Plantas trepadoras con flores rojas o escarlata. Los cultivares pertenecen a var. *coccineus* y en raras ocasiones pueden presentar flores blancas. La subespecie *coccineus* presenta una amplia distribución, que incluye zonas montañosas desde Chihuahua hasta Costa Rica.
P. coccineus subsp. *striatus*	var *striatus* var *minuticicatricatus* var *guatemalensis* var *purpurascens* var *rigidicaulis* var *pringlei* var *timilpanesis*	Variedades con flores de color lila o púrpura, en su mayoría de crecimiento postrado, excepto la var. *guatemalensis*. La distribución de *striatus* es más restringida en comparación con la otra subespecie. En México se encuentra principalmente en la Ciudad de México, Morelos y Estado de México. Fuera de México se distribuye en Guatemala y Nicaragua.

Como cultivo, *P. coccineus* es aprovechada en México, Guatemala y Honduras principalmente para autoconsumo por pequeños agricultores, y en menor medida, se comercializa en áreas urbanas. Las partes consumidas con mayor frecuencia son las semillas y vainas inmaduras ("ejotes"), pero también se usan como alimento las flores, el follaje y la raíz ("camotes"; Basurto, 2000). Comúnmente se siembra asociado a maíz y en regiones como la Sierra Norte de Puebla es posible encontrar cultivos con hasta tres especies de frijoles (*P. vulgaris, P. dumosus* y *P. coccineus,* Basurto 2000). En el rango de distribución de la especie, es frecuente que poblaciones silvestres, cultivadas y ferales se encuentren en simpatría, lo que resulta en un gradiente morfológico (Salinas, 1988).

Gracias al intercambio de productos entre el Nuevo y el Viejo Mundo, *P. coccineus* fue llevado a Europa, y se cree que España fue el país a través del cual se introdujo (Zeven et al., 1993). Actualmente el Reino Unido es el país donde más se cultiva fuera de América debido a su capacidad de crecer en temperaturas bajas, donde se le conoce como "scarlet runner bean". (Rodiño et al., 2006).

Diversidad genética del ayocote

A diferencia de *P. vulgaris*, que es una especie autógama, *P. coccineus* presenta un alto porcentaje de entrecruzamiento (Escalante et al., 1994), mediado por abejas y varias especies de colibríes, sin embargo, los polinizadores más importantes son los abejorros (Sousa-Peña, 1992). Este tipo de sistema de apareamiento promueve el mantenimiento de una alta diversidad, la cual fue reportada por Escalante et al. (1994) en poblaciones silvestres y cultivadas de México, quienes realizaron uno de los primeros estudios de diversidad genética, identificando también alta diferenciación entre las poblaciones del centro del país y de Chiapas.

Son escasos los trabajos en esta especie, y pese a que la domesticación de *P. coccineus* ocurrió en Mesoamérica, la variación genética de los cultivos europeos ha sido estudiada más ampliamente: Sicard et al. (2005) y Acampora et al. (2007) estimaron la diversidad en cultivos italianos, registrando una alta variación pese a las pocas accesiones incluídas; Nowosielski et al. (2002) y Boczkowska et al. (2012) emplearon RAPDs y AFLPs, respectivamente, para analizar poblaciones de Polonia, donde registraron una baja variabilidad, probablemente debido a los cuellos de botella por los que habían pasado los cultivares al llegar a este país. Spataro et al. (2011) analizaron accesiones de 10 países europeos y de Mesoamérica con 12 microsatélites, encontrando altos niveles de diversidad tanto en las muestras del Viejo y del Nuevo Continente, lo que sugiere que el cuello de botella por el que

pasaron las poblaciones europeas no fue severo (Tabla 2). Estos resultados fueron apoyados por los de Rodriguez et al. (2013), quienes usaron microsatélites de cloroplasto y aumentaron la cantidad de accesiones (Tabla 2).

Spataro et al. (2011) además encuentra evidencia de alta diferenciación genética entre los cultivos europeos y mesoamericanos, lo cual puede ser resultado de nuevas presiones selectivas, deriva génica y la falta de flujo genético con poblaciones silvestres (Spataro et al. 2011). Dentro de las poblaciones europeas, la diferenciación es baja, y se ha reportado una correlación entre la distancia genética y geográfica, sugiriendo flujo génico mediado por el intercambio de semillas (Rodriguez et al. 2013).

En cuanto a la historia de domesticación de esta especie, empleando microsatélites de cloroplasto, Angioi et al., (2009) detectaron dos grupos genéticos, cada uno contenía tanto accesiones silvestres como cultivadas, sugiriendo múltiples eventos de domesticación de *P. coccineus*. Sin embargo, sólo ocho accesiones silvestres y 11 cultivadas de *P. coccineus* fueron usadas en este trabajo. Rodriguez et al. (2013), con evidencia de seis microsatélites de cloroplasto, proponen dos eventos de domesticación, uno de los cuales ocurrió en la región de Guatemala y Honduras, y un segundo en México. En este caso también pocas accesiones silvestres y cultivadas de México son incluidas (siete y 31 respectivamente), y su trabajo se centra en cultivos de Europa (331 accesiones europeas).

Tabla 2. Estadísticos de diversidad de *P. coccineus*. Los resultados de microsatélites nucleares fueron obtenidos por Spataro et al. (2011), y los de cloroplasto por Rodriguez et al. (2013). Las accesiones de Mesoamérica provienen de México, Guatemala, Honduras y Costa Rica.

	Microsatélites nucleares			**Microsatélites de cloroplasto**		
	Cultivares Europa	Cultivares Mesoamérica	Silvestres Mesoamérica	Cultivares Europa	Cultivares Mesoamérica	Silvestres Mesoamérica
Tamaño de muestra	148	52	28	331	37	12
n alelos	91	101	77	19	16	17
n alelos privados	36	28	29	-	-	-
n haplotipos	-	-	-	29	15	10
n haplotipos privados	-	-	-	26	8	4
H_E	0.37	0.55	0.5	0.29	0.33	0.56

Para poder conocer a mayor profundidad la historia de domesticación de *P. coccineus*, es necesario analizar cultivos y poblaciones silvestres que representen el área de distribución de la especie, y utilizar marcadores moleculares que den una resolución a nivel del genoma completo. Así, los objetivos del presente trabajo son: 1) brindar información acerca de los patrones de diversidad y diferenciación de las poblaciones cultivadas y silvestres de *P. coccineus* en México; 2) generar información acerca de la historia de domesticación de la especie, poniendo a prueba la hipótesis de dos eventos de domesticación; 3) identificar loci candidatos que presenten firmas de selección natural y artificial (domesticación).

En el segundo capítulo se hace una revisión de los métodos, enfoques y herramientas genómicas que se emplean para hacer inferencias acerca de la historia de las especies domesticadas, así como para identificar los genes o regiones afectadas durante este proceso. En el tercer capítulo se abordan las preguntas de *P. coccineus* antes planteadas, utilizando individuos silvestres, cultivados y ferales (242 en total), y usando la herramienta de *Genotyping by Sequencing* (GBS), con la cual se hace un muestreo aleatorio del genoma, obteniendo decenas de miles de marcadores moleculares a lo largo de los genomas de los individuos.

Métodos y análisis en genómica de poblaciones

La caracterización de la historia evolutiva de las especies y la identificación de patrones de adaptación es crucial en campos como la ecología funcional, la conservación, la agronomía y la biología evolutiva. Para poder comprender los procesos evolutivos usando herramientas genómicas son necesarias las bases teóricas de la genética de poblaciones. Sin embargo, en los inicios de la genómica, ambos campos fueron desarrollados de forma independiente. El surgimiento de herramientas de secuenciación de siguiente generación (*Next Generation Sequencing,* NGS) o secuenciación masiva han incrementado la cantidad de marcadores moleculares disponibles, derivando en la creación de herramientas computacionales para su análisis. Estas herramientas frecuentemente están poco unificadas e interconectadas entre sí, lo que puede restringir la cantidad de información que se extrae de los datos.

Gran parte de los esfuerzos para generar y analizar datos del genoma se han centrado en el ser humano, en especies modelo y/o de valor económico. No obstante, es necesario trasladar los métodos y herramientas a especies no modelo para poder entender los procesos evolutivos a distintos niveles y abarcando una amplia diversidad biológica.

Dentro de las estrategias en genómica para el estudio de organismos no modelo las más usadas son las librerías de representación reducida y la resecuenciación de genomas completos (Bourgeois et al., 2016). Las librerías de representación reducida tienen la ventaja de bajar costos y permiten muestrear variantes a lo largo del genoma al secuenciar fragmentos de ADN flanqueantes a sitios de restricción. Dentro de este tipo de técnica se encuentra la secuenciación RAD (*RAD-sequencing*; Baird et al., 2008) y Genotipado por Secuenciación (*Genotyping by Sequencing,* GBS; Elshire et al., 2011). Pese a que no es forzoso contar con un genoma de referencia, el usar uno brinda información de las posiciones de las variantes, del tamaño de haplotipos y desequilibrio de ligamiento (Bourgeois et al., 2016). Por otro lado, la resecuenciación de genomas completos requiere de un genoma de referencia y es más costosa, especialmente en el caso de genomas grandes y complejos. Sin embargo, ofrece una visión completa de las variantes del genoma.

Exploración de la estructura poblacional

Identificar la estructura poblacional es un paso crucial y omitirla puede sesgar inferencias demográficas (Chikhi et al., 2010) o puede llevar a identificar erróneamente variantes bajo selección (Nielsen, 2005). Un método relativamente simple para comenzar a explorar el agrupamiento de los individuos es el Análisis de Componentes Principales (*Principal Component Analysis*; PCA), el cual se basa en analizar la varianza y covarianza entre los genotipos (Bourgeois et al., 2016). Este análisis puede llevarse a cabo en el paquete de R SNPRelate (Zheng et al., 2012) o en el programa PLINK (Chang et al., 2015).

Otro enfoque para inferir la estructura de las poblaciones es usar métodos como Structure (Pritchard et al., 2000) o fastSTRUCTURE (Raj et al., 2014), ambos basados en estadística Bayesiana; o Admixture (Alexander et al., 2009), que usa máxima verosimilitud. Estos métodos determinan la estructura al agrupar a los individuos en *clusters* sin ningún *priori* y además pueden detectar flujo genético. Computacionalmente fastSTRUCTURE y Admixture son más rápidos y permiten analizar bases de datos más grandes.

Para poner a prueba la existencia de poblaciones con una estructura jerarquizada también pueden usarse métodos basados en medidas de diferenciación (por ejemplo, F_{ST}), con los cuales es posible construir hipótesis filogenéticas (Bourgeois et al., 2016). POPTREE (Takezaki et al., 2010) permite usar varias medidas de diferenciación para inferir las relaciones filogenéticas. TreeMix (Pickerell y Pritchard, 2012) construye un árbol filogenético de las

poblaciones con base en la matriz de covarianza de las frecuencias alélicas de las poblaciones y permite rastrear eventos de flujo genético, sin embargo, requiere predefinir las poblaciones. Otros programas usan los datos de cada SNP de forma individual para construir la filogenias, como PhyML (Guindon et al., 2010) o RAxML (Stamatakis, 2014). Usando estadística Bayesiana, SNAPP (Bryant et al., 2012) hace hipótesis filogenéticas con datos de SNPs, sin embargo, requiere una gran cantidad de recursos computacionales y los tiempos de análisis son relativamente largos. Por su parte, Splitstree (Huson y Bryant, 2005) construye redes filogenéticas, permitiendo inferir eventos de flujo.

Inferir la historia de las poblaciones con coalescencia

La teoría de la coalescencia fue inicialmente planteada para modelar la genealogía de los alelos de una muestra tomada de una población. Al ir hacia atrás en el tiempo, los alelos convergen (coalescen) de forma estocástica hasta llegar al ancestro común más reciente (Kingman, 1982). Diversos métodos han usado y enriquecido esta base teórica para inferir la historia de las poblaciones y sus parámetros demográficos asociados, como los tiempos de divergencia, los tamaños efectos o flujo genético (Bourgeois et al., 2016). Dentro de los programas basados en coalescencia más usados se encuentran IMa (Hey y Nielsen, 2007) y Migrate-n (Beerli y Palczewski, 2010). Pese a que estas herramientas son muy poderosas para estimar parámetros demográficos, computacionalmente son costosas y lentas ya que requieren la evaluación completa de la función de la probabilidad asociada a un modelo, proceso que puede ser muy complejo cuando se trata de cientos o miles de marcadores moleculares. Los métodos computacionales de aproximación bayesiana (*Approximate Bayesian Computation,* ABC) evitan este problema al comparar los datos reales con datos simulados por coalescencia bajo escenarios predefinidos, y posteriormente se determina qué escenario explica mejor los datos. Dentro de este tipo de herramientas se encuentran DIYABC (Cornuet et al., 2008) y fastsimcoal2 (Excoffier y Foll, 2011).

Detección de firmas selección

Tanto los procesos demográficos como los selectivos alteran los patrones de diversidad y diferenciación en los genomas. Sin embargo, estos procesos afectan de forma diferencial: los eventos demográficos alteran por igual a todo el genoma, mientras que la selección sólo actúa sobre algunos genes y sus alrededores (Lewontin y Krakauer, 1973). De acuerdo al tipo de

selección que esté actuando, el patrón de diversidad resultante es distinto: la selección balanceadora produce altos niveles de variación; la selección purificadora disminuye la diversidad genética; finalmente la selección direccional incrementa la diferenciación (Charlesworth et al. 1997). Además, las regiones neutrales ligadas o cerca de loci bajo selección pueden ser afectadas por barrido selectivo. El tamaño de la región genómica sujeta al barrido depende de la intensidad de la selección y de la frecuencia de recombinación efectiva (Charlesworth et al. 1997).

Los métodos para identificar regiones que han estado baja selección pueden clasificarse en tres tipos de acuerdo al patrón dejado por la selección que tratan de detectar (Bourgeois et al., 2016): 1) basados en la variación de las frecuencias alélicas y polimorfismos; 2) basados en los patrones de desequilibrio de ligamiento; 3) basados en la reconstrucción de la genealogía de los alelos usando coalescencia.

La selección afecta la diversidad y diferenciación, y es posible analizar los patrones de estos atributos a lo largo del genoma usando estadísticos como la diversidad nucleotídica (π; Nei y Li, 1979) y F_{ST}. Algunos métodos de análisis de selección se basan en detectar valores extremos (*outliers*) de estos estadísticos, por ejemplo BAYESCAN (Foll y Gaggiotti, 2008) y PCAdapt (Luu et al., 2017). Algunos permiten además correlacionar factores ambientes o fenotípicos a las variantes candidatas, como BAYENV (Günther y Coop, 2013) y LFMM (Frichot et al., 2013). Es importante tener en cuenta que procesos demográficos pueden sesgar los resultados, por lo que es indispensable conocer previamente la estructura e historia de las poblaciones (Nielsen, 2005).

Por otro lado, el desequilibrio de ligamiento se incrementa y la diversidad se reduce en regiones cercanas a loci bajo selección, especialmente cuando el proceso selectivo es reciente (Charlesworth et al. 1997). Un grupo de métodos busca identificar esas regiones que muestran un exceso de haplotipos homócigos largos, como la prueba de homocigosis de haplotipos extendidos (*Extended Haplotype Homozygosity*; EHH; Sabeti et al., 2002). Este tipo de análisis requieren de un genoma de referencia para poder conocer la posición de las variantes, así como una alta densidad de marcadores (Bourgeois et al., 2016).

Finalmente están los métodos basados en coalescencia. Una vez que se tiene un locus candidato, es posible usar simulaciones de coalescencia para evaluar la fuerza de la selección y estimar el tiempo del alelo. Dentro de este grupo se encuentran los programas ARGWeaver (Rasmussen, 2014) y BALLET (DeGiorgio et al., 2014).

Capítulo 2:

Current approaches and methods in plant domestication studies

Azalea Guerra-García & Daniel Piñero

Botanical Sciences

DOI: 10.17129/botsci.1209

Review

Current approaches and methods in plant domestication studies

Azalea Guerra-García[1,2]* and Daniel Piñero[1]

Botanical Sciences
95 (3): 345-362, 2017

DOI: 10.17129/botsci.1209

Abstract

Background: The domestication process has left signatures in the genomes of domesticated species. Before the existence of molecular markers, only phenotypic traits could be used in domestication studies and breeding programs, but these approaches required long time and effort. In the last decades, the use of molecular markers dramatically increased, and the development of massive sequencing tools have enable to obtain thousands or even millions of molecular markers.
Questions: The main goal of this review us to bring a general and an integrative perpective of the data, approaches and questions that can be answered using new sequencing tools.
Species study and data description: This work focuses on domesticated plants, comparing genetic and genomic data.
Results: The use of molecular markers in the last decades has increased the efficiency and accuracy of plant breeding, allowing to access information about domestication history and to identify genes affected by domestication. Some patterns have been identified in domesticated species: (1) genetic diversity reductions due to demographic bottlenecks and artificial selection; (2) frequently, mutations related with domestication syndrome preexisted at low frequency in natural populations; (3) accumulation of deleterious mutation; (4) gene flow between wild and cultivated populations; (5) phenotypic convergence usually do not result from molecular convergence. There are several approaches that can be used in massive sequencing tools: de novo genome sequencing, whole genome resequencing, reduction of genome complexity using restriction enzymes, transcriptome analysis and epigenetic studies.
Conclusions: Despite the progress made, enormous challenges remain: storage of large databases; development of fast and accurate methods to evaluate phenotypes; identification of paralogous genes in polyploid species; and the analysis of large and highly diverse genomes.
Keywords: artificial selection, genomics, plant breeding, molecular markers

Métodos y enfoques modernos del estudio de la domesticación

Resumen

Antecedentes: La domesticación es un proceso que ha dejado huellas en los genomas de las especies domesticadas. Anterior al surgimiento de los marcadores moleculares, únicamente se utilizaban atributos fenotípicos para estudiar la domesticación y para seleccionar individuos para el mejoramiento de cultivos. En las últimas décadas, el uso de marcadores moleculares se ha incrementado dramáticamente, y las técnicas de secuenciación masiva han permitido obtener miles, e incluso millones de marcadores
Preguntas: El objetivo de esta revisión es brindar una visión general e integrativa de los distintos tipos de datos, enfoques y preguntas que se pueden contestar usando nuevas herramientas de secuenciación.
Especies de estudio y descripción de datos: Especies de plantas cultivadas, contrastando datos genéticos con datos genómicos.
Resultados: El uso de marcadores moleculares ha incrementado la eficiencia del mejoramiento de cultivos y ha permitido hacer inferencias sobre la historia de la domesticación y los genes involucrados. Se han encontrado algunos patrones en las especies domesticadas: (1) reducción de la diversidad genética, debido a cuellos de botella y selección artificial; (2) frecuentemente, las mutaciones relacionadas al síndrome de domesticación preexistían en poblaciones silvestres; (3) acumulación de alelos deletéreos; (4) flujo genético entre formas cultivadas y silvestres; (5) la convergencia fenotípica comúnmente no resulta de convergencia a nivel molecular. Hay una amplia gama de estrategias que se pueden usar con las herramientas de secuenciación masiva que van desde secuenciar genomas completos, usar enzimas de restricción para reducir la complejidad del genoma, hasta el análisis de transcriptomas y estudios epigenómicos.
Conclusiones: Pese a estos avances, aún quedan grandes retos como el análisis de grandes bases de datos; desarrollo de estrategias para obtener información fenotípica de forma rápida y precisa; identificación de genes parálogos en especies poliploides; análisis de genomas grandes y muy diversos.
Palabras clave: selección, genómica, mejoramiento, marcadores moleculares

Author Contributions
Azalea Guerra-García wrote the paper.
Daniel Piñero reviewed the drafts of the manuscript.
Running head: Approaches and methods in plant domestication studies

[1] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México
[2] Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México
* Corresponding author: azalea.guerra.g@gmail.com

The domestication process and its effect on genomes. During the Neolithic, a major change occurred in the way humans obtained food and resources. Human transitioned from hunters-gatherers to practice agriculture, a sedentary lifestyle, settlement of villages and the creation of ceramic, eventually leaded to the development of hierarchical civilizations (Gepts 2014, Larson *et al*. 2014). During this process, domestication was a key technological tool (Gepts 2014).

Domestication is an evolutionary process, in which humans promote the adaptation of wild species to agroecological niches and to local preferences (Figure 1. Casas *et al*. 1996, Larson *et al*. 2014). The evolutionary trajectory from wild to domesticated population is a complex multistage process, that involves several steps (Figure 2. Meyer & Purugganan 2013):

a) Predomestication: humans begin to purposely plant and look after wild plants with favorable traits.
b) Artificial selection in environments created by humans: at this the stage, human selective pressures promote alleles related with domestication traits.
c) Spread and diversification: propagation and local adaptation to different agroecological and cultural environments. In this stage, cultivated and wild populations diverge and domesticated traits diversify.

Figure 1. Examples of wild and domesticated forms of crops. The first image of each row is the wild relative. **a)** teosinte and maize (*Zea mays*); **b)** chilli pepper (*Capsicum annuum*); **c)** common bean (*Phaseolus vulgaris*); **d)** cotton (*Gossypium hirsutum*). Images taken from CONABIO.and CIAT and CIAT.

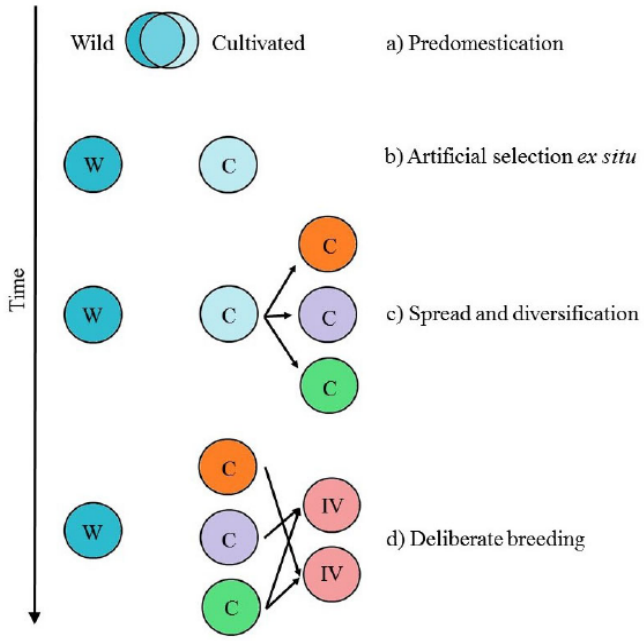


Figure 2. Evolutionary stages of domestication. W represents wild populations, C cultivated populations and IM improved lines. **a)** Predomestication; **b)** Management in environments created and controlled by humans; **c)** Spread and adaptation of cultivated populations; **d)** Deliberate breeding. It is important to notice that these stages are not mutually exclusive and can occur in presence of wild progenitors, increasing the probability of gene flow. Modified from Meyer & Purugganan (2013).

d) Deliberate breeding: this last stage has been practiced intensively in the last century. In classical breeding, controlled crosses between inbred lines are made and then individuals with desirable traits are selected. In the last decades, molecular and genetic tools have been integrated to breeding practices.

These stages are not mutually exclusive and can occur simultaneously, even today. For example, Nahuas communities in Mexico consume wild progenitors of crops like common bean, chili pepper, agave and tomato (in total 68 wild species) using techniques and tools that are available since the Archaic period (Zizumbo-Villarreal *et al*. 2012).

The domestication process affects the genomes of cultivated species and leaves a signature that can be identified. Domestication results in a pronounced alteration in the diversity and differentiation of few genes, while most of the genome reduces its variation as a result of population bottlenecks that occur in the initial predomestication stage (Charlesworth *et al*. 1997, Gepts 2004).

What do domesticated plants have in common? While the domestication process has varied in different plant groups depending on the mating system, growth form, harvested plant parts and economic importance there are several aspects that have been identified to be common to most domestication processes. These include loss of genetic diversity, the genes associated with the domestication process, an accumulation of deleterious variants, the relation between the cultivar and its wild relatives and the possibility of phenotype convergence and its molecular basis in different domestication processes.

Genetic diversity reduction.- One of the most important determinants in crop evolution is the level of genetic diversity contained in the domesticated accessions, especially with reference to the wild ancestral gene pool. Genetic diversity is important as a necessary condition for further evolution in response to selection pressures, not only in the wild but also in breeding programmes (Gepts 2002). Genetic diversity reduction has been widely described in some of the most important crops. It is due to genetic drift resulted from population bottlenecks, and to artificial selection and the consequent selective sweep (Charlesworth *et al*. 1997, Gepts & Papa 2003, Gepts 2014). The domestication process itself, the spread from the center of origin and modern breeding practices, involve population bottlenecks because only some genotypes are selected, promoting a genetic diversity reduction (Gepts & Papa 2003). This pattern has been

reported in numerous species, for example, Mesoamerican common bean (~20 % reduction; Schmutz *et al*. 2014); sunflower (~30 %; Renaut & Rieseberg 2015); soybean (~30 %; Li *et al*. 2013); maize (~17 %; Hufford *et al*. 2012); and in cultivated Agave species (from 21 to 66 %; Eguiarte *et al*. 2013). In sunflower, maize and soybean, the largest diversity decrease occurs between wild populations and landraces; on the contrary, the difference in genetic variation between landraces and improved varieties is low, indicating that breeding practices, at least in these cases, do not have major effects in terms of genetic diversity, at least in these species (Li *et al*. 2013, Hufford *et al*. 2012, Renaut & Rieseberg 2015).

However, exceptions to this pattern have been reported. For example, in carrots, high levels of gene flow between wild and domesticated populations, besides a strong inbreeding depression, have prevented from genetic variation reduction (Iorizzo *et al*. 2013). In the Andean region, cultivated common bean is more genetically diverse than wild populations (Schmutz *et al*. 2014). In central Mexico, managed populations of the cacti *Stenocereus stellatus* (Pfeiff.) Riccob., presents higher levels of genetic variation than unmanaged populations, showing that the management made by humans can potentially alter the genetic diversity of non-domesticated species (Cruse-Sanders *et al*. 2013).

Mutations in genes under domestication-. Identification and isolation of genes that underlie domestication traits provide the opportunity to identify patterns present in these loci. These genes show a wide range of functions, from transcription factors to metabolic enzymes, although many encode similar enzymes or are involved in the same metabolic pathways (Meyer & Purugganan 2013).

According to Meyer & Purugganan (2013), the most frequent functional changes occur in alleles involved in function loss and gene expression. These mutations promote major effects in phenotypes, which usually distinguish domesticated and wild populations. For example, in woodland strawberry (*Fragaria vesca* L.), a 2-bp deletion in the coding region of the *KSN* gene –which is a transcriptional factor– introduces a frameshift, resulting in continuous flowering (Iwata *et al*. 2012). Also, the function loss of *Vrs1* gene in barley (*Hordeum vulgare* L.), changed the inflorescence architecture from two-rowed to six-rowed type (Komatsuda *et al*. 2007).
Mutations in genes under domestication include single nucleotide polymorphisms (SNPs), insertions and deletions (indels), transposon insertions, gene duplication (including gene copy number) and chromosomal rearrangements. The most frequent alterations are SNPs, followed by indels (Meyer & Purugganan 2013).

An important question is whether mutations that lead to domestication phenotypes are new or preexisted in wild populations. This has implications in the nature of selective sweeps and crop species evolution dynamics; for example, a selective sweep caused by a preexisting mutation leaves weaker signatures of selection and allows rapid evolution, because it has a higher frequency than a new mutation (Gepts 2014).

