Universitat Autònoma de Barcelona
Facultat de Biociències
Dept. Biologia Animal, Biologia Vegetal i Ecologia
Estudis de Doctorat en Biologia i Biotecnologia Vegetal

PhD thesis

Fragaria vesca NIL collection: Development and Genetic Characterization of agronomical, Nutritional and Organoleptic traits.

Dissertation presented by María Urrutia Rosauro for the degree of Doctor of Biology and Plant Biotechnology by Universitat Autònoma de Barcelona (UAB)

This work was performed in Centre de Recerca en Agrigenòmica (CRAG), Bellaterra.

Thesis director	Tutor	PhD candidate
Dr. Amparo Monfort Vives	Dr. Charlotte Poschenrieder Wiens	María Urrutia Rosauro

Barcelona, July 2015

Index

Index of contents

Index of Figures_Toc425777434

Index of tables

Summaries

Woodland strawberry, *Fragaria vesca,* is a diploid species from the *Rosaceae* familiy. Its small genome (240 Mb), its huge genetic variability and its high degree of colinearity with the cultivated strawberry (*Fragaria x ananassa*), make it an ideal model to develop genetic and functional studies.

This work has the objective of delving into knowledge on the species and its variability and aims to map characters with agronomical interest and involved in fruit quality. Therefore, a highly relevant genetic tool has been developed and characterized, the near isogenic lines collection. This population consists of 41 lines that allow quantitative traits (QTL) and major genes mapping with an average resolution of 14.2 cM.

Exhaustive phenotyping of the collection, together with a thorough statistical analysis lead to mapping hundreds of QTL with agronomic, nutritional and organoleptic interest. Attending to agronomical traits, nine major genes and seven QTL were mapped controlling important characters such as fruit shape, runnering and flowering habit.

Ripe fruit nutritional study focused on its polyphenolic content, according to the high antioxidant capacity that these compounds bring to berries. Twenty-two (poly)phenols were unambiguously identified and quantified, including anthocyanins, flavonols and flavan-3-ols among others. For which 49 QTL controlling important proportion of the observed variability were mapped. Furthermore, two additional QTL explaining total antioxidant capacity were detected.

Organoleptic compounds that influence the most fruit quality are sugars and volatiles. A total of five QTL were mapped for three quantified sugars. Attending to volatiles, more than 100 different compounds were identified in the collection and 126 QTL were mapped for 81 of them. Among the most significant QTL we found some relevant metabolites for strawberry aroma like methyl 2-aminobenzoate and mesifurane.

Transcriptomic study of introgression lines covering genetic regions with a high number of QTL for fruit quality has revealed a selection of candidate genes for the most interesting organoletpic and nutritional QTL.

This work deepens in the genetic knowledge of woodland strawberry, bringing new genetic tools and QTL useful for strawberry breeding programs and open new perspectives for future studies.

Resumen

La fresa silvestre, *Fragaria vesca,* es una especie diploide del la familia *Rosaceae*. Su pequeño genoma (240 Mb), su gran variabilidad genética y su elevada colinearidad con la fresa cultivada (*Fragaria x ananassa*) la hacen un modelo ideal para desempeñar toda clase de estudios genéticos y funcionales.

En este trabajo se ha querido contribuir al conocimiento profundo de la especie y su variabilidad con el objetivo de mapear caracteres de interés agronómico y de calidad de fruto. Para ello se ha desarrollado y caracterizado una herramienta genética de elevado valor, la colección de líneas casi isogénicas. Esta población consta de 41 líneas que permiten mapear caracteres cuantitativos (QTL) y genes mayores con una resolución media de 14.2 cM.

El fenotipado exhaustivo de la colección junto con un minucioso análisis estadístico ha permitido mapear cientos de QTL con interés agronómico, nutricional y organoléptico. Atendiendo a rasgos agronómicos, se mapearon nueve genes mayores y siete QTL controlando caracteres tan importantes como la forma del fruto, la producción de estolones y el periodo de floración.

El estudio nutricional del fruto maduro se centró en su contenido en polifenoles dado el alto poder antioxidante que otorgan estos compuestos a las bayas. Se identificaron y cuantificaron inequívocamente 22 polifenoles, incluyendo antocianinas, flavonoles y flavan-3-oles entre otros, para los que se mapearon 49 QTL controlando una importante proporción de la variabilidad observada. Además se localizaron dos QTL adicionales explicando la capacidad antioxidante total.

Los compuestos organolépticos que influyen más decisivamente en la calidad del fruto son los azúcares y los volátiles. Un total de cinco QTL fueron mapeados para los tres azúcares cuantificados. En cuanto a los volátiles, más de 100 compuestos distintos fueron identificados en la colección y 126 QTL han sido mapeados para 81 de ellos. Entre los QTL más significativos encontramos algunos metabolitos de gran importancia para el aroma de la fresa como el methyl 2-aminobenzoato o el mesifurano.

El estudio transcriptómico de líneas de introgresion que cubren regiones genéticas con un alto número de QTL de calidad de fruto ha revelado una selección de genes candidatos para algunos de los QTL organolépticos y nutricionales más interesantes.

Este trabajo profundiza en el conocimiento de la genética de la fresa silvestre, proporcionando nuevas herramientas genéticas y nuevos QTL que pueden ser empleados para la mejora genética del cultivo y abren un gran abanico de oportunidades para estudios posteriores.

La maduixa silvestre*, F. vesca*, és una espècie diploid de la família *Rosaceae*. El seu petit genoma (240 Mb), la seva gran diversitat genètica i la seva elevada col·linealitat amb la maduixa cultivada (*Fragaria x ananassa*), la fan un model ideal per tota mena de estudis genètics i funcionals.

En aquest treball, s'ha volgut contribuir al coneixement profund de la espècie i la seva variabilitat amb l'objectiu de mapar caràcters de interès agronòmic i de qualitat de fruit. Per això s'ha desenvolupat i caracteritzat una eina genètica de elevat valor, la col·lecció de línies casi isogèniques. Aquesta població comprèn 41 línies que permeten mapar caràcters quantitatius (QTL) i gens majors amb una resolució mitja de 14.2 cM.

El fenotipat exhaustiu de la col·lecció juntament amb un detallat anàlisi estadístic han permès mapar centenars de QTL amb interès agronòmic, nutricional i organolèptic. Atenen als aspectes agronòmics, es van mapar nou gens majors i set QTL controlant caràcters tan importants com la forma del fruit la producció d'estolons i el període de floració.

L'estudi nutricional del fruit madur es va centrar en el seu contingut en polifenols, donat l'elevat poder antioxidant que proporcionen aquests compostos a les baies. Es van identificar i quantificar inequívocament 22 polifenols, incloent-hi antocianines, flavonoles i flavan-3-oles entre d'altres, per als quals es van mapar 49 QTL controlant una important proporció de la variabilitat observada. A més, es van localitzar dos QTL addicionals explicant la capacitat antioxidant total.

Els compostos organolèptics que influencien més decisivament la qualitat del fruit són els sucres i els volàtils. En total, cinc QTL van ser mapats per als tres sucres quantificats. En quant als volàtils, més de 100 compostos diferents van ser identificats en la col·lecció i 126 QTL van ser mapats per 81 d'ells. Entre els QTL més significatius trobem alguns metabòlits de gran importància per l'aroma de la maduixa com el methyl 2-aminobenzoat o el mesifurano.

L'estudi transcriptomic de les línies d'introgressió que cobreixen regions genètiques amb un alt nombre de QTL per la qualitat del fruit, han revelat una selecció de gens candidats per alguns dels QTL organolèptics i nutricionals més interessants.

Aquest treball, profunditza en el coneixement genètic de la maduixa silvestre i aporta noves eines genètiques i nous QTL que poden ser emprats per la millora genètica del cultiu i que obren un gran ventall d'oportunitats per a futurs estudis.

General introduction

1. *Fragaria* genus

1.1. Taxonomy and distribution

Strawberries (*Fragaria* sp.) belong to the *Rosaceae* family that comprises over 100 genera and 3000 species and includes most of the fruit trees with high commercial value. For instance peaches, cherries, apricots or plumps from the genus *Prunus*, apples from genus *Malus* and pears from genus *Pyrus*. It also includes berry crops such as raspberries, blueberries or blackberries from genus *Rubus* and ornamental plants such as roses from *Rosa* genus. Species of the Rosaceae family are cultivated and commercialized worldwide having an important economical impact in agriculture.

The precise classification of strawberry is (Potter *et al.* 2007):

Kingdom: *Viridiplantae*

Division*: Magnoliophyta*

Class: *Magniolopsida*

Subclass*: Rosids*

Order*: Rosales*

Family: *Rosacea*e

Subfamily: *Rosoideae*

Supertribe: *Rosodae*

Tribe: *Potentilleae*

Subtribe: *Fragariinae*

Genus: *Fragaria*

Fragaria genus comprises around 20 species with base chromosome number x=7 and different levels of ploidy (Folta & Davis 2006; Stewart 2011) (Table I.1). In nature diploid (2*n*=14), tetraploid (4*n*=28), hexaploid (6*n*=42), octoploid (8*n*=56) and decaploid (5*n*=70) *Fragaria* sp. exist. The genus is distributed across all temperate regions in North hemisphere (Eurasia and North America) and in western South America although not all species are homogenously spread along this area. *F. vesca* is the most widely distributed strawberry and can be found all across Eurasia and North America. The rest of the diploid species are spread along Eurasia while tetraploid species are restricted to eastern Asia. *F. moschata*, the only natural hexaploid, is typical from central Europe and Siberia and wild octoploids are native to America from where they hypothetically colonized Iturup Island (Northeast Japan). Finally cultivated strawberries (*F. x ananassa*) emerged as a natural hybrid of wild American octoploides in Western France. Although the geographic origin of the genus is not completely elucidated, the high genetic diversity present in Asia points out this region as a possible origin of the *Fragaria* genus (reviewed by Stewart 2011).

Table I.1. *Fragaria* genus species, ploidy and distribution. From Folta & Davis (2006) and Stewart (2011)

Species	Ploidy	Geographical distribution
F. bucharica	2x	Western Himalaya
F. daltoniana	2x	Eastern Himalaya
F. gracilis	2x	Northwest China
F. innumae	2x	Japan
F. mandshurica	2x	Northeastern Asia
F. nilgerrensis	2x	Central Asia
F. nipponica [a]	2x	Japan
F. nubicola	2x	Eastern Himalaya
F. pentaphylla	2x	Chinese Himalaya
F. vesca	2x	Europe, temperate Asia and America
spp. *americana*	2x	Eastern North America
spp. *bracteata*	2x	Western North America
spp. *californica*	2x	California
spp. *vesca*	2x	Europe and temperate Asia
F. viridis	2x	Northern Europe and western Asia
F. yezoensis [a]	2x	Japan
F. x bifera	2x/3x	Europe (natural hybrid of *F. vesca* and *F. viridis*)
F. corymbosa	4x	China
F. moupinensis	4x	Southwestern China
F. orientails	4x	Northeastern Asia
F. tibetica	4x	East Himalaya
F. x bringhurstii	5x/6x/9x	California (natural hybrid of *F. vesca* and *F. chiloensis*)
F. moschata	6x	Northern Europe and western Asia
F. chiloensis [b]	8x	Western America
F. virginiana [b]	8x	North America
F. iturupensis	8x/10x	Iturup Island
F. x ananassa	8x	Cultivated natural hybrid of *F. chiloensis* and *F. virgininana*
F. x vescana	10x	Artificial hybrids of *F. x ananassa* and *F. vesca*

[a] *F. nipponica* and *F. yezoensis* have been propossed for a possible merge of the two taxa due to their high similarity.

[b] These species have four subspecies each

1.2. Morphology

Strawberries are small perennial plants that can be sexually and vegetatively (by runners) reproduced. They have a short thickened stem or crown from which leaves, inflorescences, runners and roots emerge (Figure I. 1). Trifoliate leaves are borne along the crown on petioles that can be 20 cm tall. Axillary buds give rise, depending on the environmental conditions, to runners or inflorescences. Runners are involved in vegetative reproduction and can eventually produce roots and develop a new plant while inflorescences are involved in sexual reproduction. Flowers are usually bisexual, with five white petals surrounding the anthers and a yellow receptacle sustaining the pistils. Botanically, the red fleshy “fruit” or “berry” is just an enlargement of the flower receptacle and the achenes (dry ovaries) embedded in its surface are the true fruits and carry the seeds inside. For simplicity, in this work we will refer to the combination of receptacle and the achenes as fruit or berry.

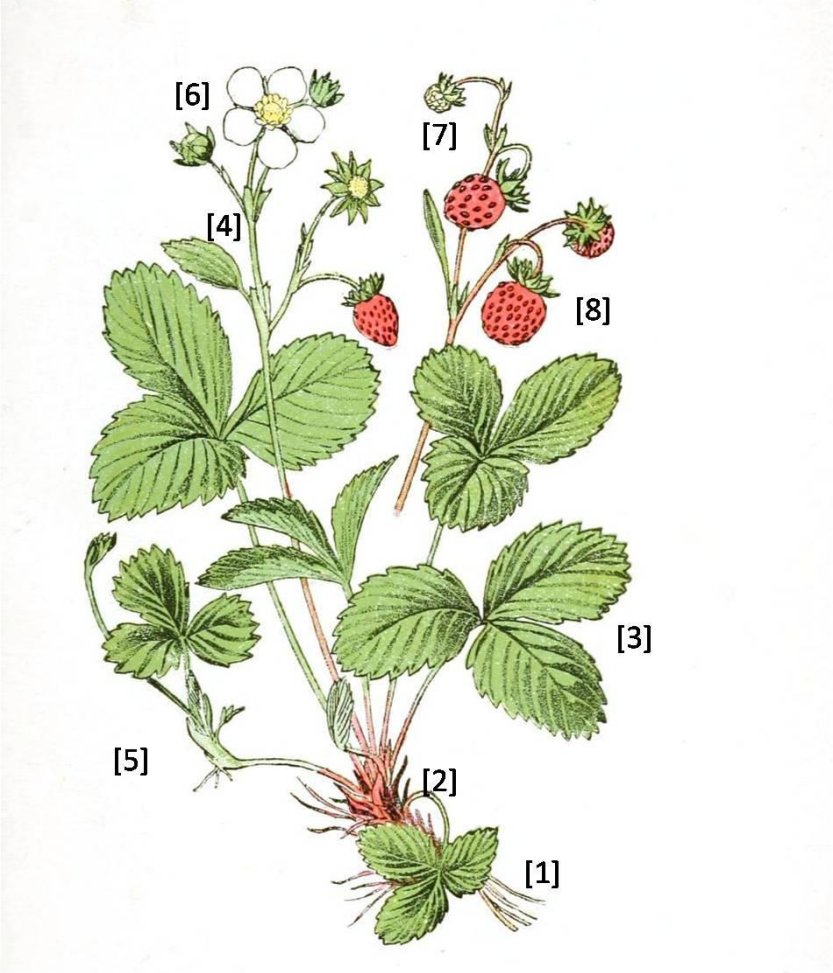


Figure I. 1 *Fragaria vesca* botanical drawing (unknown author). [1] roots, [2] thickened stem or crown, [3] leaves, [4] inflorescence, [5] runner, [6] flowers, [7] undeveloped fruit, [8] ripe fruit

1.3. Phylogeny

Nowadays, phylogenetic relationships between *Fragaria* sp. remain relatively poorly resolved. Diploid species have been classified in two main clades (Potter *et al.* 2000; Sargent *et al.* 2004b), one containing *F. vesca*, *F. nubicola*, *F. viridis* and *F. nilgerrensis* and another one with *F. nipponica*, *F. daltoniana* and *F. pentaphylla*, with *F. iinumae* remaining separated from both. Tetraploids are in general considered autopolyploidizations of diploids. Studies with low copy number nuclear genes described two main clades, suggesting two independent polyploidization events. The first clade was composed by the tetraploid *F. orientalis* and the diploids *F. vesca* and *F. mandshurica* (Potter *et al.* 2000). The second clade grouped the remaining three tetraploid species with the diploids *F. nipponica*, *F. nubicola*, *F. pentaphylla* and *F. yezoensis*, however no clear conclusion was obtained for the possible subgenome origin of the tetraploids.

The maximum controversy and efforts have been addressed to elucidate the phylogenetic origin of the octoploid species. Octoploid strawberry species are allopolyploid, this meaning

that the four subgenomes have at least two ancestral genomes. A detailed allelic study of two low-copy nuclear genes (*GBSSI-2* and *DHAR*) in a set of diploid, tetraploid, hexaploide and wild octoploid accessions suggested that both *F. iinumae* and either *F. vesca* or *F. mandshurica* were the likely subgenomes donors of the wild octoploid strawberries (Rousseau-Gueutin *et al.* 2009). A recent study using dense genetic linkage maps by Tennessen *et al.* (2014) proposed that only one of the subgenomes ancestries is *F. vesca* spp. *braceata*, another one corresponds to *F. iinumae* and two others correspond to an unknown ancestor of *F. innumae.*

1.4. Fragaria vesca

In the last years, *Fragaria vesca* has been outlined as an ideal plant model for the study of perennial plants and has become the reference model species for the *Fragaria* genus and one of the model species in the *Rosaceae* family (Slovin & Mihael 2011). Its wide distribution and genetic diversity are unique traits for the study of the genus. It is the most widely distributed *Fragaria* sp. It can be found all along temperate zone and it is the only species native from both Eurasia and America. In addition it presents a diverse germplasm base divided in four subspecies (spp. *americana*, spp. *bracteata*, spp. *californica* and spp. *vesca*) with reported phenotypic differences for fruit and flower color, floral and leaf morphology, runnering, photoperiodic sensitivity, flowering habit or crown structure among other traits (Hadonou *et al.* 2004).

Practical reasons also support its role as a model plant. Its small size is optimal to grow it in laboratory facilities. Most varieties are self-compatible, what facilitates sexual reproduction, and develop runners what allows cloning or vegetative reproduction. Besides it has a short life cycle (4-5 months) and produces many seeds by plant. In addition, *F. vesca* can be stably and transiently transformed both by *Agrobacterium* and particle bombardment (El Mansouri *et al.* 1996; Wang *et al.* 2004; Hoffmann *et al.* 2006; Pantazis *et al.* 2013).

Genetically, *F. vesca* also presents advantages as a model for the *Fragaria* genus as it has a diploid small genome (240 Mb) that is probably ancestor of one of the four subgenomes of the cultivated strawberry (*F. x ananassa*) with which shares a high degree of colinearity (Rousseau-Gueutin *et al.* 2008; Tennessen *et al.* 2014).

F. vesca spp. *vesca* var. 'Reine des vallées' is an alpine accession (PI 551824) with which most of the work presented in this thesis has been developed. It is an everbearing non-runnering diploid variety, with good-sized bright red fruits, very appreciated for its intense flavor and aroma and that is cultivated with industrial proposes (Figure I. 2).

1.5. Fragaria bucharica

F. bucharica is a diploid Himalayan species. It is known to be self-incompatible and produces tiny fruits with no commercial interest. *F. bucharica* FDP601 (PI657844) is an accession cross-compatible with *F. vesca* with which produces fertile off-spring that shows a seasonal flowering habit and that can be reproduced by runners. It has been used in molecular studies such as development of the *Fragaria* reference linkage map (Sargent *et al.* 2004a) and is used in this work as a parental of a mapping population (Figure I. 2).

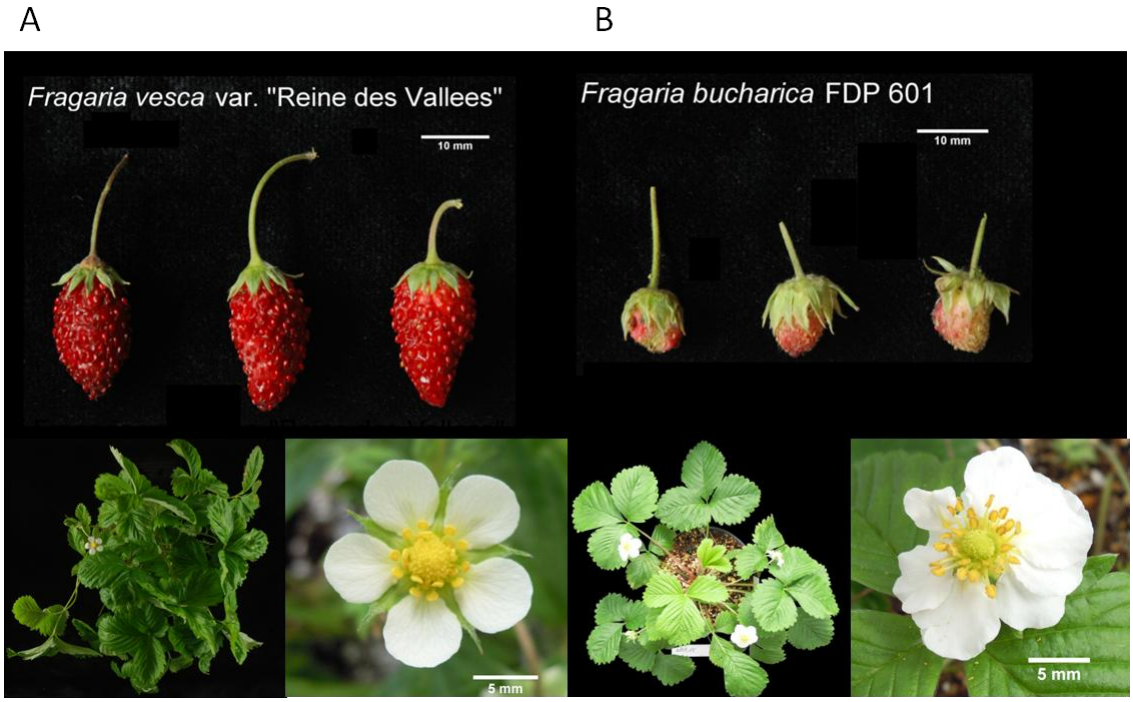


Figure I. 2 Detail of fruits, plant and flower from *F. vesca* and *F. bucharica*. A: *F. vesca* var. 'Reine des Vallées'. **B**: *F. bucharica* FDP 601

2. Resources in *Fragaria vesca*

2.1. Markers

Molecular markers are tools to identify genetic variability among individuals. Traditionally, the most abundant and widely used molecular markers in strawberry genetic studies have been microsatellites or single sequence repeats (SSRs). SSRs are based in the variable length of polynucleotide repeats. These are islands of repetitive sequence that rapidly expand or contract with time and can be amplified with specific primer sequences. SSRs are a reproducible, reliable, transferable and cheap genotyping method. In *Fragaria* several sets of SSRs have been developed from diploids *F. vesca* (James *et al.* 2003; Cipriani & Testolin 2004; Hadonou *et al.* 2004; Monfort *et al.* 2006) and *F. viridis* (Sargent *et al*. 2003). Also from octoploids *F. virginiana* (Ashley *et al.* 2003) and *F. x ananassa* (Govan *et al.* 2008; Rousseau-Gueutin *et al.* 2008). Microsatellite markers have been also developed from expressed sequence tags (EST) from both diploid and octoploid strawberries (Folta *et al.* 2005; Bassil *et al.* 2006; Gil-Ariza *et al.* 2006; Sargent *et al.* 2006; Zorrilla-Fontanesi *et al.* 2011b).

Transferability of SSRs within *Fragaria* genus is very high, with 70-100% of them being transferable (Ashley *et al.* 2003; Bassil *et al.* 2006; Monfort *et al.* 2006; Rousseau-Gueutin *et al.* 2008). They have also been successfully transferred to other genera in the *Rosaceae* family such as *Malus* and *Prunus* with 78% transferability for EST derived SSRs (Illa *et al.* 2011) and with lower success rates to *Rosa* genus (20-30%) (Zorrilla-Fontanesi *et al.* 2011b). These microsatellites sets have largely proven to be useful for comparative mapping within a genus and sometimes between genera.

2.2. Mapping populations and linkage maps

Well characterized segregating populations are useful tools for identifying and mapping phenotypic diversity among a species or a genus. Few, although extensively used, stable and well characterized mapping populations have been developed in *F. vesca*.

The first diploid strawberry reported linkage map was generated from a F_2 population derived from the cross between *F. vesca* spp. *vesca* var. 'Baron Solemacher', a non-runnering European cultivar, and *F. vesca* spp. *americana* 'W6' collected in the wild (Davis & Yu 1997). It was built using 75 randomly amplified polymporphic DNA (RAPD) markers and span over 445 cM. These markers were subsequently used in two new F_2 populations crossing *F. vesca* var. 'Yellow Wonder', an alpine European variety producing yellow fruits, with *F. vesca* 'DN1C' and *F. nubicola* 'FRA520' (that was probably *F. bucharica*) to map locus 'c', responsible for the yellow strawberry fruit color (Deng & Davis 2001).

Later, an F_2 population from the cross between *F. vesca* 'FDP815' and *F. bucharica* 'FDP601' allowed the construction of the still today's reference linkage map in *F. vesca* (Sargent *et al.* 2004a; Sargent *et al.* 2006). This map was built using 78 markers first (68 SSRs) and enhancing it later with 109 more SSRs. It has been revised at least three more times by Sargent *et al.* (2007, 2008, 2011), mapping 24 new gene-specific markers, 103 SSRs and 194 SSRs respectively. This collection has been extensively used among the *Fragaria* community and has been successful in mapping known genes (Sargent *et al.* 2006).

A loss-of-function mutant collection was generated in *F. vesca* 'Hawaii 4' using T-DNA insertion (Oosumi *et al.* 2010) that may help to functionally characterize genes of interest.

The development of new mapping populations is advisable in order to cover a wider range of genetic and phenotypic variability among the diploid *Fragaria* sp., to reinforce its role as a model species and to thoroughly characterize agronomical interesting traits that may relay undocumented in the genus.

