Hybridization as an evolutionary driver for speciation:

a case in the Southern European *Erysimum* species

UNIVERSIDAD
DE GRANADA

Carolina Inés Osuna Mascaró

Ph.D. Thesis

Editor: Universidad de Granada. Tesis Doctorales
Autor: Carolina Osuna Mascaró
ISBN: 978-84-1306-667-7
URI: ttp://hdl.handle.net/10481/63946

La doctoranda, **Carolina I. Osuna Mascaró**, y los directores de la Tesis Doctoral **Francisco Perfectti** y **Rafael Rubio de Casas**:

Garantizamos, al firmar esta Tesis Doctoral, que el trabajo ha sido realizado por la doctoranda bajo la dirección de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados cuando se han utilizado sus resultados o publicaciones.

Granada, 3 de julio de 2020

Dr Francisco Perfectti Dr Rafael Rubio de Casas

Carolina I. Osuna Mascaró

Durante el tiempo de realización de esta Tesis Doctoral he disfrutado de una Ayuda para Contratos Predoctorales del Ministerio de Economía y Competitividad (BES-2014-069022).

Las dos estancias realizadas durante esta tesis han sido financiadas por el Ministerio de Economía y Competitividad con dos ayudas para Estancias Breves (EEBB-I-16-10858 y EEBB-I-18-12920).

Este trabajo ha estado finaciado por el Ministerio de Economía y Competitividad (CGL2013-47558-P, CGL2016-79950-R and CGL2017-86626-C2-2-P), y fondos FEDER.

La investigación presentada en esta Tesis Doctoral se ha desarrollado en los departamentos de de Genética y Ecología de la Universidad de Granada.

A Amona, ¡bihar obeto!

Table of Contents

Summary

Hybridization between species is an important evolutionary phenomenon that has played an influential role across the entire tree of life. Recent studies have documented it in a diverse range of taxa, including mammals, birds, fish, fungi, and insects. However, it appears to have been especially influential in the evolution of plants, with about 25% of them derived from hybridization events. The evolutionary results of hybridization can vary widely. Hybridization between recently formed species can hinder speciation and, therefore, diversification. However, hybridization can also foster the formation of new species and increase biodiversity by facilitating the introduction of new genetic variability in a different genetic background through introgression. In addition, in many cases, hybridization events are associated with polyploidization due to the fusion of the genomes of the two species.

Hybridization leaves detectable genomic footprints that may be characterized using different approaches. However, the most widely used technique is arguably the reconstruction of molecular phylogenies. For instance, phylogenetic hypotheses based on the internal transcribed spacers (hereafter ITS) have been used to detect hybridization and polyploidization in many species. The popularity of these markers derives partly from the easy amplification of these regions by almost universal eukaryotic primers, but also from the fact that hybridization signatures can often be detected as intragenomic nucleotide polymorphisms due to incomplete concerted evolution. Organellar, especially plastidial, markers have been used as an alternative or complementary approach to nuclear regions to determine hybridization patterns at a phylogenetic/phylogeographic scale. The chloroplast genomes commonly have exclusive maternal inheritance (i.e., uniparental inheritance). Thus, they generally have low effective

population size and stable structure, which makes them extremely useful for establishing phylogenetic hypotheses not influenced by hybridization events.

With the advent of next-generation sequencing and the fast-growth of genomic tools, scientists have begun to analyze the dynamics of hybridization at a genomic scale, using whole genomes instead of single markers for phylogenetic reconstructions. Phylogenomic approaches allow for the analysis of thousands of orthologous loci or even complete (organellar) chromosomes, providing further insight into the influence of hybridization on evolution. However, whole-genome analyses for plant species remains a demanding task, particularly in taxa with large genomes. A workaround to this staggering complexity is to focus only on informative subsets of the genome. For instance, phylogenomics using transcriptomic data, particularly RNA-Seq, use only the transcribed sequences as regions of interest for phylogenomic studies. Nevertheless, the interpretation of these phylogenomic reconstructions remains challenging, especially when hybridization is expected to influence evolutionary patterns. Increasing the number of markers across the genome also increases the likelihood of retrieving regions retaining ancestral polymorphisms (i.e., deep coalescence or incomplete lineage sorting).

Erysimum L. is one of the most abundant genera of the *Brassicaceae*, comprising more than 200 species. It has been described as a taxonomically complex genus in which molecular evidence indicated a reticulated evolutionary history, with polyploidization events in some clades. This genus is distributed mainly in Eurasia, with some species in North America and North Africa. With more than a hundred species described in it, the Mediterranean region appears to be a center of diversification of *Erysimum* spp. The group is particularly abundant in the Iberian Peninsula, where twenty-one, or twenty-three, different species occur. Within this region, the Baetic Mountains in the SE appear to host a significant amount of the group’s biodiversity, with ten *Erysimum* species, seven of them endemic to this area. The Baetic Mountains are considered one of the biggest glacial refugia in Europe and a significant hotspot of biodiversity. The isolation

(allopatry) during glacial maxima followed by post-glacial secondary contacts may have facilitated interspecific gene flow and hybridization. Moreover, *Erysimum* spp. inhabiting the Baetic Mountains show characteristics that may facilitate hybridization and interspecific gene-flow, such as occasional sympatry and a generalist pollination system. Previous studies have indicated that the evolution of corolla color in Iberian *Erysimum* might have been influenced by hybridization, and also, some of the *Erysimum* species of this region have been described as polyploid. These reticulated complex evolutionary dynamics are evident in phylogenetic reconstructions of the group, which are conflicting and generally not well-resolved.

The principal objective of this Ph.D. thesis was to disentangle the role of hybridization and polyploidization in the evolution of plants, using a set of Southern European *Erysimum* species as a study case, namely, *E. nevadense* and *E. baeticum*, *E. mediohispanicum* and *E. bastetanum*, *E fitzii*, and *E. popovii*. These species inhabit the Baetic Mountains, sometimes in sympatry. Our central hypothesis was that hybridization and polyploidization might have had a significant influence on the evolution of these species, leaving a signature in their genomes. This study was addressed considering that these and other evolutionary processes take place at the population level and that therefore intraspecific variability had to be explicitly incorporated in all analyses. Additionally, we hypothesized that the hybridization signal might be more patent in species with purple corollas, as this phenotype might have been favored by adaptive introgression. First, we have studied to which degree ITS sequences were homogenized by concerted evolution in this group (**Chapter I**). We found that these nuclear sequences were not thoroughly homogenized, suggesting a recent hybridization signature for these *Erysimum* species. Second, we assembled the chloroplast genomes of these species from RNA-Seq data, demonstrating the reliability of this strategy to provide complete cpDNA genomes (**Chapter II**). Then, we provided new genomic resources for *Erysimum* spp., like *de novo* assembled and annotated transcriptomes. Following this methodology, we assembled and annotated the

chloroplast genomes, reconstructing a time-calibrated phylogeny (**Chapter III**). Finally, we studied the general hybridization scenario in the presence of ILS for these species by combining nuclear and chloroplast data (**Chapter IV**). We found that introgression was ubiquitous in the species studied here and that even diploid species showed a signature of introgression in their genomes. These results provide insights into the importance of hybridization and polyploidization for plant evolution in general and *Erysimum* in particular.

Resumen

La hibridación ha desempeñado un papel muy influyente en la evolución de muchas especies. Estudios recientes han encontrado huellas de hibridación en una diversa gama de taxones, incluidos mamíferos, aves, peces, hongos e insectos. Parece haber sido especialmente influyente en la evolución de las plantas, ya que se estima que alrededor del 25 % de ellas derivan de eventos de hibridación. Las consecuencias evolutivas de los procesos de hibridación pueden variar ampliamente. La hibridación entre especies recientes puede dificultar la especiación y por lo tanto la diversificación. Sin embargo, puede actuar también fomentando la formación de nuevas especies y aumentando la diversidad biológica, al facilitar la introducción de nueva variabilidad genética mediante la introgresión. Además, en muchos casos, los eventos de hibridación están asociados a la poliploidización debido a la fusión de los genomas de las dos especies.

La hibridación deja huellas genómicas detectables que pueden caracterizarse utilizando diferentes enfoques. La técnica más utilizada es posiblemente la reconstrucción de filogenias moleculares. En muchas especies se han utilizado hipótesis filogenéticas basadas en los espaciadores transcritos internos (en adelante, ITS) para detectar tanto hibridación como poliploidización. La popularidad de estos marcadores se debe en parte a la fácil amplificación de esas regiones, pero también al hecho de que las firmas de hibridación pueden detectarse a menudo como polimorfismos intragenómicos de nucleótidos, debido a una evolución concertada incompleta. Los marcadores provenientes de orgánulos, especialmente los cloroplastidiales, se han utilizado como un enfoque alternativo, o complementario, de las regiones nucleares para determinar las pautas de hibridación a escala filogenética/filogeográfica. Los genomas del

cloroplasto suelen tener una herencia materna (es decir uniparental), por lo general tienen un tamaño efectivo de población bajo y una estructura estable lo que los hace sumamente útiles para establecer hipótesis filogenéticas no influidas por los acontecimientos de hibridación.

Con la llegada de la secuenciación masiva y el rápido crecimiento de las herramientas genómicas, los científicos han comenzado a analizar las dinámicas de la hibridación a escala genómica, utilizando genomas completos en lugar de marcadores únicos para las reconstrucciones filogenéticas. Los enfoques filogenómicos permiten el análisis de miles de loci ortólogos o incluso de cromosomas completos (organelas), lo que permite comprender mejor la influencia de la hibridación en la evolución. Sin embargo, los análisis de genomas completos para las especies vegetales siguen siendo una tarea ardua, en particular en los taxones con genomas grandes. Una solución provisional a esta abrumadora complejidad consiste en centrarse solo en subconjuntos informativos del genoma. Así, la filogenómica que utiliza datos transcriptómicos, en particular el ARN-Seq, utiliza solamente las secuencias transcritas como regiones de interés para los estudios filogenómicos. No obstante, la interpretación de estas reconstrucciones filogenómicas sigue siendo difícil, especialmente cuando se espera que la hibridación influya en las pautas evolutivas. El aumento del número de marcadores en todo el genoma aumenta la probabilidad de recuperar regiones que conservan polimorfismos ancestrales (es decir, la coalescencia profunda o la clasificación incompleta de linajes).

Erysimum L. es uno de los géneros más abundantes de la familia *Brassicaceae*, que incluye a más de 200 especies. Se ha descrito como un género taxonómicamente complejo en el que las pruebas moleculares indican una historia evolutiva reticulada, con acontecimientos de poliploidización en algunos clados. Se distribuye principalmente en Eurasia, con algunas especies en Norteamérica y en el norte de África. Con más de un centenar de especies descritas, la región mediterránea parece ser un centro de diversificación de *Erysimum* spp. El grupo es particularmente abundante en la Península Ibérica, donde hay más de veinte especies descritas.

Dentro de esta región, las cordilleras Béticas en el SE albergan una parte significativa de la biodiversidad del grupo, con diez especies de *Erysimum*, siete de ellas endémicas de esta zona.

Las cordilleras Béticas están consideradas como uno de los mayores refugios glaciares de Europa y un importante punto caliente de biodiversidad. El aislamiento (alopatría) durante los máximos glaciares, seguido de contactos secundarios postglaciares, puede haber facilitado el flujo de genes interespecíficos y la hibridación. Además, las especies de *Erysimum* que habitan en sistema Bético muestran características que pueden facilitar la hibridación y el flujo de genes interespecíficos, como la simpatría en algunas ocasiones y un sistema de polinización generalista. Estudios previos han sugerido que la evolución del color de la corola en estas especies ha podido estar influenciado por la hibridación, y además, algunas de las especies de *Erysimum* de esta región han sido descritas como poliploides. Esta dinámica evolutiva reticulada es evidente en las reconstrucciones filogenéticas del grupo, que son conflictivas y generalmente no están bien resueltas.

El objetivo principal de esta tesis doctoral ha sido desenmarañar el papel de la hibridación y de la poliploidización en la evolución de las plantas, utilizando como caso de estudio a un conjunto de pares de especies de *Erysimum* del sur de Europa: *E. nevadense* y *E. baeticum*; *E. mediohispanicum* y *E. bastetanum*; *E fitzii* y *E. popovii*. Estas especies habitan las montañas a veces en simpatría. Nuestra hipótesis central era que estos procesos podrían haber tenido una influencia significativa en la evolución de estas especies, dejando una firma en sus genomas. Este estudio se abordó teniendo en cuenta que estos y otros procesos evolutivos tienen lugar a nivel de población, y que por ello la variabilidad intraespecífica debía incorporarse explícitamente en todos los análisis. Planteamos además la hipótesis de que la señal de hibridación podría ser más patente en las especies con corolas púrpuras, ya que este fenotipo podría haberse visto favorecido por la introgresión adaptativa. En primer lugar, hemos tratado de dilucidar en qué medida las secuencias de ITS fueron homogeneizadas por la evolución

concertada (**Capítulo I**). Encontramos que estas secuencias nucleares no fueron completamente homogeneizadas, lo que sugiere una firma de hibridación reciente para estas especies de *Erysimum*. En segundo lugar, hemos reunido los genomas del cloroplasto de estas especies a partir de los datos del ARN-Seq, demostrando la fiabilidad de esta estrategia para proporcionar genomas completos de ADNcp (**Capítulo II**). Tras ello, hemos proporcionado nuevos recursos genómicos para el estudio de las especies de *Erysimum*, como transcriptomas ensamblados y anotados *de novo*. Siguiendo esta metodología, hemos ensamblado y anotado los genomas del cloroplasto, reconstruyendo una filogenia calibrada en el tiempo (**Capítulo III**). Finalmente, hemos estudiado el escenario general de hibridación en presencia de ILS para estas especies, combinando datos nucleares y del cloroplasto (**Capítulo IV**). Nuestro trabajo indica que la introgresión ha sido ubicua en las especies estudiadas, incluso las especies diploides mostraban una firma de introgresión en sus genomas. Estos resultados inciden en la gran importancia de la hibridación y la poliploidización en la evolución del mundo vegetal en general y del género *Erysimum* en particular.

General Introduction

Hybridization in Ecology and Evolution

Hybridization among different species is an evolutionary phenomenon that has played a creative role in many lineages (Rieseberg and Carney, 1998; Coyne and Orr, 2004; Arnold, 2006; Abbott et al., 2013). Because of its pervasiveness, it has long been a subject of research (Stebbins, 1959; Anderson, 1954; Arnold et al., 1999). Nevertheless, it has remained controversial and a matter of debate among scientists. Thus, zoologists such as Mayr (1942) or Dobzhansky (1953) have interpreted hybridization as a marginal phenomenon, as opposed to the point of view of many botanists as Stebbins (1950) that considered hybridization as a widespread process (Yakimowski and Rieseberg, 2014). Recently, however, with the advent of next-generation sequencing and the fast development of genomic tools, scientists have begun to analyze the dynamics of hybridization at a genomic scale, which has undoubtedly increased our understanding of the role of hybridization in nature (Payseur and Rieseberg, 2016; Goulet et al., 2017; Taylor and Larson, 2019). Recent studies have documented hybridization in a diverse range of taxa, including mammals (Liu et al., 2015; De Manuel et al., 2016; Cahill et al., 2016; Svardal et al., 2017; Jones et al., 2018), birds (Toes et al., 2016; Lamichhaney et al., 2018; Runemark et al., 2018), fishes (Seehausen et al., 2003; Meier et al., 2017), fungi (Neafsey et al., 2010; Keuler et al., 2020), and insects (Fontaine et al., 2015; Larson et al., 2019). However, hybridization appears to have been especially influential in the evolution of plants, and about 25% of angiosperms species derive from hybridization events (Mallet, 2005; Soltis and Soltis, 2009; Abbott et al., 2013).

The evolutionary outcomes of hybridization may vary widely (Taylor and Larson, 2019). A cross between different species or even between genetically distant populations of the same species can result in the formation of a new hybrid species (Rieserberg and Willis, 2007; Soltis and Soltis, 2009). However, in some instances, the new hybrid species could present lower fitness than their parents. In these cases, hybridization might hinder speciation and, therefore, diversification (Mayr, 1992; Schemske, 2000; Mallet, 2005; Saari and Faeth, 2012; Gómez et al.,

2015a). Under these circumstances, prezygotic reproductive barriers may evolve among the parental taxa avoiding hybrid formation (in a process named reinforcement; Howard, 1993; Servedio and Noor, 2003; Ortiz-Barrientos et al., 2009; Hopkins and Rausher, 2012; Hopkins and Rausher, 2014; Lemmon et al., 2017; Calabrese and Pfennig, 2020). Nevertheless, hybridization also may play a creative role, fostering the formation of new species that show significant genotypic and phenotypic variation compared to their parent species (i.e., hybrid speciation; Rieseberg et al., 2003; Stelkens and Seehausen, 2009). In some instances, hybridization may produce novel extreme phenotypes exceeding the phenotypic range of parental lineages (i.e., transgressive segregation; Stelkens and Seehausen, 2009). One such example has been described in *Helianthus*, where at least three diploid species have arisen by hybridization (namely, *H. anomalus*, *H. deserticola*, and *H. paradoxus*; Rieseberg, 1991). The combination of alleles from parental species in these hybrids has been posited to help them colonize extreme habitats (Lexer et al., 2003). The contribution of both parental genomes to the hybrid is not necessarily uniform since some genic regions of one of the parents might well be overrepresented relative to the other parent, usually by continuous backcrossing between the hybrids and one of the parental species. This process can produce taxa with the general gene pool of one of the parent species but with some genetic material incorporated from the other one. This process is called “genetic introgression” and can produce genetic variants with considerable evolutionary consequences (Arnold, 2004; Abbot et al., 2013; Hanušová et al., 2014; Arnold and Kunte, 2017).

Introgressed regions might be instrumental in the acquisition of novel functional traits and can mediate the response to new selective pressures, a process that is understood as “adaptive introgression” (Hanušová et al., 2014; Arnold and Martin, 2009; Arnold and Kunte, 2017; Suarez-Gonzalez et al., 2018). Adaptive introgression has been well documented in some plant systems such as *Senecio* (Kim et al., 2008a), *Helianthus* (Kim and Rieseberg, 1999; Whitney et al., 2006; Whitney et al., 2010), *Iris* (Martin et al., 2006), *Phaseolus* (Rendón-Anaya et al., 2017),

Arabidopsis (Arnold et al., 2016), or *Populus* (Suarez-Gonzalez et al., 2016; Suarez-Gonzalez, 2017). However, the analysis of adaptive introgression requires detailed ecological and genetic knowledge of the natural system to demonstrate an adaptive function for the introgressed genic regions, which is often challenging (Suarez-Gonzalez et al., 2018; Taylor and Larson, 2019).

After a hybridization event, the ploidy level of the hybrid species may change. Hybridization can lead to the union of the two sets of chromosomes from both parentals, followed by genome duplication, which results in polyploidization (i.e., allopolyploidization; Soltis and Soltis, 2009, **Figure 1**).

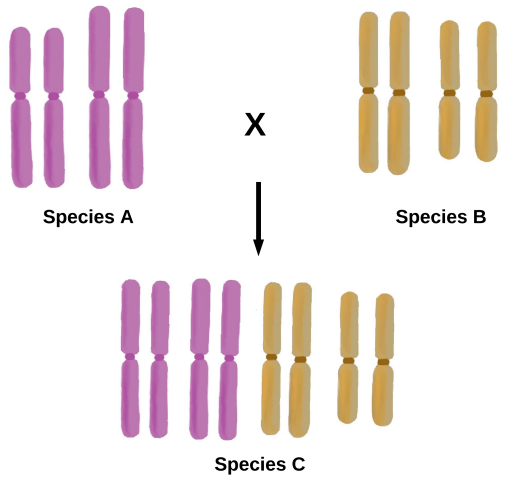

Figure 1. Formation of an allopolyploid (species C) after a hybridization event among species A and species B.

About 30%-70% of extant plant species have a polyploid origin, and the majority arose through the combined effects of interspecific hybridization and polyploidy (Soltis and Soltis, 2009; Moghe and Shiu, 2014). Nevertheless, different points of view exist about the role of polyploidization in nature. Some authors have pointed out that polyploidy tends to be an evolutionary dead-end, with few polyploid species surviving for long (Mayrose et al., 2011;

Arrigo and Baker, 2012; Mayrose et al., 2014). Several examples indicate that polyploidy can be associated with lower vigor and reduced adaptive capacity (Comai, 2005). These deleterious consequences are driven by a variety of effects on fundamental biological processes, including changes in cellular architecture (Kondorosi et al., 2000), difficulties in meiosis and mitosis (Bennett and Smith, 1972; Grandont et al., 2013; Pelé et al., 2018), regulatory changes in gene expression (Cheng et al., 2018; Ramírez-González et al., 2018) and epigenetic instability (Din and Chen., 2018; Wendel et al., 2018) among others. In contrast, other authors argue that polyploidy has been one of the most creative processes for the diversification and evolution of plants (Stebbins, 1950; Wood et al., 2009; Chen, 2010; Mayfield et al., 2011; Soltis et al., 2014). The increase of alleles for a given gene may buffer polyploids from deleterious mutations (Gu et al., 2003), make polyploidy progeny more vigorous than their diploid progenitors (i.e., heterosis; Comai, 2005; Chen et al., 2010; Birchler et al., 2010; Chen, 2013; Washburn et al., 2014; Sattler et al., 2016), foster the neofunctionalization of genomic regions (Osborn et al., 2003; Adams and Wendel., 2005) or the subfunctionalization of others (Lynch and Force., 2000; Van de Peer et al., 2017). Some polyploids have been described as better adapted to extreme environments and more tolerant to adverse conditions than their parents (Otto and Whitton, 2000; Coate et al., 2013), which indicates that under some circumstances, polyploidization can facilitate niche expansion (Moore and Purugganan, 2005). These novelties might be facilitated by introgression of specific genomic regions of adaptive value. However, despite the close evolutionary link between hybridization and polyploidization, there is limited knowledge on the interplay of adaptive introgression and polyploidization (Chapman and Abbott, 2010).

Studying hybridization

Hybridization leaves detectable genomic footprints that may be characterized using different approaches. For instance, hybrids might be distinguishable based on morphologic traits, cytologic characteristics (Rieseberg et al.,1993; Bartish et al., 2000; Thórsson et al., 2001; Cronberg and Natcheva, 2002; Radosavljević et al., 2019), and the distribution of haplotypes or allele frequencies (Dobeš et al., 2004; Palme et al., 2004; Zhang et al., 2013; Payseur and Rieseberg, 2016). However, the most widely used technique is arguably the reconstruction of molecular phylogenies (Rieseberg and Soltis, 1991; Wissemann, 2007).

In particular, phylogenetic hypotheses based on the internal transcribed spacers (hereafter ITS) have been used to detect hybridization and polyploidization in many species (Baldwin et al., 1995; Alvarez and Wendel, 2003; Hughes et al., 2006). This popularity derives from the easy amplification of these regions by almost universal eukaryotic primers (White et al., 1990) and the fast rate of concerted evolution of ITS that tends to homogenize sequences in the hybrid genome relatively quickly (Nieto-Feliner and Rosselló, 2007). However, sequence homogenization is not uniform, and hybridization signatures can often be detected as intragenomic nucleotide polymorphisms (Buckler and Holtsford, 1996; Mayol and Roselló, 2001; Alvarez and Wendel., 2003; Bailey et al., 2003; Popp and Oxelman, 2004; Won and Renner, 2005; Harpke and Peterson, 2006; Grimm and Denk, 2008; Zheng et al., 2008; Xiao et al. , 2010; Drabková et al., 2009; Soltis and Soltis, 2009). Nevertheless, as concerted evolution often occurs fast, ITS polymorphism can be detected mostly in recently formed hybrids. Therefore, the use of ITS for phylogenetic reconstruction may lead to erroneous interpretations in systems where ancient hybridization events might have taken place (Alvarez and Wendel., 2003; Hörandl et al., 2005; Winkworth et al., 2005; Whitfield and Lockhart, 2007; Grimm and Denk, 2008).

Organellar, especially plastidial, markers have been used as an alternative or complementary approach to nuclear regions to determine hybridization patterns at a phylogenetic/phylogeographic scale. The chloroplast genomes commonly have a maternal inheritance in most angiosperms (i.e., uniparental inheritance; Reboud and Zeyl, 1994). Thus, they generally have low effective population size and stable structure, which makes them extremely useful for establishing phylogenetic hypotheses not influenced by hybridization events (Harris et al., 1991; Bernhardt et al., 2017). Many studies have used specific chloroplast genes in combination with nuclear ones for phylogenetic reconstructions. Ideally, under a non-hybridization scenario, both types of markers should render the same phylogenetic results. However, when hybridization exists, a phylogenetic discordance appears among nuclear and plastidial phylogenies (i.e., cytonuclear discordance; Wendel and Doyle, 1998; Sang and Zhong, 2000; Barber et al., 2007; Kim and Donoghue, 2008b; Fehrer et al., 2007; Sun et al., 2015; Tkach et al., 2019). Unfortunately, the evolutionary rate of plant plastidial genomes is comparatively slow. Consequently, phylogenetic studies using a few plastid regions are hindered by limited resolution, and comparisons of poorly resolved phylogenetic trees may be uninformative and ineffective to trace hybridization events.

The development of high-throughput sequencing technologies has made possible the use of whole genomes instead of single markers for phylogenetic reconstructions (Ekblom and Galindo, 2011; Straub et al., 2012; Hou et al., 2016). Phylogenomic approaches allow analyzing thousands of orthologous loci or even complete (organellar) chromosomes, providing further insight into the influence of hybridization on the evolution of many groups (Delsuc et al. 2005; Ma et al., 2014; Ruhfel et al., 2014; Carbonell-Caballero et al., 2015; Guo et al., 2017; McKain et al., 2018; Morales-Briones et al., 2018). Using a larger number of informative sites can make the construction of better-resolved phylogenies possible. However, whole-genome analyses for plant species remains a demanding task, particularly in taxa with large genomes. One possible

approach to overcome the inherent complexity of working with complete genomes is to focus only on subsets of the genome a-priori deemed as informative. Phylogenomics using transcriptomic data, particularly RNA-Seq, can provide a useful alternative to complete-genome phylogenomics by using only the transcribed sequences as regions of interest for phylogenomic studies (Timme et al., 2012; Wickett et al., 2012; Yang and Smith., 2013; Léveillé-Bourret et al., 2017). However, the interpretation of these phylogenomic reconstructions remains challenging, especially when hybridization is expected to influence evolutionary patterns (Willyard et al., 2009; Shao et al., 2019). In fact, increasing the number of markers across the genome also increases the likelihood of retrieving regions retaining ancestral polymorphisms (i.e., deep coalescence or incomplete lineage sorting, hereafter ILS, **Figure 2**). Because these regions are often plesiomorphic, they can be easily confused with hybridization signals (i.e., they are shared among different taxa; Solís-Lemus and Ané, 2016; Elworth et al., 2019; Glémin et al., 2019). Hybridization studies have long ignored ILS, mainly due to methodological limitations (Maddison, 1997; Meng and Kubatko, 2009; Kubatko, 2009; Solís-Lemus et al. 2016) which might have led to wrong estimations of the number of hybridization events in certain groups (Mallet, 2005; Yu et al., 2013). This confusion is particularly problematic when studying closely-related species, in which both ILS and hybridization are expected to be relatively frequent (Gerard et al., 2011; Meng and Kubatko, 2009; Wang et al., 2018).

When the horizontal genetic transfer is frequent due to either complete hybridization -with or without polyploidization- or introgression, the representation of phylogenetic relationships as hierarchical trees might lead to a misleading interpretation of taxonomic relationships (Holland et al., 2008; Yu et al., 2011; Elworth et al., 2019). In these situations, evolutionary histories are best represented by phylogenetic networks, which extend the phylogenetic tree model by allowing for horizontal connections. Furthermore, phylogenetic network approaches can also incorporate ILS, and thus give a better picture of the cross-linking

events between taxa (Than et al., 2008; Solís-Lemus et al., 2017; Wen et al., 2018). These approaches have been strengthened by the development of several genomic tools that can be used to infer hybridization and introgression in the presence of ILS. For instance, the ABBA-BABA test has been broadly used to detect ancient hybridization and introgression despite high levels of ILS (Green et al., 2010; Durand et al., 2011; Dasmahapatra et al., 2012; Freedman et al., 2014; Pease et al., 2016; Ru et al., 2016; Malinsky et al., 2018; Lin et al., 2019). The detection of introgression events has also been dramatically improved by novel methodologies, like the f̂d statistic (Martin et al., 2015) and techniques to detect long haplotypes shared among species (Staubach et al., 2012; Harris and Nielsen, 2013; Leitwein et al., 2020). However, each method has shortcomings and requires caution in its application (Wagner et al., 2013; Wisecaver and Hackett., 2014; Huber et al., 2015). Therefore, an accurate picture of hybridization and introgression patterns requires using a combinational approach that incorporates several different methodologies (Arnold, 2015).

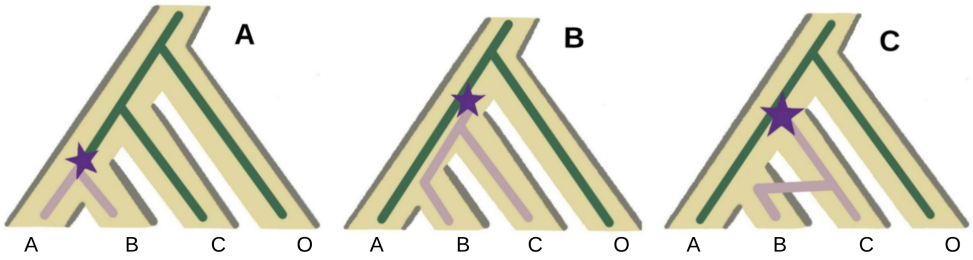

Figure 2. The beige area represents the species tree. The green line represents a single gene genealogy. The star represents a mutation changing the ancestral allele into the derived allele (purple line). (A) Gene genealogy congruent with species tree; (B) ILS: the ancestral polymorphism is produced before the divergence between C from A and B and maintained in B and C, so these share the novel, derived allele not present in A; (C) Introgression: B receives the allele from C through gene flow. In the case of ILS and introgression, the gene genealogy is not consistent with the species tree, but the two processes produce similar phylogenies, rendering discrimination of the two phenomena highly complex.

Erysimum species as a study system

Erysimum L. is a plant genus of the family Brassicaceae, distributed mainly in Eurasia, with some species in North America and North Africa (Warwick et al., 2006). Its members span a range of different habitats (Koch and Al-Shehbaz, 2016), and have been used as model systems to study plant-pollinator interactions (Gómez et al., 2008; Gómez et al., 2009; Ortigosa and Gómez, 2010; Lay et al., 2011; Gómez et al., 2015b; Valverde et al., 2014; Valverde et al., 2016; Valverde et al., 2019), responses to herbivory (Piippo et al., 2005; Lay et al., 2011), adaptation to alpine habitats (Kim and Donohue, 2011a; Kim and Donohue, 2011b; Kim and Donohue, 2013; Lay et al., 2013), drought resistance (Kim and Donohue, 2012), the evolution of chemical defenses (Züst et al., 2018; Züst et al., 2020), or the evolution of floral shape (Gómez et al., 2006; Gómez et al., 2008; Gómez and Perfectti, 2010; Savriama et al., 2012).

Erysimum is a taxonomically complex genus (Moazzeni et al., 2014), with between 150 (Al-Shehbaz and Al-Shammary, 1987; Al-Shehbaz, 2010) and 350 (Polatschek, 1986; Polatschek and Snogerup, 2002) species described. The striking taxonomical discrepancy among the number of species is mainly the result of the morphological similarities among them (Ančev and Polatschek, 2006; Gómez et al., 2006; Gómez and Perfectti, 2010; Mutlu, 2010; Ghaempanah et al., 2012; Ghaempanah et al., 2013; Abidkulova et al., 2017). For instance, cryptic species with very similar morphology but isolated ecologically and/or geographically have been described (Turner, 2006; Abdelaziz et al. 2011; Ouarmin et al., 2013; Lorite et al., 2015; Czarna et al., 2016). These cryptic species may reflect rapid radiations, including recent or incomplete speciation events (Beheregaray and Caccone, 2007; Bickford et al., 2007; Trontelj and Fišer, 2009; Nosil et al., 2009, Struck, et al., 2018). Therefore, molecular phylogenies of *Erysimum* spp. are conflicting and generally not well-resolved, which has been taken as evidence for reticulate evolution (Abdelaziz et al., 2014; Gómez et al. 2014; Moazzeni et al., 2014) and polyploidization

(Al-Shehbaz & Al-Shammary, 1987; Ančev and Polatschek, 2005; Ancev, 2006; Marhold & Lihová, 2006; Turner, 2006; Abdelaziz et al., 2011; Abdelaziz et al., 2014). For instance, Moazzeni et al. (2014) reconstructed a phylogeny, including 110 *Erysimum* species using ITS markers, obtaining low resolution in some clades. This lack of resolution could be due to the limitations of ITS as phylogenetic markers (as discussed above) or to the group's taxonomic complexity; rapid radiations may not have allowed enough time for clearcutdivergence. However, it could also be related to hybridization events, which may give a reticulated phylogenetic signal.

The Mediterranean region, with > 100 species, appears to be a center of diversification of *Erysimum* spp. (Greuter et al., 1986). The group is particularly abundant in the Iberian Peninsula, where twenty-one (Polatschek, 1978; Polatschek, 2014), or twenty-three (Nieto-Feliner, 2003; Mateo et al., 1998), different species occur. Within this region, the Baetic Mountains in the SE appear to host a significant amount of the group's biodiversity, harbouring ten *Erysimum* species, seven of them endemic to this area (Blanca et al., 1992). The Baetic Mountains are considered one of the biggest glacial refugia in Europe and a significant hot-spot of biodiversity (Médail and Diadéma, 2009). Here, as in other refugia, isolation (allopatry) during glacial maxima followed by post-glacial secondary contacts among species may have facilitated gene flow and hybridization (Nieto-Feliner, 2011). *Erysimum* species inhabiting the Baetic Mountains show characteristics that may facilitate hybridization and inter-specific gene-flow, such as occasional sympatry and a generalist pollination system. Moreover, some of the *Erysimum* species of this region have been described as polyploid (Nieto-Feliner, 2003). Therefore, phylogenies including these species are usually reticulated with low support for some clades (Abdelaziz et al., 2014; Gómez et al., 2014).

Furthermore, previous studies have indicated that the evolution of corolla color in Iberian *Erysimum* might have been influenced by hybridization (Nieto-Feliner, 2003; Abdelaziz et al., 2013; Abdelaziz et al., 2014). Most species are yellow, but some of them have purple corolla color. In particular, some pairs of *Erysimum* species from the Baetic Mountains have a very

similar general phenotype but differ in corolla color, having purple and yellow corolla. Therefore, there is a possibility that purple color has been transferred in the Iberian clade from a purple parental species by hybridization and has been maintained by natural selection. Moreover, some of these pairs of different corolla color species usually appear in phylogenies in the same clade as sister species. Specifically, *E. bastetanum* and *E. mediohispanicum* (**Figure 3**; Abdelaziz, et al., 2014; Gómez et al., 2014) and *E. nevadense* and *E. baeticum* (**Figure 3**, Abdelaziz, et al., 2014; Gómez et al., 2014). Moreover, *E. popovii* and *E. fitzii* (Nieto-Feliner, 2003) have also been described as species with similar morphology, but different corolla color (purple and yellow, respectively) and phylogenetic reconstructions were not well resolved for them.

In summary, Baetic *Erysimum* species constitute a promising system to study the influence of hybridization, polyploidization, and other genomic processes such as ILS on the reticulate evolution of a group of species. Several authors have already posited the importance of these processes in this plant group. However, clarifying these intricate evolutionary patterns will require extensive and detailed genomic resources, which are mostly lacking.

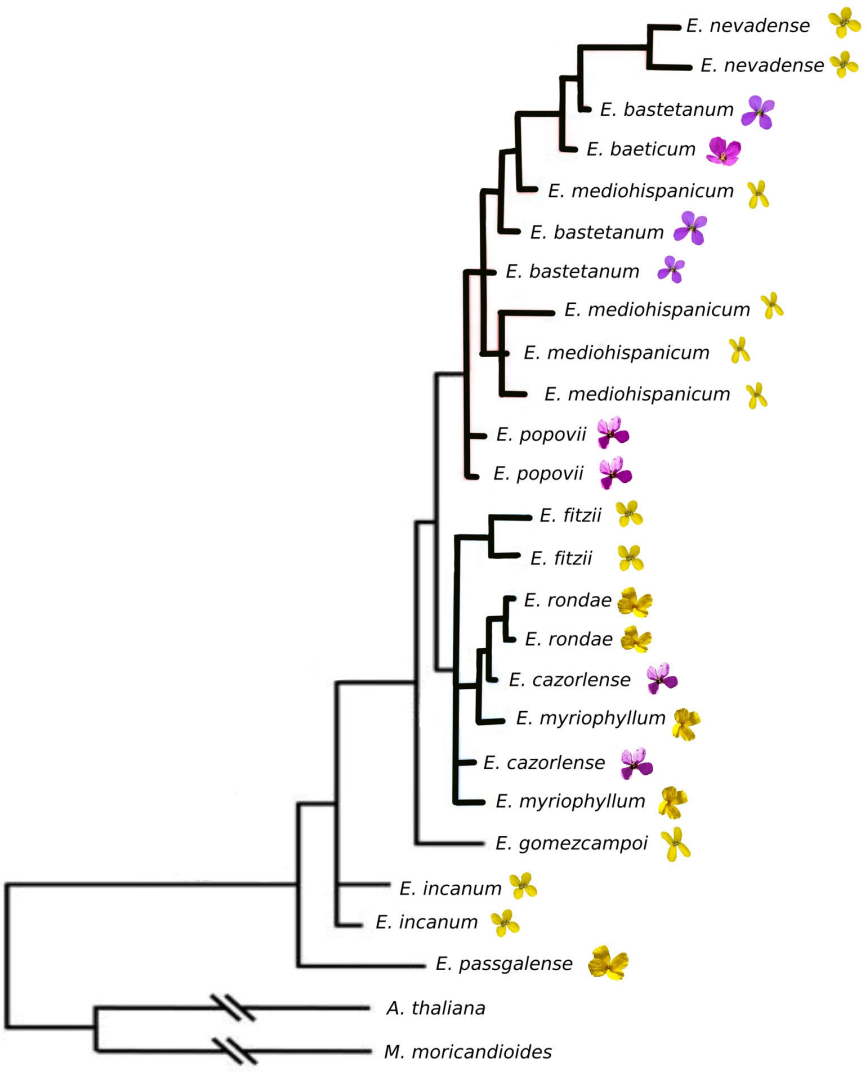

Figure 3. Phylogenetic tree depicting the lack of phylogenetic resolution (no species clustering) for some *Erysimum* species, including the studied in this thesis. Modified from Abdelaziz, et al., 2014.

Goals and general structure of this Ph.D. thesis

The principal objective of this Ph.D. thesis is to disentangle the role of hybridization and polyploidization in the evolution of several Southern European *Erysimum* species. Namely, *E. nevadense* and *E. baeticum*, *E. mediohispanicum* and *E. bastetanum*, *E fitzii* and *E. popovii*. These species inhabit the Baetic Mountains, and some of them are in sympatry. Our central hypothesis is that these processes might have had a significant influence on the evolution of these species, leaving a signature in their genomes. In some instances, these processes may have been more relevant at the population level than at the species level. Additionally, we hypothesize that the hybridization signal might be more apparent in species with purple corollas, as this phenotype might have been favored by introgression.

To test these hypotheses, we addressed the following specific objectives:

- To obtain phylogenetically informative genomic resources in the form of whole chloroplast genomes and assembled transcriptomes.
- To investigate the ploidy levels of the species studied here to infer if some have an allopolyploid origin.
- To analyze polymorphism in the ITS1 and ITS2 nuclear markers to infer whether recent hybridization's footprint can be detected in them.
- To reconstruct nuclear and chloroplast phylogenies and determine the existence of cytonuclear discordance.
- To build phylogenetic networks and infer ancient hybridization events.
- To study the potential effect of ILS on phylogenetic and evolutionary patterns.
- To investigate if the existence of introgression signatures in the anthocyanin pathway.

This Ph.D. dissertation has been structured in four chapters that address the objectives mentioned above:

In **Chapter I**, we studied the concerted evolution pattern for seven *Erysimum* species that span different ploidy levels, analyzing whether concerted evolution may have homogenized ITS1 and ITS2 sequences. We also examined whether there were differences in concerted evolution rates between diploid and polyploid species, thereby providing insight into the consequences of allopolyploidization for ITS evolution.

In **Chapter II**, we assembled chloroplast genomes for three *Erysimum* (Brassicaceae) species using three RNA-Seq samples and one genomic library of each species. We compared these assembled genomes, confirming that chloroplast genomes assembled from RNA-Seq data were highly similar to each other and those obtained from genomic libraries in terms of the overall structure, size, and sequence.

In **Chapter III**, we presented the 18 floral transcriptomes used here. These were assembled *de novo* from total RNA-Seq libraries and annotated. We also include the chloroplast genomes analyzed in Chapter II. We also presented a time-calibrated phylogeny of these species constructed with the whole chloroplast genomes.

Finally, in **Chapter IV**, we studied the evolution of these sets of species considering ILS, hybridization, and polyploidization, using multiple phylogenetic approaches. Furthermore, we explored if anthocyanin genes exhibit signatures of introgression and positive selection.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., ... & Butlin, R. K. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26 (2), 229-246.

Abdelaziz, M., Lorite, J., Muñoz-Pajares, A. J., Herrador, M. B., Perfectti, F., & Gómez, J. M. (2011). Using complementary techniques to distinguish cryptic species: a new *Erysimum* (Brassicaceae) species from North Africa. *American Journal of Botany*, 98 (6), 1049-1060.

Abdelaziz, M. (2013). How species are evolutionarily maintained? Pollinator-mediated divergence and hybridization in *Erysimum mediohispanicum* and *Erysimum nevadense* (Doctoral dissertation, Universidad de Granada).

Abdelaziz, M., Muñoz-Pajares, A. J., Lorite, J., Herrador, M. B., Perfectti, F., & Gómez, J. M. (2014). Phylogenetic relationships of *Erysimum* (Brassicaceae) from the Baetic Mountains (se Iberian peninsula). *Anales del Jardín Botánico de Madrid* (Vol. 71, No. 1, p. 005).

Abidkulova, K. T., Mukhitdinov, N. M., Ivashchenko, A. A., Ametov, A. A., & Serbayeva, A. D. (2017). Morphological characteristics of a rare endemic species, *Erysimum croceum* M. Pop. (Brassicaceae) from Trans-Ili Alatau, Kazakhstan. *Modern Phytomorphology*, 11, 131-138.

Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, 8 (2), 135-141.

Álvarez, I., & Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, 29 (3), 417-434.

Al-Shehbaz, I. A., & Al-Shammary, K. I. (1987). Distribution and chemotaxonomic significance of glucosinolates in certain Middle-Eastern Cruciferae. *Biochemical Systematics and Ecology*, 15 (5), 559-569.

Al-Shehbaz, I. A., & Windham, M. D. (2010). Flora of North America North of Mexico. *New York and Oxford*, 7, 451-453.

Ančev, M., & Polatschek, A. (2005). The genus *Erysimum* (Brassicaceae) in Bulgaria. Annalen des Naturhistorischen Museums in Wien. *Serie B für Botanik und Zoologie*, 227-273.

Ancev, M. (2006). Polyploidy and hybridization in Bulgarian Brassicaceae: distribution and evolutionary role. *Phytologia Balcanica*, 12 (3), 357-366.

Anderson, E., & Stebbins Jr, G. L. (1954). Hybridization as an evolutionary stimulus. *Evolution*, 8 (4), 378-388.

Arnold, M. L., & Meyer, A. (2006). Natural hybridization in primates: one evolutionary mechanism. *Zoology*, 109 (4), 261-276.

Arnold, M. L., Bulger, M. R., Burke, J. M., Hempel, A. L., & Williams, J. H. (1999). Natural hybridization: how low can you go and still be important?. *Ecology*, 80 (2), 371-381.

Arnold, M. L. (2004). Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right?. *The Plant Cell*, 16 (3), 562-570.

Arnold, M. L., & Martin, N. H. (2009). Adaptation by introgression. *Journal of Biology*, 8 (9), 82.

Arnold, M. L. (2015). Divergence with genetic exchange. OUP Oxford.

Arnold, B. J., Lahner, B., DaCosta, J. M., Weisman, C. M., Hollister, J. D., Salt, D. E., ... & Yant, L. (2016). Borrowed alleles and convergence in serpentine adaptation. *Proceedings of the National Academy of Sciences*, 113 (29), 8320-8325.

Arnold, M. L., & Kunte, K. (2017). Adaptive genetic exchange: a tangled history of admixture and evolutionary innovation. *Trends in Ecology & Evolution*, 32 (8), 601-611.

Arrigo, N., & Barker, M. S. (2012). Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology*, 15 (2), 140-146.

Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *The American Journal of Human Genetics*, 73 (4), 823-834.

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., & Donoghue, M. J. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden*, 247-277.

Barber, J. C., Finch, C. C., Francisco-Ortega, J., Santos-Guerra, A., & Jansen, R. K. (2007). Hybridization in Macaronesian Sideritis (Lamiaceae): evidence from incongruence of multiple independent nuclear and chloroplast sequence datasets. *Taxon*, 56 (1), 74-88.

Bartish, I. V., Rumpunen, K., & Nybom, H. (2000). Combined analyses of RAPDs, cpDNA and morphology demonstrate spontaneous hybridization in the plant genus *Chaenomeles*. *Heredity*, 85 (4), 383-392.

Beheregaray, L. B., & Caccone, A. (2007). Cryptic biodiversity in a changing world. *Journal of Biology*, 6 (4), 9.

Bennett, M. D., & Smith, J. B. (1972). The effects of polyploidy on meiotic duration and pollen development in cereal anthers. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 181 (1062), 81-107.

Bernhardt, N., Brassac, J., Kilian, B., & Blattner, F. R. (2017). Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. *BMC Evolutionary Biology*, 17 (1), 141.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K., Meier, R., Winker, K., ... & Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution*, 22 (3), 148-155.

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., & Veitia, R. A. (2010). Heterosis. *The Plant Cell*, 22 (7), 2105-2112.

Blanca, G., Morales, C. & Ruiz-Rejón, M. (1992) El género *Erysimum* L. (Cruciferae) en Andalucía (España). *Anales del Jardín Botánico de Madrid* 49: 201–214.

Buckler 4th, E. S., & Holtsford, T. P. (1996). Zea systematics: ribosomal ITS evidence. *Molecular Biology and Evolution*, 13 (4), 612-622.

Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., Shapiro, B., ... & Wayne, R. K. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Science Advances*, 2 (7), e1501714.

Calabrese, G. M., & Pfennig, K. S. (2020). Reinforcement and the Proliferation of Species. *Journal of Heredity*, 111 (1), 138-146.

Chapman, M. A., & Abbott, R. J. (2010). Introgression of fitness genes across a ploidy barrier. *New Phytologist*, 186 (1), 63-71.

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends in Plant Science*, 15 (2), 57-71.

Chen, Z. J. (2013). Genomic and epigenetic insights into the molecular bases of heterosis. *Nature Reviews Genetics*, 14 (7), 471-482.

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., & Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants*, 4 (5), 258-268.

Coate, J. E., & Doyle, J. J. (2013). Genomics and transcriptomics of photosynthesis in polyploids. *Polyploid and Hybrid Genomics*, 153-169.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics*, 6 (11), 836-846.

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M., & Dopazo, J. (2015). A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus Citrus. *Molecular Biology and Evolution*, 32 (8), 2015-2035.

Coyne, J. A., & Orr, H. A. (2004). Speciation (Sinauer Associates), *Sunderland*, MA, 276, 281.

Cronberg, N., & Natcheva, R. (2002). Hybridization between the peat mosses, *Sphagnumcapillifolium* and *S. quinquefarium* (Sphagnaceae, Bryophyta) as inferred by morphological characters and isozyme markers. *Plant Systematics and Evolution*, 234 (1-4), 53-70.

Czarna, A., Gawrońska, B., Nowińska, R., Morozowska, M., & Kosiński, P. (2016). Morphological and molecular variation in selected species of *Erysimum* (Brassicaceae) from Central Europe and their taxonomic significance. *Flora-Morphology, Distribution, Functional Ecology of Plants*, 222, 68-85.

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., ... & Salazar, C. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487 (7405), 94.

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., ... & Schmidt, J. M. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354 (6311), 477-481.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6 (5), 361-375.

Denk, T., & Grimm, G. W. (2010). The oaks of western Eurasia: traditional classifications and evidence from two nuclear markers. *Taxon*, 59 (2), 351-366.

Ding, M., & Chen, Z. J. (2018). Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. *Current Opinion in Plant Biology*, 42, 37-48.

Dobeš, C. H., Mitchell-Olds, T., & Koch, M. A. (2004). Extensive chloroplast haplotype variation indicates Pleistocene hybridization and radiation of North American *Arabis drummondii*, *A. divaricarpa*, and *A. holboellii* (Brassicaceae). *Molecular Ecology*, 13 (2), 349-370.

Dobzhansky, T. 1953. Natural hybrids of two species of *Arctostaphylos* in the Yosemite region of California. *Heredity* 7: 73-79.

Drábková, L. Z., Kirschner, J., Štěpánek, J., Záveský, L., & Vlček, Č. (2009). Analysis of nrDNA polymorphism in closely related diploid sexual, tetraploid sexual and polyploid species. *Plant Systematics and Evolution*, 278 (1-2), 67-85.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28 (8), 2239-2252.

Elworth, R. L., Ogilvie, H. A., Zhu, J., & Nakhleh, L. (2019). Advances in computational methods for phylogenetic networks in the presence of hybridization. *Bioinformatics and Phylogenetics* (pp. 317-360). Springer, Cham.

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107 (1), 1-15.

Fehrer, J., Gemeinholzer, B., Chrtek Jr, J., & Bräutigam, S. (2007). Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in *Pilosella* hawkweeds (Hieracium, Cichorieae, Asteraceae). *Molecular Phylogenetics and Evolution*, 42 (2), 347-361.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., ... & Mitchell, S. N. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347 (6217), 1258524.

Fort, A., Ryder, P., McKeown, P. C., Wijnen, C., Aarts, M. G., Sulpice, R., & Spillane, C. (2016). Disaggregating polyploidy, parental genome dosage and hybridity contributions to heterosis in *Arabidopsis thaliana*. *New Phytologist*, 209 (2), 590-599.

Freedman, A. H., Gronau, I., SchMédail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, 36 (7), 1333-1345.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., ... & Beale, H. (2014). Genome sequencing highlights the dynamic early history of dogs. *PLoS Genetics*, 10 (1).

Gerard, D., Gibbs, H. L., & Kubatko, L. (2011). Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evolutionary Biology*, 11 (1), 291.

Ghaempanah, S., Vaezi, J., Ejtehadi, H., Farsi, M., & Joharchi, M. R. (2012). Morphometric study of species *Erysimum* L. (Brassicaceae) in the provinces of North, Razavi and South Khorassan. *Journal of Taxonomy and Biosystematics*, 10, 77-94.

Ghaempanah, S., Ejtehadi, H., Vaezi, J., & Farsi, M. (2013). Seed-coat anatomy and microsculpturing of the genus *Erysimum* (Brassicaceae) in Northeast of Iran. *Phytotaxa*, 150 (1), 41-53.

Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., ... & Ranwez, V. (2019). Pervasive hybridizations in the history of wheat relatives. *Science Advances*, 5 (5), eaav9188.

Gómez, J. M., Perfectti, F., & Camacho, J. P. M. (2006). Natural selection on *Erysimum mediohispanicum* flower shape: insights into the evolution of zygomorphy. *The American Naturalist*, 168 (4), 531-545.

Gómez, J. M., Bosch, J., Perfectti, F., Fernández, J. D., Abdelaziz, M., & Camacho, J. P. M. (2008). Spatial variation in selection on corolla shape in a generalist plant is promoted by the preference patterns of its local pollinators. *Proceedings of the Royal Society B: Biological Sciences*, 275 (1648), 2241-2249.

Gómez, J. M., Abdelaziz, M., Camacho, J. P. M., Muñoz-Pajares, A. J., & Perfectti, F. (2009). Local adaptation and maladaptation to pollinators in a generalist geographic mosaic. *Ecology Letters*, 12 (7), 672-682.

Gómez, J. M., & Perfectti, F. (2010). Evolution of complex traits: the case of *Erysimum* corolla shape. *International Journal of Plant Sciences*, 171 (9), 987-998.

Gómez, J. M., Perfectti, F., & Klingenberg, C. P. (2014). The role of pollinator diversity in the evolution of corolla-shape integration in a pollination-generalist plant clade. *Phil. Trans. R. Soc. B* 369: 20130257.

Gómez, J. M., González-Mejias A, Lorite J, Abdelaziz M, Perfectti F. (2015a). The silent extinction: Climate change and the potential for hybridization-mediated extinction of endemic high-mountain plants. *Biodiversity and Conservation* 24: 1843–1857.

Gómez, J. M., Perfectti, F., Abdelaziz, M., Lorite, J., Muñoz-Pajares, A. J., & Valverde, J. (2015b). Evolution of pollination niches in a generalist plant clade. *New Phytologist*, 205 (1), 440-453.

Goulet, B. E., Roda, F., & Hopkins, R. (2017). Hybridization in plants: old ideas, new techniques. *Plant Physiology*, 173 (1), 65-78.

Grandont, L., Jenczewski, E., & Lloyd, A. (2013). Meiosis and its deviations in polyploid plants. *Cytogenetic and Genome Research*, 140 (2-4), 171-184.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... & Hansen, N. F. (2010). A draft sequence of the Neandertal genome. *Science*, 328 (5979), 710-722.

Greuter, W., Burdet, H. M., & Long, G. (1986). Dicotyledones (Convolvulaceae-Labiatae). *Med-Checklist*, 3, 106-116.

Grimm, G. W., & Denk, T. (2008). ITS evolution in Platanus (Platanaceae): homoeologues, pseudogenes and ancient hybridization. *Annals of Botany*, 101 (3), 403-419.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., & Li, W. H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421 (6918), 63-66.

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., ... & Koch, M. A. (2017). Plastome phylogeny and early diversification of Brassicaceae. *BMC genomics*, 18 (1), 176.

Hanušová, K., Ekrt, L., Vit, P., Kolář, F., & Urfus, T. (2014). Continuous morphological variation correlated with genome size indicates frequent introgressive hybridization among *Diphasiastrum* species (Lycopodiaceae) in Central Europe. *PLoS One*, 9 (6), e99552.

Harris, S. A., & Ingram, R. (1991). Chloroplast DNA and biosystematics: the effects of intraspecific diversity and plastid transmission. *Taxon*, 40 (3), 393-412.

Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, 9 (6).

Harpke, D., & Peterson, A. (2006). Non-concerted ITS evolution in *Mammillaria* (Cactaceae). *Molecular Phylogenetics and Evolution*, 41 (3), 579-593.

Holland, B. R., Benthin, S., Lockhart, P. J., Moulton, V., & Huber, K. T. (2008). Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology*, 8 (1), 202.

Hopkins, R., & Rausher, M. D. (2012). Pollinator-mediated selection on flower color allele drives reinforcement. *Science*, 335 (6072), 1090-1092.

Hopkins, R., & Rausher, M. D. (2014). The cost of reinforcement: selection on flower color in allopatric populations of *Phlox drummondii*. *The American Naturalist*, 183 (5), 693-710.

Hörandl, E., Paun, O., Johansson, J. T., Lehnebach, C., Armstrong, T., Chen, L., & Lockhart, P. (2005). Phylogenetic relationships and evolutionary traits in *Ranunculus* sl (Ranunculaceae) inferred from ITS sequence analysis. *Molecular Phylogenetics and Evolution*, 36 (2), 305-327.

Hou, C., Wikström, N., Strijk, J. S., & Rydin, C. (2016). Resolving phylogenetic relationships and species delimitations in closely related gymnosperms using high-throughput NGS, Sanger sequencing and morphology. *Plant Systematics and Evolution*, 302 (9), 1345-1365.

Howard, D. J. (1993). Reinforcement: origin, dynamics, and fate of an evolutionary hypothesis. *Hybrid Zones and the Evolutionary Process*, 46-69.

Hughes, C. E., Eastwood, R. J., & Donovan Bailey, C. (2006). From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361 (1465), 211-225.

Jones, M. R., Mills, L. S., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J., ... & Good, J. M. (2018). Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*, 360 (6395), 1355-1358.

Keuler, R., Garretson, A., Saunders, T., Erickson, R. J., Andre, N. S., Grewe, F., ... & Leavitt, S. D. (2020). Genome-scale data reveal the role of hybridization in lichen-forming fungi. *Scientific Reports*, 10 (1), 1-14.

Kim, S. C., & Rieseberg, L. H. (1999). Genetic architecture of species differences in annual sunflowers: implications for adaptive trait introgression. *Genetics*, 153 (2), 965-977.

Kim, S. T., & Donoghue, M. J. (2008a). Incongruence between cpDNA and nrITS trees indicates extensive hybridization within *Eupersicaria* (Polygonaceae). *American Journal of Botany*, 95 (9), 1122-1135.

Kim, M., Cui, M. L., Cubas, P., Gillies, A., Lee, K., Chapman, M. A., ... & Coen, E. (2008b). Regulatory genes control a key morphological and ecological trait transferred between species. *Science*, 322 (5904), 1116-1119.

Kim, E., & Donohue, K. (2011a). Population differentiation and plasticity in vegetative ontogeny: Effects on life-history expression in *Erysimum capitatum* (Brassicaceae). *American Journal of Botany*, 98 (11), 1752-1761.

Kim, E., & Donohue, K. (2011b). Demographic, developmental and life-history variation across altitude in *Erysimum capitatum*. *Journal of Ecology*, 99 (5), 1237-1249.

Kim, E., & Donohue, K. (2012). The effect of plant architecture on drought resistance: implications for the evolution of semelparity in *Erysimum capitatum*. *Functional Ecology*, 26 (1), 294-303.

Kim, E., & Donohue, K. (2013). Local adaptation and plasticity of *Erysimum capitatum* to altitude: its implications for responses to climate change. *Journal of Ecology*, 101 (3), 796-805.

Koch, M. A., & Al-Shehbaz, I. A. (2016). Molecular systematics and evolution. *Biology and breeding of crucifers* (pp. 10-27). CRC Press.

Kubatko, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology*, 58 (5), 478-488.

Kondorosi, E., Roudier, F., & Gendreau, E. (2000). Plant cell-size control: growing by ploidy?. *Current Opinion in Plant Biology*, 3 (6), 488-492.

Lamichhaney, S., Han, F., Webster, M. T., Andersson, L., Grant, B. R., & Grant, P. R. (2018). Rapid hybrid speciation in Darwin's finches. *Science*, 359 (6372), 224-228.

Larson, E. L., Tinghitella, R. M., & Taylor, S. A. (2019). Insect hybridization and climate change. *Frontiers in Ecology and Evolution*, 7, 348.

Lay, C. R., Linhart, Y. B., & Diggle, P. K. (2011). The good, the bad and the flexible: plant interactions with pollinators and herbivores over space and time are moderated by plant compensatory responses. *Annals of Botany*, 108 (4), 749-763.

