Universidad Politécnica de Valencia

Identificación de la variabilidad alélica en una colección de especies del género *Cucurbita*

Ana García Pérez

Tutores:
Joaquín Cañizares Sales
Javier Montero Pau

Instituto de Conservación y Mejora de la Agrodiversidad Valenciana

Máster en Mejora Genética Vegetal

Curso académico 2017/2018

Instituto de Conservación y Mejora de la Agrodiversidad Valenciana

Identificación de la variabilidad alélica en una colección de especies del genero *Cucurbita.*

Las especies del género *Cucurbita* (calabazas y calabacines) se encuentran entre los 10 principales cultivos de hortalizas a nivel mundial. La gran diversidad morfológica que presenta este género, lo hace interesante para el desarrollo de nuevas variedades. Sin embargo, a pesar del potencial de mejora y económico que presenta este género, los programas desarrollados para estos cultivos han sido menos numerosos en comparación con los llevados a cabo para otras hortalizas. Las tecnologías NGS, junto con la bioinformática, han resultado dos potentes herramientas para el estudio de la genética vegetal y el desarrollo de los programas de mejora. Los avances en las tecnologías NGS han ayudado al descubrimiento de millones de polimorfismos, que son un recurso invaluable para la mejora asistida por marcadores. Así mismo, la identificación y el seguimiento de la variación genética son ahora tan eficaces y económicos que se pueden genotipar miles de SNPs dentro de grandes poblaciones. El análisis de la diversidad genética existente en las poblaciones vegetales, incluyendo tanto especies domesticadas como silvestres, es un paso esencial para el descubrimiento de nuevos genes de interés con el fin de introducirlos en futuros planes de mejora.

En el presente trabajo se ha descrito la diversidad existente dentro de una población de 96 individuos de diferentes especies del género *Cucurbita*. A partir de la secuenciación de RNA de hoja joven se obtuvo un total de 702.735 SNVs de alta calidad los cuales se anotaron utilizando *SnpEff*, un programa de predicción de efectos capaz de predecir el impacto que puede ocasionar cada SNPs. La distribución de los SNVs a lo largo de los cromosomas no fue homogénea, pudiendo distinguir regiones ricas y pobres en polimorfismos. El porcentaje de heterocigosidad observado fue muy bajo, cerca del 1%. Se identificaron polimorfismos en unos 24.674 genes (78% de los genes), siendo 32 el promedio de variantes encontradas por gen. Unos 585.824 SNVs se encontraron dentro de la secuencia codificante. Más de la mitad de estos SNVs ocasionaron cambios de tipo “silencioso” y en menor medida, cambios de tipo “con sentido”. Del total de polimorfismos identificados menos del 0,5% de SNVs fueron responsables de ocasionar un impacto grave y cerca del 13% fueron responsables de causar impactos moderados. Solo cerca de 2.300 SNVs (0,4%) fueron responsables de la aparición de codones parada prematuros. A modo de ejemplo de su utilidad, esta colección de SNVs se ha utilizado para la identificación de posibles genes candidatos en QTLs relacionados con morfología y color del fruto (*IFSh_3, MaRCo_4, MLRCo_4,*

MbRCo_19, IbRCo_4), floración (*DFeF_12, DFeF_9*) y características de hoja (*Sl_12, Li_10*). Todos los SNVs y genotipos identificados han sido recogidos en un fichero VCF, el cual es de libre acceso y se encuentra disponible en RiuNet (Repositorio Institucional de la Universitat Politècnica de València) bajo el nombre *Colección_SNPs_anotados_cucurbitas.vcf*.

Palabras clave:

- SNPs: *single nucleotide polymorphism*/polimorfismos de un solo nucleótido
- *Cucurbita sp*.
- RNA
- Secuenciación
- NGS: *next generation sequencing*/secuenciación de nueva generación
- Diversidad genética
- Bioinformática
- Predicción de efectos
- Marcador molecular

Índice

Índice de tablas

Índice de figuras

1. Introducción

1.1. El género *Cucurbita*

1.1.1. Descripción general del género

El género *Cucurbita* L. pertenece a la familia de las cucurbitáceas (*Cucurbita*ceae), es originario de América y engloba un total de 13 especies (8 silvestres y 5 domesticadas) (Paris, 2016). Es un género caracterizado por sus hojas grandes y palmeadas, por sus flores amarillo-anaranjadas productoras de néctar y por sus frutos grandes, duros, principalmente esféricos e indehiscentes. En el género *Cucurbita* existen dos grupos ecológicos: las especies perennes xerofíticas con raíces de almacenamiento (*C. cordata* S. Watson*, C. foetidissima* H.B.K. y *C. pedatifolia* L. H. Bailey), y las especies mesofíticas anuales (*C. argyrosperma* C. Huber*, C. ecuadorensis* Cutler and Whitaker*, C. lundelliana, C. maxima* Duchesne*, C. moschata* Duchesne*, C. okeechobeensis* L. H. Bailey y *C. pepo* L.), entre las que puede haber perennes de vida corta que carecen de raíces de almacenamiento (Nee, 1990; Whitaker & Bemis, 1975). Los estudios filogenéticos realizados en *Cucurbita* indican que las especies xerófitas perennes fueron la base del género mientras que las anuales mesófitas y especies perennes de vida corta derivaron de ellas y formando un taxón monofilético (Jobst, King, & Hemleben, 1998; Kates, Soltis, & Soltis, 2017; Kistler et al., 2015; Montero-Pau, Blanca, Bombarely, et al., 2017; Sanjur, Piperno, Andres, & Wessel-Beaver, 2002; Wilson, Doebley, & Duvall, 1992). El clado de especies mesofíticas incluye las cinco especies domesticadas de *Cucurbita* y sus parientes silvestres más cercanos (Fig. 1). *C. pepo* y las especies silvestres (*C. pepo* subsp. *ovifera* var. *orkazana, C. pepo* subsp. *ovifera* var. *texana, C. pepo* subsp. *ovifera* var. *fraterna*), forman un clado con *C. okeechobeensis* y *C. lundelliana*. En conjunto, estas tres especies, son un clado hermano del clado formado por *C. moschata* y *C. argyrosperma* y la especie silvestre (*C. argyrosperma* subsp. *sororia*). El par de especies *C. maxima,* junto a su pariente silvestre cercano (*C. maxima* subsp. *andreana*), y *C. ecuadorensis* son hermanas del resto de especies mesofíticas (Kates et al., 2017). En cuanto a *Cucurbita ficifolia* Bouché*,* su posición en el árbol es controvertida (Montero-Pau, Blanca, Bombarely, et al., 2017). *C. ficifolia* es una especie mesofítica, pero comparte algunas características

morfológicas con las especies xerofíticas por lo que son necesarios más datos para establecer su relación con el resto de especies.

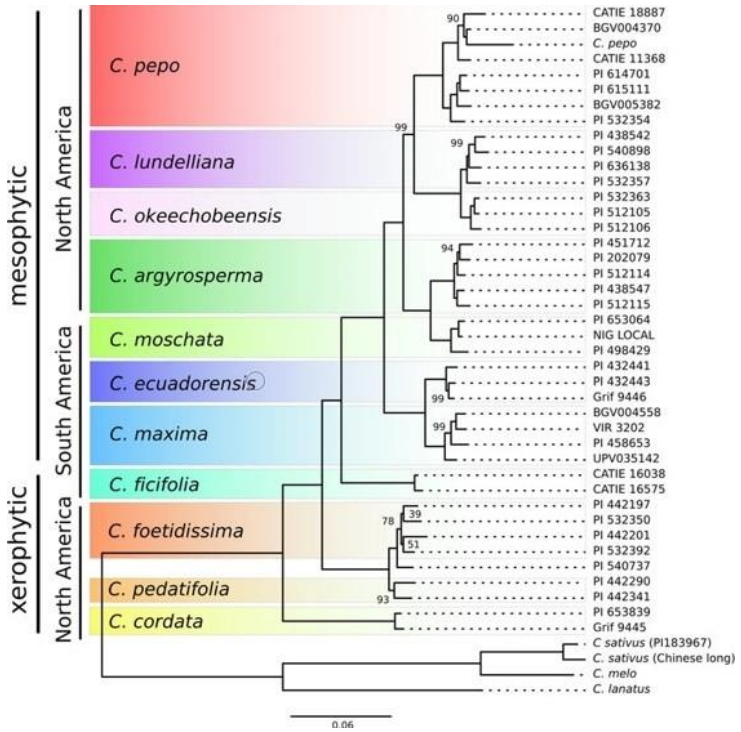

Figura 1.- Árbol filogenético del género *Cucurbita.* Árbol izquierdo, basado en el método concatenado; árbol derecho, estimación conjunta de árboles genéticos y especies. Fuente: árbol modificado a partir de Montero-Pau, Blanca, Bombarely, et al., 2017.

Las especies pertenecientes a este género son diploides y todas tienen 20 pares de cromosomas (2n = 40). Los primeros estudios citogenéticos e isoenzimáticos (Weeden, 1984), así como un análisis de sintenia (Singh, 1990) sugirieron una posible duplicación del genoma completo en este género (Esteras et al., 2012; Weeden, 1984). Estudios más recientes, basados en el análisis completo de los genomas de *C. maxima, C. moschata* (Sun et al., 2017) y *C. pepo* (Montero-Pau, Blanca, Bombarely, et al., 2017) han demostrado la existencia de un evento de duplicación completa y se ha sugerido que pueda haber sido el origen del género (Montero-Pau, Blanca, Bombarely, et al., 2017).

Las especies cultivadas de este género son apreciadas principalmente por sus frutos (botánicamente un pepo) que son una fuente significativa de carbohidratos y vitaminas (Whitaker & Davis, 1962). Bajo la denominación de “calabaza”, “calabacín”, “zapallo”, “zapatillo”, “ayote”, o “pumpkin” o “squash” en inglés se incluyen las principales especies cultivadas del género *Cucurbita*: *C*. *pepo*, *C. maxima*, *C*. *moschata, C. argyrosperma* y *C. ficifolia*. En general, el término "calabaza" como tal, suele aplicarse

para referirse al grupo conocido botánicamente como “calabazas de invierno” (Fig. 2a). Estos se caracterizan principalmente por ser cultivares con frutos redondos, que se consumen maduros y con una larga vida de poscosecha, que en algunos casos, sobrepasa los seis meses. Dentro de este grupo se encontrarían las especies *C. maxima, C. moschata, C. argyrosperma* (antes conocida como *C. mixta*) y algunos cultivares de *C. pepo* (Ferriol & Picó, 2008; Maroto Borrego, 2002). Dentro de “calabazas de verano” (Fig. 2b), se encuentran los cultivares de *C. pepo* cuyos frutos se recolectan en estado joven, sin haber alcanzado su tamaño definitivo y por tanto se consumen en estado inmaduro. La mayor parte de las variedades que se engloban dentro de “calabazas de verano” pertenecen al grupo hortícola de los “calabacines”, uno de los más importantes a nivel económico. Dentro del grupo “calabazas de verano”, cabe mencionar, aunque con menos importancia, a algunas variedades de *C. moschata* y *C. maxima* en las que sus frutos se recolectan inmaduros para su consumo (Paris, 2008). Su aplicación principal consiste en el consumo de sus frutos, fritos, cocidos, en sopas, en forma de pasteles, mermeladas, etc. También son utilizados como alimento para el ganado. En algunas zonas españolas sus semillas se consumen directamente y en determinados países asiáticos, de sus semillas, se obtiene un aceite comestible. Cabe señalar también que algunas calabazas, como *C. ficifolia* son utilizadas para elaborar bebidas alcohólicas, otras como plantas ornamentales (*C. pepo* subsp*. ovifera*), y en algunos países asiáticos y africanos, sus hojas y flores son consumidas como hierbas aromáticas (Maroto Borrego, 2002).

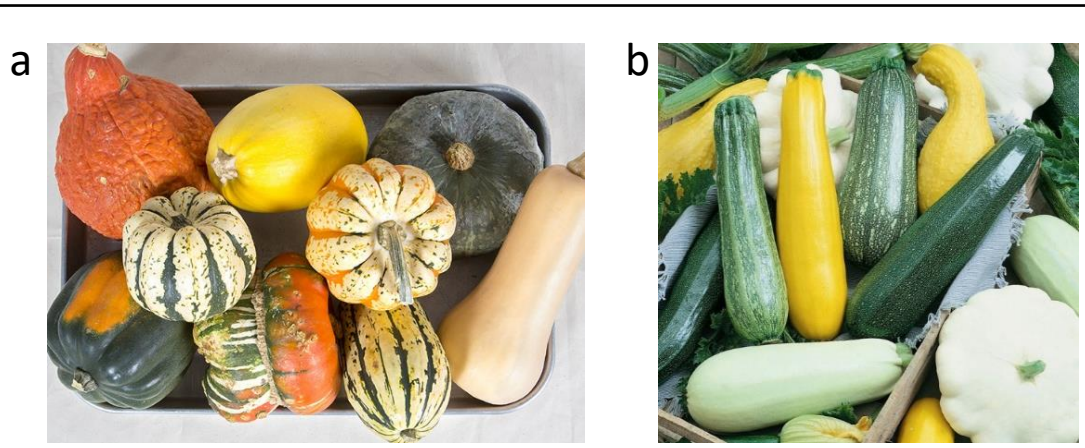

Figura 2.- Varios frutos de diferentes especies del género Cucurbita: a) frutos clasificados como calabazas de invierno y b) frutos clasificados dentro de calabazas de verano.

1.1.2. Domesticación

El origen y domesticación de las especies del género *Cucurbita* tuvo lugar en la zona de Mesoamérica. El origen de cada una de las cinco especies cultivadas del género *Cucurbita* parece ser el resultado de un suceso independiente de domesticación a partir del ancestro salvaje en diferentes lugares a lo largo del continente americano (Piperno, Holst, Wessel-Beaver, & Andres, 2002). Las culturas antiguas la cultivaban como recipiente, debido a su forma, o por sus semillas. La distribución y el uso de las especies del género *Cucurbita* se ha podido determinar principalmente gracias a los pedúnculos y semillas encontrados en yacimientos arqueológicos (Decker-Walters, D. S., and Walters, 2000).

En general, el proceso de domesticación se caracterizó por un descenso en la concentración del alcaloide cucurbitina (Paris, & Brown, 2005) y un descenso en la lignificación dando lugar a las calabazas aptas para su consumo en estado maduro. Durante el proceso de domesticación también se produjeron semillas y frutos más grandes, germinación uniforme y reducción de la dormancia de las semillas además de la adaptación a estaciones de cultivo más cortas (Lira-Saade, R., & Montes-Hernández, 1994).

C. pepo cuenta con dos centros de domesticación separados. La zona potencial de domesticación de *C. pepo* subsp. *ovifera* se extiende desde el norte de México hasta el este de EE.UU , donde se han encontrado varias poblaciones silvestres de esta subespecie (Ferriol & Picó, 2008). Estudios realizados con DNA mitocondrial indican que las subespecies silvestres de *C. pepo* (subsp. *ovifera* var. *ozarkana*, subsp. *ovifera* var. *texana* y la subsp. *fraterna*) pueden considerarse como posible origen de *C. pepo* subsp. *ovifera* (Sanjur et al., 2002), aunque *C. pepo* subsp. *ovifera* var. *texana* al presentar patrones de isoenzimas distintos, lo haga un antecesor menos válido que los otros dos (Decker-Walters, Decker-Walters, Walters, Cowan, & Smith, 1993). Por otro lado, los datos arqueológicos existentes no ayudan a resolver la cuestión de dónde se produjo la domesticación de la subespecie *ovifera*. Se han encontrado semillas y pedúnculos de tipo domesticada que provienen de una subespecie no clasificada de *C. pepo* en zonas arqueológicas situadas en el noroeste de México, muy próximas a poblaciones de *C. pepo* subsp. *fraterna* (Smith, 1997). El ancestro silvestre de *C. pepo* subsp. *pepo* es desconocido. Sanjur y colaboradores (2002) indicaron que la especie mexicana *C. pepo* subsp. *fraterna* era la población silvestre más estrechamente relacionada a *C. pepo* subsp. *pepo.* Además, información geográfica y arqueológica -- los primeros restos conocidos de *C. pepo* subsp*. pepo* se encontraron en el sur de

México-- (Fritz, 1999; Smith, 1997) apoyan esta asociación, aunque se necesitan colecciones adicionales en estas áreas. Teppner (2004) describió una subespecie adicional, *C. pepo* subsp. *gumala*, cultivada en Guatemala y México, estrechamente relacionado con las especies silvestres que podrían haber sido el punto de partida de la domesticación de *C. pepo* subsp. *pepo.*

Los primeros restos arqueológicos indicativos de la domesticación de *C.moschata* fueron descubiertos en México, que inicialmente se propuso como el centro de domesticación (Cutler, H., & Whitaker, 1967). Más tarde se desplazó hacia Perú y Guatemala según evidencias arqueológicas. Otros autores aseguran que presenta dos centros de domesticación diferentes, uno en México y otro en Sudamérica (Lira-Saade, 1995; Robinson, R. W. & Decker-Walters, 1997). Hoy en día, restos arqueológicos encontrados y la similitud morfológica con algunos cultivares de Colombia, Panamá y Bolivia, sitúan esta área como centro de domesticación o como una región secundaria de diversificación temprana (Nee, 1990; Sanjur et al., 2002). Además, también se han encontrado algunos frutos amargos en Colombia resultantes de cruces de *C. moschata* con variedades silvestres locales, lo cual respalda todavía más esta teoría. En estos momentos no se conoce el ancestro silvestre de *C. moschata*. Se propuso *C. lundelliana* de la zona del Yucatán, aunque *C. argyrosperma* es mucho más cercana (Merrick, 1990), pero debido a resultados en patrones electroforéticos y barreras reproductivas se abandonó esta idea (Sanjur et al., 2002). En estos momentos solo se puede decir que *C. moschata* solo tiene un punto de origen en algún lugar del norte de Sudamérica, a partir de un ancestro desconocido cercano a *C. argyrosperma.*

El origen de *C. argyrosperma* se sitúa en México, donde se han encontrado restos arqueológicos que sugieren que también se domesticó en esa región; (Merrick, 1990; Smith, 2005). La subespecie *sororia* es posiblemente el ancestro de la variedad cultivada, debido a su compatibilidad sexual, similitud morfológica, distribución geográfica y relación filogenética (Sanjur et al., 2002).

En cuanto a *C. maxima,* las pocas pruebas que existen la sitúan en la costa de Perú. *C. maxima* era cultivada antiguamente por las culturas precolombinas en Argentina y Paraguay. El ancestro de esta especie se considera que es *C. maxima* subsp. *andreana* (Nee, 1990; Sanjur et al., 2002). Además, los frutos de esta subespecie se han encontrado en diferentes áreas de Sudamérica, extendiendo más la zona potencial de domesticación.

Sobre *C. ficifolia*, se ha propuesto como centro de origen y domesticación la región andina (Nee, 1990; Sanjur et al., 2002) ya que solo se ha encontrado en la costa

de Perú. Sin embargo, también se cree que es de origen mesoamericano ya que el uso de sus frutos en prácticas religiosas aparece recogido en textos aztecas, pero las búsquedas de un ancestro silvestre en México no han tenido éxito (Andres, 1990). Se han obtenido algunos cruces parciales con *C. lundelliana, C. foetidissima* y *C. pedatifolia*, pero no son nada parecidas morfológicamente ni ecológicamente a *C. ficifolia*. Sin embargo sí que es parecida a *C. ecuadorensis* en los lóbulos de sus hojas, pero no en el resto de caracteres morfológicos ni ecológicos.