2.3. Monogenic traits and QTL in *F. vesca*

Few monogenic traits have been thoroughly studied in *F. vesca.* An interesting one is flowering habit as it is an important agronomical trait that determines period and number of harvests. Seasonal flowering (*SFL*) is dominant over everbearing or continuous flowering habit in *F. vesca* (Brown & Wareing 1965). Subsequent studies mapped *SFL* to LG6 (Albani *et al.* 2004; Sargent *et al.* 2004a) and it was afterwards characterized as *FvTFL1* (Iwata *et al.* 2011; Koskela *et al.* 2012). Another interesting monogenic trait is vegetative reproduction by runners as it facilitates uniformity and rapid expansion of plants. Runnering (*R*) is dominant over runner-less phenotype in *F. vesca* (Brown & Wareing 1965). This trait has been mapped to the middle region of LG2 (Sargent *et al.* 2004a).

One might suspect that fruit color is a complex trait as many metabolites contribute to pigmentation. However, natural variation between red- and white-fruited *F. vesca* varieties seems to be monogenically controlled, being red color (*C*) dominant over white color (Brown & Wareing 1965). This trait has been mapped to the end of LG1 and probably encodes a flavanone 3-hydroxylase (Deng & Davis. 2001). In addition, leaves coloration also seems to be a

monogenic trait, being dark-green phenotype (*Pg*) dominant over pale green phenotype. This trait was mapped to LG6 (Sargent *et al.* 2004a).

A quantitative trait loci (QTL) mapping strategy, in-depth phenotyping followed by interval mapping, has not been followed in diploid strawberry to our knowledge. Instead, EST and gene sequences available in public databases have been added and mapped to the *F. vesca* reference linkage map taking advantage of intron-length polymorphisms (Sargent *et al.* 2007). This is useful to build an informative and saturated linkage map; however, it does not explain phenotypic variability nor provides functional insights.

2.4. Synteny studies

High levels of conserved macrosynteny and colinearity have been reported between the diploid strawberry linkage groups and their corresponding octoploid homoeologuous using both microsatellites and new-generation sequencing technologies discarding any possible major chromosomal rearrangement during polyploidization events in *Fragaria* evolution (Rouseau-Gueutin *et al.* 2008; Tennessen *et al.* 2014).

Syntenic relations among different genera of the *Rosaceae* family have been addressed several times and described in more detail as genetic tools provided more resolution. First approach was taken by Vilanova *et al.* (2008) who reported long syntenic regions and 36 rearrangements between the *Prunus* and the *Fragaria* genomes using 71 molecular markers most of them being restriction fragment length polymorphisms (RFLP). Later Illa *et al.* (2011) compared the *Fragaria*, the *Prunus* and the *Malus* genomes after the publication of the first *Malus* genome draft. They used 806 molecular markers (148 of them anchored to the Fragaria genome) and proposed a nine chromosome common ancestor for the *Rosaceae* family. Finally in 2012, after the publication of the *Prunus* and *Fragaria* genomes, a whole genome synteny study was conducted by Jung *et al.* (2012) that reported 1399 orthologous regions between the three genomes. They also described that *Fragaria* has gone under a greater number of small-scale rearrangements than the other genomes since their divergence from a common ancestor.

Synteny studies are of great utility for transference within the *Fragaria* genus and the *Rosaceae* familiy and strengthen the suitability of *F. vesca* as a model species for its genus and familiy.

2.5. EST and BAC libraries

Before 2011, when *F. vesca* genome had not been sequenced yet and the high throughput sequencing technologies were not so wide spread, ESTs were the main source of nucleotide sequence data available for the *Fragaria* research community (Slovin & Michael 2011). NCBI nucleotide database reported in June 2015 60035 *Fragaria* EST, the majority of them being from *F. vesca* (47975) and *F. x ananassa* (10988).

The genome of *F. vesca* was accessible through a 4x-10x coverage BAC library of 18432 clones with an average insert size of 85 Kb (Bonet *et al.* 2009).

2.6. Transformation methods

Around a hundred genes had been functionally characterized in *Fragaria* sp. before *F. vesca* genome sequence publication (see Supplemental Table 10 Shulaev *et al*. 2011). Most of them were related to important metabolic functions and fruit ripening. The preferred method for functional characterization of strawberry genes has been *Agrobacterium*-mediated transformation in *F. x ananassa*. Genetic transformed lines of *Fragaria* sp. have been obtained to study gene functions (Schaart *et al.* 2004; Schaart *et al.* 2013) and to improve traits such as fruit production (Mezzetti *et al.* 2004) and plant resistance (Jiwan *et al.* 2013). However this method is time consuming, especially for ripe-fruit quality traits. Therefore, transient transformation has been extensively used in strawberry reverse-genetics studies. First reports aimed the study of endo-β-1,4-glucanase and GalUR promoters strength by transient expression mediated by *Agrobacterium* or particle bombardment (Spolaore *et al.* 2003; Agius *et al.* 2005). Later, an efficient RNAi-based transient transformation protocol for strawberry fruits was set using chalcone synthase (CHS) as a reporter gene. Agroinjection of receptacles with inoculums of *Agrobacterium* carrying CHS complementary hairpin conducted to a drastic reduction of mRNA levels and enzymatic activity of CHS, what led to pigmentation loss in ripe strawberry fruits (Hoffmann *et al.* 2006). This transient RNAi method has been successfully used afterwards for the silencing of other flavonoid-related genes as the anthocyanin glycosyltransferase *FaGT1* (Griesser *et al.* 2008), flavanone 3-hydroxylase (Jiang *et al.* 2013) and alcohol acyltransferase *FaAAT2* (Cumplido-Laso *et al.* 2012).

Transient transformation has also been used for gene overexpression alone or in combination with RNAi-silencing. For instance the simultaneous silencing of *FaCHS* and *EGS* genes overexpression, re-directioned carbon flux in the flavonoid pathway to the synthesis of phenylpropene compounds and confirmed the role of *FaEGS1* and *FaEGS2* in eugenol production (Hoffmann *et al.* 2011; Aragüez *et al.* 2013). In addition, a recent study silencing *FaEOBII* revealed that this R2R3-MYB transcription factor controlled eugenol production (Medina-Puche *et al.* 2015).

2.7. Genome sequence

A great milestone in strawberry genetics research was the publication of the first *F. vesca* sequence draft in 2011 (Shulaev *et al.* 2011). The accession sequenced was *F. vesca* ssp. *vesca* cv. 'Hawaii4' (PI551572). This was the first plant genome to be sequenced solely with short-read technologies and without a physical reference. The genome estimated size was 240 Mb and 95 % was contained in 272 scaffolds.

Scaffolds were assembled and oriented by anchoring them to the reference diploid linkage map and to its associated bin map with 390 genetic markers (Sargent *et al.* 2006, Sargent *et al.* 2008) resulting in seven pseudochromosomes assembly. Gene prediction was based in a hybrid strategy of *ab initio* prediction and transcript sequencing data resulting in 34809 predicted genes. Functional annotation was conducted on the basis of sequence similarity between predicted proteins and proteins annotated in public databases what resulted in 25050 preliminary annotated genes.

Two main updates have been done to the first genome version (v.1.1). First, a revision of the genome assembly using a dense genetic linkage map developed with an F_1 population of *Fragaria vesca* ssp. *bracteata* resulted in a new assembly version (v.2.0) (Tennessen *et al.* 2014). This new assembly contained 44 interchromosome translocations, 40 intrachromosome translocations, 39 inversions and 18 placements of previously unmapped scaffolds in relation to assembly version 1.1. Second, a new annotation (a2) based in whole transcriptome sequencing (RNAseq) data of 25 different tissues (Darwish *et al.* 2015) increased by 9139 the number of coding regions and by 2286 the number of gene models in the previous annotation.

These two updates were developed and published independently; as a result, genome assembly 2.0 conserves the original annotation (a1), while the new annotation (a2) is anchored to the original genome assembly v1.1. Merging these new resources into one enhanced genome assembly and annotation version would be of great interest for the *Fragaria* research community and would avoid unnecessary confusion.

2.8. SNP chip

New sequencing technologies, and the availability of a high quality reference genome allowed the appearance of an SNP array genotyping platform IStraw90® (Affymetrix, CA, USA) specially designed for octoploid strawberry genotyping but anchored to the diploid *F. vesca* reference genome version 1.1. It inquires a total of 95062 polymorphic sites in *F. x ananassa* genome (Bassil & Davis *et al.* 2015).

3. Economic value

Strawberries (*F. x ananassa*) are the most economically important berry; its production is twice the amount of all other berry crops combined (Figure I. 3A).

World strawberry production has been tripled in the last 20 years, growing from 2.5 to 7.5 million tonnes (data from 1992 to 2012). However, Spanish contribution has remained roughly constant over this period with an average production of 289000 tonnes per year, being the maximum peak production 377527 tonnes in 1999. The United States of America has led world production for the past two decades, however trustworthy data from China are not reported, but suspected to be higher than those of USA. Spain has traditionally occupied the second world producer (first in Europe) position but it has been surpassed in recent years by Mexico and Turkey (Figure I. 3B) that have emerged as first range producers by drastically increasing their annual production in the last few years (http://faostat3.fao.org, 2015).

Strawberries are an important agronomical product in Spain. Attending to generated income, they are the 15th agronomical product (including farm animals and crops), and the 5th fruit (after grapes, oranges, mandarins and peaches) in Spain generating 393 million dollars in 2012.

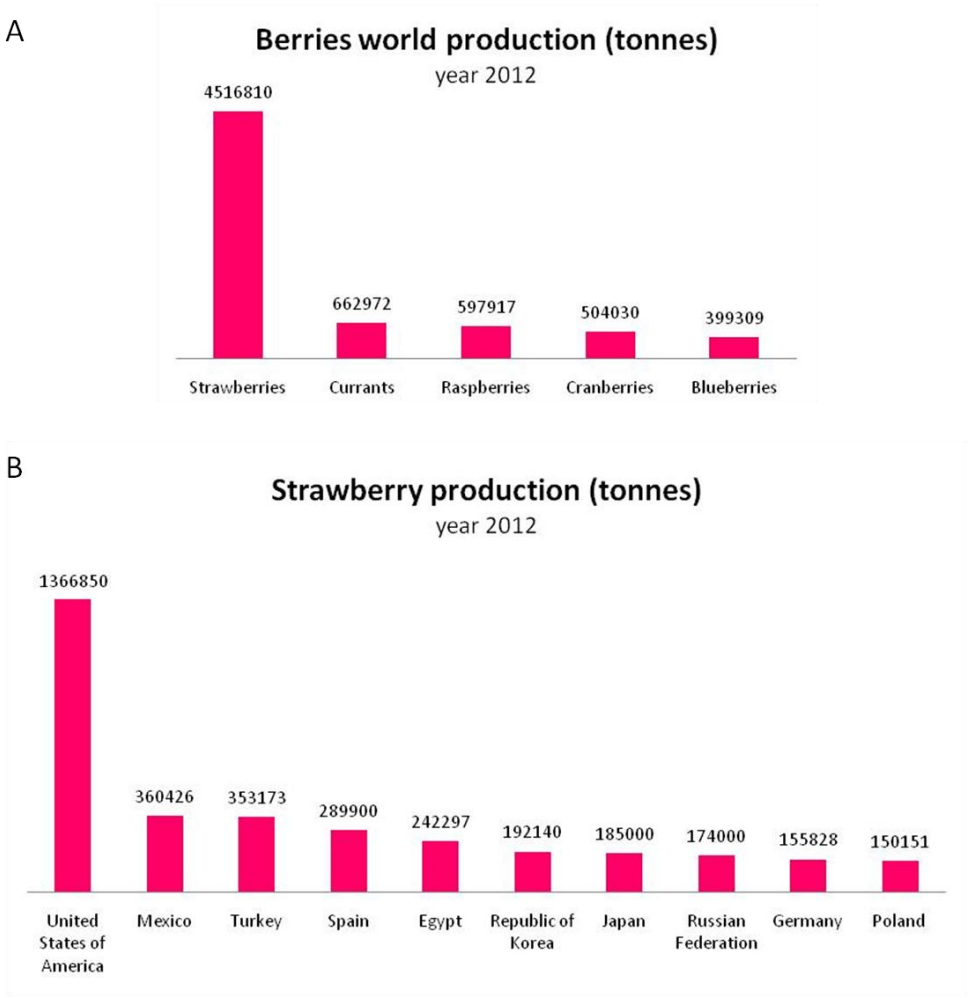


Figure I. 3 Berries and strawberries production. **A**: Global world berries production in tonnes in 2012. **B**: Top ten strawberry producers countries in 2012 (data FAO)

4. Breeding and quality traits

4.1. History

Wild *Fragaria* species have been likely known and consumed by humans since the antiquity, and probably before, although archaeobotanical evidence is limited (Kubiak-Martens 1999). In Europe, *F. vesca* has been grown in gardens at least since the Romans, as Ovidi and Virgil named it in their poems, and *F. moschata* since the 16th century (Hummer & Hancock 2009). In America, there are evidences of domestication of *Fragaria chiloensis* by the Picunche and Mapuche people of Chile over 1000 years ago (Finn *et al.* 2013).

The modern cultivated strawberry (*F. x ananassa*) did not appear until the 18th century after the introduction of America's large-fruited octoploids into Europe. The first octoploid to be introduced was *F. virginiana* that was probably brought from Virginia by colonists by the late 16th century. Around a century later, *F. chiloensis* was introduced in France from Chile by Amédée Frézier (Wilhelm & Sagen 1974). Cultivation of these species was very gradual at the beginning, especially for *F. chiloensis* as it produced few fruits for several years and was a

complete disappointment. The fact was that all five specimens brought by Frézier were pistillate and the need for cross-pollination was not well understood at the moment. Finally, by the late 18th century, first hybrids of *F. x ananassa* started to appear in the Northwest coast of France, where both *F. virginiana* and *F. chiloensis* were cultivated together. At the beginning there was controversy about the origin of those new seedlings but the French botanist Antoine Duchesne correctly determined their hybrid origin and named them *F. x ananassa* attending to their pinnaple-like aroma. Duchesne became probably the first strawberry researcher collecting specimens, producing a well-reasoned classification of species, characterizing the sexual compatibility among the different types and documenting the need for pollination in his *Histoire naturelle des fraisiers* (Duchesne 1766).

After the appearance of *F. x ananassa*, interest in cultivating strawberries increased. The first formal strawberry breeding program was initiated in England in 1817 by Thomas Knight (Wilhelm & Sagen 1974) who intercrossed and selected octoploid strawberries that still today form the background of some modern cultivars. Since then, hundreds of strawberry varieties have arose from public and private breeding programs, mainly in Europe and USA. These breeding programs have mainly addressed improvements in agronomical traits but little attention has been paid to nutritional or organoleptical traits (Darrow 1966, Stewart 2011).

Although many important and expensive strawberry programs are still today carried on by private companies, the use of marker assisted selection (MAS) is not so widely used as in other fruit crops. Main reasons for backwardness in this area are the genetic complexity of the commercial *F. x ananassa* hybrid together with the relatively cheap grow and selection in field method for strawberry. However, it is believed that reliable associated markers for interesting quantitative traits such as aroma intensity or nutritional properties would be of great use for parental selection.

4.2. Strawberry quality traits

Strawberries are among the youngest horticultural crops. It has been cultivated only since the 19th c. yet it has changed thoroughly during this period. Strawberry quality traits have been traditionally related to yield, fruit size, firmness, color, taste and resistance to pests (Darrow 1966). Yield has been more than doubled in the last decades mainly due to development of remontant cultivars that flower several times per year and not only in spring and early summers as most wild strawberries do. Fruit size has also increased in some commercial varieties, although in some cultivars with a very short season a larger amount of fruits per plant is preferred instead of bigger fruits. However, the pillar trait that triggered strawberry production and commercialization worldwide was fruit firmness (Darrow 1966). Most cultivars produce nowadays firm fruits that can be shipped and arrive to consumers' homes undamaged. Another important trait towards efforts have been addressed to is resistance to pests. Most commercial varieties are at least partially resistant to most common strawberry pests like red steele, powdery mildew, leaf spot, leaf scorch, verticillium and several viruses, however none of this diseases has been eradicated. The bright red color has been preferred in most cultivar selections, but wild strawberries naturally present a much wider palette of colors. White, yellow and a range of different intensities of red accessions could be used to produce distinctive fruits. Taste has traditionally been measured in terms of sugar-acidity ratio.

Although great efforts have been made in order to release sweet strawberry varieties, it is still the main cause of dissatisfaction among consumers (Bruhn *et al.* 1991). This might be because the sugar-acidity ratio is not enough informative in terms of overall flavor perception. A recent study revealed that flavor intensity, associated to volatile composition, contributes equally to consumer's overall-liking as sweetness (Schwieterman *et al.* 2014).

Traditional breeding has not paid attention to organoleptic or nutritional fruit quality traits. However, nowadays almost all strawberry commercial cultivars have reached high quality standards in agronomical parameters and new varieties with enhanced flavor and nutritional properties could make a difference in terms of marketability.

Attending to nutritional properties, strawberries are claimed to have an outstanding antioxidant capacity due to their high polyphenol content (Battino *et al.* 2009). Consumption of antioxidant-rich fruits has been associated with beneficial health effects in humans (Ramassamy 2006; Seeram *et al.* 2006). Regarding aoganoleptic quality, it could be easily improved by increasing specific volatile concentrations like methyl 2-aminobenzoate or furaneol, that have almost disappeared in commonly commercialized varieties but are still present in wild accessions.

4.3. (Poly)-phenols

4.3.1. Polyphenols in fruits and strawberry

Polyphenols are a large class of plant secondary metabolites derived from the phenylpropanoid pathway (Figure I. 4). They are divided in several groups including flavonoids, hydroxycinnamic and benzoic acid derivatives, hydroxystilbenes and lignans. Wide structural diversity is found within each group and especially among flavonoids family which comprises over 8000 molecules (reviewed by Ageorges *et al.* 2014). Most common flavonoids include anthocyanins, flavonols, flavan-3-ols and flavanones.

Polyphenols are present in all vascular plants, but phenolic profiles vary between species and cultivars and between tissues and developmental stages within the species.

Fruit polyphenols have been extensively studied as they have been claimed health beneficial nutritional properties including prevention of neurodegenerative diseases and cancer (Ramassamy 2006; Seeram *et al*. 2006) and contribute to major organoleptic traits such as color, sensitivity to enzymatic browning, bitterness and astringency.

Anthocyanins are common flavonoids that provide red, blue and black pigments to fruits. Their basic structure consists in six aglycones (anthocyanidins) that differ in the number of hydroxyl groups on their B-ring and their methylation pattern. Cyanidin derivatives are the most ubiquitous anthocyanins while pelargonidins are rare. However, in strawberry fruit pelargonidin-3-glucoside is the most abundant anthocyanin (Almeida *et al.* 2007; Buendia *et al.* 2010).

Flavonols are widespread in fruits but are accumulated only in small quantities. Different substitution patterns in the B-ring and nature and position of sugar substituents provide structural diversity to the flavonols group.

Flavan-3-ols comprise monomers, oligomers and polymers, called proanthocyanidins (PAs). PAs usually consist in catechin and epicatechin units but (epi)afzelechin units are also detected in strawberry. Although PAs are very abundant polyphenols in strawberry their content is usually underestimated as they are not extracted by the usual solvents (methanol). However, acid-catalysed depolymerization in presence of a nucleophile followed by HPCL showed that they are the most abundant polyphenols in strawberry (Buendía *et al.* 2010).

4.3.2.Polyphenol pathway

Phenylpropanoids are synthesized from phenylalanine by the sequential action of phenylalanine ammonia–lyase (PAL), cinnamic acid 4-hydroxylase (C4H) and p-coumarate coA ligase (4CL) (Figure I. 4). Substrates are compromised for flavonoid synthesis by the action of chalcone synthase (CHS) and chalcone isomerase (CHI) that give rise to the 15C three-ring basic flavonoid structure. The sequential action of flavanone 3β-hydroxylase (FHT), dihydroflavonol 4-reductase (DFR) and anthocyanindin synthase (ANS) enzymes leads to the synthesis of anthocyanidin pigments. Minor polyphenols, flavonols and flavan-3-ols are produced by flavonol synthase (FLS), leucoanthocyanidin reductase (LAR) and anthocyanidin reductase (ANR) enzymes. Finally numerous flavonoid glycosyltransferases (FGT) enzymes add sugar moieties that diversify flavonoids prior to their compartimentalization in the vacuole.

All commonly known enzymes in the phenylpropanoid pathway have been characterized in strawberry (*F. x ananassa*) (Halbwirth *et al.* 2006; Almeida *et al.* 2007).

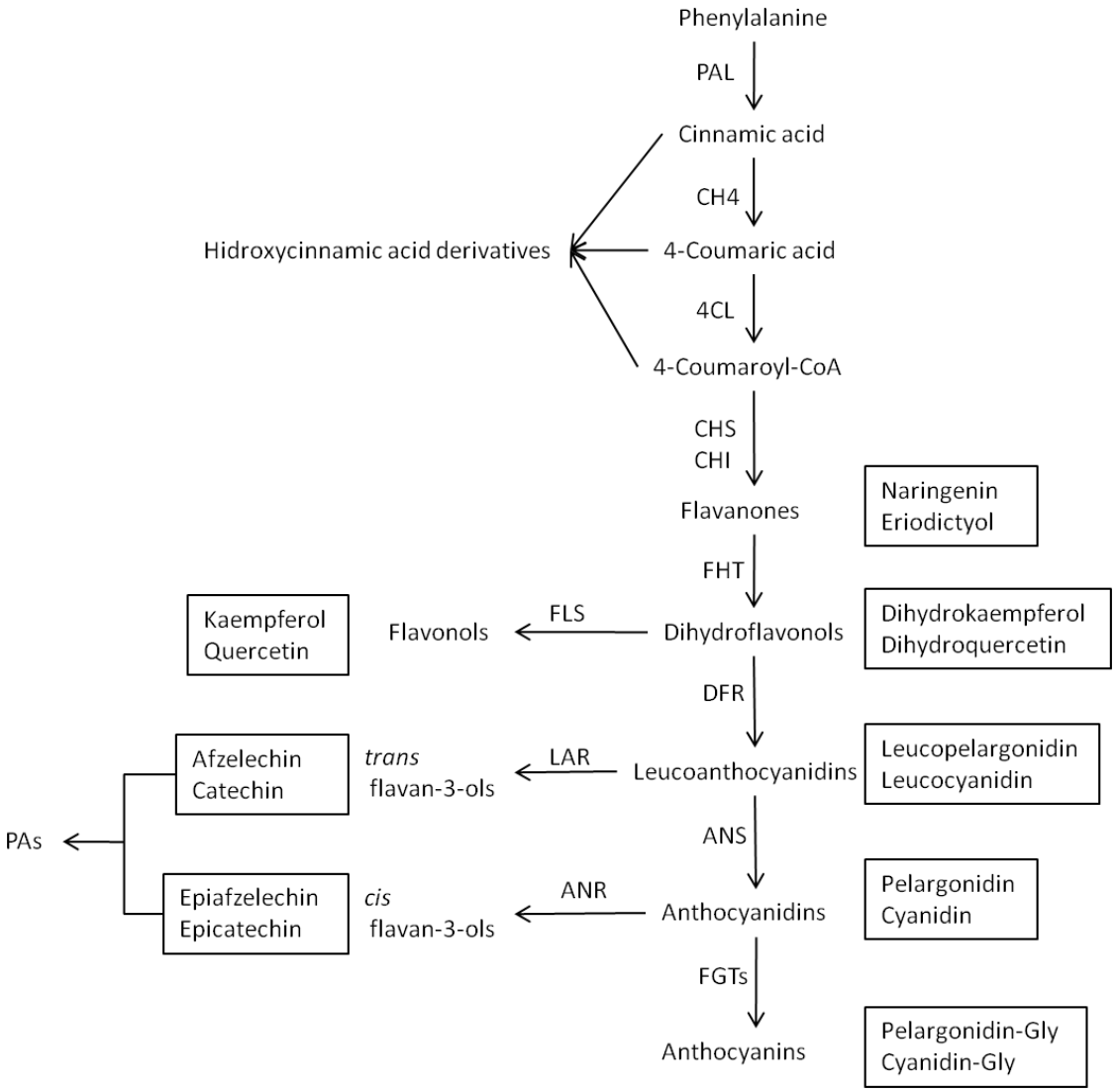


Figure I. 4 Simplified representation of phenylpropanoid metabolic route. Modified from Schwab *et al.* (2009).

4.4. Volatile compounds

4.4.1. Volatiles in fruits

Fruit aroma is a quality trait that has a major impact in consumers' preference. Fruit smell is composed by a complex mixture of volatiles, including esters, terpenoids, ketones, lactones, aldehydes and alcohols. Although these compounds account only for 10^{-7}-10^{-4}% of the fresh fruit weight, they can be detected by human's olfactory system and are regarded as the characteristic flavor of the fruits. Thousands of volatile compounds have been identified in fruits and they are known to be dependent on various factors such as fruit species and variety, maturity stage, postharvests treatment and analysis technique. Generally its production increases with ripening (Goff & Klee 2006). Aroma volatiles are represented in three main categories: fatty acid, amino acid and carbohydrate derivatives (reviewed by Schwab *et al.* 2008).

4.4.2.Fatty acid pathway

Fatty acids are the most important precursors for most fruit aroma volatiles, including straight-chain aldehydes, alcohols, esters, lactones and ketones. These compounds are synthesized mainly through the lipoxygenase (LOX) pathway and α- β-oxidation (Figure I. 5).

In the LOX pathway linoleic (18:2) and linolenic (18:3) acid are catalyzed into hydroperoxide isomers, which are further cleaved by hydroperoxyde lyase (HPL) to form hexanal and hexenal, respectively. The aldehydes are subsequently reduced to the corresponding C6 alcohols by alcohol dehydrogenase (ADH). Alcohol acyl transferase (AAT) catalyses the final linkage of an acyl moiety and an alcohol to form esters (Schwab *et al.* 2008).