There are some examples of alleles that underlie domesticated traits that appeared recently and are not present in wild populations. That is the case of *LG1* gene in rice, associated with closed panicle (Huang *et al*. 2012) or the *SUN* gene duplication in tomato, which regulates fruit shape (Rodriguez *et al*. 2011). However, there are cases of alleles that preexisted in wild populations at low frequency that were subjected to human selective pressures. For example, the *Brassica oleracea CAL* gene encodes a transcription factor that regulates floral meristem development, and a nonsense mutation leads to the proliferation of floral meristem in domesticated cauliflower and broccoli (Purugganan *et al*. 2000). This mutation is either fixed or present at a high frequency in cultivars, but it is also present at low frequencies in wild populations (Purugganan *et al*. 2000). Another example is the single-stem phenotype of domesticated maize, which is controlled by the dominant *Tb-1* allele. This allele arose before domestication by the insertion of a *Hopscotch* transposon leading to overexpression. The insertion occurred 28,000-23,000 years BP, predating domestication (Studer *et al*. 2011). This suggests that many domesticated traits arise not from new mutations but rather from mutations that segregate in wild population of crop species.

Accumulation of deleterious mutations.- Mutations can have several effects on fitness that range from lethal to neutral and advantageous, but most of these new variants in coding regions are

expected to be deleterious, because they may alter phylogenetically conserved sites or can cause loss of protein function (Kono *et al*. 2016). Under mutation-selection balance and in sexually reproducing species, the accumulation of deleterious alleles -- particularly strongly deleterious and with dominant effects -- is infrequent, because recombination brings together deleterious variants, resulting in unfit genotypes that are eliminated from the population, purging deleterious mutations by purifying selection. Weakly deleterious variants on the other hand, can be potentially maintained under some circumstances, like population size reduction and/or inbreeding, reducing the effective recombination rate and promoting the accumulation of deleterious mutations (Morrell *et al*. 2011, Renaut & Rieseberg 2015).

Besides, as a species expands into new environments (either natural or artificial), genetic drift could increase, due to the reduction of effective population size and a posterior fast growth rate. During the process of domestication and improvement, populations undergo multiple bottlenecks, accompanied by strong artificial selection and the relaxation of selective pressures on traits important in the wild (Morrell *et al*. 2011). These bottlenecks increase the probability that a deleterious variant increases its frequency and gets fixed (Mezmouk & Ross-Ibarra 2014). Also, the linkage between desirable beneficial and deleterious mutations may reduce the efficiency of selection to eliminate the latter, contributing cumulatively to fitness reduction and probably constraining crop yield (Renaut & Rieseberg 2015).

The selective and demographic processes of domestication have led to three hypotheses about the patterns of deleterious mutations (Kono *et al*. 2016):

a. There will be more deleterious alleles in domesticated than in wild relatives.
b. Deleterious variants will be enriched near loci that have been under artificial selection.
c. There will be less deleterious mutations in elite cultivars than in landraces due to strong selection for yield.

There is evidence that supports these hypotheses, as it is estimated that 20 % of non-synonymous variants in rice (Lu *et al*. 2006) and in maize (Mezmouk & Ross-Ibarra 2014) present deleterious effects. In sunflower, landraces and elite lines possess more non-synonymous SNPs compared to wild relatives, particularly regions with low recombination rates and less deleterious mutation, are present in elite cultivars compared to landraces (Renaut & Rieseberg 2015). Also, hundreds of deleterious mutations have been identified in barley and soybean cultivars, being non-sense mutations the least frequents (Kono *et al*. 2016).

Recent advances in sequencing technologies permit to use bioinformatics tools to examine the patterns of deleterious variants in populations. Pure bioinformatic approaches use sequence conservation to infer variants with a significant probability of being deleterious. As a consequence, the identification and elimination of these alleles can potentially provide a complementary approach to breeding practices if identified variants are truly deleterious (Kono *et al*. 2016).

Gene flow between domesticated plants and their wild relatives.- Gene flow between crops and their wild relatives is common, since these two types of population coexist in sympatry, and in most cases, crops and their wild progenitors belong to the same biological species (Gepts 2014). The potential consequences of gene flow are diverse. For example, gene flow between wild and domesticated populations can result in diversity recovery in cultivated forms, as has been described in maize due to continuous introgression with teosinte (Hufford *et al*. 2012). This is also the case in grapevine during cultivars spread through Europe (Myles *et al*. 2011). Also, emergence of weeds has been reported, as occurred in hybrid rice populations from China (Jiang *et al*. 2011, Xia *et al*. 2011) and the United States (Olsen *et al*. 2007).

Despite gene flow, wild and domesticated plants remain phenotypically distinct, at least with respect to domesticated syndrome traits, probably because of human selective pressures on cultivated populations, and natural selection of wild forms. For example, gene flow among wild and cultivated squash taxa have been detected in Mexico (Montes-Hernández & Eguiarte 2002). But farmers do not select seeds from the individuals that present intermediate morphological characters for seed stock (Montes-Hernández *et al*. 2005).

Following the definition in its broad sense, introgression is the transfer of genes between genetically distinguishable populations (Rieseberg & Carney 1998). Introgression can take place between populations of wild species (wild–wild) that are related and between a cultivated spe-

cies and its close wild relatives (crop–wild; Rieseberg & Carney 1998). Nevertheless, introgression is not uniform across the genome and it is unlikely to occur on regions close to genes related with domestication (Gepts 2014). Gene flow may also be asymmetric, although the direction is not consistent: for example, in common bean it mainly occurs from domesticated to wild types (Papa & Gepts 2003), but the opposite happens in maize (from *Zea mays* L. subspecies mexicana teosinte to cultivated maize; Hufford *et al*. 2013).

Introgression may produce profound effects on domestication evolutionary trajectories. The evolutionary history of rice (*Oryza sativa* L.) is an example: Molina *et al*. (2011) suggest that rice domestication occurred in China, where *japonica* variety arose. Later, *indica* variety was originated by a hybridization event between *japonica* and a putative *indica* protoform or with *O. ryfipogon* from South Asia. This resulted in the introgression of *japonica* domesticated genes.

Another case that illustrates the impact of gene flow occurs in domesticated citrus species. Wu *et al*. (2014) sequenced and compared the genomes of some *Citrus* species and showed that cultivated types derive from two progenitor species: introgression from *C. maxima* (Burm.) Merrill to ancestral mandarin species *C. reticulata* Blanco originated tangerine; sweet orange is the offspring of previously admixed individuals of these two species; and sour orange is an *F1* hybrid of pure *C. maxima* and *C. reticulata* parents. The exception is pomelo, which is the result of selection on *C. maxima* (Wu *et al*. 2014).

Phenotypic convergence.- The Law of Homologous Series, formulated by Vavilov (1922), and 'analogous variation' term used by Darwin, are based on the observation that similar phenotypes were selected during domestication. This leads to the question: phenotypic parallelisms result from molecular convergences? In other words, selection on particular traits affects the same genes in different species? If this is true, parallelisms potentially can be used in breeding programs focusing on regions previously identified to be related with important domestication traits (Pickersgill 2009). However, contradictory data has been found and there is evidence for and against this. For example, Grube *et al*. (2000) found five clusters of resistance genes common to potato, tomato and chili pepper, and seven clusters common to two of these three crops; they concluded that the position of resistance genes, is conserved in Solanaceae. However, they found only two examples in which these homologous genes controlled resistance to the same disease. Therefore, knowing the sequence of alleles and the structure of their products can be useful in the development of new pathogens or pest control methods (Pickersgill 2009). Another example is the *Shattering gene* (*Sh1*), which avoids dehiscence in maize, rice, sorghum and common bean; furthermore, there is evidence that Sh1 independently emerged three times in sorghum (Lin *et al*. 2012, Schmutz *et al*. 2014).

On the other hand, evidence against molecular convergence in crops includes two cotton species (*Gossypium hirsutum* L. and *G. barbadense* L.), which were independently domesticated, but selection by humans acted in different genetic components that control fiber development (Hovav *et al*. 2008). Another case where parallelism in phenotypes does not correspond at the gene level occurs between common bean and pea: *p* and *v* genes are partially responsible of dehiscence inhibition in common bean, whereas *dpo1* and *dpo2* genes control this trait in pea, despite *p* and *v* genes are also present (Weeden 2007). Also, in common bean –which was independently domesticated in South America and Mesoamerica–Schmutz *et al*. (2014) identified 1,835 candidate genes under artificial selection in Mexican populations, and 748 in Andean ones, however only 59 were shared genes between the two gene pools. Probably, in some cases, genetic architecture behind domestication is so diverse that similar phenotypic changes are due to selection on different genes.

Using molecular markers to study domestication and advance breeding programs. Before the development of molecular markers, farmers and breeders only had phenotypic traits to select desirable individuals to interbreed. Nevertheless, relative long periods of time and several generations were required to evaluate and select useful genotypes. Some decades ago, molecular markers started to be used in breeding programs and in the development of new cultivars or varieties (Kim *et al*. 2015).

Following the first molecular markers tools (He *et al*. 2014), Sanger DNA sequencing al-

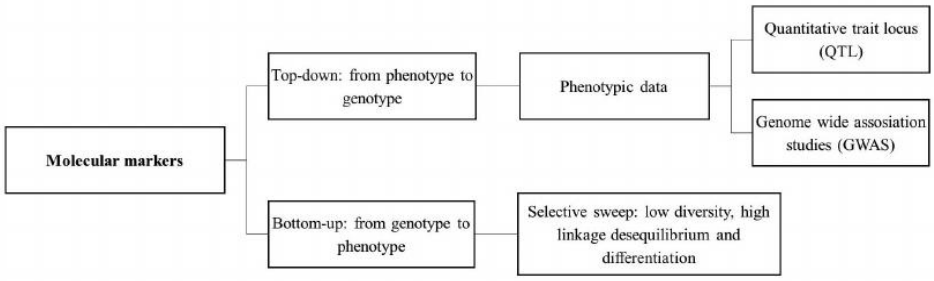


Figure 3. Overview of massive parallel sequencing (NGS) applications in crop genetics and breeding.

lowed the identification of variants at the single base pair resolution. However, sequencing small genomes or large regions from complex genomes using Sanger technique is expensive and laborious. The Human Genome Project encouraged the development of techniques for sequence whole genomes at low cost and less time-consuming. These sequencing tools are known as massive parallel sequencing (MPS), also called Next Generation Sequencing (NGS) and have allowed the routine use of single nucleotide polymorphisms (SNPs) in the last decade, even in no model species (Escalante *et al*. 2014). These massive sequencing tools are used by breeders to sequence large populations, study genetic composition of crop varieties, understand evolutionary relationships between cultivars and wild relatives, identify genotype-genotype (G × G) and genotypes-environment (G × A) interactions, and increase the resolution of gene and quantitative trait locus (QTL) discovery, providing the basis for modeling complex genotype-phenotype relationships at the whole-genome level (Figure 3. Cobb *et al*. 2013, Varshney *et al*. 2014).

Several methods are used to identify relationships between genes or genomic regions and domestication syndrome traits. These can be classified in *bottom-up* and *top-down* approaches (Figure 4).

Top-down domestication studies.- Classical studies to infer the genetic basis of domestication traits use a top-down approach, especially when used in mapping QTLs. Traditional QTL mapping studies begin with two inbred parental lines that differ in their trait values. After creating an F_1 generation by crossing the parental lines, F_1 individuals are crossed to create an F_2 mapping population. These individuals are measured for the trait of interest and genotyped at markers that segregate between parental populations. QTLs are then identified by correlations between trait value and genotype at a locus. The ability to identify a QTL using this method depends on the scale of recombination in the mapping population. For this reason, recombinant inbred lines (RILs) are often used to increase mapping resolution (Mackay *et al*. 2009).

On the other hand, genome-wide association studies (GWAS) utilize association mapping, also known as linkage disequilibrium (LD) mapping, to map QTLs by taking advantage of historic LD to identify statistically significant phenotype-genotype associations (Varshney *et al*. 2014). Unlike classical methods to identify QTLs, in GWAS population mapping is usually

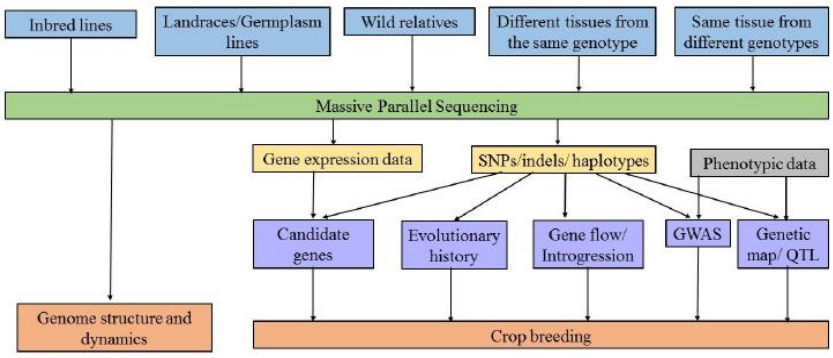


Figure 4. Approaches used to identify genes or genomic regions under selection: Top-down *vs*. bottom-up.

an outbred population and the markers used are SNPs. In general, the results are better than the ones using RILs. However, the power to detect a QTL depends of the phenotypic variance that is explained by the QTL that is determined by its effect and allele frequency. Because of this, GWAS generally has low power to detect QTL variants with small effect or that are rare in the population (Mackay *et al*. 2009). Nevertheless, GWAS have been successfully performed in several crops, and the use of massive sequencing makes possible to genotype larger populations with a higher density of markers (tens to hundreds of thousands of markers are needed in GWAS), increasing mapping resolution (Varshney *et al*. 2014, Andrews *et al*. 2016).

Results of the top-down approach have shown that many traits are controlled at least in part by a few major genes, that genetic effects predominate over environmental ones, and that some genes are linked. It is likely that this approach underestimates the number of genomic regions involved in domestication syndrome. For example, Hufford *et al*. (2012) estimated that only ~3 % of maize genome have been affected by domestication using top-down approach.

The disadvantages of the top-down approach include that phenotypic information is required, and that only few traits can be considered at the same time (Tang *et al*. 2010). A recent study using a top-down approach was performed in rice, in which 203 varieties with well-characterized phenotypes were sequenced, and three traits were mapped (Wang *et al*. 2016). For amylose content and seed length traits, this approach leads to direct identification of the previously identified causal SNPs as the major-effect loci. For pericarp color, they identified a new major-effect locus. Although previously known loci can explain color variation in the varieties of two main subspecies of Asian domesticated rice, *japonica* and *indica*, the new locus identified is unique to another domesticated rice subgroup, aus, and together with existing loci, can fully explain the major variation in pericarp color (Wang *et al*. 2016).

Bottom-up domestication studies.- Compared to classic genetic studies, in which relatively few markers are used, genomic approaches provide additional information about the diversity and LD patterns at the genome level. Therefore, genomics provides a broader view across taxa and across individual genomes (Gepts 2014). The bottom-up approach consists in population genetic screenings of genome-wide diversity, including sequence diversity departures from neutrality, scans for selective sweeps or highly divergent regions (F_{ST}). Further research, involving functional genetics, is then needed to directly establish a causal effect between candidate genes or alleles and phenotypes.

Comparison of sequence diversity combined with neutrality tests in wild and domesticated populations can identify regions that were affected by domestication. Furthermore, comparisons between landraces and wild forms focus specifically on the effect of initial domestication, whereas comparisons between landraces and improved cultivars measure the effect of subsequent selection, including modern breeding (Tang *et al*. 2010, Gepts 2014). The bottom-up approach does not require previous phenotypic data, allowing the identification of a larger amount of candidate genes or genomic regions under selective pressures, unlike the top-down approach which can only consider few genes or regions (Tang *et al*. 2010). For example, with a top-down approach, in which only ~3 % of the maize genome was estimated to be affected by domestication, using a bottom-up approach 1,764 genes (representing ~6.6 % of its genome) were inferred to be related to domestication and 1,508 genes (~5.6 %) related to breeding (Hufford *et al*. 2012).

An example of selective sweep detection using a bottom-up approach includes sorghum, where selective sweeps that encompass genes for starch synthesis enzymes, seed shattering, plant height and time to maturity, were detected using 1,000 accessions and ~265,000 SNPs (Morris *et al*. 2013). Another example was found in rice, where by resequencing 1,529 wild and domesticated accessions, 55 selective sweeps were detected, including *Bh4* (hull color), *PROG1* (tiller angle), *sh4* (seed shattering), *qSW5* (grain width) and *OsC1* (leaf sheath and apiculus color; Huang *et al*. 2012).

Challenges to overcome using a bottom-up approach include that the phenotypic traits affected by the genes or regions encompassed in selective sweeps may be and remain unknown, and that demographic events may hide selective signatures because they alter diversity patterns, and standard test typically assume that populations evolve according to the idealized Wright-Fisher model, with panmictic populations of constant size (Ross-Ibarra *et al*. 2007). When these

assumptions are inaccurate, because domestication involves genetic bottlenecks followed by demographic expansion, tests to detect selection can be inaccurate (Ross-Ibarra *et al*. 2007).

Advantages of using molecular markers.- There are two main types of genomic-assisted breeding: marker-assisted selection (MAS) and genomic selection (GS). MAS, which includes marker-assisted back-crossing (MABC), uses molecular markers that map within specific genes or QTLs known to be associated with target traits or phenotypes to select individuals that carry favorable alleles and/or discard those that do not (Mackay *et al*. 2009).

GS uses all available marker data for a population as predictor of breeding value. GS integrates marker data from a training population with phenotypic and when available, pedigree data to generate a prediction model. The model output genomic estimated breeding values (GEBVs) for all genotyped individual, which are used as a predictor of how well a plant will perform as a parent for crossing (Varshney *et al*. 2014).

The advantage of using molecular markers in breeding programs is that genotypic data obtained from a seed or seedling can be used to predict the phenotypic performance of mature individuals without the need for extensive phenotypic evaluation over long periods of time, allowing for more selection cycles and greater genetic gain per unit of time (Varshney *et al*. 2014). Previously, marker data were expensive per data point, and laborious to generate, and MAS were constrained by the number of available markers. As a result, only markers in important genome regions were utilized to predict the presence or absence of agriculturally valuable traits. By contrast, the use of massive parallel sequencing technologies provides genome-wide markers coverage at a low cost per data point, allowing to assess the inheritance of the entire genome with nucleotide-level precision (Cobb *et al*. 2013, Varshney *et al*. 2014).

Today, studies are no longer limited by our ability to genotype large populations or by the number of markers, but rather by the high cost and low throughput of phenotypic strategies for traits of interest and in environments relevant to plant breeding (Cobb *et al*. 2013).

How genomic markers help understand the history of domesticated species?.- Molecular markers that are not affected by selection bring information about demographic histories, including the history of the domestication process (Tang *et al*. 2010). Earlier models of domestication proposed a single domestication event and suggested that domestication occurred through strong selection and severe genetic bottleneck, which resulted in reproductive isolation between wild and cultivated populations (Meyer & Purugganan 2013). Then, models integrated gene flow and introgression between cultivated and wild relatives, as occurred in grape (*Vitis vinifera* L. subsp. *vinifera*), that emerged in Near East region, but evidence of introgression from the wild progenitor (*Vitis vinifera* subsp. *sylvestris*) during Europe introduction was found (Myles *et al*. 2011). High levels of genetic diversity and fast decay of LD in *vinifera* suggest a weak bottleneck at the beginning of domestication process, followed by thousands of years of vegetative propagation (Myles *et al*. 2011).

There are some cases in which multiple domestication events occurred. For example, resequencing common bean genomes -- including wild and domesticated forms -- confirmed two independent domestication events, one occurred in Mesoamerica and the other one in the Andean region. Besides, these two gene pools diverged 165,000 years ago, long before domestication started (Schmutz *et al*. 2014). Finally, an alternative domestication model proposes that crops are domesticated from interspecific hybridization, sometimes followed by clonal propagation. This is common in tree crops, as is the case of species of the genus *Citrus*, previously discussed. New groups of bananas and plantains developed when diploid domesticated bananas (genome AA) spread into the range of wild *Musa balbisiana* Colla (genome BB), producing the AAB and ABB triploids (Heslop-Harrison & Schwarzacher 2007).

How is genomic data obtained? Several massive parallel sequencing platforms are commercially available, and while the specific methods vary, they obtain millions of short DNA sequence reads from random locations in the genome (Escalante *et al*. 2014). Massive sequencing technologies have been widely used for de *novo* sequencing and several crop and other angiosperm genomes have been sequenced, which are used as reference in no model species studies

(Table 1). Some of the genome assemblies are in draft stage and extensive work is ongoing in the direction of closing the gaps and re-sequencing. In addition to the genome sequence, transcriptomes and expressions profiles are also available for many crops. The large genome size and polyploidy exhibited by many crop species make more difficult the sequencing and further analysis. In some allopolyploid crops, the genomes of progenitors species were first sequenced and then were used to assemble the polyploid genomes of the domesticated forms. This strategy was used in cotton (Wang *et al*. 2012, Li *et al*. 2014, Li *et al*. 2015), strawberry (Hirakawa *et al*. 2014) and peanut (Bertioli *et al*. 2016).

Several strategies to construct genomic libraries (collection of DNA that will be sequenced) have been developed. When the genomic library consists in fragments from whole genomes, it is called whole genome resequencing (WGRS). This approach has been used in several crop species, as common bean (Schmutz *et al*. 2014), rice (Huang *et al*. 2012, Wang *et al*. 2016), maize (Hufford *et al*. 2012, Xu *et al*. 2014); and soybean (Li *et al*. 2013) to identify selection signatures and to infer the domestication history. Nevertheless, WGRS is relatively expensive, particularly in large genomes or polyplolid species, reducing the number of samples that can be multiplexed (Kim *et al*. 2015).

In the last few years, different strategies to genotype a larger number of samples at reduced costs have been developed, including microarrays and methods to construct reduced representation libraries (RRLs). The cost per SNP using microarrays is relatively low but it is necessary to know the sequence of the variants (SNPs) that will be used (Bolger *et al*. 2014). Construction of RRLs, also known as restriction-associated DNA sequencing (RAD-seq), is a method for sequencing loci adjacent to restriction cut sites across the genome. RAD-seq targets a subset of the genome, therefore allow multiplex a higher number of samples without prior genomic information (Elshire *et al*. 2011, Andrews *et al*. 2016). The use of these

Table 1. Examples of published sequenced crop genomes. The statistics for each genome are taken from the publication, despite several model plant genomes have had significant updates to genome assemblies and gene counts.

Species	Common name	Genome size (Mb)	Assembly size	Ploidy and chromosome number	# predicted genes	Reference
Domesticated in Mexico						
Amaranthus hypochondriacus	Amaranth	466	377	2n = 32	23,059	Clouse *et al*. 2016
Capsicum annuum	Hot pepper	3,480	3,060	2n = 24	34,903	Kim *et al*. (2014)
Carica papaya	Papaya	372	271	2n = 18	24,746	Ming *et al*. (2008)
Gossypium hirsitum	Cotton	2,250-2,430	2,173	2n = 4x = 52	76,943	Li *et al*. (2015)
Phaseolus vulgaris	Common bean	587	473	2n = 22	27,197	Schmutz *et al*. (2014)
Theobroma cacao	Cocoa	430	327	2n = 20	28,798	Argout *et al*. (2011)
Zea mays	Maize	2,300	2,048	2n = 20	32,540	Schnable *et al*. (2009)
Domesticated in America						
Anana comosus	Pineapple	526	382	2n = 50	27,024	Ming *et al*. (2015)
Chenopodium quinoa	Quinoa	1,450-1,500	1,390	2n = 4x = 36	44,776	Jarvis *et al*. (2017)
Gossypium barbadense	Cotton	2,470	1,395 A subgenome 776 B subgenome	2n = 4x = 52	77,526	Liu *et al*. (2015)
Manihot esculenta	Cassava	770	532	2n = 36	30,666	Prochnik *et al*. (2012)
Nicotina tabacum	Tobacco	4,500	3,700	2n = 4x = 48	90,000	Sierro *et al*. (2014)
Solanum lycopersicum	Tomato	900	760	2n = 24	34,727	The Tomato Genome Consortium (2012)
Solanum tuberosum	Potato	844	727	2n = 12	39,031	The Potato Genome Sequencing Consortium (2011)
Vaccinium macrocarpon	Cranberry	470	420	2n = 12	36,364	Polashock *et al*. (2014)
Other important economic crops						
Glycine max	Soybean	1,115	950	2n = 20	46,430	Schmutz *et al*. (2010)
Oriza sativa indica	Rice	466	429	2n = 24	46,022-55,615	Yu *et al*. (2002)
Oriza sativa japónica	Rice	420	390	2n = 24	37,544	Goff *et al*. (2002)
Sorghum bicolor	Sorghum	730	698	2n = 20	27,640	Paterson *et al*. (2009)
Triticum aestivum	Bread wheat	17,000	3,800	2n = 6x = 42	90,000-94,000	Brenchley *et al*. (2012)

tools has been intensified to genotype large populations, for both top-down and bottom up approaches. For example, the 539 inbred lines of maize found in the International Maize and Wheat Improvement Center (*Centro Internacional de Mejoramiento de Maíz y Trigo*, CIMMYT) have been genotyped using genotyping by sequencing (GBS) to know the genetic diversity and structure (Wu *et al*. 2016).

To design a genomic study it is necessary to consider the size and complexity of the genome, and if a reference genome is available. Besides, the amount of markers that will be identified depends on the approach used to construct the library (WGRS, RAD-seq or microarray). This must be considered within the context of the biological question. For instance, to infer neutral processes, as demographic histories, it only requires from hundreds to a few thousands of molecular markers. If the goal is to identify selection signatures or functionally characterize regions, dozens of thousands to hundreds of thousands of markers are required; for example, for GWAS or gene mapping, at least hundreds of thousands of SNPs are needed (Andrews *et al*. 2016).

The functional aspect of genomics in domestication studies. The goal of functional genomics is to understand the complex relationships between the genome and the phenotype. This involves dynamic processes as gene transcription, translation, regulation of gene expression and protein interactions. These processes comprise a number of *-omics* approaches, such as transcriptomics (gene expression), proteomics (protein production), and metabolomics (characterization of metabolic products; Huang *et al*. 2016). Recently, some studies have added the functional genomics aspect of domestication to assess how many and which genes show differences in expression when wild and domesticated forms are compared.

From the *-omics*, transcriptomics has received more attention (Huang *et al*. 2016). The transcriptome is the set of messenger RNA molecules in one cell or a population of cells, and provides information about expression and genic regulation. RNA-seq (RNA sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses massive sequencing to reveal the presence and quantity of RNA in a sample at a given moment in time. RNA-seq is an alternative of genome reduction representation, that can be used to identify differential expressed genes, alternative splicing (mechanism by which different forms of mature mRNAs are generated from the same gene), and genetic regulatory networks, as well as for annotation, and gene and markers discovery (Andrews *et al*. 2016, Huang *et al*. 2016). Transcriptomics may complement genomic studies, because RNAseq focused on coding regions, while genomics approaches, as RADseq, include coding and noncoding regions (Andrews *et al*. 2016).

For example, in developing cotton fiber, ~15 % of 1,300 proteins are significant up or down-regulated (Hu *et al*. 2013). Most of the changes took place in the early developmental stages, which is consistent with human selection for earlier activation of fiber elongation in domesticated types. Nevertheless, there were a few changes that overlapped between transcripts and proteins, probably because protein abundances depend on multiple factors, as translation, post-translational modifications and degradation processes (Hu *et al*. 2013).

In maize, Swanson-Wagner *et al*. (2012) detected hundreds of genes whose expression patterns were altered during domestication, some of them are involved in biotic and abiotic stress responses. More recently, Huang *et al*. (2016) studied the transcriptomes of six teosinte accessions and found that approximately 75 % of the genes were highly conserved between maize and teosinte. Moreover, they also found 1,516 unigenes (set of transcripts located at the same loci) that were specifically expressed in teosinte, and identified 99 unigenes with strong selection signals, of which 57 might be under strong selection during maize domestication and improvement. This kind of functional studies allows an integrative understanding of the genomic effects that domestications has on species.