1.1.3. Diversidad morfológica de las especies cultivadas

Cucurbita se considera uno de los géneros morfológicos más variados en todo el reino vegetal, existe una enorme diversidad entre las 13 especies en color, tamaño y forma de la fruta (Robinson, R.W., H.M. Munger, T.W. Whitaker, 1979). A continuación se describen las principales características de las cinco especies cultivadas:

a) *C. pepo* se caracteriza por su crecimiento casi determinado, con un tallo relativamente corto, sinuoso y grueso, de sección cilíndrica y cubierto de formaciones pilosas, que lo hacen áspero al tacto. Las hojas son grandes, palmeadas y con el borde aserrado. Presentan el haz glabro y el envés piloso. Su color oscila del verde claro al verde oscuro y a veces se ve matizado por manchas blanquecinas. El pedúnculo floral es de sección poligonal. Esta especie es tal vez la más diversa de todas en cuanto a morfología del fruto se refiere (Fig. 3). Los frutos son muy diversos tanto en tamaño, forma y coloración. Una clasificación relativamente reciente que está ganando aceptación es la iniciada

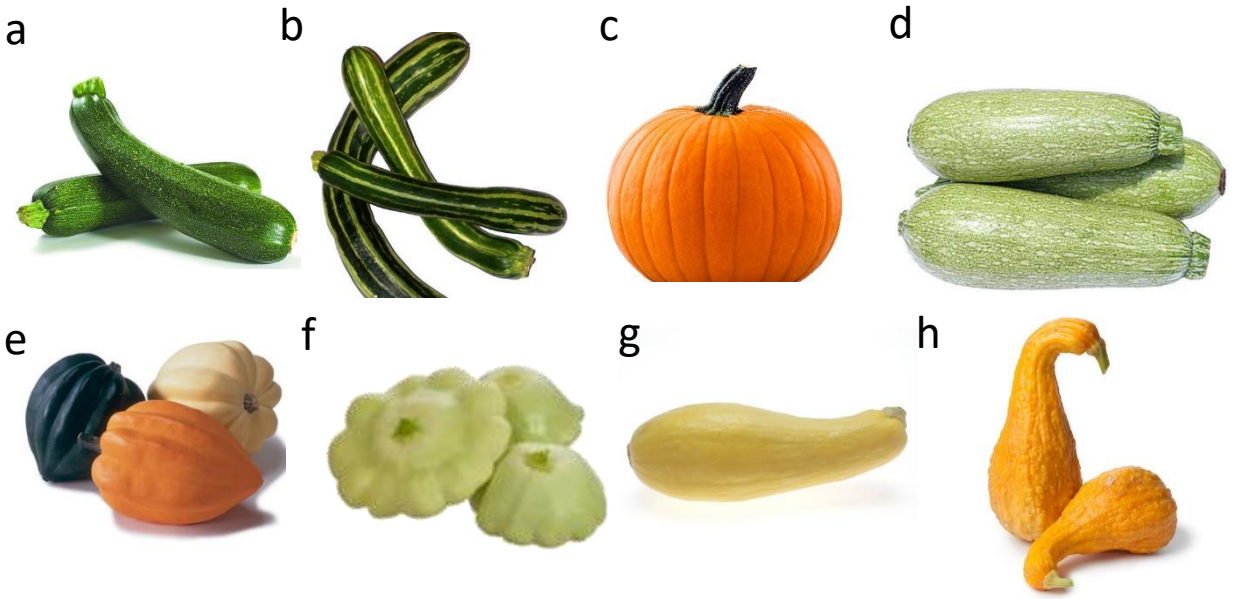

Figura 3.- Diferentes moforfotipos de *Cucurbita pepo* subsp *pepo* (a-d) y subsp *ovifera* (e-h). a) zucchini b) cocozelle c) pumpkin d) vegetable marrow e) acorn f) scallop g) straighneck h) crookneck.

por (Paris, 1986) (Fig. 3), que clasifica no en variedades botánicas (por afinidades
genéticas), sino en grupos de cultivares, de tal modo que cada subespecie posee frutos con forma y algún accidente topográfico que se reconocen como de algún grupo comercial en el mercado (Giner Martorell, Mariano, & Olivert, 2017; Josefa López Marín, 2017).

b) *C. maxima* presenta tallos de crecimiento indefinido y de sección redonda, hojas grandes, orbiculares, no lobuladas y cordadas en la base, flores amarillas y con el pedúnculo de inserción en el fruto, de forma cilíndrica y sin surcos. Los frutos suelen ser voluminosos, de color variable y carne anaranjada. Un ejemplo típico es la variedad 'Dulce de horno' también conocida como calabaza de asar (Fig. 4a) (Giner Martorell et al., 2017).

c) *C. moschata* se caracteriza por poseer tallos de crecimiento indefinido y angulosos, hojas poco enhiestas, en ocasiones aterciopeladas, poco lobuladas, con o sin manchas blanquecinas en función del cultivar y de tamaños muy variables, presentando el pedúnculo de inserción del fruto ensanchado y con surcos. Las flores son amarillas, de pétalos grandes y erectos, siendo los frutos de formas variables y color apagado. Algunos de los tipos más cultivados en España son las calabazas tipo '*Butternut'*, también denominadas 'violín' o 'cacahuete' (Fig. 4b) (Giner Martorell et al., 2017).

d) *C. ficifolia* Las hojas son pentapalmadas y de gran tamaño. Son color verde oscuro, dorso pubescente, similares a la hoja de la higuera, de donde deriva su nombre científico *ficifolia*, "de hojas de higuera" en latín. El pedúnculo del fruto es duro, de ángulos redondeados y ligeramente extendido sobre el mismo en su unión. Es relativamente homogénea, particularmente el fruto, y más o menos fácil de distinguir de los frutos de las demás especies. El color exterior puede tener básicamente 3 patrones: blanco, verde oscuro o un variegado de estos dos. A diferencia de las otras especies, las semillas suelen ser negras. Es conocida como "Cabello de ángel" o "calabaza confitera" por el gran dulzor que caracteriza su pulpa la cual se usa para elaborar gran cantidad de productos. (Fig. 4c) (Giner Martorell et al., 2017).

e) *C. argyrosperma* de tallo fuerte y angular, sin asperezas, hojas anchas, cordadas, escasamente lobuladas y, en ocasiones, con manchas blanquecinas. Presenta el pedúnculo corchoso y ancho, pero no ensanchado en la inserción del fruto. Los frutos generalmente son agrandados aunque existe diversidad en cuanto al tamaño se refiere, de carne blanda o dura y generalmente de color apagado. Las variedades más conocidas de esta especie probablemente sean

las llamadas ‘Pipián’, para consumir sus semillas molidas (México) o su fruto inmaduro como verdura de estación (América Central), y el estadounidense ‘Cushaw’, que se cultiva principalmente en el sureste de Estados Unidos. Fue conocida hasta recientemente como *C. mixta* Pangalo (Fig. 4d) (Giner Martorell et al., 2017).

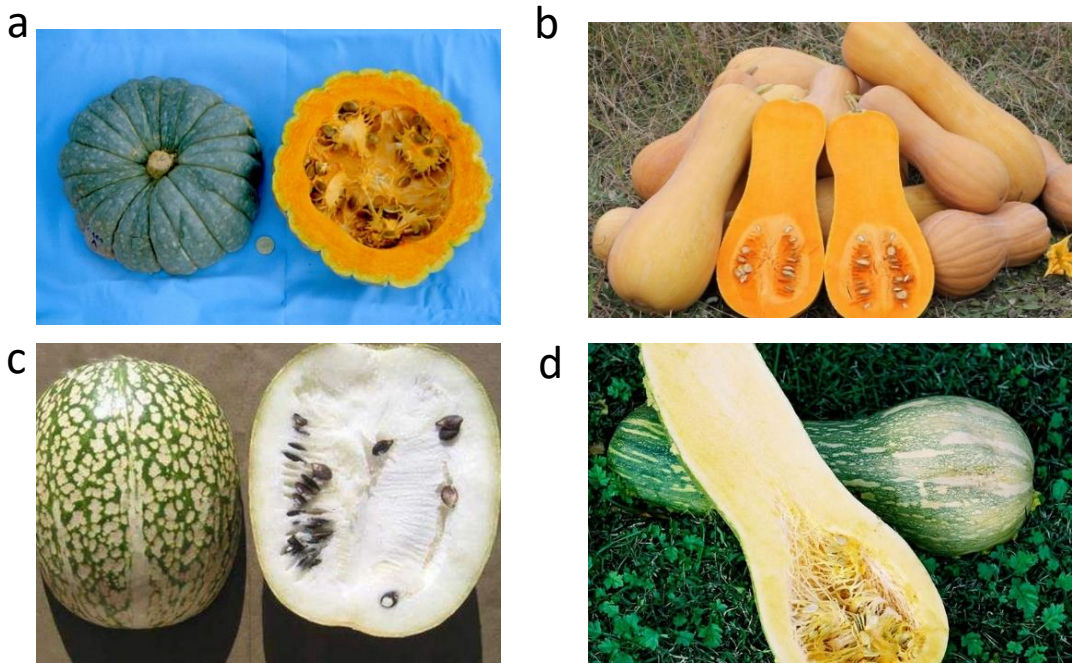

Figura 4.- Se representan frutos de diferentes especies cultivadas del género *Cucurbita.* a) *Cucurbita maxima* b*) Cucurbita moschata* c) *Cucurbita ficifolia* d) *Cucurbita argyrosperma.*

Las demandas del mercado en cuanto a una mayor diversificación del producto están aún por explotarse. *C. pepo* es una de las hortalizas que mayor variabilidad natural tiene disponible (Paris and Maynard, 2008) en cuanto a parámetros que afectan a la calidad del fruto (forma, color, sabor, calidad nutritiva), variabilidad que puede ser aprovechada para el desarrollo de futuras variedades. Así, el retraso en los programas de mejora en las especies de *Cucurbita* se debe fundamentalmente a la falta de herramientas genómicas que ha existido hasta hace relativamente poco tiempo y al gran desconocimiento sobre el control genético de los caracteres de interés para la mejora.

1.1.4. Importancia económica

Como ya se comentó anteriormente, las especies del género *Cucurbita* se dividen en dos grupos en función de sus características morfológicas y agrícolas,

pudiendo diferenciar entre “calabazas de verano” y “calabazas de invierno”. Aun así, muchas veces estos términos se superponen generando confusión, algo que también afecta a las estadísticas de producción. Por ejemplo, en la FAO (2017), solo hay una categoría para incluir a todas las especies del género *Cucurbita* (calabazas, zapayos, calabazas confiteras; pumpkins, squash and gourds, en inglés), mientras que en las estadísticas de producción del Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente (MAPAMA) se puede distinguir entre producción de calabacines (*C. pepo* subsp.*pepo*) y producción de calabazas (resto de especies).

Según la FAO (2017) (www.fao.org/faostat/es) el conjunto de especies de *Cucurbita* se clasifica entre los 10 principales cultivos de hortalizas en todo el mundo (Fig. 5a) con un volumen de producción mundial de 25.196.723 toneladas y con una superficie de cultivo total de 2.004.058 hectáreas**.** China e India lideran la producción mundial produciendo en el año 2014, 7.241.409 y 4.987.123 toneladas respectivamente (Fig. 5b). Le siguen, aunque con producciones notablemente inferiores, Rusia, Ucrania y Estados Unidos. España se encuentra en el octavo puesto con una producción de 462.266 toneladas y una superficie total dedicada a su cultivo de 10.052 hectáreas. Cabe destacar que el cultivo del calabacín es el de mayor importancia económica dentro de este grupo, aunque es difícil obtener datos de su producción por país o por superficie. En el caso concreto de España, los datos del MAPAMA (www.mapama.gob.es/es/) muestran que en 2016 la producción de total de calabacín en España fue de 581.503 toneladas**,** siendo Andalucía la principal productora de calabacín produciendo una cantidad de 477.330 toneladas de las cuales 434.195 son producidas en la provincia de Almería. Por otro lado, la producción de calabaza en España es hasta cinco veces menor que la producción de calabacín. De las 97.149 toneladas que se producen en España, un tercio es producido en la Comunidad Valenciana (32.067 toneladas). Con producciones inferiores, le siguen Canarias con 15.580 toneladas, y Navarra con 14.633 toneladas.

En general se observa que, la producción de especies del género *Cucurbita es* bastante importante a nivel mundial y a nivel nacional, especialmente *C. pepo* subsp. *pepo* (calabacín). La producción de *C. maxima* y *C. moschata,* que se incluirían dentro del grupo “calabazas de invierno”, son de gran importancia sobre todo en países sudamericanos. Mientras que la importancia de la producción de *C. ficifolia* y *C. argyrosperma* se reduce a nivel local.

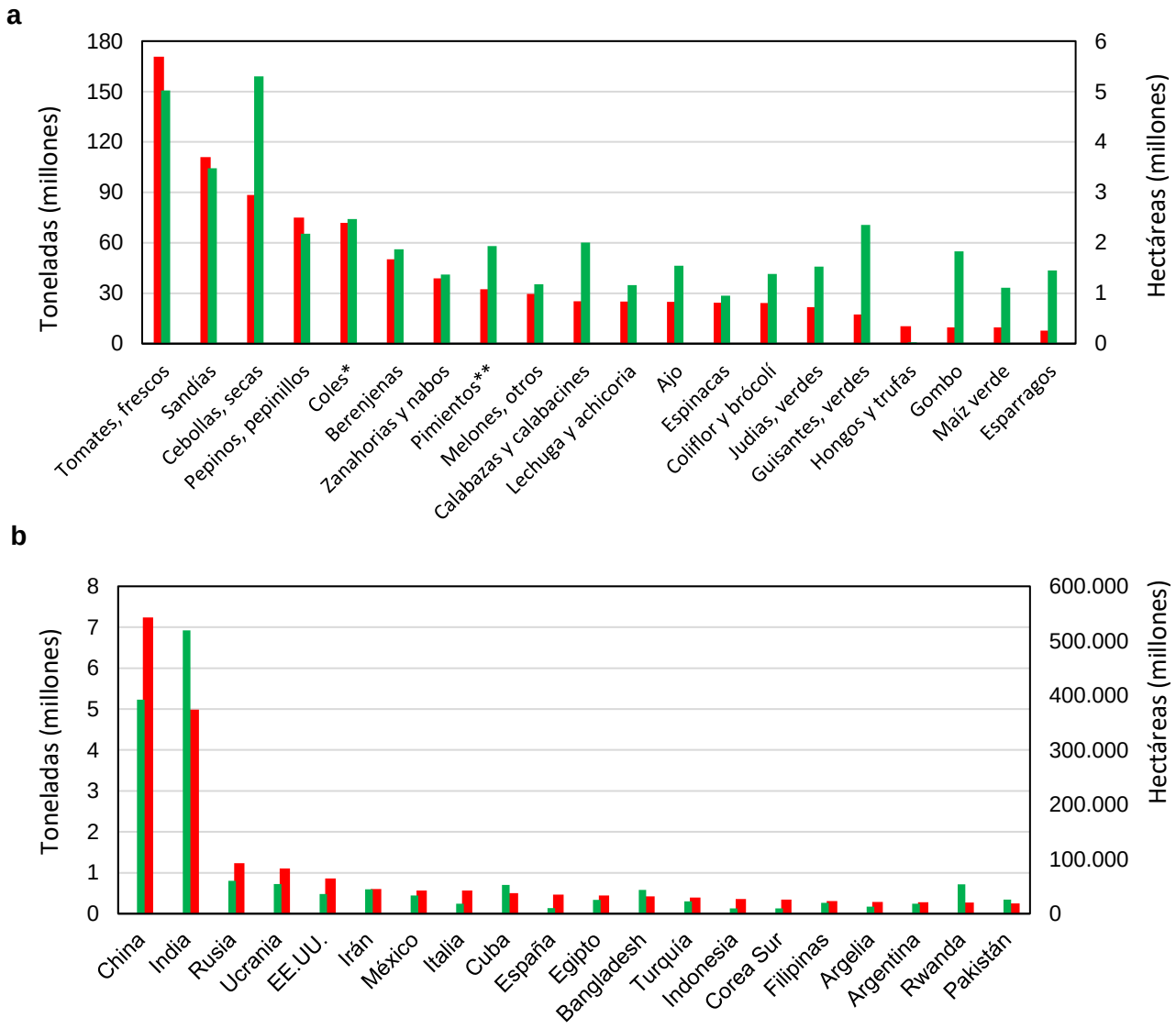

Figura 5 .- Representación gráfica de a) las principales hortalizas cultivadas en el mundo y b) principales países productores de calabazas y calabacines. En color rojo se representa la producción medida en toneladas y en color verde la superficie destinada a su cultivo, en hectáreas.
** Coles y otras crucíferas. ** Chiles, pimientos (verdes), pimientos picantes*

1.2. Diversidad genética y mejora vegetal

La diversidad genética se define como las variaciones heredables que ocurren en cada organismo, entre los individuos de una población (intrapoblacional), entre las poblaciones (interpoblacional) dentro de una especie y entre diferentes especies (interespecífica) (Glowka, Burhenne-Guilmin, & Synge, 1996). La diversidad genética es la materia prima sobre la que actúa la evolución y, ante las condiciones cambiantes del entorno, es la que determina el potencial de respuesta a la adaptación y supervivencia de los seres vivos (Glowka et al., 1996). Identificar y conservar la diversidad genética es

clave para el futuro de la mejora genética vegetal. La diversidad genética proporciona a los cultivos un seguro frente a condiciones adversas. Los recursos genéticos pueden proporcionar características útiles, tales como la resistencia a nuevas enfermedades o a la adaptabilidad de nuevas condiciones climáticas. Por lo tanto, conservar este potencial es fundamental para el desarrollo de nuevas variedades, razón demás para mantener los ecosistemas silvestres y los sistemas agrícolas tradicionales (Caruso, Broglia, & Pocovi, 2015).

Durante los procesos de domesticación de los cultivos los tamaños poblacionales disminuyeron generando un fenómeno conocido como “cuello de botella”, que tuvo como resultado una reducción drástica de su diversidad genética. Esto, junto al aumento de la consanguinidad, hace que el riesgo de extinción de estos cultivos frente a cambios ambientales sea alto. Además, los programas de mejora vegetal son responsables en gran parte de acelerar la pérdida de variabilidad genética en los cultivos. Se hace evidente la paradoja entre la generación de productos uniformes con escasa variabilidad en su interior y la necesidad de contar con variabilidad genética para continuar con los procesos de mejoramiento. Los esfuerzos de mejora de los últimos 50 años, han determinado que existe una relación entre niveles altos de productividad y niveles altos de vulnerabilidad genética, lo cual amenaza el logro de haber aumentado el potencial de rendimiento en los cultivos. La vulnerabilidad genética resulta cuando un cultivo ampliamente difundido es uniformemente susceptible a un patógeno o riesgo ambiental como resultado de su constitución genética, creando así un potencial de pérdidas generalizadas en las cosechas (Caruso et al., 2015).

Por tanto, existe la necesidad de incorporar variabilidad genética en los cultivos vegetales. La manera más eficaz de dotar a los cultivos de mejores características es cruzar variedades o especies domesticadas con germoplasma silvestre. Sin embargo, las diferencias alélicas, genotípicas y de complejos de genes entre los acervos genéticos de los materiales utilizados, reducen el éxito de la incorporación directa de variabilidad. En estos casos resultaría necesario establecer un proceso viable de transferencia genética útil, manteniendo al mismo tiempo los complejos de genes presentes en el material élite (Caruso et al., 2015). Así pues, hay una necesidad de realizar un estudio profundo de la composición genética de los materiales de trabajo y de la variabilidad genética disponible para la mejora tanto en especies próximas, como la almacenada en colecciones de germoplasma.

1.2.1. Tecnologías NGS en la mejora

La necesidad de explorar la diversidad genética en un número alto de individuos dio lugar en las últimas décadas al desarrollo de diferentes metodologías que permitieran el análisis de grandes volúmenes de muestras a un precio reducido. Así por ejemplo, se desarrollaron métodos basados en la movilidad diferencial en cromatografía o electroforesis (Q. Li, Liu, Monroe, & Culiat, 2002; Xiao & Oefner, 2001), o en la hibridación a microarrays (Borevitz et al., 2003; Tillib & Mirzabekov, 2001). El TILLING (del inglés *Targeti Induced Local Lesions in Genomes*) es un método de genética inversa de alto rendimiento y bajo costo que combina la mutagénesis química aleatoria con el cribado de regiones genéticas de interés (Colbert, 2001; McCallum, C. M., Comai, L., Greene, E. A., & Henikoff, 2000). El TILLING se inicia generando una colección de individuos mutantes con etilmetanosulfonato (EMS) u otros mutágenos, que posteriormente se analizan de forma sistemática, para identificar mutaciones puntuales en genes seleccionados. La identificación de mutaciones no requiere la secuenciación de cada genotipo, sino que se lleva a cabo a gran escala obteniendo moléculas homodúplex o heterodúplex a partir de un proceso de desnaturalización-renaturalización mezclando parejas de amplicones (Esteras Gómez & Picó Sirvent, 2011). El EcoTILLING es una variante de TILLING donde en lugar de analizar poblaciones provenientes de mutagénesis inducida, los materiales de partida que se utilizan como fuente de variabilidad suelen ser los almacenados en bancos de germoplasma, o bien ecotipos de especies no domesticadas (Comai et al., 2004). Esta tecnología permite optimizar la búsqueda de nuevos caracteres mediante el análisis de la variabilidad de los genes. Asimismo, entre sus ventajas, permite detectar nuevos alelos de resistencia a enfermedades y otros caracteres de interés en plantas. Una de las desventajas potenciales de EcoTILLING es la dificultad para la detección de un alto número de polimorfismos en un gen o fragmento de PCR, por lo que en estos casos es necesario realizar un procedimiento más costoso. Por lo tanto, una desventaja potencial de EcoTILLING es que cuando la variación es alta, la eficacia de la técnica disminuye (Barkley & Wang, 2008). Está técnica ha sido utilizada con éxito en plantas y animales. (Comai et al., 2004). En la mejora de cultivos ha sido de gran utilidad en remolacha azucarera (Frerichmann et al., 2013), en olivo (Sabetta et al., 2013), en trigo (A. Li et al., 2013), arroz (Yu et al., 2012) melón (Nieto et al., 2007), cebada (Mejlhede et al., 2006), pimiento (Ibiza, Cañizares, & Nuez, 2010) o en otros cultivos como *Populus trichocarpa* (Gilchrist et al., 2006).