Straight fatty acids can also be degraded via α- and β-oxidation pathways, however specific mechanisms in plants are not well understood. The α-oxidation catalizes free fatty acids probably through one or two long-chain aldehyde intermediates by the dual action of a α-dioxigenase/peroxidase and NAD^+ oxidoreductase (reviewed by Schwab *et al.* 2008). The β-oxidation liberates acyl-coA units from a former fatty acid by the action of acyl-coA oxidase (ACX). During the β-oxidation the breakdown of acyl-coA can be stopped between β-oxidation cycles or inside the reaction sequence as a result of many factors, resulting in the liberation of volatile lactones (Husain 2010).

4.4.3.Amino-acid pathway

Amino acids are known as important aroma volatile precursors. The main amino acids responsible for the biosynthesis of aromatic volatile compounds are branched-chain and aromatic amino acids like leucine, isoleucine, valine, alanine, phenylalanine, tyrosine and tryptophan. However, specific steps of the amino acids catabolism to form volatile compounds remain unclear in plants.

4.4.4.Carbohydrate pathway

Furanones and pyrones derive directly from hexoses and pentoses without prior degradation of the carbon skeleton. They have a very low odor threshold and contribute to fruit aroma. In strawberry, two furanones, furaneol (4-Hydroxy-2,5-dimethyl-3(2*H*)-furanone) and mesifurane (2,5-dimethyl-4-methoxy-3(2H)-furanone) are considered important flavor contributors giving sweet caramel notes to the scent (Schwab *et al.* 2008).

Terpenoids are also derived from carbohydrates. They are formed from basic 5C units of isopentenyl pyrophosphate (IPP) or its isomer dimethylallyl pyrophosphate (DMAPP). IPP and DMAPP arise from either the plastidic methylerythriol phosphate or the cytosolic mevalonate pathway. These 5C units are condensed to pyrophosphate precursors of terpenoids that are converted to final products by terpene synthases (TPS).

4.4.5.Volatile compounds in strawberry

Strawberries are very appreciated for their flavor and aroma. More than 360 volatiles have been detected in strawberry extracts (reviewed by Latrasse 1991), however not all of them are equally important for the aroma. Several studies have proved that only around 20 volatiles

contribute to strawberry smell (Schieberle & Hofmann 1997; Ulrich *et al.* 1997; Jetti *et al.* 2007; Ulrich *et al.* 2007).

Great differences are observed in aroma patterns among the *Fragaria* genus. Distinctive volatile composition of woodland strawberry (*F. vesca*) provides a more intense and fruity aroma to this wild species than to the commercial fruits (*F. x ananassa*) (Ulrich *et al.* 1997; Ulrich *et al.* 2007; Dong *et al.* 2013). However important differences have also been observed between *F. x ananassa* varieties with extreme quality phenotypes (Zorrilla-Fontanesi *et al.* 2012; Schwieterman *et al.* 2014).

Linalool, nerolidol, α-pinene and limonene are the quantitatively predominant volatile terpenes in cultivated strawberry and can reach up to 20% of total fruit volatiles (Loughrin & Kasperbauer 2002) however linalool and nerolidol are not detected in the volatile profile of wild strabwerries (Aharoni *et al.* 2004). Furaneol is a minor constituent of the total volatile compounds present in strawberry fruit but, because of its low odor threshold, it is considered as an important contributor to strawberry flavor.

Although composition of strawberry aroma has been exhaustively studied; complete volatile biosynthetic pathways have not been elucidated yet. However, a bunch of genes and enzymes associated with the synthesis of one or a group of volatile compounds have been described.

The production of linalool and nerolidol on the one side and of α-pinene on the other side have shown to be linked to specific alleles of the terpene synthases *FaNES1* and *FvPINS* respectively (Aharoni *et al.* 2004). The mesifurane synthesis has been associated to the O-methyltransferase (*FaOMT*) (Zorrilla-Fontanesi *et al*. 2012) and an omega-6 fatty acid desaturase (*FaFAD*) is linked to γ-decalactone content (Chambers *et al.* 2014; Sanchez-Sevilla *et al.* 2014). Eugenol biosynthesis is mediated by two eugenol synthases (*FaEGS1* and *FaEGS2*) and controlled by one R2R3 MYB transcription factor (*FaEOBII*) in strawberry (Aragüez *et al.* 2013; Medina-Puche *et al.* 2015). Esters production is associated with alcohol acyltransferases, SAAT in *F. x ananassa* (Aharoni *et al.* 2000) and VAAT in *F. vesca* (Beekwilder *et al.* 2004), that show high sequence similarity but different substrate preferences. Finally a lipoxygenase (*LOX*) enzyme was seen to be involved in strawberry maturity (Leone *et al.* 2006).

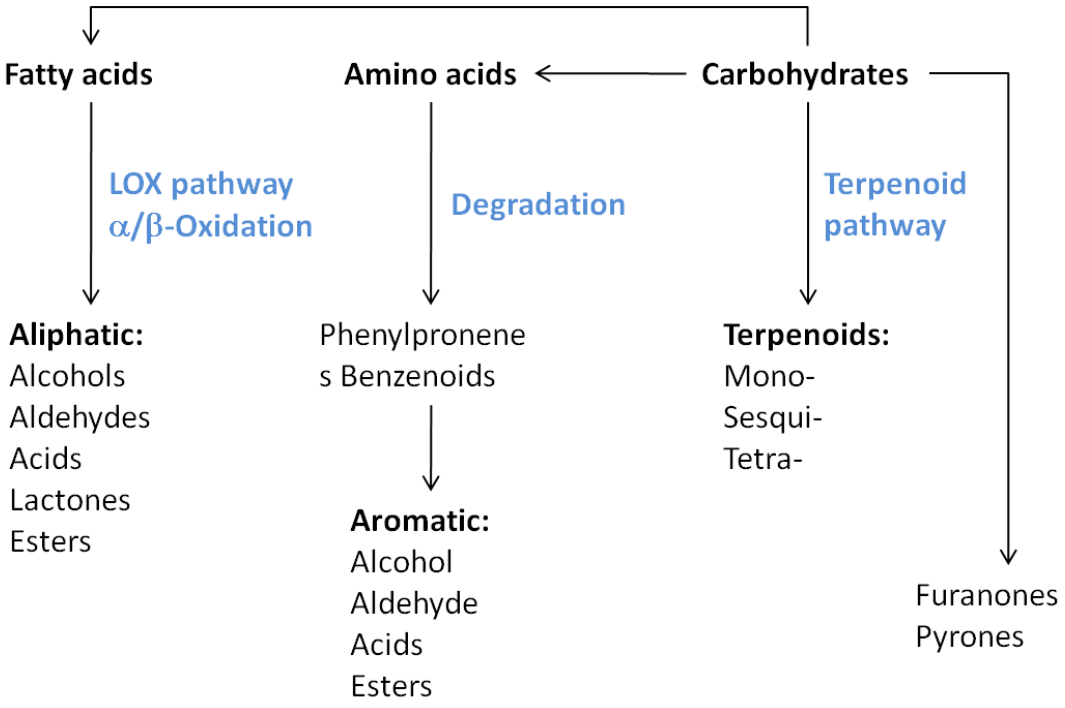


Figure I. 5 Simplified representation of volatile synthesis patways. Modified from Schwab *et al.* (2008).

Objectives

The general objective of this work is the genetic study of wild strawberry (*F. vesca*) aiming to detect QTL controlling agronomical, nutritional and organoleptic traits and propose candidate genes for further research. The specific objectives to address this main goal are:

1. Development of a Near Isogenic Lines collection in *F. vesca.*
2. Phenotypic characterization and QTL mapping of strawberry quality traits in the *F. vesca* NIL collection.
   2.1. Characterization of agronomical quality traits.
   2.2. Metabolic characterization of fruit quality traits.
      2.2.1. Nutritional traits: Identification and quantification of sugars and (poly)phenols.
      2.2.2. Organoleptic traits: Identification and relative quantification of volatile compounds, responsible of strawberry aroma.
3. Detection of differentially expressed genes in NILs harbouring high number of major QTL for organoleptic and nutritional traits.

Chapter I

A near isogenic line (NIL) collection in diploid strawberry: Development and use in genetic analysis.

Part of the content of this chapter has been published in Theoretical and Applied Genetics in co-authorship with Julio Bonet, PhD.

Theor Appl Genet
DOI 10.1007/s00122-015-2503-3

ORIGINAL PAPER

A near-isogenic line (NIL) collection in diploid strawberry and its use in the genetic analysis of morphologic, phenotypic and nutritional characters

María Urrutia[1] · Julio Bonet[1,2] · Pere Arús[1] · Amparo Monfort[1]

Received: 15 October 2014 / Accepted: 20 March 2015

Abstract

Key message **First near-isogenic line collection in diploid strawberry, a tool for morphologic, phenotypic and nutritional QTL analysis.**

Abstract Diploid strawberry (*Fragaria vesca*), with a small genome, has a high degree of synteny with the octoploid cultivated strawberry (*F.* × *ananassa*), so can be used as a simplified model for genetic analysis of the octoploid species. Agronomically interesting traits are usually inherited quantitatively and they need to be studied in large segregating progenies well characterized with molecular markers. Near-isogenic lines (NILs) are tools to dissect quantitative characters and identify some of their components as Mendelian traits. NILs are fixed homozygous lines that share the same genetic background from a recurrent parent with a single introgression region from a donor parent. Here, we developed the first NIL collection in *Fragaria*, with *F. vesca* cv. Reine des Vallées as the recurrent parent and *F. bucharica* as the donor parent. A collection of 39 NILs was identified using a set of single sequence repeat markers. The NILs had an average introgression of 32 cM (6 % of genome) and were phenotyped over several years in two locations. This collection segregates for agronomic characters, such as flowering, germination, fruit size and shape, and nutritional content. At least 16 QTLs for morphological and reproductive traits, such as round fruits and vegetative propagation, and seven for nutritional traits such as sugar composition and total polyphenol content, were identified. The NIL collection of *F. vesca* can significantly facilitate understanding of the genetics of many traits and provide insight into the more complex *F.* × *ananassa* genome.

Communicated by A. H. Schulman.

M. Urrutia and J. Bonet contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-015-2503-3) contains supplementary material, which is available to authorized users.

✉ Amparo Monfort
amparo.monfort@irta.es

[1] IRTA, Center for Research in Agricultural Genomics (CSIC-IRTA-UAB-UB), Campus UAB, 08193 Bellaterra, Barcelona, Spain

[2] Present Address: Plant Response Biotech S.L. Centro de Empresas, Parque Científico y Tecnológico Montegancedo, 28223 Pozuelo de Alarcón, Madrid, Spain

Introduction

Strawberry, one of the economically most important soft fruits, belongs to the *Fragaria* genus. This includes 23 species with different ploidy levels, from diploid wild strawberries such as *F. vesca* to octoploid cultivated (*F.* × *ananassa*) and decaploid (*F. iturupensis*). Diploid species ($2n = 2x = 14$) can be used as models for genetic analysis of cultivated strawberry, as synteny studies have demonstrated that the octoploid genome is composed of four genomes that are essentially collinear and syntenic with the genomes of the diploid species (Vilanova et al. 2008; Rousseau-Gueutin et al. 2008; Sargent et al. 2009). The reference linkage map of diploid strawberry was constructed from a cross between *F. vesca* ssp. *vesca* (FDP815) and *F. bucharica* with 68 single sequence repeat (SSR), one SCAR and six STS markers (Sargent et al. 2004) and three morphological traits: seasonal perpetual flowering, runner development and pale-dark green leaf color. Several versions increasing the number and source of markers have

Abstract

Diploid strawberry (*Fragaria vesca*), with a small genome, has a high degree of synteny with the octoploid cultivated strawberry (*F*. x *anannassa*), so can be used as a simplified model for genetic analysis of the octoploid species. Agronomically interesting traits are usually inherited quantitatively and they need to be studied in large segregating progenies well characterized with molecular markers. Near isogenic lines (NILs) are tools to dissect quantitative characters and identify some of their components as Mendelian traits. NILs are fixed homozygous lines that share the same genetic background from a recurrent parent with a single introgression region from a donor parent. Here we developed the first NIL collection in *Fragaria*, with *F. vesca* cv. Reine des Vallées as the recurrent parent and *F. bucharica* as the donor parent. A collection of 39 NILs was identified using a set of SSR markers. The NILs had an average introgression of 32 cM (6% of genome) and were phenotyped over several years in two locations. This collection segregates for agronomic characters such as flowering, germination, fruit size and shape, and nutritional content. At least 16 QTLs for morphological and reproductive traits, such as round fruits and vegetative propagation, and 7 for nutritional traits such as sugar composition and total polyphenol content, were identified. The NIL collection of *F. vesca* can significantly facilitate understanding of the genetics of many traits and provide insight into the more complex *F*. x *ananassa* genome.

Keywords: *Fragaria vesca*, QTLs, introgression, nutritional

Key message: First near isogenic line collection in diploid strawberry, a tool for morphologic, phenotypical and nutritional QTL analysis

Introduction

Strawberry, one of the most economically important soft fruits, belongs to the *Fragaria* genus. This includes 23 species with different ploidy levels, from diploid wild strawberries such as *F. vesca* to octoploid cultivated (*F.* x *ananassa*) and decaploid (*F. iturupensis*). Diploid species (2*n* = 2*x* = 14) can be used as models for genetic analysis of cultivated strawberry, as synteny studies have demonstrated that the octoploid genome is composed of four genomes that are essentially collinear and syntenic with the genomes of the diploid species (Vilanova *et al.* 2008; Rousseau-Gueutin *et al.* 2008; Sargent *et al.* 2009). The reference linkage map of diploid strawberry was constructed from a cross between *F. vesca ssp. vesca* (FDP815) and *F. bucharica* with 68 SSR, one SCAR and six STS markers (Sargent *et al.* 2004a) and three morphological traits: seasonal perpetual flowering, runner development and pale/dark green leaf colour. Several versions increasing the number and source of markers have been done (Sargent *et al.* 2006; Sargent *et al.* 2008; Vilanova *et al.* 2008; Ruiz-Rojas *et al.* 2010) until reaching its genome anchoring (Schulaev *et al.* 2011).

In this paper we describe the development of a collection of near isogenic lines (NILs) using two accessions of two diploid species: *F. vesca*, used as recurrent parent, and *F. bucharica*, as donor parent. *F. vesca,* the most widely distributed diploid, has been reported as a principal contributor of the octoploid genomes (Rousseau-Gueutin *et al.* 2009). It includes the 'woodland' strawberry *F. vesca* ssp. *vesca*, a fruit of minor economic importance, found in the wild throughout the northern hemisphere and which has been described as a predominantly selfing species (Arulsekar & Bringhurst 1981). The other diploid species, *F. bucharica,* is a self-incompatible (Bošković *et al.* 2009) stolon-propagated species, belonging to the Asiatic group, probably of hybrid origin (Staudt 2006; Staudt 2008). These two species are phylogenetically close (Rousseau-Gueutin *et al.* 2008) but morphologically divergent, allowing fertile crosses between them, with a wide range of segregating phenotypes from the offspring. The selected *F. bucharica* accession was that used by Sargent *et al.* (2004a) as one of the parents of the reference map for diploid strawberry.

NIL collections are important resources for analysing complex characters. A NIL is identical to an original genotype, except for a single DNA introgressed fragment from a donor line. Comparison of the phenotypes of a NIL and its recurrent parent means any significant difference can be attributed to genetic factors in the introgressed fragment. In addition, phenotyping can be done with high precision, as each NIL can be duplicated as many times as necessary and grown under different environments, circumventing some of the major limitations of conventional mapping populations (Paterson *et al.* 1988; Eshed & Zamir 1995). A collection of NILs covering the whole genome of a species can be considered a genomic library containing overlapping introgressions from a donor genome on a recipient or recurrent genome. NIL collections can also be used to introduce new genetic variability from wild species into elite cultivars (Tanksley & McCouch 1997; Zamir 2001), to study the genetic bases of heterosis (Melchinger *et al.* 2007; Fernández-Silva *et al.* 2009), to study interactions between different QTL and QTL with the environment (Monforte et al. 2001), or as initial materials for map-based cloning (Frary *et al.* 2000; Fridman *et al.* 2000; Fujita *et al.* 2013). Several NIL collections have been developed as genomic resources in *Arabidopsis thaliana* (Koumproglou *et al.* 2002; Keurentjes *et al.* 2007; Fletcher *et al.* 2013) and in crops such as wheat (Pestsova *et al.* 2001), lettuce (Jeuken & Lindhout 2004) and melon (Eduardo *et al.* 2005). There are also partial collections

covering specific genomic regions of *Brassica oleracea* (Ramsay *et al.* 2000), rice (Wan *et al.* 2004; Zhang *et al.* 2009), maize (Szalma *et al.* 2007), sorghum (Harris *et al.* 2007), tobacco (Lewis *et al.* 2007), barley (Marcel *et al.* 2007) and soybean (Kopisch-Obuch & Diers 2006; Bolon *et al.* 2010).

The diploid strawberry NIL collection was phenotyped for a set of characters responsible for morphological and phenotypical characteristics of different parts of the plant and for characters related with the chemical composition of the fruit, which have allowed us to find the positions of a set of Mendelian genes and QTL responsible for their variation. We propose this new resource as a tool for the broad Rosaceae community. It will be of use for the in-depth study of the genetics of important characters in this family and to complement other resources which have been developed in *F. vesca*, such as the reference linkage map (Sargent *et al.* 2004a), the collection of mutants (Oosumi *et al.* 2010) and the whole genome sequence (Shulaev *et al.* 2011).

Material and Methods

Plant materials and NIL development

One of the selected parental lines was *F. vesca* cv. 'Reine des Vallées' (RV), a French, non-runnering, day-neutral, homozygous cultivated variety with long petioles, large leaflets, small flowers, and bright-red, long conic berries with a powerful aroma and low susceptibility to the common pests in strawberry crops. The other was *F. bucharica* (PI657844, also known as FDP601) (FB), one of the parents of diploid strawberry reference map. It is a short-day variety with short petioles, small leaflets and large flowers. The berries, with no commercial value, are necked, round flattened on top, and dark wine-red. The *F. vesca* parent was chosen as the recurrent genome and *F. bucharica* parent as the donor.

Pollen from an F_1 individual of a cross between RV x FB was used as the donor to pollinate an *F. vesca* individual. The resulting backcross 1 (BC_1) seeds were germinated and DNA from the seedlings extracted and genotyped using seven molecular markers (EMFvi072, EMFvi099, EMFv029, CEL-1, EMFvi018, EMFn228 and EMFv021) all of them microsatellites, located in one of the extremes of the seven linkage groups of the diploid *Fragaria* reference map (Supplementary Figure CI. 1) and described in supplementary Table CI. 1. Individuals with less than four introgressions were selected and then genotyped for seven additional SSR (Single Sequence Repeat) markers (EMFv025, EMFv003, EMFn207, EMFv007, EMFvi108, EMFv160BC and EMFv023) mapping to the other linkage group ends (see reference in supplementary Table CI. 1). Individuals harbouring less than five of these markers in heterozygosity, therefore containing at least the same number of introgressions, were selected and genotyped for a minimum of 17 and up to 28 additional markers located in intermediate regions of the seven linkage groups (Supplementary Table CI. 1), depending on the genetic composition of each line. This gave a set of plants with a low number of introgressions covering the whole genome of the donor parental line.

Selected BC_1 plants were used as female parents for a second backcross with the *F. vesca* parental line. The resulting seeds were germinated and each family evaluated for molecular markers located at the extremes of each introgression. A reduced group of plants per family was selected for further self-pollination until lines with single introgressions were obtained. Individuals resulting from a further round of self-pollination were screened with a set of 77 segregating loci (Supplementary Table 1), well distributed across the genome, allowing characterization of the introgressed fragments and eliminating any plants with more than one introgression from the donor genome. The size of the introgressions and genome percentages were calculated by adding together the length between the two markers at the extreme of the introgression fragment, and half the distance of the two intervals between these markers and those flanking the introgression that were homozygous for the RV allele.

The crossing scheme for the development of the *F. vesca* - *F. bucharica* collection of NILs is presented in Figure CI. 1.

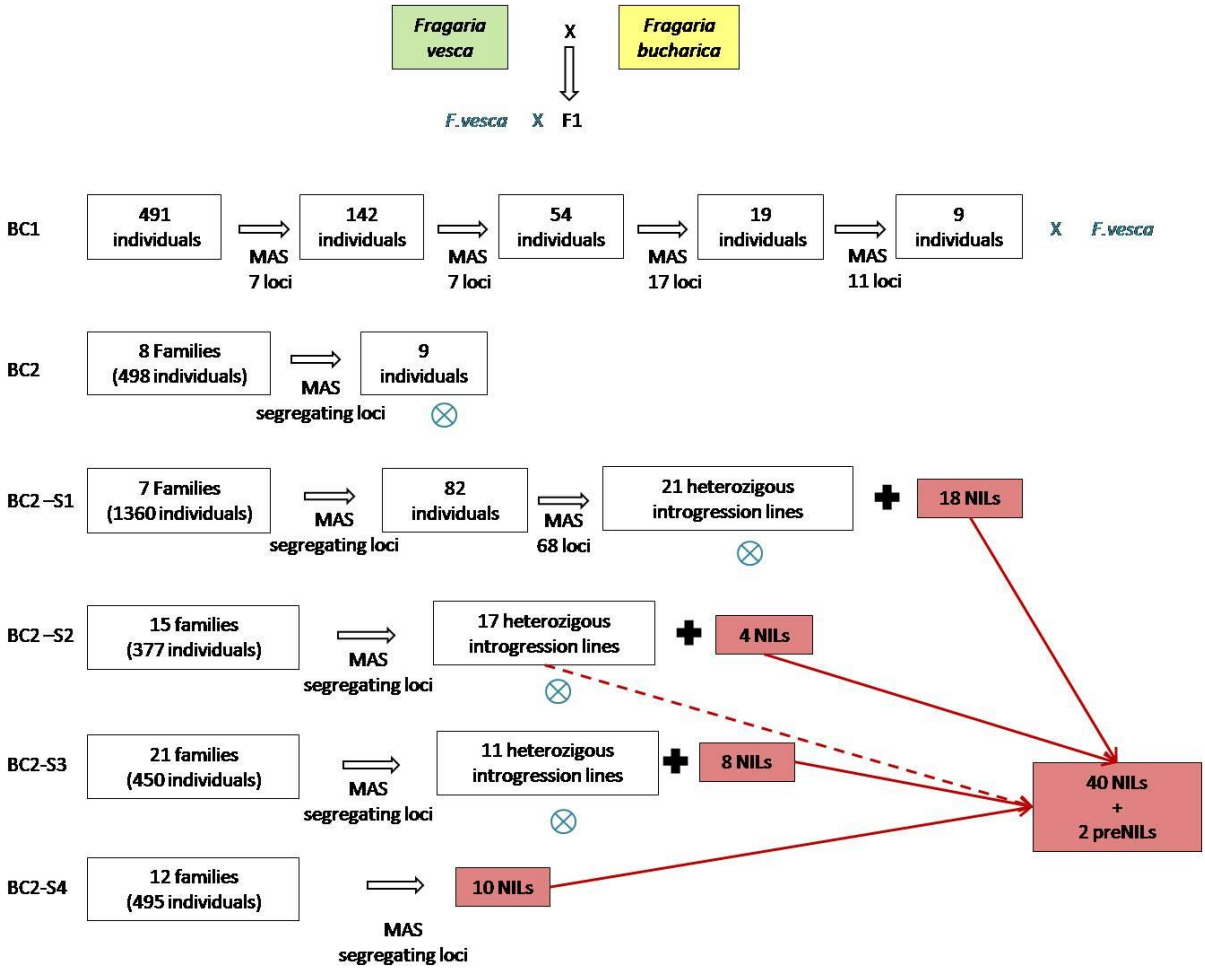


Figure CI. 1. Crossing scheme for the development of a collection of NILs in *F. vesca*. Each box contains the number of individuals chosen after a process of marker-assisted selection (MAS). Orange boxes include the lines selected as final homozygous NILs in each generation. The red box includes the complete set of NILs.

Molecular marker analysis and map construction

Tissue from young leaves was collected and disrupted by mechanical shaking using a tissue homogenizer. DNA was extracted using the method of (Doyle & Doyle 1990) modified by the addition of 2% PVP-40. For PCR, the DNA was quantified and 20 ng per reaction SSR and CAPS (Cleaved Amplified Polymorphic Sequences) markers were used to locate the introgressed regions. A set of 16 new SSR markers in the region of previously mapped RFLP markers (Vilanova et al. 2008) was developed (Supplementary Table CI. 2) and mapped, to obtain additional markers in map regions with insufficient coverage of SSR markers. The region was identified by comparing the Strawberry Genome Assembly version 8 (Genome Database for Rosaceae, www.rosaceae.org (Jung *et al.* 2014)) with RFLP original sequences (www.rosaceae.org) (Vilanova *et al.* 2008). Alignments were accepted when the estimated similarity between aligned sequences (RFLP and Genome) was statistically significant (*E* value ≤ 0.05). New SSR motifs were located near to the aligned sequences, with the

primer pairs designed using Websat (Martins *et al.* 2009) and Primer3 (Rozen & Skaletsky 1999). To reduce the cost of marker detection, forward primers were modified by adding a 17 nucleotide M13-tail (5´ GTAAAACGACGGCCAGT 3´) to the 5´extreme (Boutin-Ganache *et al.* 2001), allowing hybridization of any forward primer with a unique fluorescent-labelled primer in PCR reactions.

A new marker linked to the perpetual flowering locus in linkage group 6 (LG6) was designed to amplify a specific small indel. Using the method of (Iwata *et al.* 2011) we sequenced a 1,000 bp region on LG6, scaffold 0513102 (Shulaev *et al.* 2011), that contained a 2 bp deletion in the *F. vesca* TFL homolog which causes a perpetual flowering phenotype (Koskela *et al.* 2012). Two *FvTFL* primers were designed to specifically amplify this deletion using Primer3 (Rozen & Skaletsky 1999). The forward primer was TCGGAACCTCTAGCTGTTGG and the *FvTFL* reverse primer, GAGCTCATGTCCATTGCAGA.