Lay, C. R., Linhart, Y. B., & Diggle, P. K. (2013). Variation among four populations of *Erysimum capitatum* in phenotype, pollination and herbivory over an elevational gradient. *The American Midland Naturalist*, 169 (2), 259-273.

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P. A., & Bernatchez, L. (2020). Using haplotype information for conservation genomics. *Trends in Ecology & Evolution*, 35 (3), 245-258.

Lemmon, E. M., & Juenger, T. E. (2017). Geographic variation in hybridization across a reinforcement contact zone of chorus frogs (*Pseudacris*). *Ecology and Evolution*, 7 (22), 9485- 9502.

Léveillé-Bourret, É., Starr, J. R., Ford, B. A., Moriarty Lemmon, E., & Lemmon, A. R. (2018). Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Systematic Biology*, 67 (1), 94-112.

Lin, H. Y., Hao, Y. J., Li, J. H., Fu, C. X., Soltis, P. S., Soltis, D. E., & Zhao, Y. P. (2019). Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* sl (Theaceae). *Molecular Phylogenetics and Evolution*, 135, 1-11.

Liu, K. J., Steinberg, E., Yozzo, A., Song, Y., Kohn, M. H., & Nakhleh, L. (2015). Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences*, 112 (1), 196-201.

Lexer, C., Welch, M. E., Raymond, O., & Rieseberg, L. H. (2003). The origin of ecological divergence in *Helianthus paradoxus* (Asteraceae): selection on transgressive characters in a novel hybrid habitat. *Evolution*, 57 (9), 1989-2000.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154 (1), 459-473.

Ma, P. F., Zhang, Y. X., Zeng, C. X., Guo, Z. H., & Li, D. Z. (2014). Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe *Arundinarieae* (Poaceae). *Systematic Biology*, 63 (6), 933-950.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46 (3), 523-536.

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, 2 (12), 1940-1955.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20 (5), 229-237.

Martin, N. H., Bouck, A. C., & Arnold, M. L. (2006). Detecting adaptive trait introgression between *Iris fulva* and *I. brevicauli*s in highly selective field conditions. *Genetics*, 172 (4), 2481-2489.

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 32 (1), 244-257.

Marhold, K., & Lihová, J. (2006). Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. *Plant Systematics and Evolution*, 259 (2-4), 143-174.

Mateo, G., Villalba, M. B. C., & Udias, S. L. (1998). Acerca del orófito minusvalorado de la Sierra de Javalambre (Teruel). *Flora Montiberica*, (9), 41-45.

Mayfield, D., Chen, Z. J., & Pires, J. C. (2011). Epigenetic regulation of flowering time in polyploids. *Current Opinion in Plant Biology*, 14 (2), 174-178.

Mayol, M., & Rosselló, J. A. (2001). Why nuclear ribosomal DNA spacers (ITS) tell different stories in *Quercus*. *Molecular Phylogenetics and Evolution*, 19 (2), 167-176.

Mayr, E. (1992). A local flora and the biological species concept. *American Journal of Botany*, 79 (2), 222-238.

Mayr, E. (1999). Systematics and the origin of species, from the viewpoint of a zoologist. Harvard University Press.

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., & Otto, S. P. (2011). Recently formed polyploid plants diversify at lower rates. *Science*, 333 (6047), 1257-1257.

Mayrose, I., Zhan, S. H., Rothfels, C. J., Arrigo, N., Barker, M. S., Rieseberg, L. H., & Otto, S. P. (2015). Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al., 2014. *New Phytologist*, 206 (1), 27-35.

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., & Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, 6 (3), e1038.

Médail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, 36 (7), 1333-1345.

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, 8 (1), 1-11.

Meng, C., & Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology*, 75 (1), 35-45.

Moazzeni, H., Zarre, S., Pfeil, B. E., Bertrand, Y. J., German, D. A., Al-Shehbaz, I. A., ... & Oxelman, B. (2014). Phylogenetic perspectives on diversification and character evolution in the species-rich genus *Erysimum* (Erysimeae; Brassicaceae) based on a densely sampled ITS approach. *Botanical Journal of the Linnean Society*, 175 (4), 497-522.

Moghe, G. D., & Shiu, S. H. (2014). The causes and molecular consequences of polyploidy in flowering plants. *Annals of the New York Academy of Sciences*, 1320 (1), 16-34.

Morales-Briones, D. F., Liston, A., & Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytologist*, 218 (4), 1668-1684.

Moore, R. C., & Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology*, 8 (2), 122-128.

Mutlu, B. (2010). New morphological characters for some *Erysimum* (Brassicaceae) species. *Turkish Journal of Botany*, 34 (2), 115-121.

Neafsey, D. E., Barker, B. M., Sharpton, T. J., Stajich, J. E., Park, D. J., Whiston, E., ... & Heiman, D. (2010). Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Research*, 20 (7), 938-946.

Nieto-Feliner, G. (2003) *Erysimum* L. In: Flora iberica. Vol. IV. Cruciferae-Monotropaceae. Real Jardín Botánico, CSIC, Madrid, pp. 48–76.

Nieto-Feliner, G. N., & Rosselló, J. A. (2007). Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*, 44 (2), 911-919.

Nieto-Feliner, G. N. (2011). Southern European glacial refugia: a tale of tales. *Taxon*, 60 (2), 365-372.

Nosil, P., Harmon, L. J., & Seehausen, O. (2009). Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, 24 (3), 145-156.

Ortiz-Barrientos, D., Grealy, A., & Nosil, P. (2009). The genetics and ecology of reinforcement. *Annals of the New York Academy of Sciences*, 1168 (1), 156-182.

Ortigosa, A. L., & Gómez, J. M. (2010). Differences in the diversity and composition of the pollinator assemblage of two co-flowering congeneric alpine wallflowers, *Erysimum nevadense* and *E. baeticum*. *Flora-Morphology, Distribution, Functional Ecology of Plants*, 205 (4), 266-275.

Osborn, T. C., Pires, J. C., Birchler, J. A., Auger, D. L., Chen, Z. J., Lee, H. S., ... & Martienssen, R. A. (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics*, 19 (3), 141-147.

Otto, S. P., & Whitton, J. (2000). Polyploid incidence and evolution. *Annual Review of Genetics*, 34 (1), 401-437.

Okuyama, Y., Fujii, N., Wakabayashi, M., Kawakita, A., Ito, M., Watanabe, M., ... & Kato, M. (2004). Nonuniform concerted evolution and chloroplast capture: heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian Mitella (Saxifragaceae). *Molecular Biology and Evolution*, 22 (2), 285-296.

Ouarmim, S., Dubset, C., & Vela, E. (2013). Morphological and ecological evidence for a new infraspecific taxon of the wallflower *Erysimum cheiri* (Brassicaceae) as an indigenous endemism of the southwestern Mediterranean. *Turkish Journal of Botany*, 37 (6), 1061-1069.

Palme, A. E., Su, Q., Palsson, S., & Lascoux, M. (2004). Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among *Betula pendula, B. pubescens* and *B. nana*. *Molecular Ecology*, 13 (1), 167-178.

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., ... & Raison, J. M. (2018). A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences*, 115 (11), E2566-E2574.

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, 25 (11), 2337-2360.

Pease, J. B., Haak, D. C., Hahn, M. W., & Moyle, L. C. (2016). Phylogenomics reveals three sources of º adaptive variation during a rapid radiation. *PLoS Biology*, 14 (2).

Pelé, A., Rousseau-Gueutin, M., & Chèvre, A. M. (2018). Speciation success of polyploid plants closely relates to the regulation of meiotic recombination. *Frontiers in Plant Science*, 9, 907.

Piippo, S., Huhta, A. P., Rautio, P., & Tuomi, J. (2005). Resource availability at the rosette stage and apical dominance in the strictly biennial *Erysimum strictum* (Brassicaceae). *Canadian Journal of Botany*, 83 (4), 405-412.

Polatschek, A. (1978). Die arten der gattung *Erysimum* auf der Iberischen Halbinsel. *Annalen des Naturhistorischen Museums in Wien*, 325-362.

Polatschek, A., (1986). *Erysimum*. In: Greuter W, Burdeth H, Long G, eds., Med-Checklist 3. *Willdenowia* 16: 107–116.

Polatschek, A., Snogerup, S. (2002). *Erysimum*. In: Strid A, Tan K eds. Flora Hellenica 2. Königstein: Koeltz Scientific Books, 130–152.

Polatschek, A. (2014). Revision der gattung *Erysimum* (Cruciferae): Nachträge zu den bearbeitungen der Iberischen. Halbinsel und Makaronesiens. *Annalen des Naturhistorischen Museums in Wien. Serie B für Botanik und Zoologie*, 87-105.

Popp, M., & Oxelman, B. (2004). Evolution of an RNA polymerase gene family in *Silene* (Caryophyllaceae)—incomplete concerted evolution and topological congruence among paralogues. *Systematic Biology*, *53* (6), 914- 932.

Radosavljević, I., Bogdanović, S., Celep, F., Filipović, M., Satovic, Z., Surina, B., & Liber, Z. (2019). Morphological, genetic and epigenetic aspects of homoploid hybridization between *Salvia officinalis L*. and *Salvia fruticosa* Mill. *Scientific Reports*, 9 (1), 1-13.

Ramírez-González, R. H., Borrill, P., Lang, D., Harrington, S. A., Brinton, J., Venturini, L., ... & Khedikar, Y. (2018). The transcriptional landscape of polyploid wheat. *Science*, 361 (6403), eaar6089.

Reboud, X., & Zeyl, C. (1994). Organelle inheritance in plants. *Heredity*, 72 (2), 132-140.

Rendón-Anaya, M., Montero-Vargas, J. M., Saburido-Álvarez, S., Vlasova, A., Capella-Gutierrez, S., Ordaz-Ortiz, J. J., ... & Gabaldón, T. (2017). Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biology*, 18 (1), 60.

Rieseberg, L. H., & Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants*.

Rieseberg, L. H. (1991). Homoploid reticulate evolution in *Helianthus* (Asteraceae): evidence from ribosomal genes. *American Journal of Botany*, 78 (9), 1218-1237.

Rieseberg, L. H., Ellstrand, N. C., & Arnold, M. (1993). What can molecular and morphological markers tell us about plant hybridization?. *Critical Reviews in Plant Sciences*, 12 (3), 213-241.

Rieseberg, L. H., & Carney, S. E. (1998). Plant hybridization. *The New Phytologist*, 140 (4), 599-624.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., ... & Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, 301 (5637), 1211-1216.

Rieseberg, L. H., & Willis, J. H. (2007). Plant speciation. *Science*, 317 (5840), 910-914.

Ru, D., Mao, K., Zhang, L., Wang, X., Lu, Z., & Sun, Y. (2016). Genomic evidence for polyphyletic origins and interlineage gene flow within complex taxa: a case study of *Picea brachytyla* in the Qinghai-Tibet Plateau. *Molecular Ecology*, 25 (11), 2373-2386.

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., & Burleigh, J. G. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*, 14 (1), 23.

Runemark, A., Trier, C. N., Eroukhmanoff, F., Hermansen, J. S., Matschiner, M., Ravinet, M., ... & Sætre, G. P. (2018). Variation and constraints in hybrid genome formation. *Nature Ecology & Evolution*, 2 (3), 549-556.

Saari, S., & Faeth, S. H. (2012). Hybridization of Neotyphodium endophytes enhances competitive ability of the host grass. *New Phytologist*, 195 (1), 231-236.

Sang, T., & Zhong, Y. (2000). Testing hybridization hypotheses based on incongruent gene trees. *Systematic Biology*, 49 (3), 422-434.

Sattler, M. C., Carvalho, C. R., & Clarindo, W. R. (2016). The polyploidy and its key role in plant breeding. *Planta*, 243 (2), 281-296.

Schemske, D. W. (2000). Understanding the origin of species 1. *Evolution*, 54 (3), 1069-1073.

Seehausen, O., Koetsier, E., Schneider, M. V., Chapman, L. J., Chapman, C. A., Knight, M. E., ... & Bills, R. (2003). Nuclear markers reveal unexpected genetic variation and a Congolese-Nilotic origin of the Lake Victoria cichlid species flock. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270 (1511), 129-137.

Servedio, M. R., & Noor, M. A. (2003). The role of reinforcement in speciation: theory and data. *Annual Review of Ecology, Evolution, and Systematics*, 34 (1), 339-364.

Shao, C. C., Shen, T. T., Jin, W. T., Mao, H. J., Ran, J. H., & Wang, X. Q. (2019). Phylotranscriptomics resolves interspecific relationships and indicates multiple historical out-of-North America dispersals through the Bering Land Bridge for the genus *Picea* (Pinaceae). *Molecular Phylogenetics and Evolution*, 141, 106610.

Solís-Lemus, C., & Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS genetics*, 12 (3).

Solís-Lemus, C., Yang, M., & Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology*, 65 (5), 843-851.

Solís-Lemus, C., Bastide, P., & Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Molecular Biology and Evolution*, 34 (12), 3292-3298.

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., ... & Soltis, P. S. (2009). Polyploidy and angiosperm diversification. *American Journal of Botany*, 96 (1), 336-348.

Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, *60*, 561-588.

Soltis, P. S., Liu, X., Marchant, D. B., Visger, C. J., & Soltis, D. E. (2014). Polyploidy and novelty: Gottlieb's legacy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369 (1648), 20130351.

Staubach, F., Lorenc, A., Messer, P. W., Tang, K., Petrov, D. A., & Tautz, D. (2012). Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genetics*, 8 (8).

Stebbins Jr, C. L. (1950). Variation and evolution in plants. *Variation and evolution in plants*.

Stebbins, G. L. (1959). The role of hybridization in evolution. *Proceedings of the American Philosophical Society*, 103 (2), 231-251.

Stelkens, R., & Seehausen, O. (2009). Genetic distance between species predicts novel trait expression in their hybrids. *Evolution: International Journal of Organic Evolution*, 63 (4), 884-897.

Straub, S. C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, 99 (2), 349-364.

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., ... & Stedje, B. (2018). Finding evolutionary processes hidden in cryptic species. *Trends in Ecology & Evolution*, 33 (3), 153-163.

Suarez-Gonzalez, A., Hefer, C. A., Christe, C., Corea, O., Lexer, C., Cronk, Q. C., & Douglas, C. J. (2016). Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. átrichocarpa* (black cottonwood). *Molecular Ecology*, 25 (11), 2427-2442.

Suarez-Gonzalez, A. (2017). Adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood) (Doctoral dissertation, University of British Columbia).

Suarez-Gonzalez, A., Lexer, C., & Cronk, Q. C. (2018). Adaptive introgression: a plant perspective. *Biology Letters*, 14 (3), 20170688.

Sun, M., Soltis, D. E., Soltis, P. S., Zhu, X., Burleigh, J. G., & Chen, Z. (2015). Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Molecular Phylogenetics and Evolution*, 83, 156-166.

Svardal, H., Jasinska, A. J., Apetrei, C., Coppola, G., Huang, Y., Schmitt, C. A., ... & Weinstock, G. (2017). Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. *Nature Genetics*, 49 (12), 1705.

Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution*, 3 (2), 170-177.

Than, C., Ruths, D., & Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatic*s, 9 (1), 322.

Thórsson, Æ. T., Salmela, E., & Anamthawat-Jónsson, K. (2001). Morphological, cytogenetic, and molecular evidence for introgressive hybridization in birch. *Journal of Heredity*, 92 (5), 404-408.

Timme, R. E., Bachvaroff, T. R., & Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One*, 7 (1).

Tkach, N., Schneider, J., Döring, E., Wölk, A., Hochbach, A., Nissen, J., ... & Röser, M. (2019). Phylogeny, morphology and the role of hybridization as driving force of evolution in grass tribes *Aveneae* and *Poeae* (Poaceae). *BioRxiv*, 707588.

Toews, D. P., Taylor, S. A., Vallender, R., Brelsford, A., Butcher, B. G., Messer, P. W., & Lovette, I. J. (2016). Plumage genes and little else distinguish the genomes of hybridizing warblers. *Current Biology*, 26 (17), 2313-2318.

Trontelj, P., & Fišer, C. (2009). Perspectives: cryptic species diversity should not be trivialised. *Systematics and Biodiversity*, 7 (1), 1-3.

Turner, B. L. (2006). Taxonomy and nomenclature of the *Erysimum asperum-E. capitatum* complex (Brassicaceae). *Phytologia*, 88, 279-287.

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, 18 (7), 411.

Valverde, J., Calatayud, J., Gómez, J. M., & Perfectti, F. (2014). Variación intraestacional en los visitantes florales de *Erysimum mediohispanicum* en Sierra Nevada. *Revista Ecosistemas*, 23 (3), 83-92.

Valverde, J., Gómez, J. M., & Perfectti, F. (2016). The temporal dimension in individual-based plant pollination networks. *Oikos*, 125 (4), 468-479.

Valverde, J., Perfectti, F., & Gómez, J. M. (2019). Pollination effectiveness in a generalist plant: adding the genetic component. *New Phytologist*, 223 (1), 354-365.

Wang, K., Lenstra, J. A., Liu, L., Hu, Q., Ma, T., Qiu, Q., & Liu, J. (2018). Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. *Communications Biology*, 1 (1), 1-9.

Washburn, J. D., & Birchler, J. A. (2014). Polyploids as a "model system" for the study of heterosis. *Plant Reproduction*, 27 (1), 1-5.

Warwick, S. I., Francis, A., & Al-Shehbaz, I. A. (2006). Brassicaceae: species checklist and database on CD-Rom. *Plant Systematics and Evolution*, 259 (2-4), 249-258.

White TJ, Bruns T, Lee S, Taylor JW (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds). PCR Protocols: A Guide to Methods and Applications. Academic Press: San Diego, CA, USA, pp 315–322.

Whitfield, J. B., & Lockhart, P. J. (2007). Deciphering ancient rapid radiations. *Trends in Ecology & Evolution*, 22 (5), 258-265.

Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2006). Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. *The American Naturalist*, 167 (6), 794-807.

Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2010). Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytologist*, 187 (1), 230-239.

Wen, D., Yu, Y., Zhu, J., & Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67 (4), 735-740.

Wendel, J. F., & Doyle, J. J. (1998). Phylogenetic incongruence: window into genome history and molecular evolution. *Molecular systematics of plants II* (pp. 265-296). Springer, Boston, MA.

Wendel, J. F., Lisch, D., Hu, G., & Mason, A. S. (2018). The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Current Opinion in Genetics & Development*, 49, 1-7.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., ... & Ruhfel, B. R. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111 (45), E4859-E4868.

Willyard, A., Cronn, R., & Liston, A. (2009). Reticulate evolution and incomplete lineage sorting among the ponderosa pines. *Molecular Phylogenetics and Evolution*, 52 (2), 498-511.

Winkworth, R. C., Bryant, D., Lockhart, P. J., Havell, D., & Moulton, V. (2005). Biogeographic interpretation of splits graphs: least squares optimization of branch lengths. *Systematic Biology*, 54 (1), 56-65.

Wissemann, V. (2007). Plant evolution by means of hybridization. *Systematics and Biodiversity*, 5 (3), 243-253.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., & Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, 106 (33), 13875-13879.

Won, H., & Renner, S. S. (2005). The internal transcribed spacer of nuclear ribosomal DNA in the gymnosperm Gnetum. *Molecular Phylogenetics and Evolution*, 36 (3), 581-597.

Xiao, L. Q., Möller, M., & Zhu, H. (2010). High nrDNA ITS polymorphism in the ancient extant seed plant *Cycas*: incomplete concerted evolution and the origin of pseudogenes. *Molecular Phylogenetics and Evolution*, 55 (1), 168-177.

Yang, Y., & Smith, S. A. (2013). Optimizing *de novo* assembly of short-read RNA-Seq data for phylogenomics. *BMC Genomics*, 14 (1), 328.

Yakimowski, S. B., & Rieseberg, L. H. (2014). The role of homoploid hybridization in evolution: a century of studies synthesizing genetics and ecology. *American Journal of Botany*, 101 (8), 1247-1258.

Yu, Y., Than, C., Degnan, J. H., & Nakhleh, L. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60 (2), 138-149.

Yu, Y., Barnett, R. M., & Nakhleh, L. (2013). Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology*, 62 (5), 738-751.

Zhang, R., Liu, T., Wu, W., Li, Y., Chao, L., Huang, L., ... & Zhou, R. (2013). Molecular evidence for natural hybridization in the mangrove fern genus *Acrostichum*. *BMC Plant biology*, 13 (1), 74.

Zheng, X., Cai, D., Yao, L., & Teng, Y. (2008). Non-concerted ITS evolution, early origin and phylogenetic utility of ITS pseudogenes in *Pyrus*. *Molecular Phylogenetics and Evolution*, 48 (3), 892-903.

Züst, T., Mirzaei, M., & Jander, G. (2018). *Erysimum cheiranthoides*, an ecological research system with potential as a genetic and genomic model for studying cardiac glycoside biosynthesis. *Phytochemistry Reviews*, 17 (6), 1239-1251.

Züst, T., Strickler, S. R., Powell, A. F., Mabry, M. E., An, H., Mirzaei, M., ... & Petschenka, G. (2020). Independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum*, Brassicaceae). *eLife*, 9, e51712.

Materials and Methods

Study species: *Erysimum* spp.

Erysimum L. comprises more than 200 species, being one of the largest genera of the plant family Brassicaceae. This genus is distributed mainly in Eurasia, but some species occur in North America and North Africa (Warwick et al., 2006). Notably, more than a hundred species have been described in the Mediterranean region (Greuter et al., 1986). *Erysimum* spp. are particularly abundant in the Iberian Peninsula, where, depending on the author, between twenty-one and twenty-three species can be identified (Polatschek 1979, 2014; Nieto-Feliner, 2003; Mateo et al., 1998). The Baetic Mountains in SE Iberia are particularly rich in *Erysimum* diversity, and ten species have been described, eight of them endemic species of this area (Blanca et al., 2009; Morales-Torres 2011).

We have focused our study in the following *Erysimum* species:

E. nevadense (Reut, 1855)

Biennial or perennial, monocarpic, or polycarpic herb with yellow corolla. *E. nevadense* grows mainly in xeroacanthic formations at high altitude, on schist and limestone. It is a diploid species with 2n=14 inhabiting the Sierra Nevada (Granada, Spain) and Sierra de Gádor (Almería, Spain), 1700-2800 m.a.s.l. (Nieto-Feliner 2003).

E. mediohispanicum (Polatschek, 1979)

Biennial or perennial plant, usually monocarpic. Yellow corolla. Hipotetraploid (2n=26, 28) in the Iberian Plateau but usually diploid (2n=14) in the Baetic populations (Nieto-Feliner 2003). Cleared shrublands and fallow, moorlands, mainly on limestone. Submontane bands of the Iberian Peninsula 600-1700 m.a.s.l.

E. fitzii (Polatschek, 1979) = *Erysimum nevadense* subsp. *fitzii* (Polatschek) P.W.Ball.

Perennial, polycarpic with yellow corolla. Most often found in shrublands on limestone substrates. Microendemism inhabiting the Sierra de la Pandera (Jaén, Spain). 1200-1800 m.a.s.l. Diploid with 2n=14 (Nieto-Feliner, 2003).

E. baeticum (Heywood, 1979)

Biennial or perennial herb, usually monocarpic with purple corolla. It usually occurs in high-mountain shrublands and perennial subalpine grasslands. It is restricted to metamorphic substrates (mica schists and quartzites). Narrow endemism of the eastern part of the Sierra Nevada (Granada and Almería provinces). 1500–2600 m.a.s.l. 2n=28, 56 (Nieto-Feliner, 2003).

E. bastetanum (Blanca et al., 1992),

(Lorite, Perfectti & Gómez comb. & stat. Nov, 2015)

Biennial herb, usually monocarpic with purple corolla. *E. bastetanum* inhabits gaps of holm-oak, mixed pine forests, and shrublands, mainly on limestones (rarely on mica schists and quartzites). It is distributed across the eastern part of Baetic mountains (Sierra de Baza, Sierra de Filabres, Mencal, Sierra de María-Orce, Sierra Jureña). 800–2200 ma.s.l. Until 2015, it was considered a subspecies of *E. baeticum* (Lorite et al. 2015).

E. popovii (Rothm, 1940)

Perennial, polycarpic herb with purple corolla. It inhabits shrublands on limestone substrates. Distributed across the Baetic Mountains (Jaén, west of Granada and southeast of Córdoba: Sierra Mágina, Jabalcuz, Sierra Harana, and Horconera). 500-2000 m.a.s.l. 2n=42 (Nieto-Feliner, 2003).

E. lagascae (Rivas Goday and Bellot, 1942)

Biennial or perennial monocarpic herb with purple corolla. It grows mainly on steep, rocky slopes and rock cracks, generally on siliceous sandy soils. South-west of the Iberian Peninsula. 400-1350 m.a.s.l. 2n=14 (Nieto-Feliner, 2003).

Study area

Our main study area was the Baetic Ranges, one of the main glacial refugia in Europe, located in the South of the Iberian Peninsula (Médail and Diadema, 2009). Across this region, we sampled three different populations of *E. baeticum*, *E. bastetanum*, *E. mediohispanicum*, *E. nevadense*, and *E. popovii*, and one population of *E. fitzii*. Also, we sampled one population of *E. lagascae*, which occurs outside of this refugium region (**Table 1**, **Figure 1**). Some species population pairs occur in sympatry (**Figure 2**). Specifically: *E. nevadense* (En12) with *E. baeticum* (Ebb10), *E. bastetanum* (Ebt13) with *E. mediohispanicum* (Em71), and *E. popovii* (Ep20) with *E. mediohispanicum* (Em39).

Figure 1. Map of the Iberian Peninsula showing the location of the sampled populations. The insert shows a more detailed map of the Baetic mountains. The populations Ebt13 – Em71, Ebb10 – En12, and Em39 – Ep27 represents population pairs located in sympatry.

Species	Population	Location	Elevation	Geographical coordinates	Flower color	Sympatry with
E. baeticum	Ebb07	Sierra Nevada, Almería, Spain	2128	37°05′46″N, 3°01′01″W	purple	
	Ebb10	Sierra Nevada, Almería, Spain	2140	37°05′32″N, 3°00′40″W	purple	En12
	Ebb12	Sierra Nevada, Almería, Spain	2264	37°05′51″N, 2°58′06″W	purple	
E. bastetanum	Ebt01	Sierra de Baza, Granada, Spain	1990	37°22′52″N, 2°51′49″W	purple	
	Ebt12	Sierra de María, Almería, Spain	1528	37°41′03″N, 2°10′51″W	purple	
	Ebt13	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W	purple	Em71
E. fitzii	Ef01	Sierra de la Pandera, Jaén, Spain	1804	37°37′56″N, 3°46′46″W	yellow	
E. lagascae	Ela07	Sierra San Vicente, Toledo, Spain	516	44°05′49″N, 4°40′40″W	purple	
E. mediohispanicum	Em21	Sierra Nevada, Granada, Spain	1723	37°08′04″N, 3°25′43″W	yellow	
	Em39	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W	yellow	Ep20
	Em71	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W	yellow	Ebt13
E. nevadense	En05	Sierra Nevada, Granada, Spain	2074	37°06′35″N, 3°01′32″W	yellow	
	En10	Sierra Nevada, Granada, Spain	2321	37°06′37″N, 3°24′18″W	yellow	
	En12	Sierra Nevada, Granada, Spain	2255	37°05′37″N, 2°56′19″W	yellow	Ebb10
E. popovii	Ep16	Jabalcuz, Jaén, Spain	796	37°45′26″N, 3°51′02″W	purple	
	Ep20	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W	purple	Em39
	Ep27	Llanos del Purche, Granada, Spain	1470	37°07′46″N, 3°28′48″W	purple	

Table 1. Population code, location, and details of sympatry status for all of the populations sampled.

Biological samples

Fresh leaves

In the spring of 2015, we visited each population and collected fresh leaf tissue, about ten leaves, from five individuals. The samples were dried and preserved in silica gel until DNA extraction.

Floral buds

In the Spring of 2015, we sampled no less than five pre-anthesis flower buds from one individual per population, making sure they were at a similar development stage (based on the size and development of the buds). The buds were immediately submerged in liquid nitrogen and, after arriving to the laboratory, maintained in an ultra freezer (-80 C) until RNA extraction.

Plant rosettes

We sampled five rosettes per population during the fall of 2016. We maintained the plants in a common garden (Science Faculty facilities, Universidad de Granada).

Laboratory procedures

DNA and RNA extraction

DNA extraction

We carried out two rounds of DNA extraction from leaf samples. First, we extracted DNA for *E. baeticum*, *E. nevadense*, and *E. mediohispanicum* using an individual sample for each species to sequence total genomic DNA. Second, we extracted DNA from five individuals of three different populations of *E. baeticum*, *E. bastetanum*, *E. mediohispanicum*, *E. nevadense*, and *E. popovii*, and five individuals of one population of *E. fitzii* and *E. lagascae*, amounting a total of 85 samples.

For each sample, we disrupted ~ 60 mg of leaves using a Beadbug microtube homogenizer (Benchmark Scientific, Edison, NJ) with 2 mm steel beads. We used the GenElute

Plant Genomic DNA Miniprep kit (Sigma-Aldrich, St. Louis, MO) for total genomic DNA isolation, following the manufacturer's protocol. The quantity and the quality of the obtained DNA were checked using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States). The integrity of the extracted genomic DNA was checked on agarose gel electrophoresis.

RNA extraction

We extracted RNA from the floral samples. The buds were snap-frozen in liquid nitrogen and ground with mortar and pestle. Total RNA was isolated using the Qiagen RNeasy Plant Mini Kit following the manufacturer's protocol. The RNA obtained's quality and quantity were checked using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States) and analyzed with the Agilent 2100 Bioanalyzer system (Agilent Technologies Inc).

Library preparation and sequencing

Genomic DNA

We sent the isolated DNA to Macrogen (Macrogen Inc., Seoul, South Korea). Library preparation for deep sequencing was carried out using the TruSeq Nano DNA Library Preparation Kit (350 bp insert size). The sequencing of the *E. mediohispanicum*, *E. nevadense*, and *E. baeticum* libraries was performed using the Illumina HiSeq X platform and following the paired-end 150 bp strategy.

ITS1 and ITS2

We amplified by PCR the ITS1 and the ITS2 regions of the 45S genomic rDNA from the extracted DNA of the 85 samples. We used our own custom-tailored long primers that target the conserved flanking rDNA and show a 5’ leading sequence complementary with the adapter sequence of the Nextera XT DNA index primers to amplify these regions. See **Chapter I** for primer sequences and PCR details. We purified the amplified fragments using spin columns

(GenElute TM PCR Clean-Up Kit, Sigma-Aldrich) and checked their quality on agarose gel electrophoresis. Finally, we quantified the starting DNA using the Infinite M200 PRO NanoQuant spectrophotometer (TECAN, Männedorf, Switzerland).

We constructed independent libraries for ITS1 and ITS2 amplicons using the Nextera XT DNA Sample Preparation Kit. We tagged individual amplicons by adding a unique combination of adapter labels to the 3' and 5' ends of the DNA sequence using nine cycles of PCR (See **Chapter I** for PCR details). We purified the tagged amplicons to remove short fragments. Lastly, we quantified each amplicon's final concentration, made them equimolar, and mix in only one library per ITS region.

Sequencing of ITS1 and ITS2 libraries was carried out at the Novogene Bioinformatics Technology Co., Ltd, with an Illumina MiSeq platform (Illumina, USA) using a paired-end 150 bp sequence read run. We obtained an unexpected low sequencing output for *E. mediohispanicum* libraries. Therefore, we sequenced these libraries twice. The second sequencing was done in the Basic biology service of the "Centro de Instrumentación Científica" of the University of Granada (Spain), using the same Illumina Miseq platform and paired-end chemistry.

Transcriptomes

The 17 transcriptomic libraries of floral tissue and subsequent RNA sequencing were conducted at Macrogen Inc. (Seoul, Korea). Library preparation was performed using the TruSeq Stranded Total RNA LT Sample Preparation Kit (Plant). An rRNA-depletion (Ribo-Zero) step was used for mRNA enrichment and to avoid sequencing rRNAs. The 17 libraries' sequencing was carried out using the Hiseq 3000-4000 sequencing protocol and TruSeq 3000-4000 SBS Kit v3 reagent, following a paired-end 150 bp strategy on the Illumina HiSeq 4000 platform.

Data analyses

Read quality analyses

We analyzed the read quality of each raw library with FastQC v0.11.5 (Andrews, 2010). Then, we performed a two-step trimming analysis first trimming the adapters and then trimming the reads based on their quality. For this purpose, we used cutadapt v1.15 (Martin, 2011) and Sickle v1.33 (Joshi and Fass, 2011). Then, we checked the trimming efficiency analyzing the raw reads again with FastQC v0.11.5. See **Chapters from I to IV** for details.

ITS1 and ITS2 analyses

For ITS analyses, we paired and merged the ITS1 and ITS2 forward and reverse reads using Geneious R.11 (Kearse et al. 2012) and BBMerge v37.64 (Bushnell et al., 2017). Then, we did a cluster analysis using cd-hit v4.6 (Li and Godzik, 2006) merging sequences that were 99% similar to reduce redundancy and discard low-represented haplotypes as sequencing errors or tag switching events.

We studied the degree of sequence polymorphism for each ITS1 and ITS2 samples:

-First, we aligned the ITS1 and ITS2 sequences from each sample using MAFFT v7.450 (Katoh et al., 2014) with default parameters, generating one alignment per species/marker. Then, we trimmed the alignments using trimAl v1.2 (Capella-Gutiérrez et al., 2009), removing gaps with the "gappyout" method.

-To estimate nucleotide and haplotype diversity, we used the R package PEGAS v0.1 (Paradis, 2010). We used the "nuc.div" function to calculate nucleotide diversity (π) and the "hap.div" function to estimated haplotype diversity (Hd).

-We then used the "haplotype" and "haploNet" functions to calculate the total number of haplotypes and the haplotype frequencies distribution.

-Lastly, we used the "amova" function from the R package PEGAS v0.1 (Paradis, 2010) to perform a hierarchical analysis of molecular variance (AMOVA; Excoffier, 1992).

Transcriptome assembly and annotations

We followed a *de novo* assembly approach using Trinity v2.8.4 (Grabherr et al., 2011; Haas et al., 2013). First, to validate and reduce the number of reads, we normalized the libraries in silico before assembly, using the “insilico_read_normalization.pl” function of Trinity v2.8.4. Then, to eliminate single-occurrence k-mers that could be heavily enriched in sequencing errors, we used the parameter ‘min_kmer_cov 2’, following the approach of Haas et al. (2013). Thus, only k-mers that occur more than once were considered for contigs.

We used TransDecoder v5.2.0 (Haas et al. 2013) to predict and translate candidate open reading frames (ORFs) within transcript sequences. Then, we performed functional annotation of those Trinity transcripts with ORFs using Trinotate v3.0.1 (Haas, 2015). Sequences were examined against UniProt (UniProt Consortium, 2014), using SwissProt databases (Bairoch and Apweiler, 2000) with BLASTX and BLASTP searching and an e-value cutoff of 10. We then used the Pfam database (Bateman et al., 2004) to annotate protein domains for each predicted protein sequence. We also annotated the transcripts using the databases eggnog (Jensen et al., 2007), GO (Gene Ontology Consortium, 2004), and Kegg (Kanehisa and Goto, 2000). We used BUSCO v2.0 (Waterhouse, 2018) to validate the quality of all the assemblies, using the plant database brassicales_odb10.2019-11-20.

Chloroplast genome analyses

Chloroplast assembly

Assembly chloroplast genomes from DNA genomic libraries:

We used the NOVOPlasty pipeline v6.2.3 (Dierckxsens et al., 2017) to assemble *de novo* the chloroplast genomes of *E. mediohispanicum*, *E. nevadense*, and *E. baeticum*. In brief, this pipeline assembles a chloroplast genome from whole-genome sequencing data, starting from a related single reference sequence iteratively and bidirectionally until the circular genome is obtained. We used *Arabidopsis thaliana* chloroplast genome sequence (NC_000932.1) as a reference, and the untrimmed genomic reads as assembly elements as recommended by Dierckxsens et al., (2017). We specified the following parameters: 'automatic insert size detection', a 'genome range' from 120000 to 200000, a 'K-mer value' of 39, and an 'insert range' of 1.6, a 'strict insert range' of 1.2, and the 'paired-end reads' option.

Assembly chloroplast genomes from RNA-Seq libraries:

We used a read-mapping approach, using the RNA-Seq trimmed reads and the *E. mediohispanicum* chloroplast genome previously assembled from the genomic libraries. We used the read mapper of Geneious R.11 (Kearse et al. 2012) with the highest sensitivity and default parameters. We validated the results obtained with Geneious R.11 by performing a mapping of reads with BWA v0.7.17 (Li and Durbin, 2009).

Chloroplast genome annotation

We annotated the chloroplast genomes using the program cpGAVAS (Liu et al., 2012). The annotations were manually curated using Geneious R.11 (Kearse et al. 2012). All transfer RNA sequences (tRNA) encoded in the cp genomes were verified using tRNAscan-SE v2.0 (Lowe and Chan, 2016) with the default search settings.

Comparative analysis among chloroplast genomes and cross-validation of the methodology

To evaluate whether the chloroplast genomes assembled from RNA-Seq libraries were similar to those obtained from DNA libraries, we performed the following analyses:

1- We compared the chloroplast genomes assembled from DNA and RNA-Seq libraries. First, we used the mVISTA software (Frazer et al., 2004), which compares the sequences from different species by pairwise alignment and allows for the visualization of these alignments with annotations. We used *A. thaliana* cpDNA as a reference (NC_000932.1) and selected a RankVISTA probability threshold of 0.5, and the Shuffle-LAGAN mode for finding rearrangements (inversions, transpositions, and some duplications).

2- We investigated the degree of within-genome variation of the assembled chloroplast genomes obtained from RNA-Seq and DNA libraries. In particular, we performed a reference-guided assembly in which we remapped the quality-trimmed reads to each assembled genome using the Geneious R.11 (Kearse et al. 2012) mapper with medium-low sensitivity and default parameters. Later, we estimated the percentage of pairwise identity of each assembly.

3- We explored the degree of overall sequence variation found within the three replicas of RNA-Seq assembled genomes and the genome assembled from DNA libraries. We aligned the chloroplast genomes using MAFFT v7.450 (Katoh and Standley, 2014) with the following parameters: FFT-NS-2 fast progressive method algorithm, a scoring matrix of 200PAM/k=2, gap open penalty of 1.53, and an offset value of 0.123. Then, we estimated the nucleotide diversity using VariScan v2.0.3 (Vilella et al., 2005).

4- We studied the degree of sequence variation of some relevant chloroplast genes within only the three replicas of RNA-Seq and, after that, including the same genes assembled from DNA libraries. Thus, we extracted and assembled all the chloroplast genes using the HybPiper

pipeline v1.2 (Johnson et al., 2016). This pipeline uses BWA (Li and Durbin, 2009) to align reads to target sequences, and SPAdes (Bankevich et al., 2012) to assemble these reads into contigs. Once cpDNA genes were obtained, we selected 12 genes out of the total: rbcl, psaA, psbA, ndhK, atpA, atpH (with an important function in the photosynthesis process), rpoA, rps3, rrn16S, trnH (as self-replication genes), yfc2 (the largest plastid gene in angiosperms), and matK (the only maturase of higher plants and widely used in angiosperm systematics). Then, we aligned these genes using MAFFT v7.450 (Katoh and Standley, 2014) and calculated the percentage of pairwise identity between the genes obtained from the three RNA-Seq replicas, and the same but including those from genomic libraries.

5- We identified the size and location of repeat sequences –including palindromic, reverse, and direct repeats– and compared them between chloroplast genomes obtained from RNA-Seq and DNA libraries. We used the software REPuter (Kurtz, S., and Schleiermacher, 1999; Kurtz et al., 2001) with the following settings: Hamming distance of 3; 90% or greater sequence identity; and minimum repeat size of 30 bp. Also, simple sequence repeat (SSR) elements were detected using the Perl script MISA (Beier et al., 2017) by setting the minimum number of repeats to 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta- and hexanucleotides, respectively.

6- We analyzed the impact that sequencing depth has on the assemblage of a complete chloroplast genome from RNA-Seq data. We subsampled the transcriptome reads of *E. nevadense* at five levels of sequencing depths (1 M, 5 M, 10 M, 20 M, and 30 M), producing four different sets of transcriptomic reads for each level. These reads were processed and mapped to the cpDNA of *E. mediohispanicum* using Geneious R.11 mapper (Kearse et al. 2012) with medium-low sensitivity and default parameters as previously done. We calculated several mapping-quality indexes (coverage of bases, expected errors, mean confidence, and % of Q 40 positions) with Geneious R.11 (Kearse et al. 2012) and plotted them against the subsampling depth levels.

7- To estimate whether our proposed pipeline allowed for the recovery of complete cpDNA chromosomes from RNA-Seq libraries in other plant species, we downloaded five transcriptomes from the Sequence Read Archive website and processed them with our workflow. We downloaded two *A. thaliana* (SRR6757372; SRR6676021), one *E. cheiri* (SRR5195368), one *Moricandia suffruticosa* (SRR4296233), one *M. arvensis* (SRR4296231), one *Oriza sativa* (SRR7079258), and one *Zea mays* (ERR1407273) transcriptome. These libraries were trimmed and quality filtered using cutadapt v1.15 (Martin, 2011) and Sickle v1.33 (Joshi and Fass, 2011) with the same parameters described above and mapped using Geneious R.11 (Kearse et al. 2012) to cp genomes of the same species (or the closest relative available): Genbank accession NC_000932 for *A. thaliana*, our *E. mediohispanicum* cp genome for the *E. cheiri* sample, *Brassica napus* GQ861354 for the *Moricandia* samples, *Oriza sativa* NC_001320 for *O. Sativa*, and *Z. mays* NC_001666 for *Z. mays*.

Hybridization and introgression analyses

We used different approaches to study the general hybridization scenario for the *Erysimum* species and if there were signatures of introgression in their genomes.

Orthology inference

First, we clustered the translated sequences to reduce redundancy (the ORFs previously obtained with Transdecoder v5.2.0, see above “transcriptome assembly and annotation section”) using cd-hit v4.6 (Li and Godzik, 2006) following the steps of the pipeline described in Yang and Smith (2014). We inferred orthologs using only the CDS (we excluded UTRs and non-coding transcripts), thus limiting the possibility of including sequencing errors (Yang and Smith, 2014). Then, we identified ortholog genes using the OrthoFinder v2.3.3 pipeline (Emms and Kelly, 2015). In brief, this pipeline first searches the orthogroups (a set of protein genes descended from a single gene in the last common ancestor of all the species sampled), making a BLASTp analysis

with the protein sequences as input. In a second step, it clusters and aligns the orthologous sequences using MAFFT v7.450 (Katoh and Standley, 2013) with default parameters. Finally, we obtained a maximum-likelihood phylogenetic gene tree for each orthogroups using IQ-Tree v1.6.1 (Nguyen et al., 2014) and estimated a species tree with all of them using STAG v1.0.0 (Emms and Kelly, 2018). This consensus evolutionary hypothesis was inferred using all the orthogroups that included all species. The resulting species tree was compared with the gene trees obtained using DLCpar v1.1 (Wu et al., 2014), which considered gene duplication, losses, and incomplete lineage sorting (ILS) as potential causes of discordance among the trees.

Phylogenetic reconstructions

Time-calibrated phylogeny reconstruction:

We reconstructed a time-calibrated phylogeny for whole chloroplast genomes (**Chapter III**). We used Beast v2.0 (Drummond and Rambaut, 2007) using the substitution rate for non-codifying plastidial DNA ($1.2–1.7 \times 10^{-9}$ substitutions/site/year; Graur and Li, 2000). We conducted the Bayesian search for tree topologies and node ages during 20,000,000 generations using a strict clock model and a Yule process as prior. The MCMC was sampled every 1,000 generations, discarding a burn-in of 10%. We ascertained chain convergence by checking the MCMC traces with Tracer v1.6.1 (Rambaut et al., 2014).

Coalescent species tree:

We reconstructed a coalescent species tree using ASTRAL v5.6.3 (Mirarab et al., 2014), with default parameters. We used the unrooted gene trees from the orthologous genes previously obtained (see “Orthology inference”). To visualize and edit the species tree, we used FigTree v1.4.0 (Rambaut and Drummond, 2012).

Detecting cytonuclear discordance:

We compared the species tree based on nuclear orthologous genes with the phylogeny obtained with Beast using the whole chloroplast genomes. Statistical comparisons between the two phylogenetic hypotheses were based on the Shimodaira-Hasegawa Test (SH-Test) (Shimodaira and Hasegawa, 1999) from the phangorn v2.5.5 R package (Schliep, 2011). Both phylogenies were also compared visually, plotting them as mirror images with the function cophyloplot, using the R package ape v5.4 (Paradis et al. 2004).

Population structure analyses: Discriminant Analysis of Principal Components

To estimate the degree to which our sampling populations and taxa represented natural groups, we conducted a series of statistical population genetic analyses. First, we run a variant calling analysis, using the assembled transcriptome of *E. lagascae* as a reference because it represents the *a priori* most distant species of this group. We indexed the *E. lagascae* transcriptome using BWA v0.7.17 (Li and Durbin, 2009), and then we mapped all the trimmed raw reads against it using the BWA v0.7.17 “mem” option. We used SAMtools v1.7 (Li et al., 2009) to convert SAM to BAM files and sort the alignment files. Then, we identified the SNP positions within the aligned reads comparing with the reference transcriptome using the SAMtools “mpileup” command, which transposes the mapped data into a sorted BAM file. Lastly, we used bcftools v1.9 to filter the SNPs (Narasimhan et al., 2016), running the SAMtools v1.7 Perl script “vcfutils.pl VarFilter” with default parameters to filter down the candidate variants and to eliminate false positives. Then, we conducted a Discriminant Analysis of Principal Components (Jombart et al., 2010) for the SNP data, using the R package adegenet v2.1.3 (Jombart and Ahmed, 2011) that identifies and describes clusters of genetically related individuals. We set a range of K values from two to seven because K=7 is the number of different species in our dataset. To identify the optimal K number, we selected the model with the lowest BIC.

Introgression analyses

Phylogenetic species networks

We inferred phylogenetic species networks using the software PhyloNet v3.6.9 (Than et al., 2008; Wen et al., 2018), which accounts for incomplete lineage sorting inferring a specified number of hybridization events. To generate the input for PhyloNet, we ultrametricized the trees obtained previously with IQ-Tree v1.6.1 (see Othology inference section), using the "nnls" method in the "force.ultrametric" function within the R package phytools v0.6-99 (Revell, 2012). We inferred the species networks using a maximum pseudo-likelihood method (MPL; Yu and Nakhleh, 2015). We performed the search five times and estimated optimal networks considering a range from 0 to 15 introgression events. We then calculated the best log-likelihood of all the networks by computing the Akaike's Information Criterion (AIC) (Bozdogan, 1987) with the generic function for AIC in the R package stats v3.6.1. We also estimated the more optimal network by slope heuristic of log-likelihood values. The optimal network was then visualized with Dendroscope v3.5.10 (Huson and Scornavacca, 2019).

ABBA-BABA and Fbranches statistic

We evaluated the gene flow between species calculating the D-statistic, also known as the ABBA-BABA statistic (Durand et al., 2011) with the software Dsuite v0.1 (Malinsky, 2019) that allows the assessment of gene flow across large genomic datasets using a standard block-jackknife procedure (as in Green et al., 2010, and Durand et al., 2011). We estimated the statistic "Dmin", which gives the lowest D-statistic value in a given trio, and plotted the introgression among pairs of samples using the ruby script "plot_d.rb".

We also computed the Fbranch statistic implemented in Dsuite v0.1 (Mallinsky et al., 2018; Mallinsky et al., 2019), which allows identifying gene flow events into specific internal branches of a phylogeny. We used the whole chloroplast genomes phylogeny in newick format

specifying which species should be treated as sister taxa (i.e., as P1 and P2) and *E. lagascae* as outgroup.

Introgression in the anthocyanin biosynthetic pathway genes (ABP genes)

To verify whether introgression might have facilitated shifts in corolla color in *Erysimum*, we explored if any of the genes in the anthocyanin biosynthetic pathway (ABP) exhibited significant introgression signals. As a first step, we download the *Arabidopsis thaliana* ABP genes from TAIR (Lamesch et al., 2011): CHI (AT3G55120), CHS (AT5G13930), F3H (AT3G51240), DFR (AT5G42800), ANS (AT4G22880), and UF3GT (AT5G54060). Then, we mapped the trimmed transcriptome raw reads from all samples to the *A. thaliana* genes as a reference using BWA v0.7.17 (Li and Durbin, 2009). We imported the mapping reads to Geneious R.11 (Kearse et al., 2012), and assembled them *de novo* using the Geneious assembler with the highest sensitivity. To eliminate poorly mapped sequences, we clustered the assembled reads using cd-hit v4.6 (with parameter -c = 0.99) (Li and Godzik, 2006) and aligned the cluster reads for each sample using MAFFT v7.450 with default parameters (Katoh and Standley, 2013). We estimated the consensus sequence for each sample in Geneious R.11 using the strict threshold option (bases matching at least 50 % of the sequences) and aligned them using MAFFT v7.450, getting a single alignment per gene. We confirmed the annotations using BLAST (Johnson et al., 2008). Finally, we double-checked manually that the candidate genes appeared in the transcriptome annotations of all the taxa studied.

We estimated the f̂d statistics using the R package HybridCheck v1.0 (Ward and Oosterhout, 2016) for the ABP genes. This statistic is suitable for small genomic regions and proceeds by comparing the observed difference in the number of ABBA and BABA patterns to what would be expected in the event of complete introgression (Martin et al., 2014, Pfeifer and Kapan, 2019). We considered *E. lagascae* as the archaic population (P3), thus acting as the

possible introgression donor. Then, we estimated the f^d for each purple species (P2) and every possible combination as P1.

Signatures of selection in ABP genes

To determine if selection was favoring genomic variants caused by introgression, we estimated the pairwise ratios of synonymous and non-synonymous sites (dN/dS) in each ABP gene in which we found signatures of introgression, using the "dnds" function from the ape v5.4 package in R (Paradis et al., 2004). We also estimated the MacDonald-Kreitman test (MKT; Egea et al., 2008) for each ABP gene, dividing our samples into two groups per gene (purple vs. yellow corollas). We used the *A. thaliana* sequence of each gene as a putatively neutral outgroup (white flowers). We computed the standard MKT with Jukes Cantor's correction using the MKT website (http://mkt.uab.es/mkt/help_mkt.asp; Egea et al., 2008).

Prezygotic barriers study

Pollen tube growth

We collected plant rosettes of *E. mediohispanicum*, *E. bastetanum*, and *E. popovii* from natural populations. We cultivated these plants in a common garden (University of Granada facilities), and before blooming, we excluded them from pollinators, placing them inside a greenhouse. We hand-pollinated the flowers performing four different treatments:

1- Hybrid crosses: We tipped the anther of a flower with a small stick removing the pollen, and we placed it on the stigma of a flower from a different species previously emasculated.

2- Intra-specific crosses: We removed the pollen and placed it on an emasculated flower of the same species but from an individual coming from a different population.

3- Forced selfing crosses: Within a single individual, we emasculated some flowers and hand-pollinated them with pollen from the same individual.

4- Spontaneous selfing crosses: Some flowers were not manipulated and left to self-pollinate.

We collected the *Erysimum* pistils after 72 hours and conserved them in ethanol at 4C until pollen-tube staining. Tube staining was conducted according to the Mori et al. (2006) protocol:

- Each pistil was cleaned in 70% EtOH for 10 minutes, and then rinsed sequentially with 50% and 30% EtOH, and finally distilled water.
- Samples were softened by placing them in a small petri dish with 8 M NaOH for one hour at room temperature (as recommended in Kearns and Inouye, 1993).
- Then, the pistils were rinsed with distilled water for ten minutes, and afterward, the stigmas were incubated with 0.1 % aniline blue in phosphate buffer (pH 8.3) for two hours.
- The final slide preparations were examined under a fluorescence microscope with blue light (410 nm) to observe the pollen tube formation.

Ploidy determination

We used flow cytometry to assess genome size and to estimate DNA ploidy levels. Following Galbraith et al. (1983), we isolated the nuclei from fresh leaf tissues by chopping simultaneously with a razor blade 0.5 cm^2 of leaf and 0.5 cm^2 of an internal reference standard. As an internal reference standard, we used *Solanum lycopersicum* L. 'Stupické' with 2C = 1.96 pg or *Raphanus sativus* L. with 2C = 1.11 pg (Doležel et al., 1992). The nuclei extraction was made on a Petri dish containing 1 ml of WPB buffer (Loureiro et al., 2007). Then, we filtered the nuclear suspension using a 50 µm nylon mesh and stained the DNA 50 µg ml^{-1} of propidium iodide (PI, Fluka, Buchs, Switzerland). We also added 50 µg ml^{-1} of RNAse (Fluka, Buchs, Switzerland) to degrade dsRNA. After 5 min incubation, the samples were analyzed in a Partec CyFlow Space flow cytometer (532 nm green solid-state laser, operating at 30 mW; Partec GmbH., Görlitz, Germany). We used the Partec FloMax software v2.4d (Partec GmbH, Münster, Germany) to acquire the results in the form of four graphics: histogram of fluorescence pulse integral in linear scale (FL); forward light scatter (FS) vs. side light scatter (SS), both in logarithmic (log) scale; FL

vs. time; and FL vs. SS in log scale. To remove the effect of debris, the FL histogram was gated using a polygonal region defined in the FL vs. SS histogram. At least 5,000 particles were analyzed per sample. Only CV values of 2C peak of each sample below 5% were accepted; otherwise, a new sample was prepared and analyzed until quality standards were achieved (Greilhuber et al., 2007). In a few cases, samples produced poorer quality histograms even after repetition due to cytosolic compounds' presence. Thus, it was impossible to estimate the ploidy level and/or genome size for some individuals.

Genome size in mass units (2C in pg; sensu Greilhuber et al., 2005) was obtained using the formula: sample 2C nuclear DNA content (pg) = (sample G1 peak mean / reference standard G1 peak mean) * genome size of the reference standard. The ploidy levels were inferred for each sample based on previous chromosome counts and genome size estimates obtained in the species and genus.

Ploidy determination was performed in the laboratory of Dr. Loureiro (Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Coimbra, Portugal

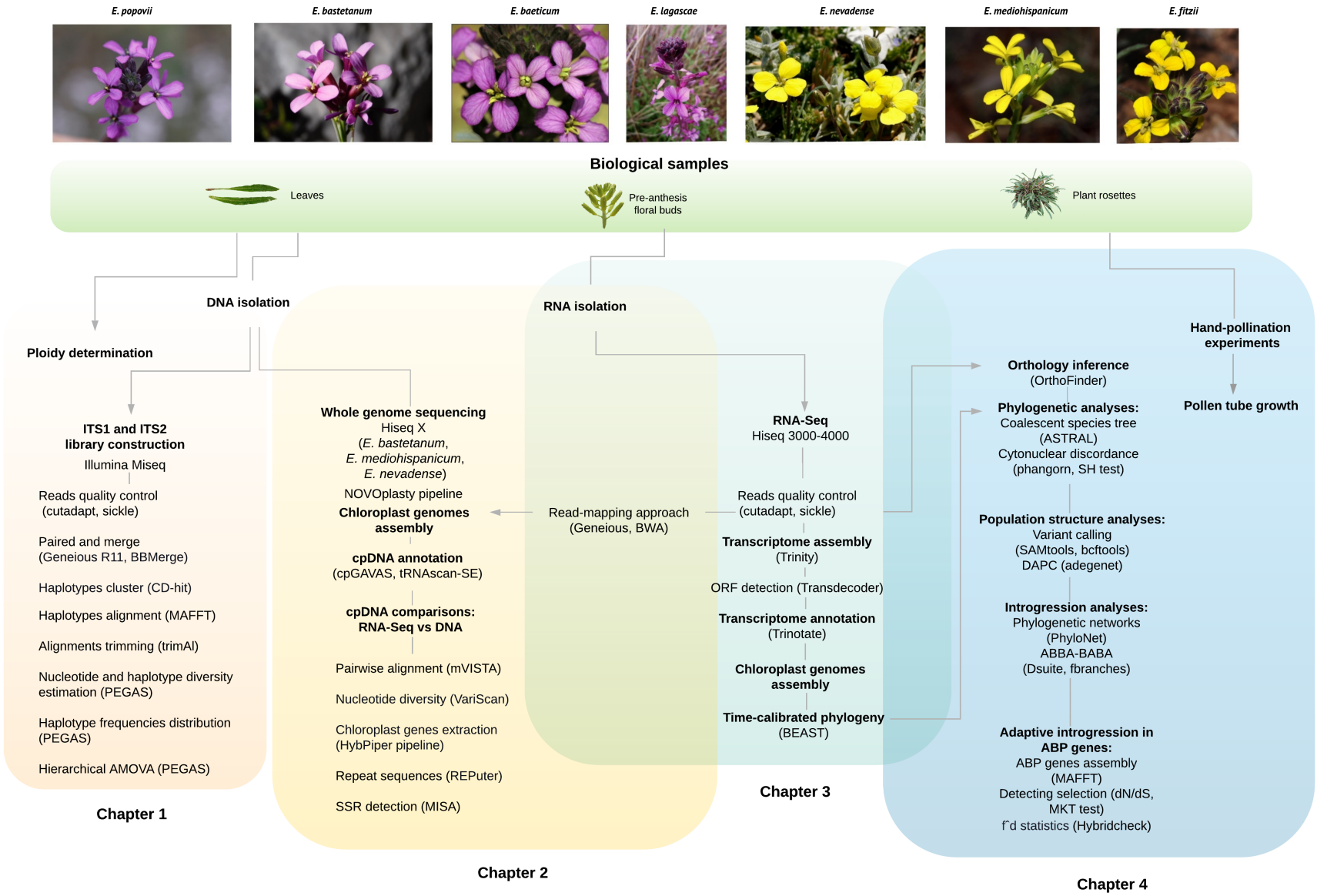

Figure 2. A summary of the materials and methods used in this thesis.

References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28 (1), 45-48.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19 (5), 455-477.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., ... & Studholme, D. J. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32 (suppl_1), D138-D141.

Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics*, 33 (16), 2583-2585.

Blanca, G., Morales, C. & Ruiz-Rejón, M. (1992) El género *Erysimum* L. (Cruciferae) en Andalucía (España). *Anales del Jardín Botánico de Madrid* 49: 201–214.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52 (3), 345-370.

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge–accurate paired shotgun read merging via overlap. *PloS One*, 12 (10).

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25 (15), 1972-1973.

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45 (4), e18-e18.

Doležel, J., Sgorbati, S., & Lucretti, S. (1992). Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiologia Plantarum*, 85 (4), 625-631.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7 (1), 214.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28 (8), 2239-2252.

Egea, R., Casillas, S., & Barbadilla, A. (2008). Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, 36 (suppl_2), W157-W162.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16 (1), 157.

Emms, D., & Kelly, S. (2018). STAG: species tree inference from all genes. *BioRxiv*, 267914.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131 (2), 479-491.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Research*, 32 (suppl_2), W273-W279.

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., & Firoozabady, E. (1983). Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*, 220 (4601), 1049-1051.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32 (suppl_1), D258-D261.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29 (7), 644.