Sin embargo, la llegada de las tecnologías de secuenciación de nueva generación o “next generation sequencing” (NGS) ha supuesto un enorme avance en la

exploración de la diversidad genética. Las tecnologías NGS engloban un conjunto de técnicas de secuenciación distintas a la secuenciación de Sanger (Service, 2006) que permiten la obtención de grandes cantidades de información genética con un coste reducido. Estas tecnologías han sido de especial utilidad para las especies denominadas huérfanas o de menor importancia (Crespi, 2012; Varshney, Close, Singh, Hoisington, & Cook, 2009) y especialmente para cultivos en los cuales apenas se disponía de recursos genómicos (Varshney, Close, et al., 2009). Gracias a la tecnología NGS, actualmente es posible secuenciar genomas completos o transcriptomas de un gran número de individuos (Causse et al., 2013; M. Perez-de-Castro et al., 2012; Qin et al., 2014; Subbaiyan et al., 2012; Thudi et al., 2016; Varshney, Nayak, May, & Jackson, 2009). La disponibilidad de un transcriptoma o un genoma de referencia y la información obtenida en la resecuenciación son herramientas básicas para obtener grandes colecciones de marcadores de alta calidad de tipo SSR (*Simple Sequence Repeat*) y SNP (*Single Nucleotide Polymorphism*). Sin embargo, la secuenciación de genomas completos sólo se ha aplicado de manera marginal aestudios filogenéticos, filogeográficos o de genética de poblaciones de plantas y animales por el elevado coste de trabajar con un tamaño muestral alto (López de Heredia, 2016). No obstante, la secuenciación masiva puede ser aplicada a bibliotecas genómicas que consigan una reducción del genoma de los organismos y a partir de la cual se obtenga un número elevado de variantes genómicas en múltiples individuos. La reducción del genoma se consigue en estas metodologías mediante el empleo de endonucleasas de restricción, que fragmentan la doble hélice en sitios específicos. Esto es la base de la técnica RAD-seq (*restriction site associated DNA markers sequencing*). A día de hoy existen numerosas modificaciones de la técnica RAD-seq original, siendo las dos más populares por en el estudio de especies de plantas y animales el genotipado por secuenciación (GBS) (Elshire et al., 2011); y *double-digested* RAD-seq (ddRAD-seq) (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). Otras modificaciones como ezRAD-seq (Toonen et al., 2013) o el genotipado por secuenciación con dos enzimas, como 2-enzyme GBS (J. A. Poland, Brown, Sorrells, & Jannink, 2012) o 2b-RAD (S. Wang, Meyer, McKay, & Matz, 2012), han tenido hasta ahora un menor uso. La técnica de GBS se desarrolló para estudios de mapeo genético en plantas de maíz (Elshire et al., 2011). La principal modificación con respecto a la técnica original consiste en la utilización de endonucleasas de restricción resistentes a metilación (originalmente *ApeK*I), de manera que se evitan regiones repetitivas del genoma, aumentando la profundidad de secuenciación de regiones con un bajo número de copias. El GBS ha tenido una gran repercusión y ha demostrado ser una herramienta válida para ser usada en la mejora genética vegetal (J. Poland et al., 2012), también hay algunos ejemplos de su uso en

estudios de diversidad, (C. Chen, Mitchell, Elshire, Buckler, & El-Kassaby, 2013; Ilut et al., 2015; Ratcliffe et al., 2015; Schilling et al., 2014). Otra metodología con potencial para el estudio de la diversidad biológica es la técnica ddRAD-seq (Peterson et al., 2012). La principal modificación que presenta respecto de la técnica original es el uso conjugado de dos endonucleasas de restricción, seleccionando únicamente los fragmentos que están flanqueados por cada una de las enzimas de restricción y que muestran un tamaño de inserto definido por el investigador. Se ha utilizado para estudios de hibridación y delimitación de especies, establecimiento de estructura poblacional o detección de regiones genómicas. En plantas ha sido utilizada en coníferas (Friedline et al., 2015) y en frondosas (Mastretta-Yanes et al., 2014).

La disponibilidad de un gran número de marcadores genéticos desarrollados a través de tecnologías NGS está facilitando el mapeo de caracteres de interés y haciendo más factible el uso de marcadores moleculares en mejora genética**.** Así mismo, el rápido desarrollo de marcadores moleculares por todo el genoma puede acelerar el proceso de mapeo por ligamiento y los estudios de asociación genómico (*genome* wide association study, GWAS) (Alagna et al., 2009). El hecho de poder secuenciar poblaciones enteras en lugar de solamente individuos, ayudará a ampliar nuestra comprensión en el campo de la genética de poblaciones. La posibilidad de secuenciar también RNA utilizando tecnologías NGS ha sido de gran utilidad para los estudios de expresión (Aslam, Khattak, Ahmed, & Asif, 2017). En la epigenética, las NGS también han jugado un papel importante, llegando a reemplazar al ChIP-chip (técnica que combina la inmunoprecipitación de la cromatina, ChIP, con el uso de microarrays) por la secuenciación ChIP. Esta técnica implica la inmunoprecipitación de la cromatina de forma tradicional seguido de la secuenciación directa (Aslam et al., 2017). Aunque esta técnica está bien establecida para el análisis del genoma humano (Mardis & Ris, 2007), actualmente está en desarrollo para sistemas vegetales y solo hay disponible información en *Arabidopsis* (Kaufmann et al., 2010)**.**

1.2.2. Bioinformática en la mejora

Las nuevas tecnologías de secuenciación son capaces de producir datos a un ritmo elevadísimo y en grandísimas cantidades. Esto está creando un desafío tanto de manejo de datos como analítico: se requiere de la generación de herramientas y métodos para poder analizar la gran cantidad de datos que se producen como resultado de los proyectos de secuenciación de genomas (Singh VK, Singh AK, Chand R, &

Kushwaha C, 2011). Se requiere software y algoritmos que permitan tratar estos datos de una manera eficiente (Shendure, Mitra, Varma, & Church, 2004), crear bases de datos, herramientas especializadas y métodos computarizados para organizar, almacenar, analizar y visualizar los datos.

La bioinformática surge como una herramienta indispensable para utilizar la gran cantidad de datos generados con el fin de encontrar soluciones a problemas biológicos. Es una ciencia interdisciplinar que surge de la interacción entre la estadística, la biología, las matemáticas y las ciencias de la computación con el fin de poder analizar genomas, datos de secuencias biológicas y predecir la estructura y función de macromoléculas. Se encarga de desarrollar programas, algoritmos, bases de datos y herramientas de análisis de datos para hacer descubrimientos e inferir la información. Dado el constante aumento de la información biológica, la bioinformática desempeña un papel fundamental en el proceso de toma de decisiones ya que hace posible la automatización de gran cantidad de procesos y análisis de la información que se genera a diario en los laboratorios (Varshney, Nayak, et al., 2009). La aplicación de varias herramientas y bases de datos permite el análisis, el almacenamiento, la anotación, la visualización y la recuperación de resultados para ayudar a la comprensión del funcionamiento de los organismos. Realizar tareas como el ensamblado de genomas, mapeo de secuencias o búsqueda de SNP, es ahora posible gracias al desarrollo de numerosos programas. Obtener un genoma, o en su defecto, un transcriptoma, de referencia es una herramienta valiosísima para cualquier estudio genómico, por ello se han invertido grandes esfuerzos en desarrollar algoritmos y programas capaces de ensamblar genomas de *novo.* Para cumplir con el requisito de una alineación eficiente y precisa de millones a miles de millones de secuencias de lectura corta contra un genoma de referencia, se han desarrollado una gama completamente nueva de alineadores. Para la búsqueda de polimorfismos se han desarrollado también diferentes softwares llamados *snp-caller*. No solo se han generado nuevas herramientas, sino que las diferentes etapas del procesamiento de datos de NGS han hecho necesario el desarrollo de nuevos formatos para almacenar diferentes tipos de información (e.g. fastq; s*equence alignment/map format,* sam*;* b*inary alignment/map format,* bam o v*ariant call format,* vcf) y herramientas que permiten su manejo (SAMTools, VCFtools o Picard-tools…).

El uso de nuevas bases de datos y herramientas en el campo de la biología molecular permitirá al investigador analizar e investigar no solo el genoma, sino también el transcriptoma, el metabaloma y el proteoma. Por citar sólo un ejemplo, los genes conocidos y almacenados en bases de datos son ya millones y los análisis de expresión génica o de secuenciación de nuevos genomas generan inmensas cantidades de datos,

que necesitan ser almacenados y analizados de forma automática para generar nuevo conocimiento (Varshney, Nayak, et al., 2009).

La bioinformática en la agricultura desempeña un papel cada vez mayor. Cada vez es mayor el número de proyectos de secuenciación de genomas para diferentes tipos de cultivos. Un ejemplo serían las especies pertenecientes a la familia de las poáceas, como el arroz o el maíz, cuya información genómica podría ser utilizada en futuros programas de mejora para obtener variedades resistentes a determinadas plagas (Aslam et al., 2017). El descubrimiento de nuevos genes utilizando programas bioinformáticos ha tenido como objetivo mejorar la calidad de las semillas, agregar o aumentar el contenido de algunos micronutrientes con el fin de mejorar la salud humana y desarrollar nuevas plantas, o bien por medio de la ingeniería genética o mediante mejora tradicional, para tratar problemas medioambientales relacionados con la acumulación de metales (fitorremediación) (Varshney, Nayak, et al., 2009). Además, el estudio de datos genéticos o genómicos en especies vegetales está siendo enfocado a encontrar genes asociados a ciertos fenotipos deseables, llamados "loci de rasgos económicos" o *"economical traits loci"* (ETL).

Las bases de datos están en constante crecimiento debido al aumento en el número de secuencias generadas que se acumulan, donde en muchos casos, éstas no aportan nueva información, por lo que existe una demanda de condensar y eliminar datos redundantes (Varshney, Nayak, et al., 2009). El desarrollo de las bases de datos junto con los avances en las herramientas de análisis ha permitido a los investigadores extraer conclusiones biológicas de datos complejos y anotar secuencias completas (D. Edwards & Batley, 2008; David Edwards & Batley, 2004). Actualmente muchos de los cultivos de mayor importancia económica disponen de su propia base de datos en las cual se va depositando información. Algunos ejemplos son el maíz (https://www.maizegdb.org/), la soja (www.soybase.org) o el trigo (https://www.wheatgenome.org). También existen bases de datos en las que se alberga información genómica de varios cultivos de un mismo género o familia botánica, como es el caso de *Solgenomics* (https://solgenomics.net/), para solanáceas o *Cucurbitgenomics* (http://cucurbitgenomics.org/) para las cucurbitáceas.

Hoy en día la comunidad bioinformática se enfrenta a varios desafíos, entre ellos asegurar el almacenamiento de gran cantidad de datos de forma eficiente e inteligente y al mismo tiempo, proporcionar un acceso a dichos datos de forma sencilla. Además, es necesario desarrollar herramientas computacionales que sean capaces de extraer y manipular esta información biológica. La utilización de herramientas y métodos bioinformáticos es fundamental para la identificación, búsqueda y anotación de genes específicos dentro de genomas de especies de interés comercial. De este modo se

espera que la secuenciación de los genomas de especies vegetales proporcione grandes beneficios para la comunidad agrícola tales como la obtención de cultivos resistentes a plagas y enfermedades, tolerantes a la sequía, más productivos y de mejor calidad.

1.3. Recursos genómicos en *Cucurbita*

El desarrollo de herramientas genómicas y su aplicación para la mejora ha sido reciente y no se ha avanzado de la misma forma en los diversos cultivos de cucurbitáceas. A pesar de la gran importancia económica, los estudios genéticos y genómicos en el género *Cucurbita* han sido limitados, en comparación con otros cultivos de la misma familia. Así para pepino (*Cucumis sativus*), melón (*Cucumis melo*) o sandía (*Citrullus lanatus*) han existido diversas herramientas como plataformas de TILLING de genética reversa (Boualem et al., 2014; González et al., 2011) colecciones de marcadores de alta calidad (Blanca et al., 2011; Guo et al., 2011; Kong, Xiang, & Yu, 2006). También se han diseñado microarrays de expresión (Mascarell-Creus et al., 2009; Wechter et al., 2008) y han elaborado mapas genéticos densos (Deleu et al., 2009). Así mismo, también se dispone de mapas de QTLs (*quantitative trait locus*) para varios caracteres de interés, sobre todo caracteres relacionados con el fruto, la floración y estructura de la planta (C. et al., 2002; Gonzalo & Monforte, 2016; Hashizume, Shimamoto, & Hirai, 2003; Wenzel, Kennard, & Havey, 1995). Además desde hace unos años también se cuenta con el genoma de estos cultivos. El genoma del pepino está disponible desde 2009 (Huang et al., 2009), y desde 2012 también están disponibles los genomas del melón (Garcia-Mas et al., 2012) y de la sandía (Guo et al., 2012). Muchos de estos recursos están disponibles en la base de datos http://cucurbitgenomics.org/). Así el desarrollo de herramientas genómicas que hoy en día son esenciales y la identificación de genes de interés son esenciales para llevar a cabo una mejora competitiva de las especies pertenecientes al género *Cucurbita*.

En *Cucurbita,* el primer mapa genético se desarrolló empleando marcadores RAPD (*Random Amplification of Polymorphic DNA*) y AFLPs (*Amplified fragment length polymorphism*) (Brown & Myers, 2002; Lee, Jeon, Hong, & Kim, 1995; Zraidi et al., 2007) por lo que prácticamente era imposible hacer comparaciones con otros mapas. En 2008 se elaboró el primer mapa basado en SSR a partir de un cruce entre los cultivares *Pumpking* x *Crookneck* de *C. pepo* (Gong, Stift, Kofler, Pachner, & Lelley, 2008). Mediante el uso de las nuevas herramientas genómicas y a partir de varios experimentos de secuenciación masiva se han realizado grandes avances en la obtención de

información para las especies de este género. En 2011, se obtuvo el primer transcriptoma de *C. pepo* con un total de 49.610 *unigenes* y se reportó una colección de aproximadamente 20.000 SNPs (Blanca et al., 2011) que sirvió para crear la primera plataforma de genotipado disponible en *C. pepo.* Esta plataforma se usó para construir el primer mapa genético basado en SNPs en una población segregante de *C. pepo* subsp.*pepo* var. Zucchini x *C. pepo* subsp.*ovifera* var. Scallop (Esteras et al., 2012). En 2017 se construyó el primer mapa saturado de SNP en esta especie anclado al mapa físico. Además, se encontraron varios QTLs relacionados con la floración temprana, la forma y la longitud del fruto, y el color de la piel y la carne, así como los genes candidatos para ellos.(Montero-Pau, Blanca, Esteras, et al., 2017). Otro de los avances logrados en esta especie es el desarrollo de la primera plataforma de genética reversa, la cual se generó para *C. pepo* (Vicente-Dólera et al., 2014). Este hecho supone un gran avance para la mejora del calabacín ya que es una herramienta de mayor flexibilidad que otras como los transgénicos o T-DNA (L. Chen et al., 2012) y potencia el estudio de genes de interés agronómico cuya función hasta ahora ha sido desconocida (T. L. Wang, Uauy, Robson, & Till, 2012).

En otras especies de interés comercial del género *Cucurbita* también ha sido posible desarrollar nuevas herramientas gracias a las tecnologías de secuenciación NGS. En el caso de *Cucurbita moschata,* fue en 2014 cuando se obtuvo por primera vez la secuencia del transcriptoma (Wu et al., 2014) el cual cuenta con un total de 47.899 *unigenes.* Así mismo, se identificó una colección de 7.814 SSR que fueron implementados en una plataforma con el fin de ser utilizados para realizar estudios de genética funcional o bien para ser usados en programas de mejora genética de la calabaza. Para *C. maxima* son menos los recursos de los que se disponen. Aun así para esta especie se disponen de mapas genéticos de alta densidad (Zhang et al., 2015) y actualmente también se dispone de la secuencia del transcriptoma (Sun et al., 2017).

El hecho de contar con la información que nos proporciona un genoma de referencia puede ser de gran utilidad para cualquier estudio genómico. El genoma de las principales especies comerciales de *Cucurbita* no ha estado disponible hasta hace relativamente poco. En 2017 se publicó la secuencia de estas tres especies (Montero-Pau, Blanca, Bombarely, et al., 2017; Sun et al., 2017). La obtención del genoma para estas especies nos da la posibilidad de analizar y comparar entre si gran cantidad de información genética de diversos individuos. Así mismo, el hecho de contar con un genoma de referencia sirve como punto de partida en el estudio del funcionamiento de los organismos.

Además existen diferentes fuentes de datos en los que todavía se sigue trabajando. Existen proyectos aún sin publicar como es el caso del *Whole genome re-*

sequencing of Cucurbita pepo varieties de Ganopoulus y colaboradores (www.researchgate.net/project/Whole-genome-re-sequencing-of-Cucurbita-pepo-varieties), cuyo objetivo es encontrar variación en la secuencia (SNP, eliminaciones e inserciones) en diferentes variedades de *Cucurbita pepo;* o información sobre 96 transcriptomas de hoja joven de diferentes especies del genero *Cucurbita* dentro de las cuales incluye especies silvestres, que fueron secuenciados por la Dra. Belén Picó.

En resumen, el género *Cucurbita* posee un gran potencial y margen para la mejora, y su importancia económica justifica el interés de explotar la gran diversidad que existe en las especies que lo componen y que todavía está por explorar. Estudiar y describir la información que nos proporcionan los recientes genomas y transcriptomas obtenidos en *Cucurbita* es el primer paso hacia el desarrollo de bases de datos y herramientas útiles para la mejora. Además, es interesante que éstas puedan ser puestas a la disposición de los mejoradores para que puedan trabajar de forma más eficiente en este cultivo. Este es exactamente el objetivo principal del presente trabajo. Para conseguir este objetivo se proponen los siguientes objetivos específicos que se mostrarán a continuación.

2. Objetivos

Los objetivos que se persiguen en el presente trabajo son:

- Describir la diversidad existente en una población formada por 96 individuos de diferentes especies del género *Cucurbita.*
- Obtener una colección de SNPs de alta calidad para el conjunto de especies del género *Cucurbita.*
- Anotar y clasificar los SNPs en función de su impacto fenotípico de manera que puedan servir como elemento de referencia a la hora de buscar genes candidatos en futuras investigaciones y proyectos de mejora.
- Mostrar la utilidad de esta aproximación mediante su aplicación a genes de interés que se encuentran dentro de regiones QTL conocidas.
- Generar y poner a disposición del público un recurso útil para la mejora.

3. Materiales y métodos

3.1. Origen de las secuencias

El presente trabajo se ha desarrollado con un conjunto de 96 transcriptomas secuenciados por la Dra. Belén Picó responsable del Laboratorio de Cucurbitáceas del Instituto de Conservación y Mejora de la Agrodiversidad Valenciana de la Universidad Politécnica de Valencia (COMAV-UPV). Las 96 accesiones secuenciadas incluyen 12 especies silvestres y cultivadas del género *Cucurbita* (ver tabla 1 y tabla suplementaria 1). Los transcriptomas se obtuvieron a partir de mRNA extraído de hoja joven y las bibliotecas de cDNA fueron secuenciadas usando un Illumina Hiseq2000 (para más detalles del protocolo ver Montero-Pau, Blanca, Bombarely, et al., 2017). Las secuencias en crudo de los transcriptomas han sido cedidas para la realización de este trabajo. Parte de estas secuencias se encuentran depositadas en el NCBI bajo el BioProject PRJNA386743.

Tabla 1.- Especies del género Cucurbita utilizadas y número de individuos utilizados por especie.

Especie	Nº muestras
C. cordata	2
C. ecuadorensis	3
C. ficifolia	2
C. lundelliana	4
C. argyrosperma subsp. *argyrosperma*	5
C. foetidissima	3
C. maxima	13
C. maxima subsp. *andreana*	1
C. moschata	27
C. okeechobensis subsp. *martinezii*	3
C. pedatifolia	3
C. pepo subsp. *ovifera*	10
C. pepo subsp. *fraterna*	2
C. pepo subsp. *pepo*	17
C. scabridifolia	1

3.2 Preprocesado de las lecturas

Antes del mapeo de las secuencias y la búsqueda de SNPs, se inspeccionó la calidad de las lecturas en crudo y se procedió a su filtrado y limpieza. Para el análisis de la calidad se empleó el programa *FastQC* (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Esta herramienta elabora un informe de control de calidad de los datos que permite detectar posibles problemas originados tanto en el secuenciador como en el material de partida. Se prestó especial atención al análisis de calidad por base, análisis de calidad de las lecturas, los datos de conteo de adaptadores y la información sobre secuencias sobrerrepresentadas. Se creó un "script" en Shell para extraer esta información y automatizar este proceso. Los resultados fueron inspeccionados a mano.

Tras el análisis de calidad, fue necesario realizar un procesamiento previo de las lecturas para eliminar adaptadores, regiones de baja calidad y lecturas cortas. Para ello se usó *Trimmomatic* versión 0.36 (Bolger, Lohse, & Usadel, 2014). Se filtró por calidad media de lectura, eliminando aquellas lecturas que tuvieran un valor de Phred promedio inferior a 15 en ventanas de 4 nucleótidos. Además, se eliminaron los nucleótidos de los extremos cuyo valor de Phred fuera menor a 20, ya que al principio y al final de las lecturas el número de errores suele ser mayor debido al propio proceso de secuenciación. Tras la aplicación de los filtros, se eliminaron aquellas lecturas muy cortas cuyo tamaño no llegaba a 50 bases (i.e., la mitad del tamaño de las lecturas crudas).

3.3. Mapeo de las lecturas

Las lecturas limpias se mapearon contra la versión 4.1 del genoma de *C. pepo* (Montero-Pau, Blanca, Bombarely, et al., 2017). Se usó el mapeador *HISAT2* v. 2.1.0 (Kim, Langmead, & Salzberg, 2015). Este es un programa de alineación rápida y sensible para mapear lecturas contra un genoma de referencia. Es un mapeador de tipo "splice-aware", es decir, reconoce y permite el mapeo considerando el *splicing*, algo que resulta útil para alinear lecturas que provienen de RNA-seq.

La calidad del mapeo se evaluó con la herramienta *stats* del paquete *Samtools* v. 1.5 (Li et al., 2009) ya que los alineamientos pueden contener errores, secuencias duplicadas o secuencias que han sido alineadas con una calidad baja. Una vez analizadas las estadísticas del alineamiento, se optó por filtrar las lecturas con una calidad de mapeo (MAPQ) inferior a 57 usando *Samtool*s, además se marcaron las lecturas duplicadas usando *MarkDuplicates* de *PicardTools* v. 2.8.1 (http://broadinstitute.github.io/picard/) y

posteriormente se eliminaron con *Samtool*s. Tras la limpieza se volvieron a realizar las estadísticas para comprobar la efectividad del filtrado.

Es común que en los extremos de las lecturas se produzcan errores de alineación sin que ello afecte al MAPQ. Esto da como resultado la detección de falsos SNPs en dichas regiones. Para evitarlo, se bajó la calidad de los 3 nucleótidos de los extremos de cada lectura mapeada usando Downgrade_*bam_edge_qual* versión 0.1, un script del paquete *ngs_crumbs* (https://github.com/JoseBlanca/ngs_crumbsl). También se modificó el Read Group (usando el programa *AddOrReplaceReadGroups* de *Picard Tools*) de todos los BAMs para introducir información relevante de cada muestra, tal y como nombre de la muestra, la librería o plataforma de secuenciación empleada. Todos los ficheros BAMs fueron ordenados e indexados usando *Samtools*.