PCR reactions were in a final volume of 10 µl containing: 10 ng of DNA, 1X PCR buffer (50 mM KCl, 10 mM Tris-HCl pH 8.3, 0.001% gelatin), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 1U of DNA polymerase Ampli*Taq* (Perkin-Elmer, IL, USA), 0.15 µM forward primer, 0.2 µM reverse primer and 0.2 µM of M13 labelled primer, with an identical sequence to the added tail on the forward primers. The choice of dye label, 6-FAM, VIC, PET or NED (Applied Biosystems, CA, USA), depended on further multiloading capillary electrophoresis. PCR amplifications were run on a PE9600 thermal-cycler (Applied Biosystems, CA, USA) as follows: 2 min of initial denaturing at 94°C; 10 cycles of 15 sec at 94°C, 15 sec at annealing temperature and 30 sec at 72°C; 25 cycles of 15 sec at 94°C, 15 sec at 50°C and 30 sec at 72°C, followed by a final extension of 5 min at 72°C.

Amplicons were visualized by capillary electrophoresis in an ABI3130xl Genetic Analyzer (Applied Biosystems, CA, USA), run using 2 µl of a mix containing three differently labelled PCR products, 0.3 µl of LIZ-500 ladder and 12 µl of deionized formamide. Data generated by capillary electrophoresis were analysed using the GENEMAPPER software application (Applied Biosystems, CA, USA). For the 4 CAPS marker, PCR products were digested as described by Ruiz-Rojas *et al.* (2010), and electrophoresed in agarose gel. Segregation data of new markers were added to the segregation matrix of the FVxFB diploid *Fragaria* reference map (Sargent *et al.* 2006, Vilanova *et al.* 2008, Ruiz-Rojas *et al.* 2010, Zorrilla-Fontanesi *et al.* 2011b). RFLP markers on the previously developed FVxFB map were excluded from the analysis.

Chi-squared tests of goodness-of-fit, to an expected segregation ratio of 1:2:1 for co-dominant markers and 3:1 for dominant markers, were carried out for new markers using the JoinMap® software (Van Ooijen 2006). Linkage analysis was performed and the map constructed using MAPMAKER 3.0 (Lander *et al.* 1987), applying the Kosambi mapping function. Linkage groups were constructed and marker order determined using a minimum LOD score threshold of 3.0. Markers selected for marker assisted selection (MAS) and new markers obtained are shown in Supplementary Figure CI. 1. Graphical genotyping software GGT2.0 (van Berloo 2008) was used to present the NILs and evaluate the genome composition of the NIL collection.

Phenotyping and genetic analysis

NIL collection seeds were germinated in *in-vitro* chambers, conditions set were: photoperiod 12h/12h light/dark, temperature 24ºC. Seeds were kept in petri-plates over humid filter paper until germination occurred (appearance of the root and both cotyledons open). Four to eight plants of each genotype were transplanted to soil pots and grown under glasshouse conditions (photoperiod 12h/12h light/dark, day temperature between 22 to 24ºC and 17ºC during night, relative humidity 40-50%.) for 8 weeks at the Center for Research in Agricultural Genomics (CRAG) in Bellaterra (latitude: 41° 29'N, longitude: 2° 06'E). When plants were two months old were moved to shaded greenhouses in Cabrils (latitude: 41º 31'N, longitude: 2º 22' E, altitude 146m above sea level, coastal Mediterranean climate) and at the Center of Torre Marimon (TM) in Caldes de Montbui (latitude: 41º 36'N, longitude: 2º 10' E, altitude 203m above sea level, pre-coastal Mediterranean climate). The shaded greenhouses where not provided with supplementary artificial light or heating so plants were exposed to natural climatic conditions and photoperiod in north Spain (latitude 41º) from march to September (when the harvest was over). The agronomical practices were the usual for strawberry-fruit production. The population was established over three years in Cabrils (2010, 2011 and 2012) and two years in TM (2011 and 2013). Fruits were collected in two different periods or harvests, in May and July.

Qualitative morphological traits and phenotypical events phenotypes were evaluated visually in all seasons. For floral traits, we evaluated the presence of pink spot on the petal base (*PSP*), extra petal number (*EPN*, above 5 petals), short and long floral stem (*SFS* and *LFS*, respectively), seasonal flowering or short day flowering (*SDF*) as plants blooming only once a year in short day as *F. bucharica* parent, and a trait in plants flowering later than one year of germination (*LLF* for late- late flowering). For other plant traits we evaluated the presence of runners (*R*), dwarfism (*DWARF*) and delayed germination (*DOG*, as more than 15 days later than in RV).

The basic requirement for accepting these traits was that they were stable throughout the studied seasons, and that all the lines containing the introgressed region had the same phenotype. Flowering time was measured as the number of days to produce the first flower after germination, with early flowering (EF) being shorter and late flowering (LF) longer than the flowering time for RV, in all lines with the same introgression.

For berry measurements, 20 fruits per line were collected during the spring (May) and summer (July) of 2011 and 2012 and measured using ImageJ (Schneider *et al.* 2012) (http://imagej.nih.gov/ij/). Fruits of *F. vesca* cv. 'Reine des Vallées' were collected as the reference parent, but not enough fruits of *F. bucharica* could be collected. Photographs were taken of all the collected fruits under standardized conditions and the longest straight line of the fruits measured lengthwise (Length) and crosswise (Width). We then divided the length by the width to obtain a parameter (called fruit index) that measured fruit shape (round fruits being close to 1 and long fruits having higher values). All the fruits collected from a NIL plant were weighed together and the average used as the value for fruit weight. Fruit traits were analyzed for five independent harvests, in two years and at two locations (2011 and 2012 in May and July in Cabrils, and 2011 in May in TM).

Three biological replicates of 5 g of fruit from each genotype, in 2011, 2012, and 2013, were collected for nutritional analyses. The fruits were immediately frozen in liquid nitrogen, ground, divided in subsamples and stored at -80ºC until extraction. The sugars in the samples were ethanol

extracted following the protocol described by Pérez *et al.* (1997), then quantified by HPLC (Agilent 1200 HPLC, Agilent Technologies, CA, USA) using a Zorbax Carbohydrate Analysis column (Agilent Technologies, CA, USA). Total polyphenolic content was determined by the Folin-Ciocalteu method following the extraction and detection protocol described by Singleton *et al.* (1999). Within the population, the distribution parameters for the fruit sugar content (fructose, glucose and sucrose) were quantified as mg/g of fresh weight, and the total polyphenol content as mg of gallic acid equivalents per gram fresh weight.

Statistical analyses were using the JMP®8.0.1 statistical package (SAS Institute, NC, USA) and R 2.15.1 (RCoreTeam 2012) with the R-studio interface (Rstudio 0.92.501, Rstudio, MA, USA). Distribution of all traits among the population was evaluated by a Shapiro-Wilk normality test, and the range of variation was also calculated by skewness and kurtosis of the distribution for every trait and harvest. The mean values and standard deviation for RV, and every line of the population, were calculated for each trait and harvest, Correlation of the quantitative traits in different years was calculated separately for nutritional and fruit shape characters. Means were compared with the recurrent parent by a Dunnett test with $\alpha \leq 0.05$ (Dunnett 1955). The QTL effects are presented in relation to the recurrent parental control. Qualitative traits and QTLs were mapped to a specific bin only when all the NILs carrying this bin had a significant effect on the phenotypic trait of study.

Results

Linkage map of FVxFB

The diploid strawberry linkage map of *F. vesca* x *F. bucharica* F_2 population, (Sargent *et al.* 2006, Ruiz-Rojas *et al.* 2010), was used to monitor the introgressions for the selection of the NIL collection. This map, including new markers developed here (14 SSRs and one CAPS) and excluding RFLPs, covered 541.3 cM, 13.2 cM more than that of Ruiz-Rojas *et al.* (2010), and had 219 microsatellites, one SCAR, 22 STS, 53 CAPS/dCAPS, one SNP, one indel and one morphological marker. This map is presented in Supplementary Figure CI. 1. Segregation distortions were observed on 48.8% of the mapped markers ($P \leq 0.05$), especially on LGs 2, 4, 5 and 7.

Some of the new loci on the map reduced the extension of the original FVxFB linkage groups because segregation changes from dominant in RFLP to codominant in SSR. The RFLP markers AC-32, AC-24 and AG-53 defined the top region of LG4 in the map of Vilanova *et al.* (2008). The PCR markers developed here to replace these RFLPs (CFV-3135, CFV-3138 and CFV-3819 respectively) were located on the central region of LG4, in a cluster previously described by Sargent *et al.* (2004a). In addition, the RFLPs MC-45 and Co-MET, replaced by CFV-3117 and CFV-3217, were moved to the central positions of LG7, allowing a 19.1 cM reduction of this group. There were minor rearrangements in the placement of some of the dominant marker loci presented by Sargent *et al.* (2006), due to a better estimation of mapping distances with the addition of the novel codominant SSR loci (e.g. EMFv164 *vs.* CFV-164).

Development of the NIL collection

The BC_1 seedlings (491 individuals) were analyzed for allele composition using seven SSRs, one for each linkage group (see Supplementary Tble CI. 1). On average, the observed heterozygosity at these loci was $H_o = 0.57$. A set of 142 individuals (27.6%) was selected for a second round of PCR, for allele characterization of seven additional markers at the opposite end of LG. The average observed heterozygosity for these loci was $H_o = 0.52$. The marker showing the highest segregation distortion was EMFv160BC ($\chi^2 = 31.11$), located at the end of LG6 with maximum of heterozigosity. A group of 54 plants (38%) was selected for further genome characterization. Twenty eight additional loci were analyzed and nine BC_1 individuals, with introgressions covering the whole *F. bucharica* genome, were selected for a second backcross and segregating families generated. From the nine BC_1 individuals selected for their low number of introgressions and maximum genome coverage, only eight BC_2 families were generated. The donor genome percentage of BC_1 lines was 35.2%, on average, ranging from 25% to 49% in 28 introgressions, distributed as follows: eight introgressions on LG2, five on LG6, four on LG3, three on LG1, LG4 and LG7, and two on LG5. Whole introgressed linkage groups were selected for the seven linkage groups of the map (35% of the selected introgressions). These selected BC_1 lines had three or four introgressions per line. BC_2 families (498 individuals) were analyzed with markers that were in the extremes of, and inside for long introgressions, the introgressed fragments detected in their BC_1 progenitors. A total of 50 markers were chosen to genotype all introgressed regions (Supplementary Table CI. 1). Nine BC_2 individuals were selected as introgression donors for further generations. Five individuals had only one introgression, one had two and three had three introgressions. The average percentage of *F. bucharica* fragments (in cM) on the selected BC_2

individuals was 14.6% (range 6.2% - 24.6%) in 15 introgressions distributed as follows: three introgressions on LG3 and LG4, two on LG1, LG2, LG5 and LG6, and a single introgression on LG7. Entire introgressed LGs were obtained for LG2, LG3, LG4, LG5, LG6 and LG7 (37.5% of the introgressions).

Selected individuals were self-pollinated and the resulting seeds germinated to generate segregating families. Between one and four selfing generations were necessary to obtain homozygous lines with only one introgression and covering together almost all of the donor parent genome. Depending on the genotype of its parental line, selfed offspring (2,682 in total) were analyzed with different subsets of 77 SSRs (see Supplementary Table CI. 1) in order to select only the homozygous target regions. This fine characterization also allowed more exact location of the introgressed regions. Lines in the BC_2 offspring with LG6 introgressed fragments blossomed more than one year after germination, and NIL extraction was delayed relative to the rest of the population. Until now, progeny obtained from line Fb6:71h-101h never fixed region around CFvCT017 marker, and it has only been possible to obtain as heterozygous introgression, and line Fb6:30-39h only has in homozygosis part of introgressed region (see Figure CI. 2).

Segregation at six of the seven loci analyzed during the first stage of screening of the BC_1 population was distorted ($p < 0.05$; $N = 491$). The top LG2 region (EMFvi099) and the distal part of LG5 (EMFvi018) had the most skewed segregations ($\chi^2 = 170.96$ and $\chi^2 = 65.92$ respectively), with a higher frequency of *F. bucharica* alleles in both cases. The linkage group with the lowest *F. bucharica* allele frequency was LG1 (only 27.4% of the plants had *F. bucharica* genotypes at LG1 loci). Biased segregation ratios were detected also in BC_2 families, especially for the markers located at the end positions of LG5 (91.8% of *F. bucharica* genotypes). The lowest frequency of *F. bucharica* alleles was observed for markers located on LG1 (39.2%). Significantly distorted ratios were also detected in the segregation of markers on LG2, LG3 and LG6.

The strawberry NIL collection consisted of a final set of 39 homozygous lines and two heterozygous lines (Figure CI. 2). Among these, 18 correspond to BC_2S_1 lines, four to BC_2S_2 lines, seven to BC_2S_3 and ten correspond to BC_2S_4 ("S_n" is the nth selfing generation) lines. Assuming no double crossovers in the introgressed fragments, each NIL contained a single homozygous introgression, with an average introgression size of 32.9cM (6% of the donor genome) (Table CI. 1). Each linkage group was represented by an average of six NILs with overlapping introgressions. Only a 19.3cM region on LG1, defined by markers UDF002 and CFV-4021, was not covered by *F. bucharica* introgressions. Furthermore, lines covering a region of LG6 (Fb**6:**30-39h and Fb**6:**71h-101h), which showed large period to flowering (more than one year), retained heterozygous introgressions.

Taking the collection of NILs of each linkage group into consideration, the ensemble of recombination breakpoints of the *F. bucharica* genome introgressed fragments define a set of identifiable regions, or bins, defined by two consecutive breakpoints with no recombination between them. This population had 37 bins (4-7 per linkage group), with an average size of 14.2 cM per bin and ranging between 3.2 and 65.7 cM. Each bin represents on average a small part, 2.6 %, of the genome (Table CI. 1). A minimal subset of nine of these lines covering the whole genome (except for the missing fragment of LG1 and for the heterozygous LG6 fragment) may be used to locate any segregating locus to its chromosome when resources are limited.

Figure CI. 2. Graphical genotypes of the *F. vesca* collection of 39 NILs and 2 heterozygous NILs. *F. bucharica* homozygous introgressions are shown in black and heterozygous introgressions in pink. The *F. vesca* genetic background is shown in green. The NIL names are indicated on the right. The first number indicates the LG carrying the introgression. The following two numbers, separated with a hyphen, indicate the marker position at the start and end point of the introgression in centiMorgans, respectively. Dotted NILs indicate the minimal set of lines covering the entire *F. bucharica* genome.

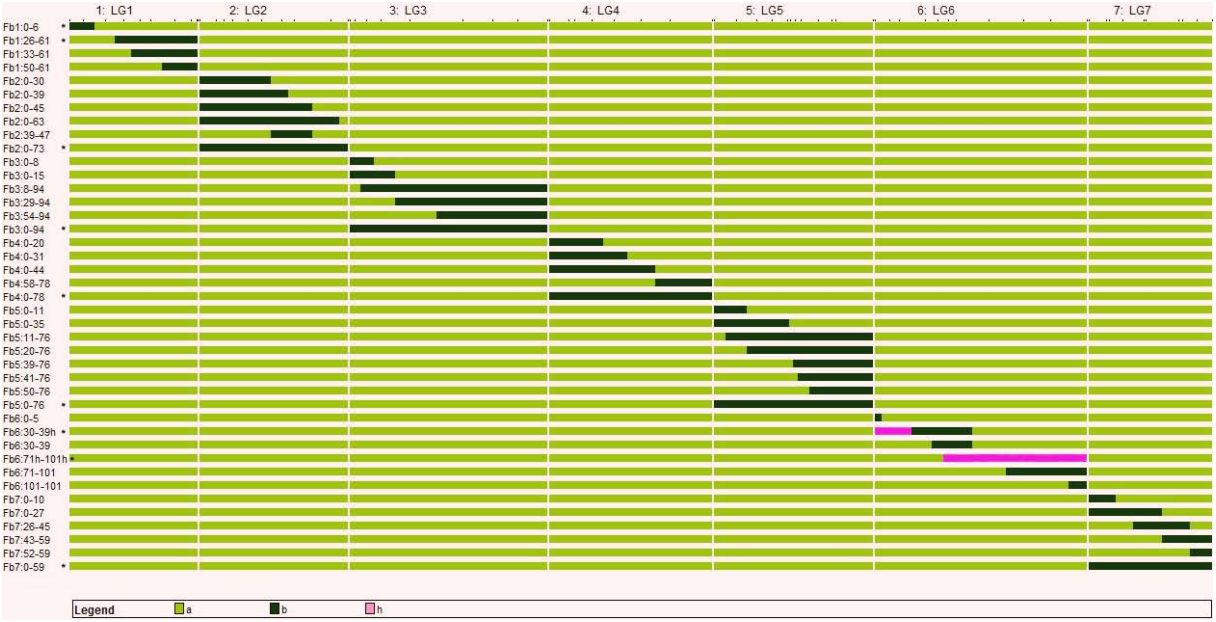


Table CI. 1. NIL collection characteristics

	n	cM	%
Nb. NILs	39		
Nb. Heterozygous NILs	2		
Nb. NILs covering whole LG	5		
Total genome covered in intogressions		522.0	96.4
Genome covered in homozygosity		479.3	88.5
Genome covered in heterozygosity		42.7	7.9
Average NILs per LG	6		
Average introgression size *		32.9	6.1
Smallest introgression *		3.2	0.6
Largest introgression *		89.3	16.5
Nb. of BINs	37		
Average BIN size		14.2	2.6
Smallest BIN		3.2	0.6
Largest BIN		65.7	12.1

*Not considering whole LG introgressions

Analysis of morphological and phenotypical traits with the NIL collection

In Table CI. 2 we present the results for all the harvests analyzed. The initial and final QTL positions indicate always the smallest region that could be assigned to a QTL, taking into account all the harvests where it was found to be significant. All QTL found were significant in overlapping NILs, the NIL carrying the shorter introgression (expressed in cM) are indicated. The effects of the QTLs are expressed as the percent mean variation between the lines carrying the QTL and the recurrent parent, RV. A single bin was determined for all qualitative morphological and phenotypical characters (Table CI. 2A).

The presence of runners has been extensively described (Sargent *et al.* 2004a), and in our population could be clearly assigned to the region between cM 39 and 47 in LG2. *F. bucharica* donor parent is causal for runner trait presence. In this region we also detected one or more genes involved with long floral stems, so that plants carrying this introgression from *F. bucharica* produced flowers that emerged from the plant canopy. This effect was recessive, not visible in parental and F_1 lines. The opposite was seen in lines carrying *F. bucharica* alleles between cM 29 and 39 in LG5: these lines produced short floral stems so that the flowers and fruits remained within the leaf canopy of the plant, as with the *F. bucharica* donor parent.

F. bucharica has a very characteristic flower that usually has more than five overlapping petals, typically eight to ten (Figure CI. 3). This particular phenotype can be seen in plants with introgressions in LG3 between cM 54 and 94. We also observed flowers with a pink blush in the inner part of their petals, a trait not shown by any of the parents but a little visible in hybrid F_1 plants. This dominant trait was located in LG1 between cM 50 and 61.

We also identified one or more genes in the region between cM 15 and 29 of LG3 causing extreme dwarfism in plants (Figure CI. 4) and leading to death before flowering. Plants carrying homozygous introgressions in this region did not grow well under our greenhouse conditions, so we have never been able to reproduce them. This is a recessive trait, not present in parents.

Flowering time, seed germination, fruit weight and fruit shape were measured as quantitative traits. Their distribution among the NIL collection is summarized in Table CI. 3A. The flowering time was studied in two consecutive years. We were able to identify two map regions that determined flowering time (Table CI. 2B), one in LG6, between cM 0 to 11 (86 days after germination on average) as early flowering (EF), and the other in LG5, cM 11 to 20 (144 days after germination on average) as late flowering (LF). The seasonal flowering behaviour typical of *F. bucharica* that blossom one time a year, opposite to the continuous flowering characteristic of RV, also segregated in the NIL population. We mapped this trait in LG6 at cM 30 to 38, in agreement with that published by Koskela *et al.* (2012) in plants with few flowering success. Another specific region for everbearing were mapped in LG5, cM 35 to 39, these plants blossom from March to June, as donor parent, and not everbearing in autumn as *F. vesca* RV.In addition, we considered the germination time for every line in 2013 (Table CI. 3A). In a first germination trial, germination was considerably delayed (around 40 days, compared with 14 days for RV) in a particular line carrying an introgression in LG5 (line Fb**5**:39-76). We repeated the experiment twice, using only a few lines with introgressions in LG5 (Fb**5**:20-76, Fb**5**:39-76, Fb**5**:41-76, Fb**5**:50-76 and Fb**5**:0-76) and RV as the control. The same results were

obtained in both additional trials (data not shown), and the delayed germination time was mapped to a short bin at LG5 cM 39 to 41.

Figure CI. 3. Flower phenotypes. From left to right, flower morphology of, *F. bucharica*, the hybrid and NILs Fb**1**:50-61 and Fb**3**:54-94.

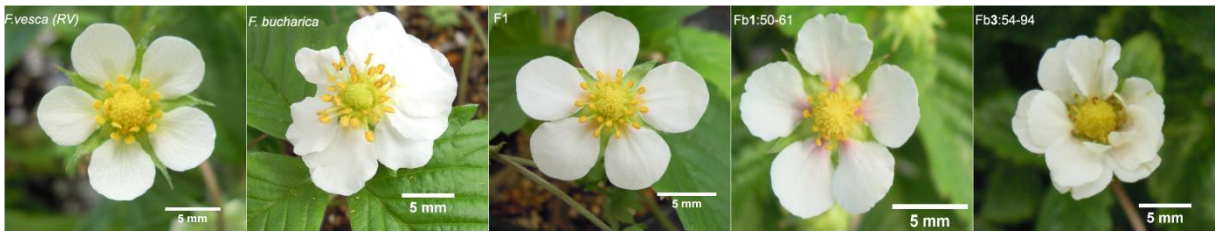
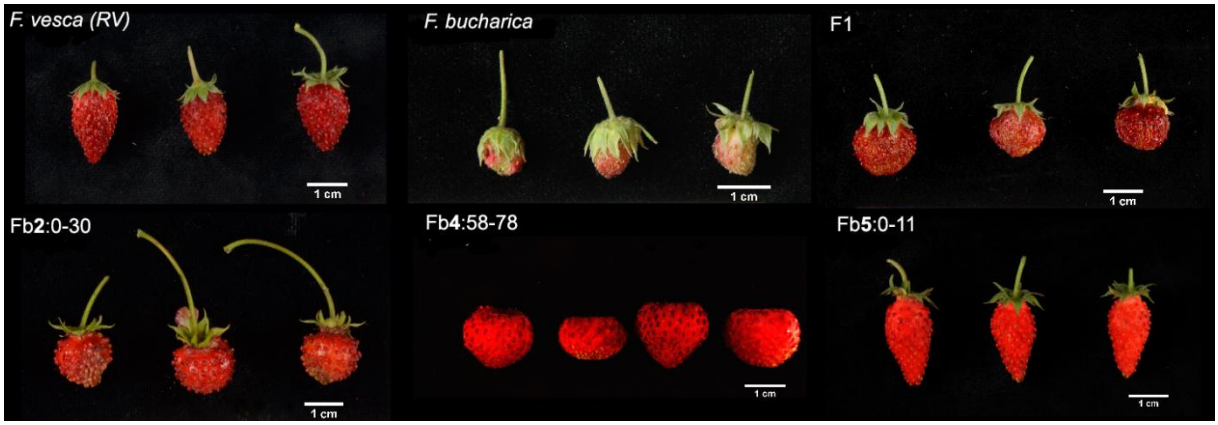


Figure CI. 4. Size phenotypes. Images show the characteristic plant size 12 weeks after germination for both parents, RV and *F. bucharica*, and NIL Fb**3**:8-94.

Figure CI. 5. Fruit phenotypes. Left to right and top to bottom, characteristic fruit morphology of the parents, Reine des Vallées and *F. bucharica*, their $F_1$and NILs Fb**2:**0-30, Fb**4:**58-78 and Fb**5:**0-11.



Fruit size was phenotyped as a quantitative trait by measuring the length and width of fruits, and their ratio (the fruit index) that corresponds to shape of fruit. We observed that the population means were lower than the mean values for RV except for the first harvest (Table CI. 3A). The mean fruit index values for both RV and the population were >1, indicating that the fruits were generally elongated (except for RV in the first harvest). However, the population fruit index ranges showed

that the phenotype among the NILs varied from rounded (0.5) to very elongated fruits (2.5) (Table CI. 3A, Figure CI. 5). We mapped two QTLs for rounded fruits, as donor parent, stable through all five harvests and present in overlapping NILs (Table CI. 2). The first one was in LG2, from cM 0 to 30 and the second one in LG4, cM 58 to 78, accounting for 24% and 35% of the total variation, respectively (Table CI. 2B). An additional QTL for elongated fruits mapped to LG5 cM 0 to 11, accounting for 25% of the variation and stable in four out of five seasons (Table CI. 2B). Also two QTLs explaining more than 70% of fruit weight reduction were detected, one in LG4, cM 58 to 78, on the same region of fruit shape QTL, and another was observed in LG1, cM 50 to 61 were fruits are smaller but elongated.

Table CI. 2. Position of Mendelian traits (A) and QTLs for quantitative agronomic traits (B). Trait, name of QTL, effect percentage means average of trait value on NIL compared to RV trait value, bin position and markers flanking the bin, stability as number of harvests observed over total measured and the shortest NIL showing trait effect.