Graur, D., and Li W.H. (2000). Fundamentals of molecular evolution. Second edition. Sunderland: Sinauer Associates, Inc.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... & Hansen, N. F. (2010). A draft sequence of the Neandertal genome. *Science*, 328 (5979), 710-722.

Greilhuber, J., Doležel, J., Lysak, M. A., & Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms ‘genome size’and ‘C-value’to describe nuclear DNA contents. *Annals of Botany*, 95 (1), 255-260.

Greilhuber, J., Temsch, E. M., & Loureiro, J. C. (2007). Nuclear DNA content measurement. *Flow Cytometry with Plant Cells: Analysis of Genes, Chromosomes and Genomes*, 67-101

Greuter, W., Burdet, H. M., & Long, G. (1986). Dicotyledones (Convolvulaceae-Labiatae). *Med-Checklist*, 3, 106-116.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & MacManes, M. D. (2013). *De novo* transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8 (8), 1494.

Haas, B. J. (2015). Trinotate: transcriptome functional annotation and analysis.

Huson, D. H., & Scornavacca, C. (2019). User Manual for Dendroscope V3.6.2.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., & Bork, P. (2007). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36 (suppl_1), D250-D254.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36 (suppl_2), W5-W9.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., ... & Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, 4 (7), 1600016.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11 (1), 94.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27 (21), 3070-3071.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28 (1), 27- 30.

Katoh, K., & Standley, D. M. (2014). MAFFT: iterative refinement and additional methods. In Multiple sequence alignment methods (pp. 131-146). *Humana Press*, Totowa, NJ.

Kearns, C. A., & Inouye, D. W. (1993). *Techniques for pollination biologists*. University press of Colorado.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... & Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28 (12), 1647-1649.

Kurtz, S., & Schleiermacher, C. (1999). REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* (Oxford, England), 15 (5), 426-427.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29 (22), 4633-4642.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., ... & Karthikeyan, A. S. (2011). The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40 (D1), D1202-D1210.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25 (14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25 (16), 2078-2079.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22 (13), 1658-1659.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, 13 (1), 715.

Lorite, J., Perfectti, F., & Gómez, J. M. (2015). A new combination in *Erysimum* (Brassicaceae) for Baetic mountains (South-eastern Spain). *Phytotaxa*, *201* (1), 103-105. Rothm, in Feddes Repert. 49: 180 (1940)

Loureiro, J., Rodriguez, E., Doležel, J., & Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Annals of Botany*, 100 (4), 875-888.

Lowe, T. M., & Chan, P. P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, 44 (W1), W54-W57.

Malinsky, M. (2019). Dsuite-fast D-statistics and related admixture evidence from VCF files. *BioRxiv*, 634477.

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, 2 (12), 1940-1955.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17 (1), 10-12.

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2014). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 32 (1), 244-257.

Mateo, G., Villalba, M. B. C., & Udias, S. L. (1998). Acerca del orófito minusvalorado de la Sierra de Javalambre (Teruel). *Flora Montiberica*, (9), 41-45.

Médail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, 36 (7), 1333-1345.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: genome- scale coalescent-based species tree estimation. *Bioinformatics*, 30 (17), i541-i548.

Morales Torres C. (2011). *Erysimum* L. in Blanca G., Cabezudo B., Cueto M., Salazar C. & Morales Torres C. (eds.). Flora Vascular de Andalucía Oriental. Universidades de Almería, Granada, Jaén y Málaga, Granada.

Mori T., Kuroiwa H., Higashiyama, T., and Kuroiwa T. (2006). Generative Cell Specific 1 is essential for angiosperm fertilization. *Nature Cell Biology* 8 (1): 64-71.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32 (11), 1749- 1751.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32 (1), 268-274.

Nieto-Feliner, G. (2003) *Erysimum* L. In: Castroviejo et al. (eds) Flora iberica. Vol. IV. Cruciferae-Monotropaceae. Real Jardín Botánico, CSIC, Madrid, pp. 48–76.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26 (3), 419-420.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20 (2), 289-290.

Pfeifer, B., & Kapan, D. D. (2019). Estimates of introgression as a function of pairwise distances. BMC *Bioinformatics*, 20 (1), 207.

Polatschek, A. (1979). Die arten der gattung *Erysimum* auf der Iberischen Halbinsel. *Annalen des Naturhistorischen Museums in Wien*, 325-362.

Polatschek, A. (2014). Revision der gattung *Erysimum* (Cruciferae): Nachträge zu den bearbeitungen der Iberischen. Halbinsel und Makaronesiens. *Annalen des Naturhistorischen Museums in Wien. Serie B für Botanik und Zoologie*, 87-105.

Rambaut, A., & Drummond, A. J. (2012). FigTree version 1. 4. 0.

Rambaut, A., Suchard, M. A., Xie, W., & Drummond, A. J. Tracer v1. 6.1. 2014.

Revell, L. J. (2012). Phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3 (2), 217-223.

Rivas Goday, S. & Bellot, F. (1942) Bol. Soc. Esp. Hist. Nat., 40: 57-71.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27 (4) 592-593.

Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16 (8), 1114-1114.

Than, C., Ruths, D., & Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9 (1), 322.

UniProt Consortium. (2014). UniProt: a hub for protein information. *Nucleic Acids Research*, 43 (D1), D204-D212.

Vilella, A. J., Blanco-Garcia, A., Hutter, S., & Rozas, J. (2005). VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, 21 (11), 2791-2793.

Ward, B. J., & Van Oosterhout, C. (2016). HybridCheck: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Molecular Ecology Resources*, 16 (2), 534-539.

Warwick, S. I., Francis, A., & Al-Shehbaz, I. A. (2006). Brassicaceae: species checklist and database on CD-Rom. *Plant Systematics and Evolution*, 259 (2-4), 249-258.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., ... & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, 35 (3), 543-548.

Chapter I

Lack of ITS sequence homogenization in *Erysimum* species with a varying ploidy level

Abstract

The ribosomal DNA internal transcribed spacers have been described as having fast concerted evolution rates, which tend to homogenize their sequences. However, in some cases, such as species of hybrid origin and/or polyploids, sequence homogenization is not complete across ITS copies. Here, we studied the concerted evolution pattern for seven species of the genus *Erysimum* (Brassicaceae), spanning ploidy levels from 2x to 10x. We sampled 85 individuals from different populations of each species and estimated concerted evolution separately in ITS1 and ITS2. Also, we examined if concerted evolution rates varied with ploidy. Our results showed a lack of sequence homogenization, especially for polyploid samples, indicating a lack of concerted evolution in these species. Lack of homogenization was more remarkable for ITS1 than for ITS2, suggesting that concerted evolution is operating more efficiently on the latter. Furthermore, the high haplotype diversity and lack of dominant haplotypes in polyploid and diploid species point to a recent hybrid origin for several *Erysimum* species.

Introduction

Concerted evolution is an evolutionary process by which sequences from the same gene family show higher sequence similarity to each other than to orthologous genes in related species (Elder, 1995; Wendel, 2003; Liao, 2004; Xu et al., 2017). Hence, genes that evolve in a concerted manner present low polymorphism in their sequences, i.e., their sequences are homogenized. Concerted evolution is particularly notable in multicopy nuclear genes, where homogenization is mainly achieved by unequal crossing over and gene conversion (Dover, 1994; Ganley and Kobayashi, 2007). One of the multicopy gene families is the 45S nuclear ribosomal DNA (nrDNA). It appears arranged as tandemly repeated units with hundreds to thousands of copies in one or several loci per genome. These units are composed of the 18S rDNA, internal transcribed spacer 1 (ITS1), 5.8 S rDNA, internal transcribed spacer 2 (ITS2), and 26S rDNA, separated by longer non-transcribed intergenic spacers (Sone et al. 1999). Among all these units, the internal transcribed spacers (ITS1 and ITS2) are the best-characterized nrDNA sequences (Long and Dawid, 1980), partly because ITS sequences show characteristics advantageous for phylogenetic studies, such as biparental inheritance, short length, and high evolution rate (Baldwin et al., 1995; Alvarez and Wendel, 2003; Ganley and Kobayashi, 2007).

ITS sequences usually show fast concerted evolution that leads to low levels of intra-genomic sequence variation. Therefore, these regions have usually been considered to show very few polymorphic positions (Xu et al., 2017; Nieto-Feliner and Rosselló, 2007). However, in some animals (e.g., Teruel et al., 2014) and especially in plants (Buckler and Holtsford, 1996; Mayol and Roselló, 2001; Popp and Oxelman, 2004; Harpke and Peterson, 2006; Grimm and Denk, 2008; Xiao et al., 2010), full sequence homogenization can remain incomplete across ITS sequences, resulting in a relatively high intra-genomic polymorphism. In some instances, the

polymorphic positions found among ITS sequences have been attributed to hybridization events (Alvarez and Wendel, 2003; Bailey et al., 2003; Won and Renner, 2005; Zheng et al. 2008; Drábková et al., 2009). After a hybridization event, different ITS sequences meet and may become homogenized after a time, although this homogenization may not be consistent among descendant lineages (Okuyama et al., 2004). As concerted evolution tends to homogenize sequences rapidly (Alvarez and Wendel., 2003), evidence of non-concerted evolution is mainly found in recently-formed hybrid species, where both parental ITS sequences may still be present. This phenomenon is particularly conspicuous in allopolyploid species, where the occurrence of different ITS sequences located in distinct chromosomes tends to retard this homogenization (Soltis and Soltis, 2009). The level of genetic variation in these sequences will involve a balance between all those processes that homogenize the sequences, including purifying selection and the non-selective processes commonly known under the generic term of molecular impulse (Dover 1982), and, on the other hand, those other events that introduce variation, mainly mutation and hybridization. Therefore, ascertaining the variation of these sequences is crucial to understand what processes are driving concerted evolution in lineages with complex evolution and to know the evolutionary history of these lineages.

Erysimum (Brassicaceae) is a genus of more than 200 species, widely distributed in the N. Hemisphere (Al-Shehbaz, 2012). The Baetic Mountains (SE Iberia), one of the most important glacial refugia in Europe, constitute a hotspot for this group, with ten *Erysimum* species occurring in this small area (Nieto-Feliner, 1993; Médail and Diadema, 2009). Previous studies have suggested that several of these species have a hybrid origin (Abdelaziz, 2013; Pajares, 2013). Furthermore, it has been described that in some instances, ploidy levels vary across these species (Nieto-Feliner, 1993), suggesting that the evolution of this group requires a detailed understating of hybridization and allopolyploidization. However, the effects of hybridization and polyploidization on the genomes of these species have never been elucidated.

In this study, we explored the homogenization degree of ITS1 and ITS2 for seven *Erysimum* species analyzing the polymorphism at the intra-species, intra-population, and intra-individual level, by sequencing both markers by NGS, attempting to recover all ITS copies (Rauscher et al., 2002; Nieto-Feliner and Roselló, 2007). Our objectives were 1) to infer the degree of sequence homogenization on ITS1 and ITS2 for *Erysimum* at the individual, population, and species-level; 2) determine the evolutionary outcome of concerted evolution operating in hybrid and polyploid *Erysimum* spp., thereby providing insight into the consequences of hybridization and polyploidy, and that of the concomitant allopolyploidization, for ITS evolution.

Materials and methods

Taxon sampling

We collected fresh leaves from five individuals of three different populations of *E. baeticum*, *E. bastetanum*, *E. mediohispanicum*, *E. nevadense*, and *E. popovii*, and five individuals of one population of the microendemic *E. fitzii*. Also, we sampled five individuals of *E. lagascae*, a species inhabiting the central Iberian Peninsula. A total of 85 samples were dried and preserved in silica gel until DNA extraction. **Table 1** shows the code, location, and ploidy levels of all the samples.

DNA extraction

We used at least 60 mg of dry plant material for each sample. We disrupted the tissues using a mortar and pestle. Then, total genomic DNA was isolated using the GenElute Plant Genomic DNA Miniprep kit (Sigma-Aldrich, St. Louis, MO) following the manufacturer's protocol. The quantity and the quality of the obtained DNA were checked using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States), and the integrity of the extracted DNA was checked on agarose gel electrophoresis.

Taxon	Population	Location	Elevation	Geographical coordinates	Ploidy level
E. baeticum	Ebb07	Sierra Nevada, Almería, Spain	2128	37°05′46″N, 3°01′01″W	8x
	Ebb10	Sierra Nevada, Almería, Spain	2140	37°05′32″N, 3°00′40″W	8x
	Ebb12	Sierra Nevada, Almería, Spain	2264	37°05′51″N, 2°58′06″W	8x
E. bastetanum	Ebt01	Sierra de Baza, Granada, Spain	1990	37°22′52″N, 2°51′49″W	4x
	Ebt12	Sierra de María, Almería, Spain	1528	37°41′03″N, 2°10′51″W	4x
	Ebt13	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W	8x
E. fitzii	Ef01	Sierra de la Pandera, Jaén, Spain	1804	37°37′56″N, 3°46′46″W	2x
E. lagascae	Ela07	Sierra de San Vicente, Toledo, Spain	516	44°05′49″N, 4°40′40″W	2x
E. mediohispanicum	Em21	Sierra Nevada, Granada, Spain	1723	37°08′04″N, 3°25′43″W	2x
	Em39	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W	2x
	Em71	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W	4x
E. nevadense	En05	Sierra Nevada, Granada, Spain	2074	37°06′35″N, 3°01′32″W	2x
	En10	Sierra Nevada, Granada, Spain	2321	37°06′37″N, 3°24′18″W	2x
	En12	Sierra Nevada, Granada, Spain	2255	37°05′37″N, 2°56′19″W	2x
E. popovii	Ep16	Jabalcuz, Jaén, Spain	796	37°45′26″N, 3°51′02″W	4x
	Ep20	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W	10x
	Ep27	Llanos del Purche, Granada, Spain	1470	37°07′46″N, 3°28′48″W	4x

Table 1. Taxonomic assignment, population code, location, and ploidy level for all of the *Erysimum* spp. populations sampled.

ITS1 and ITS2 amplification

We independently amplified ITS1 and ITS2 in each of the 85 samples. The ITS PCR reactions were performed in a total volume of 25 µl with the following composition: 5 µL 5× buffer containing $MgCl_2$ at 1.5 mmol/L (New England), 0.1 mmol/L each dNTP, 0.2 µmol/L each primer, and 0.02 U Taq high fidelity DNA-polymerase (Q5 New England). We used a set of long primers developed to have a 5' flanking sequence complementary to the Nextera XT DNA index to facilitate adaptor ligation during library construction:

>ITS1-Flabel

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCGTAGGTGAACCTGCGG

>ITS2-Rlabel

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTGCGTTCTTCATCGATGC

>ITS3-Flabel

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCATCGATGAAGAACGCAGC

>ITS4-Rlabel

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCTCCGCTTATTGATATGC.

Reactions included 30 cycles, and the thermocycling conditions were 94°C during 15 s, 60°C for 30 s, and 72°C for 30 s. Amplified fragments were purified using spin columns (GenElute TM PCR Clean-Up Kit, Sigma-Aldrich), and were checked on agarose gel electrophoresis. Finally, we quantified the starting DNA using the Infinite M200 PRO NanoQuant spectrophotometer (TECAN, Männedorf, Switzerland).

Libraries construction

We constructed two libraries, one for ITS1 amplicons and one for ITS2 amplicons. The libraries were prepared using the Nextera XT DNA Sample Preparation Kit. In brief, the DNA was tagged by adding a unique adapter label combination to the 3' and 5' ends of the DNA sequence. Then, the DNA was amplified via a nine cycle PCR. The total volume reaction was 25 µl with the

following composition: 5 µL 10× buffer at 1.0 mmol/L (New England BioLabs), 0.1 mmol/L each dNTP, 0.2 µmol/L each Nextera primer, 0.02 U Taq high fidelity DNA-polymerase (Q5 New England), and 5X Q5 High GC Enhancer (New England BioLabs). PCR thermocycling conditions were 98°C during 5s, 55°C for 10 s, and 72°C for 10 s. After that, we purified both libraries using the GenElute PCR Clean-Up Kit (Sigma) to remove short library fragments. Finally, we generated equal volumes of the libraries to prepared equimolar libraries for sequencing, and the final concentration of each library was quantified using the Infinite M200 PRO NanoQuant spectrophotometer (TECAN, Männedorf, Switzerland).

Library sequencing

ITS1 and ITS2 library sequencing was carried out by Novogene Bioinformatics Technology Co., Ltd, with an Illumina MiSeq platform (Illumina, USA) using a paired-end 150 bp sequence read run. The ITS libraries of *E. mediohispanicum* were sequenced twice due to an unexpected low sequencing output. This sequencing was done using the Illumina Miseq platform and paired-end chemistry in the Scientific Instrumentation Center of the University of Granada, Spain.

Data analysis

FASTQ files were demultiplexed, and read quality was checked in FastQC v0.11.5 (Andrews, 2010). Then, we did a trimming analysis using first cutadapt v1.15 (Martin, 2011) to trim the adapters, followed by a quality trimming using Sickle v1.33 (Joshi and Fass, 2011). Forward and reverse reads were paired in Geneious R.11 (Kearse et al. 2012). Using the function "Set pair read" with default parameters for Illumina paired-end read technology. Then the paired reads were merged using BBMerge v37.64 (Bushnell et al., 2017) with Low Merge rate to decrease false positives. Then, to reduce redundancy and reduce the noise caused by sequencing errors and tag switching events, we did a cluster analysis using "cd hit" (Li and Godzik, 2006). We clustered the sequences from each sample using an identity threshold of 0.99 (i.e., we merged sequences with similarity ≥ 99%) and discarded the clusters that included < 5 % of the total reads (Esling et

al. 2015). This step reduced the contribution of sequencing errors to the reported sequence diversity.

Then, we aligned the sequences from each sample using MAFFT v7.450 (Katoh et al., 2002) with default parameters, generating one alignment per species and marker. We trimmed the alignments using trimAl v1.2 (Capella-Gutiérrez et al., 2009), removing gaps with the "gappyout" method. Then, we used the R package PEGAS v0.1 (Paradis, 2010) to estimate population genetic parameters at intra-species, intra-population, and intra-individual level. We used the "nuc.div" function to calculate nucleotide diversity (π), estimated as the average number of nucleotide differences per site between two sequences (Nei and Li, 1979; Nei and Jin, 1989), and the "hap.div" function to estimated haplotype diversity (Hd), as the probability that two randomly chosen haplotypes from a given population were different. We then used the "haplotype" and "haploNet" functions to calculate the total number of haplotypes and the haplotype frequency distribution for each species, population, and individuals. We studied the normality of our data using the Shapiro-Wilk's method, and then as a non-normal distribution was found, we studied the distribution of them using the Mann-Whitney-Wilcoxon test. Additionally, we studied whether a correlation exists among ploidy levels and haplotype and nucleotide diversity, for ITS1, and ITS2 samples. All statistical analyses were done in R using the package stats v3.6.1 (Team, R. C., 2013). Also, as these species were described as frequently hybridizing, we studied if there were shared haplotypes among different populations from the same species and also among different species. To study that, we estimated the total number of haplotypes and their frequencies in ITS1 and ITS2 samples.

We analyzed the genetic structure of ITS1 and ITS2 samples performing a hierarchical analysis of molecular variance (AMOVA; Excoffier, 1992). We used the "amova" function from the R package PEGAS v0.1 (Paradis, 2010). Thus, we explored the genetic variation explained by populations (i.e., at the population level), among individuals within populations (i.e., at the individual level), and within individuals (i.e., at the intra-individual level). Moreover, we

analyzed the amount of genetic variation that could be explained by differences among species for ITS1 and ITS2, partitioning the variance in three levels: among species, among populations within species, and within populations (i.e., among individuals).

Results

We obtained a total of 84 samples for ITS1, and 81 samples for ITS2, as some samples were not well-sequenced (**Table S1**). The final number of sequences obtained after the quality trimming and the cd-hit clustering for each ITS1 and ITS2 samples were presented in **Table S1**. A mean of 10156 ± 1233 sequences per individual was obtained for ITS1 and 49428 ± 7678 sequences per individual for ITS2.

Polyploid species (*E. baeticum*, *E. bastetanum*, *E. popovii*) presented higher nucleotide diversities (mean π ± SE; 0.012 ± 0.007 for ITS1; 0.003 ± 0.004 for ITS2) than diploid species (0.004 ± 0.006 for ITS1; 0.002 ± 0.003 for ITS2), for both ITS1 and ITS2 samples (**Figure 1**). We have found a significant positive correlation between ploidy level and nucleotide diversity for ITS1 (Spearman's rho: 0.48, p-value: 2.10^{-6}), but a non-significant positive correlation for ITS2 (Spearman's rho: 0.20, p-value: 0.06) (**Figure 2**).

Accordingly, the polyploid population of *E. mediohispanicum* (Em71, 4X) showed higher nucleotide diversity for both ITS1 (mean π = 0.011 ± 0.006) and ITS2 (mean π = 0.003 ± 0.002) than the two diploid populations of this species (Em39: mean π = 0.006 ± 0.007 for ITS1; mean π = 0.0003 ± 0.001 for ITS2; and Em21: mean π = 0.0004 ± 0.001 for ITS1; mean π = 0.001 ± 0.002 for ITS2) (**Figure S1**).

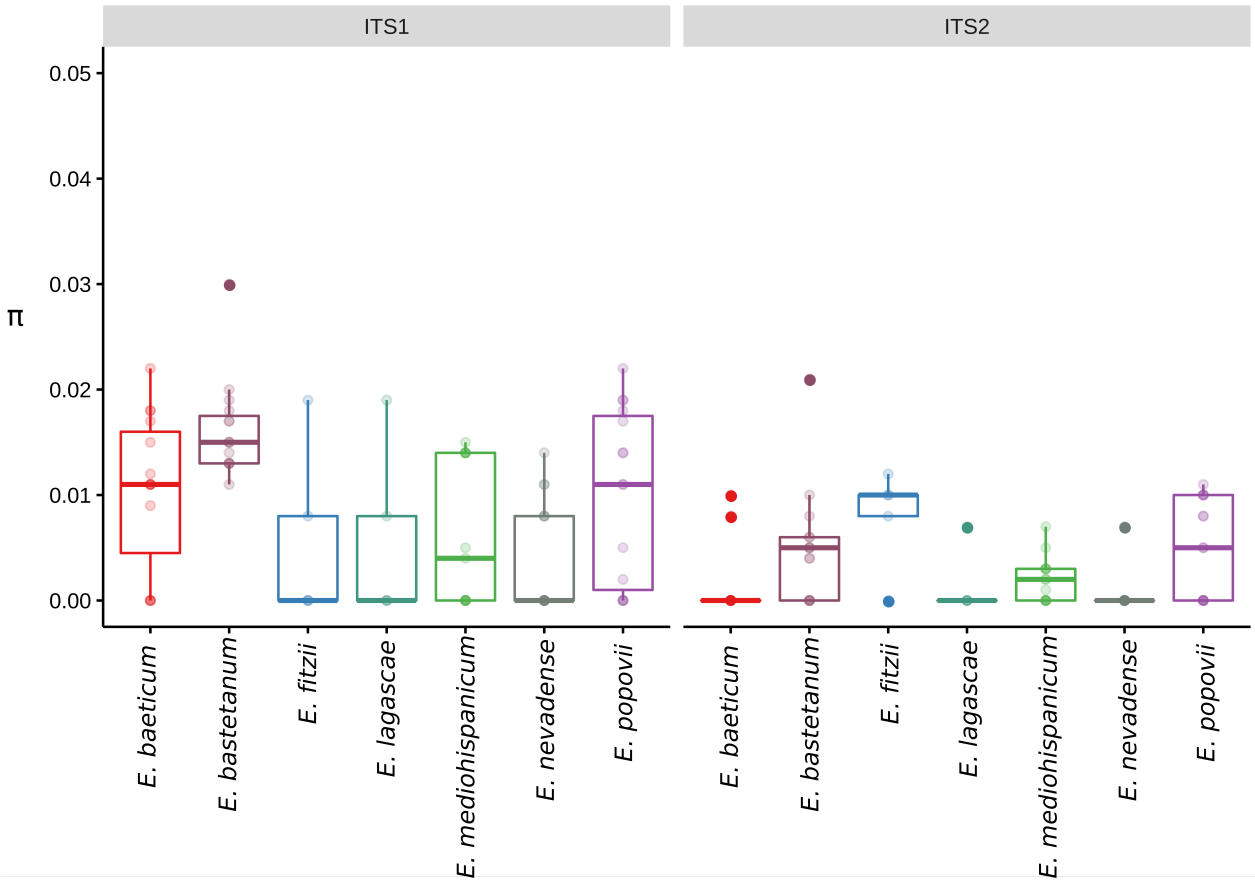

Figure 1. Boxplot depicting the nucleotide diversity (π) for ITS1 and ITS2 samples. Nucleotide polymorphism was estimated for each *Erysimum* individual as the average number of nucleotide differences per site between two sequences (Nei and Li, 1979). *E. baeticum*, *E. bastetanum*, *E. popovii*, and one population of *E. mediohispanicum* are polyploids.

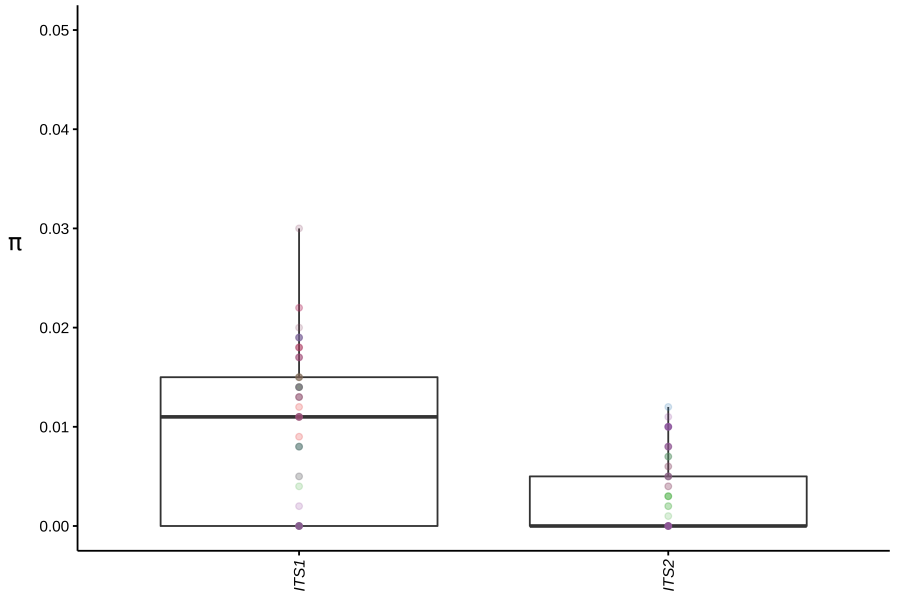

Figure 2. Boxplot depicting the nucleotide diversities (π) for ITS1 and ITS2, estimated for each individual sample.

Haplotype diversity showed a pattern similar to that of the nucleotide diversities, with higher haplotype diversity for polyploid species (mean Hd = 0.89 ± 0.38 for ITS1; mean Hd = 0.50 ± 0.49 for ITS2) than diploid species (mean Hd = 0.39 ± 0.49 for ITS1; mean Hd = 0.28 ± 0.45 for ITS2). Furthermore, we have found a significant positive correlation between ploidy level and haplotype diversity for ITS1 (Spearman's rho: 0.43, p-value: 2.96^{-5} for ITS1) and a non-significant positive correlation for ITS2 (Spearman's rho: 0.18, p-value: 0.09 for ITS2). Moreover, polyploid species showed a significant higher number of haplotypes than diploid species for ITS1 samples (Wilcoxon test = 343, p-value: 2.16^{-6}), but not significant for ITS2 samples (Wilcoxon test = 632.5, p-value = 0.059) (**Figure S2**). Haplotype absolute frequencies for all samples and at the three levels analyzed are presented in **Tables S9-S15**.

Furthermore, ITS2 presented significantly lower nucleotide diversities than ITS1 (**Figure 2**; Wilcoxon test = 5165.5, p-value: 3.33^{-7}). Moreover, we have found no polymorphism for ITS2 in 49 individuals (**Table S2-S8**), indicating the presence of only a single haplotype. This pattern

was less usual for ITS1, where only 30 individuals showed no nucleotide diversity (**Table S2-S8**). Also, ITS2 presented lower haplotype diversities when compared with ITS1 samples (Wilcoxon test = 4458, p-value 0.002; **Table 2**). Haplotype and nucleotide diversity values for ITS1 and ITS2 at the three levels analyzed for all the species studied are shown in the supplementary material (**Tables S2-S8**).

***Erysimum* species**	**ITS 1**	**ITS2**
E. baeticum	0.983 (max: 1.000; min: 0.963)	0.897 (max: 1.000; min: 0.933)
E. bastetanum	0.983 (max: 1.000; min: 0.969)	0.893 (max: 1.000; min: 0.666)
E. fitzii	0.944 (max: 1.000; min: 0)	0.872 (max: 1.000; min: 0)
E. lagascae	1.000 (max: 1.000; min: 0)	0.733 (max: 1.000; min: 0)
E. mediohispanicum	0.969 (max: 1.000; min: 0.400)	0.805 (max: 1.000; min: 0.866)
E. nevadense	0.938 (max: 1.000; min: 0.785)	0.941 (max: 1.000; min: 0.833)
E. popovii	0.984 (max: 1.000; min: 0.888)	0.943 (max: 1.000; min: 0.866)

Table 2. Average haplotype diversity (Hp) per species, estimated for ITS1 and ITS2 samples. Maximum and minimum values refer to individual samples.

For ITS1, several haplotypes were shared among different species, particularly among some populations of *E. bastetanum*, *E. fitzii*, *E. mediohispanicum*, and *E. nevadense* (see supplementary materials **Tables S10**, **S11**, **S13**, and **S14**). Specifically, we have found that the three populations of *E. bastetanum* studied here shared haplotypes with two *E. mediohispanicum* populations (Em39, Em71) and with the three populations of *E. nevadense*. In addition, *E. bastetanum* populations and one population of *E. nevadense* (En05) also shared haplotypes with the *E. fitzii* population included in the analyses.

The hierarchical AMOVAs showed that species level was a significant source of variation for both ITS sequences (**Table 3**). This level explained 52.63 % and 73.50 % of the variance for ITS1 (p-value < 0.001, Φ = 0.48) and ITS2 (p-value < 0.001, Φ = 0.70) respectively, implying ample genetic divergence between species. This analysis also showed a high value of divergence among populations (high Φ values; **Table 3**) although without reaching statistical significance.

Sequence	Source of variation	df	Variance (sigma^2)	% Variance	Φ statistics	p-value
ITS1	Species	6	1.96×10^{-5}	52.63	0.48	**< 0.01**
	Populations within species	10	2.23×10^{-6}	6.00	0.55	0.58
	Within populations	197	1.54×10^{-5}	41.36	-	-
ITS2	Species	6	1.10×10^{-4}	73.50	0.7	**< 0.01**
	Populations within species	10	1.32×10^{-5}	8.84	0.8	0.99
	Within populations	128	2.64	17.64	-	-

Table 3. Hierarchical AMOVA results for ITS1 and ITS2 regions.

When the genetic structure was separately analyzed for each species, we found a more complex results, with the bigger part of the variance (44.96 % – 100 % for ITS1; 29.12 % – 100 % for ITS2) resided at the within individuals levels (see **Table 4** for ITS1, and **Table 5** for ITS2). The among populations component (i.e., at the population level) varied from 0 % to 48.07 % for ITS and from 0 % to 70.87 % for ITS2. Then, the differentiation among populations was significant in *E. mediohispanicum*, *E. nevadense*, and *E. popovii* for ITS1 (**Table 4**) and only in *E. bastetanum* for ITS2 (**Table 5**).

Species	Source of variation	df	Variance (sigma^2)	% Variance	Φ statistics	p-value
E. baeticum	Populations	2	1.48×10^{-5}	8.93	0.11	0.09
	Individuals within populations	12	0	0	0	0.99
	Within individuals	24	1.51×10^{-4}	91.06	-	-
E. bastetanum	Populations	2	0	0	0	0.96
	Individuals within populations	12	0	0	0	0.61
	Within individuals	31	2.33×10^{-4}	100	-	-
E. fitzii	Individuals within populations	4	8.37×10^{-5}	55.03	0	0.14
	Within individuals	4	6.48×10^{-5}	44.96	-	-
E. lagascae	Individuals within populations	4	6.50×10^{-8}	0.07	0	0.55
	Within individuals	2	8.48×10^{-5}	99.92	-	-
E. mediohispanicum	Populations	2	4.58×10^{-5}	48.07	0.5	**<0.01**
	Individuals within populations	12	0	0	0.45	0.69
	Within individuals	42	4.95×10^{-5}	51.92	-	-
E. nevadense	Populations	2	1.67×10^{-5}	18.29	0.25	**<0.01**
	Individuals within populations	11	0	0	0	0.8
	Within individuals	7	7.47×10^{-5}	81.7	-	-
E. popovii	Populations	2	3.13×10^{-5}	19.01	0.2	**0.02**
	Individuals within populations	12	0	0	0.13	0.72
	Within individuals	20	1.33×10^{-4}	80.98	-	-

Table 4. ITS1 hierarchical AMOVA results for *E. baeticum*, *E. bastetanum*, *E. fitzii*, *E.lagascae*, *E. mediohispanicum*, *E. nevadense*, and *E. popovii.*

Species	Source of variation	df	Variance (sigma^2)	% Variance	Φ statistics	p-value
E. baeticum	Populations	2	1.38 x 10^{-7}	0.3	0.01	0.46
	Individuals within populations	12	0	0	0	0.99
	Within individuals	2	4.46 x 10^{-5}	99.69	-	-
E. bastetanum	Populations	2	8.05 x 10^{-5}	70.87	0.75	**<0.01**
	Individuals within populations	10	0	0	0.69	0.99
	Within individuals	29	3.30 x 10^{-5}	29.12	-	-
E. fitzii	Individuals within populations	3	0	0	0	0.9
	Within individuals	5	6.08 x 10^{-5}	100	-	-
E. lagascae	Individuals within populations	4	0	0	0	0.94
	Within individuals	4	0.06	100	-	-
E. mediohispanicum	Populations	2	0	0	0	0.97
	Individuals within populations	12	0	0	0	0.28
	Within individuals	10	5.08 x 10^{-5}	100	-	-
E. nevadense	Populations	2	0	0	0	0.99
	Individuals within populations	11	0	0	0	0.88
	Within individuals	3	5.12 x 10^{-5}	100	-	-
E. popovii	Populations	2	1.15 x 10^{-3}	7.73	0.09	0.56
	Individuals within populations	12	0	0	0	0.41
	Within individuals	12	0.01	92.26	-	-

Table 5. ITS2 hierarchical AMOVA results for *E. baeticum*, *E. bastetanum*, *E. fitzii*, *E.lagascae*, *E. mediohispanicum*, *E. nevadense*, and *E. popovii*.

Discussion

We have found a deficit of concerted evolution for the 45S rDNA regions in the *Erysimum* species studied here, manifested by a general lack of homogenization of the ITS sequences. Here, we analyzed samples taken from 85 individuals from 17 populations and 7 species, in which we have studied the degree of sequence homogenization at the individual, population, and species level. Our analyses were based on a very stringent trimming to avoid false polymorphisms due to sequencing errors. However, despite being so restrictive, we have found general high nucleotide and haplotype diversities and some genetic structure.

We have found that polyploid *Erysimum* species presented lower ITS homogenization levels than diploid species. Specifically, polyploid species presented higher genetic and haplotypic diversity and a higher number of haplotypes. A pattern that agrees with the hypothesis that polyploids harbor greater genetic diversity (Otto and Whitton, 2000). This lack of concerted evolution in polyploid species has been previously described in several study systems in which the lack of sequence homogenization is related with recent hybrid origin (Rauscher et al., 2002; Koch et al., 2003; Kovarick et al., 2004; Lunerová et al., 2017; Morales-Briones et al., 2019). The expected higher number of rDNA loci, usually located on different chromosomes, in polyploids than in diploids might have made homogenization very difficult in the polyploids. We unknown the number of rDNA loci and chromosomic locations in these *Erysimum* species. In the genome of the diploid *E. cheiranthoides* (Züst et al., 2020), the rDNA appears in eight locations in chromosomes 3, 6, 7, and 8. In polyploids, the number of rDNA loci is expected to vary with the age of polyploid formation (Weiss-Schneeweiss et al., 2013): The number of rDNA loci will coincide with the sum of those of its parents in young allopolyploids, but it will be more variable in older polyploids, where some loci are usually lost (Clarkson et al., 2005; Chester et al., 2012; Rebening et al., 2012; Weiss-Schneeweiss et al., 2013). Therefore, a higher number of rDNA loci

in polyploids than diploids probably has contributed to the lower sequence homogenization in these *Erysimum* species. Moreover, recent hybridization events may have contributed to the observed pattern of lack of ITS homogenization, since we have also found this lack of homogenization in the diploid species, that also showed high nucleotide and haplotype diversities.

Our results suggest a complex evolutionary history for these *Erysimum* species. Polyploid species show the sequence non-homogenization expected for young polyploids. Diploid species might show signatures of past hybridization events or even an ancient polyploid origin (and posterior diploidization), which might hinder concerted evolution in their genomes. However, the similar DNA content of these diploid species (our unpublished data, see **Chapter 4**) does not support an ancient polyploid origin and posterior diploidization. We think it is more plausible the hypothesis of past and recent hybridization events for these *Erysimum* species in which concerted evolution did not have time to thoroughly homogenize the ITS copies presented across chromosomes in diploid and polyploid species. The high molecular variance at the intraindividual level (**Tables 4 and 5**), which shows this level as the main reservoir of ITS variation, supports this hypothesis. The significant molecular variance among populations for some species (i.e., *E. mediohispanicum*, *E. nevadense*, *E. popovii* for ITS1) indicates some degree of genetic differentiation among populations that could be explained by the presence of different ITS copies arriving via introgression from other species.

Our results supported the lack of homogenization was more remarkable for the ITS1 than ITS2, suggesting that concerted evolution is operating more efficiently on the latter than on the former. This result was in concordance with previous studies that have shown that ITS1 is, on average more variable than ITS2, which has been described as a very conserved marker (Hershkovitz and Zimmer, 1996; Coleman, 2003; Chen et al., 2010; Buchheim et al., 2011; Wang et al., 2014; Yang et al., 2018). The AMOVA results (**Table 3**) suggest that species differences

explain a large and significant amount of genetic variance and that these markers, specially ITS2 that show a higher Φ statistics, could be useful to identify species despite the elevated intragenomic variation in the ITS sequences.

ITS sequences, specially ITS2, have for a long time been used as phylogenetic markers and barcodes in plants (Baldwin et al., 1995; Hughes et al., 2006; Wang et al., 2014; Mishra et al., 2016; Cheng et al., 2016). However, many studies have pointed out the drawbacks of using ITS for these purposes, especially in species where sequence homogenization is lacking due to hybridization or other genome rearrangement events (Alvarez and Wendel, 2003; Nieto-Feliner and Rosselló, 2007). This might be the case in our study system, where population differentiation was significant for some species (see **Table 4** and **5**). Therefore, the phylogenetic relationship of *Erysimum* cannot be resolved entirely based on ITS results alone. In this sense, the published phylogenies based on ITS sequences (Abdelaziz et al., 2014; Gómez et al., 2014; Moazzeni et al., 2014) showed a variable degree of phylogenetic incongruences, possibly because the presence of hybridization events has led to a lack of ITS sequences homogenization in the genus (Abdelaziz et al., 2014; Moazzeni et al., 2014). Therefore, our results indicate that caution is recommended when using ITS for phylogenetic studies without prior knowledge of the haplotype distribution, even for diploid species, where hybridization is generally assumed not to occur.

ITS markers have been used in several studies to identify the parental contributors of hybrid species (Koch et al., 2003; Sun et al., 2003; Nieto-Feliner et al., 2004; Grimm and Denk, 2008; Hodač et al., 2014). In our study, we have found shared haplotypes among diploid and polyploid species (specifically among *E. bastetanum* (polyploid) and the diploid species *E. fitzii*, *E. mediohispanicum*, and *E. nevadense*), which could be the results of incomplete lineage sorting or the effect of recent or ancient hybridization events. Unfortunately, we have not found clear evidence of whether these diploid species could be considered as parental species of the polyploidy taxa. Therefore, in light of our results, we think that hybridization occurring in

different populations of the same species, with multiple backcrossing events, may have resulted in a complex hybridization scenario that hides the parental species. Thus, future studies measuring the introgression among these species would bring light for these species' probable reticulated history. Moreover, studies identifying the alleles that are co-located on the same chromosome through phased haplotypes (Browning and Browning, 2011) would be ideal for identifying the parental species and trace back the hybridization events for these *Erysimum* species.

References

Abdelaziz, M. (2013). How species are evolutionarily maintained? Pollinator-mediated divergence and hybridization in *Erysimum mediohispanicum* and *Erysimum nevadense* (Doctoral dissertation, Universidad de Granada).

Abdelaziz, M., Muñoz-Pajares, A. J., Lorite, J., Herrador, M. B., Perfectti, F., & Gómez, J. M. (2014). Phylogenetic relationships of *Erysimum* (Brassicaceae) from the Baetic Mountains (se Iberian peninsula). *Anales del Jardín Botánico de Madrid* (Vol. 71, No. 1, p. 005).

Álvarez, I., & Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, 29 (3), 417-434.

Al-Shehbaz, I. A. A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon* 61: 931–954 (2012).

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Available at: www.bioinformatics.babraham.ac.uk/projects/fastqc

Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *The American Journal of Human Genetics*, 73 (4), 823-834.

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., & Donoghue, M. J. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden*, 247-277.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12 (10), 703-714.

Buchheim, M. A., Keller, A., Koetschan, C., Förster, F., Merget, B., & Wolf, M. (2011). Internal transcribed spacer 2 (nu ITS2 rRNA) sequence-structure phylogenetics: towards an automated reconstruction of the green algal tree of life. *PloS One*, 6 (2).

Buckler 4th, E. S., & Holtsford, T. P. (1996). Zea systematics: ribosomal ITS evidence. *Molecular Biology and Evolution*, 13 (4), 612-622.

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge–accurate paired shotgun read merging via overlap. *PloS One*, 12 (10).

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25 (15), 1972-1973.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., ... & Luo, K. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PloS One*, 5 (1).

Cheng, T., Xu, C., Lei, L., Li, C., Zhang, Y., & Zhou, S. (2016). Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Molecular Ecology Resources*, 16 (1), 138-149.

Chester M., Gallagher, J.P., Symonds, V.V., Cruz da Silva, A.V., Mavrodiev, E.V., et al. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc. Natl. Acad. Sci. USA* 109: 1176–1181.

Clarkson, J.J., Lim, K.Y., Kovarik, A., Chase, M.W., Knapp, S., Leitch, A.R. (2005). Long-term genome diploidization in allopolyploid *Nicotiana* section Repandae (Solanaceae). *New Phytologist* 168: 241–252.

Coleman, A. W. (2003). ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics*, 19 (7), 370-375.

Denk, T., & Grimm, G. W. (2010). The oaks of western Eurasia: traditional classifications and evidence from two nuclear markers. *Taxon*, 59 (2), 351-366.

Dover, G. (1994). Concerted evolution, molecular drive, and natural selection. *Current Biology*, 4 (12), 1165-1166.

Drábková, L. Z., Kirschner, J., Štěpánek, J., Záveský, L., & Vlček, Č. (2009). Analysis of nrDNA polymorphism in closely related diploid sexual, tetraploid sexual and polyploid species. *Plant Systematics and Evolution*, 278 (1-2), 67-85.

Elder Jr, J. F., & Turner, B. J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology*, 70 (3), 297-320.

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon- sequencing. *Nucleic Acids Research*, 43 (5), 2513-2524.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131 (2), 479-491.

Ganley, A. R., & Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research*, 17 (2), 184-191.

Gómez, J.M., Perfectti. F., Klingenberg, C.P. (2014).The role of pollinator diversity in the evolution of corolla-shape integration in a pollination-generalist plant clade. *Philosophical Transactions B* 369: 20130257 1–11.

Harpke, D., & Peterson, A. (2006). Non-concerted ITS evolution in *Mammillaria* (Cactaceae). *Molecular Phylogenetics and Evolution*, 41 (3), 579-593.

Hershkovitz, M. A., & Zimmer, E. A. (1996). Conservation patterns in angiosperm rDNA ITS2 sequences. *Nucleic Acids Research*, 24 (15), 2857-2867.

Hodač, L., Scheben, A. P., Hojsgaard, D., Paun, O., & Hörandl, E. (2014). ITS polymorphisms shed light on hybrid evolution in apomictic plants: a case study on the *Ranunculus auricomus* complex. *PLoS One*, 9 (7).

Hughes, C. E., Eastwood, R. J., & Donovan Bailey, C. (2006). From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361 (1465), 211-225.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30 (14), 3059-3066.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... & Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28 (12), 1647-1649.

Koch, M. A., Dobeš, C., & Mitchell-Olds, T. (2003). Multiple hybrid formation in natural populations: concerted evolution of the internal transcribed spacer of nuclear ribosomal DNA (ITS) in North American *Arabis divaricarpa* (Brassicaceae). *Molecular Biology and Evolution*, 20 (3), 338-350.

Kovarik, A., Matyasek, R., Lim, K. Y., Skalická, K., Koukalova, B., Knapp, S., ... & Leitch, A. R. (2004). Concerted evolution of 18–5.8–26S rDNA repeats in *Nicotiana* allotetraploids. *Biological Journal of the Linnean Society*, 82 (4), 615-625.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22 (13), 1658-1659.

Liao, D., & Jiang, C. (2004). Toward an experimental system to study the mechanism of concerted evolution. *Evolution*, 2, 3.

Long, E. O., & Dawid, I. B. (1980). Repeated genes in eukaryotes. *Annual review of biochemistry*, 49 (1), 727-764.

Lunerova, J., Renny-Byfield, S., Matyášek, R., Leitch, A., & Kovařík, A. (2017). Concerted evolution rapidly eliminates sequence variation in rDNA coding regions but not in intergenic spacers in *Nicotiana tabacum* allotetraploid. *Plant Systematics and Evolution*, 303 (8), 1043-1060.

Martin, M. Cutadapt removes adapter sequences from highthroughput sequencing reads. *EMBnet* J 17: 10–12 (2011).

Mayol, M., & Rosselló, J. A. (2001). Why nuclear ribosomal DNA spacers (ITS) tell different stories in *Quercus*. *Molecular Phylogenetics and Evolution*, 19 (2), 167-176.

Médail, F., & Diadema, K. Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, 36 (7), 1333-1345 (2009).

Mishra, P., Kumar, A., Rodrigues, V., Shukla, A. K., & Sundaresan, V. (2016). Feasibility of nuclear ribosomal region ITS1 over ITS2 in barcoding taxonomically challenging genera of subtribe Cassiinae (Fabaceae). *PeerJ*, 4, e2638.

Moazzeni, H., Zarre, S., Pfeil, B. E., Bertrand, Y. J., German, D. A., Al-Shehbaz, I. A., ... & Oxelman, B. (2014). Phylogenetic perspectives on diversification and character evolution in the species-rich genus *Erysimum* (Erysimeae; Brassicaceae) based on a densely sampled ITS approach. *Botanical Journal of the Linnean Society*, 175 (4), 497-522.

Morales-Briones, D. F., & Tank, D. C. (2019). Extensive allopolyploidy in the neotropical genus *Lachemilla* (Rosaceae) revealed by PCR-based target enrichment of the nuclear ribosomal DNA cistron and plastid phylogenomics. *American Journal of Botany*, 106 (3), 415-437.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76 (10), 5269-5273.

Nei, M., & Jin, L. (1989). Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution*, 6 (3), 290-300.

Nieto-Feliner, G. (1993) *Erysimum* L. In: Flora iberica. Vol. IV. Cruciferae-Monotropaceae. Real Jardín Botánico, CSIC, Madrid, pp. 48–76.

Nieto-Feliner, G., Gutiérrez Larena, B., & Fuertes Aguilar, J. (2004). Fine-scale geographical structure, intra-individual polymorphism and recombination in nuclear ribosomal internal transcribed spacers in *Armeria* (Plumbaginaceae). *Annals of Botany*, 93 (2), 189-200.

Nieto-Feliner, G., & Rosselló, J. A. (2007). Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*, 44 (2), 911-919.

Okuyama, Y., Fujii, N., Wakabayashi, M., Kawakita, A., Ito, M., Watanabe, M., ... & Kato, M. (2004). Nonuniform concerted evolution and chloroplast capture: heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). *Molecular Biology and Evolution*, 22 (2), 285-296.

Otto, S.P., & Whitton, J. (2000). Polyploid incidence and evolution. Annu. Rev. Genet. 34: 401– 437.

Pajares, A. J. M. (2013). *Erysimum mediohispanicum* at the evolutionary crossroad: phylogrography, phenotype, and pollinators (Doctoral dissertation, Universidad de Granada).

Paradis, E. (2010): an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26 (3), 419-420.

Popp, M., & Oxelman, B. (2004). Evolution of an RNA polymerase gene family in *Silene* (Caryophyllaceae) incomplete concerted evolution and topological congruence among paralogues. *Systematic Biology*, 53 (6), 914-932.

Rauscher, J. T., Doyle, J. J., & Brown, A. H. D. (2002). Internal transcribed spacer repeat-specific primers and the analysis of hybridization in the *Glycine* (Leguminosae) polyploid complex. *Molecular Ecology*, 11 (12), 2691-2702.

Rebernig, C.A., Weiss-Schneeweiss, H., Blöch, C., Turner, B., Stuessy, T.F., et al. (2012). The evolutionary history of the white-rayed species of *Melampodium* (Asteraceae) involved multiple cycles of hybridization and polyploidization. *American Journal of Botany*, 99: 1043–1057.

Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, 60, 561-588.

Sone, T., Fujisawa, M., Takenaka, M., Nakagawa, S., Yamaoka, S., Sakaida, M., ... & Fukuzawa, H. (1999). Bryophyte 5S rDNA was inserted into 45S rDNA repeat units after the divergence from higher land plants. *Plant Molecular Biology*, 41 (5), 679-685.

Sun, K., Ma, R., Chen, X., Li, C., & Ge, S. (2003). Hybrid origin of the diploid species *Hippophae goniocarpa* evidenced by the internal transcribed spacers (ITS) of nuclear rDNA. *Belgian Journal of Botany*, 91-96.

Team, R. C. (2013). R: A language and environment for statistical computing.

Teruel, M., Ruíz-Ruano, F. J., Marchal, J. A., Sánchez, A., Cabrero, J., Camacho, J. P., & Perfectti, F. (2014). Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper *Eyprepocnemis plorans*. *Heredity*, 112 (5), 531-542.

Wang, X. C., Liu, C., Huang, L., Bengtsson-Palme, J., Chen, H., Zhang, J. H., ... & Li, J. Q. (2015). ITS 1: a DNA barcode better than ITS 2 in eukaryotes?. *Molecular Ecology Resources*, 15 (3), 573-586.

Weiss-Schneeweiss, H., Emadzade, K., Jang, T., Schneeweiss, G. (2013). Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants. *Cytogenetic and Genome Research*, 40: 137–150.

Wendel, J. F., & Cronn, R. C. (2003). Polyploidy and the evolutionary history of cotton. *Advances in Agronomy*, 78, 139.

Won, H., & Renner, S. S. (2005). The internal transcribed spacer of nuclear ribosomal DNA in the gymnosperm *Gnetum*. *Molecular Phylogenetics and Evolution*, 36 (3), 581-597.

Xiao, L. Q., Möller, M., & Zhu, H. (2010). High nrDNA ITS polymorphism in the ancient extant seed plant *Cycas*: incomplete concerted evolution and the origin of pseudogenes. *Molecular Phylogenetics and Evolution*, 55 (1), 168-177.

Xu, B., Zeng, X. M., Gao, X. F., Jin, D. P., & Zhang, L. B. (2017). ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Scientific Reports*, 7, 40057.

Yang, R. H., Su, J. H., Shang, J. J., Wu, Y. Y., Li, Y., Bao, D. P., & Yao, Y. J. (2018). Evaluation of the ribosomal DNA internal transcribed spacer (ITS), specifically ITS1 and ITS2, for the analysis of fungal diversity by deep sequencing. *PloS One*, 13 (10).

Zheng, X., Cai, D., Yao, L., & Teng, Y. (2008). Non-concerted ITS evolution, early origin and phylogenetic utility of ITS pseudogenes in *Pyrus*. *Molecular Phylogenetics and Evolution*, 48 (3), 892-903.

Züst, T., Strickler, S.R., Powell, A.F., Mabry, M.E., An, H., Mirzaei, M., York, T., Holland, C.K., Kumar, P., Erb, M., Petschenka, G., Goméz, J.M., Perfectti, F., Müller, C., Pires, J.C., Mueller, L.A., Jander, G. (2020). Independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum*, Brassicaceae). *eLife*, 9:e51712.

Supplementary Material

Figure S1. Boxplot depicting the nucleotide diversities (π) estimated for the ITS1 and ITS2 *Erysimum mediohispanicum* samples.

Figure S2. Haplotype distribution for ITS1 (A) and ITS2 samples (B), represented as the total number of haplotypes (x-axis), and the total frequency of each haplotype (y-axis), estimated as the percentage of sequences of each individual haplotype in the total sequence pool. C) Boxplot depicting the number of ITS1 haplotypes per sample for diploid and polyploid species, D) Number of ITS2 haplotypes per sample for diploid and polyploid species.

Table S1. Number of sequences after quality trimming and after cd-hit clustering for all the samples.

Table S2. Nucleotide and haplotype diversity for *E. baeticum*, ITS1 and ITS2 samples, at the three-level analyzed.

Table S3. Nucleotide and haplotype diversity for *E. bastetanum*, ITS1 and ITS2 samples, at the three-level analyzed.

Table S4. Nucleotide and haplotype diversity for *E. fitzii*, ITS1 and ITS2 samples, at the three-level analyzed.

Table S5. Nucleotide and haplotype diversity for *E. lagascae*, ITS1 and ITS2 samples, at the three-level analyzed.

Table S6. Nucleotide and haplotype diversity for *E. mediohispanicum*, ITS1 and ITS2 samples, at the three-level analyzed.

Table S7. Nucleotide and haplotype diversity for *E. nevadense*, ITS1 and ITS2 samples, at the three-level analyzed.

Table S8. Nucleotide and haplotype diversity for *E. popovii*, ITS1 and ITS2 samples, at the three-level analyzed.

Table S9. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. baeticum* and other *Erysimum* species studied here.

Table S10. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among E. bastetanum and other Erysimum species studied here.

Table S11. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. fitzii* and other *Erysimum* species studied here.

Table S12. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. lagascae* and other *Erysimum* species studied here.

Table S13. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. mediohispanicum* and other *Erysimum* species studied here.

Table S14. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. nevadense* and other *Erysimum* species studied here.

Table S15. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. popovii* and other *Erysimum* species studied here.

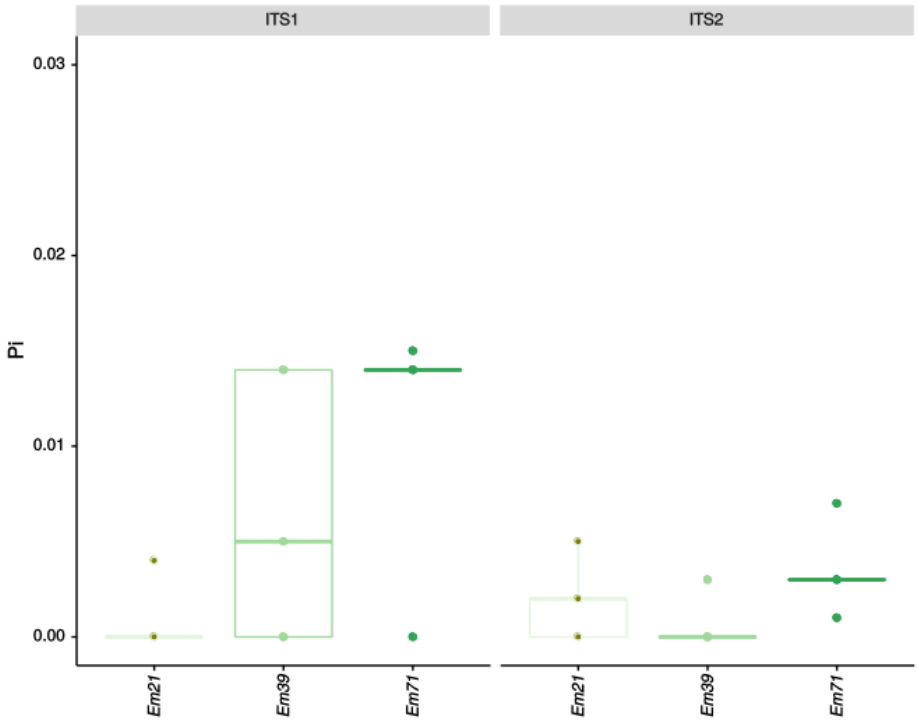

Figure S1. Boxplot depicting the nucleotide diversities (π) estimated for the ITS1 and ITS2 *Erysimum mediohispanicum* samples.

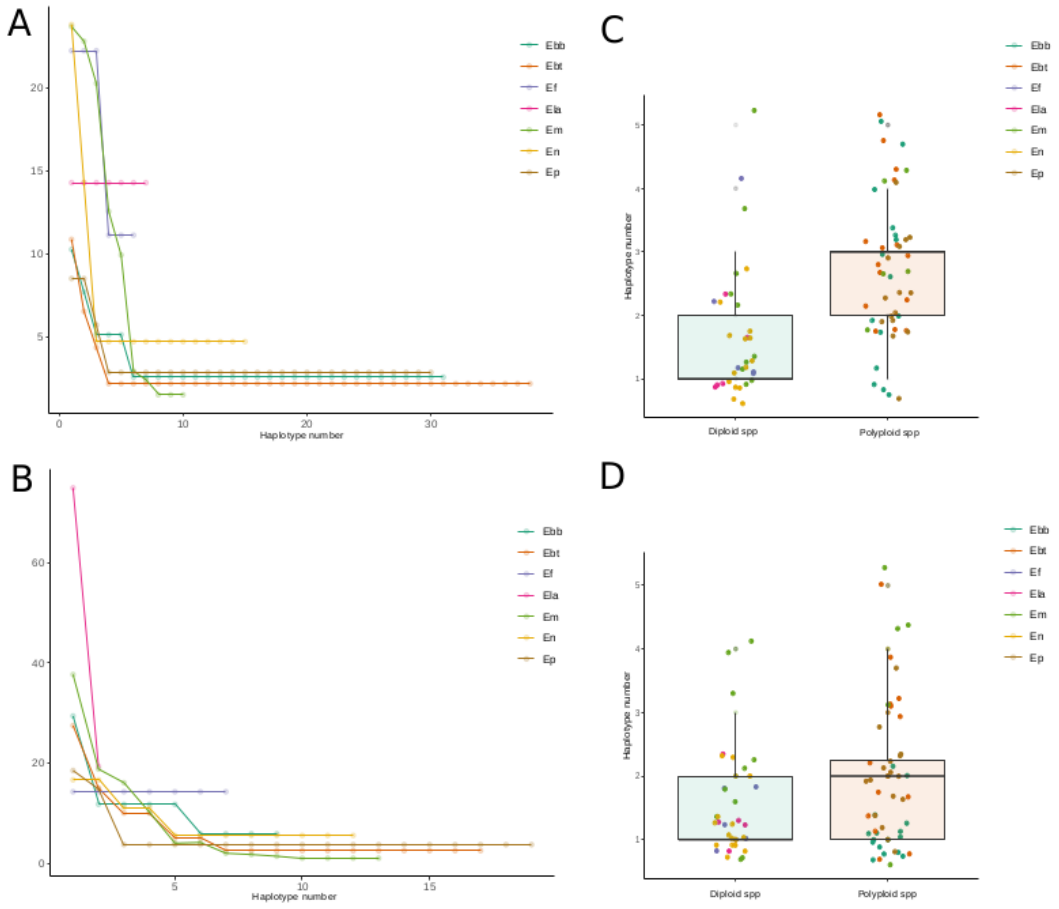

Figure S2. Haplotype distribution for ITS1 (A) and ITS2 samples (B), represented as the total number of haplotypes (x-axis), and the total frequency of each haplotype (y-axis), estimated as the percentage of sequences of each individual haplotype in the total sequence pool. C) Boxplot depicting the number of ITS1 haplotypes per sample for diploid and polyploid species, D) Number of ITS2 haplotypes per sample for diploid and polyploid species.

