Crossability barriers in *Prunus*: the role of modifiers in the regulation of the gametophytic self-incompatibility system

TESIS DOCTORAL presentada por

Juan Vicente Muñoz Sanz

Dirigida por

Dr. Carlos Romero Salvador

Dra. Marisa Badenes Catalá

Valencia, julio 2016

A Vanesa y a Pau

“Si no vas a por todo, ¿a qué vas?”
(Joe Namath)

AGRADECIMIENTOS

En primer lugar quería agradecer a mis directores de tesis, Marisa Badenes y Carlos Romero, toda la confianza puesta en mí para desarrollar este trabajo de investigación. Muchas gracias por vuestro tiempo, energía y conocimientos. Y especialmente a Carlos, que a pesar de las circunstancias siempre ha estado ahí para ayudarme y apoyarme; pero sobre todo por saber aconsejarme siempre de la mejor manera. Muchas gracias por tu amistad.

Desde luego quería dar las gracias a todos los compañeros del departamento de frutales del IVIA con los que he convivido y disfrutado durante estos años. No voy a extenderme en cada caso, pero todos en algún momento me han ayudado o aconsejado, con lo cual, este trabajo no habría sido posible. Gracias a Ana Catalá, Ana Conejero, Alba Lloret, Carmen Leida, Enzo, Fran Gil, Gabino Ríos, Inma López, José Martínez, Manuel Blasco, Mar Naval, Mati González y Pepe Palanca. Mención especial merece Elena Zuriaga, sin sus conocimientos en bioinformática esta tesis no habría llegado a buen puerto. De verdad, gracias a todos por alegrarme el camino.

No puedo olvidarme de los grupos que me acogieron durante mis estancias en México y USA. Gracias a Felipe Cruz García y a todos los compañeros de laboratorio por hacerme sentir uno más en tan poco tiempo, pero especialmente por enseñarme un pedacito de México. Gracias a Alejandra Ávila, Claudia, Edgar, Gustavo, Javier Andrés Juárez, Jorge, Lili García, Lilia, Yuridia Cruz y sobre todo a Carlos Bravo (espero que todo te vaya bien amigo). Como no a Bruce McClure y Alejandro Tovar, que me facilitaron mucho las cosas durante mi vivencia en el *midwest* y con los que tuve la suerte y el privilegio de trabajar. También agradecer a un grupo de personas que no voy a nombrar pero que, a pesar de no tener relación alguna con la tesis, me acogieron y ayudaron en sendas aventurillas. Si alguna vez leen esto, estoy convencido de que se sentirán aludidos/as.

También quería agradecer a Hugo Merle, José Blanca, Javier Forment y Peio Ziarsolo su colaboración y conocimientos relacionados con la botánica y la bioinformática. Y a Belén Picó por su inestimable ayuda como tutora.

Como no mencionar a los masters del universo. Gracias a Laura Campos, Cecilia Primo, Cristina Codes, Félix Martínez, Patricia Ballester y Rafa Aparicio. Porque sigamos juntándonos para contar las penas que pasamos gracias a la investigación.

Agradecer al Instituto Valenciano de Investigaciones Agrarias (IVIA) por haberme permitido desarrollar la tesis. También agradecer al antiguo Ministerio de Ciencia e Innovación (actual secretaría dentro del Ministerio de Economía y Competitividad) por haber financiado mi formación predoctoral (beca FPI), así como el proyecto al cual estaba adscrito. Y por extensión a todos los españoles que durante este periodo tan complicado han contribuido, mediante el pago de sus impuestos, a que yo pueda hacer realidad uno de los proyectos personales más importantes de mi vida. Espero haberles correspondido y espero poder seguir haciéndolo en el futuro.

Por otra parte, no haber formado parte intelectual de este trabajo no quiere decir que su contribución no haya sido de enorme ayuda para mí. El mejor ejemplo lo forma mi familia, incluyendo la "adquirida" de manera oficial en estos años. A todos agradecerles su inestimable ayuda, pero sin duda quienes más deben sentirse aludidos son mis padres Juan y Milagros, y mi hermano Dani. Todavía hoy, aunque ahora cada vez menos, me cuesta entender cómo se puede dar tanto sin esperar nada a cambio (gracias por los valores que me habéis inculcado, los cuales creo que se reflejan en esta tesis). También quería agradecer a mi abuelo (como otros/as que me han tenido que dejar) el haberme enseñado el otro sentido de la herencia genética. En este apartado también quería dar las gracias a la familia que afortunadamente he podido elegir. A mis amigos de toda la vida y también a los de la universidad, espero seguir contando con vuestro apoyo y amistad durante muchos más años.

Y en este momento llego a donde más ganas tenía de hacerlo. Aunque se trate de una personita que en la parte final de la tesis me ha absorbido buena parte de tiempo y energía vital, la verdad es que mediante el uso de fonemas basados en vocablos sencillos como *aahhhh* o *eehhhh* me ha convencido de mi nueva gran responsabilidad. Si algún día coges este tomo y te da por echar una ojeada (e imagino que solo leerás esta sección), sólo te diré que después de haber leído, discutido y pensado sobre ciencia "*todo lo que puedas imaginar es real*" (Pablo Picasso). Muchas gracias Pau por haber entrado en mi vida. Te quiero.

Siempre me ha gustado dejarme lo mejor para el final. Siempre he tenido claro que nunca habría estudiado una carrera de no haber sido por ti. Siempre he tenido claro que nunca habría alcanzado la confianza necesaria para llegar hasta aquí de no ser por tu ayuda. Siempre me has apoyado en los proyectos más importantes. Siempre has estado ahí. Hay cosas que necesito que permanezcan perpetuas bajo un cambio casi constante de forma o estilo. Espero que todo siga fluyendo. Mucho ánimo con lo que te queda, aunque no lo parezca, hay luz al final del túnel. Gracias Vanesa.

Puedo volver, puedo callar, puedo forzar la realidad, puedo doler, puedo arrasar, puedo sentir que no doy más.

Puedo escurrir, puedo pasar, puedo fingir que me da igual, puedo incidir, puedo escapar, puedo partirme y negociar la otra mitad.

Puedo romper, puedo olvidar, puede comerme la ansiedad, puedo salir, puedo girar, puedo ser fácil de engañar.

Puedo joder, puedo encantar, puedo llamarte sin hablar, puedo vencer, puedo palmar,
PUEDO SABER QUE SIN VOSOTROS DUELE MÁS.

"Sálvese quien pueda" (Vetusta Morla)

ABSTRACT

Self-incompatibility (SI) comprises a compendium of molecular intraspecific barriers, under the control of the *S*-locus, which enhances outcrossing and prevents inbreeding. Solanaceae, Plantaginaceae and Rosaceae exhibit the Gametophytic SI (GSI) type where specific recognition is controlled by S-RNases and *S*-locus F-box (SFB) proteins as the female and male *S*-determinants, respectively. On the other hand, unlinked *S*-locus genes known as modifier factors or modifier genes are also completely necessary for the mechanism to function. The GSI system seems to be basically preserved in *Prunus* but striking differences with Solanaceae and other Rosaceae have also been observed. On the basis of this background, this thesis is focused on the identification and characterization of modifier genes involved in *Prunus* GSI to improve our understanding of the underlying mechanism.

Previous works in apricot (*Prunus armeniaca* L.) showed that an *S*-locus unlinked mutation expressed in pollen and located at the distal end of chr. 3 (*M*-locus) confers self-compatibility (SC) in the cv. 'Canino'. In this work, another self-compatible apricot cultivar, named 'Katy', was molecular and genetically analyzed. Similarly, an *S*-locus unlinked pollen-part mutation was found to cause the loss of self-incompatible response in 'Katy'. A mapping strategy based on segregation distorted loci mapped 'Katy' mutation (referred as *m*-mutation) at the distal end of chr. 3, in a region overlapping with that identified for 'Canino' *M*-locus. A new screening was carried out to identify additional self-compatible mutants in apricot cultivar/accessions from germplasm banks. Through *S*-genotyping, three uncategorized *S*-alleles were recovered and two new mutations putatively conferring SC by affecting the male *S*-determinant *SFB* were detected. Additionally, *M*-genotyping showed that the same mutated *m*-haplotype was shared by 'Canino' and 'Katy', but also by 17 cultivars more from North-America and Western-Europe. A widely distributed haplotype $M_{1\text{-}0}$ was proposed as the putative *m*-haplotype ancestor suggesting that it arose much later in time than S_C-allele, a mutation in the *S*-locus also conferring SC in apricot.

In order to identify this mutation, an integrative genetic, genomic and transcriptomic approach based on NGS data from 'Canino', 'Katy' and the self-incompatible apricot cultivar 'Goldrich' was carried out. This approach led to identify a unique polymorphism able to explain the self-compatible phenotype, a *FaSt* insertion type of 358-bp in coupling with the *m*-haplotype within a gene encoding a disulfide bond A-like

oxidoreductase (named *PaMDOr*). *PaMDOr* was found to be differentially over expressed in mature anthers and the *FaSt* insertion is predicted to produce a truncated protein. These two findings also support *PaMDOr* as the pollen-part mutated modifier conferring SC in apricot.

Furthermore, phylogenetic analysis suggest *PaMDOr* as a putative paralogue of its contiguous gene (named *PaM*-8), that emerged after the split of the Rosaceae and Solanaceae and which function became essential for the proper functioning of the GSI system in *Prunus*. Aimed to shed light on the differences and similarities between the *S*-RNase-based GSI systems in Rosaceae and Solanaceae, orthology relationships were analyzed for modifier factors. Putative orthologues were found for *NaTrxh*, *SBP1* and *MdABCF* in *Prunus* but a more complex evolutionary pattern was detected for *120K*, *NaStEP* and *NaPCCP*. Thus, in spite of the differences, it can be hypothesized that part of the GSI modifier factors are shared by both families.

As a whole, the multidisciplinary strategy developed in this thesis has allowed us to identify a novel modifier factor (*PaMDOr*) essential for the self-incompatible response in *Prunus* as the most significant contribution. In addition, new sources of SC have been detected in apricot and the orthology analysis helped to deepen our understanding on evolutionary aspects of the *S*-RNase-based GSI system exhibited by *Prunus*.

RESUMEN

La autocompatibilidad (AI) comprende un conjunto de barreras moleculares intraespecíficas, controladas por el locus *S*, que favorecen la polinización cruzada y previenen de la endogamia. Solanáceas, Plantagináceas y Rosáceas presentan la llamada autoincompatibilidad gametofítica (AIG) donde el reconocimiento específico está controlado por ARNasas-*S* y proteínas F-box del locus *S* (SFB) como los determinantes femenino y masculino, respectivamente. Por otra parte, genes no ligados al locus *S*, conocidos como factores o genes modificadores, son también totalmente necesarios para la correcta regulación del mecanismo. El sistema AIG parece estar básicamente conservado en *Prunus* pero se han observado notables diferencias con Solanáceas y otras Rosáceas. Con estos antecedentes, el trabajo realizado en esta tesis se ha centrado en la identificación y caracterización de factores modificadores de la AIG en *Prunus* con el fin de mejorar nuestro conocimiento del mecanismo subyacente.

Trabajos previos en albaricoquero (*Prunus armeniaca* L.) mostraron la existencia de una mutación expresada en el polen y no ligada al locus *S*, que se localiza en el extremo distal del cr.3 (locus *M*) y que es capaz de conferir autocompatibilidad (AC) en el cultivar 'Canino'. En este trabajo, otro cultivar de albaricoquero autocompatible llamado 'Katy' fue genética y molecularmente analizado. De manera parecida a 'Canino', una mutación que afectaba a un factor no ligado al locus *S* expresado en el polen era el causante de la pérdida de la respuesta autoincompatible en 'Katy'. Una estrategia de mapeo genético basada en la distorsión en los ratios de segregación permitió mapear la mutación de 'Katy' en el extremo distal del cr.3 (denominada mutación *m*) en una región solapante con la identificada para 'Canino'.

Una búsqueda para la identificación de nuevo mutantes autocompatibles en cultivares y/o accesiones de albaricoquero procedentes de bancos de germoplasma fue llevada a cabo. Por medio del genotipado del locus *S*, 3 alelos *S* no clasificados con anterioridad fueron hallados, mientras que 2 nuevas mutaciones autocompatibles que parecen haber afectado al determinante S masculino SFB fueron detectadas. Adicionalmente, el genotipado para el locus *M* mostró que el mismo haplotipo *m* mutado está compartido por 'Canino' y 'Katy', pero también por 17 cultivares más del norte de América y el oeste de Europa. El haplotipo $M_{1\text{-}0}$, ampliamente distribuido, ha sido propuesto como posible ancestro del haplotipo *m*, sugiriendo que éste surgió mucho más tarde que el alelo Sc, una mutación en el locus S que también confiere AC en albaricoquero.

Con el objetivo de identificar esta mutación, un abordaje integral tanto a nivel genético como genómico y transcriptómico mediante datos NGS procedentes de 'Canino', 'Katy' y del cultivar de albaricoquero autoincompatible 'Goldrich', fue llevado a cabo. Esta aproximación sirvió para identificar un único polimorfismo capaz de explicar el fenotipo de AC, una inserción tipo *FaSt* de 358 pb en acoplamiento con el haplotipo *m* en un gen que codifica para una *disulfide bond A-like oxidoreductase* (*PaMDOr*). *PaMDOr* mostró estar diferencialmente sobre-expresado en anteras maduras, mientras que la inserción *FaSt* predice la formación de una proteína truncada. Estos dos hechos apoyan a *PaMDOr* como el factor modificador de la parte del polen que confiere AC en albaricoquero.

Adicionalmente, análisis filogenéticos sugieren que *PaMDOr* como un posible parálogo de su gen contiguo (llamado PaM-8) que surgió después de la división de Rosáceas y Solanáceas, cuya función ha llegado a ser esencial para el correcto funcionamiento del sistema autoincompatible en *Prunus*. A fin de arrojar cierta luz en las diferencias y similitudes entre los sistemas de AIG basado en ARNasas-*S* de Rosáceas y Solanáceas, las relaciones de ortología para factores modificadores fueron estudiadas. Ortólogos candidatos fueron encontrados para *NaTrxh*, *SBP1* y *MdABCF*, sin embargo, un patrón evolutivo más complejo fue observado para *NaStEP*, *120K* y *NaPCCP*. De modo que, a pesar de las diferencias, se puede hipotetizar que una parte de los modificadores de la AIG están compartidos por las dos familias.

En resumen, el estudio multidisciplinario desarrollado durante esta tesis ha permitido encontrar un novedoso factor modificador (*PaMDOr*) esencial para la respuesta autoincompatible en *Prunus*. Además, nuevas fuentes de AC han sido detectadas en albaricoquero y análisis de ortología ayudaron a profundizar en el entendimiento de los aspectos evolutivos del sistema de AIG basado en ARNasas-*S* en *Prunus*.

RESUM

L'autocompatibilitat (AI) comprèn un conjunt de barreres moleculars intraespecífiques, controlades pel locus *S*, que afavorixen la pol·linització creuada i prevé de l'endogàmia. Solanàcies, Plantaginàcies i Rosàcies presenten l'anomenada AI gametofítica (AIG) on el reconeixement específic està controlat per ARNases-S i proteïnes F-box del locus S (SFB) com a determinants femení i masculí, respectivament. Per un altra banda, gens no lligats al locus *S*, coneguts com factors o gens modificadors, són també totalment necessaris per a la correcta regulació del mecanisme. El sistema AIG pareix estar bàsicament conservat en *Prunus*, però s'han observat notables diferències amb Solanàcies i altres Rosàcies. Amb estos antecedents, el treball realitzat durant aquesta tesi se ha focalitzat en la identificació i caracterització de factors modificadors de l'AIG en *Prunus* a fi d millorar el nostre enteniment del mecanisme subjacent.

Treballs previs a l'albercoquer (*Prunus armeniaca*) mostraren l'existència d'una mutació expressada al pol·len no lligada al locus *S*, la qual està localitzada a l'extrem distal del cr.3 (locus *M*), es capaç de conferir autocompatibilitat (AC) al cultivar 'Canino'. En aquest treball, un altre cultivar d'albercoquer autocompatible anomenat 'Katy' va ser genètica i molecularment analitzat. De manera pareguda a 'Canino', una mutació que afecta a un factor no lligat al locus *S* expressat al pol·len era la causa de la perduda de la resposta autoincompatible en 'Katy'. Una estratègia de mapeig genètic basada en la distorsió en els ratis de segregació va permetre mapetjar la mutació de 'Katy' a l'extrem distal del cr.3 (denominat mutació *m*) en una regió solapant amb la identificada per a 'Canino'.

Una recerca per a la identificació de nous mutants autocompatibles en cultivars i/o accessions d'albercoquer procedents de bancs de germoplasma va ser portada a terme. Mitjançant el genotipatge del locus *S*, 3 al·lels *S* no classificats amb anterioritat van ser trobats, mestres que dos noves mutacions AC que pareixen haver afectat al determinant S masculí SFB varen ser detectades. Amés, el genotipatge del locus M va mostrar que el mateix haplotip m mutat està compartit per 'Canino' i 'Katy', però també per 17 cultivars més del nord d'Amèrica i l'oest d'Europa. El haplotip $M_{1\text{-}0}$, ampliamente distribuït, ha sigut proposat com a possible ancestre del haplotip m, sugerint que aquest va sorgir més tard que el al·lel Sc, una mutació al locus *S* que també conferix AC a l'albercoquer.

Amb l'objectiu d'identificar aquesta mutació, un abordatge integral tant a nivell genètic com genòmic i transcriptòmic mitjançant diversos tipus de dades NGS provinents de 'Canino', 'Katy' i del cultivar d'albercoquer autoincompatible 'Goldrich' va ser portat terme. Aquesta aproximació va permetre identificar un únic polimorfisme capaç d'explicar el fenotip d'AC, es tracta d'una inserció de 358 pb en adaptament amb el haplotip m en un gen que codifica per a *disulfide bond A-like oxidoreductase* (*PaMDOr*). *PaMDOr* va mostrar estar diferencialment sobre-expressat en anteres madures, mentres que la inserció FaSt prediu la formació d'una proteïna truncada. Estos dos fets recolzen a PaMDOr com al factor modificador de la part del pol·len que conferix AC en albercoquer.

A més a més, anàlisis filogenètics suggerixen que *PaMDOr* podria ser un paràlog del seu gen contigu (anomenat PaM-8) que va sorgir després de la divisió de Rosàcies i Solanàcies, en la qual la funció ha arribat a ser fonamental per al correcte funcionament del sistema d'AIG a *Prunus*. A fi de tirar certa llum en quant a les diferències i similituds entre els sistemes d'AIG basats en ARNases-S de Rosàcies i Solanàcies, les relacions d'ortologia per als factors modificadors va ser estudiat. Ortòlogs candidat van ser trobats per a *NaTrxh*, *SBP1* i *MdABCF*, no obstant, un patró evolutiu més complex va ser observat per a *NaSTeP*, *120K* i *NaPCCP*. De tal manera que, a pesar de les diferències, es pot plantejar la hipòtesi de que una part dels modificadors de l'AIG estan compartits per les dues famílies.

En resum, l'estudi multidisciplinari desenvolupat en aquesta tesi ha permés trobar un nou factor modificador (*PaMDOr*) fonamental per a la resposta autoincompatible a *Prunus*. Amés, noves fonts d'AC han sigut detectades a l'albercoquer i l'anàlisi d'ortologia varen ajudar a profunditzar a l'enteniment dels aspectes evolutius del sistema de l'AIG basada en ARNases-*S* a *Prunus*.

INDEX

LIST OF FIGURES

LIST OF TABLES

ABREVIATION LIST

AGP	Arabinogalactan protein
apricot M-locus supercontig	a*M*-supercontig
ARC1	Arm Repeat-Containing
AtPDIL	Arabidopsis thaliana Protein Disulfide Isomerase-Like
DsbA	Disulfide bond A-like
GI	General Inhibitor
GmPDIL	Glycine max Protein Disulfide Isomerase-Like
GRX	Glutaredoxin
GSI	Gametophytic Self-Incompatibility
HT-B	High-Top Band
HV	Hipervariable
MLPK	M-Locus Protein Kinase
NGS	Next Generation Sequencing
ORF	Open Reading Frame
PaM	Prunus armeniaca M-locus
PaMDOr	Prunus armeniaca M-locus Disulfide bond A-like Oxidoreductase
PCD	Progammed Cell Death
PELPIII	Pistil Extensin-Like Protein III
PPM	Pollen-Part Mutation
PrpS	Papaver rhoeas pollen S
PrsS	Papaver rhoeas stigmatic S
RBH	Reciprocal Best Hit
RHV	Region Hipervariable
SC	Self-Compatibility
SCF	Skp1/Cul1/F-box
SDL	Segregation Distortion Locus
SFB	S haplotype-specific F-box
SFBB	SFB Brothers
SI	Self-Incompatibility
SLF	S-locus F-box
SLFL	SLF-Like
SLG	S-Locus Glycoprotein
Sli	*S*-locus inhibitor
SNP	Single Nucleotide Polymorphism
SP11/SCR	S-locus Protein 11/S-locus Cysteine-Rich
SRK	S-locus Receptor Kinase
SSI	Sporophytic Self-Incompatibility
SSR	Simple Sequence Repeat
TRX	Thioredoxin
TTS	Transmitting Tract-Specific
UI	Unilateral Incompatibilty

GENERAL INTRODUCTION

1. Reproductive barriers in plants. Strategies to avoid inbreeding

Plants have a predominantly sessile lifestyle, this circumstance has carried to develop hermaphrodite flowers maintaining the capacity to reproduce without the need of a mate throughout their evolutionary history (Rea & Nasrallah, 2008). The idea behind this affirmation was hypothesized by Fisher (1941), who proposed self-fertilizing as strategy that ensures the offspring. Barrett (2002) estimated that approximately 20% of the angiosperm species use selfing as sexual reproduction strategy, which also allows a rapid colonization of unoccupied space (Pannell & Barrett, 1998). This behavior has a high cost: the inbreeding depression. Thus, plants have evolved different strategies on the basis of their ecological and biological context to prevent it.

Strategies aimed to prevent selfing in plants were already reported by Charles Darwin. In his work '*The different forms of flowers on plants of the same species*' (published in 1877) he described species that elaborate alternative floral morphologies, exemplified by *Primula vulgaris*, where two floral morphs differ reciprocally from one another in the positions in which anthers and stigmas are located in flowers (Figure In1a). This strategy, currently known as heterostyly, is part of a variety of strategies that expect to separate spatially (hercogamy) or temporally (dichogamy) mature pollination intermediaries of the plant sexual structures. In addition to heterostyly, divided in distyly (Figure In1a) and tristyly depending on the number of floral morphs that differ reciprocally, enantiostily (Figure In1b) and flexistyly (Figure In1c) are other examples also included in this group of phenomena. These strategies possess their maximum expression in plants having unisexual flowers (~10% of plant species), a condition referred as dicliny (that involve various combinations of female, male and hermaphrodite flowers at plant and population levels) (Figure In1d).

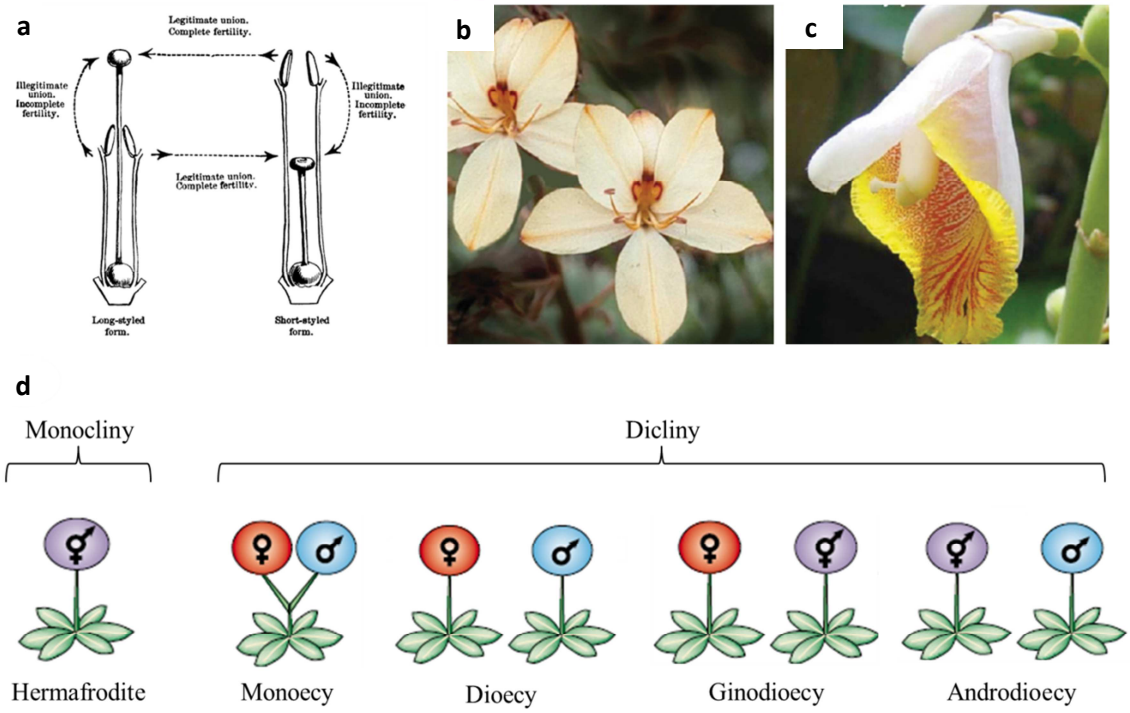


Figure In1. Morphological strategies developed by heteromorphic plants to prevent inbreeding.
a) distily, b) enantiostily, c) flexistily and d) dicliny. Image taken from Barret (2002)

Global distinct floral morphology mentioned until now refers to heteromorphic flowers which reproductive goal is to promote cross-pollination (Barrett, 2002). Notwithstanding, the majority of plant species have homomorphic flowers, that is, all flowers have exactly the same morphology. Thus, the proximity and simultaneous maturity of reproductive organs significantly increase the possibility of self-pollination. It is in this context where the commonly known as 'self-incompatibility' systems operate.

2. Historical evolution of the self-incompatibility concept

Using C. Darwin own words:

> *It is an extraordinary fact that with many species, flowers fertilised with their own pollen are either absolutely or in some degree sterile; if fertilised with pollen from another flower on the same plant, they are sometimes, though rarely, a little more fertile; if fertilised with pollen from another individual or variety of the same species, they are fully fertile; but if with pollen from a distinct species, they are sterile in all possible degrees, until utter sterility is reached. We thus have a long series with absolute sterility at the two ends;—at one end due to the sexual elements not having been sufficiently differentiated, and at the*

> *other end to their having been differentiated in too great a degree, or in some peculiar manner.*
> *The fertilisation of one of the higher plants depends, in the first place, on the mutual action of the pollen-grains and the stigmatic secretion or tissues, and afterwards on the mutual action of the contents of the pollen-grains and ovules. Both actions, judging from the increased fertility of the parent plants and from the increased powers of growth in the offspring, are favored by some degree of differentiation in the elements which interact and unite so as to form a new being.*

This fragment (p.455) from '*The effects of cross and self-fertilisation in the vegetable kingdom*' (1878) denotes how Darwin was able to value a pattern widely spread in plants pursuing to avoid self-pollination. He coined this phenomenon as *self-sterility*, describing it as a consequence of pollen-pistil interaction. By that time, Mendel had already published his results about heredity rules but they were not rediscovered until 1900. Hence, this lack of knowledge in genetics led Darwin to attribute to the environment influence the cause of self-sterility (McClure, 2009). Nevertheless, botanists and geneticists from first decades of the 20th century observed that self-sterility described by Darwin followed genetic rules proposed by Mendel. The works of Compton (1913), East & Park (1917), East & Mangelsdorf (1925), East & Yarnell (1929) and East (1932) highlighted that Darwin's self-sterility was actually a reaction of compatibility/incompatibility between pollen and pistil, laying the foundations of the currently known as self-incompatibility systems.

3. Genetics of self-incompatibility. Gametophytic and sporophytic systems.

Self-incompatibility (SI) is defined as a reproductive barrier which inhibits fertilization by either self-pollen or pollen from closed related plants preventing inbreeding and enhancing outcrossing in flowering plants (de Nettancourt 2001). SI has been reported in more than half of plant species and represents the most extended tool to avoid inbreeding in the plant kingdom (Igic & Kohn 2001).

Classic genetic studies established that most SI systems in angiosperms are controlled by a single multiallelic locus termed *S*-locus. This locus contains at least two linked genes acting as determinants, one of them specifically expressed in pollen (male *S*-determinant) and the other in the pistil (female *S*-determinant) (de Nettancourt, 2001).

Conventionally, alleles from *S*-determinants are referred to belong to the same *S*-haplotype because they are genetically linked (McCubbin & Kao, 2000). Therefore, pollen rejection during incompatible reactions is triggered when the two expressed male and female *S*-determinant alleles come from the same *S*-haplotype (Figure In2a). Instead, if *S*-alleles differ to each other, pollen tube will be potentially able to reach the ovary and fertilize the ovule (Iwano & Takayama, 2012; McClure et al., 2011; Takayama & Isogai, 2005). According to the time of gene action in the stamen, most types of self-incompatibility can be classified into the sporophytic and the gametophytic groups (de Nettancourt 2001). In the first one, the pollen phenotype is determined by the genotype of the diploid pollen-parent (Figure In2b), while in the second it is determined by the genotype of the individual microspore (Figure In2c).

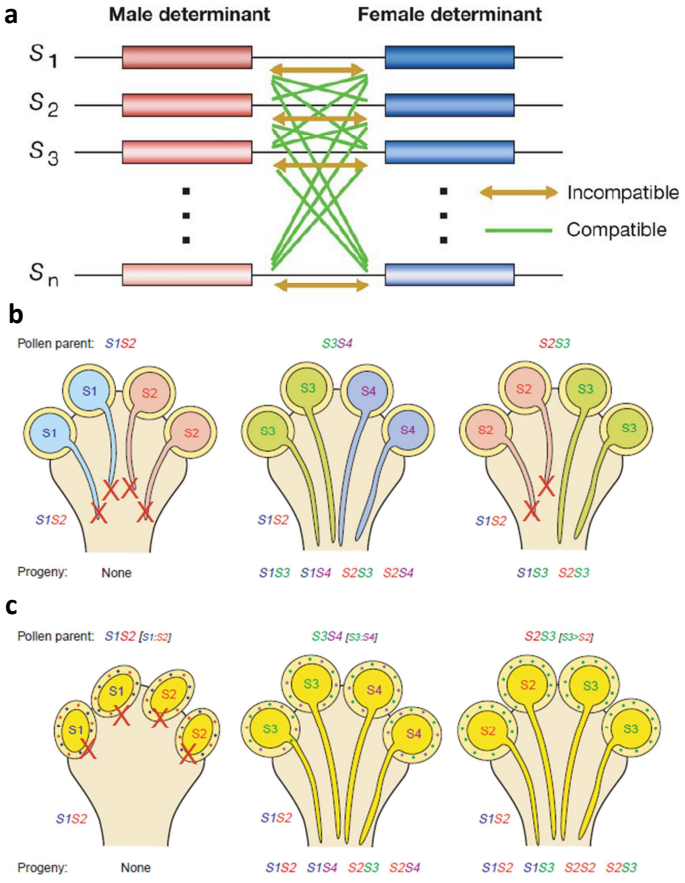


Figure In2. Self-Incompatibility genetics. a) schematic representation of the *S*-locus. *Red* and *blue rectangles* symbolize male and female *S*-determinants, while *orange arrows* and *green lines* incompatible and compatible crosses, respectively. Image taken from Takayama & Isogai (2005). b) schematic drawing of cross-compatibility response in a diploid Gametophytic Self-Incompatibility (GSI) system. Three different types of crosses are shown: incompatible, fully-compatible and semi-compatible. c) schematic drawing of cross-compatibility response in a diploid Sporophytic Self-Incompatibility (SSI) system. Interactions of co-dominance and dominance-recessiveness are indicated by *colored dots* in pollen surface. S_1 (*blue dots*) and S_2 (*red dots*), and S_3 (*green dots*) and S_4 (*purple dots*) alleles are co-dominant, whereas S_3 allele dominates over S_2 allele. Images b) and c) have been taken from Nasrallah (2005).

Incompatibility mechanisms are not only restricted to intraspecific barriers, but they have also been associated to interespecific crossability barriers, and particularly with the so-called unilateral incompatibility (UI). UI is a particular case within interspecific barriers where crosses are feasible in one direction rather than the other way round, suggesting that there are not gross differences in the requirements for pollen

tube development (Hancock et al., 2003). The UI general rule was defined as the SI x SC rule by Lewis & Crowe (1958) and means that pollen from self-compatible species is rejected by self-incompatible species meanwhile the reciprocal cross tends to be compatible. Thereafter, the phenomenon of UI has been extensively described in several plant species (Heslop-Harrison, 1982; Hiscock & Dickinson, 1993; Chen & Adelberg, 2000; Martin, 1967; Pandey, 1981; Layne & Sherman, 1986). Mechanisms controlling interspecific pollination have received less attention (McClure et al., 2000), but factors involved in SI mechanisms in Solanaceae (Li & Chetelat, 2010, 2014; Murfett et al., 1996; Tovar-Méndez et al., 2014) and Brassicaceae (Kitashiba & Nasrallah, 2014) have also been observed to be related to UI supporting a connection between SI and UI.

4. Self-incompatibility as an agronomic relevant trait

Self-incompatibility did not only generate interest among evolutionists and geneticists regarding its implication in plant evolution but also among plant breeders. Public institutions and private companies developing plant breeding programs soon focused their interest on SI. Since 1911 the John Innes Horticultural Institution studied incompatibility and sterility in plums, cherries and apples and extended the studies to pears at the end of the 1930´s (Crane and Lewis, 1942). For instance, cross-pollinations were used to define intercompatible groups in sweet cherry cultivars by Crane and Brown (1937) and, later on, a pollen irradiation program produced the first self-compatible cultivars within this strictly self-incompatible species (Lewis and Crowe, 1954). In 1940, the Japanese seed company Sakata Seed Co., introduced the F_1-hybrid cabbage cv. Suteni Kanran by using SI trait, and this success was followed by the Takii & Co. Ltd Company that introduced the cabbage cvs. Choko-1c and Choko-1cc in 1950 (Watanabe, 2008). Interest on SI trait was not only confined to fruit trees (Rosaceae) and cabbages (Brassicaceae) but also extended to other important crop species such as potato (Pushkarnath, 1942) [*Solanum tuberosum*; Solanaceae], cacao (Cope, 1962) [*Theobroma cacao* L.; Malvaceae] sunflower (Pinthus, 1959) [*Helianthus annus* L.; Asteraceae], rye (Lundqvist, 2010) [*Secale cereale* (L.) M. Bieb.; Gramineae], pummelo (Soost, 1964) [*Citrus grandis* osbeck; Rutaceae], etc. Usefulness of self-(in)compatibility trait in plant breeding and production has been proved for different objectives. Few of them are briefly summarized next.

4.1. *S*-genotyping facilitates pollination control

In crops exhibiting SI systems, cultivars serving as pollen donors "pollenizers" are commonly interspersed in the orchards since fruit set depends largely on cross-pollinations. For instance, in diploid fruit tree species having gametophytic self-incompatibility (GSI), out-crosses can be classified into three types: incompatible, semi-compatible and fully compatible when the two progenitors share both *S*-alleles, only one or none of them, respectively (Figure In2b). Obviously in semi-compatible crosses half of the available pollen grains are rejected and this fact has been shown to have a significant impact on fruit set and yield (fruit size) in different Rosaceae species such as apple, European pears and Japanese plums grown in sub-optimal regions (i.e. the Mediterranean basin) for growth and pollination (Sapir et al., 2008; Schneider et al., 2005; Zisovich et al., 2005). The use of "pollenizers" is not exclusive of stone and pome fruit trees but it is also common in other species (Woodcock, 2012). Therefore, for seed and fruit industry it is important to know how many genetically different compatibility groups exist within a particular species, since this knowledge would help to select adequately those combinations of cultivars that may work effectively in orchards settings to produce regular cropping. In those species where the *S*-determinants have already been identified, molecular genotyping has progressively replaced controlled pollination, pollen tube growth tests and enzymatic assays used to determine the *S*-genotype, accelerating the identification of new *S*-alleles, since these methods do not depend on the environmental conditions and do not require adult plants in the case of trees (Yamane & Tao, 2009).

4.2. SI as an alternative to androsterility for developing hybrids.

Heterosis or hybrid vigor is a phenomenon largely pursued by breeders because of F_1 progenies frequently show higher yields and exhibit other interesting traits favoring adaptation to production conditions including, for instance, a better response to abiotic stresses. In fact, hybrids are the most usual form of commercial cultivars in many crop species such as maize, sorghum, tomato, pepper, etc. (Kempe & Gils, 2011). However, due to the breeder´s selection, most cultivated crop species are self-compatible and, therefore, hybrid production requires an efficient pollination control system to prevent undesired self-fertilization of the female parent. Methods range from the non-biological technologies, including manual or mechanical removal of the anthers and gametocide chemical agents, to the biological systems, commonly based on nuclear

or cytoplasmic-encoded male sterility and diverse techniques to restore fertility in the F_1 hybrid (Kempe & Gils, 2011). SI has been reported as an advantageous alternative to male sterility in many cases, especially for those crop species with entomophilous pollination since pollen-collecting bees rarely visit male-sterile plants (Kaothien-Nakayama et al., 2009). Nonetheless, SI systems integrated in breeding programs lead to self-incompatible F_1 hybrids, obtained either from two self-incompatible parents or a self-incompatible female and a self-compatible male parent, and this is a handicap for those crops commercialized for their seeds (oilseed rape) or fruits (stone and pome fruits) (Figure In3). Indeed, not only self-incompatible female lines but also self-compatible F_1 hybrids are demanded by breeding programs (Kaothien-Nakayama et al., 2009).

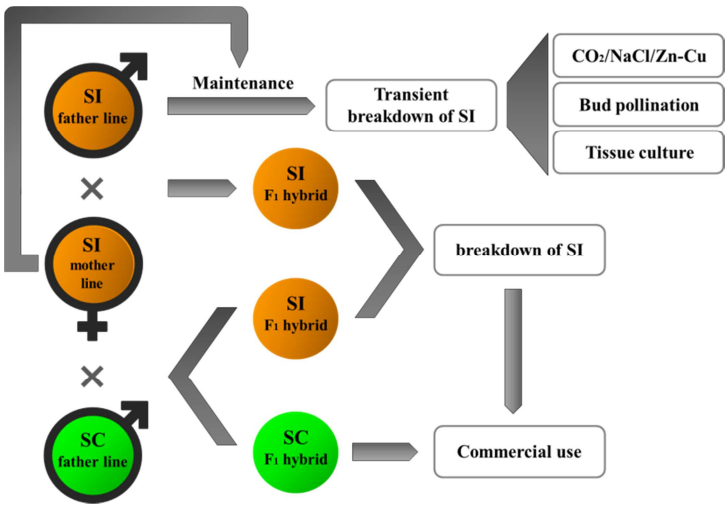


Figure In3: Systems to develop F_1 hybrids based on the use of self-incompatibility

In Brassicaceae SI has been widely used for hybrid seed production in the generally self-incompatible vegetable types of diploids *Brassica oleracea* and *B. rapa*. However, the derived amphi-diploid oilseed rape (*B. napus*) is naturally self-compatible and introgression of *S*-alleles from parental species was required to produce hybrid seeds (Rahman, 2005). Genetic modification to introgress SI in *Brassica* was already proposed by Nasrallah et al. (1991) but to date it has not been yet reported. SI is seen as a promising alternative for a hybrid breeding system in other species such as wheat (Whitford et al., 2013) and ryegrass (Pembleton et al., 2015) but until now the lack of knowledge on the *S* and *Z* SI determinants in grasses has hindered this option. In a wider sense, hybrid production might be potentially achieved by transferring *S*-determinants. Recently, Lin et al. (2015) have reached this goal conferring SI to the

self-compatible *Arabidopsis thaliana* (SSI system) throughout the use of *Papaver S*-determinants (GSI system) (de Graaf et al., 2012; Lin et al., 2015).

4.3. Removing interspersed pollinators while increasing fruit set and quality

SI restricts fertilization and fruit setting in many fruit tree crops. In terms of crop production, SC is a desired trait because it avoids the use of cross-pollinators, growing a single cultivar as a 'solid block'. In addition, it is also a crucial factor to control fruit set and yield. For instance, by definition in semi-compatible crosses half of the pollen is rejected (in contrast with fully-compatible crosses) which might reduce yield. In addition, in some genera, such as *Malus* and *Pyrus*, where many ovules could potentially be fertilized, a reduction in the number of fertilization events might result in a lower number of seeds and, subsequently, low-quality fruits. In these and other species, SC is mostly tied to satisfactory fruit set producing high yields or even over-cropping (Goldway et al., 2007). However, while SC may facilitate a reduction in the number of hives required it is generally accepted that it can not guarantee full yields in many crops (i.e. sunflower, canola, sour cherry, almond, apricot, etc.) where cross-pollination is needed to ensure maximum set (Schneider et al., 2001; Zhang & Hiratsuka, 2005). Conversely, the presence of honey bee colonies might induce 'over-pollination' when growing self-compatible stone fruits (i.e. sour cherry, peach or apricot). This phenomenon leads to an overly heavy fruit set and high yield by weight resulting in a high proportion of undersized fruits of reduced value (Woodcock, 2012).

In self-incompatible crops, commercial self-compatible cultivars are mostly the result of spontaneous style- or pollen-part mutations conferring SC, subsequently selected by growers and breeders. SC is usually the indirect result of selection for early blooming (frequently associated with early ripening) since pollinating insects and/or mates could be limited in early spring. This is the case of several cultivars in several stone fruit trees (Yamane & Tao, 2009). Unlike this process, induced mutations leading to SC by irradiating pollen with X-ray and successive crosses has given a number of self-compatible commercial cultivars in sweet cherry (Ushijima et al., 2004). In other species such as sunflower, self-compatible cultivars were introduced in the 60´s also through traditional breeding programs (Astiz et al., 2011) while in turnip (*Brassica rapa*) only a few self-compatible cultivars are available (Zhang et al., 2013).

Notwithstanding, the reported uses of SC are not based upon the knowledge of SI controlling factors. A first attempt in this regard can be credited to Broothaerts et al.

(2004) who developed a self-fertile apple cultivar by silencing the S-RNase gene-expression in the pistil (female *S*-determinant; see section 6.1) which results in un-arrested pollen tube growth and fertilization. More recently, Jung et al. (2012) developed a self-compatible *Brassica rapa* line by RNAi mediated *S*-locus gene silencing.

5. Molecular mechanisms underlying self-incompatibility

Three SI systems have been molecularly characterized to date, but a plethora of studies are being developed in many others. The capital part for the depth molecular knowledge in these 3 systems has been the elucidation of the *S*-components (Figure In4).

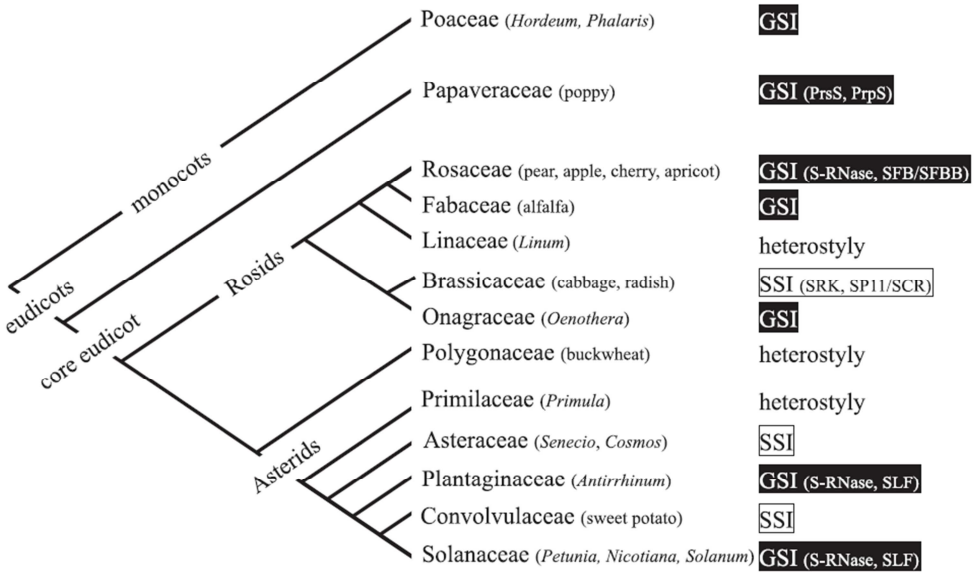


Figure In4. The different self-incompatibility systems which *S*-factors have been elucidated in angiosperms. The phylogenetic tree is based on The Angiosperm Phylogeny Group (2009). Pistil-part and pollen-part determinants are between parentheses. Image taken from Sassa (2016).

The mechanism exhibited by the Brassicaceae is the only sporophytic SI (SSI) known in depth. The *S*-locus comprises two highly polymorphic glycoproteins expressed in the papilla cells of the stigma, SLG (*S*-locus glycoprotein) (Nasrallah et al., 1987; Takayama et al., 1987) and SRK, a SLG-like protein in its extracellular domain (*S*-domain) that also contains a transmembrane domain and an intracellular serine/threonine receptor kinase domain (Stein et al., 1991). SP11/SCR is a cysteine-rich protein encoded by the *S*-locus as well but specifically expressed in the anther tapetum and pollen grains (Schopfer et al., 1999; Suzuki et al., 1999; Takayama et al., 2000). Gain-of-function assays highlighted that SRK protein was the female component

in the SSI system (Takasaki et al., 2000) while SP11/SCR was analogously demonstrated to be the male *S*-component (Schopfer et al., 1999; Shiba et al., 2001; Takayama et al., 2000; Takayama et al., 2001). SLG was proposed to act as enhancer of SRK action (Suzuki et al., 1999). In addition, interaction between SRK-SP11/SCR was determined by different biochemical approaches (Takayama et al., 2001). Despite *S*-factors and their interaction have been fully characterized, the signaling pathway triggered by ligand-binding, the receptor activation and the rapid mechanism of self-pollen rejection are not well understood. Nevertheless, some proteins involved in this process have been reported. Thus, after SRK-SP11/SCR interaction, the receptor is autophosphorylated and together with the M-Locus Protein Kinase (MLPK), a plasma membrane-tethered protein (Kakita et al., 2007; Murase et al., 2004), interact and phosphorylate the Arm repeat-Containing protein (ARC1) (Gu et al., 1998). In turn, Exo70A1 (component of the exocyst complex) was observed as an ARC1 interacting protein. Transgenic lines reducing Exo70A1 expression levels disrupted compatible pollen tube growth (Samuel et al., 2009), which has led to propose that Exo70A1 promotes compatible pollination success and that the ARC1-mediated degradation of Exo70A1 leads to self-pollen rejection by inhibiting secretion of 'compatibility' factors (Figure In5). Additionally, two Thioredoxin-*h* proteins (THL1 and THL2) were shown to bind SRK in yeast two-hybrid screening and suggested to function preventing the autophosphorylation of the SRK receptor (Mazzurco et al., 2001). In spite of these results, some works have recently questioned a specific role of the MLPK, THL and ARC1 proteins in the SSI system (Kitashiba et al., 2011; Yamamoto & Nasrallah, 2013).

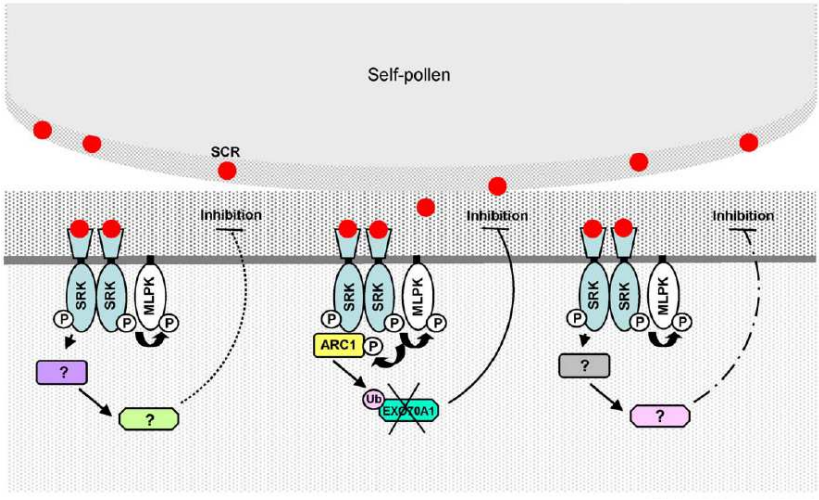


Figure In5. A model for multiple SI signaling pathways in the Brassicaceae. The diagram shows the zone of contact between a stigma epidermal cell and a self-pollen grain. SCR molecules (from diploid

tapetal tissues of pollen grain) are shown as *red circles*. The diagram shows a simplified view of the stigma–pollen interaction in which only a single SRK variant and its cognate SCR are shown. SCR-SRK interaction causes autophosphorylation of the receptor and triggers several signaling cascades within the stigma epidermal cell. MLPK is proposed as a common signaling intermediate. The middle cartoon illustrates a cascade that involves ARC1-mediated-ubiquitination of EXO70A1. The two other cartoons postulate the existence of ARC1/Exo70A1-independent signaling pathways that use currently unknown components. Image taken from Tantikanjana et al. (2010).

Two distinct GSI systems have been deeply studied from a molecular point of view: the one present in *Papaver* relying on Progammed Cell death (PCD) and that based on *S*-RNases present en several plant families (see next section for a detailed description of this latter). The GSI system characterized in *Papaver rhoeas* is, undoubtedly, the better understood physiologically. In this case the *S*-locus encodes for the PrsS female *S*-factor, a small and highly polymorphic protein secreted by stigmatic papilla cells acting as a signaling ligand (Foote et al., 1994) and for PrpS, a presumable transmembrane protein operating as male determinant (Wheeler et al., 2009). The interaction of both factors triggers an intracellular signaling network resulting in a highly specific biological events involved in PCD (Thomas & Franklin-Tong, 2004; Bosch & Franklin-Tong; Wilkins et al., 2014). *S*-determinants interaction produces an increase of free Ca^{2+} that initiates a signaling cascade (Franklin-Tong et al., 1997; Franklin-Tong et al., 1995; Franklin-Tong et al., 1993). Phosphorylation events in poppy after incompatible response are initiated from p56 protein MAPK, where different evidences have shown to be related in PCD response (Li et al., 2007; Rudd et al., 1996). Furthermore, Pr-p26.1a/b are two pollen expressed pyrophosphatases that might provide an additional inhibitory mechanism rejecting pollen tube growth (de Graaf et al., 2006). SI in poppy has also demonstrated to alter cytoskeleton throughout depolymerization of the F-actin in a Ca^{2+} signaling dependent-manner (Geitmann et al., 2000; Snowman et al., 2002). Lastly, DNA fragmentation is one of the late steps in self-incompatible response, different evidences has shown that a DEVDase/caspase-like activity is involved in SI-mediated pollen-tube inhibition and DNA fragmentation (Bosch & Franklin-Tong, 2007). Figure In6 shows in detail the complex and integrated network taking place in poppy GSI response.

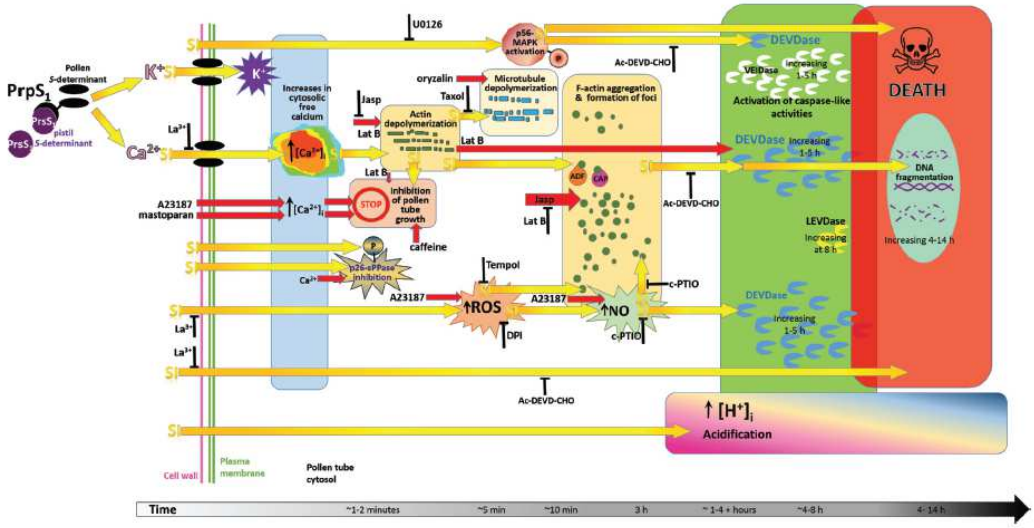


Figure In6. Cartoon showing a model of the integrated self-incompatibility (SI) programmed cell death (PCD) signalling network in *Papaver rhoeas* pollen. Image taken from Wilkins et al. (2014).

6. Gametophytic Self-Incompatibility based on S-RNases

GSI based on *S*-RNases has been found to operate in Solanaceae, Plantaginaceae, Rubiaceae and Rosaceae families being the most extended SI system in the plant kingdom (Igic & Kohn, 2001). However, though some basic features of the mechanism seem to be shared by all families striking differences have also been observed.

6.1. Female (*S*-RNases) and male (*S*-locus F-box proteins) *S*-determinants

The stylar 'component' of the *S*-locus codes for a T2-type RNase (S-RNase) shown to have ribonuclease activity in *Nicotiana alata* for the first time (Anderson et al., 1986; McClure et al., 1989). Solanaceous S-RNases contain 5 conserved domains, from C1 to C5; where C2 and C3 have histidine residues involved in RNA degradation, and the rest contribute to its hydrophobic core. In turn, two regions showing high variability, HVa and HVb (localized between C2 and C3 domains) participate in the specific recognition process (Ioerger et al., 1991; Xue et al., 1996). This set of evidences led to propose a dual role for the S-RNase. On one hand, it acts as the female factor implicated in the specific recognition and, on the other, inhibits the incompatible pollen tube growth as a consequence of its cytotoxic activity (McClure et al., 2011).

By sequencing genomic *S*-locus regions flanking S-RNases, Lai et al. (2002) identified a gene codifying for an F-Box protein (*SLF* from *S*-locus *F-box*) in *Antirrhinum* as the pollen *S*-determinant. According to the genetics of SI, linkage with

S-RNase, pollen-specific expression and high sequence diversity were the requisites to be fulfilled by pollen *S*-determinants. *SLF* showed the two first but not a high polymorphism rate. Meanwhile, pollen-expressed F-box genes linked to *Prunus* S-RNases were also cloned from almond and Japanese apricot. Proteins encoded by these genes (coined as *SFB*) showed a high amino acid variability among the different alleles (Entani et al., 2003; Ushijima K et al., 2003). Sijacic et al. 2004 demonstrated by transgenic experiments that SLF proteins were the pollen *S*-factor in Solanaceae using a distinctive feature of this family, competitive interaction generated by heteroallelic pollen (see section 6.3). Many *S*-locus *F-box* genes were also identified in *Malus* and *Pyrus* (Rosaceae) and named as *SFBB* (from *SFB* Brothers) by Sassa et al. (2007). As main features, all *S*-locus F-box proteins contain an F-Box domain in its N-terminal region and two variable (V1 and V2) and two hypervariable (HVa and HVb) regions at the C-terminal end (Ikeda et al., 2004).

6.2. *S*-locus unlinked genes controlling GSI: the modifier factors

S-locus unlinked genes are also required for the proper functioning of the SI mechanism being termed modifier genes or modifier factors. Modifiers can be classified into three different classes on the basis of their function: 1) those affecting the expression of *S*-determinants; 2) factors interacting either genetically or biochemically with the *S*-determinants being required for pollen rejection but with no wider role in pollination; 3) factors that function in pollen rejection and in other pollen-pistil interactions as well (McClure et al., 2000).

HT-B (High-Top Band) was the first non-*S*-factor identified acting in the pistil side of *Nicotiana* (McClure et al., 1999). In HT-B suppressed plants *S*-pollen rejection failed but S-RNases were normally uptaken and sequestered in vacuole compartments. Additionally, HT-B degradation was observed in compatible crosses whereas in incompatible crosses it was entirely operational. Although its role is still unknown, these evidences suggest a probable involvement in the degradation of vacuolar membranes after incompatible *S*-recognition (Goldraij et al., 2006). More recently, NaStEP, a Kunitz-type proteinase inhibitor, has been found to be crucial for *S*-specific pollen rejection and HT-B stability. This protein is expressed in stigmas of *Nicotiana spp*. being uptake into pollen tubes independently on the (in)compatibility reaction. Interestingly, non-functional transgenic lines of NaStEP showed reduced HT-B levels within pollen tubes, behavior that was retained in the wild-type preferentially in

compatible pollinations. These evidences support NaStEP as positive regulator of HT-B but the mechanism still remains to be elucidated (Busot et al., 2008; Jimenez-Duran et al., 2013). Arabinogalactan proteins (AGPs) are abundant in the transmitting tract of *Nicotiana* styles being also needed for pollen tube growth (Cheung et al., 1993). Within this diverse group of proteins, some members have shown to interact with S-RNases and enter into growing pollen tubes (Cruz-Garcia et al., 2003 and 2005). These AGPs are pistil extensin-like protein III (PELPIII), transmitting tract-specific glycoproteins (TTS) and 120K (Cheung et al., 1993; de Graaf et al., 2003). PELPIII loss of function breaks down interspecific incompatibility of *Nicotiana tabacum* (Eberle et al., 2013). Meanwhile, experiments with RNAi lines suppressing 120K in *Nicotiana alata* drove to the loss of its ability to reject *S*-specific pollen from *Nicotiana plumbaginifolia* (Hancock et al., 2005). Despite all these evidences, the exact role of these modifiers in GSI remains elusive. NaTrxh is another stylar modifier gene found in *Nicotiana* encoding a thioredoxin (TRX) from *h* group (subgroup II) shown to interact with S-RNases and AGPs *in vitro*. NaTrxh function is still unknown, but it has been proposed to participate in the transport of some of these proteins into pollen tubes or, alternatively, to release them once inside the pollen tube. Additionally, it has been argued a hypothetical interaction with NaStEP to regulate pollen rejection (Avila-Castañeda et al., 2014; Juárez-Díaz et al., 2006).

Non-*S*-factors of the pollen side have also been identified. The GSI pollen *S*-determinant SLF is proposed to be a component of the SCF E3 ubiquitin ligase complex formed by Skp1/Cul1/F-box (SCF) proteins, where additionally Cul1 interacts with Rbx1 (Hua & Kao, 2006; Huang et al., 2006; Li et al., 2014). SBP1 is a RING-finger protein (E3 ligase) expressed in a variety of tissues that binds SLFs, S-RNases, AGPs and some transcription factors in yeast two hybrid assays (Sims & Ordanic 2001). NaPCCP is an AGPs interacting protein associated with the pollen membrane and internal compartments. It has been suggested to contribute in sorting pistil proteins such as AGPs, although no evidence supporting its intervention for proteins involved in SI has been demonstrated (Lee et al., 2008 and 2009). MdABCF is the last GSI modifier discovered to date and the unique found in a non-Solanaceae species (*Malus domestica*). MdABCF is a transmembrane transporter located in the pollen tube membrane that interacts with S-RNases mediating in their transport across pollen tube in a coordinated cytoskeleton-manner. The transport of S-RNases in silenced MdABCF lines was blocked disrupting the self-incompatible response (Meng et al., 2014).

6.3. Biochemical models proposed for the S-RNase-based GSI system

Since the end of 80's several models have been proposed to explain how self-incompatible response is regulated by incorporating newly available evidences. Currently, two models seem to describe this mechanism feasibly: degradation model by collaborative non-self recognition and compartmentalization model. Both models have important aspects in common. For instance, S-RNase/SLF interaction determines (in)compatible pollination, while RNA degradation is assumed to cause pollen tube rejection in incompatible cross (McClure et al., 2011). This degradation is carried out by S-RNases that massively enter into pollen tubes from the transmitting tract style independently on their *S*-genotype (Luu et al., 2000). Nevertheless, there are also strong differences in their postulates. Degradation model proposes that S-RNase/SLF interaction displays the massive degradation of non-self S-RNases preventing their cytotoxic effect. SLF proteins are thought to be components of the SCF E3 ubiquitin ligase complex (SCF^{SLF}) that function targeting proteins by ubiquitination for their posterior degradation by 26S proteasome proteolytic pathway and S-RNases were shown to interact with the SCF^{SLF} complex (Entani et al., 2014; Hua & Kao 2006; Huang et al., 2006). *In vitro* pull down assays between allelic variants of SLF and S-RNase resulted stronger in non-self interactions than in self-interactions (Hua & Kao 2006). However, the low allelic diversity exhibited by *Petunia* SLFs in comparison to S-RNases was unexpected according to that observation. This conundrum was solved by Kubo et al. (2010), who observed that *Petunia S*-locus contain more SLF-like genes that also interact with non-self S-RNases in a collaborative manner (Figure In7a). In Solanaceous species, the loss of pollen *S*-function could only be detected when heteroallelic pollen, containing two different SLF alleles, was present in the so-called competitive interaction (Golz et al., 1999 and 2001). The cause of this phenomenon was unknown for a long time, but Kubo et al. (2010) revealed that it is produced by the collaborative action of multiple SLFs detoxifying subsets of non-self S-RNases. This observation was supported by the analysis of a natural self-compatible Japanese pear mutant (Okada et al., 2008) (Figure In7b).

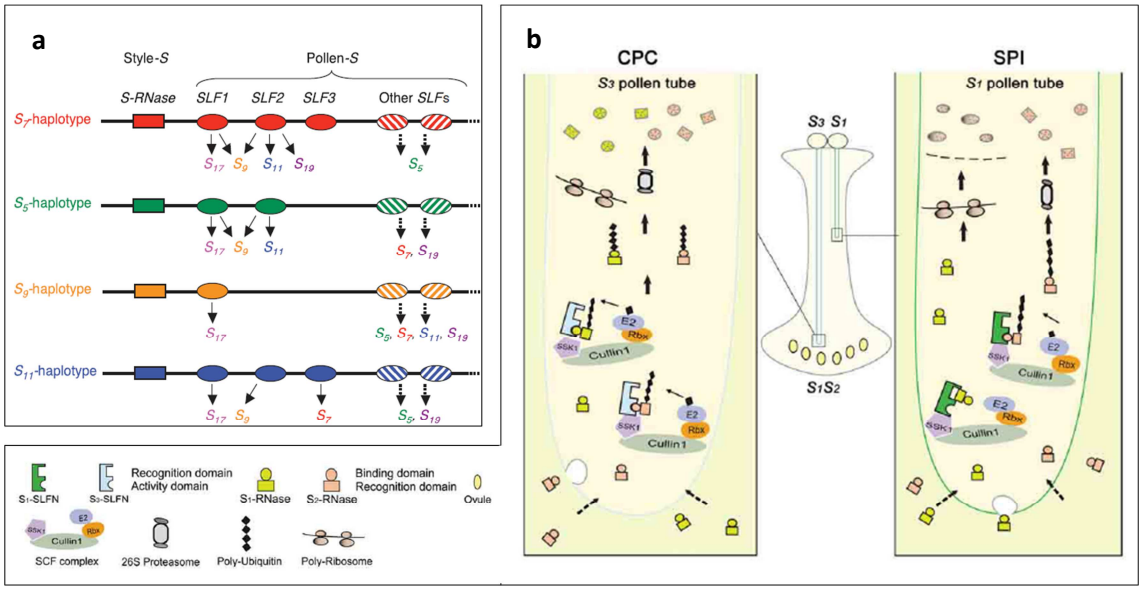


Figure In7. Degradation model by collaborative non-self recognition in S-RNase-based GSI mechanism. a) schematic representation showing Solanaceae *S*-haplotypes where single S-RNases genes and multiple SLFs are represented by *boxes* and *ovals*, respectively. Target S-RNase alleles for each SLF to detoxify are connected by *solid arrows* with their target alleles. Image taken from Kubo et al. (2010). b) In a compatible cross (CPC), S_3 pollen lands on the stigma, germinates and grows into an S_1S_2 style. Both S_1- and S_2-RNases enter the S_3 pollen tube and interact with a hypothetical activity domain of the pollen *S* in the cytosol of the pollen tube. One or more SLF (SLFN) in S_3 pollen tube form functional $SCF^{S3\text{-}SLFN}$ complexes to tag S_1-and S_2-RNases with a polyubiquitin chain, which are subsequently degraded by the 26S proteasome, escaping from S-RNase cytotoxic activity. In self-incompatible cross (SPI), self S_1 pollen lands on the S_1S_2 style and both S_1- and S_2-RNases enter the S_1 pollen tube. Similar to CPC response, non-self S_2-RNases bind to pollen *S* in the cytosol of the pollen tube. One or more SLF (SLFN) in S_1 pollen tube form $SCF^{S1\text{-}SLFN}$ complexes to tag S_2-RNase with a polyubiquitin chain, resulting in its degradation by the 26S proteasome. In contrast, the recognition domain of self S_1-RNase binds to a hypothetical recognition domain of SLF resulting in the formation of a non-functional $SCF^{S1\text{-}SLFN}$ complex, thus self S-RNase escapes degradation and acts as a cytotoxin to inhibit the pollen tube growth. Image taken from Liu et al. (2014).

In the compartmentalization model, Goldraij et al. (2006) demonstrated by immunolocalization that S-RNases are taken up into pollen tubes sequestered in vacuoles. Moreover, self S-RNases are stable in compatible and incompatible crosses, but HT-B protein is degraded in compatible crosses, maintaining the S-RNases into the vacuolar compartments. Meanwhile, HT-B levels are not affected in incompatible crosses and S-RNases are released into the cytoplasm after disruption of the vacuole membrane (Figure In8). Hence, these authors suggest that the pollen endomembrane system plays a key role in GSI and compartmentalization, instead of S-RNase degradation, is proposed to prevent pollen arrest in compatible crosses. Nevertheless, SLFs are cytoplasmic proteins and, therefore, some S-RNases should exit the luminal compartment in order to interact with the SCF^{SLF} complex, but not for its degradation. Whatever is the function of the SCF^{SLF} complex, this interaction should drive to

maintain or not the integrity of HT-B in incompatible or compatible crosses respectively (McClure et al., 2011).

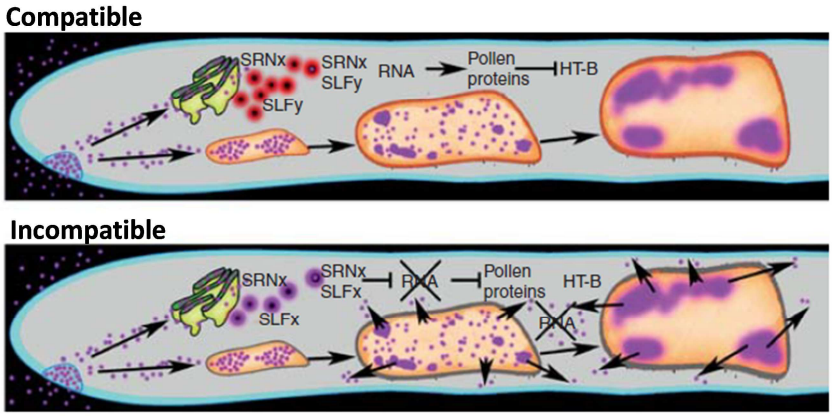


Figure In8. Compartmentalization model in S-RNase-based GSI. Pollen tubes are shown in the pistil extracellular matrix containing a single S-RNase (SRNx, *purple*); although, in a typical S-heterozygote two S-RNases would be present. Compatible (top, Sy-pollen tube in a pistil expressing Sx-RNase) and incompatible (bottom, Sx-pollen tube) pollinations are shown. S-RNase taken up by endocytosis and trafficking by default to progressively larger vacuoles in more mature regions of the pollen tube. S-RNase must exit the endomembrane system to interact with SLF; a single SLF (*red*, SLFx; *blue* SLFy) is shown. Degradation of pollen RNA (cross) in incompatible pollen tubes by exit of S-RNases from vacuolar compartments, a process that do not occur in compatible pollen tubes (no cross). HT-B is repressed in compatible cross but it remains stable in incompatible cross. Image taken from McClure et al. (2011).

7. S-RNase based GSI in *Prunus*. Is it a different mechanism?

The GSI mechanism in Rosaceae (including *Prunus*) is based on S-RNases and SLFs as in Solanaceae and Plantaginaceae. Nevertheless, GSI in *Prunus spp*. exhibit striking differences not only with Solanaceae and Plantaginaceae but also with other Rosaceae genera such *Pyrus* and *Malus*.

7.1. *S*-pollen and *S*-pistil determinants in *Prunus*

Prunus S-RNases show high allelic diversity ranging from 30% to 90% in the amino acid sequence (Ushijima et al., 1998) and maintain the five conserved regions in Solanaceae (from C1 to C5). However, instead of C4 domain, *Prunus spp* have a RC4 region, which amino acid composition and localization are slightly different. In addition, there is only one hypervariable region (RHV), unlike the two present in Solanaceae (HVa and HVb) (Ioerger et al., 1991; Xue et al., 1996). Most plant T2-type S-RNases contain only one intron present in the HVa codifying region of Solanaceae and Rosaceae. However, *Prunus* S-RNases posses an additional intron in the junction sequence between the signal peptide and the open reading frame (Figure In9). No functional analyses supporting S-RNases as female *S*-determinant have been shown in

Prunus, but mutations affecting their expression (Watari et al., 2007; Yamane et al., 2003) and structure (Tao et al., 2007) have been found to lead to SC.

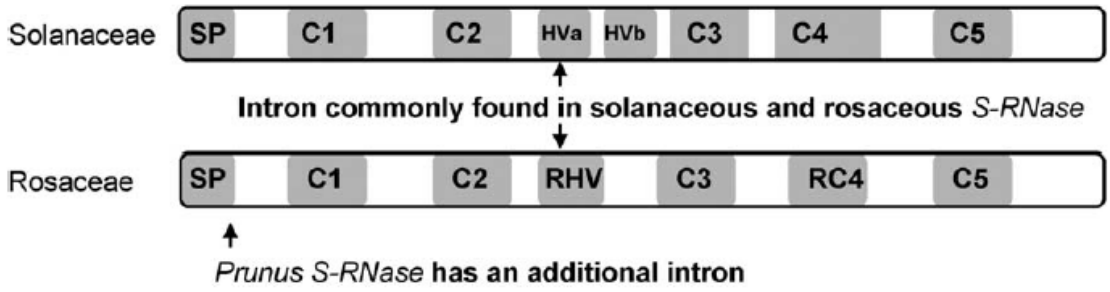


Figure In9. S-RNase structure and positions of intron sequences in S-RNase DNA sequence. Solanaceous and rosaceous S-RNase structures are schematically illustrated. Intron sequences are commonly found in the middle of the coding sequences for HVa and RHV of solanaceous and rosaceous S-RNases, respectively. In addition to this intron, there is another intron in *Prunus* S-RNase, but not in *Malus* and *Pyrus* S-RNase. SP, signal peptide; C1 to C5, conserved regions 1–5; RC4, rosaceous conserved region 4; HVa and HVb, hypervariable regions a and b; RHV, rosaceous hypervariable region. Image taken from Tao & Iezzoni (2010).

The *S*-locus *F-Box* genes of Solanaceae (*SLF*) and *Prunus* (*SFB*) have shown important differences as well. For instance, *SFB*s contain an intron in the 5'UTR region, proved useful for *S*-genotyping (Vaughan et al., 2006), that has not been found in *SLF*. But undoubtedly, the most striking difference is the distinct behavior of mutants where pollen *S*-function was lost. On the pollen side, SC in Solanaceae and Plantaginaceae is always associated with competitive interaction and no mutations affecting SLF function have been found to confer SC (Golz et al., 2001 and 1999). On the contrary, mutations disrupting SFB function and leading to SC have been widely described in *Prunus* (Hauck et al., 2006; Hegedus et al., 2012; Tao & Iezzoni, 2010; Vilanova et al., 2006) (Figure In10).

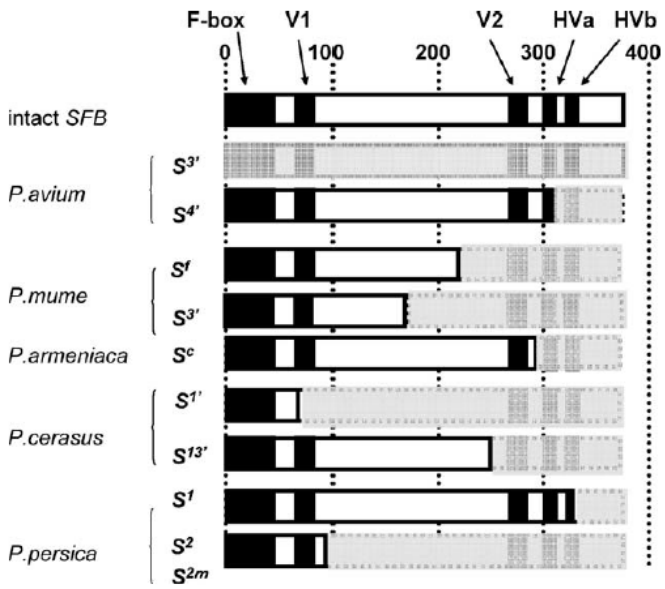


Figure In10. Schematic illustration of intact and mutated SFBs. Number of amino acid residues from the N-terminal is indicated over the intact SFB. All but *P. avium* $S^{3'}$ encode truncated SFBs. *P. avium* $S^{3'}$ is completely deleted from the genome. The truncated portion of SFB is indicated by half tone. V1 and V2, variable regions 1 and 2; HVa and HVb, hypervariable regions a and b. Image taken from Tao & Iezzoni (2010).

In this context, two different theories have been suggested for the evolution of *S*-factors involved in Solanaceae and Rosaceae GSI. One proposes that SI has evolved independently on several occasions (De Franceschi et al., 2011), while the second suggests a divergence process from a common ancestor among the eudicots before asterids and rosids division (Igic & Kohn, 2001; Steinbachs & Holsinger, 2002; Vieira et al., 2008). Recent works trying to shed some light on this point have been carried out by phylogenetic and expression analyses of Rosaceae *S*-factors. Segmental duplications seem to have occurred in a common Rosaceae ancestor, where three different *S*-loci might be involved throughout Rosaceae evolutionary history (Aguiar et al., 2015; Morimoto et al., 2015). Thus, the functional *S*-locus from *Maloideae* and *Prunoideae* are not orthologous, but they had evolved from different lineages recruiting different paralogous genes to determine each SI mechanism (Aguiar et al., 2015; Ashkani & Rees, 2015; Morimoto et al., 2015). It is also noteworthy that *F-Box* genes having the highest similarity to Solanaceous *SLF* genes, designated as *SLFL* (*SLF*-like) genes, are located in the vicinity of *Prunus S*-locus at the end of linkage group 6 (Aguiar et al., 2015; Morimoto et al., 2015) (Figure In11).

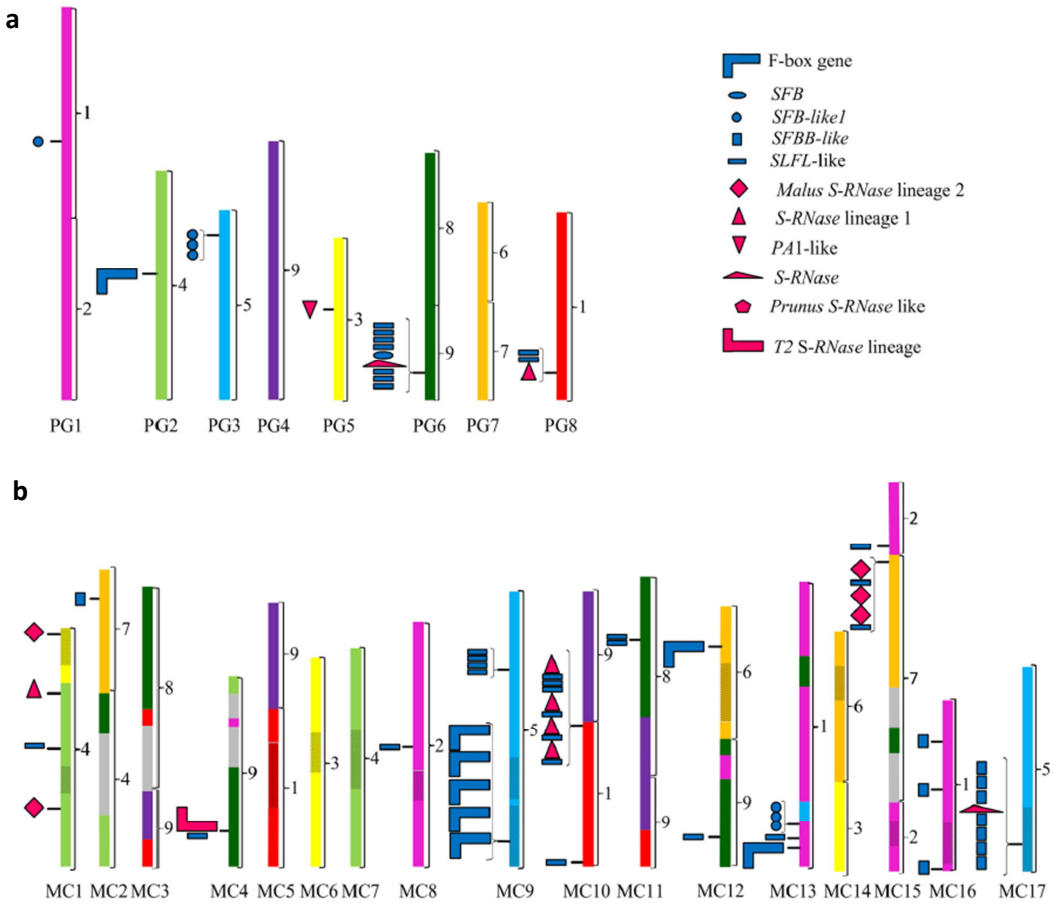


Figure In11. Chromosomal localization of the S-RNase, SFB, SFBB, and SLFL lineage genes in *P. persica* (a) and *M. domestica* (b). S-RNase lineage genes are marked in *pink* and *SFB*, *SFBB*, and *SLFL* lineage genes are marked in *blue*. *Different shapes* represent the different *S-RNase* and *Fbox SFB-*, *SFBB*, and *SLFL*- lineage genes. To represent two or more sequential genes, a *bracket at the left* of the chromosome is used. Each *Prunus* chromosome is marked in a different color: PG1- pink, PG2 light green, PG3 light blue, PG4- purple, PG5- yellow, PG6-green, PG7- orange, and PG8-red. These colors are then used to assign the synteny regions for the *M. domestica* chromosomes, according to Fig 1 in Jung et al. (2012). Regions with unknown synteny but between regions that show synteny with the same chromosome are marked in *stripes*, and regions with unknown synteny between syntenic regions from different chromosomes are marked in *grey*. *Brackets on the right* of each chromosome represent the nine ancestral synteny regions (1 to 9) according to Fig 4 in Illa et al. (2011). Image taken from Aguiar et al. (2015).