A

Trait	Name	Effect	% Effect	LG	Initial position (cM)	Final position (cM)	Stability	shortest NIL
Runners	*R* [a]	presence of runners	-	2	39 (Fvi11)	47 (EMFn134)	all	Fb**2**:39-47
Floral stem length	*LFS*	longer floral stems	-	2	39 (Fvi11)	47 (EMFn134)	all	Fb**2**:39-47
	SFS	shorter floral stems	-	5	29 (FvH4093)	39 (CFVCT024)	all	Fb**5**:20-76
Pink petals	*PSP*	pink spot on petal base	-	1	50 (UFF02F02)	61 (CFV164)	all	Fb**1**:50-61
Petals number	*EPN*	>5 petals/flower	-	3	54 (CFVCT022)	94 (CFVCT012)	all	Fb**3**:54-94
Seasonal flowering	*SDF*	short day flowering	-	5	35 (UDF006)	39 (CFVCT024)	all	Fb**5**:20-76
Flowering	*LLF*	flowering >1 year later	-	6	30 (EMFn117)	38 (EMFn017)	all	Fb**6**:30-39
Plant size	*DWARF*	dwarf plants	-	3	15 (VT398)	29 (UDF017)	all	Fb**3**:8-94
Germination	*DOG*	delay of germination	-	5	39 (CFVCT024)	41 (UDF009)	3	Fb**5**:39-76

B

Trait	QTL name	Effect	% Effect	LG	Initial position (cM)	Final position (cM)	Stability	shortest NIL
Fruit shape	FRS_2	rounded fruit	-24	2	0 (EMFvi099)	30 (BFACT002)	5/5	Fb**2**:0-30
	FRS_4	rounded fruit	-35	4	58 (CEL1)	78 (ChFaM23)	5/5	Fb**4**:58-78
	FES_5	elongated fruit	25	5	0 (CFV3072)	11 (CFV3132)	4/5	Fb**5**:0-11
Weight	FLW_1	smaller fruit	-70	1	50 (UFF02F02)	61 (CFV164)	3/3	Fb**1**:50-61
	FLW_4	smaller fruit	-79	4	58 (CEL1)	78 (ChFaM23)	2/3	Fb**4**:58-78
Flowering	EF[b]	early flowering	-30	6	0 (ARSFL007)	11 (EMFn228)	2/2	Fb**6**:0-5
	LF	late flowering	36	5	11 (CFV3132)	20 (CEL2)	2/2	Fb**5**:11-76

[a] Locus R (mapped at LG2: 45,2cM); [b] Compared to population mean

Table CI. 3. Agronomic trait distribution (A) and nutritional trait distribution (B).

A

			F. vesca "Reine des Vallées"		NIL population				
Trait	**year**	**location**	**mean**	**sd**	**mean**	**sd**	**range**	**skewness**	**kurtosis**
Fruit length (mm)	2011-may	Cabrils	11.96	1.56	13.85	2.48	7.3-22.0	0.22	3.84
	2011-may	TM	15.72	1.56	15.41	2.66	6.9-30.5	0.01	3.04
	2011-july	Cabrils	23.17	0.17	17.48	3.62	10.9-27.4	0.21	1.85
	2012-may	Cabrils	19.35	2.73	17.63	3.27	7.31-25.09	-0.22	2.20
	2012-july	Cabrils	18.64	2.10	15.72	2.10	8.09-25.07	-0.18	2.61
Fruit width (mm)	2011-may	Cabrils	14.08	1.63	12.06	1.86	5.5-17.8	-2.04	7.87
	2011-may	TM	13.52	1.56	12.42	2.29	4.6-21.4	-0.60	3.01
	2011-july	Cabrils	14.12	1.83	10.02	1.63	6.1-14.9	-0.03	2.60
	2012-may	Cabrils	12.71	1.52	12.58	1.34	7.03-18.28	-0.40	2.99
	2012-july	Cabrils	11.32	1.20	10.89	1.25	5.59-17.19	-0.90	3.15
Index (length/width)	2011-may	Cabrils	0.84	0.03	1.17	0.22	0.58-1.93	0.38	2.77
	2011-may	TM	1.18	0.20	1.28	0.25	0.57-2.42	0.51	3.48
	2011-july	Cabrils	1.65	0.18	1.77	0.30	0.94-2.6	-0.64	2.17
	2012-may	Cabrils	1.54	0.28	1.42	0.27	0.55-2.45	-0.22	2.39
	2012-july	Cabrils	1.66	0.28	1.46	0.20	0.59-2.54	-0.15	2.17
Weight (g)	2011-july	Cabrils	1.66	0.36	0.72	0.28	0.26-1.58	0.53	2.17
	2011-july	TM	2.36	0.57	0.68	0.22	0.26-1.47	0.58	2.75
	2012-july	Cabrils	0.93	0.18	0.66	0.19	0.19-1.09	-0.35	3.64
Flowering (days)	2012	Cabrils	86,00	2.87	125.20	37.58	78-264	2.22	8.54
	2013	TM	118,00	14.62	117.46	28,00	74-189	1.03	3.90
Germination (days)	2013	in-vitro	14.88	4.37	17.05	6.08	9.95-42.00	0.02	10.08

B

			F. vesca " Reine des Vallées"		Hybrid (F_1)		NILs population				
Trait	**year**	**location**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**	**range**	**skewness**	**kurtosis**
Fructose content (mg/g fw)	2011	pool [a]	27.78	12.48			20.82	5.04	8.33-33.15	0.08	3.00
	2012	Cabrils	25.57	0.79			24.80	5.57	13.05-39.94	0.24	2.90
	2013	TM	26.56	5.46	17.74	2.98	24.18	7.78	8.96-51.42	1.18	4.93
Glucose content (mg/g fw)	2011	pool [a]	24.64	11.75			17.46	5.19	1.96-30.85	-0.30	3.46
	2012	Cabrils	26.82	3.28			23.02	6.18	11.59-41.27	0.51	3.00
	2013	TM	18.15	4.22	9.90	2.10	15.71	5.64	0.875-33.20	0.63	3.93
Sucrose (mg/g fw)	2011	pool [a]	16.52	4.84			12.32	5.48	1.67-26.76	0.52	2.91
	2012	Cabrils	12.04	2.21			10.40	2.67	2.92-17.16	-0.13	2.81
	2013	TM	18.42	2.67	21.55	6.10	14.18	6.47	0-31.79	0.08	2.77
Total sugar content (mg/g fw)	2011	pool [a]	68.94	26.72			50.62	12.32	11.95-77.23	-0.52	3.65
	2012	Cabrils	64.44	5.41			58.23	11.99	29.66-92.28	0.13	3.21
	2013	TM	63.14	9.79	49.20	8.14	54.09	16.63	20-107.26	0.54	3.38
Total polyphenol content (Gae/g fw)	2011	pool [a]	4.12	1.02			4.49	1.14	2.26-8.35	0.77	3.71
	2012	Cabrils	2.09	0.16			2.23	0.73	0.95-3.96	0.51	2.45
	2013	TM	4.27	0.63	2.63	0.30	4.05	0.82	2.28-6.72	0.27	3.15

[a] Cabrils and TM

QTLs of nutritional content in the NIL collection

Results for genetic analysis of fruit nutritional characters are summarized in Table CI. 3B. Measured parameters showed a normal distribution over most of the harvests.

The sugar composition of the fruits was mainly based on fructose and glucose, with fructose more abundant in almost all the analyses. Fructose, glucose and total sugar content showed a very high correlation within each harvest (always above 0.9). These correlations existed also between harvests and between locations with the exception of Cabrils 2012 and TM 2013 harvests . Sucrose, a minor sugar in fruit, where its accumulation is very variable, had little significant correlation with other sugars for each harvest, and no significant correlation over the different years. All analyzed harvests showed that RV accumulates more sugars than the NIL collection mean. However the ranges of the population were very extreme and some lines surpassed the mean sugar values of RV (Table CI. 4B). Values for the F_1individual were collected in 2013. These fruits had lower levels of fructose and glucose than RV.

As a result of the analyses, a QTL resulting in a decrease of fructose, glucose and total sugar content, was mapped in LG2, cM 45 to 63 (Table CI. 4), accounting for 35%, 45% and 35% of the total character variation for these characters, respectively. This QTL was detected in all three years. For fructose, we found one additional QTL decreasing also its content (in LG3 cM 54 to 94), and for glucose we found also one additional QTL decreasing content (in LG5, cM 41 to 50). Sugar content was not measured in *F. bucharica* but results in F_1 fruits point to donor parent as causal of decrease.

There was a significant correlation of the total polyphenolic content among all the harvests, and the population and RV values were very similar within every harvest (Supplementary Table CI. 4). There was, however, no correlation between the sugar and the polyphenol content: the 2012 harvest had the highest concentration of fruit sugars and the lowest total polyphenol content. Two regions were associated with the variation in total polyphenolic content. One region was in LG2 cM 0 to 30, where the alleles from *F. bucharica* increased it by 26%, and another in LG5 cM 50 to 76, where the introgressed region caused a 35% decrease in total polyphenolic content (Table CI. 5). Both were detected in two of the three years.

Table CI. 4. QTL for nutritional traits. Trait, name of QTL, effect as mean percentage compared to RV, stability as number of harvests observed, bin position (LG, markers flanking bin and cM) and shortest NIL showing trait effect.

Nutritional trait	QTL	% Effect*	QTL stability	LG	Initial position (cM	Final position (cM)	shortest NIL
Fructose content	Fru_2	-35	3/3	2	45 (EMFn134)	63 (EMFxa379796)	Fb**2**:0-63
	Fru_3	-40	2/3	3	54 (CFVCT022)	94 (CFVCT012)	Fb**3**:54-94
Glucose content	Glu_2	-45	3/3	2	45 (EMFn134)	63 (EMFxa379796)	Fb**2**:0-63
	Glu_5	-28	2/3	5	41 (UDF009)	50 (CFV3821)	Fb**5**:41-76
Total sugar content	TotSug_2	-35	3/3	2	45 (EMFn134)	63 (EMFxa379796)	Fb**2**:0-63
Total polyphenol cc	Tphe_2	26	2/3	2	0 (EMFvi099)	30 (BFACT002)	Fb**2**:0-30
	Tphe_5	-35	2/3	5	50 (CFV3821)	76 (VT010)	Fb**5**:50-76

* % effect with respect to parental RV

Discussion

Fragaria vesca possesses several features that make it attractive as a model for the octoploid cultivated strawberry, the most evident being that the diploid nature of its genome allows the genetic complexity of the cultivated strawberry to be circumvented. Furthermore, wild strawberries are potential donors of genes of interest for strawberry fruit quality improvement, including disease resistances, and quality characters such as those involved in volatile compound expression and concentration. In order to enhance our capability to analyze complex traits in strawberry, we developed a near isogenic line (NIL) collection in an *F. vesca* background, using an Asiatic species *F. bucharica,* as introgression donor and *F. vesca* cv. 'Reine des Vallées', a French non-runnering variety commonly cultivated in Spain, as the recurrent parent. Considerable research has been published on *F. vesca* and *F. bucharica* genomes and maps have been updated (Sargent *et al*. 2004a; Sargent *et al.* 2006; Vilanova *et al.* 2008; Ruiz-Rojas *et al.* 2010). Here we provide an additional set of SSR markers, for a region previously covered by RFLPs, which improve the current reference map.

The 39 NILs of the collection covered 96.4 % of the *F. bucharica* genome. Each line had a single introgressed chromosome fragment from the *F. bucharica* parent genome, covering 6 % of its genome on average. The resolution capacity, 14.2 cM per bin on average, with regions defined by only 3.2 cM, makes this collection a powerful tool for QTL analysis.

This collection of introgression lines was obtained in only two generations of backcrossing and one to four of selfing. A larger number of backcross generations have been used to extract other NIL collections. Eduardo *et al.* (2005) used a group of eight double haploid lines (LDH) to produce a melon NIL collection in six BC generations, Jeuken *et al.* (2004) began from an $F_1$and needed to get to the fifth generation backcross to obtain lettuce NILs, and Eshed & Zamir (1995) needed up to 12 BC generations, using a hybrid between tomato and *Solanum pennellii*. Our strategy was to select from a large BC_1 progeny, which has the advantage of minimizing the risk of unnoticed small introgressions that may be problematic for the use of the NIL collection, and reduces the number of generations needed for its completion. In our case this was further facilitated by the low number of chromosomes of Fragaria (*x=7*) compared to melon and tomato (both with *x=12*), or lettuce (*x=9*). In Arabidopsis (*x=5*), the NIL collection was obtained after backcrossing a RIL collection, and several selfing stages (Keurentjes *et al.* 2007). We made an initial selection with only 14 molecular markers located on the ends of chromosomes, discarding over 70% of the BC_1 individuals to select a final working population of nine plants, covering the entire genome with a maximum of four introgressions per plant. Selection at the BC_1 level also facilitated extraction of five NILs containing a whole chromosome of *F. bucharica*. Backcrossing heterozygous plants for these introgressions would produce a series of recombinant individuals with diverse coverage of these chromosomes, useful for fine mapping of genes/QTL located on them. Few major genes have been mapped in strawberry: two of them, gen controlling runner formation mapped in LG2 and gen controlling continuous flowering mapped in LG6 (Sargent *et al.* 2004) segregated in our collection and were assigned to the expected position.

A study with different species of *Rosa* flowers including *Rosa x hibrida, R. chinensis,* R. *gallica* and *R. rugosa* determined that the expression pattern of the rose AGAMOUS ortholog gene (*RhAG*) is

responsible for the number of petals. Higher expression of *RhAG* has been associated with low petal number cultivars while a more restricted *RhAG* expression towards the flower centre has been associated with high petal number cultivars (Dubois *et al.* 2010). The flower developmental stages and AG expression between *F. vesca* and rose have been described as similar. The *F. vesca* AG homolog (*FvAG*) was identified as the Hybrid Gene Model #24852 (Hollender *et al.* 2012). This gene is located in *F. vesca* LG3: 3939971-3944769, within the boundaries of the exotic introgression in Fb**3**:54-94. Further studies are needed to validate if a different expression pattern of *FvAG F. bucharica* in relation to *F. vesca* could be responsible for the *EPN* phenotype.

F. vesca is a perennial plant and there are accessions with opposite photoperiodic responses. Short-day (SD) *F. vesca* accessions require short days and low temperatures to initiate flowering, in contrast, everbearing (or semperflorens) *F. vesca* accessions initiate flowering on long days and high temperatures. Control of these contrasting phenotypes was recently attributed to a single dominant gene Seasonal Flowering Locus (SFL) that was mapped and characterized as the *F. vesca* TFL1 homolog (*FvTFL1*), a photoperiodically-controlled flowering repressor (Iwata *et al.* 2011; Koskela *et al.* 2012). There is functional evidence that functional *FvTFL1* alleles cause the SDF flowering habit and that a non functional allele (2bp deletion generating a truncated protein) is associated with the everbearing flowering habit in *F. vesca*. The *F. vesca* RV parental is an everbearing species that carries the truncated allele of *FvTFL1* while the F. bucharica parental has a SD flowering habit and carries the functional allele so we observe segregation of this character in our NIL collection. *FvTFL1* is located in LG6:12836105-12837237 bp between the markers EMFn117 (30.1cM) and EMFn017 (38.8cM). We have observed that all NILs with exotic introgressions between these two markers have a SD habit that can be assumed by the functional *FvTFL1* alleles but in addition, these NILs did not flower until the 2nd or the 3rd year after germination and their blooming was reduced to 2 or 3 inflorescences. This could be caused by other independent loci, included in the same introgresion, affecting the juvenile period of the plants.

In many plant species, domestication has resulted in a modified architecture of the reproductive organs, mainly fruit and seed, and, generally, variability has been amplified by human selection. This is the case of fruit morphology, where shape and size traits have often been subjected to intense selection. The genetic control of these characters has been studied in melon (Fernández-Silva *et al.* 2010) and in tomato (Frary *et al.* 2000; Liu *et al.* 2002), where two QTLs with major effects on fruit weight (fw2.2) and shape (ovate) have been positionally cloned. Studies on the genetic basis of these traits in the Rosaceae family are few. (Zhang *et al.* 2006) have shown that the evolution of fruit size in *Pyrus pirifolia* during domestication is due to changes in the ability of cells to divide after pollination, and (Zhang *et al.* 2010) found several QTL for size and mesocarp cell number, in cherry, that suggested that both characters were positively correlated. In this work we located three QTL involved in strawberry fruit shape, two (LG2 and LG4) showing rounded fruits, and one (LG5) showing elongated fruits. The positions of these QTL in the strawberry map are not syntenic with the *Prunus* chromosome 2 (Illa *et al.* 2011) where the QTL for fruit size in cherry occurs, suggesting that they correspond to other genes involved in determining this character. Fine mapping of these QTL in future studies would be facilitated by the NIL collection presented here, which could be an incentive for the study of quantitative traits of cultivated strawberry fruit morphology and its comparison with other species of the Rosaceae family.

Given that the recurrent parent Reine des Vallées is cultivated commercially, any NIL containing an interesting new character could be rapidly evaluated as a possible new cultivar. For example, NIL Fb**2**:39-47 has the "*r*" locus that allows the plant to develop stolons. If other agronomic and organoleptic characteristics of the range remain unchanged, this line may represent a direct improvement of RV, which can be propagated vegetatively. The same could be true for the QTL of increased sugar or polyphenolic content found in this paper. For lines with interesting characters that have other undesirable traits associated, subNILs, with shorter introgressed fragments, could be obtained that would eventually eliminate this linkage drag.

Consumers now call for healthy products with enhanced nutritional and organoleptic characteristics. Strawberry is an important source of natural products such as vitamins, flavonoids and other (poly)phenolic compounds with antioxidant capacity. Sugars and organic acids of strawberries have been investigated as indicators of fruit development and ripening and as components of fruit flavour (Perez *et al.* 1997). The anticancer effects of specific phytochemical constituents of strawberries, as well as whole strawberry extracts, have been demonstrated (Seeram *et al.* 2006). The variation of nutritional parameters in different genotypes indicates a genetic control of these parameters, suggesting that the strawberry can potentially be improved for nutrition-enhanced traits, and some examples of genome regions carrying genes associated with sugar content phenotypic variation have been found in this paper: a QTL in LG2 determined a decrease (around 35%) of total sugars, fructose and glucose, a QTL for a decrease in fructose or glucose content was detected in LG3 and LG5 respectively. The polyphenol content in NIL Fb**2**:0-30 was increased in 26%, and another region (LG5) decreased total polyphenol content.

NILs have allowed cloning of genes involved in fruit sugar content (Fridman *et al.* 2000), and mapping of QTL for aroma (Tadmor *et al.* 2002) and colour (Liu *et al.* 2003) of tomato fruit. Developing melon NILs has been crucial in clarifying the genetic basis of some climacteric accessions (Vegas *et al.* 2013). Analysis of the metabolome and the dissection of its inheritance in tomato (Schauer *et al.* 2008) were thanks to a collection of introgression lines of the cultivated tomato with other wild species. Therefore, it is likely that, through the collection of introgression lines developed here, complex traits such as nutritional quality, flavour or shelf-life of the fruit can also be studied. These characters are crucial targets for commercial plant breeding in rosaceous fruit such as strawberry, apple, peach, pear, apricot and raspberry, as they are increasingly in demand by consumers. Being these fruit crops a significant part of the human diet, new cultivars based on the generated knowledge can contribute to the improvement of the health and well-being of our society.

Acknowledgements: The authors thank J. Ribes and A. Ortigosa for technical assistance. This work was funded in part by grants RTA2007-00063, AGL2010-21414 and RTA-2013-00010 from the Spanish Ministry of Science. MU was supported by a fellowship FPI from the Spanish Ministry of Education and JB was supported by a fellowship FI from the Generalitat de Catalunya.

Author contribution statement: JB obtained the first generation of NILs, phenotyped plants for the first season and was involved in edition of the manuscript. MU made crosses to increase the number of NILs, phenotyped plants for the three seasons and participated in edition of the manuscript. PA designed the crossing and selection scheme and contributed to writing the manuscript. AM lead the project, contributed to all steps of selection and phenotyping and to writing the manuscript.

Annex chap. I

Fine genotyping of the *F. vesca* NIL collection using IStraw90® Axiom® SNP array

Introduction

The *Fragaria vesca* NIL collection was developed and genotyped using mainly microsatellites (SSRs) markers evenly spread across the genome. These molecular markers are cheap, relatively fast, robust and transferable (Vilanova *et al.* 2008) what was useful for the selection and initial genetic characterization of the NIL collection. However their number and resolution was limited and therefore substantial genetic regions remained uncovered. Characterize NIL collection precisely and detect non-desired introgressions required a higher marker density throughout whole *F. vesca* genome.

The availability of a high quality reference genome sequence for *F. vesca* (Schulaev *et al.* 2011) facilitated the development of new molecular markers including single nucleotide polymorphism (SNP) markers that are very abundant and evenly distributed. Recently a SNP array for strawberry genotyping (IStraw90®) was publicly available (Bassil & Davis *et al.* 2015). The IStraw90® is a 90 K SNP array based on the Affymetrix Axiom® platform that was developed using sequencing data from octoploid (*F. x ananassa, F. virginiana* and *F. chiloensis*) and diploid (*F. vesca, F. mandshurica* and *F. iinumae*) accessions. The SNPs coordinates agree with the *F. vesca* reference genome v1.1 (Schulaev *et al.* 2011).

In order to refine the physical position of the exotic introgressions and to detect any possible non-desired introgressed regions, we genotyped the NIL collection with the IStraw90® array.

Materials and methods

DNA extraction and hybridization

DNA was extracted with DNeasy® plant mini kit (Qiagen, Hilden, Germany) from the parental lines of the NIL collection and their hybrid per duplicate (*F. vesca* var. Reine des Vallées, *F. bucharica* FDP601 and F_1), from 47 NILs, and from 7 lines from the F_2 mapping population *F. vesca* var. 815 x *F. bucharica* FDP601 that conform the BIN set (Sargent *et al.* 2008). 50µl of DNA from each sample, at concentration >= 20ng/µl, was sent for array hybridization. A total of 60 diploid samples (47 NILs, 7 BIN set and RV, FB and F_1 per duplicate) were hybridized with the IStraw90® Axiom® SNP array at CEGEN (Santiago de Compostela, Spain).

Genotyping analysis

Genotyping analysis was performed using Genotyping console™ and SNPpolisher © (both from Affymetrix, CA, USA) following manufacturers recommendations and an inbred penalty of 3 for the NIL collection. Double recombinants for one single SNP were considered as genotyping errors and the flanking allele was assumed.

Results

IStraw90® Axiom array hybridization analysis and SNP clustering

A set of 60 diploid *F. vesca* lines were hybridized with the IStraw90® including the NIL collection, its hybrid (F_1) and parental lines (RV and FB) used as controls, and the BIN set (Sargent *et al.* 2008) added to increase allelic variability. All samples passed quality control threshold.

Genotyping results revealed that the vast majority of SNPs (79922 from the initial 95063 SNPs, 84.07%) were homozygous in all analyzed lines and therefore not suitable for genotyping the NIL collection. From the remaining 15141 SNPs, 3327 (3.50%) did not surpass quality criteria (<95% call rate) and were discarded. Additionally 5747 (6.04%) were classified as off target variants (OTV) and probably represent allelic variants not considered by the SNP probes and 3079 (3.24%) did not present both homozygous alleles and were not considered for the analysis. Finally, the remaining 2988 (3.14%) were polymorphic SNPs that segregate among the set of analyzed samples and were used for genotyping. Additional filtering steps were performed including removal of non consistent duplicates and non-segregating among the NIL collection SNPs this resulted in final useful set of 1510 SNPs. After genotyping, the BIN set was not further considered and the analysis continued only with the NIL collection, the parental lines and the hybrid (RV, FB, F_1)

Physical boundaries of introgressions

Genotyping results revealed several discrepancies between the expected introgressions according to the genetic map (Urrutia *et al.* 2015) and the obtained genotypes (Figure AI. 1). This suggested that there could be some miss-assemblies in the *F. vesca* reference genome v1.1. For instance, if we look at introgressions in LG1 we can observe that there is a double recombinant event in the same position (LG1: 966.755- 1.442.404) in six NILs creating a gap in lines Fb**1**:0-6 and Fb**1**:0-6-26 and an un-desired introgression in Fb**1**:26-61, Fb**1**:33-61. Fb**1**:50-61 and Fb**1**:58-61 (Figure AI. 1). It is very unlikely that all six lines present a double recombinant event at the same position and the SNPs genotyping pattern matches segregation presented by SNPs mapped in region LG1: 4.386.861-20.747.404. These make us think that in fact the region LG1: 966.755- 1.442.404 is miss-assembled in the genome and is probably located somewhere between LG1: 4.386.861-20.747.404 (Table AI. 1). Following the same reasoning, three other regions assembled out of LG1 (LG3: 19.610.303-19.682.659, LG6: 34.750.958-35.927.392 and LG7: 610.205-792.402) that presented the same segregation pattern of SNPs in LG1: 4.386.861-20.747.404 were detected. We suggest that these regions are miss-assembled in the *Fragaria vesca* v1.1 reference genome and that their correct position is somewhere between LG1: 4.386.861-20.747.404 (Table AI. 1). According to this criteria, a total of 26 putatively miss-assembled and 2 miss-oriented regions were found and are specified in Table AI. 1.