Taxon	Sample	Number of sequences after quality trimming		Number of sequences after clustering	
		ITS1	ITS2	ITS1	ITS2
E. baeticum	Ebb07_1	9323	194745	7122	191686
	Ebb07_2	7358	25941	6539	25311
	Ebb07_3	5852	45680	5274	18641
	Ebb07_4	5198	334388	3731	132263
	Ebb07_5	6092	403632	5063	128485
	Ebb10_1	7090	608256	5289	247990
	Ebb10_2	9843	181762	9171	65007
	Ebb10_3	6710	181726	6155	109565
	Ebb10_4	9468	352662	8618	152423
	Ebb10_5	10847	381902	9190	160503
	Ebb12_1	9820	452562	8586	172469
	Ebb12_2	6487	600370	5328	262497
	Ebb12_3	6706	441938	5727	167879
	Ebb12_4	6013	493274	5403	201674
	Ebb12_5	8379	71674	6784	29745
E. bastetanum	Ebt01_1	6673	147372	5469	140547
	Ebt01_2	4309	19851	3581	18981
	Ebt01_3	6389	136348	5869	133498
	Ebt01_4	10349	10913	9876	10627
	Ebt01_5	2407	25871	1916	25497
	Ebt12_1	9387	74828	8459	73287
	Ebt12_2	3417	77159	2575	70854
	Ebt12_3	6613	31839	4897	30580
	Ebt12_4	7133	0	5651	0
	Ebt12_5	4648	0	4306	0
	Ebt13_1	5007	37	4460	52
	Ebt13_2	6116	127	5061	95
	Ebt13_3	6075	518	5443	356
	Ebt13_4	4898	180	4279	90
	Ebt13_5	7209	412	6616	311
E. fitzii	Ef01_1	2459	1144	504	666

	Ef01_2	36284	3208	27977	2992
	Ef01_3	20575	12272	18112	11221
	Ef01_4	21007	3040	19382	2819
	Ef01_5	24480	0	21220	0
E. lagascae	Ela07_1	46652	7949	38755	7498
	Ela07_2	28675	25037	26098	23325
	Ela07_3	43375	44502	43522	40688
	Ela07_4	37367	26642	34303	23085
	Ela07_5	21787	68778	18544	49696
E. mediohispanicum	Em21_1	1057	46	1027	38
	Em21_2	1517	23	1470	17
	Em21_3	1500	37	1454	34
	Em21_4	1061	54	1013	47
	Em21_5	2460	31	1801	27
	Em39_1	740	10	580	7
	Em39_2	3860	36	355	36
	Em39_3	4184	48	297	39
	Em39_4	6317	61	401	59
	Em39_5	1542	12	156	11
	Em71_1	3372	324	310	259
	Em71_2	3152	92	281	68
	Em71_3	5526	253	418	187
	Em71_4	1817	260	215	165
	Em71_5	3824	52	387	23
E. nevadense	En05_1	19550	3660	22540	3260
	En05_2	33125	3361	29703	3222
	En05_3	20691	38519	19623	36417
	En05_4	72945	9749	65114	8588
	En05_5	101	6507	72	5839
	En10_1	17010	3031	15179	2734
	En10_2	15650	70323	13255	65241
	En10_3	21251	62475	17700	60745
	En10_4	32198	66797	30531	63176
	En10_5	19828	182495	17403	175659
	En12_1	14249	101910	13363	100098
	En12_2	15973	100528	13990	97509
	En12_3	18005	6829	16465	6298

	En12_5	29274	291336	26163	287162
E. popovii	Ep16_1	10691	32788	8965	27961
	Ep16_2	2753	26133	2590	15418
	Ep16_3	5647	26132	5300	25074
	Ep16_4	6081	18519	4967	17651
	Ep16_5	6922	14860	6381	13983
	Ep20_1	19099	55464	13841	48638
	Ep20_2	11313	15011	10034	13546
	Ep20_3	4570	7092	4108	5309
	Ep20_4	5056	6522	3753	3234
	Ep20_5	5861	25346	5119	22381
	Ep27_1	7621	58789	5962	53469
	Ep27_2	6239	44639	5227	36287
	Ep27_3	15831	12787	14370	11745
	Ep27_4	6207	38248	5147	35449
	Ep27_5	25310	24543	21224	236

Table S1. Number of sequences after quality trimming and after cd-hit clustering for all the samples.

E. baeticum	Sample code	ITS1		ITS2	
		π	Hd	π	Hd
Species level	Ebb	0.013	0.983	0.006	0.897
Population level	Ebb07	0.015	1.000	0.009	0.933
	Ebb10	0.012	0.963	0.004	1.000
	Ebb12	0.015	1.000	0.008	0.933
Individual level	Ebb07 1	0.022	1.000	0	0
	Ebb07 2	0.011	1.000	0	0
	Ebb07 3	0	0	0.010	1.00
	Ebb07 4	0.012	1.000	0	0
	Ebb07 5	0.018	1.000	0	0
	Ebb10 1	0.018	1.000	0	0
	Ebb10 2	0	0	0	0
	Ebb10 3	0.011	1.000	0	0
	Ebb10 4	0	0	0	0
	Ebb10 5	0.017	1.000	0	0
	Ebb12 1	0.009	1.000	0	0
	Ebb12 2	0.011	1.000	0.008	1.000
	Ebb12 3	0	0	0	0
	Ebb12 4	0.015	1.000	0	0
	Ebb12 5	0.011	1.000	0	0

Table S2. Nucleotide and haplotype diversity for *E. baeticum*, ITS1 and ITS2 samples, at the three-level analyzed.

E. bastetanum	Sample code	ITS1		ITS2	
		π	Hd	π	Hd
Species level	Ebt	0.013	0.983	0.006	0.893
Population level	Ebt01	0.013	0.969	0.005	0.694
	Ebt12	0.012	0.991	0.008	0.900
	Ebt13	0.013	0.975	0.005	0.916
Individual level	Ebt01 1	0.013	1.000	0.005	0.694
	Ebt01 2	0.014	1.000	0	0
	Ebt01 3	0.020	1.000	0.008	1.000
	Ebt01 4	0.019	1.000	0.021	1.000
	Ebt01 5	0.017	1.000	0.005	0.666
	Ebt12 1	0.015	1.000	0.010	1.000
	Ebt12 2	0.015	1.000	0	0
	Ebt12 3	0.015	1.000	0	0
	Ebt12 4	0.013	1.000	0	0
	Ebt12 5	0.011	1.000	0	0
	Ebt13 1	0.013	1.000	0.004	1.000
	Ebt13 2	0.018	1.000	0.004	1.000
	Ebt13 3	0.017	1.000	0.006	1.000
	Ebt13 4	0.013	1.000	0.005	1.000
	Ebt13 5	0.030	1.000	0.006	1.000

Table S3. Nucleotide and haplotype diversity for *E. bastetanum*, ITS1 and ITS2 samples, at the three-level analyzed.

E. fitzii	Sample code	ITS1		ITS2	
		π	Hd	π	Hd
Species level	Ef	0.011	0.944	0.009	0.972
Individual level	Ef01 1	0	0	0.012	1.000
	Ef01 2	0	0	0.010	1.000
	Ef01 3	0.008	1.000	0.008	1.000
	Ef01 4	0	0	0.010	1.000
	Ef01 5	0.019	1.000	0	0

Table S4. Nucleotide and haplotype diversity for *E. fitzii*, ITS1 and ITS2 samples, at the three-level analyzed.

E. lagascae	Sample code	ITS1		ITS2	
		π	Hd	π	Hd
Species level	Ela	0.011	1.00	0.005	0.733
Individual level	Ela07 1	0	0	0	0
	Ela07 2	0.008	1.00	0	0
	Ela07 3	0	0	0	0
	Ela07 4	0	0	0.007	1.000
	Ela07 5	0.019	1.00	0	0

Table S5. Nucleotide and haplotype diversity for *E. lagascae*, ITS1 and ITS2 samples, at the three-level analyzed.

E. mediohispanicum	Sample code	ITS1		ITS2	
		π	Hd	π	Hd
Species level	Em	0.300	0.969	0.003	0.805
Population level	Em21	0.001	0.400	0.002	0.750
	Em39	0.001	0.955	0.001	0.333
	Em71	0.003	0.963	0.003	0.847
Individual level	Em21 1	0.004	0.812	0.005	1.000
	Em21 2	0	0	0.002	1.000
	Em21 3	0	0	0.002	1.000
	Em21 4	0	0	0	0
	Em21 5	0	0	0	0
	Em39 1	0.014	1.000	0.003	1.000
	Em39 2	0	0	0	0
	Em39 3	0	0	0	0
	Em39 4	0.005	1.000	0	0
	Em39 5	0.014	1.000	0	0
	Em71 1	0.014	1.000	0.007	1.000
	Em71 2	0.014	1.000	0.003	1.000
	Em71 3	0	0	0.003	1.000
	Em71 4	0.014	1.000	0.001	1.000
	Em71 5	0.015	1.000	0.003	1.000

Table S6. Nucleotide and haplotype diversity for *E. mediohispanicum*, ITS1 and ITS2 samples, at the three-level analyzed.

E. nevadense	Sample code	ITS1		ITS2	
		π	Hd	π	Hd
Species level	En	0.010	0.938	0.006	0.941
Population level	En05	0.008	0.785	0.010	0.933
	En10	0.004	0.800	0.007	1.000
	En12	0.009	0.952	0.003	0.833
Individual level	En05 1	0.008	1.000	0.007	1.000
	En05 2	0	0	0	0
	En05 3	0	0	0	0
	En05 4	0.008	1.000	0	0
	En05 5	0.011	1.000	0	0
	En10 1	0	0	0	0
	En10 2	0	0	0	0
	En10 3	0	0	0	0
	En10 4	0	0	0	0
	En10 5	0.008	1.000	0	0
	En12 1	0	0	0	0
	En12 2	0.014	1.000	0	0
	En12 3	0.011	1.000	0	0
	En12 4	0	0	0	0
	En12 5	0	0	0	0

Table S7. Nucleotide and haplotype diversity for *E. nevadense*, ITS1 and ITS2 samples, at the three-level analyzed.

E. popovii	Sample code	ITS1		ITS2	
		π	Hd	π	Hd
Species level	Ep	0.015	0.984	0.007	0.943
Population level	Ep16	0.013	0.977	0.007	0.977
	Ep20	0.016	1.000	0.008	0.944
	Ep27	0.004	0.888	0.004	0.866
Individual level	Ep16 1	0	0	0.010	1.000
	Ep16 2	0.014	1.000	0.011	1.000
	Ep16 3	0.011	1.000	0	0
	Ep16 4	0.019	1.000	0.005	1.000
	Ep16 5	0.022	1.000	0.008	1.000
	Ep20 1	0.019	1.000	0.010	1.000
	Ep20 2	0.017	1.000	0.010	1.000
	Ep20 3	0.014	1.000	0.005	1.000
	Ep20 4	0.011	1.000	0.010	1.000
	Ep20 5	0.018	1.000	0	0
	Ep27 1	0	0	0	0
	Ep27 2	0	0	0	0
	Ep27 3	0.005	1.000	0	0
	Ep27 4	0	0	0.008	1.000
	Ep27 5	0.002	1.000	0	0

Table S8. Nucleotide and haplotype diversity for *E. popovii*, ITS1 and ITS2 samples, at the three-level analyzed.

E. baeticum		ITS1				ITS2			
	Sample	**Number of haplotypes**	**Relative abundance**	**Hap. shared**	**Hap. shared with other spp**	**Number of haplotypes**	**Relative abundance**	**Hap. shared**	**Hap. shared with other spp**
Species level	Ebb	31	H1: 10,25 % H2: 7,69 % H3-H5: 5,12 % H6-H31: 2,56 %	3	0	9	H1: 29,41 % H2-H5: 11,76 % H6-H9: 5,88 %	4	0
Population level	Ebb07	13	H1-H13: 7,69 %	H1: Ebb12		4	H1-H2: 33,33 % H3-H4: 16,66 %	H1: Ebb12 H2: Ebb10, Ebb12	
	Ebb10	9	H1: 33,33 % H2-H9: 8,33 %	H1: Ebb12		4	H1: 40 % H2-H4: 20 %	H2: Ebb07, Ebb12 H3: Ebb12 H6: Ebb12	
	Ebb12	11	H1-H3: 14,28 % H4-H11: 7,14 %	H2: Ebb07, H3: Ebb10		4	H1: 50 % H2-H4: 16,66 %	H1: Ebb07 H2: Ebb12, Ebb07 H3: Ebb10 H4: Ebb10	
Individual level	Ebb07_1	2	H1: 64,55 % H2: 11,84 %			1	H1: 98,42 %		
	Ebb07_2	3	H1: 51,38 % H2: 24,58 % H3: 12,89 %			2	H1: 64,67 % H2: 32,90 %		
	Ebb07_3	1	H1: 90,12 %			1	H1: 40,41 %		
	Ebb07_4	3	H1: 31,89 % H2: 27,79 % H3: 12,08 %			1	H1: 39,55 %		
	Ebb07_5	4	H1: 54,89 % H2: 16,00 % H3: 6,18 % H4: 6,02 %			1	H1: 31,83 %		
	Ebb10_1	5	H1: 31,01 % H2: 18,22 % H3: 10,23 % H4: 8,29 % H5: 6,82 %			1	H1: 40,77 %		
	Ebb10_2	1	H1: 93,17 %			1	H1: 35,76 %		
	Ebb10_3	2	H1: 80,78 % H2: 10,93 %			1	H1: 60,29 %		
	Ebb10_4	1	H1: 91,02 %			1	H1: 43,22 %		
	Ebb10_5	3	H1: 61,98 % H2: 17,56 % H3: 5,18 %			1	H1: 40,02 %		
	Ebb12_1	3	H1: 55,44 % H2: 19,68 % H3: 12,30 %			2	H1: 22,50 % H2: 15,69 %		
	Ebb12_2	3	H1: 61,30 % H2: 11,94 % H3: 8,87 %			1	H1: 43,72 %		

Ebb12_3	1	H1: 85,40 %	1	H1: 37,98 %
Ebb12_4	5	H1: 38,78 % H2: 15,61 % H3: 15,20 % H4: 12,68 % H5: 7,56 %	1	H1: 40,88 %
Ebb12_5	2	H1: 42,54 % H2: 38,41 %	1	H1: 41,50 %

Table S9. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. baeticum* and other *Erysimum* species studied here.

E. bastetanum		ITS1				ITS2			
	Sample code	**Number of haplotypes**	**Relative abundance**	**Hap. shared with Ebt populations**	**Hap. shared with other spp**	**Number of haplotypes**	**Relative abundance**	**Hap. shared with Ebt populations**	**Hap. shared with other spp**
Species level	Ebt	38	H1: 10,86 % () H2: 6,52 % H3: 4,34 % H4-H38: 2,17 %	2	5	17	H1: 27,5 % H2: 15 % H3-H4: 10 % H5-H6: 5 % H7-H17: 2,5 %	2	0
Population level	Ebt01	9	H1: 25 % H2: 16,6 % H3-H9: 8,33 %	H1:Ebt12, Ebt13 H2: Ebt13	H1: Em71 H2: Em39, En05, En12 H3: Em71 H4: Em71 H5: En05	4	H1: 55,55 % H2: 22,22 % H3-H4: 11,11 %	H1: Ebt12	
	Ebt12	15	H1: 12,5 % H2: 6,25 % H3-H15: 6,25 %	H1: Ebt01, Ebt13	H1: Em71 H2: Em39, En05, En12 H5: En05	4	H1: 40 % H2-H4: 20 %	H1: Ebt01 H2: Ebt13	
	Ebt13	13	H1: 5,55 % H2-H13: 5,55 %	H1: Ebt01,Ebt12 H2: Ebt01	H1: Em71 H2: Em39, En05, En12 H3: Em71 H4: Em71	4	H1: 60,25 % H2: 17,32 % H3: 15,94 % H4: 3,82 %	H2: Ebt12	
Individual level	Ebt01_1	3	H1: 43,75 % H2: 28,68 % H3: 9,5 %			1	H1: 95,36 %		
	Ebt01_2	2	H1: 59,43 % H2: 23,67 %			1	H1: 95,61 %		
	Ebt01_3	2	H1: 61,91 % H2: 29,94 %			2	H1: 83,54 % H2: 14,36 %		
	Ebt01_4	2	H1: 74,73 % H2: 20,69 %			2	H1: 91,87 % H2: 5,51 %		
	Ebt01_5	3	H1: 46,94 % H2: 22,18 % H3: 10,46 %			2	H1: 86,81 % H2: 11,74 %		
	Ebt12_1	3	H1: 75,14 % H2: 9,05 % H3: 5,91 %			3	H1: 59,53 % H2: 31,10 % H3: 7,31 %		
	Ebt12_2	3	H1: 41,94 % H2: 21,94 % H3: 11,91 %			1	H1: 91,82 %		
	Ebt12_3	4	H1: 33,87 % H2: 17,93 % H3: 11,81 % H4: 10,43 %			1	H1: 96,04 %		

Ebt12_4	4	H1: 33,38 % H2: 25,34 % H3: 15,49 % H4: 5 %	0	0
Ebt12_5	2	H1: 84,18 % H2: 8,45 %	0	0
Ebt13_1	3	H1: 55,78 % H2: 23,92 % H3: 9,36 %	5	H1: 30 % H2: 27,27 % H3: 21,21 % H4: 15,15 % H5: 6 %
Ebt13_2	5	H1: 38,84 % H2: 20,40 % H3: 9,66 % H4: 7,66 % H5: 6,16 %	4	H1: 46 % H2: 22,10 % H3: 18,94 % H4: 12,63 %
Ebt13_3	2	H1: 53,444 % H2: 36,14 %	3	H1: 57,02 % H2: 26,40 % H3: 16,57 %
Ebt13_4	3	H1: 34,01 % H2: 27,41 % H3: 25,92 %	3	H1: 63,33 % H2: 24,44 % H3: 12,22 %
Ebt13_5	5	H1: 37,57 % H2: 31,77 % H3: 10,75 % H4: 5.85 % H5: 5,81 %	3	H1: 72,34 % H2: 14,14 % H3: 9,6 %

Table S10. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. bastetanum* and other *Erysimum* species studied here.

E. fitzii		ITS1			ITS2		
	Sample code	**Number of haplotypes**	**Relative abundance**	**Hap. shared with other spp**	**Number of haplotypes**	**Relative abundance**	**Hap. shared with other spp**
Population level	Ef01	6	H1-H3: 22,22 % H4-H6: 11,11 %	0	7	H1-H7: 14,27 %	0
Individual level	Ef01_1	1	H1: 20 %		1	H1: 83,07 %	
	Ef01_2	4	H1: 27,42 % H2: 24,04 % H3: 20,34 % H4: 5,27 %		1	H1: 91,01 %	
	Ef01_3	1	H1: 88,02 %		2	H1: 57,63 % H2: 42,97 %	
	Ef01_4	1	H1: 92,26 %		2	H1: 81,30 % H2: 10,50 %	
	Ef01_5	2	H1: 58,84 % H2: 27,83 %		1	H1: 85,16 %	

Table S11. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. fitzii* and other *Erysimum* species studied here.

E. lagascae		ITS1			ITS2		
	Sample code	**Number of haplotypes**	**Relative abundance**	**Hap. shared with other spp**	**Number of haplotypes**	**Relative abundance**	**Hap. shared with other spp**
Population level	Ela07	7	H1-H7: 14,27 %	0	2	H1: 74,98 % H2: 19,34 %	0
Individual level	Ela07_1	1	H1: 83,07 %		1	H1: 93,16 %	
	Ela07_2	1	H1: 91,01 %		1	H1: 91,42 %	
	Ela07_3	2	H1: 57,63 % H2: 42,97 %		1	H1: 97,55 %	
	Ela07_4	2	H1: 81,30 % H2: 10,50 %		2	H1: 46,94 % H2: 25,31 %	
	Ela07_5	1	H1: 85,16 %		1	H1: 94,5 %	

Table S12. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. lagascae* and other *Erysimum* species studied here.

E. mediohispanicum		ITS1				ITS2		
	Sample code	Number of haplotypes	Relative abundance	Hap. shared with En populations	Hap. shared with other spp	Number of haplotypes	Relative abundance	Hap. shared with En populations	Hap. shared with other spp
Species level	Em	10	H1: 23,67 % H2: 22,79 % H3: 20,25 % H4: 12,54 % H5: 9,92 % H6:2,92% H7: 2,41 % H8-H10: 1,5 %	1	5	13	H1: 37,70 % H2: 18,80 % H3: 16,07 % H4: 10,19 % H5: 4 % H6: 4,14 % H7: 1,98 % H8: 1,73 % H9-H13: 1 %	0	0
Population level	Em21	2	H1: 63,79 % H2: 36,20 %			4	H1: 44,85 % H2: 36,02 % H3: 11,74 % H4: 5,14 %		
	Em39	3	H1: 19 % H2: 14 % H3: 14 %	H1	H1: Em71, Ebt12,1 Ebt13, Ebt01, En12 H5: En10		H1: 96,64 %		
	Em71	5	H1-H3: 10 % H4-H5: 7 %	H1	H1: Em39, Ebt12,5 Ebt13, Ebt01, En12 H2: Ebt01. Ebt12, Ebt13 H3: Ebt13 H4: Ebt13		H1: 25,92 % H2: 19,65 % H3: 18,66 % H4: 17,52 % H5: 7,12 %		
Individual level	Em21_1	1	H1: 97,16 %			4	H1: 39,47 % H2: 31,57 % H3: 21,05 % H4: 7,89 %		
	Em21_2	1	H1: 96,90 %			2	H1: 70,58 % H2: 29,41 %		
	Em21_3	1	H1: 96,93 %			3	H1: 58,82 % H2: 32,35 % H3: 8,82 %		
	Em21_4	2	H1: 65,97 % H2: 29,50 %			4	H1: 63,82 % H2: 19,14 % H3: 8,51 % H4: 8,51 %		
	Em21_5	1	H1: 73,21 %			1	H1: 100 %		
	Em39_1	3	H1: 51,03 % H2: 39,31 % H3: 9,65 %			2	H1: 55,55 % H2: 44,44 %		
	Em39_2	5	H1: 33,34 % H2: 28,14 % H3: 17,51 % H4: 12,45 % H5: 8,51 %			2	H1: 57,14 % H2: 42,85 %		
	Em39_3	2	H1: 85,52 % H2: 14,47 %			8	H1: 80,55 % H2-H8: 2,77 %		
	Em39_4	1	H1: 100 %			1	H1: 100 %		
	Em39_5	4	H1: 48,44 %			1	H1: 100 %		

		H2: 25,77 % H3: 16,14 % H4: 9,62 %		
Em71_1	4	H1: 52,22 % H2: 36,75 % H3: 5,55 % H4: 5,54 %	1	H1: 100 %
Em71_2	4	H1: 48,06 % H2: 24,11 % H3: 18,49 % H4: 9,32 %	4	H1: 38,61 % H2: 31,27 % H3: 18,91 % H4: 11,19 %
Em71_3	3	H1: 46,24 % H2: 39,33 % H3: 13,77 %	3	H1: 52,94 % H2: 29,41 % H3: 17,64 %
Em71_4	2	H1: 74 % H2: 26 %	5	H1: 25,66 % H2: 23,52 % H3: 20,32 % H4: 17,64 % H5: 12,83 %
Em71_5	3	H1: 44,27 % H2: 41,42 % H3: 14,30 %	4	H1: 49,69 % H2: 23,63 % H3: 13,93 % H4: 12,72 %

Table S13. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. mediohispanicum* and other *Erysimum* species studied here.

E. nevadense		ITS1				ITS2			
	Sample code	Number of haplotypes	Relative abundance	Hap. shared with En populations	Hap. shared with other spp	Number of haplotypes	Relative abundance	Hap. shared with En populations	Hap. shared with other spp
Species level	En	15	H1: 23,80 %	4	H1: Ebt, Em	12	H1-H2: 16,66 %	4	0
			H2: 14,28 %		H47: Ebt		H3-H4: 11,11 %		
			H3.H15: 4,7 %		H66: Ef		H5-H12: 5,55 %		
					H81: Em				
Population level	En05	5	H1: 50 %	H1, H2: En10	H1: Ebt12	6	H1: 28,57 %	H1: En10	
			H2: 12,5 %	H3: En12	H2: Ef		H2-H6: 14,28 %	H2: En12	
			H3: 12.5 %		H4: En12, Em71, Em39, Ebt13, Ebt12, Ebt01			H4: En12	
			H4: 12,5 %						
			H5: 12,5 %						
	En10	4	H1: 50 %	H1, H2: En05	H3: Em39	7	H1-H7: 14,28 %	H1: En05	
			H2: 16,66 %	H4: En12				H3: En12	
			H3: 16,66 %						
			H4: 16,66 %						
	En12	7	H1-H7: 14,28 %	H4: En10	H4: En05, Em71,3 Em39, Ebt13, Ebt12, Ebt01		H1: 50 %	H2: En05	
				H3: En05			H2-H3: 25 %	H4: En05	
Individual level	En05_1	1	H1: 86,73 %			2	H1: 77,20 %	H5, H10	
							H2: 11,88 %		
	En05_2	2	H1: 76,73 %			1	H1: 91,81 %		
			H2: 12,93 %						
	En05_3	1	H1: 93,09 %			1	H1: 94,54 %		
	En05_4	2	H1: 59,24 %			1	H1: 95,78 %		
			H2: 30,02 %						
	En05_5	2	H1: 58,41 %			1	H1: 89,73 %		
			H2: 12,87 %						
	En10_1	1	H1: 89,23 %			2	H1: 78,78 %		
							H2: 11,41 %		

En10_2	1	H1: 84,69 %	1	H1: 92,77 %
En10_3	1	H1: 83,29 %	1	H1: 97,23 %
En10_4	1	H1: 94,82 %	1	H1: 94,57 %
En10_5	2	H1: 53,33 %	2	H1: 61,68 %
		H2: 34,43 %		H2: 34,57 %
En12_1	1	H1: 93,78 %	1	H1: 98,22 %
En12_2	2	H1: 76,02 %	1	H1: 96,99 %
		H2: 11,55 %		
En12_3	3	H1: 39,48 %	1	H1: 92,22 %
		H2: 26,18 %		
		H3: 25,77 %		
En12_4				
En12_5	1	H1: 89,37 %	1	H1: 98,56 %

Table S14. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. nevadense* and other *Erysimum* species studied here.

E. popovii		ITS1				ITS2			
	Sample code	Number of haplotypes	Relative abundance	Hap. shared with Ep populations	Hap. shared with other spp	Number of haplotypes	Relative abundance	Hap. shared with Ep populations	Hap. shared with other spp
Species level	Ep	30	H1: 8,5 % H2: 8,5 % H3: 5,7 % H4-H30: 2,85	1	0	19	H1: 18,51 % H2: 14,81 % H3-H19: 3,70 %	1	0
Population level	Ep16	12	H1-H12: 8,33 %			9	H1: 20 % H2-H9: 10 %	H1	
	Ep20	12	H1: 15,38 % H2-H12: 7,69 %	H1: Ep27		9	H1-H2: 18,18 % H3-H9: 9 %	H1	
	Ep27	7	H1: 30 % H2: 20 % H3-H7: 10 %	H1: Ep20		4	H1-H2: 33,33 % H3-H4: 16,66 %	H1	
Individual level	Ep16_1	2	H1: 65,55 % H2: 18,30 %			4	H1: 42,33 % H2: 18,43 % H3: 16,87 % H4: 7,64 %		
	Ep16_2	2	H1: 61,09 % H2: 32,98 %			3	H1: 25,23 % H2: 21,94 % H3: 11,82 %		
	Ep16_3	2	H1: 62,93 % H2: 30,91 %			1	H1: 95,95 %		
	Ep16_4	2	H1: 75,46 % H2: 6,21 %			2	H1: 85,98 % H2: 9,33 %		
	Ep16_5	4	H1: 32,72 % H2: 27,70 % H3: 26,19 % H4: 5,56 %			2	H1: 87,94 % H2: 6,15 %		
	Ep20_1	2	H1: 44,87 % H2: 27,59 %			2	H1: 78,89 % H2: 8,79 %		
	Ep20_2	3	H1: 72,67 % H2: 8,49 % H3: 7,52 %			2	H1: 78,82 % H2: 11,41 %		
	Ep20_3	2	H1: 80,70 % H2: 9,19 %			2	H1: 67,72 % H2: 7,13 %		
	Ep20_4	3	H1: 48,23 % H2: 20,90 % H3: 5,08 %			2	H1: 61,34 % H2: 28,57 %		
	Ep20_5	3	H1: 61,28 % H2: 14,43 % H3: 11,61 %			1	H1: 88,30 %		
	Ep27_1	1	H1: 78,23 %			1	H1: 90,95 %		
	Ep27_2	2	H1: 78,36 % H2: 5,41 %			2	H1: 41,02 % H2: 40,26 %		
	Ep27_3	3	H1: 49,91 % H2: 27,51 % H3: 13,34 %			2	H1: 67,41 % H2: 24,43 %		

Ep27_4	2	H1: 75,88 % H2: 7,04 %	2	H1: 84,08 % H2: 8,59 %
Ep27_5	2	H1: 69,96 % H2: 13,89 %	2	H1: 90,20 % H2: 6,19 %

Table S15. Number of total haplotypes, frequency of each haplotype (based on the total of sequences after cd-hit analysis), number of haplotypes shared among different populations from the same species, and number of haplotypes shared among *E. popovii* and other *Erysimum* species studied here.

Chapter II

Comparative assessment shows the reliability of chloroplast genome assembly using RNA-Seq

Abstract

Chloroplast genomes (cp genomes) were widely used in comparative genomics, population genetics, and phylogenetic studies. Obtaining chloroplast genomes from RNA-Seq data seems feasible due to the almost full transcription of cpDNA. However, the reliability of chloroplast genomes assembled from RNA-Seq instead of genomic DNA libraries remains thorough. In this study, we assembled chloroplast genomes for three *Erysimum* (Brassicaceae) species from three RNA-Seq replicas and one genomic library of each species, using a streamlined bioinformatics protocol. We compared these assembled genomes, confirming that assembled cp genomes from RNA-Seq data were highly similar to each other and to those from genomic libraries in terms of the overall structure, size, and composition. Although post-transcriptional modifications, such as RNA-editing, may introduce variations in the RNA-Seq data, the assembly of cp genomes from RNA-Seq appeared to be reliable. Moreover, RNA-Seq assembly was less sensitive to sources of error, such as the recovery of nuclear plastid DNAs (NUPTs). Although some precautions should be taken when producing reference genomes in non-model plants, we conclude that assembling cp genomes from RNA-Seq data is a fast, accurate, and reliable strategy.

Published in *Scientific reports*, 8 (1), 1-12

Introduction

Chloroplast genomes are an informative and valuable resource for comparative genome evolution, population genetics, and phylogenetic studies (Huang et al., 2014; Du et al., 2017; Guo et al., 2017). Their uni-parental inheritance, low effective population size, and stable structure make them extremely useful for studying plant evolution at different taxonomic levels (Henry, 2005; Petit et al., 2005; Daniell et al., 2016; Dierckxsens et al., 2017; Asaf et al., 2017; Zhang et al., 2014). Most plant species have a stable chloroplast genome size ranging from 120 kb to 160 kb (Zhang et al., 2014) with a highly conserved structure and gene content (Jasen et al., 2012; Twyford and Ness, 2017; Jennings, 2016). The typical chloroplast genome structure is quadripartite, comprising two inverted repeats (IRs) separated by a single small copy (SSC) and a single large copy (LSC) region (Palmer, 1985; Wicke et al., 2011; Wang et al., 2015; Sablok et al., 2016; Guo et al., 2017). Most chloroplast genomes contain 110–130 genes (Du et al., 2017), most of which encode proteins involved in translation and photosynthesis (Zhang et al., 2014). Several chloroplast genes exhibit conserved flanking regions, but internal variability (e.g., matK and rbcL18) and have become basic tools in plant phylogeny and phylogeography (Clegg et al., 1994; Shaw et al., 2005; Asaf et al., 2017; Jansen et al., 2017).

The development of high-throughput sequencing technologies has led to a rapid increase in the availability of chloroplast genomes (Martin et al., 2005; Yap et al., 2015; William et al., 2015; Sablok et al., 2016) making possible the use of complete molecules in phylogenomic analyses (Zhang et al., 2017; Ruhfel et al., 2014; Ma et al., 2014; Carbonell-Caballero et al., 2015). At present, more than 2,500 complete chloroplast genomes are available (Benson et al., 2018). However, complete genome sequencing to obtain reliable chloroplast genomes also poses some caveats and remains relatively expensive. Transcriptome sequencing (RNA-Seq) is

comparatively less complex because it yields only the sections of the genome that are transcribed into RNA, providing a relatively cheap and fast method to obtain large amounts of functional genomic data (Timme et al., 2012; Wickett et al., 2014; Yang and Smith, 2013; Léveillé-Bourret et al., 2017). Accordingly, global initiatives such as the 1,000 plants (1KP) project have generated a wealth of transcriptomic data for over 1,000 plant species (Matasci et al., 2014). Since the chloroplast genome appears to be fully transcribed, RNA-Seq data could be used to obtain the complete chloroplast genome (Shi et al., 2016). However, the reliability of assembling chloroplast genomes from transcriptomic versus genomic data has not been thoroughly evaluated.

In this study, we compared the reliability of RNA-Seq to genomic DNA libraries to obtain cpDNA complete sequence. For this purpose, we assembled for the first time the complete chloroplast genome of three *Erysimum* (Brassicaceae) species: *Erysimum mediohispanicum*, *E. nevadense*, and *E. baeticum* from genomic libraries. *Erysimum* constitutes an interesting case study because it is a genus that encompasses wide diversity attained through rapid and complex evolutionary processes (Ančev, 2006; Marhold and Lihová, 2006; Moazzeni et al., 2014) while being evolutionarily close enough to *Arabidopsis thaliana* to render the use of genomic and transcriptomic references from this model species relatively easy. We assembled the chloroplast genomes of these three species using different computational approaches and compared several genetic features (gene content, presence of repeats, microsatellites –SSRs–, etc.) across genomes obtained RNA-Seq or genomic DNA. Based on these results, we assessed a) the characteristics of the chloroplast genomes of *Erysimum* spp.; b) the genomic coverage provided by RNA-Seq across species and c) a bioinformatic approach to ensure reliable chloroplast genome assembly from transcriptomic data. In the light of these results, we propose a pipeline-like methodology for processing RNA-Seq reads into high-quality cp genomes.

Material and Methods

Plant materials

Fresh leaves and flower buds of *Erysimum mediohispanicum*, *E. nevadense*, and *E. baeticum* were collected from several populations located in the Baetic Mountains, in the Southern region of Spain (**Table 1** shows the code and location of all populations). Leaves were dried and preserved in silica gel until DNA extraction. Pre-opening flower buds at the same development stage were stored in liquid nitrogen for RNA extraction.

Taxon	Population	Sample	Location	Elevation	Geographical coordinates
E. baeticum	Ebb09	Leaves	Sierra Nevada, Almería, Spain	2128	37º 05' 46" N 3º 01' 01" W
	Ebb07	Buds	Sierra Nevada, Almería, Spain	2128	37º 05' 46" N 3º 01' 01" W
	Ebb10	Buds	Sierra Nevada, Almería, Spain	2140	37º 05' 32" N 3º 00' 40" W
	Ebb12	Buds	Sierra Nevada, Almería, Spain	2264	37º 5' 51.23" N 2º 58' 5.88" W
E. mediohispanicum	Em21	Leave and buds	Sierra Nevada, Granada, Spain	1723	37º 8.07' N 3º 25.71' W
	Em71	Buds	Sierra de Huétor, Granada, Spain	1352	37º 57.164' N 2º 29.393' W
	Em39	Buds	Sierra Jureña, Granada, Spain	1272	37º19.14' N 3º 33.177' W
E. nevadense	En14	Leaves	Nigüelas, Granada, Spain	2314	37º01.451' N 3º 28.141 W
	En12	Buds	Sierra Nevada, Granada, Spain	2255	37º 05.615' N 2º 56.313' W
	En10	Buds	Sierra Nevada, Granada, Spain	2321	37º 06.615' N 3º 24.306' W
	En05	Buds	Sierra Nevada, Granada, Spain	2074	37º 06' 35' N 3º 01' 32" W

Table 1. Details of the plant populations sampled: Taxon, population code, sampled tissue, location, and geographical coordinates.

We used an individual sample for each species (**Table 1**). For each sample, at least 60 mg of leaves were disrupted using a Beadbug microtube homogenizer (Benchmark Scientific, Edison, NJ) with 2 mm steel beads. Total genomic DNA was isolated using the GenElute Plant Genomic DNA Miniprep kit (Sigma-Aldrich, St. Louis, MO) following the manufacturer's protocol. The quantity and the quality of the obtained DNA were checked using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States), and the integrity of the extracted genomic DNA was checked using agarose gel electrophoresis. Isolated DNA was sent to Macrogen (Macrogen Inc., Seoul, South Korea) to perform library preparation and sequencing. Library preparation for deep sequencing was carried out using the TruSeq Nano DNA Library Preparation Kit (350 bp insert size). The sequencing of the three cDNA libraries (*E.mediohispanicum*, *E.nevadense*, and *E.baeticum*) was carried out using the Illumina HiSeq X platform and following the paired-end 150 bp strategy. A summary of sequencing statistics is shown in **Table S1** (Supporting Information).

RNA extraction and sequencing

For each population, three replicas consisting of one pre-anthesis bud each were used. They were snap-frozen in liquid nitrogen and disrupted with mortar and pestle. Total RNA was isolated using the Qiagen RNeasy Plant Mini Kit following the manufacturer's protocol. The quality and quantity of the RNA obtained was checked using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States), and analyzed with the Agilent 2100 Bioanalyzer system (Agilent Technologies Inc). The RNA was sent to Macrogen (Macrogen Inc., Seoul, South Korea) for library preparation and sequencing. We used a rRNA-depletion protocol (Ribo-Zero; Sooknanan et al., 2010) to perform a mRNA enrichment and to avoid sequencing rRNAs. Library preparation was performed using the TruSeq Stranded Total RNA LT Sample Preparation Kit (Plant). The nine libraries' sequencing was carried out using the Hiseq 3000–4000 sequencing protocol and TruSeq 3000–4000 SBS Kit v3 reagent, following a

paired-end 150 bp strategy on the Illumina HiSeq 4000 platform. A summary of sequencing statistics is shown in **Table S1** (Supporting Information).

Chloroplast genome assembly and annotation

We assembled *de novo* the chloroplast genomes of *E. mediohispanicum*, *E. nevadense*, and *E. baeticum* using the NOVOPlasty pipeline v6.2.3 (Dierckxsens et al., 2017) (**Figure 1**). Basically, through this pipeline, a cp genome is assembled from whole-genome sequencing (WGS) data, starting from a related single seed sequence iteratively extended bidirectionally until the circular genome is obtained. We used untrimmed reads as recommended by Dierckxsens et al., 2017 and *Arabidopsis thaliana* cpDNA sequence (NC_000932.1) as the seed since *Erysimum* is a close relative of *Arabidopsis* (Price et al., 2008). We specified the following parameters: automatic insert size detection, a genome range from 120000 to 200000, a K-mer value of 39, an insert range of 1.6, a strict insert range of 1.2, and the paired-end reads option.

After assembling the full chloroplast genome of *E. medihospanicum*, we proceeded to assemble the cp genomes from the RNA-Seq data by using this chloroplast genome as a reference. From the RNA-Seq libraries, we first trimmed the adapters in the raw reads using cutadapt v1.15 (Martin., 2011). For trimming adapters in 5′ and 3′ direction, we used the "-b" option, and only used the prefix of the adapter sequence that is common to all "TruSeq Indexed Adapter" sequences (AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC). In addition, we used the "-n" option to search repeatedly for the adapter sequences (28 iterations). This option ensures that the correct adapters are detected by searching in loops until any adapter match is found or until the specified number of rounds is reached. Then, we quality-filtered the reads using Sickle v1.33 (Joshi and Fass, 2011), a trimming software that uses sliding-window analyses along with quality and length thresholds to cut and discard the reads which do not fit the selected threshold values. We specified the "pe" option for paired-end reads and the "-t" to use Illumina quality values (see https://github.com/najoshi/sickle). After filtering, we used the read mapper of

Geneious R.11 (Kearse et al., 2012) with the highest sensitivity and default parameters (http://www.geneious.com) (Kearse et al., 2012) for a reference-guided assembly of the trimmed reads using the *E. mediohispanicum* reference assembly (see above). We validated the results obtained with Geneious R.11 by comparing the percentage of reads mapped with the BWA v0.7.17 read mapper (Li and Durbin, 2009).

The program cpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis, and GenBank Submission Tool; Liu et al., 2012) was used to annotate and visualize the cp genomes. This program takes as input a FASTA file containing the genome information and performs bioinformatic analyses to annotate the genome. We used the annotated *Arabidopsis thaliana* cp genome (NC_000932.1) (Sato et al., 1999). Protein coding genes were manually curated. Lastly, cpGAVAS gives as output the statistics of the annotation process, the annotated genome, and a visualization of the annotated genome. The annotations were then manually curated using Geneious R.11 (Kearse et al., 2012). All transfer RNA sequences (tRNA) encoded in the cp genomes were verified using tRNAscan-SE v2.0 (Schattner et al., 2005) with the default search settings. The step-by-step process is presented in **Figure 1.**

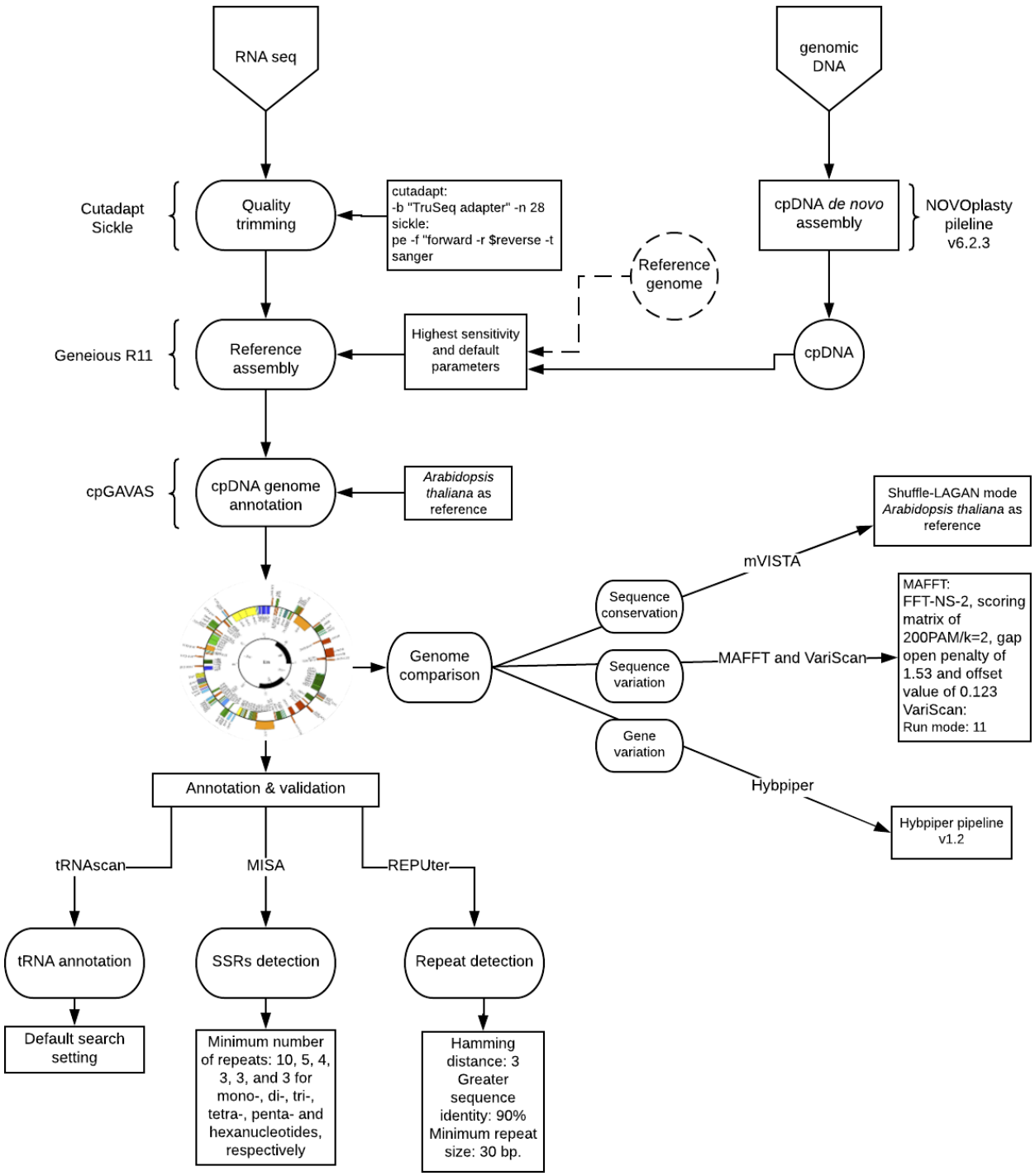

Figure 1. A flow chart depicting the bioinformatic analyses to assembly cp genomes.

Comparative analysis among cp genomes assemblies

To compare the cp genomes assembled from DNA or RNA libraries, we used the mVISTA software, part of the VISTA suite of tools for comparative genomics (http://genome.lbl.gov/vista/mvista/submit.html; Frazer et al., 2004). This software compares DNA sequences from different species by pairwise alignment and allows the visualization of these alignments with annotation information. The output allows the identification of homologies between sequences, determining the percentage of identity between them using a sliding window of predefined length. We selected default parameters, a RankVISTA probability threshold of 0.5, and the Shuffle-LAGAN mode, a global alignment algorithm for finding rearrangements (inversions, transpositions, and some duplications). We used the *A. thaliana* cpDNA as a reference (NC_000932.1) (Sato et al., 1999). The sequence conservation profiles were visualized in mVISTA plots (Frazer et al., 2004).

We investigated the degree of within-genome variation of the assembled cp genomes. In particular, we performed a reference-guided assembly in which we remapped the quality-trimmed reads (for the RNA-Seq assemblies, see above) to each assembled genome using the Geneious R.11 (Kearse et al., 2012) mapper with medium-low sensitivity and default parameters (http://www.geneious.com). Later, we estimated the percentage of pairwise identity of each assembly. This statistic gives the average identity (as %), computed by scoring a hit when all pairs of bases are identical and dividing it by the total numbers of pairs.

For each species, we explored the degree of overall sequence variation found within the three replicas of RNA-Seq assembled genomes and then, we compared the results to those of similar analyses that included the genome assembled from genomic libraries. For this purpose, we estimated the nucleotide diversity (π) among the three replicas of cp genomes assembled from RNA-Seq and then computed it again, including the corresponding genomic library. Genomes were first aligned using MAFFT v7.450 (Katoh et al., 2009) with the following parameters: FFT-

NS-2 fast progressive method algorithm, a scoring matrix of 200PAM/k = 2, gap open penalty of 1.53, and offset value of 0.123. We then estimated the cpDNA nucleotide diversity using VariScan v2.0.3 (Hutter et al., 2006).

We studied the degree of sequence variation of some relevant chloroplast genes within the three replicas of RNA-Seq, we explored it but included the genes assembled from genomic libraries. We first extracted and assembled all the chloroplast genes using the HybPiper pipeline v1.2 (Johnson et al., 2016). This pipeline uses BWA v0.7.17 (Li and Durbin, 2009) to align reads to target sequences, and SPAdes (Bankevich et al., 2012) to assemble these reads into contigs. Once cpDNA genes were obtained, we selected 12 genes out of the total: rbcl, psaA, psbA, ndhK, atpA, atpH (with an important function in the photosynthesis process; Pfannschmidt, 2003), rpoA, rps3, rrn16S, trnH (as self-replication genes; Sakaguchi, 2017), yfc2 (the largest plastid gene in angiosperms; Huang et al., 2010), and matK (the only maturase of higher plants and widely used in angiosperm systematic; Hilu et al., 2003). Then, we aligned these genes using MAFFT v7.450, as explained above. Lastly, we calculated the percentage of pairwise identity between the genes obtained from the three RNA-Seq replicas, and those from genomic libraries.

The size and location of repeat sequences, including palindromic, reverse, and direct repeats, within these cp genomes, were identified using REPuter software (Kurtz et al., 2001). Following Asaf et al., 2017 and Ni et al., 2017, REPuter was parametrized with the following settings: Hamming distance of 3; 90% or greater sequence identity; and minimum repeat size of 30 bp.

Simple sequence repeat (SSR) elements were detected using the Perl script MISA (Thiel, 2003) by setting the minimum number of repeats to 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta- and hexanucleotides, respectively.

Analysis of minimum transcriptome depth to produce quality cp genomes assemblies

To analyze the impact that sequencing depth has in the assemblage of transcriptome data into a complete cp genome, we subsampled the transcriptome reads of *E. nevadense* four times at 1 M, 5 M, 10 M, 20 M, and 30 M paired reads. These reads were processed and mapped to the cpDNA

of *E. mediohispanicum* with Geneious R.11 (Kearse et al., 2012) with medium-low sensitivity and default parameters as previously done. We calculated several mapping quality indexes (coverage of bases, expected errors, mean confidence, and % of Q40 positions) with Geneious R.11 (Kearse et al., 2012) and plotted them against the sub-sampling depth.

Cross-validation of the methodology

To estimate the recovery of complete cpDNA chromosomes from RNA-Seq libraries in other plant species, we downloaded five transcriptomes from the Sequence Read Archive website and processed them with our workflow. We downloaded two *A. thaliana* (SRR6757372; SRR6676021), one *E. cheiri* (SRR5195368), one *Moricandia suffruticosa* (SRR4296233), one *M. arvensis* (SRR4296231), one *Oriza sativa* (SRR7079258), and one *Zea mays* (ERR1407273) transcriptome. These libraries were trimmed, and quality filtered using cutadapt v1.15 (Martin, 2011) and Sickle v1.33 (Joshi and Fass, 2011) with the same parameters described above and mapped using Geneious R.11 (Kerase et al., 2012) to cp genomes of the same species (or the closest relative available): Genbank accession NC_000932 for *A. thaliana*, our *E. mediohispanicum* cp genome for the *E. cheiri* sample, *Brassica napus* GQ861354 for the *Moricandia* samples, *Oriza sativa* NC_001320 for *O. Sativa*, and *Z. mays* NC_001666 for *Z. mays*.

Results

Chloroplast genome assembly and annotation.

Genomic DNA libraries. We assembled *de novo* the whole chloroplast genomes of three *Erysimum* species. The assembled genomes were circular and had a total length of 154,599 bp, 154,660 bp, and 154,581 bp in *E. mediohispanicum*, *E. nevadense*, and *E. baeticum*, respectively (**Figures 2, S1 and S2**). These chloroplast genomes displayed the typical quadripartite structure of most angiosperms (See **Table 2**), comprising a pair of inverted repeats (IRs; 26,429 bp, 26,442

bp, and 26,429 bp respectively), the large single-copy region (LSC; 136,628 bp, 136,724 bp, and 136,625 bp respectively), and the small single-copy region (SSC; 83,853 bp, 83,804 bp, and 83, 767 respectively).

The gene content of the three chloroplast genomes was highly conserved (**Table S2**). Thus, the number of unique protein-coding genes was 124 for the three species. These chloroplast genomes contained 29 unique transfer RNA genes and eight unique ribosomal RNA genes (**Figure 3**). The number of intra-gene regions was 150 in each cp genome. We found eight split genes (rpl2, atpF, rpoC1, psaA, ycf3, clpP, ndhB, ndhA; see **Table 3**) with intronic regions for each cp genome.

The lengths of the intronic regions are shown in **Table S3**. The overall GC content was 36.6%, indicating similar conserved GC levels among the *Erysimum* chloroplast genomes. A summary of the number of sequences assembled, mean assembly coverage, and percentages of pairwise identity are shown in **Table S4** (Supporting Information).

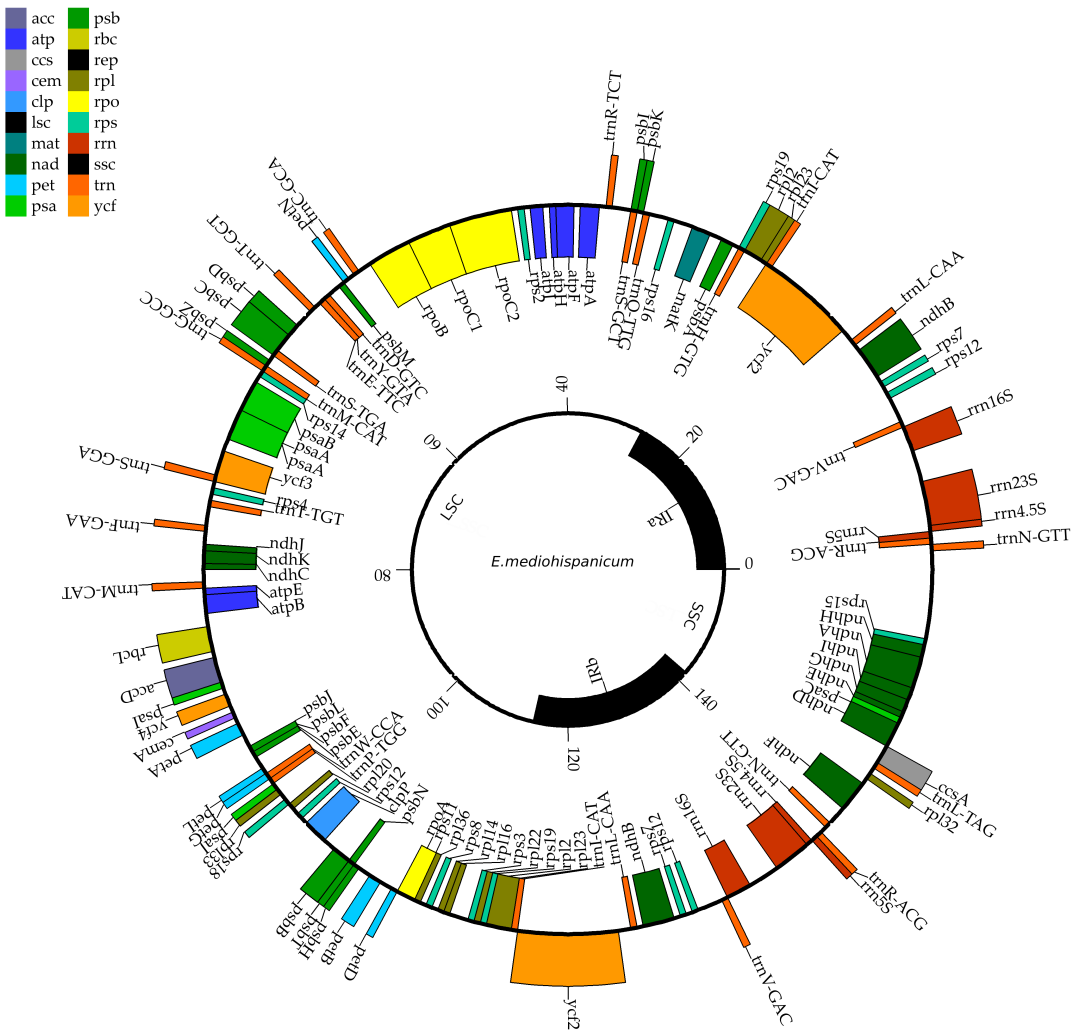

Figure 2. Chloroplast genome map of *Erysimum mediohispanicum*. Genes drawn inside the circle are transcribed clockwise, and those outside are counter-clockwise. Genes belonging to a different functional group are shown in different colors. See suplementary material for functional category of these genes.

Taxon	Population code	Type of library	Length (bp)	Assembled reads	IRa (bp)	SSC (pb)	IRb (bp)	LSC (bp)	GC %
E. baeticum	Ebb09	Genomic DNA	154,581	983,811	26,429	83,767	26,426	136,625	36.6
	Ebb07	RNA-Seq libraries	154,791	3,727,511	25,783	95,135	13,797	134,715	37.5
	Ebb10	RNA-Seq libraries	154,768	9,963,413	25,847	95,396	14,419	135,662	36.5
	Ebb12	RNA-Seq libraries	154,761	10,356,264	24,617	95,167	13,305	133,089	36.5
E.mediohispanicum	Em21	Genomic DNA	154,599	1,414,714	26,429	83,853	26,429	136,628	36.6
	Em71	RNA-Seq libraries	154,788	1,314,441	24,671	95,187	13,303	133,161	36.5
	Em39	RNA-Seq libraries	154,827	13,595,017	26,472	83,764	24,099	134,335	36.5
	Em21	RNA-Seq libraries	154,251	19,075,780	25,280	89,248	18,133	132,661	36.6
E. nevadense	En14	Genomic DNA	154,660	1,554,542	26,442	83,840	26,442	136,724	36.6
	En05	RNA-Seq libraries	153,467	12,482,406	25,863	85,139	24,831	135,833	36.7
	En10	RNA-Seq libraries	154,834	9,515,436	25,902	85,182	23,492	134,576	36.7
	En12	RNA-Seq libraries	154,747	5,338,711	25,764	84,289	24,027	134,080	36.7

Table 2. Characteristics of the chloroplast genomes of *Erysimum*: type of library (genomic DNA or RNA-Seq library), length of the cp genome (bp), number of assembled reads, length of the two inverted repeats (IR a and the IR b), length of the small single copy (SSC), and of the large single copy (LSC) region, and GC% content.