3.4. Búsqueda de SNPs y filtrado

Con los alineamientos filtrados, se procedió a realizar la búsqueda de SNPs usando *Freebayes* v. 1.1 (Garrison & Marth, 2012). Este programa es un *snp-caller* basado en haplotipos desarrollado en el marco de la estadística Bayesiana que se emplea para detectar variantes genéticas. Previo al uso de Freebayes, fue necesario eliminar las regiones intrónicas de los alineamientos. Para ello se usó *SplitNCigarReads*, una herramienta del paquete *GenomeAnalysisToolkit* v. 3.3 (*GATK*) (https://software.broadinstitute.org/gatk), desarrollada especialmente para RNA-seq que identifica las lecturas que contienen elementos N en el CIGAR y los elimina, dividiendo las lecturas en fragmentos de exones. Se ejecutó con los parámetros por defecto.

En el *snp-calling,* para reducir el número de falsos SNPs se excluyeron de la búsqueda los alineamientos cuya calidad de mapeo fuese inferior a 57, los SNPs con una cobertura inferior a 10 y los alelos que tuviesen una calidad de base inferior a 20. Por otro lado, se estableció que la frecuencia mínima para considerar un alelo en un individuo debía ser del 0,2 (i.e., debe estar presente en 2 de cada 10 lecturas). Esto permite discriminar los alelos de los errores puntuales de secuenciación. Además se consideró que no se debía asumir que la población se encontrara en equilibrio de Hardy-Weinberg. Por último, se solicitó que se calculara el valor de calidad asociado a cada genotipo y lo incluyera en el fichero VCF. Para la obtención de datos estadísticos sobre el proceso de búsqueda de variantes, tanto de las muestras individuales como del conjunto de la población se utilizó *calculate_vcf_stats* de *ngs_crumbs* v. 0.1.

Se realizó un filtrado de los archivos con la intención de eliminar aquellos SNPs de baja calidad. Se eliminaron los genotipos de las muestras determinados con baja calidad (posición leída menos de 5 veces). Se eliminaron aquellos SNPs en los que la

frecuencia del alelo minoritario fuese menor de 0,005 en la población, esta frecuencia posibilita identificar alelos raros en heterocigosis. También se eliminaron aquellos sitios en los que después de los filtrados hubiese un 100% de datos faltantes. El filtrado de los ficheros VCFs se realizó con *VCFtools* versión 0.1.14 (Danecek et al., 2011). En cada etapa de filtrado se usó la herramienta *stats* de *BCFtools* versión 1.5 (https://samtools.github.io/bcftools) para obtener estadísticas de forma rápida sobre el número de polimorfismos que iba quedando.

3.5. Anotación del efecto de los SNPs

Para la anotación de los diferentes ficheros VCF se usó *SnpEff* versión 4.3r (Cingolani et al., 2012), un programa de anotación y predicción de efectos de cambios genéticos. Este software estima el efecto que producen en un gen una variante genética (SNP, inserción, deleción o polimorfismo de múltiples nucleótidos) y los clasifica en función del impacto causado en cuatro categorías (leve, moderado, grave y modificador). Además del archivo de salida VCF anotado, *SnpEff* devuelve un fichero *snpEff_summary* en formato HTML con las estadísticas de la anotación y un fichero *snpEff_genes,* en formato texto (.txt), en el cual se listan todos los genes en los que se han identificado polimorfismos. Se utilizaron las opciones por defecto. Una de las tareas previas a la anotación del archivo VCF fue la creación de una biblioteca con información relevante al genoma de *C. pepo.* Se creó un archivo *SnpEffPredictor* en formato binario (.bin) para ubicar cada SNP dentro de regiones codificantes, regiones intrónicas, genes anotados, etc. Para ello se necesitó el archivo en formato FASTA con el genoma de referencia y el archivo GFF3 con la anotación del genoma (Montero-Pau, Blanca, Bombarely, et al., 2017).

3.6. Búsqueda de polimorfismos en regiones QTLS

Para mostrar la utilidad de la exploración de la diversidad genética en las especies del género *Cucurbita,* se decidió explorar el efecto de la colección de SNPs encontrados en regiones ligadas a rasgos de interés para la mejora. En un trabajo previo, Montero-Pau et al (2017) elaboraron un mapa genético a partir de una población RIL de calabacín que usaron para la detección de locus de rasgos cuantitativos (QTL) relacionados con 43 rasgos fenotípicos. Para algunos de dichos QTLs se indicaban posibles genes candidatos. Se seleccionaron 8 de ellos. Se escogieron aquellos que presentaron mayores valores de varianza explicada (R^2), mayores valores de LOD y que

contuviesen genes candidatos. Además se tuvo en cuenta el tamaño de cada QTL, dando prioridad a las regiones de menor tamaño. La información relativa a estas regiones se extrajo del fichero VCF anotado con *SnpEff* mediante *intersectBed,* del kit de herramientas *BEDTools* versión v2.26.0 (Quinlan & Hall, 2010), y el número de SNPs por gen en la región y su efecto predicho fue computado. Para ello se usó el conjunto de herramientas *BEDTools* y la información de anotación de genes contenida en el GFF3.

En las diferentes etapas de este proceso, es conveniente ir comprobando y analizando de forma visual el contenido de los diferentes archivos con los que se trabajó (BAMs, VCF, GFF3, etc) para identificar errores de forma rápida y obtener información gráfica de cada etapa. La aplicación que se utilizó el visor *Integrative Genomics Viewer* (IGV) versión 2.3.97 (Thorvaldsdottir, Robinson, & Mesirov, 2013).

4. Resultados

4.1. Procesado de las secuencias y mapeo

Para el análisis se utilizaron un total de 83,1 Gb obtenidas mediante secuenciación masiva del transcriptoma de las 96 muestras de *Cucurbita sp.* (Tabla 1 y Tabla suplementaria 1), lo que supone un total de 814.262.466 lecturas con una longitud de 101 nucleótidos. El número de lecturas crudas por muestra varió desde 1.892.580 (CO-54) hasta 27.822.108 (CO-84), siendo la media de lecturas en crudo por muestra de 8.481.901 (Tabla 2).

Tabla 2.- Resumen del número de lecturas en las diferentes etapas de procesado y mapeo. Se presenta el total de lecturas y el promedio, valor máximo y mínimo, mediana, desviación estándar (Desv. Est.) y coeficiente de variación (CV) para los datos del número de lecturas por muestra.

	Total	Promedio	Máximo	Mínimo	Mediana	Desv. Est.	CV
Lecturas "crudas"	814.262.466	8.481.901	27.822.108	1.892.580	7.962.209	3.494.525	41,2%
Lecturas de calidad	777.034.242	8.094.107	26.450.360	1.788.174	7.575.440	3.339.617	41,3%
Lecturas mapeadas	598.290.404	6.232.192	15.727.450	1.746.738	5.718.294	2.676.432	42,9%
Lecturas mapeadas con alta calidad	410.778.570	4.278.943	11.173.732	624.898	4.090.124	1.807.897	42,3%

Todas las muestras presentaron un elevado número de lecturas, la distribución del número de lecturas presentaba una distribución aproximadamente normal con una media en torno a los 8 millones (Fig. 6a). El grueso de las muestras (82 muestras) tenía un número de lecturas total entre 6 y 12 millones. Solamente 5 de las muestras contuvieron menos de 6 millones de lecturas, mientras que 9 muestras superaron los 12 millones. Concretamente, solo de una muestra (CO-84) se obtuvo más de 27 millones de lecturas, 20 millones más que el promedio.

Tras eliminar adaptadores, regiones de baja calidad y lecturas cortas, el número de lecturas disponibles se redujo a 785.521.864, perdiéndose aproximadamente el 4,5% del total. El promedio de lecturas por muestra fue de 8.094.107 (Tabla 2). CO-13 fue la muestra que presentó los valores de porcentaje de lecturas útiles más bajos, con un 87,7% (Fig. suplementaria 1). En el análisis de calidad realizado con FastQC se observó que esta muestra presentaba un alto número de secuencias duplicadas, y entre ellas, secuencias de adaptadores. También se observó una gran dispersión en la calidad por

base de las lecturas, sobre todo en el tercio final de la lectura. Por estos motivos es de esperar que en esta muestra el número de lecturas se redujese más que el resto. Por el contrario, CO-84 resultó ser la muestra que contaba con el mayor número de lecturas tanto crudas (27.822.108) como procesadas (26.450.360), aproximadamente unos 10 millones de lecturas más que la siguiente muestra (Fig. suplementaria 1).

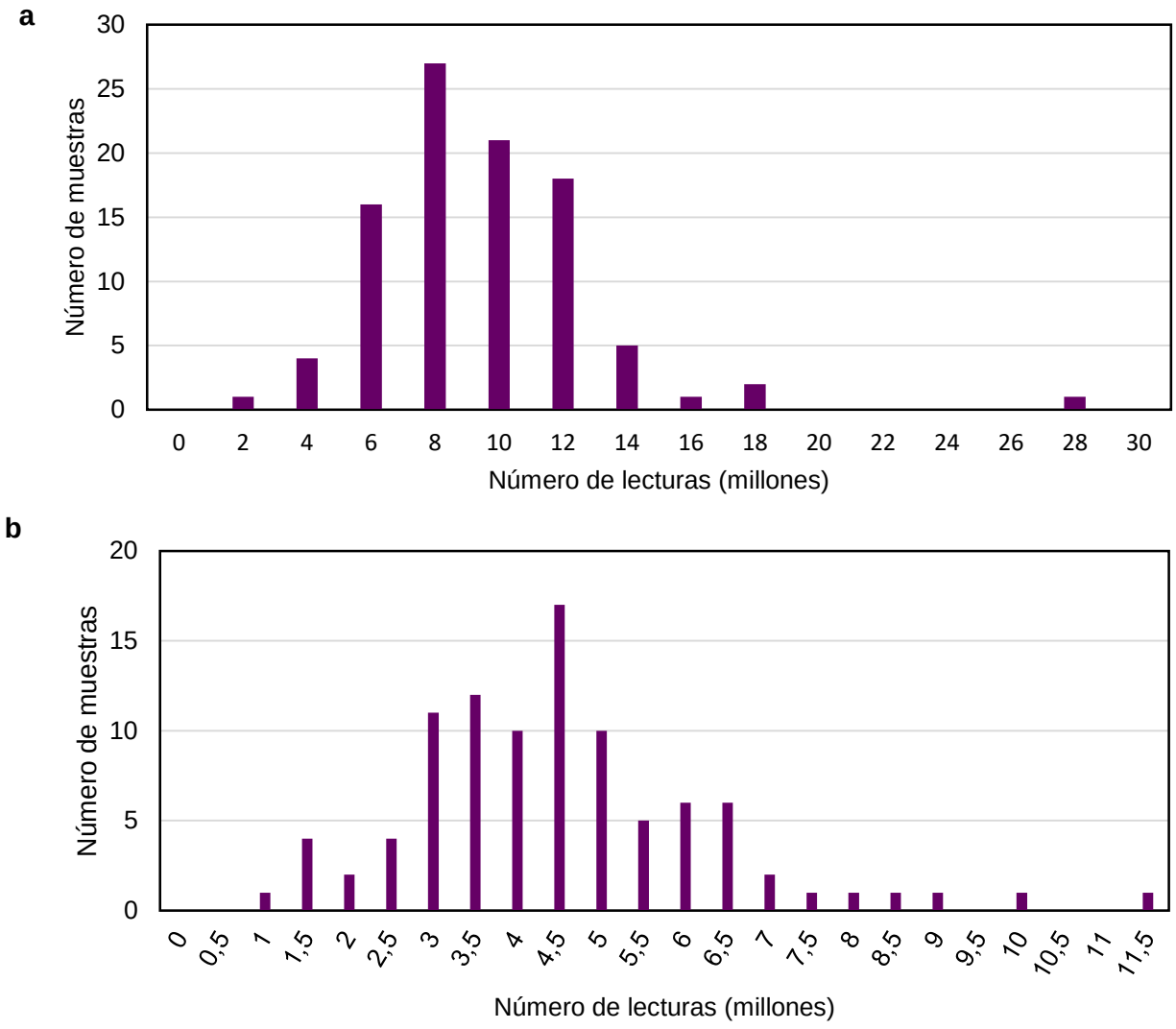

Figura 6.- Histogramas del número de muestras en función del número de lecturas a) En base a las lecturas crudas b) En base a las lecturas mapeadas de alta calidad.

Del total de lecturas procesadas, 598.290.404 lecturas (77,9%) pudieron ser alineadas contra el genoma de referencia. El número promedio de lecturas mapeadas por muestra fue de 6.232.191 (Tabla 2), donde los datos varían desde 15.727.450 (CO-48) hasta 1.746.738 (CO-54). CO-54 fue la muestra que presentó el mayor porcentaje de lecturas alineadas con un 97%, mientras que CO-65 fue la menor, con un porcentaje

de lecturas alineadas del 37%. La muestra CO-54 también fue la que perdió el menor número de lecturas durante el mapeo, algo menos de medio millón, por lo que a pesar de ser la muestra que menor número de lecturas en crudo obtuvo, prácticamente todas ellas fueron mapeadas. Por el contrario, la muestra CO-84 a pesar de ser la que mayor número de lecturas crudas y procesadas presentó, solo se mapeó un 46,4%, aunque el número de lecturas seguía siendo de los más altos (1.2927.264) (Fig. suplementaria 1). La muestra CO-84 junto a las muestras CO-83 y CO-92 se corresponden a distintas accesiones de *C. pedatifolia,* una de las especies dentro de este género más alejadas de *C. pepo*. Se observó que el porcentaje de mapeo que presentaron estas dos muestras (46% y 47%, respectivamente) es muy similar al observado en CO-84.

Tras realizar el filtrado de las lecturas mapeadas, se descartaron lecturas duplicadas y lecturas con baja calidad de mapeo. El número de lecturas descendió a 410.778.570, lo que indica que el 69,3% de las lecturas que se mapearon lo hicieron con una alta calidad. El valor medio de lecturas mapeadas con una alta calidad por muestra fue de 4.278.943, con valores que oscilaron entre 624.898 (CO-13) y 11.173.732 (CO-48) lecturas (Tabla 2). Más de la mitad de las muestras consiguió alinear entre 3 y 5 millones de lecturas con una alta calidad (Fig. 6b). Se observa que la distribución también se asemeja a una normal, centrándose la mayoría de los datos en torno a los 4,5 millones de lecturas.
Coincide también que CO-13 es la muestra que menor porcentaje de lecturas mapeadas con alta calidad presenta, con un 16%, mientras que CO-73 es la de mayor porcentaje (83,5%) (Fig. suplementaria 1). Como ya se mencionó anteriormente, en las estadísticas de calidad para la muestra CO-13 se observó un gran número de lecturas duplicadas. Una vez realizado el mapeo, el 41% de las lecturas alineadas resultaron ser lecturas duplicadas, por lo que es de esperar que al eliminarlos, el porcentaje de lecturas mapeadas se redujese drásticamente.

4.2. Búsqueda y anotación de SNPs

En una búsqueda de SNPs se deben identificar todas las diferencias entre el genoma de referencia y las lecturas almacenadas en archivos BAM para generar un archivo VCF (*Variant Call Format*) en el que se muestren estos cambios y sus posiciones. A este proceso de búsqueda de variantes se le conoce como *SNP-calling.* Una vez hecha la búsqueda, se obtuvieron un total de 2.717.825 SNVs. Este número de SNVs incluye tanto mutaciones como errores de lectura. Para reducir el número de posibles errores y retener aquellos cambios con mayor calidad, se aplicaron una serie

de filtros descritos *Materiales y métodos*. Como se puede observar en la tabla 3, el número de SNVs se fue reduciendo con cada filtro hasta obtener un total de SNVs 702.735 de alta calidad.

Tabla 3.- Número de SNVs identificados y tras la aplicación de los diferentes filtros de calidad.

	Número de SNVs
Sin filtrado	2.717.825
Eliminando genotipos con cobertura inferior a 5	2.717.825
Eliminando SNVs con alelos monomórficos	1.883.757
Eliminando datos faltantes	702.735

El número de SNVs no varía al eliminar los genotipos que no superaron un valor de cobertura de 5, ya que se puede dar el caso de que no se tenga suficiente información para determinar con precisión el genotipo de las muestras pero tomando toda la información en conjunto se pueda determinar que es variable. Filtrar por una frecuencia del alelo minoritario mayor del 0,005 nos permitió eliminar alelos monomórficos y conservar alelos muy raros, incluso si estos se detectan una única vez en heterocigosis en un único individuo. Tras la eliminación de estos SNVs, se perdió cerca de un millón de SNVs. Es posible que los SNVs que no poseen ningún dato de genotipo se sigan conservando, por lo que deben ser eliminados ya que carecen de interés. La mayor pérdida de SNVs se produjo al eliminar estos SNVs en los que el porcentaje de datos faltantes fuese del 100%, es decir, aquellos sitios donde no hubiese ningún dato de genotipo. Al final se obtuvo una colección de 702.735 SNVs, que fueron considerados de alta calidad para realizar los posteriores estudios.

Una vez que se identificaron todos los SNVs de alta calidad, se procedió a analizar su distribución a lo largo del genoma (Fig. 7). El número medio de SNVs por Mb fue de 3237 (des. est = 450) y fue similar para todos los cromosomas. Además, se vio una correlación positiva entre el tamaño del cromosoma con el número de SNVs (r = 0,92), siendo el cromosoma 1 el que mayor número de SNVs presentó. Sin embargo, existen cromosomas como el 6 y el 7 donde el número de SNVs encontrados fue bajo en comparación con su tamaño. En cuanto a la distribución de los SNVs a lo largo de los cromosomas, se puede observar que los SNVs no se distribuyen homogéneamente, sino que se pueden distinguir regiones con mayor densidad y regiones con una densidad menor. Las zonas en las cuales se observa una baja densidad de SNVs pueden corresponderse con regiones de heterocromatina, donde la actividad transcripcional es escasa o nula, como por ejemplo las regiones centroméricas. Por el contrario, zonas de

mayor densidad de SNVs corresponden con regiones con alta actividad transcripcional. La distribución de SNVs dentro de los cromosomas es muy similar, observándose en todos ellos regiones con alta y baja densidad de SNVs.

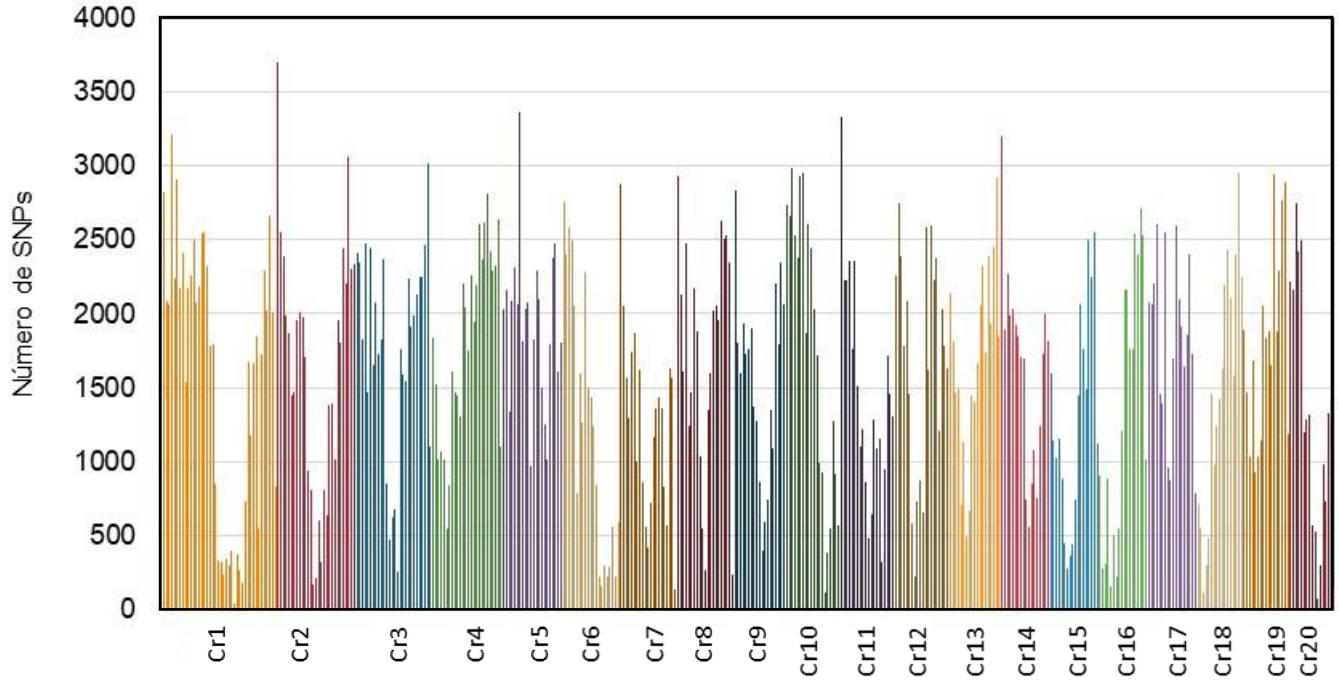

Figura 7.- Distribución de SNVs a lo largo del genoma. Los SNVs están agrupados en ventanas de 500Kb. La densidad de SNVs para cada cromosoma (Cr) se ha indicado de colores diferentes.