Figure AI. 1 F. vesca NIL collection SNP genotyping (v1.1). NIL collection genotyping using IStraw90® according to reference genome version 1.1

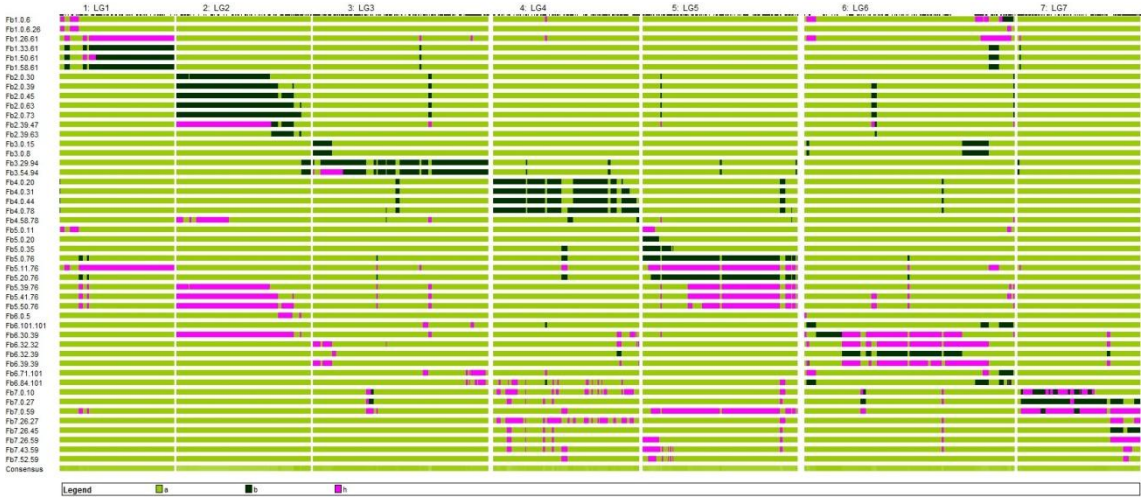
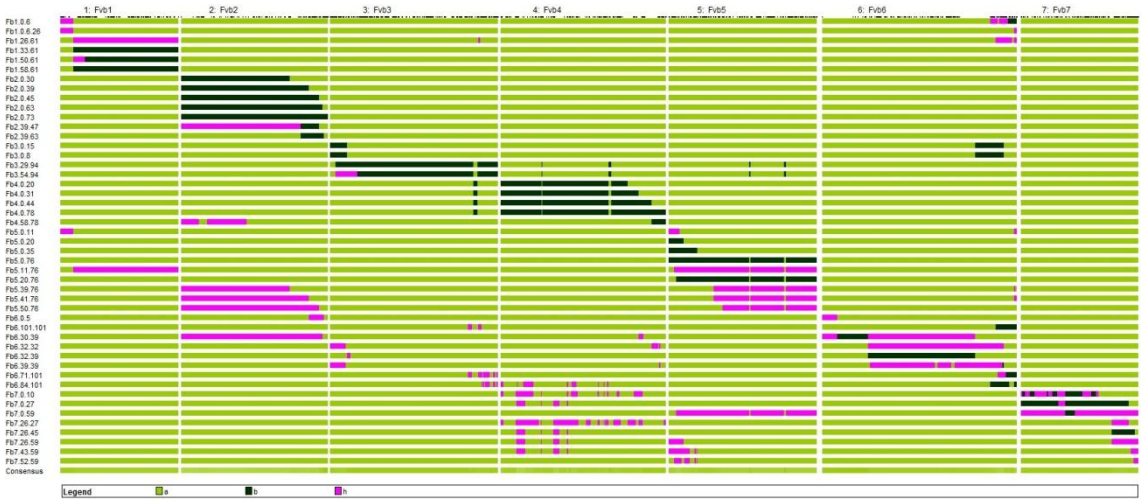


Figure AI. 2 F. vesca NIL collection SNP genotyping (v2.0). NIL collection genotyping using IStraw90® according to reference genome version 2.0



Simultaneously, we mapped all SNPs probes in the IStraw90® array to the new genome assembly v2.0 (Tennessen *et al.* 2014) and found that 25 of the 28 positional discrepancies were resolved (Figure AI. 2, Table AI. 1), confirming that they were true miss-assemblies. Not confirmed exceptions were three regions assembled to LG4 (LG4: 6.629.502-6.708.323 and LG4: 21.501.885-21.673.366) and LG5 (LG5: 28.233.300-28.278.585) in version 1.1, that we propose to allocate somewhere in LG3: 5.668.456-27.749.642 and that remain unchanged in v2.0. In addition, one of the rearrangements that we suggest (relocate LG3: 11.669.089-11.801.654 somewhere between LG5: 8.530.103-27.902.872) is reassembled in v2.0, but the rearrangement has included a bigger region than we propose and has generated a new discrepancy (Figure AI. 2). Finally genome version 2.0 reorders region LG4: 8071996-8544775, that our data suggest to be correctly placed, to Fvb3: 28352016-28824795 generating a new discrepancy (Figure AI. 2).

Genotyping results allowed a more precise drawing of the introgression's boundaries (Table AI. 2). Most NIL were fixed in homozygosity and did not harbor (or harbor only short) undesired introgressions from *F. bucharica* in the genetic background of *F. vesca*. However, exotic introgressions of 16 of the 47 NIL analyzed were genotyped as heterozygous and 10 of them (Fb**5**:0-11, Fb**5**:11-76, Fb**5**:39-76, Fb**5**:41-76, Fb**5**:50-76, Fb**6**:0-5, Fb**6**:30-39, Fb**6**:32-32, Fb**6**:39-39 and Fb**7**:0-59) presented large undesired introgressions in heterozygosity. Surprisingly, these unwanted introgressions perfectly match exotic introgressions in other NIL (Figure AI. 2, Table AI. 2). For instance, lines Fb**5**:39-76, Fb**5**:41-76 and Fb**5**:50-76 harbor unwanted introgressions in LG2 with the exact coordinates of the introgressions in NIL Fb2:0-30, Fb2:0-39 and Fb**2**:0-45 respectively (Table AI. 2, Figure AI. 2). The fact that these unwanted introgressions mirror other's NIL introgressions make us suspect that these are artifacts or sample contaminations. In addition, other complementary data presented below, point in the same direction. All NILs, and their progenies (from auto-pollination) have been exhaustively genotyped with molecular markers (SSRs) and neither heterozygous nor unwanted introgressions were detected (data not shown). In addition, clear phenotypic differences are maintained between the NIL. For instance a dominant locus in LG2:39-45 cM controls runners production, what is a clear morphological marker that can be phenotyped in heterozygosity. None of the NILs detected as harboring an unwanted introgression in LG2 covering this region (Fb**5**:50-76, Fb**6**:0-5, Fb**6**:30-39) or their offspring has ever produced runners, what probably means that they do not truly harbor *F. bucharica* alleles in this region.

Many technical reasons could be behind this incoherent result. First, IStraw90K® was designed for octoploid strawberry genotyping. Allelic variability between diploid strawberry accessions was not a priority in probes selection and was barely tested (Bassil & Davis *et al.* 2015). In addition *F. bucharica* accessions were not considered in probes designing or hybridization tests. Furthermore the number of samples that we hybridized (60) was below the minimum recommended for clustering algorithm (96). All these together led to a high proportion of homozygous SNPs (>84%) and SNPs discarded for quality reasons when evaluating the NIL collection. Even among SNPs classified as good-quality polymorphic SNPs (PolyHihgResolution) by the analytical software Genotyping console™ (Affymetrix, CA, USA) we found poorly resolved genotyping clusters that were far from the expected diploid pattern (Figure AI. 3). The fact that our NIL collection is not a real segregating population and the expected clusters are highly disproportional may also difficult the genotyping task. Finally a contamination of samples while manipulating DNA cannot be discarded.

Table AI. 1 NIL introgression position discrepancies between microsatellite and IStraw90® genotyping

F. vesca reference genome v1.1			NIL collection (Istraw90)				*F. vesca* reference genome v2.0					
position			Proposed position interval				position					
Chr	start[a]	end[a]	Chr	putative start[a]	putative end[a]	orientation[c]	scf	Chr	start[b]	end[b]	orientation[c]	confirmation
LG1	966.755	1.442.404	LG1	4.386.861	20.747.404	0	scf0513151c	Fvb1	19.366.257	20.068.377	1	confirmed by Tennessen et al 2014
LG3	19.610.303	19.682.659	LG1	4.386.861	20.747.404	0	scf0513125a	Fvb1	20.078.378	20.323.817	1	confirmed by Tennessen et al 2014
LG6	34.750.958	35.927.392	LG1	4.386.861	20.747.404	0	scf0512991b	Fvb1	9.198.415	10.545.245	1	confirmed by Tennessen et al 2014
LG7	610.205	792.402	LG1	4.386.861	20.747.404	0	scf0513170b	Fvb1	19.129.116	19.356.256	1	confirmed by Tennessen et al 2014
LG3	21.014.819	21.542.491	LG2	92.740	16.919.611	0	scf0513125c	Fvb2	5.853.827	6.556.313	0	confirmed by Tennessen et al 2014
LG6	13.214.824	13.831.033	LG2	17.392.459	17.626.895	0	scf0513196a	Fvb2	23.019.705	23.707.348	0	confirmed by Tennessen et al 2014
LG2	21.489.169	22.663.853	LG2	>23.000.000		-1	scf0513194a	Fvb2	27.690.570	28.944.653	-1	confirmed by Tennessen et al 2014
LG2	22.718.421	24.523.482	LG3	5.668.456	27.749.642	0	scf0513194b	Fvb3	20.778.152	22.607.077	-1	confirmed by Tennessen et al 2014
LG5	14.391.546	14.773.705	LG3	5.668.456	27.749.642	0	scf0513106b	Fvb3	19.761.650	20.162.852	-1	confirmed by Tennessen et al 2014
LG7	199.390	583.090	LG3	5.668.456	27.749.642	0	scf0513170a	Fvb3	8.843.731	9.442.893	1	confirmed by Tennessen et al 2014
LG3	15.214.631	15.743.622	LG4	579.729	24.200.396	0	scf0513053	Fvb4	18.170.580	18.574.896	-1	confirmed by Tennessen et al 2014
LG5	25.287.249	26.202.863	LG4	579.729	24.200.396	0	scf0513187b	Fvb4	4.643.413	5.635.634	0	confirmed by Tennessen et al 2014
LG4	14.097.814	15.001.370	LG4	26.554.022	27.173.808	0	scf0512935	Fvb4	31.516.049	32.494.384	-1	confirmed by Tennessen et al 2014
LG4	13.297.334	14.039.986	LG5	3.377.329	6.557.340	0	scf0513150b	Fvb5	3.457.498	4.220.883	1	confirmed by Tennessen et al 2014
LG1	3.365.259	4.102.736	LG5	8.530.103	27.902.872	0	scf0513114	Fvb5	24.681.213	25.460.500	-1	confirmed by Tennessen et al 2014
LG3	11.669.089	11.801.654	LG5	8.530.103	27.902.872	0	scf0513115	Fvb5	22.538.780	23.396.617	-1	confirmed by Tennessen et al 2014
LG6	19.624.821	19.703.361	LG5	8.530.103	27.902.872	0	scf0512952b	Fvb5	29.099.740	29.214.561	0	confirmed by Tennessen et al 2014
LG7	16.493.182	-	LG6	8.467.792	27.394.698	0	scf0513113b	Fvb6	18.717.510	19.053.068	0	confirmed by Tennessen et al 2014
LG4	23.065.977	23.800.559	LG6	8.467.792	27.394.698	0	scf0513004b	Fvb6	25.452.010	26.329.913	-1	confirmed by Tennessen et al 2014
LG4	10.084.335	10.114.581	LG6	33.982.439	37.008.825	0	scf0513124_4b	Fvb6	36.292.487	36.541.872	1	confirmed by Tennessen et al 2014
LG6	1.589.394	1.951.977	LG6	33.982.439	36.467.630	1	scf0512983	Fvb6	35.771.599	36.191.045	1	confirmed by Tennessen et al 2014
LG6	11.278.703	11.823.122	LG7	821.088	14.152.037	0	scf0513185b	Fvb7	1.089.753	1.753.650	-1	confirmed by Tennessen et al 2014
LG3	9.824.073	10.908.787	LG7	821.088	14.152.037	0	scf0513156	Fvb7	7.149.241	8.375.257	-1	confirmed by Tennessen et al 2014
LG7	20.230.789	22.511.020	LG7	20.230.789	22.511.020	-1	scf0513190	Fvb7	20.356.950	23.627.137	-1	confirmed by Tennessen et al 2014
LG3	47.000	1.420.634	LG3	47.000	1.420.634	-1	scf0513171	Fvb3	1	1.497.153	-1	confirmed by Tennessen et al 2014
LG4	6.629.502	6.708.323	LG3	5.668.456	27.749.642	0	scf0513163	Fvb4	5.645.635	8.519.973	0	not confirmed
LG4	21.501.885	21.673.366	LG3	5.668.456	27.749.642	0	scf0513154	Fvb4	21.139.143	22.808.238	0	not confirmed
LG5	28.233.300	28.278.585	LG3	5.668.456	27.749.642	0	scf0512963	Fvb5	15.870.954	16.207.446	0	not confirmed

[a] position in bp in reference genome *F. vesca* v1.1

[b] position in bp in reference genome *F. vesca* v2.0

[c] orientation of fragments or scaffolds relative to *F. vesca* v1.1 assembly. 1=same orientation, -1=inverted orientation, 0= no information about orientation.

Connection between genetic and physical distance

All *F. vesca* genome is covered with overlapping introgressions summing a total of 210 Mb, except a region of maximum 1.5 Mb in LG1 (LG1:1882448-3401778), where introgressions do not overlap. Higher marker density also revealed that even lines Fb**3**:0-15 and Fb**3**:29-54 that were considered to harbor non-overlapping introgressions with microsatellites (Urrutia *et al.* 2015) had in fact overlapping fragments. Physical introgression sizes, without considering whole chromosome introgressions, range from 1 Mb (0.4% of genome) to 31 Mb (15% of genome) with an average size of 16 Mb (7.6% of total coverage). Genome fragments identified by overlapping regions or bins have an average size of roughly 5 Mb (2.3% of genome), ranging from 0.16 to 29.30 Mb (0.07-13% of total coverage).

Comparing genetic and physical distances of introgressions along NIL collection an important lack of proportion was detected. For instance in LG7, a relatively short introgression of 10 cM extends over more than half the chromosome (LG7: 0-14.169.385 bp) (1.42 Mb/cM) while the remaining 49 cM in LG7 occupy a region of 8.341.635 bp (0.17 Mb/cM). Other example, at the end of LG4 where a region of only 4.54 Mb corresponds to a genetic distance of 48 cM (0.09 Mb/cM) while at the beginning of the LG4, 20 cM match to 22.63 Mb (1.13 Mb/cM). A clear disproportion is also found in LG2 were intervals defining different exotic introgressions have similar genetic sizes (10 to 20 cM), however physical distances vary from the 16.92 Mb of the 30 cM (0.56Mb/cM) introgression in Fb**2**:0-30 to the 0.33 Mb of the 18 cM interval between Fb**2**:0-45 and Fb**2**:0-63 (0.02Mb/cM).

This lack of proportion between genetic and physical distance discloses regions with low (LG2:0-30, LG4:0-20, LG5:50-76, LG7:0-10) and high (LG2:30-73, LG4:20-44, LG7:10-59) recombination rates.

Figure AI. 3 SNP clustering example. A: Ideal SNP clustering for diploid segregation. B: Off target variant (OTV) classified as a polymorphic high resolution SNP. C: Poorly resolved clusters.

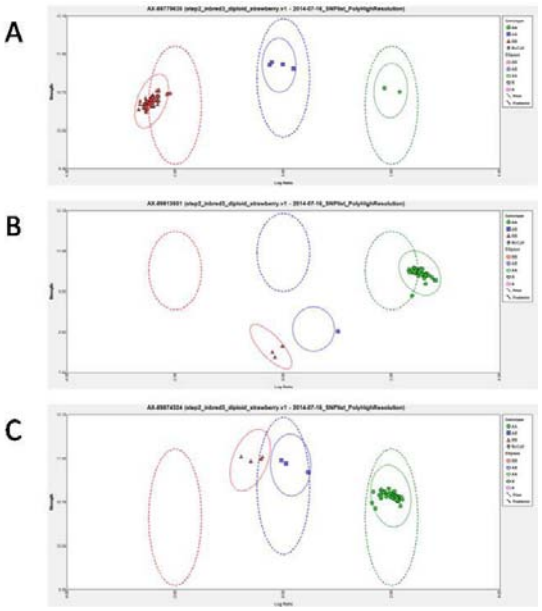


Table AI. 2 Exotic introgression coordinates in *F. vesca* NIL collection lines according to strawberry reference genome version 2.0

	introgression (cM)			introgression (bp)								
NIL	LG	start	end	Fvb	start	end	Fvb	start	end	Fvb	start	end
Fb1:0-6	1	0	6	1	0	3.401.778	6	33.528.472	38.738.438			
Fb1:0-6-26	1	0	6	1	0	3.401.778	6	38.128.254	38.738.438			
Fb1:26-61	1	26	61	1	1.882.448	23.185.958	6	34.537.486	38.738.438			
Fb1:33-61	1	33	61	1	1.882.448	23.185.958						
Fb1:50-61	1	50	61	1	1.882.448	23.185.958						
Fb1:58-61	1	58	61	1	1.882.448	23.185.958						
Fb2:0-30	2	0	30	2	0	21.514.539						
Fb2:0.39	2	0	39	2	0	25.241.015						
Fb2:0-45	2	0	45	2	0	27.203.637						
Fb2:0-63	2	0	63	2	0	27.961.648						
Fb2:0-73	2	0	73	2	0	28.911.401						
Fb2:39-47	2	39	47	2	0	27.203.637						
Fb2:39-63	2	39	63	2	23.421.616	28.120.868						
Fb3:0-8	3	0	8	3	0	3.561.105	6	27.860.882	36.145.707			
Fb3:0-15	3	0	15	3	0	3.561.105	6	27.860.882	36.145.707			
Fb3:29-94	3	29	94	3	1.152.350	32.864.194						
Fb3:54-94	3	54	94	3	1.152.350	32.864.194						
Fb4:0-20	4	0	20	4	0	26.081.626						
Fb4:0-31	4	0	31	4	0	27.599.553						
Fb4:0-44	4	0	44	4	0	30.845.487						
Fb4:0-78	4	0	78	4	0	32.449.957						
Fb4:58-78	4	58	78	4	28.348.238	32.449.957	2	0	3.880.862	2	4.624.566	13.283.763
Fb5:0-11	5	0	11	5	0	2.662.821	1	0	3.401.778			
Fb5:0-20	5	0	20	5	0	3.239.069						
Fb5:0-35	5	0	35	5	0	5.953.948						
Fb5:0-76	5	0	76	5	0	29.183.100						
Fb5:11-76	5	11	76	5	1.248.665	29.183.100	1	1.882.448	23.185.958			
Fb5:20-76	5	20	76	5	1.532.188	29.183.100						
Fb5:39-76	5	39	76	5	8.945.335	29.183.100	2	0	21.514.539			
Fb5:41-76	5	41	76	5	8.945.335	29.183.100	2	0	25.241.015			
Fb5:50-76	5	50	76	5	10.683.361	29.183.100	2	0	27.203.637			
Fb6:0-5	6	0	5	6	0	6.734.870	2	25.146.779	28.120.868			
Fb6:30-39	6	30	39	6	0	27.812.416	2	0	27.961.648	4	26.777.088	28.136.898
Fb6:32-32	6	32	32	6	9.325.727	36.145.707	3	0	3.326.960	4	28.348.238	31.457.022
Fb6:32-39	6	32	39	6	9.325.727	33.528.472	3	3.334.836	4.239.677			
Fb6:39-39	6	39	39	6	9.998.570	36.145.707	3	0	3.326.960			
Fb6:71-101	6	71	101	6	34.882.837	38.738.438	3	fragmented				
Fb6:84-101	6	84	101	6	33.528.472	38.738.438	3	fragmented		4	fragmented	
Fb6:101-101	6	101	101	6	34.537.486	38.738.438	3	26.903.811	27.840.802	3	28.919.386	29.637.517
Fb7:0-10	7	0	10	7	0	15.652.556	4	fragmented				
Fb7:0-27	7	0	27	7	0	21.595.502	4	fragmented				
Fb7:26-27	7	26	27	7	17.735.496	21.595.502	4	fragmented				
Fb7:26-45	7	26	45	7	17.735.496	22.738.269	4	fragmented				
Fb7:26-59	7	26	59	7	17.735.496	23.459.679	5	0	3.239.069	4	fragmented	
Fb7:43-59	7	43	59	7	21.867.401	23.459.679	5	0	5.953.948	4	fragmented	
Fb7:52-59	7	52	59	7	22.435.032	23.459.679	5	1.248.665	5.953.948			
Fb7:0-59	7	0	59	7	0	23.459.679	5	1.532.188	29.183.100			

Discussion

The final set of useful SNPs (1510) represents a clear improvement in genotyping density over the previous set of 72 microsatellites makers. It has revealed that almost all *F. vesca* genome (210 of 240 Mb attending to Schulaev *et al.* (2011)) is covered with *F. bucharica* overlapping introgressions. The only non-covered region in the genome is located in LG1 and spreads over 19 cM in genetic distance (Urrutia *et al.* 2015), however it is smaller than was suspected in physical distance and only has 1.5 Mb. Regions with high and low recombination rates were identified, despite of this, average proportion of total genome covered in physical distance by introgressions (7.6%) and defined by bins (2.3%) are very similar to those predicted with genetic distances (6.1% and 2.6% respectively, Urrutia *et al.* 2015).

The IStraw90® genotyping has been useful for a precise definition of the exotic introgressions of *F. bucharica* into the genetic background of *F. vesca* throughout the NIL collection and to detect miss-assemblies in the *F. vesca* reference genome v1.1 as was later confirmed by Tennessen *et al.* (2014). However, for technical reasons, it has not been useful for detecting non-desired introgressions or to determine if the introgressions are fixed in homozygosity. Therefore only genotypic results validated with complementary data (like SSRs) can be trusted.

More specific genotyping strategies should be considered in order to improve NIL collection genotyping quality and accuracy in further studies such as an SNP set designed *ad-hoc* for *F. vesca* and *F. bucharica* or genotyping by sequencing (GBS).

Chapter II

Genetic dissection of the polyphenol profile of diploid strawberry (*Fragaria vesca*) fruits using a NIL collection

Part of the content of this chapter is accepted for publication in Plant Science (July 2015)

Acceptance confirmation

ELSEVIER

- Home
- Products
- Alerts
- User Resources
- About Us
- Support & Contact
- Elsevier Websites

Search Advanced Product Search

Author's Home > Track your accepted article

TRACK YOUR ACCEPTED ARTICLE

Welcome! Login to get personalized options. New user? Register | Why register? Help

Your article's details and status are shown in the following table:

Article status

Article title:	Genetic dissection of the (poly)phenol profile of diploid strawberry (Fragaria vesca) fruits using a NIL collection
Reference:	PSL9246
Journal title:	Plant Science
Corresponding author:	Dr. Amparo Monfort
First author:	Dr. Maria Urrutia
Received at Editorial Office:	20 May 2015
Article revised:	24 Jul 2015
Article accepted for publication:	25 Jul 2015
Received at Elsevier:	25 Jul 2015
PDF offprint:	Yes
DOI information:	10.1016/j.plantsci.2015.07.019
Status comment:	Your article has been received at Elsevier.
	You will receive a PDF file of your published article free of charge.

Track another article

Title: Genetic dissection of the (poly)phenol profile of diploid strawberry (*Fragaria vesca*) fruits using a NIL collection

Authors: Maria Urrutia[1], Wilfried Schwab[2], Thomas Hoffmann[2], Amparo Monfort[1].

[1] IRTA. Center for Research in Agricultural Genomics (CSIC-IRTA-UAB-UB), Campus UAB, 08193 Bellaterra, Barcelona, Spain
[2] Biotechnology of Natural Products, Technische Universität München, Liesel-Beckmann-Str. 1, 85354 Freising, Germany

Keywords: Flavonoids, phenylpropanoids, fruit nutritional quality, metabolomic QTL, Near Isogenic Lines, strawberry

Abstract

Over the last few years, diploid strawberry (*Fragaria vesca*) has been recognized as a model species for applied research of cultivated strawberry (*Fragaria x ananassa*) that is one of the most economically important crops. Berries, particularly strawberries, are known for their high antioxidant capacity due to a high concentration of (poly)phenolic compounds. Studies have already characterized the phenolic composition of fruits from sets of cultivated strawberries but the quantification of phenolics in a *Fragaria* mapping population has not been reported yet. The metabolite profiling of a *F. vesca* near isogenic line (NIL) collection by LC-MS allowed the unambiguous identification of 22 (poly)-phenols, including anthocyanins, flavonols, flavan-3-ols, flavanones, hydroxycinnamic acid derivatives and ellagic acid, a furanone and citric acid in the diploid strawberry fruit. The variability in the collection revealed that the genetic factor was more decisive than the environmental factor for the accumulation of 18 of the 24 compounds. Genotyping the NIL collection with the Axiom® IStraw90® SNPs array, we were able to map 76 stable QTLs controlling accumulation of the (poly)-phenolic compounds. They provide a powerful new tool to characterize candidate genes to increase the antioxidant capacity of fruits and produce healthier strawberries for consumers.

Introduction

Strawberries are one of the most popular fruits among consumers worldwide, breeding programs have traditionally focussed on fruit agronomic traits such as firmness or colour and resistance to pests, but little attention has been paid to fruit nutritional quality. Strawberries (and other berries) have interesting nutritional properties and a very high antioxidant capacity mainly due to their high proportion of (poly)-phenolic compounds. There is evidence that consumption of berries and other phenolic-rich products has beneficial effects on human health (Ross & Kasum 2002; Scalbert *et al.* 2005; Ramassamy 2006; Battino *et al.* 2009). Breeding programs aiming to improve fruit nutritional quality could have a positive impact on consumer health. The (poly)-phenols that account for most of this antioxidant capacity belong to two branches of the phenylpropanoid pathway: flavonoids and hydroxycinnamic acid derivatives. Phenylpropanoids are a large family of compounds usually classified as "secondary metabolites" although they carry out important functions for plant survival and fitness such as seed dispersion and resistance to biotic and abiotic stress (Gould & Lister 2006).

Phenylpropanoids arise from the phenylalanine derived from the shikimate pathway, and can lead to hydroxycinnamic acid derivatives and to flavonoids (anthocyanins, flavonols, flavanones and flavan-3-os) whereas chalcone synthase (CHS) represents the branching point. The phenylpropanoid biochemical pathway has been extensively studied and described in several plant species, including *Arabidopsis*, poplar, tomato and rose (Vogt 2010). Regulation of the pathway is modulated by the development stage and environmental factors, involving both structural genes and transcription factors (Lepiniec *et al.* 2006 and references herein). In strawberry, many of the most relevant enzymes have been characterized and the regulation described (Almeida *et al.* 2007; Carbone *et al.* 2009).