7.2. Modifier factors identified in *Prunus*

Numerous genetic evidences supporting pollen modifiers have been reported in *Prunus* mainly relying on *S*-locus unlinked pollen-part mutations conferring SC. These type of mutations were firstly reported in sweet cherry (*Prunus avium*) cv. 'Cristobalina' (Wunsch & Hormaza, 2004) and apricot (*Prunus armeniaca*) cv. 'Canino' (Vilanova et al., 2006) (see section 8), and more recently in Japanese apricot (*Prunus mume*) cv. 'Zaohong' (Wang et al., 2013), Japanese plum (*Prunus salicina*) cv. 'Methley' (Beppu et al., 2015) and sweet cherry cvs. 'Son Miró' and 'Talegal Ahín' (Cachi & Wünsch, 2014). Nevertheless, none of these putative mutated modifiers have

been identified to date. On the other hand, Matsumoto et al. (2012) have successfully identified E3 ubiquitin ligase components (PavSSK1 and PavCul1) in *Prunus avium* that, as a necessary part of the SCF^{SLF} complex, can also be considered modifiers. In addition, homologs to SCF^{SLF} components have also been found in *Malus* and *Pyrus* (Minamikawa et al., 2014; Xu et al., 2013; Yuan et al., 2014). More recently, *Prunus* orthologous gene to *SBP*1 has been identified, but this has not been shown to interact with F-box proteins or S-RNases (Matsumoto & Tao, 2016).

7.3.Biochemical model proposed for *Prunus* GSI

The main difference between GSI mechanisms proposed for Solanaceae, Plantaginaceae and *Maloideae* on one hand, and *Prunus* on the other, concerns to the expected function for the male *S*-determinant. Reinforcing this point, competitive interaction, commonly detected in Solanaceae, is absent in *Prunus*. In fact, heteroallelic pollen does not drive to GSI breakdown in tetraploid *Prunus cerasus* self-compatible cultivars. This response in uniquely obtained by the accumulation of non-functional *S*-alleles like in diploid *Prunus spp*. (Hauck et al., 2006). This distinct behavior has led to propose a self-recognition mechanism in *Prunus* (equivalent to those in Brassicaceae SSI and Papaver GSI) instead of a non-self recognition mechanism as it has been established for the rest of species exhibiting S-RNase-based GSI (Matsumoto & Tao, 2016; Sassa, 2016; Tao & Iezzoni, 2010). In this sense, Tao & Iezzoni (2010) already proposed a model where S-RNases are not the substrate for SCF complex but a S-RNase inhibitor (general inhibitor, GI) that reversibly interacts with and inactivates the S-RNase. Matsumoto & Tao (2016) have recently proved that SCF complex binds $SLFL_2$ protein, which interacts *in vitro* with all S-RNases tested. Hence, they hypothesize that $SLFL_2$ is a good candidate for being the GI, which polyubiquitinates both self- and non-self S-RNases. Meanwhile, SFBs should recognize its cognate self-S-RNase and protect it from degradation by the GI releasing self-S-RNases to accomplish their cytotoxic activity. Nonetheless, this model needs to be tested *in planta* in order to validate SLFL function and identify SFB interacting protein. Furthermore, few modifiers have been identified in *Prunus* when compared with Solanaceae and they are crucial for the characterization of the distinct biochemical model operating in *Prunus*.

8. The self-compatible apricot cultivar 'Canino'. A case of study

8.1. Genetic and molecular analysis

The apricot cultivar 'Canino' (S_2S_C) was found to contain two different types of mutations conferring SC. On one hand, the S_C-haplotype bears an insertion of 358-bp in the SFB_C gene that produces a truncated protein leading to the loss of pollen *S*-function. On the other, a mutation in a modifier gene gametophytically expressed in the pollen side provokes, independently, the loss of pollen *S*-activity. Segregation analysis of *S*-genotypes performed in different controlled crosses using 'Canino' as male and female parent showed that segregation rates fit with the expected rates for a mutation in heterozygosis, outside of the *S*-locus and expressed in pollen conferring SC (Table In1) (Vilanova et al., 2006). Molecular analyses discarded mutations in the specific *S*-determinants (*SFB* and *S-RNase*), as well as miss-expression of these genes or even allele duplications (heteroallelic pollen) as possible causes of the SC phenotype. The locus containing this mutation was named *M*-locus (from modifier) and belongs to the group 2 of modifier types (required for pollen rejection but have no wider role in pollination) (Vilanova et al., 2006).

Table In1. a) Expected gamete and seedling genotypes formed from the outcross 'Goldrich' (S_1S_2) X 'Canino' (S_2S_C) and the selfing of 'Canino' (S_2S_C) considering 'Canino' heterozygous for a pollen-part mutation unlinked to the *S*-locus (*Mm*). b) Segregation of the *S-RNase* alleles in the progenies of controlled field crosses and self-pollinations *S*-genotypes were determined by PCR. Observed *S-RNase* genotypes, expected segregation ratios, and χ^2 values obtained for each population are indicated. Tables taken from Vilanova et al. (2006).

a

Female Goldrich (S_1S_2 *MM*)/ Male Canino (S_2S_C *Mm*)	S_2M[a]	S_2m	S_CM	S_Cm[a]
S_1M	X[b]	S_1S_2 *Mm*	S_1S_C *MM*	S_1S_C *Mm*
S_2M	X	S_2S_2 *Mm*	S_2S_C *MM*	S_2S_C *Mm*
Female/Male Canino (S_2S_C Mm)	S_2M[a]	S_2m	S_CM	S_Cm[a]
S_2M	X	S_2S_2 *Mm*	S_2S_C *MM*	S_2S_C *Mm*
S_2m	X	S_2S_2 *mm*	S_2S_C *Mm*	S_2S_C *mm*
S_CM	X	S_2S_C *Mm*	S_CS_C *MM*	S_CS_C *Mm*
S_Cm	X	S_2S_C *mm*	S_CS_C *Mm*	S_CS_C *mm*

b

Seed Parent (*S*-Genotype)	Pollen Parent (*S*-Genotype)	*S*-Genotypes Observed						Total	Expected Segregation Ratio	χ^{2d} (*P* Value)
		S_1S_C	S_2S_C	S_1S_2	S_2S_2	S_1S_1	S_CS_C			
Goldrich (S_1S_2)	Currot (S_CS_C)	31	39	–	–	–	–	70	1:1	0.91 (0.339)
Goldrich (S_1S_2)	Canino (S_2S_C)	66	55	28	22	–	–	171	2:2:1:1[b]	2.98 (0.394)
Canino (S_2S_C)	Canino (S_2S_C)	–	53	–	11	–	35	99	3:1:2[b]	2.20 (0.333)
GC-8 (S_2S_C)[a]	GC-8 (S_2S_C)	–	14	–	4	–	6	24	3:1:2[b]	0.83 (0.659)
GC-10 (S_2S_C)[a]	GC-10 (S_2S_C)	–	14	–	–	–	10	24	1:1[c]	0.67 (0.414)
GC-80 (S_1S_C)[a]	GC-80 (S_1S_C)	15	–	–	–	3	6	24	3:1:2[b]	1.50 (0.472)
GC-86 (S_1S_C)[a]	GC-86 (S_1S_C)	13	–	–	–	–	11	24	1:1[c]	0.17 (0.683)

[a]Seedlings derived from the cross Goldrich (S_1S_2) × Canino (S_2S_C). [b]Expected ratios for a single mutation unlinked to the *S*-locus. [c]Expected ratios for nonmutated GC seedlings. [d]Observed ratios do not differ significantly from expected at $P < 0.05$ in any case.

8.2. Mapping of the *S*-locus unlinked pollen-part mutation conferring SC

Zuriaga et al. (2012) fine-mapped the *M*-locus at the distal end of chr. 3 using 141 individuals from the 'Goldrich' x 'Canino' population. To identify the chromosome bearing *m*-mutation, a strategy based on the identification of molecular markers exhibiting distorted segregations was carried out. According to this premise, molecular markers with the highest distortion were localized in LG3 and LG6 (which contains *S*-locus). Taking into account that *M*- and *S*-loci are unlinked, LG3 was the most probable localization for the *M*-locus. Afterwards, 120 SSR markers designed from chr. 3 were tested in both progenitors and 25 were successfully mapped to construct LG3 in 'Canino'. *M*-locus was flanked by PGS3.71 and PGS3.96 markers in an interval of 1,8 cM that comprised the PGS3.62 marker co-segregating with the *m*-mutation. On the basis of the high collinearity between apricot and peach maps, and according to the peach genomic sequence, a contig was obtained through the identification of BACs from the SI apricot cultivar 'Goldrich'. This contig encompassed approximately 364 Kb and 59 ORFs regarding the syntenic peach region (Figure In12).

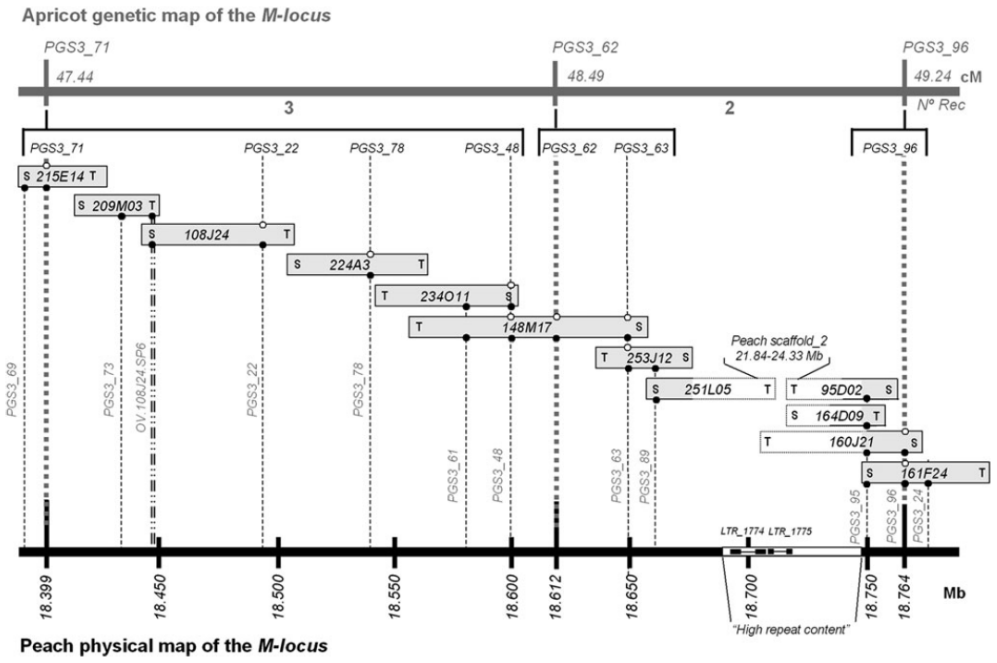


Figure In12. Contig constructed with 'Goldrich' BACs covering the *M*-locus region on the distal part of apricot chr.3 (not to scale). Aligned BACs showing their BAC-ends Sp6 (S) and T7 (T) are represented by *grey boxes*. Miss-aligned fragments are shown as *white boxes*. SSRs amplified from BACs are indicated by *black dots* and those anchored into the 'Goldrich' genetic map are indicated by *white dots*. *Dashed-lines* indicate the SSR positions corresponding to the apricot genetic map and the peach physical map. Distances in centimorgan (cM) are shown at the top for the 'Goldrich' genetic map and those in megabases (Mb) are shown down below for the peach physical map. *Nº Rec* indicates the number of recombinants found in 'GxC-01' corresponding to 'Goldrich'. Image taken from Zuriaga et al. (2012).

MAIN OBJECTIVES

The general aim of this thesis was to investigate Gametophytic Self-Incompatibility (GSI) system in *Prunus* by studying modifier factors involved in the underlying mechanism. For this purpose, the following specific objectives were addressed:

1. Genetic and molecular characterization of the self-compatible apricot cultivar 'Katy'. Fine-mapping of the mutation conferring this phenotype.

2. To screen for new mutations conferring self-compatibility in apricot by phenotyping this trait and genotyping the *S*- and *M*-locus in a set of cultivars/accessions with distinct geographic origins.

3. Identification and cloning of the apricot *M*-locus modifier gene by using an integral strategy based on NGS genomic and transcriptomic data.

4. To perform a comparative study of the *S*-RNase based GSI system in Rosaceae and Solanaceae by analyzing orthology relationships between modifier factors.

Chapter 1:
An *S*-locus independent pollen factor confers Self-Compatibility in 'Katy' apricot

This work has been published in:

Zuriaga E*, Muñoz-Sanz JV*, Molina L, Gisbert AD, Badenes ML, et al. (2013) An S-Locus Independent Pollen Factor Confers Self-Compatibility in 'Katy' Apricot. PLoS ONE 8(1): e53947. doi:10.1371/journal.pone.0053947.

and it has been presented in the international congress:

Muñoz-Sanz JV, Zuriaga E, Badenes ML and Romero C. "Self-compatibility associated with pollen(s) modifier genes in apricot". Rosaceous Genomics Conference 6 (Trento, Italy). 02/10/2012.

Abstract

Loss of pollen-*S* function in *Prunus* self-compatible cultivars has been mostly associated with deletions or insertions in the *S*-haplotype-specific F-box (*SFB*) genes. However, self-compatible pollen-part mutants defective for non-*S*-locus factors have also been found, for instance, in the apricot (*Prunus armeniaca*) cv. 'Canino'. In the present study, we report the genetic and molecular analysis of another self-compatible apricot cv. termed 'Katy'. *S*-genotype of 'Katy' was determined as S_1S_2 and *S-RNase* PCR-typing of selfing and outcrossing populations from 'Katy' showed that pollen gametes bearing either the S_1- or the S_2-haplotype were able to overcome self-incompatibility (SI) barriers. Sequence analyses showed no SNP or indel affecting the SFB_1 and SFB_2 alleles from 'Katy' and, moreover, no evidence of pollen-*S* duplication was found. As a whole, the obtained results are compatible with the hypothesis that the loss-of-function of a *S*-locus unlinked factor gametophytically expressed in pollen (*M'*-locus) leads to SI breakdown in 'Katy'. A mapping strategy based on segregation distortion loci mapped the *M'*-locus within an interval of 9.4 cM at the distal end of chr.3 corresponding to ~1.29 Mb in the peach (*Prunus persica*) genome. Interestingly, pollen-part mutations (PPMs) causing self-compatibility (SC) in the apricot cvs. 'Canino' and 'Katy' are located within an overlapping region of ~273 Kb in chr.3. No evidence is yet available to discern if they affect the same gene or not, but molecular markers seem to indicate that both cultivars are genetically unrelated suggesting that every PPM may have arisen independently. Further research will be necessary to reveal the precise nature of 'Katy' PPM, but fine-mapping already enables SC marker-assisted selection and paves the way for future positional cloning of the underlying gene.

Introduction

Gametophytic self-incompatibility (GSI) is a widespread mechanism in the plant kingdom that prevents inbreeding (de Nettancourt, 2001). In Solanaceae, Plantaginaceae and Rosaceae GSI is controlled by the *S*-locus that contains at least two genes coding for S-RNase and F-box proteins. S-RNases are style-specific expressed and their ribonuclease activity is essential for self-pollen rejection (McClure et al., 1989; Boskovic et al., 1996; Xue et al., 1996). In turn, the *S*-locus F-box proteins (*SLF* or *SFB*) are the pollen *S*-determinants (Lai et al., 2002; Sijacic et al., 2004; Ushijima et al.,

2003). Evidence accumulated in *Petunia* and *Antirrhinum* supports a model in which SLFs are components of a SCF E3 ubiquitin ligase complex that interacts with non-self S-RNases leading to their ubiquitination and degradation by the 26S proteasome proteolytic pathway (Hua & Kao, 2006; Huang et al., 2006). Alternately, the compartmentalization model proposed by Goldraij et al. (2006) in *Nicotiana* explains the resistance to non-self S-RNases by their sequestration in vacuolar compartments of pollen compatible tubes. A hypothetical S-RNase endosome sorting model involving both S-RNase degradation and compartmentalization has been recently proposed (Chen et al., 2010), but many pieces of the puzzle remain elusive.

Spontaneous and induced self-compatible mutants have been particularly important to support *S-RNase* and *S*-locus *F-box* genes as the *S*-determinants in *Prunus* (Rosaceae) since other functional approaches based on transgenic experiments are seriously hindered in this genus. For instance, a *Mu*-like element insertion upstream of the S_{6m}-*RNase* in sour cherry (*Prunus cerasus*) (Yamane et al., 2003) and a similar mutation in the Japanese plum (*Prunus salicina*) S^{e}-*RNase* (Watari et al., 2007) reduce the *S-RNase* expression level leading to a insufficient accumulation of S-RNase in the pistil which breaks the rejection mechanism. Modifications affecting the S-RNase structure and conferring self-compatibility (SC) have also been found in peach (*Prunus persica*) where the S^{2m}-RNase shows a reduced stability as a consequence of the cysteine residue replacement by a tyrosine in the C5 domain (Tao et al., 2007). Regarding the pollen-part mutations (PPM), self-compatible mutants with non-functional *SFB* genes have been identified in sweet cherry (*Prunus avium*) (Ushijima et al., 2004; Sonnelveld et al., 2005; Marchese et al., 2007), apricot (*Prunus armeniaca*) (Vilanova et al., 2006), sour cherry (Hauck et al., 2006), Japanese apricot (*Prunus mume*) (Ushijima et al., 2004) and peach (Tao et al., 2007), supporting their role as the pollen-*S* determinants in this genus. In most of these cases, the self-compatible phenotype was associated with indels in the *SFB* codifying region causing a frame-shift in translation that produces a non-functional truncated protein (Yamane & Tao, 2009). This seems to be a specific feature of the S-RNase based GSI system operating in *Prunus*, since in Solanaceae the only pollen-side mutations found to cause SC are due to the *S*-heteroallelic pollen effect (Golz et al., 1999). Therefore, *SLF* mutations were initially suggested to confer SI or lethality, but recent findings provide an alternative explanation since in the non-self recognition by multiple factors SI system, shown to operate in Solanaceae (Kubo et al., 2010) and *Pyrus* (Rosaceae) (Kakui et al., 2011), the

loss of pollen-*S* function does not lead to SC. In contrast, all loss-of-function mutations found in *Prunus SFB* cause SC which may support differences in the self-recognition mechanism where the SFB target would be an S-RNase 'inhibitor' instead of the S-RNase itself (Tao & Iezzoni, 2010). Nevertheless, even considering the discrepancies, major similarities (i.e. *S-RNase* and *SLF/SFB* as *S*-specificity determinants) are still more striking and the model as a whole might be preserved across families (McClure et al., 2011).

As reported above, self-compatible accessions found in Rosaceae are mostly related to mutations in pistil and pollen *S*-locus determinants (Yamane & Tao, 2009). However, mutations in non *S*-locus factors have also been associated with SC in sweet cherry (Wünsch & Hormaza, 2004), almond (*Prunus amygdalus*) (Fernández et al., 2009) and diploid strawberries (*Fragaria spp*.) (Boskovic et al., 2010). Genetic evidence for *S*-locus unlinked factors required for GSI, also called modifier genes, was previoulsy accumulated in Solanaceae. For instance, Ai et al. (1991) showed that the self-compatible *Petunia hybrida* cv. 'Strawberry Daddy' (S_OS_X) accumulates a non-functional *S*-allele (S_O) and a stylar mutation in an additional factor necessary for SI. Later studies in *Nicotiana* revealed that the so called 4936 stylar factor is also required for SI (McClure et al., 2000). Moreover, mutations in modifier loci affecting the pollen-*S* function have been suggested to explain SI breakdown in *Solanum tuberosum* (Thompson et al., 1991) and *Petunia axillaris* (Tsukamoto et al., 2003). More intriguing is the behaviour of the PPM found in *Solanum chacoense* that predicts a *S*-locus inhibitor (*Sli*) gene acting as a single dominant factor that displays sporophytic inhibition of SI (Hosaka & Hanneman, 1998a; 1998b) . More recently, some stylar modifier factors have been identified and successfully cloned in *Nicotiana*, such as the small asparagine-rich protein HT-B (McClure et al., 1999), the 120K glycoprotein (Hancock et al., 2005) and the Kunitz-type proteinase inhibitor NaStEP (Busot et al., 2008) but their role in SI still has not been completely elucidated. Pollen modifier factors have also been identified in the Solanaceae, such as the *Petunia* pollen-expressed Skp1-like protein PhSSK1 proposed to be acting as adaptor in the SCF complex (Zhao et al., 2010). Interestingly, Matsumoto et al. (2012) have identified a similar SFB-interacting Skp1-like protein (PavSSK1) in sweet cherry and suggest that it could also be a functional component of the SCF complex. Nevertheless, the identification of additional GSI modifier factors will be necessary to dissect completely the underlying mechanism in *Prunus*.

In apricot, the cv. 'Canino' (S_2S_CMm) was found to contain two different mutations conferring SC, an insertion in the SFB_C gene that produces an SFB_C truncated protein and a mutation in a modifier gene (*m*) unlinked to the *S*-locus, both independently causing the loss of pollen-*S* function (Vilanova et al., 2006; Zuriaga et al., 2012). In this work, we have analyzed the self-compatible apricot cv. 'Katy' using genetic and molecular approaches, and the compiled evidence suggest that the loss of function of an *S*-locus unlinked factor (*M*'-locus) is also involved in pollen-*S* function breakdown in this case. According to the current knowledge on GSI in *Prunus* the possible roles for the mutated modifier gene are discussed. In addition, we have paved the way for future positional cloning of the 'Katy' pollen-part modifier gene by fine-mapping the *M'*-locus to the distal part of apricot chr. 3. Macro- and micro-synteny of this region has been studied by comparing with the *M*-locus in 'Canino' and by analyzing the ORFs comprised in the peach syntenic region according to the peach genome v1.0 (International Peach Genome Initiative - IPGI; http://www.rosaceae.org/peach/genome).

Results

'Katy' is an apricot self-compatible cultivar with *S*-genotype S_1S_2

'Katy' is an apricot variety developed by Zaiger´s Genetics (Modesto, CA, USA) and reported as self-fruitful (Rusell, 1998). In this study, SC of this cultivar was confirmed by self-pollination in the field (Table 1.1). To determine the *S*-genotype of 'Katy', fragments containing the first intron of the *S-RNases* were PCR-amplified using the SRc-F/SRc-R primers (Figure 1.1a). These fragments were assigned to S_1 and S_2-alleles by comparison with known *S*-genotypes, following the nomenclature established by Burgos et al. (1998). This *S*-genotype was confirmed by the amplification of the second intron using the primers Pru-C2/Pru-C4R (Tao et al., 1999) since fragment sizes obtained were coincident with those expected for the S_1 and S_2-alleles (Figure 1.1b). In addition, PCR-amplified fragments spanning the first intron, were sequenced and compared with GenBank accessions, being identical to the already identified *Prunus armeniaca S*-RNases 1 and 2. The alignment of their deduced amino acid sequences (44 aa) showed the presence of the C1 and C2 *Prunus S*-RNase conserved domains along with the hypervariable region HV1 located between them (Romero et al., 2004).

Table 1.1. Segregation of the S-RNase alleles in progenies of self-pollinations and outcrosses performed with the self-compatible cultivar 'Katy'. Observed *S-RNase* genotypes, expected segregation ratios and χ^2 values obtained for each population are indicated.

Seed parent (*S*-genotype)[a]	**Pollen parent (*S*-genotype)**	**Population name**	**N[c]**								**Exp. Ratio[d]**	**χ^{2e} *P*-value**
				***S*$_1$*S*$_C$**	***S*$_2$*S*$_C$**	***S*$_1$*S*$_2$**	***S*$_2$*S*$_2$**	***S*$_1$*S*$_1$**	***S*$_1$*S*$_4$**	***S*$_2$*S*$_4$**		
Katy (S_1S_2)	Katy (S_1S_2)	'K×K'[b]	94	--	--	45	33	16	--	--	2:1:1	6.32 (0.04)
Katy (S_1S_2)	Goldrich (S_1S_2)	---	0	--	--	--	--	--	--	--	--	--
Goldrich (S_1S_2)	Katy (S_1S_2)	'G×K'	26	--	--	12	10	4	--	--	2:1:1	2.92 (0.23)
Harcot (S_1S_4)	Katy (S_1S_2)	'H×K'	44	--	--	20	--	4	7	13	2:1:1:2	3.68 (0.30)
Katy (S_1S_2)	Canino (S_2S_C)	'K×C'	50	15	19	6	10	--	--	--	2:2:1:1	1.49 (0.69)
Canino (S_2S_C)	Katy (S_1S_2)	'C×K'	88	32	15	29	12	--	--	--	2:1:2:1	0.74 (0.86)

[a] *S*-genotypes for 'Goldrich', 'Harcot' and 'Canino' were prevously reported by Vilanova et al. (2005) and the *S*-genotype for 'Katy' was determined in this work

[b] 'K×K' data correspond to three combined F_2 populations obtained by self-pollinating 'Katy' in 2005, 2006 and 2010.

[c] Obtained seedlings

[d] Expected ratios for a single mutation unlinked to the *S*-locus

[e] Observed ratios do not differ significantly from expected at $P < 0.05$ (barring 'Katy' self-pollination at $P > 0.01$)

SC in 'Katy' is associated with a PPM unlinked to the *S*-locus

To analyze the nature of SC in 'Katy', this cultivar was self-pollinated and reciprocally crossed with 'Goldrich', a self-incompatible cultivar sharing the same *S*-genotype. *S-RNase* genotyping of the progenies derived from the 'Katy' (S_1S_2) self-pollination (Figure 1.1c) and the 'Goldrich' (S_1S_2) × 'Katy' (S_1S_2) outcross (Figure 1.1d) revealed three different *S*-genotypes (S_1S_1:S_1S_2:S_2S_2) in both cases (Table 1.1). In turn, the 'Katy' (S_1S_2) × 'Goldrich' (S_1S_2) cross did not produce any seedling. Thus, 'Katy' pollen is able to grow through the 'Goldrich' pistil meanwhile 'Goldrich' pollen is rejected in the 'Katy' styles. According to these results, SI breakdown in 'Katy' may be due to a pollen-part mutation since 'Katy' is completely functional as a female parent. Indirect evidence supporting this hypothesis was also compiled from the *S*-genotype segregation ratio in 'K×C', because the number of S_2 bearing genotypes is lower than that expected for a non-functional pistil-S_2 determinant (Table 1.1). Moreover, both 'Katy' *S*-alleles are able to grow in 'Goldrich' and 'Katy' styles suggesting that the PPM is unlinked to the *S*-locus.

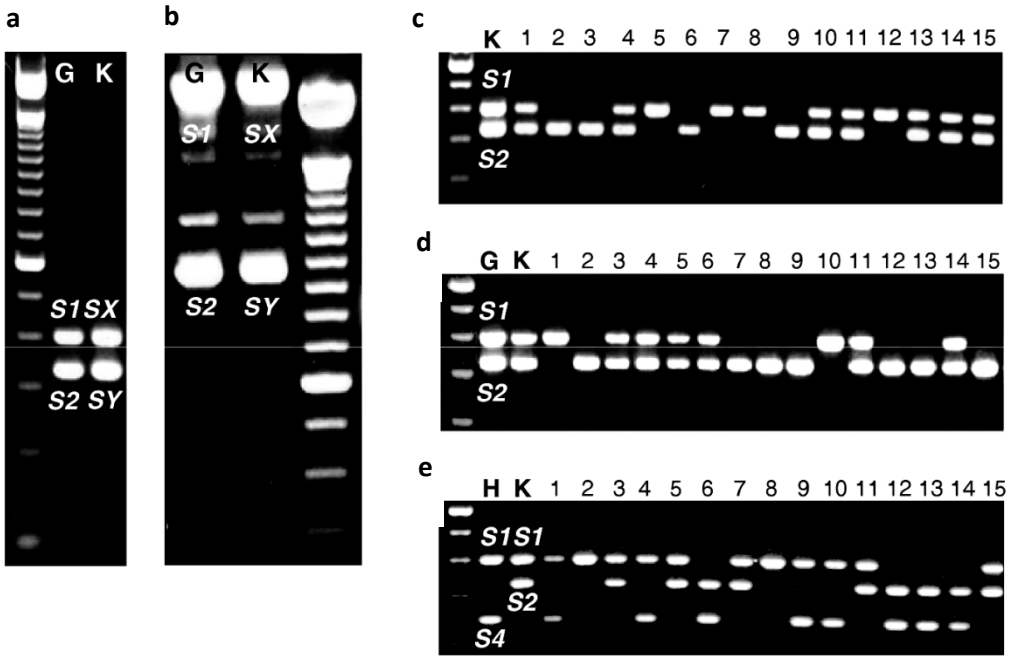


Figure 1.1. Determination of the 'Katy' *S*-genotype and analysis of *S*-alleles segregation in selfing and outcrossing populations derived from 'Katy'. PCR amplification of apricot genomic DNA with consensus primers for the 1st (a) and 2nd (b) *S-RNase* intron. Samples in (a) and (b) are as follows: (G) Goldrich (S_1S_2) and (K) Katy (S_xS_y). *S*-RNase allele fragments PCR-amplified with SRc-F/SRc-R primers from the 'K×K' (c), 'G×K' (d) and 'H×K' (e) progenies. Samples are as follows: (K) Katy (S_1S_2), (G) Goldrich (S_1S_2) and (H) Harcot (S_1S_4) and 15 seedlings from each cross.

To complement these observations, we performed additional crosses with cultivars having different *S*-genotypes. Figure 1.1e shows the *S-RNase* genotyping of the 'Harcot' (S_1S_4) × 'Katy' (S_1S_2) population where *S*-genotypes fell into four classes (S_1S_1:S_1S_2:S_1S_4:S_2S_4) (Table 1.1). Two of these *S*-genotypes were unexpectedly obtained (S_1S_1 and S_1S_4) since pollen tubes carrying the S_1-haplotype from 'Katy' were expected to be incompatible in 'Harcot' styles. On the other hand, reciprocal crosses with the cv. 'Canino' (S_2S_C *Mm*) produced four *S*-genotype classes (S_2S_C:S_2S_2:S_1S_C:S_1S_2). According to the two unlinked PPMs associated wtih SC in 'Canino' (S_C and *m*), these four *S*-genotypes were expected for the 'K×C' progeny (Table 1.1). Nevertheless, since pollen tubes having the S_2-haplotype should be arrested in S_2-styles, the S_2S_C and S_2S_2 genotypes observed in the 'C×K' progeny were unexpected. The observed ratios for *S*-genotype segregations in 'H×K' and 'C×K' fit with that expected in a model where 'Katy' carries a heterozygous PPM affecting pollen-*S* function that is unlinked to the *S*-locus (2:2:1:1) with χ^2 values of 3.68 and 0.74 (P=0.30 and P=0.86) (Tables 1.1 and 1.2). On the contrary, if we consider an heterozygous PPM linked in coupling to the incompatible *S*-allele or an homozygous PPM (linked or unlinked to the *S*-locus) the

expected ratios (1:1:1:1) do not fit with the observed data with χ^2 values of 13.6 and 13.5, respectively (P<0.004).

All performed crosses were shown to be compatible, barring 'Katy × Goldrich' cross, and fruit set ranged approximately from 15% ('K×K') to 34% ('C×K'). Differences in germination rate and seedling fitness were striking. Only 59% of the 'K×K' inbred seeds produced healthy plants while this percentage increased to 82-96% in the outcrossed seeds.

Table 1.2. Expected gamete and seedling genotypes formed from the outcross 'Harcot' (S_1S_4) x 'Katy' (S_1S_2) and the selfing of 'Katy' (S_1S_2) considering 'Katy' heterozygous for a pollen-part mutation unlinked to the *S*-locus (*M'm'*)

Female gametes	**Male gametes** 'Katy' (S_1S_2 *M'm'*)			
'Harcot' (S_1S_4 *M'M'*)	***S*$_1$*M'*** [b]	***S*$_1$*m'***	***S*$_2$*M'***	***S*$_2$*m'***
S*$_1$*M'	X [a]	S_1S_1 *M'm'*	S_1S_2 *M'M'*	S_1S_2 *M'm'*
S*$_4$*M'	X	S_1S_4 *M'm'*	S_2S_4 *M'M'*	S_2S_4 *M'm'*
'Katy' (S_1S_2 *M'm'*)	***S*$_1$*M'***	***S*$_1$*m'***	***S*$_2$*M'***	***S*$_2$*m'*** [b]
S*$_1$*M'	X	S_1S_1 *M'm'*	X	S_1S_2 *M'm'*
S*$_1$*m'	X	S_1S_1 *m'm'*	X	S_1S_2 *m'm'*
S*$_2$*M'	X	S_1S_2 *M'm'*	X	S_2S_2 *M'm'*
S*$_2$*m'	X	S_1S_2 *m'm'*	X	S_2S_2 *m'm'*

[a] Pollen incompatibility

[b] If *m'* was linked in coupling with S_2 the S_2M' and S_1m' gametes from 'Katy' would not be formed, and conversely if *m'* was linked in coupling with S_1 the S_1M' and S_2m' gametes would not be formed.

Molecular analysis of the self-compatible cv. 'Katy' (S_1S_2)

To test whether the 'Katy' pollen tubes are not rejected in pistils bearing a matching *S*-allele as a consequence of SNPs or indels affecting SFB_1 and SFB_2, genomic DNA fragments containing both alleles were cloned and sequenced. Genomic sequences of S_1 and S_2-haplotype regions from the self-incompatible cv. Goldrich (S_1S_2) were used as references (Romero et al., 2004). No changes were found in the nucleotide sequences of the two cloned fragments (approximately 1.3 and 1.9 kb, respectively) containing the complete SFB_1 and SFB_2 open reading frames as well as their 5' and 3' adjacent flanking regions (~110/390 and ~70/470 bp from the 5' and 3' SFB_1/SFB_2 flanking regions, respectively).

PPMs identified in Solanaceae are mostly associated with *S*-allele duplications caused by polyploidy or induced mutations (Golz et al., 2001). To discard this reason,

we first examined the ploidy level in 'Katy' by flow cytometry analysis. The peaks of nuclei isolated from 'Katy' were coincident with those detected in the control diploid plant ('Goldrich'), indicating that 'Katy' is a diploid (data not shown). A hypothetical duplication of the *SFB* alleles in 'Katy' was also tested by a real-time PCR-based gene dosage assay, but the relative DNA amounts detected for SFB_1 and SFB_2 were not significantly different between 'Katy' and the self-incompatible cv. 'Goldrich' (Figure 1.2).

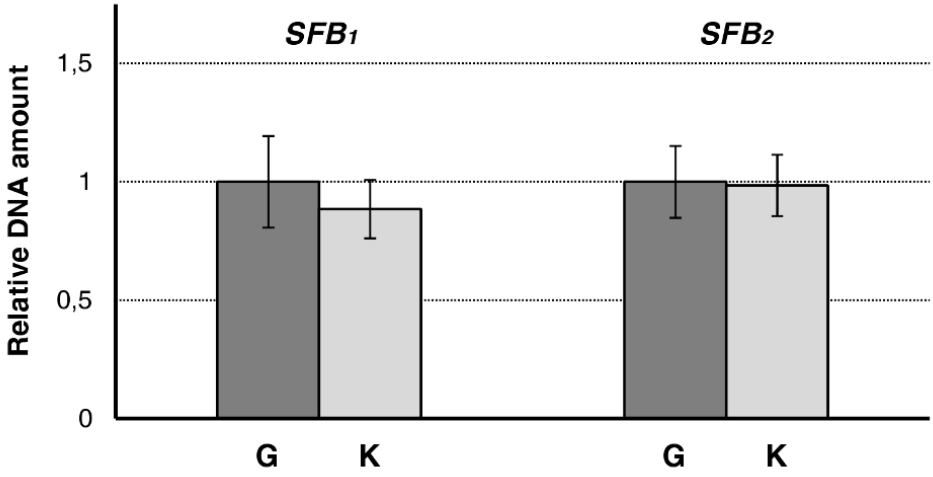


Figure 1.2. Relative DNA amount of SFB_1 and SFB_2 in 'Goldrich' (G) and 'Katy' (K). Quantities correspond to the average of two independent biological replicates repeated three times and were determined using *actin* as endogenous control. Bars indicate standard deviations.

Gene expression analysis showed that SFB_1 and SFB_2 alleles are specifically expressed in pollen in 'Katy' and 'Goldrich' (data not shown). Furthermore, relative transcript abundance of SFB_1 and SFB_2 in 'Katy' and 'Goldrich' was quantified by real-time RT-PCR using *actin* as endogenous control to normalize transcription values. No significant differences in the transcript levels were found for any of the two *SFB* alleles between 'Katy' and the self-incompatible cv. 'Goldrich' (Figure 1.3) discarding transcriptional repression of *SFB*s as the cause of SC.

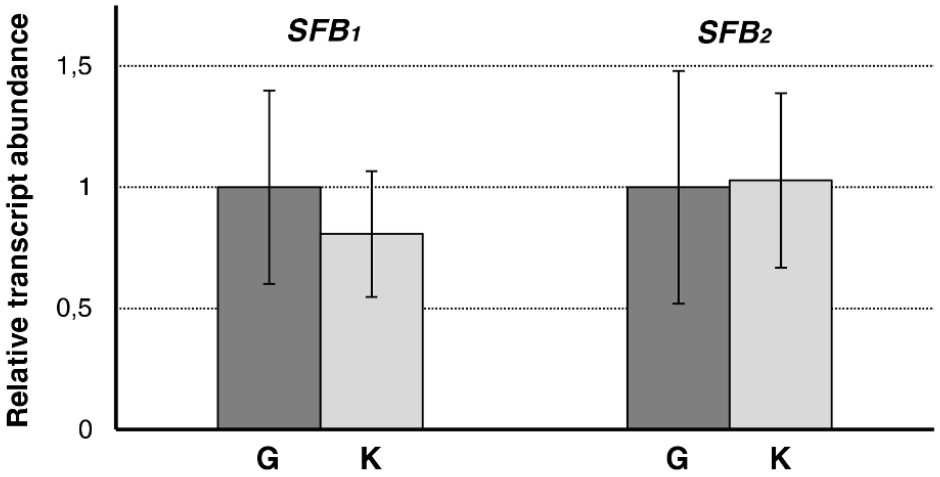


Figure 1.3. Relative transcript abundance of *SFB$_1$* and *SFB$_2$* in 'Goldrich' (G) and 'Katy' (K). Quantities correspond to the average of three independent biological replicates repeated three times. Bars indicate standard deviations.

S-locus unlinked PPM conferring SC in 'Katy' is located on linkage group 3

Overall, genetic and molecular evidence support a model where 'Katy' is heterozygous for a PPM unlinked to the *S*-locus that confers SC. The locus containing this PPM in 'Katy' was referred as *M'*-locus to distinguish it from the *M*-locus previously reported in 'Canino' (Zuriaga et al., 2012). Thus, according to the *S*- and *M'*-locus genotypes, 'Katy' was designated as S_1S_2 *M'm'* (Table 1.2). Under the proposed genetic model, SSR markers linked to the *M'*-locus in 'Katy' selfing populations should be highly distorted, since only seedlings derived from 'Katy' pollen gametes carrying the *m'*-allele (S_1*m'* or S_2*m'*) could be obtained (Table 1.2). Thus, the expected ratio for a SSR marker segregating independently of the *M'*-locus in the F_2 populations is 1:2:1 while that for an absolutely linked SSR is 1:1. On this assumption, genome-wide distributed SSR markers were tested to look for associations with the *M'*-locus. Thereby, 118 SSR markers distributed across the eight *Prunus* chromosomes (ranging from 9 in LG7 to 34 in LG3) were selected for mapping (Tables S1.1 and S1.2). Fifty-five of these SSRs (47%) were found to be polymorphic in 'Katy' and, subsequently, tested in the 'K×K_{05}' and 'K×K_{06}' progenies (Table 1.3). According to the genetic maps constructed for each group, the maximum genetic distance estimated between any pair of markers was ~52cM in LG5 (Table 1.3). In terms of the physical distance, determined from the peach genome sequence, the major gap was found in LG1 (~23Mb). Considering the estimated sizes for the peach genome (~290 Mb) and for the *Prunus* general map (519cM) (Zhebentyayeva et al., 2008), the relationship between

physical and genetic distances is ~0.56 Mb/cM on average. Accordingly, the LG1 23Mb gap should correspond to <45 cM. Consequently, in the most unfavourable scenario, distance to *M'*-locus should be lower than 25cM and recombination frequency lower than 0.25. In this hypothetical case, the expected ratio for a SSR linked to the *M'*-locus would be 1:4:3, and only markers located on LG3 and LG6 fulfill this prediction and show skewed segregations (χ^2> 5.99 with P< 0.05 for 2 d.f.) (Table 1.3).

Table 1.3. Identification of segregation distortion SSR loci distributed throughout the eight linkage groups (LG) of 'Katy' using the 'K×K_{05}' and 'K×K_{06}' populations.

LG	Locus	Peach Mb[a]	Apricot cM[b]	Seg. type[c]	A	H	B	Total	χ^2 (*P*-value)[d]
1	Gol051	4,69	00,0 (0,26)	<abxab>	12	22	12	46	0,09 (0,96)
1	EPPCU0027	9,51	30,7 (0,00)	<abxab>	17	19	9	45	3,93 (0,14)
1	pchcms4	9,51	30,7 (0,36)	<abxab>	18	19	9	46	4,91 (0,09)
1	CPPCT045	32,02	77,5	<abxab>	7	30	9	46	4,44 (0,11)
2	ssrPaCITA19	13,01	00,0 (0,17)	<abxab>	18	18	10	46	4,96 (0,08)
2	UDP98-411	20,17	17,2 (0,18)	<abxab>	13	24	9	46	0,78 (0,67)
2	CPSCT021	23,74	36,9 (0,03)	<abxab>	10	27	9	46	1,44 (0,49)
2	CPSCT031	25,15	40,3	<abxab>	10	26	10	46	0,78 (0,68)
3	ssrPaCITA23	02,70	00,0 (0,17)	<abxab>	8	25	13	46	1,44 (0,49)
3	UDAp468	04,85	18,0 (0,24)	<abxab>	20	16	9	45	9,13 (0,01)[e]
3	PGS3_03	16,41	44,7 (0,23)	<abxab>	5	20	21	46	11,91 (0,003)[e]
3	EPPCU7190	19,78	69,0	<abxab>	18	25	2	45	12,29 (0,002)[e]
4	UDP96-003	08,76	00,0 (0,12)	<abxab>	9	25	12	46	0,74 (0,69)
4	BPPCT040	06,46	12,0 (0,13)	<abxab>	10	27	9	46	1,44 (0,49)
4	UDAp404	---	25,4	<abxab>	12	26	8	46	1,48 (0,48)
5	PGS5_02	00,48	00,0 (0,39)	<abxab>	8	24	12	44	1,09 (0,58)
5	UDAp452	13,76	52,3 (0,35)	<abxab>	8	23	15	46	2,13 (0,34)
5	CPSCT006	11,53	95,1	<abxab>	10	25	11	46	0,39 (0,82)
6	PGS6_04	04,95	00,0 (0,20)	<abxab>	5	23	16	44	5,59 (0,06)
6	UDAp420	08,14	21,6 (0,10)	<abxab>	6	20	20	46	9,30 (0,01)[e]
6	UDAp489	16,82	31,9 (0,09)	<abxab>	18	21	7	46	5,61 (0,06)
6	Ma027a	20,90	41,3 (0,23)	<abxab>	16	25	4	45	6,96 (0,03)[e]
6	ssrPaCITA12	27,84	64,3 (0,03)	<abxab>	7	22	17	46	4,44 (0,11)
6	Locus-*S*	26,45	67,6	<abxab>	6	23	17	46	5,26 (0,07)
7	CPSCT026	10,98	00,0 (0,00)	<abxab>	13	23	10	46	0,39 (0,82)
7	CPPCT022	10,23	00,0 (0,26)	<abxab>	13	23	10	46	0,39 (0,82)
7	CPSCT042	17,08	29,2	<abxab>	10	20	16	46	2,35 (0,31)
8	PGS8_02	03,28	00,0 (0,03)	<abxab>	7	24	7	38	2,63 (0,27)
8	PGS8_05	07,39	03,4 (0,04)	<abxab>	8	25	11	44	1,23 (0,54)
8	UDAp401	10,50	07,2 (0,00)	<abxab>	10	23	12	45	0,20 (0,90)
8	UDAp470	12,61	07,2 (0,05)	<abxab>	10	24	12	46	0,26 (0,88)
8	M6a	15,03	11,8	<abxab>	9	25	11	45	0,73 (0,69)

[a] Marker position (Mb) within the corresponding peach genome scaffolds which sizes were estimated by IPGI (scaffold_1, 46.88Mb; _2, 26.81Mb; _3, 22.02Mb; _4, 30.53Mb; _5, 18.50Mb; _6, 28.90Mb; _7, 22.79Mb and _8, 21.83Mb)

[b] Map position (cM) and rec. frequencies (in brackets) estimated by JoinMap 3.0

[c] Segregation type as per JoinMap 3.0
[d] Chi-square test was performed for the expected ratio 1:2:1 (<abxab>)
[e] Observed ratios differ significantly from expected at $P < 0.05$ for 2 degrees of freedom

In agreement with the segregation of the *S*-genotypes in the analyzed populations, the *M'*-locus is proposed to be unlinked to the *S*-locus (Table 1.1). Therefore, LG3 or a region far from the LG6 distal end, where the *S*-locus is located, are likely positions for the *M'*-locus. To discern between these two possibilities, a more detailed SDL analysis was performed in LG3 (Table 1.4a) and LG6 (Table S1.1) by including the 'K×K_{10}' population and additional markers. On one side, this analysis showed that LG6 distorted markers are partially linked to the *S*-locus (i.e. Ma027a shows a recombination frequency of 0.26 at LOD 3.3 with the *S*-locus). On the other, the magnitude of the segregation distortion detected in LG6 (χ^2= 15.28 with P= 5×10^{-4} for PGS6_07) lower than that found in LG3 (χ^2= 31.30 with P= 1.6×10^{-7} for PGS3_23). This is due to the lower imbalance between homozygous genotypes found in PGS6_07 (7B against 32A) when compared with PGS3_23 (0B against 37A) (Table 1.4a and Table S1.1). It is inferred from the model that pollen gametes carrying SSR alleles linked in repulsion phase with the PPM would not grow into incompatible styles. Therefore, homozygous genotypes for these SSR alleles should not be obtained in the progeny, as observed for the LG3 SSR distorted markers and particularly for PGS3_23 (Table 1.4a). Thus, both arguments support LG3 as the most likely location for the *M'*-locus allowing us to discard LG6.

High-density mapping of the *M'*-locus on chr.3

To construct a high-density map of the *M'*-locus region on chr.3, 102 SSRs identified from the peach scaffold_3 sequence by Zuriaga et al. (2012) (Table S1.2) and 18 additional SSRs available from the GDR website (Jung et al., 2004) were tested in 'Katy'. A higher percentage of these SSRs did not amplify or produced multi-band patterns in 'Katy' (40%) when compared with both 'Goldrich' and 'Canino' (~30%). However, polymorphism of amplified SSRs was similar between 'Goldrich' and 'Katy' (~55%) and significantly higher than that found in 'Canino' (23%) (Table S1.2). Polymorphic SSRs in 'Katy' were tested in 87 trees from the 'K×K' F_2 population. Sixteen of them were mapped, forming a LG3 genetic map of 72cM with an average marker density of 0.22 marker/cM (Table 1.4a). This marker density increased up to 0.62 marker/cM in the region flanked by the most distorted markers PGS3_12 and

AMPA119 (Table 1.4a). An additional LG3 map obtained from the outcrossing population 'C×K' was found to be essentially collinear with the 'K×K' map (sharing >80% markers), except for a single order change between AMPA119 and PGS3_32 (data not shown). The SDL associated with the *M'*-locus region were confirmed by analyzing 60 additional seedlings derived from the outcrosses 'H×K', 'G×K' and 'C×K' for all sixteen LG3 markers (Table 1.4b). These seedlings were selected by their *S*-genotypes, so that they could only be derived from the fertilization with a 'Katy' pollen gamete carrying the PPM (*m*') and, therefore, directly assigned to the *M' m'* genotype (Table 1.2). Skewed segregations in selfing (F_2) and outcrossing populations suggested that the *M'*-locus is roughly located between PGS3_22 and PGS3_28 (Table 1.4).

Table 1.4. Identification of segregation distortion SSR loci distributed throughout the 'Katy' LG3. a) Data corresponding to the 'K×K' F_2 population. b) Data corresponding to the subsets carrying the PPM from the outcrossing populations 'H×K', 'G×K' and 'C×K'.

a)

Locus	Peach Mb[a]	Apricot cM[b]	Seg. type[c]	A	H	B	Total	χ^2 (*P*-value)[d]
MA066a	02,40	00,0 (0,03)	<abxab>	15	46	25	86	2,74 (0,25)
ssrPaCITA23	02,70	02,3 (0,10)	<abxab>	16	44	27	87	2,79 (0,25)
UDAp468	04,85	12,1 (0,08)	<abxab>	16	38	31	85	6,25 (0,04)[e]
BPPCT039	05,80	19,6 (0,30)	<abxab>	13	42	30	85	6,81 (0,03)[e]
PGS3_03	16,41	39,2 (0,07)	<abxab>	4	46	35	85	23,19 (9×10^{-6})[e]
PGS3_12	17,38	46,3 (0,01)	<abxab>	4	44	35	83	23,46 (8×10^{-6})[e]
PGS3_15	17,71	46,9 (0,03)	<abxab>	4	45	32	81	20,36 (4×10^{-5})[e]
PGS3_22	18,49	49,2 (0,03)	<abxab>	3	45	35	83	25,27 (3×10^{-6})[e]
PGS3_23	18,61	51,1 (0,05)	<abxab>	0	48	37	85	33,64 (5e-8)[e]
PGS3_28	19,14	55,1 (0,02)	<abxab>	3	49	31	83	21,60 (2×10^{-5})[e]
PGS3_32	19,60	56,8 (0,00)	<abxab>	4	48	31	83	19,60 (6×10^{-5})[e]
PGS3_33	19,66	56,9 (0,03)	<abxab>	4	50	30	84	19,14 (7×10^{-5})[e]
AMPA119	20,00	59,0 (0,00)	<abxab>	4	47	35	86	23,09 (9×10^{-6})[e]
EPPCU7190	19,78	59,1 (0,10)	<abxab>	4	47	33	84	21,21 (2×10^{-5})[e]
CPDCT027	21,67	67,1 (0,12)	<abxab>	9	40	32	81	13,07 (0,001)[e]
EPPCU0532	22,00	72,0	<abxab>	12	42	21	75	3,24 (0,20)

b)

Locus	Peach Mb[a]	Population	Seg. type[c]	-c	-d	-e	-g	-n	-p	Total	χ^2 (*P*-value)[d]
MA066a	02,40	H×K-/G×K[f]	<efxeg>/<nnxnp>			5	6	11	14	36	0,44 (0,50)
ssrPaCITA23	02,70	H×K/G×K	<efxeg>			16	20			36	0,44 (0,50)
UDAp468	04,85	H×K/C×K	<efxeg>			21	14			35	1,40 (0,24)
BPPCT039	05,80	H×K/C×K	<abxcd>/<efxeg>	3	8	13	9			33	2,46 (0,12)
PGS3_03	16,41	H×K/C×K	<efxeg>			33	2			35	27,46 (1.6e-7)[e]
PGS3_12	17,38	All three	<efxeg>/<nnxnp>			23	1	34	2	60	48,60 (0,00)[e]
PGS3_15	17,71	C×K	<efxeg>			24	0			24	24,00 (9.6e-7)
PGS3_22	18,49	All three	<efxeg>/<nnxnp>			36	0	24	0	60	60,00 (0,00)[e]
PGS3_23	18,61	All three	<efxeg>/<nnxnp>			36	0	24	0	60	60,00 (0,00)[e]
PGS3_28	19,14	All three	<nnxnp>					60	0	60	60,00 (0,00)[e]
PGS3_32	19,60	All three	<efxeg>/<nnxnp>			11	0	48	1	60	56,07 (0,00)[e]
PGS3_33	19,66	All three	<abxcd>/<efxeg>	0	11	48	1			60	56,07 (0,00)[e]
AMPA119	20,00	All three	<efxeg>			59	1			60	56,07 (0,00)[e]
EPPCU7190	19,78	All three	<efxeg>			59	1			60	56,07 (0,00)[e]
CPDCT027	21,67	All three	<abxcd>/<nnxnp>	31	3			19	5	58	30,41 (3e-8)[e]
EPPCU0532	22,00	H×K/G×K	<efxeg>/<nnxnp>			11	0	21	3	35	24,03 (9.5e-7)[e]

[a] Marker position (Mb) within the peach genome scaffold_3 which size estimated by IPGI was 22.02Mb

[b] Map position (cM) and rec. frequencies (in brackets) estimated by JoinMap 3.0

[c] Segregation type as per JoinMap 3.0

[d] Chi-square test was performed for the expected ratios 1:2:1 (<abxab>) (a) and 1:1 (<nn×np>/<ef×eg>/<ab×cd>) (b)

[e] Observed ratios differ significantly from expected at $P < 0.05$ for 2 (a) or 1 degrees of freedom (b)

[f] *S*-genotypes of the selected seedlings were: S_1S_1 and S_1S_4 in 'H×K'; S_2S_2 and S_CS_2 in 'C×K'; S_1S_1, S_1S_2 and S_2S_2 in 'G×K'

To define the *M'*-locus location more consistently, not only considering distortions but also on the basis of genotyping data, an additional mapping strategy was performed. As described above, all 'K×K-F_2' trees could only be derived from pollen gametes with genotype S_1m' or S_2m', having either the *M'm'* or the *m'm'* genotype. To discriminate between these two genotypes the screening of F_3 offsprings was necessary. Thereby, twelve 'K×K-F_2' individuals, with recombination breakpoints mapping to the LG3 region between UDAp468 and CPDCT027, were self-pollinated to obtain F_3 populations. Six of them (K05-15, K05-21, K06-18, K06-25, K06-34 and K06-37) were finally discarded for the analysis due to the low number of embryos obtained (less than 7 in four cases) or because they were redundantly represented (other F_2 individuals with larger F_3 populations have identical SSR genotypes in this genomic region). The six F_3 populations obtained from the remaining F_2 recombinants (K05-12, K05-24; K06-05, K06-06, K06-17 and K06-21) were tested for a subset of 6 SSRs encompassing the *M'*-locus (PGS3_13/PGS3_32 interval) (Table 1.5). Those SSR markers heterozygous in the F_2 recombinant (H) were expected to segregate 1:1 in the F_3 population when the F_2

recombinant had the *M'm'* genotype and 1:2:1 if it had the *m'm'* genotype (Table 1.5). According to the segregation of these markers (A, H or B as per JoinMap 3.0 notation) the *M'*-locus was proposed to be flanked by PGS3_22 and EPPCU7190 markers within an interval of 9.4 cM. Graphical ordering of genotype data enables the positioning of recombination breakpoints to confirm map order (Figure 1.4a).

Table 1.5. *M'*-locus genotyping of trees belonging to the 'K×K05' and 'K×K06' F_2 populations. *M'*-genotypes were determined by PCR-based amplification of SSR markers (PGS3_13, PGS3_15, PGS3_22, PGS3_23, PGS3_28 and PGS3_32) in the F_3 progenies. Number of embryos falling into each genotypic class (A, H or B) are indicated and *bold lines* represent recombination breakpoints.

SSR genotypes of F_3 progenies from 'K×K_{05}' and 'K×K_{06}' F_2 trees													
K05-12	**Gen[a]**	**A**	**H**	**B**	**χ^2 (*P*-value)**	***M'*-locus**	**K06-05**	**Gen**	**A**	**H**	**B**	**χ^2 (*P*-value)**	***M'*-locus**
PGS3_12	H						PGS3_12	H					
PGS3_13							PGS3_13						
PGS3_15							PGS3_15						
PGS3_22	H	0	14	15	0,03 (0,85)	*M'm'*	PGS3_22	H	0	12	6	2,00 (0,16)	*M'm'*
PGS3_23							PGS3_23						
PGS3_28							PGS3_28						
PGS3_32							PGS3_32						
EPPCU7190	B						EPPCU7190	H					
K05-24	**Gen**	**A**	**H**	**B**	**χ^2 (*P*-value)**	***M'*-locus**	**K06-06**	**Gen**	**A**	**H**	**B**	**χ^2 (*P*-value)**	***M'*-locus**
PGS3_12	A						PGS3_12	A					
PGS3_13							PGS3_13						
PGS3_15	A	63	0	0	0,02 (0,90)	*M'm'*	PGS3_15	A	24	0	0	1,50 (0,22)	*M'm'*
PGS3_22							PGS3_22						
PGS3_23							PGS3_23						
PGS3_28	H	0	31	32			PGS3_28	H	0	9	15		
PGS3_32							PGS3_32						
EPPCU7190	H						EPPCU7190	H					
K06-17	**Gen**	**A**	**H**	**B**	**χ^2 (*P*-value)**	***M'*-locus**	**K06-21**	**Gen**	**A**	**H**	**B**	**χ^2 (*P*-value)**	***M'*-locus**
PGS3_12	H						PGS3_12	H					
PGS3_13	H	10	8	3	5,85 (0,05)	*m'm'*	PGS3_13	H	5	15	9	1,14 (0,57)	*m'm'*
PGS3_15							PGS3_15						
PGS3_22							PGS3_22						
PGS3_23	B	0	0	21			PGS3_23	B	0	0	29		
PGS3_28							PGS3_28						
PGS3_32							PGS3_32						
EPPCU7190	B						EPPCU7190	B					

[a] 'Gen' indicates the SSR genotype for each F_2 recombinant

[b] Chi-square χ^2 and *P* values for the expected segregation ratios 1:2:1 (*m'm'*) and 1:1 (*M'm'*) obtained from each independent F_3 population.

Macro- and microsynteny analysis of the *M'*-locus in apricot

Eleven out of the sixteen SSR markers contained in the 'Katy' LG3 map had been previously mapped in the 'Canino' LG3 (Zuriaga et al., 2012). As a whole, these markers were found to be collinear between both maps (8 out of 11) but some order changes regarding PGS3_33, AMPA119 and EPPCU0532 were observed at the distal chromosome end (data not shown). In turn, marker order in the 'Katy' LG3 map was completely collinear with the physical position of the markers in the peach genome (Table 1.4a and Figure 1.4). Unfortunately, most of the markers surrounding the *M*-locus in 'Canino' LG3 were found to be monomorphic in 'Katy' and therefore could not be mapped (Table S1.3). Genetic differences between 'Katy' and 'Canino' were detected across the whole genome, they share only 38,8 % of their SSR alleles and show a Nei´s genetic distance of 0,83 (Table S1.4). Indeed, only a few collinear markers, such as PGS3_12, PGS3_15 and EEPCU7190, were useful to define a syntenic region between both apricot maps containing the *M*- and *M'*-loci and corresponding to a physical interval between 17.38-19.78 Mb in the peach genome (Figure 1.4a). The PGS3_22/EEPCU7190 interval comprising the *M'*-locus in 'Katy' corresponds to ~1.29 Mb in the peach syntenic genomic region (between 18.490-19.780 Mb positions). Meanwhile, in 'Canino' the *M*-locus was predicted to be flanked by PGS3_71 and PGS3_96 markers within an interval of 1.8 cM corresponding to ~364 Kb in the peach genome (between 18.399-18.763 Mb positions) (Zuriaga et al., 2012). Therefore, there is an overlapping interval between these two regions spanning ~273 kb. To have a complementary view of the predicted positions for the *M*- and *M'*-loci, the relative frequency of individuals lacking SSR alleles in coupling phase with the PPM (expected to be zero in those markers absolutely linked) was represented graphically on the peach chr.3 (Figure 1.4b). To do this, only individuals carrying the 'Canino' *m* mutated allele from the 'G×C-01' population (Zuriaga et al., 2012) or the *m'* allele from 'K×K' and 'Katy' outcrossing populations were computed. This analysis showed frequency values of zero in shorter overlapping intervals: PGS3_23 (18.61 Mb) in 'K×K', PGS3_22/PGS3_28 (18.49-19.14 Mb, ~650 Kb) in 'Katy' outcrosses and PGS3_44/PGS3_62 (18.29-18.61 Mb, ~320 Kb) in 'G×C-01'.

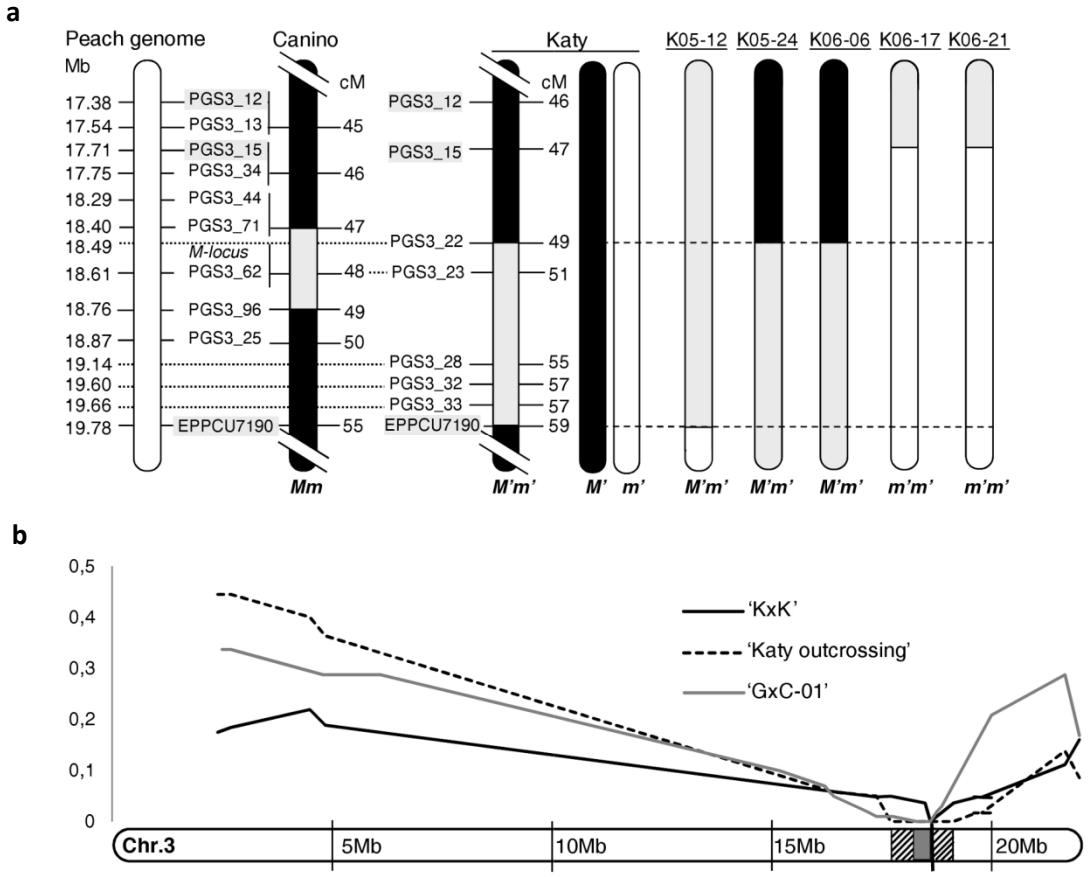


Figure 1.4. Mapping of the *M'*-locus and macro-synteny within *Prunus*. (a) Graphical LG3 maps of K×K-recombinant hybrids at the *M'*-locus. The corresponding map region between markers PGS3_12 and EPPCU7190 is shown for 'Katy' and 'Canino'. Distances in centimorgan (cM) are shown on the right of the apricot maps and their corresponding positions in megabases (Mb) on the peach genome are shown on the left. *Black vertical* bars represent self-incompatible (*M'M'*) chromosomal regions, while *grey* (*M'm'*) and *white* bars (*m'm'*) correspond to self-compatible chromosomal regions. Recombinant seedlings are numbered at the top. (b) Predicted positions for the *M*- and *M'*-loci on the peach chr.3 according to the relative frequency of individuals lacking SSR alleles in coupling phase with the PPM (*Y-axis*). The *black line* represents data corresponding to the 'K×K', the *dashed line* to 'Katy' outcrossing populations ('H×K', 'G×K' and 'C×K') and the *grey line* to the 'G×C-01' population.

The genomic landscape of the ~1.29 Mb peach region syntenic to the apricot *M'*-locus contains 223 predicted gene transcripts as annotated by IPGI. Forty-two of these transcripts (located in the overlapping interval) were shared in common with the 'Canino' *M*-locus. BLASTP analysis of the ORFs against The Arabidopsis Information Resource (TAIR) database, with an exp. value cut-off $<1e^{-6}$, was used by IPGI to predict gene functions based on homology to *Arabidopsis*. Table S1.5 includes the results of the BLASTP analysis for the ORFs comprised in the *M'*-locus region (IPGI) and indicates those *Prunus*/*Arabidopsis* gene pairs that are best-reciprocal BLASTP hits identifying putative orthologues. According to the large-scale gene expression analysis performed by Wang et al. (2008) in *Arabidopsis* mature pollen, hydrated pollen and pollen tubes

using Affymetrix ATH1 Genome Arrays, up to 53 of these *Arabidopsis* homologues were found to be pollen-expressed (Table S1.5).

Discussion

Loss of function of an *S*-locus external factor is responsible for SI breakdown in 'Katy'(S_1S_2)

In this work the North-American apricot cv. 'Katy', released by Zaiger´s Genetics (Modesto, CA, USA) in 1978 (Russell, 1998), was confirmed as self-fruitful and its *S*-genotype was determined as S_1S_2 following the nomenclature established by Burgos et al. (1998). However, previous reports assigned to 'Katy' the *S*-genotypes S_8S_C (Feng et al., 2006) and S_1S_8 (Wu et al., 2010). In addition, these two manuscripts referred 'Katy' as a spontaneous cultivar native to Europe and lately introduced to China. Therefore, both the *S*-genotype and the geographic origin proposed by these authors suggest that the cultivars they analyzed might be different from the cv. 'Katy' we describe here. Wu et al. (2010) also suggest that SC in 'Katy' is associated with PPMs that, according to the segregation of *S*-genotypes, seem to exert a polygenic control. Again, this is not the case in the Zaiger´s 'Katy' where SC is associated with a single PPM, however a sort of kinship between the two cultivars can not be discarded.

To investigate the genetics of SC, 'Katy' (S_1S_2) was self-pollinated and reciprocally crossed with the self-incompatible cv. 'Goldrich' (S_1S_2) (Egea & Burgos, 1996; Alburquerque et al., 2002). 'Katy' pollen tubes bearing either the S_1- or the S_2-haplotype were able to grow in 'Katy' and 'Goldrich' pistils and to complete fertilization, producing the three *S*-genotype classes expected for an F_2 population (S_1S_1:S_1S_2:S_2S_2). However, no progeny was obtained in the reciprocal cross using 'Katy' as female parent. These results would support a PPM unlinked to the *S*-locus as the cause for SC. Crosses performed with other cvs. such as 'Harcot' (S_1S_4) and 'Canino' (S_2S_C) reinforce this conclusion, since seedlings carrying the 'Katy' S_1- (when crossing with 'Harcot') and the S_2-haplotype (when crossing with 'Canino') were also obtained. Moreover, segregation ratios in all performed crosses fit with a model where 'Katy' is heterozygous for the PPM conferring SC (*M'm'*) (see Table 1.2).

Interestingly, in the 'K×K' and 'G×K' populations the number of seedlings homozygous for the S_1-haplotype (20) is significantly lower than that for the S_2-

haplotype (Tao et al., 1999) (see Table 1.1). Similar deviations were observed by Wünsch and Hormaza (2004) when the sweet cherry cv. 'Cristobalina' was self-pollinated. Following their reasoning, several causes might explain these deviations such as postzygotic selection against homozygous embryos, linkage in coupling between the mutated allele of the modifier factor (*m*) and the S_2-allele or differences in the pollen competitive capacity to grow through the style (depending on the *S*-haplotype). In this particular case, a hypothetical effect of postzygotic selection would explain the reduced number of S_1S_1 but not the high number of S_2S_2 genotypes. Regarding the second reason, neither the segregation ratios observed in different populations nor the SDL analysis support a linkage between the *M'*-and the *S*-locus. Therefore, a lower growth capacity for pollen gametes bearing the S_1-haplotype is regarded as the most acceptable hypothesis to explain this discrepancy.

SC caused by loss of pollen-*S* function has been usually found to be associated with mutations (mainly indels) of the *SFB* genes in different *Prunus* species such as sweet cherry (Ushijima et al., 2004; Sonnelveld et al., 2005; Marchese et al., 2007), apricot (Vilanova et al., 2006), Japanese apricot (Ushijima et al., 2004), peach (Tao et al., 2007) and sour cherry (Hauck et al., 2006). However, sequence analysis revealed no mutations or indels affecting any of the two 'Katy' *SFB* alleles discarding this as the cause of SI breakdown. In Solanaceae, self-compatible PPMs may arise from *S*-allele duplications located in a centric fragment, in a non-S chromosome or linked to the *S*-locus leading to the formation of *S*-heteroallelic pollen (Golz et al., 2001). According to the segregations obtained in the performed crosses, *S*-allele duplications did not seem probable in 'Katy' (all descendants should have had the S_1S_2 genotype), even so, we discarded that possibility showing that *SFB* gene dosage is equivalent between 'Katy' and the self-incompatible cv. 'Goldrich'. *S*-allele duplications may also result from polyploidy but 'Katy' was confirmed as diploid by flow cytometry analysis and by marker segregation and mapping in all crosses. These results rule out competitive interaction resulting from *S*-heteroallelic pollen as the cause of SC in 'Katy'.

Altogether, it can be hypothesized that the loss-of-function of a *S*-locus unlinked factor gametophytically expressed in pollen causes breakdown of SI in 'Katy'. Moreover, according to the relative abundance of SFB_1 and SFB_2 transcripts in 'Katy', when compared with the reference cv. 'Goldrich', the hypothetical defective factor in 'Katy' does not seem to affect their expression. These characteristics of the self-compatible mutant 'Katy' resemble those of other self-compatible pollen-part mutants

defective for non *S*-locus factors already found in *Prunus*. For instance, gene duplications and modified transcription levels of the *S*-locus genes were also discarded as the cause of SC in the *Prunus avium* cv. 'Cristobalina' (Wünsch & Hormaza, 2010) and the *Prunus armeniaca* cv. 'Canino' (Vilanova et al., 2006). According to the classification established by McClure et al. (2000) the modifier factor in 'Katy' would belong to the group of modifier genes required for pollen rejection but with no wider role in pollination. Although no direct evidence is available about its possible function, last findings in *Prunus* may provide some clue in this respect. For instance, the PavSSK1 and PavCul1 proteins recently identified by Matsumoto et al. (2012) in *Prunus avium* are proposed to form the SCF^{SFB} E3 ubiquitin ligase complex involved in *S*-RNases degradation. Therefore, the loss-of-function of any of them would predictably lead to SC. However, none of these two genes is located in LG3 where the *M*'-locus region is found and so they can be discarded as a possible cause of SC in 'Katy'. On the other hand, Tao & Iezzoni (2010) proposed an alternative model for the GSI in *Prunus* where a *S*-RNase inhibitor would be the target for the SCF^{SFB} ubiquitination complex instead of the *S*-RNases. If the modifier factor found in 'Katy' was this hypothetical inhibitor, its loss-of-function would lead to SI and not to SC what also rules out this possibility. Further research will therefore be necessary to reveal the SI related function affected by the PPM in 'Katy'.

PPMs conferring SC in 'Katy' and 'Canino' apricots are both located at the chr. 3 distal end

To facilitate future identification and cloning, the 'Katy' GSI mutated modifier gene locus (*M'*-locus) was mapped following a two-steps strategy. First, we hypothesized that those markers linked with the *M'*-locus should be highly distorted in the populations obtained from crosses where 'Katy' was the pollen parent, since only 'Katy' pollen tubes carrying the *m'*-allele would be able to grow. In other words, the *M'*-locus genomic region should correspond to a segregation distortion locus (SDL), a chromosomal region that causes distorted segregation ratios (Zhu & Zhang, 2007). To identify this kind of regions, '$K \times K_{05}$' and '$K \times K_{06}$' populations, which all trees carry the PPM, were tested for genome-wide distributed SSRs to detect SDL by examining changes in genotypic frequencies. Attending to segregation of pollen alleles, two SDL were found in LG3 and LG6 but a deeper analysis showed that LG6 markers were partially linked to the *S*-locus and only moderately distorted. Consequently, LG3 was

predicted as the most likely location for the *M'*-locus. Distortion in LG6 seems more plausibly related to the different capacity of S_1 and S_2-pollen gametes for growing through the style. Further analyses are in progress to confirm this point.

In a second step, to refine *M'*-locus mapping, chr.3 specific SSRs were analyzed to estimate their segregation distortion ratios in selfing (F_2) and outcrossing populations obtained by using 'Katy' as pollen parent. Additionally, indirect *M'*-locus genotyping was performed by analyzing linked SSRs in the F_3 offspring of six selected 'K×K' F_2 trees. Recombination breakpoints in five of these trees defined a 9.4 cM interval for the 'Katy' *M'*-locus that corresponds to ~1.29 Mb in the peach genome (18.49-19.78 Mb) and overlaps ~273 Kb with that established for the *M*-locus in 'Canino' (Zuriaga et al., 2012). A non *S*-locus PPM conferring SC to the *P. avium* cv. 'Cristobalina' was also mapped on the LG3 by Cachi and Wünsch (2010). However, it was tentatively predicted to be downstream the EMPaS02 marker (~20,0 Mb) and therefore, if confirmed, the position for this locus is not coincident with those for the *M*- and *M'*-loci in apricot. Different map locations for PPMs would support different defective genes as responsible for SC in sweet cherry and apricot, but this point still requires confirmation. Particularly in apricot, SSR markers showing the highest distortion values associated with the PPMs in 'Canino' (PGS3_62) and 'Katy' (PGS3_23) are located in very close positions (18.612 and 18.608 Mb, respectively). Thus, in the light of the similarities found between the apricot cvs. 'Katy' and 'Canino' (i.e. genetics of SC, *M*- and *M'*-locus mapping positions, etc.) it is tempting to speculate that both PPMs causing SC might be affecting the same gene, however no conclusive evidence is yet available on this point. Only 42 genes are shared in common between *M*- and *M'*-locus (Zuriaga et al., 2012) and, if this was the case, the availability of two different PPMs would be very helpful to identify the modifier gene. Interestingly, both cultivars have different geographic origins (i.e. 'Katy' is a North-American apricot selection (Russell, 1998) and 'Canino' is a local Spanish apricot (Vilanova et al., 2006) and, according to the analysis of gemome-wide distributed SSRs, they seem to be genetically unrelated. This prompts us to speculate that both PPMs (being or not the same) may have arisen independently.

According to the peach syntenic genome region annotated by IPGI, the apricot *M'*-locus is predicted to contain about 223 gene transcripts. Based on sequence similarity, putative *Arabidopsis* orthologues were suggested for many of these *Prunus* genes (Zheng et al., 2005) and, according to Movahedi et al. (2011), a consistent tissue-

specific expression might be expected for the reported gene pairs. Under this general rule, a high number of genes scattered throughout the *M'*-locus (up to 53) might be pollen-expressed fulfilling one of the main requirements for the SI 'Katy' modifier gene. Nevertheless, those genes whose orthologues are not pollen-expressed should not be discarded because inferred orthologues do not always have the same biological function (Movahedi et al., 2011). Gene function annotation might also be helpful to select candidate genes for the SI 'Katy' modifier gene. Unfortunately, the hypothetical roles suggested for this factor are still merely speculative hindering this approach. In view of the limitations for these strategies and considering the high number of ORFs comprised within the *M'*-locus, narrowing down the mapping region will be an essential step to identify the SI modifier gene in 'Katy'. In summary, 'Katy' does not only provide an additional *S*-locus unlinked source of SC, a desired trait for apricot breeding programs, but also becomes a very useful tool to dissect the molecular genetics behind pollen-pistil interactions in *Prunus*.

Materials and Methods

Plant material

Four apricot cvs. 'Goldrich', 'Canino', 'Harcot' and 'Katy', the progenies derived from the outcrosses 'Goldrich × Katy - 2005' ('G×K'), 'Canino × Katy - 2007' ('C×K'), 'Katy × Canino - 2007' ('K×C') and 'Harcot × Katy - 2005' ('H×K'), and the F_2 populations obtained by selfing 'Katy' in 2005 ('K×K_{05}') (*N*=16), 2006 ('K×K_{06}') (*N*=37) and 2010 ('K×K_{10}') (*N*=41) were used in this study (Table 1). 'K×K' population was formed by pooling all the individuals from these three latter F_2 populations. All these trees are maintained at the collection of the Instituto Valenciano de Investigaciones Agrarias (IVIA) in Valencia (Spain). Additionally, 12 independent F_3 seed populations (ranging from *N*=2 to *N*=77) were obtained after self-pollination of 'K×K_{05}' and 'K×K_{06}' trees. Selfing populations from 'Katy' (F_2 and F_3) were obtained by putting insect-proof bags over several branches (containing 200-250 flower buds) before anthesis to prevent cross-pollination. Outcrossing populations were obtained by pollinating balloon-stage flowers. Fruits were collected about three months later. F_3 seed-derived embryos were dissected from the rest of the seed tissue and stored at –20ºC.

Nucleic acids extraction

Two leaf discs of each selection were collected and stored at -80ºC before DNA isolation. Genomic DNA was extracted following the method of Doyle & Doyle (1987). DNA quantification was performed by NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE) and integrity was checked by comparison with lambda DNA (Promega, Madison, WI, USA). Embryo DNA was extracted by incubating for 10 min at 95ºC with 20 µl of TPS (100 mM Tris-HCl, pH 9.5; 1 M KCl; 10 mM EDTA) isolation buffer (Thomson and Henry, 1995). Total RNA was extracted from mature anthers (contaning mature pollen grains) of balloon-stage flowers using the UltraClean Plant RNA Isolation Kit (MoBio, Carlsbad, CA, USA).

PCR-amplification, cloning and sequencing of *S-RNase* gene fragments and the complete *S*-locus *F-box* alleles from 'Katy'

Fragments comprising the *S-RNase* first intron were PCR-amplified with primers SRc-F (Romero et al., 2004) and Pru-C2R (Tao et al., 1999) (Table S1.6) using 'Katy' genomic DNA as template. Cycling conditions were as follows: an initial denaturing step of 94ºC for 2 min; 30 cycles of 94ºC for 30 s, 55ºC for 60 s and 72ºC for 1 min 30 s; and a final extension of 72ºC for 10 min (GeneAmp®PCR System 9700, Perkin-Elmer, Fremont, CA). PCR products were electrophoresed in 1% (w/v) agarose gel, purified using the QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany) and cloned into the pGEM T-Easy vector (Promega, Madison, WI). DNA sequences from four independent clones were determined with an ABI3730 equipment using the Big Dye Terminator v.3.1. cycle sequencing kit (Applied Biosystems, Foster City, CA). Sequences were assembled and edited with the Staden package v1.4 (Bonfield, 2004) and homology searches were performed with BLASTX (Altschul et al., 1990). *S-RNase* fragments comprising the second intron were amplified with primers Pru-C2/Pru-C4R (Tao et al., 1999) (Table S1.6) using PCR-conditions described by Sonneveld et al. (2003). Genomic fragments containing the complete coding sequence of SFB_1 and SFB_2 (as well as their 3'/5' flanking regions) were PCR-amplified with the haploytpe-specific primer pairs FBf-Hap1/FBr-Hap1 (this work) and FBf-Hap2/FBr-Hap2 (Vilanova et al., 2006) respectively (Table S1.6), using 'Katy' genomic DNA as template. PCR conditions and methods for isolating, cloning, and sequencing these fragments were the same used for the *S-RNase* fragments.