Taxon	Population	Library	PCG	tRNA	mRNA	rRNA	Exons	CDS	Gene introns	Repeats	Forward repeats	Reverse repeats	Palindrome repeats	Complemented repeats	Repeats in IRa	Repeats in SSC	Repeats in IRb	Repeats in LSC
E. baeticum	Ebb09	Genomic DNA	124	29	87	8	136	99	8	65	38	0	25	0	16 (24.61%)	42 (64.61%)	7 (10.79%)	0
	Ebb07	RNA-Seq	124	29	87	8	136	99	8	90	37	21	32	2	22 (24.44%)	66 (73.3%)	0	0
	Ebb10	RNA-Seq	124	29	87	8	136	99	8	76	38	0	36	2	12 (15.78%)	64 (84.2%)	0	0
	Ebb12	RNA-Seq	124	29	87	8	136	99	8	86	38	0	46	2	15 (17.44%)	71 (82.55%)	0	0
E. mediohispanicum	Em21	Genomic DNA	124	29	86	8	135	98	8	64	38	0	24	2	12 (18.75%)	48 (75%)	4 (6.25%)	0
	Em71	RNA-Seq	124	29	87	8	135	98	8	85	40	0	43	2	14 (16.4%)	70 (83.35%)	1 (1.17%)	0
	Em39	RNA-Seq	124	29	87	8	135	98	8	70	40	0	28	2	30 (42.85%)	41 (58%)	7 (10%)	0
	Em21	RNA-Seq	124	29	87	8	135	98	8	80	44	1	33	2	16 (20%)	63 (78.75%)	1 (1.25%)	0
E. nevadense	En14	Genomic DNA	124	29	87	8	136	99	8	74	51	1	25	2	14 (18.89%)	55 (74.32%)	5 (6.75%)	3 (4.40%)
	En05	RNA-Seq	124	29	86	8	136	99	8	90	51	0	37	2	28 (31%)	57 (63%)	5 (5.55%)	0
	En10	RNA-Seq	124	29	87	8	136	99	8	90	43	18	29	0	19 (21%)	66 (73.3%)	5 (5.55%)	0
	En12	RNA-Seq	124	29	87	8	136	99	8	90	53	0	36	1	15 (16.6%)	71 (78.8%)	4 (4.44%)	0

Table 3. Comparison of RNA-Seq vs. genomic assembly of chloroplast genomes. Number of protein-coding genes (PCG), tRNA, mRNA, rRNA, exons, coding sequences (CDS), genes with introns, repeat sequences, and number of repeats in different chloroplast regions (IRa, SSC, IRb, and LSC) for chloroplast genomes obtained from genomic DNA and chloroplast genomes obtained from RNA-Seq libraries are presented. The eight genes showing introns were rpl2, atpF, rpoC1, psaA, ycf3, clpP, ndhB, and ndhA.

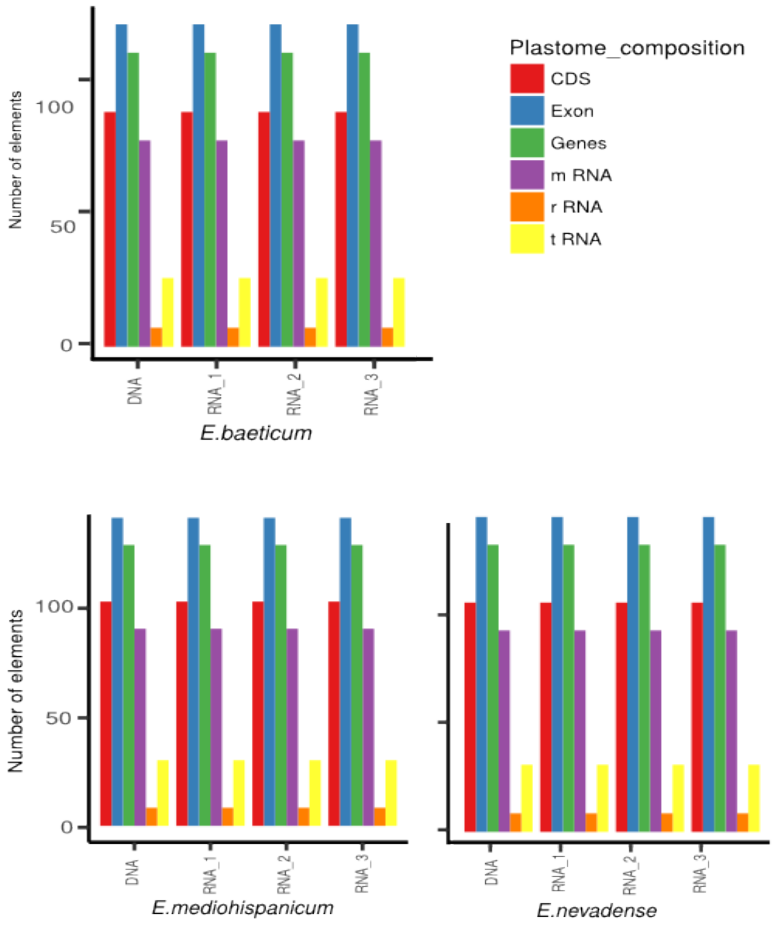

Figure 3. Composition of *Erysimum baeticum*, *E. mediohispanicum, and E. nevadense* cp genomes, obtained from genomic data and for the three RNA-Seq replicas.

RNA-Seq libraries. We assembled the chloroplast genomes from three different replicas for each of the three species. We recovered high-quality complete chloroplast genomes that were very similar to those obtained from genomic DNA. In particular, for *E. mediohispanicum* the retrieved chloroplast genome sizes were 154,788 bp, 154,827 bp, and 154, 251 bp; for *E. nevadense* genome sizes were 153,467 bp, 154,834 bp, and 154,747 bp; and for *E. baeticum* were 154,791 bp, 154,768 bp, and 154,761 bp. The IR, LSC, and SSC contents (See **Table 2**), as well as the protein-coding gene contents, tRNAs, and rRNAs, were very similar between species replicates but when comparing with the chloroplast genomes obtained from genomic DNA, we found that the IRb regions were shorter and SSC regions were slightly larger (see **Figure S3** for chloroplast borders comparison). We found the same eight split genes with introns regions that were found in cp genomes obtained from genomic DNA (rpl2, atpF, rpoC1, psaA, ycf3, clpP, ndhB, ndhA; see **Table 3**). The lengths of all intronic regions are shown in **Table S3**. We found the same number of intra-gene regions using RNA-Seq and genomic libraries (150 in each cp genome). The overall GC was 36%. A summary of assembly statistics, including the bases assembled, mean assembly coverage, and percentages of pairwise identity, are shown in **Table S4**. The results of the mapping assembly using BWA v0.7.17 were highly similar (**Table S5**).

Repeat and SSRs analyses.

The total number of repeats was 64, 78, and 65 in *E. mediohispanicum*, *E. nevadense*, and *E. baeticum*, respectively. Forward repeats were the most common across the three species, followed by palindromic repeats. Reverse and complement repeats were found in low abundance (**Table 3**). In particular, *E. mediohispanicum* contained 38 forward, 24 palindromic, and two complement repeats; *E. nevadense* contained 51 forward, 25 palindromic, and one complement repeats; and *E. baeticum* contained 38 forward, 25 palindromic, and two complement repeats, respectively. In addition, the repeats from the three species had a sequence identity higher than 90%. The length of these repeats ranged for all the species from 30 to 26,429 bp, and the most common copy

length had 30 bp. The number of repeats in the chloroplast genome assembled from RNA-Seq data was similar to that obtained from genomic DNA (**Table 3**). The average number of repeats was 78.3, 90, and 84, for the three replicas of *E. mediohispanicum*, *E. nevadense*, and *E. baeticum*, respectively. Forward repeats were the most common, followed by palindrome repeats, with lower levels of reverse and complemented repeats. The repeats of these population samples had a sequence identity greater than 90% for each species. The length of these repeats reached from 30 to 14, 353 bp, with the units with 30 bp being also the most common. The SSRs contained in the three chloroplast genomes were analyzed using the MISA Perl script (**Figure 4**).

The number of detected SSRs were 78, 83, and 81, for *E. mediohispanicum*, *E. nevadense*, and *E. baeticum*, respectively. Among them, most of the SSRs were mononucleotide repeats, followed by dinucleotide and tetranucleotide repeats. The hexanucleotides were the less frequent type. Among these SSRs, mononucleotide A/T repeat units were the most represented, with a proportion of 58% in *E. mediohispanicum*, 59% in *E. nevadense*, and 59% in *E. baeticum*. The number of SSRs identified in cp genomes assembled from RNA-Seq was lower than the number identified in cp genomes obtained from genomic libraries. We found a total of 61, 66, and 68 SSRs in the each of the three *E. mediohispanicum* population samples; 68, 67, and 68 in the three *E. baeticum* samples, and 69, 68, and 60, in the three *E. nevandense* samples. **Table S6** shows the numbers of SSR's that were quantitatively different between cp genomes assembled from genomic and RNA-Seq libraries. Among them, most of the SSRs were also mononucleotide repeats, with A/T repeats showing the highest proportion in the three replicas per species.

Genomic comparison.

Results from mVISTA plots revealed a high similarity, with 99% of shared sequence identity in pairwise comparisons, between chloroplast genomes from genomic libraries and those from RNA-Seq libraries (See **Figure 5**; the top and bottom percentage bounds are shown to the right of every row). These plots also showed a high degree of synteny between the three *Erysimum* species. In addition, the two IR regions were similar to the LSC and SSC regions in all these

species. Lastly, non-coding regions reveal a higher divergence than coding regions. Nucleotide diversity (π) was lower among the three replicas assembled from RNA-Seq. In contrast, nucleotide diversity increased dramatically (~ three orders of magnitude) when including the cp genome from genomic libraries in the alignments (0.35988 vs. 0.00008 for *E. mediohispanicum*; 0.36617 vs. 0.00037 for *E. nevadense*; and 0.36068 vs. 0.00123 for *E. baeticum*). Percentages of pairwise identity were always higher than 99% when comparing genes assembled from the different RNA-Seq replicas, and this similarity did not decrease when including genes assembled from genomic libraries (see **Table S7**).

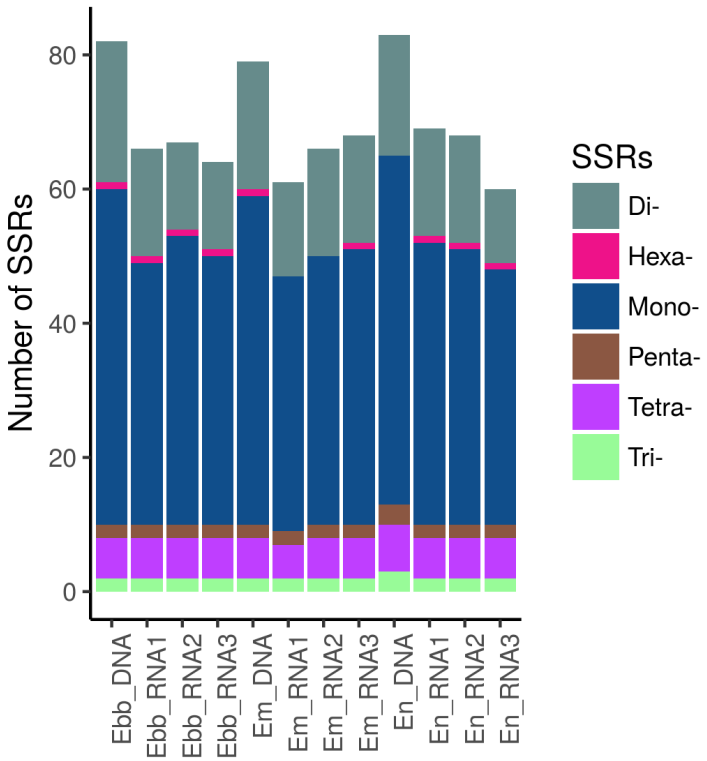

Figure 4. The number of single small repeats (SSRs) sequences in the chloroplast genomes of *Erysimum* species, obtained from genomic data and for the three RNA-Seq replicas.

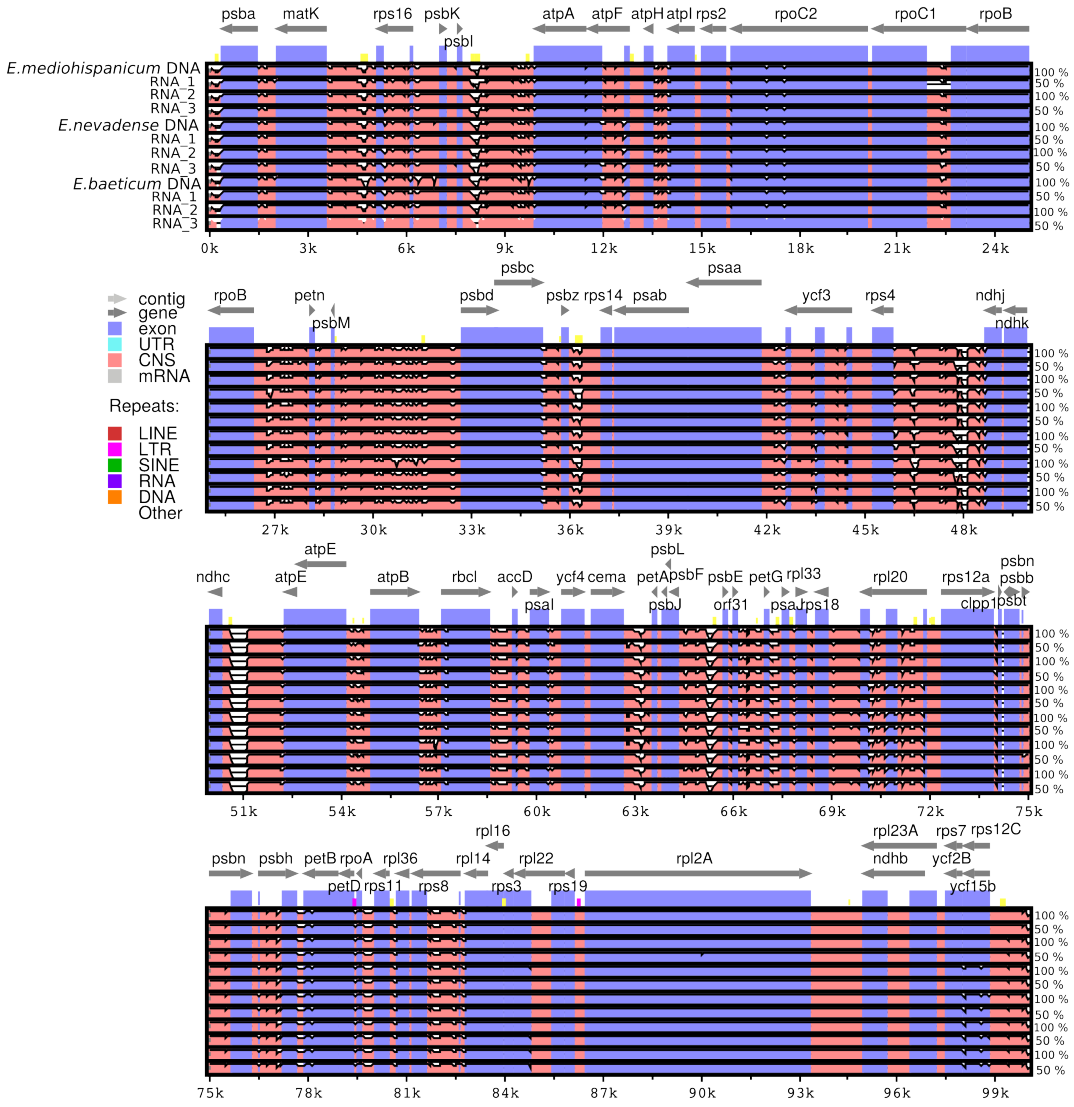

Figure 5. Sequence identity plots among the *Erysimum* chloroplast genomes, with *Arabidopsis thaliana* as a reference. Annotated genes are displayed on the top. A cut-off of 50% identity was used for the plot. The vertical scale represents the percent identity between 50 and 100%. Genome regions are color-coded as CNS (conserved non-coding sequences), exons, and introns. The color legend is summarized in the upper left-hand corner.

Effect of sequencing depth. We assembled the chloroplast genomes from four different resamplings of an *E. nevadense* transcriptome at 1 M, 5 M, 10 M, 20 M, and 30 M paired reads. With this particular transcriptome, chloroplast genomes were obtained with coverage >95% from libraries of only 1 M reads, and with coverage >99% from sequencing depths >5 M reads (see supplementary information **Figure S4**). As expected, all metrics of mapping quality as well as the mean coverage at each position ($p < 0.0001$; $R2 = 0.921$; from ~170 K to close to 200 K) increased significantly with sequencing depth (see supplementary information **Figure S4**).

Cross-validation results. We assembled the cp genomes from RNA-Seq data of five species: *A. thaliana*, *E. cheiri*, *M. suffruticosa*, *M. arvensis*, *O. sativa*, and *Z. mays*. All estimated parameters (assembly consensus length, confidence mean, Q20, Q30 and Q40 sequence quality scores, the total number of assembled reads, percentage of pairwise identity, mean coverage, and coverage concerning the reference sequence) showed that assembling cpDNA from RNA-Seq data was feasible, albeit the reliability of the assembly was dependent on the RNA-Seq reads used (see **Table S8**).

Discussion

Our results showed that complete chloroplast genomes could be reliably assembled from transcriptomic data. We studied some *Erysimum* species as a proof-of-concept and obtained genomes congruent in structure and sequence with previously published chloroplast genomes (Do et al., 2013; Du et al., 2017; Guo et al., 2017). Both the chloroplast genomes assembled from transcriptomic and genomic libraries exhibit the typical quadripartite structure, low GC content, and are mainly composed of polythymine (polyT), and polyadenine (polyA) repeats (Kuang et al., 2011). Chloroplast genomes assembled from RNA-Seq data are highly similar in terms of SSRs, the number of repeats, and plastome composition (CDS, exons, genes, rRNA, and tRNA) to those assembled from genomic libraries. Moreover, the similarity of the genomic and RNA-Seq assemblages validates that chloroplast genomes were fully transcribed. This is in line with

findings from Shi et al., (2016) who showed full transcription of the chloroplast genome in photosynthetic eukaryotes using several tissues (flowers, complete seedlings, and seedlings shoots). Here, we show that chloroplast genomes of flower buds, the tissues we have used to obtain the RNA-Seq libraries, are also fully transcribed. Therefore, chloroplasts appear to be fully transcribed across organs and development stages in angiosperms, at least in samples containing functional plastids.

We have found significant differences in nucleotide diversity when comparing both kinds of assemblages (RNA-Seq vs. genomic libraries). This may be explained by post-transcriptional modifications, i.e., by RNA-editing (Gutman et al., 2017). However, we found that nucleotide diversity significantly increased when including the assemblies from genomic libraries into the alignments. Accordingly, nucleotide diversity was lower when only comparing the three replicates of the RNA-Seq data. This implies that genomic assemblies were more heterogeneous or noisier than transcriptomic ones. Since both libraries were obtained using similar Illumina platforms, it appears that the genomic libraries were intrinsically more heterogeneous. This heterogeneity is likely caused by segments of chloroplast DNA transferred to the nuclear genome (i.e., nuclear plastic DNA or NUPT) that may potentially be incorporated during the mapping procedure introducing heterogeneity (i.e., within-genome polymorphism) into the cpDNA genomic assemblies (Sato et al., 1999; Kim et al., 2015). However, NUPTs are generally fragmented and eliminated from the nuclear genome and, therefore, not transcribed at low level (Matsuo et al, 2005; Noutsos et al., 2005; Scarcelli et al., 2016), they should not be recovered in the RNA-Seq libraries. Moreover, the lack of differences in pairwise identity when comparing genes from RNA-Seq to those from genomic libraries may be a consequence of NUPTs located in the intergenic regions, as found in previous studies (e.g., only 25% of NUPTs in Arabidopsis thaliana are located in genes; Richly et al., 2004). NUPTs are well documented in plants (Arthofer et al., 2010), and they often represent a significant part of the nuclear genome (Michalovova et al., 2013; Yoshida et al., 2013). Because of the maternal inheritance in most plant genera (Connett

et al., 1986), cpDNA is widely used for the inference of relationships among plants. Therefore, the presence of NUPT into cp genomes may lead to erroneous phylogenetic inferences (Arthofer et al., 2010). According to our results, using cpDNA assembled from transcriptomes might reduce the problems due to NUPT inclusion when using cpDNA in phylogenomics. Alternatively, methods specifically designed to correct these assembly errors have been developed for genomic data, such as the dnaLCW method (Kim et al., 2015), and should be considered whenever possible. However, validating that NUPTs is a source of error in cp genome assembly requires comparison with a reference genome, which is currently not available for *Erysimum*. Therefore, the potential misleading mapping caused by this type of genetic element will require further studies.

We found that assembling chloroplast genomes from RNA-Seq data is a relatively fast and flexible approach when we tested our methodology across several plant species. In light of these results, we put forward a pipeline-like procedure in the hope that it can be useful to other researchers (**Figure 1**). In addition, we showed that, although the chloroplast genome coverage increased with the number of reads used for the assembly, 1 M reads was sufficient to obtain a 95% coverage of the cp genome. These results corroborate that the chloroplast could be fully transcribed and easily assembled from transcriptomic data at low-medium coverage. Moreover, cross-validation (Supplementary **Table S8**) showed that assembling the cp genome using transcriptomes from the SRA database is feasible even though the reliability of the assemblage is always a function of the tissue and methodology used. For example, *Arabidopsis thaliana* cp genomes, assembled from RNA-Seq data coming from different libraries (SRR667021 and SRR6757372), produced different assembly results that were related to differences in coverage and number of reads. Furthermore, the genome of *E. cheiri* cpDNA was surprisingly not fully assembled despite being closely related to the *Erysimum* species used in this study (Moazzeni et al., 2014). However, this result may be explained by the fact that this *E. cheiri* transcriptome was obtained from petals. The reliability of our results is probably attributable to careful sample

preparation (our RNA-Seq samples were submitted to a treatment that depleted rRNA implying that the samples were enriched in the other types of RNA), and because sequencing depth, at least over a minimum threshold ~5 M reads (see **Figure S4**), does not appear to be a crucial factor. Therefore, as a general rule, samples obtained from photosynthetic tissues, depleted in rRNA, and high quality sequenced (as indicated by quality scores) are likely to be trustworthy.

We conclude that assembling cp genomes from good quality transcriptomic data (either obtained *de novo* or downloaded from public databases such as the SRA database) may be a straightforward approach in plant systematics and phylogeny. In fact, this approach may reduce the risk of incorporating NUPTs, avoiding posterior phylogenetic incongruences. However, precautions must be taken due to the possibility of RNA editing, and alternative methods (Kim et al., 2015) could also be used to minimize the assembly of NUPTs or other nuclear DNA into cp genomes. In summary, we think the pipeline presented here is an accessible and time-saving approach to produce high-quality cp genomes that could complement other genomic approaches.

Data Accessibility

Chloroplast genomes were submitted to GenBank with the following accession numbers: *E. baeticum*: Ebb07 (MH414570), Ebb09 (MH414571), Ebb10 (MH414572), Ebb12 (MH414573); *E. mediohispani- cum*: Em21(MH414574), Em21 (MH414581), Em39 (MH414575), Em71 (MH414576); *E. nevandese*: En14 (MH414577), En10 (MH414578), En12 (MH414579), En05 (MH414580). RNA-Seq and genomic raw reads were submitted to Sequence Read Archive with the project accession number SRP149044, and the following samples accession number: *E. baeticum*: Ebb07 (SRR7223707), Ebb09 (SRR7223704), Ebb10 (SRR7223700), Ebb12 (SRR7223699); *E. mediohispanicum*: Em21(SRR7223703), Em39 (SRR7223701), Em71 (SRR7223702), Em21 (SRR7223709). *E. nevandese*: En14 (SRR7223710), En10 (SRR7223706), En12 (SRR7223705), En05 (SRR7223708).

References

Arthofer, W., Schueler, S., Steiner, F. M., & Schlick-steiner, B. C. (2010). Chloroplast DNA-based studies in molecular ecology may be compromised by nuclear-encoded plastid sequence. *Molecular Ecology*, 19 (18), 3853-3856.

Ancev, M. (2006). Polyploidy and hybridization in Bulgarian Brassicaceae: distribution and evolutionary role. *Phytologia Balcanica*, 12 (3), 357-366.

Asaf, S., Khan, A. L., Khan, M. A., Waqas, M., Kang, S. M., Yun, B. W., & Lee, I. J. (2017). Chloroplast genomes of *Arabidopsis halleri* ssp. gemmifera and *Arabidopsis lyrata* ssp. petraea: Structures and comparative analysis. *Scientific Reports*, 7 (1), 1-15.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology,* 19 (5), 455-477.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2018). GenBank. *Nucleic Acids Research*, 46 (D1), D41-D47.

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M., & Dopazo, J. (2015). A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus*. *Molecular Biology and Evolution*, 32 (8), 2015-2035.

Clegg, M. T., Gaut, B. S., Learn, G. H., & Morton, B. R. (1994). Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences*, 91 (15), 6795-6801.

Connett, M. B. (1986). Mechanisms of maternal inheritance of plastids and mitochondria: developmental and ultrastructural evidence. *Plant Molecular Biology Reporter*, 4 (4), 193-205.

Daniell, H., Lin, C. S., Yu, M., & Chang, W. J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology,* 17 (1), 134.

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45 (4), e18-e18.

Do, H. D. K., Kim, J. S., & Kim, J. H. (2013). Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of *Veratrum patulum* O. Loes. (Melanthiaceae). *Gene*, 530 (2), 229-235.

Du, Y. P., Bi, Y., Yang, F. P., Zhang, M. F., Chen, X. Q., Xue, J., & Zhang, X. H. (2017). Complete chloroplast genome sequences of Lilium: insights into evolutionary dynamics and phylogenetic analyses. *Scientific Reports*, 7 (1), 1-10.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Research*, 32 (suppl_2), W273-W279.

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., ... & Koch, M. A. (2017). Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics*, 18 (1), 176.

Gutmann, B., Royan, S., & Small, I. (2017). Protein complexes implicated in RNA editing in plant organelles. *Molecular Plant*, 10 (10), 1255-1257.

Huang, H., Shi, C., Liu, Y., Mao, S. Y., & Gao, L. Z. (2014). Thirteen Camelliachloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evolutionary Biology*, 14 (1), 151.

Henry, R. J. (2005). Plant diversity and evolution: genotypic and phenotypic variation in higher plants. *Cabi Publishing.*

Hollingsworth, M. L., Andra Clark, A. L. E. X., Forrest, L. L., Richardson, J., Pennington, R. T., Long, D. G., ... & Hollingsworth, P. M. (2009). Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources*, 9 (2), 439-457.

Huang, J. L., Sun, G. L., & Zhang, D. M. (2010). Molecular evolution and phylogeny of the angiosperm ycf2 gene. *Journal of Systematics and Evolution*, 48 (4), 240-248.

Hilu, K. W., Borsch, T., Müller, K., Soltis, D. E., Soltis, P. S., Savolainen, V., ... & Sauquet, H. (2003). Angiosperm phylogeny based on matK sequence information. *American Journal of Botany*, 90 (12), 1758-1776.

Hutter, S., Vilella, A. J., & Rozas, J. (2006). Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, 7(1), 409.

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Depamphilis, C. W., Leebens-Mack, J., ... & Chumley, T. W. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences*, 104 (49), 19369-19374.

Jansen, R. K., & Ruhlman, T. A. (2012). Plastid genomes of seed plants. In Genomics of chloroplasts and mitochondria (pp. 103-126). *Springer*, Dordrecht.

Jennings, W. B. (2016). Phylogenomic data acquisition: principles and practice. *CRC Press*.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., ... & Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, 4 (7), 1600016.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. In Bioinformatics for DNA sequence analysis (pp. 39-64). *Humana Press.*

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... & Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28 (12), 1647-1649.

Kim, K., Lee, S. C., Lee, J., Yu, Y., Yang, K., Choi, B. S., ... & Jang, W. (2015). Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Scientific Reports*, 5, 15655.

Kuang, D. Y., Wu, H., Wang, Y. L., Gao, L. M., Zhang, S. Z., & Lu, L. (2011). Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome*, 54 (8), 663-673.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29 (22), 4633-4642.

Léveillé-Bourret, É., Starr, J. R., Ford, B. A., Moriarty Lemmon, E., & Lemmon, A. R. (2018). Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Systematic Biology*, 67 (1), 94-112.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25 (14), 1754-1760.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, 13 (1), 715.

Ma, P. F., Zhang, Y. X., Zeng, C. X., Guo, Z. H., & Li, D. Z. (2014). Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe *Arundinarieae* (Poaceae). *Systematic Biology*, 63 (6), 933-950.

Marhold, K., & Lihová, J. (2006). Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. *Plant Systematics and Evolution*, 259 (2-4), 143-174.

Martin, W., Deusch, O., Stawski, N., Grünheit, N., & Goremykin, V. (2005). Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. *Trends in Plant Science*, 10 (5), 203-209.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17 (1), 10-12.

Matasci, N., Hung, L. H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., ... & Burleigh, J. G. (2014). Data access for the 1,000 Plants (1KP) project. *Gigascience*, 3 (1), 2047-217X.

Matsuo, M., Ito, Y., Yamauchi, R., & Obokata, J. (2005). The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast–nuclear DNA flux. *The Plant Cell*, 17 (3), 665-675.

Michalovová, M., Vyskot, B., & Kejnovsky, E. (2013). Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity*, 111 (4), 314-320.

Moazzeni, H., Zarre, S., Pfeil, B. E., Bertrand, Y. J., German, D. A., Al-Shehbaz, I. A., ... & Oxelman, B. (2014). Phylogenetic perspectives on diversification and character evolution in the species-rich genus *Erysimum* (Erysimeae; Brassicaceae) based on a densely sampled ITS approach. *Botanical Journal of the Linnean Society*, 175 (4), 497-522.

Ni, L., Zhao, Z., Xu, H., Chen, S., & Dorje, G. (2017). Chloroplast genome structures in *Gentiana* (Gentianaceae), based on three medicinal alpine plants used in Tibetan herbal medicine. *Current Genetics*, 63 (2), 241-252.

Noutsos, C., Richly, E., & Leister, D. (2005). Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Research*, 15 (5), 616-628.

Palmer, J. D. (1985). Comparative organization of chloroplast genomes. *Annual Review of Genetics*, 19 (1), 325-354.

Petit, R. J., Duminil, J., Fineschi, S., Hampe, A., Salvini, D., & Vendramin, G. G. (2005). Invited review: comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology*, 14 (3), 689-701.

Pfannschmidt, T. (2003). Chloroplast redox signals: how photosynthesis controls its own genes. *Trends in Plant Science*, 8 (1), 33-41.

Price, A. M., Orellana, D. F. A., Salleh, F. M., Stevens, R., Acock, R., Buchanan-Wollaston, V., ... & Rogers, H. J. (2008). A comparison of leaf and petal senescence in wallflower reveals common and distinct patterns of gene expression and physiology. *Plant Physiology*, 147 (4), 1898-1912.

Richly, E., & Leister, D. (2004). NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Molecular Biology and Evolution*, 21 (10), 1972-1980.

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., & Burleigh, J. G. (2014). From algae to angiosperms–inferring the phylogeny of green plants (*Viridiplantae*) from 360 plastid genomes. *BMC Evolutionary Biology*, 14 (1), 23.

Sablok, G., Mudunuri, S. B., Edwards, D., & Ralph, P. J. (2016). Chloroplast genomics: Expanding resources for an evolutionary conserved miniature molecule with enigmatic applications. *Current Plant Biology*, 7, 34-38.

Schattner, P., Brooks, A. N., & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*, 33 (suppl_2), W686-W689.

Sakaguchi, S., Ueno, S., Tsumura, Y., Setoguchi, H., Ito, M., Hattori, C., ... & Lannuzel, G. (2017). Application of a simplified method of chloroplast enrichment to small amounts of tissue for chloroplast genome sequencing. *Applications in Plant Sciences*, 5 (5), 1700002.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., & Tabata, S. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research*, 6 (5), 283-290.

Scarcelli, N., Mariac, C., Couvreur, T. L. P., Faye, A., Richard, D., Sabot, F., ... & Vigouroux, Y. (2016). Intra-individual polymorphism in chloroplasts from NGS data: Where does it come from and how to handle it?. *Molecular Ecology Resources*, 16 (2), 434-445.

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., ... & Small, R. L. (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany,* 92 (1), 142-166.

Shi, C., Wang, S., Xia, E. H., Jiang, J. J., Zeng, F. C., & Gao, L. Z. (2016). Full transcription of the chloroplast genome in photosynthetic eukaryotes. *Scientific Reports*, 6, 30135.

Sooknanan, R., Pease, J., & Doyle, K. (2010). Novel methods for rRNA removal and directional, ligation-free RNA-Seq library preparation. *Nature Methods*, 7 (10), i-ii.

Timme, R. E., Bachvaroff, T. R., & Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One*, 7 (1).

Thiel, T. (2003). MISA—Microsatellite identification tool. Website http://pgrc. ipk-gatersleben. de/misa/.

Twyford, A. D., & Ness, R. W. (2017). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources*, 17 (5), 858-868.

Wang, M., Cui, L., Feng, K., Deng, P., Du, X., Wan, F., ... & Nie, X. (2015). Comparative analysis of Asteraceae chloroplast genomes: structural organization, RNA editing and evolution. *Plant Molecular Biology Reporter*, 33 (5), 1526-1538.

Wicke, S., Schneeweiss, G. M., Depamphilis, C. W., Müller, K. F., & Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology*, 76 (3-5), 273-297.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., ... & Ruhfel, B. R. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111 (45), E4859-E4868.

Williams, A. V., Boykin, L. M., Howell, K. A., Nevill, P. G., & Small, I. (2015). The complete sequence of the *Acacia ligulata* chloroplast genome reveals a highly divergent clpP1 gene. *PLoS One*, 10 (5).

Yang, Y., & Smith, S. A. (2013). Optimizing *de novo* assembly of short-read RNA-Seq data for phylogenomics. *BMC Genomics*, 14 (1), 328.

Yap, J. Y. S., Rohner, T., Greenfield, A., Van Der Merwe, M., McPherson, H., Glenn, W., ... & Wilkins, M. R. (2015). Complete chloroplast genome of the wollemi pine (*Wollemia nobilis*): structure and evolution. *PloS One*, 10 (6).

Yoshida, T., Furihata, H. Y., & Kawabe, A. (2014). Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species. *DNA Research*, 21 (2), 127-140.

Zhang, Y., Li, L., Yan, T. L., & Liu, Q. (2014). Complete chloroplast genome sequences of *Praxelis* (*Eupatorium catarium* Veldkamp), an important invasive species. *Gene*, 549 (1), 58-69.

Zhang, Y., Iaffaldano, B. J., Zhuang, X., Cardina, J., & Cornish, K. (2017). Chloroplast genome resources and molecular markers differentiate rubber dandelion species from weedy relatives. *BMC Plant Biology*, 17 (1), 34.

Supplementary Material

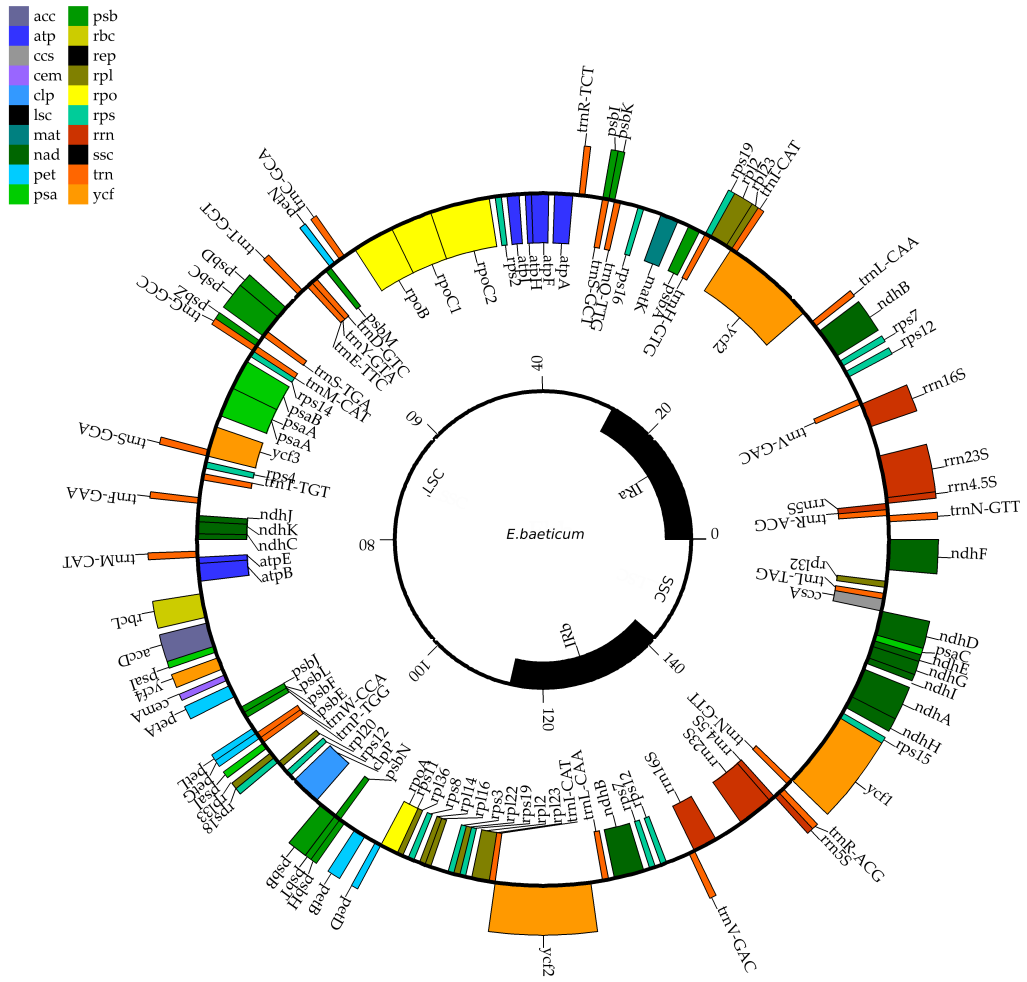

Figure S1. Chloroplast genome map of *Erysimum baeticum*. Genes depicted inside the circle are transcribed clockwise, and those outside are counterclockwise. Genes belonging to different functional groups are shown in different colors.

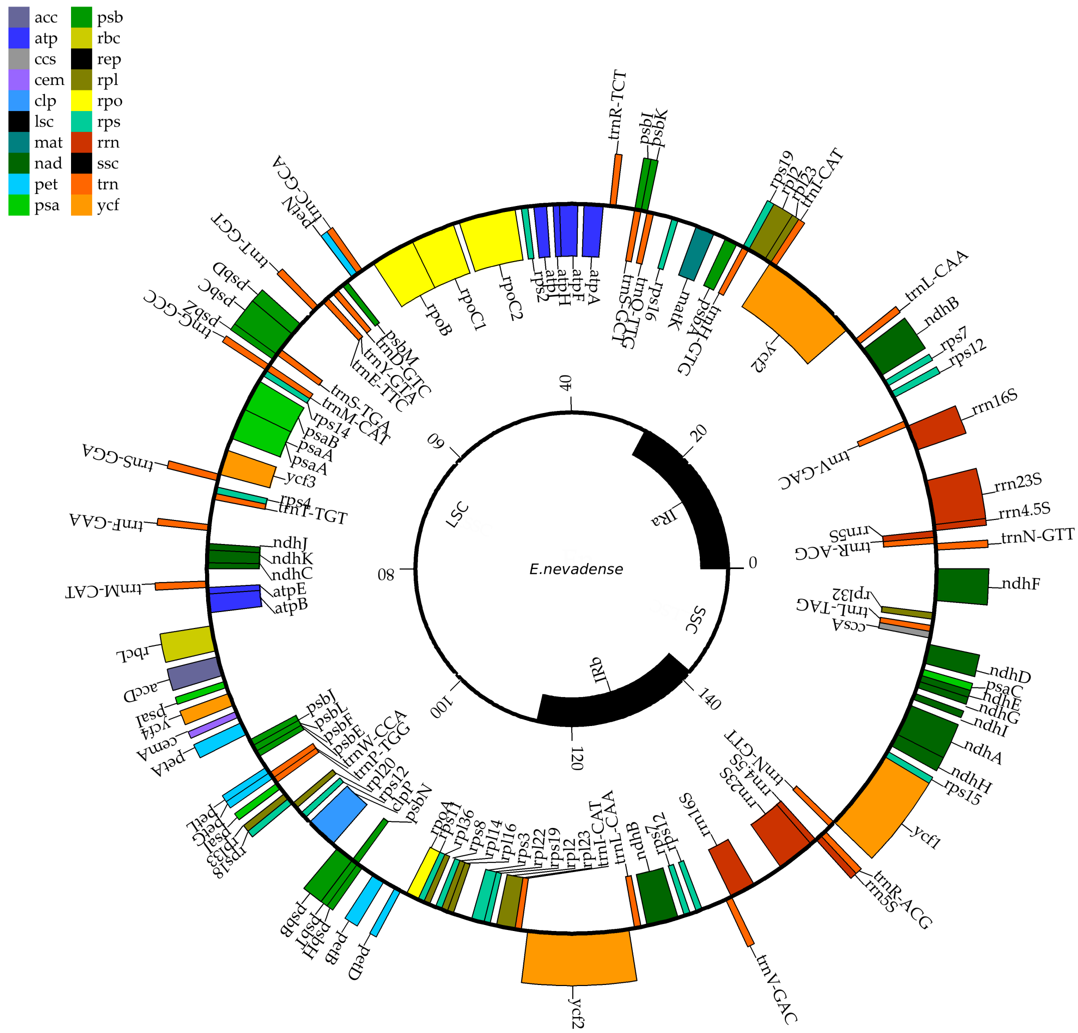

Figure S2. Chloroplast genome map of *Erysimum nevadense.* Genes depicted inside the circle are transcribed clockwise, and those outside are counterclockwise. Genes belonging to different functional groups are shown in different colors.

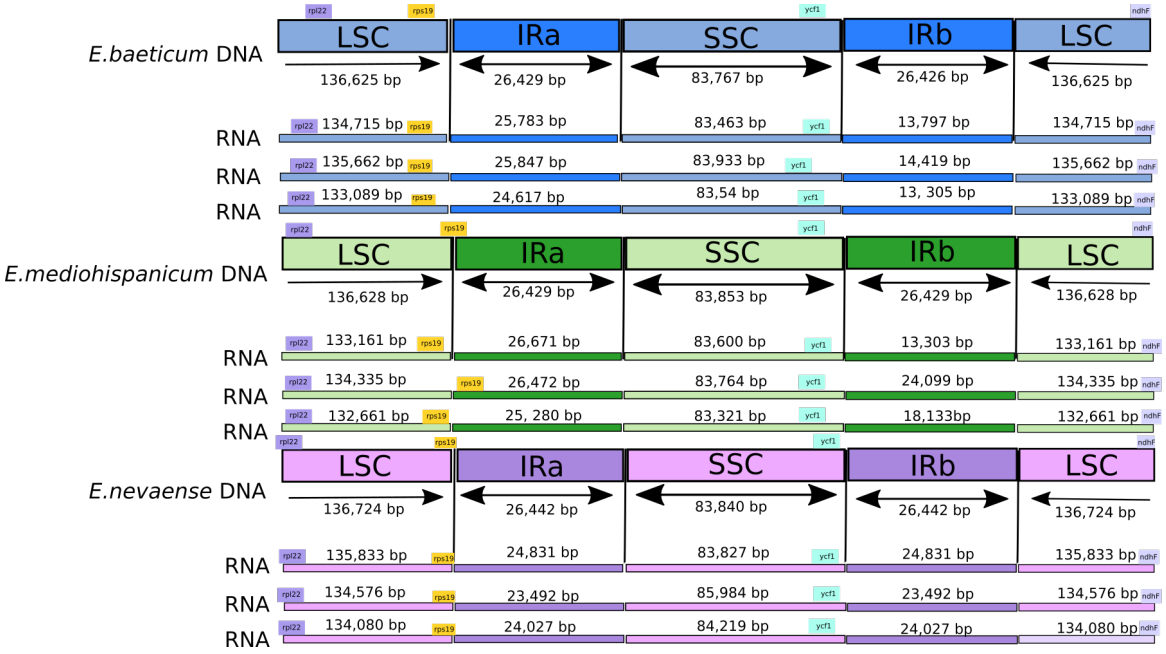

Figure S3. Comparison of the borders of LSC, SSC and IR chloroplast regions among the three replicates of RNA-Seq (RNA-Seq) and the genomic libraries. Selected genes or portions of genes are indicated by boxes.

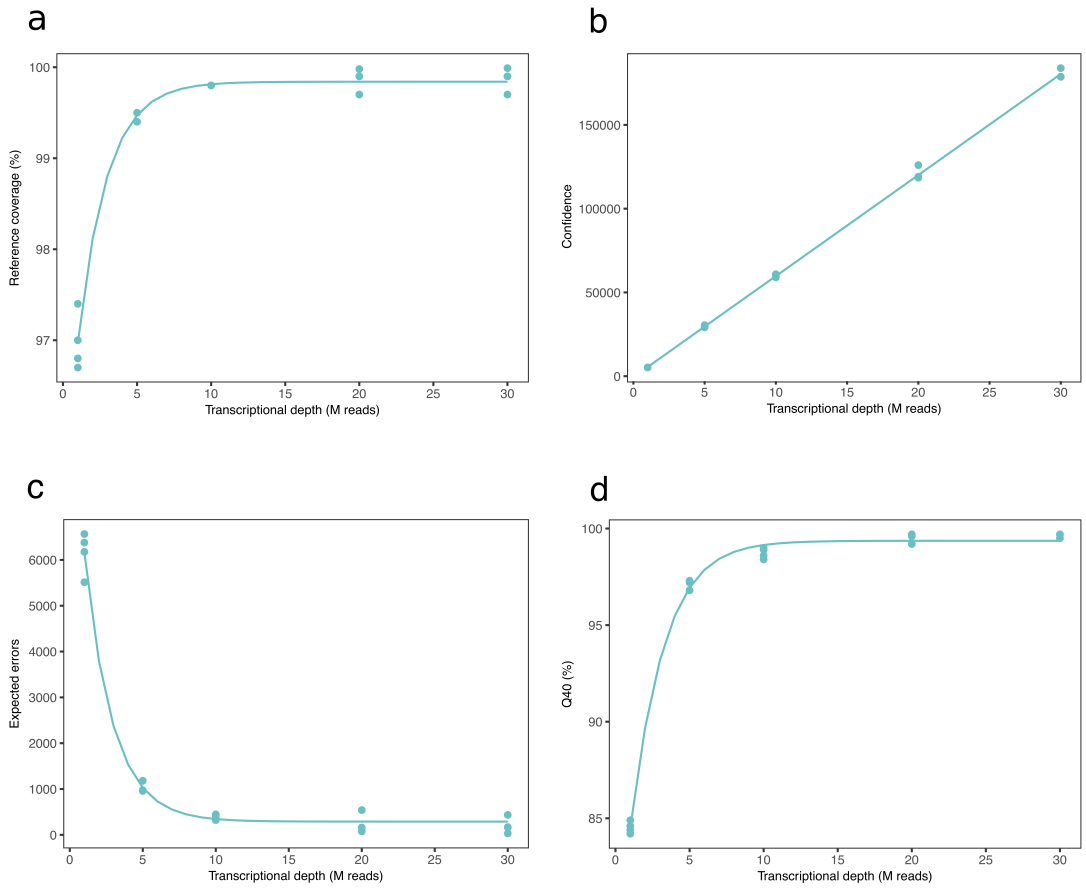

Figure S4. Mapping metrics vary with transcriptomic depth. To analyze the impact that sequencing depth has in the assemblage of transcriptome data into a complete cpDNA genome, we subsampled the RNA-Seq of *E. nevadense* four times at 1 M, 5 M, 10 M, 20 M and 30 M paired reads and mapped them to the cpDNA of *E. mediohispanicum.* **(a)** Percentage of reference coverage in the consensus sequence. Fit curve was adjusted to an asymptotic grown function. **(b)** Mean confidence of the mapping based on quality scores of the reads. This is a measure of quality, with higher values indicating that a base call is more likely to be correct. Line fit was adjusted to a linear function. **(c)** Expected errors. Quality scores of the reads give the approximate number of errors that are statistically expected in mapping. The expected error value is then calculated by summing up the error rates for each base. Curve fit was adjusted to a decreasing exponential function. **(d)** Percentage of Q40 positions. Curve fit was adjusted to an asymptotic grown function.

Library	Species	Population code	Reads (bp)	Gb
DNA genomic	*E. baeticum*	Ebb09	403,325,778	36
	E. mediohispanicum	Em21	451,077,550	43
	E. nevadense	En14	419,863,768	37
RNA-Seq	*E. baeticum*	Ebb07	68,491,616	5.1
	E. baeticum	Ebb10	134,987,506	9.9
	E. baeticum	Ebb12	136,653,914	10.1
	E. mediohispanicum	Em39	153,806,948	11.2
	E. mediohispanicum	Em21	67,787,608	4.7
	E. mediohispanicum	Em71	67,833,508	5.0
	E. nevadense	En05	147,450,544	10.8
	E. nevadense	En10	159,239,340	11.4
	E. nevadense	En12	67,164,590	5.1

Table S1. Summary of the sequencing statistics.

Category	Group of Genes	Name of Genes
Photosynthesis	Subunits of ATP synthase	atpA, atpB, atpE, atpF, atpH, atpI
	Subunits of protochlorophyllide reductase	chl
	Subunits of NADH-dehydrogenase	ndhA,ndhB,ndhC, ndhE, ndhG, ndhH, ndhI, ndhJ, ndhK
	Subunits of cytochrome b/f complex	petA, petB, petD, petG, petL, petN
	Subunits of photosystem I	psaA, psaB, psaC, psaI, psaJ
	Subunits of photosystem II	psbA, psbB, psbC, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
	Subunit of rubisco	rbcl
Self replication	Large subunit of ribosome	rpl 2, 14, 16, 20, 22, 23, 32, 33, 36
	DNA dependent RNA polymerase	rpo A, B, C1, C2
	Small subunit of ribosome	rps 3, 4, 7, 8, 11, 12, 14, 15, 16, 18, 19
	rRNA Genes	rrn rrn16S, rrn23S, rrn4.5S, rrn5S
	tRNA Genes	trn trnC-GCA, trnD-GTC, trnE-TTC, trnF-GAA,trnG-GCC, trnH-GTG, trnI-CAT, trnI-CAT, trnL-CAA, trnL-CAA, trnL-TAG, trnM-CAT, trnM-CAT, trnN-GTT, trnN-GTT, trnP-TGG, trnQ-TTG, trnR-ACG, trnR-ACG, trnR-TCT, trnS-GCT, trnS-GGA, trnS-TGA, trnT-GGT, trnT-TGT, trnV-GAC, trnV-GAC, trnW-CCA, trnY-GTA
Unknown function	Conserved open reading frames	ycf1, ycf2, ycf3, ycf4, ycf5
Other	Subunit of Acetyl-CoA-carboxylase	aacD
	c-type cytochrom synthesis gene	ccsA
	Envelop membrane protein	cemA
	Protease	clpP
	Translational initiation factor	infA
	Maturase	matK
	Elongation factor	tuf

Table S2. List of genes found in the *Erysimum* cpDNA genome.

Gene	Intron length
ndhB	685
rpl2	682
atpF	723
rpoC1	799
psaA	30
ycf3	710
clpP	873
ndhA	1070

Table S3. Length of the intronic regions for the eight split genes.

Taxon	Population	Library	N. of bases (bp)	Mean Cov.	P.wise (%)
E. baeticum	Ebb07	DNA genomic	3,742,777	4,513	97.3
	Ebb09	RNA-Seq	14,099,423	12,827.6	99.7
	Ebb10	RNA-Seq	9,995,492	12,171	96.4
	Ebb12	RNA-Seq	9,692,811	11,468	92.9
E. mediohispanicum	Em21	DNA genomic	43,187,499	39,097.9	98.9
	Em21	RNA-Seq	3,257,794	5,454.6	97.1
	Em71	RNA-Seq	5,441,636	8,809.3	95.8
	Em39	RNA-Seq	13,294,086	17,269.6	95.8
E. nevadense	En14	DNA genomic	56,955,629	52,891.8	98.6
	En05	RNA-Seq	9,342,246	12,909.8	96.1
	En10	RNA-Seq	12,271,273	17,956.2	96.1
	En12	RNA-Seq	5,627,498	7,805.9	96.4

Table S4. Summary of the assembly statistics: Number of bases (the number of bases assembled), mean coverage (the mean of the coverage for each base in the consensus sequence), and the pairwise identity (the average % of identity -considering ambiguity characters- over the alignment, with 100 indicating identical agreement).

Taxon	Population	Type of library	Length	Assembled reads	IRa	SSC	IR b
E. baeticum	Ebb09	Genomic DNA	154,581	983,811	26,429	83,767	26,426
	Ebb07	RNA-Seq libraries	154,691	3,395,688	26,429	95,235	13,675
	Ebb10	RNA-Seq libraries	154,564	8,836,821	25,702	95,456	14,444
	Ebb12	RNA-Seq libraries	154,761	9,126,450	24,987	95,167	13,304
E. mediohispanicum	Em21	Genomic DNA	154,599	1,414,714	26,429	98,752	13,303
	Em71	RNA-Seq libraries	154,788	4,999,161	24,617	98,165	14,919
	Em39	RNA-Seq libraries	154,687	11,730,858	26,356	98,248	18,345
	Em21	RNA-Seq libraries	154,251	1,988,198	25,280	89,248	18, 333
E. nevadense	En14	Genomic DNA	154,661	1,554,542	26,442	83,840	26,442
	En05	RNA-Seq libraries	153, 467	10,925,889	25,634	83,357	24, 833
	En10	RNA-Seq libraries	154,832	8,326,981	25,900	83,212	22, 256
	En12	RNA-Seq libraries	154,701	4,751,517	25,760	80,119	23, 027

Table S5. Summary of characteristics of referenced-based of *Erysimum* cp genomes assembly using BWA. Type of library (genomic DNA or RNA-Seq library), length of the cp genome (bp), number of assembled reads, length of the two inverted repeats (IR a and the IR b), length of the small single copy (SSC) and of the large single copy (LSC) regions, and GC% content.

Taxon	Population code	Library	A	C	G	T	AT	TA	TTA	ATAG	CAAA	TAAA	TTAA	ATTAG	ATAGAA
E. baeticum	Ebb07	DNA genomic	23	2	1	24	11	8	1	1	2	1	1	1	1
	Ebb09	RNA-Seq	18	2	1	23	8	5	1	0	1	2	1	1	1
	Ebb10	RNA-Seq	17	2	1	20	11	5	1	0	1	2	1	1	1
	Ebb12	RNA-Seq	17	2	1	21	11	5	1	0	1	2	1	1	1
E. mediohispanicum	Em21	DNA genomic	19	2	1	27	11	6	1	1	1	2	1	1	1
	Em21	RNA-Seq	16	1	0	21	9	5	1	0	1	1	1	1	0
	Em71	RNA-Seq	16	2	1	21	11	5	1	0	0	1	2	1	1
	Em39	RNA-Seq	17	2	1	21	11	5	1	0	1	2	1	1	1
E. nevadense	En14	DNA genomic	24	2	1	25	12	6	2	1	2	1	1	2	0
	En05	RNA-Seq	18	2	1	21	11	5	1	0	1	2	1	1	1
	En10	RNA-Seq	17	2	1	21	11	5	1	0	1	2	1	1	1
	En12	RNA-Seq	17	2	1	18	8	3	1	0	1	2	1	1	1

Table S6. Numbers of SSR's that differed quantitatively between RNA-Seq and genomic assemblages of cp genomes.

Category	**Gen**	***E. nevadense***		***E. mediohispanicum***		***E. baeticum***	
		RNA-Seq	**RNA-Seq + genomic**	**RNA-Seq**	**RNA-Seq + genomic**	**RNA-Seq**	**RNA-Seq + genomic**
Photosynthesis genes	rbcl	100	99.9	100	100	100	100
	psaA	100	99.9	100	100	99.9	99.8
	psbA	100	99.9	99.9	99.9	99.98	99.97
	ndhK	99.9	99.8	99.8	99.9	100	100
	atpA	100	99.8	100	100	99.9	99.9
	atpH	99.2	99.4	100	100	99.6	99.5
Self replication	rpoA	100	99.8	100	99.97	99.9	99.9
	rps3	99.97	99.9	99.97	99.9	99.8	99.9
	rrn16S	100	99.8	100	100	99.9	99.9
	trnH	99.9	100	99.96	99.97	99.9	99.9
Other Genes	matK	100	100	100	100	100	100
	ycf2	100	99.9	100	100	99.9	99.9

Table S7. Percentage of pairwise identity in gene comparisons across the three replicates of RNA-Seq (RNA-Seq) and including the genomic libraries (RNA-Seq + genomic).

Species	RNA-Seq library	Tissue	Genbank chloroplast genome reference	Consensus lenght (bp)	Confidence mean	Q20 (%)	Q30 (%)	Q40 (%)	Assembled reads	Pairwise identity (%)	Mean coverage	Ref-Seq (%)
Arabidopsis thaliana	SRR6676021	leaf	*A. thaliana* NC_001666 (140,384 bp)	150,757	1,711.1	80.7	77.6	71.6	608,988	89.3	315.1	96.0
Arabidopsis thaliana	SRR6757372	leaf	*A. thaliana* NC_001666 (140,384 bp)	138,157	24,592.1	85.4	83.2	80.9	6,618,957	99.8	1184.1	91.4
Erysimum cheiri	SRR5195369	petal	*E. mediohispanicum* MH414570 (154,599 bp)	79,757	1,585.2	66.8	64.8	57.0	54,362	76.2	136.1	86.1
Moricandia arvensis	SRR4296231	leaf	*Brassica napus* GQ861354 (152,860 bp)	139,300	4,030.8	67.9	66.9	60.7	361,750	94.0	353.9	80.6
Moricandia suffruticosa	SRR4296233	leaf	*Brassica napus* GQ861354 (152,860 bp)	135,131	2,834.7	63.7	62.1	54.5	231,408	88.4	225.0	76.8
Oryza sativa	SRR7079258	seedling	*O. sativa* NC_001320 (134,525 bp)	103,417	514.7	49.5	43.0	39.5	79,647	93.7	114.6	79.8
Zea mays	ERR1407273	seedling root	*Z. mays* NC_001666 (140,384 bp)	112,606	1,741.5	68.3	66.7	61.9	58,591	71.4	211.1	78.7

Table S8. cp genome assembly statistics from *A. thaliana*, *E. cheiri*, *M. arvensis*, *M. sufructicosa*, *O. sativa* and *Z. mays* RNA-Seq reads using the proposed bioinformatic approach.

Chapter III

Genomic resources for *Erysimum* spp. (Brassicaceae):

Transcriptome and chloroplast genomes

Abstract

We have sequenced the floral transcriptomes of 18 populations from seven species of the genus *Erysimum* (Brassicaceae). Transcriptomes were *de novo* assembled and annotated. Moreover, we assembled the whole chloroplast genomes from nine RNA-Seq libraries and reconstructed a calibrated phylogeny for these species. Here, for these 18 floral transcriptomes, we present the RNA-Seq raw reads, the sets of assembled unigenes, their predicted coding sequences and proteins, and the annotation of these unigenes. The resources presented here represent reliable reference sequences for whole-transcriptome and proteome studies for other Brassicaceae, from primer design to phylotranscriptomics.