Conclusions

The domestication process and breeding have modified the genomes of the plant species that we consume on a daily basis. The recent development of massive parallel sequencing tools allows us to access to information from whole genomes, accelerating identification of genes affected by domestication and to correlate domestication syndrome traits with their molecular basis.

Because many crop species represent a good model for evolutionary studies and have high economic value, there are detailed genetic studies of the most important crops and their wild relatives. From these studies, general patterns have been inferred and in some species, at least partially, the dynamics of the evolutionary process associated to domestication and diversification have been elucidated. Nevertheless, it is necessary to expand our knowledge to no model species, particularly to perennial plants, including tree crops, which have received far less attention. Besides, more functional genomics studies are needed to have an integrative knowledge of the changes that have affected crop species.

Despite major advances of the last years in genomics, important challenges still remain. One important challenge is the analysis of big databases that are generated because computer clusters are frequently needed. Also, the analysis of large, complex and diverse genomes as is the case of conifers is still a relevant problem. We also need to develop efficient approaches to include and analyze repetitive regions. In addition, the identification of paralogues, and the study of polyploid species, remain as some of the most critical problems when analyzing massive sequencing data. Besides, in GWAS, phenotyping is a major operational bottleneck that limits the power and resolution of genetic analysis. There is an urgent need for high throughput, cost-effective, and precise phenotyping methods.

Given the importance of wild progenitors to analyze changes derived from domestication, more attention should be paid to the genetic diversity and adaptation of the wild forms. Besides, in the context of climate change, this approach facilitates the development of climate-tolerant cultivars. In order to achieve these goals, it is necessary to integrate training and research across scientific fields, including genetics, plant breeding, molecular biology, evolution, statistics and bioinformatics.

Acknowledgements

We thank Luis E. Eguiarte for encourage us to make this review, and Alicia Mastretta for reviewing the manuscript. AGG acknowledge the scholarship provided by Consejo Nacional de Ciencia y Tecnología (CONACyT) and thanks the Posgrado en Ciencias Biológicas, from Universidad Nacional Autónoma de México (UNAM). This work is supported by CONACyT PDCPN2014-01, project 247730.

Literature cited

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* **17**: 81-92. DOI:10.1038/nrg.2015.28

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, Boccara M, Risterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C. 2011. *The genome of Theobroma cacao*. *Nature Genetics* **43**: 101-108. DOI: 10.1038/ng.736

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, Liu X, Gao D, Clevenger J, Dash S, Ren L, Moretzsohn MC, Shirasawa K, Huang W, Vidigal B, Abernathy B, Chu Y, Niederhuth CE, Umale P, Araújo AC, Kozik A, Kim KD, Burow MD, Varshney RK, Wang X, Zhang X, Barkley N, Guimarães PM, Isobe S, Guo B, Liao B, Stalker HT, Schmitz RJ, Scheffler BE, Leal-Bertioli SC, Xun X, Jackson SA, Michelmore R, Ozias-Akins P. 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics* **48**: 438-446. DOI: 10.1038/ng.3517

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF. 2014. Plant genome sequencing – application for crop improvement. *Current opinion in Biotechnology* **26**: 31-37. DOI: 10.1016/j.copbio.2013.08.019

Brenchley R, Spannag M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan

Hu G, Koh J, Yoo M, Grupp K, Chen S, Wendel JF. 2013. Proteomic profiling of developing cotton fibers from wild and domesticated *Gossypium barbadense*. *New Phytologist* **200**: 570-582. DOI: 10.1111/nph.12381

Huang J, Gau Y, Jia H, Zhang Z. 2016. Characterization of the teosinte transcriptome reveals adaptive sequence divergence during maize domestication. *Molecular Ecology Resources*. DOI: 10.1111/1755-0998.12526

Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Lui K, Lu H, Li W, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497-501. DOI: 10.1038/nature11532

Hufford MB, Lubinsky P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. 2013. The genomic signatures of crop-wild introgression in maize. *PLOS Genetics* **9**: e1003477. DOI: 10.1371/journal.pgen.1003477

Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner RA, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J. 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**: 808-811. DOI: 10.1038/ng.2309

Iorizzo M, Senalik DA, Ellison SL, Grzebelus D, Cavagnaro PF, Allender C, Brunet J, Spooner DM, Van Deynze A, Simon PW. 2013. Genetic structure and domestication of carrot (*Daucus carota* subsp. *sativus*) (Apiaceae). *American Journal of Bonaty* **100**: 930-938. DOI: 10.3732/ajb.1300055

Iwata H, Gaston A, Remay A, Thouroude T, Jeauffre J, Kawamura K, Oyant LH, Araki T, Denoyes B, Foucher F. 2012. The *TFL1* homologue *KSN* is a regulator of continuous flowering in rose and strawberry. *The Plant Journal* **69**: 116-125. DOI: 10.1111/j.1365-313X.2011.04776.x

Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, Ohyanagi H, Mineta K, Michell CT, Saber N, Kharbatia NM, Rupper RR, Sharp AR, Dally N, Boughton BA, Woo YH, Gao G, Schijlen EGWM, Guo X, Momin AA, Negrão S, Al-Babili S, Gehring C, Roessner U, Jung C, Murphy K, Arold ST, Gojobori T, van der Linden CG, van Loo EN, Jellen EN, Maughan PJ, Tester M. 2017. *The genome of Chenopodium quinoa*. *Nature* **542**: 304-312. DOI :10.1038/nature21370

Jiang Z, Xia H, Basso B, Lu B-R. 2011. Introgression from cultivated rice influences genetic differentiation of weedy rice population at local spatial scale. *Theorical and Applied Genetics* **124**: 309-322. DOI: 10.1007/s00122-011-1706-5

Kim C, Guo H, Kong W, Chandnani R, Shuang LS, Paterson AH. 2015. Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science* **242**: 14-22. DOI: 10.1016/j.plantsci.2015.04.016

Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT, Jung K, Lee GW, Oh SK, Bae C, Kim SB, Lee HY, Kim SY, Kim MS, Kang BC, Jo YD, Yang HB, Jeong HJ, Kang WH, Kwon JK, Shin C, Lim JY, Park JH, Huh JH, Kim JS, Kim BD, Cohen O, Paran I, Suh MC, Lee SB, Kim YK, Shin Y, Noh SJ, Park J, Seo YS, Kwon SY, Kim HA, Park JM, Kim HJ, Choi SB, Bosland PW, Reeves G, Jo SH, Lee BW, Cho HT, Choi HS, Lee MS, Yu Y, Do Choi Y, Park BS, van Deynze A, Ashrafi H, Hill T, Kim WT, Pai HS, Ahn HK, Yeam I, Giovannoni JJ, Rose JK, Sørensen I, Lee SJ, Kim RW, Choi IY, Choi BS, Lim JS, Lee YH, Choi D. 2014. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genetics* **46**: 270–278. DOI: 10.1038/ng.2877

Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, Lundqvist U, Fujimura T, Matsuoka M, Matsumoto T, Yano M. 2007. Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proceedings of the National Academy of Sciences of the United States of America* **104**:1424-1429. DOI: 10.1073/pnas.0608580104

Kono TJK, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrel PL. 2016. The role of deleterious substitutions in crop genomes. *Molecular Biology and Evolution* **33**: 2307-2317. DOI: 10.1093/molbev/msw102

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arrollo-Kalin M, Barton L, Vigueira CC, Denham T, Dobney K, Doust AN, Gepts P, Gilbert MT, Gremillion KJ, Lucas L, Lukens L, Marshall FB, Olsen KM, Pires JC, Richerson PJ, Rubio de Casas R, Sanjur OI, Thomas MG, Fuller DQ. 2014 Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 6139-6146. DOI: 10.1073/pnas.1323964111

Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G,

Hovav R, Chaudhary B. Udall JA, Flagel L, Wendel JF. 2008. Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. *Genetics* **179**: 1725-1733. DOI: 10.1534/genetics.108.089656

Hu G, Koh J, Yoo M, Grupp K, Chen S, Wendel JF. 2013. Proteomic profiling of developing cotton fibers from wild and domesticated *Gossypium barbadense*. *New Phytologist* **200**: 570-582. DOI: 10.1111/nph.12381

Huang J, Gau Y, Jia H, Zhang Z. 2016. Characterization of the teosinte transcriptome reveals adaptive sequence divergence during maize domestication. *Molecular Ecology Resources*. DOI: 10.1111/1755-0998.12526

Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Lui K, Lu H, Li W, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497-501. DOI: 10.1038/nature11532

Hufford MB, Lubinsky P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. 2013. The genomic signatures of crop-wild introgression in maize. *PLOS Genetics* **9**: e1003477. DOI: 10.1371/journal.pgen.1003477

Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner RA, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J. 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**: 808-811. DOI: 10.1038/ng.2309

Iorizzo M, Senalik DA, Ellison SL, Grzebelus D, Cavagnaro PF, Allender C, Brunet J, Spooner DM, Van Deynze A, Simon PW. 2013. Genetic structure and domestication of carrot (*Daucus carota* subsp. *sativus*) (Apiaceae). *American Journal of Bonaty* **100**: 930-938. DOI: 10.3732/ajb.1300055

Iwata H, Gaston A, Remay A, Thouroude T, Jeauffre J, Kawamura K, Oyant LH, Araki T, Denoyes B, Foucher F. 2012. The *TFL1* homologue *KSN* is a regulator of continuous flowering in rose and strawberry. *The Plant Journal* **69**: 116-125. DOI: 10.1111/j.1365-313X.2011.04776.x

Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, Ohyanagi H, Mineta K, Michell CT, Saber N, Kharbatia NM, Rupper RR, Sharp AR, Dally N, Boughton BA, Woo YH, Gao G, Schijlen EGWM, Guo X, Momin AA, Negrão S, Al-Babili S, Gehring C, Roessner U, Jung C, Murphy K, Arold ST, Gojobori T, van der Linden CG, van Loo EN, Jellen EN, Maughan PJ, Tester M. 2017. *The genome of Chenopodium quinoa*. *Nature* **542**: 304-312. DOI :10.1038/nature21370

Jiang Z, Xia H, Basso B, Lu B-R. 2011. Introgression from cultivated rice influences genetic differentiation of weedy rice population at local spatial scale. *Theorical and Applied Genetics* **124**: 309-322. DOI: 10.1007/s00122-011-1706-5

Kim C, Guo H, Kong W, Chandnani R, Shuang LS, Paterson AH. 2015. Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science* **242**: 14-22. DOI: 10.1016/j.plantsci.2015.04.016

Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT, Jung K, Lee GW, Oh SK, Bae C, Kim SB, Lee HY, Kim SY, Kim MS, Kang BC, Jo YD, Yang HB, Jeong HJ, Kang WH, Kwon JK, Shin C, Lim JY, Park JH, Huh JH, Kim JS, Kim BD, Cohen O, Paran I, Suh MC, Lee SB, Kim YK, Shin Y, Noh SJ, Park J, Seo YS, Kwon SY, Kim HA, Park JM, Kim HJ, Choi SB, Bosland PW, Reeves G, Jo SH, Lee BW, Cho HT, Choi HS, Lee MS, Yu Y, Do Choi Y, Park BS, van Deynze A, Ashrafi H, Hill T, Kim WT, Pai HS, Ahn HK, Yeam I, Giovannoni JJ, Rose JK, Sørensen I, Lee SJ, Kim RW, Choi IY, Choi BS, Lim JS, Lee YH, Choi D. 2014. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genetics* **46**: 270–278. DOI: 10.1038/ng.2877

Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, Lundqvist U, Fujimura T, Matsuoka M, Matsumoto T, Yano M. 2007. Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proceedings of the National Academy of Sciences of the United States of America* **104**:1424-1429. DOI: 10.1073/pnas.0608580104

Kono TJK, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrel PL. 2016. The role of deleterious substitutions in crop genomes. *Molecular Biology and Evolution* **33**: 2307-2317. DOI: 10.1093/molbev/msw102

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arrollo-Kalin M, Barton L, Vigueira CC, Denham T, Dobney K, Doust AN, Gepts P, Gilbert MT, Gremillion KJ, Lucas L, Lukens L, Marshall FB, Olsen KM, Pires JC, Richerson PJ, Rubio de Casas R, Sanjur OI, Thomas MG, Fuller DQ. 2014 Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 6139-6146. DOI: 10.1073/pnas.1323964111

Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G,

Wang K, Xu X, Yu JZ, Zhu Y, Yu S. 2015. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnology* **33**: 524-530. DOI: 10.1038/nbt.3208

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W, Liang X, Shang H, Liu W, Shi C, Xiao G, Gou C, Ye W, Xu X, Zhang X, Wei H, Li Z, Zhang G, Wang J, Liu K, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nature Genomics* **46**: 467-472. DOI: 10.1038/ng.2987

Li YH, Zhao SC, Ma JX, Li D, Yan L, Li J, Qi XT, Guo XS, Zhang L, He WM, Chang RZ, Liang QS, Guo Y, Ye C, Wang XB, Tao Y, Guan RX, Wang JY, Liu YL, Jin LG, Zhang XQ, Liu ZX, Zhang LJ, Chen J, Wang KJ, Nielsen R, Li RQ, Chen PY, Li WB, Reif JC, Purugganan M, Wang J, Zhang MC, Wang J, Qiu LJ. 2013. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14**: 1-12. DOI: 10.1186/1471-2164-14-579

Lin Z, Li X, Shannon LM, Yeh C-T, Wang ML, Bai G, Peng Z, Li J, Trick HN, Clemente TE, Doebley J, Schnable PS, Tuinstra MR, Tesso TT, White F, Yu J. 2012. Parallel domestication of the *Shattering1* genes in cereals. *Nature Genetics* **44**: 720-724. DOI: doi:10.1038/ng.2281

Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, Yang CQ, Chen JD, Chen JJ, Chen DY, Zhang L, Zhou Y, Wang LJ, Guo WZ, Bai YL, Ruan JX, Shangguan XX, Mao YB, Shan CM, Jiang JP, Zhu YQ, Jin L, Kang H, Chen ST, He XL, Wang R, Wang YZ, Chen J, Wang LJ, Yu ST, Wang BY, Wei J, Song SC, Lu XY, Gao ZC, Gu WY, Deng X, Ma D, Wang S, Liang WH, Fang L, Cai CP, Zhu XF, Zhou BL, Jeffrey Chen Z, Xu SH, Zhang YG, Wang SY, Zhang TZ, Zhao GP, Chen XY. 2015. *Gossypium barbadense* genome sequence provides insight into evolution of extra-long staple fiber and specialized metabolites. *Scientific Reports* **5**: 1-14. DOI: 10.1038/srep14139

Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics* **22**: 126-131. DOI: 10.1016/j.tig.2006.01.004

Mackay TF, Stone E, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**: 565-577. DOI: 10.1038/nrg2612

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* **14**: 840-852. DOI: 10.1038/nrg3605

Mezmouk S, Roos-Ibarra J. 2014. The pattern and distribution of deleterious mutations in maize. *G3: Genes, Genomes, Genetics* **4**: 163-171. DOI: 10.1534/g3.113.008870

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Pérez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**: 991–996. DOI: 10.1038/nature06856

Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E, Zhang J, Huang L, Zhang L, Miao W, Zhang J, Ye Z, Miao C, Lin Z, Wang H, Zhou H, Yim WC, Priest HD, Zheng C, Woodhouse M, Edger PP, Guyot R, Guo HB, Guo H, Zheng G, Singh R, Sharma A, Min X, Zheng Y, Lee H, Gurtowski J, Sedlazeck FJ, Harkess A, McKain MR, Liao Z, Fang J, Liu J, Zhang X, Zhang Q, Hu W, Qin Y, Wang K, Chen LY, Shirley N, Lin YR, Liu LY, Hernandez AG, Wright CL, Bulone V, Tuskan GA, Heath K, Zee F, Moore PH, Sunkar R, Leebens-Mack JH, Mockler T, Bennetzen JL, Freeling M, Sankoff D, Paterson AH, Zhu X, Yang X, Smith JA, Cushman JC, Paull RE, Yu Q. 2015. The pineapple genome and the evolution of CAM photosynthesis. *Nature genetics* **47**: 1435-1442. DOI: 10.1038/ng.3435

Molina J, Sikora M, Garud N, Flowers JM, Rubinsteina S, Reynolds A, Huang P, Jackson S, Schaal BA, Bustamante CD, Boyko AR, Purugganan MD. 2011. Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences of the United States of America* **7**: 8351–8356. DOI: 10.1073/pnas.1104686108

Montes-Hernández S, Eguiarte LE. 2002. Genetic structure indirect estimates of gene flow in three taxa of *Cucurbita* (Cucurbitaceae) in western Mexico. *American Journal of Botany* **89**: 1156-1163. DOI: 10.3732/ajb.89.7.1156

Montes-Hernández S, Merrick LC, Eguiarte LE. 2005. Maintenance of squash (*Cucurbita* spp.) landrace diversity by farmers' activities in Mexico. *Genetic Resources and Crop Evolution* **52**: 697-707. DOI:10.1007/s10722-003-6018-4

Morrell PL, Bucker ES, Ross-Ibarra J. 2011. Crop genomics: advances and applications. *Nature Reviews Genetics* **13**: 85-96. DOI: 10.1038/nrg3097

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE *et al*. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 453-458. DOI: 10.1073/pnas.1215985110

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, Bustamente CD, Buckler ES. 2011. Genetic domestication history of the grape. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 3530-3535. DOI: 10.1073/pnas.1009363108

Olsen KM, Caicedo AL, Jia Y. 2007. Evolutionary of genomics of weedy rice in the USA. *Journal of Integrative Plant Biology* **49**: 811-816. DOI: 10.1111/j.1744-7909.2007.00506.x

Papa R, Gepts P. 2003. Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theoretical and Applied Genetics* **106**: 239-250. DOI: 10.1007/s00122-002-1085-z

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**: 551-556. DOI: 10.1038/nature07723

Pickersgill B. 2009. Domestication of plants revisited – Darwin to the present day. *Botanical Journal of the Linnean Society* **161**: 203-212. DOI: 10.1111/j.1095-8339.2009.01007.x

Polashock J, Zelzion E, Fajardo D, Zalapa J, Georgi L, Bhattacharya D, Vorsa N. 2014. The American cranberry: first insights into the whole genome of a species adapted to bog habitat. *BMCPlant Biology* **14**: 165–181. DOI: 10.1186/1471-2229-14-165

Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S. 2012. The cassava genome: current progress, future directions. *Tropical Plant Biology* **5**: 88-94. DOI:10.1007/s12042-011-9088-z

Purugganan MD, Boyles AL, Suddith J. 2000. Variation and señection at the *CAULIFLOWER* floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. *Genetics* **155**: 855-862.

Renaut S, Rieseberg LH. 2015. The accumulation of deleterious mutation as a consequence of domestication and improvement in sunflowers and other Compositae crops. *Molecular Biology and Evolution* **32**: 2273-2283. DOI: 10.1093/molbev/msv106

Rieseberg LH, Carney SE. 1998. Plant hybridization. *New Phytologist* **140**: 599–624. DOI:10.1046/j.1469-8137.1998.00315.x

Rodriguez GR, Muños S, Anderson C, Sim SC, Michel A, Causse M, Gardener BB, Francis D, van der Knaap E. 2011. Distribution of *SUN*, *OVATE*, *LC* and *FAS* in the tomato germplasm and the relationships to fruit shape diversity. *Plant physiology* **156**: 275-285. DOI: 10.1104/pp.110.167577

Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the gentic basis of adaptation. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 8641-8648. DOI: 10.1073/pnas.0700643104

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**:178–183. DOI: 10.1038/nature08670

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MM, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS, Jackson SA. 2014. A reference genome for common bean and genome-wide analysis of dual domestication. *Nature genetics* **46**: 707-713. DOI: 10.1038/ng.3008

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A,

Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. 2009. The B73 maize genome: Complexity, diversity and dynamics. *Science* **326**: 1112–1115. DOI: 10.1126/science.1178534

Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, Geopfert S, Peitsch MC, Ivanov NV. 2014. The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications* **5**: 1-9. DOI: 10.1038/ncomms4833.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics* **43**: 1160-1163. DOI: 10.1038/ng.942

Swanson-Wagner R, Briskine R, Schaefer R, Schaefer R, Hufford MB, Ross-Ibarra J, Myers CL, Tiffin P, Springer NM. 2012. Reshaping of the maize transcriptome by domestication. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 11878–11883. DOI: 10.1073/pnas.1201961109

Tang H, Sezen U, Paterson AH. 2010. Domestication and plant genomes. *Current Opinion in Plant Biology* **13**: 160-166. DOI: 10.1016/j.pbi.2009.10.008

The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189–195. DOI: 10.1038/nature10158

The Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**: 635–641. DOI: 10.1038/nature11119

Varshney RK, Terauchi R, McCouch SR. 2014. Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLOS Biology* **12**: 1-8. DOI: 10.1371/journal.pbio.1001883

Vavilov NI. 1922. The law of homologous series in variation. *Journal of Genetics* **12**: 47–89

Wang H, Xu X, Vieira FG, Xiao Y, Li Z, Wang J, Neilsen R, Chu C. 2016. The power of inbreeding: NGS based GWAS of rice reveals convergent evolution during rice domestication. *Molecular Plant* **9**: 975-85. DOI: 10.1016/j.molp.2016.04.018

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C, Li Q, Yuan Y, Lu C, Wei H, Gou C, Zheng Z, Yin Y, Zhang X, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genetics* **44**: 1098-1103. DOI: 10.1038/ng.2371

Weeden NF. 2007. Genetic changes accompanying the domestication of *Pisum sativum*: is there a common basis to the domestication syndrome for legumes?. *Annals of Bonaty* **100**: 1017-1025. DOI: 10.1093/aob/mcm122

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J, Takita MA, Labadie K, Poulain J, Couloux A, Jabbari K, Cattonaro F, Del Fabbro C, Pinosio S, Zuccolo A, Chapman J, Grimwood J, Tadeo FR, Estornell LH, Muñoz-Sanz JV, Ibanez V, Herrero-Ortega A, Aleza P, Pérez-Pérez J, Ramón D, Brunel D, Luro F, Chen C, Farmerie WG, Desany B, Kodira C, Mohiuddin M, Harkins T, Fredrikson K, Burns P, Lomsadze A, Borodovsky M, Reforgiato G, Freitas-Astúa J, Quetier F, Navarro L, Roose M, Wincker P, Schmutz J, Morgante M, Machado MA, Talon M, Jaillon O, Ollitrault P, Gmitter F, Rokhsar D. 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology* **32**: 656–662. DOI: 10.1038/nbt.2906

Wu Y, San Vicente F, Huang K, Dhliwayo T, Costich DE, Semagn K, Sudha N, Olsen M, Prasanna BM, Zhang X, Babu R. 2016. Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theoretical and Applied Genetics* **129**: 753–765. DOI: 10.1007/s00122-016-2664-8

Xia HB, Wang W, Xia H, Zhao W, Lu BR. 2011. Conspecific crop-weed introgression influences evolution of weedy rice (*Oryza sativa* f. *spontanea*) across a geographical range. *PLoS ONE* **6**: 1-11. DOI: 10.1371/journal.pone.0016189

Xu J, Yuan Y, Xu Y, Zhang G, Guo X, Wu F, Wang Q, Rong T, Pan G, Cao M, Tang Q, Gao S, Liu Y, Wang J, Lan H, Lu Y. 2014. Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC Plant Biololy* **14**: 1-15. DOI:10.1186/1471-2229-14-83.

Received:
January 17th, 2017

Accepted:
May 12th, 2017

Yu J, Hu S, Wang J, Wong KS, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79–92. DOI: 10.1126/science.1068037

Zizumbo-Villarreal D, Flores-Silva A, Colunga-García Marín P. 2012. The archaic diet in Mesoamerica: Incentive for milpa development and species domestication. *Economic Botany* **66**: 328-343. DOI: 10.1007/s12231-012-9212-5

Capítulo 3:

Domestication genomics of the open-pollinated Scarlet Runner Bean (*Phaseolus coccineus* L.)

Azalea Guerra-García, Marcio Suárez-Atilano, Alicia Mastretta-Yanes, Alfonso Delgado-Salinas & Daniel Piñero

Frontiers in Plant Science

DOI: 10.3389/fpls.2017.01891

frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 November 2017
doi: 10.3389/fpls.2017.01891

Domestication Genomics of the Open-Pollinated Scarlet Runner Bean (*Phaseolus coccineus* L.)

Azalea Guerra-García[1,2], Marco Suárez-Atilano[3], Alicia Mastretta-Yanes[4], Alfonso Delgado-Salinas[5] and Daniel Piñero[2]*

[1] Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México, Ciudad de México, Mexico, [2] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico, [3] Departamento de Ecología de la Biodiversidad, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico, [4] CONACYT-CONABIO, Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Ciudad de México, Mexico, [5] Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

The runner bean is a legume species from Mesoamerica closely related to common bean (*Phaseolus vulgaris*). It is a perennial species, but it is usually cultivated in small-scale agriculture as an annual crop for its dry seeds and edible immature pods. Unlike the common bean, *P. coccineus* has received little attention from a genetic standpoint. In this work we aim to (1) provide information about the domestication history and domestication events of *P. coccineus*; (2) examine the distribution and level of genetic diversity in wild and cultivated Mexican populations of this species; and, (3) identify candidate loci to natural and artificial selection. For this, we generated genotyping by sequencing data (42,548 SNPs) from 242 individuals of *P. coccineus* and the domesticated forms of the closely related species *P. vulgaris* (20) and *P. dumosus* (35). Eight genetic clusters were detected, of which half corresponds to wild populations and the rest to domesticated plants. The cultivated populations conform a monophyletic clade, suggesting that only one domestication event occurred in Mexico, and that it took place around populations of the Trans-Mexican Volcanic Belt. No difference between wild and domesticated levels of genetic diversity was detected and effective population sizes are relatively high, supporting a weak genetic bottleneck during domestication. Most populations presented an excess of heterozygotes, probably due to inbreeding depression. One population of *P. coccineus* subsp. *striatus* had the greatest excess and seems to be genetically isolated despite being geographically close to other wild populations. Contrasting with previous studies, we did not find evidence of recent gene flow between wild and cultivated populations. Based on outlier detection methods, we identified 24 domestication-related SNPs, 13 related to cultivar diversification and eight under natural selection. Few of these SNPs fell within annotated loci, but the annotated domestication-related SNPs are highly expressed in flowers and pods. Our results contribute to the understanding of the domestication history of *P. coccineus*, and highlight how the genetic signatures of domestication can be substantially different between closely related species.

Keywords: domestication, genotyping by sequencing, *Phaseolus coccineus*, adaptative variation, population genomics

OPEN ACCESS

Edited by:
Alejandro Casas, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Peter J. Prentis, Queensland University of Technology, Australia
Gonzalo Gajardo, University of Los Lagos, Chile

****Correspondence:***
Azalea Guerra-García
azalea.guerra@iecologia.unam.mx

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

Received: *31 July 2017*
Accepted: *18 October 2017*
Published: *15 November 2017*

Citation:
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

INTRODUCTION

The scarlet runner bean (*Phaseolus coccineus* L.) is one of the five *Phaseolus* species that were domesticated in Mesoamerica, and it is the third-most economically important, after *P. vulgaris* L. and *P. lunatus* L. The domestication process of this species continues today both in the Americas and Europe, where it was introduced by the Spaniards. One of its main characteristics is its ability to tolerate cooler climates than other *Phaseolus* and up to date it is an important food source for smallholders and indigenous groups in Mexico (Salinas, 1988). Despite the cultural value, economic importance, and agronomic potential of *P. coccineus*, little is known about its domestication history and the genetic variability of its wild and cultivated forms.