Del total de genotipos obtenidos (Fig. 8), el porcentaje promedio de heterocigotos fue del 1,5% (Fig. 8a) y prácticamente la totalidad de las muestras presentaban porcentajes de heterocigosidad inferiores al 5%. Solamente dos de las muestras superaron el 5% de los heterocigotos y más concretamente, solo una de ellas llegó al 8%. Cabe destacar que de los dos híbridos presentes en este estudio, ninguno de ellos presentó un alto porcentaje de heterocigosidad: el híbrido *C. foetidissima x C. scabridioflia* tuvo un 3%, mientras que el híbrido *C. pedatifolia x C. foetidissima* apenas llegó al 2% de genotipos heterocigotos. Por el contrario, el porcentaje medio de genotipos homocigotos para el alelo de referencia (Fig. 8b) se situó en torno al 84%, mientras que el porcentaje promedio de homocigotos para el alelo alternativo, fue de un 14%. Para un número alto de muestras la práctica totalidad de sus genotipos son homocigotos para el alelo de referencia. En la distribución para el alelo alternativo, se observan dos modas, una en torno al 5% de genotipos homocigotos para el alelo alternativo y otra en torno al 20%. Como cabría esperar, las muestras de la especie *C*

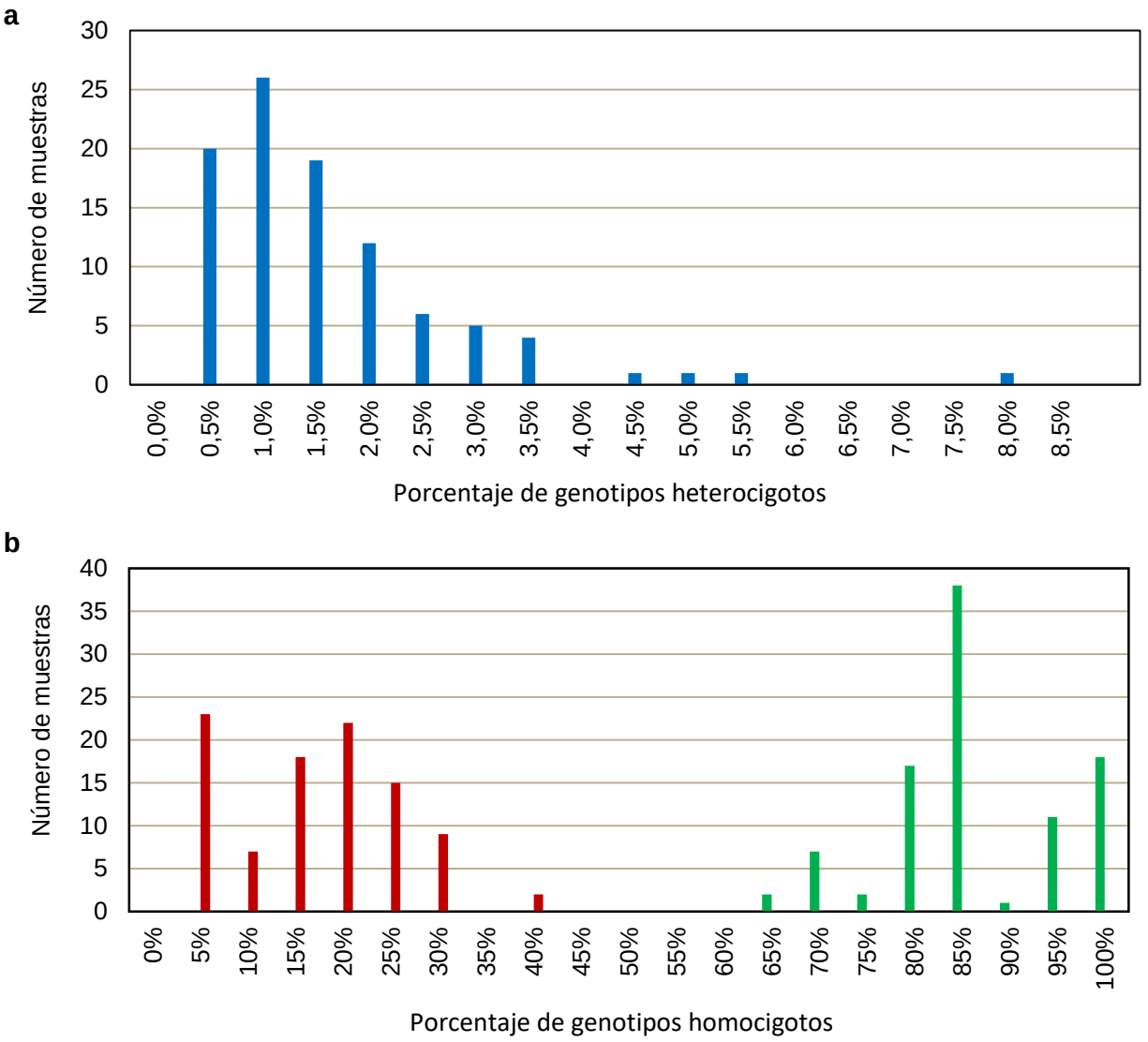

Figura 8.- Histograma del número de muestras en función del porcentaje de genotipos a) heterocigotos o b) homocigotos de referencia (verde) y homocigotos alternativos (rojo).

. *pepo* fueron las que presentaron los mayores porcentajes de homocigosidad para el alelo de referencia (> 94%), mientras que los porcentajes más bajos variaban entre el 62-70% y correspondían con las especies *C. cordata, C. pedatifolia* y *C. foetiisima* (Fig. suplementaria 2). Se observó una correlación entre la proximidad filogenética y el porcentaje de genotipos homocigotos para el alelo de referencia, de tal forma que cuanto mayor es la cercanía entre la muestra y el genoma de referencia (*C. pepo*), mayor es este porcentaje (Fig. suplementaria 2). El efecto contrario se observó al analizar el porcentaje de genotipos homocigotos para el alelo alternativo, donde las especies más

alejadas de *C. pepo* presentaron los mayores porcentajes, y ninguna muestra de *C. pepo* superó el 8% (Fig. suplementaria 2).

Dada la diversidad de especies muestreada, especialmente al haber incluido especies alejadas filogenéticamente entre sí y del genoma de referencia (*C. pepo*), es de esperar que el número de SNVs detectados sea amplio, sin embargo su número será mucho menor si consideramos especies concretas. Por tanto, se decidió analizar, además de la colección de SNVs global, la variabilidad presente únicamente en las tres especies de *Cucurbita* de mayor importancia económica e interés comercial *(C. maxima, C. moschata* y C. *pepo*). Los SNVs se clasificaron en SNPs, MNP: polimorfismos de múltiples nucleótidos, Inserciones y Deleciones utilizando *SnpEff.* Para el conjunto de SNVs de alta calidad de la población de especies de *Cucurbita* se obtuvo un total de 779.979 variantes, de las cuales, un 91% resultaron ser SNPs (Tabla 4). La diferencia entre el número de polimorfismos totales obtenidos con *SnpEff* y variantes incluidas en el fichero VCF (702.735) se debe a que en el formato VCF una única variante puede ser compuesta (e.g., una variante que suponga dos sustituciones nucleotídicas) mientras que el *SnpEff* toma en consideración cada cambio individualmente. Además, una variante puede tener más de una anotación debido a múltiples transcripciones (isoformas) de un gen o debido a genes superpuestos en la ubicación genómica de la variante. Se observó que el número de variantes decreció (entorno a un 90% menos) en los tres grupos de especies cultivadas en comparación con la población principal de cucurbitas (Tabla 4). Sin embargo *C. pepo* y *C. moschata* presentan un 31% y un 47% de polimorfismos más que *C. maxima*. Esta diferencia podría deberse a que el número de variantes encontradas depende del número de individuos presentes en cada análisis; cuanto mayor sea el número de individuos con el que se trabaje, mayores serán las probabilidades de encontrar SNPs en una población. En *C. pepo* y *C. moschata* se han analizado 27 individuos en cada una, mientras que *C. maxima* cuenta con la mitad de individuos (13 individuos). No obstante, los porcentajes para los distintos tipos de variantes fueron muy similares entre los tres grupos de especies de interés comercial (Tabla 4). En todos los grupos, la presencia de SNPs fue mucho mayor que la de cualquier otro tipo de polimorfismo (inserción, deleción, etc). En comparación con la población global, se encontró un porcentaje de SNPs mayor y de MNP menor en las tres especies. Para el resto de tipos de polimorfismos los porcentajes encontrados fueron muy similares en los cuatro grupos.

Tabla 4.- Número total de polimorfismos y porcentaje de los diferentes tipos de polimorfismos encontrados en la población de *Cucurbita sp*. y en las especies cultivadas *C. maxima*, *C. moschata* y *C. pepo*. N: número de muestras en cada grupo.

	Cucurbita sp. **N = 96**	***C. maxima*** **N = 13**	***C. moschata*** **N = 27**	***C. pepo*** **N = 27**
Polimorfismos totales	779.979	55.123	104.765	80.987
SNP	91,5%	97,2%	96,3%	96,3%
MNP	6,8%	1,3%	2,0%	1,6%
INS	0,8%	0,7%	0,8%	0,9%
DEL	0,9%	0,7%	0,9%	1,2%
MIXED	0,1%	0,0%	0,1%	0,1%

*SNPs, polimorfismos de nucleótido simple ('A'='T'); MNP, polimorfismos de múltiples nucleótidos ('ATA'= 'GTC'); INS, inserciones ('A' = 'AT'); DEL, deleciones ('AC'= 'C'); MIXED, combinación de estos polimorfismos ('ATA'= 'GTCAGT').

El posible efecto fenotípico de los SNVs fue predicho mediante el programa *SnpEff.* Este programa, basándose en la anotación de los genes y su estructura, clasifica los SNVs por tipo y magnitud de su efecto. Con respecto al número y tipo de polimorfismos en regiones génicas (Tabla 5), se encontró que de los 34.230 genes anotados y descritos en *C.* pepo, se identificó al menos un polimorfismo en 24.674 de ellos, siendo 32 el promedio de variantes encontradas por gen. Algo más bajo fue el número de genes con polimorfismos que se identificó en las poblaciones de *C. pepo, C. maxima* y *C. moschata*. Sin embargo, sí que fue bastante menor el número promedio de polimorfismos que se encontraron por gen.

Tabla 5.- Número y porcentaje de polimorfismos dentro de genes para la población de *Cucurbita sp*. y en las especies cultivadas *C. maxima*, *C. moschata* y *C. pepo*. N: número de muestras en cada grupo.

	Cucurbita sp. **N = 96**	***C. maxima*** **N = 13**	***C. moschata*** **N = 27**	***C. pepo*** **N = 27**
Nº de polimorfismos totales	779.979	55.123	104.765	80.987
Nº de genes con polimorfismo	24.674	20.950	22.707	22.297
Porcentaje de genes con polimorfismos	72%	61%	66%	65%
Nº promedio de polimorfismos por gen	32	3	5	4

SnpEff también es capaz de predecir el impacto que puede ocasionar cada SNVs. Para ello este *software* hace una estimación del impacto que produce el polimorfismo y lo clasifica en diferentes categorías. Se considera que una variante es la causante de un impacto grave cuando ésta es la responsable de cambios estructurales drásticos en la

proteína (e.g., la ganancia de un codón stop o a la pérdida de un codón de inicio de traducción). Los cambios como la sustitución de un aminoácido por otro, que no afectan drásticamente a la función de la proteína, aunque pueden afectar a su eficacia, se clasifican como efectos moderados. Efectos leves son aquellos que son inocuos o que muy improbablemente cambien el comportamiento de la proteína. Por ejemplo, un cambio sinónimo. Por último, se clasifican como polimorfismos responsables de provocan un efecto modificador a todos los cambios, tanto en regiones codificantes como en no codificantes, donde las predicciones son complejas o donde no hay evidencia de impacto, como por ejemplo, cambios en regiones intrónicas o en regiones UTR. Hay que notar que los diferentes alelos para un mismo SNVs pueden ocasionar efectos distintos, por lo que el número de efectos registrados no tiene por qué coincidir con el número de SNVs.

La gran mayoría de los efectos encontrados fue de tipo modificador, mientras que los efectos graves no pasaron del 1,2% para ningún grupo de los considerados (Tabla 6). Estos resultados se observaron tanto en la población principal como en los tres grupos de especies, aunque como era de esperar, el número de efectos encontrados fue mucho mayor en la población principal que en el resto de grupos.

Tabla 6.- Número total de impactos registrados y porcentaje de los diferentes tipos de impactos que se han encontrado en la población de *Cucurbita sp* . y en las especies cultivadas *C. maxima*, *C. moschata* y *C. pepo*. N: número de muestras en cada grupo.

	Cucurbita sp. **N = 96**	***C. maxima*** **N = 13**	***C. moschata*** **N = 27**	***C. pepo*** **N = 27**
Grave	0,3%	0,5%	1,1%	0,7%
Moderados	12,7%	1,5%	16,2%	13,1%
Leves	21,4%	24,6%	31,9%	20,1%
Modificador	65,6%	73,5%	50,8%	66,2%
Número de impactos*	1.862.178	115.144	101.429	194.238

*Suma de todos los impactos identificados en el estudio.

Con respecto a su distribución en genes (Tabla 7), en la población principal se identificaron un total de 3.833 genes (16%) con al menos un polimorfismo causante de un efecto grave. Este porcentaje es superior al del resto de grupos debido probablemente al alto número de individuos que lo conforman y la diversidad entre ellos. Se puede observar que en la población global casi todos los genes contenían al menos un SNVs que ocasionaba un efecto de tipo modificador. En el resto de grupos, el porcentaje fue algo inferior, donde el porcentaje de genes con un al menos un SNVs que ocasione un efecto de tipo modificador fue de entre el 85% y el 89%.

Tabla 7.- Número de genes que presentan al menos un polimorfismo causante de un impacto grave, moderado, leve o de tipo modificador. Se representan los valores y porcentaje de los diferentes efectos encontrados en la población de *Cucurbita sp*., *C. maxima*, *C. moschata* y *C. pepo*. N: número de muestras en cada grupo.

	Cucurbita sp. **N = 96**		***C. maxima*** **N = 13**		***C. moschata*** **N = 27**		***C. pepo*** **N = 27**	
Graves	3.833	16%	519	2%	1.004	4%	1.189	5%
Moderados	14.382	58%	7.719	37%	10.219	45%	9.516	43%
Leves	14.806	60%	9.408	45%	11.378	50%	10.678	48%
Modificadores	23.115	94%	17.876	85%	20.175	89%	19.725	88%
Nº genes con polimorfismos*	24.674		20.950		22.707		22.297	

*Número de genes en los que al menos se ha identificado un polimorfismo.

Con respecto a la localización de los cambios, se observó que en la población global un 78% de los cambios implicaban directamente a la secuencia de la proteína. Porcentajes algo superiores, en torno al 80%, se observaron en el los otros tres grupos. El resto de polimorfismos se encontrarían en regiones UTR. Dentro de población principal se observó que un 64% de las variantes producían cambios silenciosos. El resto de variantes encontradas ocasionaban cambios "con sentido" que daban lugar a la sustitución de un aminoácido por otro. Por último mencionar que se observó un porcentaje de polimorfismos inferior al 0,5% que eran responsables de la aparición de codones parada prematuros (Tabla 8).

Tabla 8.- Número de polimorfismos y porcentaje de las diferentes clases de efectos que modifican directamente la secuencia de la proteína encontrados en la población de *Cucurbita sp*. y en las especies cultivadas *C. maxima*, *C. moschata* y *C. pepo*. N: número de muestras en cada grupo.

	Cucurbita sp **N = 96**	***C. maxima*** **N = 13**	***C. moschata*** **N = 27**	***C. pepo*** **N = 27**
"Con sentido"	35,6%	37,1%	38,1%	39,0%
"Sin sentido"	0,4%	0,6%	0,7%	1,1%
"Silencioso"	64,0%	62,3%	61,2%	59,9%
Nº de polimorfismos	585.824	44.221	81.635	63.143

*Número de polimorfismos que se ha identificado como responsables de un cambio en la secuencia de la proteína.

4.3. Búsqueda de genes candidatos asociados a QTLs

Uno de los objetivos del presente trabajo es mostrar la utilidad de crear una base de datos anotada de diversidad genética del género *Cucurbita* para la localización de genes de interés. A continuación se muestran dos posibles aplicaciones de la base de datos desarrollada para el estudio de genes de interés agronómico: 1) determinar el

efecto de los cambios observados cuando se dispone de una lista de genes candidatos para un QTL y 2) orientar la búsqueda de genes de interés cuando no se dispone de ningún gen candidato para un QTL.

A partir de ocho de los QTLs implicados en floración, morfología y color del fruto de plantas de *C. pepo* descritos por Montero-Pau et al. (Montero-Pau, Blanca, Esteras, et al., 2017) (tabla suplementaria 2), se exploró el número y tipo de cambios que afectaban a los genes incluidos en dichas regiones. En la tabla 9 se muestra el número de genes con SNVs y el número de genes con cambios graves y moderados. Solamente los SNVs que ocasionen un efecto de estas magnitudes son capaces de producir un cambio en la secuencia de la proteína que altere su estructura y pueda modificar su función. Por este motivo se obvió la información sobre los SNVs que ocasionaban un impacto leve o de tipo modificador. Como era de esperar, no todos los genes presentes en la región QTL presentaron polimorfismos. Por lo general, el número de genes en los que se identificó al menos un SNVs responsable de un efecto grave, fue bajo, aunque muy variable entre QTLs, no superando el tercio en ningún caso. Por el contrario, prácticamente la totalidad de los genes en los que se identificó un SNVs, presentaban al menos un SNVs responsable de un efecto de tipo moderado. Este tipo de aproximación, permite reducir al número de genes potenciales a aquellos que presentan cambios sustanciales.

Tabla 9.- Número de genes con SNVs para cada QTL. Se muestra el nombre del QTL, el número de genes dentro de cada región, número de genes en los que hay al menos un SNVs, número de genes en los que se han encontrado al menos un cambio responsable de un impacto grave y moderado.

QTL	Nº genes	Nº genes con SNVs	Nº de genes con impactos graves	Nº genes con impactos moderados
Li_10	683	14	5	13
DFeF_9	2050	46	10	42
DFeF_12	2556	52	9	51
MaRCo_4	1910	34	11	33
MLRCo_4	805	25	3	20
MbRCo_19	2928	65	17	64
IFSh_3	2456	50	6	49
IbRCo_4	27285	566	138	534
Sl_12	1118	24	2	24

Para aquellos QTLs en los que se había descrito un gen candidato (Tabla 10), se decidió identificar y analizar los polimorfismos existentes en dichos genes. En primer lugar se determinó si el gen se había secuenciado. Se debe tener en cuenta de que esta herramienta ha sido generada con información procedente de transcriptoma de hoja, por lo que solo tendremos información de aquellos genes que se estén expresando en el momento de recogida de la muestra.

Cuando se encontraron transcritos, se procedió a identificar los polimorfismos en los genes candidatos. Se ha analizado el contenido de polimorfismos de dichos genes y se cuantificó el número de SNVs responsables de provocar un impacto alto y moderado. Hemos observado que el número de SNVs que ocasionan un efecto grave en el gen es muy reducido, casi inexistente, mientras que el número de SNVs responsables de provocar un impacto moderado es mayor. Además, se identificó las especies en las que se encuentran dichos polimorfismos. Por ejemplo, para para la *fitoeno sintasa* presente en el QTL *MbRCo_19* localizado en el cromosoma 14 de *C. pepo* y que está asociado al parámetro b de color de las rayas del fruto maduro. Este gen está implicado en la biosíntesis de carotenoides y ha sido propuesto como uno de los genes responsable de la pigmentación de la piel y de la carne en frutos maduros (Harry S Paris & Padley, 2014). Este gen (Fig. 9) presenta tanto SNVs responsables de provocar impactos graves como moderados y no presenta un número excesivo de polimorfismos, motivos por los que se ha elegido este gen para explorar la diversidad presente en esta población.

Se ha identificado un solo SNV responsable de ocasionar un impacto grave y 10 responsables de ocasionar un impacto moderado (tabla 11). El SNVs responsable de ocasionar un impacto grave detectado se traduce en la aparición de un codón de parada prematuro y solo se ha identificado en *C. pepo* subsp. *ovifera* (CO-58). Desafortunadamente, no disponemos de la información fenotípica para esta entrada,

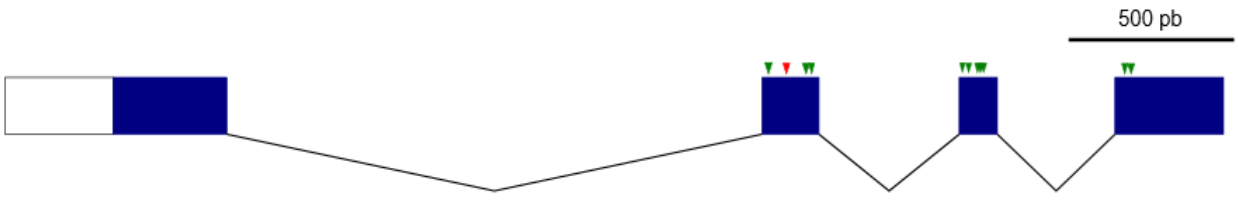

Figura 9.- Representación esquemática de la estructura del gen de la *Fitoeno sintasa* (Cp4.1LG14g03180) con indicación de las posiciones de lo polimorfismos. En rojo se representan los SNPs responsables de ocasionar un impacto grave y en verde los responsables de causar un impacto moderado. Los exones se representan como rectángulos, siendo la parte azul la región codificante.

Tabla 10.- Número de SNVs que se encuentran dentro de genes candidatos para cada QTL. Se muestra el nombre del QTL, nombre de los genes candidatos propuestos para cada QTL, identificador del gen candidato, presencia o ausencia de lecturas mapeadas para cada gen candidato y número de SNVs responsables de ocasionar un impacto grave y moderado.