Genomic PCRs for *S*-genotyping

S-genotyping of populations and cultivars was performed by PCR-amplification of the *S-RNase* first intron with the primer pair SRc-F/SRc-R (Romero et al., 2004) (Table S1.6) following the protocol described by Vilanova et al. (2006).

Ploidy level determination

Ploidy level was determined using the *Partec CyStain UV precise P* reagent kit (Partec PAS, Münster, Germany) for nuclei extraction and DNA staining of nuclear DNA from plant tissues. Approximately 0.5 cm^2 leaf tissue was chopped using a sharp razor blade in 400 µl extraction buffer and filtered through a *Partec 50 µm CellTrics* disposable filter. Samples were then incubated for 60 seconds in the staining solution and analyzed in the *Partec* flow cytometer Ploidy Analyzer PA (Partec, Münster, Germany) in the blue fluorescence channel.

Real time RT-PCR for SFB_1 and SFB_2

cDNA was obtained from total RNA isolated from mature anthers of the cvs. 'Goldrich' and 'Katy' using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen, Carlsbad, CA, USA). Genomic DNA traces were previously removed from RNA samples by treatment with DNAse I (Invitrogen, Carlsbad, CA, USA). *SFB* allele-specific PCR-primer pairs were designed in this work to amplify SFB_1 and SFB_2 (RT-SFB1-for/RT-SFB1-rev1 and RT-SFB2-for/RT-SFB2-rev2, respectively) (Table S1.6). Primer allele-specificity was tested by PCR-amplifying both alleles from genomic DNA and comparing fragment sizes with known *S*-genotypes in agarose gels after electrophoresis. The *actin* gene was used as endogenous control and the specific PCR primers Act3 and Act4 designed from the peach genome sequence (Gabino Ríos personal comm.) were used for amplification (Table S1.6). Specificity of *actin* PCR reaction was tested through size estimation of the amplified product by gel electrophoresis. Real-time PCR reactions were performed using an Applied Biosystems StepOnePlus Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) in a final volume of 20 µl, containing 10 µl of the SYBR Premix Ex Taq (Takara, Foster City, CA, USA), 0.4 µl of ROX reference dye, 0.375 µM of each primer and 2 µl of cDNA template diluted 1:15 from a total of 20 µl synthesized from 2 µg of total RNA. Cycling conditions were as follows: an initial denaturing step of 95ºC for 30 s; 40 cycles of 95ºC for 5 s, 60ºC for 30 s and 72ºC for 1 min. Relative expression of SFB_1

and SFB_2 from 'Katy' and 'Goldrich' RNA of mature anthers was measured by the standard curve method. Threshold cycle (C_T) values were automatically determined by StepOne v. 2.0 software (Applied Biosystems, Foster City, CA, USA). PCR reaction specificity was assessed after the amplification by confirming the presence of a single peak in the dissociation curve analysis. Results were the average of three independent biological replicates repeated three times.

Real-time PCR-based gene dosage assay for SFB_1 and SFB_2

SFB allele-specific PCR primers used to determine gene dosage of SFB_1 and SFB_2 from genomic DNA of cvs. 'Goldrich' and 'Katy' were also RT-SFB1-for/RT-SFB1-rev1 and RT-SFB2-for/RT-SFB2-rev2. *Actin* was used as endogenous control and the specific primers used to amplify this gene were Act3/Act4 (see previous sections). Real-time PCR reactions were performed using the same PCR mixtures (except for 2 µl of gDNA as a template), cycling conditions and thermocycler previously reported for real-time RT-PCR. Relative DNA quantity corresponding to SFB_1 and SFB_2 alleles from 'Katy' and 'Goldrich' was measured by the standard curve method. C_T values and PCR reaction specificity were also determined as for the real-time RT-PCR. Results were the average of two independent biological replicates repeated three times.

SSR marker analysis

A total of 118 SSR markers, spread over the 8 *Prunus* chromosomes, were tested to perform a genome-wide screen for the PPM (Table S1.7). Those SSRs amplifying in 'Katy', 'Goldrich' and 'Canino' (85) (Table S1.3) were used to estimate Nei´s genetic distance between the three cultivars (Nei, 1972) by means of GENETIX v.4.05 software (Belkhir et al., 2004). One hundred and two additional SSRs developed by Zuriaga et al. (2012) were tested to construct the 'Katy' LG3 map (Table S1.2). SSR amplifications were performed in a GeneAmp® PCR System 9700 thermal cycler (Perkin–Elmer, Freemont, CA, USA) in a final volume of 20 µl, containing 75 mM Tris–HCl, pH 8.8; 20 mM $(NH_4)_2SO_4$; 1.5 mM $MgCl_2$; 0.1 mM of each dNTP; 20 ng of genomic DNA and 1 U of Taq polymerase (Invitrogen, Carlsbad, CA). Each polymerase chain reaction was performed by the procedure of Schuelke (2000) using three primers: the specific forward primer of each microsatellite with M13(-21) tail at its 5' end at 0.4 µM, the sequence-specific reverse primer at 0.8 µM, and the universal fluorescent-labeled M13(-21) primer at 0.4 µM. The following temperature profile was used: 94°C for 2 min, then

35 cycles of 94°C for 45 s, 50–60°C for 1 min, and 72°C for 1 min and 15 s, finishing with 72°C for 5 min. Allele lengths were determined using an ABI Prism 3130 Genetic Analyzer with the aid of GeneMapper software, version 4.0 (Applied Biosystems).

M'-locus fine mapping

Segregation distortion locus (SDL) associated with the PPM was detected using JoinMap 3.0 software (Van Ooijen & Voorrips, 2001) by analyzing χ^2 values of selected SSRs spread over the *Prunus* genome in the 'K×K_{05}' and 'K×K_{06}' F_2 populations. Genetic maps for each linkage group were roughly estimated using these two populations. The logarithm of odds (LOD) grouping threshold was established at ≥ 3.0 for LG2, LG4, LG7 and LG8 but < 3.0 for the rest. Comparative mapping with other apricot cvs. was used to support grouping of markers in these latter cases.

Linkage maps of 'Katy' chr.3 were constructed using SSR markers segregating in 'K×K' and 'C×K' populations. Calculations were performed by JoinMap 3.0 software (Van Ooijen & Voorrips, 2001) using the Kosambi mapping function (Kosambi, 1944) to convert recombination units into genetic distances. In the 'C×K' population, LG3 was established following the "two-way pseudo test-cross" model of analysis Grattapaglia and Sederoff (1994) under a LOD grouping threshold of 5.0 and a recombination frequency parameter below 0.4. According to the single LG3 map obtained for 'Katy' from 'C×K', LOD score was relaxed to 2.0 for merging, two separated groups (at LOD > 5.0) in the 'K×K' population to construct LG3.

M'-locus genotyping of K×K-F_2 individuals was indirectly performed by analyzing segregation ratios of heterozygous SSR markers linked to the PPM (according to the SDL analysis) in the F_3 progenies. A χ^2 test was performed to check whether the observed ratios fit a 1:2:1 ratio, corresponding to the *m'm'* genotype, or a 1:1 ratio, corresponding to the *M'm'* genotype.

Acknowledgements

We thank Dr. Gabino Ríos for helpful suggestions on Real-Time RT-PCR and Real-Time PCR-based gene dosage assay and Pepe Juarez for his help in the ploidy level determination.

Chapter 2:
Pollen-part mutated *m*-haplotype is associated with self-compatibility and widely distributed in apricot germplasm

Abstract

Apricot (*Prunus armeniaca* L.) is basically considered as a self-incompatible species where numerous self-compatible exceptions occur, mainly linked to the mutated S_C-haplotype. However, more recently *S*-locus unlinked pollen-part mutations (PPMs) *m* and *m'* have also been reported to confer self-compatibility (SC) in apricot cultivars 'Canino' and 'Katy', respectively. This work was aimed to explore whether other additional mutations might explain SC in apricot as well. To do this, a set of 67 cultivars/accessions with different geographic origins and pedigrees were genetically analyzed by PCR-screening of *S*- and *M*-genotypes, contrasting results with the available phenotype data. As first finding, *m* and *m'*, initially described as independent PPMs, were found to be within the same haplotype. Results also indicate that this pollen-part mutated *m*-haplotype is tightly associated with SC in apricot germplasm. Its prevalence was higher than expected but lower than that for S_C, either in frequency or geographic distribution. In addition, two new putative mutations conferring SC were pointed out. Overall, results led to conclude that, despite a number of different mutations can be behind SC in apricot, the affected *loci* are restricted to two as occurring in other *Prunus* species. Reasons that could be underlying this behavior are discussed.

Introduction

Gametophytic self-incompatibility (GSI) is a widely distributed system in the plant kingdom (Igic and Khon 2001) that prevents self-fertilization favoring outcrossing (De Nettancourt 2001). GSI specific recognition is under the control of a multi-allelic locus, termed *S*-locus, containing at least two linked genes: a pistil expressed S-RNase (McClure et al., 1989) and the pollen expressed *S*-locus *F-box* (Lai et al., 2002; Sijacic et al., 2004; Ushijima et al., 2003). *S*-locus F-Box proteins are thought to be components of E3 ubiquitin ligase complexes that recognize non-self *S*-RNases promoting their ubiquitination and degradation by the 26S proteasome proteolytic pathway (Hua and Kao, 2006; Huang et al. 2006). Recently, the collaborative model proposed in Solanaceae (*Petunia*) suggests that several F-box proteins are necessary to recruit *S*-RNases for degradation (Kubo et al., 2010). This system seems to be extended to other plant families exhibiting GSI such as Rosaceae and particularly the Maloideae

subfamily (Kakui et al., 2011). However, interestingly, *Prunus* does not seem to follow this model since knock-out of the SFB leads to SC in contrast with the observations in Solanaceae. Reasons behind this singular behavior have been speculated for a long time but only recently evidences supporting a 'general inhibitor' distinct from SFB in *Prunus* have been provided (Matsumoto and Tao 2016).

In general, SC trait predominates in stone fruits (*Prunus* genus) in accordance with the high level of heterozygosity showed by these species. However, different 'degrees' of SC have been detected in this genus ranging from almost strict SI in sweet cherry, with a few exceptions, to complete SC in peach (Tao and Iezzoni 2010). SC sources are also variable but mostly related to mutations in the *S*-locus genes. In fact, mutations affecting both genes have been detected in many *Prunus* species (Tao and Iezzoni 2010; Hegedüs et al., 2012). Particularly, in apricot (*Prunus armeniaca* L.) the S_C-allele known to confer SC has been well characterized showing that a 358-bp insertion in the *SFB* gene leads to a putative truncated protein lacking the two essential 3´-hypervariable domains HVa and HVb (Vilanova et al., 2006). Furthermore, the origin and dissemination of S_C has also been reported, identifying the non-mutated ancestor S_8-allele and detecting its presence in cultivars from different geographic areas (Halazs et al., 2007). In fact, most *S*-genotyped self-compatible apricot cultivars have been shown to carry the S_C-allele (Vilanova et al., 2005; Halázs et al., 2007 and 2010; Kodad et al., 2013).

Along with the *S*-locus specific products, other *S*-locus unlinked factors are also necessary for the GSI system to work. These factors known as 'modifiers' were firstly identified in Solanaceae (McClure et al., 2000; Chen et al., 2010). Nevertheless, genetic evidence supporting modifiers have also been accumulated in other species including *Prunus spp*. (Wünsch and Hormaza 2004; Vilanova et al., 2006.). In apricot, pollen-part mutations (PPMs) conferring SC by putatively affecting modifiers have been identified in the Spanish local cultivar 'Canino' and the North-American one 'Katy'. Both PPMs were mapped at the distal end of chr.3 within the so-called *M*- and *M*'-loci, respectively (Zuriaga et al., 2012 and 2013). This led us to hypothesize that these two PPMs could have arisen independently (according not only to the origin but also to the genetic distance between both cultivars) affecting the same locus. In the same line, we thought that additional similar *S*-locus unlinked mutations might be present in apricot germplasm.

In this work, we have genotyped molecularly the *S*- and *M*-haplotypes using the 5'UTR *SFB* intron and the 1st/2nd *S*-RNase introns, on one side, and *M*-locus linked SSR markers, on the other. This approach has allowed us to dissect SC causes and distribution in apricot germplasm by using a wide set of cultivars from different geographic origins.

Results

Self-incompatibility vs. self-compatibility in apricot

In this study a set of 67 apricot cultivars was analyzed. They were selected trying to represent a wide range of geographic origins as well as different phenotypes regarding two main traits: self-(in)compatibility and blooming time (Table 2.1). Information about pedigree was only available for a few of them but it was quite useful to reinforce genotyping data. According to the blooming time phenotype reported in the literature (footnotes on Table 2.1) apricot cultivars were grossly classified into 5 classes: early, mid-early, mid, mid-late and late. Self-incompatible vs. self-compatible phenotypes were assigned to the different cultivars according to literature reports (Table 2.1) and our own data when adult trees were available (Table 2.2).

Table 2.1. Apricot cultivars analyzed in this study. Material sources, origin, pedigree as well as self-(in)compatibility and blooming time phenotypes are indicated.

	Cultivar	**Source**	**Geographic area**	**Country of origin**	**Pedigree****	**SI/SC****	**Blooming time****
1	Alba	CAPA	Western-Eur.	Italy	Unknown[1]	SC[11]	Mid[13]
2	ASP*	FM	Western Eur.	Spain (V)	Unknown (isolated tree)[2]	?***	Mid[13]
3	Aurora	CAPA	North-Am.	USA	RR17-62 × NJA-13[3]	SI[12]	Mid[15]
4	Bebecou	MAGRAMA	S-Eur/N-Afr	Greece	Unknown[4]	SC[12]	Mid-early[4]
5	Bergeron	MAGRAMA	Western Eur.	France	Unknown[4]	SC[12]	Mid-late[4,16]
6	Budapest	St. Istvan	Eastern-Eur.	Hungary	'Nancy' × ('Acme', 'Hungarian Best', 'Kései Rózsa')[5]	SC[5]	Mid-late[17]
7	Búlida	MAGRAMA	S-Eur/N-Afr	Spain (M)	Unknown[4]	SC[12]	Mid[4]
8	Canino	IVIA	Western-Eur.	Spain (V)	Unknown[4]	SC[12]	Mid[4]
9	Canino 9-7	IVIA	Western-Eur.	Spain (V)	Clonal selection from Canino[6]	?	Mid[6]
10	Canino 14-4	IVIA	Western-Eur.	Spain (V)	Clonal selection from Canino[6]	?	Mid[6]
11	Canino 14-6	IVIA	Western-Eur.	Spain (V)	Clonal selection from Canino[6]	?	Mid[6]
12	Castlebrite	CEBAS	North-Am.	USA	o.p. (Perfection × Castleton)[7]	SC[12]	Mid-late[13]
13	Castleton	CEBAS	North-Am.	USA	Perfection × Newcastle[7]	SC[7]	Mid[4]
14	Cegledi Orias	St. Istvan	Eastern-Eur.	Hungary	Unknown (local selection)[5]	SI[12]	Mid[17]
15	Colorao	CEBAS	S-Eur/N-Afr	Spain (M)	Unknown[4]	↗ sterile[12]	Mid[4]

16	Corbató	MAGRAMA	Western-Eur.	Spain (V)	Unknown[2]	?	Mid-late[15]
17	Cow-1	CAPA	Western-Eur.	France	INRA	?	Mid[8]
18	Cow-2	CAPA	Western-Eur.	France	INRA	?	?
19	Cristalí	FM	Western-Eur.	Spain (V)	Unknown[2]	?	Mid[15]
20	Currot	IVIA	Western-Eur.	Spain (V)	Unknown[2]	SC[12]	Early[15]
21	Dulcinea	CAPA	Western-Eur.	Italy	Unknown (Toscana variety)[1]	SC[12]	Mid[8]
22	Effect	St. Istvan	Eastern-Eur.	Ukraine	Krupnolodnyi o.p.[5]	SC[5]	Late[18]
23	Ezzine	CAPA	S-Eur/N-Afr	Tunisia	INRAT	?	Early[13]
24	Fergani	St. Istvan	Eastern-Eur.	Former USSR	Unknown	?	?
25	Galta Roja*	CAPA	Western-Eur.	Spain (V)	Unknown[2]	SC[12]	Mid-early[4]
26	GVV	FM	Western-Eur.	Spain (V)	Unknown (isolated tree)[2]	?	?
27	Gandía	FM	Western-Eur.	Spain (V)	Unknown[2]	?	Mid-early[15]
28	Gavatxet	FM	Western-Eur.	Spain (V)	Unknown[2]	?	Mid[15]
29	Ginesta	IVIA	Western-Eur.	Spain (V)	Unknown[2]	SC[13]	Mid-early[15]
30	Goldrich	IVIA	North-Am.	USA	Sunglo × Perfection[7]	SI[12]	Mid[4,16]
31	Gonci Magyar	St. Istvan	Eastern-Eur.	Hungary	Clone or hybrid of Hungarian Best[5]	SC[5]	Mid-late[17]
32	Harcot*	IVIA	North-Am.	Canada	[(Geneva × Naramata) × Morden 604] ×NJA1 (Phelps × Perfection)[7]	SI[12]	Mid[4]
33	Hargrand	St. Istvan	North-Am.	Canada	V51092 [(Reliable × o.p.) × o.p.] × NJA1 (Phelps × Perfection)[7]	SI[12]	Late[13]
34	Harlayne	IVIA	North-Am.	Canada	V51092 [(Reliable × o.p.) × o.p.] × Sunglo[7]	SC[12]	Mid-late[4,18]
35	Henderson	IVIA	North-Am.	USA	Unknown[7]	SC[14]	Mid-late[4,18]
36	Katy	IVIA	North-Am.	USA	Zaiger´s genetics (USA)[8]	SC[8]	Early[13]
37	Kech-pshar	St. Istvan	Eastern-Eur.	Uzbekistan	Unknown (local selection)[5]	?	?
38	Konservnyi Pozdnii	St. Istvan	Eastern-Eur.	Ukraine	Unknown (chance seedling)[5]	SC[5]	Mid-late[17]
39	Lambertin-1	CEBAS	North-Am.	USA	[Perfection × (Royal x Blush)] o.p. × (Perfection × Royal)[8]	SI[12]	Mid[8]
40	Lito	IVIA	S-Eur/N-Afr	Greece	SEO × Tyrinthos[9]	SC[9]	Mid-late[13]
41	Manrí	FM	Western-Eur.	Spain (V)	Clonal selection from Rojo de Carlet[2]	?	Mid-early[15]
42	Mari de Cenad	St. Istvan	Eastern-Eur.	Romania	Unknown[5]	SC[5]	?
43	Mariem	CAPA	S-Eur/N-Afr	Tunisia	2nd generation (Bergeron × Ouardi) × (Carraut × Crossa-Raynaud)[8]	?	Mid[8]
44	Martinet	FM	Western-Eur.	Spain (V)	Unknown[2]	?	Mid[15]
45	Mitger	IVIA	Western-Eur.	Spain (V)	Unknown[2]	SC[13]	Mid[8]
46	Moniquí	CEBAS	S-Eur/N-Afr	Spain (M)	Unknown[4]	SI[12]	Mid[4]
47	Ninfa	CAPA	Western-Eur.	Italy	Ouardi × Tyrinthos[8]	SC[12]	Mid-early[16]
48	Orange Red	IVIA	North-Am.	USA	Lasgerdi Mashhad × NJA2[8]	SI[12]	Mid[4,16]
49	Ouardi*	IVIA	S-Eur/N-Afr	Tunisia	Canino × Hamidi[4]	SI[12]	Mid-early[4]
50	Palabras	IVIA	Western-Eur.	Spain (V)	Unknown[2]	SC[13]	Mid-early[15]
51	Palau	IVIA	Western-Eur.	Spain (V)	Unknown[2]	SC[13]	Mid-early[15]
52	Patterson	CEBAS	North-Am	USA	F_2 seedling Perfection ×	SC[12]	Mid-late[7]

					unknown[7]		
53	Perla	CAPA	Western-Eur.	Italy	ANFIC (Italy)	?	Mid-early[8]
54	Portici*	FM	Western-Eur.	Italy	Unknown[4]	SC[12]	Mid-early[4]
55	Rojo de Carlet	IVIA	Western-Eur.	Spain (V)	Unknown[2]	SC[13]	Mid[15]
56	Rozsakajszi	St. Istvan	Eastern-Eur.	Hungary	Local selection Nagykórös[5]	SC[5]	Mid-late[17]
57	Sayeb*	CEBAS	S-Eur/N-Afr	Tunisia	Canino × Hamidi[4]	SC[12]	Mid-early[4]
58	SEO*	IVIA	North-Am.	USA	Unknown[7]	SI[12]	Mid-late[4]
59	Shalah*	St. Istvan	Eastern-Eur.	Armenia	Unknown	SC/SI?[12]	Late[18]
60	Stella	IVIA	North-Am.	USA	Unknown[4]	SI[12]	Late[4, 18]
61	Szegedi Mammut	St. Istvan	Eastern-Eur.	Hungary	Unknown (local selection)[10]	SI[12]	Mid[17]
62	Tadeo	IVIA	Western-Eur.	Spain (V)	Unknown[2]	SC[12]	Mid-late[15]
63	Tirynthos	IVIA	S-Eur/N-Afr	Greece	Unknown[4]	SC[12]	Early[4]
64	Trevatt	CEBAS	North-Am.	Australia	Unknown[8]	SC[12]	Mid-late[4]
65	Veecot	IVIA	North-Am.	Canada	o.p. from Reliable[7]	SI[12]	Mid[4]
66	Velázquez	MAGRAMA	S-Eur/N-Afr	Spain (M)	Unknown	SI[12]	Late[13]
67	Xirivello	CAPA	Western-Eur.	Spain (V)	Unknown[2]	?	Mid-late[15]

* Synonyms and acronyms: ASP '*Albaricoquero Sin Polen*'; Galta Roja 'Galta Roja de Mitger' or 'Palau'; GVV 'Galta Vermella Valenciana'; Ouardi 'Priana'; Portici 'Portici-6'; Sayeb 'Beliana'; SEO 'Stark Early Orange' and Shalah 'Erevani'.

** References: [1] CRA Consiglio per la Ricerca e la Sperimetazione in Agricoltura; [2] García et al. (1985) [3] Halász et al. (2005); [4] Della Strada et al., (1989); [5] Halász et al., (2007); [6] Badenes et al. (1993); [7] Brooks and Olmo (1997); [8] Russell (1998); [9] Syrgianidis and Mainou (1993); [10] Nyéki et al. (1999); [11] Egea and Ruiz (2014); [12] Burgos et al., (2004); [13] IVIA (see footnotes in Table 2.3); [14] www.gb-online.co.uk [15] Badenes et al. (1997); [16] Massai (2010); [17] Nyujtó et al. (1982); [18] Mehlenbacher et al. (1991).

*** No reports on the SI/SC or blooming phenotype were found for these cultivars (?).

Self-pollinations were used to determine (9) or to confirm (15) self-(in)compatibility phenotypes and to check progeny *S*-genotypes searching for *S*-locus unlinked mutations. A total of 19 cultivars/accessions out of the 21 included in Table 2.2 were self-pollinated in this work (2008). Data suggest that 6 of them are self-incompatible ('Aurora', 'Cow-2', 'Mariem', 'Perla', 'Veecot' and 'Velázquez') while the remaining 15 show variable fruit-setting ranging from 0.5% ('Búlida') to 55% ('Ninfa') being recorded as self-compatible.

Table 2.2. Self-pollination assays. Data about dates, bagged flowers, fruit-setting and inferred phenotype are included.

Cultivar	**Year**	**Flowers**	**Setting**	**%**	**Phenotype**	**Progeny**
Alba	2008	355	37	10.4	SC	1
Aurora	2008	350	0	0	SI	--
Bebecou	2009/2013	760	108	14.2	SC	96

Búlida	2009	200	1	0.5	SC?*	--
Canino	2005	412	99	24.0	SC	99
Castlebrite	2009	300	36	12.0	SC	2
Corbató	2009/2014	320	52	16.3	SC	44
Cow-1	2008	200	8	4.0	SC	7
Cow-2	2008	315	0	0	SI	--
Cristalí	2009	200	24	12.0	SC	13
Dulcinea	2008/2013	850	111	13.1	SC	104
Ezzine	2008	350	160	45.7	SC	21
Galta Roja	2008/2013	775	106	13.7	SC	106
Katy	2005	731	80	10.9	SC	80
Mariem	2013	450	0	0	SI	0
Ninfa	2008	400	221	55.3	SC	12
Perla	2008	370	0	0	SI	--
Portici-6	2008/2013	850	63	7.4	SC	59
Tadeo	2008	375	14	3.7	SC	5
Veecot	2013	450	0	0	SI	--
Velázquez	2009	270	0	0	SI	--

In general, fruit-setting percentages below 2-3% should not be undoubtedly associated to SC since some degree of pollen contamination can not be fully discarded.

S-genotyping: identification of new *S*-alleles and new *S*-genotypes

All the 67 cultivars/accessions were *S*-genotyped using primer pairs amplifying different fragments of the *S*-haplotype region: the first and second *S-RNase* introns as well as the 5'-UTR *S*-locus *Fbox* (*SFB*) intron (Table 2.3).

Table 2.3. *S*- and *M*-haplotypes of the apricot cultivars analyzed in this study

	Cultivar	SI/SC	*S*-genot[a]	*M*-genot[b]		Cultivar	SI/SC	*S*-genot	*M*-genot
1	Alba	SC^{od}	$S_1/\boldsymbol{S_C}$	$M_{1\text{-}1}/M_3$	35	Henderson	SC	S_{29}/S_{31}	$M_{2\text{-}0}/M_{16}$
2	ASP	↗sterile^{od}	$S_5/\boldsymbol{S_C}$	$M_{4\text{-}0}/M_{5\text{-}1}$	36	Katy*	SC/SC^{od}	S_1/S_2	$M_3/\boldsymbol{m_{0\text{-}0}}$
3	Aurora*	SI	S_1/S_{29}	M_3/M_{11}	37	Kech-pshar*	?	S_{15}/S_Z	$M_?/M_?$
4	Bebecou	SC	$S_6/\boldsymbol{S_C}$	$M_{4\text{-}0}/M_{4\text{-}2}$	38	Konservnyi P.*	SC	$S_2/\boldsymbol{S_C}$	$M_{4\text{-}1}/M_{8\text{-}1}$
5	Bergeron*	SC	$S_2/\boldsymbol{S_C}$	$M_{4\text{-}0}/M_{8\text{-}0}$	39	Lambertin-1*	SI	S_1/S_2	$M_{1\text{-}0}/M_{17}$
6	Budapest*	SC	$S_2/\boldsymbol{S_C}$	$M_{1\text{-}1}/M_{12}$	40	Lito	SC^{od}	$S_6/\boldsymbol{S_C}$	M_6/M_{10}
7	Búlida	SC	$S_5/\boldsymbol{S_C}$	$M_{4\text{-}1}/M_{5\text{-}1}$	41	Manrí	?	$\boldsymbol{S_C/S_C}$	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}0}}$
8	Canino*	SC/SC^{od}	$S_2/\boldsymbol{S_C}$	$M_{1\text{-}0}/\boldsymbol{m_{0\text{-}0}}$	42	Mari de Cenad*	SC	$\boldsymbol{S_C}/S_?^c$	$M_{8\text{-}0}/M_{18}$
9	Canino 9-7	?	$S_2/\boldsymbol{S_C}$	$M_{1\text{-}0}/\boldsymbol{m_{0\text{-}0}}$	43	Mariem	SC/SC^{od}	S_7/S_{20}	$M_{1\text{-}0}/M_{8\text{-}2}$
10	Canino 14-4	SC^{od*}	$S_2/\boldsymbol{S_C}$	$M_{1\text{-}0}/\boldsymbol{m_{0\text{-}0}}$	44	Martinet	?	S_2/S_2	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}1}}$
11	Canino 14-6	SC^{od*}	$S_2/\boldsymbol{S_C}$	$M_{1\text{-}0}/\boldsymbol{m_{0\text{-}0}}$	45	Mitger	SC^{od*}	$\boldsymbol{S_C/S_C}$	$M_{5\text{-}0}/M_{5\text{-}0}$
12	Castlebrite	SC	S_2/S_2	$M_3/\boldsymbol{m_{0\text{-}0}}$	46	Moniquí*	SI	S_2/S_6	$M_{4\text{-}1}/M_{14\text{-}0}$
13	Castleton	SC	S_1/S_2	$M_3/\boldsymbol{m_{0\text{-}0}}$	47	Ninfa	SC	$S_7/\boldsymbol{S_C}$	$M_{7\text{-}0}/M_{10}$

14	Cegledi Orias*	SI	S_8/S_9	M_{12}/M_{13}	48	Orange Red	SI	S_6/S_{29}	$M_{2\text{-}0}/M_{2\text{-}1}$
15	Colorao*	↗sterile	$S_5/\boldsymbol{S_C}$	$M_{4\text{-}0}/M_{4\text{-}0}$	49	Ouardi*	SI	S_2/S_7	$M_{1\text{-}0}/M_{7\text{-}0}$
16	Corbató	SC^{od}	S_2/S_5	$M_{5\text{-}0}/\boldsymbol{m_{0\text{-}0}}$	50	Palabras	SC/SC^{od*}	$\boldsymbol{S_C/S_C}$	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}0}}$
17	Cow-1	SC^{od}	S_1/S_{30}	$M_3/\boldsymbol{m_{0\text{-}0}}$	51	Palau*	SC/SC^{od*}	$\boldsymbol{S_C/S_C}$	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}0}}$
18	Cow-2	SI^{od}	S_{20}/S_{30}	$M_{1\text{-}0}/\boldsymbol{m_{0\text{-}0}}$	52	Patterson	SC	$S_1/\boldsymbol{S_C}$	$M_{1\text{-}3}/M_3$
19	Cristalí	SC^{od}	$S_{20}/\boldsymbol{S_C}$	$M_{5\text{-}0}/\boldsymbol{m_{0\text{-}0}}$	53	Perla	SI^{od}	S_2/S_{20}	$M_{1\text{-}2}/M_{15\text{-}1}$
20	Currot*	SC/SC^{od*}	$\boldsymbol{S_C/S_C}$	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}0}}$	54	Portici	SC/SC^{od}	S_2/S_{20}	$M_{1\text{-}4}/\boldsymbol{m_{0\text{-}0}}$
21	Dulcinea	SC	$S_2/\boldsymbol{S_C}$	$M_{7\text{-}3}/M_{14\text{-}1}$	55	Rojo Carlet	SC^{od*}	$\boldsymbol{S_C/S_C}$	$M_{4\text{-}0}/M_{5\text{-}0}$
22	Effect*	SC	$S_8/\boldsymbol{S_C}$	$M_{8\text{-}0}/M_{12}$	56	Rozsakajszi*	SC	$S_2/\boldsymbol{S_C}$	$M_{1\text{-}0}/M_{12}$
23	Ezzine	SC^{od}	$S_{24}/\boldsymbol{S_C}$	$M_{1\text{-}0}/M_{7\text{-}1}$	57	Sayeb*	SC	$S_7/\boldsymbol{S_C}$	$M_{1\text{-}0}/M_{7\text{-}2}$
24	Fergani	SC	S_V/S_X	$M_?/M_?$	58	Shalah	SC	S_5/S_{11}	$M_{8\text{-}1}/M_{19}$
25	GaltaRoja	SC/SC^{od}	$\boldsymbol{S_C/S_C}$	$M_{4\text{-}0}/M_{5\text{-}2}$	59	SEO	SI	S_6/S_{29}	$M_{2\text{-}0}/M_6$
26	GVV	?	$S_2/\boldsymbol{S_C}$	$M_{4\text{-}0}/M_{5\text{-}2}$	60	Stella	SI	S_6/S_{20}	M_9/M_9
27	Gandía	?	$\boldsymbol{S_C/S_C}$	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}0}}$	61	Szegedi M.	SI	S_8/S_9	M_{12}/M_{13}
28	Gavatxet	?	$S_{20}/\boldsymbol{S_C}$	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}0}}$	62	Tadeo	SC	$S_{20}/\boldsymbol{S_C}$	$M_{15\text{-}0}/\boldsymbol{m_{0\text{-}0}}$
29	Ginesta*	SC/SC^{od*}	$\boldsymbol{S_C/S_C}$	$\boldsymbol{m_{0\text{-}0}/m_{0\text{-}0}}$	63	Tirynthos	SC	$\boldsymbol{S_C/S_C}$	M_{10}/M_{10}
30	Goldrich*	SI/SI^{od}	S_1/S_2	$M_{1\text{-}0}/M_{2\text{-}0}$	64	Trevatt	SC	$S_2/\boldsymbol{S_C}$	$M_{1\text{-}0}/\boldsymbol{m_{0\text{-}0}}$
31	Gonci Magyar*	SC	$S_8/\boldsymbol{S_C}$	$M_{8\text{-}0}/M_{12}$	65	Veecot	SI/SI^{od}	S_2/S_{20}	$M_{2\text{-}0}/M_3$
32	Harcot*	SI/SI^{od}	S_1/S_4	$M_{1\text{-}0}/M_{2\text{-}2}$	66	Velázquez	SI	S_5/S_{20}	$M_{4\text{-}0}/M_{4\text{-}1}$
33	Hargrand*	SI	S_1/S_2	$M_{1\text{-}0}/M_{2\text{-}0}$	67	Xirivello	?	$\boldsymbol{S_C/S_C}$	$M_{4\text{-}0}/M_{4\text{-}1}$
34	Harlayne*	SC	S_{20}/S_{31}	$M_{2\text{-}0}/M_9$					

* Cultivars previously *S*-genotyped

[od] Own data on self-(in)compatibility phenotype (see Table 2.2)

[od*] Moderate fruit setting was also observed (though not quantified) across several years (since 2000) for a set of cultivars grown under insect-proof screen house at IVIA: 'Rojo de Carlet', 'Mitger', 'Palabras', 'Palau', 'Currot', 'Ginesta', 'Canino 14-4' and 'Canino 14-6'.

[a] *S*-allele nomenclature is proposed according to Vilanova et al. (2005), Halász et al. (2005 and 2010) and Wu et al. (2009). *S*-haplotype associated with self-compatibility (S_C) is written in bold. Except for S_4 (Vilanova et al. 2005) S_{11} and S_{15} (Halász et al. 2005 and 2010), those *S*-alleles detected only once were named with letters (S_V, S_X and S_Z). Allele fragment sizes corresponding to S_V and S_X could not be established since both were found in the same cultivar, while S_{15} and S_Z could only be clearly distinguished on the basis of the second intron size (see Table S2.2).

[b] *M*-haplotypes were named with two digits. The first one corresponds to the *M*-haplotype itself and the second to the variant type. *M*-haplotype variants associated with self-compatibility ($m_{0\text{-}0}$ and $m_{0\text{-}1}$) are written in bold.

[c] $S_?$ is S_{19} according to the nomenclature reported by Halász et al. (2010) and Kodad et al. (2013).

Fragment analyses of the PCR products obtained with 5 primer pair combinations allowed us to detect up to 19 different *S*-alleles. Fourteen out of them had already been identified in apricot (S_1, S_2, S_4, S_5, S_6, S_7, S_8, S_9, S_{11}, S_{15}, S_{20}, S_{24}, S_C and S_Z). Three more (present in at least two different cultivars) have been shown to be new and named as S_{29}, S_{30} and S_{31} according to the nomenclature established by Vilanova et al. (2005), Halász et al. (2005 &2010) and Wu et al. (2009). Intron sizes and sequence analysis (data not shown) strongly supports that none of these three corresponds to any previously described *S*-allele (Figure 2.1). Lastly, three additional *S*-alleles were detected only once and preliminary suggested as S_V, S_X and S_Z (this latter also already

identified) (see next section). Variability in size of the 5'-UTR *SFB* intron containing fragments ranged from 189 (S_{29}) to 210 bp (S_1) and facilitated the identification of 12 *S*-alleles (S_1, S_2-S_{11}, S_6, S_7, S_{15}, S_{29}, S_{30}-S_Z, S_{31} and S_C-S_8) but it was not useful to distinguish the remaining 7 (S_4, S_5, S_9, S_{20}, S_{24}, S_V and S_X). On his side, the amplification of first *S-RNase* intron with a single primer pair (SR1-F/SR1-R) distinguished up to 13 *S*-alleles ranging from 260 (S_4) to 427 bp (S_{29}) (Table S2.2. and Figure 2.1). Exceptions were S_1-S_7 and S_C-S_8 pairs having exactly the same fragment sizes and S_{20} and S_X that could not be PCR amplified with any of the four different primer combinations tested (data not shown). Moreover, an additional primer pair had to be used to amplify S_6 and S_{24}-alleles (Pru-T2/SR1-R). Sizes for the second *S-RNase* intron were approximately determined ranging from 300 (S_{24}) to 2800 bp (S_C) (by agarose gel electrophoresis) and allowed to define up to 15 *S*-alleles with a single primer pair (Pru-C2/Pru-C6R). Exceptions were S_5-S_6, S_8-S_C and S_{20}-S_V pairs sharing the same fragment size, and S_{31} that could not be amplified with any of the two primer combinations used. Altogether, combined data allowed us to identify unambiguously 17 *S*-alleles (Table S2.2). Lastly, according to the nomenclature reported by Halász et al. (2010) and Kodad et al. (2013) $S_?$ should be S_{19}, but under the experimental conditions used in this work it was not possible to detect other *S*-allele than S_C in 'Mari de Cenad'.

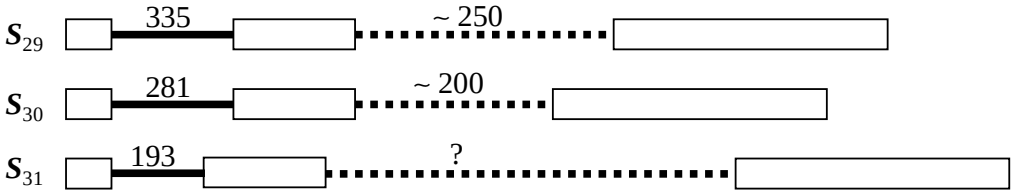


Figure 2.1. Identification of new apricot *S*-alleles. Genomic structure of the *Prunus armeniaca* S_{29}, S_{30} and S_{31}-*RNAse* alleles. *White boxes* and *lines* represent exons and introns, respectively (not to scale). *Dashed lines* indicate not sequenced introns and numbers estimated base pair sizes (? could not be determined).

M-locus genotyping

The apricot *M*-locus was previously identified in cultivars 'Canino' and 'Katy' as carrying an *S*-locus unlinked pollen-part mutation conferring SC (Zuriaga et al., 2012 and 2013). Up to 35 SSRs were found in the peach syntenic region of the *M*-locus (defined according to the 'Canino' genetic map) but just five of them could be finally mapped due to several reasons (i.e. multiloci patterns, monomorphism, non-amplification, etc.). Three were mapped using the 'G×Ca' population (PGS3_71, PGS3_62 and PGS3_96) and two with 'G×K' (PGS3_22 and PGS3_23). Linkage

phases determined by JoinMap 3.0 were used to identify alleles in coupling (haplotypes). Additionally, these 5 markers were PCR-amplified from *M*-locus homozygotes individuals derived from the self-pollination of 'Canino' [CC-77 (*MM*) and CC-67 (*mm*)] and 'Katy' [K06-17 (*mm*)] confirming linkage phases. This procedure also allowed us to incorporate two additional SSRs (AGS.20 and AGS.30) into the *M*-haplotypes. As a whole, these results lead to define four different *M*-haplotypes ($M_{1\text{-}0}$, $M_{2\text{-}0}$, M_3 and $m_{0\text{-}0}$ according to nomenclature as shown in Table 2.3) from 'Canino', Katy' and the reference cv. 'Goldrich' (Figure 2.2).

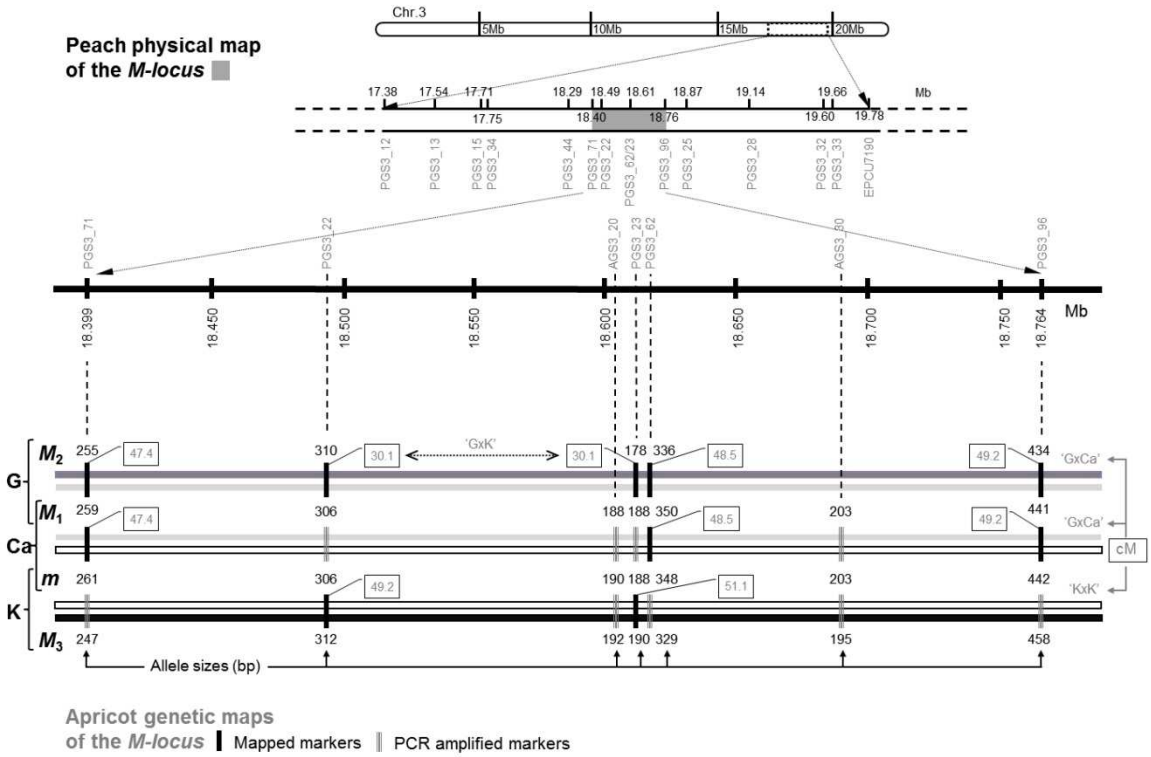


Figure 2.2. Apricot *M*-locus haplotypes structure. The peach syntenic region at the distal end of chr. 3 (*black box*) comprising the *M*-locus (*grey box*) is zoomed twice and shows SSRs (*dashed lines*) PCR-amplified in apricot cultivars 'Goldrich' (G), 'Canino' (Ca) and 'Katy' (K). SSR positions in peach genome (Kb) and allele sizes (bp) determined in apricot are indicated. *White*, *grey*, *dark grey* and *black thick line*s represent apricot $m_{0\text{-}0}$, $M_{1\text{-}0}$, $M_{2\text{-}0}$ and M_3 haplotypes, respectively. SSR anchoring positions are shown in centimorgans (cM) (*boxed numbers*) according to the available mapping populations ('GxCa', 'KxK' and 'GxK').

These seven SSR markers were subsequently PCR-genotyped in the rest of cultivars studied (Table S2.3). Structures of up to 34 additional *M*-haplotypes were statistically inferred (see Mats. and Methods for details) and those of the first four already identified fully confirmed (Table S2.4). To facilitate the graphical representation of their relative frequencies, haplotypes were grouped together into 'classes' when they differ in no more than three SSR alleles (resulting in a total of 20

M-haplotype 'classes': M_1 to M_{19} and *m*) (Figure 2.3). Pedigree, when known, was helpful in order to confirm assignments.

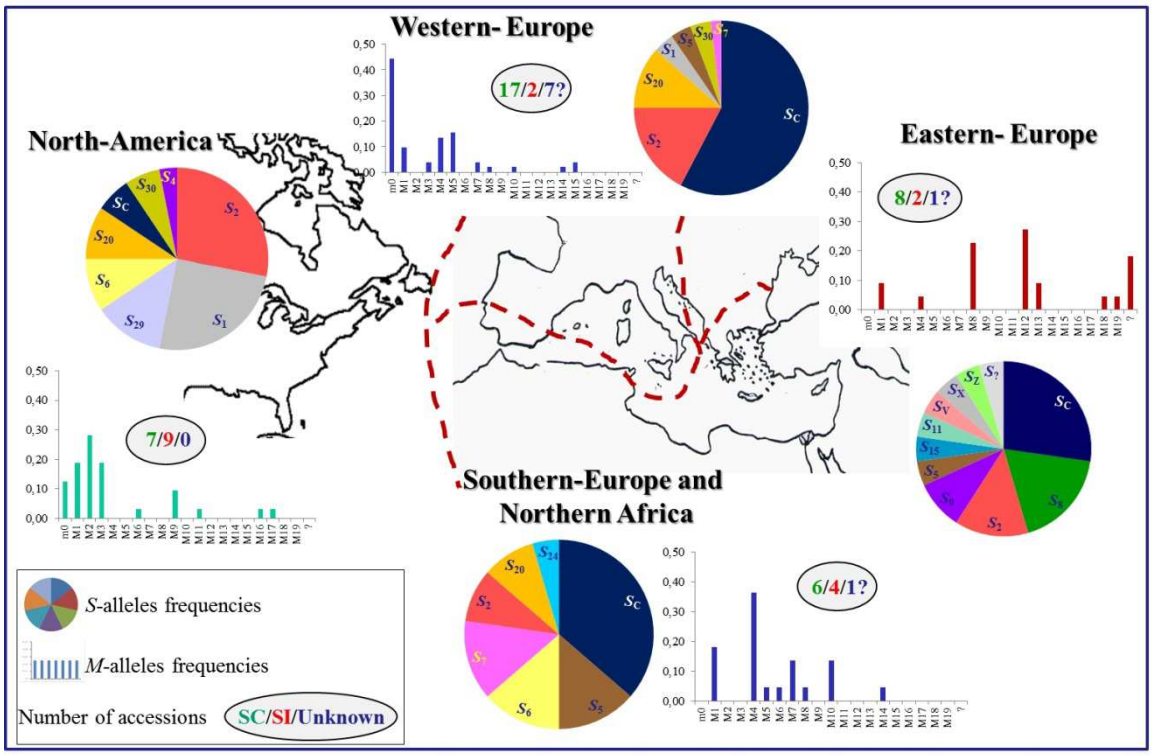


Figure 2.3. *S*- and *M*-locus haplotypes distribution according to geographic areas. Apricot cultivars analyzed in this study were grouped in four arbitrarily defined geographic areas represented by the bottom map: Western-Europe (WE), North-America (NA), Southern-Europe and North-Africa (SE/NAf) and Eastern-Europe (EE). Accordingly, relative frequencies for *S*- (pie charts) and *M*-haplotypes (histograms), respectively, are shown for each area (clonal sibs from 'Canino' were excluded from estimations). For the sake of simplicity *M*-locus haplotypes are represented in their 'main' classes. Number of accessions showing self-(in)compatible phenotype are encircled (*green* color means SC, *red* SI and *blue* undetermined phenotype).

Mutations conferring self-compatibility in apricot

Based on literature reports and our own results, 41 out of the 67 cultivars/accessions analyzed in this work were self-compatible, 16 self-incompatible, 2 male-sterile and the self-(in)compatibility phenotype could not be determined for 8 more (Table 2.3). The S_C-haplotype, characterized by Vilanova et al. (2006) as conferring SC, was found in 38 cvs. (being homozygous in 7). Thirty out of the 38 were self-compatible cultivars (including 'Búlida' that produced only one fruit under self-pollination test conditions, see Table 2.2), 6 have undetermined phenotype and two were male-sterile. On the other side, the *m*-haplotype, also associated with SC by Zuriaga et al. (2012), was detected in a total of 19 cvs. (being homozygous in 8): 8 were shown previously to be self-compatible, 5 were undetermined and one was classified as self-incompatible ('Cow-2') since it did not produce any fruit under self-pollination test

(Tables 2.2 and 2.3). As a whole, all self-compatible cultivars analyzed have at least one of these two haplotypes already known to confer SC except for 'Harlayne' and 'Henderson'. The number of self-compatible haplotypes (S_C and m_0) carried by each cultivar ranged from 0 to 1, 2, 3 and 4 (with 21, 27, 12, 3 and 4 cultivars, respectively). *S*-genotypes segregation observed in the progeny obtained from self-pollination of two self-compatible cultivars carrying a single copy of the *m*-haplotype ('Portici' and 'Corbató') suggested a mutation outside the *S*-locus as the cause for the phenotype in both cases. Progenies from self-compatible cultivars ('Dulcinea' and 'Bebecou') not carrying the *m*-haplotype were used as controls (Table 2.4). Similar results were observed in other cultivars (with a modest progeny) carrying ('Cow-1' and 'Cristalí') and not carrying ('Ezzine' and 'Ninfa') the *m*-haplotype (data not shown). Moreover, distortion ratios detected in the segregation of SSR markers known to be tightly linked to the *M*-locus (PGS3_62 and PGS3_23) point out that the mutation is located at the *M*-locus in accordance with the haplotypes previously assigned by genotyping (Table 2.4). Interestingly, *S*-genotypes segregation found in the 'Portici' progeny might also indicate the presence of another mutation affecting the S_2-haplotype and conferring SC. Analysis of genomic DNA fragments containing the complete sequence of S_2-*RNase* and SFB_2 alleles from 'Portici' revealed only one mismatch (A/G) with the functional SFB_2 allele located at position 1.296. This change leads to a non-synonymous substitution (lysine by arginine) in the hypervariable region HVb (data not shown).

Table 2.4. Segregation of *S-RNase* alleles in controlled self-pollinations

Cultivar	**Progeny *S*-genotypes**			**Total**	**Ratio 1:1**[a]	**Ratio 1:2:1**[a]	**Ratio 1:1**[b]
***S*-genotype**	**A**	**H**	**B**				
Bebecou (S_6/S_C)	0 (S_6/S_6)	33 (S_6/S_C)	45 (S_C/S_C)	96 (78)	**0.93 (0.34)**	30.8 (2.9e-7)	n.d.[c]
Dulcinea (S_C/S_2)	36 (S_C/S_C)	37 (S_C/S_2)	0 (S_2/S_2)	104 (73)	**0.01 (0.93)**	24.1 (6.0e-10)	n.d.
Corbató (S_5/S_2)	10 (S_2/S_2)	17 (S_5/S_2)	9 (S_5/S_5)	44 (36)	---	**0.08 (0.96)**	**0.08 (0.77) [24]**
Portici (S_2/S_{20})	16 (S_2/S_2)	24 (S_2/S_{20})	4 (S_{20}/S_{20})	59 (44)	---	**4.28 (0.12)**	**0.45 (0.50) [40]**

[a] χ^2 and (*P*) values for *S*-genotypes expected ratios considering a single mutation unlinked (1:2:1) or linked to the *S*-locus (1:1). Significant *P*-values are written in bold.

[b] χ^2 and (*P*) values for *PGS3_23* and *AGS.20* genotypes [*n*] expected ratios considering a single mutation located at the *M*-locus (1:1). Significant *P*-values are written in bold.

[c] n.d. Not determined.

Regarding 'Harlayne' and 'Henderson' both are the only two self-compatible cultivars sharing the S_{31}-allele what suggest this *S*-haplotype could be non-functional.

Mapping of the 'Harlayne' genomic sequence against peach genome allowed us to decipher the sequence of the S_{31}-*RNase* and SFB_{31} alleles detecting a putative deletion at the 3´end of this latter that might be the cause of SC (Figure 2.4). In total, 4 different mutations conferring SC have been recorded (two of them just putatively): three affecting the *S*-haplotype (and more concretely *SFB*) and one affecting the *M*-haplotype. All self-compatible cultivars can be assigned to any of these mutated haplotypes. Lastly, characterization of clonal sibs from 'Canino' (Canino 14-4; 14-6 and 9-7) did not reveal any difference in genotype with the original cultivar.

Figure 2.4. Identification of new apricot *S*-alleles. Alignment of 'Harlayne' S_{20}/S_{31}-RNase (upper) and *SFB* (below) genomic Illumina sequences against peach genome v1.0 reference sequence as reported by CLC Genomics Workbench 8.0.1 software. Haplotypes were identified by CLCbio tools using represented SNPs. ORFs, coverage and conservation are also indicated according to peach genome v1.0. *Blue* and *yellow* arrows define gene and transcript annotation. *Blue* and *red* boxes zooms in adjacent regions at the *SFB* terminal end containing both S_{20}/S_{31} and only S_{31} haplotype reads, respectively.

Geographical distribution of self-(in)compatibility and blooming time traits

Cultivars/accessions studied in this work can be grouped as belonging to four big geographic areas according to the country of origin, pedigree (see Table 2.1) and data about dissemination (Mehlenbacher et al., 1991; Faust et al., 1998; Bourguiba et al., 2012): North-America (NA), West Europe (WE), East Europe (EE) and South

Europe/North Africa (SE/NAf). Self-(in)compatibility phenotype is distributed across all four groups but frequencies varied from 2/19 in WE to 9/16 in NA. Similarly, blooming time types from early/mid-early to mid-late/late are present in all four groups, but frequency ranged from 9/24 in WE to 1/16 in NA (Table 2.1). In general it could be said that most early/mid-early classified cultivars are self-compatible (13/15) but this phenotype is also generally found in many mid-late/late classified cultivars (12/17). Therefore, assuming limitations in sampling and about phenotypic characterization, it seems that there is no correlation between SC and early blooming.

Frequencies of the *S*-alleles and *M*-haplotypes also varied between the different regions analyzed. For instance, S_2 and S_C are present in all four groups but, on the contrary, S_8 and S_9 (as well as S_{15} and S_Z) are only present in EE cultivars, S_{31} in NA and S_7 in SE/NAf (Figure 2.3). It should be highlighted that meanwhile the S_C-allele (conferring SC) is widely distributed, its ancestor S_8 seems to be restricted to EE cultivars.

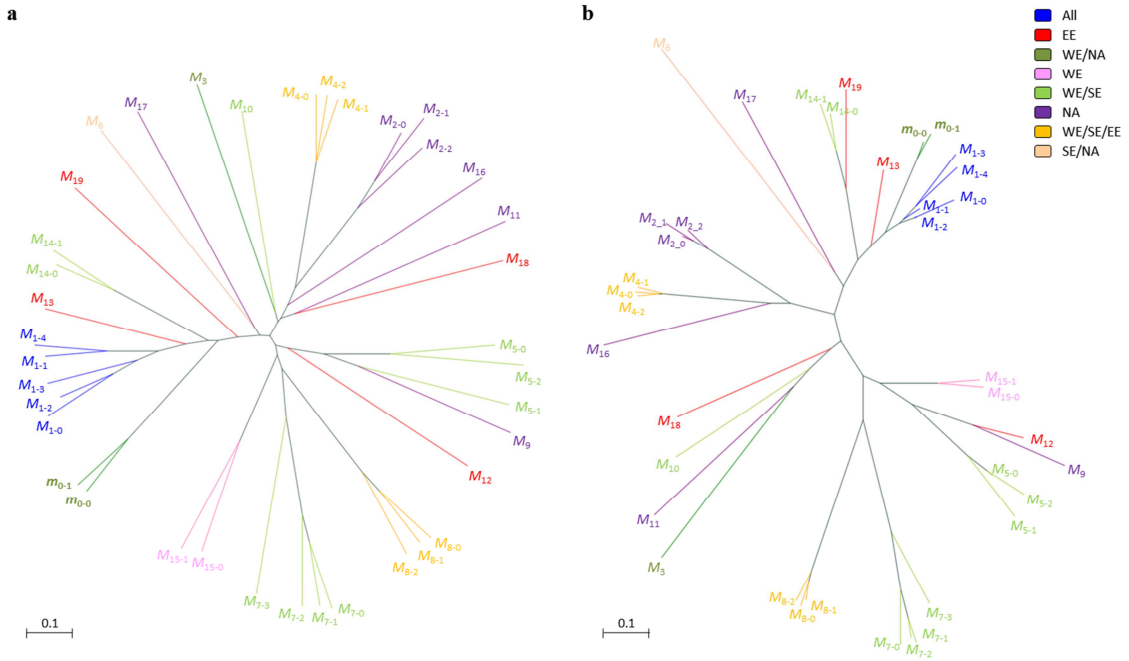


Figure 2.5. Clustering analysis of apricot *M*-locus haplotypes based on genetic distances. a) Clustering obtained by Neighbor-Joining algorithm using Jaccard´s distance. b) Clustering obtained by Neighbor-Joining algorithm using Bruvo´s distance. Colors represent geographic areas where the distinct *M*-locus haplotypes were detected (see legend).

Similarly, distribution of the 38 *M*-haplotypes that could be finally inferred is assymetrical, with M_1 present in all four groups while others such as M_{12} and M_2 were exclusive for EE and NA cultivars, respectively (Figure 2.3). Particularly, at odds with

the S_C allele, the *m*-haplotype associated with SC is only present in the WE and NA groups, while its putative ancestor M_1 is widely distributed (Figure 2.5).

Discussion

Self-compatibility and genotyping: uncovering new elements behind the phenotype

Part of the materials analyzed in this work could not be self-pollinated because adult trees were not available. However, though phenotype could not be directly assessed in these cases, according to the almost perfect association between SC and S_C/*m*-alleles it can be inferred that all cultivars carrying whatever of these two alleles will also be self-compatible. However, some incongruities between phenotype and genotype should be highlighted. For instance, self-pollination of cv. 'Cow-2' (S_{20}/S_{30}-$M_{1\text{-}0}/m_{0\text{-}0}$) did not produce any fruit as should be expected. This behavior could be related to fruit-set problems, as exemplified by the self-compatible cv. 'Búlida' (Burgos et al., 2004) which setting was also nearly null. Different were the cases for the self-compatible cvs. 'Mariem' (S_7/S_{20}-$M_{1\text{-}0}/M_{8\text{-}2}$), 'Shalah' ($S_5/S_{11}$-$M_{8\text{-}1}/M_{19}$), 'Harlayne' ($S_{20}/S_{31}$-$M_{2\text{-}0}/M_9$) and 'Henderson' ($S_{29}/S_{31}$-$M_{2\text{-}0}/M_{16}$). None of these *S*- and *M*-haplotypes has been previously associated with SC. In addition, most of them are also present in self-incompatible cultivars and therefore can be considered 'functional' (such as S_7, S_{20}, S_{11}, S_5, S_{29}, on one side, and $M_{1\text{-}0}$ and $M_{2\text{-}0}$, on the other). This reason (albeit with reservations) might be used to discard them as a putative source for SC. Regarding the rest, $M_{8\text{-}1}$ and $M_{8\text{-}2}$ belong to the M_8-haplotype subclass always detected in self-compatible cultivars (a total of 7 in this work and mostly East-European), but no progenies are available and therefore it could not be tested if M_8 is mutated (non-functional). Something similar can be said for M_{16} and M_{19} but in this case they were only detected once in the set of cultivars analyzed. Overall, other *S*/*M*-loci unlinked mutations can not be discarded for 'Mariem' and 'Shalah'. Evidences are different for S_{31}, since this rare allele is shared by two of the few North-American self-compatible cultivars. In fact, sequence analysis point out a putative indel within the SFB_{31} 3´-end as a plausible cause for SC, similarly to many other cases reported in *Prunus* (Tao and Iezzoni 2010). Lastly, genetic analysis suggests the presence of a SNP mutation within SFB_2 HVb region in 'Portici' that could also be associated with SC. It could be speculated that a single non-synonymous change within a SFB hypervariable region might alter its specificity, since these domains (strongly hydrophobic and under positive selection) were already suggested to have a role in the specific recognition of S-RNases

(Ikeda et al., 2004). As a whole and in the light of these results, it is conceivable that additional mutations conferring SC may remain 'hidden' in the bunch of unanalyzed plant material.

New *S*-alleles (S_{29}, S_{30} and S_{31}) were identified and named basically according to the nomenclature previously adopted by Vilanova et al. (2005) [S_1-S_7 and S_C], Halázs et al. (2005) [S_8-S_{16}] and Wu et al. (2009) [S_{17}-S_{28}]. However, some exceptions to this nomenclature can be appreciated. For instance, the S_{20}-allele was found to be highly similar (99%) to the sequence reported by Zhang et al. (2008). To our knowledge S_{20}- and S_{24}-alleles are two of the few *S*-alleles found in Chinese cultivars also present in Mediterranean germplasm (S_{24} only in 'Ezzine'). Particularly, S_{20}-allele is surprisingly widely distributed across all the geographic areas studied except for Eastern-Europe. Lastly, according to the *S*-genotyping results S_5 reported in this work is proposed to be the same that S_{13} reported by Halász et al. (2010) in the Armenian cv. 'Shalah'. This finding is relevant since connects this low frequent allele, mainly found in Armenian, Eastern-Turkish and Moroccan cultivars (Halász et al., 2010; Kodad et al., 2013) with Southern-Spanish cultivars (Burgos et al., 1998; Vilanova et al., 2005; this work) supporting the Southwest-Mediterranean diffusion route for apricot, from the Irano-Caucasian gene pool, proposed by Bourguiba et al. (2013).

Pollen-part mutated *m*-haplotype origin and dissemination

Pollen-part mutated *m*-haplotype had been previously associated with SC in 'Canino' and 'Katy' cultivars (Zuriaga et al., 2012 and 2013). In this work the *m*-haplotype has been detected in 17 additional cultivars (excluding 'Canino' clonal sibs) mainly Spanish (12 in total) but also from USA, Australia, France and Italy. Fifteen of them were confirmed as self-compatible (exceptions were 'Cow-2' described above as well as 'Gandía', 'Gavatxet', 'Manrí' and 'Martinet' with undetermined phenotype). The *m*-haplotype was frequently accompanied by the S_C-allele (9 cases), suggesting that mutations conferring SC might tend to accumulate once the system is broken. However, it was also found alone in 6 self-compatible cultivars. The analysis of progenies from two of them ('Portici' and 'Corbató') fully confirmed the association with SC in apricot germplasm.

Beside the *m*-haplotype, 37 additional *M*-haplotypes were identified by SSR analysis being grouped in 19 'main' classes. According to these results, it was not so surprising to find that heterozygosity was higher for the *M*-locus (0.82) than for the *S*-

locus (0.77). High variability at the *S*-locus is thought to be a requirement for the control of GSI specific recognition (Ushijima et al., 2003). Thus, lower variability should be expected for the *M*-locus if, as expected, it is not a specificity factor. This discrepancy may be explained by the bigger size of the defined *M*-locus and the high mutation rate associated with SSR markers.

To reinforce results on microsatellite haplotype distances, two alternative methods were used for clustering analysis. The first one relies on the proportion of shared alleles assuming independence and ignoring mutational processes which can bias distances, particularly when loci are highly polymorphic. This method was based on the similarity coefficient for binary data developed by Jaccard (1901). The second takes into account the stepwise mutation model considering higher likelihood for small than for large changes in microsatellite repeat number and it is based on the Bruvo´s distance (Bruvo et al., 2004). Results obtained were equivalent with both methods. Regarding the distribution of the *m*-haplotype it seems to be restricted to North-American and Western-European cultivars. However, according to the clustering analyses the closest *M*-haplotype (putative founder) is $M_{1\text{-}0}$ which is widely distributed in all geographic areas studied (the second one was M_{13} only detected in Eastern-European cultivars).

Meanwhile, the mutated S_C-allele is widely distributed in all geographic areas (Vilanova et al., 2005; Halász et al., 2007 and 2010, Kodad et al., 2013) but the ancestor S_8-allele was only detected in Hungarian cultivars. Altogether, these results suggest that the mutated *m*-haplotype arose much later in time, after apricot was established as a regular crop in Europe.

Forces selecting for self-compatibility in apricot

It is possible to conclude that SC is quite common in apricot. However, distribution of SI vs. SC is not uniform across geographic areas. SI is the prevalent phenotype in three out of the four major eco-geographical groups for apricot (centers of origin): Central Asian, Irano-Caucasian and Dzhungar-Zailij and also in the later proposed Chinese groups (Mehlenbacher et al., 1991). In fact, most recent works with Chinese material do not report self-compatible cultivars (Jie et al., 2005; Zhang et al., 2008; Wu et al., 2009). On the contrary, SC predominates in the European group but some disequilibrium can also be observed. Among the materials analyzed in this work, SI is frequent in commercial North-American cultivars (two thirds) but unusual in West and East-European countries (one fifth). This might point out a non-European SI

ancestral donor for North-American cultivars apricots as previously suggested for PPV resistance trait (Zhebentyayeva et al., 2008). Nonetheless, it is generally accepted that apricot genetic diversity decreases from east to south-west (Mehlenbacher et al., 1991). Supporting this hypothesis, a marked domestication bottleneck from the Irano-Caucasian gene pool was detected using SSRs by Bourguiba et al. (2013). In this context, a question arise regarding self-(in)compatibility: is it one of the causes of this bottleneck or is it just a result of selection? The reduction in apricot genetic diversity has been pointed out as a consequence of selection for high and reliable yield (Halász et al., 2007). Interestingly, the relationship between SC and yield has already been proved in stone fruits (Goldway et al., 2007). Moreover, SC by itself does not seem to be an essential requirement for growing apricots (and in general *Prunus spp.*) though it favors the removing of interspersed pollinators. In fact, when most of the 'European traditional' cultivars were selected (mid of the past century) apricots were commonly grown in mixed orchards and therefore this should not be a real problem. Regarding blooming time, results of this work do not even support a direct association between SC and earliness or late blooming. However, SC presence is ubiquitous in apricot cultivars despite their geographic origin. Thus, SC in apricot seems more to be the indirect result of selection linked to other main traits for breeding such as yield or earliness.

In this work we have confirmed the presence of at least three different PPMs conferring self-compatibility that affect two loci (S_C/S_{31} and *m*). These PPMs seem to be originated independently in different geographical contexts and the (putative) reasons why these mutations were selected have been depicted above. According to Hegedüs et al. (2012) a total of 27 non-functional *S*-haplotypes (including natural and induced mutations) and two mutated modifiers (Wünsch and Hormaza 2004; Vilanova et al., 2006) have been identified in *Prunus*. Their frequencies seem to depend largely on the clonal propagation process of stone fruit cultivars, where dissemination of self-compatible mutations is far from that expected for a panmictic population. On the other hand, it is noticeable that all reported *Prunus* self-compatible mutations to date affect the same two loci (*S* and *M*) (Tao and Iezzoni 2010; Hegedüs et al., 2012). Does it have any biological meaning? This is currently a matter of speculation but some hypothesis could be suggested. It can be argued that a more exhaustive screening is necessary to discard the presence of additional (and novel) mutations but the number of accessions evaluated in all *Prunus* species is already large enough. Assuming this point, if other modifiers are participating in the control of the GSI system in *Prunus* they should be

redundant in the genome, since no new *S*-locus unlinked mutation has been recovered. On the other hand, loss of function of some factors might lead to SI hindering their detection such as that recently predicted for SLFL2 (Matsumoto and Tao 2016). Lastly, proteins encoded by *S* and *M* loci might be operating in the same pathway. In other words, the *M*-modifier could be somehow modulating the *S*-determinants. Further research is necessary to clarify these questions in order to dissect the molecular mechanism underlying GSI in *Prunus* and to better understand GSI implications from an evolutionary point of view.

Material and methods

Plant Material and self-pollination test

Sixty seven apricot cultivars/accessions from diverse geographic origins were used in this study (Table 2.1). Most are currently kept at the collections of the *Instituto Valenciano de Investigaciones Agrarias* (IVIA) in Valencia (Spain). Part of these collections was kindly provided by *Frutales Mediterráneo S.A.* (FM) company, by the *Consellería de Agricultura, Pesca y Alimentación* (CAPA) and by the *Ministerio de Agricultura, Alimentación y Medio Ambiente* of Spain (MAGRAMA). Other materials were provided by the *Departamento de Mejora y Patología Vegetal del* CEBAS-CSIC in Murcia (Spain) and by the University of St. Istvan (Budapest, Hungary).

Trees from different cultivars/accessions (Table 2.2) were tested for self-compatibility by self-pollination in the field. Before anthesis, insect-proof bags were put over several branches, containing approximately 200-250 flower buds in total per cultivar *ad minimum*, to prevent cross pollination. Subsequent fruit set was recorded and fruits collected about 3 months later.

DNA extraction

Two leaf discs were collected from each accession, frozen in liquid N_2 and stored at -80ºC before DNA isolation. Genomic DNA was extracted following the method of Doyle and Doyle (1987) with some modifications. DNA quantification was performed by NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE) and integrity was checked by comparison with lambda DNA (Promega, Madison, WI, USA). Embryo DNA was extracted by incubating for 10 min

at 95ºC with 20 ml of TPS (100 mM Tris-HCl, pH 9.5; 1 M KCl; 10 mM EDTA) isolation buffer (Thomson and Henry 1995).

_S_-locus genotyping

Apricot *S*-alleles were identified by PCR-amplifying fragments comprising first and second introns of the *S-RNase* as well as the 5´UTR F-box intron, respectively. Genomic DNA isolated from cultivars listed in Table 2.1 was used as PCR template. PCRs were performed in a final volume of 20 μL containing 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.25 μM of each primer, 20 ng of genomic DNA and 1 U of Taq polymerase (Thermo Fisher Scientific, Waltham, MA, USA). Previously developed primers designed from conserved regions of *Prunus armeniaca S-RNase* genomic sequences, SRc-F and SRc-R (Romero et al. 2004; Vilanova et al., 2005) and from *Prunus avium S-RNase*-cDNA sequences, Pru-T2 and Pru-C2R (Tao et al., 1999), were used to amplify the first intron (Table S2.1) in all four possible combinations. The amplification was carried out using a temperature profile with an initial denaturing of 95 ºC for 3 min; 35 cycles of 95 ºC for 30 s, 54 ºC for 45 s and 72 ºC for 1 min 15 s; and a final extension of 72 ºC for 10 min (UNO96, VWR, Radnor, PA, USA). Each polymerase chain reaction was performed by the procedure of Schuelke (2000) using three primers: the specific forward primer with M13(-21) tail at its 5´ end at 0.4 mM, the reverse primer at 0.8 mM, and the universal fluorescent-labeled M13(-21) primer at 0.4 mM. Allele lengths were determined using an ABI Prism 3130 Genetic Analyzer with the aid of GeneMapper software, version 4.0 (Applied Biosystems, Foster City, CA, USA).

The second intron was amplified using two sets of primers designed from *Prunus avium S-RNase*-cDNA sequences (Tao et al., 1999; Vilanova et al., 2003) Pru-C2/Pru-C4R and Pru-C2/Pru-C6R (Table S2.1). PCRs were performed using the program previously described by Sonneveld et al. (2003) to amplify long PCR products. PCR products for the second intron were electrophoresed in 0.8 % (w/v) agarose gels using 1 x TBE (89 mM Tris, 89 mM boric acid, and 2 mM EDTA (pH 8.0)) buffer, stained with ethidium bromide (0.8 μg/mL) and visualized under UV light. Molecular sizes of amplified fragments were estimated using a 100-bp ladder (Life Technologies, Rockville, Md.)

The 5´UTR F-box intron was also amplified using the degenerate primer pair (F-BOX5´A/F-BOXintronR) developed by Vaughan et al. (2005) from sweet cherry

sequences (Table S2.1). PCR components, thermo-cycler conditions and detection procedure were identical to that described above for the first *S-RNase* intron.

Two additional primers (RFBc-F/SFBins-R), designed from the consensus sequence of the *Prunus SFB* alleles (Vilanova et al., 2006), were used to amplify the SFB_C insertion from genomic DNA of several apricot cultivars in order to distinguish S_C and S_8-alleles (Table S2.1).

_M_-locus genotyping

Seven SSR markers comprised within (or flanking) the *M*-locus were genotyped: PGS3_71, PGS3_22, PGS3_62, PGS3_23 and PGS3_96 (Zuriaga et al., 2012 and 2013) and AGS.20 and AGS.30 (see Chapter III). SSR amplifications were performed in a GeneAmp® PCR System 9700 thermal cycler (Perkin–Elmer, Freemont, CA, USA) in a final volume of 20 ml, containing 75 mM Tris–HCl, pH 8.8; 20 mM (NH4)2SO4; 1.5 mM MgCl2; 0.1 mM of each dNTP; 20 ng of genomic DNA and 1 U of Taq polymerase (Invitrogen, Carlsbad, CA). PCRs were performed as described above by the procedure of Schuelke (2000). The following temperature profile was used: 94ºC for 2 min, then 35 cycles of 94ºC for 45 s, 50-60ºC for 1 min, and 72ºC for 1 min and 15 s, finishing with 72ºC for 5 min. Allele lengths were determined using an ABI Prism 3130 Genetic Analyzer with the aid of GeneMapper software, version 4.0 (Applied Biosystems).

Sequence analysis of PCR products containing _S-RNase_ introns

PCR products containing the first and second introns of S_{20} and S_{24}-*RNase* alleles previously obtained from genomic DNA of cv. 'Ezzine' (S_CS_{24}) on one side and cvs. 'Cow-2' ($S_{20}S_{30}$), 'Harlayne' ($S_{20}S_{31}$), Cristalí ($S_{20}S_C$), Gavatxet ($S_{20}S_C$), Mariem ($S_{20}S_7$), Perla ($S_{20}S_2$), Portici ($S_{20}S_2$), Stella ($S_{20}S_6$), Tadeo ($S_{20}S_C$), Veecot ($S_{20}S_2$) and Velázquez ($S_{20}S_5$) on the other (see Table 2.3) were sequenced to check their identities. Similarly, PCR products associated to the S_{29}, S_{30}, and S_{31}-*RNase* alleles previously obtained from cvs. 'Orange Red' ($S_{29}S_6$), 'SEO' ($S_{29}S_6$), 'Cow-1' (S_1S_{30}), 'Cow-2' ($S_{20}S_{30}$), 'Harlayne' ($S_{20}S_{31}$) and 'Henderson' (S_dS_{31}) were sequenced to confirm they were new alleles. Primer combinations SRc-F/SRc-R and Pru-C2/Pru-C4R were used in all cases for the first and second introns respectively except for S_{24} (Pru-T2/SRc-R) and S_{30} (Pru-C2/Pru-C6R) (see Table 2.3). PCR products were electrophoresed in 0.8% or 2% (w/v) agarose gels (second or first intron, respectively) stained with RedSafe Nucleic acid Staining Solution (iNtRON Biotechnology, Korea) and using TBE 1x buffer. Molecular sizes of

the amplified fragments were estimated using GeneRuler 100bp DNA ladder (Thermo Fisher Scientific). Fragments were extracted and purified from the agarose gels using the Zymoclean Gel DNA Recovery kit (Zymo Research, Irvine, CA, USA). Sequences were determined automatically using an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems) and the Big Dye Terminator Cycle Sequencing Kit v3.1 (Applied Biosystems) following the manufacturer's instructions. Homology searches were performed against the NCBI Genbank database using the BLASTN program (Altschul et al., 1990).

S_2- and S_W-loci sequence analysis

Specific primers designed from the apricot S_2-haplotype sequence (Vilanova et al., 2006) were used to amplify genomic fragments containing the complete *S_2-RNase* (Sf-Hap2/Sr-Hap2) and *SFB_2* coding sequences (FBf-Hap2/FBr-Hap2) (Table S2.1) using 'Portici' (S_2S_{20}) genomic DNA as template. PCR conditions (LD-PCR Techne) and methods for isolating and sequencing (FBF1, SFBc-F, FBF5, FBF6 and FBr-Hap2-2) these bands (Table S2.1) were the same reported above for fragments containing the *S-RNase* second intron.

Whole-genome sequencing of the cv. 'Harlayne' was conducted on an Illumina HiSeq2000 platform, using 100-bp paired-end reads, at genomic facilities of the DHMRI (David H. Murdock Research Institution, Kannapolis, NC, USA; http://www.dhmri.org) and later on kindly provided by Dr. Chris Dardick (USDA-ARS Appalachian Fruit Research Station, USA). 'Harlayne' Illumina sequences were mapped against the peach v1.0 genome sequence (IPGI, http://www.rosaceae.org/species/prunus_persica/genome_v1.0) by using CLC Genomics Workbench 8.0.1 software (http://www.clcbio.com) (Aarhus, Denmark). *S-RNAse* and *S*-locus *F-box* genes corresponding to the S_{31}-haplotype were identified using the variant calling tool (CLCbio) and the S_{20} sequence already published in *Prunus armeniaca* as reference (Zhang et al., 2008).

Clustering analysis

Reference $m_{0\text{-}0}$, $M_{1\text{-}0}$, $M_{2\text{-}0}$ and M_3-haplotypes were established using genetic maps from 'GxCa', 'KxK' and 'GxK' populations through the automatic determination of linkage phases by JoinMap 3.0 (Van Ooijen and Voorrips 2001). Remaining *M*-haplotypes were inferred from SSR genotypes by comparing with the references and

confirmed by using the EM algorithm (Excoffier and Slatkin 1995) implemented in PowerMarker V3.25 software (Liu and Muse 2005). Similarities between *M*-haplotypes were estimated by using Jaccard´s similarity coefficient (Jaccard 1901) through Phyltools 1.32 free-software (Buntjer 1997) and Bruvo´s genetic distance (Bruvo et al., 2004) through a hand-made script. Clustering to reconstruct phylogenetic trees was performed using the Neighbor-Joining algorithm (Saitou and Nei 1987) and HyperTree software was used to visualize the obtained trees (Bingham and Sudarsanam 2000).

Acknowledgements

This work was supported by two grants from the Ministerio de Economia y Competitividad del Gobierno de España (AGL19018-2010) and the Instituto Nacional de Investigaciones Agrarias (RF2011-00020-C02-02). The authors want to thank Laura Ramirez and Mati González for their technical assistance.

Chapter 3:
The *Prunus armeniaca M*-locus Disulfide bond A-like Oxidoreductase (*PaMDOr*) gene is an essential pollen factor for self-incompatibility

This work has been presented in the international congress:

Muñoz-Sanz JV, Zuriaga E, Badenes ML and Romero C. “Prunus Gametophytic Self Incompatibility. A NGS technology approach to identify modifier genes”. XIV International Eucarpia Symposium on Fruit Breeding and Genetics (Bologna, Italy), 07/2015.

Abstract

Prunus spp exhibit a Gametophytic Self-incompatibility (GSI) mechanism, where *S*-RNases (pistil *S*-determinant) and *S*-haplotype-specific F-box (pollen *S*-determinant) genes control specific recognition. However, non-*S*-factors (modifier factors) are also known to be completely necessary for the mechanism to function as pointed out by Pollen Part Mutations (PPMs) conferring self-compatibility in apricot cultivars 'Canino' and 'Katy'. Both PPMs map in an overlapping region at the distal end of chr. 3 named *M*-locus. This work was aimed to the identification of this *S*-locus unlinked PPM using a strategy based on genomic and transcriptomic NGS data. Firstly, an apricot *M*-locus supercontig was obtained after *de novo* assembly of BAC clones from the self-incompatible apricot cv. 'Goldrich'. Next, new recombinant hybrids and molecular markers were used to narrow down the *M*-locus region leading to a physical map of ~134 Kb. On the basis of RNAseq data (from mature anthers, styles and leaves) this refined *M*-locus region was shown to contain 15 genes, four of which over-expressed differentially in mature anthers. Finally, comparative screening of non-synonymous polymorphisms (called by *Illumina* WGS data) in 'Canino' (M_1m), 'Katy' (M_3m) and 'Goldrich' (M_1M_2) *M*-locus led to identify a 358-bp insertion segregating in coupling with the *m*-haplotype in self-compatible apricots. This insertion corresponds to a *FaSt* transposable mutator element and, presumably, leads to a premature stop-codon that produces a truncated protein lacking the C-terminus. The mutated gene codes for a pollen-expressed Disulfide bond A-like Oxidoreductase (named *PaMDOr* from *Prunus armeniaca* *M*-locus Disulfide bond A-like Oxidoreductase). Phylogenetic analysis suggested that *PaMDOr* might have occurred from tandem duplication and its function became essential for the *Prunus S*-RNase-based GSI system. Altogether, evidences support *PaMDOr* as the first non-*S*-factor identified in *Prunus* essential for the GSI mechanism to function.