Wild *P. coccineus* are perennial climbing plants, occurring mostly at mid-high elevations (1,000–3,000 m.a.s.l.), from northern Mexico (Chihuahua) to Panama (Salinas, 1988). It has 11 pairs of chromosomes and an estimated genome size of 660 Mb (Plant DNA C-values database). Contrasting with the autogamous common bean, the scarlet runner bean is an open-pollinated species. The high morphological diversity of this species has been classified under two subspecies (Freytag and Debouck, 2002): *P. coccineus* subsp. *coccineus* (mostly with red flowers), including 11 wild varieties and the domesticated form, and *P. coccineus* subsp. *striatus* (purple or mauve flowers), conformed by eight wild varieties. No genetic evidence supports these subspecies and varieties, but given the environmental and cultural heterogeneous landscape where *P. coccineus* occurs, it is expected that the species should be genetically structured.

As a cultivated species, *P. coccineus* is currently grown in Mexico, Guatemala, Honduras and Costa Rica, and in lesser degree in South America. In Europe, it is mostly cultivated in the United Kingdom, Netherlands, Italy, and Spain (Rodiño et al., 2006). In Mexico, the scarlet runner bean is cultivated both as a self-sufficiency crop by smallholder farmers (<5 ha) and also commercially for urban areas. Besides its native cultivars, in Mexico there is one breeding line (Blanco Tlaxcala) developed using a multi linear method (Vargas-Vázquez et al., 2012). Feral populations are common, but it is unknown if they originated from hybridization between wild and domesticated populations, or if they escaped from cultivation. Wild, feral and domesticated distributions overlap in Mesoamerica, suggesting that there are plenty of opportunities for gene flow to occur, making the domestication history of *P. coccineus* difficult to disentangle without high resolution genetic markers.

The domestication history of the scarlet runner bean has been explored previously with low resolution molecular markers, and multiple domestication events were suggested. Specifically, chloroplast and nuclear SSRs of *P. coccineus* accessions including European domesticated populations, Mesoamerican landraces and wild samples from Mexico, Guatemala, and Honduras (Angioi et al., 2009; Spataro et al., 2011; Rodriguez et al., 2013) suggest that *P. coccineus* domestication took place in the Guatemala-Honduras area, or that alternatively another domestication event occurred in Mexico followed by extensive hybridization with the cultivated populations from Guatemala and Honduras. However, few Mesoamerican samples were included in these studies, and they focused on European domesticated populations. Phylogenetic analyses including more samples from the wide distribution of *P. coccineus* could bring clues about the number of domestication events that took place in this species. For example, if cultivars are grouped in one monophyletic clade, it would suggest one domestication event.

Another interesting feature of *P. coccineus* domestication history is that similar levels of genetic variation have been reported in wild and cultivated populations (Escalante et al., 1994; Spataro et al., 2011; Rodriguez et al., 2013). This contradicts the population genetics models that predicts a genetic diversity reduction and increased divergence between wild and domesticated forms due to demographic factors and selection at target loci (Meyer and Purugganan, 2013). This pattern has been described in crops like sunflower (∼30%; Renaut and Rieseberg, 2015); soybean (∼30%; Li et al., 2013); maize (∼17%; Hufford et al., 2012); and in cultivated *Agave* species (from 21 to 66%; Eguiarte et al., 2013). Also, in the Mesoamerican common bean a ∼20% reduction in genetic variation has been reported (Schmutz et al., 2014). However, the amount of genetic diversity that is lost along domestication depends on several factors, including the severity and the number of bottlenecks, the strength of selection and human management (Gepts, 2014). To properly assess impact of domestication on the genetic diversity of *P. coccineus,* genomic data comparing wild and cultivated populations is necessary.

The use of genomic tools also allows to characterize diversity and differentiation patterns across genomes. Regions or variants that departure from neutral predictions are probably influenced by selective pressures and are tagged as candidates. Applying this approach to crop species and their wild relatives allows to distinguish loci affected during domestication, whereas comparisons between landraces and/or improved cultivars measure the effect of subsequent selection (Tang et al., 2010; Gepts, 2014). Furthermore, hypotheses about phenotypic convergence in crops can be tested. In other words, if the same genes or genomic regions were affected during the domestication process of different species.

Here, we aim to deal with the previous knowledge gaps by using genomic data to (1) provide information about the domestication history of *P. coccineus* and its current evolutionary dynamic in Mexico, in particular to analyze the occurrence of a single or multiple domestication events in Mexico; (2) examine the extent of the domestication bottleneck in this species by comparing the levels of genetic diversity and geographic patterns of the wild, feral and domesticated Mexican populations; and (3) identify candidate loci under natural and artificial selection in *P. coccineus* genome.

MATERIALS AND METHODS

Plant Material and SNP Genotyping

Phaseolus coccineus individuals from 10 wild, three feral and 11 cultivated Mexican populations and one cultivar from Spain were analyzed, as well as plants from the breeding line Blanco Tlaxcala. Taxonomic and wild/feral/domesticated categories were

assigned based on morphology and habitat observations. Only one of the wild populations that were sampled corresponds to subsp. *striatus*, the rest belong to subsp. *coccineus*. A population was classified as feral if it was growing out of cultivation and presented intermediate traits between wild and domesticated forms. The Mexican samples cover the species distribution and main cultivation areas at the national level. As outgroups, samples from the closely related species *P. vulgaris* (three wild and one cultivated) and *P. dumosus* (seven cultivated) were included (Supplementary Table S1). For the three species, the samples size of each population varied between three to 16 individuals.

Sampling was performed during September–December of 2014 and 2015. In the case of the wild populations, tissue from young leaves was collected and stored in silica until processed. Seeds from cultivars were collected and germinated at the Instituto de Ecología, UNAM. DNA was extracted using DNeasy Plant Mini Kit (Qiagen). DNA samples were genotyped at the Institute for Genomic Diversity at Cornell University (Services | Institute of Biotechnology, 2017). Sequencing libraries were constructed using enzymes PstI and BfaI following the Genotype by Sequencing (GBS) protocol of Elshire et al. (2011). A total of 326 samples were processed in four plates of ninety six samples each, multiplexed and sequenced on four lanes of Illumina HiSeq 2500 (100 bp, single-end reads).

Reads were aligned to *P. vulgaris* reference genome v1.0 (Phytozome) DOE-JGI and USDA-NIFA, http://phytozome.jgi.doe.gov/ (Phytozome, 2017) using bwa v 0.7.8-r455; (Li and Durbin, 2009). Demultiplexing, initial quality control, assembly and SNP discovery were made with TASSEL pipeline v3.0.174 (Glaubitz et al., 2014). Assembly and SNP discovery were performed independently for two sets of data, one containing samples from *P. vulgaris*, *P. dumosus,* and *P. coccineus* (VDC group), which are the domesticated species of the Vulgaris clade (Delgado-Salinas et al., 2006); and the other data set only including *P. coccineus* samples. SNPs were filtered in VCFtools 0.1.15 (Danecek et al., 2011) using the following parameters for the two data sets: (1) VDC group: maximum missingness threshold 20% per individual; minimum mean depth 10X; minimum allele frequency (MAF) 0.01; minimum allele count 90%; and only SNPs mapped in chromosomes. (2) *Phaseolus coccineus*: maximum missingness threshold 30% per individual; minimum mean depth 5X; MAF 0.02; minimum allele count 80%; and only SNPs mapped in chromosomes.

Filtered SNP data, species occurrence data and scripts used for the analyses are available at Dryad Repository under the identifier doi: 10.5061/dryad.q343c.

Inferring Population Structure and Phylogenetic Relationships

We inferred the population structure of *P. coccineus* because different genetic clusters are expected to occur due to the isolation and environmental and cultural heterogeneity in which this species occurs. For this, the software Admixture v1.3 (Alexander et al., 2009) was used to infer population structure of *P. coccineus*. Values of K ranging from one to twenty were tested, and the value that exhibited the lowest cross-validation error was chosen. Then, we examined the phylogenetic relationships between the genetic groups, both cultivated and wild, and if each cluster forms a monophyletic clade. This phylogenetic analysis was also used as a preliminary approach to identify the plausible number of domestication events for the Mexican cultivated *P. coccineus* (see below for other analyses). Specifically we examined if the cultivated samples was recovered as a monophylogenetic group. For the phylogenetic analysis, wild and cultivated samples of *P. coccineus*, *P. vulgaris,* and *P. dumosus* were analyzed under three schemes:

First, a Maximum-Likelihood based approach was carried out with the FastTree software (Price et al., 2009). For this, a mix of Nearest-Neighbor Interchange and Subtree-Prune and Regraft moves (NNI+SPR) was considered for topology and branch-length optimization and the General-Time Reversible with a single rate per site model (GTR+CAT) was included as nucleotide substitution model. Because FastTree only considers those SNPs identified as fixed within individuals (i.e., homozygous), but polymorphic among individuals, only the 82% of the total VDC subset (41,223 SNPs) were considered in this analysis. Second, a phylogenetic network based on the Neighbor-net algorithm and Patristic Distances with GTR+I+G correction was estimated with SplitsTree (Huson and Bryant, 2006) software. Lastly, we employed a Bayesian multispecies coalescent model (Rannala and Yang, 2003) to estimate the phylogenetic relationships among well-supported clades within *P. coccineus* solely. We used the program SNAPP 1.3.0 (Bryant et al., 2012), included in the package BEAST 2.4.5 (Bouckaert et al., 2014) to infer species trees directly from biallelic genetic data. We used the eight main genetic clusters (see section Results) inferred by Admixture as *a priori* designated species and the Wild-TMVB cluster was partitioned in two, taking into account the ML topology of that cluster. Because SNAPP does not incorporate missing data, we selected a subset of our taxonomic sampling that maximized the number of SNPs available. The final analysis retained a total of 600 SNPs under linkage equilibrium; without any missing data and considering a minimum of five individuals from each cluster of the designated species. We used SNAPP's default settings and ran the analysis for 1,000,000 generations sampling every 1,000 generations. We evaluated the convergence (i.e., short variation in -lnL scores, ESS $>$ 100) from our runs by examining log files with the program Tracer 1.5 (Drummond and Rambaut, 2007). We analyzed the tree files with SNAPP-TreeSetAnalyser 2.4.5, to identify species trees that were contained in the 95% highest posterior density (HPD) set and using 10% of topologies as burn-in. Resulted tree files (cloudgrams) were visualized using DensiTree (Bouckaert, 2010).

Population Genetics Statistics

To evaluate the existence and degree of the domestication bottleneck on *P. coccineus* we estimated genetic diversity and differentiation indices of the genetic groups inferred by the Admixture analysis (see section Results). Specifically, we used the Hierfstat package (Goudet, 2005) in R (R Core Team, 2017) to estimate per site heterozygosity and F_{IS}, as well as pairwise F_{ST} among groups, performing a bootstrap (1,000) to obtain confidence intervals. To test the hypothesis that $n_i = n_j$ (where n_i is the number of loci of the cluster i where $H_{Ei} > H_{Ej}$, and

n_2 is the number of loci of the cluster j where $H_{Ej} > H_{Ei}$) we used a pairwise χ^2 tests with Bonferroni correction to avoid false positive results (Sokal and Rohlf, 1995). Also, we estimated the heterozygosity and F_{IS} at the sampling location (*P. coccineus* dataset) and at the species level (VDC dataset) applying the same test.

Multiple vs. Single Domestication Events Test

In order to confirm the hypothesis of a single domestication event in Mexico suggested by our phylogenetic analyses (see section Results) we applied the Approximate Bayesian computation (ABC; Beaumont et al., 2002) method implemented in DIYABC 2.04 (Cornuet et al., 2014). Preliminary tests included comparisons among three scenarios with 3×10^6 simulated datasets (1×10^6 each scenario) in which the position of the Wild-Sierra Madre Occidental (Wild-SMOCC) clade was evaluated (see section Results, Supplementary Figure S1). Our final estimation included 4×10^6 simulated datasets (2×10^6 each scenario) considering the Wild-SMOCC population fixed as sister clade of the Wild-Trans-Mexican Volcanic Belt (Wild-TMVB) populations (see section Results). The number of domestication events was tested as follows: multiple events (Scenario 1, Supplementary Figure S2) vs. a single one (Scenario 2, Supplementary Figure S2). The DIYABC approach was also applied to estimate the time at which domestication occurred, as well as other demographic parameters such as effective population size (*Ne*). A subsample from the SNAPP dataset (279 SNPs) and the scheme of eight clusters were used to set populations in DIYABC (**Figure 2B**). Priors were set as follow: log-uniform distributions across all parameters, *Ne* ranging from 100 to 100,000 individuals, mutation rate set to 10^{-8}–10^{-6} across SNPs, and divergence times among populations set to 10–100,000 generations ago (**Table 1**).

We compared the fit of the single vs. multiple domestication events scenarios by estimating their posterior probabilities: with the obtained reference tables from each scenario, we ranked the simulated datasets in order of increasing distance to the observed data considering direct and logistic approaches (Beaumont et al., 2002; Cornuet et al., 2014). Distance between datasets was based on summary statistics, estimated from the empirical and simulated sets. We performed a pre-evaluation step using a principal components analysis (PCA), to ensure that at least one (or more) scenarios would produce simulated datasets close enough to the empirical data. The PCA was based on a set of 5,000 simulated datasets, generated from the parameters' prior distributions (Supplementary Figure S3).

Identifying Candidate Loci

We used the wild and cultivated samples of *P. coccineus* to identify candidate loci related to domestication, to cultivar diversification, and to natural selection. Before the candidate SNPs analysis, an additional filter based on linkage disequilibrium (LD) was applied. To determine the threshold distance at which there is no LD, we estimate the inter-variant allele correlations (r^2) using PLINK 1.9 (Chang et al., 2015). To distinguish LD due to physical distance (bp), the r^2 was estimated for SNPs located in the same and in different chromosomes. The distance threshold was established in 3,000 bp, so that SNPs closer than this distance were removed.

This LD-filtered dataset was analyzed with two different approaches for outlier detection: the R package pcadapt (Luu et al., 2017) and BayeScan 2.1 (Foll and Gaggiotti, 2008). Only loci identified by the pcadapt and BayeScan methods were considered as candidate loci. Pcadapt detects candidate SNPs assuming that these are outliers with respect to how they are related to population structure. By contrast to population-based approaches, pcadapt does not require grouping individuals into populations and handles admixed individuals (Luu et al., 2017). BayeScan instead uses differences in allele frequencies of pre-defined populations, in this case the genetic clusters previously established by Admixture.

In both approaches, three separate analyses were performed with each method to detect signatures of different types of selective pressure. First, to detect candidate domestication loci, wild and cultivated samples of *P. coccineus* were included, and feral individuals were removed. In this case, for the pcadapt analysis, only the first principal component was assessed because it explains the difference between wild and cultivated populations (see section Results). Also, an additional SNPs filter was made and MAF were adjusted to consider SNPs present in at least five individuals. For this dataset, that is MAF = 0.023. For Bayescan no additional filter was made. Second, to identify loci related to diversification in the context of domestication, only cultivated

TABLE 1 | Estimations of effective population sizes of the best-fit DIYABC model (single domestication) for *Phaseolus coccineus* in Mexico.

Genetic group	Minimum prior value	Maximum prior value	Average posterior value	95%CI
Wild-SUR-CH	100	1×10^5	1.0×10^5	9.98×10^4–1×10^5
Wild-SMOCC	100	1×10^5	8.94×10^4	8.01×10^4–9.5×10^4
Wild-TMVB	100	1×10^5	8.94×10^4	8.01×10^4–9.5×10^4
Wild-striatus	100	1×10^5	9.68×10^4	8.9×10^4–1×10^5
Cult-OV	100	1×10^5	8.83×10^4	7.5×10^4–1×10^4
Cult-SUR-CH	100	1×10^5	6.39×10^4	5.7×10^4–9.38×10^4
Cult-TMVB	100	1×10^5	9.36×10^3	8.39×10^3–1.54×10^4
Cult-SMOCC	100	1×10^5	8.57×10^3	6.01×10^3–1.2×10^4

Please refer to the text to understand what the acronyms stand for.

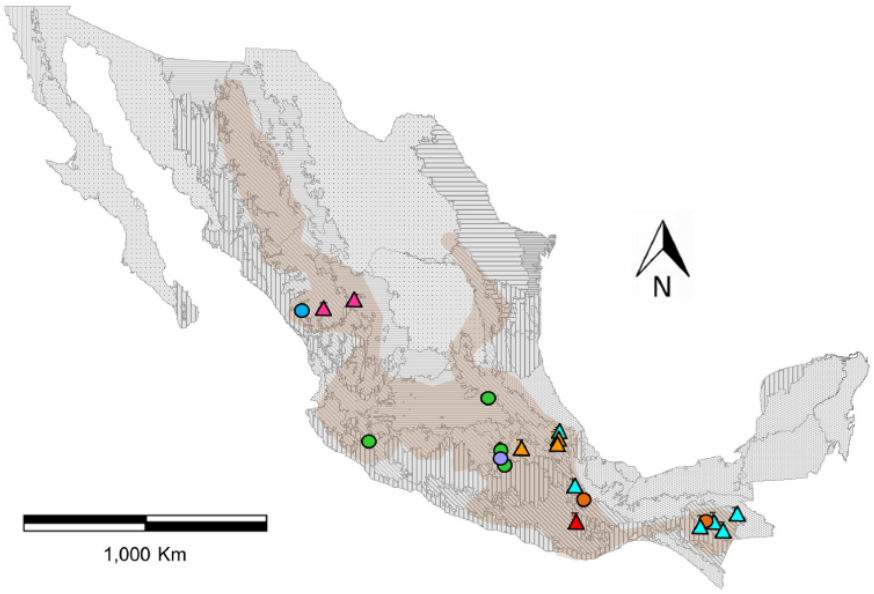


FIGURE 1 | Distribution map of genotyped populations of *Phaseolus coccineus* in Mexico, assigned to eight genetic clusters with the program Admixture (colors). Circles indicate wild populations and triangles show cultivated populations (**Figure 2**). Frames indicate the seven first-level ecoregions (9–15) present in Mexico (level 1) and boundaries represent 21 second-level nested ecoregions. Shaded area shown the potential geographical distribution of *P. coccineus* in Mexico (López-Soto et al., 2005).

samples were analyzed. In the pcadapt analyses, the first six components were assessed because they explain the genetic structure of populations, and MAF threshold was set to 0.038 to excluded alleles present in less than five individuals. Notice that in this case, diversification refers to the phase that follows initial domestication and involves the spread and adaptation to different agro-ecological and socio-cultural environments (Meyer and Purugganan, 2013). Lastly, to detect natural selection signatures, we focused both methods on wild samples. Again, for the pcadapt analyses the first six components were assessed and the MAF threshold was set 0.055 to exclude SNPs present in less than five individuals. In all cases, no additional filter was made for BayeScan.

The false discovery rate threshold applied in pcadapt and BayeScan were 0.005 and 0.05, respectively. To compare how genetic variance is explained by candidate SNPs and by data set LD filtered, PCAs were made using the SNPrelate package (Zheng et al., 2012).

Using Phytozome's JBrowser, the putative function and tissue of expression of these loci was examined by looking for the annotation of the selected SNPs in *P. vulgaris* genome v 2.1 (DOE-JGI and USDA-NIFA[1]). For each annotated loci we looked for homologous proteins with the highest similarity in other plants, and examined if the homolog genes in *Glycine max* (soybean) were among the domestication-related loci associated with flowering time and seed size in this species (Zhou et al., 2015).

[1]http://phytozome.jgi.doe.gov/

RESULTS

Sampling and SNP Genotyping

A total of 296 individuals representing four ecoregions of Mexico (as defined in Instituto Nacional de Estadística, Geografía e Informática (INEGI), Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), and Instituto Nacional de Ecología (INE), 2008) were sampled and successfully genotyped (**Figure 1**). After assembly and SNP discovery, the VDC group dataset contains 241 individuals of *P. coccineus*, 20 of *P. vulgaris* and 35 of *P. dumosus*, 50 273 SNPs, 2.24% mean missing data per individual, and a mean depth per site of 58.63. The *P. coccineus* dataset includes 242 individuals (91 wild; 20 feral; 131 cultivated), 42,548 SNPs, 3.97% mean missing data per individual, and a mean depth per site of 50.41.

Inferring Population Structure and Phylogenetic Relationships

The *K*-value that presents the lower error rate in Admixture analysis was eight (Supplementary Figure S4). Half of the genetic groups correspond to the cultivars from the Trans-Mexican Volcanic Belt (Cult-TMVB), Sierra Madre del Sur and Chiapas Highlands (Cult-SUR-CH), Sierra Madre Occidental (Cult-SMOCC) and Oaxaca Valley (Cult-OV). The other half of the genetic clusters belong to wild populations from the Trans-Mexican Volcanic Belt (Wild-TMVB), Sierra Madre del Sur and Chiapas Highlands (Wild-SUR-CH), Sierra Madre

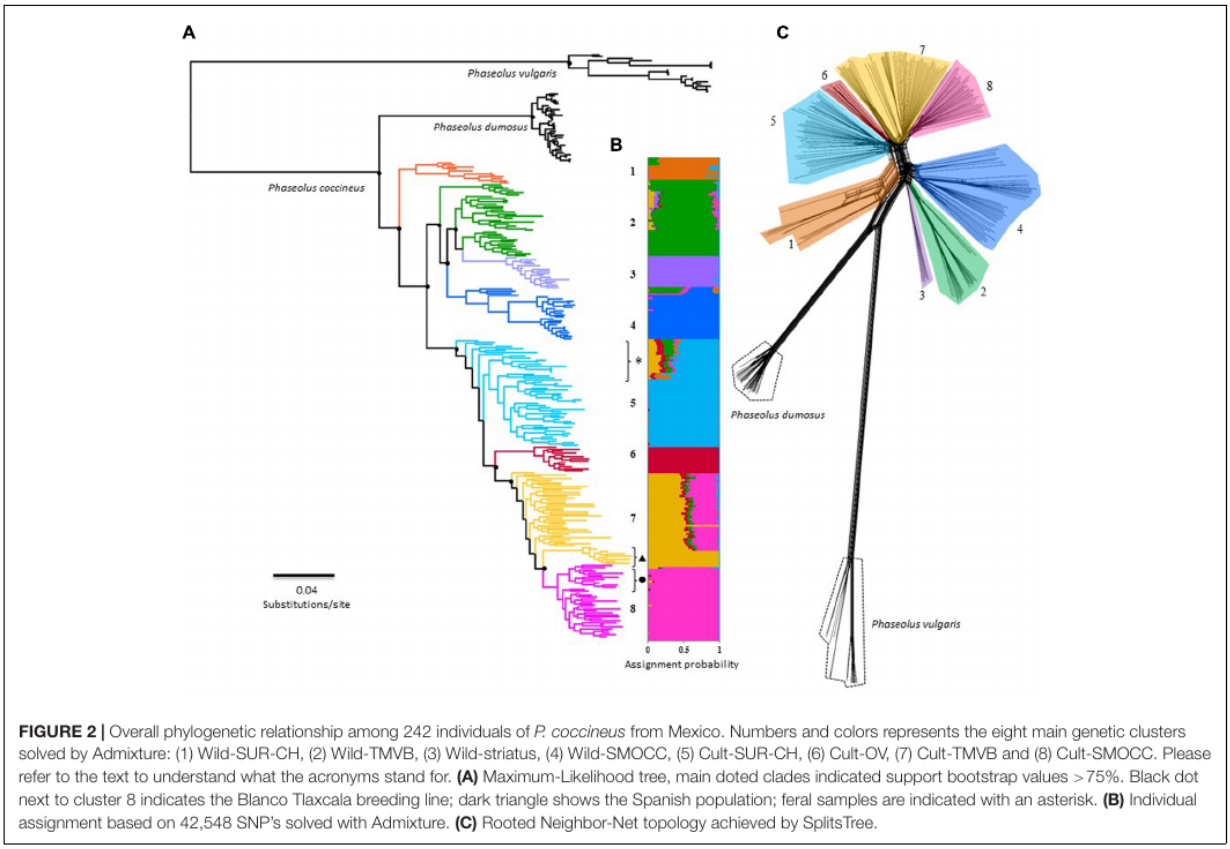


FIGURE 2 | Overall phylogenetic relationship among 242 individuals of *P. coccineus* from Mexico. Numbers and colors represents the eight main genetic clusters solved by Admixture: (1) Wild-SUR-CH, (2) Wild-TMVB, (3) Wild-striatus, (4) Wild-SMOCC, (5) Cult-SUR-CH, (6) Cult-OV, (7) Cult-TMVB and (8) Cult-SMOCC. Please refer to the text to understand what the acronyms stand for. **(A)** Maximum-Likelihood tree, main doted clades indicated support bootstrap values >75%. Black dot next to cluster 8 indicates the Blanco Tlaxcala breeding line; dark triangle shows the Spanish population; feral samples are indicated with an asterisk. **(B)** Individual assignment based on 42,548 SNP's solved with Admixture. **(C)** Rooted Neighbor-Net topology achieved by SplitsTree.

Occidental (Wild-SMOCC) and subsp. *striatus* population, located in the TMVB (Wild-*striatus*; **Figure 2**). The genetic clusters seem to be related to geographic distances (**Figure 1**), except the population Wild-*striatus*, which is geographically close to populations of *P. coccineus* subsp. *coccineus* but seems genetically isolated. Samples from the Spanish population (**Figure 2B**, triangle) were assigned to the Cult-TMVB genetic group, but unlike the individuals of this cluster, samples from Spain do not present a mixed ancestry. Regarding samples of the breeding line Blanco Tlaxcala (**Figure 2B**, circle), they are grouped with landraces from Cult-SMOCC cluster.

The phylogenetic hypotheses constructed with FastTree and SplitsTree (**Figures 2A,C**) are consistent with the Admixture genetic groups (**Figure 2B**). Nevertheless, both analysis suggested the Wild-TMVB group as a paraphyletic clade. ML topology revealed a finer-scale structure, identifying three paraphyletic clades within this genetic cluster, and Wild-*striatus* cluster is a nested clade differentiated from the rest of the Wild-TMVB group (**Figure 2**). Remarkably, the domesticated populations integrate a monophyletic clade statistically well supported, suggesting a unique domestication event for the Mexican populations. Nevertheless, these phylogenetic hypotheses do not allow to distinguish the genetic pool from which domestication took place, although the Wild-SUR-CH genetic cluster can be discarded.

The ML and Neighbor-Net topologies in which *P. dumosus* and *P. vulgaris* were included, positioned *P. dumosus* as a sister group of *P. coccineus* (**Figure 2A**). However, the SplitsTree method indicated a basal reticulate pattern among *P. dumosus*, *P. coccineus,* and *P. vulgaris* (**Figure 2C**), suggesting ancestral gene flow, but not recent. Furthermore, there is no evidence of recent gene flow between wild and cultivated groups, but only within genetic clusters (**Figure 2C**).

Regarding SNAPP cloudgram (**Figures 3B,C**), 53 single topologies summarize the 95% HPD consensus tree, indicating a different divergence pattern in which Wild-TMVB populations are the closest clade to the domesticated group. Nevertheless, the complex assignment of individuals within Wild-TMVB and Wild-*striatus* are shown in a non-solved pattern within the cloudgram as well as in low values of nodal support in the consensus topology (**Figure 3C**). Despite these main inconsistencies between ML and Neighbor-Net vs. SNAPP topologies, all hypotheses favor the occurrence of a single domestication event.