QTL	Gen candidato	ID-gene	Existencia de transcriptos	Nº impactos graves	Nº impactos moderados
Li_10	*Homeobox-leucine zipper*	Cp4.1LG10g04300	No	-	-
	Homeobox-leucine ziper	Cp4.1LG10g04310	No	-	-
DFeF_9	*WUSCHEL related homeobox 8*	Cp4.1LG19g06930.1	No	-	-
	Auxin response factor 4	Cp4.1LG19g06940.1	Sí	0	32
	Ethylene-responsive transcription factor 4-like	Cp4.1LG19g07200	Sí	0	12
DFeF_12	*MADS-box factor 6*	Cp4.1LG17g02480.1	No	-	-
	MADS-box factor 50	Cp4.1LG17g02550	Sí	0	23
	TCP family transcription facto	Cp4.1LG17g02400.1	Sí	0	27
MaRCo_4	*Zeaxanthin epoxidase*	Cp4.1LG05g05530	Sí	0	24
MLRCo_4	*APRR2-like (ARABIDPOSIS PSEUDO RESPONSE REGULATOR2-LIKE*	Cp4.1LG05g02060	No	-	-
	APRR2-like (ARABIDPOSIS PSEUDO RESPONSE REGULATOR2-LIKE	Cp4.1LG05g02070	No	-	-
MbRCo_19	*Phytoene synthase*	Cp4.1LG14g03180	Sí	1	11
IFSh_3	*Ovate family protein 4*	Cp4.1LG03g03420	No	-	-
IbRCo_4	*F-box/kelch-repeat protein*	Cp4.1LG05g05320	Sí	2	20
	F-box/kelch-repeat protein	Cp4.1LG05g05330	Sí	0	1

pero sería muy interesante poder comprobar el efecto en el rasgo fenotípico de dicho cambio. Todos los SNVs responsables de ocasionar un impacto moderado consistieron en la sustitución de un aminoácido por otro. Se han detectado cuatro SNVs con más de un alelo alternativo. Para 3 de los SNVs, el segundo alelo alternativo daba lugar a un cambio "silencioso", de tal forma que el cambio de nucleótido no producía un cambio de aminoácido. En la búsqueda de SNPs responsables de ocasionar un cambio fenotípico, estos alelos carecerán de importancia debido a que no causan un efecto en la secuencia de la proteína. Para el último SNVs trialélico que se encontró, ambos alelos alternativos ocasionaban el mismo cambio de aminoácido, por lo que si estamos estudiando un cambio en el fenotipo ambos alelos serán de interés. Haciendo un breve análisis de las especies en las que se han hallado estos polimorfismos, se puede ver que estos SNPs han sido identificados en las principales especies comerciales, siendo *C. maxima* la especie que cuenta con mayor número de muestras en las que se han identificado SNVs. Una vez que se ha descrito la diversidad presente en esta población de cucurbitas para el gen *fitoeno sintasa* sería de gran interés analizar más en profundidad aquellos individuos en los que se han identificado polimorfismos con el fin de hallar una correlación entre el fenotipo y el genotipo.

Como hemos comentado anteriormente, esta herramienta también puede ser de utilidad para hacer una primera aproximación de los posibles genes que pueden ser los responsables de ocasionar un fenotipo determinado. Utilizando la información existente en una determinada región candidata podemos seleccionar aquellos genes que presenten polimorfismos con impacto grave y, o bien, buscar evidencia bibliográfica de su implicación en dicho fenotipo, como por ejemplo búsqueda de genes homólogos de otras especies en los que ya se ha descrito el efecto que se está estudiando. O bien, intentar buscar individuos con el mismo cambio y estudiar el fenotipo. A modo de ejemplo de esta segunda aplicación, se analizaron los polimorfismos existentes dentro de la región *Sl_12* relacionada con *silver leaf* en el cromosoma 17 para la que no se han propuesto genes candidatos. En esta región se detectaron 24 genes del total de 1.118 genes que hay en dicha región (tabla suplementaria 3). De los genes en los que se ha obtenido información, solo dos presentan cambios que producen un efecto grave (tabla 12). El primer gen en el que se ha identificado un SNPs que ocasiona un impacto grave es el gen Cp4.1LG17g08860, un gen de la familia de proteínas *Phox* (PX). El dominio PX es un dominio estructural de unión a fosfoinosítidos que participa en tráfico las proteínas hacia las membranas celulares (Ellson, Andrews, Stephens, & Hawkins, 2002).

Tabla 11.- Descripción de los polimorfismos encontrados en el gen de la *fitoeno sintasa*. Se muestra la posición del polimorfismo, el cambio nucleótido, el efecto producido en la secuencia codificante y el correspondiente cambio en la secuencia proteica y el impacto ocasionado, genotipos que poseen el primer alelo alternativo en heterocigosis, genotipos que poseen el primer alelo alternativo en homocigosis, genotipos que poseen el segundo alelo alternativo en heterocigosis, genotipos que poseen el segundo alelo alternativo en homocigosis. N: número de individuos.

Posición	SNPs	Tipo de cambio	Impacto	Genotipos heterocigotos (0/1)	Genotipo homocigoto (1/1)	Genotipos heterocigotos (0/2)	Genotipo homocigoto (2/2)
2474036	A → G	"Con sentido" (Glu122Gly)	Moderado	*C. pepo pepo (N=1)*	-	-	-
2474091	G → A	Ganancia de codón de parada	Grave	*C. pepo ovifera (N=1)*	-	-	-
2474155	G → A	"Con sentido" (Asp162Asn)	Moderado	-	*C. ecuadorensis (N=1)* *C. moschata (N=1)*	-	-
2474158	A → T	"Con sentido" (Thr163Ser)	Moderado	*C. pepo pepo (N=1)*	-	-	-
2474629	G → C	"Con sentido" (Glu179Gln)	Moderado	*C. pepo pepo (N=1)*	-	-	-
2474647	T → A,C	"Con sentido"/ "Silencioso" (Leu185Met/Leu185Leu)	Moderado/Leve	*C. maxima (N=1)*	*C. maxima (N=6)*	-	*C. ficifolia (N=1)*
2474665	G → A	"con sentido" (Glu191Lys)	Moderado	*C. moschata (N=2)* *C. maxima (N=2)*	*C. moschata (N=2)*	-	-
2474676	C → T,A	"Con sentido"/ "Silencioso" (Asp194Glu/Asp194Asp)	Moderado/Leve	*C. pedatifolia (N=1)*	-	*C. foetidissima (N=1)*	-
2474685	CC → TC, TG	"Con sentido"/ "Silencioso" (Leu198Val / Tyr197Tyr)	Moderado/Leve	-	*C. foetidissima (N=2)* *C. pedatifolia (N=1)*	*C. pedatifolia (N=1)*	-
2475112	C →A	"con sentido" (Asp221Glu)	Moderado	-	*C. pepo pepo (N=1)*	-	-
2475126	AGAAA → GGAAA,GGAAG	"Con sentido"/"Con sentido" (Glu226Gly/Glu226Gly)	Moderado/Moderado	-	*C. moschata (N=14)* *C. maxima (N=13)* *C. argyrosperma (N=3)* *C. pepo orkazana (N=1)* *C. pepo pepo (N=3)*	-	*C. scabridifolia (N=1)* *C. foetidissima (N=1)* *C. pedatifolia (N=2)*

El polimorfismo ha sido identificado en *C. ecuadorensis* (CO-74) y el efecto se traduce en la aparición de un codón de parada prematuro. El segundo gen en el que se ha identificado un SNPs que provoque un impacto grave es en Cp4.1LG17g08790, un gen de la familia de los receptores de kinasa. En *Arabidopsis*, se ha observado que genes pertenecientes a esta familia están involucrados en vías de respuesta hormonal, diferenciación celular, crecimiento y desarrollo de plantas, autoincompatibilidad y reconocimiento de simbiontes y patógenos (Morris & Walker, 2003). En este gen se ha identificado un SNPs en un individuo de *C.maxima* (CO-33), el cual da lugar a un cambio en la pauta de lectura.

Tabla 12.- Se indican los genes que contienen al menos un cambio responsable de un efecto grave en la región *Sl_12*. Se indica su posición, el polimorfismo y el efecto que produce el cambio y en los genotipos en los que se han encontrado los polimorfismos. N: número de individuos.

ID gen	Posición	SNPs	Tipo de cambio	Genotipos heterocigotos (0/1)	Genotipo homocigotos (1/1)
Cp4.1LG17g08860	5066347	T → G	Ganancia de codón de parada	-	*C. ecuadorensis (N=1)*
Cp4.1LG17g08790	5145930	CTT→CTTT	Cambio en la pauta de lectura	-	*C. maxima (N=1)*

5. Discusión

Los cultivares del género *Cucurbita* tienen una gran importancia económica y su consumo va en aumento. La gran diversidad morfológica que presenta este género, lo hace interesante para el desarrollo de nuevas variedades. Sin embargo, a pesar del potencial de mejora, los programas de mejora desarrollados para estos cultivos han sido menos numerosos en comparación con los llevados a cabo para otras hortalizas. Es más, hasta hace relativamente poco tiempo no se disponían de apenas recursos genéticos y genómicos. Hoy en día ya se cuentan con mapas genéticos, plataformas de TILLING y EcoTILLING y colecciones de SNPs. Además, ya se dispone del genoma completo de las principales especies cultivadas. Actualmente, las nuevas técnicas NGS junto con el análisis bioinformático están haciendo posible el análisis detallado de la diversidad genética existente en las poblaciones vegetales, incluyendo tanto especies domesticadas como silvestres, lo cual es un paso esencial para el descubrimiento de nuevos genes de interés con el fin de introducirlos en futuros planes de mejora. En el presente trabajo se ha analizado la diversidad existente en una población muy diversa de *Cucurbita* a partir de transcriptomas de 96 muestras lo que ha permitido obtener una colección de SNPs de alta calidad y anotados funcionalmente, que puede ser usado para el desarrollo de marcadores y la detección de genes de interés.

Durante el proceso de limpieza y mapeado de las secuencias cruda, se pudo observar como el porcentaje de lecturas eliminadas tras el procesado inicial de limpieza de calidad fue muy similar en todas las muestras, mientras que el porcentaje de mapeo varió considerablemente. Durante el proceso de secuenciación pueden ocurrir errores que afectaran indistintamente a cualquier muestra. Por ello se espera que los posibles errores ocurridos durante la secuenciación se repartan de forma homogénea por todas las lecturas, ya que es algo intrínseco del propio proceso y no está condicionado por el origen de la muestra. Por el contrario la pérdida de lecturas en el mapeo sí que está relacionada con el origen de la muestra. Como era de esperar las muestras que correspondían con transcriptomas de *C. pepo* fueron las que tuvieron porcentajes de mapeo más altos. Las especies más próximas filogenéticamente (*C. okeechobeensis* y *C. lundelliana*) también tuvieron porcentajes altos de mapeo, situándose la mayoría de estas muestras por encima de la media (77%). Las muestras correspondientes con especies más alejadas a *C. pepo* como *C. ficifolia, C. cordata, C. foetidissima o C. pedatifolia* fueron las que menores porcentajes de mapeo registraron. El resto de

especies tienen porcentajes intermedios, siendo más frecuente encontrar porcentajes más altos en muestras de *C. argyrosperma* y *C. moschata* (más próximas filogenéticamente), que de *C. ecuadorensis* y *C. maxima*. Por lo tanto, un factor a tener en cuenta a la hora de obtener una colección de polimorfismos es el genoma de referencia utilizado para ello, ya que las diferencias entre dicho genoma y las muestras a analizar condicionarán el número de SNPs encontrados.

En un principio se espera que, cuanto mayor sean las diferencias entre los genomas comparados, mayor será el número de polimorfismos encontrados. Sin embargo, las diferencias entre genomas pueden llegar al punto en el que existan segmentos cromosómicos relativamente grandes presentes en alguna especie y no en el resto, y viceversa. La pérdida de gran cantidad de lecturas en el mapeo, puede ocasionar un descenso importante en el número de SNPs que pueden llegar a ser identificados. En especies silvestres como *C. cordata, C. foetidissima o C. pedatifolia* en la que además, se dispone de muy pocos individuos en este estudio, este hecho supone una limitación para la búsqueda de polimorfismos en regiones de interés. En estos casos nos será imposible discriminar si la falta de SNPs en una determinada región es debida a la ausencia de la región de interés o si se debe a un problema de mapeo. Al mismo tiempo, el bajo número de individuos analizados hace imposible saber si un SNPs es propio de especie o si es SNP que solo se encuentra en un determinado individuo. Todo esto supone una limitación a la hora de buscar polimorfismos en regiones de interés en estas especies. Sin embargo, el germoplasma silvestre es fuente de gran variedad genética, por lo que se deben elaborar estrategias complementarias para aprovechar al máximo recursos genéticos que nos proporcionan las especies silvestres.

A pesar de las limitaciones citadas, las especies silvestres aportaron gran parte de los SNPs identificados. El alto número de SNPs encontrados es debido principalmente al gran número de especies utilizadas en este análisis, más que en el número de individuos en sí. De hecho, en este estudio se ha podido ver como la contribución de polimorfismos en conjunto de las tres principales especies cultivadas, un total de 67 individuos, suponían solamente un tercio del total de polimorfismos encontrados mientras que el resto de SNVs fueron aportados por los individuos de las 9 especies restantes (29 individuos en total).

Otro hecho a tener en cuenta a la hora de obtener colecciones de SNPs es la estructura poblacional a partir de la cual se han extraído. Los SNPs son una

representación de la variabilidad presente en dicho grupo y su número puede variar en función del número de individuos y de especies implicadas. Por otro lado, también se debe tener en cuenta que los transcriptomas con los que se trabajó correspondían a un único individuo de cada una de las 96 accesiones. Las especies del género *Cucurbita* son alógamas, por lo que una "accesión" (muestra de semillas mantenida en un banco de germoplasma procedente de una determinada colecta) no tiene porqué ser una colección de individuos homocigotos. Según Lewontin (Lewontin, 1979), para una correcta evaluación de la variabilidad genética existente en una determinada población, el número de individuos estudiados debe de ser suficiente y representativo de la población natural de la que procedan.

Un hecho bastante llamativo es la baja heterocigosidad que ha sido detectada. Teniendo en cuenta lo diversa que es la población de estudio y más aún, tratándose de especies que son alógamas, se obtienen porcentajes muy altos homocigosidad. Sin embargo, las entradas utilizadas en este estudio provienen de diferentes de bancos de germoplasma, los cuales perpetúan sus semillas por medio de la autofecundación. Esto, junto al hecho de que se obtienen pocos frutos por planta, ha propiciado que la diversidad presente en los bancos de germoplasma se vaya reduciendo progresivamente. Hasta hace aproximadamente 90 años, los cultivares de *Cucurbita* se caracterizaban por una alta variabilidad genética atribuida en parte a la tendencia a cruzar. Sin embargo la demanda de uniformidad dio como resultado un aumento en la homocigosis. De este modo se desarrollaron líneas puras, las cuales se han usado para desarrollar híbridos, que eran más uniformes y homogéneos que los cultivares polinizados abiertos anteriores (Paris, 1989).

Tras la anotación de SNVs, se observó que el porcentaje de polimorfismos responsables de ocasionar un efecto grave fue muy bajo (aprox. 0,3%). Este tipo de mutaciones pueden resultar trascendentales en la supervivencia de los individuos, motivo por el cual aparecen en tan bajo porcentaje. En otros trabajos en los cuales se ha realizado la anotación empleando *SnpEff* se han observado porcentajes similares: 0,1% en tomate (Causse et al., 2013) y 0,046% en melocotón (Martínez-García et al., 2013). No obstante, el número de genes con efectos graves es menor al observado en el presente estudio posiblemente porque ambos estudios usaron un número menor de individuos y con muestras filogenéticamente más próximas. Algo mayores fueron los porcentajes de cambios que producen un efecto moderado (12,7%) y bajo (21%), mientras que en el estudio realizado en tomate los porcentajes de estos cambios fueron

bastante menores (1,5% en moderados y 1,3% bajos). Los cambios de tipo modificador fueron los más comunes en los estudios en tomate y melocotón (98% y 86% respectivamente) (Causse et al., 2013; Martínez-García et al., 2013), sin embargo, en *Cucurbita* fue menor (65%). Esto puede ser debido a que como efecto modificador se catalogan aquellos SNPs que se encuentran en regiones intergénicas o en regiones intrónicas. Recordamos que en dichos estudios se trabajó con DNA genómico, motivo por el cual han obtenido porcentajes tan altos.

Con respecto a la región codificante, en *Cucurbita* sp. se observó un porcentaje muy bajo de aparición de codones de parada prematuros (0,4%). Porcentajes similares se observaron en otros estudios (1,7% tomate y 0,3% melocotón). El porcentaje de cambios no sinónimos y sinónimos (NS/S) fueron el 35% y el 64%, siendo el ratio de cambios no sinónimo/sinónimo de 0,55. En melocotón también fueron mayores los SNPs que ocasionaban un cambio no sinónimo que los que daban lugar al mismo aminoácido, siendo el porcentaje de cambios sinónimos del 56,3% y no sinónimos del 43,3% (Martínez-García et al., 2013), lo que supondría un ratio *NS/S* de 0,7. Sin embargo, en tomate se observó un mayor porcentaje de cambios no sinónimos (56%) que de cambios sinónimos (40%), siendo el ratio *NS/S* de 1,36. Ratios similares se han observado en otros cultivos como en el arroz (1,2) (Subbaiyan et al., 2012) o la soja (1,36) (Lam et al., 2010) .

Los programas de predicción de efectos tales como *SnpEff* han demostrado ser de gran utilidad estudios genómicos. Algunas de sus aplicaciones han sido para la búsqueda de genes asociados a determinados caracteres (Martínez-García et al., 2013b) o la identificación de polimorfismos relacionados con introgresiones y eventos de mejora (Causse et al., 2013). Programas como *SnpEff* se basan en algoritmos bioinformáticos que combinan la información de la secuencia del genoma de referencia junto a su anotación para realizar la predicción. De este modo, la calidad de la anotación del genoma y el mismo genoma de referencia serán un factor limitante para la predicción de efectos y anotación de SNPs. Por lo que los datos obtenidos mediante predicción serán de gran utilidad para hacernos una idea inicial de diversidad alélica de una población, de qué genes están implicados en una determinada característica o de qué mutación es la responsable de un efecto concreto. Aunque una vez realizada esta preselección de genes o mutaciones candidatas se debería de comprobar experimentalmente. Del mismo modo, también es fundamental validar los SNPs identificados mediante análisis computacionales. Se ha observado que el número de polimorfismos obtenidos por predicción es mayor al real (Cingolani et al., 2012), de tal

modo que su comprobación es fundamental para eliminar falsos positivos. Aun así, la predicción de SNPs obtenida mediante análisis *in silico* no es tan diferente al número de SNPs reales. Por ejemplo en tomate, se validaron un 96,2% de los SNPs identificados (Causse et al., 2013).

En el presente trabajo se ha obtenido una colección de SNVs amplia, de alta calidad y distribuidos por todo el genoma. Además, esta colección representa gran parte de la diversidad existente dentro del género *Cucurbita*. Por estos motivos se propone esta colección de polimorfismos como un repositorio para la selección de marcadores moleculares. El desarrollo de marcadores moleculares a partir de la información disponible en el fichero VCF puede ser de gran ayuda tanto en la mejora y desarrollo de nuevas variedades, como en la investigación y búsqueda de genes de interés. En mejora genética vegetal pueden resultar de interés para la selección temprana de materiales. También son fundamentales para seguir la trazabilidad de los genomas a la hora de realizar hibridaciones entre diferentes parentales, algo esencial para el desarrollo de poblaciones experimentales (poblaciones MAGIC, poblaciones NAM, líneas de introgresión, etc). Así mismo, el hecho de disponer de polimorfismos de 12 de las especies de este género, incluyendo especies silvestres, facilitaría el desarrollo de estas de poblaciones. No solo esta colección resultaría útil como fuente de marcadores moleculares, sino que también puede ser usada como plataforma de EcoTILLING. Al disponer de la anotación de los SNVs, puede servir de base de datos para buscar polimorfismos ya descritos en determinados genes. De este modo, cualquier investigador o mejorador podría consultar esta información con el fin de averiguar si en sus genes de interés existen polimorfismos en estas entradas. Así se puede identificar variantes alélicas en genes relacionados con resistencias, implicados en la maduración, floración, etc.

Por otro lado, en este trabajo se ha hablado de SNPs individuales, centrándonos principalmente en aquellos responsables de ocasionar efectos graves y moderados ya que estos efectos son los que producen un verdadero efecto en la región codificante afectando más adelante a la proteína. Sin embargo, no se suele trabajar con SNPs individuales ya que estos polimorfismos por si solos resultan poco informativos, siendo lo más habitual trabajar con haplotipos. Podemos definir haplotipo como un conjunto de marcadores genéticos estrechamente vinculados presentes en un cromosoma que tienden a ser heredados juntos (no fácilmente separables por recombinación) (Andersen & Lübberstedt, 2003). Sabido esto, se cree que la identificación de unos cuantos alelos

de un bloque haplotípico puede identificar de manera inequívoca el resto de *loci* polimórficos de la región. Dicha información es de gran utilidad para investigar la genética de los caracteres complejos. El conjunto de SNPs que forman un determinado haplotipo no tienen por qué ocasionar todos ellos un impacto grave o moderado. De modo que en un futuro, para completar la información obtenida en este trabajo, se podrían buscar variantes haplotípicas en regiones de interés y comprobar si éstas están asociadas algún fenotipo en particular.