Introduction

Erysimum (Brassicaceae) is a genus of more than 200 species (Al-Shehbaz, 2012). It is widely distributed in the N. Hemisphere and has been the focus of active research in ecology, evolution, and genetics (e.g., Gómez and Perfectti, 2010; Gómez, 2012; Valverde et al., 2016). Despite this long-standing interest in *Erysimum*, its taxonomy remains to be established appropriately, partly due to a complex and reticulated evolutionary history that renders phylogenetic reconstructions highly challenging (Ancev, 2006; Marhold and Lihová, 2006; Abdelaziz et al., 2014; Moazzeni et al., 2014; Gómez et al., 2014; Züst et al., 2019).

The Baetic Mountains (SE Iberia) are among the most important glacial refugia in Europe; the waxing and waning of plant populations following climatic fluctuations have likely complicated the distribution and genetic variation of extant diversity. Isolation and posterior secondary contact among different species may have favored hybridization and introgression (Médail and Diadema., 2009). The species of *Erysimum* that inhabit these mountains have been a particularly fruitful system to address plant evolutionary ecology (e.g., Gómez et al., 2006; Gómez et al., 2008; Gómez and Perfectti., 2010; Gómez, 2012; Valverde et al., 2016). However, the relationships among these species remain unresolved, hampering comparative and evolutionary studies. Genome duplications, incomplete lineage sorting, and hybridization have compromised the phylogenetic reconstructions within *Erysimum* (Marhold and Lihová., 2006). Additionally, clarifying the complex evolution of this group requires extensive and detailed genomic resources currently being produced but are mostly lacking.

The fast development of high-throughput sequencing technologies leads to a rapid increase in the availability of genomic and transcriptomic information for many plant species (Dong et al., 2004; Duvick et al., 2007; Sundell et al., 2015; Boyles et al., 2019). However, obtaining complete genome sequencing remains a challenge with large, repetitive-DNA enriched genomes. Transcriptome sequencing is comparatively more accessible, providing a relatively

cheap and fast method to obtain large amounts of functional genomic data (Timme et al., 2012; Wickett et al., 2014; Yang and Smith., 2013; Léveillé-Bourret et al., 2017). Accordingly, global initiatives such as the 1,000 plants (1KP) project have generated transcriptomic resources for over 1,000 plant species (Matasci et al., 2014; Leebens-Mack et al., 2019). The use of RNA-Seq could be particularly useful for obtaining complete chloroplast genomes in a reliable and accessible way, making the use of complete molecules in phylogenomic analyses (Smith, 2013; Osuna-Mascaró et al., 2018; Morales-Briones et al., 2019).

Here, we report 18 floral transcriptomes *de novo* assembled from total RNA-Seq libraries and annotated, and nine chloroplast genomes from *Erysimum* populations belonging to seven species inhabiting the Baetic Mountains. The chloroplast genomes were assembled from total RNA-Seq data following a reference assemble approach, previously validated in Osuna-Mascaró et al., 2018 (Osuna-Mascaró et al., 2018). The data presented here represent reliable genomic resources for transcriptomic, proteomic, and phylotranscriptomic studies. These data contribute to the *Erysimum* genus' resources, being the only genomic resources for these species coming from flower buds.

Generation of the datasets

We sampled pre-anthesis flower buds at the same development stage from three different *Erysimum mediohispanicum*, *E. nevadense*, *E. popovii*, and *E baeticum*, four populations of *E. bastetanum*, and one population of *E. lagascae*, and *E. fitzii* each (see **Table 1** for details). The samples were stored in liquid nitrogen until RNA extraction. Then, the buds were ground with mortar and pestle in liquid nitrogen. Total RNA was isolated using the Qiagen RNeasy Plant Mini Kit following the manufacturer's protocol. Library preparation and RNA sequencing were conducted at Macrogen Inc. (Seoul, Korea). We used rRNA-depletion (Ribo-Zero) for mRNA enrichment and to avoid sequencing rRNAs. Library preparation was performed using the TruSeq Stranded Total RNA LT Sample Preparation Kit (Plant). The sequencing of the 18 libraries was

carried out using the Hiseq 3000-4000 sequencing protocol and TruSeq 3000-4000 SBS Kit v3 reagent, following a paired-end 150 bp strategy on the Illumina HiSeq 4000 platform. A summary of sequencing statistics appears in **Table 2**.

Taxon	Population	Location	Elevation	Geographical coordinates
E. baeticum	Ebb07	Sierra Nevada, Almería, Spain	2128	37°05′46″N, 3°01′01″W
	Ebb10	Sierra Nevada, Almería, Spain	2140	37°05′32″N, 3°00′40″W
	Ebb12	Sierra Nevada, Almería, Spain	2264	37°05′51″N, 2°58′06″W
E. bastetanum	Ebt01	Sierra de Baza, Granada, Spain	1990	37°22′52″N, 2°51′49″W
	Ebt12	Sierra de María, Almería, Spain	1528	37°41′03″N, 2°10′51″W
	Ebt13	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W
	Ebt22	Sierra Morena, Ciudad Real, Spain	483	38°26′44″N, 3°56′23″W
E. fitzii	Ef01	Sierra de la Pandera, Jaén, Spain	1804	37°37′56″N, 3°46′46″W
E. lagascae	Ela07	Sierra de San Vicente, Toledo, Spain	516	44°05′49″N, 4°40′40″W
E. mediohispanicum	Em21	Sierra Nevada, Granada, Spain	1723	37°08′04″N, 3°25′43″W
	Em39	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W
	Em71	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W
E. nevadense	En05	Sierra Nevada, Granada, Spain	2074	37°06′35″N, 3°01′32″W
	En10	Sierra Nevada, Granada, Spain	2321	37°06′37″N, 3°24′18″W
	En12	Sierra Nevada, Granada, Spain	2255	37°05′37″N, 2°56′19″W
E. popovii	Ep16	Jabalcuz, Jaén, Spain	796	37°45′26″N, 3°51′02″W
	Ep20	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W
	Ep27	Llanos del Purche, Granada, Spain	1470	37°07′46″N, 3°28′48″W

Table 1. Population code, location, and details of sympatry status for all of the populations sampled.

Taxon	Population code	Total read bases (bp)	Total number of reads (bp)	GC (%)	AT (%)	Q20 (%)	Q30 (%)	Gb	SRA accession numbers
E. baeticum	Ebb07	10,342,234,016	68,491,616	47.38	52.61	96.13	91.88	5.1	SRX4130235
	Ebb10	20,383,113,406	134,987,506	47.25	52.75	96.63	92.11	9.9	SRX4130242
	Ebb12	20,634,741,014	136,653,914	47.30	52.70	96.71	92.24	10.1	SRX4130243
E. bastetanum	Ebt01	11,454,840,974	75,859,874	44.88	55.11	99.89	99.10	5.0	SRX7756231
	Ebt12	10,417,333,262	68,988,962	45.71	54.28	98.22	95.24	5.2	SRX7756232
	Ebt13	10,777,576,680	71,374,680	45.47	54.52	98.58	95.91	5.4	SRX7756233
	Ebt22	10,404,933,746	68,906,846	45.49	54.51	98.23	95.280	5.2	SRX7756234
E. fitzii	Ef01	10,156,440,596	67,261,196	47.09	52.91	98.94	96.66	4.6	SRX7756235
E. lagascae	Ela07	7,201,775,508	71,304,708	43.50	56.49	96.47	93.68	5.4	SRX7756236
E. mediohispanicum	Em21	10,235,928,808	67,787,608	48.86	51.14	96.05	91.72	4.7	SRX4130233
	Em39	23,224,849,148	153,806,948	47.49	52.51	96.79	92.43	11.2	SRX4130241
	Em71	10,242,859,708	67,833,508	46.51	53.48	98.18	95.14	5.0	SRX4130240
E. nevadense	En05	22,265,032,144	147,450,544	46.73	53.27	96.65	92.20	10.8	SRX4130234
	En10	24,045,140,340	159,239,340	44.94	55.06	96.90	92.72	11.4	SRX4130236
	En129	10,141,853,090	67,164,590	47.49	52.50	96.17	91.96	5.1	SRX4130237
E. popovii	Ep16	10,556,586,670	69,911,170	46.26	53.74	96.09	91.78	5.7	SRX7756237
	Ep20	10,860,711,844	71,925,244	45.98	54.01	98.12	94.96	5.5	SRX7756238
	Ep27	25,555,407,006	169,241,106	47.38	52.62	96.64	92.14	12.5	SRX775623

Table 2. Summary of the sequencing statistics.

Data processing and transcriptome analyses

We analyzed the fastq files containing the raw cDNA for each library using FastQC v0.11.5 (Andrews, 2010). Then, we trimmed the adapters using cutadapt v1.15 (Martin, 2011), specifying the "-b" option for trimming the adapters in 5′ and 3′ and the "-n" option to search repeatedly for the adapter sequences. Following, we trimmed the reads by quality using Sickle v1.33 (Joshi, 2011), using the "pe" option for paired-end reads and the "-t" to use Illumina quality values (see https://github.com/najoshi/sickle). To assemble contigs from the resulting high-quality cleaned reads, we followed a *de novo* approach using Trinity v2.8.4 (Grabherr et al., 2011), due to the absence of a published assembled genome available for *Erysimum*. Each library was normalized in silico before assembly to validate and reduce the number of reads using the "insilico_read_normalization.pl" function in Trinity v2.8.4 (Haas et al., 2013). Then we used the parameter 'min_kmer_cov 2' to eliminate single-occurrence k-mers that are profoundly enriched in sequencing errors, following the approach of Haas et al. (2013) (UniProtConsortium, 2014). Thus, only k-mers that occur more than once were considered for contigs. Candidate open reading frames (ORF) within transcript sequences were predicted and translated using TransDecoder v5.2.0 (Haas et al., 2013). We performed functional annotation of Trinity transcripts with ORFs using Trinotate v3.0.1 (Haas et al., 2015), an annotation suite designed for automatic functional annotation of *de novo* assembled transcriptomes. Sequences were searched against UniProt (UniProtConsortium, 2014), using SwissProt databases (with BLASTX and BLASTP searching and an e-value cutoff of 10; Bairoch and Apweiler, 2000). We then used the Pfam database (Bateman et al., 2004) to annotate protein domains for each predicted protein sequence. We also annotated the transcripts using the databases eggnog (Jensen et al., 2007), GO (GeneOntologyConsortium, 2004), and Kegg (Kanehisa and Goto, 2000). The summary statistics of the assembled transcriptomes are presented in **Table 3**. The unigenes from the assembled sequences using different databases are shown in **Table 4**. We used BUSCO v2.0 (Seppey et al.,

2019) to validate the quality of all the assemblies, using the plant database brassicales_odb10.2019-11-20.

Chloroplast genome assembly and annotation

We assembled the whole chloroplast genome from each *Erysimum* species RNA-Seq library. We used a reference assembly approach, using Geneious R.11 (Kearse et al., 2012) with the *A. thaliana* cp genome as reference (NC_000932.1, Sato et al., 1999). We annotated the chloroplast genomes using cpGAVAS (Liu et al., 2012) (see **Table 5**). The chloroplast genomes of *E. mediohispanicum*, *E. nevadense*, and *E. baeticum* were used to validate the reliability of chloroplast genome assembly using RNA-Seq in Osuna-Mascaró et al. 2018 (Osuna-Mascaró et al., 2018).

Time-calibrated phylogeny reconstruction

To reconstruct a dated phylogeny, we used Beast 2.0 (Bouckaert et al., 2014) using the substitution rate for non-codifying plastidial DNA (1.2–1.7 × 109 substitutions/site/year (Graur and Li, 2000). The Bayesian search for tree topologies and node ages were conducted during 20,000,000 generations in BEAST using a strict clock model and a Yule process as prior. MCMC was sampled every 1,000 generations, discarding a burn-in of 10%. We checked the MCMC trace files generated using Tracer v1.6.1 (Rambaut et al., 2014). The time-calibrated phylogeny is shown in Figure 1.

Data records

The raw sequence read data for all the transcriptomes were deposited in the NCBI Sequence Read Archive (Data citation 1). Furthermore, we have created a project on figshare, containing for freely download: the assembled transcriptomes (Data citation 2), the transcriptome annotations (Data citation 3), the set of assembled unigenes, their annotations, and the predicted amino acid sequences (Data citation 4), the chloroplast genomes assemblies and their annotations (Data

citation 5), and chloroplast genome resources as trn's, rrn's, mrn's, genes, trna validation results, and annotation report files (Data citation 6). The chloroplast genome sequences are also deposited in GenBank (Data citation 7).

Taxon	Population code	Total Trinity genes	Total Trinity transcripts	Contig N50	Median contig length	Average contig length	Total assembled bases (bp)
E. baeticum	Ebb07	116,006	188,787	983	423	678.35	128,063,827
	Ebb10	235,313	382,286	859	381	617.37	236,012,489
	Ebb12	171,950	335,960	958	417	663.71	222,979,601
E. bastetanum	Ebt01	164,708	291,831	991	438	688.00	200,778,797
	Ebt12	123,268	212,255	1,088	444	723.11	153,483,728
	Ebt13	186,374	278,526	938	382	639.83	126,262,947
	Ebt22	119,364	197,415	1,156	426	732.89	144,683,034
E. fitzii	Ef01	77,047	130,076	1,502	620	957.81	124,588,696
E. lagascae	Ela07	106,811	203,045	1,361	540	865.68	175,772,242
E. mediohispanicum	Em21	66,162	104,486	1,362	579	888.57	92,843,559
	Em39	93,902	238,394	1,495	662	977.67	233,071,647
	Em71	93,154	160,490	1,128	490	764.13	122,635,013
E. nevadense	En05	92,897	235,515	1,504	663	981.53	231,165,137
	En10	217,656	368,656	1,436	496	867.83	319,929,639
	En12	75,088	126,245	1,212	561	829.65	104,739,622
E. popovii	Ep16	107,329	186,398	1,130	477	757.88	141,267,189
	Ep20	199,665	300,036	738	353	566.56	169,989,446
	Ep27	123,780	288,344	1,249	529	819.94	236,425,328

Table 3. Summary statistics of the transcriptome assembly.

Number of unigenes								
Taxon	**Population code**	**Sprot_top_BLASTX_hit**	**Sprot_top_BLASTP_hit**	**Pfam**	**GO_Pfam**	**GO_BLAST**	**eggog**	**Kegg**
E. baeticum	Ebb07	112543	1844	58464	37009	41713	40129	37760
	Ebb10	197069	111229	97272	61805	184103	172826	164226
	Ebb12	190835	108248	22484	22480	175134	166701	158092
E. bastetanum	Ebt01	165626	98309	85207	52817	149999	142448	135739
	Ebt12	128004	76939	68066	42817	116348	111477	105855
	Ebt13	159869	86090	74989	46573	146455	140640	133137
	Ebt22	132907	85868	76201	48546	118802	113696	108497
E. fitzii	Ef01	92184	65200	57552	37661	83190	79488	75459
E. lagascae	Ela07	132907	85868	76201	48546	118802	113696	108497
E. mediohispanicum	Em21	71606	51019	45366	29311	64679	62559	59930
	Em39	170089	119187	105585	68625	154101	148134	140544
	Em71	159869	86090	74989	46573	146455	140640	133137
E. nevadense	En05	167406	117596	103913	67649	151343	145238	138224
	En10	189900	127250	113593	75293	175240	164625	156297
	En12	85222	59138	52245	33974	77857	74795	71130
E. popovii	Ep16	114873	72485	64212	40623	104434	99335	94405
	Ep20	164455	79736	68739	42568	152320	145732	137445
	Ep27	184887	116140	102230	65460	167458	161254	153256

Table 4. Annotation summary of the assembled transcripts using different databases: SwissProt (BLASTX and BLASTP), Pfam, eggnog, GO, and Kegg.

Taxon	**Population**	**Plastome size (bp)**	**Genes**	**t-RNA**	**r-RNA**	**m-RNA**	**GC %**	**Validation of t-RNA**	**GenBank AN**

E. baeticum	Ebb07	154,791	124	29	8	87	37.5	29	MH414570
	Ebb10	154,768	124	29	8	87	36.5	29	MH414572
	Ebb12	154,761	124	29	8	87	36.5	29	MH414573
E. bastetanum	Ebt01	136,941	111	28	8	75	37.4	28	MT150122
	Ebt12	152,625	94	29	8	94	37.5	29	MT150121
	Ebt13	152,625	83	29	8	75	37.8	29	MT150114
	Ebt22	136,427	75	28	8	75	36.5	28	MT150115
E. fitzii	Ef01	136,877	74	28	8	75	37.4	28	MT150118
E. lagascae	Ela07	136,740	75	28	8	75	37.4	28	MT150116
E. mediohispanicum	Em21	154,251	124	29	8	87	36.5	29	MH414581
	Em39	154,827	124	29	8	87	36.6	29	MH414575
	Em71	154,788	124	29	8	87	36.6	29	MH414576
E. nevadense	En05	153,467	124	29	8	87	36.7	29	MH414580
	En10	154,834	124	29	8	87	36.7	29	MH414578
	En12	154,747	124	29	8	87	36.7	29	MH414579
E. popovii	Ep16	136,812	96	28	8	60	36.8	28	MT150117
	Ep20	136,820	111	28	8	75	37.4	28	MT150119
	Ep27	135,108	104	28	8	75	37.5	28	MT150120

Table 5. Length (bp), number of genes, t-RNA, m-RNA, r-RNA, GC %, and number of t-RNA using a validation method for the chloroplast genomes assembled.

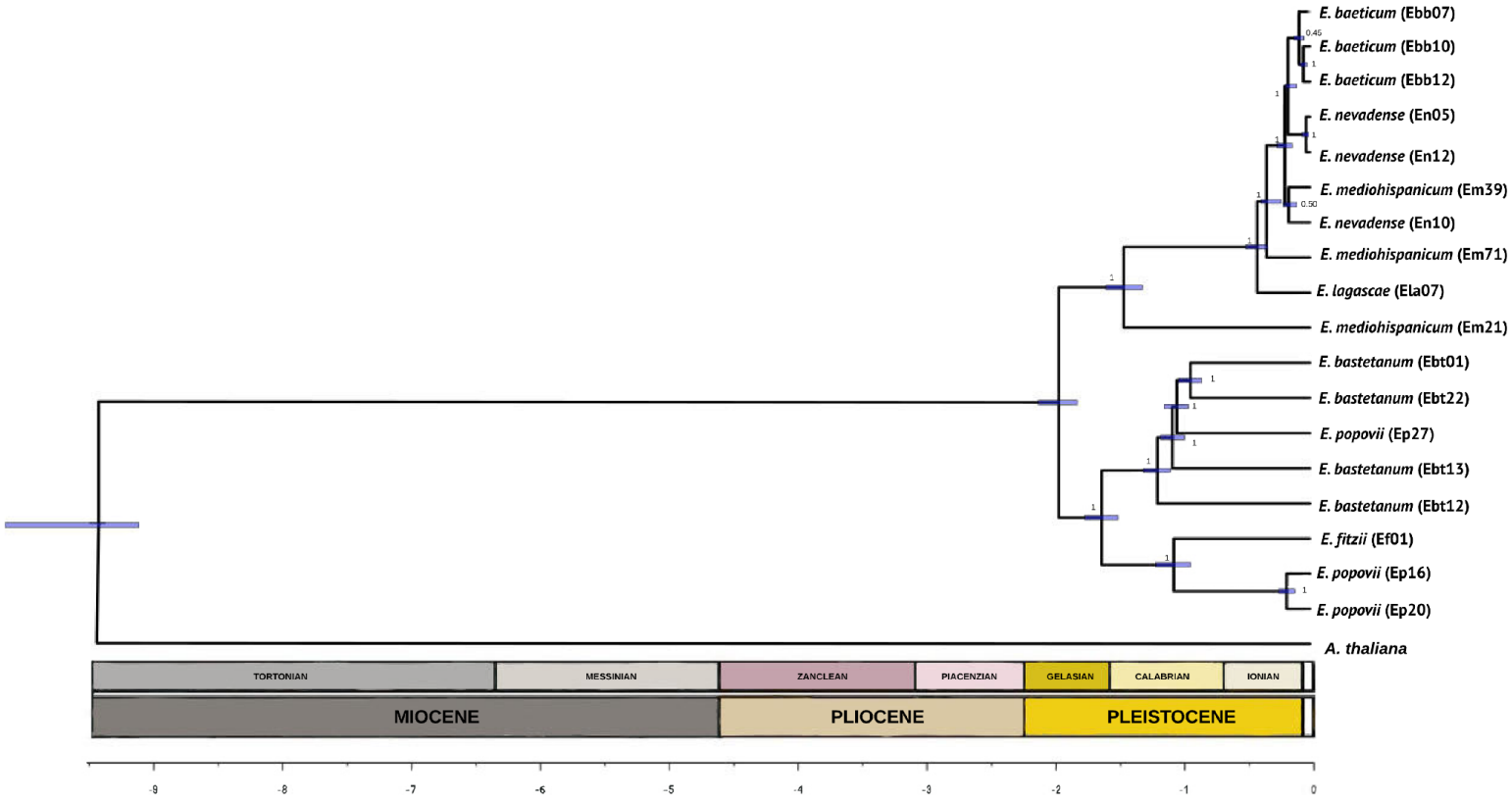

Figure 1. A time-calibrated phylogeny for the complete cpDNA of the different populations of the *Erysimum* species analyzed here. Note the reticulated position of some populations, probably due to hybridization events.

Usage notes

Erysimum is a genus for which phylogenetic relationships have not yet been fully established. Therefore, this dataset's main future uses will likely lie in analyses of molecular evolution aimed at disentangling the taxonomy and biogeography of these and related species. Moreover, since the primary data are transcriptomes, they could be useful in plant evo-devo and physiological studies. They can also be incorporated into comparative studies aimed at identifying the differential expression of the genes expressed in the tissues sequenced in this work (i.e., flower buds).

Technical Validation

Extraction and RNA integrity

The quality and quantity of the RNA obtained was checked using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States) and analyzed with the Agilent 2100 Bioanalyzer system (Agilent Technologies Inc).

Quality trimming

After trimming, we used FastQC (Andrews, 2000) again to verify the trimming efficiency. The summary of the number of reads after the quality trimming is represented in **Table 6**.

Transcriptome completeness

A high level of single-copy orthologous retrieval was noted for the 18 assemblies, with at least a 65% ratio, as shown in **Figure 2**. The least complete case exhibited < 14.93 % missing orthologs.

Chloroplast genome annotation

The annotations were manually curated using Geneious R.11 (Kearse et al., 2012). All transfer RNA sequences (tRNA) encoded in the cp genomes were verified using tRNAscan-SE v2.0 (Schattner et al., 2005) and ARAGORN v1.2.38 (Laslett et al., 2004) with the default search settings.

Taxon	Population code	Total number of reads (bp)	GC (%)
E. baeticum	Ebb07	60,985,000	46
	Ebb10	119,520,613	46
	Ebb12	121,103,833	46
E. bastetanum	Ebt01	75,854,979	44
	Ebt12	65,757,353	45
	Ebt13	65,765,324	45
	Ebt22	31,503,434	45
E. fitzii	Ef01	66,245,223	46
E. lagascae	Ela07	65,740,933	45
E. mediohispanicum	Em21	60,130,101	48
	Em39	136,858,973	47
	Em71	64,619,740	46
E. nevadense	En05	130,509,040	46
	En10	142,305,123	46
	En12	59,838,058	46
E. popovii	Ep16	62,135,880	45
	Ep20	68,273,570	46
	Ep27	149,733,872	46

Table 6. Number of reads after the quality trimming.

Figure 2. BUSCO assessment results for the 18 assembled transcriptomes.

Data citations

Data citation 1: NCBI Sequence Read Archive, BioProject PRJNA607615 under the following accession numbers: SRX7756239, SRX7756238, SRX7756237, SRX7756236, SRX7756235, SRX7756234, SRX7756233, SRX7756232, SRX7756231, and BioProject PRJNA473238 under the following accession numbers: SRX4130243, SRX4130242, SRX4130241, SRX4130240, SRX4130237, SRX4130236, SRX4130235, SRX4130234, SRX4130233.

Data citation 2: figshare https://doi.org/10.6084/m9.figshare.11877786.v3 (2020).

Data citation 3: figshare https://doi.org/10.6084/m9.figshare.11866389.v3 (2020).

Data citation 4: figshare https://doi.org/10.6084/m9.figshare.11873937.v1 (2020).

Data citation 5: figshare https://doi.org/10.6084/m9.figshare.11881656.v2 (2020).

Data citation 6: figshare https://doi.org/10.6084/m9.figshare.11881419.v2 (2020).

Data citation 7: Chloroplast genome sequences deposited in GenBank under the following accession numbers: MH414570, MH414572, MH414573, MH414581, MH414575, MH414576, MH414578, MH414579, MH414580, MT150114, MT150115, MT150116, MT150117, MT150118, MT150119, MT150120, MT150121, MT150122.

References

Al-Shehbaz, I. A. (2012). A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon*, 61 (5), 931-954.

Ančev, M. (2006). Polyploidy and hybridization in Bulgarian Brassicaceae: distribution and evolutionary role. *Phytologia Balcanica*, 12 (3), 357-366.

Abdelaziz, M., Muñoz-Pajares, A. J., Lorite, J., Herrador, M. B., Perfectti, F., & Gómez, J. M. (2014, June). Phylogenetic relationships of *Erysimum* (Brassicaceae) from the baetic mountains (se Iberian Peninsula). *Anales del Jardín Botánico de Madrid* (Vol. 71, No. 1, p. 005).

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28 (1), 45-48.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., ... & Studholme, D. J. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32 (suppl_1), D138-D141.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., ... & Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computer Biology*, 10 (4), e1003537.

Boyles, R. E., Brenton, Z. W., & Kresovich, S. (2019). Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments. *The Plant Journal*, 97 (1), 19-39.

Dong, Q., Schlueter, S. D., & Brendel, V. (2004). PlantGDB, plant genome database and analysis tools. *Nucleic Acids Research*, 32 (suppl_1), D354-D359.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., ... & Brendel, V. (2007). PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research*, 36 (suppl_1), D959-D965.

Gómez, J. M. (2003). Herbivory reduces the strength of pollinator-mediated selection in the Mediterranean herb *Erysimum mediohispanicum*: consequences for plant specialization. *The American Naturalist*, 162 (2), 242-256.

Gómez, J. M., & Perfectti, F. (2010). Evolution of complex traits: the case of *Erysimum* corolla shape. *International Journal of Plant Sciences*, 171 (9), 987-998.

Gomez, J. M., Munoz-Pajares, A. J., Abdelaziz, M., Lorite, J., & Perfectti, F. (2014). Evolution of pollination niches and floral divergence in the generalist plant *Erysimum mediohispanicum*. *Annals of Botany*, 113 (2), 237-249.

Gómez, J. M., Perfectti, F., & Camacho, J. P. M. (2006). Natural selection on *Erysimum mediohispanicum* flower shape: insights into the evolution of zygomorphy. *The American Naturalist*, 168 (4), 531-545.

Gómez, J. M., Bosch, J., Perfectti, F., Fernández, J. D., Abdelaziz, M., & Camacho, J. P. M. (2008). Association between floral traits and rewards in *Erysimum mediohispanicum* (Brassicaceae). *Annals of Botany*, 101 (9), 1413-1420.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29 (7), 644-652.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32 (suppl_1), D258-D261.

Graur, D., & Li, W. H. (2000). Fundamentals of molecular evolution. Sinauer Associations. Inc. Sunderland MA.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & MacManes, M. D. (2013). *De novo* transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8 (8), 1494-1512.

Haas, B. J. (2015). Trinotate: Transcriptome functional annotation and analysis.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., & Bork, P. (2007). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36 (suppl_1), D250-D254.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28 (1), 27-30.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... & Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28 (12), 1647-1649.

Léveillé-Bourret, É., Starr, J. R., Ford, B. A., Moriarty Lemmon, E., & Lemmon, A. R. (2018). Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Systematic Biology*, 67 (1), 94-112.

Transcriptomes, O. T. P. (2019). I: One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574, 679-685.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., & Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, 13 (1), 1-7.

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32 (1), 11-16.

Marhold, K., & Lihová, J. (2006). Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. *Plant Systematics and Evolution*, 259 (2-4), 143-174.

Moazzeni, H., Zarre, S., Pfeil, B. E., Bertrand, Y. J., German, D. A., Al-Shehbaz, I. A., ... & Oxelman, B. (2014). Phylogenetic perspectives on diversification and character evolution in the species-rich

genus *Erysimum* (Erysimeae; Brassicaceae) based on a densely sampled ITS approach. *Botanical Journal of the Linnean Society*, 175 (4), 497-522.

Médail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, 36 (7), 1333-1345.

Matasci, N., Hung, L. H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., ... & Burleigh, J. G. (2014). Data access for the 1,000 Plants (1KP) project. *Gigascience*, 3 (1), 2047-217X.

Morales-Briones, D. F., Kadereit, G., Tefarikis, D. T., Moore, M., Smith, S. A., Brockington, S. F., ... & Yang, Y. (2019). Disentangling Sources of Gene Tree Discordance in Phylotranscriptomic Datasets: A Case Study from Amaranthaceae sl. *BioRxiv*, 794370.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17 (1), 10-12.

Osuna-Mascaró, C., de Casas, R. R., & Perfectti, F. (2018). Comparative assessment shows the reliability of chloroplast genome assembly using RNA-Seq. *Scientific reports*, 8 (1), 1-12.

Rambaut, A., Drummond, A. J., & Suchard, M. (2014). Tracer v1. 6 http://beast. bio. ed. ac. uk. Tracer (Online 2015, May 29).

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., & Tabata, S. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research*, 6 (5), 283-290.

Schattner, P., Brooks, A. N., & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*, 33 (suppl_2), W686-W689.

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Gene Prediction* (pp. 227-245). Humana, New York, NY.

Smith, D. R. (2013). RNA-Seq data: a goldmine for organelle research. *Briefings in Functional Genomics*, 12 (5), 454-456.

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y. C., Sjödin, A., ... & Street, N. R. (2015). The plant genome integrative explorer resource: PlantGenIE. org. *New Phytologist*, 208 (4), 1149-1156.

Timme, R. E., Bachvaroff, T. R., & Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One*, 7 (1), e29696.

UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, 43 (D1), D204-D212.

Valverde, J., Gómez, J. M., & Perfectti, F. (2016). The temporal dimension in individual-based plant pollination networks. *Oikos*, 125 (4), 468-479.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., ... & Ruhfel, B. R. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111 (45), E4859-E4868.

Yang, Y., & Smith, S. A. (2013). Optimizing *de novo* assembly of short-read RNA-Seq data for phylogenomics. *BMC Genomics*, 14 (1), 328.

Züst, T., Strickler, S. R., Powell, A. F., Mabry, M. E., An, H., Mirzaei, M., ... & Petschenka, G. (2020). Independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum*, Brassicaceae). *Elife*, 9, e51712.

Chapter IV

Hybridization and introgression are prevalent in Southern European *Erysimum* (Brassicaceae) species and may mediate corolla color changes

Abstract

Hybridization rules plant evolution. One of their outcomes is introgression, which is considered the transfer of a small amount of the genome from one taxon to another by hybridization and repeated backcrosses. Here, we studied the genomic signature of hybridization and introgression in several *Erysimum* (Brassicaceae) species with purple and yellow corolla from the South of the Iberian Peninsula. Previous studies suggested that purple *Erysimum* species may have a hybrid origin. Accordingly, there is the possibility that purple color has been transferred in the Iberian clade from a purple parental species by hybridization and has been maintained by natural selection. We have sequenced full transcriptomes of yellow and purple *Erysimum* species, and then, we have studied the general hybridization scenario for these species by using different phylogenetic approaches. Furthermore, we have explored if anthocyanin genes have signatures of introgression and positive selection. Our results suggest that the purple and also yellow *Erysimum* species studied here have a strong signature of hybridization and introgression. Moreover, we have found signatures of introgression on two anthocyanin genes of a purple parental species into one purple species and the three yellow species studied here. Overall, all results support a scenario of multiple hybridization events involving not only the purple species but also the yellow ones.

Introduction

Hybridization is widespread across the tree of life, determining the branching and diversification patterns of many taxonomic groups (Rieseberg and Carney, 1998; Coyne and Orr, 2004; Arnold, 2006; Abbott et al., 2013). Because of its pervasiveness, it has been a subject of research for a long time (Stebbins, 1959; Anderson, 1953; Arnold et al., 1999). However, only recently, with the advent of next-generation sequencing, scientists have started to analyze the dynamics of hybridization at a genomic scale, thus rekindling interest in the evolutionary relevance of this phenomenon. This renewed interest has undoubtedly increased our understanding of the role of hybridization in nature (Payseur and Rieseberg, 2016; Goulet et al., 2017; Taylor and Larson, 2019). Nevertheless, many of the factors that determine the incidence (i.e., the probability of new cases arising in a population or taxon) and consequences of hybridization remain unexplored.

Hybridization is particularly relevant for plant evolution, with around 25% of plant species showing evidence of having a hybrid origin (Mallet, 2005; Soltis and Soltis, 2009). The evolutionary outcomes of hybridization may vary widely. Thus, interspecific hybridization may hinder speciation and therefore diversification (Mayr, 1992; Schemske, 2000; Mallet, 2005; Saari and Faeth, 2012; Gómez et al. 2015a), but hybridization may also foster the formation of new species (Rieseberg et al., 2003; Stelkens and Seehausen, 2009) or the introgression of novel genetic variations (by hybridization and repeated backcrossing) (Anderson and Hubricht, 1938; Anderson, 1953; Rieseberg and Wendel, 1993). In addition, in many cases, hybridization events lead to changes in ploidy level, as a result of the fusion of the genomes of the two hybridizing species (Soltis et al., 2014). However, despite some evidence showing introgression on polyploid species (e.g., gene flow between diploid and tetraploid species of *Senecio*; Chapman and Abbott, 2010), there is limited knowledge on the interplay of introgression and polyploidization.

Notably, in some instances, the introgressed genetic variations produce significant evolutionary changes in the new species (Arnold, 2004; Arnold and Kunte, 2017). Indeed, when introgression is adaptive, a process understood as adaptive introgression, these introgressed regions might be instrumental for the acquisition of novel functional traits and could facilitate responses to new selective pressures (Hanušová et al., 2014; Arnold and Martin, 2009; Arnold and Kunte, 2017; Suarez-Gonzalez et al., 2018). However, the analysis of adaptive introgression requires detailed ecological and genetic knowledge to demonstrate an adaptive function for the introgressed genic regions, which is often challenging (Suarez-Gonzalez et al., 2018; Taylor and Larson, 2019). Despite these difficulties, adaptive introgression has been well documented in some plant species such as *Senecio* (Kim et al., 2008), *Helianthus* (Kim and Rieseberg, 1999; Whitney et al., 2006; Whitney et al., 2010), *Iris* (Martin et al., 2006), *Phaseolus* (Rendón-Anaya et al., 2017), *Arabidopsis* (Arnold et al., 2016), or *Populus* (Suarez-Gonzalez et al., 2016; Suarez-Gonzalez, 2017). Nonetheless, given that natural hybridization is a common and widespread phenomenon, many other systems showing adaptive introgression may remain to be identified and elucidated.

Erysimum L. is one of the largest genera of the Brassicaceae, comprising more than 200 species (Polatschek, 1986), and has been described as a taxonomically complex genus in which molecular evidence indicated a reticulated evolutionary history, with polyploidization events in some clades (Al-Shehbaz & Al-Shammary, 1987; Ancev, 2006; Marhold & Lihová, 2006; Turner, 2006; Abdelaziz et al., 2011; Abdelaziz, 2013; Pajares, 2013; Abdelaziz et al., 2014). This genus is distributed mainly in Eurasia, with some species in North America and North Africa (Warwick et al., 2006). Notably, more than a hundred species have been described in the Mediterranean region (Greuter et al., 1986), being particularly abundant in the Iberian Peninsula, with twenty-one (Polatschek, 1979; Polatschek, 2014), or twenty-three (Nieto-Feliner, 1993; Mateo et al., 1998) different species described. Most of the Iberian *Erysimum* species are yellow, but six

species have purple corollas (Nieto-Feliner, 1993). Interestingly, previous phylogenetic studies suggested that these purple corolla species have appeared in different events in the Iberian lineages, suggesting the evolutionary convergence in this trait (Gómez et al., 2015b). However, the process underlying the origin of the color in these species is still unknown. Furthermore, previous studies suggested that some Iberian purple species may have a hybrid origin (Nieto-Feliner, 1993; Abdelaziz et al., 2014). Accordingly, there is the possibility that purple color has been transferred in the Iberian clade from a purple parental species by hybridization and has been maintained by natural selection. However, it is also unclear the adaptive advantages that this color may confer. The purple color is produced by anthocyanin pigments, and have a crucial role as a signal for pollinators (Vaidya et al., 2018). Also may confer fitness advantage under some abiotic and biotic factors, such as temperature, drought stress, exposure to ultraviolet radiation, and resistance to herbivory (ChalkerScott, 1999; Schemske & Bierzychudek, 2001; Warren and Mackenzie, 2001; Coberly & Rausher, 2003; Irwin et al., 2003; Strauss et al., 2004; Truetter, 2006; Dick et al., 2011; Arista et al., 2013).

Here we studied the hybridization scenario for six species of *Erysimum* that inhabit the Baetic Mountains (one of the principal glacial refugia in Europe in the south the Iberian Peninsula; Médail and Diadema, 2009). These species have yellow (*E. mediohispanicum*, *E. nevadense*, *E. fitzii*) or purple (*E. popovii*, *E. baeticum*, *E. bastetanum*) corollas. Moreover, these *Erysimum* species show characteristics that may facilitate hybridization, such as growing in sympatry in some locations and having a generalist pollination system that may facilitate gene flow among different species. Therefore, the main goals of this study are to disentangle the general hybridization scenario for the *Erysimum* species complex studied here, investigate whether a signature of introgression is detectable in purple species and if so, test if existed a signature of introgression on the anthocyanin biosynthetic pathway genes and if these genes are maintained by natural selection.

Material and Methods

Plant samples

We have studied six species of the genus *Erysimum* collected in the Baetic Mountains, South of Spain (**Table 1**; **Figure 1 from the general Material and Methods**). Specifically, we sampled three different individuals (each from one different population) for *E. mediohispanicum*, *E. nevadense*, *E. popovii*, *E. bastetanum*, and *E. baeticum*, and one individual for *E. fitzii*. Some of these species appear in sympatry in some localities that were included in the sampling (**Table 1**). Additionally, we sampled *E. lagascae*, an allopatric species with purple corollas inhabiting Central Spain that has been posited as one potential parental species of the species studied here (Nieto-Feliner, 1993). In each species, we sampled pre-anthesis flower buds for transcriptomic analyses (five buds per individual) and plant leaves for flow cytometric analyses (ten leaves per individual).

Flow cytometry analyses

Flow cytometry was used to assess genome size and estimate DNA ploidy levels. The nuclei were isolated from fresh leaf tissues by chopping simultaneously with a razor blade 0.5 cm2 of leaf and 0.5 cm2 of an internal reference standard (Galbraith et al., 1983). As internal reference standard we used *Solanum lycopersicum* L. 'Stupické' with 2C = 1.96 pg or *Raphanus sativus* L. with 2C = 1.11 pg (Doležel et al., 1992). The nuclei extraction was made on a Petri dish containing 1 ml of WPB buffer (Loureiro et al., 2007). Then, the nuclear suspension was filtered using a 50 µm nylon mesh and DNA was stained with 50 µg ml-1 of propidium iodide (PI, Fluka, Buchs, Switzerland). Also, 50 µg ml-1 of RNAse (Fluka, Buchs, Switzerland) were added to degrade dsRNA. After 5 min incubation, the samples were analysed in a Partec CyFlow Space flow cytometer (532 nm green solid-state laser, operating at 30 mW; Partec GmbH., Görlitz, Germany).

Species	Population	Location	Elevation	Geographical coordinates	Flower color	Sympatry with
E. baeticum	Ebb07	Sierra Nevada, Almería, Spain	2128	37°05′46″N, 3°01′01″W	purple	
	Ebb10	Sierra Nevada, Almería, Spain	2140	37°05′32″N, 3°00′40″W	purple	En12
	Ebb12	Sierra Nevada, Almería, Spain	2264	37°05′51″N, 2°58′06″W	purple	
E. bastetanum	Ebt01	Sierra de Baza, Granada, Spain	1990	37°22′52″N, 2°51′49″W	purple	
	Ebt12	Sierra de María, Almería, Spain	1528	37°41′03″N, 2°10′51″W	purple	
	Ebt13	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W	purple	Em71
E. fitzii	Ef01	Sierra de la Pandera, Jaén, Spain	1804	37°37′56″N, 3°46′46″W	yellow	
E. lagascae	Ela07	Sierra de San Vicente, Toledo, Spain	516	44°05′49″N, 4°40′40″W	purple	
E. mediohispanicum	Em21	Sierra Nevada, Granada, Spain	1723	37°08′04″N, 3°25′43″W	yellow	
	Em39	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W	yellow	Ep20
	Em71	Sierra Jureña, Granada, Spain	1352	37°57′10″N, 2°29′24″W	yellow	Ebt13
E. nevadense	En05	Sierra Nevada, Granada, Spain	2074	37°06′35″N, 3°01′32″W	yellow	
	En10	Sierra Nevada, Granada, Spain	2321	37°06′37″N, 3°24′18″W	yellow	
	En12	Sierra Nevada, Granada, Spain	2255	37°05′37″N, 2°56′19″W	yellow	Ebb10
E. popovii	Ep16	Jabalcuz, Jaén, Spain	796	37°45′26″N, 3°51′02″W	purple	
	Ep20	Sierra de Huétor, Granada, Spain	1272	37°19′08″N, 3°33′11″W	purple	Em39
	Ep27	Llanos del Purche, Granada, Spain	1470	37°07′46″N, 3°28′48″W	purple	

Table 1. Population code, location, and details of sympatry status for all of the populations sampled.

Results were acquired using Partec FloMax software v2.4d (Partec GmbH, Münster, Germany) in the form of four graphics: histogram of fluorescence pulse integral in linear scale (FL); forward light scatter (FS) vs. side light scatter (SS), both in logarithmic (log) scale; FL vs. time; and FL vs. SS in log scale. To remove debris, the FL histogram was gated using a polygonal region defined in the FL vs. SS histogram. At least 5,000 particles were analyzed per sample. Only CV values of 2C peak of each sample below 5% were accepted, otherwise a new sample was prepared and analyzed until quality standards were achieved (Greilhuber et al., 2007). In a few cases, samples produced histograms of poorer quality even after repetition due to presence of cytosolic compounds. Thus, it was not possible to estimate ploidy level and/or genome size for some individuals (**Table 2**).

Genome size in mass units (2C in pg; sensu Greilhuber et al., 2005) was obtained using the formula: sample 2C nuclear DNA content (pg) = (sample G1 peak mean / reference standard G1 peak mean) * genome size of the referencedate introgression events standard. The ploidy levels were inferred for each sample based on chromosome counts and genome size estimates obtained in the species and genus.

Species	Population	DNA Ploidy level		Genome size (2C, pg)					
		2n	N	Mean	SD	CV	Min	Max	N
E. baeticum	Ebb07	8x	5	2.08	0.08	3.85	1.93	2.17	2
	Ebb10	8x	6	2.07	0.09	4.35	1.93	2.17	5
	Ebb12	8x	-	-	-	-	-	-	-
E. bastetanum	Ebt01	4x	4	1.06	0.06	5.66	0.97	1.10	4
	Ebt12	4x	2	1.06	0.12	11.32	0.97	1.15	2
	Ebt13	8x	64	1.96	0.06	3.06	1.87	2.17	60
E. fitzii	Ef01	2x	3	0.44	0.004	0.91	0.44	0.45	3
E. lagascae	Ela07	2x	10	0.46	0.02	4.35	0.44	0.50	10
E. mediohispanicum	Em21	2x	2	0.44	0.01	2.27	0.43	0.44	2
	Em39	2x	21	0.46	0.02	4.35	0.43	0.49	19
	Em71	4x	59	0.98	0.04	4.08	0.93	1.13	59
E. nevadense	En05	2x	-	-	-	-	-	-	-
	En10	2x	-	-	-	-	-	-	-
	En12	2x	3	0.45	0.03	6.67	0.42	0.47	3
E. popovii	Ep16	4x	3	0.98	0.02	2.041.86	0.95	1.00	3
	Ep20	10x	15	2.49	0.06	2.416	2.40	2.60	10
	Ep27	4x	39	0.96	0.04	4.17	0.92	1.05	9

Table 2. Genome size estimates and DNA ploidy levels obtained in populations of *Erysimum*. The following data are given for each population and ploidy level: mean, the standard deviation of the mean (SD), coefficient of variation (CV, %), minimum (Min) and maximum values (Max) of the holoploid genome size (2C, pg) followed by sample size for genome size estimates (N); DNA ploidy level (2n) and respective sample size (N) for ploidy estimates. DNA ploidy levels: 2x, diploid; 4x, tetraploid; 8x, octoploid; 10x, decaploid. For Ebb12, En05, and En10 samples were not possible to estimate the ploidy levels, and we have used the described in Blanca et al. (1992).

RNA extraction and sequencing

Details of the sampling and RNA extraction and sequencing were detailed in **Chapter three**. Here we used the transcriptomes from the samples represented in **Table 1**. The collected flower buds of each individual were stored in liquid nitrogen until RNA extraction. Floral buds were ground with a mortar and a pestle in liquid nitrogen. Total RNA was isolated using the Qiagen RNeasy Plant Mini Kit following the manufacturer's protocol. The quality and quantity of the

RNA were checked using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, United States). Library preparation and RNA sequencing were conducted at Macrogen Inc. (Seoul, Korea). Before sequencing, the quality of the RNA was analyzed with the Agilent 2100 Bioanalyzer system (Agilent Technologies Inc), and an rRNA-depletion protocol (Ribo-Zero) was used to perform an mRNA enrichment and to avoid sequencing rRNAs. Library preparation was performed using the TruSeq Stranded Total RNA LT Sample Preparation Kit (Plant). The sequencing of the 17 libraries (one per individual) was carried out using the Hiseq 3000-4000 sequencing protocol and TruSeq 3000-4000 SBS Kit v3 reagent, following a paired-end 150 bp strategy on the Illumina HiSeq 4000 platform. A summary of sequencing statistics is shown in **Table S1** (Supporting Information).

Data processing and transcriptome analyses

Reads quality control and trimming

We used FastQC v0.11.5 (Andrews, 2010) for analyzing the quality of the raw reads of each library. Then, we trimmed the adapters in the raw reads using cutadapt v1.15 (Martin, 2011). Specifically, we used the "-b" option for trimming the adapters in 5′ and 3′ direction, using the prefix of the adapter sequence that is common to all "TruSeq Indexed Adapter" sequences (AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC). In addition, we used the "-n" option to search repeatedly for the adapter sequences (28 iterations). This option ensures that the correct adapters are detected by searching in loops until any adapter match is found or until the specified number of rounds is reached (Martin, 2011). Then, we quality-filtered the reads using Sickle v1.33 (Joshi and Fass, 2011). This trimming software uses sliding-window analyses along with quality and length thresholds to cut and discard the reads which do not fit the selected threshold values. We specified the "pe" option for paired-end reads and the "-t" to use Illumina quality values (see https://github.com/najoshi/sickle). After trimming, we used FastQC v0.11.5 (Andrews, 2010) again to verify the trimming efficiency.

De novo transcriptome assembly

To assemble the resulting high-quality, cleaned reads into contigs, we followed a *de novo* approach using Trinity v 2.8.4 (Grabherr et al. 2011). Each library was normalized in silico before assembly to validate and reduce the number of reads using the "insilico_read_normalization.pl" function in Trinity (Haas et al., 2013). Then we used the parameter 'min_kmer_cov 2' to eliminate singly occurring k-mers that are heavily enriched in sequencing errors, following the approach of Haas et al. (2013). Thus, only k-mers that occur more than once were considered for contigs.

Transcriptome annotation

Candidate open reading frames (ORF) within transcript sequences were predicted and translated using TransDecoder v5.2.0 (Haas et al. 2013). We performed functional annotation of Trinity transcripts with ORFs using Trinotate v3.0.1 (Haas, 2015), an annotation suite designed for automatic functional annotation of *de novo* assembled transcriptomes. Sequences were searched against UniProt (UniProt Consortium, 2014), using SwissProt databases (Bairoch and Apweiler, 2000) (with BLASTX and BLASTP searching and an e-value cutoff of 10^{-5}). We then used Pfam database (Bateman et al., 2004) to annotate protein domains for each predicted protein sequence. Transcripts were also searched through the eggnog (Jensen et al., 2007), GO (Gene Ontology Consortium, 2004), and Kegg (Kanehisa and Goto, 2000) annotation databases.

Orthology inference

To reduce redundancy, we clustered the translated sequences using cd-hit v4.6 (Li and Godzik, 2006) following the steps of the pipeline described in Yang and Smith (2014). For the orthologs inference, we excluded UTRs and non-coding transcripts, using only the CDS, avoiding the possibility of including sequencing errors (Yang and Smith, 2014). We identified ortholog genes using the OrthoFinder v2.3.3 pipeline (Emms and Kelly, 2015). In brief, this pipeline first made a

BLASTp analysis with the protein sequences as input for searching the orthogroups (a set of protein genes that are descended from a single gene in the last common ancestor of all the species sampled), then clustered and aligns the orthologous sequences using MAFFT v7.450 (Katoh and Standley, 2013) with default parameters. Finally, we obtained the maximum-likelihood phylogenetic gene trees for all orthogroups using IQ-Tree v1.6.1 (Nguyen et al., 2014). Then, each orthogroup with all species present was used to infer a species tree using STAG v1.0.0 (Emms and Kelly, 2019). Then, we used DLCpar v1.1 (Wu et al., 2014) to reconcile the species tree with the gene trees obtained, considering gene duplication, losses, and incomplete lineage sorting (ILS) as potential causes of discordance among the trees.

Phylogenetic reconstruction

We obtained a coalescent species tree using ASTRAL v 5.6.3 (Mirarab et al., 2014), with default parameters. This method reconstructs species trees from unrooted gene tree topologies. We used the gene trees previously obtained by maximum likelihood by IQ-Tree v1.6.1 as input. We used FigTree v. 1.4.0 (Rambaut and Drummond, 2012) to visualize and edit the species tree. Then, we compared the alternative tree topologies from the obtained species tree with a whole chloroplast genomes phylogeny for the same species samples from **Chapter III**. For that, we used the Shimodaira-Hasegawa Test (SH-Test) (Shimodaira and Hasegawa, 1999) from the R package phangorn v2.5.5 (Schliep, 2011). Both phylogenies were also compared visually, plotting them as mirror images with the function cophyloplot, using the R package ape v5.4 (Paradis et al. 2004).

Population structure analyses

Variant calling

To study the population structure, we first run a variant calling analysis, using the *E. lagascae* transcriptome as reference. We indexed the *E. lagascae* transcriptome using BWA v0.7.17 (Li and Durbin, 2009) to create a reference, and then we mapped all the trimmed raw reads against it

using the BWA v0.7.17 “mem” option. We used SAMtools v1.7 (Li et al., 2009) to convert and sort the alignment files. Then, we identified the SNP positions within the aligned reads comparing with the reference transcriptome using the SAMtools v1.7 “mpileup” command, which transposes the mapped data into a sorted BAM file. Lastly, we used bcftools v 1.9 to filter the SNPs (Narasimhan et al., 2016), running the SAMtools v1.7 Perl script “vcfutils.pl VarFilter” with default parameters to filter down the candidate variants and to eliminate false positives.

Discriminant Analysis of Principal Components (DAPC)

To investigate population structure, we conducted a Discriminant Analysis of Principal Components (DAPC; Jombart et al., 2010) for the SNP data, using the R package adegenet v2.1.3 (Jombart and Ahmed, 2011). DAPC is a multivariate method that identifies and describes clusters of genetically related individuals from large datasets. It provides a measure of the optimal number of genetic clusters (K) across a range of K values by using the Bayesian Information Criterion (BIC). We set a range of K values from two to seven because K=7 is the number of different species in our dataset. To identify the optimal number of K, we selected the model with the lowest BIC.

Introgression analyses

Phylogenetic inference of introgression

We inferred phylogenetic species networks to evaluate introgression events. This approach extends the phylogenetic tree model, representing the gene flow by edges connecting the samples with signatures of introgression in the tree. We used the software PhyloNet v 3.6.9 (Than et al., 2008; Wen et al., 2018), which implements a phylogenetic network method based on the frequencies of rooted trees accounting for incomplete lineage sorting (ILS). To generate the input for PhyloNet, we first ultrametricize the trees obtained previously with IQ-Tree v1.6.1, using the "nnls" method in the "force.ultrametric" function within the R package phytools v0.6-99 (Revell,

2012). Due to computational limitations, we inferred the species networks using a maximum pseudo-likelihood method (MPL) (Yu and Nakhleh, 2015). We performed the search five times, to avoid getting stuck at local optima. We estimated optimal networks among a range from 0 to 15 introgression events, calculating all the networks' best log-likelihood by computing the Akaike's Information Criterion (AIC) (Bozdogan, 1987) with the generic function for AIC in R package stats v3.6.1. As AIC may not provide precise values when using pseudo-likelihood phylogenetic networks (Cao et al., 2019), we also estimated the more optimal network by slope heuristic of log-likelihood values. The optimal network was then visualized with Dendroscope v3.5.10 (Huson and Scornavacca, 2019).

ABBA-BABA statistic

To assess gene flow between species, we calculated the D-statistics, also known as the ABBA-BABA statistics (Durand et al., 2011). In order to evaluate introgression among the seven species, we used the software Dsuite v0.1 (Malinsky, 2019) that allows the assessment of gene flow across large datasets and directly from a variant call format (VCF) file. This algorithm computes the D statistic by considering multiple groups of four populations: P1, P2, P3, and O, related by the asymmetric tree (((P1, P2), P3), O). The site patterns are ordered such that the pattern BBAA refers to P1 and P2 sharing the derived allele (B-derived allele, A-ancestral allele), ABBA to P2 and P3 sharing the derived allele, and BABA to P1 and P3 sharing the derived allele. The ABBA and BABA patterns are expected to occur with equal frequencies, assuming no gene flow (null hypothesis), while a significant deviation from that suggests possible introgression. To assess whether D is significantly different from zero, D-suite uses a standard block-jackknife procedure (as in Green et al., 2010, and Durand et al., 2011), obtaining approximately normally-distributed standard errors. As recommended by Malinsky (2019), we used a conservative approach estimating the statistic Dmin, which gives the lowest D-statistic value in a given trio. We used the ruby script “plot_d.rb” to plot into a heatmap the introgression among all the pairs of samples. To

complement these analyses, we computed the Fbranch statistic implemented in Dsuite v0.1 (Mallinsky et al., 2018, Mallinsky et al., 2019). This statistic allows identifying gene flow events into specific internal branches of a phylogeny. Thus, evaluating the excess sharing of alleles between one species and the descendant or ancestral species, helping to understand when the gene flow happened. We used the whole chloroplast genomes phylogeny from **Chapter III** in Newick format specifying which species should be treated as sister species (i.e., as P1 and P2) and *E. lagascae* as outgroup.

Detecting introgression in the anthocyanin biosynthetic pathway genes

We looked for evidence of introgression in the genes coding for enzymes of the anthocyanin biosynthetic pathway (ABP). To obtain complete gene sequences, we used a read-mapping approach. First, we downloaded the *Arabidopsis thaliana* ABP genes from TAIR (Lamesch et al., 2011): CHI (AT3G55120), CHS (AT5G13930), F3H (AT3G51240), DFR (AT5G42800), ANS (AT4G22880), and UF3GT (AT5G54060). Then, we mapped the trimmed raw reads from all samples to the *A. thaliana* genes as a reference using BWA v0.7.17 (Li and Durbin, 2009). We imported the mapping reads to Geneious R.11 (Kearse et al., 2012), and assembled them *de novo* using the Geneious assembler with the highest sensitivity. To eliminate poorly mapped sequences, we clustered the assembled reads using cd-hit v4.6 (with parameter -c = 0.99) (Li and Godzik, 2006) and aligned the cluster reads for each sample using MAFFT v7.450 with default parameters (Katoh and Standley, 2013). We estimated the consensus sequence for each sample in Geneious R.11 using the strict threshold option (bases matching at least 50 % of the sequences). We confirmed the annotations using BLAST (Johnson et al., 2008). Finally, we double-checked manually that the candidate genes appeared in the transcriptome annotations of all the species studied here.

f^d statistics in ABP genes

We estimated the f^d statistics and Patterson's D in the ABP genes using the R package HybridCheck v1.0 (Ward and Oosterhout, 2016). The f^d statistic is not prone to false positives (Pfeifer and Kapan, 2019) and is suitable for small genomic regions, estimating the fraction of the genome shared through introgression by comparing the observed difference in the number of ABBA and BABA patterns between P2 and P3, and P1 and P3 (Martin et al., 2014).We used a block-jackknifing correction to overcome the problem of non-independence between loci (Green et al., 2010). We considered *E. lagascae* as the basal population (P3), thus acting as the possible introgression donor. We then estimated the f^d for each purple species as P2 and all possible combinations of species as P1.

Signatures of selection

To estimate if the genes with signatures of introgression were under positive selection, we estimated the pairwise ratios of synonymous and non-synonymous sites (dN/dS), using the "dnds" function from the R package ape v5.4 (Paradis et al., 2004). Because synonymous sites are presumed to be neutral while non-synonymous sites might be under selection, dN/dS =1 is expected under neutrality, dN/dS >1 is expected when natural selection promotes changes in the protein sequence (positive selection) and dN/dS < 1 is expected when natural selection suppresses protein changes (purifying selection) (Kryazhimskiy and Plotkin, 2008)

We also performed the MacDonald-Kreitman test (MKT; Egea et al., 2008). This test examines the neutral prediction of equal ratios of non-synonymous to synonymous variation for polymorphic sites and fixed differences between species by estimating the Neutrality Index (NI). Under neutrality, we expected that NI = 1, if NI < 1, there is an excess fixation of non-neutral replacements due to positive selection, and if NI > 1, negative selection is preventing the fixation of harmful mutations (Egea et al., 2008). We used the *A. thaliana* sequence of each gene as a

putatively neutral outgroup (white flowers). We computed the standard MKT with Jukes Cantor's correction using the MKT website (http://mkt.uab.es/mkt/help_mkt.asp; Egea et al., 2008).

Pollen tube growth

To explore the existence of prezygotic barriers, we carried out a preliminary experiment on the growth of pollen tubes on a reduced set of species. We collected 20 individuals plants of each *E. mediohispanicum, E. bastetanum*, and *E. popovii* species from natural populations. We set the plants into a common garden (University of Granada facilities), and before blooming, we moved them to a greenhouse excluded from pollinators. When the flowers were open, we performed hand-pollination experiments by tipping the anther with a small stick removing the pollen and placing it on the stigma of a flower from different species previously emasculated (hybrid crosses), or of a flower from the same species but different populations previously emasculated (intra-specific crosses). Moreover, we emasculated some flowers and hand-pollinated them with their pollen (forced selfing crosses), and some flowers were not manipulated and left for spontaneous self-pollination (spontaneous selfing crosses).