In regards of the ABC-based computations, the model comparisons in preliminary trials indicated scenarios where the Wild-SMOCC population that are paraphyletic to Wild-TMVB

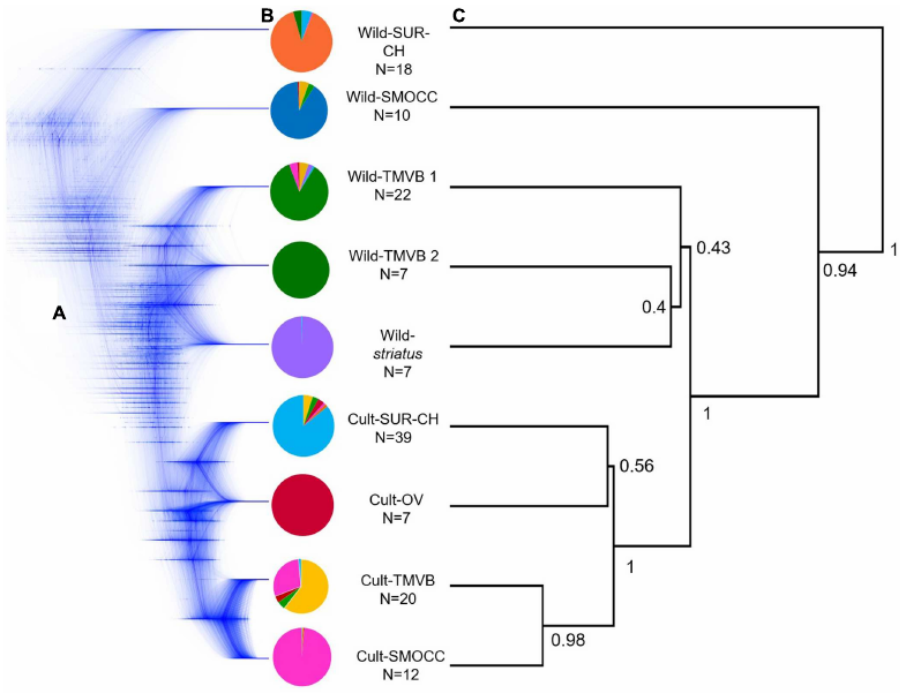


FIGURE 3 | (A) Cloudgram depicting topologies of 9,999 species trees obtained from an analysis of 600 single nucleotide polymorphism loci from 124 *P. coccineus* using SNAPP; **(B)** average assignment probability achieved by Admixture of selected individuals considered in species tree analyses based on nine groups; **(C)** associated root canal depicting a consensus topology from SNAPP analysis. Nodal support values on the root canal are posterior probabilities that correspond to strongly supported nodes designated *a priori* in the species tree analysis.

yielded a higher probability in both direct and logistic approaches (Supplementary Figures S1, S2). A final test indicated that the most likely scenario was a single domestication event, being the Wild-TMVB group the closest to the domesticated clade (**Figure 4**; Scenario 2, direct $P = 0.786$, logistic $P = 1.0$), which is congruent with the results of SNAPP phylogenetic analyses. Evaluation of the posterior predictions via PCA indicated that parameter values and summary statistics from the simulated datasets based on Scenario 1 closely matched the empirical data (Supplementary Figure S3).

Wild and Domesticated Population Genetics Statistics

High levels of genetic diversity were found in wild and cultivated populations (**Figure 5**). At the genetic cluster level, the Wild-TMVB group presented the highest diversity and the Cult-OV group the lowest. No clear pattern in the amount of diversity was observed between wild and cultivated clusters. There were cultivated groups with high genetic variance (Cult-SUR-CH and Cult-TMVB), and wild clusters that presented lower diversity than cultivated populations (Wild-SMOCC). At the location level (Supplementary Table S2), the samples from Spain ($H_E = 0.134$) and Oaxaca Valley ($H_E = 0.148$) presented the lowest diversity, and the highest was found in wild population located in Tlalpan, Mexico City ($H_E = 0.208$). Regarding species, *P. coccineus* showed the highest diversity and *P. dumosus* the lowest.

Outstandingly, H_O was greater than H_E in all the genetic groups except in the Wild-SUR-CH cluster, resulting in negative values of F_{IS}. Within the groups with an excess of observed heterozygosity, Wild-*striatus* had the lowest inbreeding coefficient (**Figure 5**). On the contrary, at the species level *P. vulgaris* showed a deficit of heterozygotes, showing a high F_{IS}. The inbreeding coefficient is positive when estimated taking into account all *P. coccineus* samples. This is caused by the Wahlund effect, which is the reduction of heterozygosity due to subpopulation structure. Regarding pairwise differentiation index, F_{ST} values ranged from 0.022 (Cult-TMVB vs. Cult-SMOCC) to 0.178 (Cult-OV vs. Wild-*striatus*; **Figure 6**). As expected, the pair F_{ST} values are greater between wild genetic groups than between cultivated genetic clusters (**Figure 6**).

Cultivated populations of *P. coccineus* show smaller effective population sizes than wild populations. In some cases, like in Cult-TMVB and Cult-SMOCC, *Ne* was one order of magnitude smaller than in the rest of the populations. On the contrary, the genetic cluster Wild-SUR-CH had the biggest *Ne* (**Table 1**). The most recent split was estimated to happen 3.9×10^3 generations ago, and occurred between the Cult-SMOCC and the Cult-TMVB clusters. On the contrary, the oldest split event was dated in 4.95×10^5 generations ago between the Wild-SUR-CH and the

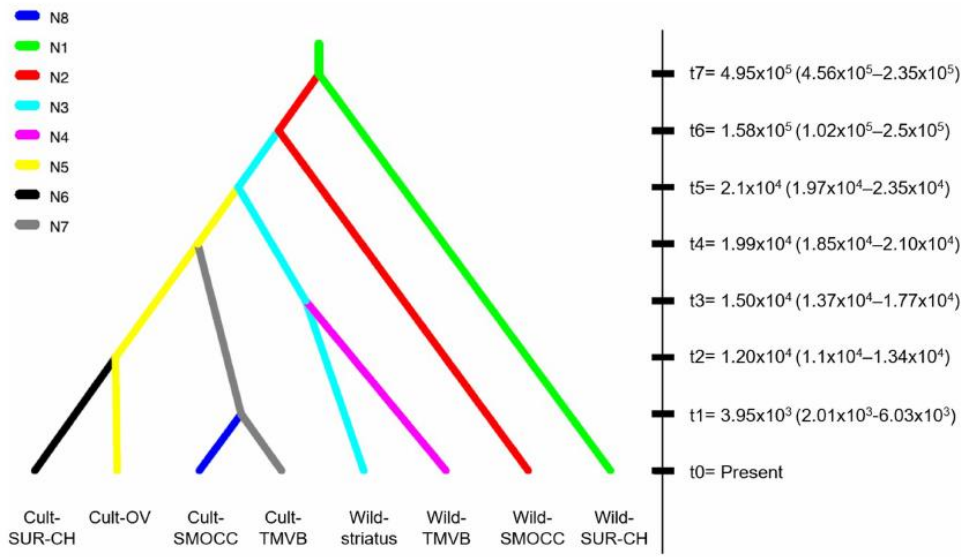


FIGURE 4 | Best-fitted domestication scenario of *P. coccineus* achieved with DIYABC. Split times in generations (tn) indicated the average posterior value estimated after Bayesian Computations (95% CI).

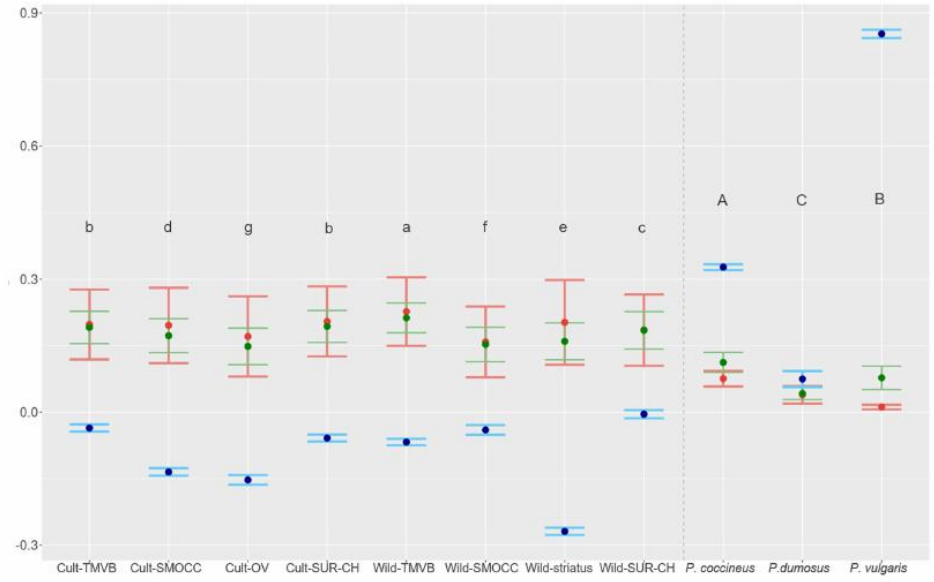


FIGURE 5 | H_E in green (variance interval), H_O in red (variance interval) and inbreeding coefficient in blue (95% IC) estimated for genetic clusters of *P. coccineus* and for species of VDC group. Letters show groups that are statistically different and are decreasingly ordered according H_E.

rest of *P. coccineus* clade. The split event that separates wild and domesticated samples was dated about 2.1×10^4 generations ago (**Figure 4**). Since *P. coccineus* is usually treated as an annual when cultivated, that represents 21,000 years ago. In the case of wild, perennial plants, one generation could be more than a year.

Identifying Candidate Loci

Before LD filtering, the mean r^2 value among SNPs located in the same chromosome separated by a maximum distance of 10,000 bp was 0.151. After eliminating SNPs closer than 3,000 bp, the mean r^2 was 0.063 (Supplementary Figure S5). In the case of

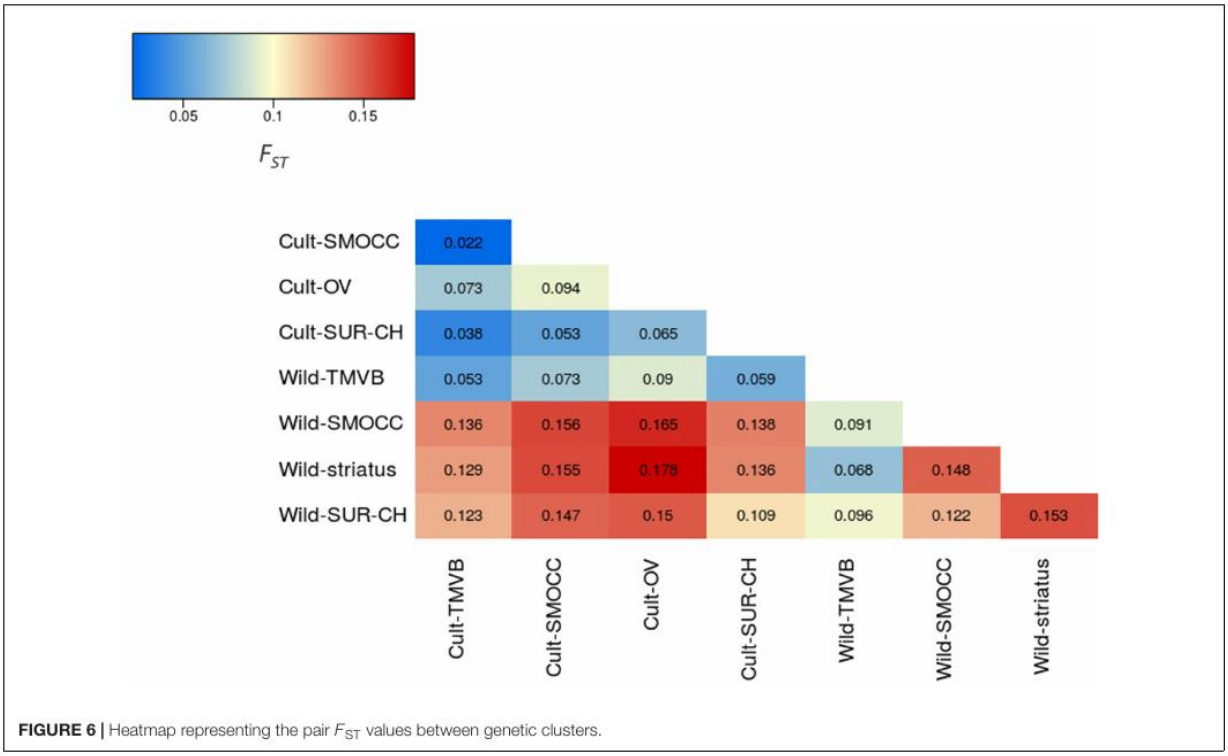


FIGURE 6 | Heatmap representing the pair F_{ST} values between genetic clusters.

SNPs from different chromosomes, the mean r^2 was 0.022. This low LD is not due to the closeness, but rather by factors like populations structure. Interestingly, the pattern in the decay of LD differed between genetic groups, with the fastest decay and lowest r^2 in cultivated and wild populations from the TMVB. Meanwhile, Wild-*striatus*, Wild-SURCH and Cult-OV had the slowest LD decay and highest r^2 values (Supplementary Figure S5). After filtering, the data set for candidate loci contained 11,693 SNPs distributed across the 11 chromosomes. In the central region of most of the chromosomes, there is a reduction in SNP density, probably due to centromeres (Supplementary Figure S6).

Using the pcadapt package, 47 SNPs were identified as candidate domestication loci; 342 involved in cultivar diversification; and 1,030 potentially under natural selection. Despite the great number of candidate SNPs that were identified, few are shared among selection types (Supplementary Figure S7). In the case of the BayeScan analyses, 469 candidate SNPs for domestication were identified; 16 related to cultivar diversification; and 12 candidates associated with natural selection. None of these SNPs were shared among the three BayeScan analysis.

Twenty four SNPs related to domestication, 13 to cultivar diversification and eight to natural selection were detected by both approaches and considered as candidate loci for further analyses (Supplementary Table S3). The genetic variance explained by the candidate SNPs compared to the 11,693 SNPs used previously changed dramatically (**Figure 7**). Notably, the genetic and geographic structure of wild and cultivated groups can be recovered by these few candidate SNPs (**Figures 7B,C**) and a clear separation of wild and domesticated populations is observed (**Figure 7A**).

Four SNPs of the candidate domestication loci were found to be annotated in *P. vulgaris* genome, one of the candidate loci under natural selection and none of the candidate loci for cultivar diversification (Supplementary Table S3). Three of the annotated candidate domestication loci (Phvul.001G232200, Phvul.007G256000, Phvul.009G156400) are highly expressed in flowers, flower buds or young pods, and the remaining locus (Phvul.002G145600) is highly expressed in green mature pods. All these loci have their highest similarity homologs in *G. max* genome v2.0 (Schmutz et al., 2010), but none of these correspond to the domestication-related loci previously identified by Zhou et al. (2015). The annotated candidate locus for natural selection (Phvul.003G197500) is highly expressed in roots and steam and corresponds to a calmodulin binding protein-like, which also has an homolog in *G. max*.

DISCUSSION

A Single Domestication Event for Mexican *P. coccineus* in the TMVB

Spataro et al. (2011) and Rodriguez et al. (2013), using SSR data, suggested two domestications events of *P. coccineus*, one in Mexico and the other in Guatemala-Honduras. The genomic

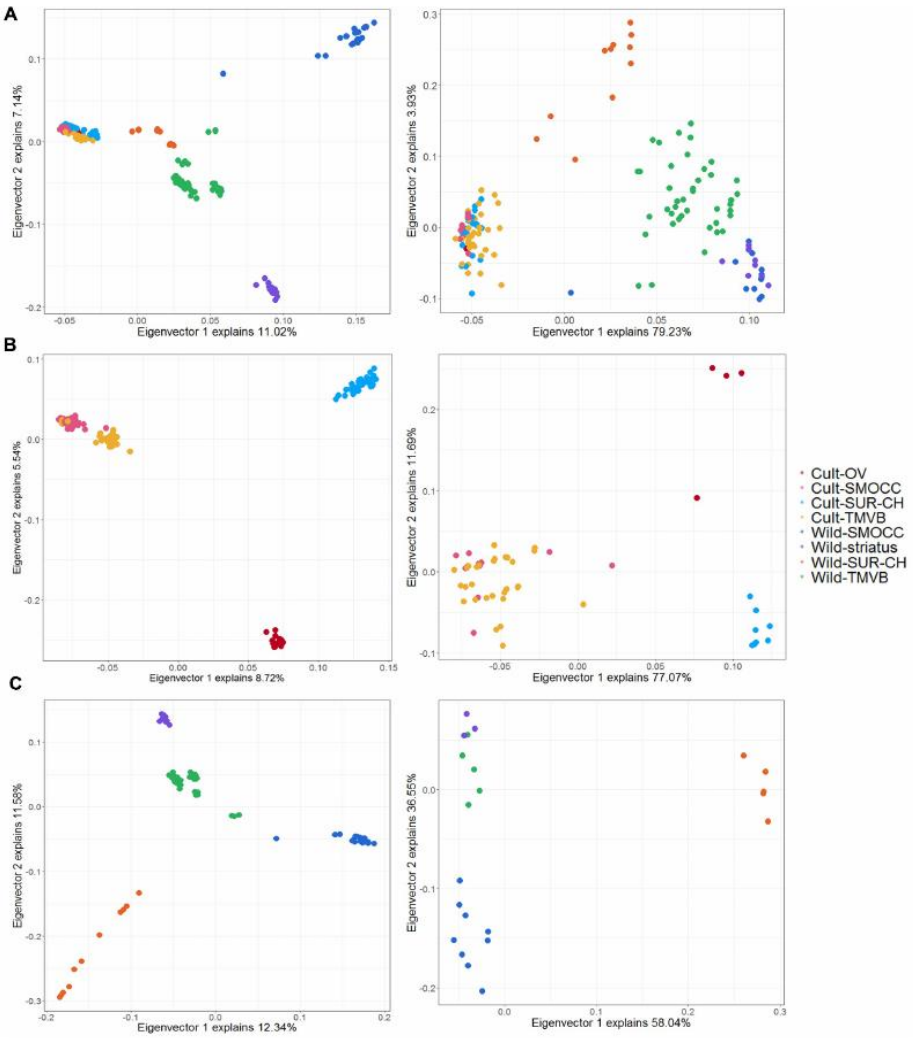


FIGURE 7 | Principal components analysis plot for the first two principal components using LD dataset (left) and candidate SNPs identify in both in PCAdapt and BayeScan analysis (right). **(A)** Analysis including cultivated and wild samples (no feral) to detect domestication-related SNPs. **(B)** Analysis of cultivated populations to distinguish cultivar diversification-related loci. **(C)** Analysis of wild samples to detect signatures of natural selection.

data generated in this work indicates a unique domestication event for the cultivated populations from Mexico (**Figures 2**, **3**). This includes Chiapas populations (Cult-SUR-CH), which are geographically and culturally closer to Guatemala than to Central and Northern Mexico. However, no samples from Guatemala and Honduras were included, therefore a second domestication event in this area cannot be discarded with the present data. Nonetheless, based on the results from SNAPP and DIYABC analyses, we were able to identify Wild-TMVB as the genetic pool from which domestication started in Mexico (**Figure 3**).

The most recent divergence time, that corresponds to the separation between cultivated groups of SMOCC and TMVB, was dated in 3,950 generations ago (**Figure 4**, t1). Assuming one generation per year in cultivated populations, this represents 3,950 years. But divergence between the cultivated and wild clades was dated in 21,000 generations (**Figure 4**, t5). This date is out of range of any plant domestication event and it seems unlikely. There are evolutionary processes that may affect these estimations. Processes like selection, population subdivision and incomplete lineage sorting may result in an overestimations of divergence times because increase the time to coalescence, that is, the time it takes for the two sequences to find their common ancestor (Albrechtsen et al., 2010; Angelis and Dos Reis, 2015). In *P. coccineus*, the selection made by humans during domestication and the high population structure in wild and domesticated groups probably has resulted in overestimated divergence times.

Also, it has to be considered that wild populations are perennial and thus generation times may be longer than a year.

The genetic findings suggest that *P. coccineus* domestication likely occurred from TMVB's material, pinpointing the domestication of this species to a particular region within the large Mexican territory where it is cultivated nowadays. Other sources of information could be incorporated to confirm this, using our findings as a geographic reference. If confirmed, identifying the TMVB as the area where domestication started for this species is interesting and important from an evolutionary, cultural and conservation perspective. The TMVB is the most recent mountainous region of Mexico, a biodiversity hotspot and it has a complex bio- and phylogeographic history characterized by following a sky-island dynamic during the last 2 Myr (Mastretta-Yanes et al., 2015). Culturally it became prominent during the Mexica Empire, and has been the most populated part of Mexico since little before the Spanish conquest (Bataillon, 1972). This has derived in other important cases of domestication to occur in this region. For instance, this central region was where the introgression of *Zea mays* ssp. *parviglumis* and *Z. mays* ssp. *mexicana* occurred during the domestication of maize (van Heerwaarden et al., 2011). However, human occupation in this area is also a concern for conservation, because the growth of urbanization and high-input agriculture in this area threat both *P. coccineus* landraces and wild populations (CONABIO and IUCN, 2016).

Besides genetic data, a Mexican domestication origin of *P. coccineus* is also supported by the several names that this bean has among different cultures. For instance, it is called tekómari in Chihuahua (Tarahumara indigenous language); tasukhu in Hidalgo and Puebla (Otomi); ayocote in central states of Mexico (Nahuatl); shaushana or xaxana in Veracruz (Totonaco); ma-má-ja (Mazateco) in Oaxaca; and botil or shbotil chenec in Chiapas (Tzeltal) (Salinas, 1988). Associated to these groups, there is also considerable traditional knowledge regarding the cultivation and use of *P. coccineus* species (e.g., Monroy and Quezada-Martínez, 2010).

Historic and Recent Gene Flow among Wild, Feral and Domesticated Populations

The individuals identified as feral clustered in the domesticated clade (**Figure 2A**), suggesting that they are escaped cultivars. This questions the hypothesis of an hybrid origin between wild and cultivated populations (Salinas, 1988) and contrasts with previous studies of feral *P. vulgaris* populations in Mexico (Papa and Gepts, 2003), where weedy populations appear to be genetically intermediate between domesticated and wild populations, and not cultivar escapees. Interestingly, the three collected feral populations belonged to the same genetic cluster (Cult-SUR-CH) and presented high levels of mixed ancestry, of which only a small proportion corresponds to wild clusters (**Figure 2B**). Since little evidence of gene flow was found in SplitsTree (**Figure 2C**), probably the mixed ancestry is due to shared polymorphisms or ancestral gene flow, rather than recent introgression events.

The breeding line Blanco Tlaxcala grouped with SMOCC landraces. Probably, breeding practices have acted over specific regions rather than over all the genome. The individuals of this breeding line did not present mixed ancestry, despite that Blanco Tlaxcala was developed using a multi linear method (Vargas-Vázquez et al., 2012). This suggests that all lines used to generate Blanco Tlaxcala belonged to the same genetic cluster (Cult-SMOCC), and they were submitted to several rounds of strong selection, decreasing genetic variation.

Contrary to what was reported by Spataro et al. (2011) and Rodriguez et al. (2013), samples from Spain clustered within the TMVB landraces, indicating that this European population was originated by the introduction of individuals of the Cult-TMVB group into Spain. Nevertheless, because just one European population was analyzed, no general pattern can yet be inferred. Notably, Spanish samples did not present mixed ancestry, meanwhile the rest of the individuals of this genetic group did (**Figure 2B**). Probably the genetic bottleneck that originated European populations and the isolation from wild relatives and American landraces, have decreased the amount of shared ancestral polymorphisms between cultivars from TMVB and Spain.

It has been suggested that hybridization and introgression have played a major role in *P. coccineus* evolution, both in cultivated and wild populations (Escalante et al., 1994; Angioi et al., 2009; Spataro et al., 2011; Rodriguez et al., 2013). Our results showed mixed ancestry both in wild and cultivated clusters. However, little evidence of introgression and hybridization was detected, and mixed ancestry can also be due to shared ancestral polymorphisms. Nevertheless, wild and cultivated populations frequently coexist, therefore hybridization cannot be discarded and a formal test considering the number and size of introgressed regions and the direction of gene flow must be done.

Phaseolus coccineus Is Highly Diverse and Structured

Phaseolus coccineus wild populations are divided in four genetic clusters that show considerable population differentiation. Similar levels of differentiation have been observed in several other highland species, which has been related to the high environmental variability and the complex geologic and climatic history of Mexico (Mastretta-Yanes et al., 2015). The extent of this differentiation in crop wild relative species has been mostly done with low resolution neutral makers (Bellon et al., 2009; Piñero et al., 2009) so it still needs to be further explored with genomic data. However, the present study and analyses in teosinte (van Heerwaarden et al., 2011; Aguirre-Liguori et al., 2017), highlight that there is high diversity contained in the genetic pools of crop wild relatives from Mexico.

Besides the diversity contained in wild relatives, one of the most important determinants in crop evolution is the level of genetic diversity contained in the domesticated populations, especially with reference to the wild ancestral gene pool. Genetic diversity reduction has been widely described in crop domestication (Hufford et al., 2013; Li et al., 2013; Schmutz

et al., 2014; Renaut and Rieseberg, 2015). This reduction of genetic diversity is caused by genetic drift resulting from population bottlenecks, and by artificial selection (Gepts, 2014). This phenomenon was also described in *P. vulgaris* (Schmutz et al., 2014) but in *P. coccineus* no clear pattern of genetic reduction was found between the wild or cultivated genetic groups (**Figure 5**). The Wild-TMVB cluster presented the highest genetic variation, followed by the Cult-TMVB and Cult-SUR-CH groups. On the contrary, the Cult-OV and Wild-SMOCC clusters showed the lowest H_E. Regarding effective population sizes, these were greater in wild than in cultivated genetic clusters, which is expected due to the genetic bottlenecks associated to domestication process. Nevertheless, *Ne* estimations of domesticated groups are in the order of 10^3–10^4. Taking together all results, these suggest that the genetic bottleneck during domestication was not severe. Other factors that may favor the maintenance of genetic diversity in *P. coccineus* are its high outcrossing rate (Escalante et al., 1994) and the fact that the genetic cluster from which domestication started (Wild-TMVB) presents the highest diversity. Little evidence of recent gene flow was detected, but early gene flow could also favor the amount of genetic diversity in cultivars.

Analyzing the genetic variance at the location level, Spanish samples presented the lowest diversity (Supplementary Table S2), which may be due to the recent demographic bottleneck that occurred during its introduction to Europe. Nevertheless, Oaxaca Valley also showed low genetic variation (Supplementary Table S2) and the ancestry analysis (**Figure 2B**) suggests that it has been genetically isolated from the other genetic clusters.

Regarding the inbreeding coefficient, the wild and cultivated genetic clusters presented negative F_{IS} values, indicating an excess of heterozygotes, except in the Wild-SUR-CH group. A possible explanation for this pattern is inbreeding depression, which effect in progeny has been studied in cultivars from Spain, finding that selfing affected germination, survival rate and seed weight (González et al., 2014). Also, a negative correlation was found between outcrossing rate and seed abortion in wild populations studied by Escalante et al. (1994). In the case of domesticated populations, the bottlenecks that they suffered during domestication may promote the accumulation of deleterious alleles and the increase of inbreeding depression, resulting in lower values of the inbreeding coefficient (Morrell et al., 2011). Opposite to what was expected, in *P. coccineus* the population with the lowest F_{IS} was a wild cluster (Wild-striatus). This population was previously studied by Búrquez and Sarukhán (1984), who found evidence of self-incompatibility, which is congruent with our results. A possible explanation for this pattern is the accumulation of deleterious alleles in the Wild-striatus cluster. Notably, no mixed ancestry was detected in this genetic group, indicating that it is genetically isolated from other populations despite being geographically close to other wild and cultivated TMVB populations. It is necessary to evaluate other populations of *P. coccineus* subsp. *striatus* to know if this is a common pattern and to explore the ecological and genetic causes and consequences of it.

Adaptative Variation in Wild and Domesticated Populations

Mexico is an environmentally and culturally heterogeneous country, which favored crop genetic diversity. The distribution of *Phaseolus*, both cultivated and wild, involves an interaction with a wide range of different cultures, and isolated populations are exposed to diverse environmental conditions. For example, compared to *P. vulgaris*, *P. coccineus* grows in more humid environments, at cooler temperatures and at higher altitudes. Nevertheless, there are few studies that aim to elucidate the genetic basis of adaptation, especially for the wild populations of *Phaseolus* crop species (Bitocchi et al., 2017). Our outlier analyses listed some candidate SNPs that could be under artificial selection during the domestication and diversification stages, and others that could be under natural selection. Although most of these outliers are still not annotated, they could serve as a base for identifying population differentiation in adaptive variation, which is a needed step for genetic resources and crop wild relatives conservation (Maxted et al., 2012). Our study is based on GBS data, so *P. coccineus* genome is not fully saturated, and likely there are loci under selection that we did not sample. Nevertheless, this set of outliers are a first approximation to identify candidate loci to domestication and natural selection in runner bean.