Para finalizar, mencionar que el origen de la secuencia con la cual se ha trabajo es un factor clave a la hora de identificar polimorfismos. En este caso, se han identificado SNPs a partir de RNA obtenido de hoja joven. El RNA solo nos dará información sobre un determinado momento, por lo que se estará perdiendo gran parte de la información. Aquí encontramos una limitación de esta herramienta como posible fuente de búsqueda de genes. Aquellos genes que se expresen en otro estadio vegetal o que se expresen en otro órgano (caracteres de floración, morfología del fruto, etc) serán difícil o no se podrán identificar, puesto que lo más probable es que no se estén expresando. Precisamente, esto supuso una limitación al buscar SNPs dentro de genes candidatos. Recordamos que en algunos QTLs solo se disponía información del 2% del total de genes existentes en esa región. Para evitar este problema, la mejor opción es buscar polimorfismos a partir de DNA. Trabajar con DNA genómico no solo nos hubiese permitido tener una representación de la totalidad de genes del organismo sin vernos limitados por el tipo de tejido, órgano o estado de desarrollo, sino que el número de SNPs identificados habría sido mucho mayor. Además, se hubiese obtenido SNPs, no solo en todos los genes presentes en el genoma, sino que también se hubiese obtenido información de regiones no codificantes, la cuales pueden ser de interés (como en regiones intrónicas). Sin embargo, a pesar de esta limitación, se ha obtenido una colección amplia de SNPs que cubre prácticamente la totalidad del genoma. Así, se espera que esta colección de SNPs haya servido para aumentar los recursos disponibles en *Cucurbita.*

6. Conclusiones

1. En el presente trabajo se ha descrito la diversidad existente dentro de una población de 96 individuos de diferentes especies del género *Cucurbita.* Los polimorfismos de alta calidad (702.735 SNPs) fueron anotados utilizando un programa de predicción de efectos que fuese capaz de predecir el impacto que puede ocasionar cada SNPs.
2. Se observó un porcentaje de heterocigosidad muy bajo, cerca del 1%. Se identificaron polimorfismos en unos 78% de los genes, siendo 32 el promedio de variantes encontradas por gen.
3. Menos del 0,5% de SNPs fueron responsables de ocasionar un impacto grave y cerca del 13% fueron responsables de causar impactos moderados.
4. Unos 585.824 SNPs se encontraban dentro de la secuencia codificante. Más de la mitad ocasionaron cambios de tipo “silencioso” y en menor medida, cambios de tipo “con sentido”. Solo cerca del 0.4% de los SNPs fueron responsables de la aparición de codones de parada prematuros.
5. Como ejemplo práctico, se decidió mostrar la utilidad de crear una base de datos anotada de diversidad genética del género *Cucurbita* para la localización de genes de interés.
6. Toda la diversidad encontrada en la población de cucurbitas se recogió en un fichero VCF, el cual es de acceso libre.

7. Bibliografía

Alagna, F., D'Agostino, N., Torchia, L., Servili, M., Rao, R., Pietrella, M., … Perrotta, G. (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics*, *10*(1), 399.

Andersen, J. R., & Lübberstedt, T. (2003). Functional markers in plants. *Trends in Plant Science*, *8*(11), 554–560.

Andres, T. (1990). Biosystematics, theories on the origin, and breeding potential of Cucurbita cifolia. In *Biology and Utilization of the Cucurbitaceae* (pp. 102–119).

Aslam, Z., Khattak, J. Z. K., Ahmed, M., & Asif, M. (2017). A Role of Bioinformatics in Agriculture. In *Quantification of Climate Variability, Adaptation and Mitigation for Agricultural Sustainability* (pp. 413–434). Cham: Springer International Publishing.

Barkley, N. A., & Wang, M. L. (2008). Application of TILLING and EcoTILLING as Reverse Genetic Approaches to Elucidate the Function of Genes in Plants and Animals. *Current Genomics*, *9*(4), 212–26.

Blanca, J. M., Cañizares, J., Ziarsolo, P., Esteras, C., Mir, G., Nuez, F., … Picó, M. B. (2011). Melon Transcriptome Characterization: Simple Sequence Repeats and Single Nucleotide Polymorphisms Discovery for High Throughput Genotyping across the Species. *The Plant Genome Journal*, *4*(2), 118.

Borevitz, J. O., Liang, D., Plouffe, D., Chang, H.-S., Zhu, T., Weigel, D., … Chory, J. (2003). Large-Scale Identification of Single-Feature Polymorphisms in Complex Genomes. *Genome Research*, *13*(3), 513–523. https://doi.org/10.1101/gr.541303

Boualem, A., Fleurier, S., Troadec, C., Audigier, P., Kumar, A. P. K., Chatterjee, M., … Bendahmane, A. (2014). Development of a Cucumis sativus TILLinG Platform for Forward and Reverse Genetics. *PLoS ONE*, *9*(5), e97963.

Brown, R. N., & Myers, J. R. (2002). A Genetic Map of Squash (Cucurbita sp.) with Randomly Amplified Polymorphic DNA Markers and Morphological Markers. *J. AMER. SOC. HORT. SCI*, *127*(1274).

C., P., L., H., N., G., D., B., C., D., & M., P. (2002). Genetic control of fruit shape acts prior to anthesis in melon (Cucumis melo L.). *Molecular Genetics and Genomics*, *266*(6), 933–941.

Caruso, G., Broglia, V., & Pocovi, M. (2015). Diversidad genética. Importancia y aplicaciones en el mejoramiento vegetal Genetic diversity. Importance and applications in plant breeding.

Causse, M., Desplat, N., Pascual, L., Le Paslier, M.-C., Sauvage, C., Bauchet, G., … Bouchet, J.-P. (2013). Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics*, *14*(1), 791.

Chen, C., Mitchell, S. E., Elshire, R. J., Buckler, E. S., & El-Kassaby, Y. A. (2013). Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genetics & Genomes*, *9*(6), 1537–1544.

Chen, L., Huang, L., Min, D., Phillips, A., Wang, S., Madgwick, P. J., … Hu, Y.-G. (2012). Development and Characterization of a New TILLING Population of Common Bread Wheat (Triticum aestivum L.). *PLoS ONE*, *7*(7), e41570.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92.

Colbert, T. (2001). High-Throughput Screening for Induced Point Mutations. *Plant Physiology*.

Comai, L., Young, K., Till, B. J., Reynolds, S. H., Greene, E. A., Codomo, C. A., … Henikoff, S. (2004). Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *The Plant Journal*, *37*(5), 778–786.

Crespi, M. (2012). *Root genomics and soil interactions*. Wiley-Blackwell.

Cutler, H., and Whitaker, T. (1967). Cucurbits from the Tehuacan caves. In *The Prehistory of the Tehuacán Valley* (pp. 212–219).

Decker-Walters, D. S., and Walters, T. W. (2000). Squash. In *The Cambridge World History of Food*.

Decker-Walters, D. S., Decker-Walters, D. S., Walters, T. W., Cowan, C. W., & Smith, B. D. (1993). Isozymic characterization of wild populations of Cucurbita pepo. *Journal of Ethnobiology.*, *13*, 55–72.

Deleu, W., Esteras, C., Roig, C., González-To, M., Fernández-Silva, I., Gonzalez-Ibeas, D., … Garcia-Mas, J. (2009). BMC Plant Biology A set of EST-SNPs for map saturation and cultivar identification in melon. *BMC Plant Biology*, *9*(90).

Edwards, D., & Batley, J. (2004). Plant bioinformatics: from genome to phenome. *Trends in Biotechnology*, *22*(5), 232–237.

Edwards, D., & Batley, J. (2008). *Bioinformatics: Fundamentals and applications in plant genetics, mapping and breeding. In Chittaranjan Kole and Albert G. Abbott (Ed.), Principles and practices of plant genomics* (Vol. 1). Science Publishers.

Ellson, C. D., Andrews, S., Stephens, L. R., & Hawkins, P. T. (2002). The PX domain: a new phosphoinositide-binding module. *Journal of Cell Science*, *115*(Pt 6), 1099–105.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, *6*(5), e19379.

Esteras, C., Gomez, P., Monforte, A. J., Blanca, J., Vicente-Dolera, N., Roig, C., … Pico, B. (2012). High-throughput SNP genotyping in Cucurbita pepo for map construction and quantitative trait loci mapping. *BMC Genomics*, *13*(1), 80.

Esteras Gómez, C., & Picó Sirvent, M. B. (2011). Nuevas Estrategias de Análisis de la Diversidad Genética Natural: Identificación de Variantes Alélicas en Genes de Interés Mediante Ecotilling.

Ferriol, M., & Picó, B. (2008). Pumpkin and Winter Squash. *Handbook of Plant Breeding - Vegetables 1: Asteraceae, Brassicaceae, Chenopodicaceae and Cucurbitaceae*, 317–349.

Frerichmann, S. L., Kirchhoff, M., Müller, A. E., Scheidig, A. J., Jung, C., & Kopisch-Obuch, F. J. (2013). EcoTILLING in Beta vulgaris reveals polymorphisms in the FLC-like gene BvFL1 that are associated with annuality and winter hardiness. *BMC Plant Biology*, *13*(1), 52.

Friedline, C. J., Lind, B. M., Hobson, E. M., Harwood, D. E., Mix, A. D., Maloney, P. E., & Eckert, A. J. (2015). The genetic architecture of local adaptation I: the genomic landscape of foxtail pine (Pinus balfouriana Grev. & Balf.) as revealed from a high-density linkage map. *Tree Genetics & Genomes*, *11*(3), 49.

Fritz, G. J. (1999). Gender and the Early Cultivation of Gourds in Eastern North America. *American Antiquity*, *64*(03), 417–429.

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., González, V. M., … Puigdomènech, P. (2012). The genome of melon (Cucumis melo L.). *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11872–7.

Gilchrist, E. J., Haughn, G. W., Ying, C. C., Otto, S. P., Zhuang, J., Cheung, D., … Cronk, Q. C. B. (2006). Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of Populus trichocarpa. *Molecular Ecology*, *15*(5), 1367–1378.

Giner Martorell, A., Mariano, J., & Olivert, A. (2017). Calabaza. *Cultivos hortícolas al aire libre*.

Glowka, L., Burhenne-Guilmin, F., & Synge, H. (1996). Guía del Convenio sobre la Diversidad Biológica UICN – Unión Mundial para la Naturaleza.

Gong, L., Stift, G., Kofler, R., Pachner, M., & Lelley, T. (2008). Microsatellites for the genus Cucurbita and an SSR-based genetic linkage map of Cucurbita pepo L. *Theoretical and Applied Genetics*, *117*(1), 37–48.

González, M., Xu, M., Esteras, C., Roig, C., Monforte, A. J., Troadec, C., … Picó, B. (2011). Towards a TILLING platform for functional genomics in Piel de Sapo melons. *BMC Research Notes*, *4*.

Gonzalo, M. J., & Monforte, A. J. (2016). Genetic Mapping of Complex Traits in Cucurbits (pp. 269–290). Springer, Cham.

Guo, S., Liu, J., Zheng, Y., Huang, M., Zhang, H., Gong, G., … Xu, Y. (2011). Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles. *BMC Genomics*, *12*, 454.

Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., … Xu, Y. (2012). The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nature Genetics*, *45*(1), 51–58.

Hashizume, T., Shimamoto, I., & Hirai, M. (2003). Construction of a linkage map and QTL analysis of horticultural traits for watermelon [Citrullus lanatus (THUNB.) MATSUM & NAKAI] using RAPD, RFLP and ISSR markers. *Theoretical and Applied Genetics*, *106*(5), 779–785.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., … Li, S. (2009). The genome of the cucumber, Cucumis sativus L. *Nature Genetics*, *41*(12), 1275–1281.

Ibiza, V. P., Cañizares, J., & Nuez, F. (2010). EcoTILLING in Capsicum species: searching for new virus resistances. *BMC Genomics*, *11*, 631.

Ilut, D. C., Sanchez, P. L., Costich, D. E., Friebe, B., Coffelt, T. A., Dyer, J. M., … Gore, M. A. (2015). Genomic diversity and phylogenetic relationships in the genus Parthenium (Asteraceae). *Industrial Crops and Products*, *76*, 920–929.

Jobst, J., King, K., & Hemleben, V. (1998). Molecular Evolution of the Internal Transcribed Spacers (ITS1 and ITS2) and Phylogenetic Relationships among Species of the Family Cucurbitaceae. *Molecular Phylogenetics and Evolution*, *9*(2), 204–219.

Josefa López Marín. (2017). Calabacín. *Cultivos hortícolas al aire libre*.

Kates, H. R., Soltis, P. S., & Soltis, D. E. (2017). Evolutionary and domestication history of Cucurbita (pumpkin and squash) species inferred from 44 nuclear loci. *Molecular Phylogenetics and Evolution*, *111*, 98–109.

Kaufmann, K., Muiño, J. M., Østerås, M., Farinelli, L., Krajewski, P., & Angenent, G. C. (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nature Protocols*, *5*(3), 457–472.

Kistler, L., Newsom, L. A., Ryan, T. M., Clarke, A. C., Smith, B. D., & Perry, G. H. (2015). Gourds and squashes (Cucurbita spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(49), 15107–12.

KONG, Q., XIANG, C., & YU, Z. (2006). Development of EST-SSRs in Cucumis sativus from sequence database. *Molecular Ecology Notes*, *6*(4), 1234–1236.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., … Zhang, G. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, *42*(12), 1053–1059.

Lee, Y. H., Jeon, H. J., Hong, K. H., & Kim, B. D. (1995). Use of random amplified polymorphic DNA for linkage group analysis in an interspecific cross hybrid F2 generation of Cucurbita. *Journal of Korean Society for Horticultural Science*.

Lewontin, R. (1979). *La Base Genética de la Evolución*.

Li, A., Yang, W., Lou, X., Liu, D., Sun, J., Guo, X., … Zhang, A. (2013). Novel Natural Allelic Variations at the *Rht-1* Loci in Wheat. *Journal of Integrative Plant Biology*, *55*(11), 1026–1037.

Li, Q., Liu, Z., Monroe, H., & Culiat, C. T. (2002). Integrated platform for detection of DNA sequence variants using capillary array electrophoresis. *ELECTROPHORESIS*, *23*(10), 1499.

Lira-Saade, R., and Montes -Hernández, S. (1994). Cucurbita (Cucurbita spp.). In

Neglected Crops : 1492 from a Different Perspective. Food & Agriculture Org. (pp. 63–70).

Lira-Saade, R. (1995). *Estudios taxonomicos ecogeograficos de las Cucurbitaceae Latinoamericanas de importancia economica. Systematic and Ecogeographic Studies on Crop Genepools 9*.

López de Heredia, U. (2016). Las técnicas de secuenciación masiva en el estudio de la diversidad biológica. *Munibe Ciencias Naturales*, *64*.

M. Perez-de-Castro, A., Vilanova, S., Canizares, J., Pascual, L., M. Blanca, J., J. Diez, M., … Pico, B. (2012). Application of Genomic Tools in Plant Breeding. *Current Genomics*, *13*(3), 179–195.

Mardis, E. R., & Ris, K. (2007). ChIP-seq: welcome to the new frontier. *Nature Methods*, *4*(8), 613–614.

Maroto Borrego, J. V. (2002). *Horticultura herbácea especial*. Ediciones Mundi-Prensa.

Martínez-García, P. J., Fresnedo-Ramírez, J., Parfitt, D. E., Gradziel, T. M., & Crisosto, C. H. (2013a). Effect prediction of identified SNPs linked to fruit quality and chilling injury in peach [Prunus persica (L.) Batsch]. *Plant Molecular Biology*, *81*(1–2), 161–174.

Martínez-García, P. J., Fresnedo-Ramírez, J., Parfitt, D. E., Gradziel, T. M., & Crisosto, C. H. (2013b). Effect prediction of identified SNPs linked to fruit quality and chilling injury in peach [Prunus persica (L.) Batsch]. *Plant Molecular Biology*, *81*(1–2), 161–174.

Mascarell-Creus, A., Cañizares, J., Vilarrasa-Blasi, J., Mora-García, S., Blanca, J., Gonzalez-Ibeas, D., … Caño-Delgado, A. I. (2009). An oligo-based microarray offers novel transcriptomic approaches for the analysis of pathogen resistance and fruit quality traits in melon (Cucumis melo L.). *BMC Genomics*, *10*(1), 467.

Mastretta-Yanes, A., Zamudio, S., Jorgensen, T. H., Arrigo, N., Alvarez, N., Piñero, D., & Emerson, B. C. (2014). Gene Duplication, Population Genomics, and Species-Level Differentiation within a Tropical Mountain Shrub. *Genome Biology and Evolution*, *6*(10), 2611–2624.

McCallum, C. M., Comai, L., Greene, E. A., & Henikoff, S. (2000). Targeting induced locallesions in genomes (TILLING) for plant functional genomics. *Plant Physiology*, *123*(2), 439–442.

Mejlhede, N., Kyjovska, Z., Backes, G., Burhenne, K., Rasmussen, S. K., & Jahoor, A. (2006). EcoTILLING for the identification of allelic variation in the powdery mildew resistance genes mlo and Mla of barley. *Plant Breeding*, *125*(5), 461–467.

Merrick, L. C. (1990). Systematics and evolution of a domesticated squash, Cucurbita argyrosperma, and its wild and weedy relatives. In *Biology and utilization of the Cucurbitaceae* (pp. 77–95).

Montero-Pau, J., Blanca, J., Bombarely, A., Ziarsolo, P., Esteras, C., Martí-Gómez, C., … Cañizares, J. (2017). De-novo assembly of zucchini genome reveals a whole genome duplication associated with the origin of the Cucurbita genus. *BioRxiv*.

Montero-Pau, J., Blanca, J., Esteras, C., Martínez-Pérez, E. M., Gómez, P., Monforte, A. J., … Picó, B. (2017). An SNP-based saturated genetic map and QTL analysis of fruit-related traits in Zucchini using Genotyping-by-sequencing. *BMC Genomics*, *18*(1), 94.

Morris, E. R., & Walker, J. C. (2003). Receptor-like protein kinases: the keys to response. *Current Opinion in Plant Biology*, *6*(4), 339–42. Retrieved from

Nee, M. (1990). The domestication ofcucurbita (Cucurbitaceae). *Economic Botany*, *44*(S3), 56–68.

Nieto, C., Piron, F., Dalmais, M., Marco, C. F., Moriones, E., Gómez-Guillamón, M. L., … Bendahmane, A. (2007). EcoTILLING for the identification of allelic variants of melon eIF4E, a factor that controls virus susceptibility. *BMC Plant Biology*, *7*(1), 34.

Paris, H. S., & Brown, R. N. (2005). The genes of pumpkin and squash. *HortScience*, *40*(6), 1620–1630.

Paris, HS., and Maynard, D. (2008). Cucurbita. In *The Encyclopedia of Fruits and Nuts*.
Paris, H. S. (1986). A proposed subspecific classifiaction for Cucurbita pepo. *Phytologia*.
Paris, H. S. (1989). Historical records, origins, and development of the edible cultivar groups of Cucurbita pepo (Cucurbitaceae). *Economic Botany*, *43*(4), 423–443.
Paris, H. S. (2008). Summer squash. *Handbook of Plant Breeding - Vegetables 1: Asteraceae, Brassicaceae, Chenopodicaceae and Cucurbitaceae*, *Volume*, *1*, 351–381.
Paris, H. S. (2016). Genetic Resources of Pumpkins and Squash, Cucurbita spp. (pp. 111–154). Springer, Cham.
Paris, H. S., & Padley, L. D. (2014). Gene List for Cucurbita species. Retrieved from
Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, *7*(5), e37135.
Piperno, D. R., Holst, I., Wessel-Beaver, L., & Andres, T. C. (2002). Evidence for the control of phytolith formation in Cucurbita fruits by the hard rind (Hr) genetic locus: Archaeological and ecological implications. *Proceedings of the National Academy of Sciences*.
Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE*, *7*(2), e32253.
Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., … Jannink, J.-L. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome Journal*, *5*(3), 103.
Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., … Zhang, Z. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(14), 5135–40.
Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B., & El-Kassaby, Y. A. (2015). A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered SNP imputation methods. *Heredity*, *115*(6), 547–55.
Robinson, R. W. and Decker-Walters, D. S. (1997). *Cucurbits*.
Robinson, R.W., H.M. Munger, T.W. Whitaker, and G. W. B. (1979). New Genes for the Cucurbitaceae. *Cucurbit Genetics Cooperative Report*, *4*, 49–53.
Sabetta, W., Blanco, A., Zelasco, S., Lombardo, L., Perri, E., Mangini, G., & Montemurro, C. (2013). Fad7 gene identification and fatty acids phenotypic variation in an olive collection by EcoTILLING and sequencing approaches. *Plant Physiology and Biochemistry*, *69*, 1–8.
Sanjur, O. I., Piperno, D. R., Andres, T. C., & Wessel-Beaver, L. (2002). Phylogenetic relationships among domesticated and wild species of Cucurbita (Cucurbitaceae) inferred from a mitochondrial gene: Implications for crop plant evolution and areas of origin. *Proceedings of the National Academy of Sciences*, *99*(1), 535–540.
Schilling, M. P., Wolf, P. G., Duffy, A. M., Rai, H. S., Rowe, C. A., Richardson, B. A., & Mock, K. E. (2014). Genotyping-by-Sequencing for Populus Population Genomics: An Assessment of Genome Sampling Patterns and Filtering Approaches. *PLoS ONE*, *9*(4), e95292.
Service, R. F. (2006). Gene sequencing. The race for the $1000 genome. *Science (New York, N.Y.)*, *311*(5767), 1544–6. https://doi.org/10.1126/science.311.5767.1544
Shendure, J., Mitra, R. D., Varma, C., & Church, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics*, *5*(5), 335–344.
Singh, A. . (1990). Cytogenetics and evolution in the Cucurbitaceae. *Biology and Utilization of the Cucurbitaceae*, 10–28.
Singh VK, Singh AK, Chand R, & Kushwaha C. (2011). Role Of Bioinformatics In

Agriculture And Sustainable Development. *International Journal of Bioinformatics Research*, *3*(2), 975–3087.
Smith, B. D. (1997). Reconsidering the Ocampo Caves and the Era of Incipient Cultivation in Mesoamerica. *Latin American Antiquity*, *8*(04), 342–383.
Smith, B. D. (2005). Reassessing Coxcatlan Cave and the early history of domesticated plants in Mesoamerica. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9438–45.
Subbaiyan, G. K., Waters, D. L. E., Katiyar, S. K., Sadananda, A. R., Vaddadi, S., & Henry, R. J. (2012). Genome-wide DNA polymorphisms in elite *indica* rice inbreds discovered by whole-genome sequencing. *Plant Biotechnology Journal*, *10*(6), 623–634.
Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., … Xu, Y. (2017). Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes. *Molecular Plant*, *10*(10), 1293–1306.
Teppner, H. (2004). Notes on Lagenaria and Cucurbita (Cucurbitaceae) ± Review and New Contributions.
Thudi, M., Khan, A. W., Kumar, V., Gaur, P. M., Katta, K., Garg, V., … Varshney, R. K. (2016). Whole genome re-sequencing reveals genome-wide variations among parental lines of 16 mapping populations in chickpea (Cicer arietinum L.). *BMC Plant Biology*, *16*(S1), 10.
Tillib, S. V, & Mirzabekov, A. D. (2001). Advances in the analysis of DNA sequence variations using oligonucleotide microchip technology. *Current Opinion in Biotechnology*, *12*(1), 53–8.
Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, *1*, e203.
Varshney, R. K., Close, T. J., Singh, N. K., Hoisington, D. A., & Cook, D. R. (2009). Orphan legume crops enter the genomics era! *Current Opinion in Plant Biology*, *12*(2), 202–210.
Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, *27*(9), 522–530.
Vicente-Dólera, N., Troadec, C., Moya, M., del Río-Celestino, M., Pomares-Viciana, T., Bendahmane, A., … Gómez, P. (2014). First TILLING Platform in Cucurbita pepo: A New Mutant Resource for Gene Function and Crop Improvement. *PLoS ONE*, *9*(11), e112743.
Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, *9*(8), 808–810. https://doi.org/10.1038/nmeth.2023
Wang, T. L., Uauy, C., Robson, F., & Till, B. (2012). TILLING *in extremis*. *Plant Biotechnology Journal*, *10*(7), 761–772.
Wechter, W. P., Levi, A., Harris, K. R., Davis, A. R., Fei, Z., Katzir, N., … Trebitsh, T. (2008). Gene expression in developing watermelon fruit. *BMC Genomics*, *9*(1), 275.
Weeden, N. F. (1984). *Isozyme studies indicate that the genus Cucurbita is an ancient tetraploid*. *Report: Cucurbit genetics cooperative (USA)*.
Wenzel, G., Kennard, W. C., & Havey, M. J. (1995). Quantitative trait analysis of fruit quality in cucumber: QTL detection, confirmation, and comparison with mating-design variation. *Theoretical and Applied Genetics*, *91*(1), 53–61.
Whitaker, T. W., & Bemis, W. P. (1975). Origin and Evolution of the Cultivated Cucurbita. *Bulletin of the Torrey Botanical Club*, *102*(6), 362.
Whitaker, T. W., & Davis, G. N. (1962). Cucurbits. Botany, cultivation, and utilization. *Cucurbits. Botany, Cultivation, and Utilization.* Retrieved from
Wilson, H., Doebley, J., & Duvall, M. (1992). Chloroplast DNA diversity among wild and cultivated members of Cucurbita (Cucurbitaceae). *Theoretical and Applied*

Genetics, *84–84*(7–8), 859–865.