We collected the *Erysimum* pistils after 72 hours and conserved them in ethanol in the fridge at 4C until pollen tube staining. Them, we followed the Mori et al. (2006) protocol with small modifications. In brief, each pistil was cleaned in 70% EtOH for ten minutes, and then we moved them to 50, 30% EtOH, and finally distilled water. We softened the samples, moving them into a small petri dish of 8 M NaOH for one hour at room temperature (as recommended in Kearns and Inouye, 1993). Then, we changed the pistils to distilled water for ten minutes, and afterward, the stigmas were incubated with 0.1 % aniline blue in phosphate buffer (pH 8.3) for two hours. The final slide preparations were examined under a fluorescence microscope with blue light (410 nm) to observe the pollen tube formation.

Results

All the sampled yellow corollas species were diploid, except the Em71 population of *E. mediohispanicum* (4x). However, the purple corollas species except *E. lasgacae* showed ploidy levels higher than 2X (**Table 2)**. Populations within species showed different ploidy levels for *E. bastetanum* (4x and 8x), and *E. popovii* (4x and 10x). For Ebb12, En05, and En10 samples were not possible to estimate the ploidy levels, and we have used the described in Blanca et al. (1992) (**Table 2**).

Transcriptome assembly and orthology inference

The summary of the sequenced statistic is shown in the **Table S1** (Suporting Information). After assembling, we obtained between 104K and 382K different Trinity transcripts, producing between 66K and 235K Trinity isogenes. The total assembled bases ranged from 92 Mbp (in Em21 population of *E. mediohispanicum*) to 319 Mbp (in En10 population of *E. nevadense*). The summary statistics of the assembled transcriptomes is presented in **Supplementary Table S2**. We annotated the Trinity unigenes using different protein databases (see **Table S3**). Among the annotated unigenes, the highest proportion was annotated using BLASTX search against the SwissProt reference database, and the number of genes annotated ranges between 71,606 (*E. nevadense*, En12) and 197,069 (*E. baeticum*, Ebb10); mean value 146,314.35. OrthoFinder assigned 1,519,064 protein gene sequences (96.4% of total) to 92,984 gene families (orthogroups) (**Table S4**). Among them, 16,941 orthogroups were shared by all species, and their corresponding gene trees were used for further analyses.

Phylogenetic trees and population clustering

We inferred a coalescence tree using the 16,941 maximum likelihood gene trees obtained with IQTREE as input for ASTRAL (**Figure S1**). This species tree was almost entirely resolved,

having only four nodes with low quartet scores results (BS for these nodes: 0.78, 0.77, 0.70, and 0.53; see **Figure S1**). The earliest diverging sample was Em71, the 4x *E. mediohispanicum* population. Three clades, although with low supports, were evident. A clade composed of the individuals of *E. bastetanum* and *E. baeticum* (purple corollas); another clade including *E. fitzii* and the three individuals of *E. popovii*; and the last clade including the individuals of *E. nevadense* and the 2x individuals of *E. mediohispanicum* (yellow corollas). We did not find any evidence of within-species clustering, which is compatible with interspecific hybridization. Moreover, when comparing the species tree with the whole chloroplast genomes phylogeny, we have found a signature of cyto-nuclear discordance (**Figure 1**) with significant SH test result (p-value < 0.01, Diff -ln L= 345426.4). This lack of congruence among both phylogenies also supports the hybridization hypothesis.

The discriminant analysis revealed K=4 and K=5 as the most likely number of genetic clusters (**Figure S2**), both of them with very similar BIC values (K=4, BIC= 189.99; K=5, BIC= 188.99). The clusters corresponding to K=4 produced the same clusters that appeared in the coalescence tree (**Figure S2**). However, the clusters corresponding to K=5, included three monospecific ones (for *E. lagascae*, *E. fitzii* and *E. popovii*), one for the yellow diploid species (*E. nevadense*, the three individuals, and *E. mediohispanicum*, Em21 and Em39), and the last one including the individuals of *E. baeticum*, *E. bastetanum*, *E. popovii* and the *E. mediohispanicum* 4X (Em71).

Analysis of introgression

Based on the AIC values for the log-likelihood of the networks (**Table S5**), the network with 13 reticulation instances was the most reliable, indicating frequent hybridization events in the genealogy of these populations. The estimates of slope heuristic of log-likelihood values also supported the network with 13 reticulation instances as the most reliable network estimated. **Figure 3** shows the probable introgression events by edges connecting the tree branches between

different individuals. Moreover, the most likely network included edges connecting non-terminal branches (see **Figure 3**), which indicates reticulations with past extinct taxa or incomplete sampled taxa (“ghost species”).

This scenario of frequent hybridization is also supported by the ABBA-BABA statistics, even using the D-statistic conservative approach of D-min. We summarized the tested topologies and the inferred D-statistics with corrected p-values for all triplet combinations in the supplementary **Table S6** and in the **Figure S3**. The highest signal of introgression occurred between *E. fitzii* and individuals of *E. baeticum* (Ebb12, and Ebb10) and *E. popovii* (Ep16), and between *E. popovii* (Ep16) and individuals of *E. bastetanum* (Ebt12) and *E. baeticum* (Ebb07, Ebb10, Ebb12). The Fbranches results also supported several events of inter-specific gene flow **(Figure S4).** Specifically, we found the highest signal of gene flow between *E. bastetanum* (Ebt12) and individuals of *E. mediohispanicum* (Em71, Em21), *E. nevadense* (En12, En05), and *E. baeticum* (Ebb07, Ebb10), between *E. bastetanum* (Ebt13) and individuals of *E. mediohispanicum* (Em71), and *E. popovii* (Ep20), between *E. bastetanum* (Ebt01) and individuals of *E. mediohispanicum* (Em21, Em39) and *E. nevadense* (En12). Moreover, we have found many gene flow events provided by ancestral or non-sampled taxa (**Figure S4)**.

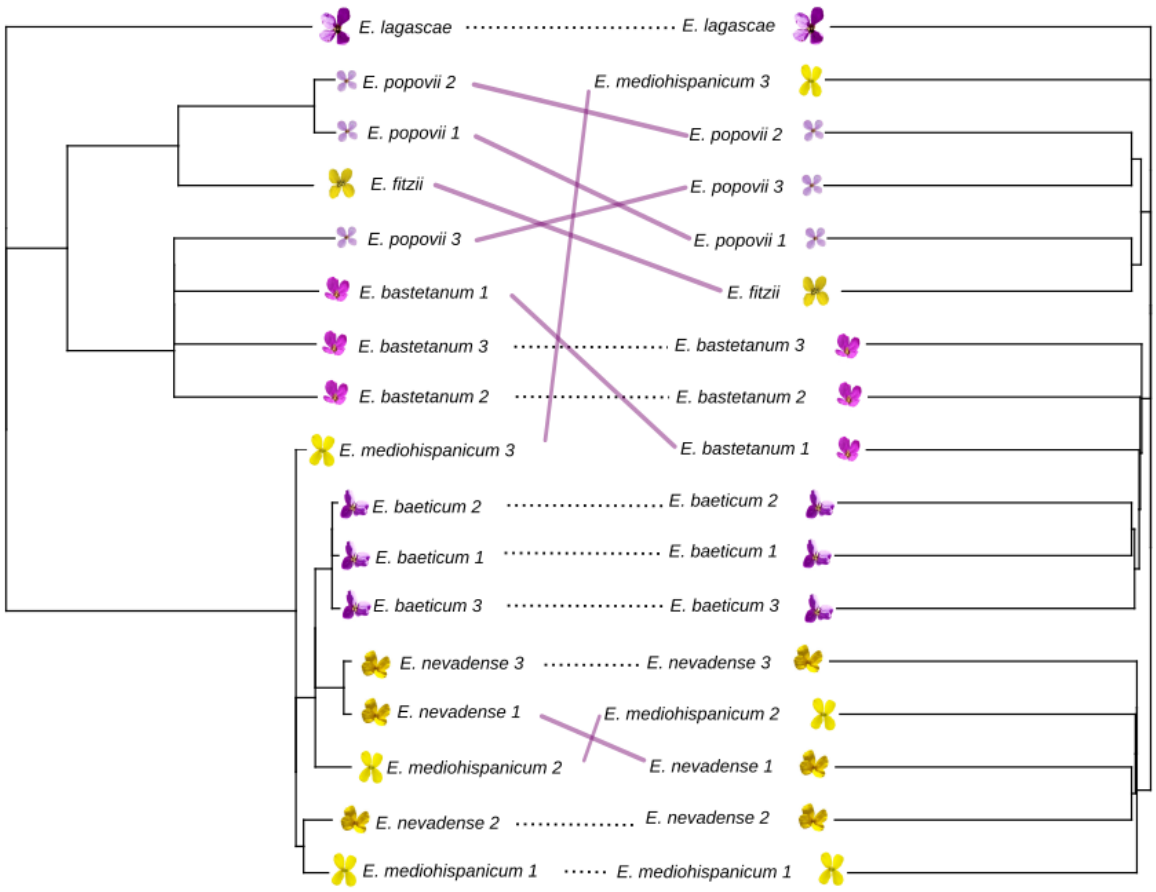

Figure 1. Cyto-nuclear discordance among the nuclear species tree (on the right) and the chloroplast phylogeny from Chapter III (on the left).

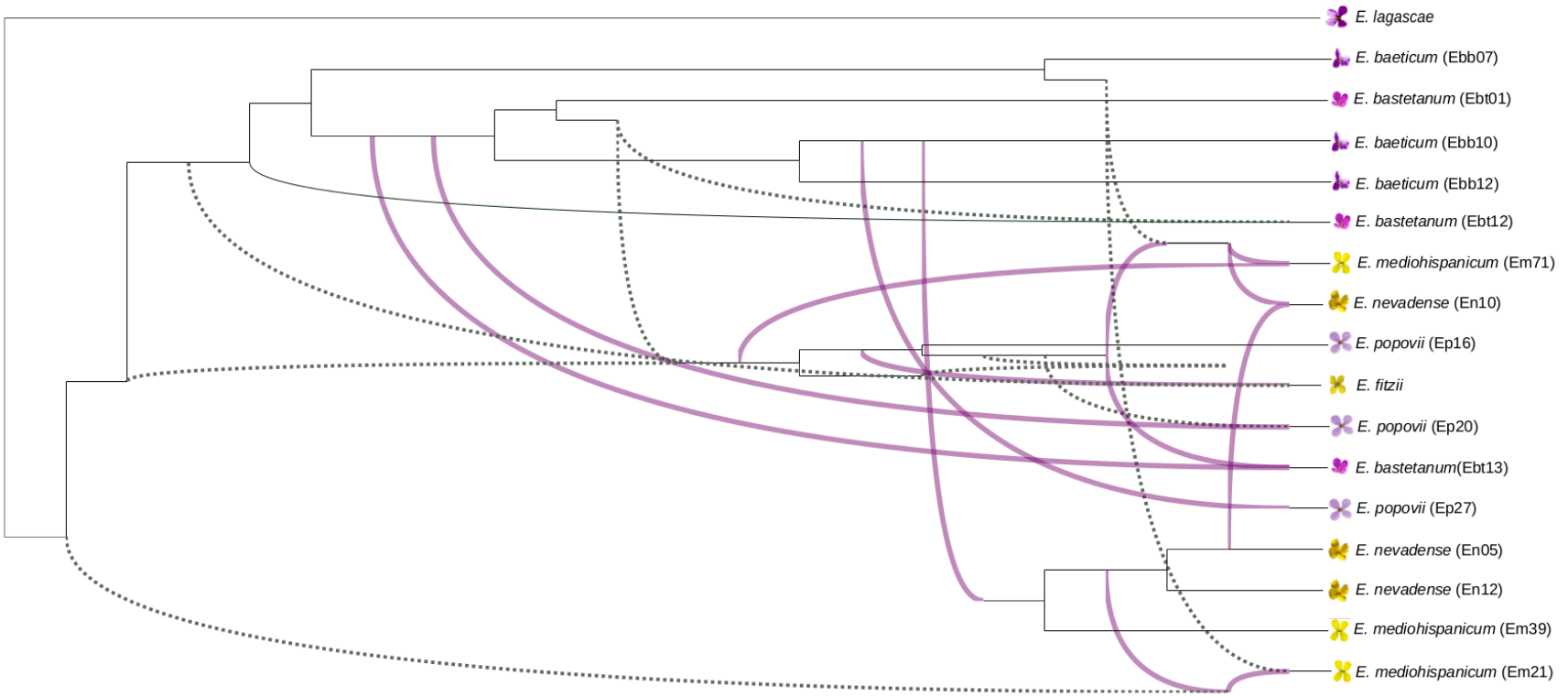

Figure 2. The optimal species network inferred using the PhyloNet software. The result is maximum pseudo-likelihood (MPL) tree with 13 reticulations. High-confidence tree structure is represented in black and introgression events are represented in purple. The discontinuous lines represent the "ghost species" (i.e., reticulations between past extinct taxa or incomplete sampled taxa). The code of the populations appears in parenthesis.

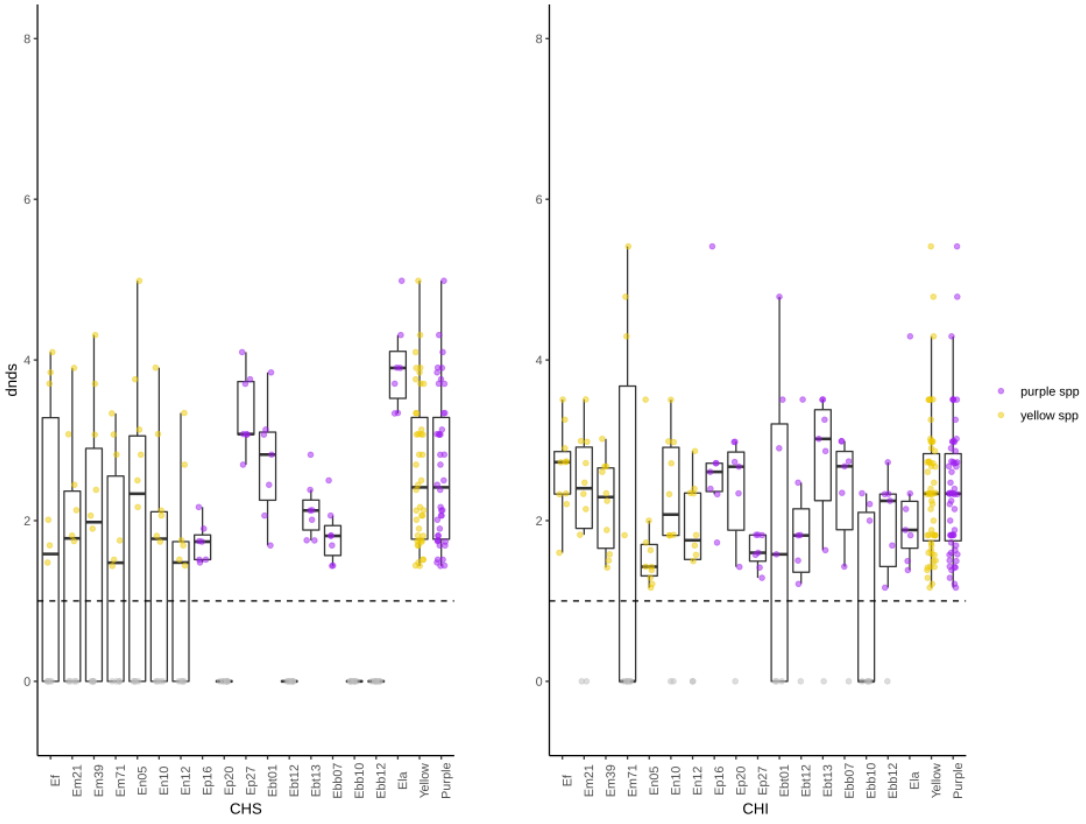

Figure 3. Boxplots depicting dN/dS ratios for CHS and CHI genes. On the right of each x-axis is shown the average of dN/dS ratio for yellow and purple corolla samples. The dashed line represents unity. Thus, dN/dS < 1 means purifying selection, dN/dS > 1 means positive selection and dN/dS = 1 neutrality. “Na” values represented cases where we do not find synonymous mutations, and the dN/dS ratio could not be calculated.

Evolution of ABP genes

We have found introgression signatures of *E. lagascae* into *E. bastetanum* for CHS, CHI, with significant fˆd values (**Tables S7 and S8**, respectively). The most robust signature of introgression was for the CHI gene, in which we have found signatures of introgression in *E. bastetanum* when using all the possibles P1 combinations. Moreover, we have also found signatures of the introgression of *E. lagascae* for the CHI gene into the yellow species *E. mediohispanicum*, *E. nevadense*, and *E. fitzii* with significant fˆd values. Patterson's D values supported the gene flow in all the cases. The F3H gene showed a similar introgression pattern than the CHI gene for *E. bastetanum*, *E. mediohispanicum*, *E. nevadense*, and *E. fitzii*, but with not significant fˆd values (**Table S9**). However, we have not found introgression in the other genes studied (DFR, ANS, and UF3GT; **Tables S10-S12**)

The dN/dS results showed that CHS and CHI genes were under positive selection in almost all the species studied (**Figure 4**). However, in some instances, we have not found synonymous mutations for the CHI gene, and the dN/dS ratio could not be calculated for some populations (*E. popovii*-Ep20, *E. bastetanum*-Ebt12, *E. baeticum*-Ebb10, and Ebb12). Moreover, for the CHS gene, the Em71 population of *E. mediohispanicum* and Ebb10 of *E. baeticum*, seems to be under purifying selection. Overall, the McDonald–Kreitman test (MKT) results showed NI values under one, supporting positive selection for both genes in *E. bastetanum*, *E. popovii*, *E. mediohispanicum*, and *E. nevadense* (only in CHS gene). However, none of the p values were significant (**Table 3**). We could not estimate the MKT in *E. fitzii* as we have only one population sampled, and the test requires more than a sequence to be estimated.

Species	CHI	CHS
E. baeticum	NI: null	NI: 0,00
	p-value: 0.78	p-value: 0.17
E. bastetanum	NI: 0.51	NI: 0.18
	p-value: 0.54	p-value: 0.10
E. popovii	NI: 0.32	NI: 0.81
	p-value: 0.16	p-value: 0.86
E. mediohispanicum	NI: 0.21	NI: 1.72
	p-value: 0.16	p-value: 0.63
E. nevadense	NI: 0.15	NI: null
	p-value: 0.09	p-value: 0.16
E. fitzii	NI: null	NI: null
	p-value: null	p-value: null

Table 3. Results of the MacDonald-Kreitman test (MKT). NI corresponds to the Neutrality Index. Under neutrality, we expected that NI = 1, if NI < 1, there is an excess of fixation of non-neutral replacements due to positive selection, and if NI > 1, negative selection is preventing the fixation of harmful mutations. The divergence was corrected by the Jukes & Cantor model. Gene codes: CHS: Chalcone synthase, CHI: Chalcone flavone isomerase.

Prezygotic barriers

Pollen tube growth

A total of 103 *Erysimum* pistils preparations were examined: 52 from hybrid crosses, 24 from forced selfing crosses, 16 from spontaneous selfing crosses, and 11 from intra-specific crosses. Our results showed that pollen tube full growth (until the ovary) in 51,92 % of the hybrid crosses (n=52), 29,16 % of the forced selfing crosses (n=24), 25,16 % of the spontaneous selfing crosses (n= 16), and 63,33 % of the intra-specific crosses (n= 11) (**Figure S5**). Cases in which the pollen tube grew but did not reach the ovary were treated as non-growing, and we could not estimate whether the tube grew to that point or did not have enough time to develop completely.

Discussion

Our results suggest that the *Erysimum* species studied here have a strong signature of hybridization and introgression. Thus, by using several methodological approaches that allow studying hybridization while accounting for incomplete lineage sorting (e.g., PhyloNet (Than et al., 2008; Wen et al., 2018), and ABBA-BABA (Durand et al., 2011)), we have found a significant incidence of hybridization events, thus suggesting that reproductive barriers among the species studied here were weak (Abbott et al., 2013). In fact, this is supported by the pollen tube growth experiments where pollen tube was able to grow until the ovary in hybrid crosses. Moreover, we have found that purple species studied here were polyploid and may suggest an allopolyploid origin. Nevertheless, we also have found a signature of hybridization on the yellow species, that were mostly diploid (except one population of *E. mediohispanicum*), suggesting that yellow species could have signatures of past hybridization events or an ancient polyploid origin, i.e., being the result of a diploidization event. Furthermore, our results suggest that one purple species (*E. bastetanum*) of the three studied here have signatures of introgression from a purple ancestral (*E. lagascae*) in two anthocyanin genes. However, we have found that the three yellow species studied (*E. mediohispanicum*, *E. nevadense*, and *E. fitzii*) also have a signature of introgression from a purple parental in an anthocyanin gene. Overall, all results support a scenario of multiple hybridization events involving not only the purple species but also the yellow ones.

However, we have not found a consistent hybridization pattern for these species. In particular, we have found that individuals from a given species might hybridize with multiple different species. Both ABBA-BABA test (with many introgression events found) and PhyloNet (with 13 reticulations as the most optimal network) support these promiscuous hybridization patterns. In the same vein, DAPC results supported a scenario with no species identity clustering

in which the species were grouped by geography. In other words, populations from different species, placed in the same geographical area (in the same mountain range in this case), were grouped independently of their species identity. These results are in line with the previous findings in other systems, which suggested differential rates of hybridization for the same species across different geographical areas (Payton et al., 2019; Sujii et al., 2019; Wang et al., 2019). These hybridization patterns may have varied due to ecological pressures (Grant & Grant 2002, Borge et al. 2005, Seehausen et al. 2008; Ortego et al., 2014; Harrison and Larson., 2014; Kay et al., 2018). Nevertheless, the historical dynamics of genetic isolation and gene flow might have also played a role (Albaladejo and Aparicio, 2007; Rifkin et al., 2019; Zielinski et al., 2019). In fact, these species studied were located in a well known glacial refugium (Médail and Diadema, 2009; Hughes and Woodward, 2017), and thus, the isolation and then re-establishment of gene flow (i.e., secondary contact zones) among populations of different species may have favored locally specific hybridization patterns (Coyne, 2004; Harrison and Larson, 2014; Arnold, 2015). Therefore, in line with our results, a knowledge of the historical dynamics of species populations and past ranges overlap is required to fully understand the genomic pattern of divergence between closely related species, for instance by using macroecological methods combining niche models with a phylogenetic approach (Folk et al., 2018).

Furthermore, we have found a ghost introgression signature. Thus, ancestral species may have influenced the hybridization scenario of the species studied here. This result was first supported by the cytonuclear discordance patterns that we have found which might be due to organellar introgression from extinct species (Huang et al., 2014; Pereira et al., 2016; Folk et al., 2017; Forsythe et al., 2018; Seixas et al., 2018; Lee-Yaw et al., 2019; Lin et al., 2019; Rakotoarivelo et al., 2019; Zhang et al., 2019; Dufresnes et a., 2020). Furthermore, we have found an ancestral introgression signature in the phylogenetic species network, in which some of the reticulations that we have found were coming from ghost introgression events. Also,

Fbranches results supported a very similar scenario. Specifically, we have found that some ancestral of *E. popovii* (purple) could be related to *E. fitzii* (yellow). Also, we have found evidence of gene flow between an ancestral of *E. mediohispanicum* (yellow; Em21) with *E. bastetanum* (purple) and *E. baeticum* (purple). Moreover, the results showed that many past gene flow events could have occurred between *E. baeticum* (purple), *E. nevadense* (yellow), and *E. bastetanum* (purple). In light of our results, some ancestral species may have played a role as introgression sources for purple and yellow species studied here. However, in this study, we did not include some *Erysimum* species that also inhabit the Baetic Mountains and may have acted as a source of introgression. Accordingly, we may be mistaking the signal of the unsampled species for ancestral species. Nevertheless, the number of studies reporting ghost introgression events are increasing (Sušnik et al., 2007; Green et al., 2010; Meyer et al., 2012; Ai et al., 2015; Barlow et al., 2018; Gokhman et al., 2019; Kuhlwilm et al., 2019; Zhang et al., 2019; Liu et al., 2019;) and novel methodological approaches are appearing, which allow the search into the extant genome for signatures of introgression from extinct species (e.g., S* statistic (Racimo et al., 2015; De Manuel et al., 2016), hidden Markov models (Skov et al., 2018; Foote et al., 2019) and demographic models (Gronau et al., 2011). Therefore, further research about the ghost introgression's influence on *Erysimum* evolution, including all the Iberian species and high-quality genome assemblies, would be required to understand the hybridization scenario thoroughly.

Here, in addition to studying the general signature of hybridization on several *Erysimum* species, we have studied the introgression and selection patterns of the anthocyanin genes' evolution. Our results are partially compatible with the possibility that that purple color of some *Erysimum* species might have appeared as a consequence of an introgression event by hybridization from a purple parental species. Thus, our results supported signatures of introgression (f^d) for some species in the CHS and CHI genes, even though this statistic tends to

underestimate the rate of introgression (Pfeifer and Kapan, 2019). However, we have not found introgression in the other genes studied (F3H, DFR, ANS, and UF3GT), maybe because these downstream anthocyanin genes may have lost the introgression signal (if it would exist) as they have an accelerated rate of evolution compared to genes upstream (Rausher et al., 1999; Lu and Rausher, 2003). Specifically, considering *E. lagascae* (purple) as the introgression source, we have found introgression signatures in *E. bastetanum* (purple) for these two genes. Nevertheless, we have also found introgression signatures for the CHI gene in the yellow species (*E. mediohispanicum*, *E. nevadense*, *E. fitzii*). These results may suggest that yellow species studied have signatures of introgression from a purple diploid parental (*E. lagascae*), and may have appeared by hybridization from ancestral species with purple and yellow color. These results were congruent with previous studies that suggested that yellow color appears in the Iberian species secondarily, from a purple ancestral species (Gómez et al., 2015b). Moreover, we have not found signatures in *E. baeticum* and *E. popovii* from *E. lagascae*. Therefore, in light of our results, which suggested that these two species have hybridization signatures, these purple species may have a different origin, possibly resulting from different hybridization events with another purple species as a parental not sampled or extinct. Further studies using new methodologies that allow the reconstruction of ancestral characters (Gokhman et al., 2019) would be required for unraveling the ancestral corollas color for these species (Ottenburghs, 2020). Moreover, the results of the selection test showed signatures of positive selection in the CHS and CHI gene for yellow and purple species according to both the dN/dS and MKT tests, but this last with no significant results. However, in closely related species, differences between sequences may not represent fixation events. Rather, there is a segregation of pre-existing polymorphism among descendant populations (Kryazhimskiy and Plotkin, 2008; Keightley and Eyre-Walker, 2012) and here selection tests may be biased by relatedness among *Erysimum* species. Overall, our approach has some limitations, and we may have lost variant information for these species as using a consensus sequence of each gene. Therefore, in further studies, it would be ideal to use a

haplotype phasing approach with long reads to capture all the allelic variants from each chromosome in the diploid and polyploid species (Zhang et al., 2019, Schrinner et al., 2020).

The introgression that we have found on yellow and purple species may have an adaptive function. The fact that yellow species have signatures of introgression from purple species may suggest that anthocyanin genes introgressed may not have played a role as pollinator attractors. Interestingly, Gómez et al. 2015b found that purple *Erysimum* species were not more specialized in pollinator's attraction than yellow species. Thus, it seems that pollinators did not play an essential role in the appears purple on the *Erysimum* species studied here, and anthocyanin production may be related to other functions. In particular, we have found introgression of CHI and CHS genes, which participated in flavonoid biosynthesis (Winkel-Shirley, 2001), and were related to responses to environmental plant stress (Treutter, 2006; Dao et al., 2011; Khlestkina, 2013). The species studied here inhabit high mountains, where these genes may confer adaptive advantages. Accordingly, adaptive introgression events related to tolerance to high mountain ambients have been described in *Cupressus* species (Ma et al., 2019). Moreover, the introgression of genetic variants that confers adaptations to ecological pressures has been described in several plant systems. Thus, for instance, has been related with herbivory resistance (*Helianthus* spp; Kim and Rieseberg, 1999; Whitney et al., 2006; Whitney et al., 2010), tolerance to flooding ambients (*Iris*; Martin et al., 2006), drought tolerance (*Arabidopsis* spp; Arnold et al., 2006; *Quercus* spp; Khodwekar and Gailing, 2017; *Populus* spp; Chhatre et al., 2018), cold tolerance (*Populus* spp; Soolanayakanahally et al., 2009), and with photoperiodic regulation (*Populus* spp; Chhatre et al., 2018). Furthermore, adaptive introgression related to environmental changes has been described in animal systems. For instance, hares species (Jones et al., 2018) and wolf species (Anderson et al., 2009) are good examples of systems in which adaptive introgression of allelic variant confer changes in color to the receiving species that helps them to better adapt to environmental changes. Nevertheless, to assume that the introgression that we have found in CHI

and CHS are a candidate region for adaptive introgression, a direct measurement of a fitness effect of the introgressed region of these species is required.

Hence, to fully understand how the purple color has evolved in these species, it would be necessary for studying all the *Erysimum* species from the Iberian Peninsula with purple and yellow corollas. Moreover, studying the whole introgression pattern for these species and dating the polyploid appears, may help us to understand whether the success of the polyploid species is a consequence of adaptive introgression in key genes. For that, a whole-genome sequencing approach would be required, considering the influence of ghost species and the date and directionality of the introgression events. Thus, a better understanding of the hybridization and adaptive introgression pattern on these species located in a glacial refugee may help us to understand how species cope with climate fluctuations.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., ... & Butlin, R. K. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26 (2), 229-246.

Abdelaziz, M., Lorite, J., Muñoz-Pajares, A. J., Herrador, M. B., Perfectti, F., & Gómez, J. M. (2011). Using complementary techniques to distinguish cryptic species: a new *Erysimum* (Brassicaceae) species from North Africa. *American Journal of Botany*, 98 (6), 1049-1060.

Abdelaziz, M. (2013). How species are evolutionarily maintained? Pollinator-mediated divergence and hybridization in *Erysimum mediohispanicum* and *Erysimum nevadense* (Doctoral dissertation, Universidad de Granada).

Abdelaziz, M., Muñoz-Pajares, A. J., Lorite Moreno, J., Herrador, M. B., Perfectti Álvarez, F., & Gómez Reyes, J. M. (2014). Phylogenetic relationships of *Erysimum* (Brassicaceae) from the Baetic Mountains (se Iberian peninsula). *Anales del Jardín Botánico de Madrid* 71 (1): e005 2014.

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., ... & Yang, J. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics*, 47 (3), 217.

Albaladejo, R. G., & Aparicio, A. (2007). Population genetic structure and hybridization patterns in the Mediterranean endemics *Phlomis lychnitis* and *P. crinita* (Lamiaceae). *Annals of Botany*, 100 (4), 735-746.

Al-Shehbaz, I. A., & Al-Shammary, K. I. (1987). Distribution and chemotaxonomic significance of glucosinolates in certain Middle-Eastern Cruciferae. *Biochemical Systematics and Ecology*, 15 (5), 559-569.

Ancev, M. (2006). Polyploidy and hybridization in Bulgarian Brassicaceae: distribution and evolutionary role. *Phytologia Balcanica*, 12 (3), 357-366.

Anderson, E., & Hubricht, L. (1938). Hybridization in Tradescantia. III. The evidence for introgressive hybridization. *American Journal of Botany*, 25 (6), 396-402.

Anderson, E. (1953). Introgressive hybridization. *Biological Reviews*, 28 (3), 280-307.

Anderson, T. M., Candille, S. I., Musiani, M., Greco, C., Stahler, D. R., Smith, D. W., ... & Ostrander, E. A. (2009). Molecular and evolutionary history of melanism in North American gray wolves. *Science*, 323 (5919), 1339-1343.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arista, M., Talavera, M., Berjano, R., & Ortiz, P. L. (2013). Abiotic factors may explain the geographical distribution of flower colour morphs and the maintenance of colour polymorphism in the scarlet pimpernel. *Journal of Ecology*, 101 (6), 1613-1622.

Arnold, M. L., Bulger, M. R., Burke, J. M., Hempel, A. L., & Williams, J. H. (1999). Natural hybridization: how low can you go and still be important?. *Ecology*, 80 (2), 371-381.

Arnold, M. L. (2004). Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right?. *The Plant Cell*, 16 (3), 562-570.

Arnold, M. L., & Martin, N. H. (2009). Adaptation by introgression. *Journal of Biology*, 8 (9), 82.

Arnold, M. L. (2015). Divergence with genetic exchange. OUP Oxford.

Arnold, B. J., Lahner, B., DaCosta, J. M., Weisman, C. M., Hollister, J. D., Salt, D. E., ... & Yant, L. (2016). Borrowed alleles and convergence in serpentine adaptation. *Proceedings of the National Academy of Sciences*, 113 (29), 8320-8325.

Arnold, M. L., & Kunte, K. (2017). Adaptive genetic exchange: a tangled history of admixture and evolutionary innovation. *Trends in Ecology & Evolution*, 32 (8), 601-611.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28 (1), 45-48.

Barlow, A., Cahill, J. A., Hartmann, S., Theunert, C., Xenikoudakis, G., Fortes, G. G., ... & García-Vázquez, A. (2018). Partial genomic survival of cave bears in living brown bears. *Nature Ecology & Evolution*, 2 (10), 1563-1570.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., ... & Studholme, D. J. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32 (suppl_1), D138-D141.

Borge, T., Lindroos, K., Nadvornik, P., Syvänen, A. C., & Sætre, G. P. (2005). Amount of introgression in flycatcher hybrid zones reflects regional differences in pre and post-zygotic barriers to gene exchange. *Journal of Evolutionary Biology*, 18 (6), 1416-1424.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52 (3), 345-370.

Cao, Z., Liu, X., Ogilvie, H. A., Yan, Z., & Nakhleh, L. (2019). Practical aspects of phylogenetic network analysis using phylonet. *BioRxiv*, 746362.

Chapman, M. A., & Abbott, R. J. (2010). Introgression of fitness genes across a ploidy barrier. *New Phytologist*, 186 (1), 63-71.

Chalker-Scott, L. (1999). Environmental significance of anthocyanins in plant stress responses. *Photochemistry and Photobiology*, 70 (1), 1-9.

Coberly, L. C., & Rausher, M. D. (2003). Analysis of a chalcone synthase mutant in Ipomoea purpurea reveals a novel function for flavonoids: amelioration of heat stress. *Molecular Ecology*, 12 (5), 1113-1124.

Coyne, J. A., & Orr, H. A. (2004). Speciation Sinauer Associates. *Sunderland, MA*, 276, 281.

Dao, T. T. H., Linthorst, H. J. M., & Verpoorte, R. (2011). Chalcone synthase and its functions in plant resistance. *Phytochemistry Reviews*, 10 (3), 397.

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., ... & Schmidt, J. M. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354 (6311), 477-481.

Dick, C. A., Buenrostro, J., Butler, T., Carlson, M. L., Kliebenstein, D. J., & Whittall, J. B. (2011). Arctic mustard flower color polymorphism controlled by petal-specific downregulation at the threshold of the anthocyanin biosynthetic pathway. *PLoS One*, 6 (4).

Doležel, J., Sgorbati, S., & Lucretti, S. (1992). Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiologia Plantarum*, 85 (4), 625-631.

Dufresnes, C., Nicieza, A. G., Litvinchuk, S. N., Rodrigues, N., Jeffries, D. L., Vences, M., ... & Martínez-Solano, Í. (2020). Are glacial refugia hotspots of speciation and cytonuclear discordances? Answers from the genomic phylogeography of Spanish common frogs. *Molecular Ecology*, 29 (5), 986-1000.

Doležel, J., Greilhuber, J., & Suda, J. (Eds.). (2007). Flow cytometry with plant cells: analysis of genes, chromosomes and genomes. John Wiley & Sons.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28 (8), 2239-2252.

Egea, R., Casillas, S., & Barbadilla, A. (2008). Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, 36 (suppl_2), W157-W162.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16 (1), 157.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology,* 20 (1), 1-14.

Folk, R. A., Mandel, J. R., & Freudenstein, J. V. (2017). Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Systematic Biology*, 66 (3), 320-337.

Folk, R. A., Visger, C. J., Soltis, P. S., Soltis, D. E., & Guralnick, R. P. (2018). Geographic range dynamics drove ancient hybridization in a lineage of angiosperms. *The American Naturalist*, 192 (2), 171-187.

Foote, A. D., Martin, M. D., Louis, M., Pacheco, G., Robertson, K. M., Sinding, M. H. S., ... & Barlow, J. (2019). Killer whale genomes reveal a complex history of recurrent admixture and vicariance. *Molecular Ecology*, 28 (14), 3427-3444.

Forsythe, E. S., Nelson, A. D., & Beilstein, M. A. (2018). Biased gene retention in the face of massive nuclear introgression obscures species relationships. *BioRxiv*, 197087.

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., & Firoozabady, E. (1983). Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*, 220 (4601), 1049-1051.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database an informatics resource. *Nucleic Acids Research*, 32 (suppl_1), D258-D261.

Chhatre, V. E., Evans, L. M., DiFazio, S. P., & Keller, S. R. (2018). Adaptive introgression and maintenance of a trispecies hybrid complex in range-edge populations of Populus. *Molecular Ecology*, *27* (23), 4820-4838.

Gokhman, D., Mishol, N., de Manuel, M., de Juan, D., Shuqrun, J., Meshorer, E., ... & Carmel, L. (2019). Reconstructing denisovan anatomy using DNA methylation maps. *Cell*, 179 (1), 180-192.

Gómez, J. M., González-Mejias A, Lorite J, Abdelaziz M, Perfectti F. (2015a). The silent extinction: Climate change and the potential for hybridization-mediated extinction of endemic high-mountain plants. *Biodiversity and Conservation* 24: 1843–1857.

Gómez, J. M., Perfectti, F., & Lorite, J. (2015b). The role of pollinators in floral diversification in a clade of generalist flowers. *Evolution*, 69 (4), 863-878.

Goulet, B. E., Roda, F., & Hopkins, R. (2017). Hybridization in plants: old ideas, new techniques. *Plant Physiology*, 173 (1), 65-78.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29 (7), 644.

Grant, P. R., & Grant, B. R. (2002). Unpredictable evolution in a 30-year study of Darwin's finches. *Science*, 296 (5568), 707-711.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... & Hansen, N. F. (2010). A draft sequence of the Neandertal genome. *Science*, 328 (5979), 710-722.

Greilhuber, J., Doležel, J., Lysak, M. A., & Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms ‘genome size’and ‘C-value’to describe nuclear DNA contents. *Annals of Botany*, 95 (1), 255-260.

Greuter, W., Burdet, H. M., & Long, G. (1986). Dicotyledones (Convolvulaceae-Labiatae). *Med-Checklist*, 3, 106-116.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43 (10), 1031.Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & MacManes, M. D. (2013). *De novo* transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8 (8), 1494.

Haas, B. J. (2015). Trinotate: transcriptome functional annotation and analysis.

Hanušová, K., Ekrt, L., Vit, P., Kolář, F., & Urfus, T. (2014). Continuous morphological variation correlated with genome size indicates frequent introgressive hybridization among *Diphasiastrum* species (Lycopodiaceae) in Central Europe. *PLoS One*, 9 (6), e99552.

Harrison, R. G., & Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105 (S1), 795-809.

Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J., & Cronk, Q. C. (2014). Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related

balsam poplars, *Populus balsamifera* and *P. trichocarpa* (S alicaceae). *New Phytologist*, 204 (3), 693-703.

Hughes, P. D., & Woodward, J. C. (2017). Quaternary glaciation in the Mediterranean mountains: a new synthesis. *Geological Society, London, Special Publications*, 433 (1), 1-23.

Huson, D. H., & Scornavacca, C. (2019). User Manual for Dendroscope V3.6.2.

Irwin, R. E., Strauss, S. Y., Storz, S., Emerson, A., & Guibert, G. (2003). The role of herbivores in the maintenance of a flower color polymorphism in wild radish. *Ecology*, 84 (7), 1733-1743.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., & Bork, P. (2007). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36 (suppl_1), D250-D254.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36 (suppl_2), W5-W9.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11 (1), 94.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27 (21), 3070-3071.

Jones, M. R., Mills, L. S., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J., ... & Good, J. M. (2018). Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*, 360 (6395), 1355-1358.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28 (1), 27-30.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30 (4), 772-780.

Kay, K. M., Woolhouse, S., Smith, B. A., Pope, N. S., & Rajakaruna, N. (2018). Sympatric serpentine endemic Monardella (Lamiaceae) species maintain habitat differences despite hybridization. *Molecular Ecology*, 27 (9), 2302-2316.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... & Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28 (12), 1647-1649.

Keightley, P. D., & Eyre-Walker, A. (2012). Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *Journal of Molecular Evolution*, 74 (1-2), 61-68.

Kim, S. C., & Rieseberg, L. H. (1999). Genetic architecture of species differences in annual sunflowers: implications for adaptive trait introgression. *Genetics*, 153 (2), 965-977.

Kim, M., Cui, M. L., Cubas, P., Gillies, A., Lee, K., Chapman, M. A., ... & Coen, E. (2008). Regulatory genes control a key morphological and ecological trait transferred between species. *Science*, 322 (5904), 1116-1119.

Khlestkina, E. K., & Shoeva, O. Y. (2014). Intron loss in the chalcone-flavanone isomerase gene of rye. *Molecular Breeding*, 33 (4), 953-959.

Khodwekar, S., & Gailing, O. (2017). Evidence for environment-dependent introgression of adaptive genes between two red oak species with different drought adaptations. *American Journal of Botany*, 104 (7), 1088-1098.

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genetics*, 4 (12), e1000304.

Kuhlwilm, M., Han, S., Sousa, V. C., Excoffier, L., & Marques-Bonet, T. (2019). Ancient admixture from an extinct ape lineage into bonobos. *Nature Ecology & Evolution*, 3 (6), 957-965.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., ... & Karthikeyan, A. S. (2011). The *Arabidopsis* Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, 40 (D1), D1202-D1210.

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., & Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytologist*, 221 (1), 515-526.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22 (13), 1658-1659.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25 (14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25 (16), 2078-2079.

Lin, H. Y., Hao, Y. J., Li, J. H., Fu, C. X., Soltis, P. S., Soltis, D. E., & Zhao, Y. P. (2019). Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* sl (Theaceae). *Molecular Phylogenetics and Evolution*, 135, 1-11.

Liu, L., Bosse, M., Megens, H. J., Frantz, L. A., Lee, Y. L., Irving-Pease, E. K., ... & Madsen, O. (2019). Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nature Communications*, 10 (1), 1-9.

Loureiro, J., Rodriguez, E., Doležel, J., & Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Annals of Botany*, 100 (4), 875-888.

Lu, Y., & Rausher, M. D. (2003). Evolutionary rate variation in anthocyanin pathway genes. *Molecular Biology and Evolution*, 20 (11), 1844-1853.

Ma, Y., Wang, J., Hu, Q., Li, J., Sun, Y., Zhang, L., ... & Mao, K. (2019). Ancient introgression drives adaptation to cooler and drier mountain habitats in a cypress species complex. *Communications Biology*, 2 (1), 1-12.

Malinsky, M. (2019). Dsuite-fast D-statistics and related admixture evidence from VCF files. *BioRxiv*, 634477.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20 (5), 229-237.

Mateo, G., Villalba, M. B. C., & Udias, S. L. (1998). Acerca del orófito minusvalorado de la Sierra de Javalambre (Teruel). *Flora Montiberica*, (9), 41-45.

Marhold, K., & Lihová, J. (2006). Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. *Plant Systematics and Evolution*, 259 (2-4), 143-174.

Martin, N. H., Bouck, A. C., & Arnold, M. L. (2006). Detecting adaptive trait introgression between *Iris fulva* and *I. brevicauli*s in highly selective field conditions. *Genetics*, 172 (4), 2481-2489.

Martin, M. (2011). Cutadapt removes adapter sequences from highthroughput sequencing reads. *EMBnet J*, 17: 10–12.

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2014). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 32 (1), 244-257.

Mayr, E. (1992). A local flora and the biological species concept. *American Journal of Botany*, 79 (2), 222-238.

Médail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, 36 (7), 1333-1345.

Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., ... & Sudmant, P. H. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338 (6104), 222-226.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30 (17), i541-i548.

Mori, T., Kuroiwa H., Higashiyama, T., and Kuroiwa T. (2006). Generative Cell Specific 1 is essential for angiosperm fertilization. *Nature Cell Biology* 8 (1): 64-71.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32 (11), 1749-1751.

Nieto-Feliner, G. (1993) *Erysimum* L. In: Flora iberica. Vol. IV. Cruciferae-Monotropaceae. Real Jardín Botánico, CSIC, Madrid, pp. 48–76.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32 (1), 268-274.

Ortego, J., Gugger, P. F., Riordan, E. C., & Sork, V. L. (2014). Influence of climatic niche suitability and geographical overlap on hybridization patterns among southern Californian oaks. *Journal of Biogeography*, 41 (10), 1895-1908.

Ottenburghs, J. (2020). Ghost Introgression: Spooky Gene Flow in the Distant Past. *BioEssays*.

Pajares, A. J. M. (2013). *Erysimum mediohispanicum at the evolutionary crossroad: phylogrography, phenotype, and pollinators* (Doctoral dissertation, Universidad de Granada).

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20 (2), 289-290.

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, 25 (11), 2337-2360.

Payton, A. C., Naranjo, A. A., Judd, W., Gitzendanner, M., Soltis, P. S., & Soltis, D. E. (2019). Population genetics, speciation, and hybridization in *Dicerandra* (Lamiaceae), a North American Coastal Plain endemic, and implications for conservation. *Conservation Genetics*, 20 (3), 531-543.

Pereira, R. J., Martínez-Solano, I., & Buckley, D. (2016). Hybridization during altitudinal range shifts: nuclear introgression leads to extensive cyto-nuclear discordance in the fire salamander. *Molecular Ecology*, 25 (7), 1551-1565.

Pfeifer, B., & Kapan, D. D. (2019). Estimates of introgression as a function of pairwise distances. *BMC Bioinformatics*, 20 (1), 207.

Polatschek, A. (1978). Die arten der gattung *Erysimum* auf der Iberischen Halbinsel. *Annalen des Naturhistorischen Museums in Wien*, 325-362.

Polatschek A. (1986). *Erysimum*. In Strid A. [ed.], Mountain flora of Greece, l, 239–247. Cambridge University Press, Cambridge, UK.

Polatschek, A. (2014). Revision der gattung *Erysimum* (Cruciferae): Nachträge zu den bearbeitungen der Iberischen. Halbinsel und Makaronesiens. *Annalen des Naturhistorischen Museums in Wien. Serie B für Botanik und Zoologie*, 87-105.

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16 (6), 359-371.

Rambaut, A., & Drummond, A. J. (2012). FigTree version 1. 4. 0.

Rakotoarivelo, A. R., O'Donoghue, P., Bruford, M. W., & Moodley, Y. (2019). An ancient hybridization event reconcilesmito-nuclear discordance among spiral-horned antelopes. *Journal of Mammalogy*, 100 (4), 1144-1155.

Rausher, M. D., Miller, R. E., & Tiffin, P. (1999). Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology and Evolution*, 16 (2), 266-274.

Rendón-Anaya, M., Montero-Vargas, J. M., Saburido-Álvarez, S., Vlasova, A., Capella-Gutierrez, S., Ordaz-Ortiz, J. J., ... & Gabaldón, T. (2017). Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biology*, 18 (1), 60.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3 (2), 217-223.

Rieseberg, L. H., & Wendel, J. F. (1993). Introgression and its consequences in plants. *Hybrid Zones and the Evolutionary Process*, 70, 109.

Rieseberg, L. H., & Carney, S. E. (1998). Plant hybridization. *The New Phytologist*, 140 (4), 599-624.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., ... & Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, 301 (5637), 1211-1216.

Rifkin, J. L., Castillo, A. S., Liao, I. T., & Rausher, M. D. (2019). Gene flow, divergent selection and resistance to introgression in two species of morning glories (Ipomoea). *Molecular Ecology*, 28 (7), 1709-1729.

Saari, S., & Faeth, S. H. (2012). Hybridization of Neotyphodium endophytes enhances competitive ability of the host grass. *New Phytologist*, 195 (1), 231-236.

Schemske, D. W. (2000). Understanding the Origin of Species 1. *Evolution*, 54 (3), 1069-1073.

Schemske, D. W., & Bierzychudek, P. (2007). Spatial differentiation for flower color in the desert annual *Linanthus parryae*: was Wright right?. *Evolution: International Journal of Organic Evolution*, *61* (11), 2528-2543.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27 (4) 592-593.

Schrinner, S., Mari, R. S., Ebler, J. W., Rautiainen, M., Seillier, L., Reimer, J., ... & Klau, G. W. (2020). Haplotype Threading: Accurate Polyploid Phasing from Long Reads. *BioRxiv*.

Seehausen, O. L. E., Takimoto, G., Roy, D., & Jokela, J. (2008). Speciation reversal and biodiversity dynamics with hybridization in changing environments. *Molecular Ecology*, 17 (1), 30-44.

Seixas, F. A., Boursot, P., & Melo-Ferreira, J. (2018). The genomic impact of historical hybridization with massive mitochondrial DNA introgression. *Genome Biology*, 19 (1), 91.

Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16 (8), 1114-1114.

Skov, L., Hui, R., Shchur, V., Hobolth, A., Scally, A., Schierup, M. H., & Durbin, R. (2018). Detecting archaic introgression using an unadmixed outgroup. *PLoS Genetics*, 14 (9), e1007641.

Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, 60, 561-588.

Soltis, D. E., Visger, C. J., & Soltis, P. S. (2014). The polyploidy revolution then… and now: Stebbins revisited. *American Journal of Botany*, 101 (7), 1057-1078.

Soolanayakanahally, R. Y., Guy, R. D., Silim, S. N., Drewes, E. C., & Schroeder, W. R. (2009). Enhanced assimilation rate and water use efficiency with latitude through increased photosynthetic capacity and internal conductance in balsam poplar (*Populus balsamifer*a L.). *Plant, Cell & Environment*, 32 (12), 1821-1832.

Stebbins, G. L. (1959). The role of hybridization in evolution. *Proceedings of the American Philosophical Society*, 103 (2), 231-251.

Stelkens, R., & Seehausen, O. (2009). Genetic distance between species predicts novel trait expression in their hybrids. *Evolution: International Journal of Organic Evolution*, 63 (4), 884-897.

Strauss, S. Y., Irwin, R. E., & Lambrix, V. M. (2004). Optimal defence theory and flower petal colour predict variation in the secondary chemistry of wild radish. *Journal of Ecology*, 92 (1), 132-141.

Suarez-Gonzalez, A., Hefer, C. A., Christe, C., Corea, O., Lexer, C., Cronk, Q. C., & Douglas, C. J. (2016). Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. átrichocarpa* (black cottonwood). *Molecular Ecology*, 25 (11), 2427-2442.

Suarez-Gonzalez, A. (2017). Adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood) (Doctoral dissertation, University of British Columbia).

Suarez-Gonzalez, A., Lexer, C., & Cronk, Q. C. (2018). Adaptive introgression: a plant perspective. *Biology Letters*, 14 (3), 20170688.

Sujii, P. S., Cozzolino, S., & Pinheiro, F. (2019). Hybridization and geographic distribution shapes the spatial genetic structure of two co-occurring orchid species. *Heredity*, 123 (4), 458-469.

Sušnik, S., Weiss, S., Odak, T., Delling, B., Treer, T., & Snoj, A. (2007). Reticulate evolution: ancient introgression of the Adriatic brown trout mtDNA in softmouth trout *Salmo obtusirostris* (Teleostei: Salmonidae). *Biological Journal of the Linnean Society*, *90* (1), 139-152.

Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution*, 3 (2), 170-177.

Than, C., Ruths, D., & Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9 (1), 322.

Treutter, D. (2006). Significance of flavonoids in plant resistance: a review. *Environmental Chemistry Letters*, 4 (3), 147.

Turner, B. L. (2006). Taxonomy and nomenclature of the *Erysimum asperum-E. capitatum* complex (Brassicaceae). *Phytologia*, *88*, 279-287.

UniProt Consortium. (2014). UniProt: a hub for protein information. *Nucleic Acids Research*, 43 (D1), D204-D212.

Vaidya, P., McDurmon, A., Mattoon, E., Keefe, M., Carley, L., Lee, C. R., ... & Anderson, J. T. (2018). Ecological causes and consequences of flower color polymorphism in a self-pollinating plant (Boechera stricta). *New Phytologist*, 218 (1), 380-392.

Wang, D., Wang, Z., Kang, X., & Zhang, J. (2019). Genetic analysis of admixture and hybrid patterns of *Populus hopeiensis* and *P. tomentosa*. *Scientific Reports*, 9 (1), 1-13.

Ward, B. J., & Van Oosterhout, C. (2016). HybridCheck: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Molecular Ecology Resources*, 16 (2), 534-539.

Warwick, S. I., Francis, A., & Al-Shehbaz, I. A. (2006). Brassicaceae: species checklist and database on CD-Rom. *Plant Systematics and Evolution*, 259 (2-4), 249-258.

Warren, J., & Mackenzie, S. (2001). Why are all colour combinations not equally represented as flower-colour polymorphisms?. *New Phytologist*, 151 (1), 237-241.

Wen, D., Yu, Y., Zhu, J., & Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67 (4), 735-740.

Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2006). Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. *The American Naturalist*, 167 (6), 794-807.

Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2010). Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytologist*, 187 (1), 230-239.

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiology*, 126 (2), 485-493.

Wu, Y. C., Rasmussen, M. D., Bansal, M. S., & Kellis, M. (2014). Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Research*, 24 (3), 475-486.

Yang, Y., & Smith, S. A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, 31 (11), 3081-3092.

Yu, Y., & Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16 (10), S10.

Zhang, D., Tang, L., Cheng, Y., Hao, Y., Xiong, Y., Song, G., ... & Lei, F. (2019). “Ghost Introgression” As a Cause of Deep Mitochondrial Divergence in a Bird Species Complex. *Molecular Biology and Evolution*, 36 (11), 2375-2386.

Zhang, X., Wu, R., Wang, Y., Yu, J., & Tang, H. (2019). Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal*.

Zieliński, P., Dudek, K., Arntzen, J. W., Palomar, G., Niedzicka, M., Fijarczyk, A., ... & Babik, W. (2019). Differential introgression across newt hybrid zones: Evidence from replicated transects. *Molecular Ecology*, 28 (21), 4811-4824.

Supplementary Material

Figure S1. Map of the Iberian Peninsula showing the location of the sampled populations.

Figure S2. Cyto-nuclear discordance among the nuclear species tree and the plastidial phylogeny.

Figure S3. The optimal species network inferred using the PhyloNet software.

Figure 4. Boxplots depicting dN/dS ratios for CHS and CHI genes

Table S1. Summary of sequencing statistics.

Table S2. Transcriptome assembly statistics.

Table S3. Results of the annotation of the assembled transcripts using different protein databases.

Table S4. Summary of OrthoFinder results.

Table S5. Log likelihood and AIC for all the estimated phylogenetic species network (range of 0 to 15 reticulations).

Table S6. D-statistics with corrected p-values for all triplets combinations (attached as .xls).

Table S7-S12. Summary of the f^d statistics for the ABP genes.

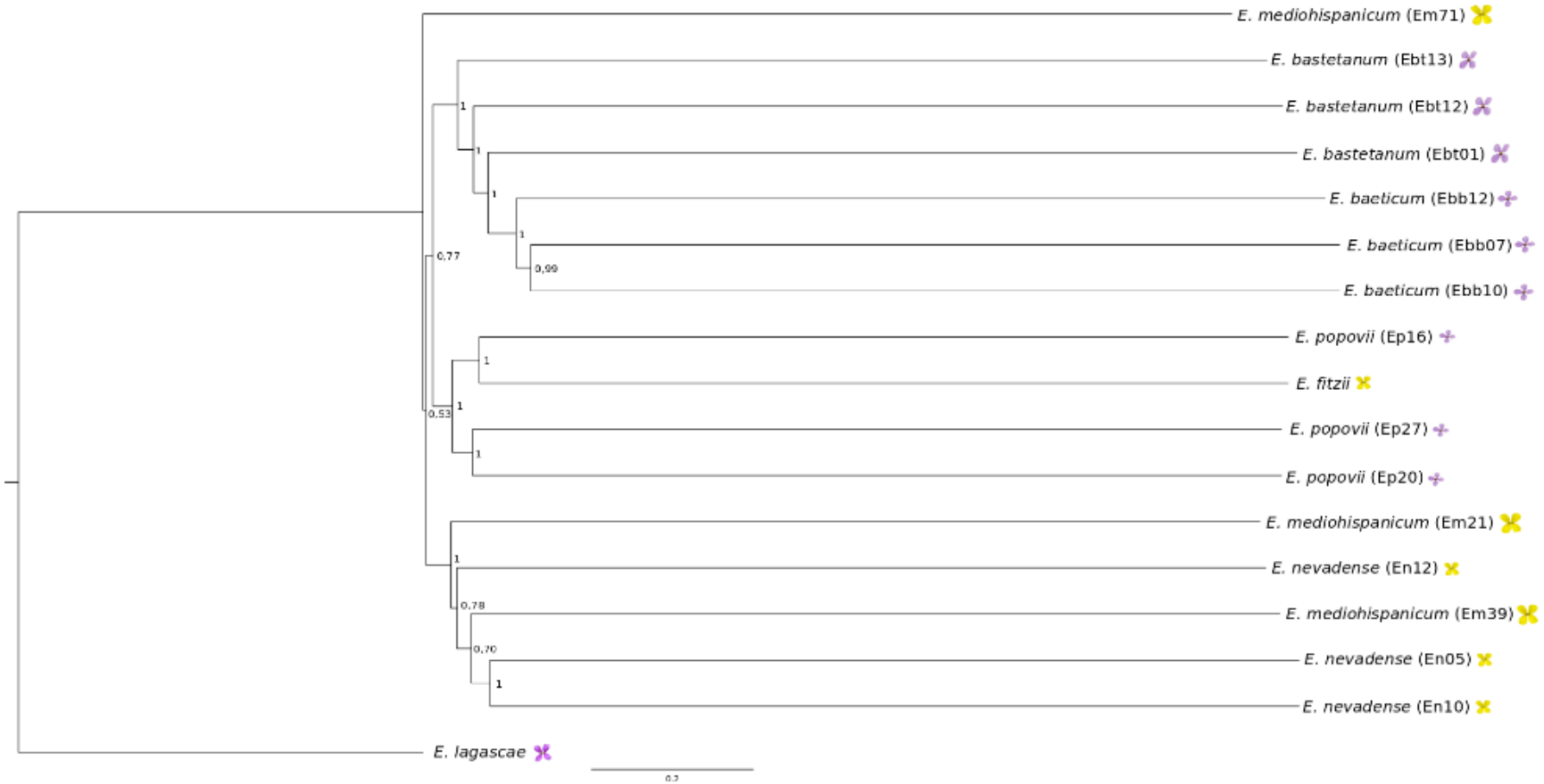

Figure S1. Phylogeny inferred after a species tree analysis (ASTRAL) of 16,941 gene trees. We rooted the obtained species tree to *E. lagascae*, considered the more phylogenetically distant to the rest of species in our data set. Branch support (BS) values are in coalescent units and are a direct measure of the amount of discordance in the gene trees. The code of the populations appears inside parentheses.