The fact that no loci overlapped between domestication, diversification and natural selection categories shows that different selective processes were detected. This is to be expected because, in general, loci under natural selection and artificial selection related to domestication and diversification are expected to differ across the genome (Meyer and Purugganan, 2013).

The loci involved in domestication are expected to be specially related to the phenotypic changes of the domestication syndrome (Koinange et al., 1996), that is modifications in morphological and physiological traits like seed dispersal, seed dormancy, gigantism, increased harvest index and flowering time (Hammer, 1984). Most of the domestication-related loci identified here are still of unknown function, but the four that are annotated are highly expressed in flowers or pods (Supplementary Table S3). This is interesting because in the soybean, another legume, several domestication-related loci associated with flowering time have been identified (Zhou et al., 2015). However, no overlap among those loci and the ones identified here was found.

CONCLUSION

The SNPs generated in this work provided high resolution data to understand the domestication of *P. coccineus*. Results suggest one domestication event for Mexico, which started from the wild genetic pool from TMVB. Furthermore, wild and domesticated populations are highly diverse and presented high values of *Ne*, suggesting that the demographic bottleneck due to domestication was not severe. These genomic analyses allow to highlight how the genetic signatures of domestication can be substantially different even between species of the same genus domesticated in the same geographic area. Common bean and

scarlet runner bean are closely related species, nevertheless their reproductive strategies and domestication histories seem to be different: *P. vulgaris* tends to self-crossing, which theoretically facilities the domestication process, and it also suffered a severe domestication bottleneck. On the contrary, *P. coccineus* is an open pollinated species that presents high levels of genetic diversity and population structure, and its domestication did not result in a strong demographic bottleneck.

Our findings also show that both wild and domesticated populations of *P. coccineus* are highly structured. Most of the genetic clusters presented an heterozygotes excess, showing evidence of inbreeding depression. Interestingly, the population identified as *P. coccineus* subsp. *striatus* shows the greatest excess of heterozygotes and seems to be genetically isolated from other wild and cultivated populations. Contrasting with previous studies, our data shows that gene flow within and between wild and cultivated populations is not a common process. Fully testing this represents an area where further research is needed.

The levels of diversity and population differentiation found here support that the runner bean is a potential source of variability for several traits for plant breeding (Schwember et al., 2017). The data presented here highlights that for a better characterization of *P. coccineus* wild and cultivated forms there is still a need of more sampling, specially including Central American populations. Complete and annotated genomes of *Phaseolus* and other legume crops will facilitate not only comparative genomics, but will give a better knowledge of the evolution and domestication of this group of plants that has been independently domesticated by several human groups across its distribution.

AUTHOR CONTRIBUTIONS

AG-G, DP, and AD-S designed the study. AG-G made the molecular procedures. AG-G, AM-Y, and MS-A conducted the analyses. All authors revised the results and wrote the manuscript.

FUNDING

This work was supported by Consejo Nacional de Ciencia y Tecnología through the Ph.D. scholarship number 440709 to AG-G and CONACYT Grant 247730 to DP.

ACKNOWLEDGMENTS

We thank Idalia Rojas, Myriam Campos, Erick García, Verónica González, Alfredo Villarruel, Nancy Gálvez, and Rocío González for fieldwork assistance, Tania Garrido for laboratory technical assistance and Ernesto Campos Murillo for bioinformatic assistance to execute analyses in a cluster environment. We acknowledge funding from the CONACYT grant number 247730 and IEUNAM to DP. Statistical analyses were carried out in the 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. This work constitutes a partial fulfillment of the Posgrado en Ciencias Biológicas at the Universidad Nacional Autónoma de México (UNAM) for AG-G. Finally, we thank to all farmers that share with us their seeds and knowledge.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2017.01891/full#supplementary-material

REFERENCES

Aguirre-Liguori, J. A., Tenaillon, M. I., Vázquez-Lobo, A., Gaut, B. S., Jaramillo-Correa, J. P., Montes-Hernandez, S., et al. (2017). Connecting genomic patterns of local adaptation and niche suitability in teosintes. *Mol. Ecol.* 26, 4226–4240. doi: 10.1111/mec.14203

Albrechtsen, A., Nielsen, F. C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27, 2534–2547. doi: 10.1093/molbev/msq148

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

Angelis, K., and Dos Reis, M. (2015). The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. *Curr. Zool.* 61, 874–885. doi: 10.1093/czoolo/61.5.874

Angioi, S. A., Desiderio, F., Rau, D., Bitocchi, E., Attene, G., and Papa, R. (2009). Development and use of chloroplast microsatellites in *Phaseolus* spp. and other legumes. *Plant Biol.* 11, 598–612. doi: 10.1111/j.1438-8677.2008.00143.x

Bataillon, C. (1972). *La Ciudad y el Campo en el México Central.* Available at: https://books.google.co.in/books/about/La_ciudad_y_el_campo_en_el_M%C3%A9xico_Centr.html?hl=&id=h_grAAAAMAAJ&redir_esc=y

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–2035.

Bellon, M. R., Barrientos-Priego, A. F., Colunga-García, M. P., Perales, H., Reyes-Agüero, J. A., Rosales-Serna, R., et al. (2009). Diversidad y conservación de recursos genéticos en plantas cultivadas. *Capital Nat. México* 2, 355–382.

Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M. L., Gioia, T., et al. (2017). Beans (*Phaseolus* ssp.) as a model for understanding crop evolution. *Front. Plant Sci.* 8:722. doi: 10.3389/fpls.2017.00722

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for bayesian evolutionary analysis. *PLOS Comp. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Bouckaert, R. R. (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26, 1372–1373. doi: 10.1093/bioinformatics/btq110

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29, 1917–1932. doi: 10.1093/molbev/mss086

Búrquez, A., and Sarukhán, J. (1984). Biología floral de poblaciones silvestres de *Phaseolus coccineus* L. II. Sistemas reproductivos. *Bol. Soc. Bot. México* 46, 3–12.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. doi: 10.1186/s13742-015-0047-8

CONABIO and IUCN (2016). *Resultados del Segundo Taller del Proyecto Salvaguardando los Parientes Silvestres de Plantas Cultivadas*. Gland: Unión Internacional para la Conservación de la Naturaleza.

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

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Delgado-Salinas, A., Bibler, R., and Lavin, M. (2006). Phylogeny of the genus phaseolus (Leguminosae): a recent diversification in an ancient landscape. *Syst. Bot.* 31, 779–791. doi: 10.1600/036364406779695960

Drummond, A. J., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Eguiarte, L. E., Aguirre-Planter, E., Aguirre, X., Colín, R., González, A., Rocha, M., et al. (2013). From isozymes to genomics: population genetics and conservation of *Agave* in México. *Bot. Rev.* 79, 483–506. doi: 10.1007/s12229-013-9123-x

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Escalante, A. M., Coello, G., Eguiarte, L. E., and Pinero, D. (1994). Genetic structure and mating systems in wild and cultivated populations of *Phaseolus coccineus* and *P. vulgaris* (Fabaceae). *Am. J. Bot.* 81:1096. doi: 10.2307/2445471

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

Freytag, G. F., and Debouck, D. G. (2002). *Taxonomy, Distribution, and Ecology of the Genus Phaseolus (Leguminosae-Papilionoideae) in North America, Mexico and Central America*. Forth Worth, TX: Botanical Research Institute of Texas (BRIT).

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

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLOS ONE* 9:e90346. doi: 10.1371/journal.pone.0090346

González, A. M., De Ron, A. M., Lores, M., and Santalla, M. (2014). Effect of the inbreeding depression in progeny fitness of runner bean (*Phaseolus coccineus* L.) and it is implications for breeding. *Euphytica* 200, 413–428. doi: 10.1007/s10681-014-1177-2

Goudet, J. (2005). hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 5, 184–186. doi: 10.1111/j.1471-8286.2004.00828.x

Hammer, K. (1984). Das Domestikationssyndrom. *Kulturpflanze* 32, 11–34. doi: 10.1007/BF02098682

Hufford, M. B., Lubinksy, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., and Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLOS Genet.* 9:e1003477. doi: 10.1371/journal.pgen.1003477

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

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Instituto Nacional de Estadística, Geografía e Informática (INEGI), Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), and Instituto Nacional de Ecología (INE) (2008). *Ecorregiones Terrestres de México*. Available at: http://www.conabio.gob.mx/informacion/gis/

Koinange, E. M. K., Singh, S. P., and Gepts, P. (1996). Genetic control of the domestication syndrome in common bean. *Crop Sci.* 36, 1037–1045. doi: 10.2135/cropsci1996.0011183x003600040037x

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, Y.-H., Zhao, S.-C., Ma, J.-X., Li, D., Yan, L., Li, J., et al. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* 14:579. doi: 10.1186/1471-2164-14-579

López-Soto, J. L., Ruiz-Corral, J. A., Sánchez-González, J. J., and Lépiz-Ildelfonso, R. (2005). Adaptación climática de 25 especies de frijol silvestre (*Phaseolus* spp) en la República Mexicana. *Rev. Fitotec. Mex.* 28, 221–230.

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

Mastretta-Yanes, A., Moreno-Letelier, A., Piñero, D., Jorgensen, T. H., and Emerson, B. C. (2015). Biodiversity in the Mexican highlands and the interaction of geology, geography and climate within the Trans-Mexican Volcanic Belt. *J. Biogeogr.* 42, 1586–1600. doi: 10.1111/jbi.12546

Maxted, N., Kell, S., Ford-Lloyd, B., Dulloo, E., and Toledo, Á. (2012). Toward the systematic conservation of global crop wild relative diversity. *Crop Sci.* 52, 774–785. doi: 10.2135/cropsci2011.08.0415

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

Monroy, R., and Quezada-Martínez, A. (2010). *Estudio Etnobotánico del frijol Yepatlaxtle (Phaseolus coccineus L.), en el Área Natural Protegida Corredor Biológico Chichinautzin, Morelos, México*. Cuernavaca: Universidad Autónoma del Estado de Morelos.

Morrell, P. L., Buckler, E. S., and Ross-Ibarra, J. (2011). Crop genomics: advances and applications. *Nat. Rev. Genet.* 13, 85–96. doi: 10.1038/nrg3097

Papa, R., and Gepts, P. (2003). Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theor. Appl. Genet.* 106, 239–250. doi: 10.1007/s00122-002-1085-z

Phytozome (2017). Available at: http://phytozome.jgi.doe.gov/ [accessed June 26, 2017].

Piñero, D., Caballero-Mellado, J., Cabrera-Toledo, D., et al. (2009). "La diversidad genética como instrumento para la conservación y el aprovechamiento de la biodiversidad: estudios en especies mexicanas," in *Capital Natural de México, p. 619. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad*, eds J. Sarukhán, J. Soberón, G. Halffter and J. Llorente Bousquets (Mexico: Comisión Nacional para el Conocimiento y Uso de la Biodiversidad).

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rannala, B., and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164, 1645–1656.

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

Rodiño, A. P., Paula Rodiño, A., Lema, M., Pérez-Barbeito, M., Santalla, M., and De Ron, A. M. (2006). Assessment of runner bean (*Phaseolus coccineus* L.) germplasm for tolerance to low temperature during early seedling growth. *Euphytica* 155, 63–70. doi: 10.1007/s10681-006-9301-6

Rodriguez, M., Rau, D., Angioi, S. A., Bellucci, E., Bitocchi, E., Nanni, L., et al. (2013). European *Phaseolus coccineus* L. landraces: population structure and adaptation, as revealed by cpSSRs and phenotypic analyses. *PLOS ONE* 8:e57337. doi: 10.1371/journal.pone.0057337

Salinas, A. D. (1988). "Variation, taxonomy, domestication, and germplasm potentialities in phaseoluscoccineus," in *Genetic Resources of Phaseolus Beans. Current Plant Science and Biotechnology in Agriculture*, Vol. 6, ed. P. Gepts (Dordrecht: Springer), 441–463. doi: 10.1007/978-94-009-2786-5_18

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

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

Schwember, A. R., Carrasco, B., and Gepts, P. (2017). Unraveling agronomic and genetic aspects of runner bean (*Phaseolus coccineus* L.). *Field Crops Res.* 206, 86–94. doi: 10.1016/j.fcr.2017.02.020

Services | Institute of Biotechnology (2017). Available at: http://www.biotech.cornell.edu/brc/genomic-diversity-facility/services [accessed June 26, 2017].

Sokal, R. R., and Rohlf, F. J. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd Edn. New York: W. H. Freeman and Co.

Spataro, G., Tiranti, B., Arcaleni, P., Bellucci, E., Attene, G., Papa, R., et al. (2011). Genetic diversity and structure of a worldwide collection of *Phaseolus coccineus* L. *Theor. Appl. Genet.* 122, 1281–1291. doi: 10.1007/s00122-011-1530-y

Tang, H., Sezen, U., and Paterson, A. H. (2010). Domestication and plant genomes. *Curr. Opin. Plant Biol.* 13, 160–166. doi: 10.1016/j.pbi.2009.10.008

van Heerwaarden, J., Doebley, J., Briggs, W. H., Glaubitz, J. C., Goodman, M. M., de Jesus Sanchez Gonzalez, J., et al. (2011). Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1088–1092. doi: 10.1073/pnas.1013011108

Vargas-Vázquez, L. P., Muruaga-Martínez, J. S., Lépiz-Ildefonso, R., and Pérez-Guerrero, A. (2012). La colección INIFAP de frijol ayocote (*Phaseolus coccineus* L.) I. Distribución geográfica de sitios de colecta. *Rev. Mex. Cien. Agríc.* 3, 1247–1259.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi: 10.1038/nbt.3096

Conflict of Interest Statement: 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.

The handling Editor declared a shared affiliation and past co-authorship, though no other collaboration, with the authors and states that the process nevertheless met the standards of a fair and objective review.

Capítulo 4:

Discusión

La domesticación es un proceso complejo en el que interactúan fuerzas naturales y antropogénicas, y que ha dejado huellas en el genoma de las especies manejadas por el humano, las cuales es posible rastrear para obtener información sobre la historia de este proceso. Sin duda, el cúmulo de información genética y evolutiva de las especies domesticadas es significativo, especialmente en el caso de los cultivos de gran importancia económica, los cuales se han convertido en modelos de estudio. Sin embargo, aún existen lagunas en el conocimiento y dos de las principales inquietudes de Darwin con respecto a las especies domesticadas siguen vigentes: cuáles son las causas de la variación y cuál es el efecto de la selección.

El desarrollo de herramientas tecnológicas para la secuenciación masiva ha acelerado la obtención de información genómica, lo que ha permitido poner a prueba las hipótesis sobre las consecuencias de la domesticación en los genomas, y ha facilitado la identificación de genes o mutaciones responsables de los cambios fenotípicos observados en los cultivos. No obstante, grandes retos permanecen: la obtención de información fenotípica (fenotipificar) a gran escala de forma rápida y precisa; aumentar las capacidades informáticas para el almacén y análisis de los datos generados; el estudio de genomas grandes y complejos (ej., especies poliploides); el manejo de genes parálogos y regiones repetitivas; la identificación de variantes causales en atributos poligénicos; y expandir el campo de estudio a especies no modelo y a los parientes silvestres.

El presente trabajo contribuye a llenar el vacío de información del último de los retos antes mencionados. *Phaseolus coccineus* es una especie de frijol domesticada en Mesoamérica que ha recibido relativamente poca atención y de la que se sabe poco desde el punto de vista genético. Con la finalidad de conocer los patrones de diversidad y diferenciación genética de esta especie, y con ésto hacer inferencias acerca de su historia de domesticación, se colectaron más de 240 individuos pertenecientes a poblaciones silvestres, ferales y cultivos de México, además de muestras de un cultivo proveniente de España. A partir de estas muestras, se obtuvieron decenas de miles de marcadores moleculares usando herramientas de secuenciación masiva.

Estructura poblacional de *Phaseolus coccineus* y el origen de sus cultivos

Los datos presentados aquí muestran que las poblaciones cultivadas y silvestres de *P. coccineus* presentan una gran estructuración, identificándose cuatro grupos genéticos silvestres (Wild-TMVB, Wild-striatus, Wild-SMOCC y Wild-SUR) y cuatro domesticados (Cult-TMVB, Cult-SMOCC, Cult-OV y Cult-SUR).

Todas las muestras de cultivos integran un clado monofilético, lo que sugiere un evento de domesticación único, al menos en los cultivares de México. Este resultado contradice lo propuesto por Spataro et al. (2011) y Rodriguez et al. (2013), quienes usando marcadores de menor resolución (microsatélites nucleares y de cloroplasto, respectivamente), sugieren dos eventos de domesticación, uno en Guatemala y un segundo en México. Sin embargo, Spataro et al. (2011) y Rodriguez et al. (2013) incluyen un bajo número de muestras de cultivos mexicanos (31 individuos) y de parientes silvestres (siete individuos). No obstante, para poder descartar una segunda domesticación en Centroamérica es necesario en futuros trabajos incluir muestras de cultivos y poblaciones silvestres de esta región.

El grupo genético integrado por las poblaciones silvestres de la Faja Volcánica Transmexicana (Wild-TMVB) es el más cercano a los cultivos, lo que indica que la domesticación ocurrió a partir de estas poblaciones. Esto es consistente con la alta diversidad genética presente en los cultivos de esta región y con los resultados del DIYABC, en cual el mejor escenario también establece a las poblaciones de la Faja Volcánica como las más cercanas a los cultivos. Sin embargo, un muestreo más amplio de los parientes silvestres, incluyendo muestras de Centroamérica, modelar posibles eventos de flujo genético y considerar únicamente loci putativamente neutrales puede ayudar a profundizar más en la historia de domesticación y evolutiva de la especie.

Fueron clasificados como ferales aquellos individuos que presentaban características de domesticados pero que se encontraban creciendo fuera de campos de cultivo. Se ha hipotetizado que las poblaciones ferales pueden ser el resultado de flujo genético entre poblaciones domesticadas y silvestres (Salinas, 1988), lo cual se ha reportado en *P. vulgaris* (Papa & Gepts, 2003). Por el contrario, los datos obtenidos en este trabajo sitúan a los ferales colectados dentro del clado de los cultivados, sugiriendo que los ferales son cultivos escapados.

En el caso de los individuos del cultivo de España, se encontró que pertenecen al grupo genético de los domesticados de la Faja Volcánica Transmexicana (Cult-TMVB). Lo anterior sugiere que estos cultivares europeos se originaron a partir de poblaciones domesticadas del

centro de México. Sin embargo, con un solo cultivar europeo incluído no es posible hacer inferencias generales de los cultivos de *P. coccineus* del Viejo Continente.

Flujo genético e introgresión

El flujo genético puede tener un rol importante en los patrones de diversidad y diferenciación en las formas cultivadas y silvestres de las especies domesticadas, especialmente si se considera la acción combinada del flujo asimétrico (Papa & Gepts, 2003) y la selección sobre ciertos loci (Papa et al., 2005). En el caso de *P. coccineus*, se ha sugerido que la introgresión y la hibridación han jugado un papel muy importante tanto en las formas silvestres como en las cultivadas (Escalante et al., 1994; Angioi et al., 2009; Spataro et al., 2011; Rodriguez et al., 2013).

La red filogenética obtenida con SplitsTree, el cual permite hacer inferencia de hibridación, muestra reticulaciones basales, especialmente en los grupos genéticos Cult-SUR y Wild-SUR. Lo anterior junto con la cercanía de Wild-SUR con los cultivares en el Análisis de Componentes Principales (PCA por sus siglas en inglés) podría sugerir flujo genético entre las poblaciones domesticadas y las silvestres del sur de México. Por otro lado, el análisis de ancestría realizado (Admixture) muestra que es frecuente la ancestría mezclada en los grupos genéticos de las formas silvestres y domesticadas. Los anterior no es prueba directa de flujo genético, ya que otros procesos, como polimorfismos ancestrales compartidos, pueden explicar el patrón de ancestría encontrado.

Para poder conocer el papel de la hibridación e introgresión en la historia evolutiva y de domesticación de *P. coccineus* es necesario llevar a cabo análisis específicos que pongan a prueba el flujo genético entre las poblaciones. Dentro de los análisis que pueden llevarse a cabo está TreeMix desarrollado por Pickrell y Pritchard (2012) y el cual se basa en los cambios en las frecuencias alélicas. Un análisis más es la D de Patterson, también conocido como ABBA-BABA (Green et al., 2010; Durand et al., 2011), el cual determina si la proporción de estados derivados está influenciada por flujo genético.

Diversidad genética y el cuello de botella de la domesticación

En especies alógamas, como en el maíz, se espera que la domesticación produzca un cuello de botella menos severo comparado con especies autógamas, como es el caso del frijol común (Schmutz et al., 2014). Esto ha sido confirmado en especies autógamas como la soya (*Glycine*

max; Lam et al., 2010) y el arroz (*Oryza sativa*; Xu et al., 2012), en donde se ha reportado una importante reducción de la diversidad genética resultado del proceso de domesticación. En el caso de *P. coccineus*, que presenta un sistema de polinización abierto, los resultados indican que el cuello de botella consecuencia de la domesticación no fue severo, lo que permitió el mantenimiento de una variación genética relativamente alta en los cultivos.

Cabe destacar que cada grupo genético parece tener su propia historia y no hay un patrón claro al comparar los niveles de diversidad de las poblaciones silvestres y cultivadas. Por ejemplo, los cultivos de la Faja Volcánica (Cult-TMVB) poseen niveles de heterocigosidad similares a poblaciones silvestres. Por el contrario, los grupos genéticos silvestres de la Sierra Madre Occidental (Wild-SMOCC) y de *P. coccineus* subsp. *striatus* (Wild-striatus) presentan menor variación genética que algunos cultivos de la misma especie.

Las poblaciones que muestran la menor heterocigosis son los cultivos de Oaxaca (Tabla S2; H_E = 0.148) y de España (H_E = 0.134). Pese a no haber un patrón claro de pérdida de diversidad resultado de la domesticación, es probable que los bajos niveles de heterocigosis de estas poblaciones sean resultado de un segundo cuello de botella, particularmente en el caso de la población española, que además de pasar por un cuello de botella subsecuente al inicial de la domesticación, ha estado aislada de parientes silvestres, anulando la posibilidad de flujo genético.

Los tamaños efectivos estimados con DIYABC son en general mayores en los grupos genéticos silvestres en comparación con los cultivados, lo cual puede ser explicado por el cuello de botella asociado a la domesticación. Sin embargo, las estimaciones de tamaños poblaciones efectivos, así como los tiempos de divergencia puede estar sesgados por procesos de selección natural y/o artificial, así como por flujo genético. Por tanto, estimaciones considerando loci neutrales y probando escenarios con flujo genético son necesarios para obtener valores más precisos.

La historia demográfica de las poblaciones tiene un gran efecto en los patrones de diversidad de los genomas. Por tanto, para poder entender el proceso de domesticación e identificar regiones genómicas bajo selección, es necesario conocer los procesos demográficos por los que han pasado las poblaciones silvestres y cultivadas.

Huellas de selección en el genoma de *Phaseolus coccineus*

Uno de los objetivos de este trabajo fue detectar firmas de selección natural y artificial en el genoma de *P. coccineus*. Para esto, se implementaron dos métodos (pcadapt y Bayescan)

basados en la detección valores extremos (*outliers*) de diferenciación. Bayescan encuentra altas diferencias en las frecuencias alélicas de acuerdo a estadísticos derivados de F_{ST} (Foll y Gaggiotti, 2008). Dentro de las desventajas que presenta Bayescan es que no toma en cuenta la estructura jerárquica que pueden presentar las poblaciones y es necesario preasignar a los individuos a grupos. Por otro lado, pcadapt asume que los marcadores excesivamente relacionados con la estructura poblacional son candidatos a adaptación local (Luu et al., 2017). pcadapt recupera la estructura de las poblaciones de acuerdo a la agrupación de los individuos en el Análisis de Componentes Principales, aún si la estructura es jerárquica; además los individuos no son asignados a grupos y puede analizar datos de poblaciones con ancestrías mezcladas y/o flujo genético (Luu et al., 2017).

Los métodos basados en detección de *outliers* presentan una alta tasa de falsos positivos, si no se cumplen con los supuestos como estructura poblacional no anidada, no cambios en los tamaños poblaciones, y no flujo genético. Por ejemplo, en el caso de Bayescan se ha registrado hasta 40% de falsos positivos cuando existe flujo genético entre las poblaciones (Luu et al., 2017). Es por esto que se tomó una postura conservadora y únicamente se consideraron como candidatos aquellos loci que fueron detectados por ambos algoritmos.

El conjunto de datos fue subdividido para poder identificar loci relacionados con tres procesos selectivos distintos: adaptación, domesticación y diversificación de cultivos. Se detectaron 24 SNPs relacionados con domesticación, 13 con diversificación de cultivos, y ocho con selección natural. Las diferencias entre la cantidad de marcadores detectados en cada método pueden deberse a que sus algoritmos y supuestos son distintos. Por ejemplo, la estructura jerárquica que presentan las poblaciones, particularmente las silvestres, no es recuperada usando Bayescan; y un número importante de individuos presentan ancestría mezclada, posiblemente debido a flujo genético o a polimorfismos compartidos ancestrales. Posteriores análisis en donde se tomen en cuenta escenarios con flujo genéticos, cambios demográficos, y que representan mejor la compleja estructura de las poblaciones son necesarios para identificar con mayor precisión regiones que han estado bajo procesos selectivos.

Cuatro de los SNPs relacionados a domesticación corresponden a genes codificantes que se han reportado altamente expresados en flores y vainas, estructuras que han presentado grandes cambios durante la domesticación. Por otro lado, sólo uno de los loci relacionados a selección natural estaba anotado, y su expresión se ha reportado en la raíz y en el tallo. Sin embargo, la mayoría de los loci con señales de selección no presentan funciones anotadas.

Este estudio tiene varias limitantes para la detección de firmas de selección: 1) la baja densidad de SNPs que se deriva de la técnica de GBS ya que sólo se muestrea una fracción del genoma y es posible que no se tengan representadas regiones que están bajo procesos selectivos; 2) no se cuenta con un genoma de *P. coccineus*, por lo que únicamente se analizaron variantes que se alinearon contra el genoma del frijol común; 3) Bayescan asume que las poblaciones o grupos contrastados son discretos, lo cual no se cumplirían en caso de haber hibridación. Pese a las limitantes metodológicas, este es el primer trabajo en *P. coccineus* que, usando datos genómicos, busca identificar regiones del genoma afectadas por selección.

Conclusiones y perspectivas

Para poder aprovechar los recursos genéticos de las especies cultivadas es necesario tener un conocimiento profundo de su historia evolutiva y, por tanto, estudiar a sus parientes silvestres. Los trabajos de genética de poblaciones en especies domesticadas permiten: 1) conocer la cantidad de diversidad disponible y que puede ser utilizada con fines de mejoramiento; 2) conocer las bases genéticas de atributos fenotípicos; 3) entender procesos evolutivos y generar conocimiento que puede extrapolarse a especies no modelo. Este tipo de estudios son especialmente importante en países como México, donde se originaron un número importante de cultivos, donde formas silvestres y cultivadas coexisten, y donde los humanos han interactuado con otras especies por miles de años de diversas formas.