The octoploid nature of commercial varieties of strawberry (mainly varieties from the *Fragaria x ananassa*) is a drawback for breeding programs. However the diploid wild relative *Fragaria vesca* (woodland or wild strawberry) can be used as a model for breeding purposes, taking advantage of the many tools that have been developed recently for its study and which demonstrate that *F. vesca* and the commercial hybrid *F. x ananassa* share a very high degree of synteny (Tennessen *et al*. 2014). Saturated genetic linkage maps (Ruiz-Rojas *et al.* 2010) and synteny studies between *F. vesca*, *F. x ananassa*, and other model species from the *Rosaceae* family (*Prunus persica* and *Malus x domestica)* available facilitate transfer of the results between species. This is especially important considering the long intergenerational period of fruit trees (Illa *et al.* 2011, Jung *et al.* 2012, Rousseau-Gueutin *et al.* 2008, Vilanova *et al.* 2008). Besides, the genome and the annotation of *F. vesca* var. Hawaii 4 have been published (Shulaev *et al.* 2011), and interesting population resources have been developed in *F. vesca*, including a bin mapping population (Sargent *et al.* 2008) a mutant collection (Veilleux *et al.* 2012) and a near isogenic lines (NILs) collection (Urrutia *et al.* 2015). A new genotyping array, IStraw90® (Bassil & Davis *et al.* 2015), has been also recently relised. So there are both quality and genetic reasons for using *F. vesca* as a model species for the *Fragaria genus* and as a source of variability for *Fragaria* breeding programs.

The mapping population that we used in this study is described in detail in Urrutia *et al*. (2015). NIL collections are mapping populations based on introgression lines. All the lines share a

common genetic background (usually from an elite variety) and have introgressions (ideally fixed in homozygosity) from a donor parental (usually an exotic relative). Together, all the lines must cover the background genome with overlapping exotic introgressions (Esed & Zamir 1995) These introgressions can be tracked using molecular markers such as microsatellites (SSRs) or single nucleotide polymorphisms (SNPs).

The main advantage of NIL collections compared to other mapping populations is that they isolate quantitative trait loci (QTL) simplifying their inheritance, the estimation of their effect and their dissection by positional cloning using sub-NILs. NIL collections have been developed and successfully used for QTL mapping in many crop species, such as tomato (Monforte & Tanksley 2000; Monforte *et al.* 2001), melon (Eduardo *et al.* 2005; Moreno *et al.* 2008; Vegas *et al.* 2013), wheat (Pestsova *et al.* 2001), lettuce (Jeuken & Lindhout 2004) and also in the model species *A. thaliana* (Koumproglou *et al.* 2002; Fletcher *et al.* 2013).

Multiple studies have characterized the phenolic composition of fruits from sets of strawberry varieties, mainly from octoploid accessions (Buendia *et al.* 2010; Muñoz *et al.* 2011; Aaby *et al.* 2012) however there has never been a rigorous characterization of phenolic content in a *Fragaria* mapping population. In this study we characterised in detail the phenolic composition of ripe strawberry fruit in a *F. vesca* NIL collection in order to map the metabolomic QTL (mQTL) responsible for the variability of metabolites, or groups of metabolites, and to quantify the effect of the genotype and the environment on the phenolic composition in wild strawberry as a tool for increase nutritional quality in cultivated strawberry breeding programs.

Materials and methods

Chemicals

All chemicals, solvents and reference compounds were obtained from Sigma, Aldrich, Fluka, Riedel de Haën (all Taufkirchen, Germany), Merck (Darmstadt, Germany) or Roth (Karlsruhe, Germany).

Plant material and sample extraction

Six to eight plant replicates, per genotype, of the available *Fragaria vesca* NILs collection (Urrutia *et al.* 2015) 42 lines in total (Supplemental Figure CII. 1), the recurrent parental *F. vesca* var. *Reine des vallées* (RV), the hybrid (*F. vesca* x *F. bucharica*) F_1 parental of the whole collection, and the white fruited variety of *F. vesca* var. 'Yellow wonder' (YW), were planted in a shade greenhouse in Caldes de Montbui (latitude: 41º 36'N, longitude: 2º 10' E, altitude 203m above sea level, pre-coastal Mediterranean climate) for two consecutive years (2012 and 2013). As plants were maintained in outfield conditions we can consider each harvest year as an independent experiment. Three to five biological replicates of ripe strawberry fruits were harvested. Each biological replicate, a mix of ten to 15 fruits (> 5 g), was individually and immediately frozen in liquid nitrogen and ground to a fine powder. An aliquot of 500 mg of the powder from every biological replicate was used for extraction. A biochanin A solution (250 µl in methanol; 0.2 mg mL^{-1}) was added as internal standard (IS) yielding 50 µg of IS in each sample. After addition of 250 µL methanol, 1 min vortexing and sonication for 10 min, the sample was centrifuged at 16000 g for 10min. The supernatant was removed and the residue re-extracted with 500 µL methanol. The supernatants were combined, concentrated to dryness in a vacuum concentrator and re-dissolved in 35 µL water. After 1 min vortexing, 10 min sonication and 10 min centrifugation at 16000 g, the clear supernatant was used for LC-MS-analysis. Each extract was injected twice as a technical replicate.

LC-ESI-MS^n analysis

Samples were analyzed on an Agilent 1100 HPLC/UV-system (Agilent Technologies, Waldbronn, Germany) equipped with a reversed phase column (Luna 3 µ C18(2) 100A 150x2 mm, Phenomenex, Aschaffenburg, Germany) and connected to a Bruker esquire3000_{plus} ion trap mass spectrometer (Bruker Daltonics, Bremen, Germany). Solutions of 0.1% formic acid in water (A) and 0.1% formic acid in methanol (B) were used as mobile phases. The injection volume was 5 µL and the flow rate 0.2 mL min^{-1}. The gradient was from 0% B to 50% B in 30 min. Then to 100% B in the next 5 min, maintained at these conditions for 15 min and decreased to 0% B over 5 min. The initial conditions were held for 10 min for system equilibration. The electrospray ionization voltage of the capillary was set to -4000 V and the end plate to -500 V. Nitrogen was used as nebulizer gas at 30 psi and as dry gas at 330 ºC and 9 L min^{-1}. MS spectra were recorded in alternating polarity mode. The full scan ranged from 100 – 800 $m\,z^{-1}$. Ions were accumulated until an ion charge control (ICC) target, 20000 (positive mode) and 10000 (negative mode), was achieved or the maximum time of 200 ms was reached. Tandem mass spectrometry was carried out using helium as the collision gas at 4×10^{-6} mbar and a collision voltage of 1 V. Data were analyzed with Data Analysis 5.1 software (Bruker Daltonics, Bremen, Germany). Metabolites were identified by comparing their

retention times and mass spectra (MS and MS2) with that of measured authentic reference compounds. The major known phenolic metabolites were quantified in the positive and negative MS mode by the internal standard method, using QuantAnalysis2.0 (Bruker Daltonics, Bremen, Germany), and expressed as part per ten thousand equivalents of fresh weight (fw) assuming a response factor of 1.

Data and QTL analysis

All lines that set fruit each year were processed and analyzed by LC-MS and taken into account for QTL mapping, however, for the exploratory analysis only those genotypes that gave more than 5 g fruit production both years were considered. All statistical analyses and graphical representations were using the free source software R v.2.15.1 (RCoreTeam 2012) with the Rstudio v.0.92.501 interface (RStudio 2012) unless otherwise specified. Pearson's correlation between the metabolites and the genotypes was calculated using the *rcorr* function from the Hmisc package (Harrell 2014). The analysis of variance (ANOVA), fitting the model G+E+GxE, was calculated using the *Anova* function from the car package for R (Fox 2011) and the omega squared values (ω^2) were obtained from ANOVA parameters according to the following formula: $(SS_i - df_i * MS_{error}) * (SS_t + MS_{error})^{-1}$

Principal components analysis (PCA) was calculated with the *prcomp* function and scaled values. Hierarchical clustering analysis (HCA) was calculated considering Euclidean distance and complete linkage clustering method. Cluster network analysis (CNA) was using the correlation values with the *qgraph* function from the qgraph R package (Epskamp 2012) The significance tests were recursively calculated using the *t.test* function, comparing each genotype mean value with the RV mean value in the same harvest. The p-values obtained were corrected by false discovery rate (FDR) with the *p.adjust* function. Significant threshold was set at p-value < 0.05 after correction. QTLs were mapped to a specific bin only when all the NILs carrying an exotic introgression in this bin had a significant effect, and in the same direction, for the specific metabolite of study. To confirm these QTLs and estimate their effect, an interval mapping analysis was performed using MapQTL v.6 (Van Ooijen 2009) Stable regions that explained around 20% or more of the variability and had LOD scores >1.8 were considered strong QTLs. The graphical representation of the mQTLs was using MapChart 2.2 (Voorrips 2002).

Genotyping analysis

DNA from the parental lines of the NIL collection and their hybrid (*F. vesca* var. Reine des Vallées, *F. bucharica* FDP60147 and F1), and from 47 NILs was extracted using QIAGEN Plant DNA kit (Qiagen, Hilden, Germany). A total of 53 diploid samples (47 NILs, parentals and the hybrid per duplicate) were hybridized with the IStraw90® SNP array (Affymetrix, CA, USA) (Bassil & Davis *et al.* 2015) at CEGEN (Santiago de Compostela, Spain). Genotyping analysis was performed using Genotyping console™ and SNPpolisher© (both from Affymetrix, CA, USA) following manufacturers recommendations and an inbred penalty of 3 for the NIL collection. Double recombinants for one single SNP were considered as genotyping errors and the flanking allele was assumed.

Results

Global metabolite profiling analysis

To evaluate the different levels of phenolic compounds identified by LC-MS and detect genomic regions responsible for their control we analyzed the metabolite profiles of fully ripe strawberry fruits from 42 interspecific introgression lines of a *F. vesca* NIL collection (Urrutia *et al*. 2015). The recurrent parental, *F. vesca* var. Reine des Vallées (RV) was used as population control. Additionally the white-fruited *F. vesca* var. Yellow Wonder (YW) served as an external control or out-group. Over a period of two consecutive years (2012 and 2013), the metabolite profiling analyses and the QTL mapping was carried out with all the genotypes that set enough fruit each year. However, only those 25 genotypes that could be analyzed with at least three biological replicates in both years were taken into account for the exploratory statistical analysis (Supplemental Figure CII.1).

Among the 936 putative metabolites observed (data not shown), 24 were unambiguously identified by comparison with commercial standards run under the same conditions. The known compounds were quantified in positive and negative MS mode with an internal standard, assuming a response factor of 1. The results are expressed as part per ten thousand (mg/10g) of fresh weight (fw) (Supplemental Table CII.1) and graphically represented in Figure 1. Various compounds were identified. This included three anthocyanins, pelargonidin-3-glucoside, pelargonidin-3-glucoside-malonate and cyanidin-3-glucoside; five flavonols, kaempferol- and quercetin- glucosides and glucuronides and kaempferol-coumaryl-glucoside; six flavan-3-ols, procyanidin -B1 and - B3, catechin, two isomers of (epi) catechin dimers and a (epi) afzelechin - (epi) catechin dimer; two flavanones, isomers of eriodictyol; five hydroxycinnamic acid derivatives, cinnamoyl-glucose, p-coumaroyl-glucose and – glucoside, feruloyl-glucose and caffeoyl-glucose. In addition, there were three other compounds not belonging to any of the previous families: ellagic acid, 4-hydroxy-2,5-dimethyl-3(2H)-furanyl glucoside (HDMF-glucoside) and citric acid all previously described in strawberry (Aaby *et al.* 2012; Buendía *et al.* 2010; Määttä-Riihinen *et al.* 2004; Fait *et al.* 2008; Medina-Puche *et al.* 2014; Muñoz *et al.* 2011; Ring *et al.* 2013). For abbreviations on these compounds names see Table CII.1

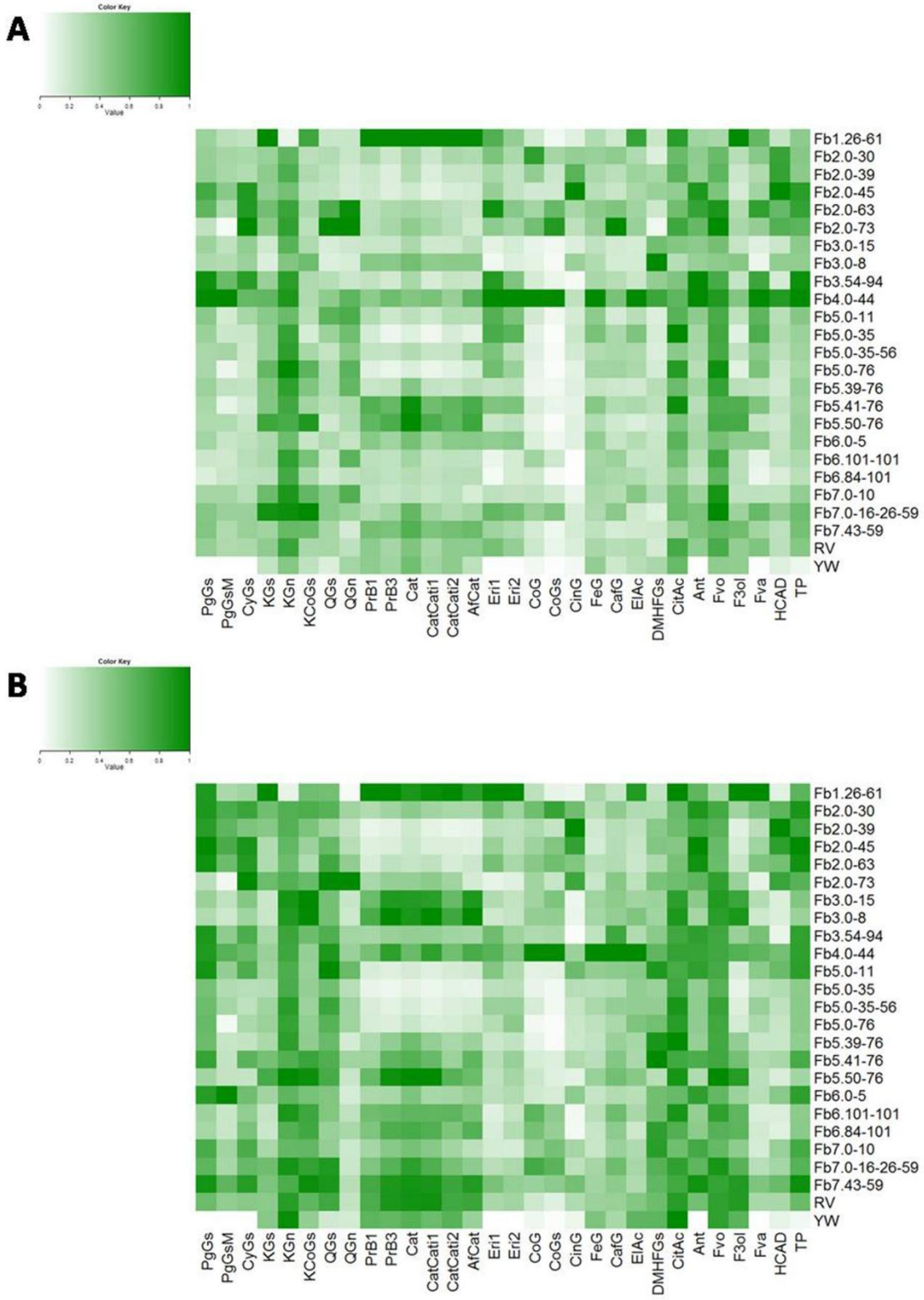
We also quantified the content of chemical families by summing up the metabolites identified: anthocyanins (Ant), flavonols (Fvo), flavanones (Fva), flavan-3-ols (F3ol) and hydroxycinnamic acid derivatives (HCAD). Additionally the total phenolics (TP) accumulation was evaluated by summing all identified compounds, except for citric acid and HDMF-glucoside because they are not phenolic compounds.

Table CII.1 Phenolic compound accumulation summary. Data presented correspond to mean (expressed as mg/10g of fw), standard deviation (sd) and coefficient of variation (cv) of each identified compound or chemical family quantified in the recurrent parental F. vesca (RV) and in the average NIL collection for 2012 and 2013 harvests. The range (min and max values) are given for the NIL collection in both harvests.

		parental *F. vesca* (RV)						**NIL collection average**									
		2012			**2013**			**2012**					**2013**				
Compound	**Abv**	**mean (mg/10 g fw)**	**sd**	**cv**	**mean (mg/10 g fw)**	**sd**	**cv**	**mean (mg/10 g fw)**	**sd**	**cv**	**range**		**mean (mg/10 g fw)**	**sd**	**cv**	**range**	
Pelargonidin-3-glucoside	PgGs	8,46	1,65	0,20	13,24	8,32	0,63	8,83	5,55	0,63	1,08	33,26	15,57	8,33	0,54	0,48	40,48
Pelargonidin-3-glucoside-malonate	PgGsM	1,81	0,39	0,22	3,54	1,37	0,39	1,71	1,31	0,76	0,00	8,20	3,32	2,65	0,79	0,11	18,68
Cyanidin-3-glucoside	CyGs	5,37	1,23	0,23	11,72	7,06	0,60	6,76	5,18	0,77	1,38	32,09	14,40	8,52	0,59	1,68	44,20
Kaempferol-glucoside	KGs	0,16	0,06	0,37	0,13	0,05	0,42	0,16	0,09	0,52	0,04	0,63	0,11	0,06	0,56	0,03	0,37
Kaempferol-glucuronide	KGn	1,50	0,40	0,27	1,10	0,31	0,28	1,36	0,49	0,36	0,11	3,03	0,96	0,32	0,35	0,09	2,15
Kaempferol-coumaryl-glucoside	KCoGs	0,11	0,05	0,45	0,18	0,07	0,42	0,13	0,08	0,60	0,02	0,46	0,17	0,09	0,53	0,03	0,45
Quercetin-glucoside	QGs	0,29	0,11	0,39	0,13	0,04	0,30	0,29	0,19	0,65	0,08	1,00	0,11	0,06	0,51	0,03	0,32
Quercetin-glucuronide	QGn	0,08	0,06	0,70	0,04	0,02	0,41	0,13	0,11	0,87	0,00	0,53	0,06	0,06	0,88	0,00	0,34
Procyanidin B1	PrB1	0,31	0,11	0,34	0,23	0,12	0,53	0,24	0,15	0,64	0,05	0,80	0,12	0,10	0,83	0,01	0,51
Procyanidin B3	PrB3	0,20	0,08	0,40	0,22	0,14	0,62	0,18	0,14	0,78	0,03	0,73	0,12	0,11	0,93	0,00	0,63
Catechin	Cat	0,29	0,09	0,32	0,29	0,17	0,60	0,24	0,16	0,64	0,05	0,92	0,16	0,12	0,78	0,01	0,63
(epi)catechin dimers iso1	CatCati1	0,21	0,09	0,45	0,24	0,14	0,58	0,18	0,15	0,79	0,03	0,76	0,12	0,12	0,95	0,00	0,68
(epi)catechin dimers iso2	CatCati2	0,33	0,12	0,38	0,23	0,11	0,46	0,24	0,16	0,65	0,03	0,80	0,12	0,11	0,84	0,00	0,57
(epi)afzelechin-(epi)catechin dimers	AfCat	0,04	0,01	0,35	0,03	0,02	0,56	0,03	0,02	0,71	0,00	0,12	0,02	0,02	0,94	0,00	0,09
Eriodictyol iso1	Eri1	0,39	0,20	0,50	0,18	0,08	0,43	0,34	0,25	0,73	0,00	1,15	0,18	0,13	0,70	0,00	0,73
Eriodictyol iso2	Eri2	0,11	0,05	0,46	0,05	0,02	0,35	0,10	0,06	0,66	0,02	0,41	0,04	0,03	0,67	0,00	0,18

		parental *F. vesca* (RV)						NIL collection average									
		2012			2013			2012					2013				
Compound	**Abv**	**mean (mg/10 g fw)**	**sd**	**cv**	**mean (mg/10 g fw)**	**sd**	**cv**	**mean (mg/10 g fw)**	**sd**	**cv**	**range**		**mean (mg/10 g fw)**	**sd**	**cv**	**range**	
p-Coumaroyl-glucose ester	CoG	0,08	0,04	0,47	0,04	0,01	0,33	0,13	0,12	0,91	0,02	0,66	0,07	0,07	0,95	0,00	0,61
p-Coumaroyl-glucoside	CoGs	0,01	0,01	1,06	0,01	0,01	1,02	0,03	0,05	1,29	0,00	0,26	0,03	0,04	1,13	0,00	0,32
Cinnamoyl-glucose ester	CinG	0,14	0,04	0,28	0,76	0,70	0,93	0,28	0,35	1,27	0,00	2,35	0,89	0,90	1,01	0,00	4,60
Feruloyl-glucose ester	FeG	0,04	0,01	0,36	0,04	0,01	0,42	0,05	0,03	0,54	0,01	0,21	0,03	0,02	0,80	0,01	0,23
Caffeoyl-glucose ester	CafG	0,03	0,02	0,58	0,02	0,01	0,37	0,04	0,03	0,65	0,01	0,13	0,02	0,02	0,74	0,00	0,17
Ellagic acid	ElAc	0,10	0,03	0,27	0,06	0,03	0,47	0,12	0,08	0,69	0,02	0,69	0,05	0,03	0,75	0,00	0,23
DMHF-glucoside	DMHFGs	0,17	0,06	0,34	0,80	0,86	1,07	0,29	0,38	1,33	0,02	2,65	0,78	0,64	0,83	0,06	4,22
Citric acid	CitAc	12,09	3,47	0,29	5,97	1,82	0,31	10,02	4,48	0,45	0,00	23,63	5,18	1,91	0,37	2,52	11,32
Anthocyanins	Ant	15,63	2,71	0,17	28,50	16,10	0,57	17,31	10,86	0,63	4,15	68,47	33,30	17,20	0,52	6,29	97,02
Flavonols	Fvo	2,14	0,62	0,29	1,58	0,42	0,26	2,08	0,74	0,35	0,79	5,03	1,41	0,46	0,32	0,61	2,73
Flavan-3-ols	F3ol	1,38	0,48	0,35	1,24	0,68	0,55	1,11	0,74	0,66	0,25	3,90	0,66	0,56	0,84	0,04	2,94
Flavanones	Fva	0,50	0,24	0,47	0,22	0,09	0,39	0,43	0,30	0,69	0,03	1,56	0,23	0,15	0,67	0,02	0,84
Hydroxycinnamic acid deriv.	HCAD	0,31	0,09	0,27	0,87	0,69	0,79	0,53	0,41	0,77	0,11	2,53	1,05	0,88	0,84	0,08	4,70
Total phenolics	TP	20,07	3,29	0,16	32,47	16,24	0,50	21,57	10,98	0,51	6,98	72,60	36,69	17,60	0,48	8,11	104,76

Figure CII. 1 Relative metabolite profiling in 2012 (A) and 2013 (B). Detected compounds are at the bottom and the NIL analyzed in both harvests on the right. The white (0%) to green (100%) colour scale represents the relative accumulation of the compound in each NIL, RV and YW line.



General distribution of metabolites in *F.vesca* RV

The general distribution parameters (mean, standard deviation and coefficient of variation) of the metabolites in independent harvests (2012 and 2013) were evaluated for RV and the NIL collection separately (Table CII.1). The range calculated for the NIL collection went from undetectable concentrations to a 9-fold change in the mean of the collection.

Correlation coefficients between the compounds in different harvests were very high (Table CII.2). Moreover 21 out of 24 compounds (87%) presented a significant positive correlation between the two years. Significant correlation values ranged between 0.42 and 0.77, being the median greater than 0.70. Considering the chemical families, all except the flavonols presented significant correlation values. These results indicate a genotypic effect in the pattern of inheritance of phenolic expression.

The relative contribution of the individual compounds to their chemical family, and that of the chemical families and ellagic acid to total phenolics content was calculated for both harvests independently (Table CII.3). Both absolute and relative values for phenolics concentration were very similar between RV and the average *F. vesca* NIL collection within each harvest (Tables CII. 1 and 3), this was expected considering that all lines from the NIL collection share their genetic background with the recurrent parental RV (Supplemental Figure CII. 1), (Urrutia *et al.* 2015) Therefore, the global RV (poly)-phenol compounds distribution was the reference to evaluate the significance of the differences observed in the NILs.

Anthocyanins were the major contributors to total phenolic content, representing 78% in 2012 and 88% in 2013 (Table CII.3). There was also a difference in anthocyanin accumulation in absolute terms between the two years (16 mg/10g and 28 mg/10g of fw respectively) (Table CII.1), emphasizing a possible environmental effect. The concentration of individual anthocyanins (pelargonidin- and cyanidin-3-glucoside, and pelargonidin-3-glucoside-malonate) was also higher in the 2013 harvest, although their proportional contribution to total anthocyanin accumulation was the same in both years (pelargonidin-3-glucoside, 52%, cyanidin-3-glucoside, 37% and pelargonidin-glucoside-malonate, 11% of total anthocyanins on average), indicating a genotypic control for a specific metabolite balance (Table CII.3).

Flavonols were the second major contributors to total phenolic content (11% in 2012 and 5% in 2013) (Table CII.3). The accumulation was significantly different in the two years (2.14 mg/10g and 1.58 mg/10g of fw in 2012 and 2013, respectively, Table CII.1) but, as in the case of anthocyanins, the individual flavonol ratios remained constant. Kaempferol-glucuronide is the major flavonol, accounting for an average 70% of total flavonols while the kaempferol-, kaempferol-coumaryl- and quercetin-glucoside accounted for around 10% of total flavonols each. The least abundant flavonol was quercetin-glucuronide, representing about 3% of total flavonols (Table CII.3). This finding is in agreement with results in *F.vesca* published by Muñoz *et al.* (2011) but contrasts with other publications where quercetin-glucuronide has been described as the most abundant flavonol in octoploid strawberry varieties (Aaby *et al.* 2012, Buendía *et al.* 2010)

The contribution of other phenolics (flavan-3-ols, flavanones, hydroxycinnamic acid derivatives and ellagic acid) to total phenolics concentration is minor (usually <5%, Table CII.3). The

average concentration of flavan-3-ols was slightly higher in 2012 (1.38 mg/10g of fw) than in 2013 (1.24 mg/10g of fw). The individual flavan-3-ols contributed equally to total flavan-3-ols accumulation (around 20% each) except the (epi) afzelechin-(epi) catechin dimers, which accounted for 3% (Table CII.3).