Introduction

A common feature in angiosperms is the close proximity between male (anther) and female (pistil) reproductive organs increasing the probability of selfing (Barret 2002). This may generate a hazardous situation compromising genetic variability within a species as a consequence of the long-term deleterious effect derived from inbreeding. To escape this problem and enhance outcrossing plants have adopted several strategies

including self-incompatibilty (SI), a molecular mechanism widely spread in plant kingdom. SI is controlled by a multiallelic locus (*S*-locus) encompassing at least two linked transcriptional units acting as female and male *S*-determinants (de Nettancourt, 2001). Rosaceae, Solanaceae and Plantaginaceae are proposed to share the same Gametophytic Self-Incompatibility (GSI) system based on S-RNases as the female *S*-factor (McClure et al., 1989), while the male *S*-determinant codes for an F-Box protein (SLF in Solanaceae and Plantaginaceae, SFBB in *Pyrus* and *Malus* and SFB in *Prunus*) (Lai et al., 2002; Sijacic et al., 2004; Ushijima et al., 2003). S-RNases are style-specific expressed proteins proposed to specifically recognize and reject self-pollen growth by its cytotoxic activity, while unrelated pollen would be able to reach the ovary (Boskovic et al. 1996; Xue et al. 1996).

It still remains unclear how the *S*-RNases and *S*-locus F-box proteins interact with each other. Nevertheless, different proofs support that *S*-locus F-Box proteins are components of a conventional E3 ubiquitin ligase complex aimed to recognize non-self S-RNases promoting their ubiquitination and posterior degradation by 26S proteasome proteolytic pathway (Hua & Kao, 2006; Huang et al. 2006). More recently, a refinement of the inhibitor model (the so-called collaborative model) where several SLFs work together to recognize non-self *S*-RNases has been proposed (Kubo et al., 2010; Entani et al., 2014). Alternatively, Goldraij et al. (2006) proposed the compartmentalization model in *Nicotiana* where pollen endomembrane system plays a key role since *S*-RNases access into pollen tubes via vacuolar compartments, being released or not depending on cross compatibility. Interestingly, *Prunus* GSI mechanism exhibits several differences regarding Solanaceae, Plantaginaceae and even *Maloideae*, but remarkably pollen part mutations (PPMs) truncating *SFB* genes lead to the loss of SI in contrast with the collaborative model (Tao & Iezzoni, 2010; Hegedűs et al., 2012). Matsumoto & Tao (2016) have proposed that SLF-like2 factor acts as a 'general inhibitor' instead of SFB, whose role would be to protect self-S-RNases from degradation. Nevertheless, this working model needs to be more carefully tested.

Apart from *S*-specific factors, other *S*-locus unlinked genes are required for pollen rejection. These non-*S*-specific factors are commonly named as modifier factors or modifier genes. Some of them have been isolated through biochemical studies. For instance, stylar modifier factors identified in *Nicotiana* include HT-B, a small asparagine-rich protein presumably involved in S-RNase discharge by vacuole degradation (McClure et al., 1999;Goldraij et al., 2006); NaStEP, a proteinase inhibitor

that positively regulates HT-B stability (Busot et al., 2008; Jimenez-Duran et al., 2013); 120K, an arabinogalactan protein (AGP) that binds *S*-RNases (Hancock et al., 2005); and NaTrxh, which might function reducing *S*-RNases and/or other proteins from extracellular matrix of the transmitting tract, such as AGPs (Avila-Castañeda et al., 2014; Juárez-Díaz et al., 2006). Pollen non-*S*-factors have been described as well, these include SCF E3 ubiquitin ligase complex components in *Petunia* mentioned above, such as PhSSK1 (Zhao et al., 2010), Rbx1 and Cullin1 (Li et ., 2014). SBP1, another E3 ubiquitin ligase (RING-finger) protein that interacts with SLFs and S-RNases has been identified in *Petunia* (Sims & Ordanic, 2001). NaPCCP protein has been suggested to be involved in AGPs transport through endomembrane system (Lee et al. 2008) and an ABCF transporter has been shown to interact with *S*-RNases and mediates their transport across pollen tube in *Malus domestica* (Meng et al., 2014). In *Prunus spp*, uniquely orthologues to Skp1 and Cullin1 by protein-protein interaction analysis have been proposed (Matsumoto et al. 2012).

Genetic evidence [mainly based on mutations conferring self-compatibility (SC)] has also been compiled for other modifier genes, although they have not been yet identified, for instance, in Solanaceae, the *Nicotiana* 4936 stylar factor (Mc Clure et al., 2000), two pollen *S*-function modifiers in *Solanum tuberosum* (Thompson et al., 1991) and *Petunia axilaris* (Tsukamoto et al., 2003), or the *S*-locus inhibitor (*Sli*) factor from *Solanum chacoense* (Hosaka & Hanneman, 1998a,b). Meanwhile, in Rosaceae, mutations in non-*S*-locus factors conferring SC have been characterized in *Pyrus spp*. (Wu et al., 2013), *Prunus mume* (Wang et al., 2013), *Prunus salicina* (Beppu et al., 2015), *Prunus avium* (Cachi & Wünsch 2011 and 2014) and *Prunus armeniaca* (Vilanova et al., 2006).

Particularly in apricot (*Prunus armeniaca* L.), two different self-compatible cultivars bearing non-*S*-locus mutations have been genetically characterized in deep. On one side, 'Canino' ($S_2S_cM_1m$) carries two independent mutations conferring SC: an insertion in the *SFB* allele leading to a putative truncated protein (S_C-haplotype) and a mutation in a modifier gene (named *m*-allele) (Vilanova et al., 2006). This latter has been mapped in a 364 Kb interval (according to the syntenic peach genome region) at the distal end of chr.3, referred as *M*-locus (Zuriaga et al., 2012). A similar genetic scenario was observed in 'Katy' ($S_1S_2M_3m$), excepting that a unique mutation in a non-*S*-factor was responsible of self-compatible phenotype. Similarly, SC in 'Katy' was found to be due to a unique mutation in a modifier gene located within a 1,2 Mb region

of chr.3 overlapping in 274 Kb with the 'Canino' *M*-locus (Zuriaga et al., 2013). Interestingly, ulterior genetic analysis showed that *m*-haplotype structure defined by SSR alleles is shared by these two and 17 additional self-compatible apricot cultivars from different geographic origins. Furthermore, segregation analysis of the progenies in two of them ('Portici' and 'Corbató') confirmed the presence of the *m*-haplotype (Chapter II). Overall, these results strongly support that *m*-haplotype contains a mutation in a modifier gene conferring SC present in a number of apricot cultivars.

This work was aimed to identify the mutated modifier gene comprised within the *m*-haplotype. For this purpose, and partly due to the intrinsic limitations of working with woody species, an approach mainly based on Next Generation Sequencing (NGS) technologies was performed. Identification of modifier factors required for the GSI system to function is a key step in order to dissect the underlying molecular mechanism in *Prunus*, but also to improve our knowledge on the evolution of GSI within the Rosaceae.

Results

To identify variants within the *m*-haplotype associated with the loss of pollen *S*-function in apricot, a four stepwise strategy based on NGS data was used as follows (Figure 3.1): 1) To get an apricot reference sequence for the *M*-locus from a self-incompatible cultivar. 2) To develop a high resolution map narrowing down the previous *M*-locus map. 3) Gene annotation and differential tissue expression analysis by RNAseq data. 4) Polymorphism screening using genomic *Illumina* data by comparing variants from self-compatible vs. self-incompatible cultivars.

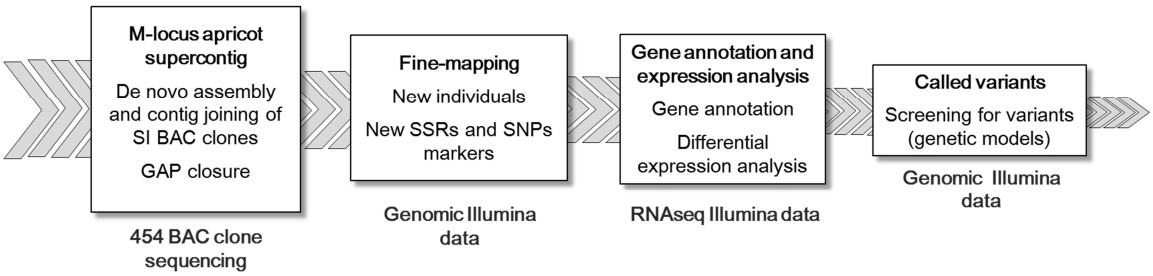


Figure 3.1. Schematic workflow for *m*-mutation identification using NGS data.

Apricot *M*-locus sequence assembly

For the first step, twelve 'Goldrich' (M_1M_2) BAC clones, previously reported to cover 'Canino' *M*-locus (Zuriaga et al., 2012), were pyrosequenced (Figure 3.2 and Table S3.1). Cleaned 454 BAC sequences were assembled, ordered and oriented using *Prunus persica* (peach v1.0 IPGI) and *Prunus mume* (NCBI BioProject under accession PRJNA171605 and http://prunusmumegenome.bjfu.edu.cn) genome sequences as references. Contigs were mostly located at the distal end of the *P. persica* scaffold 3, between 18.380.006 and 18.815.966 positions, except for those belonging to the 251L05 BAC clone 3' half with a high-repeat content, which match regions of the scaffold 2 between 21.831.979 and 25.460.858 positions. When *P. mume* genome was used as reference, *M*-locus contigs obtained per BAC clone were located at the chr. 4 (corresponding to peach chr. 3) between 19.113.622 and 23.060.140 positions (a much larger region than that in peach) whereas contigs with high-repeat content were located in chr. 5 (chr. 2 in peach). A total of thirty contigs spanning *M*-locus region were obtained from self-incompatible cv. 'Goldrich'. Twenty-five were assigned to the M_2-haplotype and the rest to the M_1-haplotype (Figure S3.1 and Table S3.2) as confirmed by PCR-genotyping of SSRs according to the nomenclature reported for the *M*-haplotype in Chapter II (Figure 3.2 and Table S3.4). Contigs from overlapping BAC clones from the same *M*-haplotype (234O11, 148M17, 253J12 and 251L05, on one side, and 160J21, 95D02 and 159P08 on the other) were successfully joined by *GAP4 software*. Afterwards, overlapping contigs derived from different *M*-haplotypes were joined using synteny criteria (see Materials and Methods for detail) remaining 15 unsolved GAPs (Figure S3.1). Subsequently, GAP closure was performed to refine the reference sequence. GAPs known to be within PGS3 series SSRs (8 out of 15) were resolved by editing the sequence and five more by specific PCR-amplification and Sanger sequencing (Table S3.3). Only GAP-13 and GAP-15 could not be resolved. Hence, three major contigs (*M*-locus_contig-1, 2 and 3 with 311.575, 3.193 and 120.995 bp sizes, respectively) were obtained and then joined by indeterminations (N) resulting in a 435.961 bp supercontig, named as apricot *M*-locus supercontig (a*M*-supercontig), that constituted the reference sequence used for subsequent analysis (Figure 3.2).

Narrowing-down the apricot *M*-locus

Zuriaga et al. (2012) defined the *M*-locus in cv. 'Canino' within an interval flanked by PGS3.71 and PGS3.96 SSR markers. Both SSRs are located in positions

18.860 and 380.760 bp within the a*M*-supercontig previously obtained encompassing 361.900 bp. In 'Katy' the *M'*-locus was flanked by PGS3.22 (position 115.850 in the a*M*-supercontig) and EPPCU7190 (Zuriaga et al., 2013) distantly located at the supercontig 3'-end (over 1 Mb according to peach syntenic region). *M*- and *M'*-loci overlap within an interval of 264.940 Kb (between PGS3.22 and PGS3.96) where PGS3.62 and PGS3.23 (Figure 3.2), showed to be fully linked to this mutation in 'Canino' and 'Katy', respectively. This overlapping region encompasses 42 ORFs according to the peach gene annotation. To reduce the number of candidate genes, a new fine-mapping was accomplished for both cultivars by using new recombinant hybrids and new molecular markers identified using the a*M*-supercontig and genomic *Illumina* data (Figure 3.2). In the 'G×C-08' outcross population only one new recombinant was found, M-54 ($S_1S_CM_2m$), which breakpoint is located between PGS3.71 and PGS3.62, like GC-98 recombinant. Two additional 'Katy' F_1 recombinants (K06-18 and K06-37) were also added to narrow down the 'Katy' *M*-locus map (Figure S3.2). New molecular markers were identified in two phases. Firstly, up to 40 SSRs (coded in AGS and 160J21 series) were identified in the a*M*-supercontig (Table S3.4), and secondly, SNPs called from 'Canino', 'Goldrich' and 'Katy' genomic *Illumina* data were used to refine SSR mapping results (Table S3.6). In 'Canino', AGS.20 marker was the unique SSR that could be mapped (the rest were not polymorphic, did not amplify, showed multiband patterns or did not fulfill 'Canino' genetic requirements) and it co-segregated with the mutation (Figure 3.2 and Table S3.4). Regarding SNPs, 5.297 and 5.104 variants were found for 'Canino' and 'Goldrich' respectively (against the a*M*-supercontig), but only 30.37% and 76.23% were heterozygous (Table S3.5). Five loci (SNPCaMmap1 to 5) from the a*M*-supercontig were tested but only one recombination breakpoint (corresponding to SNPCaMmap1) was observed in M-54 15.007 bp downstream of PGS3.71 (Figure 3.2 and Tables S3.5 and S3.6). Regarding 'Katy' fine-mapping, 62.5% of new SSRs markers proved to be polymorphic (Table S3.4) leading to find individuals that recombined between AGS.12 (position 136.201, individuals K05-24 and K06-06) and AGS.30 (position 309.620, individual K06-18), improving substantially previous 'Katy' map (Figure 3.2). In 'Katy', 6.597 variants were found and the 59.83% showed to be heterozygous SNPs (Table S3.5). Eight SNPs, 4 between SSRs AGS.12/AGS.14 and 4 between AGS.27/AGS.30 were tested (Figure 3.2, and Tables S3.5 and S3.6). Recombination breakpoints were observed in positions 142.155 (SNPKaMmap1) and 276.184

(SNPKaMmap7) regarding a*M*-supercontig. This new physical map delimited by the SNPs refines 'Katy' map flanked by SSRs from ~173 Kb to ~134 Kb, a region comprised within the 'Canino' map (Figure 3.2). According to results shown in Chapter II, 'Canino' and 'Katy' share the same *m*-haplotype and therefore this region delimited by SNPKaMmap1/7 in the 'Katy' high-resolution map was considered for subsequent analysis.

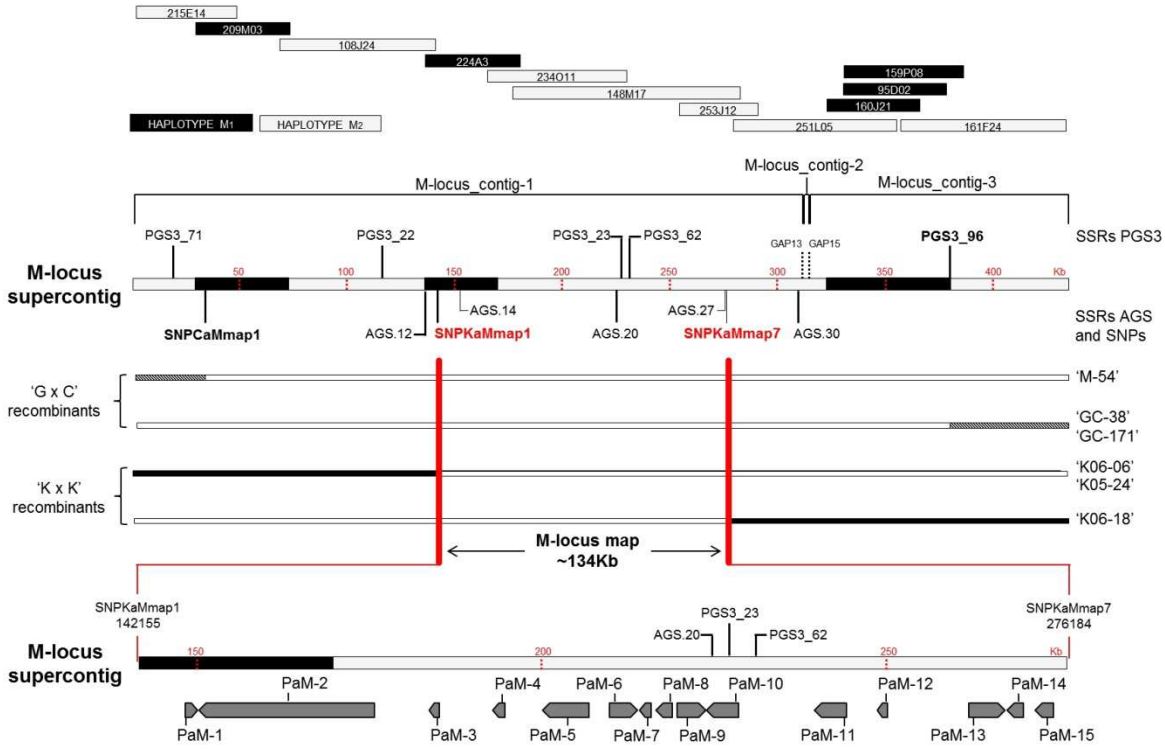


Figure 3.2. Graphical representation of a*M*-supercontig assembly, *M*-locus high resolution map and gene annotation. Upper *black* and *grey rectangles* represent BAC clones from the self-incompatible apricot cv. 'Goldrich' BAC library used for *de novo M*-locus reference sequence assembly (*M*-haplotype is indicated). Resulting a*M*-supercontig assembly (formed by contigs M-locus_contig-1,-2 and -3, and the interspersed GAP-13 and GAP-15) is shown pointing out both *M*-haplotypes (*grey* and *black rectangle*s). The scale in Kb for a*M*-supercontig is shown with *red dotted lines*. SSRs from PGS3 series used in previous works to define 'Canino' and 'Katy' genetic maps are shown above of a*M*-supercontig, while SSRs from AGS series and SNPs markers developed in this work are shown below. *Molecular makers in bold* delimit 'Canino' *M*-locus physical map; recombinant hybrids of 'G×C' populations are shown in *stripped* and *black thin rectangles*. Whereas molecular markers defining 'Katy' *M*-locus physical map are shown in *red bold*; recombinant hybrids of 'Katy' self-pollination are indicated in *white* and *black thin rectangles*. *Vertical red thick lines* show definitive positions delimiting the ~134Kb *M*-locus high resolution physical map. A zoom in this ~134Kb *M*-locus map is shown below where gene annotation results are indicated by *dark grey arrows*. SSRs AGS.20, PGS3.23 and PGS3.62 linked to the mutation are shown as well.

Gene annotation and differential expression analysis

Illumina RNAseq data from mature anthers, mature styles and leaves of apricot cvs. 'Canino', 'Katy' and 'Goldrich' were used for gene annotation and tissue expression analysis (Table S3.1). Trimmed data from all three 'Goldrich'

transcriptomes were aligned through 'Transcript discovery 2.0' (module included in CLC Genomics Workbench 8.0.1). Our own gene annotation for the ~134 Kb region within the a*M*-supercontig was reviewed and manually curated using *P. persica* (peach v1.0 and v2.0 IPGI) and *P. mume* (NCBI BioProject under accession PRJNA171605) annotations as references. A total of 15 ORFs were annotated and named as *PaM*-1 to *PaM*-15 (from *Prunus armeniaca M*-locus) (Figure 3.2). An additional gene homologous to ppa004594m (peach v1.0 IPGI) was excluded of the subsequent analysis because the *M*-locus recombination breakpoint (SNPKaMmap1) was located in the middle of this gene. Up to 4 and 3 genes annotated in *P. persica* (peach v1.0 and v2.0, respectively) and 7 in *P. mume* were not found in our apricot gene annotation (Table 3.1). However, the genomic sequence of the 3 *P. persica* genes (peach v2.0) is highly conserved in the a*M*-supercontig suggesting that they were not expressed in the analyzed tissues used for apricot gene annotation. Apricot annotated genes showed a high homology rate with *P. persica* and *P. mume* predicted CDS and protein sequences, ranging from 93 to 100% and 85 to 100%, respectively (except for ppa011450m, which lower homology might be due to a wrong annotation in peach v1.0) (Table 3.1).

Table 3.1. Apricot *M*-locus high resolution map gene content and homology rate sequence with corresponding putative orthologues of *P. persica* (v1.0 and v2.0, IPGI) and *P.mume* (NCBI BioProject, accession PRJNA171605) for CDS and predicted protein sequences. Start and end positions for each gene within a*M*-supercontig are indicated as well as the gene and protein sizes.

P.armeniaca gene annotation*	Start position	End position	Size gene (nt)/ protein (aa)	Putative orthologue in **P. persica v1.0**	Homology rate: CDS/protein	Putative orthologue in **P. persica v2.0**	Homology rate: CDS/protein	Putative orthologue in **P.mume**	Homology rate: CDS/protein
PaM-1	141885	141885	1901/182	ppa012139m	99,27/100	Prupe.3G248300.1	99,27/100	Pm015410	99,27/99,45
---	---	---	---	---	---	---	---	Pm015409	---
PaM-2	141885	141885	25271/4966	ppa000002m	98,53/98,45	Prupe.3G248400.1	98,53/98,45	Pm015408	99,33/98,98
---	---	---	---	ppa026731m	---	---	---	---	---
---	---	---	---	---	---	---	---	Pm015407	---
---	---	---	---	ppa023507m	---	Prupe.3G248500.1	---	Pm015406	---
PaM-3	141885	141885	1626/424	ppa005351m	96,81/94,02	Prupe.3G248600.1	96,81/94,02	Pm015405	99,6/98,57
PaM-4	141885	141885	1594/360	ppa011450m	47,37/32,76	Prupe.3G248700.1	93,7/94,82	Pm015403	98,15/98,61
PaM-5	141885	141885	6668/786	ppa001620m	95,08/99,24	Prupe.3G248800.1	95,08/99,24	Pm015402	93,3/98,98
PaM-6	141885	141885	4360/227	ppa011007m	98,68/99,12	Prupe.3G248900.1	98,68/99,12	Pm015401	100/100
PaM-7	141885	141885	1842/212	ppa017665m	97,34/96,23	Prupe.3G249000.1	97,34/96,23	Pm015400	99,06/98,58
PaM-8	141885	141885	2688/216	ppa011285m	98,92/99,54	Prupe.3G249100.1	98,92/99,54	Pm015399	99,08/99,54
PaM-9	141885	141885	4183/477	ppa005069m	99,09/99,16	Prupe.3G249200.1	99,09/99,16	Pm015398	99,37/98,95

PaM-10	141885	141885	4663/269	ppa010249m	98,05/95,33	Prupe.3G249300.1	97,04/85,88	Pm015397	98,77/97,03
---	---	---	---	---	---	---	---	Pm015396	---
---	---	---	---	---	---	---	---	Pm015395	---
PaM-11	141885	141885	4769/245	ppa010548m	99,05/98,78	Prupe.3G249400.1	99,05/98,78	Pm015394	99,73/99,18
PaM-12	141885	141885	1320/112	ppa026503m	98,82/99,11	Prupe.3G249500.1	98,82/99,11	Pm015393	99,12/100
---	---	---	---	ppa016385m	---	Prupe.3G249600.1	---	Pm015392	---
---	---	---	---	ppa1027219m	---	Prupe.3G249700.1	---	Pm015391	---
PaM-13	141885	141885	5381/579	ppa003386m	98,79/99,31	Prupe.3G249800.1	98,79/99,31	Pm015390	99,77/99,65
PaM-14	141885	141885	2433/356	ppa007756m	99,07/99,44	Prupe.3G249900.1	99,07/99,44	Pm015389	99,72/99,16
PaM-15	141885	141885	2621/464	ppa005994m	99,08/99,07	Prupe.3G250000.1	99,2/99,28	Pm015388	99,64/99,78

*Gene annotatioin for ~134 Kb region within the a*M*-supercontig

RNAseq data from all tissues among the three cultivars were compared by using Trinity (Haas et al., 2013). Seven out of the fifteen apricot annotated genes were differentially over-expressed in mature anthers regarding leaves (*PaM*-6, -7, -8, -9, -11, -14 and -15) in 'Canino', 'Goldrich' and 'Katy', and 5 (*PaM*-6, -7, -9, -10 and -14) with regards to mature styles in 'Canino' and 'Goldrich' (no mature styles RNAseq data were available for 'Katy') (Figure 3.3). Additional RT-PCR analysis using mature anther RNAs obtained from recombinant hybrids homozygotes for the *M*- and *m*-haplotypes detected gene-expression for all apricot annotated genes (data not shown).

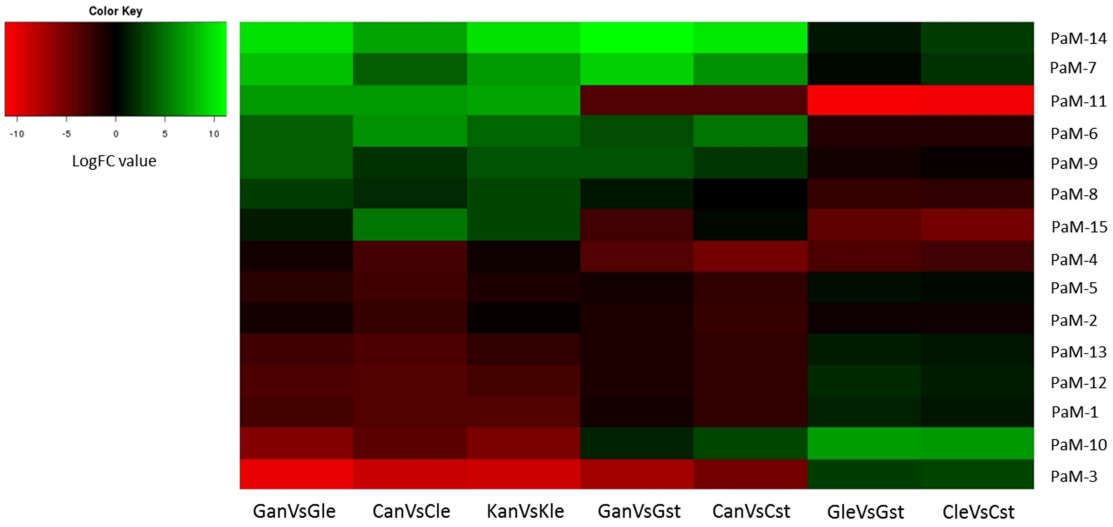


Figure 3.3. Heat map illustrating log fold-change (logFC) values of *M*-locus genes in the pairwise tissue comparison for each apricot cultivar. Colour key indicates logFC (rows) from over-expressed (*green boxes*) to down-expressed (*red boxes*) genes for each cultivar/tissue pairwise comparison (columns). LogFC is calculated between first cultivar/tissue sample against (indicated with Vs) second cultivar/tissue sample. Thus, positive logFC value means a higher expression in the first cultivar/tissue sample regarding second for corresponding gene (green box), negative logFC value (red box) represents a lower expression following the same order of comparison. Cultivar/tissue sample is as follows: (Gan) 'Goldrich' anthers, (Gle) 'Goldrich' leaves, (Gst) 'Goldrich' styles, (Can) 'Canino' anthers, (Cle) 'Canino' leaves, (Cst) 'Canino' styles, (Kan) 'Katy' anthers and (Kle) 'Katy' leaves.

Variant calling

A screening for identifying non-synonymous polymorphisms was carried out for the annotated ~134 Kb region within the a*M*-supercontig (Figure 3.1). Assuming the hypothesis suggesting that 'Canino' and 'Katy' have the same mutation (Chapter II), both cultivars should share a variant in heterozygosis. In addition, one of the two alleles of this variant should be absent in 'Goldrich' leading to a non-synonymous change in the predicted protein in 'Canino' and 'Katy'. Under these assumptions, low stringent conditions and different types of analysis were used for variant calling (SNPs, Indels and Structural Variants) in the three cultivars (Table S3.7). A total of 1.118, 1.136 and 1.365 variants from genomic *Illumina* data were called for 'Goldrich', 'Canino' and 'Katy', respectively, in the ~134 Kb region within the a*M*-supercontig. Of these, 249 and 414 were variants uniquely found in 'Canino' and 'Katy' and not in 'Goldrich', respectively. However, only 39,4% of the 'Canino' variants are heterozygous while the percentage in 'Katy' reaches out the 92,5%. The number of heterozygous variants differing with respect to 'Goldrich' within annotated genes (excluding intergenic regions) decreased noticeably for both cultivars to 20 and 133 (6 and 33 only in exons) for 'Canino' and 'Katy', respectively. A total of 5 and 27 variants in 'Canino' and 'Katy', respectively, lead to non-synonymous changes for predicted proteins (Table 3.2). Lastly, only one of these variants was found to be shared by 'Canino' and 'Katy': an insertion of undetermined size between positions 214.578-214.587 within *PaM-7* (also detected in the transcriptomic alignments).

Table 3.2. Comparative polymorphism screening between self-compatible apricots (Canino' and 'Katy') against self-incompatible apricot cultivar 'Goldrich'. Data corresponding to distinct heterozygous variants regarding 'Goldrich' in annotated *M*-locus genes and exons leading to non-synonymous changes.

Cultivar	Total variants[a]	Variants differing with 'Goldrich'	Heterozygote variants different from 'Goldrich'			
			Total	Annotated genes	Annotated exons	Non-Synonymous
Goldrich	1118	---	---	---	---	---
Canino	1136	249	98	20	6	5
Katy	1365	414	383	133	33	27

[a] All variants found in the ~134 Kb region within the a*M*-supercontig

The presence of the insertion in 'Canino' and 'Katy' and the absence in 'Goldrich' was confirmed by specific PCR-amplification. An extra-band, ~350bp larger

in size than the band shared by the three cultivars and putatively containing the insertion, was only detected in ‘Canino’ and ‘Katy’ (Figure 3.4).

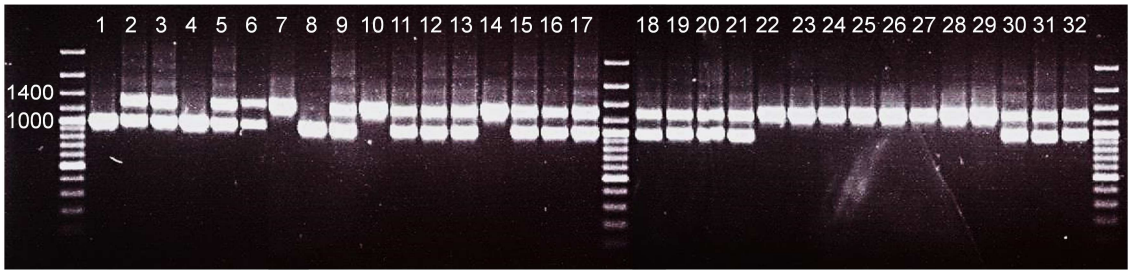
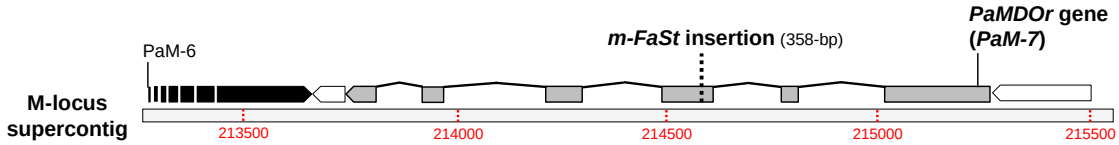


Figure 3.4. *FaSt* insertion genotyping in recombinant hybrids and apricot cultivars bearing *m*-haplotype. PCR amplification of apricot gDNA with primer pair (PaMDsb_F2/PaMDsb_R2) containing the insertion. Samples are as follows: (1) ‘Goldrich’ (M_1M_2), (2) ‘Canino’ (M_1m), (3) ‘GC-38’ (M_1m), (4) ‘GC-98’ (M_1M_2), (5) ‘GC-171’ (M_2m), (6) ‘M-54’ (M_2m), (7) ‘CC-67’ (*mm*), (8) ‘CC-77’ (M_1M_1), (9) ‘Katy’ (M_3m), (10) ‘K05-06’ (*mm*), (11) ‘K05-15’ (M_3m), (12) ‘K05-24’ (M_3m), (13) ‘K06-06’ (M_3m), (14) ‘K06-17’ (*mm*), (15) ‘K06-18’ (M_3m), (16) ‘Castelbrite’ (M_3m), (17) ‘Castleton’ (M_3m), (18) ‘Corbató’ (M_5m), (19) ‘Cow-1’ (M_3m), (20) ‘Cow-2’ (M_1m), (21) ‘Cristalí’ (M_5m), (22) ‘Currot’ (*mm*), (23) ‘Gandia’ (*mm*), (24) ‘Gavatxet’ (*mm*), (25) ‘Ginesta’ (*mm*), (26) ‘Manrí’ (*mm*), (27) ‘Martinet’ (*mm*), (28) ‘Palabras’ (*mm*), (29) ‘Palau’ (*mm*), (30) ‘Portici’ (M_1m), (31) ‘Tadeo’ ($M_{15}m$) y (32) ‘Trevatt’ (M_1m).

A *FallingStone* (*FaSt*) mutator element within *PaM*-7 is in coupling with *m*-haplotype and leads to a putative premature stop-codon

In order to check and characterize whether this insertion might result in a non-functional protein, specific amplification of whole *PaM*-7 genomic region was carried out. Genomic DNA from ‘CC-77’ (M_1M_1 ‘Canino’ self-pollinated recombinant hybrid) was used for *PaM*-7 *M*-allele sequencing, whereas ‘CC-67’ and ‘K06-17’ (*mm* ‘Canino’ and ‘Katy’ self-pollinated recombinant hybrids) were used for *PaM*-7 *m*-allele sequencing. Both sequences proved to be identical except for a 358-bp insertion between positions 332/690 of the *PaM*-7 *m*-allele coding region within the third exon (Figure 3.5a). The 358-bp insertion leads to a substitution of a TTT codon (Phenylalanine) by TGA producing a premature stop-codon of the predicted translated protein in the amino acid position 111 (Figure 3.5b). This insertion is identical in size and highly similar in sequence (86,3%) to the one found in SFB_C by Vilanova et al. (2006). SFB_C insertion was characterized as a Miniature Inverted-repeat Transposable Element (MITE) type named *Falling Stone* (*FaSt*) (Halász et al., 2014). *FaSt* elements contain Target Site Duplications (TSDs), short AT-reach segments and Terminal Inverted Repeats (TIRs). Both structural elements are also present in the *PaM*-7 *m*-allele insertion. TIR elements are well conserved between *PaM*-7 *m*-allele and SFB_C while TSDs differ in AT-repeats content. In addition, the two TSDs observed in *PaM*-7 *m*-allele differ in the first nucleotide (Figure 3.5b).




```
PaMDOr_CC-77(MM)          1   ATGGCTGGAATCAAGAAAAAGCTTGTGGAAATTGATATAATTGCAGACACTGTGTGTCCATGGTGCTTTGTGGGCAAAAGAAACCTTGAC 90
                          1   M  A  G  I  K  K  K  L  V  E  I  D  I  I  A  D  T  V  C  P  W  C  F  V  G  K  R  N  L  D   30
PaMDOr_CC-67/K06-17(mm)   1   ATGGCTGGAATCAAGAAAAAGCTTGTGGAAATTGATATAATTGCAGACACTGTGTGTCCATGGTGCTTTGTGGGCAAAAGAAACCTTGAC 90
                          1   M  A  G  I  K  K  K  L  V  E  I  D  I  I  A  D  T  V  C  P  W  C  F  V  G  K  R  N  L  D   30

PaMDOr_CC-77(MM)         91  AAAGCTCTAGTGGAAGGTAATGATCGATACGAGTTTGAGCTCAGATGGCATCCATTTCAAATTGATCCTGAAGTCCCTAAAGAAGGCATT 180
                         31  K  A  L  V  E  G  N  D  R  Y  E  F  E  L  R  W  H  P  F  Q  I  D  P  E  V  P  K  E  G  I   60
PaMDOr_CC-67/K06-17(mm)  91  AAAGCTCTAGTGGAAGGTAATGATCGATACGAGTTTGAGCTCAGATGGCATCCATTTCAAATTGATCCTGAAGTCCCTAAAGAAGGCATT 180
                         31  K  A  L  V  E  G  N  D  R  Y  E  F  E  L  R  W  H  P  F  Q  I  D  P  E  V  P  K  E  G  I   60

PaMDOr_CC-77(MM)        181 TACAAGAAAGAGTTTTATGATACAAAGATGGGCGCTGATGTAGCTGAAGTGTTTCAGACCCGTATGGCGGATATCTTTTCAAACCATGAC 270
                         61  Y  K  K  E  F  Y  D  T  K  M  G  A  D  V  A  E  V  F  Q  T  R  M  A  D  I  F  S  N  H  D   90
PaMDOr_CC-67/K06-17(mm) 181 TACAAGAAAGAGTTTTATGATACAAAGATGGGCGCTGATGTAGCTGAAGTGTTTCAGACCCGTATGGCGGATATCTTTTCAAACCATGAC 270
                         61  Y  K  K  E  F  Y  D  T  K  M  G  A  D  V  A  E  V  F  Q  T  R  M  A  D  I  F  S  N  H  D   90

PaMDOr_CC-77(MM)        271 ATGACCTATAAGATCGAGGGACTCACGGGAAATACTATTAAGAGTCACAGGCTTATATATT---------------------------- 331
                         91  M  T  Y  K  I  E  G  L  T  G  N  T  I  K  S  H  R  L  I  Y  F                              111
PaMDOr_CC-67/K06-17(mm) 271 ATGACCTATAAGATCGAGGGACTCACGGGAAATACTATTAAGAGTCACAGGCTTATATATTGAGATATTTAGTAATATACCCATTCTTAG 360
                         91  M  T  Y  K  I  E  G  L  T  G  N  T  I  K  S  H  R  L  I  Y  *                               110

PaMDOr_CC-77(MM)        331 ---------------------------------------------------------------------------------------- 331
                        111                                                                                           111
PaMDOr_CC-67/K06-17(mm) 361 CACTAAAACTATAAATAAACCCTACATCATTTCATTTTTATAAACATACCCAAAAAAAACCCAAAAAATGACAAATTGTACTATTAAATT 450
                        110                                                                                           110

PaMDOr_CC-77(MM)        331 ---------------------------------------------------------------------------------------- 331
                        111                                                                                           111
PaMDOr_CC-67/K06-17(mm) 451 TAATTTTAATTATTAAATTACTTTGATACCCTATTGAGTGTTTTGGGTTTTTTTATGAAATTTTGGGTTGGGTTTGTTTTAAGAAATCAA 540
                        110                                                                                           110

PaMDOr_CC-77(MM)        331 ---------------------------------------------------------------------------------------- 331
                        111                                                                                           111
PaMDOr_CC-67/K06-17(mm) 541 TGGCAATTTTGTAATTTAAAAGAAGTTAAAAGCTTTTTTGTTATGTTATAAATGGGCTTTGGGTGTATTTATAAATTCCATTTCATATAG 630
                        110                                                                                           110

PaMDOr_CC-77(MM)        331 ----------------------------------------------------------TTGCTGGGCTACAGGATCATGATAAGCAGCA 362
                        111                                                            A  G  L  Q  D  H  D  K  Q  Q  121
PaMDOr_CC-67/K06-17(mm) 631 GGTTTTTTTATAATTTGGGCTCATATATTGGGTATATTAGTGAATCTCCCATATATATTTTGCTGGGCTACAGGATCATGATAAGCAGCA 720
                        110                                                                                           110
```

Figure 3.5 *FaSt* insertion within *PaM-7 m*-allele. a) Schematic representation of *PaM*-7 gene structure in the a*M*-supercontig. *Light grey rectangle* represents region of a*M*-supercontig containing *PaM*-7 gene and *dotted red lines* show positions in Kb. *PaM*-6 is indicated with *black arrow*, *white arrows* symbolize *PaM*-7 UTR regions and *dark grey arrows PaM*-7 exons (reverse orientation). *PaM*-7 *m*-allele *FaSt* insertion is indicated in the third exon. b) Partial CDS (predicted by NGS) nucleotide alignment of *PaM*-7 *m*- and *M*-alleles as well as codifying amino acid sequence (homozygous recombinant hybrids sequenced are indicated: CC-67, K06-17 and CC-77). *FaSt* insertion starts after position 331 (*blue line*) and lead to a premature stop-codon (shown by an *asterisk*). Target site duplications and terminal inverted repeats in *FaSt* transposable elements are indicated by *red letters* and *grey shading*, respectively. Dicysteinic redox motif typically conserved in DsbA proteins is shown with *bold letters* and into *black frame* of the amino acid sequence (CPWC domain).

The *PaM*-7 *m*-allele insertion was PCR-amplified in all self-compatible apricot cultivars (other than ‘Canino’ and ‘Katy’) known to carry the *m*-haplotype (Chapter II) and recombinant hybrids derived from ‘Canino’ and ‘Katy’. Heterozygote (*Mm*) and homozygote (*MM* and *mm*) cultivars/accessions were included (Figure 3.4). The presence of the *PaM*-7 *m*-allele insertion was detected in all self-compatible cultivars bearing the *m*-haplotype, and homozygote recombinants confirmed that the insertion is in coupling with the *m*-haplotype.

***PaM*-7 codes for a Disulfide bond A-like Oxidoreductase (PaMDOr). Analysis of the phylogenetic relationships.**

The predicted translated protein PaM-7 is homologous to oxidoreductases that contain a Disulfide bond A-like (DsbA-like) domain (IPR001853; PF01323) and belong to the large Thioredoxin (TRX)-like superfamily (proteins containing a Thioredoxin fold; IPR012336) (Table 3.3). Accordingly, *PaM*-7 was renamed as *PaMDOr* (*Prunus armeniaca M*-locus DsbA-like Oxidoreductase). These proteins have a characteristic motif containing two active cysteines, separated by 2 amino acids (CXXC), responsible for the redox state in target proteins; the motif sequence in PaMDOr is CPWC (Figure 3.5b).

The direct BLASTP analysis performed against NCBI 'non-redundant protein sequences' database using PaMDOr translated protein as query found homologues from different plant species with no more than one-two hits per specie with E-values lower than 10^{-3}. Table 3.3 shows direct BLASTP hits belonging to families where SI molecular mechanism has been partially elucidated (Rosaceae, Solanaceae and Brassicaceae) excluding Papaveraceae for which no hits were found. First two hits were Pm015400 and ppa017665m from *P. mume* and *P. persica*, respectively (Table 3.3). Third hit ppa011285m corresponds to a putative protein contiguous to ppa017665 in *P. persica* highly homologous to the apricot PaM-8 (Figure 3.2 and Table 3.1). Nevertheless, there is a significant difference in homology with ppa017665m and ppa011285m decreasing from $6{,}32e^{-148}$ to $1{,}2e^{-96}$ E-values, respectively.

Table 3.3. PaMDOr 'Direct BLASTP'. 'NCBI accession name' column shows selected hits of PaMDOr BLASTP output using PaMDOr predicted protein as query. 'Genome database accession name' refers to the annotated ID in the corresponding protein database (see Material and Methods). In 'NCBI description' column, specie which accession belongs to is indicated in square brackets

NCBI accession name	Genome database accession name	Max score	Query coverage	E-value	Identity	NCBI description
XP_008230297.1	Pm15400	429	100%	7,0E-154	98%	uncharacterized protein LOC103329582 [Prunus mume]
XP_007216055.1	ppa017665m	421	100%	9,0E-151	96%	hypothetical protein PRUPE_ppa017665mg [Prunus persica]
XP_007215948.1	ppa011285m	291	100%	1,0E-99	62%	hypothetical protein PRUPE_ppa011285mg [Prunus persica]
XP_008379454.1	MDP0000233548	291	100%	3,0E-99	62%	uncharacterized protein LOC103442449 [Malus domestica]

XP_008230295.1	Pm15399	291	100%	1,0E-99	62%	uncharacterized protein LOC103329581 [Prunus mume]
XP_004304201.1	gene04226-v1.0-hybrid	284	99%	2,0E-96	60%	uncharacterized protein LOC101310818 [Fragaria vesca subsp. vesca]
XP_008341809.1	MDP0000148485	283	100%	3,0E-96	59%	uncharacterized protein LOC103404656 [Malus domestica]
XP_004306054.1	gene04224-v1.0-hybrid	262	98%	5,0E-88	62%	uncharacterized protein LOC101292444 [Fragaria vesca subsp. vesca]
XP_009774377.1		256	98%	5,0E-85	53%	uncharacterized protein LOC104224422 [Nicotiana sylvestris]
XP_009608607.1		256	97%	5,0E-85	54%	uncharacterized protein LOC104102573 isoform X2 [Nicotiana tomentosiformis]
XP_009608606.1		256	97%	6,0E-85	54%	uncharacterized protein LOC104102573 isoform X1 [Nicotiana tomentosiformis]
XP_004232135.1	Solyc02g089230.2.1	255	97%	6,0E-85	54%	uncharacterized protein LOC101251049 [Solanum lycopersicum]
NP_198706.1	AT5G38900.1	251	97%	1,0E-87	55%	Thioredoxin superfamily protein [Arabidopsis thaliana]

To search for putative *PaMDOr* orthologues a three step approach was followed: 1) 'Reciprocal Best BLASTP Hit' (RBH) analysis; 2) Syntenic blocks identification across species; 3) Inference of phylogenetic relationships on the basis of clustering (tree-based) methods. Proteins from Rosaceae subfamilies *Maloideae*, *Potintilleae* and *Prunoideae/Amygdaloideae* along with proteins from Solanaceae and Brassicaceae families were used as queries for RBH identification using NCBI database (Figure 3.6). Within *Prunus*, RBHs were detected between Pm015400/ppa017665 and Pm015399/ppa011285m. RBHs were also detected for the pair Pm015399/ppa011285m in all the rest of species but not for Pm015400/ppa017665 (Figure 3.6).

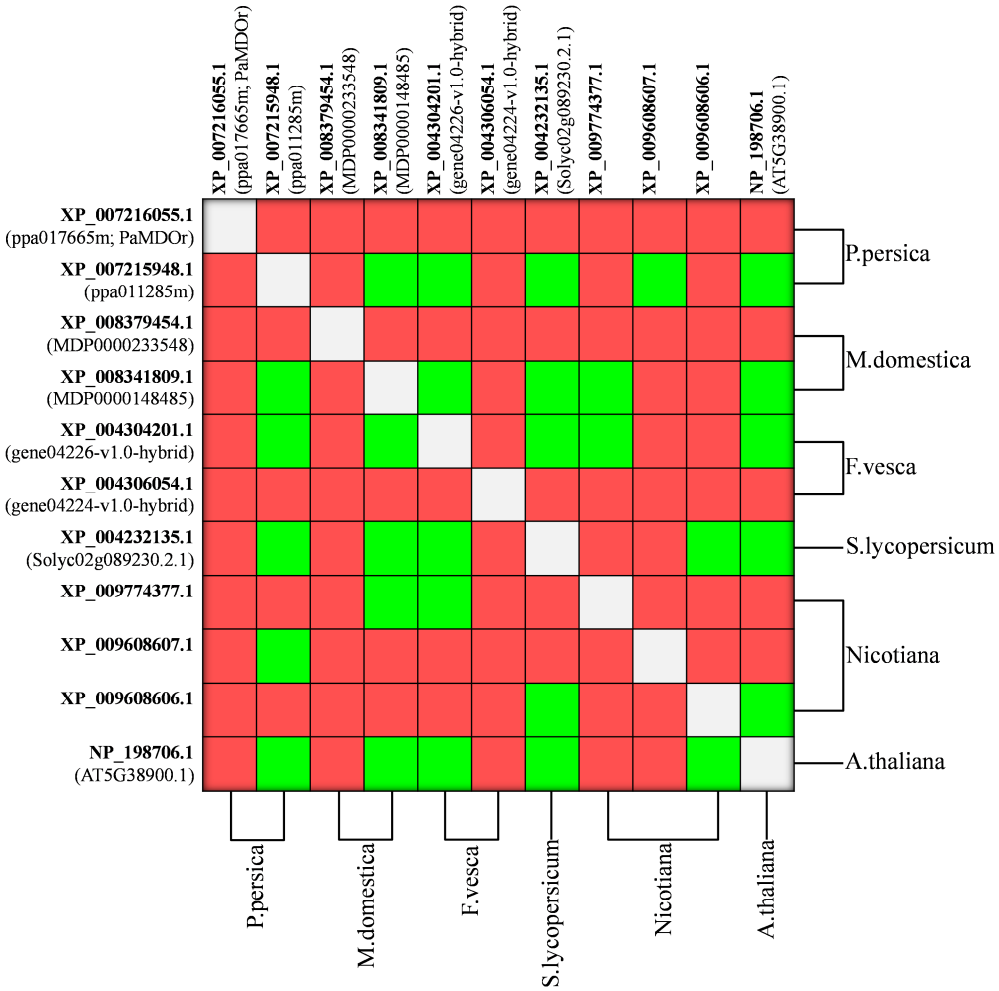


Figure 3.6. PaMDOr/PaM-8 RBH results. Reciprocal Best Hit outcome of accessions from Table 3.3 per pairwise comparison is shown in each box, where *green boxes* refer to RBHs, while *red boxes* to non-RBHs. NCBI (*bold*) and genome protein database (within *brackets*) (see Material and Methods) accessions are indicated in the left and upper side of RBH square, whereas specie which belongs each accession is indicated in the right and lower sides. For values of similarity from BLASTP analysis consult Table S3.8.

Syntenic blocks for the *P. persica M*-locus region between ppa001611m and ppa001157m (~18,4-18,8 Mb in peach scaffold_3 v1.0 IPGI) were found in *Malus domestica* chrs. 9 and 17 (~4,3-4,7 Mb and ~4,9-5,3 Mb, respectively), *Solanum lycopersicum* chr. 2 (~45,6-45,8 Mb) and *Arabidopsis thaliana* chrs. 3 (regions ~0,4-0,5 Mb; ~5,1-5,2 Mb; ~11,2-12,0 Mb) and 5 (~15,5-15,6 Mb) (Figure 3.7).

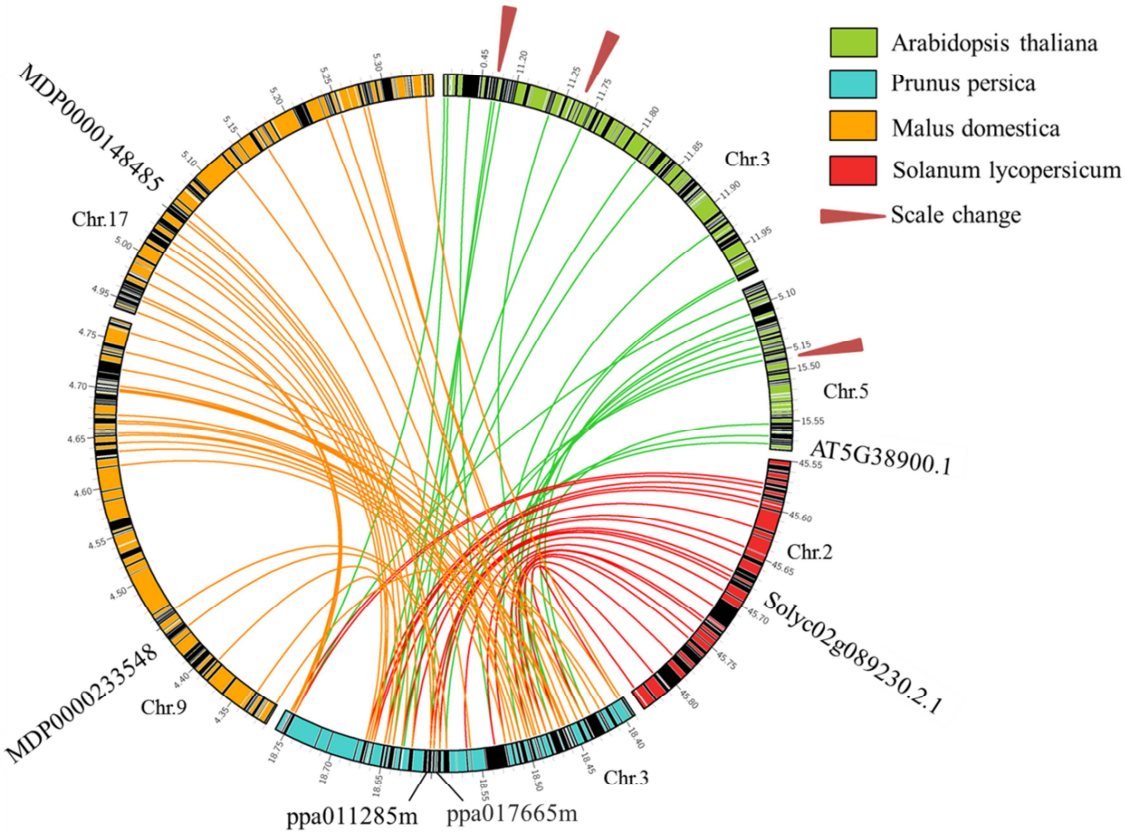


Figure 3.7. *M*-locus syntenic blocks among Rosaceae, Solanaceae and Brassicaceae. *M*-locus syntenic blocks in *Prunus persica* (*Pruinodeae*; SFB/S-RNase SI mechanism; blue) and *Malus domestica* (*Maloideae*; SFBB/S-RNase SI mechanism; orange) Rosaceae subfamilies, and *Solanum lycopersicum* (Solanaceae; SLF/S-RNase SI mechanism; red) and *Arabidopsis thaliana* (Brassicaceae; Sporophytic SI mechanism; green) non-Rosaceae families. *Black rectangles* within circular genome regions represent gene annotation in scale, *orange lines* are anchors between *P. persica* and *M. domestica*, *red lines* are anchors between *P. persica* and *S. lycopersicum* and *green lines* are anchors between *P. persica* and *A. thaliana*. *Red triangles* indicate a scale change. The putative *PaMDOr* and PaM-8 orthologues are shown for each specie.

Figure 3.8 shows phylogenetic relationships among proteins identified by direct BLASTP analysis. Brassicaceae and Solanaceae proteins grouped separately but both groups clustered with a third one including *Maloideae*, *Prunus* (orthologous to PaM-8) and *Fragaria* (XP_004304201.1/gene04226-v1.0-hybrid) proteins. According to the conserved dicysteinic site, this major group was named $CPWC_1$, whereas *Prunus* proteins group with higher similarity to PaMDOr was named $CPWC_2$. *F. vesca* XP_004306054.1/gene04224-v1.0-hybrid protein branched separately from $CPWC_1$ and $CPWC_2$ groups but closer to second. This protein contains a SPWC domain losing the first Cys residue, which is thought to be the most important of the two for redox activity (Grauschopf et al., 1995).

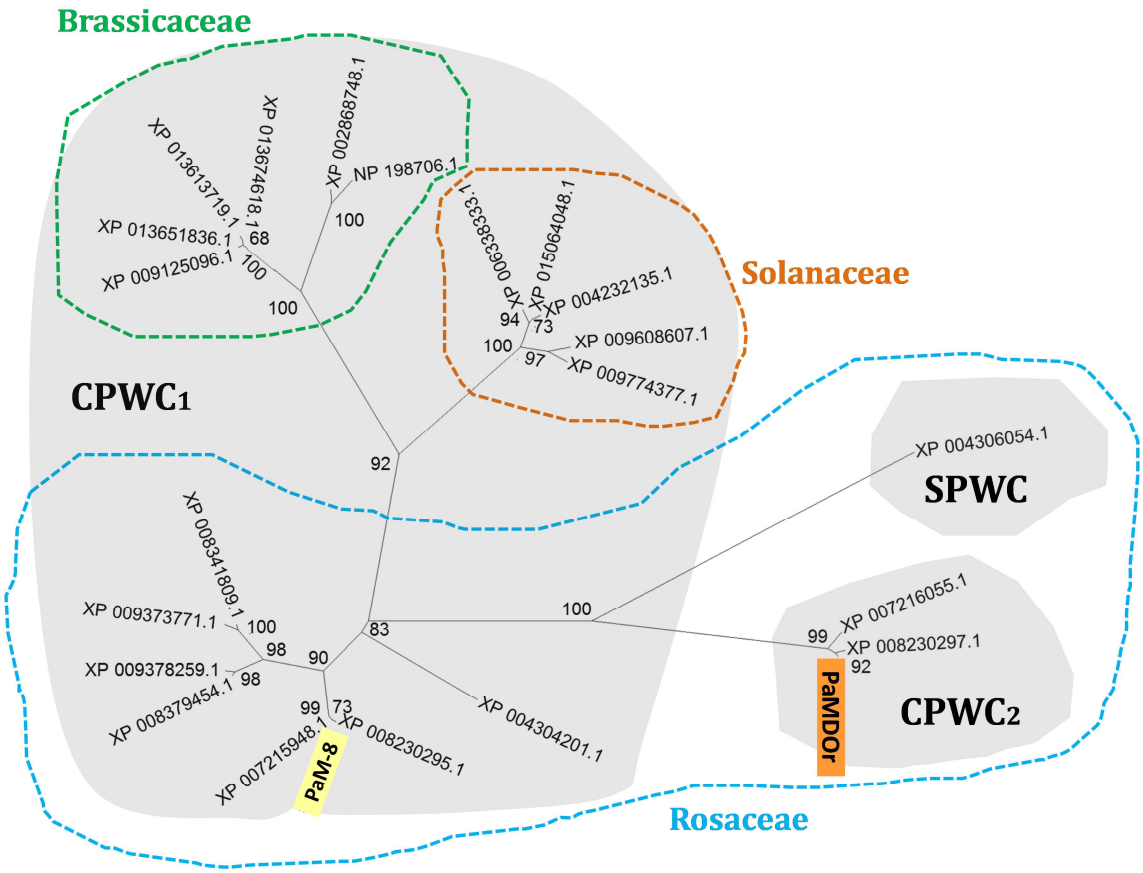


Figure 3.8. PaMDor and PaM-8 phylogenetic tree analysis. Maximum Likelihood phylogenetic tree for selected accessions from 'Direct BLASTP' output as well as PaMDor (*orange square*) and PaM-8 (*yellow square*) proteins. Bootstrap values are shown for every node. Three groups ($CPWC_1$, $CPWC_2$ and SPWC) correspond to proteins homologous to PaM-8, *PaMDOr* and XP_004306054.1//gene04224-v1.0-hybrid, respectively (*grey shading*). *Dashed lines* group Brassicaceae (*green*), Solanaceae (*orange*) and Rosaceae (*blue*) accessions.

Discussion

Paving the way for the identification of the mutated *M*-locus modifier gene

Previous works identified two heterozygous PPMs in apricot cvs. 'Canino' and 'Katy' within *M*- and *M*'-loci, respectively. Interestingly, both mutations mapped in an overlapping region at the distal end of chr. 3, pointing out that the same gene might be affected. However, according to peach genome annotation, defined intervals comprised ~60 ORFs in 'Canino' and more than a hundred in 'Katy' (Zuriaga et al., 2012 and 2013). Therefore, mapping refinement was required before starting a positional cloning strategy. Incorporation of new molecular markers was necessary to achieve this goal but the peach genome reference sequence had already been exhausted. Thus, 'Goldrich' BAC clones covering the *M*-locus were sequenced and assembled to get an apricot

reference sequence useful for this purpose. Three major contigs were obtained and GAPs were joined by indeterminations defining the a*M*-supercontig.

This strategy provided 40 new SSRs and first apricot SNPs for the analyzed region. Unfortunately, only one SSR, co-segregant with the mutation, could be mapped in 'Canino' (AGS.20). Similarly, only one third of the SNP variants found in 'Canino' were heterozygous. Finally, from the 5 selected variants tested in 'GxC' recombinants only one was useful to reduce the mapping region ~15 kb downstream PGS3.71 through the newly identified recombinant M-54. Fine-mapping results were much more successful in 'Katy', largely due to the higher number of recombinants and markers that could be incorporated. Thus, new recombination breakpoints were detected by SSR analysis in AGS.12 and AGS.30 encompassing an interval of ~174 Kb. This map was further improved by using SNPs such as SNPKaMmap1 and SNPKaMmap7 reducing the interval to ~134 Kb. Interestingly, this region is included within the 'Canino' *M*-locus fine-map and excludes the fragment containing high-repeat content.

Moreover, recent findings suggest that 'Canino' and 'Katy' share the same *m*-haplotype supporting the hypothesis that the two share the same PPM as well (Chapter II). Altogether, this apricot ~134 Kb *M*-locus region was decided to be screened for the identification of the PPM.

Candidate genes: discrimination from expression patterns

Fifteen genes were annotated in the apricot *M*-locus region (~134 Kb) using RNAseq data and all of them were found to be highly conserved in other *Prunus spp.* according to collinearity and homology rates. Three additional ORFs were consistently predicted in *P. persica* and *P. mume* suggesting that these genes might be not expressed in the apricot sequenced tissues. Discrepancies in gene content were also found between *P. mume* and *P. persica* (v1.0 and v2.0) but this might result from the annotation methodology used in each case. In fact, RNAseq data and EST collections have been used in *P. persica* whereas in *P. mume* an *ab initio* prediction approach was employed (Zhang et al., 2012; Verde et al., 2013).

The 15 genes showed to be expressed in all tissues, therefore no specific pollen-expressed genes are contained in this region. However, four of these (*PaM*-6, -7, -9 and -14) showed higher differential over-expression in mature anthers with regards to other tissues and therefore may be considered as candidate genes. Furthermore, neither RNAseq data nor RT-PCR analysis support miss-expression as the cause of SC since

positive expression were detected for all 15 genes in homozygote *MM* and *mm* recombinants.

A unique variant within the *M*-locus fulfills genetic requirements

In parallel to gene-expression analysis, variants of any nature, from SNPs to structural variants, were called for the apricot *M*-locus region in the three reference self-incompatible/self-compatible cultivars. The number of variants different to those found in 'Goldrich' identified in 'Katy' (414) was almost twice than in 'Canino' (249). This might be explained because 'Canino' (M_1m) shares the M_1-haplotype with 'Goldrich' (M_1M_2) whereas 'Katy' (M_3 m) does not (Chapter II).

To fulfill the genetic requirements of *S*-locus unlinked SC in 'Canino and 'Katy' both PPMs should be in heterozygosis, hence homozygote called variants were discarded. 'Canino' showed a much lower percentage of heterozygote variants compared with 'Katy', confirming previous results and suggesting that currently cultivated 'Canino' apricot might have arisen from self-pollination event/s (Zuriaga et al., 2012 and 2013). According to genetic and genomic background data in 'Canino' it might be anticipated a low number of variants within the *M*-locus fulfilling the exposed requirements. In fact, only 27 and 5 variants leading to non-synonymous changes in exons were found in 'Katy' and 'Canino', respectively. More restrictively, only one of them was present in both, an insertion detected by using genomic and transcriptomic alignments based on NGS data. It was also confirmed by PCR-amplification in these two cultivars showing an extra-band ~350 bp larger in size than the band shared with 'Goldrich'. This insertion is located within *PaM*-7, very close to microsatellite markers AGS.20, PGS3.23 and PGS3.62 previously shown to be fully linked to the PPM (Zuriaga et al., 2012 and 2013; Chapter II). Finally, the 358-bp insertion was found to be in coupling with the *m*-haplotype not only in the recombinant hybrids analyzed but also in many other cultivars (Chapter II). Therefore, this variant was the only one fulfilling all genetic requirements for being the cause of SC within the *m*-haplotype. *PaM*-7 was fully sequenced for *M*/*m*-alleles and the 358-bp insertion was found to putatively lead to a premature stop-codon in the predicted protein lacking 4 out of the 6 exons. Furthermore, *PaM*-7 was one of the four genes differentially over-expressed in pollen, in agreement with the tissue-specific expression expected for the *M*-locus mutated modifier gene.

This insertion has identical size and shows high sequence similarity with the SFB_C-insertion previously found to confer SC in 'Canino' (Vilanova et al., 2006). SFB_C-insertion was characterized as an active non-autonomous mutator (transposable) element [named *FallingStone* (*FaSt*)] containing structural features that have been proved to be also present in the *PaM-7* insertion (Halász et al. 2014). *FaSt* elements are suggested to have a recent specific occurrence in *Prunoideae* subfamily, being accumulated in gene-rich regions of the *Prunus* genome and authors advanced further effects that would be likely identified in the future (Halász et al., 2014). This work seems to confirm their hypothesis. However, these are not isolated cases. Most of the mutations conferring SC in *Prunus* have been associated with transposable elements. For instance, the insertions of a 615-bp *Ds*-like element into $SFB_{1'}$ (Hauck et al., 2006) and that of a 2,6 Kb Mu element upstream of the S_{6m2}-RNase (Yamane et al., 2003) in *P. cerasus*, the insertions of 115-bp and 5-bp direct repeats within the SFB_1 HVb region and between the SFB_2 V1 and V2 regions in *P. persica* (Tao et al., 2007), or the insertion of 6,8 Kb in the SFB_f coding region in *P. mume* (Ushijima et al., 2004).

PaMDOr is an oxidoreductase essential for the *Prunus* GSI system to function

PaM-7 encodes an oxidoreductase that contains a Thioredoxin fold domain (IPR012336). Proteins having this domain form a large and diverse protein superfamily characterized by a CXXC motif, which confers the thiol-disulfide redox activity essential for folding, stability and function in target proteins (Hogg, 2003; Schmidt et al., 2006). Proteins containing this domain have been associated with a wide range of events during sexual plant reproduction, from gametophyte formation to seed setting (either for their redox activity or as signaling factors) specially under the control of thioredoxin (TRXs) and glutaredoxin (GRXs) proteins (Traverso et al., 2013). More specifically, several TRXs participating in SI systems have already been identified. In *Phalaris coerulescens*, a protein containing a TRX motif in the C-terminal end and expressed in mature pollen grains has been shown to be essential for the SI response (Li et al., 1996). In *Brassica*, pistil TRX proteins THL-1 and THL-2 have been related to the SSI system being suggested to prevent SRK autophosphorilation (Cabrillac et al., 2001). Functional analysis showed that antisense transgenic lines led to reduced levels of SI supporting their requirement for the system to function (Haffani et al., 2004). Lastly, NaTrxh is a TRX protein with a novel Nβ signal peptide (that leads to use the endoplasmic reticulum, Golgi apparatus and vesicles for secretion) localized in the

extracellular matrix of the *Nicotiana* stylar transmitting tract that interacts with S-RNases reducing them (Juárez-Díaz et al., 2006; Avila-Castañeda et al., 2014).

All these TRXs taking part in SI belong to h-type group of TRX proteins (up to 8 groups have been classified in plants to date) and normally work reducing target proteins (Meyer et al., 2012). However, on the contrary, *PaM-7* does not codes for neither a TRX type h nor other TRX type but for a protein containing a Disulfide bond A-like (DsbA-like) domain (IPR001853; PF01323). DsbA-like proteins were firstly identified in *Escherichia coli* as disulfide bond introducers in the periplasm, a necessary process for protein folding (Depuydt et al., 2011). Therefore, DsbA-like proteins are not usually reducing enzymes such as TRX proteins but oxidizing. However, proteins of the TRX superfamily are intrinsically bidirectional, thus can catalyze either oxidation or reduction depending on the redox states in which they are maintained (Ito & Inaba, 2008). Accordingly, *PaM-7* was renamed as *Prunus armeniaca M*-locus Disulfide bond A-like Oxidorreductase (*PaMDOr*), which dicysteinic motif CPWC is located at the protein N-terminal end (Cys19-PW-Cys22). Interestingly, *E.coli* DsbA defective mutants (*dsba*-) showed partially and fully restored function in heterologous systems expressing *Arabidopsis thaliana* Protein Disulfide Isomerase-like (AtPDIL) factors (proteins containing 2 TRX fold domains) (Yuen et al., 2013). Particularly, AtPDIL1-1 has also been associated with the regulation of Programmed Cell Death (PCD) (Onda, 2013). Concretely, AtPDIL-1 acts as a redox-sensitive regulator of the activity of noncaspase-type proteins by the prevention of their premature activation during embryogenesis and to control the timing of the onset of PCD by protease activation (Cho et al., 2011). PCD is well known to be involved in pollen rejection in *Papaver* (Bosch & Franklin-Tong, 2008) but evidences also suggest a role for PCD in the *Pyrus* S-RNase based GSI system (Wang & Zhang, 2011). In this context, a possible role of PaMDOr regarding PCD may be speculated. On the other hand, it is well known that pollen-expressed *S*-locus F-box proteins are cytoplasmic, thus some S-RNases must exit the luminal compartment, possibly after retrograde transport to the endoplasmic reticulum, in order to interact (McClure et al., 2011). Interestingly, AtPDIL1-1 is orthologous to the *Glycine max* GmPDIL-1 which showed, along with GmPDIL-2, the capability to refold denatured RNaseA by oxidative activity in recombinant versions of both genes expressed in *E. coli* (Kamauchi et al., 2008). Since stylar non-*S*-factor NaTrxh was observed to reduce S-RNase *in vitro*, it has been suggested that protein three-dimensional structure might be altered to favor its trafficking in pollen tubes (may

be through the pollen endomembrane system) (McClure et al., 2011) and, consequently, affecting its function. This function should be restored by protein refolding to the active state in an oxidative-dependent manner. AtPDILs have been detected in endoplasmic reticulum (Yuen et al., 2013), therefore, S-RNase refolding (after releasing) might be carried out by an oxidizing protein like PaMDOr similarly to GmPDIL-1/AtPDIL1-1 restore RNase/DsbA activity in the endoplasmic reticulum.

A hypothesis on the evolutionary history of PaMDor in *Prunus*

Reciprocal Best Hits for PaMDOr *P. persica* homologous protein (ppa017665) were detected in *Prunus spp*. but not in non-*Prunus* species (including *Maloideae, Potentilleae, Solanaceae* and *Brassicaceae*). Best hit in all these latter cases was ppa011285m protein (peach v1.0 IPGI) homologous to apricot PaM-8. However, the best direct BLASTP hit of the *Fragaria vesca* accession XP_004306054.1 was the *P. mume* accession XP_008230297.1 found to be homologous to PaMDOr. Since PaM-8 seems to have orthologous in all considered species, whereas PaMDOr has not, a tandem duplication in Rosaceae is suggested where PaMDOr and PaM-8 might be paralogues. Regarding this point, it is relevant to highlight that *PaM-8* is differentially over-expressed in reproductive tissues (anthers and pistils) but not in leaves whereas *PaMDOr* expression is largely specific to anthers. Interestingly, no *Maloideae* protein was found to be homologous to PaMDOr. This might suggest that *Maloideae* species may have lost the PaMDOr orthologue throughout its recent evolutionary history, but sooner than genome duplication event occurred in this subfamily (Velasco et al., 2010). In contrast, *F. vesca* has $CPWC_1$ and SPWC, putatively derived from a $CPWC_2$ ancestor, suggesting that tandem duplication might have taken place before subfamily split in the Rosaceae (~62 MYA). The long period of time elapsed might explain the high divergence between *Fragaria*/*Prunus* CPWC putative orthologues. In agreement with this, it has been shown that *Malus* and *Prunus* S-RNases and SFBs evolved from different lineages supporting a convergent evolution at the GSI system for these two genera (Aguiar et al., 2015; Akagi et al., 2016; Morimoto et al., 2015). This observation is also supported by differences in the GSI system, since a single SFB controls self-recognition in *Prunoideae* and multiple SFBB determine non-self-recognition in *Malus* (Matsumoto & Tao, 2016; Sassa, 2016; Tao & Iezzoni, 2010). A different scenario might imply an alleged duplication occurred uniquely in *Prunus* after splitting from

Malus (~32 MYA) and therefore XP_004306054.1 and PaMDOr would not be orthologous.

Syntenic blocks for *M*-locus region were clearly observed in *M. domestica* and *S. lycopersicum*. As expected, according to a recent genome duplication event, two different regions were shown to be syntenic in *Malus* (Velasco et al., 2010). Remarkably, syntenic blocks in *Arabidopsis* were not well defined at odds with *Malus* and *Solanum*, despite *Arabidopsis* is phylogenetically closer to *Prunus* (Igic and Kohn 2001). It could be speculated that these results from the conservation of regions containing factors needed for a common pathway or mechanism, since Solanaceae and Rosaceae share the same GSI system but *Arabidopsis* exhibits SSI (Takayama & Isogai, 2005).

Overall orthologue study supports a divergent evolution for *M*-locus DsbA proteins in the Rosaceae family. However, putative paralogues ($CPWC_2$ and $CPWC_1$) arose from gene duplication in tandem, being the function of the $CPWC_2$ type proteins specifically related to GSI. In this sense, $CPWC_1$/$CPWC_2$ divergence process might shed some light of *Malus/Prunus* GSI evolution as well.

Material and methods

Plant material

Two self-compatible apricot cvs. 'Canino' ($S_2S_CM_1m$) and 'Katy' ($S_1S_2M_3m$), and the self-incompatible control 'Goldrich' ($S_1S_2M_1M_2$) were selected for genomic and transcriptomic Next Generation Sequencing. Recombinant hybrids 'GC-38', 'GC-98' and 'GC-171' from the outcross populations 'Goldrich × Canino-01' and 'M-54' from 'Goldrich × Canino-08' were used for 'Canino' *M*-locus fine-mapping (Zuriaga et al., 2012). F_1 recombinant hybrids 'K05-24', 'K06-06', 'K06-18' and 'K06-37' from 'Katy' self-pollination (Zuriaga et al., 2013) were used for 'Katy' *M*-locus fine-mapping. All these trees are maintained at the collection of the Instituto Valenciano de Investigaciones Agrarias (IVIA) in Valencia (Spain).

'Canino' self-pollinated hybrids homozygous for the *M*-locus 'CC-67' (S_2S_2mm) and 'CC-77' ($S_CS_CM_1M_1$), F_1 self-pollinated 'Katy' recombinant hybrids 'K05-06', 'K05-15', 'K05-24', 'K06-06', 'K06-17', 'K06-18' (Zuriaga et al., 2013) and apricot cvs. 'Castelbrite', 'Castleton', 'Corbató', 'Cow-1', 'Cow-2', 'Cristalí', 'Currot', 'Gandia', 'Gavatxet', 'Ginesta', 'Manrí', 'Martinet', 'Palabras', 'Palau', 'Portici',

'Tadeo' y 'Trevatt' bearing *m*-haplotype (Chapter II) were used for *FaSt* insertion PCR-genotyping.

Nucleic acids extraction

Two leaf discs were collected from each sample and frozen at -80ºC before genomic DNA isolation following the method of Doyle & Doyle (1987). Total RNA was extracted from leaves, mature anthers (containing mature pollen grains) and mature styles of balloon-stage flowers using the RNeasy Plant Mini Kit (Qiagen), including column DNase treatment (Qiagen RNase free DNase). DNA and RNA quantification was performed by NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE) and integrity was checked by gel electrophoresis. For BAC clone DNA delivery, frozen BAC clones were grown for 20 h at 37ºC in liquid LB medium with cloranphenicol. Each BAC clone was inoculated in 1.5 ml solid LB medium and sent to sequencing. Macrogen Inc. kept on DNA BAC clone isolation and purification.

Next Generation Sequencing and cleaning

Apricot BAC clones 215E14, 209M03, 108J24, 224A3, 234O11, 148M17, 253J12, 251L05, 159P08, 95D02, 160J21 and 161F24 from the self-incompatible cv. 'Goldrich' BAc library (Vilanova et al., 2003) were pyrosequenced by 454 GS-FLX Titanium NGS technology (Roche), commercially conducted by Macrogen Inc. (Seoul, South Korea). Whole genome sequencing (WGS) of 'Canino' and 'Katy' apricot genomes was conducted on an Illumina HiSeq2000 platform, using 100-bp paired-end reads, commercially conducted by Macrogen Inc. 'Goldrich' WGS kindly provided by Chris Dardick was also generated on an Illumina HiSeq2000 platform using 100-bp paired-end reads, at genomic facilities at DHMRI (David H. Murdock Research Institution, Kannapolis, NC, USA; http://www.dhmri.org). RNA sequencing (RNAseq) data were obtained using RNA isolated from mature anthers, mature styles and leaves from cvs. 'Canino', 'Katy' and 'Goldrich' (except for 'Katy' styles that were not collected). Two biological and two technical replicates per biological replicate were generated for each tissue and cultivar with the exception of anthers samples, where three biological replicates were obtained (Table S3.1). RNAs were sequenced by Illumina paired-end (100 bp). Sequencing was conducted by UCLA Neurosciences genomic Core (University of California, CA, USA).

454 raw data from BAC clones were filtered by CLC Genomics Workbench 8.0.1 (http://www.clcbio.com), trimming those sequences with a 'Quality limit' of 0.05 and 'Ambiguous limit' of 3, maintaining sequences with a minimum length of 50 bp and a maximum of 500. Additionally, trimmed sequences were aligned against pBeloBAC11 (cloning vector used to develop 'Goldrich' BAC library) in order to remove those sequences coming from the cloning vector ('Mismatch cost = 2', 'Insertion cost' = 3, 'Deletion cost' = 3, 'Length fraction' = 0.5, 'Similarity fraction' = 0.8). WGS *Illumina* raw data was filtered by CLC Genomics Workbench 8.0.1. Sequences with a 'Quality limit' of 0.05 and 'Ambiguous limit' of 2, and sequences with a lower length of 20 bp were trimmed. *Illumina* RNAseq raw data were processed using FastQC v.0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) software to assess the quality of raw and clean read sets. Reads were quality trimmed using FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit) with a minimum quality score of 25 and a minimum length of 40. Adaptor sequences were trimmed using the 'trim_blast_short' script available as part of seq_crumbs (http://bioinf.comav.upv.es/seq_crumbs/).