Resultado de este trabajo se generaron datos genómicos de poblaciones silvestres y cultivadas de *P. coccineus* en México, que hasta ahora era una especie poco estudiada desde el punto de vista genético. Los resultados muestran que el ayocote es una especie con alta estructura poblacional y diversidad genética. Por otro lado, los datos sugieren un evento de domesticación, el cual tuvo lugar en la Faja Volcánica Transmexicana. Finalmente, se identificaron regiones candidatas a selección natural y artificial.

Esta tesis ofrece un primer acercamiento y un panorama general de los patrones de diversidad, diferenciación e historia de domesticación de *P. coccineus*, y abre la puerta a nuevas preguntas evolutivas: 1) ¿Cuál ha sido el papel del flujo genético en la historia de *P. coccineus*? Esta especie presenta un sistema de polinización abierto, lo que podría facilitar el flujo entre poblaciones silvestres, y entre cultivos y parientes silvestres. 2) ¿Hace cuánto tiempo comenzó la domesticación de la especie? Estimar la divergencia de las poblaciones silvestres y cultivadas usando marcadores en regiones no codificantes, evitando sesgos dados por procesos selectivos. 3) ¿Coinciden el tiempo y región geográfica de domesticación del ayocote con el de otras especies de frijoles?.

Los procesos demográficos afectan a todo el genoma, por lo que no es necesario contar con una alta densidad de marcadores genéticos para hacer inferencias acerca de la historia de las poblaciones. Por el contrario, los procesos selectivos únicamente tienen efecto sobre ciertas regiones y sus alrededores. Para poder mejorar la detección de firmas de selección, tanto natural como artificial, contar con una mayor cantidad de marcadores moleculares que brinden un panorama más amplio y preciso de los patrones de variación en el genoma de *P. coccineus* permitiría identificar nuevas regiones candidatas a selección. Lo anterior podría lograrse empleando técnicas como resecuenciación de genomas completos y contando con un genoma

de referencia de la especie. Otro enfoque que podría usarse para identificar genes afectados en la domesticación es realizar experimentos de expresión diferencial, especialmente en tejidos que han sido modificados en los cultivos, como las vainas y flores.

Estudios genéticos previos sugieren dos eventos de domesticación, uno de ellos en Guatemala y el segundo en México. Contrario a esto, nuestros resultados apoyan un único evento de domesticación, el cual comenzó a partir del grupo genético silvestre de la Faja Volcánica Transmexicana. Sin embargo, para poner a prueba la hipótesis del segundo evento en Centroamérica es necesario analizar cultivos y parientes silvestres de esa región. De igual forma, para poder conocer más acerca del origen de los cultivos de europeos e inferir si una o varias introducciones tuvieron lugar, se requiere incluir cultivares de diversas regiones del Viejo Continente.

Los siete eventos de domesticación que han ocurrido en cinco especies de *Phaseolus* ofrecen una gran oportunidad para estudiar la domesticación como proceso evolutivo y para tratar de encontrar paralelismos, convergencias y patrones genéticos comunes. Generar datos genómicos para las cinco especies, especialmente en las especies poco estudiadas como *P. dumosus* y *P. acutifolius*, permitiría desarrollar estudios de genómica comparada y responder preguntas como si los mismos genes fueron afectados en los diferentes eventos de domesticación o si a través de vías génicas distintas es posible llegar fenotipos similares.

Referencias

Abbo, S., Pinhasi van-Oss, R., Gopher, A., Saranga, Y., Ofner, I., & Peleg, Z. (2014). Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends Plant Sci.* 19, 351–360. doi:10.1016/j.tplants.2013.12.002.

Acampora, A., Ciaffi, M., De Pace, C., Paolacci, A. R., & Tanzarella, O. A. (2007). Pattern of variation for seed size traits and molecular markers in Italian germplasm of *Phaseolus coccineus* L. *Euphytica* 157, 69–82. doi:10.1007/s10681-007-9397-3.

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

Andueza-Noh, R. H., Martínez-Castillo, J., y Chacón-Sánchez, M. I. (2015). Domestication of small-seeded lima bean (Phaseolus lunatus L.) landraces in Mesoamerica: evidence from microsatellite markers. *Genetica* 143, 657–669. doi:10.1007/s10709-015-9863-0.

Andueza-Noh, R. H., Serrano-Serrano, M. L., Chacón Sánchez, M. I., del Pino, I. S., Camacho-Pérez, L., Coello-Coello, J., et al. (2012). Multiple domestications of the Mesoamerican gene pool of lima bean (Phaseolus lunatus L.): evidence from chloroplast DNA sequences. *Genet. Resour. Crop Evol.* 60, 1069–1086. doi:10.1007/s10722-012-9904-9.

Angioi, S. A., Desiderio, F., Rau, D., Bitocchi, E., Attene, G., y Papa, R. (2009). Development and use of chloroplast microsatellites in Phaseolus spp. and other legumes. *Plant Biol.* 11, 598–612. doi:10.1111/j.1438-8677.2008.00143.x.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3, e3376. doi:10.1371/journal.pone.0003376.

Basurto Peña, F.A. (2000) Aspecto etnobotánicos de *Phaseolus coccineus* L. y *Phaseolus polyanthus* Greeman en la Sierra Norte de Puebla, México. Tesis Maestría en Ciencias (Biología). UNAM, Facultad de Ciencias, División de Estudios de Posgrado.

Beebe, S., Rengifo, J., Gaitan, E., Duque, M. C., y Tohme, J. (2001). Diversity and Origin of Andean Landraces of Common Bean. *Crop Sci.* 41, 854. doi:10.2135/cropsci2001.413854x.

Beerli, P., y Palczewski, M. (2010). Unified Framework to Evaluate Panmixia and Migration Direction Among Multiple Sampling Locations. *Genetics* 185, 313–326. doi:10.1534/genetics.109.112532.

Bellucci, E., Bitocchi, E., Ferrarini, A., Benazzo, A., Biagetti, E., Klie, S., et al. (2014). Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean. *Plant Cell* 26, 1901–1912. doi:10.1105/tpc.114.124040.

Bitocchi, E., Bellucci, E., Giardini, A., Rau, D., Rodriguez, M., Biagetti, E., et al. (2013). Molecular analysis of the parallel domestication of the common bean (Phaseolus vulgaris) in Mesoamerica and the Andes. *New Phytol.* 197, 300–313. doi:10.1111/j.1469-8137.2012.04377.x.

Blair, M. W., Pantoja, W., & Carmenza Muñoz, L. (2012). First use of microsatellite markers in a

large collection of cultivated and wild accessions of tepary bean (*Phaseolus acutifolius* A. Gray). *Theor. Appl. Genet.* 125, 1137–1147. doi:10.1007/s00122-012-1900-0.

Boczkowska, M., Bulińska-Radomska, Z., & Nowosielski, J. (2012). AFLP analysis of genetic diversity in five accessions of Polish runner bean (*Phaseolus coccineus* L.). *Genet. Resour. Crop Evol.* 59, 473–478. doi:10.1007/s10722-012-9798-6.

Bourgeois, Y., Hazzouri, K. M., y Warren, B. (2016). Going down the rabbit hole: a review on methods characterizing selection and demography in natural populations. doi:10.1101/052761.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., y RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29, 1917–1932. doi:10.1093/molbev/mss086.

Casas, A. A., Viveros, J. L., and Caballero, J. (1994). Etnobotánica mixteca: sociedad, cultura y recursos naturales en la montaña de Guerrero. *Ciencias*. 39, 61-63.

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. *Hum. Ecol.* 24, 455–478. doi:10.1007/bf02168862.

Casas, A., Otero-Arnaiz, A., Pérez-Negrón, E., & Valiente-Banuet, A. (2007). *In situ* management and domestication of plants in Mesoamerica. *Ann. Bot.* 100, 1101–1115. doi:10.1093/aob/mcm126.

Chacón-Sánchez, M. I. (2018). “The Domestication Syndrome in *Phaseolus* Crop Plants: A Review of Two Key Domestication Traits,” en Pontarotti, P. (Ed.) *Origin and Evolution of Biodiversity*, 37–59, Springer. doi:10.1007/978-3-319-95954-2_3.

Chacón-Sánchez, M. I., Pickersgill, B., y Debouck, D. G. (2005). Domestication patterns in common bean (Phaseolus vulgaris L.) and the origin of the Mesoamerican and Andean cultivated races. *Theor. Appl. Genet.* 110, 432–444. doi:10.1007/s00122-004-1842-2.

Chacón-Sánchez, M. I., y Martínez-Castillo, J. (2017). Testing Domestication Scenarios of Lima Bean (Phaseolus lunatus L.) in Mesoamerica: Insights from Genome-Wide Genetic Markers. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.01551.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., y Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi:10.1186/s13742-015-0047-8.

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., y Beaumont, M. A. (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* 186, 983–995. doi:10.1534/genetics.110.118661.

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., et al. (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* 24, 2713–2719. doi:10.1093/bioinformatics/btn514.

de Candolle, A. (1886). Origin of Cultivated Plants. doi:10.5962/bhl.title.20259. Available at: https://books.google.com/books/about/Origin_of_Cultivated_Plants.html?hl=&id=kqcMAA

AAYAAJ.

DeGiorgio, M., Lohmueller, K. E., y Nielsen, R. (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet.* 10, e1004561. doi:10.1371/journal.pgen.1004561.

Delgado-Salinas, A. D. (1988). “Variation, Taxonomy, Domestication, and Germplasm Potentialities in *Phaseolus coccineus*,” en *Current Plant Science and Biotechnology in Agriculture*, 441–463. doi:10.1007/978-94-009-2786-5_18.

Delgado-Salinas, A., Bibler, R., and Lavin, M. (2006). Phylogeny of the Genus *Phaseolus* (Leguminosae): A Recent Diversification in an Ancient Landscape. *Syst. Bot.* 31, 779–791. doi:10.1600/036364406779695960.

Delgado-Salinas, A., Caballero, J., & Casas, A. (2004). “Crop Domestication in Mesoamerica,” en *Encyclopedia of Plant and Crop Science*, 310–313. doi:10.1081/e-epcs-120017097.

Durand, E. Y., Patterson, N., Reich, D., y Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.* 28, 2239–2252. doi:10.1093/molbev/msr048.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379. doi:10.1371/journal.pone.0019379.

Escalante, A. M., Coello, G., Eguiarte, L. E., & Pinero, D. (1994). Genetic Structure and Mating Systems in Wild and Cultivated Populations of *Phaseolus coccineus* and *P. vulgaris* (Fabaceae). *Am. J. Bot.* 81, 1096. doi:10.2307/2445471.

Excoffier, L., y 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.

Fay, J. C., y Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10880498.

Foll, M., y 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.

Freytag, G. F., & Debouck, D. G. (2002). *Taxonomy, Distribution, and Ecology of the Genus Phaseolus (Leguminosae-Papilionoideae) in North America, Mexico and Central America*. BRIT Press Available at: https://books.google.com/books/about/Taxonomy_Distribution_and_Ecology_of_the.html?hl=&id=8M2Mwi85IIIC.

Frichot, E., Schoville, S. D., Bouchard, G., y François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–1699. doi:10.1093/molbev/mst063.

Fuller, D. Q., Denham, T., Arroyo-Kalin, M., Lucas, L., Stevens, C. J., Qin, L., et al. (2014). Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. *Proc. Natl. Acad. Sci. U. S. A.* 111, 6147–6152. doi:10.1073/pnas.1308937110.

Garvin, D. F., y Weeden, N. F. (1994). Isozyme Evidence Supporting a Single Geographic Origin for Domesticated Tepary Bean. *Crop Sci.* 34, 1390. doi:10.2135/cropsci1994.0011183X003400050045x.

Gaut, B.S. et al., 2018. Demography and its effects on genomic variation in crop domestication. Nature plants, 4(8), pp.512–520. doi: 10.1038/s41477-018-0210-1.

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.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi:10.1126/science.1188021.

Guerra-García, A., Suárez-Atilano, M., Mastretta-Yanes, A., Delgado-Salinas, A., & 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.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., y Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi:10.1093/sysbio/syq010.

Gujaria-Verma, N., Ramsay, L., Sharpe, A. G., Sanderson, L.-A., Debouck, D. G., Tar'an, B., et al. (2016). Gene-based SNP discovery in tepary bean (Phaseolus acutifolius) and common bean (P. vulgaris) for diversity analysis and comparative mapping. *BMC Genomics* 17, 239. doi:10.1186/s12864-016-2499-3.

Günther, T., y Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–220. doi:10.1534/genetics.113.152462.

Harlan, J. R. (1971). Agricultural origins: centers and noncenters. *Science* 174, 468–474. doi:10.1126/science.174.4008.468.

Hey, J., y Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. U. S. A.* 104, 2785–2790. doi:10.1073/pnas.0611164104.

Huson, D. H., y Bryant, D. (2005). Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* 23, 254–267. doi:10.1093/molbev/msj030.

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. doi:10.1093/biosci/bix114.

Kaplan, L., & Lynch, T. F. (1999). Phaseolus (Fabaceae) in Archaeology: AMS. *Econ. Bot.* 53, 261–272. doi:10.1007/bf02866636.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* 13, 235–248. doi:10.1016/0304-4149(82)90011-4.

Kwak, M., Kami, J. A., y Gepts, P. (2009). The Putative Mesoamerican Domestication Center of Is Located in the Lerma–Santiago Basin of Mexico. *Crop Sci.* 49, 554. doi:10.2135/cropsci2008.07.0421.

Ladizinsky, G. (1985). Founder effect in crop-plant evolution. *Econ. Bot.* 39, 191–199.

doi:10.1007/bf02907844.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42, 1053–1059. doi:10.1038/ng.715.

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.

Lewontin, R. C., y Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74, 175–195. Available at: https://www.ncbi.nlm.nih.gov/pubmed/4711903.

Luu, K., Bazin, E., y 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.

Mamidi, S., Rossi, M., Annam, D., Moghaddam, S., Lee, R., Papa, R., et al. (2011). Investigation of the domestication of common bean (Phaseolus vulgaris) using multilocus sequence data. *Funct. Plant Biol.* 38, 953. doi:10.1071/FP11124.

Martínez-Castillo, J., Camacho-Pérez, L., Villanueva-Viramontes, S., Andueza-Noh, R. H., y Chacón-Sánchez, M. I. (2014). Genetic structure within the Mesoamerican gene pool of wild Phaseolus lunatus (Fabaceae) from Mexico as revealed by microsatellite markers: Implications for conservation and the domestication of the species. *Am. J. Bot.* 101, 851–864. doi:10.3732/ajb.1300412.

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez G, J., Buckler, E., y Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6080–6084. doi:10.1073/pnas.052125199.

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

Mina-Vargas, A. M., McKeown, P. C., Flanagan, N. S., Debouck, D. G., Kilian, A., Hodkinson, T. R., et al. (2016). Origin of year-long bean (Phaseolus dumosus Macfady, Fabaceae) from reticulated hybridization events between multiple Phaseolus species. *Ann. Bot.* doi:10.1093/aob/mcw138.

Motta-Aldana, J. R., Serrano-Serrano, M. L., Hernández-Torres, J., Castillo-Villamizar, G., Debouck, D. G., y Chacón-Sánchez., M. I. (2010). Multiple Origins of Lima Bean Landraces in the Americas: Evidence from Chloroplast and Nuclear DNA Polymorphisms. *Crop Sci.* 50, 1773. doi:10.2135/cropsci2009.12.0706.

Nei, M., y Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273.

Nielsen, R. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575. doi:10.1101/gr.4252305.

Nowosielski, J., Podyma, W., & Nowosielska, D. (2002). Molecular research on the genetic diversity of Polish varieties and landraces of *Phaseolus coccineus* L. and *Phaseolus vulgaris* L. using the RAPD and AFLP methods. *Cell. Mol. Biol. Lett.* 7, 753–762. Available at: https://www.ncbi.nlm.nih.gov/pubmed/12378235.

Papa, R., y Gepts, P. (2003). Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean *(Phaseolus vulgaris* L.) from Mesoamerica. *Theor. Appl. Genet.* 106, 239–250. doi:10.1007/s00122-002-1085-z.

Papa, R., Acosta, J., Delgado-Salinas, A., & Gepts, P. (2005). A genome-wide analysis of differentiation between wild and domesticated *Phaseolus vulgaris* from Mesoamerica. *Theor. Appl. Genet.* 111, 1147–1158. doi:10.1007/s00122-005-0045-9.

Papa, R., Bellucci, E., Rossi, M., Leonardi, S., Rau, D., Gepts, P., et al. (2007). Tagging the signatures of domestication in common bean (Phaseolus vulgaris) by means of pooled DNA samples. *Ann. Bot.* 100, 1039–1051. doi:10.1093/aob/mcm151.

Pickersgill, B. (2007). Domestication of plants in the Americas: insights from Mendelian and molecular genetics. *Ann. Bot.* 100, 925–940. doi:10.1093/aob/mcm193.

Pickrell, J. K., y 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., Ranere, A. J., Holst, I., Iriarte, J., y Dickau, R. (2009). Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proceedings of the National Academy of Sciences* 106, 5019–5024. doi:10.1073/pnas.0812525106.

Pritchard, J. K., Stephens, M., y Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10835412.

Raj, A., Stephens, M., y 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.

Rasmussen, J. (2014). *Defending the Correspondence Theory of Truth*. doi:10.1017/cbo9781107415102.

Rendón-Anaya, M., Herrera-Estrella, A., Gepts, P., y Delgado-Salinas, A. (2017a). A new species of Phaseolus (Leguminosae, Papilionoideae) sister to Phaseolus vulgaris, the common bean. *Phytotaxa* 313, 259. doi:10.11646/phytotaxa.313.3.3.

Rendón-Anaya, M., Montero-Vargas, J. M., Saburido-Álvarez, S., Vlasova, A., Capella-Gutierrez, S., Ordaz-Ortiz, J. J., et al. (2017b). Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biol.* 18, 60. doi:10.1186/s13059-017-1190-6.

Rodiño, A. P., Paula Rodiño, A., Lema, M., Pérez-Barbeito, M., Santalla, M., & De Ron, A. M. (2006). Assessment of runner bean (*Phaseolus coccineus* L.) germplasm for tolerance to low temperature during early seedling growth. *Euphytica* 155, 63–70. doi:10.1007/s10681-006-9301-6.

Rodriguez, M., Rau, D., Angioi, S. A., Bellucci, E., Bitocchi, E., Nanni, L., et al. (2013). European *Phaseolus coccineus* L. landraces: population structure and adaptation, as revealed by cpSSRs and phenotypic analyses. *PLoS One* 8, e57337. doi:10.1371/journal.pone.0057337.

Rodriguez, M., Rau, D., Bitocchi, E., Bellucci, E., Biagetti, E., Carboni, A., et al. (2016). Landscape genetics, adaptive diversity and population structure in Phaseolus vulgaris.

New Phytol. 209, 1781–1794. doi:10.1111/nph.13713.

Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 104 Suppl 1, 8641–8648. doi:10.1073/pnas.0700643104.

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.

Schinkel, C., y Gepts, P. (1988). Phaseolin Diversity in the Tepary Bean, Phaseolus acutifolius A. Gray. *Plant Breed.* 101, 292–301. doi:10.1111/j.1439-0523.1988.tb00301.x.

Schmit, V., y Debouck, D. G. (1991). Observations on the origin of phaseolus polyanthus Greenman. *Econ. Bot.* 45, 345–364. doi:10.1007/BF02887077.

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.

Serrano-Serrano, M. L., Andueza-Noh, R. H., Martínez-Castillo, J., Debouck, D. G., y Chacón S, M. I. (2012). Evolution and Domestication of Lima Bean in Mexico: Evidence from Ribosomal DNA. *Crop Sci.* 52, 1698. doi:10.2135/cropsci2011.12.0642.

Serrano-Serrano, M. L., Hernández-Torres, J., Castillo-Villamizar, G., Debouck, D. G., y Sánchez, M. I. C. (2010). Gene pools in wild Lima bean (Phaseolus lunatus L.) from the Americas: evidences for an Andean origin and past migrations. *Mol. Phylogenet. Evol.* 54, 76–87. doi:10.1016/j.ympev.2009.08.028.

Sicard, D., Nanni, L., Porfiri, O., Bulfon, D., & Papa, R. (2005). Genetic diversity of *Phaseolus vulgaris* L. and *P. coccineus* L. landraces in central Italy. *Plant Breed.* 124, 464–472. doi:10.1111/j.1439-0523.2005.01137.x.

Smith, B. D. (1997). The Initial Domestication ofCucurbita pepoin the Americas 10,000 Years Ago. *Science* 276, 932–934. doi:10.1126/science.276.5314.932.

Spataro, G., Tiranti, B., Arcaleni, P., Bellucci, E., Attene, G., Papa, R., et al. (2011). Genetic diversity and structure of a worldwide collection of *Phaseolus coccineus* L. *Theor. Appl. Genet.* 122, 1281–1291. doi:10.1007/s00122-011-1530-y.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Takezaki, N., Nei, M., y Tamura, K. (2010). POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. *Mol. Biol. Evol.* 27, 747–752. doi:10.1093/molbev/msp312.

Toledo, V. M. (2001). “Indigenous Peoples and Biodiversity,” en *Encyclopedia of Biodiversity*, 451–463. doi:10.1016/b0-12-226865-2/00157-7.

Vavilov, N. (1951) The Origin, Variation, Immunity & Breeding of Cultivated Plants. *Chronica Botanica* 13, 1–366.

Villaseñor, J. L. (2003). Diversidad y distribución de las Magnoliophyta de México. *Interciencia* 28, 160–167. Available at:

http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S0378-18442003000300008.

Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., et al. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. Theor. Appl. Genet. 129, 753–765.

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., et al. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30, 105–111. doi:10.1038/nbt.2050.

Zeven, A. C., Mohamed, H. H., Waninge, J., & Veurink, H. (1993). Phenotypic variation within a Hungarian landrace of runner bean (*Phaseolus coccineus* L.). *Euphytica* 68, 155–166. doi:10.1007/bf00024164.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., y Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi:10.1093/bioinformatics/bts606.

Material Suplementario

Domestication Genomics of the open-pollinated Scarlet Runner Bean (*Phaseolus coccineus* L.)

Azalea Guerra-García, Marco Suárez-Atilano, Alicia Mastretta-Yanes, Alfonso Delgado-Salinas, Daniel Piñero

*** Correspondence:** Azalea Guerra-García: azalea.guerra.g@gmail.com

Table S1. Sampling information

Species	Status	State and location	Latitude	Longitude	Altitude
P. coccineus subsp. *coccineus*	Cultivated	Oaxaca, Cuilapam	16.99	-96.78	1577
P. coccineus subsp. *coccineus*	Cultivated	Durango	NA	NA	NA
P. coccineus subsp. *coccineus*	Cultivated	Veracruz, Frijol Colorado	19.59	-97.35	2419
P. coccineus subsp. *coccineus*	Cultivated	Chiapas, González León	16.51	-92.06	1579
P. coccineus subsp. *coccineus*	Cultivated	España	NA	NA	NA
P. coccineus subsp. *coccineus*	Cultivated	Chiapas, Nahá	16.94	-91.59	917
P. coccineus subsp. *coccineus*	Cultivated	Veracruz, Orilla del Monte	19.66	-97.29	2402
P. coccineus subsp. *coccineus*	Cultivated	Chaipas, Oxchuc	16.80	-92.32	2001
P. coccineus subsp. *coccineus*	Cultivated	Puebla, Tlalancaneca	19.36	-98.51	2372
P. coccineus subsp. *coccineus*	Cultivated	Durango, Regocijo	23.68	-105.12	2566
P. coccineus subsp. *coccineus*	Cultivated	Durango, Villa Unión	23.97	-104.04	1901
P. coccineus subsp. *coccineus*	Cultivated	Puebla	NA	NA	NA
P. coccineus subsp. *coccineus*	Feral	Veracruz, Altotonga	19.75	-97.25	1959
P. coccineus subsp. *coccineus*	Feral	Oaxaca, Huautla	18.10	-96.83	1746
P. coccineus subsp. *coccineus*	Feral	Chiapas, Zicanantán	16.75	-92.73	2408
P. coccineus subsp. *coccineus*	Breeding line	Blanco Tlaxcala	NA	NA	NA
P. coccineus subsp. *coccineus*	Wild	Jalisco, Ciudad Gúzman	19.58	-103.53	2100
P. coccineus subsp.	Wild	Oaxaca,	17.55	-96.53	2876

coccineus		Comaltepec			
P. coccineus subsp. *coccineus*	Wild	Durango, Espinazo del Diablo	23.64	-105.82	1422
P. coccineus subsp. *coccineus*	Wild	Durango, Regocijo	23.68	-105.12	2566
P. coccineus subsp. *coccineus*	Wild	Ciudad de México, REPSA	19.32	-99.20	2328
P. coccineus subsp. *coccineus*	Wild	Querétaro, San Joaquín	20.93	-99.56	2381
P. coccineus subsp. *coccineus*	Wild	Chiapas, San Cristóbal	16.70	-92.60	2229
P. coccineus subsp. *coccineus*	Wild	Morelos, Tepoztlán	19.00	-99.13	1953
P. coccineus subsp. *coccineus*	Wild	Ciudad de México, Tlalpan	19.29	-99.19	2334
P. coccineus subsp. *striatus*	Wild	Morelos, Tres Marías	19.10	-99.21	3024
P. dumosus	Cultivated	Veracruz, Altotonga	19.75	-97.25	1959
P. dumosus	Cultivated	Chiapas, Chiquinivalvo	16.71	-92.86	1467
P. dumosus	Cultivated	Puebla, Cuetzalan	19.99	-97.54	1684
P. dumosus	Cultivated	Oaxaca, Huautla	18.10	-96.83	1746
P. dumosus	Cultivated	Chiapas, Motozintla	15.43	-92.33	2671
P. dumosus	Cultivated	Chiapas, Talquián	15.09	-92.08	1728
P. vulgaris	Cultivated	España	NA	NA	NA
P. vulgaris	Wild	Jalisco	NA	NA	NA
P. vulgaris	Wild	Ciudad de México, REPSA	19.32	-99.20	2328
P. vulgaris	Wild	Morelos, Yautepec	18.95	-99.08	1386

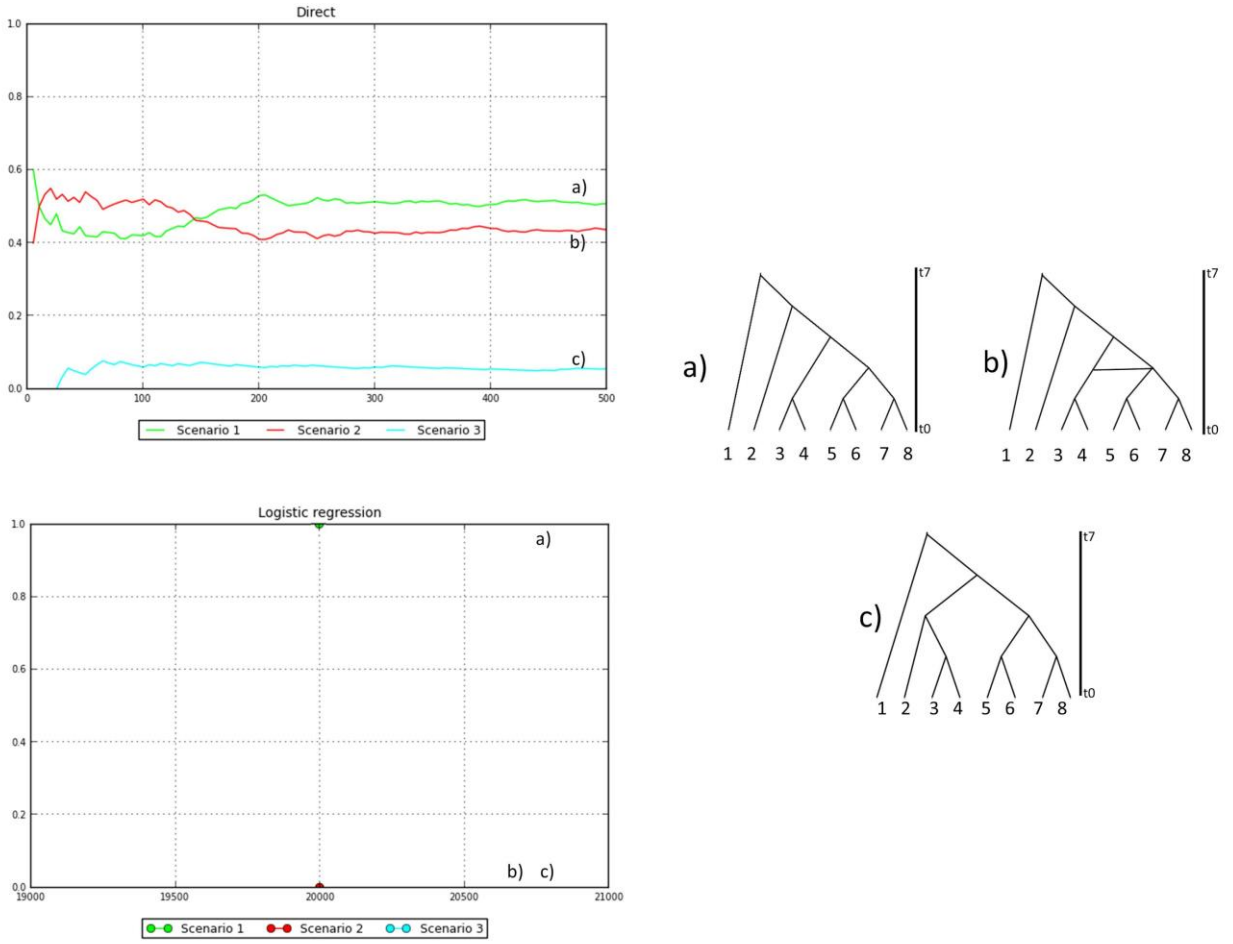


Figure S1. Posterior probabilities (Direct and Logistic) of the three main domestication scenarios tested with DIYABC, among eight wild and cultivated populations of Phaseolus coccineus in Mexico. In this preliminary evaluation: a) monophyletic origin of Wild populations b) and c) paraphyletic relationships among Wild populations.