Wu, T., Luo, S., Wang, R., Zhong, Y., Xu, X., Lin, Y., & Huang, H. (2014). The first Illumina-based de novo transcriptome sequencing and analysis of pumpkin (*Cucurbita moschata* Duch.) and SSR marker development. *Molecular Breeding*, *34*(3), 1437–1447.

Xiao, W., & Oefner, P. J. (2001). Denaturing high-performance liquid chromatography: A review. *Human Mutation*, *17*(6), 439–474.

Yu, S., Liao, F., Wang, F., Wen, W., Li, J., Mei, H., & Luo, L. (2012). Identification of Rice Transcription Factors Associated with Drought Tolerance Using the Ecotilling Method. *PLoS ONE*, *7*(2).

Zhang, G., Ren, Y., Sun, H., Guo, S., Zhang, F., Zhang, J., & Li, H. (2015). A high-density genetic map for anchoring genome sequences and identifying QTLs associated with dwarf vine in pumpkin (*Cucurbita maxima* Duch.). *BMC Genomics*, *16*(1), 1101.

Zraidi, A., Stift, G., Pachner, M., Shojaeiyan, A., Gong, L., & Lelley, T. (2007). A consensus map for *Cucurbita pepo*. *Molecular Breeding*, *20*(4), 375–388.

8. Anexos

Tabla Suplementaria 1.- Información sobre los individuos utilizados en este estudio. Se muestra el nombre de la muestra, el código del banco de germoplasma, la especie, la subespecie, el país de origen y las observaciones pertinentes.

Nombre	Códigos Banco	Especie	Subespecie	Origen	Observaciones
CO-1	CATIE 7223	*C. moschata*		Panamá	
CO-10	PI 419083	*C. moschata*		China	
CO-100	CATIE 11869	*C. pepo*		Guatemala	
CO-101	PI 202079	*C. argyrosperma*	*argyrosperma*	México	
CO-12	PI 458746	*C. moschata*		Guatemala	
CO-13	PI 482527	*C. moschata*		Zimbawe	
CO-14	PI 498429	*C. moschata*		Colombia	
CO-15	PI 512150	*C. moschata*		México	
CO-16	PI 543218	*C. moschata*		Bolivia	
CO-17	PI 550689	*C. moschata*		Canadá	
CO-18	MENINA	*C. moschata*		Portugal	
CO-19	PI 604506	*C. moschata*		Estados Unidos	
CO-2	BGV000205	*C. moschata*		Marruecos	
CO-20	PI 560946	*C. moschata*		Bolivia	
CO-21	CATIE 9243	*C. moschata*		México	
CO-22	CATIE 12054	*C. moschata*		Nicaragua	
CO-24	BGV002748	*C. moschata*		España	
CO-25	BGV004565	*C. moschata*		República Dominicana	
CO-26	BGV005863	*C. moschata*		Ecuador	
CO-27	BGV004560	*C. moschata*		Cuba	
CO-29	UPV 035143	*C. moschata*		Angola	
CO-3	BGV000818	*C. moschata*		España	
CO-30	SUD-CU 6	*C. maxima*		Argentina	
CO-31	VIR 1860	*C. maxima*		Australia	
CO-32	VIR 2422	*C. maxima*		Repblica Africana	
CO-33	VIR 3202	*C. maxima*		Chile	
CO-34	VIR 4273	*C. maxima*		Perú	
CO-35	VIR 4381	*C. maxima*		Perú	
CO-36	VIR 4656	*C. maxima*		Perú	
CO-37	BGV000832	*C. maxima*		España	
CO-38	CATIE 9824	*C. maxima*		Colombia	
CO-39	MAX 306/98	*C. maxima*		Argentina	
CO-4	NIG LOCAL	*C. moschata*		Nigeria	
CO-40	PI 543227	*C. maxima*		Bolivia	
CO-41	BGV005953	*C. maxima*		Ecuador	
CO-42	UPV 035142	*C. máxima*		Angola	

Tabla Suplementaria 1.- (continuación)

CO-43	Styriam pumpkin	*C. pepo*	*pepo*	Austria
CO-44	PI 7	*C. pepo*	*pepo*	Turquía
CO-45	CATIE 18887	*C. pepo*	*pepo*	México
CO-46	CATIE 11368	*C. pepo*	*pepo*	Guatemala
CO-47	CATIE 9394	*C. pepo*	*pepo*	México
CO-48	CATIE 11079	*C. pepo*	*pepo*	Honduras
CO-49	BGV004119	*C. pepo*	*pepo*	España
CO-5	CATIE 10913	*C. moschata*		Honduras
CO-50	BGV001325	*C. pepo*	*pepo*	España
CO-51	BGV003842	*C. pepo*	*pepo*	España
CO-52	BGV005299	*C. pepo*	*pepo*	España
CO-53	BGV000216	*C. pepo*	*pepo*	Marruecos
CO-54	BGV004370	*C. pepo*	*pepo*	Murcia
CO-55	BGV005382	*C. pepo*	*ovifera*	España
CO-56	BGV005329	*C. pepo*	*ovifera*	España
CO-57	NSL 5227	*C. pepo*	*ovifera*	Estados Unidos
CO-58	NSL 42793	*C. pepo*	*ovifera*	Estados Unidos
CO-59	PI 518687	*C. pepo*	*ovifera*	Estados Unidos
CO-6	PI 80596	*C. moschata*		Japón
CO-60	PI 615111	*C. pepo*	*ovifera*	Estados Unidos
CO-61	NSL 5206	*C. pepo*	*ovifera*	Estados Unidos
CO-62	NSL 32665	*C. pepo*	*ovifera*	Estados Unidos
CO-63	BGV004617	*C. pepo*	*ovifera*	Estados Unidos
CO-64	BGV005389	*C. pepo*	*ovifera*	Estados Unidos
CO-65	PI 653839	*C. cordata*		México
CO-66	Grif 9445	*C. cordata*		México
CO-67	PI 512114	*C. argyrosperma*	*argyrosperma*	Nicaragua
CO-68	PI 512115	*C. argyrosperma*	*argyrosperma*	Guatemala
CO-69	PI 532363	*C. okeechobensis*	*martinezii*	México
CO-7	PI 264551	*C. moschata*		Guatemala
CO-70	PI 512105	*C. okeechobensis*	*martinezii*	México
CO-71	PI 512106	*C. okeechobensis*	*martinezii*	México
CO-72	PI 432441	*C. ecuadorensis*		Ecuador
CO-73	PI 432443	*C. ecuadorensis*		Ecuador
CO-74	Grif 9446	*C. ecuadorensis*		Ecuador
CO-75	PI 438542	*C. lundelliana*		Belize
CO-76	PI 540898	*C. lundelliana*		Honduras
CO-77	PI 532354	*C. pepo*	*fraterna*	México
CO-78	PI 614701	*C. pepo*	*okarzana*	México
CO-8	PI 369346	*C. moschata*		Costa Rica
CO-80	PI 442197	*C. foetidissima*		México
CO-81	PI 532350	*C. foetidissima*		México

Tabla Suplementaria 1.- (continuación)

CO-83	PI 442341	*C. pedatifolia*		México	
CO-84	PI 442290	*C. pedatifolia*		México	
CO-85	PI 458653	*C. maxima*	*andreana*	Argentina	
CO-86	PI 532392	*C. scabridifolia*		México	Híbrido *C. foetidissima x C. scabridifolia*
CO-87	PI 653064	*C. moschata*		Nigeria	
CO-88	CATIE 16038	*C. ficifolia*		Guatemala	
CO-89	PI 532357	*C. lundelliana*		México	
CO-9	PI 381814	*C. moschata*		India	
CO-90	PI 636138	*C. lundelliana*		Belize	
CO-91	PI 442201	*C. foetidissima*		México	
CO-92	PI 540737	*C. pedatifolia*		México	Híbrido *C. pedatifolia x C. foetidissima*
CO-93	ISI 1	*C. pepo*	*pepo*	Italia	
CO-94	ISI 2	*C. pepo*	*pepo*	Italia	
CO-95	ISI 3	*C. pepo*	*pepo*	Italia	
CO-96	ISI 4	*C. pepo*	*pepo*	Italia	
CO-97	CATIE 16575	*C. ficifolia*		Guatemala	
CO-98	PI 451712	*C. argyrosperma*	*argyrosperma*	México	
CO-99	PI 438547	*C. argyrosperma*	*argyrosperma*	Belize	

Tabla Suplementaria 2.- Descripción de cada uno de los QTLs elegidos. Se indica el nombre del QTL, el fenotipo estudiado, el cromosoma en el que se encuentran y las posiciones físicas en pb de inicio y fin de cada QTL.

QTL	Fenotipo	Cromosoma	Posición de inicio	Posición de fin
Li_10	Incisión de la hoja	10	1.401.713	1.584.421
Sl_12	*Silver Leaf*	17	5.039.143	5.325.849
DFeF_12	Días para la floración femenina	17	1.777.698	2.582.093
DFeF_9	Días para la floración femenina	19	6.938.098	7.285.995
MaRCo_4	Color de la piel en frutos maduros, parámetro a	5	3.234.344	3.545.709
MLRCo_4	Color de la piel en frutos maduros, parámetro L	5	912.616	1.299.236
MbRCo_19	Color de la piel en frutos maduros, parámetro b	14	2.426.465	3.268.099
IFSh_3	Forma de fruto inmaduro	3	24.344	492.185
IbRCo_4	Color de la piel en frutos inmaduros, parámetro b	5	912.616	7.691.848

Tabla Suplementaria 3.- Listado de genes en la regi´on del *Sl_12.* Se indican los 24 genes que se detectaron en la región *Sl_12*. Se muestra el identificador del gen, el nombre del gen y el número de impactos graves y moderados encontrados en cada gen.

ID-gene	Nombre del gen	Nº impactos graves	Nº impactos moderados
Cp4.1LG17g08380	*Major facilitator superfamily protein*	0	15
Cp4.1LG17g08450	*Leucine-rich repeat receptor-like protein kinase family protein*	0	2
Cp4.1LG17g08500	*Protein of Unknown Function (DUF239)*	0	19
Cp4.1LG17g08520	*DNA glycosylase superfamily protein*	0	17
Cp4.1LG17g08530	*Protein phosphatase 2C family protein*	0	25
Cp4.1LG17g08540	*ATP-dependent zinc metalloprotease FtsH*	0	2
Cp4.1LG17g08550	*Serinc-domain containing serine and sphingolipid biosynthesis protein*	0	4
Cp4.1LG17g08560	*Unknown protein*	0	15
Cp4.1LG17g08570	*AP2-like ethylene-responsive transcription factor*	0	6
Cp4.1LG17g08580	*Ribulose bisphosphate carboxylase/oxygenase activase, chloroplastic*	0	14
Cp4.1LG17g08600	*Tudor/PWWP/MBT superfamily protein isoform 4 [Theobroma cacao]*	0	23
Cp4.1LG17g08620	*40S ribosomal protein S12*	0	10
Cp4.1LG17g08630	*Unknown protein*	0	7
Cp4.1LG17g08640	*HXXXD-type acyl-transferase family protein*	0	54
Cp4.1LG17g08650	*Pre-mRNA-splicing factor ISY1 homolog*	0	11
Cp4.1LG17g08670	*ATP-dependent RNA helicase DED1*	0	2
Cp4.1LG17g08710	*Glycine--tRNA ligase*	0	16
Cp4.1LG17g08720	*PPPDE putative thiol peptidase family protein*	0	1
Cp4.1LG17g08770	*F-box/LRR-repeat protein*	0	1
Cp4.1LG17g08790	*Receptor-like protein kinase*	1	9
Cp4.1LG17g08860	*Phox (PX) domain-containing protein*	1	2
Cp4.1LG17g08890	*Tetratricopeptide repeat (TPR)-like superfamily protein*	0	4
Cp4.1LG17g08940	*Early nodulin 16 precursor, putative [Ricinus communis]*	0	55
Cp4.1LG17g08950	*Methyl-CPG-binding domain 11*	0	32

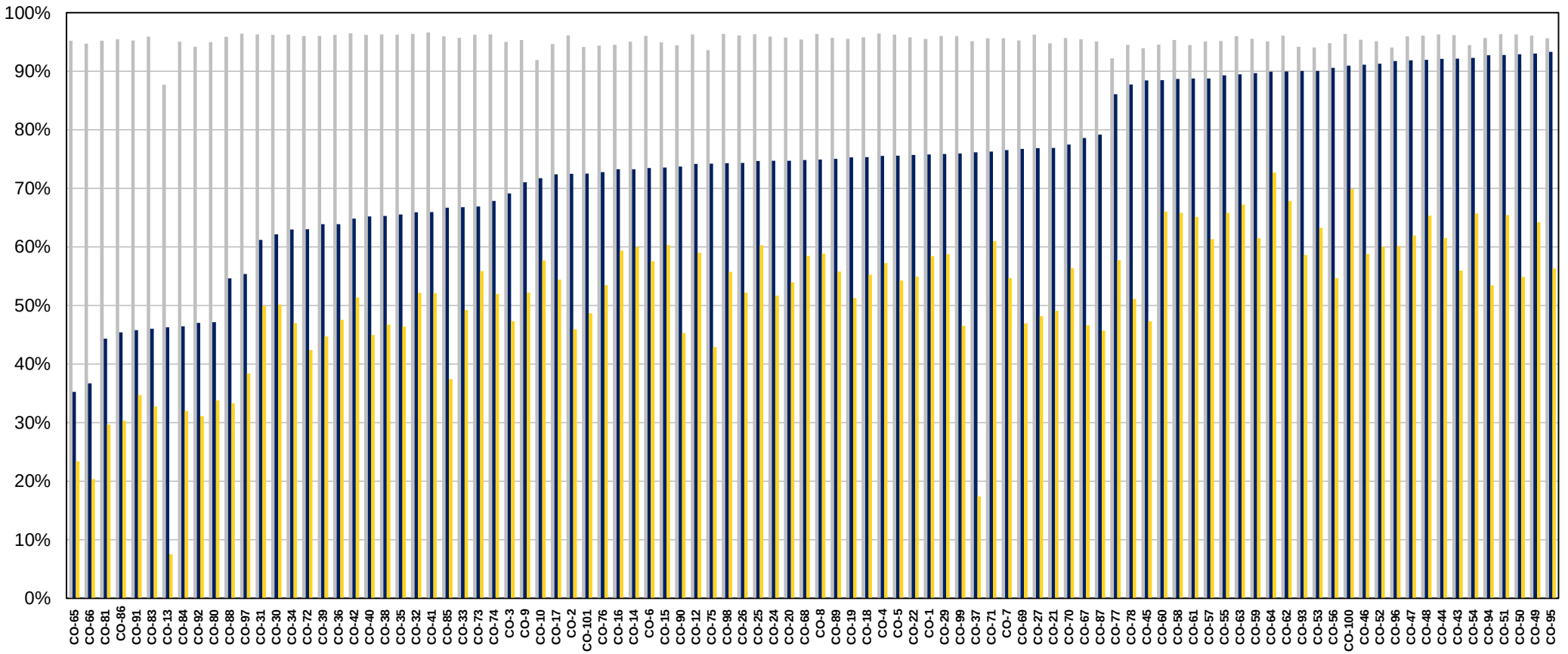

Figura suplementaria 1.- Representación gráfica del porcentaje de lecturas útiles para cada muestra. En gris se muestra el porcentaje de lecturas de buena calidad, en azul el porcentaje de lecturas mapeadas y en amarillo, el porcentaje de lecturas que se mapearon con una calidad óptima.

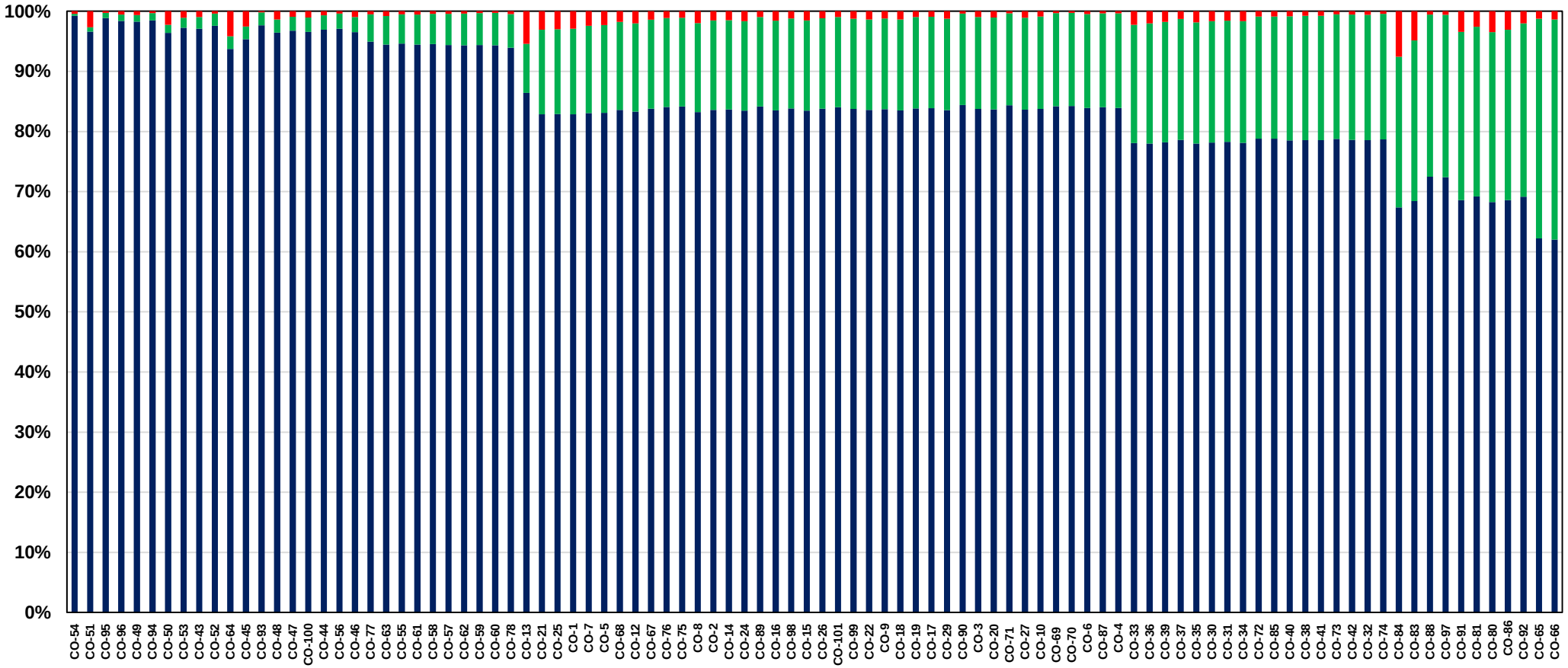

Figura suplementaria 2.- Representación gráfica del porcentaje de genotipos homocigotos para el alelo de referencia (azul), para el alelo alternativo (verde) y heterocigotos (rojo) en cada una de las muestras.