Apricot *M*-locus supercontig. BAC clones *de novo* assembly

Cleaned 454 sequences were used for *de novo* assembly of each BAC clone by CLC Genomics Workbench 8.0.1 using following parameters: k-mer = 45; automatic bubble size; minimum contig length = 100; 'Map reads back to contigs (slow)'; mismatch cost = 2; insertion cost = 3; deletion cost = 3; length fraction = 0.75 and similarity fraction = 0.9. Additionally, Macrogen Inc. provided a *de novo* assembly per BAC clone performed by GS De Novo Assembler v.2.8 (Roche) using default values for set parameters. Consensus contig sequences from both *de novo* assemblies were determined by *Staden software package* (http://www.sanger.ac.uk/Software/production/staden/). Resulting contigs were correctly oriented and ordered through *'microbial genome finishing module'* included in CLC Genomics Workbench 8.0.1 and using both *P. persica* (peach v1.0 International Peach Genome Initiative 2010 http://www.rosaceae.org/peach/genome) and *P. mume* (NCBI BioProject under accession PRJNA171605 and http://prunusmumegenome.bjfu.edu.cn) genomes as references. Contigs from overlapping BAC clones belonging to the same 'Goldrich' *M*-haplotype were joined by *GAP4* included in *Staden software package.* Those contigs from overlapping BAC

clones that could not be joined by *GAP4* were considered to belong to different *M*-haplotypes. To confirm this hypothesis SSR genotyping was performed for each BAC clone. DNA from BAC clones 215E14, 224A3, 234O11, 148M17, 253J12, 161F24 was PCR-amplified for SSRs PGS3_71, PGS3_47, PGS3_23, PGS3_62, PGS3_63 and PGS3_96, respectively; AGS_6 was PCR-amplified in BACs 209M03 and 108J24, and 160J21-2 in BACs 251L05, 159P08, 95D02 and 160J21. 'Goldrich' gDNA was used as control. PCR conditions were the same described for SSR amplification in section 'SSR identification and analysis' (see below) and *M*-haplotype BAC clone was coined according to the nomenclature reported in Chapter II. To join overlapping contigs from different haplotypes, two stepwise conditions were considered: 1) if a putative gene based on *P. persica* and *P. mume* annotation might be present in overlapping *P. armeniaca* contigs, that contig with highest homology rate to both genes inferred by ClustalW (Thompson et al., 1994) would be chosen. 2) If there were no putative genes in overlapping contigs, that contig belonging to the *M*-haplotype more numerous would be chosen. For unjoined contiguous contigs, GAP closure was carried out as follows: 1) a BLASTN (Altschul et al., 1990) analysis (cutoff e-value < 10^{-3}) using primer pair sequences for SSRs from PGS3 series was performed against contig sequences in order to identify those GAPs that matched with PGS3 microsatellites. If each primer of a pair confirmed to blast to ending sequence of contiguous contigs, then both contigs were joined. 2) For remaining unsolved GAPs, specific primers (Table S3.3) flanking ending contiguous contigs were designed by Primer3 v.0.4.0 (Untergasser et al., 2012) and PCR-amplified. PCR conditions were: initial denaturing step of 95ºC for 2 min; 35 cycles of 95ºC for 30 s, 52ºC for 30 s and 72ºC for 1 min; and a final extension of 72ºC for 10 min, using as template corresponding BACs DNA. Four independent replicates were amplified and PCR products were checked by gel electrophoresis. The four replicates were mixed together and purified by DNA Clean&Concentrator-5 Kit (Zymo Research, Irvine, CA). Purified PCR products were sequenced by Sanger, commercially conducted by Sistemas Genomicos S.L. (Paterna, Valencia, Spain). Resulting sequences were assembled through *Staden software package.*

WGS Illumina data alignment

Cleaned WGS *Illumina* reads from cvs. 'Canino', 'Katy' and 'Goldrich' were aligned separately (CLC Genomics Workbench 8.0.1 software) against a 'hybrid' reference sequence of *P. persica* genome sequence (peach v1.0 IPGI) where *M*-locus

region (scaffold_3 from 18.380.006 to 18.833.026 positions) was replaced by *P. armeniaca M*-locus supercontig sequence using following set of parameters: 'Mistmatch cost = 2', 'Cost of insertions and deletions = Affine gap cost', 'Insertion cost = 3', 'Deletion cost = 3', 'Insertion open cost = 6', 'Insertion extent cost = 1', Deletion open cost = 6', 'Deletion extent cost = 1', 'Length fraction = 0.6', 'Similarity fraction = 0.8', 'Global alignment = No' and 'Non-specific match handling = Map randomly'. Apricot *M*-locus supercontig alignments (between positions 18.380.006 and 18.833.026 of scaffold_3) were extracted and 'Local realignment' module was applied in order to improve alignment results ('Realign unaligned ends = Yes', 'Multi-pass realignment = 3').

SSR identification and analysis

New SSR markers were identified in *P. armeniaca M*-locus supercontig by Repeat Masker software (Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*. 2008-2015 http://www.repeatmasker.org). Primer pairs flanking microsatellite repeat motifs were designed using Primer3 (Table S3.4). SSR amplifications were performed in a GeneAmp_PCR System 9700 thermal cycler (Perkin–Elmer, Freemont, CA, USA) in a final volume of 20 µl, containing 75 mM Tris– HCl, pH 8.8; 20 mM $(NH_4)2SO_4$; 1.5 mM $MgCl_2$; 0.1 mM of each dNTP; 20 ng of genomic DNA and 1 U of DreamTaq polymerase (Thermo Scientific, Waltham, MA). Each polymerase chain reaction was performed by the procedure of Schuelke, (2000) using three primers: the specific forward primer of each microsatellite with M13(-21) tail at its 50 end at 0.4 µM, the sequence-specific reverse primer at 0.8 µM, and the universal fluorescent-labeled M13(-21) primer at 0.4 µM. The following temperature profile was used: 94 ºC for 2 min, then 35 cycles of 94 ºC for 45 s, 50-60 ºC for 1 min, and 72 ºC for 1 min and 15 s, finishing with 72 ºC for 5 min. Allele lengths were determined using an ABI Prism 3130 Genetic Analyzer with the aid of GeneMapper software, version 4.0 (Applied Biosystems).

SNP identification and analysis

SNPs and small InDels were called with CLC Genomics Workbench 8.0.1 using 'Basic Variant Detection' algorithm through the following parameters: 'Ploidy level = 2', 'Ignore positions with coverage above = 2000', 'Minimum coverage = 4', 'Minimum count = 1' and 'Minimum frequency = 25%' (Table S3.5). Primer pairs flanking SNPs were designed using Primer3 for selected variants (Table S3.6).

Amplifications were performed in a GeneAmp_PCR System 9700 thermal cycler (Perkin–Elmer, Freemont, CA, USA) in a final volume of 20 µl, containing 75 mM Tris– HCl, pH 8.8; 20 mM $(NH_4)2SO_4$; 1.5 mM $MgCl_2$; 0.1 mM of each dNTP; 20 ng of genomic DNA and 1 U of DreamTaq polymerase (Thermo Scientific, Waltham, MA) using corresponding gDNA as template. Cycling conditions were as follows: an initial denaturing step of 95ºC for 2 min; 35 cycles of 95ºC for 30 s, 52ºC for 30 s and 72ºC for 1 min; and a final extension of 72ºC for 10 min. Four independent replicates were amplified and PCR products were checked by gel electrophoresis. An admixture of four replicates was purified by DNA Clean&Concentrator-5 Kit (Zymo Research, Irvine, CA). Purified PCR products were sequenced by Sanger, commercially conducted by Sistemas Genomicos S.L. (Paterna, Valencia, Spain). Sequence chromatogram visualization for SNP confirmation was performed by *BioEdit software* (Hall, 1999).

Gene and transcript annotation of apricot *M*-locus physical map

'Goldrich' RNAseq data from mature styles, mature anthers and leaves transcriptomes were aligned to the 'hybrid' peach genome sequence containing the apricot *M*-locus syntenic region using 'Large gap mapper' tool included in 'Transcript discovery Plug-in 2.0' of CLC Genomics Workbench 8.0.1 with these set of parameters: 'Maximum number of hits for a segment = 10', 'Maximum distance from seed = 20000', 'Multi match mode = random', Mismatch cost = 2', 'Insertion cost = 3', 'Deletion cost = 3', 'Similarity = 0.9', 'Length fraction = 0.9' and 'Overside default distances = Yes'. Then, transcript discovery tool was used to produce mRNA and gene annotations (parameters: 'Strand specific = No', 'Extend existing annotations = Yes', 'Splice sites = All', 'Exclude uncertain splice sites = Yes', 'Ignore duplicate reads = Yes', 'Ignore non-specific matches = Yes', 'Minimum unique observations (un-spliced) = 1, 'Minimum coverage ratio (un-spliced) = 0.03', 'Minimum unique observations (spliced) = 1, 'Minimum coverage ratio (spliced) = 0.03', 'Exclude internal un-spliced events = Yes', 'Exclude external un-spliced events = Yes', 'Maximum distance between events = 1000', 'Minimum observations in gene = 3', 'Minimum length of gene = 150', 'Genes with spliced transcripts only = No', 'Maximum joining distance = 250', 'Minimum length = 100'). Resulting annotations were manually curated by comparing with predicted annotations for *P. persica* (v1.0 and v2.0 of IPGI) and *P. mume* by gene, CDS and predicted protein alignment study through Clustal Omega aligner (McWilliam et al., 2013).

RNAseq differential expression analysis

Cleaned reads from mature anthers, styles and leaves of 'Goldrich', 'Canino' and 'Katy' were aligned against 'hybrid' peach genome sequence containing the apricot *M*-locus syntenic region through the 'align_and_estimate_abundance.pl' script available using the Bowtie aligner as part of Trinity software (Langmead et al., 2009). From these alignments, transcript quantification was performed with RSEM (Li and Dewey 2011). Estimated fraction of fragments that are derived from a gene were further used for differential expression analysis. Raw counts generated from RSEM were imported into the edgeR package (Robinson et al., 2010) from Bioconductor v.2.11 (Gentleman et al., 2004) in order to determinate the significance level of gene-level expression. To filter out the likely transcript artifacts and lowly expressed transcripts, only were maintained those with a value of counts per million (cpm) > 1 at least in all replicates of one condition. A between-sample normalization was made taking into account the total number of reads by library. To observe the relations between samples and replicates, both technical and biological, a multidimensional scaling (MDS) plot was made. False discovery rate (FDR) <= 0.05 was used to determine the threshold of the P-value in multiple tests. Heat-map was performed using a hand-made script by R.

Polymorphism screening

Variants from positions 142.155 to 276.184 (a*M*-supercontig) called in 'SNP identification and analysis' section were used for this analysis (Table S3.5). Additionally, 'InDels and Structural Variants detection' algorithm through CLC Genomics Workbench 8.0.1 software was also used in the a*M*-supercontig realignment of the three cultivars (from 'WGS Illumina data alignment' section), parameters settings: 'P-value threshold = 0.0001', 'Maximum number of mistmatches = 3' and 'Minimum number of reads = 2'. Overall variants (variants called from 'Basic variant detection' and 'InDels and Structural Variants detection') between positions 142.155-276.184 from self-compatible cultivars ('Canino' and 'Katy') were compared against self-incompatible cultivar 'Goldrich' using 'Compare sample variants' ('Keep variants that are different' option was choice) and 'Amino acid changes' tools included in CLC Genomics Workbench 8.0.1 in order to identify those polymorphisms present in self-compatible cultivars and absent in the self-incompatible cultivar that led to non-synonymous changes in the predicted proteins (Table S3.7).

PaMDOr gene amplification and sequencing from gDNA

Overlapping fragments comprising *PaMDOr* gene (considering that has a reverse orientation in the apricot *M*-locus supercontig sequence annotation) were PCR amplified with specific primer pairs PaMDsb_F1 (5'-GTTCTCTTGCCGGATATCTAATATGT-3', -1741 bp from start-codon)/PaMDsb_R1 (5'-ACGGTTGGGTTGACATTAAAAC-3', +169 bp from start-codon) and PaMDsb_F2 (5'- TTTGGCCTGTTTTGGAACC-3', -1179 bp from start-codon)/PaMDsb_R2 (5'- ATACAAAGATGGGCGCTGA-3', -199 bp from start-codon) using 'CC-67' (*mm*) and 'K06-17' (*mm*) gDNA as template for *m*-allele sequencing and 'CC-77' (*MM*) gDNA as template for *M*-allele sequencing. PCR amplifications were performed in a final volume of 20 µl containing 75 mM Tris– HCl, pH 8.8; 20 mM $(NH_4)_2SO_4$; 1.5 mM $MgCl_2$; 0.1 mM of each dNTP; 20 ng of genomic DNA and 1 U of DreamTaq polymerase (Thermo Scientific, Waltham, MA). Cycling conditions were as follows: an initial denaturing step of 94ºC for 2 min; 30 cycles of 94ºC for 30 s, 55ºC for 60 s and 72ºC for 1 min 30 s; and a final extension of 72ºC for 10 min (GeneAmp®PCR System 9700, Perkin-Elmer, Fremont, CA). Four independent replicates were amplified and PCR products were checked by electrophoresis and finally mixed together to purify by DNA Clean&Concentrator-5 Kit (Zymo Research, Irvine, CA). Purified PCR products were sequenced by Sanger by the Bioinformatics Service at the IBMCP (http://www.ibmcp.upv.es) and resulting sequences were assembled through *Staden software package*.

PaMDOr m-FaSt insertion genotyping

Specific primer pair PaMDsb_F2/PaMDsb_R2 was used for PaMDsb *m-FaSt* insertion genotyping. PCR conditions were the same used for *PaMDOr* gene amplification from gDNA in previous section (see above). PCR products were electrophoresed in 1% (w/v) agarose gel.

BLASTP Reciprocal Best Hit (RBH) analysis

BLASTP analysis of the PaMDOr predicted protein was performed against 'non-redundant protein sequences' database of NCBI (Altschul et al., 1990). Some accessions related to SI from this 'BLASTP direct' result were selected for 'BLASTP reciprocal' analysis against *Prunus* (taxid: 3754) database of NCBI. These proteins were XP 008230295.1 and XP 008230297.1 (corresponding to Pm015399 and Pm015400

respectively in *Prunus mume* protein database), XP_007215948.1 and XP 007216055.1 (corresponding to ppa011285m and ppa017665m respectively in *Prunus persica*, v1.0 IPGI, protein database), XP_008341809.1 and XP 008379454.1 (corresponding to MDP0000148485 and MDP0000233548 respectively in *Malus domestica*, v1.0, protein database (Velasco et al., 2010)), XP 004304201.1 and XP 004306054.1 (corresponding to gene04226-v1.0-hybrid and gene04224-v1.0-hybrid respectively in *Fragaria vesca*, v1.0, protein database (Shulaev et al., 2011)), XP_004232135.1 (corresponding to Solyc02g089230.2.1 in *Solanum lycopersicum*, ITAG 2.3 protein database), XP_009774377.1 (*Nicotiana Sylvestris* accession), XP_009608607.1 (*Nicotiana tomentosiformis* accession) and NP_198706.1 (corresponding to AT5G38900.1 in *Arabidospsis thaliana*, TAIR10, protein database). All BLASTP analysis were carried out with an E value cut-off < $1e^{-3}$.

Apricot *M*-locus syntenic block analysis

Assembled genome sequence and predicted protein collections of *Prunus persica* (Verde et al., 2013), *Malus domestica* (Velasco et al., 2010), *Solanum lycopersicum* (Consortium et al., 2012) and *Arabidopsis thaliana* (Arabidopsis genome initiative, 2000) were used to localize *M*-locus syntenic blocks in the four genomes. Data were retrieved from the GDR database (www.rosaceae.org; Prunus_persica_v1.0_scaffolds.fa, Prunus_persica_v1.0_scaffolds.gff3, Prunus_persica_v1.0_peptide.fa, Malus_x_domestica.v1.0.contigs.fa, Malus_x_domestica.v1.0.consensus.gff and Malus_x_domestica.v1.0.consensus_peptide.fa), the SolGenomics Network (www.solgenomics.net; S_lycopersicum_chromosomes.2.30.fa, ITAG2.3_gene_models.gff3 and ITAG2.3_proteins.fasta) and from TAIR database (TAIR10_chr_all.fas, TAIR10_GFF3_genes.gff and TAIR10_pep_20101214.txt). Proteins from ppa001611m (position 18.391.171) to ppa001157m (position 18.769.249) of scaffold_3 in *Prunus persica*, encompassing 62 genes, were used as queries for RBH analysis against the other 3 predicted protein databases. Thus, best hit after 'Direct BLASTP' analysis was used as query against *Prunus persica* protein database; whether best hit from 'Reciprocal BLASTP' analysis matched with initial *Prunus persica* protein used as query, this positive RBH was used as anchor in syntenic block identification. RBH analysis against *Malus domestica*, *Solanum lycopersicum* and *Arabidopsis thaliana* databases was carried out through custom-made python scripts

using executable gffutils (www.pythonhosted.org/gffutils) and blastall (Altschul et al., 1990) packages. RBH results supporting syntenic blocks identification was visualized by *Circos software* (Krzywinski et al., 2009).

Phylogenetic tree-based analysis

Phylogenetic analysis was conducted by *MEGA6* (Tamura et al., 2013). Amino acid sequences of accessions selected in section 'BLASTP Reciprocal Best Hit (RBH) analysis' and the protein accessions XP_09373771.1 and XP_009378259.1 (*Pyrus* accessions), XP_006338333.1 and XP_015064048.1 (corresponding to Sotub02g033650.1.1 and Sopen02g034020.1 respectively in *Solanum tuberosum* (protein database) and *Solanum pennelli*, PGSC DM v3.4 and *S.pennelli* protein databases of), XP_013674618.1 and XP_013651836.1 (*Brassica napus* accessions), XP_009125096.1 (*Brassica rapa* accession), XP_013613719.1 (*Brassica oleracea* accession) and XP_002868748.1 (*Arabidopsis lyrata* accession) were aligned by ClustalW (Thompson et al., 1994). Phylogenetic relationship tree was constructed by the Maximum Likelihood method (Felsenstein, 1981) and phylogenetic test was done based on 1,000 bootstrap replicates. LG+G was employed as the best fits for protein alignment.

Acknowledgements

This work was supported by a grant from the Ministerio de Economia y Competitividad del Gobierno de España (AGL19018-2010). The authors want to thank Inmaculada López for its technical assistance and Javier Forment for bioinformatic assistance.

Chapter 4:
Comparative study of the GSI system in Rosaceae and Solanaceae by analyzing orthology relationships for modifier factors

Abstract

Self-Incompatibility (SI) is the reproductive barrier widest spread in plant kingdom, which has evolved differently at molecular level to prevent inbreeding and assist genetic exchange. SI is genetically controlled by the *S*-locus, where at least two linked multiallelic genes are expressed in the pollen and pistil sides determining the specific pollen-pistil recognition. In Rosaceae and Solanaceae, S-RNases (pistil) and S-locus F-Box (pollen) proteins are these *S*-determinants. In spite of sharing *S*-factors, recent studies have spotlighted a different origin for *Prunus* F-Box proteins, scenario that matches with a different recognition behavior observed in this genus. Nevertheless, factors unlinked to the *S*-locus named modifiers are also fully required to reject self-pollen, but they have received less attention than *S*-determinants. Most of described modifiers have only been identified in Solanaceae and remain elusive in Rosaceae. This work was particularly aimed to identify in *Prunus* putative orthologues for Solanaceae and *Maloideae* modifiers in order to deepen into the evolutionary history of the GSI system in both families. Six genes *120K*, *NaTrxh*, *NaStEP*, *SBP1*, *NaPCCP* (Solanaceae) and *MdABCF* (*Maloideae*) have been assessed using an *in silico* method to predict orthology based on reciprocal best hits, genomic synteny and clustering analyses. Putative *Prunus* orthologues were found for *NaTrxh*, *MdABCF* and *SBP1* suggesting a divergent evolution from a common ancestor before splitting of Rosaceae and Solanaceae. Regarding *120K*, tandem duplication and subsequent functional specialization in pollen-pistil interaction might have occurred in *asterids* after *eudicots* division. Meanwhile, putative gene losses, duplications and/or chromosomal rearrangements draw a convoluted evolutionary history for *NaStEP* and *NaPCCP*.

Introduction

Solanaceae and Rosaceae families apparently share equivalent factors involved in the recognition and rejection of pollen genetically related to avoid inbreeding in the so-called Self-Incompatibility (SI) mechanism. These factors are enclosed into the *S*-locus that comprises two or more linked genes (de Nettancourt, 2001). In the gametophytic self-incompatibility (GSI) mechanism, one of these genes encodes a ribonuclease specifically expressed in the style (S-RNases) (Anderson et al., 1986; McClure et al., 1989), while the other/s codes an F-Box containing-domain protein

(termed SLF, SFB and SFBB in Solanaceae, *Prunoideae* and *Maloideae*, respectively) (Entani et al., 2003; Lai et al., 2002; Ushijima et al., 2004). Both *S*-genes are multiallelic, thus, if the variants for each tissue-specific gene displayed during pollination derive from the same haplotype, the pollen tube growth is arrested (McCubbin & Kao, 2000).

Phylogenetic studies with *S*-RNases have pointed out a single evolutionary origin about 120 MYA for eudicots (Igic & Kohn, 2001; Steinbachs & Holsinger, 2002). Notwithstanding, extended recent studies have led to propose that *S*-pollen factors in *Prunus* and *Malus* might have evolved from different paralogues derived from a common Rosaceae ancestor (Aguiar et al., 2015; Akagi et al., 2016; Morimoto et al., 2015). This proposal might explain differences regarding effects of pollen-part mutations affecting *S*-locus F-box proteins in *Prunus* and *Malus* (Tao & Iezzoni, 2010).
S-locus unlinked genes participating in the GSI system based on *S*-RNases are also required for the mechanism to function. Some of these genes, known as modifier genes (McClure et al., 2011), have been identified. HT-B, a small asparagine-rich protein (McClure et al., 1999); NaStEP, a Kunitz-type proteinase inhibitor (Busot et al., 2008; Jimenez-Duran et al., 2013); 120K, an arabinogalactan protein (AGP) abundant in the stylar transmitting tract (Cruz-García et al., 2003 and 2005), and NaTrxh, a thioredoxin (TRX) protein from *hII* group (Avila-Castañeda et al., 2014; Juárez-Díaz et al., 2006) are stylar modifiers found in *Nicotiana* species. Pollen side non-*S*-factors have also been identified, mostly related to the SCF-like E3 ubiquitin ligase complex formed by Skp1/Cul1/F-box proteins where additionally Cul1 interacts with Rbx1 (Zhang et al., 2009). All SCF components have been cloned, firstly in Petunia (Hua & Kao, 2006; Huang et al., 2006; Li et al., 2014) and later in other species including *Malus*, *Pyrus* and *Prunus* (Matsumoto et al., 2012; Minamikawa et al., 2014; Xu et al., 2013; Yuan et al., 2014). SBP1, a RING-finger E3 ligase (Sims & Ordanic, 2001); NaPCCP from *Nicotiana alata* (Lee et al., 2008; 2009) and MdABCF, a transmembrane transporter identified in *Malus domestica* (Meng et al., 2014) are other pollen modifier factors found to be likely involved in the GSI system.

SCF components from *Prunus* and *Malus/Pyrus* have shown to be orthologous to Solanaceae components by phylogenetic relationship studies. Therefore, more genes descending from a common ancestor for remaining modifiers might be expected to be found within Rosaceae. In addition, due to recent advances in high-throughput sequencing technologies and bioinformatics, genome sequences of many plant species

have been released publicly including *Solanum lycopersicum* (Consortium, 2012), *Nicotiana benthamiana* (Bombarely et al., 2012), *Malus domestica* (Velasco et al., 2010) and *Prunus persica* (Verde et al., 2013). This constitutes a powerful source of information for comparative genomics and gene orthology analysis (Gabaldon & Koonin, 2013; Gabaldón, 2008; Sonah et al., 2011).

In order to shed some light on the evolutionary history of the Rosaceae and Solanaceae GSI system, we performed a screening to identify putative orthologues in *Prunus* for modifier factors already reported in Solanaceae and Maloideae.

Results

A three pillar strategy was designed to identify putative orthologues in *Prunus* for *120K*, *NaTrxh*, *NaStEP*, *MdABCF*, *SBP1* and *NaPCCP* modifiers through Rosaceae and Solanaceae (using Brassicaceae as outgroup) comparison (Figure 4.1). Orthology was evaluated at three different levels corresponding to each requirement, thus two genes might be considered as orthologues if they fulfilled the three conditions. The starting point before orthology study was to perform a BLASTP analysis ('Direct BLASTP') using the target modifiers as query against Solanaceae (*Solanum lycopersicum* and *Nicotiana benthamiana*), Rosaceae (*Prunus persica* and *Malus domestica*) and Brassicaceae (*Arabidopsis thaliana*) protein databases (Table 4.3, see Material and methods). Best scoring matches from direct BLASTP outcome were used to manage the three different orthology analyses. First one relies on Reciprocal Best BLASTP Hit (RBH) search comparing all-to-all protein databases (Zheng et al., 2005). The second is an orthologue mapping approach to identify syntenic blocks (Zheng et al., 2005) and the third one is aimed to infer phylogenetic relationships on the basis of clustering (tree-based) methods using selected accessions (Yang et al., 2012).

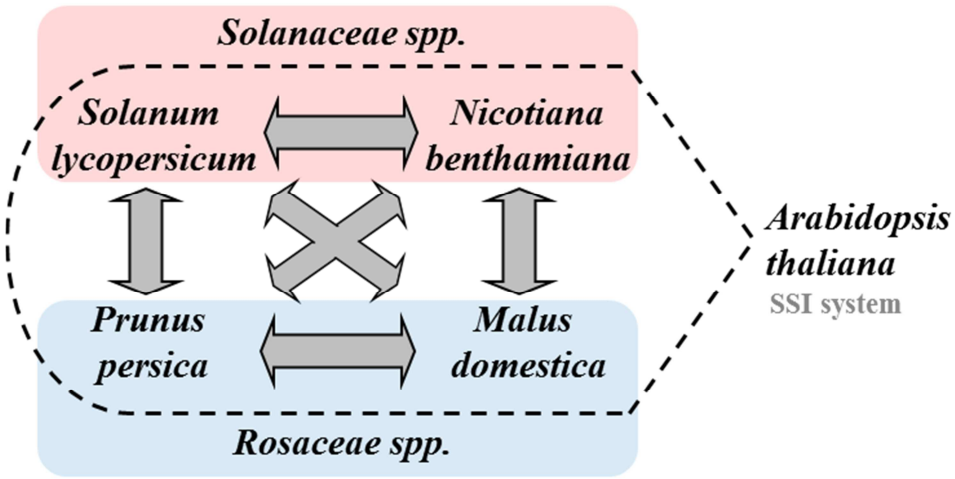


Figure 4.1. Schematic representation of genome and protein databases used to study orthology relationships between modifiers from Rosaceae and Solanaceae families.

NaTrxh

Direct BLASTP and RBH analyses

Best hits obtained by direct BLASTP using NaTrxh (AAY42864) as query in each protein database resulted RBHs in almost all comparisons. MDP0000752795 blast against *S.lycopersicum* protein database was the unique exception, where Solyc05g006830.2.1 was the best hit in spite of having the same e-value than second hit (Solyc02g087630.2.1, RBH in the other comparisons). Therefore, ppa011576m, MDP0000752795 (MDP0000448333), Solyc02g087630.2.1, NbS00020764g0013.1 and AT5G39950.1 were the RBHs of NaTrxh protein among *P. persica*, *M. domestica*, *S. lycopersicum*, *N. benthamiana* and *A. thaliana*, respectively (Figure 4.2a; Tables 4.1a and S4.1).

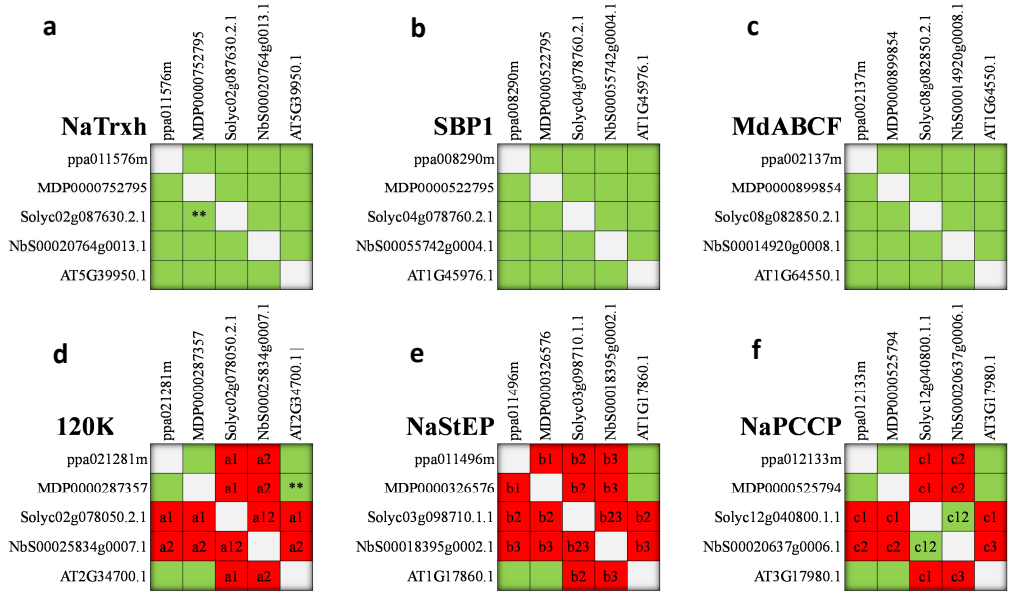


Figure 4.2. Schematic representation of RBH results for NaTrxh (a), SBP1 (b), MdABCF (c), 120K (d), NaStEP (e) and NaPCCP (f). First hits from direct BLASTP analysis (Table 4.1) are shown in horizontal and vertical orientation. Squares show the results of pairwise comparison of first hits, where *Green squares* indicate that both proteins confirm to be RBHs, while *red squares* do not. Protein accessions from *red square* comparisons confirmed to have a different protein accession as RBH. These proteins and corresponding RBH are indicated below by an alphanumeric code. See supplemental data of chapter 4 for more detailed information.
[**]First and second hits have similar rates of homology (see Tables S4.1 and S4.4).
[a1] RBH of ppa021281m, MDP0000287357, AT2G34700.1 and NbS00008703g0009.1 in *S. lycopersicum* genome is Solyc02g078040.2.1.
[a2] RBH of ppa021281m, MDP0000287357, AT2G34700.1 and Solyc02g078040.2.1 in *N. benthamiana* genome is NbS00008703g0009.1.
[b1] RBH of MDP0000326576 in *P. persica* genome is ppa011448m.
[b2] RBH of ppa011496m, MDP0000326576, AT1G17860.1 and NbS00009480g00031.1 in *S. lycopersicum* is Solyc03g020010.1.1.
[b3] RBH of ppa011496m, MDP0000326576, AT1G17860.1 and Solyc03g020010.1.1 in *N. benthamiana* is NbS00009480g00031.1.
[c1] RBH of ppa012133m, MDP0000525794, AT3G17980.1 and NbS0009334g0006.1 in *S. lycopersicum* is Solyc03g118720.2.1.
[c2] RBH of ppa012133m, MDP0000525794, AT3G17980.1 and Solyc03g118720.2.1 in *N. benthamiana* is NbS0009334g0006.1.
[c3] RBH of AT3G17980.1 in *N. benthamiana* genome is NbS00020637g0006.1.

Identification of syntenic blocks

Genome segments having the RBH identified for *P. persica*, *S. lycopersicum* and *A. thaliana* at chrs.3, 2 and 5, respectively, showed distinct degrees of synteny. The major number of anchors were obtained between *P. persica* and *S. lycopersicum* encompassing a total of 23 in regions 19,75-20,09 Mbs and 44,43-44,74 Mbs, respectively. Meanwhile, *A. thaliana* and *P. persica* shared 17 anchor points (from 15,93 to 16,1 Mb in *Arabidopsis* genome), but *P. persica* syntenic region spans a larger distance (19,70-20,29 Mbs) regarding *S. lycopersicum* syntenic block. In addition, bordering ppa011576m region had certain synteny with a region of *A. thaliana* at chr.3 (10,66-10,79 Mbs). Between *A. thaliana* and *S. lycopersicum* only 4 anchors were positively used to connect both regions, mainly due to neighbouring Solyc02g087630.2.1 proteins are homologous to *A. thaliana* proteins of 4,87-4,91 Mbs region at chr.5 (Figure 4.3).

Table 4.1. BLASTP direct results for NaTrxh (a), SBP1 (b), MdABCF (c), 120K (d), NaStEP (e) and NaPCCP (f) protein accessions (see Figure 4.1) against *Prunus persica* (blue colors), *Malus domestica* (orange colors), *Solanum lycopersicum* (red colors), *Nicotiana benthamiana* (yellow colors) and *Arabidopsis thaliana* (green colors) protein databases. Three first hits are indicated (from *darker* to *lighter colors*) matching with results shown in Tables S4.1 to S4.6.

a)

Hit	P. persica				M. domestica				S. lycopersicum				N. benthamiana				A. thaliana			
	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val
1st	ppa011576m	135	67	4E-64	MDP0000752795	106	69	7E-43	Solyc02g087630.2.1	104	88	2E-51	NbS00020764g0013.1	116	100	2E-65	AT5G39950.1	109	69	1E-43
2nd	ppa013299m	130	55	6E-52	MDP0000448333	108	68	5E-40	Solyc05g006830.2.1	108	72	2E-44	NbS00027633g0013.1	116	93	2E-59	AT1G19730.1	108	47	4E-28
3rd	ppa013161m	110	49	2E-37	MDP0000391509	110	59	1E-37	Solyc05g006860.2.1	108	67	3E-42	NbS00010261g0005.1	123	60	1E-40	AT1G45145.1	107	50	9E-28

b)

Hit	P. persica				M. domestica				S. lycopersicum				N. benthamiana				A. thaliana			
	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val
1st	ppa008290m	321	70.09	2E-167	MDP0000522795	319	69.59	3E-159	Solyc04g078760.2.1	316	91.14	0.0	NbS00055742g0004.1	331	90.63	0.0	AT1G45976.1	327	63.30	2E-141
2nd	ppa008184m	251	39.44	2E-51	MDP0000717791	323	64.71	8E-138	Solyc05g005210.2.1	234	38.03	6E-46	NbS00016021g0006.1	331	88.22	0.0	AT1G60610.3	243	38.68	2E-45
3rd	ppa007884m	237	30.80	2E-25	MDP0000650075	316	64.56	3E-133	Solyc03g112860.2.1	214	35.51	4E-29	NbS00020248g0004.1	234	40.17	3E-48	AT1G60610.2	243	38.68	2E-45

c)

Hit	P. persica				M. domestica				S. lycopersicum				N. benthamiana				A. thaliana			
	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val
1st	ppa002137m	711	93,95	0	MDP0000170302	711	99,16	0	Solyc08g082850.2.1	713	80,65	0	NbS00014920g0008.1	535	77,38	0	AT1G64550.1	713	81,77	0
2nd	ppa002097m	576	45,14	2E-154	MDP0000899854	711	97,47	0	Solyc07g008610.1.1	562	45,73	4E-156	NbS00001134g0008.1	311	77,17	3E-167	AT5G60790.1	571	42,91	2E-158
3rd	ppa003175m	545	42,94	4E-153	MDP0000477774	545	43,3	2E-153	Solyc11g069090.1.1	545	43,85	4E-156	NbS00011489g0001.1	565	45,49	5E-155	AT3G54540.1	567	45,15	3E-153

d)

Hit	P. persica				M. domestica				S. lycopersicum				N. benthamiana				A. thaliana			
	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val
1st	ppa021281m	129	42.64	4E-19	MDP0000287357	131	44.27	5E-21	Solyc02g078050.2.1	281	56.23	4E-68	NbS00009580g0020.1	197	71.07	5E-36	AT2G33790.1	197	35.53	2E-19
2nd					MDP0000165381	129	44.96	1E-19	Solyc02g078060.1.1	112	65.18	2E-40	NbS00025834g0007.1	134	55.22	5E-33	AT2G34700.1	140	40.00	2E-19
3rd					MDP0000423907	129	31.01	3E-08	Solyc02g078100.2.1	145	53.10	2E-35	NbS00007980g0003.1	167	46.71	3E-32	AT3G09925.1	149	26.17	4E-09

e)

Hit	P. persica				M. domestica				S. lycopersicum				N. benthamiana				A. thaliana			
	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val
1st	ppa011496m	190	42.11	8E-33	MDP0000326576	203	35.96	2E-22	Solyc03g098710.1.1	203	47.29	8E-45	NbS00018395g0002.1	246	68.29	1E-92	AT1G17860.1	237	33.33	2E-25
2nd	ppa011653m	159	40.25	2E-23	MDP0000619608	203	35.96	4E-22	Solyc06g072230.1.1	225	39.11	1E-34	NbC24872723g0001.1	220	60.45	2E-77	AT1G73260.1	192	35.94	2E-22
3rd	ppa011448m	207	35.75	3E-23	MDP0000635659	193	35.75	9E-22	Solyc03g019690.1.1	226	40.71	2E-34	NbS00018395g0011.1	243	58.44	3E-69	AT1G73325.1	229	32.31	3E-15

f)

Hit	P. persica				M. domestica				S. lycopersicum				N. benthamiana				A. thaliana			
	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val	Match	length	% id	E-val
1st	ppa012133m	187	75.94	4E-99	MDP0000525794	187	75.94	2E-100	Solyc12g040800.1.1	165	93.33	1E-111	NbS00020637g0006.1	188	93.09	3E-119	AT3G17980.1	172	76.74	1E-92
2nd	ppa012128m	187	75.94	4E-99	MDP0000776395	184	75.54	6E-98	Solyc06g068940.2.1	166	83.73	9E-98	NbS00009698g0010.1	183	83.61	2E-102	AT1G48590.1	164	77.44	2E-88
3rd	ppa012140m	187	75.94	4E-99	MDP0000259616	164	70.73	1E-82	Solyc03g118720.2.1	166	80.12	4E-97	NbS00020564g0001.1	166	84.94	1E-101	AT1G48590.2	164	77.44	2E-88

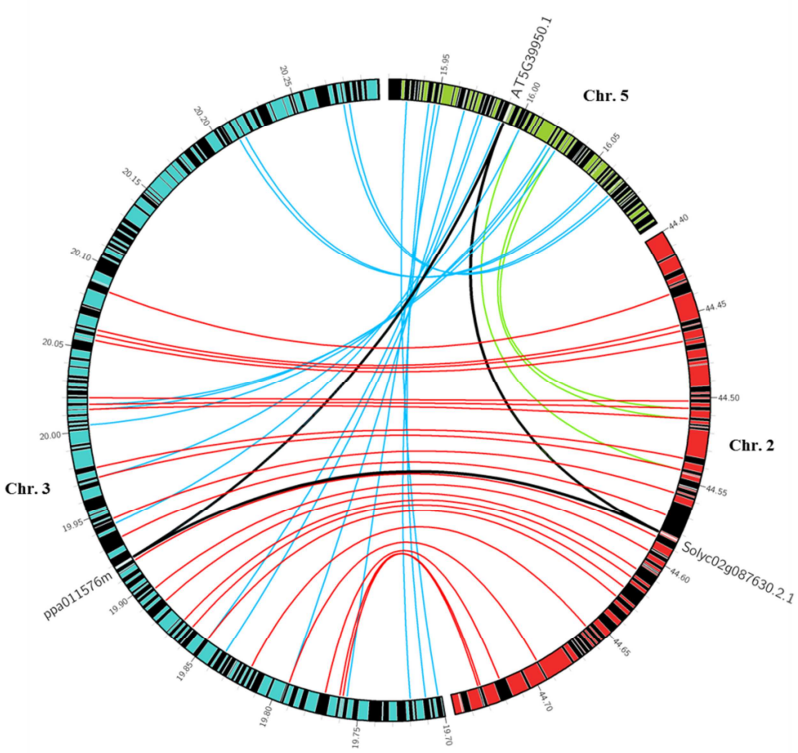


Figure 4.3. Syntenic comparative analysis of regions containing NaTrxh RBHs (anchors) with *A. thaliana* (green), *S. lycopersicum* (red) and *P. persica* (blue) genomes. *Black rectangles* within circular genome regions represent gene annotation in scale, *red lines* are anchors between *P. persica* and *S. lycopersicum*, *blue lines* are anchors between *A. thaliana* and *P. persica*, and *green lines* are anchors between *P. persica* and *A. thaliana*. RBHs accessions are shown for each species in the corresponding genome regions and are connected by *black lines*. *Red triangles* represent a change of scale.

Phylogenetic analysis based on clustering methods

TRX proteins are currently classified into 8 different groups (Meyer et al., 2012). NaTrxh belongs to the *h* group and the II subgroup (there are 3 subgroups in class *h*) (Juárez-Díaz et al., 2006). For the phylogenetic analysis of NaTrxh, accessions from TRX groups *f*, *m*, *x* and *o* as well as proteins from subgroups I, II and III of group *h* (previously used by Juárez-Díaz et al. (2006) for NaTrxh classification) were accordingly employed along with the first three hits obtained in direct BLASTP analysis for each species. LG+G was the model that best fit with NaTrxh data alignment. All TRX groups were successfully clustered including *h* subgroups, except for AAL54858.1 that grouped with Trxh subgroup III, when in fact belongs to subgroup I. All RBHs grouped with NaTrxh accession independently on the species they came from. Additionally, MDP0000448333 and NbS00027633g0013.1 (second hits) were also included in this group, occurrence expected for apple but not for *N. benthamiana*. The

rest of second and third best hits from BLASTP analysis also clustered in *h*II group (but in a separate branch) or in *h*I group (Figure 4.4).

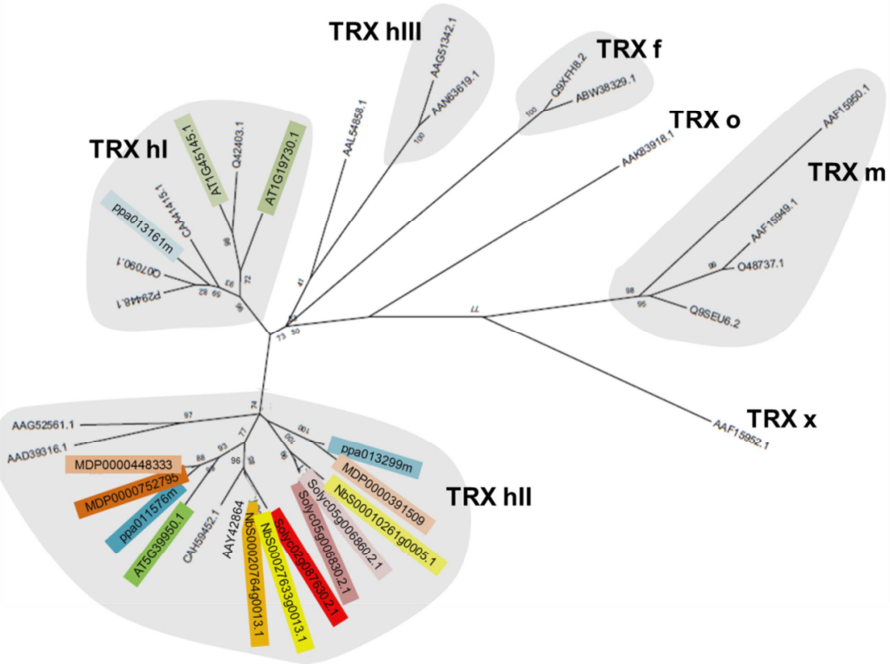


Figure 4.4. NaTrxh phylogenetic tree analysis. Maximum Likelihood phylogenetic tree for accessions from 'Direct BLASTP' output, RBH results and protein accessions from TRX hI (P29448.1 *A. thaliana*, Q07090.1 *N. tabacum*, CAA41415.1 *N. tabacum*, Q42403.1 *A. thaliana*), TRX hII (AAG52561.1 *A. thaliana*, CAH59452.1 *Plantago major*, AAY42864.1 NaTrxh), TRX hIII (AAL54858.1 *N. tabacum*, AAG51342.1 *A. thaliana*, AAN63619.1 *N. tabacum*), TRX f (Q9XFH8 *A. thaliana*, AAN63619.1 *N. tabacum*), TRX o (AAK83918.1 *A. thaliana*), TRX m (AAF15949.1 *A. thaliana*, AAF15950.1 A. thaliana, Q9SEU6.2 *A. thaliana* and O48737.1 *A. thaliana*) and TRX x (AFF15952.1 *A. thaliana*) groups taken by Juárez-Díaz et al. (2006). These groups are shown in *shading grey*. Bootstrap values are shown for every node. Accessions highlighted by *colors* refer to 'Direct BLASTP' results (see Table 4.1a).

SBP1

Direct BLASTP and RBH analyses

AAF28357 accession from *Petunia hybrida* was used as query in direct BLASTP analysis. The similarity rate between best hits and the rest of matches found was very high in all protein databases studied except for *M. domestica* (which e-value difference with second hit was not as high as observed for other species) and *N. benthamiana* (where first and second hits had similar e-values and percentage of similarity, but significantly separated of the third hit) (Table 4.1b). The most remarkable outcome was that first hits from direct BLASTP analyses resulted to be RBHs all-to-all (Figure 4.2b and Table S4.2).

Identification of syntenic blocks

Regions containing AT1G45976.1, ppa008290m and Solyc04g078760.2.1 RBHs proteins were analysed to anchor syntenic blocks. *P. persica* and *S. lycopersicum* showed to have a high synteny, 32 anchor sites led to define syntenic regions between chr1: 35,95-36,25 of *P. persica* and chr.4: 60,83-61,22 of *S. lycopersicum*. On the contrary, AT1G45976.1 region (chr.1: 17,08-17,24) presented less positive anchors with *P. persica* (13) and *S. lycopersicum* (15) encompassing larger regions (3 and 2,7 Mbs, respectively) than those found for both *P. persica* and *S. lycopersicum* (Figure 4.5).

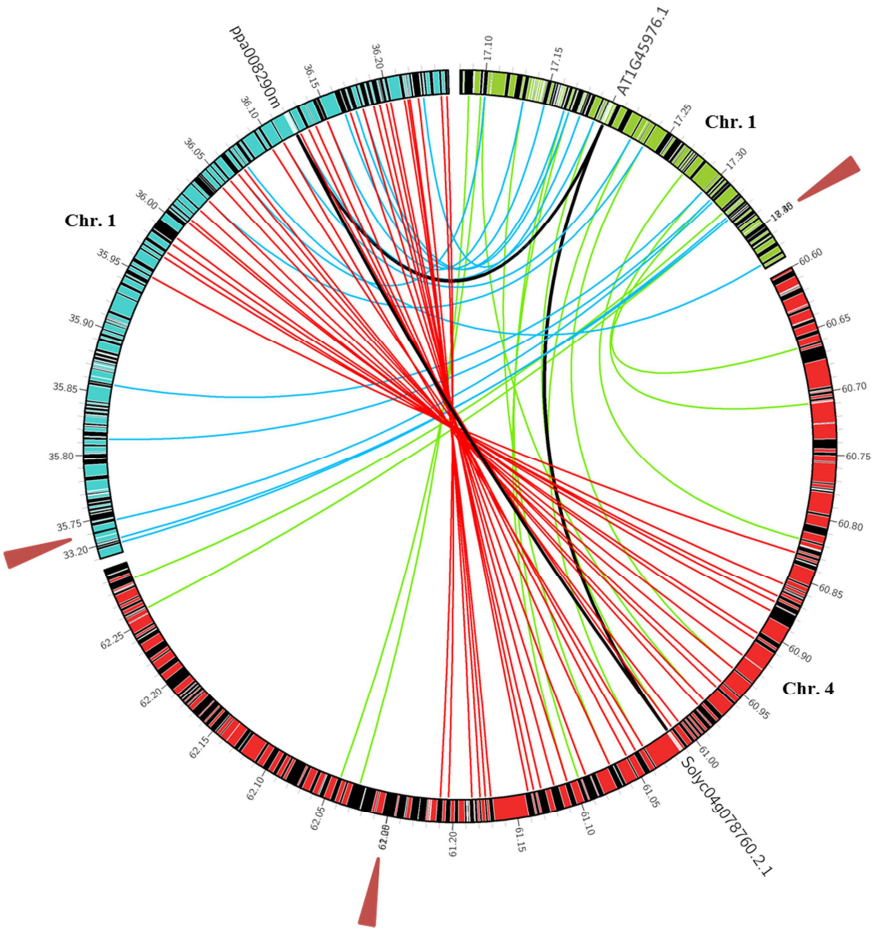


Figure 4.5. Syntenic comparative analysis of regions containing SBP1 RBHs (anchors) with *A. thaliana* (green), *S. lycopersicum* (red) and *P. persica* (blue) genomes. *Black rectangles* within circular genome regions represent gene annotation in scale, *red lines* are anchors between *P. persica* and *S. lycopersicum*, *blue lines* are anchors between *A. thaliana* and *P. persica*, and *green lines* are anchors between *P. persica* and *A. thaliana*. RBHs accessions are shown for each species in the corresponding genome regions and are connected by *black lines*. *Red triangles* represent a change of scale.

Phylogenetic analysis based on clustering methods

PhSBP1 is an E3 ubiquitin ligase having a RING domain where two types have been described, H2 (C3H2C3) and HC (C3HC4). PhSBP1 contains this last motif that has been, in turn, divided in 2 subtypes in *A. thaliana*, HCa (SBP1) and HCb, according to Stone et al. (2005). Furthermore, they also established a wide number of groups in order to classify RING proteins. *A. thaliana* best hit, AT1G45976.1, belongs to group 6. Keeping in mind this information, phylogeny analysis for SBP1 was performed including proteins with H2, HCa and HCb RING domains either from group 6 as well as from other different groups (excluding type HCb/group 6). The substitution model that best fit with SBP1 data alignment was JTT+G, which phylogenetic tree is shown in Figure 4.6. Three well differentiated clusters were obtained. RING-type E3 proteins having HCa motif of group 6 (HCa/6) were clustered together but into two separated subgroups, whereby those accessions annotated as SBP1 proteins and the RBHs identified in this work did it in the same subgroup. Notwithstanding, proteins of subtype H2 and group-6 (H2/6) did not cluster with HCa/6 group; they were clustered with H2 proteins of group-1 (although in two subgroups properly separated). Meanwhile, proteins of subtypes HCa and HCb from groups 24 and 11.1 respectively were also clustered in two separated subgroups but sharing a common ancestor-branch.

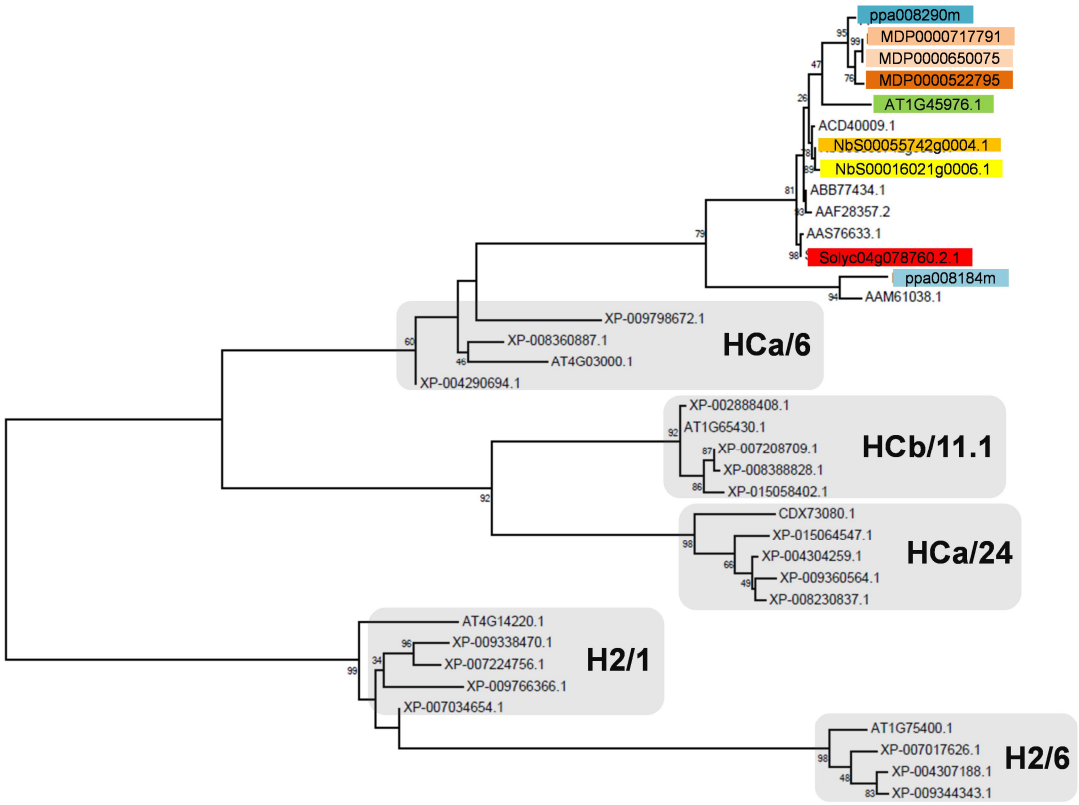


Figure 4.6. SBP1 phylogenetic tree analysis. Maximum Likelihood phylogenetic tree for accessions from 'Direct BLASTP' output, RBH results and protein accessions AT4G03000.1 (*A. thaliana*), XP_008360887.1 (*M. domestica*), XP_004290694.1 (*F. vesca*), XP_009798672.1 (*N. sylvestris*), AT1G75400.1 (*A. thaliana*), XP_004307188.1 (*F. vesca*), XP_009344343.1 (*P. bretschneideri*), XP_007017626.1 (*T. cacao*), AAF28357.2 (*P. hybrida*), ACD40009.1 (*N. alata*), AAS76633.1 (*S.chacoense*), ABB77434.1 (*Petunia inflata*), AT4G14220.1 (*A. thaliana*), XP_007034654.1 (*T. cacao*), XP_009338470.1 (*P. bretschneideri*), XP_007224756.1 (*P. persica*), XP_009766366.1 (*N. sylvestris*), AT1G65430.1 (*A. thaliana*), XP_002888408.1 (*Arabidopsis lyrata*), XP_007208709.1 (*P. persica*), XP_008388828.1 (*M. domestica*), XP_015058402.1 (*S. pennellii*), CDX73080.1 (*B. napus*), XP_015064547.1 (*S. pennellii*), XP_009360564.1 (*P. hybrida*), XP_004304259.1 (*F. vesca*), XP_008230837.1 (*P. mume*) and AAM61038.1 (*A. thaliana*). RING domain/groups (according to Stone et al., 2005 classification) are included in *grey shading*. Bootstrap values are shown for every node. Accessions highlighted by *colors* refer to 'Direct BLASTP' results (see Table 4.1b).

MdABCF

Direct BLASTP and RBH analyses

MdABCF transporter was identified in *M. domestica* (protein accession MDP0000170302). Direct BLASTP analysis against *P. persica* and *A. thaliana* showed a marked difference in the e-value threshold for highest scoring hits regarding 2nd and 3rd hits. Either way, best hits found in all protein databases (ppa002137m, Solyc08g082850.2.1, NbS00014920g0008.1 and AT1G64550.1) demonstrated to be RBHs all-to-all. Solely when apple protein database was used as subject, MDP0000899854 protein was the best match in all cases (excluding *M. domestica*). But this did it exhibiting the same e-value than MDP0000170302 (even MDP0000170302 presented higher percentages of similarity). This result may be due to the recent genome duplication occurred in apple (Figure 4.2c and Tables 4.1c and S4.3).

Identification of syntenic blocks

Contiguous locations for *P. persica*, *S. lycopersicum* and *A. thaliana* RBHs were studied for the identification of reciprocal syntenic blocks. Twenty anchors between AT1G64550.1 region (chr.1 at 23,85-24,06 Mb) and ppa002317m (chr.5 at 11,79-12,80) were found. *P. persica* region also showed synteny with two different regions of *A. thaliana* at chr.4 (6,61-6,68 and 12,45-12,52). Less anchors (12) were found between AT1G64550.1 and Solyc08g082850.2.1 (chr.8 at 62,15-63 Mb). *S. lycopersicum* region also had synteny with chr.4 (6,62-6,69 and 12,45-12,53 Mb) and 5 (16,78-16,82) of *A. thaliana*. On the other hand, more anchors between *P. persica* and *S. lycopersicum* were observed (23) encompassing shorter regions for both (approximately 350 Kb for *P.*

persica and 333 Kb for *S. lycopersicum*) in comparison to the corresponding orthologous segments observed for *A. thaliana* (Figure 4.7).

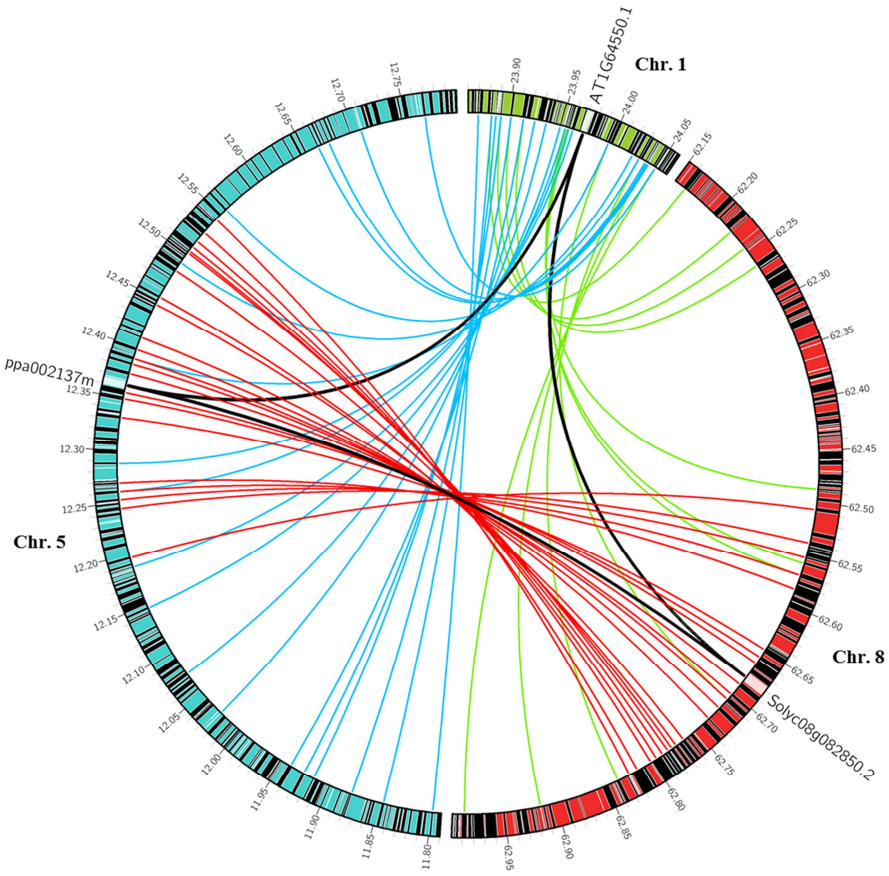


Figure 4.7. Syntenic comparative analysis of regions containing MdABCF RBHs (anchors) with *A. thaliana* (green), *S. lycopersicum* (red) and *P. persica* (blue) genomes. *Black rectangles* within circular genome regions represent gene annotation in scale, *red lines* are anchors between *P. persica* and *S. lycopersicum*, *blue lines* are anchors between *A. thaliana* and *P. persica*, and *green lines* are anchors between *P. persica* and *A. thaliana*. RBHs accessions are shown for each species in the corresponding genome regions and are connected by *black lines*. *Red triangles* represent a change of scale.

Phylogenetic analysis based on clustering methods

ABC transporter-types in plants have been recently reviewed and classified into 7 groups from A to G (Verrier et al., 2011). Particularly, pollen-expressed *M. domestica* S-RNase transporter belongs to group F (Meng et al., 2014). Proteins from Solanaceae, Rosaceae and Brassicaceae families of each ABC group, in addition to the RBHs, were used for phylogenetic analysis. LG+G+I model substitution was the best for resulting alignment. Figure 4.8 shows that all proteins were clustered in corresponding ABC group and RBHs identified in this work for the four families were grouped with MdABCF. Interestingly, three subgroups were obtained within group F. All RBHs and

MdABCF accession were included in one of them, while the other two subgroups were composed by second and third hits obtained from BLASTP analysis (some hits with an e-value threshold lower than $1e^{-5}$ not shown in Table 4.1c were also analysed and found in both subgroups). According to these results, ABCF subgroups were classified from ABCF-1 to ABCF-3, where ABCF-1 includes MdABCF and their RBHs.

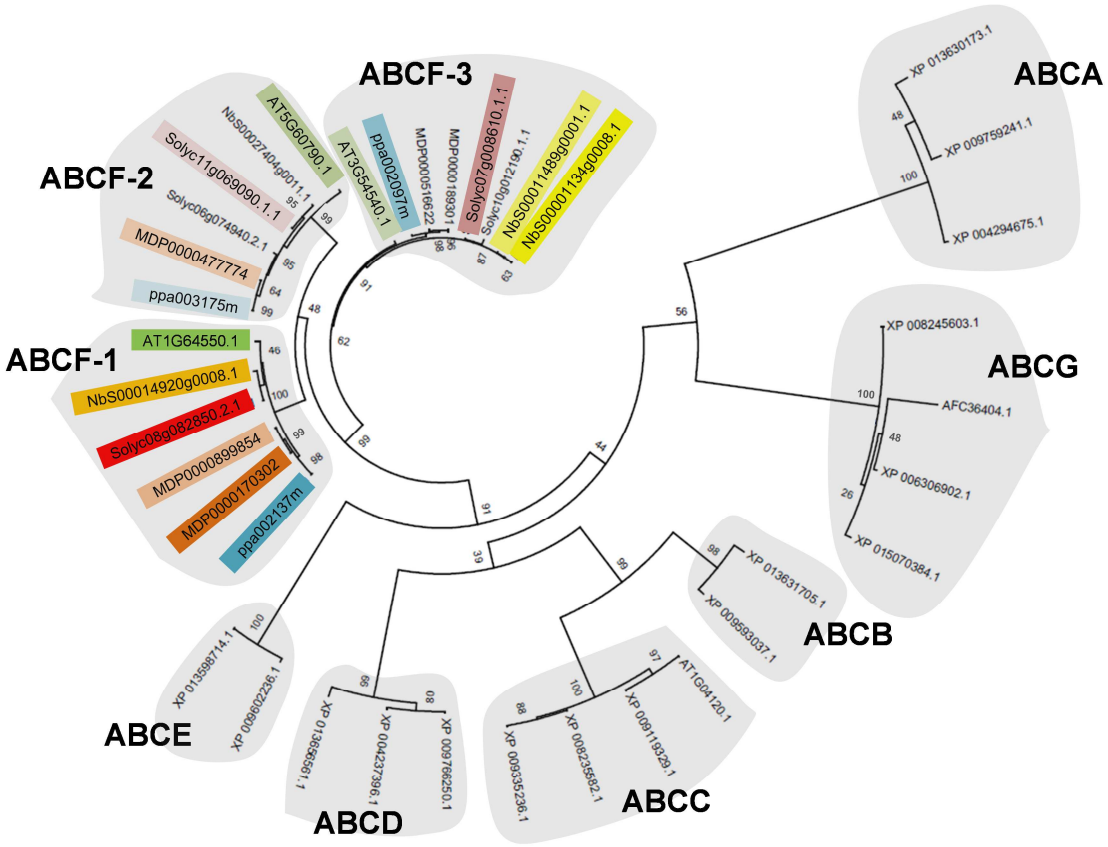


Figure 4.8. MdABCF phylogenetic tree analysis. Maximum Likelihood phylogenetic tree for accessions from 'Direct BLASTP' output, RBH results and protein accessions from ABCA (XP_013630173.1 *B. oleracea*, XP_009759241.1 *N. sylvestris* and XP_004294675.1 *F. vesca*), ABCB (XP_013631705.1 *B. oleracea* and XP_009593037.1 *Nicotiana tomentosiformis*), ABCC (AT1G04120.1 *A. thaliana*, XP_008235582.1 *P. mume*, XP_009335236.1 *Pyrus x bretschneideri* and XP_009119329.1 *B. rapa*), ABCD (XP_009766250.1 *N.sylvestris*, XP_004237396.1 *S. lycopersicum* and XP_013656561.1 *B.napus*), ABCE (XP_009602236.1 *N. tomentosiformis* and XP_013598714.1 *B. oleracea*) and ABCG (AFC36404.1 *P. hybrida*, XP_008245603.1 *P. mume*, XP_015070384.1 *Solanum pennellii* and XP_006306902.1 *Capsella rubella*) groups. These groups and ABCF group, divided in three subgroups from I to III, are included in *grey shading*. Bootstrap values are shown for every node. Accessions highlighted by *colors* refer to 'Direct BLASTP' results (see Table 4.1c).

120K

Direct BLASTP and RBH analyses

120K (AAC15893 accession from *Nicotiana alata*) was used as query for BLASTP analysis. Against *P. persica* protein database a unique significant hit was

obtained, ppa021281m (e-value = $4e^{-19}$). Similarly, in *M. domestica* only two hits (MDP0000287357 and MDP0000165381) scored e-values<$1e^{-5}$, ($5e^{-21}$ and $1e^{-19}$ e-values respectively). On the other hand, in *S. lycopersicum* and *N. benthamiana* there were more hits with low e-values. Solyc02g078050.2.1 protein had lowest e-value ($4e^{-68}$) followed by Solyc02g078060.1.1, Solyc02g078100.2.1 and Solyc02g078040.2.1 (having $2e^{-40}$, $2e^{-35}$ and $2e^{-32}$ values respectively). Whereas in *N. benthamiana* a total of 9 proteins had e-value threshold less than $1e^{-5}$; among them, the first three matches NbS00009580g0020.1, NbS00025834g0007.1 and NbS00007980g0003.1 presented similar e-values ($5e^{-36}$, $5e^{-33}$ and $5e^{-32}$). Meanwhile, AT2G33790.1 and AT2G34700.1 *A. thaliana* proteins had the same (lowest) e-values ($2e^{-19}$) (Table 4.1d).

The first direct BLASTP hit of each species was used as query to search RBHs in the five protein databases, except for *A. thaliana* where AT2G34700.1 was used. RBHs between *P. persica* and *M. domestica* were confirmed and coincided with highest-scoring hits found in BLASTP. AT2G34700.1 protein was also RBH of the Rosaceae 120K RBHs identified. However, RBHs of ppa021281m, MDP0000287357 and AT2G34700.1 in *S. lycopersicum* and *N. benthamiana* were Solyc02g078040.2.1 and NbS00006956g0008.1 respectively, and not the best BLASTP matches found for both species (Solyc02g078050.2.1 and NbS00025834g0007.1). This result shows that non-Solanaceae 120K RBHs are the first BLASTP hit either for Solyc02g078040.2.1 and NbS00006956g0008.1 as well as for Solyc02g078050.2.1 and NbS00025834g0007.1. Analogously, *S. lycopersicum* and *N. benthamiana* showed to have distinct RBHs between them in regards to BLASTP results (Solyc02g078100.2.1 and NbS00008703g0009.1) (Figure 4.2d and Table S4.4).

Identification of syntenic blocks

A. thaliana region containing AT2G33790.1 and AT2G34700.1 genes was used because both accessions are located in close proximity at chr.2 (positions 14,29 and 14,63 Mbs respectively). Meanwhile, the four *S. lycopersicum* proteins found in BLASTP analysis are positioned in tandem between 37,43 and 37,46 Mb positions at chr.2. On the other hand, a region that contains ppa21281m (2,33 Mbs) at chr.4 was included in 120K syntenic study as well. Thirty-one anchors between ppa21281m and Solyc02g078050.2.1 regions were determined, which led to define conserved syntenic blocks between *P. persica* and *S. lycopersicum* encompassing approximately 270 and 380 Kbs respectively. In turn, 40 and 22 anchors covering more extended regions for

ppa21281m (chr.4: 2,15-6,65 Mbs) and Solyc02g078050.2.1 (chr.2: 34,5-37,65 Mbs) have also shown to maintain conserved blocks of synteny with *A. thaliana*, even though these are dispersed in 2 different chromosomes (chr.1: 9,8-10; chr.2: 14,2-14,8) (Figure 4.9).

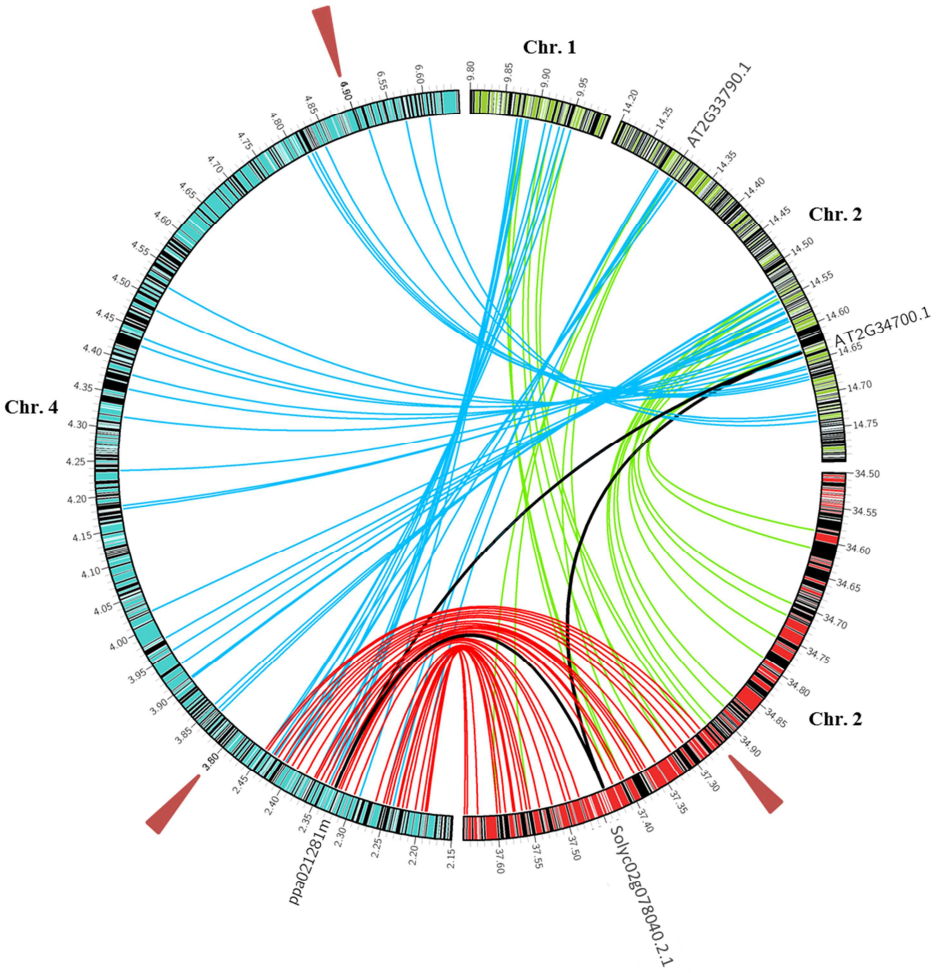


Figure 4.9. Syntenic comparative analysis of regions containing 120K RBHs (anchor) with *A. thaliana* (green), *S. lycopersicum* (red) and *P. persica* (blue) genomes. *Black rectangles* within circular genome regions represent gene annotation in scale, *red lines* are anchors between *P. persica* and *S. lycopersicum*, *blue lines* are anchors between *A. thaliana* and *P. persica*, and *green lines* are anchors between *P. persica* and *A. thaliana*. RBHs accessions are shown for each species in the corresponding genome regions and are connected by *black lines*. *Red triangles* represent a change of scale.

Phylogenetic analysis based on clustering methods

120K is an AGP from the *N.alata* stylar transmitting tract that binds to S-RNases (Cruz-García et al., 2005); however other AGPs, such as pistil extensin-like protein III (PELPIII) and Transmitting Tract-Specific glycoproteins (TTS) have been described to share conserved domains with 120K and interact with S-RNases as well (Cheung et al.,

1993;Cruz-García et al., 2005). These AGPs, Solanaceae proteins from direct BLASTP analysis and the non-Solanaceae RBHs previously identified were used for 120K tree clustering reconstruction. WAG + G substitution model was the best fit on 120K data alignment. Maximum Likelihood tree showed two differentiated groups, in one of them 3 subgroups corresponding to 120K, PELPIIIs and TTS *Nicotiana* proteins were found, with 120K and PELPs subgroups branched together. *S. lycopersicum* proteins Solyc02g078050.2.1, Solyc02g078060.1.1 and Solyc02g078100.2.1 were included in 120K, PELP and TTS subgroups respectively. The other big group was formed in one hand by the rest of proteins considered in this analysis belonging to *P. persica*, *M. domestica* and *A. thaliana* genomes, which clustered in one branch. Whereas in the other branch Solyc02g078040.2.1, a PELP protein from *Solanum nigrum* (ADW66159.1) and remaining *Nicotiana* proteins (NbS00008703g0009.1, NbS00003320g0020.1 and NbS00006956g0008.1) were clustered together (Figure 4.10).

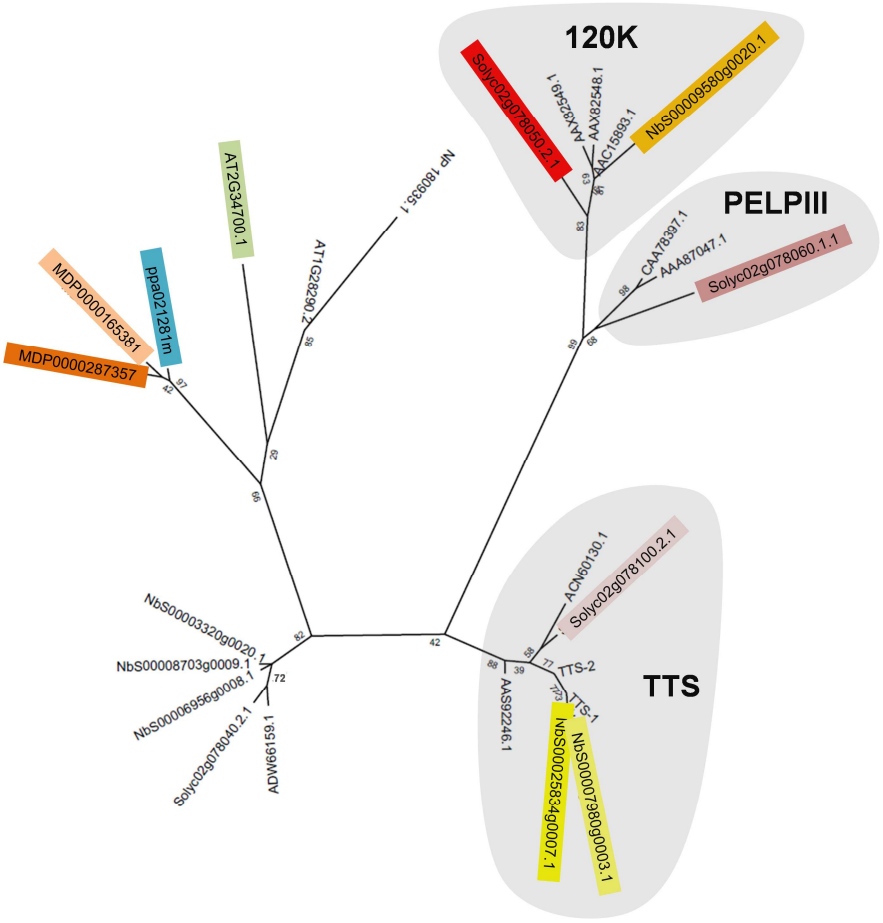


Figure 4.10. 120K phylogenetic tree analysis. Maximum Likelihood phylogenetic tree for accessions from 'Direct BLASTP' output, RBH results and protein accessions AAC15893.1 (*N. alata* 120K), AAX82548.1 (*Nicotiana tabacum* 120K), AAX82549.1 (*Nicotiana plumbaginifolia* 120K), AAA87047.1 (*N. alata* PELPIII), ADW66159.1 (*Solanum nigrum* PELP), CAA78397.1 (*N. tabacum* PELP), ACN60130.1 (*Petunia x hybrida*), AAS92246.1 (*Capsicum annum*), TTS-1 (translated protein from

transcript accession Z16403) and TTS-2 (translated protein from transcript accession Z16404). Bootstrap values are shown for every node. Three groups that correspond to proteins that branch with 120K, PELPIII and TTS accessions are included in *grey shading*. Accessions highlighted by *colors* refer to 'Direct BLASTP' results (see Table 4.1d).

NaStEP

Direct BLASTP and RBH analyses

NaStEP direct BLASTP analysis gave a noteworthy outcome. In non-Solanaceae species 3 or 4 hits having e-values < $1e^{-5}$ were found, whereas in *S. lycopersicum* and *N. benthamiana* 16 and 10 proteins were obtained, respectively (data not shown). Ten out of sixteen tomato proteins are contiguously located (from Solyc03g098670.1.1 to Solyc03g098790.1.1 including Solyc03g098710.1.1, which was the best hit in this analysis). Solyc06g072210.1.1, Solyc06g072220.1.1 and Solyc06g072230.1.1 (second hit) are also located together, and Solyc03g019690.1.1 and Solyc03g020010.1.1 (third and four hits respectively) are in close positions. Additionally, first hit in every specie showed significant lower e-values regarding second hits, with the exception of *M. domestica* where the three first hits presented similar e-values ($2e^{-22}$, $4e^{-22}$ and $9e^{-22}$) (Table 4.1e).

In RBH outcome is also important to notice that *P. persica* and *M. domestica* best hits from direct BLASTP were not RBHs. Best match for ppa011496m, ppa011653m and ppa011448m *P. persica* proteins (best three hits in this order) in *M. domestica* protein database was MDP0000326576, and best match for this protein in peach protein database was ppa011448m. Hence, MDP0000326576 and ppa011448m were RBHs. Meanwhile, AT1G17860.1 from *A. thaliana* was RBH of ppa011496m and MDP0000326576. In turn, Solyc03g098710.1.1 and NbS00018395g0002.1 (first hit in *N. benthamiana* after BLASTP analysis) were not RBHs in non-Solanaceae species, which RBHs were Solyc03g020010.1.1 and NbS00009480g0031.1 (Solyc03g020010.1.1 was also the *N. benthamiana* RBH) (Figure 4.2e and Table S4.5).

Identification of syntenic blocks

Three regions were considered for *S. lycopersicum*, those containing Solyc03g098710.1.1 (chr.3 at position 54,5 Mb), Solyc06g072230.1.1 (chr.6 at position 40,89 Mb) and Solyc03g020010.1.1 (chr.3 at position 6,8 Mb) accessions. While *P. persica* proteins obtained in direct BLASTP analysis are located in close positions at

chr.4. Lastly, AT1G17860.1 bordering region at chr.1 (6,15 Mb) was studied. Solyc03g098710.1.1 and Solyc06g072230.1.1 regions showed to be syntenyc with two different regions of the same chromosome in both *P. persica* and *A. thaliana* genomes (13,62-13,98 and 17,91-18,46 Mbs at chr.5 of *P. persica* and 6,03-6,64 and 17,96-18,03 Mbs at chr.3 of *A. thaliana*). Solyc03g020010.1.1 region also had anchor sites with 18,33-18,43 Mbs region at chr.5 of *P. persica* such as Solyc03g098710.1.1 and Solyc06g072230.1.1 regions, but additionally showed certain degree of synteny with positions 8,41-8,87 Mbs of this chromosome and with chr.2 (25,22-25,27 Mb). Furthermore, chr.1 of *A. thaliana* in both regions 6,14-7,23 and 27,57-27,60 Mbs were syntenyc with this *S. lycopersicum* region. Against these results, ppa011496m-ppa011448m region was clearly syntenic with chr.8 (59,66-59,77 Mbs) of *S. lycopersicum* and with several regions from chr.1 and 4 of *A. thaliana*. Meanwhile, AT1G17860.1 has certain degree of synteny with chr.3 (2,52-2,91) and 5 of *P. persica* between positions 17,46-18,45 (segments also syntenic with *S. lycopersicum* blocks containing Solyc03g098710.1.1 and Solyc06g072230.1.1). AT1G17860.1 showed anchors with chrs.3 and 6 of *S. lycopersicum* in locations separated by several Mbs of Solyc03g098710.1.1 and Solyc06g072230.1.1 (Figure 4.11).