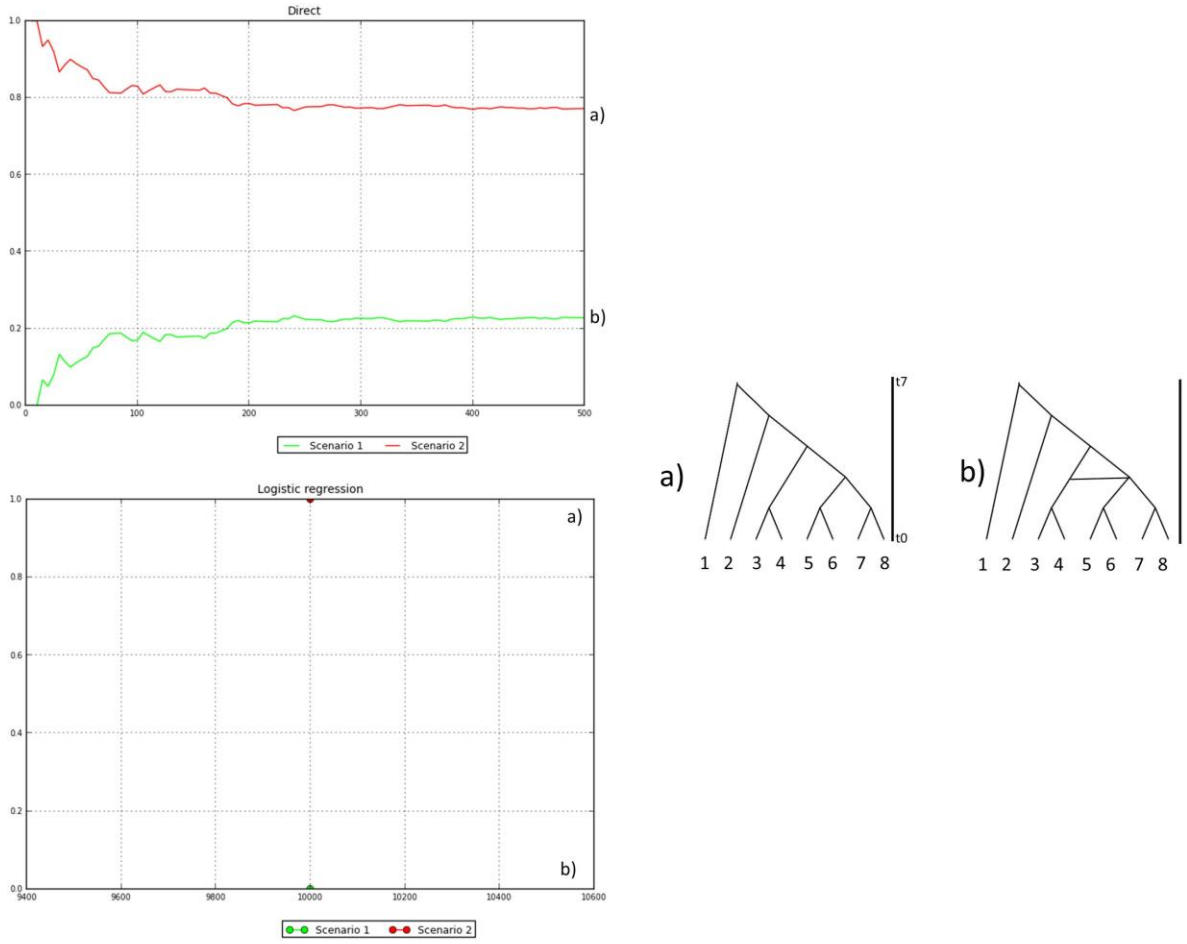
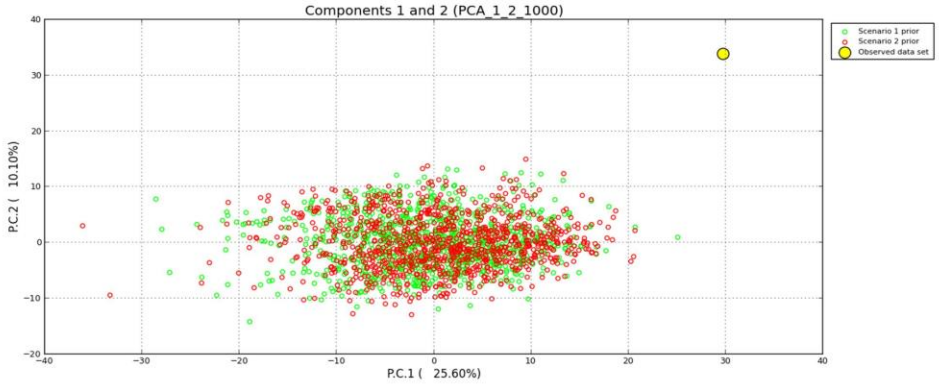


Figure S2. Posterior probabilities (Direct and Logistic) of the two previously selected domestication scenarios tested with DIYABC among eight wild and cultivated populations of Phaseolus coccineus in Mexico. Both scenarios reflect a paraphyletic relationship among Wild populations and a) with one contribution in time of Wild Populations into Cultivated ancestral genetic pool and b) with several contribution events in time of Wild Populations into Cultivated ancestral genetic pool

1)



2)

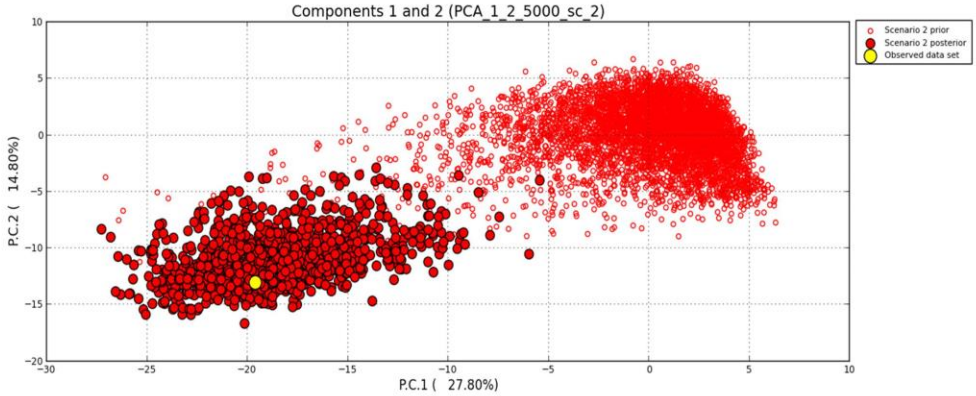


Figure S3. Principal component analyses of the a) Pre-evaluate scenario prior combinations of two selected scenarios in which paraphyletic relationship among Wild populations are represented and b) Model checking of the selected scenario indicating a single contribution in time of Wild Populations into Cultivated ancestral genetic pool

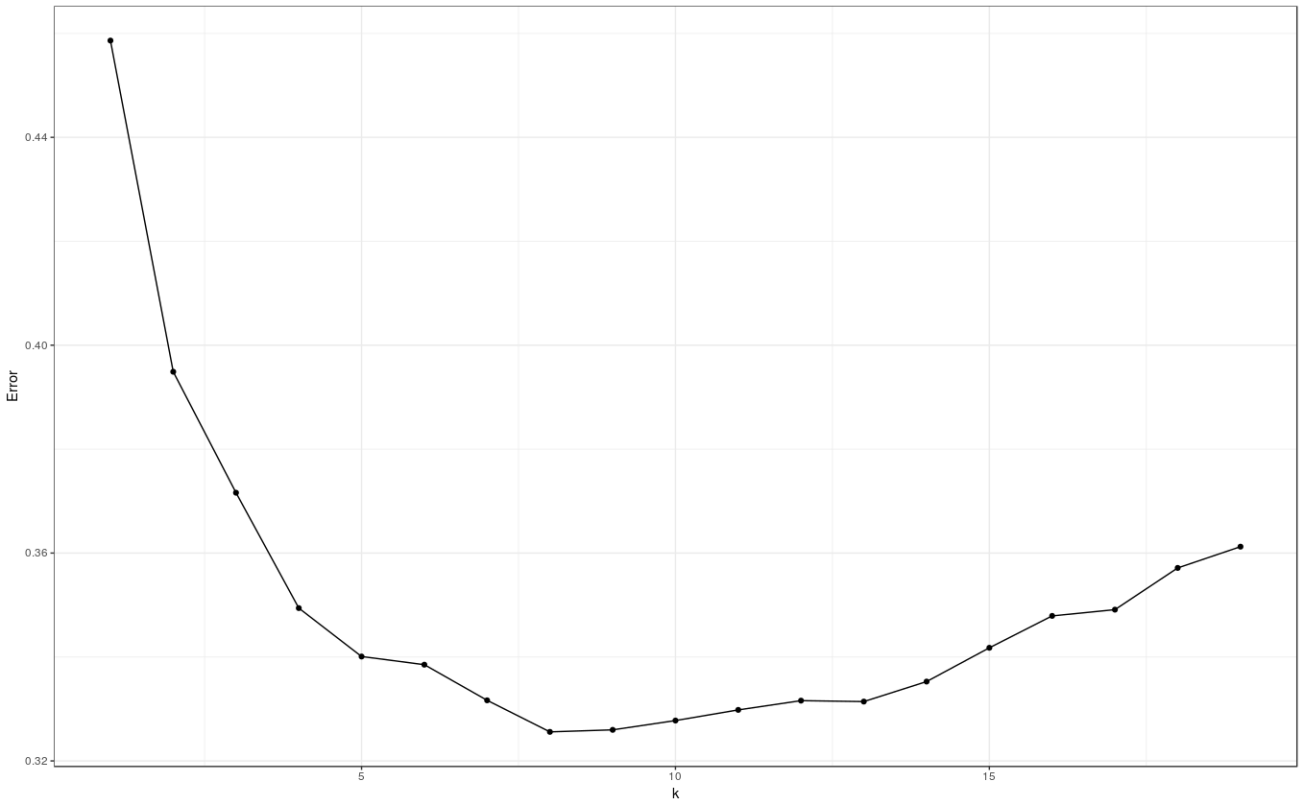


Figure S4. Cross-validation errors among *K* values using Admixture software for the *P. coccineus* data set.

Table S2. Heterozygosity and inbreeding coefficient statistics of *P. coccineus* samples grouped by population.

Genetic cluster	Population	H_O	H_E	F_{IS} (95% CI)
Cult-OV	Cuilapam	0.171	0.148	-0.152 (-0.164--0.142)
Cult-SMOCC	Blanco Tlaxcala	0.194	0.167	-0.159 (-0.168--0.150)
	Durango	0.218	0.173	-0.262 (-0.270--0.253)
	Regocijo	0.205	0.160	-0.280 (-0.293--0.273)
	Villa Unión	0.179	0.167	-0.072 (-0.083--0.063)
Cult-SUR-CH	Altotonga	0.215	0.177	-0.216 (-0.225--0.208)
	González de León	0.200	0.176	-0.135 (-0.144--0.126)
	Huautla	0.219	0.199	-0.103 (-0.112--0.095)
	Nahá	0.196	0.167	-0.171 (-0.182--0.162)
	Oxchuc	0.200	0.178	-0.124 (-0.133--0.114)
	Zicanantán	0.196	0.191	-0.029 (-0.091--0.064)
Cult-TMVB	España	0.166	0.134	-0.237 (-0.253--0.230)
	Frijol Colorado	0.193	0.191	-0.009 (-0.018-0.000)
	Orilla del Monte	0.221	0.195	-0.133 (-0.142--0.125)
	Tlalancaneca	0.199	0.190	-0.044 (-0.053--0.036)
	Puebla	0.206	0.191	-0.079 (-0.089--0.070)
Wild-SMOCC	Espinazo del Diablo	0.150	0.118	-0.274 (-0.293--0.268)
	Regocijo-S	0.165	0.141	-0.173 (-0.189--0.166)
Wild-striatus	Tres Marías	0.203	0.160	-0.267 (-0.277--0.261)
Wild-SUR-CH	Comaltepec	0.189	0.157	-0.200 (-0.231--0.207)
	San Cristobal	0.182	0.164	-0.105 (-0.125--0.102)
Wild-TMVB	Ciudad Guzman	0.169	0.176	-0.072 (-0.088--0.056)
	REPSA	0.237	0.202	-0.175 (-0.183--0.168)
	San Joaquín	0.195	0.160	-0.221 (-0.235--0.216)
	Tepoztlán	0.229	0.188	-0.217 (-0.225--0.210)
	Tlalpan	0.247	0.208	-0.184 (-0.192--0.177)

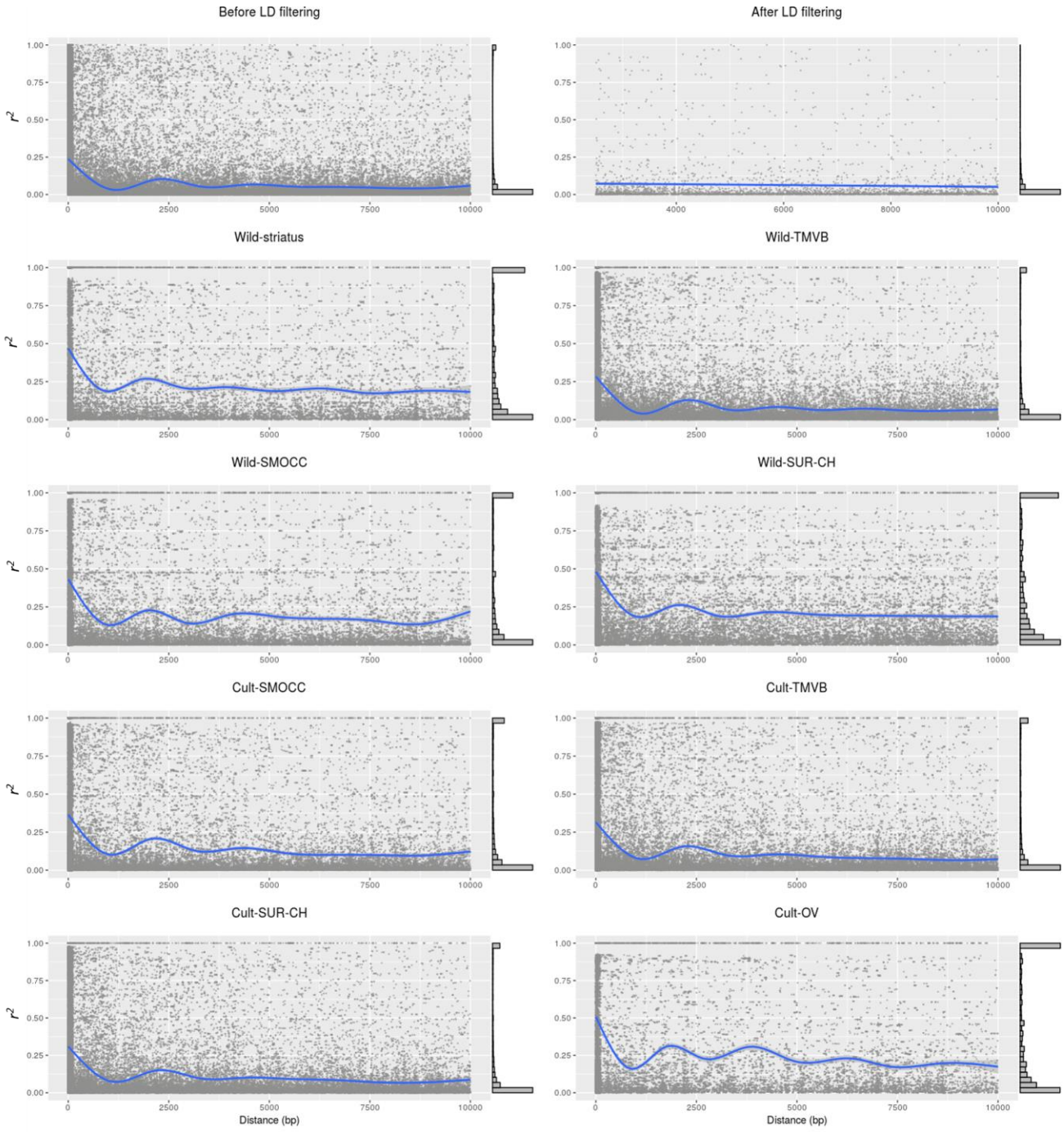


Figure S5. Pair r^2 value among SNPs located in the same chromosome separated by a maximum distance of 10 000 bp. At the right side of each plot there is a histogram showing the frequency of r^2 values.

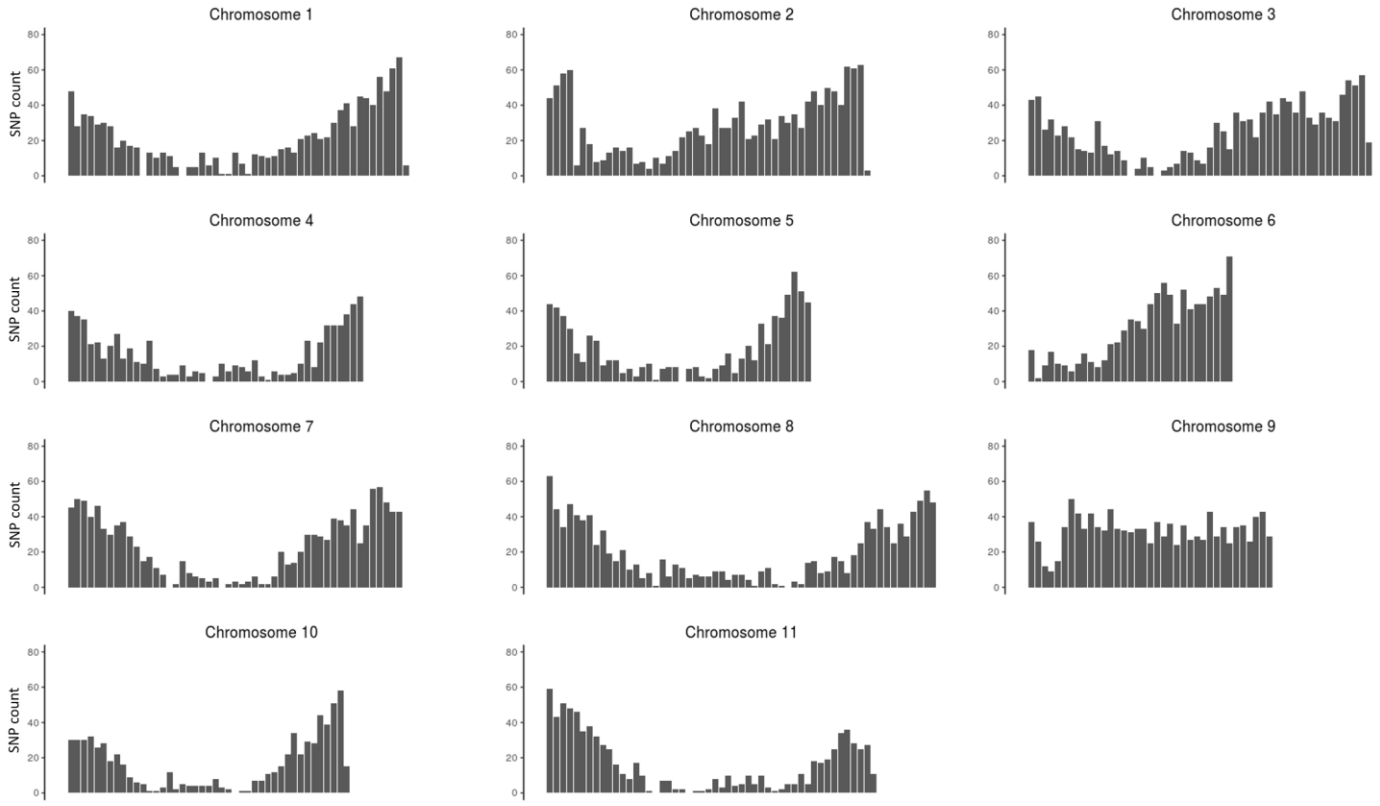


Figure S6. Distribution of 11 693 SNPs detected in *P. coccineus* (after LD filtering) present in each chromosome. Each bar represents a region of one Mb. The differences in plots sizes show the discrepancies in chromosomes lengths.

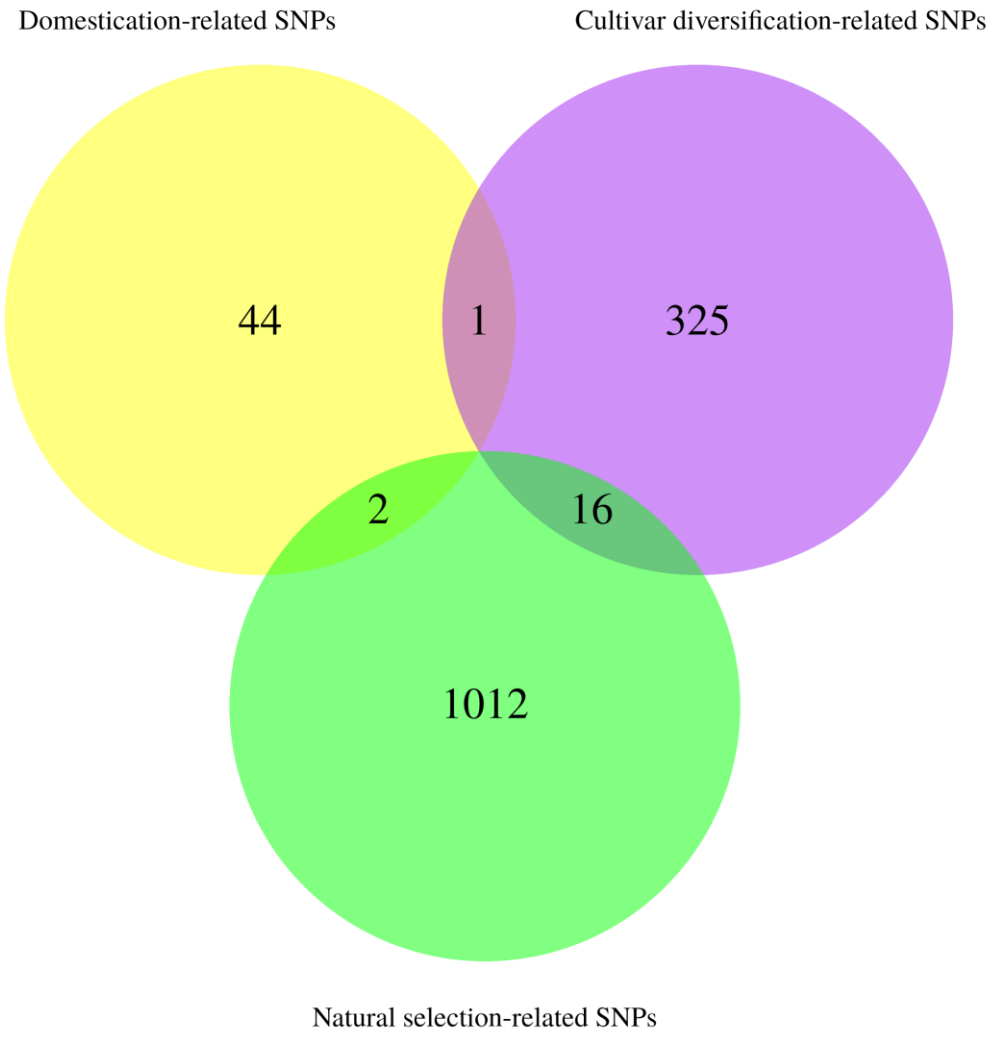


Figure S7. Venn diagram showing share candidate SNPs detected in the three PCAdapt analysis to detect natural and artificial selective pressures.

Table S3. Candidate loci identified by the pcadapt and BayeScan methods.

Chr	SNP ID	Position	*P. vulgaris* transcript name	Annotation	High expression tissue	Protein homologs *in G. max*
Domestication-related candidate						
	1S1_15176129	15176129				
	1S1_48602792	48602792	Phvul.001G232200	Phosphomethylpyrimidine kinase	Flower, buds	Glyma.11G218700
	2S1_56416681	4211050				
	2S1_79642304	27436673				
	2S1_81673246	29467615	Phvul.002G145600	TIR domain	Green and mature pods	Glyma.12G135600
	2S1_85681762	33476131				
	2S1_86298230	34092599				
	3S1_128138350	26891681				
	3S1_145377297	44130628				
	6S1_248268696	7958113				
	6S1_250412008	10101425				
	6S1_252913073	12602490				
	6S1_259648113	19337530				
	6S1_262918075	22607492				
	7S1_310062177	37774238	Phvul.007G256000	DnaJ domain	Flowers, young pods	Glyma.02G179900
	7S1_313970760	41682821				
	8S1_329571814	5525253				
	8S1_350061425	26014864				
	8S1_350157505	26110944				
	8S1_375823077	51776516				
	9S1_406871596	23162403	Phvul.009G156400	Dehydrogenase E1 component	Flower buds, flowers	Glyma.04G212100
	11S1_467512979	3058827				
	11S1_470314329	5860177				
	11S1_472148556	7694404				
Natural selection-related candidate windows						
	3S1_103319613	2072944				
	3S1_108854438	7607769				
	3S1_126647289	25400620				
	3S1_138412433	37165764				
	3S1_141480201	40233532				
	3S1_143433256	42186587	Phvul.003G197500	Calmodulin binding protein-like	Root, stem	Glyma.17G092700
Cultivar diversification-related candidate loci						
	1S1_45827637	45827637				
	3S1_146749207	45502538				
	4S1_178058430	24527352				
	4S1_178200752	24669674				
	4S1_178417950	24886872				
	4S1_181150184	27619106				
	4S1_183310744	29779666				

5S1_219843009	20351812
5S1_220112744	20621547
5S1_225806704	26315507
6S1_271294756	30984173
7S1_301822240	29534301
9S1_388863253	5154060