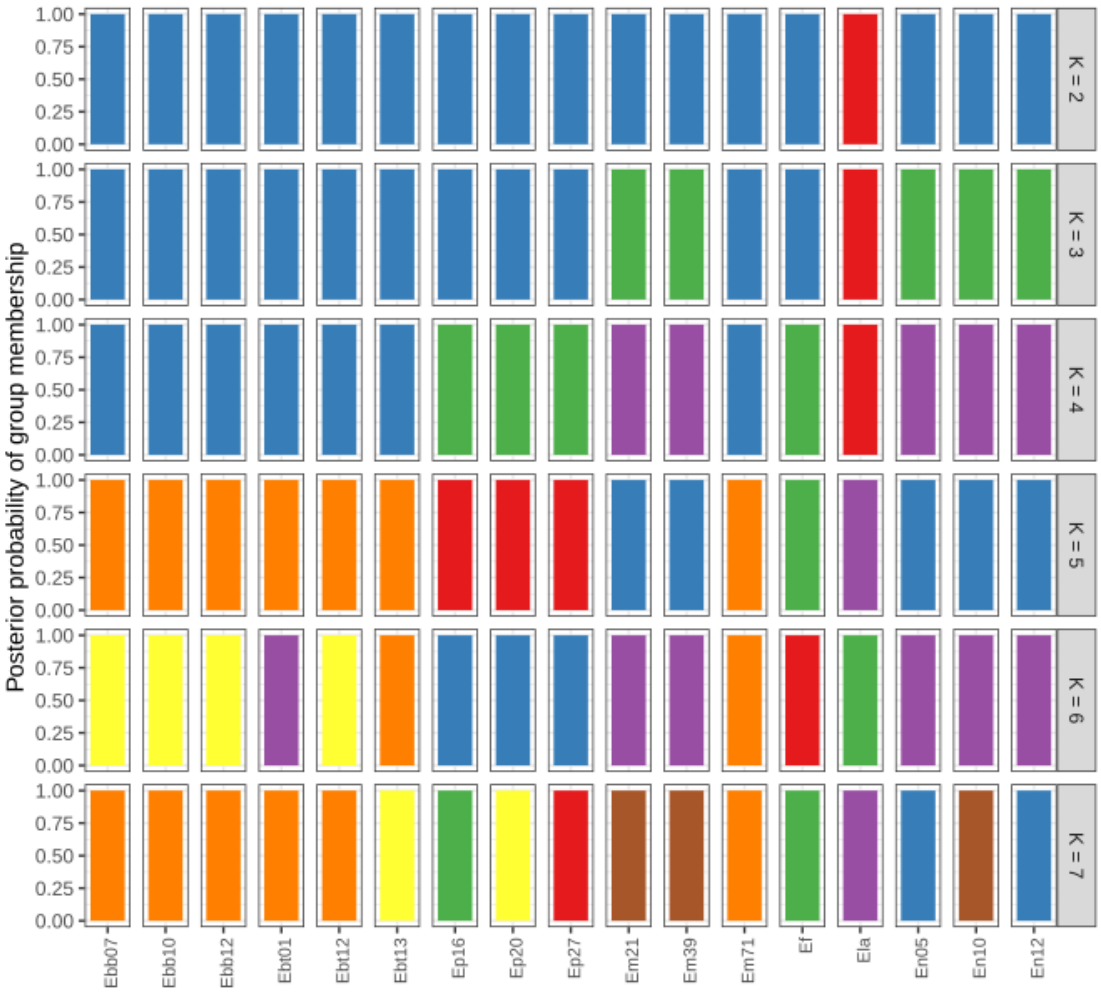

Figure S2. Membership probability plot showing the DAPC results representing the populations grouped into predetermined different clusters (ranging from K=2 to K=7, where each color represent a cluster). The Bayesian analyses (BIC) revealed K = 5 as the most likely number of genetic clusters.

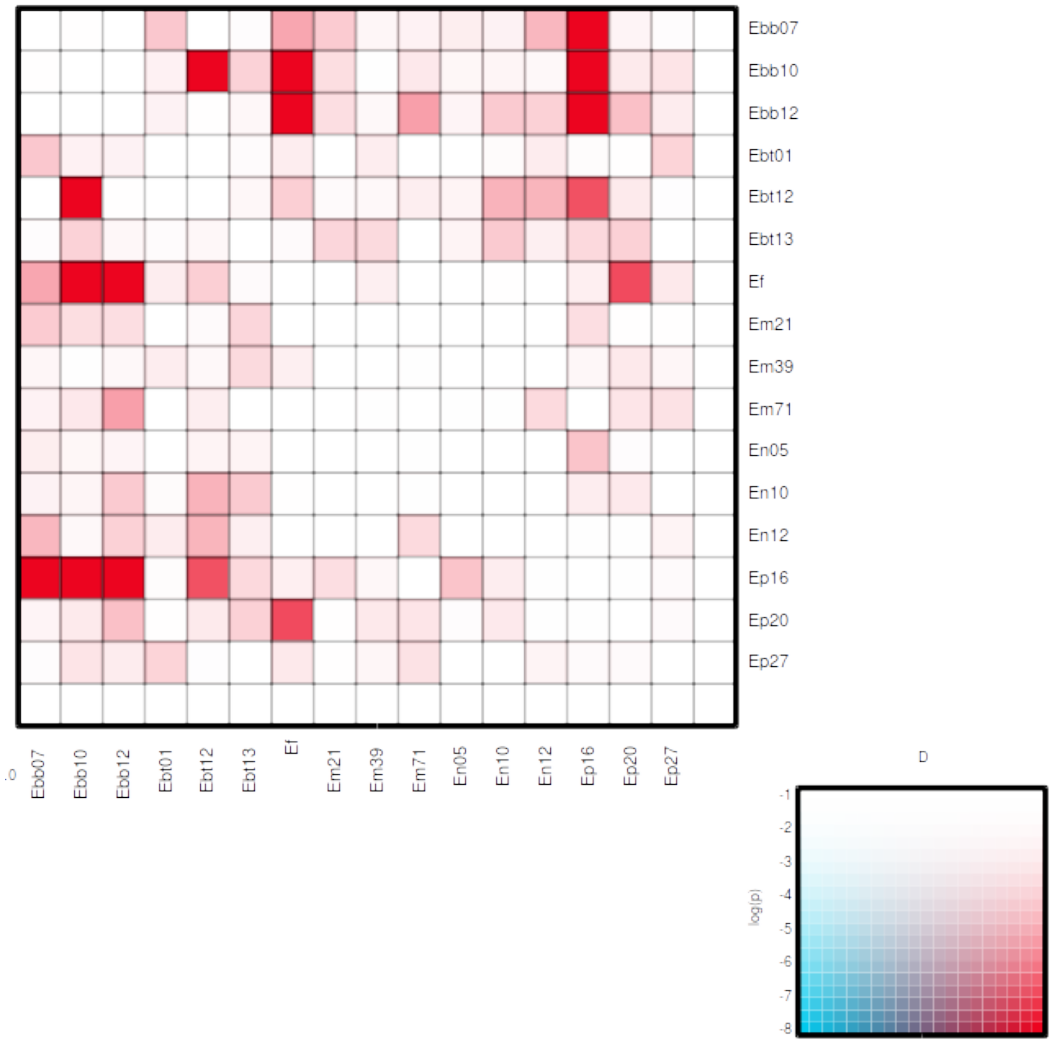

Figure S3. Heatmap depicting the introgression found by Dsuite. The colors of this heatmap show the D-statistic as well as its p-value. Red colors indicate higher D-statistics, and generally more saturated colors indicate greater significance. Thus, the strongest signals for introgression are shown with saturated red, as in the bottom right of the color legend.

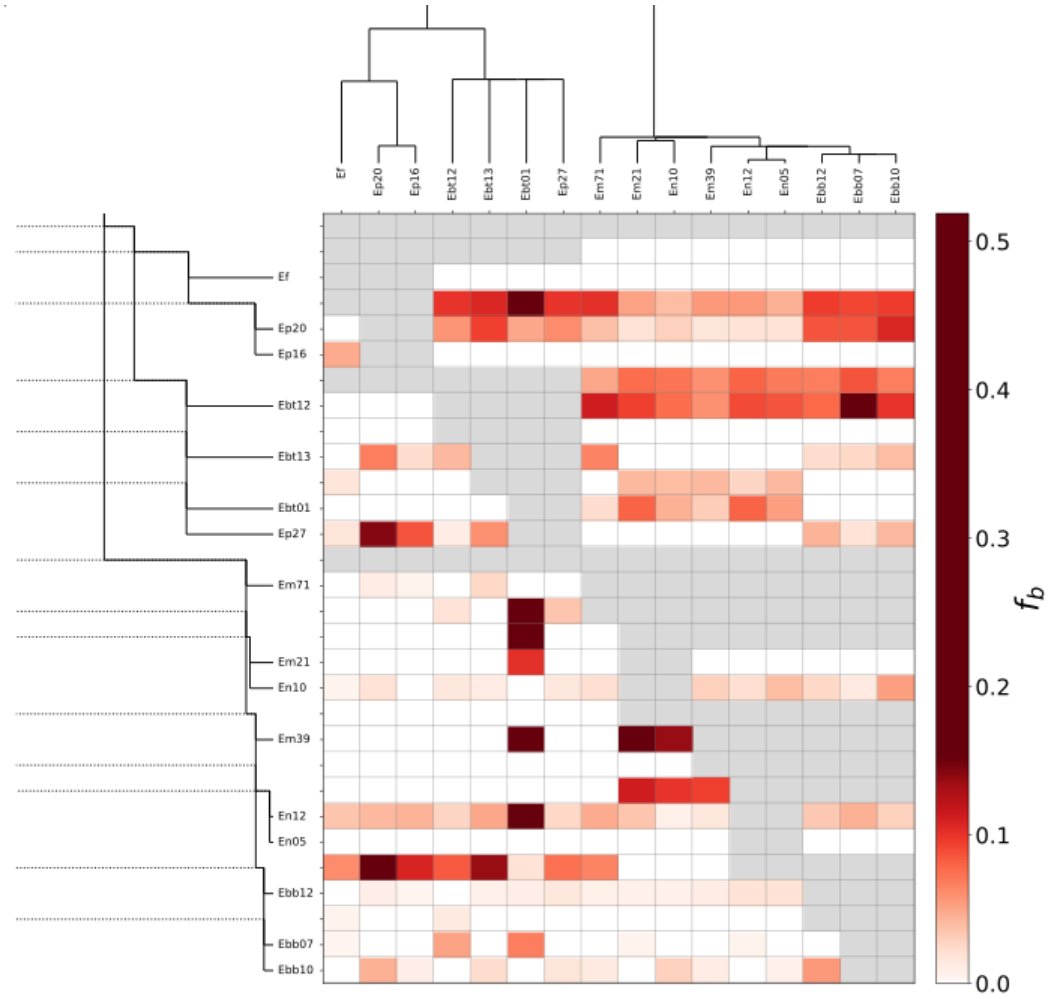

Figure S4. Heatmap depicting the gene flow among phylogeny branches estimated with the fbranch statistic. The phylogenetic tree is shown along the x and y axes (in 'laddered' form along the y-axis). The matrix shows the inferred f-branch statistics, showing the gene flow between the branch of the 'laddered' tree on the y-axis and the species identified on the x-axis. In the y-axis appears some branches with depicting gene flow that not corresponds to any sampled taxon (horizontal dotted lines). These branches represented ancestral taxa or non-sampled taxa that have contributed to the gene flow.

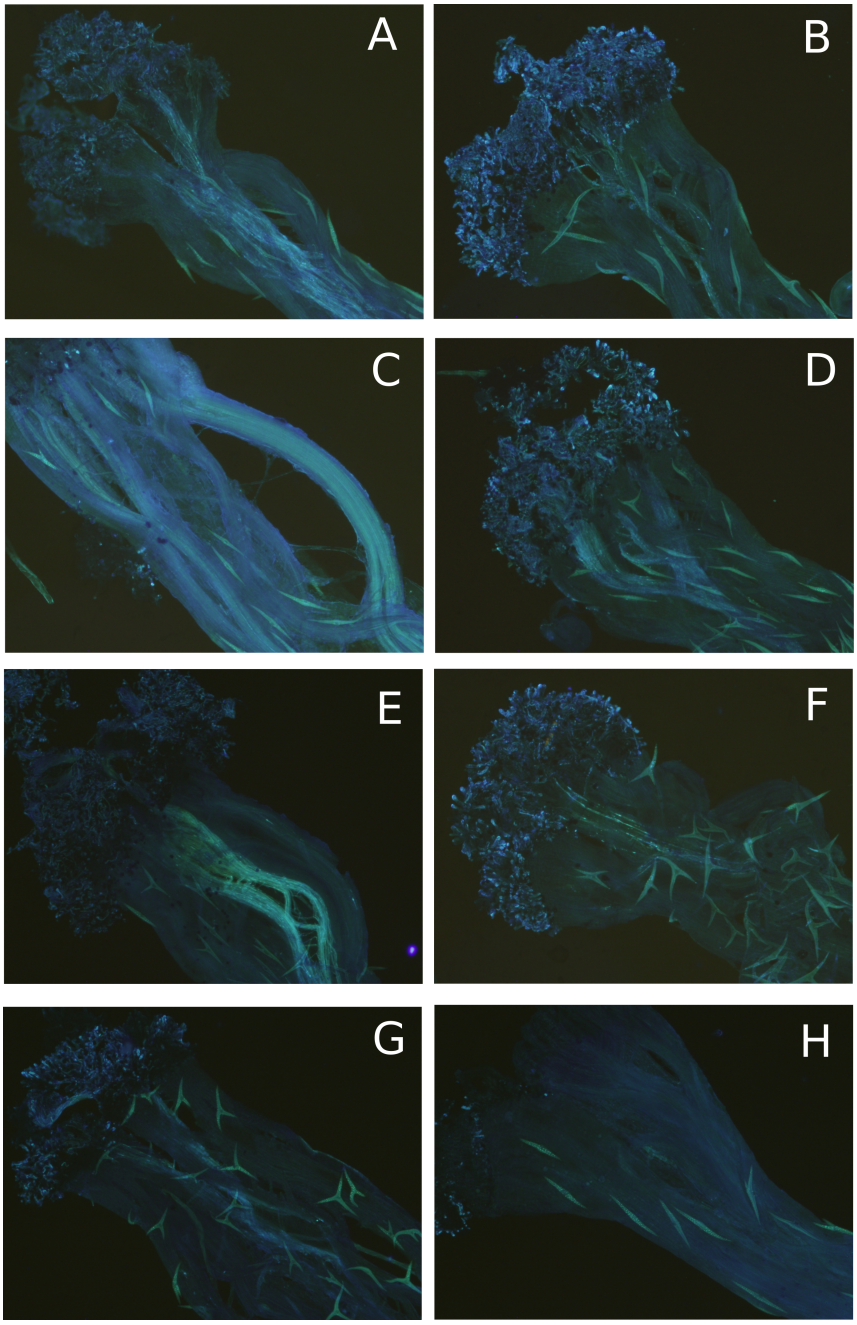

Figure S5. Pollen tube growth as a result of hand pollination crosses. A and B, pollen tube formation after a hybrid cross of *E. bastetanum* with *E. mediohispanicum* (A) and with *E. popovii* (B); C and D, pollen tube formation after a hybrid cross of *E. popovii* with *E. mediohispanicum* (C) and *E. bastetanum* (D); E and F, pollen tube formation after a hybrid cross of *E. mediohispanicum* with *E. popovii* (E) and with *E. bastetanum* (F), F pollen tube formation after an intra-specific cross of *E. popovii* (Ep27) with other population (Ep16) of the same species, and G, shows a lack of growth of the pollen tube.

Taxon	Population	Total read bases (bp)	Total number of reads (bp)	GC (%)	Q20 (%)	Q30 (%)	Gb	Reads after trimming (bp)
E. baeticum	Ebb07	10,342,234,016	68,491,616	47.38	96.13	91.88	5.1	60,985,000
E. baeticum	Ebb10	20,383,113,406	134,987,506	47.25	96.63	92.11	9.9	119,520,613
E. baeticum	Ebb12	20,634,741,014	136,653,914	47.30	96.71	92.24	10.1	121,103,833
E. bastetanum	Ebt01	11,454,840,974	75,859,874	44.88	99.89	99.10	5.0	75,854,979
E. bastetanum	Ebt12	10,417,333,262	68,988,962	45.71	98.22	95.24	5.2	65,757,353
E. bastetanum	Ebt13	10,777,576,680	71,374,680	45.47	98.58	95.91	5.4	65,765,324
E. fitzii	Ef01	10,156,440,596	67,261,196	47.09	98.94	96.66	4.6	66,245,223
E. lagascae	Ela07	7,201,775,508	71,304,708	43.50	96.47	93.68	5.4	65,740,933
E. mediohispanicum	Em21	10,235,928,808	67,787,608	48.86	96.05	91.72	4.7	60,130,101
E. mediohispanicum	Em39	23,224,849,148	153,806,948	47.49	96.79	92.43	11.2	136,858,973
E. mediohispanicum	Em71	10,242,859,708	67,833,508	46.51	98.18	95.14	5.0	64,619,740
E. nevadense	En05	22,265,032,144	147,450,544	46.73	96.65	92.20	10.8	130,509,040
E. nevadense	En10	24,045,140,340	159,239,340	44.94	96.90	92.72	11.4	142,305,123
E. nevadense	En12	10,141,853,090	67,164,590	47.49	96.17	91.96	5.1	59,838,058
E. popovii	Ep16	10,556,586,670	69,911,170	46.26	96.09	91.78	5.7	62,135,880
E. popovii	Ep20	10,860,711,844	71,925,244	45.98	98.12	94.96	5.5	68,273,570
E. popovii	Ep27	25,555,407,006	169,241,106	47.38	96.64	92.14	12.5	149,733,87

Table S1. Summary of the sequencing statistics.

Taxon	Population	Total Trinity genes	Total Trinity transcripts	Contig N50	Median contig length	Average contig length	Total assembled bases (bp)
E. baeticum	Ebb07	116,006	188,787	983	423	678.35	128,063,827
E. baeticum	Ebb10	235,313	382,286	859	381	617.37	236,012,489
E. baeticum	Ebb12	171,950	335,960	958	417	663.71	222,979,601
E. bastetanum	Ebt01	164,708	291,831	991	438	688.00	200,778,797
E. bastetanum	Ebt12	123,268	212,255	1,088	444	723.11	153,483,728
E. bastetanum	Ebt13	186,374	278,526	938	382	639.83	126,262,947
E. fitzii	Ef01	77,047	130,076	1,502	620	957.81	124,588,696
E. lagascae	Ela07	106,811	203,045	1,361	540	865.68	175,772,242
E. mediohispanicum	Em21	66,162	104,486	1,362	579	888.57	92,843,559
E. mediohispanicum	Em39	93,902	238,394	1,495	662	977.67	233,071,647
E. mediohispanicum	Em71	93,154	160,490	1,128	490	764.13	122,635,013
E. nevadense	En05	92,897	235,515	1,504	663	981.53	231,165,137
E. nevadense	En10	217,656	368,656	1,436	496	867.83	319,929,639
E. nevadense	En12	75,088	126,245	1,212	561	829.65	104,739,622
E. popovii	Ep16	107,329	186,398	1,130	477	757.88	141,267,189
E. popovii	Ep20	199,665	300,036	738	353	566.56	169,989,446
E. popovii	Ep27	123,780	288,344	1,249	529	819.94	236,425,328

Table S2. Summary of the transcriptome assembly statistics.

Number of unigenes								
Taxon	**Population code**	**Sprot_top_BLASTX_hit**	**Sprot_top_BLASTP_hit**	**Pfam**	**GO_Pfam**	**GO_BLAST**	**eggog**	**Kegg**
E. mediohispanicum	Em21	71606	51019	45366	29311	64679	62559	59930
E. mediohispanicum	Em39	170089	119187	105585	68625	154101	148134	140544
E. mediohispanicum	Em71	159869	86090	74989	46573	146455	140640	133137
E. nevadense	En05	167406	117596	103913	67649	151343	145238	138224
E. nevadense	En10	189900	127250	113593	75293	175240	164625	156297
E. nevadense	En12	85222	59138	52245	33974	77857	74795	71130
E. popovii	Ep16	114873	72485	64212	40623	104434	99335	94405
E. popovii	Ep20	164455	79736	68739	42568	152320	145732	137445
E. popovii	Ep27	184887	116140	102230	65460	167458	161254	153256
E. bastetanum	Ebt01	165626	98309	85207	52817	149999	142448	135739
E. bastetanum	Ebt12	128004	76939	68066	42817	116348	111477	105855
E. bastetanum	Ebt13	159869	86090	74989	46573	146455	140640	133137
E. baeticum	Ebb07	112543	1844	58464	37009	41713	40129	37760
E. baeticum	Ebb10	197069	111229	97272	61805	184103	172826	164226
E. baeticum	Ebb12	190835	108248	22484	22480	175134	166701	158092
E. lagascae	Ela07	132907	85868	76201	48546	118802	113696	108497
E. fitzii	Ef01	92184	65200	57552	37661	83190	79488	75459

Table S3. Annotation summary of the assembled transcripts using different databases: SwissProt (BLASTX and BLASTP), Pfam, eggnog, GO, and Kegg.

Total number of Orthogroup	92,984
Mean Orthogroup size	16.3
Median Orthogroup size	7.0
Number of protein genes	1,574,983
N. of genes in Orthogroups	1,519,064
N. of unassigned protein genes	55,919
Percentage of genes in Orthogroups	96.4
Percentage of unassigned protein genes	3.6
Number of Orthogroup with all species present	16,941

Table S4. Summary of OrthoFinder results.

	Log likelihood	AIC
1	-1.0224071648028404E7	20448145
2	-1.022316938865218E7	20446343
3	-1.0222191339585347E7	20444389
4	-1.0221615102393162E7	20443238
5	-1.022131616996915E7	20442642
6	-1.0220690757836848E7	20441394
7	-1.0220980513494018E7	20441975
8	-1.0223136397700764E7	20446289
9	-1.0220246575984979E7	20440511
10	-1.0220445144143904E7	20440910
11	-1.0220701238338212E7	20441414
12	-1.0219898642573046E7	20439821
13	-1.0219513355204735E7	*20439053
14	-1.021989388332834E7	20439816
15	-1.0220243030447358E7	20440516

Table S5. Log likelihood and AIC for all the estimated phylogenetic species network (range of 0 to 15 reticulations). The network with the lower AIC is depicted by an asterisk.

Gene	P1	P2	P3	f^d P1,P3	f^d P2,P3	p-value	D
CHS	En	Ebb	Ela	0.005	0.001	> 0.001	0
CHS	En	Ebt	Ela	0	0.10	> 0.001	0
CHS	En	Ep	Ela	0.05	0	> 0.001	0
CHS	Em	Ebt	Ela	0	0.03	> 0.001	0
CHS	Em	Ebb	Ela	0.01	0.03	> 0.001	0
CHS	Em	Ep	Ela	0.04	0	> 0.001	0
CHS	Ef	Ebb	Ela	0.09	0	> 0.001	0
CHS	Ef	Ebt	Ela	0	0.51	* <0.001	0.56
CHS	Ef	Ep	Ela	0.13	0	> 0.001	-0.43
CHS	Ebb	Ebt	Ela	0	0.19	* < 0.001	0.92
CHS	Ebt	Ep	Ela	0.03	0	> 0.001	0
CHS	Ebb	Ep	Ela	0.04	0	> 0.001	0

Table S7. Summary of the f^d statistics and Patterson D results using a block-jackknifing correction, for the Chalcone synthase (CHS). The test was performed with *E. lagascae* (Ela) as archaic population (P3). P1 and P2 code populations are: Ebt (*E. bastatenanum*), Ebb (*E. baeticum*), Ep (*E. popovii*), Em (*E. mediohis-panicum*), En (*E. nevadense*), and Ef (*E. fitzii*). The significatnt p-values are depicted by an asterisk.

Gene	P1	P2	P3	f^d P1,P3	f^d P2,P3	p-value	D
CHI	En	Ebb	Ela	0.25	0	* < 0.001	-0.88
CHI	En	Ebt	Ela	0.22	0	* < 0.001	-0.47
CHI	En	Ep	Ela	0.30	0	* < 0.001	-0.87
CHI	Em	Ebt	Ela	0.10	0	* < 0.001	-0.51
CHI	Em	Ebb	Ela	0.27	0	* < 0.001	-0.93
CHI	Em	Ep	Ela	0.33	0	* < 0.001	-0.92
CHI	Ef	Ebb	Ela	0.17	0	* < 0.001	-0.97
CHI	Ef	Ebt	Ela	0.16	0	* < 0.001	1
CHI	Ef	Ep	Ela	0.23	0	* < 0.001	1
CHI	Ebb	Ebt	Ela	0	0.19	* < 0.001	0.92
CHI	Ebt	Ep	Ela	0.24	0	* < 0.001	-0.93
CHI	Ebb	Ep	Ela	0.04	0	> 0.001	-0.29

Table S8. Summary of the f^d statistics and Patterson D results using a block-jackknifing correction, for the Chalcone flavone isomerase (CHI). The test was performed with *E. lagascae* (Ela) as archaic population (P3). P1 and P2 code populations are: Ebt (*E. bastatenanum*), Ebb (*E. baeticum*), Ep (*E. popovii*), Em (*E. mediohispanicum*), En (*E. nevadense*), and Ef (*E. fitzii*). The significatnt p-values are depicted by an asterisk.

Gene	P1	P2	P3	f^d P1,P3	f^d P2,P3	p-value	D
F3H	En	Ebb	Ela	0.11	0	> 0.001	-0.87
F3H	En	Ebt	Ela	0	0.40	> 0.001	0.79
F3H	En	Ep	Ela	0.10	0	> 0.001	0.29
F3H	Em	Ebt	Ela	0	0.03	> 0.001	0.15
F3H	Em	Ebb	Ela	0.75	0	> 0.001	-0.58
F3H	Em	Ep	Ela	0.10	0	> 0.001	0.29
F3H	Ef	Ebb	Ela	0.18	0	> 0.001	-0.66
F3H	Ef	Ebt	Ela	0.01	0	> 0.001	-0.14
F3H	Ef	Ep	Ela	0	0	> 0.001	-
F3H	Ebb	Ebt	Ela	0	0.20	> 0.001	0.53
F3H	Ebt	Ep	Ela	0	0	> 0.001	-
F3H	Ebb	Ep	Ela	0	0	> 0.001	-

Table S9. Summary of the f^d statistics and Patterson D results using a block-jackknifing correction, for the Flavanone 3-hydroxylase (F3H). The test was performed with *E. lagascae* (Ela) as archaic population (P3). P1 and P2 code populations are: Ebt (*E. bastatenanum*), Ebb (*E. baeticum*), Ep (*E. popovii*), Em (*E. mediohispanicum*), En (*E. nevadense*), and Ef (*E. fitzii*).

Gene	P1	P2	P3	f^d P1,P3	f^d P2,P3	p-value	D
DFR	En	Ebb	Ela	0	0	> 0.001	-
DFR	En	Ebt	Ela	0.01	0	> 0.001	0
DFR	En	Ep	Ela	0	0	> 0.001	-
DFR	Em	Ebt	Ela	0.005	0	> 0.001	-
DFR	Em	Ebb	Ela	0.005	0	> 0.001	-
DFR	Em	Ep	Ela	0	0	> 0.001	-
DFR	Ef	Ebb	Ela	0	0	> 0.001	-
DFR	Ef	Ebt	Ela	0	0	> 0.001	-
DFR	Ef	Ep	Ela	0	0	> 0.001	-
DFR	Ebb	Ebt	Ela	0	0	> 0.001	-
DFR	Ebt	Ep	Ela	0	0	> 0.001	-
DFR	Ebb	Ep	Ela	0	0	> 0.001	-

Table S10. Summary of the f^d statistics and Patterson D results using a block-jackknifing correction, for the Dyhydroflavonol 4-reductase (DFR). The test was performed with *E. lagascae* (Ela) as archaic population (P3). P1 and P2 code populations are: Ebt (*E. bastatenanum*), Ebb (*E. baeticum*), Ep (*E. popovii*), Em (*E. mediohispanicum*), En (*E. nevadense*), and Ef (*E. fitzii*).

Gene	P1	P2	P3	f^d P1,P3	f^d P2,P3	p-value	D
ANS	En	Ebb	Ela	0	0.12	> 0.001	-
ANS	En	Ebt	Ela	0	0	> 0.001	-
ANS	En	Ep	Ela	0	0	> 0.001	-
ANS	Em	Ebt	Ela	0	0	> 0.001	-
ANS	Em	Ebb	Ela	0	0	> 0.001	-
ANS	Em	Ep	Ela	0	0	> 0.001	-
ANS	Ef	Ebb	Ela	0	0	> 0.001	-
ANS	Ef	Ebt	Ela	0	0	> 0.001	-
ANS	Ef	Ep	Ela	0	0	> 0.001	-
ANS	Ebb	Ebt	Ela	0	0	> 0.001	-
ANS	Ebt	Ep	Ela	0	0.6	> 0.001	-
ANS	Ebb	Ep	Ela	0	0	> 0.001	-

Table S11. Summary of the f^d statistics and Patterson D results using a block-jackknifing correction, for the Anthocyanidin synthase (ANS). The test was performed with *E. lagascae* (Ela) as archaic population (P3). P1 and P2 code populations are: Ebt (*E. bastatenanum*), Ebb (*E. baeticum*), Ep (*E. popovii*), Em (*E. mediohispanicum*), En (*E. nevadense*), and Ef (*E. fitzii*).

Gene	P1	P2	P3	f^d P1,P3	f^d P2,P3	p-value	D
UF3GT	En	Ebb	Ela	0	0	> 0.001	0
UF3GT	En	Ebt	Ela	0.16	0.22	> 0.001	0.31
UF3GT	En	Ep	Ela	0.28	0.02	> 0.001	0.12
UF3GT	Em	Ebt	Ela	0	0.40	> 0.001	0.81
UF3GT	Em	Ebb	Ela	0	0.2	> 0.001	-
UF3GT	Em	Ep	Ela	0	0.3	> 0.001	0.33
UF3GT	Ef	Ebb	Ela	0	0.13	> 0.001	-
UF3GT	Ef	Ebt	Ela	0	0.13	> 0.001	-
UF3GT	Ef	Ep	Ela	0	0.22	-	-
UF3GT	Ebb	Ebt	Ela	0	0	-	-
UF3GT	Ebt	Ep	Ela	0	0	> 0.001	0.52
UF3GT	Ebb	Ep	Ela	0	0	> 0.001	0.20

Table S12. Summary of the f^d statistics and Patterson D results using a block-jackknifing correction, for the UDP-glucose (UF3GT). The test was performed with *E. lagascae* (Ela) as archaic population (P3). P1 and P2 code populations are: Ebt (*E. bastatenanum*), Ebb (*E. baeticum*), Ep (*E. popovii*), Em (*E. mediohis-panicum*), En (*E. nevadense*), and Ef (*E. fitzii*).

General Discussion

Throughout this Ph.D. thesis, we have explored and characterized the hybridization scenario for some Southern Iberian *Erysimum* using a broad approach and providing new resources and molecular tools. First, we studied to which degree ITS sequences are homogenized by concerted evolution (**Chapter I**). We found that these nuclear sequences were not thoroughly homogenized in this group of species, suggesting that hybridization is frequent and has occurred relatively recently and might even be ongoing. Second, we assembled the chloroplast genomes of these species from RNA-Seq data, demonstrating the reliability of this strategy to provide complete cpDNA genomes (**Chapter II**). Then, we used this approach to develop genomic resources for the study of evolution in *Erysimum* and other taxa assembling several cpDNA genomes. *de novo* and annotating them In **Chapter III**, we used these genomes to reconstruct a time-calibrated phylogeny. Finally, we combined the cpDNA and nuclear genomic and transcriptomic data with ecological and cytogenetic information to analyze the general hybridization scenario in *Erysimum* spp. (**Chapter IV**). Our results indicated that introgression is ubiquitous for all the species studied here, including even diploid species. These multiple layers of evidence strongly support the hypothesis that recurrent hybridization is a major driver of diversification in *Erysimum* spp..

Hybridization and introgression patterns in *Erysimum* species

In this Ph.D. thesis, we have analyzed a group of species with a possible history of hybridization, applying different genetic methodologies and accounting for other phenomena such as ILS. Our results supported a scenario of both recent and ancient hybridization for the *Erysimum* species studied here. The phylogenetic incongruence between the evolutionary scenarios reconstructed from nuclear and plastidial markers is the first indication of hybridization and introgression. Nuclear sequences such as ITS have usually been used to infer recent gene flow (Lexer et al., 2005; Burgarella et al., 2009; Pinheiro et al., 2010). These sequences are assumed to be under concerted evolution that reduces nucleotide and haplotypic diversity (Álvarez and Wendel, 2003).

However, we have encountered a lack of homogenization in ITS markers in *Erysimum* spp. , which varies depending on ploidy (**Chapter I**). A reduction in sequence homogenization is expected after an allopolyploidization event since two different genomes are combined into a single one, with a subsequent and sudden increase of intragenomic diversity. Later, concerted evolution homogenizes variation at ITS sequences in a process that depends on both time since hybridization and nucleotidic differences among the initial ITS sequences. However, additional events of introgression could maintain a relatively high ITS variability, in a cycle of accretion by hybridization and reduction by concerted evolution that will only come to an end after complete reproductive isolation between the hybrid and its parentals. In the case of diploid (2X) species, such as *E. nevadense*, *E. fitzii*, and some populations of *E. mediohispanicum*, our findings of intragenomic variation suggest that interspecific introgression events have influenced even the configuration of diploid genomes. Therefore, our results may be interpreted as the product of relatively recent hybridization, in which concerted evolution has not had enough time to erase sequence polymorphism after hybridization.

We have also studied the complete sequence of the chloroplast genomes of several species, to try and identify signals of introgression. Comparison of nuclear and plastid phylogenies has been previously used to describe ancient episodes of introgression (Palmé et al., 2004; Heuertz et al., 2006; Palma-Silva et al., 2011). We found clear cytonuclear discordances that support an scenario of ancient hybridization (see **Figure 1**, **Chapter IV**). Moreover, we found indications of ghost introgressions, i.e., signatures of ancient admixture with unsampled or now-extinct lineages (**Chapter IV**). Presumably, after hybridization, some of the parental hybrid species must have either gone (locally) extinct or at least become very rare, but the introgression signature persists in the genomes of extant species. Thus, this Ph. D. thesis's results highlight the importance of incorporating hybridization events in the analysis of plant evolution. Moreover, we have showed that hybridization should not be treated as a once-only phenomenon, at least in *Erysimum* spp., and that both recent and ancient hybridization events, including ghost

introgression, inform current genetic diversity. The evolution of complete reproductive isolation barriers may take numerous generations and hybridization and introgression can occur recursively at different stages of speciation.

Plant evolution appears to have been strongly influenced not only by hybridization but also by polyploidization. Therefore, we also addressed the potential role of polyploidy in our system. *Erysimum* is one of Brassicaceae's genera with a wider variation in chromosome number, with species ranging from 2n = 12 to 2n = 84, and including extensive variation below the species level (Marhold & Lihová, 2006; Warwick & Al-Shehbaz, 2006). In the particular case of Iberian *Erysimum*, several authors have reported ample variation at the intraspecific level, with remarkable cases in the Northern species *E. duriaei* and for *E. popovii* in the South (Clot 1992; Blanca et al., 1992). In the species and populations we sampled, the species with purple corolla were all polyploids, except *E. lagascae*, which we had sampled outside the Baetic Mountains precisely because we assumed it would be a diploid (see **Chapter IV**). However, all the yellow species were diploids (except a tetraploid population of *E. mediohispanicum*, **Chapter IV)**. This species has been previously studied as having diploid and polyploid populations in the Iberian Peninsula, with a clear geographical division between northern (4n) and southern (2n) populations (Nieto-Feliner 2003; Muñoz-Pajares, 2013). Em71, the 4x population of *E. mediohispanicum*, is located in the North of the Granada province, coinciding with the South limit of the geographical distribution of 4n *E. mediohispanicum*. We have also found intra-specific differences in ploidy levels for *E. popovii* and *E. bastetanum*. Cytotypes of *E. bastetanum* and especially *E. popovii* are more geographically intermixed than those of *E. mediohispanicum*, with interspersed population showing 28, 40, 42, 56, and 70 chromosomes (4X to 10X; Fernández et al., 2012, 2105; Gómez et al., unpublished). It has been estimated that around 15 % of plant species show intra-specific ploidy variation (Rice et al., 2015). This variation has been described in several species (Petit et al., 1999; Mráz et al., 2012; Eidesen et al., 2013; Hülber et al., 2015; Zozomová-Lihová et al., 2015; Rejlová et al., 2019) often related to

historical demographic dynamics. When allopatric but recently diverged species coincide spatially again, they can hybridize and produce new alloploid species and introgressed populations. This dynamic has been associated with range expansions and contractions caused by glacial cycles in the Pleistocene. As a consequence, glacial refugia are often regions with high intraspecific variation in ploidy (Tribsch and Schönswetter, 2003; Schönswetter et al., 2005; Sonnleitner et al., 2010). *Erysimum* species studied here are located in a well-known glacial refugium (Médail and Diadema, 2009), which might have influenced the hybridization and introgression dynamics among these species. In light of our results, the diversification of these species may have happened in the Pleistocene, when the last glacial oscillations occurred (**Figure 1**, **Chapter III**). Thus, recurrent isolation (i.e., allopatry) and subsequent admixture after secondary contact have probably been driven by range expansion and contraction during glacial cycles. Secondary contacts might have repeatedly created hybrid zones and therefore facilitated multiple bouts of hybridization (Coyne and Orr, 2004; Harrison and Larson, 2014; Arnold, 2016), leading to the different ploidy-levels described here (Kolář et al., 2017). In fact, a hybrid zone for *E. mediohispanicum* and *E. nevadense* has been described previously (Abdelaziz 2013) and sympatry is common even under present-day conditions. These results underscore the importance of incorporating multiple individuals and populations per species when exploring polyploidization to account for intra-specific differences in ploidy level. Our findings clearly show that it is untenable to assume that all individuals from a given species have the same ploidy level. In complex hybrid scenarios, individuals from the same species may have different introgression signatures in their genomes, and a population- or individual-based approach is recommended.

Moreover, in this Ph.D. thesis, we have described signatures of hybridization and introgression in diploid species, which were not presumed to have a hybrid origin (see **Chapter I** and **Chapter IV**). An explanation of this introgression could be a hybrid origin with no changes

in the ploidy level (i.e., homoploid hybrid speciation; Hersch-Green, 2012; Schumer et al., 2014; Nieto-Feliner et al., 2017). Homoploid hybrid speciation, or at least patterns congruent with it, has been described in several plant species (Arnold, 1993; Wolfe et al., 1998; Wang et al., 2001; James and Abbott, 2005; Pan et al., 2007; Brennan et al., 2012; Gao et al., 2012; Taylor et al., 2013). Nevertheless, it remains an underexplored and controversial evolutionary phenomenon (Schumer et al., 2014). Another possible explanation is that the introgression that we found was a signature of alloploid hybrid speciation followed by diploidization. This process has been described in several systems (Tate et al., 2009; Mandáková et al., 2010; Mandáková and Lysak, 2018) resulting in chromosomal rearrangements, genome-wide reorganization, and a reduction in chromosome number (Dodsworth et al., 2016). However, this process is challenging to demonstrate, although the influence of diploidization on plant speciation has recently been vindicated (Dodsworth et al., 2016). In any case, the results of this Ph.D. thesis emphasize that even species that were not presumed to have a hybrid origin may harbor introgressed genomic regions from other species. Under these circumstances, reconstructing a reliable phylogeny for species that could hybridize might be difficult without accounting for the possibility of hybrid origins or, at least without considering the possibility of introgression in specific chromosomic segments.

The potential role of adaptive introgression in *Erysimum* spp.

We have found signatures of introgression for the anthocyanin genes in the species studied here. Specifically, considering *E. lagascae* (a 2x species with purple corolla) as a representative of the introgression source, we have found introgression signatures in *E. bastetanum* (purple corolla). Moreover, we have also found signatures of introgression in the yellow corolla species *E. mediohispanicum*, *E. nevadense*, and *E. fitzii*. These results suggest that the studied yellow-corolla species show signatures of introgression from a purple diploid species (*E. lagascae*). Therefore, these species may have appeared by the hybridization between ancestral species of

purple and yellow color. These results were concordant with the substantial introgression that we have found in diploid species (discussed above) which suggest at least a partially hybrid origin for extant diversity. Moreover, our results are also congruent with a previous study that suggested that yellow color appears in the Iberian *Erysimum* species secondarily, possibly from ancestral purple species (Gómez et al., 2015). However, to investigate the introgression of particular genes in these species thoroughly, further research is needed, ideally accounting for all the *Erysimum* species inhabiting the Iberian Peninsula. It would also be ideal to use long-read sequence approaches and chromosome phasing to recover all the gene copies in the polyploid species. Therefore, our results encourage taking ancient introgression, even from extinct lineages, into consideration when studying the evolution of potentially adaptive traits. .

We tried to account for the adaptive potential of the introgressed genes, assuming that corolla color might influence pollinator attraction. However, in light of our results, we cannot conclude that the introgression we detected confers an adaptive advantage, and further research is needed. The fact that we found that yellow species have signatures of introgression from purple species may suggest that anthocyanins and corolla color play only a minor role in pollination. Interestingly, Gómez et al. 2015 did not find a clear differentiation in the pollination niche of purple and yellow *Erysimum* species. Thus, anthocyanin production may be related to other functions, such as responses to abiotic stress, as described in other systems (Treutter, 2006; Dao et al., 2011; Khlestkina, 2013). The species studied here occur at high altitudes, where these genes may confer adaptive advantages by providing photoprotection against UV-light. Examples of introgression of genes that confer adaptation to abiotic factors have been reported in several plant species (Rieseberg, 1998; Martin et al., 2006; Whitney et al., 2006; Soolanayakanahally et al., 2009; Whitney et al., 2010; Arnold et al., 2016; Khodwekar and Gailing, 2017; Chhatre et al., 2018). Notably, our results are in line with a study of *Cupressus* spp. in which gene introgression in the anthocyanin pathway was linked to adaptation to high altitude habitats (Ma et al., 2019).

Our results indicate the need to study the role of introgression in the adaptation of species to climate and other abiotic selective factors, particularly in the context of rapid climate change. Our results also encourage the study of adaptive introgression in wild polyploid species, specifically in glacial refugia, which is presently understudied. The possibility that adaptive introgression helping plants to cope with climate fluctuations can be associated with the success of polyploid species deserves attention.

Future perspectives

This Ph. D. thesis's results provide insights into hybridization dynamics, which may have had a strong influence on the speciation in the genus *Erysimum*. However, many questions remain unanswered. For instance, addressing hybridization considering all the *Erysimum* species of the Iberian Peninsula, and even including species of the Western Mediterranean, would allow us to have a more comprehensive knowledge of the role played by this process in the evolutionary history of this genus. Here, we have used nuclear markers, whole transcriptome, and chloroplast genomes. Presently, whole-genome sequencing and long-read sequencing technologies (third-generation sequencing) are taking over as sequencing techniques, and appear to be particularly suited in the study of plant genomes (Dumschott et al., 2020). Thus, further research using whole genomes and long reads would allow for more detailed and in-depth analyses of hybridization patterns. Also, haplotype phasing seems to be a promising approach to study hybridization and polyploidization, as it allows assembling all chromosomes and recovering all the intragenomic variation (Browning and Browning, 2011). Furthermore, the fast development of novel molecular tools might enable studies in the near future that consider the influence of extinct species as introgression donors and reconstruct their ancestral characters. Dating of the hybridization and polyploidization events would be ideal for disentangling the reticulated history of these species. Also, studies with a more ecological perspective, addressing prezygotic and postzygotic reproductive barriers across populations with different ploidy levels, would help us understand

whether the different cytotypes are reproductively isolated. At the same time, the introgression signatures that we found in diploid species require further exploration to understand the ecological factors that may favor the formation of a diploid or a polyploid genome.

We also believe that not only the results of this Ph.D. but also the scientific tools described here can benefit other researchers and be extrapolated to other study systems. We encourage the use of several complementary methods to detect hybridization in the presence of ILS. For instance, using chloroplast genomes combined with nuclear data was proven helpful to disentangle complex hybridization scenarios with a strong signature of introgression. Moreover, phylogenies, phylogenetic networks, and introgression tests are promising approaches; however, all of them have specific drawbacks, and using them in combination to extract patterns that are method-independent is recommended. Finally, we emphasize that when studying the evolution of organismal groups where hybridization is not deemed necessary *a priori*, an in-depth knowledge of the genomic of the specie studied is advisable.

General Conclusions

- The genomic and transcriptomic data presented in this thesis, currently the only derived from flower tissues, constitute a valuable genetic resource for the genus *Erysimum*. They also contribute to the existing resources for Brassicaceae and could be used as reliable reference sequences for comparative and phylotranscriptomic studies.
- Ploidy levels vary widely among the *Erysimum* species studied. In some instances, cytotypes differ even among populations of the same species (i.e., intra-specific ploidy variation). This indicates dynamic genome re-configuration processes that might have been facilitated by historical demographic processes, such as Pleistocene glacial cycles.
- Both polyploid and diploid species present strong signatures of introgression. However, these introgression signals vary across populations of the same species, which seriously complicates the reconstruction of a reliable phylogeny.
- We have shown that these species present a deficit of sequence homogenization for ITS regions, suggesting that concerted evolution had not enough time to fully homogenize ribosomal DNA arrays, this lack of homogenization is more patent in polyploid species.
- The deficit of homogenization in ITS sequences is congruent with rampant hybridization and introgression events among different lineages.
- We have demonstrated that whole-chloroplast genomes can be reliably assembled from total RNA-Seq data, resulting in genomes that are highly similar to those assembled from genomic libraries in terms of the overall structure, size, and sequence.

- We detected conspicuous cytonuclear discordance in *Erysimum* spp., which suggests that hybridization is a major evolutionary feature in this group.
- Hybridization appears to have been a recurrent phenomenon in the presence of ILS.
- We detected a signature of introgression in some anthocyanin genes (CHS/CHI). The adaptive value of the introgression of these genes is unclear as of yet, but it may confer adaptive advantages in high mountain habitats and/or be associated with changes in pollinator attraction between purple and yellow corollas.

References

Abdelaziz M. (2013). How species are evolutionarily maintained: Pollinator mediated divergence and hybridization in *Erysimum mediohispanicum* and *E. nevadense*. PhD Thesis. Universidad de Granada.

Álvarez, I., & Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, 29 (3), 417-434.

Arnold, M. L. (1993). *Iris nelsonii* (Iridaceae): origin and genetic composition of a homoploid hybrid species. *American Journal of Botany*, 80 (5), 577-583.

Arnold, M. L. (2016). Anderson's and Stebbins' prophecy comes true: genetic exchange in fluctuating environments. *Systematic Botany*, 41 (1), 4-16.

Arnold, B. J., Lahner, B., DaCosta, J. M., Weisman, C. M., Hollister, J. D., Salt, D. E., ... & Yant, L. (2016). Borrowed alleles and convergence in serpentine adaptation. *Proceedings of the National Academy of Sciences*, 113 (29), 8320-8325.

Blanca G, Morales C, Ruiz Rejón M. 1992. El género *Erysimum* L. (Cruciferae) en Andalucía (España). *Anales del Jardín Botánico de Madrid* 49: 201-214.

Burgarella, C., Lorenzo, Z., Jabbour-Zahab, R., Lumaret, R., Guichoux, E., Petit, R. J., ... & Gil, L. (2009). Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). *Heredity*, 102 (5), 442-452.

Brennan, A. C., Barker, D., Hiscock, S. J., & Abbott, R. J. (2012). Molecular genetic and quantitative trait divergence associated with recent homoploid hybrid speciation: a study of *Senecio squalidus* (Asteraceae). *Heredity*, 108 (2), 87-95.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12 (10), 703-714.

Chhatre, V. E., Evans, L. M., DiFazio, S. P., & Keller, S. R. (2018). Adaptive introgression andmaintenance of a trispecies hybrid complex in range-edge populations of *Populus*. *Molecular Ecology*, 27 (23), 48204838.

Clot B. (1992). Caryosystématique de quelques *Erysimum* L. dans le nord de la Peninsule Ibérique. *Anales del Jardín Botánico de Madrid*, 49: 215-229

Coyne, J. A., & Orr, H. A. (2004). Speciation (Sinauer Associates), *Sunderland*, MA, 276, 281.

Dao, T. T. H., Linthorst, H. J. M., & Verpoorte, R. (2011). Chalcone synthase and its functions in plant resistance. *Phytochemistry Reviews*, 10 (3), 397.

Dodsworth, S., Chase, M. W., & Leitch, A. R. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?. *Botanical Journal of the Linnean Society*, 180 (1), 1-5.

Dumschott, K., Schmidt, M. H. W., Chawla, H. S., Snowdon, R., & Usadel, B. (2020). Oxford Nanopore Sequencing: New opportunities for plant genomics?. *Journal of Experimental Botany*.

Eidesen, P. B., Müller, E., Lettner, C., Alsos, I. G., Bender, M., Kristiansen, M., ... & Verweij, K. F. (2013). Tetraploids do not form cushions: association of ploidy level, growth form and ecology in the High Arctic *Saxifraga oppositifolia* L. s. lat.(Saxifragaceae) in Svalbard. *Polar Research*, 32 (1), 20071.

Fernández JD, Nieto B, Bosch J, Gómez JM. (2012). Pollen limitation in a narrow endemic plant: geographical variation and driving factors. *Oecologia*, 170: 421-431.

Fernández JD, Lorite J, Bosch J, Gómez JM. (2015). Variation in the reproductive success of a narrow endemic plant: Effects of geographical distribution, abiotic conditions and pollinator community composition. *Basic and Applied Ecology*, 16: 375-385.

Gao, J. I. E., Wang, B., Mao, J. F., Ingvarsson, P., Zeng, Q. Y., & Wang, X. R. (2012). Demography and speciation history of the homoploid hybrid pine *Pinus densata* on the Tibetan Plateau. *Molecular Ecology*, 21 (19), 4811-4827.

Gómez, J. M., Perfectti, F., & Lorite, J. (2015). The role of pollinators in floral diversification in a clade of generalist flowers. *Evolution*, 69 (4), 863-878.

Hahn, M. W., & Hibbins, M. S. (2019). A three-sample test for introgression. *Molecular Biology and Evolution*, 36 (12), 2878-2882.

Harrison, R. G., & Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105 (S1), 795-809.

Hersch-Green, E. I. (2012). Polyploidy in Indian paintbrush (Castilleja; Orobanchaceae) species shapes but does not prevent gene flow across species boundaries. *American Journal of Botany*, 99 (10), 1680-1690.

Heuertz, M., Carnevale, S., Fineschi, S., Sebastiani, F., Hausman, J. F., Paule, L., & Vendramin, G. G. (2006). Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp. (Oleaceae): roles of hybridization and life history traits. *Molecular Ecology*, 15 (8), 2131-2140.

Hibbins, M. S., Gibson, M. J., & Hahn, M. W. (2020). Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. *BioRxiv*.

Hülber, K., Sonnleitner, M., Suda, J., Krejčíková, J., Schönswetter, P., Schneeweiss, G. M., & Winkler, M. (2015). Ecological differentiation, lack of hybrids involving diploids, and asymmetric gene flow between polyploids in narrow contact zones of *Senecio carniolicus* (syn. *Jacobaea carniolica*, Asteraceae). *Ecology and Evolution*, 5 (6), 1224-1234.

James, J. K., & Abbott, R. J. (2005). Recent, allopatric, homoploid hybrid speciation: the origin of *Senecio squalidus* (Asteraceae) in the British Isles from a hybrid zone on Mount Etna, Sicily. *Evolution*, 59 (12), 2533-2547.

Khlestkina, E. K., & Shoeva, O. Y. (2014). Intron loss in the chalcone-flavanone isomerase gene of rye. *Molecular Breeding*, 33 (4), 953-959.

Khodwekar, S., & Gailing, O. (2017). Evidence for environment-dependent introgression of adaptive genes between two red oak species with different drought adaptations. *American Journal of Botany*, 104 (7), 1088-1098.

Kolář, F., Čertner, M., Suda, J., Schönswetter, P., & Husband, B. C. (2017). Mixed-ploidy species: progress and opportunities in polyploid research. *Trends in Plant Science*, 22 (12), 1041-1055.

Lexer, C., Kremer, A., & Petit, R. J. (2006). Comment: shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Molecular Ecology*, 15 (7), 2007-2012.

Ma, Y., Wang, J., Hu, Q., Li, J., Sun, Y., Zhang, L., ... & Mao, K. (2019). Ancient introgression drives adaptation to cooler and drier mountain habitats in a cypress species complex. *Communications Biology*, 2 (1), 1-12.

Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., & Lysak, M. A. (2010). Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *The Plant Cell*, 22 (7), 2277-2290.

Mandáková, T., & Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, 42, 55-65.

Marhold K, Lihová J. (2006). Polyploidy, hybridization and reticulate evolution: Lessons from the Brassicaceae. *Plant Systematics and Evolution*, 259: 143–174.

Martin, N. H., Bouck, A. C., & Arnold, M. L. (2006). Detecting adaptive trait introgression between *Iris fulva* and *I. brevicauli*s in highly selective field conditions. *Genetics*, 172 (4), 2481-2489.

Médail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography*, 36 (7), 1333-1345.

Mráz, P., Španiel, S., Keller, A., Bowmann, G., Farkas, A., Šingliarová, B., ... & Müller-Schärer, H. (2012). Anthropogenic disturbance as a driver of microspatial and microhabitat segregation of cytotypes of *Centaurea stoebe* and cytotype interactions in secondary contact zones. *Annals of Botany*, 110 (3), 615-627.

Muñoz-Pajares, A. J. (2013). *Erysimum mediohispanicum* at the evolutionary crossroad: phylogrography, phenotype, and pollinators (Doctoral dissertation, Universidad de Granada).

Nieto-Feliner, G., Álvarez, I., Fuertes-Aguilar, J., Heuertz, M., Marques, I., Moharrek, F., ... & Villa-Machío, I. (2017). Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity*, 118 (6), 513

Palma Silva, C., Wendt, T., Pinheiro, F., Barbará, T., Fay, M. F., Cozzolino, S., & Lexer, C. (2011). Sympatric bromeliad species (*Pitcairnia* spp.) facilitate tests of mechanisms involved in species cohesion and reproductive isolation in Neotropical inselbergs. *Molecular Ecology*, 20 (15), 3185-3201.

Palme, A. E., Su, Q., Palsson, S., & Lascoux, M. (2004). Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among *Betula pendula*, *B. pubescens* and *B. nana*. *Molecular Ecology*, 13 (1), 167-178.

Pan, J., Zhang, D., & Sang, T. (2007). Molecular phylogenetic evidence for the origin of a diploid hybrid of *Paeonia* (Paeoniaceae). *American Journal of Botany*, 94 (3), 400-408.

Petit, C., Bretagnolle, F., & Felber, F. (1999). Evolutionary consequences of diploid–polyploid hybrid zones in wild species. *Trends in Ecology & Evolution*, 14 (8), 306-311.

Pinheiro, F., de Barros, F., Palma-silva, Meyer, D., Fay, M. F., Suzuki, R. M., ... & Cozzolino, S. (2010). Hybridization and introgression across different ploidy levels in the Neotropical orchids *Epidendrum fulgens* and *E. puniceoluteum* (Orchidaceae). *Molecular Ecology*, 19 (18), 3981-3994.

Rejlová, L., Chrtek, J., Trávníček, P., Lučanová, M., Vit, P., & Urfus, T. (2019). Polyploid evolution: The ultimate way to grasp the nettle. *PloS One*, 14 (7).

Rieseberg, L. H., & Carney, S. E. (1998). Plant hybridization. *The New Phytologist*, 140 (4), 599-624.

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N. M., Salman-Minkov, A., ... & Mayrose, I. (2015). The Chromosome Counts Database (CCDB)–a community resource of plant chromosome numbers. *New Phytologist*, 206 (1), 19-26.

Schönswetter, P., Stehlik, I., Holderegger, R., & Tribsch, A. (2005). Molecular evidence for glacial refugia of mountain plants in the European Alps. *Molecular Ecology*, 14 (11), 3547-3555.

Schumer, M., Rosenthal, G. G., & Andolfatto, P. (2014). How common is homoploid hybrid speciation?. *Evolution*, 68 (6), 1553-1560.

Soolanayakanahally, R. Y., Guy, R. D., Silim, S. N., Drewes, E. C., & Schroeder, W. R. (2009). Enhanced assimilation rate and water use efficiency with latitude through increased photosynthetic capacity and internal conductance in balsam poplar (*Populus balsamifer*a L.). *Plant, Cell & Environment*, 32 (12), 1821-1832.

Soltis, D. E., Buggs, R. J., Doyle, J. J., & Soltis, P. S. (2010). What we still don't know about polyploidy. *Taxon*, 59 (5), 1387-1403.

Sonnleitner, M., Flatscher, R., Escobar García, P., Rauchová, J., Suda, J., Schneeweiss, G. M., ... & Schönswetter, P. (2010). Distribution and habitat segregation on different spatial scales among diploid, tetraploid and hexaploid cytotypes of *Senecio carniolicus* (Asteraceae) in the Eastern Alps. *Annals of Botany*, 106 (6), 967-977.

Taylor, S. J., Rojas, L. D., Ho, S. W., & Martin, N. H. (2013). Genomic collinearity and the genetic architecture of floral differences between the homoploid hybrid species *Iris nelsonii* and *one* of its progenitors, *Iris hexagona*. *Heredity*, 110(1), 63-70.

Tate, J. A., Joshi, P., Soltis, K. A., Soltis, P. S., & Soltis, D. E. (2009). On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Biology*, 9 (1), 80.

Treutter, D. (2006). Significance of flavonoids in plant resistance: a review. *Environmental Chemistry Letters*, 4 (3), 147.

Tribsch, A., & Schönswetter, P. (2003). Patterns of endemism and comparative phylogeography confirm palaeo-environmental evidence for Pleistocene refugia in the Eastern Alps. *Taxon*, 52 (3), 477-497.

Wang, X. R., Szmidt, A. E., & Savolainen, O. (2001). Genetic composition and diploid hybrid speciation of a high mountain pine, *Pinus densat*a, native to the Tibetan Plateau. *Genetics*, 159 (1), 337-346.

Warwick SI, Al-Shehbaz IA. (2006). Brassicaceae: Chromosome number index and database on CD-Rom. *Plant Systematics and Evolution*, 259: 237–248.

Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2006). Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. *The American Naturalist*, 167 (6), 794-807.

Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2010). Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytologist*, 187 (1), 230-239.

Wolfe, A. D., Xiang, Q. Y., & Kephart, S. R. (1998). Diploid hybrid speciation in *Penstemon* (Scrophulariaceae). *Proceedings of the National Academy of Sciences*, 95 (9), 5112-5115.

Zozomová-Lihová, J., Malánová-Krásná, I., Vít, P., Urfus, T., Senko, D., Svitok, M., ... & Marhold, K. (2015). Cytotype distribution patterns, ecological differentiation, and genetic structure in a diploid–tetraploid contact zone of *Cardamine amara*. *American Journal of Botany*, 102 (8), 1380-1395.