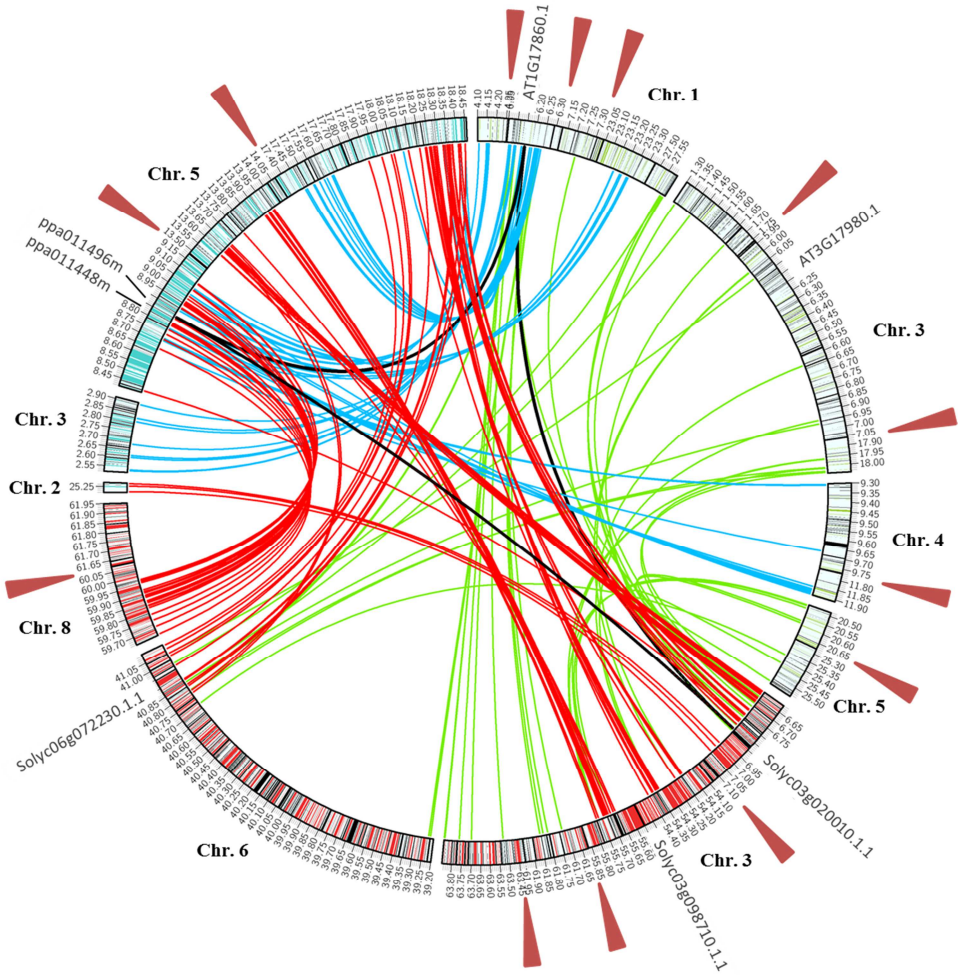


Figure 4.11. Syntenic comparative analysis of regions containing NaStEP RBHs (anchors) with *A. thaliana* (green), *S. lycopersicum* (red) and *P. persica* (blue) genomes. *Black rectangles* within circular genome regions represent gene annotation in scale, *red lines* are anchors between *P. persica* and *S. lycopersicum*, *blue lines* are anchors between *A. thaliana* and *P. persica*, and *green lines* are anchors between *P. persica* and *A. thaliana*. RBHs accessions are shown for each species in the corresponding genome regions and are connected by *black lines*. *Red triangles* represent a change of scale.

Phylogenetic analysis based on clustering methods

Busot et al. (2008) analysed different Kunitz-type (serine, aspartic and cysteine) proteinase inhibitors from I3 family (including NaStEP) and established a total of 6 clades where NaStEP was grouped in clade V. Thus, in the phylogenetic analysis for NaStEP some of proteinase inhibitor proteins used by Busot et al. (2008), including NaSoEP (a protein similar to NaStEP but showing a distinct expression pattern) were incorporated in this analysis. The best substitution model that fit with the alignment was WAG+G. All proteins extracted from Busot et al. (2008) work grouped in the same clades, only AAG38519.1 (clade IV) branched with clades II and IV, but was not included in any of both despite being closer to clade IV. *P. persica* and *A. thaliana*

RBHs clustered in clade I, plus NbS00009480g0031.1, Solyc03g020010.1.1, Solyc06g072220.1.1, Solyc06g072230.1.1, ppa011653m and AT1G73260.1, supporting RBH results. While Solyc03g098710.1.1 clustered with NaStEP protein and separately of NaSoEP, which grouped with NbS00018395g0002.1, NbS00018395g0011.1 and NbC24872723g0001.1 *N. benthamiana* proteins. In this clade V, Solyc03g019690.1.1 protein (close to Solyc03g020010.1.1 in *S. lycopersicum* genome) was also included but in a different branch of NaStEP and NaSoEP subclades. Furthermore, *M. domestica* best hits and ppa011448m protein grouped together in a separate clade of the six classified by Busot et al. (2008) This clade, classified as VII, was phylogenetically closer to NaStEP clade than clade I containing the rest of *P. persica* proteins used in this analysis (Figure 4.12).

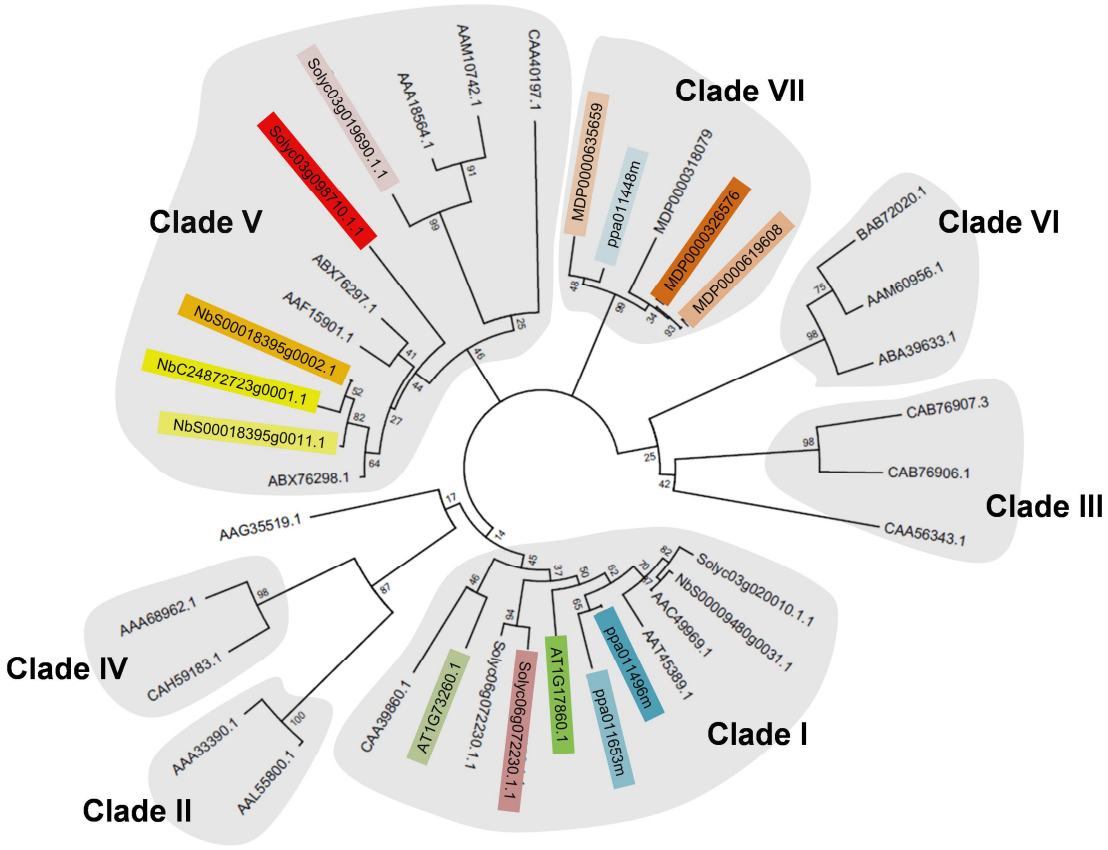


Figure 4.12. NaStEP phylogenetic tree analysis. Maximum Likelihood phylogenetic tree for accessions from 'Direct BLASTP' output, RBH results and protein accessions ABX76297.1 (*N. alata* NaStEP), ABX76298.1 (*N. alata* NaSoEP), AAF15901.1 (*Nicotiana glutinosa*), CAA40197.1 (*Solanum tuberosum*), AAA18564.1 (*S. tuberosum*), AAM10742.1 (*S. tuberosum*), BAB72020.1 (*Raphanus sativus*), AAM60956.1 (*A. thaliana*), ABA39633.1 (*Brassica oleracea*), AAC49969.1 (*N. tabacum*), AAT45389.1 (*Medicago trucatula*), CAA39860.1 (*Theobroma cacao*), AAL55800.1 (*Ipomoea batatas*), AAA33390.1 (*I.batatas*), CAB76907.3 (*Cicer arietinum*), CAB76906.1 (*C. arietinum*), CAA56343.1 (*Glycine max*), AAG38519.1 (*Citrus paradisi*), CAH59183.1 (*Populus tremula*), AAA68962.1 (*Salix viminalis*) taken from Busot et al. (2008). Six clades (from I to VI) according to Busot et al. (2008) classification and clade VII (defined in this work) are shown in *grey shading*. Bootstrap values are shown for every node. Accessions highlighted by *colors* refer to 'Direct BLASTP' results (see Table 4.1e).

NaPCCP

Direct BLASTP and RBH analyses

NaPCCP direct BLASTP analysis produced several hits under $1e^{-5}$ e-value threshold against each protein database. The first three *P. persica* hits (which are alternative transcripts of ppa012133m) had an equal e-value of $4e^{-99}$, fourth (ppa025944) and fifth (ppa012484m) hits presented $1e^{-85}$ and $9e^{-81}$ e-values, respectively. In *M. domestica*, from first (MDP0000525794) to fifth (MDP0000206691) hits e-values ranged from $2e^{-100}$ to $5e^{-80}$. Similarly to *M. domestica*, from *S. lycopersicum* first (Solyc12g040800.1.1) to fourth (Solyc02g091520.2.1) hit, e-values were under $6e^{-80}$. The five first hits in *N. benthamiana* presented the lowest e-values of all protein databases analysed (from NbS00020637g0006.1 ($3e^{-119}$) to NbS00051736g0004.1 ($6e^{-98}$)), while *A. thaliana* the highest (from first hit AT3G17980.1 to fifth hit AT5G37740.1 e-values ranged from $1e^{-92}$ to $2e^{-75}$, where second and third hits are alternative transcripts) (Table 4.1f and data not shown).

All non-Solanaceae best-scoring matches obtained in previous analysis were confirmed to be RBHs to each other; meanwhile Solanaceae first hits after BLASTP analysis were also corroborated as RBHs between *S. lycopersicum* and *N. benthamiana* species. Nevertheless *P. persica*, *M. domestica* and *A. thaliana* RBHs in this two species (Solyc03g118720.2.1 and NbS00009334g0006.1) were different to *S. lycopersicum* and *N. benthamiana* RBH pairs (excluding *N. benthamiana* RBH of *A. thaliana* that was NbS00020564g0001.1). It must be held that best hits in RBH analysis for non-Solanaceae species against *S. lycopersicum* and *N. benthamiana* protein databases had e-values that ranged in close thresholds (Figure 4.2f and Table S4.6).

Identification of syntenic blocks

Dissimilarities between Solanaceae and the rest of families observed in RBH study were once again apparent in macrosynteny analysis. A small number of anchors were found for Solyc12g040800.1.1 (Solanaceae RBH) region with both *P. persica* and *A. thaliana* genomes, and these are scattered in distanced regions. On the other hand, Solyc03g118720.2.1 (*P. persica, M. domestica* and *A. thaliana* RBH) region at chr.3 (61,46-61,91 Mb) showed to conserve an orthologous genomic segment with ppa012133m region at chr.5 (17,40-17,86 Mb) of *P. persica*, supported by a total of 37

anchor sites. Whereas distinct regions containing AT3G17980.1 and AT1G48590.2 (second BLASTP hit) proteins at chrs. 3 (6,06-6,25 Mb) and 1 (6,20-6,31/17,94-17,99/27,65-27,75 Mb) were also syntenic between *S. lycopersicum* and *A. thaliana* (29 anchors). These *A. thaliana* regions were syntenic with ppa012133m *P. persica* region held up by 41 anchors (Figure 4.13).

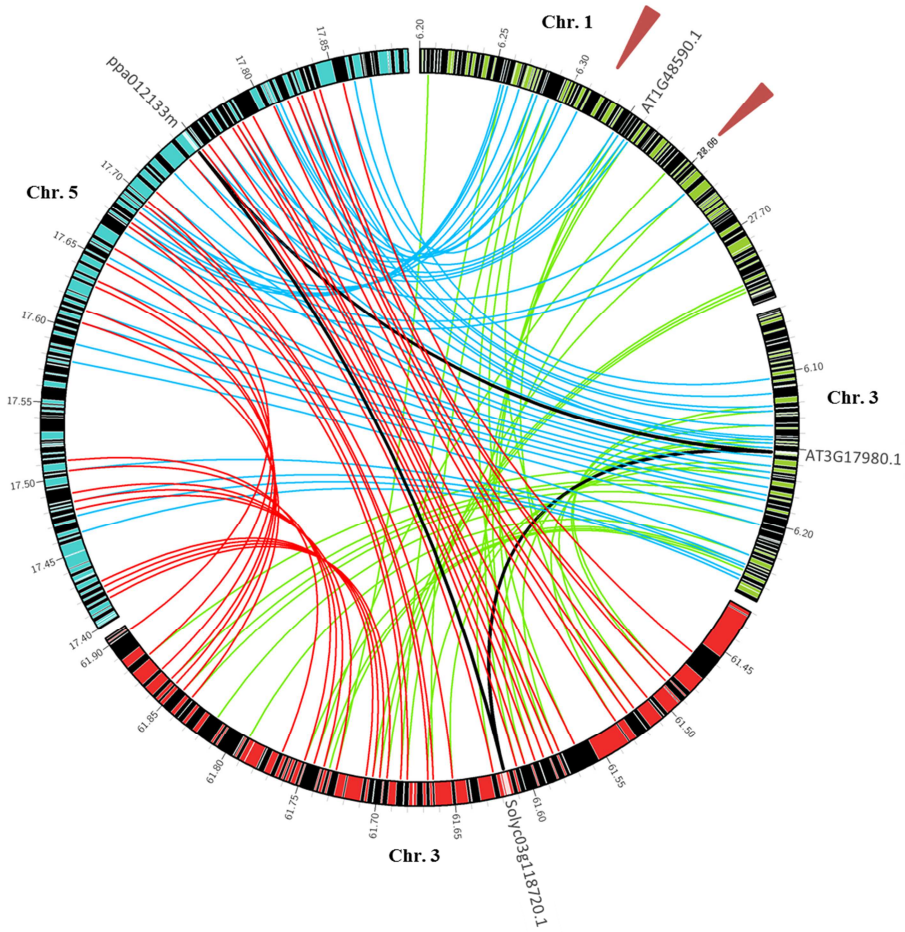


Figure 4.13. Syntenic comparative analysis of regions containing NaPCCP RBHs (anchors) with *A. thaliana* (green), *S. lycopersicum* (red) and *P. persica* (blue) genomes. *Black rectangles* within circular genome regions represent gene annotation in scale, *red lines* are anchors between *P. persica* and *S. lycopersicum*, *blue lines* are anchors between *A. thaliana* and *P. persica*, and *green lines* are anchors between *P. persica* and *A. thaliana*. RBHs accessions are shown for each species in the corresponding genome regions and are connected by *black lines*. *Red triangles* represent a change of scale.

Phylogenetic analysis based on clustering methods

Lastly, phylogenetic NaPCCP reconstruction was carried out with proposed RBHs, some proteins from BLASTP analysis with low e-values and protein accessions from BLASTP search into NCBI database. In this alignment, LG+G was the substitution model that best fit. Maximum Likelihood tree-clustering showed that NbS00020637g0006.1 (*N. benthamiana* RBH between Solanaceae species),

NbS00009698g0010.1 (2nd hit in BLASTP analysis) and Solyc12g040800.1.1 (*S. lycopersicum* RBH between Solanaceae species) grouped with NaPCCP protein. A subgroup encompassing Solyc06g068940.2.1 and NbS00020564g0001.1 (2nd and 3rd best hits in BLASTP analysis) was the nearest subgroup to NaPCCP. *A. thaliana* accessions AT3G17980.1 and AT1G48590.1 (AT1G48590.2) proteins along with the *Brassica* accessions were next subgroup in proximity, followed by another subgroup formed by Solanaceae accessions that contains Solyc03g118720.2.1 protein (RBH of *P. persica*, *M. domestica* and *A. thaliana* species). A different subgroup formed by ppa012133m, MDP0000525794 (MDP0000776395) and accessions from *Fragaria vesca* and *Prunus mume* was the subgroup integrated by Rosaceae accessions closest to NaPCCP subgroup. All these subgroups defined one of the two larger groups found in this analysis. The other group containing a smaller number of proteins, compared to the previous one, encompassed Solyc02g091520.2.1, AT5G37740.1 and the rest of Rosaceae proteins obtained in BLASTP analysis (Figure 4.14).

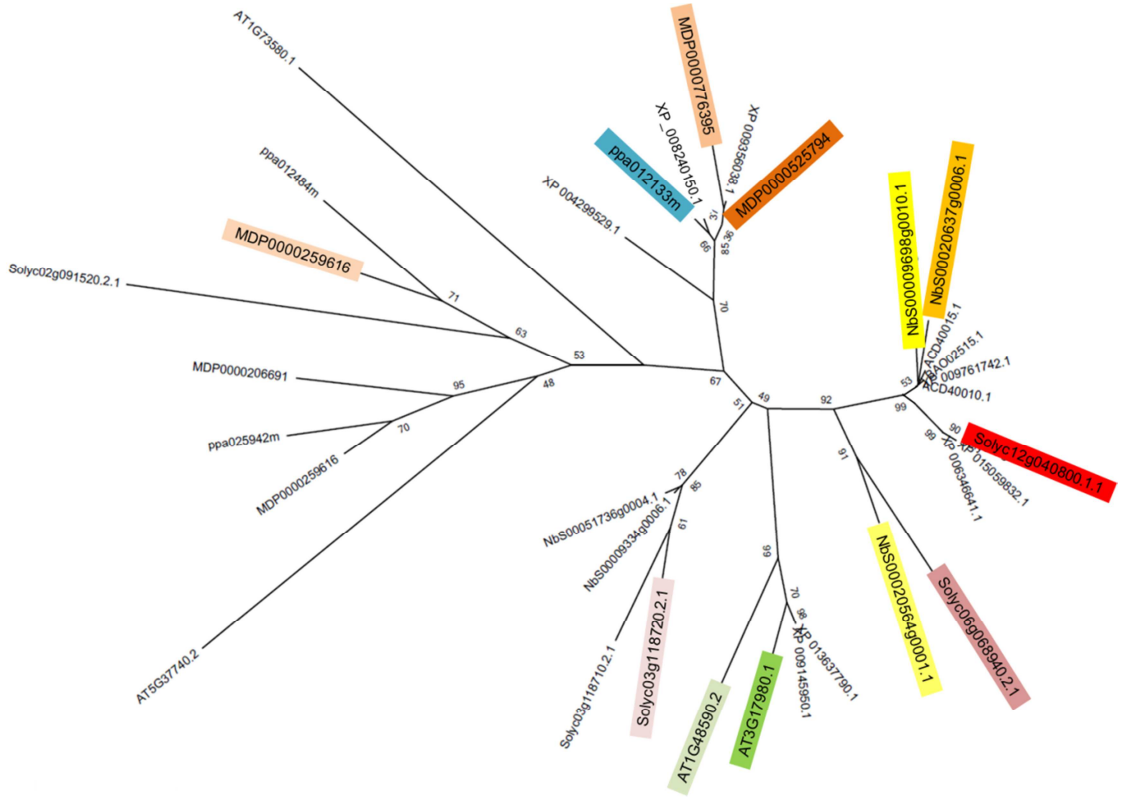


Figure 4.14. NaPCCP phylogenetic tree analysis. Maximum Likelihood phylogenetic tree for accessions from 'Direct BLASTP' output, RBH results and protein accessions ACD40010.1 (*N. alata*), XP_009761742.1 (*N. sylvestris*), BAO02515.1 (*N. alata*), ACD40015.1 (*Nicotiana bonariensis*), XP_006346641.1 (*S. tubersosum*), XP_015059832.1 (*S. pennellii*), XP_008240150.1 (*P. mume*), XP_009356038.1 (*P. bretschneideri*), XP_004299529.1 (*F. vesca*), XP_009145950.1 (*B. rapa*) and XP_013637790.1 (*B. oleracea*). Bootstrap values are shown for every node. Accessions highlighted by *colors* refer to 'Direct BLASTP' results (see Table 4.1f).

Discussion

Gametophytic Self-Incompatibility in Rosaceae and Solanaceae families is based on an allele-specific recognition mechanism between the same *S*-determinant types. In addition, some non-*S*-factors involved in this mechanism have been demonstrated to have equivalent functions among species of both families such as for instance the components of SCF^{SLF} complexes (Huang & Kao 2006; Matsumoto & Tao 2012). Therefore, it would not be surprising that other modifiers descending from common ancestors may have been preserved throughout the evolution in both families.

Several methods have been developed to identify orthologous genes on the basis of sequence similarity. Reciprocal Best Hit (RBH) approach is a tool typically employed for this purpose (Zheng et al., 2005), however this methodology may be misleading because duplication events and the consequent emergence of paralogues or co-orthologues might produce erroneous assessments or fail to capture complex relationships (Wolf & Kooning 2012). On the other hand, phylogenetic tree-clustering is comparable to RBH since both are based on pairwise similarity. Notwithstanding, phylogenetics leads to estimate the history of divergence and may also offer valuable information regarding protein structure and conserved domains. This information, hardly ascertainable by RBH strategy, is very useful for orthologue discrimination. The disadvantage of both approaches is how to face complex duplication events, such as segmental or tandem duplications and transposition events (Kong et al., 2007), as well as the loss of genes across evolution (Gabaldon & Kooning 2013). This handicap can be solved, or partially solved, by the identification of genomic syntenic blocks which might shed some light on the divergent evolutionary pathways occurred among species (Zheng et al., 2005).

In this work, a complementary study based on the 3 distinct approaches described above was carried out allowing us to identify putative orthologues for some of the GSI modifiers previously reported in the literature and to propose alternative scenarios when they could not be detected.

Putative orthologues for NaTrxh, MdABCF and SBP1 were found in *Prunus*

First hits obtained in direct BLASTP analysis for NaTrxh, MdABCF and SBP1 demonstrated to be RBHs to each other in almost all performed comparisons. Thus, on the basis of the RBH analyses, ppa011576m, ppa002137m and ppa008290m accessions

were proposed as putative orthologues for NaTrxh, MdABCF and SBP1 modifiers in *Prunus*, respectively.

Additionally, comparative structural genomics highlighted the existence of orthologous segments among regions containing *P. persica*, *S. lycopersicum* and *A. thaliana* RBHs for NaTrxh, MdABCF and SBP1. It is noteworthy that *P. persica* and *S. lycopersicum* (sharing *S*-RNase-based GSI system) have a higher degree of synteny (based on the number of anchor points) in comparison with *A. thaliana* (SSI outgroup system) despite Rosaceae and Brassicaceae are phylogenetically closer than Rosaceae and Solanaceae (Igic and Kohn 2001). It could also be mentioned that ppa011576m (NaTrxh *Prunus* orthologue) and ppa017665m (PaMDOr; Chapter III) are separated by 1,3 Mb at the *P. persica* chr.3. While the corresponding syntenic regions in *S. lycopersicum* for both genes are located in similar distances at chr.2, pointing out the conservation of extended orthologous segments in both families.

Lastly, these modifiers belong to distinct and large protein families which have been widely studied and classified. This information was used accordingly for phylogenetic tree clustering. TRX proteins are classified in 8 groups (where h group is, in turn, divided in 3 subgroups from I to III) (Meyer et al., 2012). In ABC transporters 7 groups have been established (Verrier et al., 2011) and RING-HC proteins have also been categorized in a large number of groups in *A. thaliana* (Stone et al., 2005). Phylogenetic history reconstruction confirmed a close relationship between candidates and modifiers in each case. Interestingly, in MdABCF phylogeny study, ABCF group was branched in three well defined subgroups, where orthologous candidate genes and this modifier were included in the same subgroup (ABCF-1). Similarly, TRX hII subgroup (to which NaTrxh belongs) was also divided into three subgroups, putative orthologues and NaTrxh were included in the same subgroup. Meanwhile, in SBP1 two HCa/6 subgroups were found, and SBP1 accessions were properly classified with candidates within the same HCa/6 subgroup.

A similar function might be expected for these orthologues in both families. In this sense, SBP1 role in GSI has not been fully elucidated, but *N. alata* SBP1 (NaSBP1) has been shown to interact with both *S*-determinants and AGP proteins. In a recent work, the putative SBP1 orthologue in *Prunus avium* (PavSBP1) has been cloned (Matsumoto & Tao, 2016) (which is in turn orthologous to ppa008290m accession described in this work). The protein-protein interaction analyses carried out in PavSBP1 did not detect any of the interacting-NaSBP1 factors. This evidence does not support

PavSBP1 as a modifier factor in *Prunus* GSI system in spite of its evolutionary history suggests a common origin with Solanaceae SBP1 accessions. Regarding NaTrxh and MdABCF both have shown to interact with S-RNases in *Nicotiana alata* and *Malus domestica*, respectively, but no evidence is currently available about their functions in the counterpart family.

Orthology relationships could not be inferred for 120K, NaStEP and NaPCCP

No *Prunus persica* predicted peptide was found to fulfill the requirements established in this work for being considered orthologous to 120K, NaStEP and NaPCCP modifiers. Notwithstanding, some considerations can be extracted that might help to glimpse the orthology relationships.

Most likely, the orthology study performed in 120K reflects the most understandable evolutionary framework of these three modifiers. RBH and phylogenetic analyses highlight a presumably distinct origin for those Rosaceae and Solanaceae proteins similar to 120K. Rosaceae and Solanaceae RBHs do not match and phylogeny results not only exclude Rosaceae RBHs from the 120K subgroup, but also from the S-RNase-binding AGPs subgroups (PELPIII and TTS). However, *S. lycopersicum* region that contains 120K orthologue protein (Solyc02g078050.2.1) also contains proteins homologous to PELPIII and TTS (on the basis of phylogeny outcome) and the RBH for the *P. persica* ppa021281m (Solyc02g078040.2.1). Moreover, this *S. lycopersicum* region and the one in *P. persica* containing ppa021281m are syntenic. Therefore, comparative genomics might suggest a common origin for this region that suffered tandem duplications in Solanaceae after Asteridae and Rosidae splitting. This scenario is supported by recent findings in *Nicotiana spp.* where stylar AGPs have been proposed to have a common origin, with initial intron insertion followed by two gene duplication events. It is well known that 120K is taken up into the cytoplasm of pollen tubes and required for proper S-recognition in *N. alata* (Cruz-García et al., 2003 and 2005), but PELPIII is also necessary for *N. tabacum* interspecific incompatibility (Smith et al., 2013) and TTS promotes pollen tube growth in this specie as well (Cheung et al., 1995). Thus, different stylar AGPs may operate in the incompatibility reaction during pollen-pistil interaction. In this general context, it cannot be fully discarded that ppa021821m diverged from the stylar AGP ancestor.

NaStEP orthology screening draws a complex evolutionary pattern. In this case, no RBHs for the first direct BLASTP hits were found, including between *S.*

lycopersicum and *N. benthamiana*. On the other hand, syntenic block identification between *P. persica* and *S. lycopersicum* was difficult to interpret since several segmental duplications have been detected. In addition, the first three *P. persica* hits obtained in BLASTP analysis seem to have emerged from tandem duplications. Lastly, a transposition event in Rosaceae or gene loss in Solanaceae might have occurred since syntenic *S. lycopersicum* region to the *P. persica* segment has not genes with high sequence similarity to NaStEP. Furthermore, phylogenetic tree clustering does not either support that best *P. persica* hit (in BLASTP analysis) is close related to NaStEP, because this has been clustered in clade I and not into clade V where NaStEP is included (Busot et al., 2008). Notwithstanding, clade coined as VII (containing the *Malus* and *Prunus* RBH proteins including ppa011448m) was the closest clade to clade V under phylogenetic reconstruction. Overall, according to these results it is difficult to infer orthology relationships regarding NaStEP.

NaPCCP first hits found in direct BLASTP analysis for Solanaceae species (Solyc12g040800.1.1 and NbS00020637g0006.1) showed to be RBH to each other. However, both accessions were not RBHs of the non-Solanaceae first hits obtained in direct BLASTP analysis. These non-Solanaceae accessions had different RBHs in *S. lycopersicum* and *N. benthamiana* (Solyc03g118720.2.1 and NbS0009334g0006.1). Accordingly, the region encompassing the *S. lycopersicum* RBH for non-Solanaceae species (Solyc03g118720.2.1) was highly conserved in *P. persica* (ppa012133m) and *A. thaliana* (AT3G17980.1) regions. But the region containing *S. lycopersicum* first hit from direct BLASTP analysis (Solyc12g040800.1.1) did not show synteny with any of both *P. persica* and *A. thaliana* genomes. Moreover, phylogenetic analysis showed two large groups, and the Rosaceae subgroup was the farthest related to NaPCCP subgroup. Summarizing, orthology results do not support a *Prunus* orthologue candidate for the NaPCCP modifier.

A complex evolutionary pattern is predicted for the GSI modifiers as a whole

In short, the orthologue screening performed for 120K, NaTrxh, NaStEP, MdABCF, SBP1 and NaPCCP modifier factors resulted in different degrees of fulfillment. *P. persica* ppa011576m, ppa002137m and ppa008290m accessions are putative orthologues for NaTrxh, MdABCF and SBP1. Furthermore, ppa021281m and ppa011448m cannot be discarded as 120K and NaStEP orthologues, while a no clear

orthologue has been found for NaPCCP. Orthology analysis outcome is shortly shown in Table 4.2.

Table 4.2. Summary results from the orthology relationships analysis

	Pistil modifiers			Pollen modifiers		
	120K	**NaTrxh**	**NaStEP**	**MdABCF**	**SBP1**	**NaPCCP**
Reciprocal Best Hit (RBH)	**Failed**	**Fulfilled**	**Failed**	**Fulfilled**	**Fulfilled**	**Failed**
Orthology mapping	**Fulfilled**	**Fulfilled**	**Failed**	**Fulfilled**	**Fulfilled**	**Failed**
Phylogenetic analysis	**Failed**	**Fulfilled**	**Partially fulfilled**	**Fulfilled**	**Fulfilled**	**Failed**
	ppa021281m	**ppa011576m**	**ppa011448m**	**ppa002137m**	**ppa008290m**	**ppa012133m**

Two main scenarios could explain these observations: a divergent process from an ancestral mechanism leading to different mechanisms where gene duplications and losses might have occurred, or a convergent process where proteins from different lineages (or even families) have been recruited to control the SI mechanism in a similar manner though intermediate scenarios can not be discarded. Nevertheless, compiled evidences from this work are not robust enough to support a particular model. In any case, this work was primarily aimed to lay the foundation for the identification in *Prunus* of those genes accomplishing the function of the Solanaceae GSI modifiers. Orthology is a helpful tool for this purpose but it does not necessarily imply that functions are preserved (Gabaldón & Kooning, 2013). Thus, further molecular and sequence analyses will be needed to achieve this goal.

Material and Methods

Data

Assembled and annotated genomes from tomato (*Solanum lycopersicum*) (Consortium 2012), *Nicotiana benthamiana* (Bombarely et al., 2012), peach (*Prunus persica*) (Verde et al., 2013), apple (*Malus x domestica*) (Velasco et al., 2010) and *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), as well as their corresponding protein databases, were used. Thus, four databases available from the GDR database (www.rosaceae.org) (Prunus_persica_v1.0_scaffolds.fa, Prunus_persica_v1.0_scaffolds.gff3, Prunus_persica_v1.0_peptide.fa and

Malus_x_domestica.v1.0.consensus_peptide.fa), four more from the SolGenomics Network (www.solgenomics.net) (S_lycopersicum_chromosomes.2.30.fa, ITAG2.3_gene_models.gff3, ITAG2.3_proteins.fasta and Niben.genome.v0.4.4.proteins.fasta) and three from TAIR database (TAIR10_chr_all.fas, TAIR10_GFF3_genes.gff and TAIR10_pep_20101214.txt) were used for the different screenings of orthology.

Direct BLASTP analysis and Reciprocal Best Hit (RBH) methodology

Putative modifier factors of the S-RNase-based GSI system not identified in *Prunus* species were included in the analysis (Table 4.3). Nucleotide and amino acid sequences were retrieved from NCBI database. Protein sequences were blasted (Altschul et al., 1990) as queries against the five protein databases mentioned above using available stand-alone BLAST version 2.2.28+ software (http://www.ncbi.nlm.nih.gov) with an expected value cut-off $<1e^{-5}$. Resulting hits from previous 'direct BLASTP analysis' were used in turn as queries for RBH identification by BLASTP and handmade scripts through simultaneous comparisons of all-against-all references (using the same set parameters than "BLASTP analysis") (Figure 4.1).

Table 4.3. Summary data of modifier genes (already reported) used for orthology screening

	Modifier factor	Specie	Localization	Interacting/binding protein	Accession	Protein ID	Reference
Pistil modifiers	**120K**	Nicotiana alata	ECM of STT	S-RNase/SBP1/ NaPCCP	U88587	AAC15893	Lind et al., 1996
	NaTrxh	Nicotiana alata	ECM of STT	S-RNase	DQ021448	AAY42864	Juarez-Díaz et al. 2006
	NaStEP	Nicotiana alata	Stigma	HT-B	EU253563	ABX76297	Busot et al., 2008
Pollen modifiers	**MdABCF**	Malus domestica	pollen tube membrane	S-Rnase	MDP0000170302	MDP0000170302	Meng et al., 2014
	SBP1	Petunia hybrida	Pollen	S-RNase/SLF/ Cul1/120K	AF223395	AAF28357	Sims and Ordanic 2001
	NaPCCP	Nicotiana alata	Mature pollen	NaTTS/Na120K	EU591515	ACD40010	Lee et al., 2008

Orthology mapping

Genome sequence data, gene annotation (.gff) files and predicted protein databases of *Prunus persica*, *Solanum lycopersicum* and *Arabidopsis thaliana* were used to localize mutual syntenic blocks among three genomes. The translated amino acid sequence of twenty-five genes upstream and downstream of the RBHs found in previous section (51 in total) either for *S. lycopersicum*, *P. persica* and *A. thaliana* were used as queries to identify the RBHs pairs (performed with the same set parameters than

in “Direct BLASTP analysis and RBH methodology” section) comparing all-against-all. A RBH was considered as anchor site; the presence of several anchors led to define blocks of synteny. The number of neighboring genes was employed (and not genomic distances) because the intergenic sizes for every species are highly variable to each other. All these analyses were carried out by custom-made python scripts using executable gffutils (www.pythonhosted.org/gffutils) and blastall (Altschul et al., 1990) packages. Anchors supporting syntenic blocks were visualized by *Circos software* (Krzywinski et al., 2009).

Phylogenetic tree-based analysis

Phylogenetic analyses were conducted by *MEGA6* (Tamura et al., 2013). Amino acid sequences were aligned by ClustalW (Thompson et al., 1994) and model substitution that best fit with corresponding alignment was determined. Phylogenetic relationship tree was constructed by the Maximum Likelihood method (Felsenstein, 1981) and phylogenetic test was done based on 1,000 bootstrap replicates.

GENERAL DISCUSSION

An *S*-locus independent pollen factor confers self-compatibility in Katy apricot

In this work the North-American apricot cv. 'Katy', released by Zaiger´s Genetics (Modesto, CA, USA) in 1978 (Russell, 1998), was confirmed as self-fruitful and its *S*-genotype was determined as S_1S_2 following the nomenclature established by Burgos et al. (1998). To investigate the genetics of self-compatibility (SC), 'Katy' (S_1S_2) was self-pollinated and reciprocally crossed with the self-incompatible cv. 'Goldrich' (S_1S_2) (Egea & Burgos, 1996; Alburquerque et al., 2002). 'Katy' pollen tubes bearing either the S_1- or the S_2-haplotype were able to grow in 'Katy' and 'Goldrich' pistils and to complete fertilization, producing the three *S*-genotype classes expected for an F_2 population (S_1S_1: S_1S_2: S_2S_2). However, no progeny was obtained in the reciprocal cross using 'Katy' as female parent. These results would support a PPM unlinked to the *S*-locus as the cause for SC.

SC caused by loss of pollen-*S* function has been usually found to be associated with mutations (mainly indels) of the *SFB* genes in different *Prunus* species such as sweet cherry (Ushijima et al., 2004; Sonnelveld et al., 2005; Marchese et al., 2007), apricot (Vilanova et al., 2006), Japanese apricot (Ushijima et al., 2004), peach (Tao et al., 2007) and sour cherry (Hauck et al., 2006). However, sequence analysis revealed no mutations or indels affecting any of the two 'Katy' *SFB* alleles discarding this as the cause of SI breakdown.

In Solanaceae, self-compatible PPMs may arise from *S*-allele duplications located in a centric fragment, in a non-S chromosome or linked to the *S*-locus leading to the formation of *S*-heteroallelic pollen (Golz et al., 2001). According to the segregations obtained in the performed crosses, *S*-allele duplications did not seem probable in 'Katy' (all descendants should have had the S_1S_2 genotype), even so, we discarded that possibility showing that *SFB* gene dosage is equivalent between 'Katy' and the self-incompatible cv. 'Goldrich'. *S*-allele duplications may also result from polyploidy but 'Katy' was confirmed as diploid by flow cytometry analysis and by marker segregation and mapping in all crosses. These results rule out competitive interaction resulting from *S*-heteroallelic pollen as the cause of SC in 'Katy'.

Altogether, it can be hypothesized that the loss-of-function of a *S*-locus unlinked factor gametophytically expressed in pollen causes breakdown of SI in 'Katy'.

Moreover, according to the relative abundance of *SFB*$_1$ and *SFB*$_2$ transcripts in 'Katy', when compared with the reference cv. 'Goldrich', the hypothetical defective factor in 'Katy' does not seem to affect their expression.

The *M'*-locus genomic region should correspond to a segregation distortion locus (SDL), a chromosomal region that causes distorted segregation ratios (Zhu & Zhang, 2007). To identify this kind of regions, 'K×K$_{05}$' and 'K×K$_{06}$' populations, which all trees carry the PPM, were tested for genome-wide distributed SSRs to detect SDL by examining changes in genotypic frequencies. Attending to segregation of pollen alleles, two SDL were found in LG3 and LG6 but a deeper analysis showed that LG6 markers were partially linked to the *S*-locus and only moderately distorted. Consequently, LG3 was predicted as the most likely location for the *M'*-locus.

In a second step, to refine *M'*-locus mapping, chr.3 specific SSRs were analyzed to estimate their segregation distortion ratios in selfing (F_2) and outcrossing populations obtained by using 'Katy' as pollen parent. Additionally, indirect *M'*-locus genotyping was performed by analyzing linked SSRs in the F_3 offspring of six selected 'K×K' F_2 trees. Recombination breakpoints in five of these trees defined a 9.4 cM interval for the 'Katy' *M'*-locus that corresponds to ~1.29 Mb in the peach genome (18.49-19.78 Mb) and overlaps ~273 Kb with that established for the *M*-locus in 'Canino' (Zuriaga et al., 2012). Interestingly, both cultivars have different geographic origins [(i.e. 'Katy' is a North-American apricot selection (Russell, 1998) and 'Canino' is a local Spanish apricot (Vilanova et al., 2006)] and, according to the analysis of gemome-wide distributed SSRs, they seem to be genetically unrelated. This prompts us to speculate that both PPMs (being or not the same) may have arisen independently.

Pollen-part mutated *m*-haplotype is associated with self-compatibility and widely distributed in apricot germplasm

Though phenotype could not be directly assessed in some cases, according to the almost perfect association between SC and S_C/*m*-alleles it can be inferred that all cultivars carrying whatever of these two alleles will also be self-compatible. Two new mutations putatively conferring SC have been found. S_{31} is shared by two of the few North-American self-compatible cultivars and sequence analysis point out a putative indel within the *SFB*$_{31}$ 3´-end as a plausible cause for SC, similarly to many other cases reported in *Prunus* (Tao and Iezzoni 2010). Lastly, genetic analysis suggests the presence of a SNP mutation within *SFB*$_2$ HVb region in 'Portici' that could also be

associated with SC. It could be speculated that a single non-synonymous change within a SFB hypervariable region might alter its specificity, since these domains (strongly hydrophobic and under positive selection) were already suggested to have a role in the specific recognition of S-RNases (Ikeda et al., 2004).

In addition to S_{31}, two new *S*-alleles (S_{29} and S_{30}) were identified and named basically according to the nomenclature previously adopted by Vilanova et al. (2005) [S_1-S_7 and S_C], Halázs et al. (2005) [S_8-S_{16}] and Wu et al. (2009) [S_{17}-S_{28}]. According to the *S*-genotyping results S_5 reported in this work is proposed to be the same that S_{13} reported by Halász et al. (2010) in the Armenian cv. 'Shalah'. This finding is relevant since connects this low frequent allele, mainly found in Armenian, Eastern-Turkish and Moroccan cultivars (Halász et al., 2010; Kodad et al., 2013) with Southern-Spanish cultivars (Burgos et al., 1998; Vilanova et al., 2005; this work) supporting the Southwest-Mediterranean diffusion route for apricot, from the Irano-Caucasian gene pool, proposed by Bourguiba et al. (2013).

Pollen-part mutated *m*-haplotype had been previously associated with SC in 'Canino' and 'Katy' cultivars (Zuriaga et al., 2012 and 2013). In this work the *m*-haplotype has been detected in 17 additional cultivars (excluding 'Canino' clonal sibs) mainly Spanish (12 in total) but also from USA, Australia, France and Italy. Fifteen of them were confirmed as self-compatible. The analysis of progenies from two of them ('Portici' and 'Corbató') fully confirmed the association with SC in apricot germplasm. Beside the *m*-haplotype, 37 additional *M*-haplotypes were identified by SSR analysis being grouped in 19 'main' classes. Regarding the distribution of the *m*-haplotype it seems to be restricted to North-American and Western-European cultivars. However, according to the clustering analyses the closest *M*-haplotype (putative founder) is $M_{1\text{-}0}$, which is widely distributed in all geographic areas studied (the second one was M_{13} only detected in Eastern-European cultivars). Meanwhile, the mutated S_C-allele is widely distributed in all geographic areas (Vilanova et al., 2005; Halász et al., 2007 and 2010, Kodad et al., 2013) but the ancestor S_8-allele was only detected in Hungarian cultivars. Altogether, these results suggest that the mutated *m*-haplotype arose much later in time, after apricot was established as a regular crop in Europe.

The *Prunus armeniaca M*-locus Disulfide bond A-like Oxidoreductase (*PaMDOr*) gene is an essential pollen factor for self-incompatibility

'Goldrich' BAC clones covering the *M*-locus were sequenced and assembled to get an apricot reference sequence useful for this purpose. Three major contigs were obtained and GAPs were joined by indeterminations defining the a*M*-supercontig. This strategy provided 40 new SSRs and first apricot SNPs for the analyzed region. In addition, new recombinants were used to refine the available maps. Altogether, an apricot ~134 Kb *M*-locus region was decided to be screened for the identification of the PPM. Fifteen genes were annotated in the apricot *M*-locus region (~134 Kb) using RNAseq data and all of them were found to be highly conserved in other *Prunus spp.* according to collinearity and homology rates. The 15 genes showed to be expressed in all tissues, therefore no specific pollen-expressed genes are contained in this region. However, four of these (*PaM*-6, -7, -9 and -14) showed higher differential over-expression in mature anthers with regards to other tissues and therefore may be considered as candidate genes. In parallel to gene-expression analysis, variants of any nature, from SNPs to structural variants, were called for the apricot *M*-locus region in the three reference self-incompatible/self-compatible cultivars. Only one variant (indel) fulfilled all genetic requirements for being the cause of SC within the *m*-haplotype. This insertion is located within *PaM*-7, very close to microsatellite markers AGS.20, PGS3.23 and PGS3.62 previously shown to be fully linked to the PPM (Zuriaga et al., 2012 and 2013). *PaM*-7 was fully sequenced for *M*/*m*-alleles and the 358-bp insertion was found to putatively lead to a premature stop-codon in the predicted protein lacking 4 out of the 6 exons. Furthermore, *PaM*-7 was one of the four genes differentially over-expressed in anthers in agreement with the tissue-specific expression expected for the *M*-locus mutated modifier gene. This insertion was characterized as an active non-autonomous mutator (transposable) element [named *FallingStone* (*FaSt*)] containing structural features that have been proved to be also present in the *PaM*-7 insertion (Halász et al. 2014).

PaM-7 codes for an oxidoreductase that contains a Thioredoxin fold domain (IPR012336). Proteins having this domain form a large and diverse protein superfamily characterized by a CXXC motif (two cysteines separated by 2 amino acids), which confers the thiol-disulfide redox activity essential for folding, stability and function in target proteins (Hogg, 2003; Schmidt et al., 2006). Proteins containing this domain have been associated with a wide range of events during sexual plant reproduction, from

gametophyte formation to seed setting (either for their redox activity or as signaling factors) specially under the control of thioredoxin (TRXs) and glutaredoxin (GRXs) proteins (Traverso et al., 2013). However, *PaM-7* does not code neither for a TRX type *h* nor for other TRX type but for a protein containing a Disulfide bond A-like (DsbA-like) domain (IPR001853; PF01323). DsbA-like proteins were firstly identified in *Escherichia coli* as disulfide bond introducers in the periplasm, a necessary process for protein folding (Depuydt et al., 2011). Therefore, DsbA-like proteins are not usually reducing enzymes such as TRX proteins but oxidizing. However, proteins of the TRX superfamily are intrinsically bidirectional, thus can catalyze either oxidation or reduction depending on the redox states in which they are maintained (Ito & Inaba, 2008). Accordingly, *PaM-7* was renamed as *Prunus armeniaca M*-locus Disulfide bond A-like Oxidoreductase (PaMDOr), which dycisteinic motif CPWC is located at the protein N-terminal end (Cys19-PW-Cys22). Overall orthologue study supports a divergent evolution for *M*-locus DsbA proteins in the Rosaceae family. However, putative paralogous ($CPWC_2$ and $CPWC_1$) arose from gene duplication in tandem, being the function of the $CPWC_2$ type proteins specifically related to GSI. In this sense, $CPWC_1$/$CPWC_2$ divergence process might shed some light of *Malus/Prunus* GSI evolution as well.

Comparative study of the GSI system in Rosaceae and Solanaceae by analyzing orthology relationships for modifier factors

Gametophytic Self-Incompatibility in Rosaceae and Solanaceae families involves an allele-specific recognition mechanism between the same *S*-factor types, which are essential to the (in)compatibility response. Similarly, non-*S*-factors required for this mechanism have demonstrated to be functionally equivalents among species of both families (for instance, components of SCF^{SLF} complex) and descend from a common ancestor (Huang & Kao 2006; Matsumoto & Tao 2012). Therefore, it would not be surprising that other modifiers descending from common ancestors may have persisted throughout the evolution, including the speciation process, in both families.

First hits obtained in direct BLASTP analysis for NaTrxh, MdABCF and SBP1 demonstrated to be RBHs to each other in almost all performed comparisons. Thus, on the basis of the RBH analyses, ppa011576m, ppa002137m and ppa008290m accessions were proposed as putative orthologues for NaTrxh, MdABCF and SBP1 modifiers in *Prunus*, respectively. Comparative structural genomics highlighted the existence of

orthologous segments among *Prunus*, *Solanum* and *Arabidopsis* for the three potential orthologues. It is noteworthy that *P. persica* and *S. lycopersicum* (sharing *S*-RNase-based GSI system) have a higher degree of synteny (based on the number of anchor points) in comparison with *A. thaliana* (SSI outgroup system) despite Rosaceae and Brassicaceae are phylogenetically closer than Rosaceae and Solanaceae (Igic and Kohn 2001). Lastly, these modifiers belong to distinct and large protein families which have been widely studied and classified. This information was used accordingly for phylogenetic tree clustering, where results also supported these *P. persica* accessions as orthologues for NaTrxh, MdABCF and SBP1.

On the contrary, no *P. persica* predicted peptide was found to fulfill the requirements established in this work for being considered orthologous to 120K, NaStEP and NaPCCP modifiers, likely due to a complex evolutionary pattern. Notwithstanding, relationships between ppa021281m and ppa011448m with 120K and NaStEP, respectively, cannot be discarded.

As a whole, a divergent process from an ancestral mechanism leading to different mechanisms where gene duplications and losses might have occurred, or a convergent process where proteins from different lineages (or even families) have been recruited to control the SI mechanism in a similar manner, might be plausible. Nevertheless, compiled evidences from this work are not robust enough to support a particular model. In any case, this work was primarily aimed to lay the foundation for the identification in *Prunus* of those genes accomplishing the function of the Solanaceae GSI modifiers. Further molecular and sequence analyses will be needed to achieve this goal.

CONCLUSIONS

- The self-compatible apricot cv. 'Katy' was molecular and genetically characterized. An *S*-locus unlinked pollen-part mutation (PPM) was found to be responsible of this phenotype. Fine-mapping located this mutation at the distal end of chr.3 within a region overlapping with that corresponding to the *M*-locus genetic map previously constructed for the self-compatible apricot cv. 'Canino'.

- *S*-genotyping of a set with 67 apricot cultivar/accessions allowed us to identify three new *S*-alleles and two putatively new mutations conferring self-compatibility (SC). Both mutations (associated with a SNP and an Indel) affect the male *S*-determinant *SFB* as reported for most of the non-functional *S*-haplotypes in *Prunus*.

- *M*-genotyping showed that the same mutated *m*-haplotype was shared by 'Canino' and 'Katy' but also by 17 additional cultivars. Genetic analysis of two of these self-compatible cultivars, 'Portici' and 'Corbató', confirmed our results. The *m*-haplotype was only found in North-American and Western-European cultivars. Haplotype distance analysis points out the widely distributed $M_{1\text{-}0}$ as the putative ancestor suggesting that *m*-haplotype arose much later in time than S_C-allele.

- A strategy based on genomic and transcriptomic NGS data allowed us to narrow down the apricot *M*-locus region leading to a physical map of ~134 Kb. Comparative screening of non-synonymous polymorphisms in this region led to identify a 358-bp *FaSt* insertion type, segregating in coupling with the *m*-haplotype in self-compatible apricots, as the unique polymorphism fulfilling genetic requirements for the PPM conferring SC.

- According to gene annotation, the *m*-haplotype *FaSt* insertion is located in the third exon of the *PaM*-7 gene that putatively encodes a Disulfide bond A-like Oxidoreductase (named as *PaMDOr*). *FaSt* insertion is predicted to lead to a premature stop-codon producing a truncated protein lacking the 3´end. *PaMDOr* is differentially over-expressed in mature anthers with regards to leaves and styles. Altogether, evidences suggest *PaMDOr* as the pollen-part mutated modifier conferring SC in apricot.

- Phylogenetic analysis suggest *PaMDOr* as a putative paralogue (of *PaM*-8) emerged after the split of the Rosaceae and Solanaceae (shared by *Prunus and Fragaria* but later lost in *Malus*) which function became essential for the proper functioning of the GSI system in *Prunus*.

- The analysis of orthology relationships between GSI modifier factors in Solanaceae and Rosaceae allowed the identification of putative orthologues for *NaTrxh*, *SBP1* and *MdABCF* in *Prunus*. On the contrary, a more complex evolutionary pattern was found for *120K*, *NaStEP* and *NaPCCP*. Overall, results allow thinking that at least part of the GSI regulating factors might be shared by both families.

REFERENCES

Adler PN, Holt CE (1975). Mating type and the differentiated state in Physarum polycephalum. Developmental Biology, 43(2), 240–253.

Aguiar B, Vieira J, Cunha AE, Fonseca NA, Iezzoni A, Van Nocker S, Vieira CP (2015). Convergent evolution at the gametophytic self-incompatibility system in Malus and Prunus. PLoS ONE, 10(5).

Ai Y, Kron E, Kao TH (1991) S-alleles are retained and expressed in a self-compatible cultivar of Petunia hybrida. Mol Gen Genet 230: 353-358.

Akagi T, Henry IM, Morimoto T, Tao R. (2016). Insights into the Prunus-Specific S-RNase-Based Self-Incompatibility System from a Genome-Wide Analysis of the Evolutionary Radiation of S Locus-Related F-box Genes. Plant Cell Physiol. doi:10.1093/pcp/pcw077

Alburquerque N, Egea J, Pérez-Tornero O, Burgos L (2002) Genotyping apricot cultivars for self-(in)compatibility by means of RNases associated with S alleles. Plant Breed 121: 343-347.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Anderson, M. A., Cornish, E. C., Mau, S.-L., Williams, E. G., Hoggart, R., Atkinson, A., Clarke, A. E. (1986). Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in Nicotiana alata. Nature, 321(6065), 38–44.

Andrés MV and Durán JM (1998) self-incompatibility in Spanish clones of apricot (Prunus armeniaca L.) tree. Euphytica 101: 349-355

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.

Ashkani, J., & Rees, D. J. G. (2015). A Comprehensive Study of Molecular Evolution at the Self-Incompatibility Locus of Rosaceae. J Mol Evol. 82 (2):128-145.

Astiz, V., Iriarte, L. A., Flemmer, A., Hernández, L. F. (2011). Self-compatibility in modern hybrids of sunflower (Helianthus annuus L.) fruit set in open and self-pollinated (bag isolated) plants grown in two different locations. Helia, 34(54), 129–138.

Avila-Castañeda, A., Juárez-Díaz, J. A., Rodríguez-Sotres, R., Bravo-Alberto, C. E., Ibarra-Sánchez, C. P., Zavala-Castillo, A., Cruz-García, F. (2014). A novel motif in the NaTrxh N-terminus promotes its secretion, whereas the C-terminus participates in its interaction with S-RNase in vitro. BMC Plant Biol, 14, 147.

Badenes ML, Llácer G, Asins MJ, Guerri J, García S (1993) Caracterización pomológica de variedades y clones de albaricoquero. Invest Agr: Prod Prot Veg 8: 55-65

Badenes ML, Martínez-Cavo J, García-Carbonell S, Villarrubia D, Llácer G (1997) Descripción de variedades autóctonas valencianas de albaricoquero. Generalitat Valenciana (Consellería de Agricultura, Pesca y alimentación) Series Divulgación Técnica pp: 1-61

Barrett, S. C. H. (2002). The evolution of plant sexual diversity. Nature Reviews. Genetics, 3(4), 274–284.

Bedinger PA, Chetelat RT, McClure BA, Moyle LC, Rose JKC, Stack SM, van der Knaap E, Baek YS, Lopez-Casado G, Covey PA, Kumar A, Li W, Nunez R, Cruz-García F and Royer S (2011) Insterspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation. Sex plant Reprod 24: 171-187

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France).

Beppu, K., Kumai, M., Yamane, H., Tao, R., Kataoka, I., (2012). Molecular and genetic analyses of the s haplotype of the self-compatible Japanese plum (Prunus salicina). J. Hort. Sci. Biotechnol 87, 493–498.

Bingham J and Sudarsanam S (2000) Visualizing large hierarchical clusters in hyperbolic space. Bioinformatics. 16(7): 660-661

Bolger, A. et al. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat. Genet. 46, 1034–1038 (2014).

Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller L, Martin G: A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Mol Plant Microbe Interact 2012, 25:1523-1530.

Bonfield J (2004) Staden package, version 1.4. URL http://staden.sourceforge.net Accessed 27 July 2011.

Bosch, M., & Franklin-Tong, V. E. (2007). Temporal and spatial activation of caspase-like enzymes induced by self-incompatibility in Papaver pollen. Proc Natl Acad Sci U S A, 104(46), 18327–32.

Bosch, M., & Franklin-Tong, V. E. (2008). Self-incompatibility in Papaver: signalling to trigger PCD in incompatible pollen. J Exp Bot, 59(3), 481–90.

Boskovic R, Sargent DJ, Tobutt KR (2010) Genetic evidence that two independent S-loci control RNase-based self-incompatibility in diploid strawberry. J Exp Bot 61: 755-763.

Boskovic, R., Tobutt, K. R., Mailing, W., & Me, K. (1996). Correlation of stylar ribonuclease zymograms with incompatibility sweet cherry alleles in. Euphytica, 90(1939), 245–250.

Bourguiba H, Audergon JM, Krichen L, Trifi-Farah N, Mamouni A, Trabelsi S, D´Onofrio C, Asma BM, Santoni S and Khadari B (2012) Loss of genetic diversity as a signature of apricot domestication and diffusion into de Mediterranean Basin. BMC Plant Biol 12: 49-64

Brooks RM and Olmo HP (1997) The Brooks and Olmo Register of Fruit and Nut Varieties. ASHS Press 3rd ed., Alexandria, VA.

Broothaerts, W., Keulemans, J., & Van Nerum, I. (2004). Self-fertile apple resulting from S-RNase gene silencing. Plant Cell Reports, 22(7), 497–501.

Bruvo R, Michiels NK, D'Souza TG and Schulenburg H (2004) A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. Mol Ecol. 13: 2101-2106

Buntjer JB (1997) Phylogenetic computer tools (PhylTools). Version 1.32 for Windows. Wageningen: Laboratory of Plant Breeding, Wageningen University

Burgos L, Alburquerque N and Egea J (2004) Review. Flower biology in apricot and its implications for breeding. Span J Agric Res 2(2): 227-241

Burgos L, Pérez-Tornero O, Ballester J, Olmos E (1998) Detection and inheritance of stylar ribonucleases associated with incompatibility alleles in apricot. Sex Plant Reprod 11: 153-158.

Busot GY, McClure B, Ibarra-Sánchez CP, Jiménez-Durán K, Vázquez-Santana S, et al. (2008) Pollination in Nicotiana alata stimulates synthesis and transfer to the stigmatic surface of NaStEP, a vacuolar Kunitz proteinase inhibitor homologue. J Exp Bot 59: 3187-3201.

Busot, G. Y., McClure, B., Ibarra-Sánchez, C. P., Jiménez-Durán, K., Vázquez-Santana, S., & Cruz-García, F. (2008). Pollination in Nicotiana alata stimulates synthesis and transfer to the stigmatic surface of NaStEP, a vacuolar Kunitz proteinase inhibitor homologue. J Exp Bot, 59(11), 3187–201.

Cabrillac D, Cock JM, Dumas C, Gaude T (2001). The *S*-locus receptor kinase is inhibited by thioredoxins and activated by pollen coat proteins. Nature 410: 220–223

Cachi AM, Wünsch A (2011) Characterization and mapping of non-S gametophytic self-compatibility in sweet cherry (Prunus avium L.). J Exp Bot 62: 1847-1856.

Cachi, A.M., Wünsch, A., 2014. Characterization of self-compatibility in sweet cherry varieties by crossing experiments and molecular genetic analysis. Tree Genet. Genomes 10, 1205–1212.

Chen G, Zhang B, Zhao Z, Sui Z, Zhang H, et al. (2010) 'A life or death decision' for pollen tubes in S-RNase-based self-incompatibility. J Exp Bot 61: 2027-2037

Cheung, A. Y., May, B., Kawata, E. E., Gu, Q., & Wu, H. M. (1993). Characterization of cDNAs for stylar transmitting tissue-specific proline-rich proteins in tobacco. Plant J, 3(1), 151–60.

Cheung, A.Y., Wang, H. and Wu, H. (1995). A floral transmitting tissue-specific glycoprotein attracts pollen tubes and stimulates their growth. Cell, 82, 383–393.

Cho, E. J., Yuen, C. Y. L., Kang, B.-H., Ondzighi, C. A., Staehelin, L. A., & Christopher, D. A. (2011). Protein disulfide isomerase-2 of Arabidopsis mediates protein folding and localizes to both the secretory pathway and nucleus, where it interacts with maternal effect embryo arrest factor," Molecules and Cells, vol. 32, no. 5, pp. 459–475, 2011.

Compton, R. H. (1913). Phenomena and problems of self-sterility. New Phytol, 12(6), 197–206.

Consortium, T. G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 485(7400), 635–41.

Cope FW. (1962). The mechanism of pollen incompatibility in Theobroma cacao. Heredity, 17, 157–182.

Crane MB and Lewis D. (1942). Genetical studies in pears. III Incompatibility and sterility. J Genet, 43, 31–49.

Crane MB, AG Brown (1937). Incompatibility and sterility in the sweet cherry, *Prunus avium* L. J. Pomol. Hort. Sci. 15:86-116

Cruz-Garcia, F., Hancock, C. N., & McClure, B. (2003). S-RNase complexes and pollen rejection. J Exp Bot, 54(380), 123–130.

Cruz-Garcia, F., Nathan Hancock, C., Kim, D., & McClure, B. (2005). Stylar glycoproteins bind to S-RNase in vitro. Plant J, 42(3), 295–304.

Darwin, C. (1868). The Variation of Animals and Plants Under Domestication, Volumen 2.

Darwin, C. (1876). The Effects of Cross and Self Fertilisation in the Vegetable Kingdom.

Darwin CR. 1878. The effects of cross and self-fertilisation in the vegetable kingdom, 2nd edn. London: John Murry.

De Franceschi, P., Pierantoni, L., Dondini, L., Grandi, M., Sansavini, S., & Sanzol, J. (2011). Evaluation of candidate F-box genes for the pollen S of gametophytic self-incompatibility in the Pyrinae (Rosaceae) on the basis of their phylogenomic context. Tree Genet Genomes, 7(4), 663–683.

de Graaf, B. H. J., Knuiman, B. A., Derksen, J., & Mariani, C. (2003). Characterization and localization of the transmitting tissue-specific PELPIII proteins of Nicotiana tabacum. J Exp Bot, 54(380), 55–63.

de Graaf, B. H. J., Rudd, J. J., Wheeler, M. J., Perry, R. M., Bell, E. M., Osman, K., Franklin-Tong, V. E. (2006). Self-incompatibility in *Papaver* targets soluble inorganic pyrophosphatases in pollen. Nature, 444(7118), 490–3.

de Graaf BH, Vatovec S, Juárez-Díaz JA, Chai L, Kooblall K, Wilkins KA, Zou H, Forbes T, Franklin FC, Franklin-Tong VE. (2012). The *Papaver* self-incompatibility pollen *S*-determinant, PrpS, functions in *Arabidopsis thaliana*. Current Biology 22, 154–159.

De Nettancourt D (2001) Incompatibility and incongruity in wild and cultivated plants. Springer-Verlag, Berlin

Dee, J. (1966). Multiple Alleles and Other Factors Affecting Plasmodium Formation in the True Slime Mold Physarum polycephalum Schw *. The Journal of Protozoology, 13(4), 610–616.

Della Strada G, Pennone F, Fideghelli C, Monastra F and Cobianchi D (1989) Monografía di cultivar di albicocco. Ministerio dell´Agricoltura e delle Foreste. Direzione Generale della Produzione Agricola. 239 pp.

Depuydt M, Messens J, Collet JF (2010). How proteins form disulfide bonds. Antioxid Redox Signal

Distefano, G., Las Casas, G., La Malfa, S., Gentile, A., Tribulato, E., (2009). Pollen tube behavior in different mandarin hybrids. J. Am. Soc. Hortic. Sci. 134, 583–588.

Doyle JJ, Doyle JL (1987). A rapid isolation procedure for small quantities of fresh leaf tissue. Phyto Bull 19: 11-15.

East, E. M. (1908). Inbreeding in corn. Rep. Conn. Agric. Exp. Stn, 1907, 419–428.

East, E. M. (1932). Studies on Self-Sterility. IX. the Behavior of Crosses between Self-Sterile and Self-Fertile Plants. Genetics, 17(2), 175–202.

East, E. M., & Mangelsdorf, A. J. (1925). A New Interpretation of the Hereditary Behavior of Self-Sterile Plants. Proc Natl Acad Sci U S A, 11(2), 166–71.

East, E. M., & Park, J. B. (1917). Studies on Self-Sterility I. the Behavior of Self-Sterile Plants. Genetics, 2(6), 505–609.

East, E. M., & Yarnell, S. H. (1929). Studies on Self-Sterility. VIII. Self-Sterility Allelomorphs. Genetics, 14(5), 455–87.

Eberle, C. A., Anderson, N. O., Clasen, B. M., Hegeman, A. D., & Smith, A. G. (2013). PELPIII: the class III pistil-specific extensin-like Nicotiana tabacum proteins are essential for interspecific incompatibility. Plant J, 74(5), 805–14.

Egea J, Burgos L (1996) Detecting cross-incompatibility of three North American apricot cultivars and establishing the first incompatibility group in apricot. J Am Soc Hort Sci 12: 1002-1005.

Entani, T., Iwano, M., Shiba, H., Che, F.-S., Isogai, A., & Takayama, S. (2003). Comparative analysis of the self-incompatibility (S-) locus region of Prunus mume: identification of a pollen-expressed F-box gene with allelic diversity. Genes to Cells, 8(3), 203–213.

Entani, T., Kubo, K., Isogai, S., Fukao, Y., Shirakawa, M., Isogai, A., & Takayama, S. (2014). Ubiquitin-proteasome-mediated degradation of S-RNase in a solanaceous cross-compatibility reaction. Plant J, 78(6), 1014–21.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368-376.

Feng J, Chen X, Wu Y, Liu W, Liang Q, et al. (2006) Detection and transcript expression of S-RNase gene associated with self-incompatibility in apricot (*Prunus armeniaca* L.). Mol Biol Rep 33: 215-221.

Fernández A, Hanada T, Alonso JM, Yamane H, Tao R, et al. (2009) A modifier locus affecting the expression of the S-RNase gene could be the cause of breakdown of self-incompatibility in almond. Sex Plant Reprod 22: 179-186.

Fisher, R.A. (1941). Average excess and average effect of a gene substitution. Annals of Eugenics, 11(1), 53–63.

Foote, H. C., Ride, J. P., Franklin-Tong, V. E., Walker, E. A., Lawrence, M. J., & Franklin, F. C. (1994). Cloning and expression of a distinctive class of self-incompatibility (S) gene from Papaver rhoeas L. Proc Natl Acad Sci U S A, 91(6), 2265–9.

Franklin-Tong V, Franklin F and de Graaf (2010) Engineering of plants to exhibit self-incompatibility. Patent WO 2010/061181 A1. World Intellectual Property Organization.

Franklin-Tong, V. E., Hackett, G., & Hepler, P. K. (1997). Ratio-imaging of Ca2+i in the self-incompatibility response in pollen tubes of Papaver rhoeas. Plant J, 12(6), 1375–1386.

Franklin-Tong, V. E., Ride, J. P., & Franklin, F. C. H. (1995). Recombinant stigmatic self-incompatibility (S-) protein elicits a Ca2+ transient in pollen of Papaver rhoeas. Plant J, 8(2), 299–307.

Franklin-Tong, V. E., Ride, J. P., Read, N. D., Trewavas, A. J., & Franklin, F. C. H. (1993). The self-incompatibility response in Papaver rhoeas is mediated by cytosolic free calcium. Plant J, 4(1), 163–177.

Gabaldon, T. and E. V. Koonin (2013). Functional and evolutionary implications of gene orthology. Nat Rev Genet 14: 360-366.

Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. Genome Res. 8, 779–790.

Gambetta, G., Gravina, A., Fasiolo, C., Fornero, C., Galiger, S., Inzaurralde, C., & Rey, F. (2013). Self-incompatibility, parthenocarpy and reduction of seed presence in "Afourer" mandarin. Scientia Hort, 164, 183–188.

Geitmann, A., Snowman, B. N., Emons, A. M., & Franklin-Tong, V. E. (2000). Alterations in the actin cytoskeleton of pollen tubes are induced by the self-incompatibility reaction in *Papaver rhoeas*. Plant cell, 12(7), 1239–51.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5(10):R80

Goldraij A, Kondo K, Lee CB, Hancock CN, Sivaguru M, et al. (2006) Compartmentalization of S-RNase and HT-B degradation in self-incompatible Nicotiana. Nature 439: 805-810.

Goldway M, Sapir G and Stern A (2007) Molecular basis and horticultural application of the gametophytic self-incompatibility system in Rosaceous tree fruits. In: Plant Breeding Reviews 28. pp 215-237

Goldway M, Sapir G and Stern A (2007). Molecular basis and horticultural application of the gametophytic self-incompatibility system in Rosaceous tree fruits. In: Plant Breeding Reviews 28. pp 215-237.

Golz JF, Oh HY, Su V, Kusaba M, Newbigin E (2001) Genetic analysis of Nicotiana pollen-part mutants is consistent with the presence of an S-ribonuclease inhibitor at the S locus. Proc Natl Acad Sci USA 98: 15372-15376.

Golz JF, Su V, Clarke AE, Newbigin E (1999) A molecular description of mutations affecting the pollen counterpart of the Nicotiana alata S locus. Genetics 152: 1123-1135.

Grauschopf U, Winther JR, Korber P, Zander T, Dallinger P, and Bardwell JC. Why is DsbA such an oxidizing disulfide catalyst? Cell 83: 947–955, 1995.

Grattapaglia D, Sederoff RR (1994) Genetic linkage maps of Eucalyptus grandis and E. urophylla using a pseudotest-cross strategy and RAPD markers. Genetics 137: 1121-1137.

Gu C, Wu J, Du YH, Yang YN and Zhang SL (2013) Two different *Prunus* SFB alleles have the same function in the self-incompatibility reaction. Plant Mol Biol Rep. 31:425–434

Gu, T., Mazzurco, M., Sulaman, W., Matias, D. D., & Goring, D. R. (1998). Binding of an arm repeat protein to the kinase domain of the S-locus receptor kinase. Proc Natl Acad Sci USA, 95(1), 382–7.

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F and Petit RJ (2011) Current trends in microsatellite genotyping. Mol Ecol Res. 11: 591-61

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, and Regev A (2013) De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 8(8): 10.1038/nprot.2013.084.

Haffani, Y. Z., Gaude, T., Cock, J. M., & Goring, D. R. (2004). Antisense suppression of thioredoxin h mRNA in Brassica napus cv. Westar pistils causes a low level constitutive pollen rejection response. Plant Mol Biol, 55(5), 619–30.

Hajilou J, Grigorian V, Mohammadi SA, Nazemmieh A, Romero C, Vilanova S and Burgos L (2006) Self- and cross-(in)compatibility between important apricot cultivars in northwest Iran. J Hort Sci Biotech. 81(3): 513-517

Halász J, Hegedüs A, Hermán R, Stefanovits-Bányai E and Pedryc A (2005) New self-incompatibility alleles in apricot (Prunus armeniaca L.) revealed by stylar ribonuclease assay and S-PCR analysis. Euphytica 145: 57-66

Halász J, Pedryc A and Hegedüs A (2007) Origin and dissemination of the pollen-part mutated SC haplotype which confers self-compatibility in apricot (Prunus armeniaca) New Phytol. 176: 792-803

Halász J, Pedryc A, Ercisli S, Yilmaz KU and Hegeddus A (2010) S-genotyping supports the genetic relationships between Turkish and Hungarian apricot germplasm. J Amer Soc Hort Sci. 135(5):410–417

Halász, J., Kodad, O., & Hegedüs, A. (2014). Identification of a recently active Prunus-specific non-autonomous Mutator element with considerable genome shaping force. Plant J, 79(2), 220–231.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment program for Windows 95/98/NT. Nucleic Acids Symposium Series, 41, 95–98.

Hancock CN, Kent L, McClure BA (2005) The 120K glycoprotein is required for S-specific pollen rejection in Nicotiana. Plant J 43: 716-723.

Hancock CN, Kondo K, Beecher B, McClure BA (2003) The S-locus and unilateral incompatibility. Philos Trans R Soc Lond B Biol Sci 358:1133–1140

Harada, Y., Takagaki, Y., Sunagawa, M., Saito, T., Yamada, L., Taniguchi, H., Sawada, H. (2008). Mechanism of self-sterility in a hermaphroditic chordate. Science, 320(5875), 548–50.

Hauck NR, Yamane H, Tao R, Iezzoni AF (2006) Accumulation of nonfunctional S-haplotypes results in the breakdown of gametophytic selfincompatibility in tetraploid Prunus. Genetics 172: 1191–1198.

Haussmann BIG, Parzies HK, Presterl T, Suši Z and Miedaner T (2004) Plant genetic resources in crop improvement. Plant Genet Res. 2(1): 3-21

Hegedüs A, Lénárt J and Halász J (2012) Sexual incompatibility in Rosaceae fruit tree species: molecular interactions and evolutionary dynamics. Biol Plant. 56(2): 201-209

Heslop-Harrison, J. (1982). Pollen-stigma interaction and cross-incompatibility in the grasses. Science, 215(4538), 1358–64.

Hiscock SJ and Dickinson HG (1993) Unilateral incompatibility within the Brassicaceae further evidence for the involvement of the self-incompatibility (S) locus. Theor. Appl. Genet. 86: 744-753.

Hogg PJ, (2003). Disulfide bonds as switches for protein function. Trends Biochem Sci 28: 210-214.

Hosaka K, Hanneman RE (1998a). Genetics of self-compatibility in a selfcompatible wild diploid potato species Solanum chacoense.1. Detection of an S-locus inhibitor (Sli) gene. Euphytica 99: 191-197.

Hosaka K, Hanneman RE (1998b). Genetics of self-compatibility in a selfcompatible wild diploid potato species Solanum chacoense. 2. Localization of an S-locus inhibitor (Sli) gene on the potato genome using DNA markers. Euphytica 103: 265-271.

Hua Z, Kao T-h (2006) Identification and characterization of components of a putative Petunia S-locus F-box-containing E3 ligase complex involved in S-RNase based self-incompatibility. Plant Cell 18: 2531-2553.

Huang J, Zhao L, Yang Q, Xue Y (2006) AhSSK1, a novel SKP1-like protein that interacts with the S-locus F-box protein SLF. Plant J 46: 780-793.

Igic B, Kohn JR (2001) Evolutionary relationships among selfincompatibility RNases. Proc Natl Acad Sci USA 98: 13167–13171

Ikeda, K., Igic, B., Ushijima, K., Yamane, H., Hauck, N. R., Nakano, R., Tao, R. (2004). Primary structural features of the S haplotype-specific F-box protein, SFB, in Prunus. Sex Plant Reprod , 16(5), 235–243.

Illa E, Sargent DJ, Lopez Girona E, Bushakra J, Cestaro A, Crowhurst R, Pindo M, Cabrera A, Van der Knapp E, Iezzoni A, Gardiner S, Velasco R, Arús P, Chagné D, Troggio M (2011) Comparative analysis of rosaceous genomes and the reconstruction of a putative ancestral genome for the family. BMC Evol Biol 11:9

Ioerger, T. R., Gohlke, J. R., Xu, B., & Kao, T.-H. (1991). Primary structural features of the self-incompatibility protein in solanaceae. Sex Plant Reprod , 4(2).

Ito K and Inaba K. (2008). The disulfide bond formation (Dsb) system. Curr Opin Struct Biol 18: 450–458.

Iwano M, Takayama S (2012). Self/non-self discrimination in angiosperm selfincompatibility. Curr Opin Plant Biol, 15:78-83.

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Societé Vaudoise des Sciences Naturelles 37: 547-579

Janssens G, Broekaert W, van Nerum L, Broothaerts W, Dreesen R, D. M. and K. J. (2006). Self-fertile apple resulting from S-RNase gene silencing. US Patent No 20060123514.

Jie Q, Shupeng G, Jixiang Z, Manru G and Huairui S (2005)Identification of self-incompatibility genotypes of apricot (Prunus armeniaca L.) by S-allele-specific PCR analysis. Biotechnol Lett. 27(16): 1205-1209

Jimenez-Duran, K., McClure, B., Garcia-Campusano, F., Rodriguez-Sotres, R., Cisneros, J., Busot, G., & Cruz-Garcia, F. (2013). NaStEP: A Proteinase Inhibitor Essential to Self-Incompatibility and a Positive Regulator of HT-B Stability in *Nicotiana alata* Pollen Tubes. Plant Physiol, 161(1), 97–107.

Juárez-Díaz,J.A., McClure,B., Vázquez-Santana,S., Guevara-García,A., León-Mejía,P., Márquez-Guzmán,J., et al. (2006). A novel thioredoxin h is secreted in *Nicotiana alata* and reduces S-RNase *in vitro*. J. Biol.Chem. 281, 3418–3424.

Jung S, Jesudurai C, Staton M, Du Z, Ficklin S, et al. (2004) GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. BMC Bioinformatics 5: 130.

Jung, H. J., Jung, H. J., Ahmed, N. U., Park, J. I., Kang, K. K., Hur, Y., Nou, I. S. (2012). Development of Self-Compatible B. rapa by RNAi-Mediated S Locus Gene Silencing. PLoS ONE, 7(11).

Jung, S., Cestaro, A., Troggio, M., Main, D., Zheng, P., Cho, I., Sargent, D. J. (2012). Whole genome comparisons of Fragaria, Prunus and Malus reveal different modes of evolution between Rosaceous subfamilies. BMC Genomics, 13(1), 129.

Kakita, M., Murase, K., Iwano, M., Matsumoto, T., Watanabe, M., Shiba, H., Takayama, S. (2007). Two distinct forms of M-locus protein kinase localize to the plasma membrane and interact directly with S-locus receptor kinase to transduce self-incompatibility signaling in Brassica rapa. Plant Cell, 19(12), 3961–73.

Kakui H, Kato M, Ushijima K, Kitaguchi M, Kato S, et al. (2011) Sequence divergence and loss-of-function phenotypes of S locus F-box brothers genes are consistent with non-self recognition by multiple pollen determinants in self-incompatibility of Japanese pear (Pyrus pyrifolia). Plant J 68: 1028-1038.

Kamauchi, S., Wadahama, H., Iwasaki, K., Nakamoto, Y., Nishizawa, K., Ishimoto, M.,Urade, R. (2008). Molecular cloning and characterization of two soybean protein disulfide isomerases as molecular chaperones for seed storage proteins. The FEBS Journal, 275(10), 2644–58.

Kaothien-Nakayama P, Isogai A and Takayama S (2010) Self-incompatibility systems in flowering plants. In: Plant Developmental Biology – Biotechnological Perspectives Vol I. (Eds. Pua E-C and Davey MR). Springer-Verlag, Berlin, Heidelberg.

Kempe, K., & Gils, M. (2011). Pollination control technologies for hybrid breeding. Mol. Breeding 27: 417-437.

Kitashiba H, Liu P, Nishio T, Nasrallah JB, Nasrallah ME (2011). Functional test of *Brassica* self-incompatibility modifiers in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 2011, 108:18173-18178.

Kitashiba, H., & Nasrallah, J. B. (2014). Self-incompatibility in Brassicaceae crops: lessons for interspecific incompatibility. Breeding Science, 64(1), 23–37.

Kodad O, Hegedűs A, Socias i Company R and Halász J (2013) Self-(in)compatibility genotypes of Moroccan apricots indicate differences and similarities in the crop history of European and North African apricot germplasm. BMC Plant Biol. 13:196

Kong, et al., (2007). Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth, Plant J. 50: 873–885.

Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12: 172-175.

Krzywinski M, et al. (2009). Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–1645.

Kubo K, Entani T, Tanaka A, Wang N, Fields AM, et al. (2010) Collaborative non-self recognition system in S-RNase-based self-incompatibility. Science 330: 796-799

Lai Z, Ma W, Han B, Liang L, Zhang Y, et al. (2002) An F-box gene linked to the self-incompatibility (S) locus of Antirrhinum is expressed specifically in pollen and tapetum. Plant Mol Biol 50: 29-42.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3):R25.

Lee C.B, Kim S, & McClure B. (2009). A pollen protein, NaPCCP, that binds pistil arabinogalactan proteins also binds phosphatidylinositol 3-phosphate and associates with the pollen tube endomembrane system. Plant Physiol, 149(2)

Lee CB, Swatek KN, McClure BA (2008). Pollen proteins interact with the C-terminal domain of *Nicotiana alata* pistil arabinogalactan proteins. J Biol Chem 283: 26965–26973

Lewis D and Crowe LK. (1954). The induction of self-fertility in tree fruits. J Hort Sci, 29, 220–225.

Lewis, D., & Crowe, L. K. (1958). Unilateral interspecific incompatibility in flowering plants. Heredity, 12(2), 233–256.

Li B, Dewey CN. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12:323

Li S, Samaj J, Franklin-Tong VE. (2007). A mitogen-activated protein kinase signals to programmed cell death induced by self-incompatibility in Papaver pollen. Plant Physiol, 145(1), 236–45.

Li, S., Sun, P., Williams, J. S., & Kao, T. (2014). Identification of the self-incompatibility locus F-box protein-containing complex in Petunia inflata. Plant reproduction (Vol. 27).

Li, W., & Chetelat, R. T. (2010). A pollen factor linking inter- and intraspecific pollen rejection in tomato. Science, 330(6012), 1827–30.

Li, W., & Chetelat, R. T. (2014). The role of a pollen-expressed Cullin1 protein in gametophytic self-incompatibility in Solanum. Genetics, 196(2), 439–442.

Li, X., Nield, J., Hayman, D., & Langridge, P. (1996). A self-fertile mutant of Phalaris produces an S protein with reduced thioredoxin activity. Plant J, 10(3), 505–513.

Lin, Z., Eaves, D. J., Sanchez-Moran, E., Franklin, F. C. H., & Franklin-Tong, V. E. (2015). The *Papaver rhoeas S* determinants confer self-incompatibility to *Arabidopsis thaliana* in planta. Science, 350(6261), 684–7.

Liu K and Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21(9): 2128-2129

Liu W, Fan J, Li J, Song Y, Li Q, Zhang Y et al (2014) SCF^{SLF}-mediated cytosolic degradation of S-RNase is required for cross pollen compatibility in S-RNase-based self-incompatibility in *Petunia hybrida*. Front Genet 5:228

Lundqvist, A. (2010). Studies on self-sterility in rye, Secale cereale L. Hereditas, 40(3-4), 278–294.

Luu, D. T., Qin, X., Morse, D., & Cappadocia, M. (2000). S-RNase uptake by compatible pollen tubes in gametophytic self-incompatibility. Nature, 407(6804), 649–51.

Manzanares C, Barth S, Thorogood D, Byrne SL, Yates S, Czaban A, Studer B. (2016). A Gene Encoding a DUF247 Domain Protein Cosegregates with the S Self-Incompatibility Locus in Perennial Ryegrass. Mol Biol Evol, 33(4), 870–84.

Marchese A, Boskovic R, Caruso T, Raimondo A, Cutuli M, et al. (2007) A new self-incompatibility haplotype in sweet cherry 'Kronio', S5´ attributable to a pollen-part mutation in the SFB gene. J Exp Bot 58: 4347-4356.

Martin, F. W. (1967). The Genetic Control of Unilateral Incompatibility between Two Tomato Species. Genetics, 56(3), 391–8.

Massai R (2010) Variability of apricot cultivars traits inside the 'List of recommended fruit varieties' project. Acta Hort. 862: 129-136

Matsumoto D and Tao R (2016). Recognition of a wide-range of S-RNases by S locus F-box like 2, a general-inhibitor candidate in the Prunus-specific S-RNase-based self-incompatibility system. Plant Mol Biol.

Matsumoto D, Yamane H, Abe K, Tao R (2012) Identification of a Skp1-like protein interacting with SFB, the pollen S determinant of the gametophytic self-incompatibility in Prunus. Plant Physiol 159: 1252-1262.

Mazzurco, M., Sulaman, W., Elina, H., Cock, J. M., & Goring, D. R. (2001). Further analysis of the interactions between the Brassica S receptor kinase and three interacting proteins (ARC1, THL1 and THL2) in the yeast two-hybrid system. Plant Mol Biol, 45(3), 365–376.

Mc Clure, B. a, Beecher, B., & Sulaman, W. (2000). Factors affecting inter- and intra-specific pollen rejection in Nicotiana. Annals of Botany, 85, 113–123.

McClure B, Cruz-García F, Romero C (2011) Compatibility and incompatibility in S-RNase-based systems. Ann Bot 108: 647-658.

McClure B, Mou B, Canevascini S, Bernatzky R (1999) A small asparagine-rich protein required for S-allele-specific pollen rejection in Nicotiana. Proc Natl Acad Sci USA 96: 13548-13553.

McClure BA, Cruz-García F, Beecher B, Sulaman W (2000) Factors affecting inter- and intra-specific pollen rejection in Nicotiana. Ann Bot 85: 113-123.

McClure BA, Haring V, Ebert PR, Anderson MA, Simpson RJ, et al. (1989) Style self-incompatibility gene products of Nicotiana alata are ribonucleases. Nature 342: 955-957.

McClure, B. (2009). Darwin's foundation for investigating self-incompatibility and the progress toward a physiological model for S-RNase-based SI. J Exp Bot, 60(4), 1069–1081.

McClure, B., Mou, B., Canevascini, S., & Bernatzky, R. (1999). A small asparagine-rich protein required for S-allele-specific pollen rejection in Nicotiana. Proc Natl Acad Sci USA, 96(23), 13548–53.

McCubbin AG, Kao TH. (2000). Molecular recognition and response in pollen and pistil interactions. A. Rev. Cell Dev. Biol. 16, 333–364.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., et al. (2013). Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 41, W597–W600.

Mehlenbacher SA, Cociu V and Hough LF (1991) Apricots (Prunus), p. 65–107. In: J.N.Moore and J.R. Ballington (eds.). Genetic resources of temperate fruit and nut crops, Chap. 2. International Society for Horticultural Science, Wageningen, The Netherlands.

Meng, D., Gu, Z., Li, W., Wang, A., Yuan, H., Yang, Q., & Li, T. (2014). Apple MdABCF assists in the transportation of S-RNase into pollen tubes. Plant J, 78(6), 990–1002.

Meyer Y, Belin C, Delorme-Hinoux V, Reichheld JP, Riondet, C. (2012). Thioredoxin and glutaredoxin systems in plants: molecular mechanisms, crosstalks, and functional significance. Antioxid Redox Signal. 17, 1124–1160.

Miao HX, Qin YH, Jaime A. Teixeira da Silva, Ye ZX, Hu GB (2011). Cloning and expression analysis of S-RNase homologous gene in Citrus reticulataBlanco cv. Wuzishatangju. Plant Sci., 180: 358-367.

Minamikawa, M. F., Koyano, R., Kikuchi, S., Koba, T., & Sassa, H. (2014). Identification of SFBB-containing canonical and noncanonical SCF complexes in pollen of apple (Malus × domestica). PloS One, 9(5), e97642.

Morimoto T, Akagi T, Tao R (2015) Evolutionary analysis of genes for S-RNase-based self-incompatibility reveals *S* locus duplications in ancestral Rosaceae. Hortic J 84:233–242

Movahedi S, Van de Peer Y, Vandepoele K (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. Plant Physiol 156: 1316-1330.

Murase, K., Shiba, H., Iwano, M., Che, F.-S., Watanabe, M., Isogai, A., & Takayama, S. (2004). A membrane-anchored protein kinase involved in Brassica self-incompatibility signaling. Science, 303(5663), 1516–9.

Murfett, J., Strabala, T. J., Zurek, D. M., Mou, B., Beecher, B., & McClure, B. A. (1996). S RNase and Interspecific Pollen Rejection in the Genus Nicotiana: Multiple Pollen-Rejection Pathways Contribute to Unilateral Incompatibility between Self-Incompatible and Self-Compatible Species. Plant Cell, 8(6), 943–958.

Nasrallah JB (2005) Recognition and rejection of self in plant self-incompatibility: comparisons to animal histocompatibility. Trends Immunol 26(8):412–418

Nasrallah, J. B., Kao, T.-H., Chen, C.-H., Goldberg, M. L., & Nasrallah, M. E. (1987). Amino-acid sequence of glycoproteins encoded by three alleles of the S locus of Brassica oleracea. Nature, 326(6113), 617–619.

Nasrallah, J. B., Nishio, T., & Nasrallah, M. E. (1991). The Self-Incompatibility Genes of Brassica: Expression and Use in Genetic Ablation of Floral Tissues. Annual Review of Plant Physiol and Plant Mol Biol, 42(1), 393–422.

Nei M (1972) Genetic distance between populations. Am Nat 106: 283-292.

Nei, M (1973) Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Soc. USA. 70: 3321-3323.

Nyéki J, Szabó Z, Andrásfalvy A and Erdõs Z (1999). Morphological properties and phenology of the giant ("Óriás") type apricot varieties and their fertility relations. Acta Hort 488: 173-178

Nyujtó F, Brozic S, Nyéki J and Brozic S (1982) Flowering and fruit set in apricot varieties grown in Hungary, and combination of varieties within the plantation. Acta Hort. 121: 159-165

Okada, K., Tonaka, N., Moriya, Y., Norioka, N., Sawamura, Y., Matsumoto, T., Takasaki-Yasuda, T. (2008). Deletion of a 236 kb region around S 4-RNase in a stylar-part mutant S 4sm-haplotype of Japanese pear. Plant Mol Biol, 66(4), 389–400.

Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I. & Koonin, E. V (2010). Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol. Direct 5, 31.

Onda Y. (2013). Oxidative protein-folding systems in plant cells. Int. J. Cell Biol. 2013 585431

Pandey, KK. (1981). Evolution of unilateral incompatibility in flowering plants: further evidence in favour of twin specities controlling intra- and interspecific incompatibility. New Phytol, 89(4), 705–728.

Pannell, J. R., & Barrett, S. C. H. (1998). Baker's law revisited: reproductive assurance in a metapopulation. Evolution, 52(3), 657–668.

Pembleton LW, Shinozuka H, Wang J, Spangenberg GC, Forster JW, Cogan NOI. (2015). Design of an F1 hybrid breeding strategy for ryegrasses based on selection of self-incompatibility locus-specific alleles. Frontiers in Plant Science, 6, 764.

Pinthus MJ. (1959). Seed set of self-fertilized sunflower heads. Agronomy Journal, 51, 626.

Pushkarnath M. (1942). Studies on sterility in potatoes I. The genetics of self- and cross-incompatibilities. Ind J Genet, 2, 11–19.

Rahman, M. H. (2005). Resynthesis of Brassica napus L. for self-incompatibility: Self-incompatibility reaction, inheritance and breeding potential. Plant Breeding, 124(1), 13–19.

Rea, A. C., & Nasrallah, J. B. (2008). Self-incompatibility systems: barriers to self-fertilization in flowering plants. Int J Dev Biol, 52(5-6), 627–636.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1):139-40.

Romero C, Vilanova S, Burgos L, Martínez-Calvo J, Vicente M, et al. (2004) Analysis of the S-locus structure in Prunus armeniaca L. Identification of S-haplotype specific S-RNase and F-box genes. Plant Mol Biol 56: 145-157.

Ru, S., Main, D., Evans, K., & Peace, C. (2015). Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. Tree Genet & Genomes, 11(1), 8.

Rudd, J. J., Franklin, F., Lord, J. M., & Franklin-Tong, V. E. (1996). Increased Phosphorylation of a 26-kD Pollen Protein Is Induced by the Self-Incompatibility Response in Papaver rhoeas. Plant Cell, 8(4), 713–724.

Russell D (1998) The stonefruit cultivar system (A database of worldwide stonefruit cultivars and rootstocks). Department of Primary Industries, Queensland, Australia

Saitou N and Nei M (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol Biol Evol 4(4):406-425

Samuel, M. A., Chong, Y. T., Haasen, K. E., Aldea-Brydges, M. G., Stone, S. L., & Goring, D. R. (2009). Cellular pathways regulating responses to compatible and self-incompatible pollen in Brassica and Arabidopsis stigmas intersect at Exo70A1, a putative component of the exocyst complex. Plant Cell, 21(9), 2655–71.

Sanabria, N. M., van Heerden, H., and Dubery, I. A. (2012). Molecular characterisation and regulation of a *Nicotiana tabacum S*-domain receptor-like kinase gene induced during an early rapid response to lipopolysaccharides. Gene 501, 39–48

Sapir G, Goldway M, Shafir S, Stern RA. (2007). Multiple introduction of honeybee colonies increases cross-pollination, fruit-set and yield of 'Black Diamond' Japanese plum (*Prunus Salicina* Lindl.). J. Hortic. Sci. Biotechnol.82: 590–596.

Sassa, H. (2016). Molecular mechanism of the S-RNase-based gametophytic self-incompatibility in fruit trees of Rosaceae. Breed Sci. 2016 Jan; 66(1): 116–121.

Sassa, H., Kakui, H., Miyamoto, M., Suzuki, Y., Hanada, T., Ushijima, K., Koba, T. (2007). S locus F-box brothers: multiple and pollen-specific F-box genes with S haplotype-specific polymorphisms in apple and Japanese pear. Genetics, 175(4), 1869–81.

Schmidt B, Ho P, Hogg PJ (2006). Allosteric disulfide bonds. Biochemistry 45: 7429-7433.

Schneider, D., Stern, R. A., & Goldway, M. (2005). A Comparison between Semi- and Fully Compatible Apple Pollinators Grown under Suboptimal Pollination Conditions. HortScience, 40(5), 1280–1282.

Schneider, D., Stern, R. A., Eisikowitch, D., & Goldway, M. (2001). Determination of the self-fertilization potency of 'Golden Delicious' apple. J Hort Sci Biotech 76, 259-263.

Schopfer, C. R. (1999). The Male Determinant of Self-Incompatibility in Brassica. Science, 286(5445), 1697–1700.

Schueler, S., Tusch, A., & Scholz, F. (2006). Comparative analysis of the within-population genetic structure in wild cherry (*Prunus avium* L.) at the self-incompatibility locus and nuclear microsatellites. Mol Ecol, 15(11), 3231–43.

Schuelke M (2000). An economic method for the fluorescent labelling of PCR fragments. Nat Biotechnol 18: 233-234.

Sherman, R. L. and W. (1986). Interspecific hybridization of Prunus. HortSci, 21, 48–51.

Shiba, H. (2001). A Pollen Coat Protein, SP11/SCR, Determines the Pollen S-Specificity in the Self-Incompatibility of Brassica Species. Plant Physiol, 125(4), 2095–2103.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L.,Folta, K. M. (2011). The genome of woodland strawberry (Fragaria vesca). Nat Genet, 43(2), 109–16.

Sijacic P, Wang X, Skirpan AL, Wang Y, Dowd PE, et al. (2004) Identification of the pollen determinant of S-RNase-mediated self-incompatibility. Nature 429: 302-305.

Sims, T. L., & Ordanic, M. (2001). Identification of a S-ribonuclease-binding protein in Petunia hybrida. Plant Mol Biol, 47(6), 771–783.

Snowman, B. N., Kovar, D. R., Shevchenko, G., Franklin-Tong, V. E., & Staiger, C. J. (2002). Signal-mediated depolymerization of actin in pollen during the self-incompatibility response. Plant Cell, 14(10), 2613–26.

Sonneveld T, Tobbutt KR, Vaughan SP, Robbins TP (2005) Loss of pollen-S function in two self-compatible selections of Prunus avium is associated with deletion/mutation of an S haplotype-specific F-box gene. Plant Cell 17: 37-51.

Sonneveld T, Tobutt KR, Robbins TP (2003) Allele-specific PCR detection of sweet cherry self-incompatibility (S) alleles S1 to S16 using consensus and allele-specific primers. Theor. Appl. Genet. 107(6):1059-1070

Soost RK. (1964). Self-incompatibility in Citrus grandis osbeck. Proc. Am. Soc. Hort. Sci., 442(84), 137–140.

Stein JC, Howlett B, Boyes DC, Nasrallah ME, Nasrallah JB. (1991). Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of Brassica oleracea. Proc Natl Acad Sci USA, 88(19), 8816–8820.

Steinbachs JE, Holsinger KE. (2002). S-RNase-mediated gametophytic self-incompatibility is ancestral in eudicots. Mol Biol Evol, 19(6), 825–9.

Stern RA, Goldway M, Zisovich AH, Shafir S, Dag A. (2004). Sequential introduction of honeybee colonies increases cross-pollination, fruit-set and yield of "Spadona" pear (Pyrus communis L.). J Hortic Sci Biotechnol, 79(4), 652–658.

Stern R, Eisikowitch D, Dag A. (2001). Sequential introduction of honeybee colonies and doubling their density increases cross-pollination, fruit-set and yield in "Red Delicious" apple. J Hortic Sci Biotechnol, 76(1), 17–23.

Stone SL, Hauksdottir H, Troy A, Herschleb J, Kraft E, Callis J. (2005). Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis. Plant Physiol 137: 13–30.

Suzuki, G., Kai, N., Hirose, T., Fukui, K., Nishio, T., Takayama, S., Hinata, K. (1999). Genomic Organization of the S Locus: Identification and Characterization of Genes in SLG/SRK Region of S9 Haplotype of Brassica campestris (syn. rapa). Genetics, 153(1), 391–400.

Syrgianidis GD and Mainou AC (1993) Deux nouvelles varieties d´abricotier resistances à la maladie à virus de la sharka (Plum pox) issues de croissements. In: Deuxième rencontre sur l´abricotier. Avignon (France). EUR 15009R.

Takasaki, T., Hatakeyama, K., Suzuki, G., Watanabe, M., Isogai, A., & Hinata, K. (2000). The S receptor kinase determines self-incompatibility in Brassica stigma. Nature, 403(6772), 913–6.

Takayama, S. and Isogai, A. (2005) Self-incompatibility in plants. Annu. Rev. Plant Biol. 56, 467–489

Takayama, S., Isogai, A., Tsukamoto, C., Ueda, Y., Hinata, K., Okazaki, K., & Suzuki, A. (1987). Sequences of S-glycoproteins, products of the Brassica campestris self-incompatibility locus. Nature, 326(6108), 102–105.

Takayama, S., Shiba, H., Iwano, M., Shimosato, H., Che, F. S., Kai, N., Isogai, A. (2000). The pollen determinant of self-incompatibility in Brassica campestris. Proc Natl Acad Sci U S A, 97(4), 1920–5.

Takayama, S., Shimosato, H., Shiba, H., Funato, M., Che, F. S., Watanabe, M., Isogai, A. (2001). Direct ligand-receptor complex interaction controls Brassica self-incompatibility. Nature, 413(6855), 534–8.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Bio Evol, 30(12), 2725–9.

Tantikanjana T, Nasrallah ME, Nasrallah JB (2010). Complex networks of self-incompatibility signaling in the Brassicaceae. Curr Opin Plant Biol, 13(5), 520–526.

Tao R and Iezzoni AF (2010) The S-RNase-based gametophytic self-incompatibility system in Prunus exhibits distinct genetic molecular features. Sci Hort 124:423-433

Tao R, Watari A, Hanada T, Habu T, Yaegaki H, et al. (2007) Self-compatible peach (Prunus persica) has mutant versions of the S haplotypes found in selfincompatible Prunus species. Plant Mol Biol 63: 109–123.

Tao R, Yamane H, Sugiura A (1999) Molecular typing of S-alleles through identification, characterization and cDNA cloning for S-RNAses in sweet cherry. J Am Soc Hort Sci 124: 224-233.

Thomas SG, Franklin-Tong VE. 2004. Self-incompatibility triggers programmed cell death in Papaver pollen. Nature 429, 305–309.

Thompson RD, Uhrig H, Hermsen JGT, Salamini F, Kaufmann H (1991) Investigation of a self-compatible mutation in Solanum tuberosum clones inhibiting S-allele activity in pollen differentially. Mol Gen Genom 226: 283-288.

Thompson JD, Higgins DG and Gibson TJ (1994), 'Clustal-W – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice', Nucleic Acids Res., Vol. 22, pp. 4673–4680.

Thomson D and Henry R (1995) Single-step protocol for preparation of plant tissue for analysis by PCR. Biotechniques 19: 394–400

Tovar-Méndez, A., Kumar, A., Kondo, K., Ashford, A., Baek, Y. S., Welch, L., McClure, B. A. (2014). Restoring pistil-side self-incompatibility factors recapitulates an interspecific reproductive barrier between tomato species. Plant J, 77(5), 727–36.

Traverso JA, Pulido A, Rodríguez-García MI, Alché JD (2013). Thiol-based redox regulation in sexual plant reproduction: new insights and perspectives. Front. Plant Sci.4: 465.

Tsukamoto T, Ando T, Takahashi K, Omori T, Watanabe H, et al. (2003) Breakdown of self-incompatibility in a natural population of Petunia axillaris caused by loss of pollen function. Plant Physiol 131: 1903–1912.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. (2012). Primer3-new capabilities and interfaces. Nucleic Acids Research, 40(15), 1–12.

Ushijima K, Sassa H, Dandekar AM, Gradziel TM, Tao R and Hirano H (2003) Structural and transcriptional analysis of the self-incompatibility locus of almond: Identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. Plant Cell 15:771-781

Ushijima K, Yamane H, Watari A, Kakehi E, Ikeda K, Hauck NR, Iezzoni AF, Tao R (2004) The S haplotype-specific F-box protein gene, SFB, is defective in self-compatible haplotypes of Prunus avium and P. mume. Plant J 39:573–586

Ushijima K, Sassa H, Tao R, Yamane H, Dandekar AM, Gradziel TM, Hirano H. (1998). Cloning and characterization of cDNAs encoding S-RNases from almond (Prunus dulcis): primary structural features and sequence diversity of the S-RNases in Rosaceae. Mol Genet Genom, 260(2-3), 261–8.

Van Ooijen JW and Voorrips RE (2001) JoinMap®3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands

Van Tuyl JM and De Jeu MJ (2005) Methods for overcoming interspecific crossing barriers. In: Pollen Biotechnology for Crop Production and Improvement (Ed. V. K. Sawhney & K. R. Shivanna). Cambridge University Press, Cambridge, UK. pp 273-292

Vaughan, S. P., Russell, K., Sargent, D. J., & Tobutt, K. R. (2006). Isolation of S-locus F-box alleles in Prunus avium and their application in a novel method to determine self-incompatibility genotype. Theoretical and Applied Genetics, 112(5), 856–866.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A,Viola R. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). Nat Genet, 42(10), 833–9.

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Rokhsar DS (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet, 45(5), 487–94.

Verrier PJ, Bird D, Buria B, Dassa E, Forestier C, Geisler M, Klein M, Kolukisaoglu U, Lee Y, Martinoia E, Murphy A, Rea PA, Samuels L, Schulz B, Spalding EJ, Yazaki K, Theodoulou FL (2008). Plant ABC proteins – a unified nomenclature and updated inventory. Trends Plant Sci. 13, 151–159.

Vieira J, Fonseca NA, Vieira CP (2008). An S-RNase-based gametophytic self-incompatibility system evolved only once in eudicots. J Mol Evol, 67(2), 179–90.

Vilanova S, Badenes ML, Burgos L, Martínez-Calvo J, Llácer G, Romero C (2006) Self-compatibility of two apricot selections is associated with two pollen-part mutations of different nature. Plant Physiol 142(2):629-641

Vilanova S, Romero C, Abernathy D, Abbott AG, Burgos L, Llacer G and Badenes ML (2003) Construction and application of a bacterial artificial chromosome (BAC) library of Prunus armeniaca L. for the identification of clones linked to the self-incompatibility locus. Mol Genet Genom. 269(5): 685-691

Vilanova S, Romero C, Llácer G and Badenes ML (2005) Identification of self-(in)compatibility alleles in apricot by PCR and sequence analysis. J Am Soc Hort Sci. 130(6): 893-898

Wang Y, Zhang W-Z, Song L-F, Zou J-J, Su Z, et al. (2008) Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis. Plant Physiol 148: 1201-1211.

Wang CL, Zhang SL (2011).A cascade signal pathway occurs in self-incompatibility of *Pyrus pyrifolia*. PlantSignal.Behav. 6, 420–421.

Wang PP, Gao ZH, Ni ZJ, Zhang Z, Cai BH (2013). Self-compatibility in “Zaohong” Japanese apricot is associated with the loss of function of pollen *S* genes. Mol Biol Rep, 40(11), 6485–93.

Watanabe M, S. G. and T. S. (2008). Milestones identifying self-incompatibility genes in Brassica species: from old stories to new findings. In: Self-incompatibility in flowering plants – evolution, diversity, and mechanisms. Springer Verlag Berlin Heidelberg, 151–.

Watari, A., Hanada, T., Yamane, H., Esumi, T., Tao, R., Yaegaki, H., Kataoka, I. (2007). A Low Transcriptional Level of Se-RNase in the Se-haplotype Confers Self-compatibility in Japanese Plum. J. Amer. Soc. Hort. Sci., 132(3), 396–406.

Wheeler, M. J., de Graaf, B. H. J., Hadjiosif, N., Perry, R. M., Poulter, N. S., Osman, K., Franklin-Tong, V. E. (2009). Identification of the pollen self-incompatibility determinant in Papaver rhoeas. Nature, 459(7249), 992–5.

Whitford, R., Fleury, D., Reif, J. C., Garcia, M., Okada, T., Korzun, V., & Langridge, P. (2013). Hybrid breeding in wheat: Technologies to improve hybrid wheat seed production. J Exp Bot, 64(18), 5411–5428.

Wilkins, K. A., Poulter, N. S., & Franklin-Tong, V. E. (2014). Taking one for the team: Self-recognition and cell suicide in pollen. J Exp Bot, 65(5), 1331–1342.

Woodcock, T. (2012). Pollination in the Agricultural Landscape. Best Management Practices for Crop Pollination. Canadian Pollination Initiative (NSERC-CANPOLIN), (Univ. of Guelph).

Wolf, Y. I. & Koonin, E. V. (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol. Evol. 1286–1294.

Wu J, Gu C, Du Y-H, Wu H-Q, Liu W-S, et al. (2010) Self-compatibility of ‘Katy’ apricot (Prunus armeniaca L.) is associated with pollen part mutations. Sex Plant Reprod 24: 23-35.

Wu J, Gu C, Zhang SL, Zhang SJ, Wu HQ and Heng W (2009) Identification of S-haplotype specific S-RNase and SFB alleles in native Chinese apricot (Prunus armeniaca L.). J Hortic Sci Biotechnol, 84: 645–652

Wu, J., Li, M., & Li, T. (2013). Genetic features of the spontaneous self-compatible mutant, “Jin Zhui” (Pyrus bretschneideri Rehd.). PloS One, 8(10), e76509.

Wu, J., Wang, S., Gu, Y., Zhang, S., Publicover, S. J., & Franklin-Tong, V. E. (2011). Self-incompatibility in Papaver rhoeas activates nonspecific cation conductance permeable to Ca2+ and K+. Plant Physiol, 155(2), 963–73.

Wünsch A and Hormaza JI (2004) Genetic and molecular analysis in Cristobalina sweet cherry, a spontaneous self-compatible mutant. Sex Plant Reprod 17:203-210

Wünsch A, Tao R, Hormaza JI (2010) Self-compatibility in ‘Cristobalina’ sweet cherry is not associated with duplications or modified transcription levels of S-locus genes. Plant Cell Rep 29: 715-721.

Xu, C., Li, M., Wu, J., Guo, H., Li, Q., Zhang, Y., Xue, Y. (2013). Identification of a canonical SCF(SLF) complex involved in S-RNase-based self-incompatibility of Pyrus (Rosaceae). Plant Mol Biol, 81(3), 245–57.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P.,Visser, R. G. F. (2011). Genome sequence and analysis of the tuber crop potato. Nature, 475(7355), 189–95.

Xue Y, Carpenter R, Dickinson HG, Coen ES (1996) Origin of allelic diversity in Antirrhinum S locus RNases. Plant Cell 8: 805-814.

Yamamoto M, Nasrallah JB. (2013). In planta assessment of the role of thioredoxin h proteins in the regulation of *S*-locus receptor kinase signaling in transgenic *Arabidopsis thaliana*. Plant Physiol 163: 1387–1395.

Yamane H, Ikeda K, Hauck NR, Iezzoni AF, Tao R (2003) Self-incompatibility (S) locus region of the mutated S6-haplotype of sour cherry (Prunus cerasus) contains a functional pollen S allele and a non-functional pistil S allele. J Exp Bot 54: 2431-2437.

Yamane H, Tao R (2009) Molecular basis of self-(in)compatibility and current status of S-genotyping in Rosaceous fruit trees. J Jpn Soc Hort Sci 78: 137-157.

Yuan, H., Meng, D., Gu, Z., Li, W., Wang, A., Yang, Q., Li, T. (2014). A novel gene, MdSSK1, as a component of the SCF complex rather than MdSBP1 can mediate the ubiquitination of S-RNase in apple. J Exp Bot, 65(12)

Yuen, C. Y. L., Matsumoto, K. O., & Christopher, D. A. (2013). Variation in the Subcellular Localization and Protein Folding Activity among Arabidopsis thaliana Homologs of Protein Disulfide Isomerase. Biomolecules, 3(4), 848–69.

Zhang L, Chen X, Chen X, Zhang C, Liu X, Ci Z, Zhang H, Wu C and Liu C (2008) Identification of self-incompatibility (S-) genotypes of Chinese apricot cultivars. Euphytica 160: 241–248

Zhang X, Ma C, Yin D, Zhu W, Gao C, Z. J. and F. T. (2013). Characterization of S haplotype in a new self-compatible Brassica rapa cultivar Dahuangyoucai. Czech J. Genet. Plant Breed, 49(4), 157–163.

Zhang Q, Chen W, Sun L, Zhao F, Huang B et al. (2012) The genome of *Prunus mume*. Nat Commun 3: 1318.

Zhang, S. L., & Hiratsuka, S. (2005). Analyses of pollen-tube growth and biological action of S-RNase in the style of self-compatible Japanese pear. Scientia Horticulturae, 104(2), 169–178.

Zhang, S.-W., Huang, G.-X., Ding, F., He, X.-H., & Pan, J.-C. (2012). Mechanism of seedlessness in a new lemon cultivar “Xiangshui” [Citrus limon (L.) Burm. F]. Sex Plant Reprod , 25(4), 337–45.

Zhao L, Huang J, Zhao Z, Li Q, Sims TL, et al. (2010) The Skp-like protein SSK is required for cross-pollen compatibility in S-RNase-based self-incompatibility. Plant J 62: 52-63.

Zhebentyayeva TN, Swire-Clark G, Georgi LL, Garay L, Jung S, et al. (2008) A framework physical map for peach, a model Rosaceae species. Tree Genet Genomes 4: 745–756.

Zheng XH, Lu F, Wang Z-Y, Zhong F, Hoover J, et al. (2005) Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. Bioinformatics 21: 703-710.

Zhu C, Zhang YM (2007) An EM algorithm for mapping distortion segregation loci. BMC Genetics 8: 82.

Zisovich A, Stern R, Shafir S, Goldway M (2005). Fertilisation efficiency of semi- and fully-compatible European pear (*Pyrus communis* L.) cultivars. J Hort Sci Biotech 80, 143-146

Zuriaga E , Molina L, Badenes ML, Romero C (2012) Physical mapping of a pollen modifier locus controlling self-incompatibility in apricot and synteny analysis within the Rosaceae. Plant Mol Biol 79: 229-242.

Zuriaga E, Muñoz-Sanz JV, Molina L, Gisbert AD, Badenes ML and Romero C (2013) An *S*-locus independent pollen factor confers self-compatibility in “Katy” apricot. PloS One, 8(1), e53947

SUPPLEMENTAL INFORMATION

Supporting information chapter 1

Table S1.1 Identification of segregation distortion SSR loci distributed throughout the 'Katy' LG6 using the F_2 population 'K×K'. χ^2 and *P* values estimated for each SSR, considering the expected segregation ratio 1:2:1 are indicated.

LG	Locus	Peach Mb[a]	Seg. Type[b]	A	H	B	Total	χ^2 (*P*-value)[c]
6	PGS6_01	00,12	<abxab>	28	45	12	85	6,32 (0,04)[d]
6	PGS6_04	04,95	<abxab>	29	45	11	85	7,92 (0,02)[d]
6	UDAp420	08,14	<abxab>	34	41	12	87	11,41 (0,03)[d]
6	PGS6_07	09,33	<abxab>	32	47	7	86	15,28 (0,0004)[d]
6	UDAp489	16,82	<abxab>	34	41	11	86	12,49 (0,002)[d]
6	Ma027a	20,90	<abxab>	33	43	9	85	13,57 (0,001)[d]
6	BBPCT025	21,13	<abxab>	35	42	10	87	14,47 (0,0007)[d]
6	UDP98-412	24,75	<abxab>	28	43	12	83	6,28 (0,04)[d]
6	Locus-S	26,45	<abxab>	31	40	16	87	5,74 (0,06)
6	ssrPaCITA12	27,84	<abxab>	31	38	16	85	6,24 (0,04)[d]

[a] Marker position (Mb) within the peach genome scaffold_6 which size estimated by IPGI was 28.90 Mb
[b] Segregation type as per JoinMap 3.0
[c] Chi-square test was performed for the expected ratio 1:2:1 (<abxab>)
[d] Observed ratios differ significantly from expected at $P < 0.05$ for 2 degrees of freedom

Table S1.2 SSR primers developed from the peach genomic sequence corresponding to the scaffold _3. Primer position on the scaffold (Mb) and SSR allele sizes amplified in apricot cvs. 'Goldrich', 'Canino' and 'Katy' are indicated.

Name	Start on scaffold_3	Goldrich alleles	Canino alleles	Katy alleles	Name	Start on scaffold_3	Goldrich alleles	Canino alleles	Katy alleles
PGS3.01	16,21	176	168/176	N.A.	PGS3.52	17,96	194/208	194	194/204
PGS3.02	16,30	163/174	166/174	163/174	PGS3.53	18,07	199/203	199	197/199
PGS3.03	16,41	162/178	156/162	162/178	PGS3.54	18,12	280/282	282	262/282
PGS3.04	16,52	N.A.[a]	N.A.	N.A.	PGS3.55	18,19	449	449	N.A.
PGS3.05	16,64	171	171	N.A.	PGS3.56	18,22	251/258	258	251/267
PGS3.06	16,70	152/160	152	ML	PGS3.57	18,23	N.A.	N.A.	N.A.
PGS3.07	16,84	141	141	141	PGS3.58	18,24	95/99	95	N.A.
PGS3.08	16,91	157	157	157	PGS3.59	18,50	211/215	215	215/217
PGS3.09	17,01	204	204	204	PGS3.60	18,51	293/317	317	N.A.
PGS3.10	17,16	130	129/130	129	PGS3.61	18,58	173	173	173
PGS3.11	17,22	N.A.	156	N.A.	PGS3.62	18,61	336/350	348/350	N.A.
PGS3.12	17,38	148	148/162	148/156	PGS3.63	18,65	218/220	220	212/220
PGS3.13	17,54	230	202/230	215/230	PGS3.64	18,70	ML	ML	ML
PGS3.14	17,63	N.A.	N.A.	N.A.	PGS3.65	18,31	ML	ML	ML
PGS3.15	17,71	266/267	267/274	266/267	PGS3.66	18,34	322/328	328	322/328
PGS3.16	17,80	N.A.	N.A.	N.A.	PGS3.67	18,35	210/216	210	201/210
PGS3.17	17,98	N.A.	N.A.	N.A.	PGS3.68	18,37	311/338	338	338

PGS3.18	18,06	147	147	147	PGS3.69	18,38	202/227	202	N.A.
PGS3.19	18,12	147	199	N.A.	PGS3.70	18,40	200/201	200	198/200
PGS3.20	18,25	148	N.A.	150	PGS3.71	18,40	255/259	259/261	245/259
PGS3.21	18,40	229/244	244	217	PGS3.72	18,41	ML	ML	ML
PGS3.22	18,49	306/310	306	306/312	PGS3.73	18,43	368	368	368
PGS3.23	18,61	179/188	188	188/190	PGS3.74	18,46	262/265	265	265
PGS3.24	18,77	ML[b]	ML	ML	PGS3.75	18,48	N.A.	N.A.	351/352
PGS3.25	18,87	180/192	186/192	186	PGS3.76	18,49	192	192	192/198
PGS3.26	18,94	155/166	160/166	ML	PGS3.77	18,50	273/276	273	273/276
PGS3.27	19,03	N.A.	N.A.	N.A.	PGS3.78	18,54	161/164	164	164/166
PGS3.28	19,14	141	141	141/143	PGS3.79	18,57	460	460	N.A.
PGS3.29	19,25	159	159	159	PGS3.80	18,59	474	474	N.A.
PGS3.30	19,34	247	242	242	PGS3.81	18,61	227	227	227
PGS3.31	19,45	ML	ML	ML	PGS3.82	18,61	148	148	148
PGS3.32	19,60	256	256	256/270	PGS3.83	18,61	ML	ML	N.A.
PGS3.33	19,66	133/139	133/139	129/133	PGS3.84	18,61	N.A.	N.A.	N.A.
PGS3.34	17,75	185/191	185/193	N.A.	PGS3.85	18,62	ML	ML	197/199
PGS3.35	17,85	N.A.	N.A.	N.A.	PGS3.86	18,63	170/179	170	170/179
PGS3.36	17,95	148/162	325/327	N.A.	PGS3.87	18,63	335/340	340	N.A.
PGS3.37	18,00	189/202	189	189/202	PGS3.88	18,65	278	N.A.	284
PGS3.38	18,05	246/264	264	264	PGS3.89	18,66	404/405	405	405
PGS3.39	18,07	165	165	165	PGS3.90	18,69	ML	ML	185
PGS3.40	18,14	128	177	177	PGS3.91	18,70	285/null	N.A.	N.A.
PGS3.41	18,18	N.A.	N.A.	N.A.	PGS3.92	18,70	ML	ML	ML
PGS3.42	18,22	206/224	206	206/209	PGS3.93	18,73	206/207	206	206
PGS3.43	18,24	N.A.	N.A.	N.A.	PGS3.94	18,74	N.A.	N.A.	N.A.
PGS3.44	18,29	295/307	307/309	264/309	PGS3.95	18,75	344	344	344
PGS3.45	18,33	237	N.A.	237	PGS3.96	18,76	434/441	441/442	N.A.
PGS3.46	18,47	163/173	163	163/181	PGS3.97	18,80	N.A.	N.A.	180/237
PGS3.47	18,52	242/246	242	N.A.	PGS3.98	18,81	N.A.	N.A.	N.A.
PGS3.48	18,60	256/264	264	256/264	PGS3.99	18,81	N.A.	N.A.	N.A.
PGS3.49	18,63	N.A.	N.A.	N.A.	PGS3.100	18,84	239	239	239
PGS3.50	18,83	N.A.	N.A.	N.A.	PGS3.101	18,84	N.A.	N.A.	222
PGS3.51	17,83	ML	ML	149/192	PGS3.102	18,85	251	251	251

[a] N.A. Not amplified

[b] ML. Multi-loci pattern

Table S1.3 SSR allele composition for apricot cvs. 'Goldrich', 'Canino' and 'Katy'. Start position on the corresponding scaffold (Number_Mb) and SSR allele sizes (bp) are indicated.

Name	Start on scaffold	Goldrich alleles	Canino alleles	Katy alleles	Name	Start on scaffold	Goldrich alleles	Canino alleles	Katy alleles
Gol051	1_04,69	177/181	181/183	181/183	PGS4.05	4_12,03	171/171	171/171	171/171
EPPUC0027	1_09,51	174/176	170/176	170/174	PGS4.07	4_18,14	179/179	176/176	179/179
pchcms4	1_09,51	246/248	242/248	242/246	UDAp404	4_--	165/165	188/188	165/177
UDAp414	1_26,52	179/183	167/187	179/179	PGS5.02	5_00,48	178/183	183/183	178/181
EPPCU1589	1_31,81	160/176	176/182	182/182	PGS5.03	5_05,06	138/138	136/138	138/138
CPPCT045	1_32.02	125/134	125/128	128/134	SsrPaCITA21	5_10,78	241/241	239/241	241/241
SsrPaCITA7	1_32,02	221/227	209/227	227/227	CPSCT006	5_11,53	136/138	136/138	136/140
Gol004	1_45,40	221/231	208/221	231/231	BPPCT037	5_12,31	132/137	153/163	132/132
SsrPaCITA16	2_03,76	130/130	130/152	130/130	pchgms4	5_12,67	168/168	168/195	168/168
PGS2.03	2_04,36	177/177	177/177	177/177	UDAp452	5_13,76	190/197	190/197	189/201
SsrPaCITA19	2_13,01	156/167	132/167	156/167	PGS5.08	5_15,30	219/219	219/219	219/219
CPSCT038	2_14,47	207/207	216/226	207/207	PGS6.01	6_00,12	284/289	289/289	284/289
BPPCT001	2_16,13	115/116	115/116	115/115	PGS6.02	6_01,14	265/269	263/263	255/263
CPSCT044	2_17,22	209/211	199/211	209/209	PGS6.03	6_01,23	221/225	227/227	227/233
UDP98-411	2_20,17	177/184	181/192	181/184	PGS6.04	6_04,95	149/158	149/155	149/158
CPSCT021	2_23,74	148/156	148/152	152/156	UDAp420	6_08,14	187/196	187/196	185/187
CPSCT031	2_25,15	203/205	205/205	203/205	PGS6.07	6_09,33	183/185	185/201	186/196
CPSCT023	2_25,34	210/235	216/216	210/216	PGS6.08	6_09,35	236/238	238/238	228/238
CPSCT034	2_26,35	216/231	212/231	216/216	BPPCT008	6_10,28	86/107	86/105	86/113
MA066a	3_02,40	131/131	125/125	131/133	UDAp489	6_16,82	164/181	181/181	148/181
SsrPaCITA23	3_02,70	157/161	155/165	157/165	Ma027a	6_20,90	148/180	152/170	154/180
BPPCT007	3_02,74	185/185	165/171	165/165	BPPCT025	6_21,13	150/160	148/160	150/160
UDAp446	3_04.50	166/179	166/168	166/179	UDP98-412	6_24,75	104/110	82/110	82/104
UDAp468	3_04,85	170/182	158/170	170/182	Locus-*S*	6_26,45	S1/S2	S2/SC	S1/S2
BPPCT039	3_05,80	182/187	165/187	182/187	SsrPaCITA12	6_27,84	162/168	165/168	162/168
EPPCU2256	3_06,14	162/254	162/172	162/254	CPSCT004	7_6,68	143/143	143/145	143/143
EPDCU3083	3_06,46	149/149	149/151	149/149	CPPCT022	7_10,23	257/289	269/289	252/257
UDA002	3_10,85	175/186	147/147	175/175	UDP98-405	7_10,94	120/120	120/124	120/120
SsrPaCITA10	3_14,16	191/193	191/191	193/193	CPSCT026	7_10,98	218/218	206/209	211/218
UDAp499	3_14,71	120/239	120/266	120/239	PGS7.05	7_13,08	217/221	205/215	217/217
SsrPaCITA4	3_14,81	152/158	166/166	152/152	CPPCT033	7_16,70	142/142	142/142	138/142
EPPCU9343	3_16,70	194/202	194/194	194/202	CPSCT042	7_17,08	187/192	181/187	181/192
EPPCU7190	3_19,78	212/214	212/214	214/226	CPSCT018	8_00,12	160/160	165/166	160/160
AMPA119	3_20,00	114/118	114/118	118/128	UDAp423	8_00,18	151/184	151/151	151/184
UCDCH19	3_20,03	143/143	135/143	133/135	PGS8.02	8_03,92	326/343	330/330	326/343
CPDCT027	3_21,67	169/null	166/null	166/167	PGS8.05	8_07,39	251/251	243/251	245/253
EPPCU0532	3_22,00	182/182	182/186	182/184	UDAp401	8_10,50	220/222	222/222	218/222
PGS4.01	4_03,46	348/348	356/356	348/356	UDAp470	8_12,61	110/120	120/120	114/120

PGS4.02	4_03,49	169/175	169/169	169/175	CPPCT006	8_13,66	200/200	200/200	200/204
CPDCT045	4_06,21	125/134	125/128	128/134	M6a	8_15,03	222/222	196/222	204/222
BPPCT040	4_06,46	151/158	151/160	158/160	UDP98-409	8_17,78	148/152	152/168	148/148
UDP96-003	4_08,76	112/126	112/112	112/126	Ma035a	8_21,83	170/180	180/182	180/180
PGS4.04	4_11,99	278/278	265/265	288/288					

Table S1.4 Genetic distances among apricot cvs. 'Katy', 'Canino' and 'Goldrich' estimated according to Nei (1972) (below diagonal) and % of shared SSR alleles (above diagonal).

	Katy	Canino	Goldrich
Katy	---	38,8	61,2
Canino	0,83	---	44,7
Goldrich	0,39	0,73	---

Table S1.5 Gene content of the *M'*-locus peach syntenic region. Position and length of the ORFs as well as the first BLASTP match on the TAIR database annotated by IPGI are shown. Overlap length (amino acids), percent id and E-value are indicated for each *Prunus*/*Arabidopsis* gene pair. *Arabidopsis* homologues with detectable expression in mature pollen, hydrated pollen and pollen tubes (+/-) and those with altered transcription during pollen germination (PG) or pollen tube growth (PTG) are also indicated according to the results reported by Wang et al. (2008) using Affymetrix ATH1 Genome Arrays.

Peach Gene ID	Transcript Length	Start	Stop	TAIR	Description	Overlap length	% id	E-value	Pollen expr.	Altered transc.
ppa022538m	18491554	18493033	1480	AT5G15720.1	GLIP7; carboxylesterase/ lipase	340	44,71	1E-76	-	-
ppa006182m	18495042	18498361	3320	AT5G15730.2[b]	serine/threonine protein kinase, putative	428	66,36	3E-155	-	-
ppa002721m	18498827	18500772	1946	AT2G03890.1	phosphatidylinositol 3- and 4-kinase family protein	660	63,79	0	-	-
ppa002370m	18503945	18508593	4649	AT3G30300.1[b]	unknown protein	690	68,41	0	-	-
ppa008856m	18510031	18513228	3198	AT1G69010.1	BIM2 (BES1-interacting Myc-like protein 2); DNA binding / transcription factor	273	44,32	5E-43	-	-
ppa004594m	18515587	18518005	2419	AT5G15740.1	unknown protein	508	69,09	0	-	-
ppa012139m	18522087	18523783	1697	AT5G15750.1[b]	RNA-binding S4 domain-containing protein	167	80,84	1E-80	-	-
ppa000002m	18524173	18545067	20895	AT3G02260.1[b]	BIG (BIG); binding / ubiquitin-protein ligase/ zinc ion binding	5008	68,11	0	+	/
ppa026731m	18545106	18546555	1450	AT3G25270.1	nucleic acid binding	149	28,86	1E-10	-	-
ppa023507m	18559505	18560630	1126	AT4G37180.1	myb family transcription factor	52	67,31	2E-14	+	downPTG
ppa005351m	18563326	18565121	1796	AT5G15780.1	pollen Ole e 1 allergen and extensin family protein	168	58,93	9E-41	-	-
ppa011450m	18574543	18575507	965	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa001620m	18581982	18588641	6660	AT5G38880.1[b]	unknown protein	794	74,31	0	-	-
ppa011007m	18590528	18594247	3720	AT5G15790.2[b]	zinc finger (C3HC4-type RING finger) family protein	231	60,61	3E-74	-	-
ppa017665m	18594777	18596276	1500	AT5G38900.1	DSBA oxidoreductase family protein	207	55,07	2E-67	+	/

ppa011285m	18597087	18599509	2423	AT5G38900.1	DSBA oxidoreductase family protein	214	63,08	6E-77	+	/
ppa011289m	18597087	18599313	2227	AT5G38900.1	DSBA oxidoreductase family protein	214	63,08	6E-77	+	/
ppa011302m	18597087	18599360	2274	AT5G38900.1[b]	DSBA oxidoreductase family protein	214	63,08	6E-77	+	/
ppa012296m	18597087	18599509	2423	AT5G38900.1	DSBA oxidoreductase family protein	213	55,4	6E-63	+	/
ppa005069m	18600204	18604076	3873	AT3G02300.1[b]	regulator of chromosome condensation (RCC1) family protein	463	77,97	0	+	/
ppa010249m	18604427	18608655	4229	AT1G69120.1	AP1 (APETALA1); DNA binding / protein binding / protein heterodimerization/ transcription activator/ transcription factor	152	72,37	1E-52	-	-
ppa010548m	18619188	18623759	4572	AT3G02310.1	SEP2 (SEPALLATA 2); DNA binding / protein binding / transcription factor	251	74,9	1E-95	-	-
ppa010577m	18619188	18623759	4572	AT3G02310.1[b]	SEP2 (SEPALLATA 2); DNA binding / protein binding / transcription factor	250	75,2	5E-97	-	-
ppa026503m	18629652	18630269	618	AT5G15802.1[b]	unknown protein	110	56,36	9E-30	-	-
ppa016385m	18630946	18633249	2304	AT3G30210.1[b]	MYB121 (MYB DOMAIN PROTEIN 121); DNA binding / transcription factor	217	51,15	2E-52	-	-
ppa013380m	18636187	18636866	680	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa003386m	18638891	18643879	4989	AT3G02320.1[b]	RNA binding / tRNA (guanine-N2-)-methyltransferase	558	78,85	0	-	-
ppa007756m	18645262	18647647	2386	AT3G02230.1[b]	RGP1 (REVERSIBLY GLYCOSYLATED POLYPEPTIDE 1); cellulose synthase (UDP-forming)	337	91,1	0	+	/
ppa005994m	18649075	18651638	2564	AT3G30180.1[b]	BR6OX2 (BRASSINOSTEROID-6-OXIDASE 2); monooxygenase/ oxygen binding	464	67,24	0	+	Up PTG
ppa007503m	18660657	18662638	1982	AT1G13680.1[b]	phospholipase C/ phosphoric diester hydrolase	348	62,64	1E-129	-	-
ppa014104m	18664030	18664766	737	AT3G29970.1[b]	germination protein-related	86	69,77	5E-34	+	/
ppa024465m	18666252	18668102	1851	AT5G15630.1	IRX6	397	80,1	0	-	-
ppa005522m	18668649	18671935	3287	AT5G60920.1	COB (COBRA)	435	74,25	0	-	-
ppb023073m	18676574	18678411	1838	AT3G02210.1	COBL1 (COBRA-LIKE PROTEIN 1 PRECURSOR)	180	42,22	1E-33	-	-
ppb020721m	18704742	18706234	1493	AT1G56440.1	serine/threonine protein phosphatase-related	56	62,5	7E-16	-	-
ppa022025m	18716879	18718491	1613	AT2G21660.2	CCR2 (COLD, CIRCADIAN RHYTHM, AND RNA BINDING 2); RNA binding / double-stranded DNA binding / single-stranded DNA binding	55	65,45	5E-15	-	-
ppa007463m	18748304	18751578	3275	AT5G15640.1[b]	mitochondrial substrate carrier family protein	319	79,31	4E-147	+	Up PTG
ppa010178m	18748304	18749962	1659	AT5G15640.1	mitochondrial substrate carrier family protein	210	80,95	2E-98	+	Up PTG
ppa015454m	18752057	18753454	1398	AT4G01240.1	unknown protein	329	58,66	3E-111	-	-
ppa011807m	18754276	18755073	798	AT3G02220.1[b]	unknown protein	167	69,46	7E-63	-	-
ppa013367m	18755340	18756208	869	AT1G24140.1[a]	Matrixin family protein	92	32	7E-6	+	Up PTG
ppa019352m	18760652	18763276	2625	AT3G51550.1	FER (FERONIA); kinase/ protein kinase	908	45,15	0	-	-
ppa001157m	18766568	18769249	2682	AT5G38990.1	protein kinase family protein	837	46,95	1E-175	-	-
ppa001190m	18774840	18777497	2658	AT3G51550.1	FER (FERONIA); kinase/ protein kinase	825	49,58	0	-	-
ppa018922m	18793680	18794048	369	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa020589m	18806038	18808848	2811	AT3G51550.1	FER (FERONIA); kinase/ protein kinase	812	51,6	0	-	-
ppa013018m	18809760	18810532	773	AT2G33775.1[a]	RALFL19 ralf-like19	231	32	0.007	-	-

ppa003761m	18812093	18814176	2084	AT3G51550.1	FER (FERONIA); kinase/ protein kinase	414	44,69	1E-84	-	-
ppa001413m	18815789	18818489	2701	AT3G51550.1	FER (FERONIA); kinase/ protein kinase	394	53,3	4E-108	-	-
ppa016279m	18829367	18831840	2474	AT3G51550.1	FER (FERONIA); kinase/ protein kinase	815	47,48	0	-	-
ppa006461m	18834674	18838207	3534	AT5G15610.2[b]	proteasome family protein	413	68,52	3E-177	-	-
ppa017965m	18839676	18840153	478	AT1G69230.2	SP1L2 (SPIRAL1-LIKE2)	97	48,45	1E-14	-	-
ppa007173m	18841314	18844509	3196	AT3G29770.1[b]	MES11 (METHYL ESTERASE 11); hydrolase	310	77,42	1E-125	-	-
ppa007590m	18849392	18852062	2671	AT3G29760.1	NLI interacting factor (NIF) family protein	146	58,22	1E-49	-	-
ppa000735m	18858946	18862792	3847	AT1G74160.1	unknown protein	1090	35,6	1E-116	-	-
ppa007206m	18863483	18864819	1337	AT5G15570.1[b]	unknown protein	391	38,87	1E-64	-	-
					DOMAIN/s: Bromodomain transcription factor					
					BEST Arabidopsis thaliana protein match is: DNA binding					
ppa007243m	18867715	18869807	2093	AT3G02150.2[b]	PTF1 (PLASTID TRANSCRIPTION FACTOR 1); transcription factor	75	88	3E-36	-	-
ppa021495m	18874448	18875851	1404	AT3G29635.1	transferase family protein	443	40,41	4E-74	-	-
ppa020932m	18881667	18882581	915	AT1G21280.1[a]	DOMAIN/s: Retrotransposon gag protein	294	27	4E-06	-	-
ppa005255m	18885375	18887152	1778	AT5G39090.1	transferase family protein	470	37,45	2E-78	-	-
ppa026050m	18893531	18896984	3454	AT5G39080.1	transferase family protein	436	38,99	4E-78	-	-
ppa024873m	18897669	18899334	1666	AT3G29635.1	transferase family protein	429	28,9	2E-39	-	-
ppa021452m	18899747	18901162	1416	AT5G39090.1	transferase family protein	448	41,07	9E-85	-	-
ppa022904m	18911571	18913109	1539	AT5G39080.1	transferase family protein	431	27,61	1E-33	-	-
ppa026936m	18914045	18915457	1413	AT5G39090.1	transferase family protein	472	39,41	5E-88	-	-
ppa005502m	18936429	18937883	1455	AT5G39090.1	transferase family protein	471	39,92	3E-85	-	-
ppa020665m	18940430	18941821	1392	AT5G39080.1[b]	transferase family protein	465	40,43	9E-88	-	-
ppa016949m	18957619	18959010	1392	AT5G39080.1	transferase family protein	465	40,22	9E-86	-	-
ppa025189m	18961426	18962349	924	AT5G39090.1	transferase family protein	214	40,19	5E-39	-	-
ppa018052m	18963290	18963619	330	AT5G39080.1	transferase family protein	64	45,31	8E-10	-	-
ppa016023m	18968998	18970547	1550	AT5G39090.1	transferase family protein	294	32,99	1E-31	-	-
ppa005488m	18972684	18974060	1377	AT5G39080.1	transferase family protein	468	39,53	5E-89	-	-
ppa020216m	18979935	18982632	2698	AT5G39080.1	transferase family protein	228	43,86	4E-46	-	-
ppa019904m	18992254	18993632	1379	AT3G29635.1	transferase family protein	429	38,46	1E-72	-	-
ppa016299m	18995785	18997447	1663	AT5G39090.1	transferase family protein	477	39,41	1E-80	-	-
ppa019320m	18998050	18999448	1399	AT5G39080.1	transferase family protein	222	37,84	8E-33	-	-
ppa005348m	19000743	19005122	4380	AT5G60980.2	nuclear transport factor 2 (NTF2) family protein / RNA recognition motif (RRM)-containing protein	386	48,7	8E-83	-	-
ppa007375m	19007660	19009944	2285	AT3G29575.4[b]	AFP3 (ABI FIVE BINDING PROTEIN 3)	84	82,14	5E-38	+	/
ppa000986m	19011230	19014865	3636	AT3G02130.1[b]	RPK2 (RECEPTOR-LIKE PROTEIN KINASE 2); ATP binding / kinase/ protein serine/threonine kinase	953	67,89	0	-	-
ppa009937m	19018505	19020134	1630	AT3G02125.1[b]	unknown protein	191	33,51	1E-12	+	Up PTG
ppa012522m	19020442	19021206	765	AT5G39210.1[b]	CRR7 (CHLORORESPIRATORY REDUCTION 7)	162	46,91	3E-37	-	-
ppa016098m	19021821	19022271	451	AT3G02120.1[b]	hydroxyproline-rich glycoprotein family protein	61	67,21	6E-16	-	-
ppa014248m	19022519	19023590	1072	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa017321m	19028160	19032318	4159	AT3G02110.1[b]	scpl25 (serine carboxypeptidase-like 25); serine-type carboxypeptidase	473	74,63	0	-	-
ppa005110m	19034922	19036600	1679	AT3G29400.1	ATEXO70E1 (exocyst subunit EXO70 family protein E1); protein binding	279	50,9	3E-71	-	-
ppa004810m	19038382	19044888	6507	AT3G29390.1[b]	RIK (RS2-Interacting KH protein); RNA binding	490	54,08	1E-119	+	/
ppa007817m	19047717	19051663	3947	AT1G13820.1	hydrolase, alpha/beta fold family protein	301	66,78	3E-117	-	-

ppa005967m	19052928	19057394	4467	AT5G15550.1[b]	transducin family protein / WD-40 repeat family protein	434	72,81	0	-	-
ppa000951m	19058877	19065497	6621	AT1G69830.1	AMY3 (ALPHA-AMYLASE-LIKE 3); alpha-amylase	403	59,8	3E-149	-	-
ppa000160m	19075922	19083048	7127	AT5G07980.1	dentin sialophosphoprotein-related	412	39,81	4E-71	-	-
ppa020830m	19084297	19085579	1283	AT3G10330.1	transcription initiation factor IIB-2 / general transcription factor TFIIB-2 (TFIIB2)	314	50	5E-82	+	/
ppa021805m	19089131	19090244	1114	AT3G10330.1	transcription initiation factor IIB-2 / general transcription factor TFIIB-2 (TFIIB2)	313	45,05	2E-68	+	/
ppa024744m	19091119	19092216	1098	AT3G02100.1	UDP-glucoronosyl/UDP-glucosyl transferase family protein	264	42,05	1E-47	-	-
ppa020867m	19098001	19099672	1672	AT3G02100.1	UDP-glucoronosyl/UDP-glucosyl transferase family protein	465	46,24	5E-122	-	-
ppa019116m	19105243	19106720	1478	AT3G02100.1[b]	UDP-glucoronosyl/UDP-glucosyl transferase family protein	465	47,1	3E-123	-	-
ppa020470m	19107803	19109506	1704	AT3G02100.1	UDP-glucoronosyl/UDP-glucosyl transferase family protein	464	45,47	2E-119	-	-
ppa006338m	19110526	19112487	1962	AT3G02100.1	UDP-glucoronosyl/UDP-glucosyl transferase family protein	464	43,97	8E-110	-	-
ppa025965m	19114187	19114441	255	AT5G39240.1[b]	unknown protein	90	41,11	5E-11	-	-
ppa010438m	19119067	19120153	1087	AT5G39250.1[b]	F-box family protein	252	67,06	2E-93	+	/
ppa024239m	19120527	19121460	934	AT5G39300.1	ATEXPA25 (ARABIDOPSIS THALIANA EXPANSIN A25)	224	65,18	2E-89	-	-
ppa000125m	19123369	19135325	11957	AT5G15540.1[b]	EMB2773 (EMBRYO DEFECTIVE 2773); binding / protein binding / zinc ion binding	1563	67,5	0	-	-
ppa009593m	19136301	19138411	2111	AT5G15530.1[b]	BCCP2 (BIOTIN CARBOXYL CARRIER PROTEIN 2); biotin binding	289	47,06	2E-55	-	-
ppa004059m	19138952	19142613	3662	AT3G02090.1[b]	mitochondrial processing peptidase beta subunit, putative	480	78,75	0	+	/
ppa011540m	19142598	19144436	1839	AT5G61170.1	40S ribosomal protein S19 (RPS19C)	139	90,65	1E-72	-	-
ppa012105m	19142598	19144436	1839	AT5G61170.1	40S ribosomal protein S19 (RPS19C)	120	90	5E-62	-	-
ppa020426m	19145946	19149339	3394	AT3G01015.1[b]	unknown protein DOMAIN/s: Targeting for Xklp2	496	52,82	6E-101	+	/
ppb012981m	19154167	19156382	2216	AT3G26922.1	unknown protein DOMAIN/s: Cyclin-like F-box, Leucine-rich repeat 2 BEST Arabidopsis thaliana protein match is: F-box family protein	286	27,27	4E-13	-	-
ppa004981m	19158038	19160714	2677	AT5G15490.1	UDP-glucose 6-dehydrogenase, putative	482	87,55	0	+	/
ppa004991m	19158038	19160714	2677	AT5G15490.1	UDP-glucose 6-dehydrogenase, putative	482	87,55	0	+	/
ppa005006m	19158038	19160714	2677	AT5G15490.1	UDP-glucose 6-dehydrogenase, putative	482	87,55	0	+	/
ppa005897m	19162719	19164403	1685	AT1G26580.1	unknown protein BEST Arabidopsis thaliana protein match is: myb family transcription factor / ELM2 domain-containing protein	475	32,21	1E-43	+	/
ppa016813m	19167289	19168905	1617	AT1G26610.1	zinc finger (C2H2 type) family protein	322	28,26	5E-16	+	/
ppa000827m	19171126	19177833	6708	AT3G29320.1	glucan phosphorylase, putative	516	74,42	0	-	-

ppa010903m	19180586	19181564	979	AT3G29310.1[a]	Calmodulin-binding protein related	378	29	0.023	-	-
ppa023614m	19182400	19185223	2824	AT4G38180.1	FRS5 (FAR1-related sequence 5); zinc ion binding	624	45,03	9E-163	-	-
ppa013993m	19185732	19187433	1702	AT5G61220.1[b]	complex 1 family protein / LVR family protein	83	55,42	1E-22	-	-
ppa014984m	19187929	19189674	1746	AT5G39350.1[b]	pentatricopeptide (PPR) repeat-containing protein	583	58,15	0	-	-
ppa004045m	19193460	19197883	4424	AT5G15470.1[b]	GAUT14 (Galacturonosyltransferase 14); polygalacturonate 4-alpha-galacturonosyltransferase/ transferase, transferring glycosyl groups / transferase, transferring hexosyl groups	532	87,59	0	+	Down PTG
ppa013551m	19199756	19202490	2735	AT3G01050.1[b]	MUB1 (MEMBRANE-ANCHORED UBIQUITIN-FOLD PROTEIN 1 PRECURSOR)	117	64,1	1E-42	-	-
ppa014205m	19203844	19204267	424	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa007471m	19205856	19207724	1869	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa000855m	19208678	19213930	5253	AT5G15450.1[b]	CLPB3 (CASEIN LYTIC PROTEINASE B3); ATP binding / ATPase/ nucleoside-triphosphatase/ nucleotide binding / protein binding	953	85,41	0	+	/
ppa017256m	19214613	19216111	1499	AT3G29280.1[b]	unknown protein	150	75,33	9E-59	-	-
ppa010470m	19216695	19219176	2482	AT5G39360.1[b]	EDL2 (EID1-like 2)	248	83,87	7E-126	-	-
ppa007111m	19220088	19221740	1653	AT5G39380.1[b]	calmodulin-binding protein-related	275	46,55	1E-42	+	/
ppa005586m	19228019	19230695	2677	AT3G01060.1[b]	unknown protein	455	79,78	0	-	-
ppa010047m	19231344	19232618	1275	AT3G29270.2[b]	ubiquitin-protein ligase	266	64,29	2E-85	-	-
ppa002126m	19241079	19244334	3256	AT5G15410.1[b]	DND1 (DEFENSE NO DEATH 1); calcium channel/ calmodulin binding / cation channel/ cyclic nucleotide binding / intracellular cAMP activated cation channel/ intracellular cyclic nucleotide activated cation channel/ inward rectifier potassium channel	728	74,86	0	-	-
ppa006437m	19246165	19248673	2509	AT5G39400.1[b]	PTEN1; phosphatase	401	67,58	2E-159	+	Down PTG
ppa018982m	19249067	19251699	2633	AT3G59170.1[a]	F-box/RNI like superfamily protein	108	62	9E-07	+	/
ppa000705m	19253035	19259243	6209	AT5G15400.1[b]	U-box domain-containing protein	1042	77,26	0	-	-
ppa020057m	19270904	19272288	1385	AT1G61500.1	S-locus protein kinase, putative	230	54,35	1E-61	-	-
ppb011385m	19274222	19275245	1024	AT3G47570.1[a]	Leucine-rich repeat protein kinase family	174	42	8E-07	-	-
ppa015373m	19281903	19282828	926	AT2G44970.2	lipase-related	168	63,69	1E-55	+	/
ppa022595m	19284362	19285014	653	AT1G61480.1	S-locus protein kinase, putative	113	56,64	5E-27	-	-
ppa015584m	19293768	19295622	1855	AT5G44940.1[a]	F-box/RNI like superfamily protein	492	26	1E-05	-	-
ppa008118m	19296139	19298487	2349	AT5G15390.1[b]	tRNA/rRNA methyltransferase (SpoU) family protein	273	72,53	2E-114	-	-
ppa027121m	19299176	19300573	1398	AT4G15280.1	UGT71B5 (UDP-GLUCOSYL TRANSFERASE 71B5); UDP-glycosyltransferase/ quercetin 3-O-glucosyltransferase/ transferase, transferring glycosyl groups	491	31,16	6E-49	-	-
ppa015845m	19304869	19306169	1301	AT4G15280.1	UGT71B5 (UDP-GLUCOSYL TRANSFERASE 71B5); UDP-glycosyltransferase/ quercetin 3-O-glucosyltransferase/ transferase, transferring glycosyl groups	443	32,96	8E-45	-	-

ppa005427m	19307491	19308892	1402	AT4G01070.1	GT72B1; UDP-glucosyltransferase/ UDP-glycosyltransferase/ transferase, transferring glycosyl groups	477	31,24	2E-45	+	Down PG
ppa005544m	19309086	19311255	2170	AT5G39410.1[b]	binding / catalytic	444	71,4	1E-170	-	-
ppa001868m	19311528	19313894	2367	AT4G21300.1	pentatricopeptide (PPR) repeat-containing protein	664	34,49	2E-114	-	-
ppa020905m	19318081	19318432	352	AT3G07490.1	AGD11 (ARF-GAP domain 11); calcium ion binding	54	62,96	2E-14	+	Up PG/PTG
ppa013269m	19328280	19329440	1161	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa025951m	19355024	19356771	1748	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa022986m	19372463	19374770	2308	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa018209m	19378653	19379885	1233	AT1G13940.1	unknown protein	275	30,18	2E-10	-	-
ppa026630m	19387745	19390960	3216	AT1G26620.1	unknown protein	934	26,45	4E-43	-	-
ppa010430m	19396610	19397388	779	AT3G29240.2[b]	unknown protein	237	71,31	5E-89	-	-
ppa017633m	19398358	19400169	1812	AT3G29230.1[b]	pentatricopeptide (PPR) repeat-containing protein	570	64,56	0	-	-
ppa008650m	19401227	19404376	3150	AT3G29200.1[b]	CM1 (CHORISMATE MUTASE 1); L-ascorbate peroxidase/ chorismate mutase	340	62,06	3E-107	-	-
ppa012248m	19405752	19407075	1324	AT5G15350.1[b]	plastocyanin-like domain-containing protein	151	57,62	3E-41	-	-
ppa022822m	19407697	19409782	2086	AT5G15870.1	glycosyl hydrolase family 81 protein	670	60,9	0	+	Up PTG
ppa020122m	19410871	19411919	1049	AT2G27035.1	plastocyanin-like domain-containing protein	112	45,54	3E-28	-	-
ppa017224m	19412641	19413914	1274	AT1G69390.1	ATMINE1 (Arabidopsis homologue of bacterial MinE 1); protein binding	215	54,88	9E-55	-	-
ppa006467m	19414851	19417373	2523	AT3G29185.1[b]	unknown protein	382	68,85	9E-152	-	-
ppa003809m	19419766	19423127	3362	AT1G13960.1	WRKY4; DNA binding / transcription factor	478	52,72	4E-107	+	Up PTG
ppa002153m	19425256	19428803	3548	AT5G39420.1[b]	cdc2cAt (Arabidopsis thaliana cdc2c); ATP binding / kinase/ protein kinase/ protein serine/threonine kinase	602	53,99	0	-	-
ppa003330m	19430187	19433082	2896	AT3G29180.1[b]	unknown protein	489	62,99	2E-176	+	/
ppa005577m	19444944	19447021	2078	AT3G29180.1	unknown protein	358	49,16	1E-84	+	/
ppa019328m	19452185	19454055	1871	AT1G61420.1	S-locus lectin protein kinase family protein	180	51,11	2E-37	-	-
ppa024828m	19455075	19455374	300	AT1G48940.1	plastocyanin-like domain-containing protein	57	68,42	7E-19	-	-
ppa013579m	19460383	19462427	2045	AT3G29170.1[b]	unknown protein	121	68,6	5E-30	+	Up PTG
ppa004347m	19463421	19468043	4623	AT3G01090.2[b]	AKIN10 (Arabidopsis SNF1 kinase homolog 10); protein binding / protein kinase	498	86,35	0	+	/
ppa025010m	19468943	19470217	1275	AT5G19790.1	RAP2.11 (related to AP2 11); DNA binding / transcription factor	77	76,62	1E-19	-	-
ppa002110m	19474159	19480251	6093	AT3G01100.1[b]	HYP1 (HYPOTHETICAL PROTEIN 1)	701	62,91	0	+	/
ppa019884m	19483949	19489940	5992	AT3G01100.1	HYP1 (HYPOTHETICAL PROTEIN 1)	747	60,91	0	+	/
ppa011182m	19491797	19494720	2924	AT5G39510.1[b]	SGR4 (SHOOT GRAVITROPSIM 4); receptor	219	73,52	3E-87	+	/
ppa020933m	19494965	19496866	1902	AT5G15340.1[b]	pentatricopeptide (PPR) repeat-containing protein	628	58,92	0	-	-
ppa010375m	19497382	19498954	1573	AT5G39530.1[b]	unknown protein	257	40,08	5E-47	-	-
ppa008873m	19499720	19502738	3019	AT3G29090.1[b]	pectinesterase family protein	312	80,77	5E-151	-	-
ppb013941m	19503679	19505433	1755	AT3G29075.1	glycine-rich protein	98	51,02	1E-17	+	/
ppa005604m	19505477	19508028	2552	AT3G29075.1[b]	glycine-rich protein	141	51,06	4E-23	+	/
ppa004696m	19508444	19510885	2442	AT1G74630.1	pentatricopeptide (PPR) repeat-containing protein	507	37,28	4E-98	-	-
ppb009532m	19510291	19512202	1912	AT1G14010.1	emp24/gp25L/p24 family protein	204	63,73	3E-61	-	-
ppa001662m	19513691	19518100	4410	AT3G29060.1[b]	unknown protein DOMAIN/s: EXS, C-terminal, SPX, N-terminal	824	58,01	0	+	Down PTG

ppa008419m	19518793	19520882	2090	AT5G15330.1[b]	SPX4 (SPX DOMAIN GENE 4)	305	65,57	3E-96	-	-
ppa004232m	19522035	19525797	3763	AT3G01120.1[b]	MTO1 (METHIONINE OVERACCUMULATION 1); cystathionine gamma-synthase	394	87,82	0	+	Up PTG
ppa006769m	19536562	19538472	1911	AT5G15310.1[b]	ATMYB16 (MYB DOMAIN PROTEIN 16); DNA binding / transcription factor	396	53,54	5E-91	-	-
ppa007883m	19548346	19549836	1491	AT5G61430.1	ANAC100 (ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 100); transcription factor	363	57,3	2E-103	-	-
ppa003946m	19554779	19556419	1641	AT5G15300.1[b]	pentatricopeptide (PPR) repeat-containing protein	543	60,22	0	-	-
ppa013251m	19560340	19561689	1350	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa014933m	19562764	19563770	1007	AT5G15300.1	pentatricopeptide (PPR) repeat-containing protein	257	42,02	2E-47	-	-
ppa018277m	19564757	19566145	1389	AT5G49610.1[a]	F-box family	744	23	1E-04	+	/
ppa021574m	19567105	19570089	2985	AT5G15280.1[b]	pentatricopeptide (PPR) repeat-containing protein	978	46,63	0	-	-
ppa014463m	19575091	19575831	741	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa003758m	19576121	19579601	3481	AT5G15270.1[b]	KH domain-containing protein	480	67,08	2E-160	-	-
ppa004961m	19583491	19587280	3790	AT3G01150.1[b]	PTB1 (POLYPYRIMIDINE TRACT-BINDING PROTEIN 1); RNA binding / nucleic acid binding / nucleotide binding	349	85,96	0	-	-
ppa010654m	19591405	19592794	1390	AT2G03090.1	ATEXPA15 (ARABIDOPSIS THALIANA EXPANSIN A15)	227	80,62	9E-77	-	-
ppa015390m	19595392	19597506	2115	AT5G39680.1[b]	EMB2744 (EMBRYO DEFECTIVE 2744)	688	54,22	0	-	-
ppa011701m	19598419	19599223	805	AT5G39670.1[b]	calcium-binding EF hand family protein	203	48,77	4E-38	-	-
ppa014215m	19600973	19601914	942	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa004451m	19604860	19608044	3185	AT5G39660.2[b]	CDF2 (CYCLING DOF FACTOR 2); DNA binding / protein binding / transcription factor	527	43,64	2E-81	+	/
ppa010896m	19611500	19612889	1390	AT3G01170.1[b]	structural constituent of ribosome	224	59,82	4E-64	-	-
ppa002164m	19614676	19616973	2298	AT5G50390.1[b]	pentatricopeptide (PPR) repeat-containing protein	714	61,9	0	-	-
ppb020889m	19617307	19617753	447	AT1G26800.1	zinc finger (C3HC4-type RING finger) family protein	47	55,32	0,0000001	-	-
ppa008635m	19623160	19626937	3778	AT3G28970.1[b]	AAR3 (antiauxin-resistant 3)	304	54,28	9E-80	+	Down PTG
ppa001734m	19627228	19632018	4791	AT3G01180.1$_b$	AtSS2 (starch synthase 2); transferase, transferring glycosyl groups	742	68,19	0	-	-
ppa020648m	19633174	19633872	699	AT1G09157.1[b]	unknown protein	187	71,12	2E-70	-	-
ppa002364m	19634693	19637215	2523	AT5G15250.1[b]	FTSH6 (FTSH PROTEASE 6); ATP-dependent peptidase/ ATPase/ metallopeptidase/ peptidase/ zinc ion binding	651	78,8	0	+	/
ppa004522m	19637658	19639175	1518	AT2G36730.1[b]	pentatricopeptide (PPR) repeat-containing protein	489	62,58	0	-	-
ppa006034m	19639635	19641354	1720	AT3G28960.1[b]	amino acid transporter family protein	405	62,96	2E-144	-	-
ppa016771m	19647980	19650114	2135	AT5G37820.1	NIP4;2 (NOD26-LIKE INTRINSIC PROTEIN 4;2); water channel	241	48,96	1E-58	-	-
ppa000362m	19650301	19659117	8817	AT5G48600.1[b]	ATSMC3 (ARABIDOPSIS THALIANA STRUCTURAL MAINTENANCE OF CHROMOSOME 3); ATP binding / transporter	1238	74,88	0	-	-
ppa016732m	19661696	19662608	913	AT2G02960.5	zinc finger (C3HC4-type RING finger) family protein	80	55	3E-22	+	/

ppa013692m	19664452	19665301	850	AT5G15230.1[b]	GASA4 (GAST1 PROTEIN HOMOLOG 4)	92	68,48	3E-21	-	-
ppa014185m	19665910	19666850	941	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppa008908m	19674589	19675917	1329	AT3G28920.1[b]	AtHB34 (ARABIDOPSIS THALIANA HOMEOBOX PROTEIN 34); DNA binding / transcription factor	256	52,73	6E-48	-	-
ppa013998m	19683618	19684281	664	AT3G28917.1[b]	MIF2 (MINI ZINC FINGER 2); DNA binding	99	69,7	7E-22	-	-
ppa008366m	19691186	19693083	1898	AT3G28910.1[b]	MYB30 (MYB DOMAIN PROTEIN 30); DNA binding / transcription factor	361	51,8	4E-87	-	-
ppa014261m	19703691	19703931	241	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ppb011184m	19705415	19707177	1763	AT1G26880.1[b]	60S ribosomal protein L34 (RPL34A)	95	95,79	1E-48	+	/
ppa005314m	19707885	19711622	3738	AT5G39830.1[b]	DEG8; peptidase/ serine-type peptidase	447	74,72	0	-	-
ppa016710m	19715699	19717197	1499	AT1G26870.1	FEZ (FEZ); transcription factor	216	71,76	5E-82	-	-
ppa001533m	19717958	19720378	2421	AT5G39840.1[b]	ATP-dependent RNA helicase, mitochondrial, putative	808	64,73	0	-	-
ppa017567m	19721911	19727462	5552	AT3G28880.1[b]	protein binding	410	46,1	9E-66	-	-
ppa010320m	19731857	19735527	3671	AT1G18800.1	NRP2 (NAP1-RELATED PROTEIN 2); DNA binding / chromatin binding / histone binding	226	61,5	1E-74	-	-
ppa000039m	19735991	19744143	8153	AT3G55160.1[b]	unknown protein DOMAIN/s: HEAT	2217	58,86	0	-	-
ppa001982m	19750021	19753347	3327	AT1G69670.1	CUL3B (CULLIN 3B); protein binding / ubiquitin-protein ligase	733	81,99	0	+	/
ppa011733m	19756541	19758934	2394	AT5G39850.1[b]	40S ribosomal protein S9 (RPS9C)	179	92,74	5E-96	-	-
ppa027069m	19759943	19760335	393	AT5G15190.2[b]	unknown protein	98	41,84	0,00000001	-	-
ppa005626m	19761253	19763026	1774	AT1G50010.1	TUA2; structural constituent of cytoskeleton	435	98,39	0	+	/
ppa000359m	19767597	19774531	6935	AT3G28860.1[b]	ABCB19; ATPase, coupled to transmembrane movement of substances / auxin efflux transmembrane transporter	1247	90,3	0	-	-

[a] BLASTP TAIR matches with an E-value >$1e^{-6}$ not included by IPGI (E-value <0.05)

[b] Gene pairs that are 'best-reciprocal BLASTP hits' between *Prunus* and *Arabidopsis*

Table S1.6 Primers used in this study to amplify by PCR different fragments corresponding to *S-RNase*, *SFB* and *actin* genes.

Primer	Sequence	6.1 Reference
SRc-F	5'-CTC GCT TTC CTT GTT CTT GC-3'	Romero *et al*. (2004)
SRc-R	5'-GGC CAT TGT TGC ACA AAT TG-3'	Romero *et al*. (2004)
PruC2	5'-CTT TGG CCA AGT AAT TAT TCA AAC C-3'	Tao *et al*. (1999)
PruC2R	5'-GGT TTG AAT AAT TAC TTG GCC ATA G-3'	Tao *et al*. (1999)
PruC4R	5'-GGA TGT GGT ACG ATT GAA GCG-3'	Tao *et al*. (1999)
FBf-Hap1	5'-TGG AAG CAC CAA TTT ATT TCC T-3`	This work
FBr-Hap1	5'-TGA TTG AAG GAT CGA TCA TCT TGG-3'	This work
FBf-Hap2	5´-GCC CAA TTA CTT GGT CAC TG-3´	Vilanova *et al*. (2006)
FBr-Hap2	5´-CAC CCA CTT GAC TTG TCA GC-3´	Vilanova *et al*. (2006)
RT-SFB1-for	5´-GGC AGC TCG AGT TTT GTT AGC ATA C-3´	This work
RT-SFB1-rev1	5´-GGA ACC CGA ATT GGA GAG AAA CGA G-3´	This work
RT-SFB2-for	5´-TTG GCA GCT CAA GTT TTG TTA GTG C-3´	This work
RT-SFB2-rev2	5´-GCA GAA CCC ATA AGT CAG CTT TTC G-3´	This work
Act3	5´-CTT CTT ACT GAG GCA CCC CTG AAT-3´	Gabino Ríos personal comm.
Act4	5´-AGC ATA GAG GGA GAG AAC TGC TTG-3´	Gabino Ríos personal comm.

Table S1.7 SSR markers tested for the PPM screening on the whole 'Katy' genome.

Acronyme	Species	*Number of SSR*	*6.2 Reference*
BPPCT	*P. persica*	7 (4)[a]	Dirlewanger *et al*. (2002)
Gol	*P. armeniaca*	2 (1)	Vera-Ruiz *et al*. (2010)
CPDCT	*P. dulcis*	2 (2)	Mnejja *et al*. (2005)
CPPCT	*P. persica*	5 (4)	Aranzana *et al*. (2002)
CPSCT	*P. salicina*	14 (6)	Mnejja *et al*. (2004)
EPPCU/EPDCU	*P. persica*/*P. dulcis*	10 (5)	Howad *et al*. (2005), http://www.rosaceae.org/
M/MA	*P. persica*	7 (3)	Yamamoto *et al*. (2002)
pchcms/pchgms	*P. persica*	2 (1)	Sosinski *et al*. (2000)
SsrPaCITA	*P. armeniaca*	8 (3)	Lopes *et al*. (2002)
UDAp	*P. armeniaca*	20 (10)	Messina *et al*. (2004), http://www.rosaceae.org/
UDP	*P. persica*	6 (3)	Cipriani *et al*. (1999) and Testolin *et al*. (2000)
UDA	*P. dulcis*	4 (0)	Testolin *et al.* (2004)
AMPA	*P. armeniaca*	1 (1)	Hagen et al. (2004)
UCD-CH	*P. avium*	1 (1)	Struss et al. (2003)
PGS	*P. persica*	29 (11)	Zuriaga et al. (2012)
TOTAL		118 (55)	

[a] Number of polymorphic SSRs in 'Katy' is indicated between brackets.

References SupData

Aranzana MJ, Carbo J, Arus P. (2002) Microsatellite variability in peach [*Prunus persica* (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. Theor Appl Genet 106: 1341–1352.

Cipriani G, Lot G, Huang WG, Marrazzo MT, Peterlunger E, et al. (1999) AC/GT and AG/CT microsatellite repeats in peach [*Prunus persica* (L) Batsch]: isolation, characterisation and cross species amplification in Prunus. Theor Appl Genet 99: 65–72.

Dirlewanger E, Cosson P, Tavaud M, Aranzana J, Poizat C, et al. (2002) Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.). Theor Appl Genet 105: 127-138.

Hagen LS, Chaib J, Fady B, Decroocq V, Bouchet JP, et al. (2004) Genomic and cDNA microsatellites from apricot (*Prunus armeniaca* L.). Mol Ecol Notes 4: 742–745.

Howad W, Yamamoto T, Dirlewanger E, Testolin R, Cosson P, et al. (2005) Mapping with a few plants: using selective mapping for microsatellite saturation of the Prunus reference map. Genetics 171: 1305–1309.

Lopes MS, Sefc KM, Laimer M, da Camara Machado A (2002) Identification of microsatellite loci in apricot. Mol Ecol Notes 2: 24–26.

Messina R, Lain O, Marrazzo MT, Cipriani G, Testolin R (2004) New set of microsatellite loci isolated in apricot. Mol Ecol Notes 4: 432–434.

Mnejja M, García-Mas J, Howad W, Badenes L, Arus P (2004) Simple-sequence repeat (SSR) markers of Japanese plum (*Prunus salicina* Lindl.) are highly polymorphic and transferable to peach and almond. Mol Ecol Notes 4: 163-166.

Mnejja M, García-Mas J, Howad W, Arus P (2005) Development and transportability across *Prunus* species of 42 polymorphic almond microsatellites. Mol Ecol Notes 5: 531-535.

Sosinski B, Gannavarapu M, Hager LD, Beck LE, King GJ, et al. (2000) Characterization of microsatellite markers in peach (*Prunus persica* L. Batsch). Theor Appl Genet 101: 421–424.

Struss D, Ahmad R, Southwick SM, Boritzki M (2003) Analysis of sweet analysis of sweet cherry (*Prunus avium* L.) cultivars using SSR and AFLP markers. J Am Sic Hort Sci 128: 904-909.

Testolin R, Marrazzo MT, Cipriani G, Quarta R, Verde I, et al. (2000) Microsatellite DNA in peach (*Prunus persica* L. Batsch) and its use in fingerprinting and testing the genetic origin of cultivars. Genome 43: 512–520.

Testolin R, Messina R, Lain O, Marrazzo MT, Huang WG, et al. (2004) Microsatellites isolated in almond from an AC-repeat enriched library. Mol Ecol Notes 4: 459–461

Vera-Ruiz EM, Soriano JM, Romero C, Zhebentyayeva T, Terol J, et al. (2010) Narrowing down the apricot *plum pox virus* resistance locus and comparative analysis with the peach genome syntenic region. Mol Plant Pathol 12: 535-47.

Yamamoto T, Mochida K, Imai T, Shi YZ, Ogiwara I, et al. (2002) Microsatellite markers in peach [*Prunus persica* (L.) Batsch] derived from an enriched genomic and cDNA libraries. Mol Ecol Notes 23: 298-301.

Zuriaga E , Molina L, Badenes ML, Romero C (2012) Physical mapping of a pollen modifier locus controlling self-incompatibility in apricot and synteny analysis within the Rosaceae. Plant Mol Biol 79: 229–242.

Supporting information chapter 2

Table S2.1. Primers used in this study

Name	Sequence	Reference
SRc-F	5´-CTC GCT TTC CTT GTT CTT GC -3´	Romero et al. (2004)
SRc-R	5´-GGC CAT TGT TGC ACA AAT TG -3´	Vilanova et al. (2005)
Pru-T2	5´-GTT CTT GCT TTT GCT TTC TTC-3´	Tao et al. (1999)
Pru-C2	5´-CTT TGG CCA AGT AAT TAT TCA AAC C-3´	Tao et al. (1999)
Pru-C2R	5´-GGT TTG AAT AAT TAC TTG GCC ATA G-3´	Tao et al. (1999)
Pru-C4R	5´-GGA TGT GGT ACG ATT GAA GCG-3	Tao et al. (1999)
Pru-C6R	5´-CAT TGC CAC TTT CCA CGT C-3´	Vilanova et al. (2003)
F-BOX5´A	5´-TTK SCH ATT RYC AAC CKC AAA AG -3´	Vaughan et al. (2005)
F-BOXintronR	5´-CWG GTA GTC TTD SYA GGA TG- 3´	Vaughan et al. (2005)
RFBc-F	5´-GAG GAG TGC TAC AAA CTA AGC-3´	Vilanova et al. (2006)
SFBins-R	5´-TCA AGA ACT TGG TTG GAT TCG-3´	Vilanova et al. (2006)
Sf-Hap2	5´-CGC TAG AAA TCA AAG CCA CAG-3´	Vilanova et al. (2006)
Sr-Hap2	5´-GGC GTA AGC AAG TGG AAA AG-3´	Vilanova et al. (2006)
FBf-Hap2	5´-GCC CAA TTA CTT GGT CAC TG-3´	Vilanova et al. (2006)
FBr-Hap2	5´-CAC CCA CTT GAC TTG TCA GC-3´	Vilanova et al. (2006)
FBF1	5´-GCC CAA TTA CTT GGT CAC TG-3´	unpublished
SFBc-F	5´-TCG ACA TCC TAG TAA GAC TAC CTG C-3´	Romero et al. (2004)
FBF5	5´-TAG GAC CCC TCA AAT GAG C-3´	unpublished
FBF6	5´-TGG GTT CTG CAA GAA AAA CG-3´	unpublished
FBr-Hap2-2	5´-AAA AGC AAC AGC CAC CAA AG-3´	unpublished

Table S2.2. Fragment sizes determined for *S-RNase* and *SFB* alleles

S-allele	SR1-F/SR1-R	PruT2/SR1-R	PruC2/PruC4R[d]	PruC2/PruC6R[d]	F-Box	Refs.[e]
S_1	419	407	~2100	~2000	210	[1]
S_2	345	n.a.	~950	~800	205	[1]
S_4	260	244	~400	~350	n.a.	[1]
S_5	396	384	~1300	~1200	n.a.	[1]
S_6	n.a.[c]	429	~1300	~1200	197	[1]
S_7	419	407	~900	~700	199	[1]
S_8	371	359	~3000	~2800	204	[2]
S_9	218	206	~600	~400	n.a.	[2]
S_{11}	320	308	n.a.	~1500	208	[2]
S_{20}	n.a.	n.a.	~1900	~1800	n.a.	[3]
S_{24}	n.a.	266	~450	~300	n.a.	[4]
S_{29}	427	417	~750	~550	189	[5]
S_{30}	373	n.a.	n.a.	~500	206	[5]
S_{31}	285	n.a.	n.a.	n.a.	207	[5]
S_C	371	359	~3000	~2800	204	[1]
S_V/S_X[a]	260/n.a.	248/n.a.	~2000/550	~1800/350	n.a./n.a.	[5]
S_{15}/S_Z[b]	346/338	334/326	~750 (S_{15})/400 (S_Z)	~550 (S_{15})/200(S_Z)	198/206	[2]

[a] Allele fragment sizes corresponding to S_V and S_X could not be established since both were found only once and in the same cultivar ('Fergani') (see Table 3.3).
[b] Only allele fragment sizes corresponding to the second intron could be unambiguously assigned to S_{15} and S_Z according to Halász et al. (2005).
[c] n.a. Not amplified.
[d] Fragment sizes for the second *S-RNase* intron were approximately estimated (~) from agarose gels.
[e] References reporting *S*-allele molecular sizes for the first time: [1] Vilanova et al. (2005); [2] Halász et al. (2005); [3] Zhang et al. (2008); [4] Gu et al. (2013); [5] This work.

Table S2.3. Characteristics of SSR primers developed from peach (PGS) and apricot genome sequences (AGS) located at the *M*-locus.

Name	F/R	Primer sequence	Repeat motif	Start on scaffold_3 (Mb)[3]	ORF (Prupe)	Size range (bp)	No. of alleles	*H*[4]
PGS3.22[1]	F	TCTGATTGCAGGTAAGGACAG	$(CT)_{25}$	18,49	3G247600	304-328	8	0,61
	R	TATCTTGATATCGGCCTGGA			Put. Prot.			
PGS3.23[1]	F	TGACTTTCTGCATCTTGACCT	$(AG)_{24}$	18,61	3G249300	164-190	8	0,68
	R	CTTTGCTTCCGTTAATCCAA			MADS-box			
PGS3.62[1]	F	AGCTTCCTCTATTCTTGGTGGT	$(CT)_{22}$	18,61	----	321-356	10	0,71
	R	GCTTTTCCCCGAGCTAATTC						
PGS3.71[1]	F	ACCACCCCCTATCCCTATTG	$(CT)_{13}$	18,4	----	233-269	10	0,65
	R	ACTTGCAAACCCCCTTGATT						
PGS3.96[1]	F	TGGCCACAATTAATGGGAGA	$(CT)_{14}$	18,76	----	431-474	15	0,82
	R	TCGGAGAACTTCTTGTGCAT						
AGS3.20[2]	F	CGAACGAGAGGGAAAAATGA	$(TA)_{10}$	18,61	3G249300	178-202	8	0,62
	R	AACTGATTCCGAACCACAGG			MADS-box			
AGS3.30[2]	F	CCGCACGGCTATACTGTCTAA	$(AT)_{13}$	18,71	----	193-207	7	0,38
	R	ACAGGCTGGATGCTTTGTCT						

[1] Zuriaga et al. (2013)
[2] This work
[3] Positions according to the peach v1.0 genome sequence (IPGI)
[4] Heterozygosity was estimated according to Nei (1973): $H = 1-\Sigma p_i^2$, where p_i is the frequency of the i^{th} allele. Clonal sibs from 'Canino' were excluded from estimations.

Table S2.4. Allele sizes for SSR markers comprised in the *M*-haplotypes

	PGS3_71	PGS3_22	PGS3_62	PGS3_23	AGS_20	AGS_30	PGS3_96
*m*0-0	261	306	348	188	190	203	442
*m*0-1	261	306	348	188	190	203	444
*M*1-0	259	306	350	188	188	203	441
*M*1-1	257	306	348	188	188	203	441
*M*1-2	259	306	348	188	188	203	441
*M*1-3	259	306	348	178	188	203	441
*M*1-4	257	306	348	184	188	203	441
*M*2-0	255	310	336	178	188	203	434
*M*2-1	255	310	336	178	188	201	434
*M*2-2	255	310	336	178	188	203	436
*M*3	247	312	329	190	192	195	458
*M*4-0	259	306	336	178	188	195	456
*M*4-1	259	306	336	178	188	195	458
*M*4-2	259	306	336	178	188	195	460
*M*5-0	233	310	354	184	190	203	466
*M*5-1	233	310	356	184	192	n.a.	466
*M*5-2	233	310	354	184	190	n.a.	464
*M*6	269	304	321	164	202	203	443
*M*7-0	251	316	332	186	198	193	464
*M*7-1	233	316	332	186	198	193	464
*M*7-2	233	316	332	186	196	193	464
*M*7-3	251	316	332	186	194	195	466
*M*8-0	259	328	354	186	198	195	431
*M*8-1	259	328	354	186	196	195	431
*M*8-2	259	328	356	186	196	195	431
*M*9	261	312	356	184	196	n.a.	466
*M*10	247	312	354	178	188	203	474
*M*11	233	314	332	190	188	199	474
*M*12	259	312	356	184	194	203	468
*M*13	255	310	348	188	188	203	441
*M*14-0	255	306	348	188	188	195	466
*M*14-1	255	306	348	188	188	195	450
*M*15-0	259	306	356	186	192	203	466
*M*15-1	257	306	354	186	192	203	466
*M*16	243	310	337	180	188	195	442
*M*17	256	308	348	168	178	203	434
*M*18	257	310	354	190	188	207	468
*M*19	255	306	334	188	190	n.a.	n.a.

Supporting information chapter 3

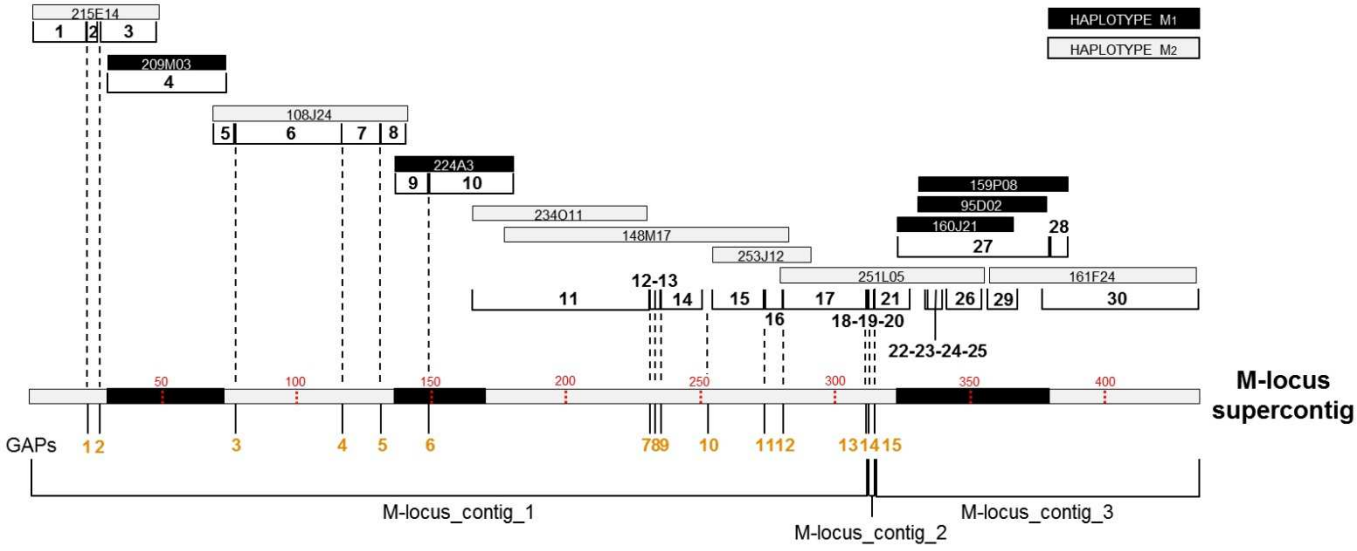


Figure S3.1. BAC clone *de novo* assembly, joining of overlapping BAC contigs and GAP closure. *Boxed* codes represent BAC clones belonging to *M*-locus from a BAC library of the SI apricot cv. 'Goldrich' (Zuriaga et al., 2012) used for *de novo M*-locus sequence assembly. *Grey* and *black* colors correspond to M_1 and M_2-haplotypes, respectively. Contigs obtained per BAC clone (BAC contigs) were numbered correlatively from 1 to 30 (*open white boxes*; see Table S3.2). GAPs (previous to GAP closure; see Table S3.3) between contiguous contigs are shown below a*M*-supercontig with *orange numbers*. Contigs conforming a*M*-supercontig after GAP closure (M-locus_contig_1,2 and 3; *open white boxes*) are shown below GAPs. The scale in Kb of a*M*-supercontig is indicated with *red dotted lines*.

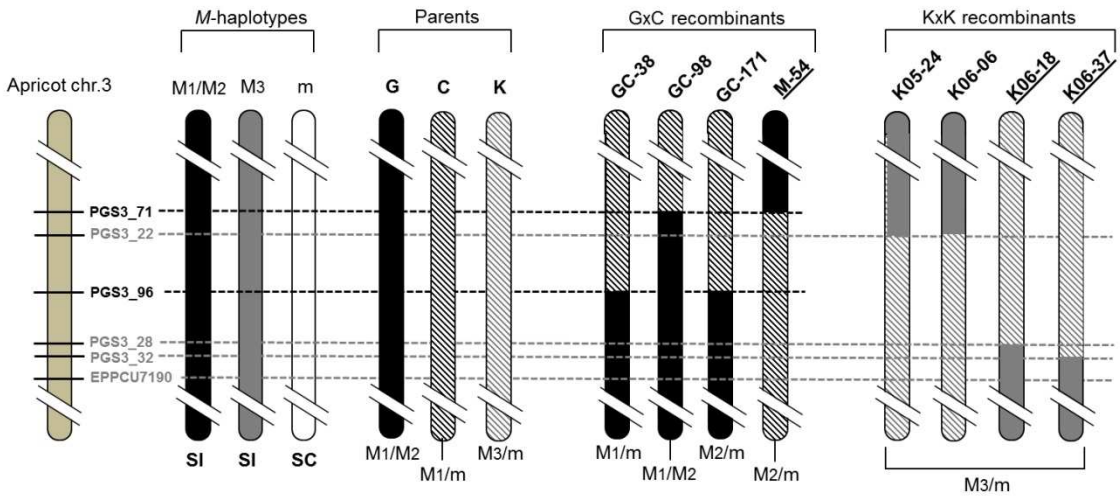


Figure S3.2. Graphical *M*-locus map representation of parents and recombinant hybrids used for 'Canino' ('GxC') and 'Katy' ('KxK') fine-mapping. *Brown vertical bar* represents *M*-locus region in apricot chr. 3. Molecular markers delimiting 'Canino' and 'Katy' *M*-loci are written in *black* and *grey* in the apricot chr.3, respectively. *Black*, *grey* and *white vertical bars* represent M_1/M_2, M_3 and *m*-haplotypes, respectively. *Black and white stripped lines* inside *vertical bars* symbolize heterozygote M_1m or M_2m *M*-locus genotype in the corresponding region, while *grey and white stripped lines* symbolize heterozygote M_3m genotype. *M*-locus genotype for the parents 'Goldrich' (G), 'Canino' (C) and 'Katy' (K), as well as 'GxC' and 'KxK' recombinant hybrids are indicated below each *vertical bar*. Recombinant breakpoints for 'GxC' and 'KxK' recombinants hybrids delimiting 'Canino' and 'Katy' *M*-locus maps are shown by *horizontal black* and *grey dotted lines*, respectively. New recombinant hybrids incorporated in this work are underlined.

Table S3.1. Summary of Next Generation Sequencing (NGS) data. NGS platform, DNA source, Sample and Tissue (for RNAseq data) as well as the Number of sequences and Average size of raw data and cleaned data after trimming are indicated.

NGS platform	DNA source	Sample	Tissue	Raw data		Cleaned data	
				No. sequences	Average size (bp)	No. sequences	Average size (bp)
454	BAC clone	215E14	-	29334	399,79	21909	377,86
454	BAC clone	209M03	-	39135	408,92	28137	387,91
454	BAC clone	108J24	-	27174	393,91	20890	370,36
454	BAC clone	224A3	-	46679	402,86	34777	385,5
454	BAC clone	234O11	-	16768	384,81	13094	363,84
454	BAC clone	148M17	-	24553	299,37	23738	296,9
454	BAC clone	253J12	-	9262	287,63	8969	284,8
454	BAC clone	251L05	-	16348	292,08	15718	288,9
454	BAC clone	160J21	-	19125	306,26	18481	301,7
454	BAC clone	95D02	-	9374	299,29	9065	299,2
454	BAC clone	159P08	-	9937	295,54	9676	291,6
454	BAC clone	161F24	-	10233	293,92	9931	291,6
illumina	gDNA	Goldrich'	-	137954275	101	136391075	92,7
illumina	gDNA	Canino'	-	373801518	101	371672380 (129438652)[a]	99,3 (99,6)
illumina	gDNA	Katy'	-	69669448	101	69042494	98,8
illumina	RNA	Goldrich'	mature anthers*	122397834	107	122338874	106,46
illumina	RNA	Goldrich'	mature styles**	122313850	107	122268676	106,52
illumina	RNA	Goldrich'	leaves**	135741242	107	135688624	106,59
illumina	RNA	Canino'	mature anthers*	159854696	107	159774857	106,47
illumina	RNA	Canino'	mature styles**	123966784	107	124313835	106,33
illumina	RNA	Canino'	leaves**	110887662	107	110843339	106,56
illumina	RNA	Katy'	mature anthers*	130685722	107	130624896	106,51
illumina	RNA	Katy'	leaves**	103966200	107	103922997	106,57

[a]Due to the high coverage, 1/3 of cleaned sequences were randomly selected

* 3 biological replicates / 2 technical replicates per biological replicate

** 2 biological replicates / 2 technical replicates per biological replicate

Table S3.2. *De novo* assembly results per BAC clone (previous to join overlapping contigs and GAP closure).

BAC clone	Contig name	Code*	size (bp)	BAC clone	Contig name	Code*	size (bp)
215E14	Contig_215E14-1	1	19926	148M17/253J12/251L05	Consensus_3	16	6640
	Contig_215E14-2	2	5830		Consensus_4	17	31108
	Contig_215E14-3	3	37296	251L05	Contig_251L05-1	18	1196
209M03	Contig_209M03	4	45069		Contig_251L05-2	19	2480
108J24	Contig_108J24-1	5	8134		Contig_251L05-3	20	13053
	Contig_108J24-2	6	39955		Contig_251L05-4	21	1077
	Contig_108J24-3	7	13930		Contig_251L05-5	22	314
	Contig_108J24-4	8	10284		Contig_251L05-6	23	2304
224A3	Contig_224A3-1	9	13179		Contig_251L05-7	24	2707
	Contig_224A3-2	10	32027		Contig_251L05-8	25	1833
234O11/148M17	Consensus_1	11	66623		Contig_251L05-9	26	19476
148M17	Contig_148M17-1	12	1783	160J21/95D02/159P08	Consensus_5	27	57003
	Contig_148M17-2	13	2034		Contig_159P08-1	28	7160
	Contig_148M17-3	14	15728	161F24	Contig_161F24-1	29	12056
148M17/253J12	Consensus_2	15	19959		Contig_161F24-2	30	58817

*Code for contigs obtained per BAC clone (BAC contigs) represented in Figure S3.1

Table S3.3. GAP closure. All GAPs between contiguous contigs and their position in *M*-locus supercontig sequence are indicated. GAPs matching with SSR markers from PGS3 series are also indicated. For remaining GAPs, markers primer pair sequences used for amplification are shown.

GAP	**Position**	**SSR coincidence**	**oligo sequence-forward / reverse**
GAP-1	20031	PGS3_21	-
GAP-2	25831	-	GGAATTGTGGAAATGGGAGA / AGGTTGCGTGAGCTCTCTTT
GAP-3	76026	-	TTCGGCTTCCAATCATAAGG / AATGCGGGACTATGAAGACG
GAP-4	115850	PGS3_22	-
GAP-5	130249	PGS3_77	-
GAP-6	148469	PGS3_47	-
GAP-7	231180	PGS3_62	-
GAP-8	232832	-	CCCAACACTCATCGAACCTT / TTGAGGAGGTCAATCCCATC
GAP-9	234847	PGS3_84	-
GAP-10	251854	PGS3_49	-
GAP-11	273812	PGS3_88	-
GAP-12	280476	-	ACTGCCATGTTACGCAATGA / AATTGGTGTGGAGCATGTGA
GAP-13	311576	-	CGACCGGCTATACACTGTCTTT / GCTTGTAGAACCTCTAGGAACTATCG
GAP-14	313614	-	CCTACGTACCCTACTAAGGGATCAA / CTAATCGTATGTGGCGCAAA
GAP-15	314867	-	AGGTGGAAGTTTTGGGGAAT / GGTTCCACTCCTGTCAATCG

Table S3.4. SSR primers developed from the apricot *M*-locus supercontig sequence. Primer position, sequence, repeat motif and SSR allele sizes amplified in apricot cvs. 'Goldrich', 'Canino' and 'Katy' are indicated.

Name	F/R	Primer sequence	Repeat motif	Start on	'Goldrich' alleles	'Canino' alleles	'Katy' alleles
AGS.3	F	AAAATGTTGGGCTCCCTTTC	(TTC)7	37610	164	164	164
	R	TGAACGACTTGGGGGAATAG					
AGS.4	F	TTGGCATCTCTGGTGCAAT	(AT)7	32804	438/470	470	441/470
	R	ACAATGAGGTTGCCTTCGTC					
AGS.6	F	GAGTGGCCGATACCTGTTCT	(AATT)4	70573	238/241	241	238/241
	R	AATGATGGGTTTTGGGTGTG					
AGS.7	F	TTCGGCTTCCAATCATAAGG	(TC)14	76026	N.A.	N.A.	N.A.
	R	AGAAATGGAGGTGTCGTTGG					
AGS.8	F	TTCGTAGCATTCTGGGGTTT	(GA)10	102236	215/230	250	238/250
	R	GGGGGCTTGAATGATAGGAT					
AGS.9	F	AGGCATGTGTGTTTGACACC	(AAT)5	103440	225/226	225	225
	R	AATGTGGACATGAAGCACCA					
AGS.10	F	CTCCCATGGAAAACCTCAAA	(CT)30	115967	198/203	198	198/203
	R	GGGGCATTTCTGATGGTAAA					
AGS.11	F	TTTGCCTTCATACACCTAGCC	(AT)10	116715	261/269	261	261
	R	CACAAGCATGAGACCATCCA					
AGS.12	F	ACGATGAATTTGAAGACGATGA	(CT)9+(TTG)6	136201	193/210	193/210	193/210
	R	ACCTTCACTGCCAAATTCCCTATC					
AGS.14	F	AGAAGGCCCTGCACCTAAAT	(CCT)6	152359	219/236	219/236	219/236
	R	CATAAACTCAGGGGCTTGGA					
AGS.17	F	AAAAACACCTCTCCCGACAA	(TA)6	177634	190/199	199	190/199
	R	AGCGGCGATACTCGTTTTAC					
AGS.18	F	CAATGGACGAGTAGGGGTGT	(AT)12	175974	387/389	389	387/389
	R	TTGGGTTTGGAGAGGTTTTG					
AGS.19	F	TATCATGCGTCGCTCTCAAG	(AT)10	208676	235/251	251	235/251
	R	CACAATTGGATGTCGAAACG					
AGS.20	F	CGAACGAGAGGGAAAAATGA	(AT)10	224961	189	189/191	191/193
	R	AACTGATTCCGAACCACAGG					
AGS.21	F	TGTGTCCCTCGATCCTTACC	(TA)10	236517	514/518	518	503/505
	R	CTATCCGATTTCCAATCCGACA					
AGS.22	F	AGTTCAAGCGGCTTTCAGAT	(TA)4+5	244325	171	171	171
	R	AATGCCAGTCCTTCGATGAG					
AGS.23	F	TACAATCAATGGCGGATTCA	(TA)8	250156	N.A.	N.A.	N.A.
	R	TTTCTTCGTCTGAGCCTTTGA					
AGS.24	F	TCCAAAAGAAGCAACGTCAA	(GA)23	250524	N.A.	N.A.	N.A.
	R	CCATGCTTGGGTTAAAGTGG					
AGS.26	F	AATATTGGTCCCCCTCCAAG	(GTT)4+5	252416	240/258	240/258	240/258
	R	GCAAGAGAAACGAAAAGCTCA					
AGS.27	F	GTTGCACGGAAATTCCAGAT	(AG)14	275668	175/182	175	175/182
	R	GTGTGCGTCTGTGTGGGTAG					
AGS.28	F	GGGTCCTCAACAGACCAAAG	(GA)9	276953	179/182	179/182	N.A.
	R	AGGTGCACGTGGATAGACCT					
AGS.29	F	ACGTCGTTTTGGCAATGTTT	(ATA)4	295820	N.A.	N.A.	N.A.
	R	ACATGTGCCCTTTGTTTGTG					
AGS.30	F	CCGCACGGCTATACTGTCTAA	(AT)13	309620	203	203	195/203
	R	ACAGGCTGGATGCTTTGTCT					
AGS.31	F	AATTGCCCCCGTCTATCAC	(CT)5+5	313614	194/196	196	194/196
	R	GAGAATGGGTGGGGTAGGAC					
AGS.32	F	CCCAGCTGAAATGGGAATAC	(AT)11	315102	282/297	282/297	282/297
	R	GCATGCATCATGTTTTCCTG					
AGS.33	F	CACCCCCTCCCTCTCTTTTA	(CT)10	319478	ML	ML	ML

	R	CATGTTGGTCGATTTGTAGCC					
AGS.34	F	TCACCAGCTGACGTGGTAGT	(CT)16	320217	N.A.	N.A.	N.A.
	R	CAATTCCTCATCTGGGCAGT					
160J21-7	F	ACTTGAGATTGATGCTCCCATT	(AT)11	332558	164	164	164
	R	ACCAACAGCTCCAAATTAAACC					
AGS.35	F	CAGGCCTCAAAGGCAAAAC	(CTAGGCGGCT)21	335737	N.A.	N.A.	N.A.
	R	CACCCCCTCCCTCTCTTTTA					
160J21-6	F	CCTTCACCAACTTCAAACCCTA	(GA)6	336648	189	189/191	191/193
	R	TTGTTCCTATTTTCGATACCCG					
160J21-5	F	CTACTGCTGAACGACCAAAACA	(AT)11+(CA)6	337460	223/238	223/238	223/238
	R	AACGGATTTTCATGGTAGATGC					
160J21-4	F	CCTCTCTCACTCAACCTGCTCT	(TC)18	338507	N.A.	N.A.	N.A.
	R	AAGCGTTTAGCCAAGGAACTAA					
AGS.36	F	ACCCAGAGGTACCCTTCGAG	(CTAGGCGGCT)14	343974	ML	ML	ML
	R	ACTTCCATCACCCTTCGTCA					
160J21-3	F	TGTGAAGGTCATGGGTTTACAA	(GT)9	347903	N.A.	397/403	397/403
	R	ACGGTTTTCCAAGTACAACGTC					
160J21-2	F	GGTTGGACTGCTTTTCATTCTT	(TAA)16	349292	350/352	350/354	352/354
	R	ATTTCTTTGGAGTTGAGGTGGA					
AGS.37	F	TCAAATCTCTTGGGCCAATC	(GGT)6	350097	256/264	264	264/270
	R	ATTCACTACCCCCACAACCA					
AGS.38	F	CATCATGTACGGAAGCACCA	(AT)12	353246	ML	219/221	219/221
	R	CCGTTGGACATTCCTTTTTC					
AGS.39	F	CTCGCGAAACCCTAACATTT	(TC)9+9	338506	N.A.	N.A.	N.A.
	R	ACCGGGAGAAAACGACAGT					
AGS.40	F	CATCATGTACGGAAGCACCA	(AT)12	353246	N.A.	222	220/222
	R	CCGTTGGACATTCCTTTTTC					
AGS.41	F	ATGGAAGATGATTGCCCAAC	(AT)15	366351	ML	342	342/344
	R	TTGTCATGTTGATGCCCTGT					

Table S3.5. Example of variant calling results from a*M*-supercontig using the alignment of genomic *Illumina* data from apricot cvs. 'Goldrich' (a), 'Canino' (b) and 'Katy' (c). First and last five variants are shown from the total variant calling. Variant position, type ('SNV'=Single-Nucleotide Variant; 'MNV'=Multi-Nucleotide Variant; 'Insertion'; 'Deletion'; 'Replacement'), reference allele in the a*M*-supercontig sequence ('Reference' column), variant allele identified ('Allele' column), zigosity, the number of sequences supporting variant allele ('Count' column), total of sequences aligned in variant position ('Coverage' column), proportion of forward/reverse sequences supporting variant allele ('Forward/reverse' column) and Phred score average value for variant position ('Average quality' column) are shown for each variant.

a

SNP number	Reference Position	Type	Length	Reference	Allele	Zygosity	Count	Coverage	Frequency	Forward/ reverse	Average quality
1	354	SNV	1	A	C	Heterozygous	5	11	45,45	0,00	32,00
2	356	SNV	1	A	C	Heterozygous	6	12	50,00	0,00	32,00
3	542	SNV	1	A	T	Heterozygous	12	23	52,17	0,33	37,67
4	631	SNV	1	T	A	Heterozygous	10	20	50,00	0,40	38,10
5	969	SNV	1	T	C	Heterozygous	13	24	54,17	0,46	36,23
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:
5100	435483	SNV	1	G	A	Heterozygous	24	67	35,82	0,00	36,17
5101	435489	SNV	1	G	T	Heterozygous	22	63	34,92	0,00	34,95
5102	435495	SNV	1	A	G	Heterozygous	20	63	31,75	0,00	37,95
5103	435522	MNV	2	CT	AC	Heterozygous	14	52	26,92	0,00	38,18
5104	435771	SNV	1	A	T	Heterozygous	3	11	27,27	0,00	6,33

b

SNP number	Reference Position	Type	Length	Reference	Allele	Zygosity	Count	Coverage	Frequency	Forward/ reverse	Average quality
1	1320	Deletion	1	T	-	Homozygous	48	48	100,00	0,40	36,42
2	1339	Deletion	1	G	-	Homozygous	48	48	100,00	0,42	35,27
3	1349	Deletion	1	C	-	Homozygous	44	45	97,78	0,43	33,39
4	9749	Insertion	1	-	T	Homozygous	30	30	100,00	0,48	36,53
5	13757	Insertion	1	-	T	Homozygous	34	34	100,00	0,32	36,41
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:
5293	435564	SNV	1	C	T	Heterozygous	23	65	35,38	0,04	39,65
5294	435565	Deletion	1	A	-	Heterozygous	36	65	55,38	0,28	36,75
5295	435571	SNV	1	C	A	Heterozygous	20	62	32,26	0,00	39,50
5296	435576	MNV	2	CC	TT	Heterozygous	17	57	29,82	0,00	39,32
5297	435592	SNV	1	T	C	Heterozygous	14	53	26,42	0,00	39,14

c

SNP number	Reference Position	Type	Length	Reference	Allele	Zygosity	Count	Coverage	Frequency	Forward/ reverse	Average quality
1	354	SNV	1	A	C	Heterozygous	5	9	55,56	0,20	34,60
2	356	SNV	1	A	C	Heterozygous	5	9	55,56	0,20	36,60
3	542	SNV	1	A	T	Heterozygous	9	20	45,00	0,40	38,11
4	631	SNV	1	T	A	Heterozygous	9	20	45,00	0,22	39,00
5	891	SNV	1	A	C	Heterozygous	13	30	43,33	0,46	38,00
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:
6593	435522	MNV	2	CT	AC	Heterozygous	14	35	40,00	0,07	35,89
6594	435559	SNV	1	C	T	Heterozygous	13	29	44,83	0,00	38,92
6595	435565	Deletion	1	A	-	Heterozygous	10	26	38,46	0,30	37,10
6596	435576	SNV	1	C	T	Heterozygous	8	24	33,33	0,00	39,50
6597	435592	SNV	1	T	C	Heterozygous	5	20	25,00	0,00	38,80

Table S3.6. Selected SNPs for 'Canino' and 'Katy' fine-mapping. Primer position, sequence and allele composition observed by Sanger sequencing in apricot cvs. 'Goldrich', 'Canino' and 'Katy' are indicated.

Name	F/R	Primer sequence	Position	'Goldrich' alleles	'Canino' alleles	'Katy' alleles
SNPCaMmap1	F	TAATGTGAGTCTTGGACGTG	33687	33687 (T/G), 33688 (C/T)	33687 (G), 33688 (C/T)	-
	R	CTGTCCTTTTTGGATTCCTGA				
SNPCaMmap2	F	AGAAACGCCACACCACACTA	112413	No heterozygote	No heterozygote	-
	R	ATTGGGACTGGTGTCTGAGC				
SNPCaMmap3	F	TTGGGGATAAGTGGAGTTGG	184272	G/T	T	-
	R	TCAGCTGGGTTCTTCACCTT				
SNPCaMmap4	F	CGAAAGGCCTCTCTATGCTG	260267	Indel	Indel	-
	R	TCGTGCACCAAGTGCATTAT				
SNPCaMmap5	F	TGGCTCTGTGTACCATCCAA	326175	326175 (A/G), 326182 (G/T)	326175 (G/G), 326182 (G/T)	-
	R	TTTGTGGGCAGTTAACACCA				
SNPKaMmap1	F	CAAGCAAGGGGCAATTAACA	142155	-	-	A/G
	R	CGCTAACACCAGAGGAAACTG				
SNPKaMmap2	F	GGTGTTCATCAGAAGCAGCA	146316	-	-	C/T
	R	CATGTTCATTCAACGGCATA				
SNPKaMmap3	F	ACGTCTCATTTCATCCCTGGT	153066	-	-	T/C
	R	GGCTGCAGAAAGAACATGAAG				
SNPKaMmap4	F	GCAAGAGGTCAACACCAAAAG	164682	-	-	A/C
	R	CTCAAAAGGCTGTTGCTCTGT				
SNPKaMmap5	F	TGCCGACTATCAACAGTAAACC	171071	-	-	G/A
	R	GACATGCATCTTCCTTGAGA				
SNPKaMmap6	F	AGCCACCATGCACCCTATAC	273675	-	-	A/G
	R	TCACATGGTAACCAAGCTCCT				
SNPKaMmap7	F	CACGAGGGCCTCTATTTTGT	276184	-	-	T/C
	R	CTCCTTTTGGTGCATGTGTG				
SNPKaMmap8	F	AATGTGTTTGGACAAGTCACG	277881	-	-	C/T
	R	CACACTTCACTCCAACCGAAT				
SNPKaMmap9	F	GGCTAATGTGCAAGAGGTTTG	285823	-	-	C/T
	R	GGGAGAGAAGTATGCAGAGCA				
SNPKaMmap10	F	CCCGTTTTGGAGAATAGAAGAC	295239	-	-	A/G
	R	CCTATGGAGATAGGTTCCTTGA				

Table S3.7. Example of the comparison among 'Canino', 'Katy' and 'Goldrich' variants, and the predicted amino acid changes, in the ~134 Kb *M*-locus region within the a*M*-supercontig. Extracted variants of those found between positions 142.155 and 276.184 (~134 Kb *M*-locus region) within the a*M*-supercontig for cvs. (a) 'Goldrich', (b) 'Canino' and (c) 'Katy'. Variant position, type ('SNV'=Single-Nucleotide Variant; 'MNV'=Multi-Nucleotide Variant; 'Insertion'; 'Deletion'; 'Replacement'), algorithm used ('Analysis' column; 'Basic'=Basic variant detection algorithm and 'InDel'/'SV'=Structural Variant detection algorithm), reference allele in a*M*-supercontig sequence ('Reference' column), variant allele identified ('Allele' column) and zygosity are indicated. In variants containing 'SV' in 'Analysis' column, zygosity can not be determined, therefore, these variants were considered as heterozygous. Furthermore, 'Canino' and 'Katy' variants distinct to 'Goldrich' ('d' and 'e', respectively) are shown. These both tables contain the same columns mentioned previously for (a), (b) and (c), and additionally two columns indicating whether each variant is located in a gene ('Coding region change' column) and also if this leads to a non-synonymous change ('Non-synonymous' column). A variant included in a gene but no amino acid change is described means that this variant is located in an intronic or UTR region. In all tables, the three first and last variants as well as the *FaSt* insertion (*grey shadow*) and some of their bordering variants are shown.

a

Reference Position	Type	Analysis	Length	Reference	Allele	Zygosity
142155	SNV	Basic	1	A	G	Heterozygous
142169	SNV	Basic	1	G	A	Heterozygous
142173	SNV	Basic	1	A	G	Heterozygous
:	:	:	:	:	:	:
:	:	:	:	:	:	:
215454	Insertion	Basic	3	-	TTT	Heterozygous
215454	SNV	Basic	1	T	A	Heterozygous
215470	SNV	Basic	1	G	C	Heterozygous
215615	Replacement	Basic	2	G	TT	Heterozygous
215616	Insertion	Basic	1	-	T	Heterozygous
215628	SNV	Basic	1	C	T	Heterozygous
:	:	:	:	:	:	:
:	:	:	:	:	:	:
276044	Insertion	Basic	2	-	TA	Heterozygous
276238	SNV	Basic	1	A	C	Heterozygous
276415	SNV	Basic	1	A	T	Heterozygous

b

Reference Position	Type	Analysis	Length	Reference	Allele	Zygosity
147311	SNV	Basic	1	G	A	Heterozygous
147357	Deletion	Basic	1	A	-	Heterozygous
147406	SNV	Basic	1	T	C	Heterozygous
:	:	:	:	:	:	:
:	:	:	:	:	:	:
214011	SNV	Basic	1	C	T	Homozygous
214079	SNV	Basic	1	T	G	Homozygous
214578^214587	Insertion	SV	0			---
214600	SNV	Basic	1	C	T	Homozygous
214955	SNV	Basic	1	T	G	Homozygous
214988	SNV	Basic	1	G	A	Homozygous
:	:	:	:	:	:	:
:	:	:	:	:	:	:
275668	Deletion	Basic	6	AGAGAG	-	Homozygous
275707	Replacement	Basic	2	GG	T	Homozygous
276044	Insertion	Basic	2	-	TA	Homozygous

c

Reference Position	Type	Analysis	Length	Reference	Allele	Zygosity
142155	SNV	Basic	1	A	G	Heterozygous
142169	SNV	Basic	1	G	A	Heterozygous
142173	SNV	Basic	1	A	G	Heterozygous
:	:	:	:	:	:	:
:	:	:	:	:	:	:
214011	SNV	Basic	1	C	T	Heterozygous
214079	SNV	Basic	1	T	G	Heterozygous
214578^214587	Insertion	SV	0			---
214600	SNV	Basic	1	C	T	Heterozygous
214895	Insertion	Basic	1	-	C	Heterozygous
214955	SNV	Basic	1	T	G	Heterozygous
:	:	:	:	:	:	:
:	:	:	:	:	:	:
276027	SNV	Basic	1	T	C	Heterozygous
276044	Insertion	Basic	2	-	TA	Heterozygous
276184	SNV	Basic	1	T	C	Heterozygous

d

Region	Type	Analysis	Reference	Allele	Length	Zygosity	Coding region change	Non-synonymous
159426	SNV	Basic	A	C	1	Heterozygous	PaM-2	Yes
169270^169271	Insertion	InDel	-	TGTGG…	128	Heterozygous	PaM-2	-
169383	SNV	Basic	A	G	1	Heterozygous	PaM-2	Yes
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
208675^208676	Insertion	Basic	-	ATATAT	6	Heterozygous		-
210974^210975	Insertion	InDel	-	TTT	3	Homozygous	PaM-6	-
214578^214587	Insertion	SV			0	-	PaM-7	Yes
215954^215955	Insertion	Basic	-	(AT)11	22	Homozygous		-
215982	SNV	Basic	C	A	1	Homozygous		-
218991	SNV	Basic	C	T	1	Heterozygous		-

:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
273812..273825	Deletion	Basic	(GA)7	-	14	Heterozygous		-
273812..273829	Deletion	InDel	(GA)9	-	18	Heterozygous		-
273851..273868	Deletion	InDel	(GA)9	-	18	Heterozygous		-

e

Region	Type	Analysis	Reference	Allele	Length	Zygosity	Coding region change	Non-synonymous
142598	SNV	Basic	G	C	1	Heterozygous		-
142909	SNV	Basic	T	C	1	Heterozygous		-
143345	SNV	Basic	G	C	1	Heterozygous		-
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
208675^208676	Insertion	Basic	-	ATATAT	6	Heterozygous		-
208676..208677	Deletion	Basic	AT	-	2	Heterozygous		-
214578^214587	Insertion	SV			0	---	PaM-7	Yes
214894^214895	Insertion	Basic	-	C	1	Heterozygous	PaM-7	-
218222	SNV	Basic	C	A	1	Heterozygous	PaM-8	Yes
218268	SNV	Basic	T	A	1	Heterozygous		-
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
275250..275251	MNV	Basic	AA	TT	2	Heterozygous		-
276027	SNV	Basic	T	C	1	Heterozygous		-
276184	SNV	Basic	T	C	1	Heterozygous		-

Table S3.8. RBH results for selected proteins from the PaMDOr BLASTP output. Hits with e-values lower than 10^{-3} from BLASTP analysis of the proteins shown in Table 4.3 against *P. persica*, *M. domestica* and *F.vesca* (a), and *S. lycopersicum*, *Nicotiana* and *A. thaliana* (b) taxids from NCBI protein database are shown in columns 'Hit' for each specie (e-value and identity are indicated). *Bold* NCBI accession ID is shown in column 'Query'; the ID for each accession in corresponding genome protein database (see materials and methods) is shown below NCBI accession ID inside *brackets*.

a

Specie		Prunus persica			Malus domestica			Fragaria vesca		
	Query	Hit	E-value	Ident	Hit	E-value	Ident	Hit	E-value	Ident
Prunus persica	**XP_007216055.1**	XP_007216055.1	1,E-156	100%	XP_008379454.1	4,E-98	62%	XP_004304201.1	9,E-98	61%
	(ppa017665m)	XP_007215948.1	3,E-99	62%	XP_008341809.1	3,E-95	59%	XP_004306054.1	3,E-87	63%
	(PaMDOr)									
	XP_007215948.1	XP_007215948.1	9,E-161	100%	XP_008341809.1	8,E-141	87%	XP_004304201.1	9,E-131	80%
	(ppa011285m)	XP_007216055.1	3,E-99	62%	XP_008379454.1	7,E-139	86%	XP_004306054.1	4,E-85	60%
Malus domestica	**XP_008379454.1**	XP_007215948.1	3,E-139	86%	XP_008379454.1	1,E-159	100%	XP_004304201.1	4,E-128	80%
	(MDP0000233548)	XP_007216055.1	2,E-98	62%	XP_008341809.1	4,E-146	91%	XP_004306054.1	4,E-83	58%
	XP_008341809.1	XP_007215948.1	4,E-141	87%	XP_008341809.1	4,E-161	100%	XP_004304201.1	2,E-124	75%
	(MDP0000148485)	XP_007216055.1	2,E-95	59%	XP_008379454.1	4,E-146	91%	XP_004306054.1	1,E-83	58%
Fragaria vesca	**XP_004304201.1**	XP_007215948.1	8,E-131	80%	XP_008379454.1	7,E-128	80%	XP_004304201.1	3,E-166	100%
	(gene04226-v1.0-hybrid)	XP_007216055.1	8,E-98	61%	XP_008341809.1	4,E-124	75%	XP_004306054.1	2,E-82	58%
	XP_004306054.1	XP_007216055.1	3,E-87	63%	XP_008341809.1	2,E-83	58%	XP_004306054.1	9,E-165	100%
	(gene04224-v1.0-hybrid)	XP_007215948.1	4,E-85	60%	XP_008379454.1	8,E-83	58%	XP_004304201.1	2,E-82	58%
Solanum lycopersicum	**XP_004232135.1**	XP_007215948.1	8,E-109	68%	XP_008341809.1	5,E-110	67%	XP_004304201.1	5,E-101	63%
	(Solyc02g089230.2.1)	XP_007216055.1	5,E-85	54%	XP_008379454.1	1,E-108	69%	XP_004306054.1	6,E-72	51%
Nicotiana	**XP_009774377.1**	XP_007215948.1	3,E-108	70%	XP_008341809.1	6,E-109	70%	XP_004304201.1	2,E-101	65%
		XP_007216055.1	2,E-85	54%	XP_008379454.1	1,E-107	69%	XP_004306054.1	9,E-73	51%
	XP_009608607.1	XP_007215948.1	8,E-110	70%	XP_008341809.1	1,E-108	68%	XP_004304201.1	2,E-101	64%
		XP_007216055.1	4,E-86	54%	XP_008379454.1	6,E-108	69%	XP_004306054.1	4,E-74	51%
	XP_009608606.1	XP_007215948.1	1,E-109	70%	XP_008341809.1	3,E-108	68%	XP_004304201.1	3,E-101	63%
		XP_007216055.1	6,E-86	54%	XP_008379454.1	1,E-107	69%	XP_004306054.1	1,E-73	51%
Arabidopsis thaliana	**NP_198706.1**	XP_007215948.1	2,E-103	65%	XP_008341809.1	5,E-100	63%	XP_004304201.1	7,E-101	63%
	(AT5G38900.1)	XP_007216055.1	9,E-84	55%	XP_008379454.1	2,E-97	63%	XP_004306054.1	1,E-68	49%

b

Specie		Solanum lycopersicum			Nicotiana			Arabidopsis thaliana		
	Query	Hit	E-value	Ident	Hit	E-value	Ident	Hit	E-value	Ident
Prunus persica	**XP_007216055.1**	XP_004232135.1	6,E-85	54%	XP_009608607.1	1,E-85	54%	NP_198706.1	2,E-83	55%
	(ppa017665m)				XP_009608606.1	2,E-85	54%			
	(PaMDOr)				XP_009774377.1	6,E-85	54%			
	XP_007215948.1	XP_004232135.1	1,E-108	68%	XP_009608607.1	2,E-109	70%	NP_198706.1	5,E-103	65%
	(ppa011285m)				XP_009608606.1	4,E-109	70%			
					XP_009774377.1	8,E-108	70%			
Malus domestica	**XP_008379454.1**	XP_004232135.1	7,E-109	69%	XP_009608607.1	1,E-107	69%	NP_198706.1	3,E-97	63%
	(MDP0000233548)				XP_009608606.1	2,E-107	69%			
					XP_009774377.1	2,E-107	69%			
	XP_008341809.1	XP_004232135.1	3,E-110	67%	XP_009774377.1	9,E-109	70%	NP_198706.1	6,E-100	63%
	(MDP0000148485)				XP_009608607.1	2,E-108	68%			
					XP_009608606.1	4,E-108	68%			
Fragaria vesca	**XP_004304201.1**	XP_004232135.1	5,E-101	63%	XP_009608607.1	5,E-101	64%	NP_198706.1	1,E-100	63%
	(gene04226-v1.0-hybrid)				XP_009774377.1	7,E-101	65%			
					XP_009608606.1	8,E-101	63%			
	XP_004306054.1	XP_004232135.1	7,E-72	51%	XP_009608607.1	1,E-73	51%	NP_198706.1	3,E-68	49%
	(gene04224-v1.0-hybrid)				XP_009608606.1	3,E-73	51%			
					XP_009774377.1	3,E-72	51%			
Solanum lycopersicum	**XP_004232135.1**	XP_004232135.1	4,E-177	100%	XP_009608606.1	3,E-157	89%	NP_198706.1	1,E-104	63%
	(Solyc02g089230.2.1)				XP_009608607.1	5,E-150	91%			
					XP_009774377.1	6,E-150	92%			
Nicotiana	**XP_009774377.1**	XP_004232135.1	2,E-150	92%	XP_009774377.1	7,E-162	100%	NP_198706.1	9,E-103	65%
					XP_009608606.1	1,E-153	95%			
					XP_009608607.1	1,E-153	95%			
	XP_009608607.1	XP_004232135.1	2,E-150	91%	XP_009608607.1	5,E-162	100%	NP_198706.1	7,E-104	65%
					XP_009608606.1	7,E-162	100%			
					XP_009774377.1	1,E-153	95%			
	XP_009608606.1	XP_004232135.1	1,E-157	89%	XP_009608606.1	2,E-177	100%	NP_198706.1	4,E-104	63%
					XP_009608607.1	8,E-162	100%			
					XP_009774377.1	1,E-153	95%			
Arabidopsis thaliana	**NP_198706.1**	XP_004232135.1	8,E-105	63%	XP_009608606.1	6,E-104	63%	NP_198706.1	2,E-161	100%
	(AT5G38900.1)				XP_009608607.1	9,E-104	65%			
					XP_009774377.1	1,E-102	65%			

Supporting information chapter 4

Table S4.1. RBH results for NaTrxh. Proteins obtained from 'direct BLASTP' analysis (see Table 4.1) are the queries (column 'protein') for RBH. The result of all-vs-all BLASTP analysis with their corresponding e-values and percentages of identity (three first hits) are shown. Color assignments for hits are the same than indicated in table 4.1.

Specie			P. persica			M. domestica			S. lycopersicum			N. benthamiana			A. thaliana		
	Protein	Hit	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val
P. persica	ppa011576m	1st	ppa011576m	100	0	MDP0000752795	82	6E-54	Solyc02g087630.2.1	63	7E-41	NbS00020764g0013.1	68	1E-46	AT5G39950.1	65	8E-45
	205aa	2nd	ppa011861m	100		MDP0000448333	76	6E-48	Solyc05g006830.2.1	59	9E-40	NbS00027633g0013.1	63	5E-39	AT1G45145.1	49	5E-28
		3rd	ppa013299m	55	6E-37	MDP0000391509	55	1E-37	Solyc05g006860.2.1	58	1E-38	NbS00010261g0005.1	55	3E-35	AT1G19730.1	44	9E-27
M. domestica	MDP0000752795	1st	ppa011576m	75	4E-54	MDP0000752795	100	3E-61	Solyc05g006830.2.1	58	4E-39	NbS00020764g0013.1	59	7E-44	AT5G39950.1	71	7E-43
	133aa	2nd	ppa013299m	57	3E-35	MDP0000448333	84	4E-53	Solyc02g087630.2.1	66	5E-39	NbS00027633g0013.1	54	4E-38	AT1G45145.1	52	8E-29
		3rd	ppa011861m	69	2E-31	MDP0000391509	51	5E-35	Solyc05g006860.2.1	56	5E-38	NbS00010261g0005.1	59	3E-37	AT1G69880.1	46	1E-27
S. lycopersicum	Solyc02g087630.2.1	1st	ppa011576m	63	2E-36	MDP0000752795	60	6E-35	Solyc02g087630.2.1		2E-77	NbS00020764g0013.1	86	3E-53	AT5G39950.1	64	9E-37
	169aa	2nd	ppa013299m	56	4E-30	MDP0000448333	50	2E-33	Solyc05g006860.2.1		5E-35	NbS00027633g0013.1	80	1E-40	AT1G19730.1	47	2E-20
		3rd	ppa011861m	60	5E-28	MDP0000391509	50	5E-28	Solyc05g006830.2.1		7E-35	NbS00018815g0006.1	57	4E-33	AT5G42980.1	49	2E-19
N. benthamiana	NbS00020764g0013.1	1st	ppa011576m	70	3E-43	MDP0000752795	69	2E-41	Solyc02g087630.2.1	88	2E-51	NbS00020764g0013.1	100	2E-65	AT5G39950.1	69	1E-43
	142aa	2nd	ppa013299m	62	8E-40	MDP0000448333	68	4E-40	Solyc05g006830.2.1	72	2E-44	NbS00027633g0013.1	93	2E-59	AT1G19730.1	47	4E-28
		3rd	ppa013161m	47	1E-29	MDP0000391509	59	8E-38	Solyc05g006860.2.1	67	3E-42	NbS00010261g0005.1	64	6E-40	AT1G45145.1	50	8E-28
A. thaliana	AT5G39950.1	1st	ppa011576m	63	1E-45	MDP0000752795	63	1E-43	Solyc02g087630.2.1	62	2E-45	NbS00020764g0013.1	63	2E-48	AT5G39950.1	100	8E-74
	133aa	2nd	ppa013299m	53	4E-39	MDP0000448333	64	1E-42	Solyc05g006830.2.1	48	2E-37	NbS00027633g0013.1	64	3E-39	AT1G59730.1	41	6E-27
		3rd	ppa011861m	62	1E-27	MDP0000391509	52	3E-37	Solyc05g006860.2.1	46	9E-36	NbS00010261g0005.1	53	6E-35	AT1G69880.1	44	2E-26

Table S4.2. RBH results for SBP1. Proteins obtained from 'direct BLASTP' analysis (see Table 4.1) are the queries (column 'protein') for RBH. The result of all-vs-all BLASTP analysis with their corresponding e-values and percentages of identity (three first hits) are shown. Color assignments for hits are the same than indicated in table 4.1.

Specie			P. persica			M. domestica			S. lycopersicum			N. benthamiana			A. thaliana		
	Protein	Hit	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	% id	Match	% id	e-val
P. persica	ppa008290m	1st	ppa008290m	100.00	0.0	MDP0000522795	88.27	0.0	Solyc04g078760.2.1	67.35	4E-165	NbS00055742g0004.1	68.71	1E-169	AT1G45976.1	63.69	8E-149
	338aa	2nd	ppa008184m	39.04	1E-47	MDP0000717791	82.40	0.0	Solyc05g005210.2.1	40.00	5E-47	NbS00016021g0006.1	66.37	7E-161	AT1G60610.3	36.21	3E-40
		3rd	ppa007884m	30.43	2E-28	MDP0000650075	77.88	1E-170	Solyc03g112860.2.1	37.43	1E-26	NbS00020248g0004.1	38.89	4E-46	AT1G60610.2	36.21	3E-40
M. domestica	MDP0000522795	1st	ppa008290m	87.10	0.0	MDP0000522795	100.00	0.0	Solyc04g078760.2.1	66.18	2E-154	NbS00055742g0004.1	68.80	2E-160	AT1G45976.1	63.16	2E-141
	447aa	2nd	ppa008184m	40.93	1E-44	MDP0000717791	85.09	0.0	Solyc05g005210.2.1	40.36	6E-45	NbS00016021g0006.1	65.89	5E-152	AT1G60610.3	38.30	1E-39
		3rd	ppa007884m	31.53	9E-29	MDP0000650075	81.58	3E-180	Solyc03g112860.2.1	35.08	6E-25	NbS00020248g0004.1	40.09	2E-43	AT1G60610.2	38.30	1E-39
S. lycopersicum	Solyc04g078760.2.1	1st	ppa008290m	68.22	2E-165	MDP0000522795	65.60	7E-154	Solyc04g078760.2.1	100.00	0.0	NbS00055742g0004.1	87.57	0.0	AT1G45976.1	61.45	3E-137
	338aa	2nd	ppa008184m	36.84	7E-51	MDP0000717791	61.16	6E-132	Solyc05g005210.2.1	34.73	4E-49	NbS00016021g0006.1	84.91	0.0	AT1G60610.3	37.86	7E-45
		3rd	ppa007884m	30.43	1E-25	MDP0000650075	60.06	3E-127	Solyc03g112860.2.1	36.62	1E-29	NbS00020248g0004.1	40.17	1E-48	AT1G60610.2	37.86	7E-45
N. benthamiana	NbS00055742g0004.1	1st	ppa008290m	68.42	2E-168	MDP0000522795	68.51	1E-158	Solyc04g078760.2.1	87.57	0.0	NbS00055742g0004.1	100.00	0.0	AT1G45976.1	62.08	8E-138
	337aa	2nd	ppa008184m	37.89	3E-51	MDP0000717791	63.95	1E-137	Solyc05g005210.2.1	34.41	2E-46	NbS00016021g0006.1	94.96	0.0	AT1G60610.3	38.06	2E-45
		3rd	ppa007884m	31.49	7E-27	MDP0000650075	63.80	5E-134	Solyc03g112860.2.1	35.05	2E-28	NbS00020248g0004.1	40.08	9E-48	AT1G60610.2	38.06	2E-45
A. thaliana	AT1G45976.1	1st	ppa008290m	63.78	2E-148	MDP0000522795	63.24	1E-140	Solyc04g078760.2.1	61.96	5E-136	NbS00055742g0004.1	60.40	3E-137	AT1G45976.1	100.00	0.0
	325aa	2nd	ppa008184m	38.84	9E-44	MDP0000717791	59.69	9E-120	Solyc05g005210.2.1	38.30	4E-45	NbS00016021g0006.1	59.69	7E-129	AT1G60610.3	36.86	2E-43
		3rd	ppa007884m	29.57	2E-26	MDP0000650075	56.04	5E-107	Solyc08g007120.2.1	33.72	1E-30	NbS00020248g0004.1	38.89	6E-48	AT1G60610.2	36.86	2E-43

Table S4.3. RBH results for MdABCF. Proteins obtained from 'direct BLASTP' analysis (see Table 4.1) are the queries (column 'protein') for RBH. The result of all-vs-all BLASTP analysis with their corresponding e-values and percentages of identity (three first hits) are shown. Color assignments for hits are the same than indicated in table 4.1.

Specie			P. persica			M. domestica			S. lycopersicum			N. benthamiana			A. thaliana		
	Protein	Hit	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val
P. persica	ppa002137m	1st	ppa002137m	100	0	MDP0000899854	93,67	0	Solyc08g082850.2.1	82,05	0	NbS00014920g0008.1	78,69	0	AT1G64550.1	82,7	0
	711aa	2nd	ppa002097m	45,14	2E-154	MDP0000170302	94,8	0	Solyc11g069090.1.1	44,59	3E-157	NbS00014920g0008.1	77,81	5E-169	AT5G60790.1	43,43	1E-159
		3rd	ppa003175m	43,49	2E-153	MDP0000477774	43,85	5E-155	Solyc07g008610.1.1	45,5	5E-157	NbS00014920g0008.1	29,73	2E-10	AT3G54540.1	42,56	8E-156
M. domestica	MDP0000170302	1st	ppa002137m	94,8	0	MDP0000170302	100	0	Solyc08g082850.2.1	80,93	0	NbS00014920g0008.1	77,38	0	AT1G64550.1	82,19	0
	711aa	2nd	ppa002097m	45,31	9E-155	MDP0000899854	97,47	0	Solyc07g008610.1.1	45,84	2E-156	NbS00014920g0008.1	77,81	4E-169	AT5G60790.1	43,08	2E-159
		3rd	ppa003175m	42,94	2E-152	MDP0000477774	43,49	4E-154	Solyc11g069090.1.1	43,85	5E-155	NbS00014920g0008.1	30,63	2E-10	AT3G54540.1	45,33	7E-154
S. lycopersicum	Solyc08g082850.2.1	1st	ppa002137m	82,05	0	MDP0000899854	79,94	0	Solyc08g082850.2.1	100	0	NbS00014920g0008.1	92,96	0	AT1G64550.1	79,33	0
	716aa	2nd	ppa003175m	42,15	1E-148	MDP0000170302	80,93	0	Solyc06g074940.2.1	43,53	6E-153	NbS00014920g0008.1	87,5	0	AT5G60790.1	44,08	6E-154
		3rd	ppa002097m	42,71	4E-148	MDP0000477774	42,7	8E-151	Solyc11g069090.1.1	43,43	1E-152	NbS00014920g0008.1	31,73	5E-10	AT3G54540.1	42,77	3E-152
N. benthamiana	NbS00014920g0008.1	1st	ppa002137m	78,69	0	MDP0000899854	76,26	0	Solyc08g082850.2.1	92,96	0	NbS00014920g0008.1	100	0	AT1G64550.1	76,11	0
	815aa	2nd	ppa003175m	41,4	2E-84	MDP0000170302	77,38	0	Solyc06g074940.2.1	43,55	1E-87	NbS00027404g0011.1	42,74	8E-87	AT5G60790.1	43,55	7E-86
		3rd	ppa002097m	38	4E-81	MDP0000477774	42,2	3E-86	Solyc07g008610.1.1	39,23	2E-87	NbS00050078g0006.1	42,74	1E-86	AT3G54540.1	39,64	3E-85
A. thaliana	AT1G64550.1	1st	ppa002137m	82,7	0	MDP0000899854	81,49	0	Solyc08g082850.2.1	79,33	0	NbS00014920g0008.1	76,11	0	AT1G64550.1	100	0
	715aa	2nd	ppa003175m	43,2	3E-152	MDP0000170302	82,19	0	Solyc11g069090.1.1	44,3	7E-155	NbS00014920g0008.1	77,32	4E-169	AT5G60790.1	44,04	2E-154
		3rd	ppa002097m	43,03	1E-148	MDP0000477774	43,93	4E-155	Solyc07g008610.1.1	41,31	2E-154	NbS00014920g0008.1	32,43	7E-11	AT3G54540.1	42,02	1E-151

Table S4.4. RBH results for 120K. Proteins obtained from 'direct BLASTP' analysis (see Table 4.1) are the queries (column 'protein') for RBH. The result of all-vs-all BLASTP analysis with their corresponding e-values and percentages of identity (three first hits) are shown. Color assignments for hits are the same than indicated in table 4.1.

Specie			P. persica			M. domestica			S. lycopersicum			N. benthamiana			A. thaliana		
	Protein	Hit	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val
P. persica	ppa021281m	1st	ppa021281m	100.00	3E-115	MDP0000287357	86.06	3E-85	Solyc02g078040.2.1	62.99	8E-44	NbS00006956g0008.1	55.84	6E-46	AT2G34700.1	51.35	3E-46
	166aa	2nd	ppa024679m	29.59	1E-11	MDP0000165381	66.84	4E-71	Solyc02g078100.2.1	41.40	2E-30	NbS00008703g0009.1	58.52	1E-42	AT1G28290.2	52.99	4E-39
		3rd	ppb017797m	32.80	7E-09	MDP0000617024	34.46	3E-17	Solyc02g078050.2.1	42.86	7E-20	NbS00025834g0007.1	54.62	2E-33	AT1G28290.1	52.99	8E-39
M. domestica	MDP0000165381	1st	ppa021281m	86.06	4E-82	MDP0000287357	100.00	0.0	Solyc02g078040.2.1	58.11	4E-40	NbS00006956g0008.1	52.83	6E-43	AT2G33790.1	42.54	1E-40
	223aa	2nd	ppa024679m	28.65	5E-08	MDP0000165381	76.26	4E-84	Solyc02g078100.2.1	39.15	5E-31	NbS00008703g0009.1	54.09	4E-39	AT2G34700.1	51.75	3E-40
		3rd	ppa022106m	29.23	1E-07	MDP0000423907	34.90	9E-15	Solyc02g078050.2.1	42.86	3E-18	NbS00003320g0020.1	51.90	2E-34	AT1G28290.1	54.41	6E-39
S. lycopersicum	Solyc02g078050.2.1	1st	ppa021281m	43.55	6E-19	MDP0000287357	42.86	4E-19	Solyc02g078050.2.1	100.00	0.0	NbS00008703g0009.1	46.85	2E-26	AT2G34700.1	40.77	2E-19
	363aa	2nd	ppa022106m	32.11	6E-08	MDP0000263610	34.55	1E-09	Solyc02g078060.1.1	57.66	6E-33	NbS00006956g0008.1	46.31	5E-26	AT2G33790.1	37.69	5E-18
						MDP0000423907	31.82	6E-09	Solyc02g078100.2.1	48.03	7E-31	NbS00007980g0003.1	42.77	2E-24	AT3G09925.1	30.71	1E-09
N. benthamiana	NbS00025834g0007.1	1st	ppa021281m	48.55	5E-34	MDP0000287357	45.22	2E-38	Solyc02g078100.2.1	60.36	1E-85	NbS00025834g0007.1	100.00	0.0	AT2G33790.1	41.54	2E-32
	311aa	2nd	ppa024679m	28.39	7E-07	MDP0000165381	52.55	7E-33	Solyc02g078040.2.1	58.21	6E-43	NbS00007980g0003.1	75.09	2E-122	AT2G34700.1	40.25	2E-30
		3rd	ppb017797m	31.82	3E-05	MDP0000617024	29.61	2E-08	Solyc02g078050.2.1	48.99	6E-28	NbS00006956g0008.1	55.28	6E-45	AT1G28290.1	43.28	1E-22
A. thaliana	AT2G34700.1 |	1st	ppa021281m	53.44	5E-44	MDP0000287357	52.67	5E-39	Solyc02g078040.2.1	51.52	4E-37	NbS00006956g0008.1	47.80	1E-34	AT2G34700.1	100.00	5E-126
	175aa	2nd	ppa024679m	29.29	4E-09	MDP0000165381	49.62	7E-39	Solyc02g078100.2.1	42.36	1E-30	NbS00008703g0009.1	50.35	3E-34	AT1G28290.2	49.64	3E-35
		3rd	ppa019712m	39.29	5E-08	MDP0000547052	29.63	2E-11	Solyc02g078050.2.1	38.89	2E-19	NbS00003320g0020.1	48.23	7E-31	AT1G28290.1	49.64	5E-35

Table S4.5. RBH results for NaStEP. Proteins obtained from 'direct BLASTP' analysis (see Table 4.1) are the queries (column 'protein') for RBH. The result of all-vs-all BLASTP analysis with their corresponding e-values and percentages of identity (three first hits) are shown. Color assignments for hits are the same than indicated in table 4.1.

Specie			P. persica			M. domestica			S. lycopersicum			N. benthamiana			A. thaliana		
	Protein	Hit	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val
P. persica	ppa011496m	1st	ppa011496m	100.00	8E-151	MDP0000326576	38.39	7E-35	Solyc03g020010.1.1	65.00	1E-88	NbS00009480g0031.1	66.04	8E-93	AT1G17860.1	60.30	2E-80
	208aa	2nd	ppa011653m	89.38	2E-99	MDP0000619608	39.15	9E-35	Solyc06g072230.1.1	53.06	8E-72	NbS00017403g0024.1	63.06	5E-90	AT1G73260.1	47.98	1E-48
		3rd	ppa011448m	40.09	2E-37	MDP0000318079	39.22	7E-32	Solyc06g072220.1.1	57.39	5E-70	NbS00049946g0001.1	60.44	2E-81	AT1G73325.1	39.15	1E-27
M. domestica	MDP0000326576	1st	ppa011448m	81.34	8,00E-123	MDP0000326576	100.00	3E-146	Solyc03g020010.1.1	36.50	7E-32	NbS00009480g0031.1	34.45	5E-30	AT1G17860.1	36.00	5E-25
	209aa	2nd	ppa011496m	38.39	7,00E-35	MDP0000619608	99.04	2E-145	Solyc06g072230.1.1	34.16	4E-24	NbS00021566g0007.1	37.36	2E-27	AT1G73260.1	31.16	1E-17
		3rd	ppa011653m	42.57	1,00E-25	MDP0000635659	67.65	3E-90	Solyc06g072220.1.1	32.40	5E-23	NbS00017403g0024.1	33.18	3E-27	AT1G73325.1	30.36	4E-09
S. lycopersicum	Solyc03g098710.1.1	1st	ppa011496m	32.37	3E-23	MDP0000635659	30.85	1E-14	Solyc03g098710.1.1	100.00	3E-166	NbS00018395g0002.1	39.82	2E-31	AT1G17860.1	30.88	1E-15
	224aa	2nd	ppa011653m	31.14	1E-13	MDP0000318079	30.00	1E-13	Solyc06g072230.1.1	33.86	1E-24	NbC24872723g0001.1	38.22	2E-26			
		3rd	ppa011448m	29.17	2E-13	MDP0000326576	26.94	5E-13	Solyc03g019690.1.1	38.39	2E-24	NbS00018395g0011.1	37.44	1E-25			
N. benthamiana	NbS00018395g0002.1	1st	ppa011496m	40.86	3E-32	MDP0000635659	32.61	1E-14	Solyc03g020010.1.1	39.67	2E-35	NbS00018395g0002.1	100.00	9E-175	AT1G17860.1	34.78	2E-24
	209aa	2nd	ppa011653m	39.58	5E-21	MDP0000326576	30.85	3E-14	Solyc03g019690.1.1	42.15	2E-35	NbC24872723g0001.1	86.79	1E-131	AT1G73260.1	34.74	3E-24
		3rd	ppa011448m	30.43	7E-15	MDP0000619608	30.85	6E-14	Solyc06g072230.1.1	37.10	4E-32	NbS00018395g0011.1	82.43	7E-124	AT1G73325.1	32.17	4E-15
A. thaliana	AT1G17860.1	1st	ppa011496m	60.30	2E-80	MDP0000326576	36.00	4E-25	Solyc03g020010.1.1	56.10	7E-74	NbS00009480g0031.1	51.14	2E-69	AT1G17860.1	100.00	9E-144
	196aa	2nd	ppa011653m	59.09	8E-57	MDP0000619608	35.50	6E-24	Solyc06g072220.1.1	55.62	3E-59	NbS00017403g0024.1	50.45	3E-69	AT1G73260.1	37.56	2E-36
		3rd	ppa011448m	35.32	9E-29	MDP0000635659	34.24	1E-22	Solyc06g072230.1.1	50.51	1E-58	NbS00049946g0001.1	53.04	2E-62	AT1G73325.1	30.52	6E-16

Table S4.6. RBH results for NaPCCP. Proteins obtained from 'direct BLASTP' analysis (see Table 4.1) are the queries (column 'protein') for RBH. The result of all-vs-all BLASTP analysis with their corresponding e-values and percentages of identity (three first hits) are shown. Color assignments for hits are the same than indicated in table 4.1.

Specie			P. persica			M. domestica			S. lycopersicum			N. benthamiana			A. thaliana		
	Protein	Hit	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val	Match	% id	e-val
P. persica	ppa012133m	1st	ppa012133m	100.00	6E-131	MDP0000525794	95.58	3E-123	Solyc03g118720.2.1	78.79	2E-97	NbS00009334g0006.1	81.21	4E-98	AT3G17980.1	74.86	1E-92
	182aa	2nd	ppa012128m	100.00	6E-131	MDP0000776395	92.78	2E-121	Solyc03g118710.2.1	81.21	2E-96	NbS00051736g0004.1	80.61	1E-97	AT1G48590.2	74.85	2E-88
		3rd	ppa012140m	100.00	6E-131	MDP0000259615	73.94	2E-88	Solyc12g040800.1.1	76.83	2E-91	NbS00020564g0001.1	76.69	2E-93	AT1G48590.1	74.85	3E-88
M. domestica	MDP0000525794	1st	ppa012133m	95.58	3E-123	MDP0000525794	100.00	3E-131	Solyc03g118720.2.1	80.61	7E-100	NbS00009334g0006.1	83.03	1E-100	AT3G17980.1	78.16	3E-95
	182aa	2nd	ppa012128m	95.58	3E-123	MDP0000776395	91.76	2E-121	Solyc03g118710.2.1	83.03	5E-99	NbS00051736g0004.1	82.42	3E-100	AT1G48590.2	76.65	7E-91
		3rd	ppa012140m	95.58	3E-123	MDP0000259615	75.15	9E-91	Solyc12g040800.1.1	78.66	4E-95	NbS00020564g0001.1	77.30	5E-95	AT1G48590.1	76.65	8E-91
S. lycopersicum	Solyc12g040800.1.1	1st	ppa012133m	76.83	2E-91	MDP0000525794	78.66	3E-95	Solyc12g040800.1.1	100.00	2E-118	NbS00020637g0006.1	87.88	2E-100	AT1G48590.2	78.40	3E-88
	166aa	2nd	ppa012128m	76.83	2E-91	MDP0000776395	77.44	5E-92	Solyc03g118720.2.1	76.83	5E-94	NbS00020564g0001.1	81.82	2E-97	AT1G48590.1	78.40	3E-88
		3rd	ppa012140m	76.83	2E-91	MDP0000259616	70.12	9E-83	Solyc06g068940.2.1	80.49	1E-92	NbS00009334g0006.1	81.71	3E-96	AT3G17980.1	75.61	3E-87
N. benthamiana	NbS00020637g0006.1	1st	ppa012133m	69.15	1E-87	MDP0000525794	70.21	3E-89	Solyc12g040800.1.1	87.88	2E-100	NbS00020637g0006.1	100.00	5E-133	AT3G17980.1	71.18	3E-81
	196aa	2nd	ppa012128m	69.15	1E-87	MDP0000776395	70.81	1E-87	Solyc06g068940.2.1	78.92	1E-87	NbS00009698g0010.1	81.87	5E-98	AT1G48590.1	71.95	1E-78
		3rd	ppa012140m	69.15	1E-87	MDP0000259616	66.46	5E-73	Solyc03g118720.2.1	75.30	1E-86	NbS00020564g0001.1	78.92	3E-91	AT1G48590.2	71.95	1E-78
A. thaliana	AT3G17980.1	1st	ppa012133m	74.86	1E-92	MDP0000525794	78.16	3E-95	Solyc03g118720.2.1	74.55	3E-92	NbS00020564g0001.1	79.14	5E-93	AT3G17980.1	100.00	4E-127
	177aa	2nd	ppa012128m	74.86	1E-92	MDP0000776395	71.43	1E-93	Solyc03g118710.2.1	74.55	3E-89	NbS00051736g0004.1	77.58	1E-92	AT1G48590.2	85.03	2E-98
		3rd	ppa012140m	74.86	1E-92	MDP0000259616	70.12	1E-80	Solyc06g068940.2.1	74.55	9E-88	NbS00009334g0006.1	77.58	1E-92	AT1G48590.1	85.03	3E-98