Table CII.2 Compound correlation between harvests. Correlation values were calculated using average genotype values (Supplemental table1) for both harvests. Asterisk indicates significance thresholds for correlation coefficients p-value <0.05 '*', p-value <0.01 '**', p-value <0.001 '***', p-value>0.05 '-'.

compound	corr	p-value
Pelargonidin-3-glucoside	0.71	***
Pelargonidin-3-glucoside-malonate	0.54	**
Cyanidin-3-glucoside	0.76	***
Kaempferol-glucoside	0.74	***
Kaempferol-glucuronide	0.50	*
Kaempferol-coumaryl-glucoside	0.42	*
Quercetin-glucoside	0.64	***
Quercetin-glucuronide	0.61	**
Procyanidin B1	0.74	***
Procyanidin B3	0.73	***
Catechin	0.71	***
(epi)catechin dimers iso1	0.71	***
(epi)catechin dimers iso2	0.73	***
(epi)afzelechin-(epi)catechin dimers	0.63	***
Eriodictyol iso1	0.70	***
Eriodictyol iso2	0.43	*
p-Coumaroyl-glucose ester	0.77	***
p-Coumaroyl-glucoside	0.61	**
Cinnamoyl-glucose ester	0.71	***
Feruloyl-glucose ester	0.75	***
Caffeoyl-glucose ester	0.25	-
Ellagic acid	0.67	***
DMHF-glucoside	0.00	-
Citric acid	0.03	-
Anthocyanins	0.68	***
Flavonols	0.21	-
Flavan-3-ols	0.73	***
Flavanones	0.67	***
Hydroxycinnamic acid deriv.	0.80	***
Total phenolics	0.69	***

Table CII.3 Relative phenolic accumulation. Percentage contribution of chemical families and ellagic acid (first column) to total phenolics content and percentage contribution of specific metabolites to their chemical families (second column) in 2012 and 2013 harvests for RV, average NIL collection and YW.

Chemical families	(poly)-phenolic compounds	RV		NIL collection		YW	
		2012	2013	2012	2013	2012	2013
Anthocyanins		0,78	0,88	0,80	0,91	0,00	0,00
	Pelargonidin-3-glucoside	0,54	0,46	0,51	0,47	-	-
	Pelargonidin-3-glucoside-malonate	0,12	0,12	0,10	0,10	-	-
	Cyanidin-3-glucoside	0,34	0,41	0,39	0,43	-	-
Flavonols		0,11	0,05	0,10	0,04	0,44	0,56
	Kaempferol-glucoside	0,07	0,08	0,08	0,08	0,06	0,08
	Kaempferol-glucuronide	0,70	0,70	0,66	0,68	0,70	0,80
	Kaempferol-coumaryl-glucoside	0,05	0,11	0,06	0,12	0,06	0,08
	Quercetin-glucoside	0,14	0,08	0,14	0,08	0,09	0,02
	Quercetin-glucuronide	0,04	0,03	0,06	0,04	0,09	0,02
Flavan-3-ols		0,07	0,04	0,05	0,02	0,47	0,33
	Procyanidin B1	0,23	0,18	0,21	0,18	0,22	0,22
	Procyanidin B3	0,15	0,18	0,16	0,18	0,15	0,16
	Catechin	0,21	0,23	0,22	0,24	0,23	0,22
	(epi)catechin dimers iso1	0,15	0,19	0,16	0,19	0,16	0,16
	(epi)catechin dimers iso2	0,24	0,19	0,22	0,19	0,23	0,22
	(epi)afzelechin-(epi)catechin dimers	0,03	0,03	0,03	0,03	0,02	0,02
Flavanones		0,03	0,01	0,02	0,01	0,00	0,00
	Eriodictyol iso1	0,78	0,79	0,78	0,82	0,62	1,00
	Eriodictyol iso2	0,22	0,21	0,22	0,18	0,38	0,00
HCAD		0,02	0,03	0,02	0,03	0,05	0,09
	p-Coumaroyl-glucose ester	0,26	0,05	0,24	0,07	0,15	0,05
	p-Coumaroyl-glucoside	0,04	0,01	0,07	0,03	0,01	0,01
	Cinnamoyl-glucose ester	0,46	0,87	0,53	0,85	0,34	0,68
	Feruloyl-glucose ester	0,13	0,04	0,09	0,03	0,35	0,19
	Caffeoyl-glucose ester	0,11	0,03	0,08	0,02	0,16	0,08
Ellagic acid		0,01	0,00	0,01	0,00	0,03	0,03

There were higher concentrations of flavanones eriodictyol isomers in 2012 (0.39 and 0.11 mg/10g of fw isomers 1 and 2 respectively) than in 2013 (0.18 and 0.05 mg/10g of fw isomers 1 and 2 respectively) (Table CII.1), but their relative abundance was stable 80:20 and favourable to isomer 1 (Table CII.3).

The average concentration of the hydroxycinnamic acid derivatives was higher in 2013 (0.87 mg/10g of fw) than in 2012 (0.31 mg/10g of fw). This was mainly due to cinnamoyl-

glucose ester that increased its concentration from 0.14 mg/10g of fw in 2012 to 0.76 mg/10g of fw in 2013. In contrast, the content and relative accumulation of the other hydroxycinnamic acid derivatives (p-coumaroyl-glucose ester and -glucoside, feruloyl- and caffeoyl-glucose ester) was reduced or maintained (Table CII.1) in 2013 compared with 2012 (Table CII.3).

Ellagic acid was a minor component, more abundant in the 2012 (0.10 mg/10g of fw) than in 2013 harvest (0.06 mg/10g of fw). HDMF-glucoside was more abundant in the 2013 harvest (0.80 mg/10g of fw) than in 2012 (0.17 mg/10g of fw). Finally, although citric acid is not a phenolic compound, it was among the most abundant metabolites. Its absolute average concentration significantly varied between years (10.09 mg/10g of fw in 2012 and 5.97 mg/10g of fw in 2013) (Table CII.1).

The phenolic compounds accumulation profile in the yellow wild strawberry, out-group *F.vesca* YW, differed compared with RV or the NIL collection (Table CII.3), characterized by the absence of anthocyanins, flavanones and a low content in p-coumaroyls and cinnamoyl-glucose (Figure 1).

Concentrations of all detected metabolites (except p-coumaroyl-glucoside) varied between years in RV and in the overall NIL collection, highlighting an environmental factor affecting accumulation (Table CII.1). However, the relative contribution to chemical families remained constant for anthocyanins, flavonols, flavan-3-ols and flavanones.

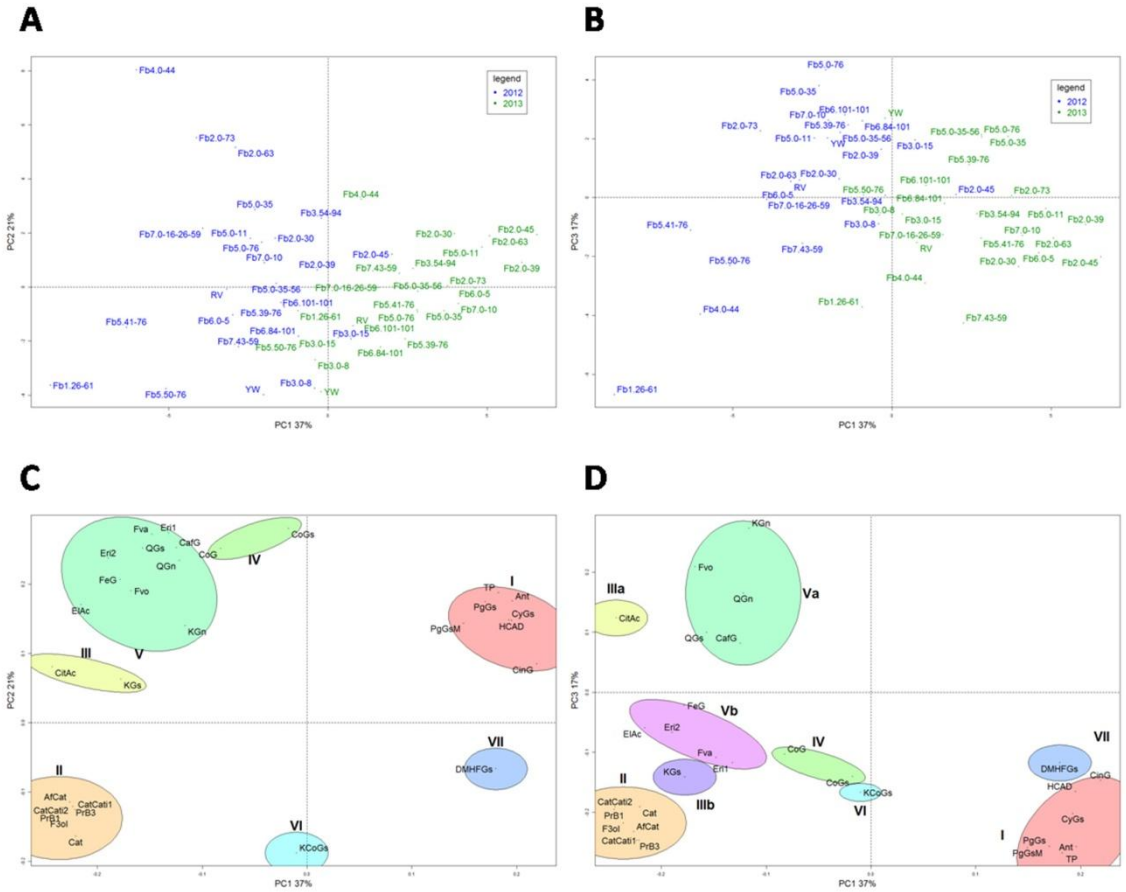
Variance of metabolite among the *F. vesca* NIL collection

To further understand the structure of the phenolic data set, we carried out a principal component analysis (PCA). Genotypes were distributed in the first three components (PC1, PC2 and PC3 explaining 37%, 21% and 17% of the variance, respectively) without forming defined groups, gathering around the axis and the recurrent parental (RV). This is in accordance with the isogenicity of the NIL collection, as they share most of their genetic background (Urrutia *et al.* 2015). However, samples tended to separate according to the different harvest years, which points to an environmental effect (Figure 2A and 2B). Despite the tendency for samples to split by the harvest year, genotypes had a very high correlation between harvests (0.95 on average), demonstrating that the year to year accumulation pattern was similar within each NIL. Only three lines had a correlation data below 0.9, NILs Fb**5**:0-35 and Fb**5**:0-76 with 0.85 and the NIL Fb**5**:41-76 with a minor value of 0.78.

Compounds contributing to samples differentiation formed 7 clear groups according to principal components 1 and 2 (Figure 2C). Group I included anthocyanins and cinnamoyl-glucose; group II contained flavan-3-ols; group III was formed by citric acid and kaempferol glucoside; groups IV, V and VI gathered hydroxycinnamic acid derivatives, flavanones most flavonols and ellagic and citric acids. There were also two additional single-compound groups, the kaempferol-coumaryl-glucoside (group VI) and DMHF glucoside (group VII). Both groups III and V were split into two sub-groups attending to PC3 (Figure 2D). Citric acid and kaempferol glucoside formed subgroups IIIa and IIIb respectively. Quercetin derivatives, kaempferol glucuronide and caffeoyl glucose formed subgroup Va, while eriodictyol isomers, ellagic acid and feruloyl glucose ester formed subgroup Vb.

According to the scores plot (Figure 2A and 2B), the 2013 harvest was globally characterized by the accumulation of compounds from group I and VII, and the 2012 harvest by compounds from group IIIa and Va. Compounds from groups I, IIIa, Va and VII were the most conditioned by the environment.

Figure CII. 2 PCA scores and loadings plot: PCA scores plot for PC1 vs. PC2 (A) and PC1 vs. PC3 (B). Genotypes represented in blue correspond to the 2012 harvest and in green to the 2013 harvest. PCA loadings plot for PC1 vs. PC2 (C) and PC1 vs. PC3 (D).



Genotypic and environmental effect over metabolite accumulation

To examine if the accumulation of metabolites was affected by the genotype (G) and the environment (E), we further evaluated their significance and quantified their effect and interaction on the accumulation of phenolics. The analysis of variance (ANOVA), following the model G+E+GxE (harvests in different years are considered different environments), revealed that the metabolites were significantly affected by both factors and their interaction (p-value<0.05), except for p-coumaroyl-glucoside that did not have significant variation between years (Supplemental Table CII.2). Although both factors contributed significantly to variance, their influence differed between metabolites (Table CII.4). Those most influenced by environment were citric acid and quercetin-glucoside while the genotype had most effect on p-coumaroyl-glucose ester and –glucoside (Figure 3).

Figure CII. 3 Interaction plots Individual plots for p-coumaroyl-glucoside, citric acid and chemical-families representing average values (mg/10g of fw) for each line in the NIL collection for both harvests (blue, 2012; green, 2013). Parallel lines indicate a major effect of the genotype. Non parallel lines reflect the effect of the environment, and crosses of the lines for specific genotypes reflect the effect of GxE interaction.

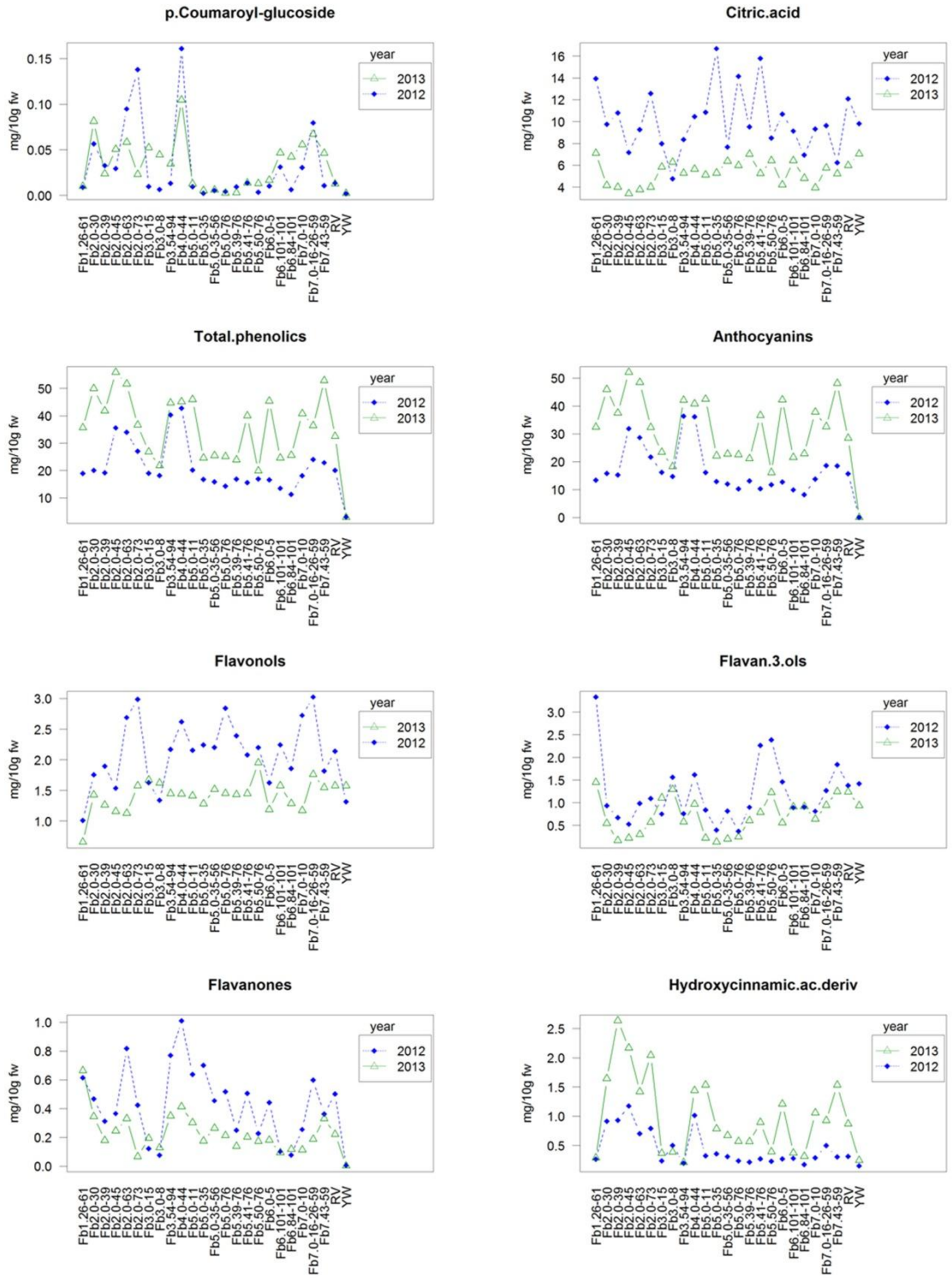


Table CII.4 G+E+GxE Omega squared values (ω2). Percentage of the variance accounted for by each of the studied factors (G, E and GxE), calculated using the omega squared values from the ANOVA results.

compound	ω^2G (%)	ω^2E (%)	ω^2GxE (%)
Pelargonidin-3-glucoside	32,38	15,24	3,61
Pelargonidin-3-glucoside-malonate	35,80	12,38	9,53
Cyanidin-3-glucoside	34,60	19,21	3,32
Kaempferol-glucoside	35,23	7,76	5,88
Kaempferol-glucuronide	17,68	19,87	7,78
Kaempferol-coumaryl-glucoside	28,97	6,40	8,58
Quercetin-glucoside	25,12	31,68	18,66
Quercetin-glucuronide	28,39	14,29	10,59
Procyanidin B1	38,62	16,39	6,86
Procyanidin B3	41,76	4,80	6,32
Catechin	41,04	7,80	7,55
(epi)catechin dimers iso1	41,58	4,19	6,55
(epi)catechin dimers iso2	37,80	16,22	6,91
(epi)afzelechin-(epi)catechin dimers	38,54	8,71	8,22
Eriodictyol iso1	37,63	11,71	9,50
Eriodictyol iso2	19,45	23,89	13,03
p-Coumaroyl-glucose ester	49,42	4,69	11,31
p-Coumaroyl-glucoside	43,92	0,00	11,94
Cinnamoyl-glucose ester	30,69	17,60	9,04
Feruloyl-glucose ester	26,11	11,21	2,20
Caffeoyl-glucose ester	19,31	10,16	16,62
Ellagic acid	19,93	25,48	10,59
DMHF-glucoside	4,53	15,03	4,41
Citric acid	7,37	35,36	9,76
Anthocyanins	29,89	19,69	3,90
Flavonols	10,76	23,84	10,55
Flavan-3-ols	33,94	15,00	10,00
Flavanones	41,00	9,90	6,90
Hydroxycinnamic.ac.deriv	37,80	14,09	6,73
Total phenolics	32,49	9,99	4,14

The influence of genotype on total phenolics accumulation was high, but that of environmental conditions low (Table CII.4). Between family compounds, only flavonols were highly influenced by environment, confirming the results of the principal component analysis (group Va) and the non-significant correlation data of this family. The other chemical families were more influenced by the genotype (Figure CII. 3, Table CII.4). Attending to specific metabolites, variations on anthocyanins accumulation were highly influenced by the genotype (32-36%) while the effects of the environment were mild (12-19%) (Supplemental Figure CII. 2). Flavan-3-ols accumulations were highly affected by the genotype (38-42%) and the effect of the environment ranged from low (4-9% for procyanidin B3, catechin, (epi) catechin dimer iso1

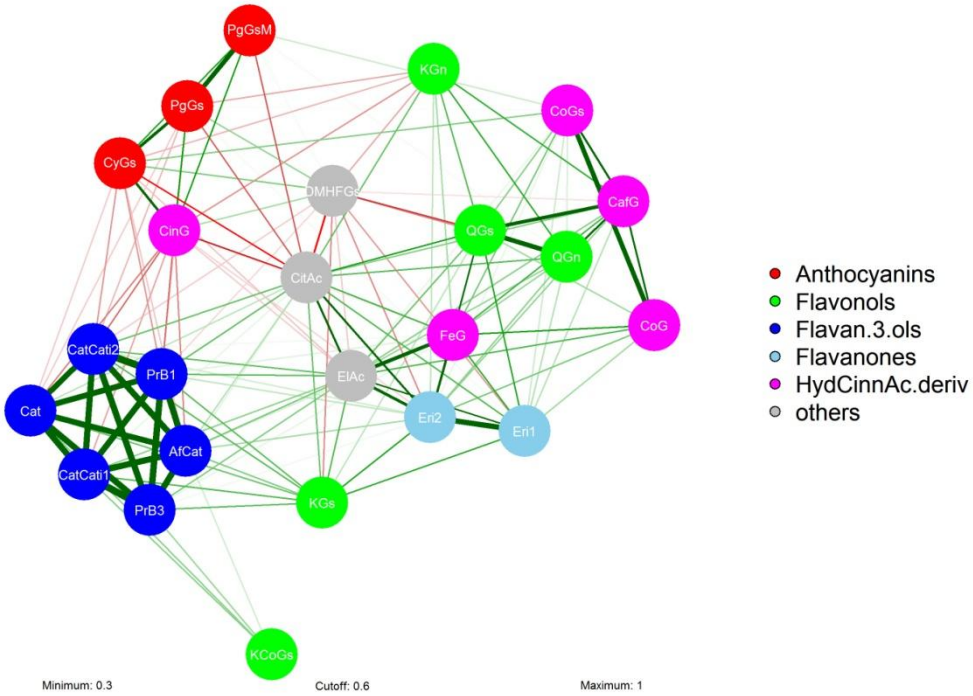
and (epi) afzelechin-(epi) catechin) to mild (16% for procyanidin B1 and (epi)catechin dimer iso2). The GxE effect was low (<10%) for all anthocyanins and flavan-3-ols. The effect of the three factors was very different between the flavonols, with kaempferol-glucoside and kaempferol-coumaroyl-glucoside mainly influenced by the genotype (35 and 29% respectively) with little effect of the environment, while kaempferol-glucuronide was equally affected by the genotype (18%) and the environment (20%). The influence of GxE was low (<10%) for all three kaempferol derivatives. In contrast, quercetin-glucoside and quercetin-glucuronide were both mildly affected by the GxE factor (19 and 11% respectively) and more by the genotype (25 and 28% respectively), but quercetin glucoside was more influenced by the environment (32%) than quercetin glucuronide (14%). There was also a great difference in how the factors affected the accumulation of flavanones. Eriodictyol isomer 1 was more influenced by the genotype (G: 38%, E: 12%) and isomer 2 more by the environment (E: 24%, G: 19%). The effect of GxE was similar in both cases (9 and 13% respectively). Various behaviours were also observed among hydroxycinnamic acid derivatives. Both p-coumaroyl-glucose ester and glucoside were highly affected by the genotype (49 and 44% respectively), while p-coumaroyl glucoside was independent of the environment and p-coumaroyl glucose very little affected by it (5%). Cinnamoyl- and feruloyl-glucose esters were both mostly influenced by the genotype (31 and 26% respectively) although there was some influence of the environment (18 and 11% respectively). The influence of all three factors on the level of caffeoyl-glucose ester was mild, with the genotype and the GxE accounting for most of the variability (19 and 17% respectively). Citric acid and DMHF-glucoside accumulation were more affected by the environment (35 and 15% respectively) than by the genotype (7 and 5% respectively), although DMHF-glucoside variation was barely affected by either of these factors.

Metabolite's relations

As there is an important genetic factor in phenolics accumulation in strawberry fruits, and the detailed regulation mechanism of plant secondary metabolism has not yet been fully understood, we wanted to study the relationships between these metabolites in order to detect possible co-regulation. The correlation between metabolites of the independent harvests was studied (Supplemental Table CII.3) and represented as a correlation network analysis (CNA), (Figure CII. 4). Correlation between metabolites was revealed to be higher and more stable between compounds belonging to the same chemical family or sharing known biological pathways, such as correlations within anthocyanins, flavan-3-ols or flavanones that were positive and strong between all the compounds in the chemical families. Flavan-3-ols shared correlations ranging from 0.88 to 1, the correlation coefficient of eriodictyol isomers was 0.86 on average and, in the case of anthocyanins, correlation was stronger between the pelargonidins (avg. 0.81) and weaker between PgGsM and CyGs. However, HCAD and flavonols did not show clear a positive relation between all their metabolites. Among HCAD, cinnamoyl-glucose ester did not correlate with the rest of the compounds in the chemical family, while there was positive correlation between p-coumaroyls, caffeoyl- and feruloyl-glucose ester (although correlation between FeG and the others was very weak). Especially strong correlations were between CoG and CoGs (avg. 0.88), and between CoG and CafG (avg. 0.60). The flavonols family established the weakest relations between its compounds, as there was only significant correlation in both harvests between quercetins (avg. 0.72), not between any of the other flavonols.

We also found stable relations between compounds belonging to different chemical families (Figure CII. 4). All flavan-3-ols were positively correlated with kaempferol-glucoside and kaempferol-coumaryl-glucoside (these correlations were significant in both years for KCoGs), and kaempferol-glucoside was also positively linked to flavanones, ellagic and citric acid. Anthocyanins correlated positively with cinnamoyl-glucose ester (correlation between CyGs and CinG was significant both years) and negatively with citric acid. There was a positive correlation between HCAD caffeoyl-glucose ester and the quercetin derivatives, while feruloyl-glucose ester was linked to eriodictyol isomer 2 and ellagic acid (Supplemental Table CII.3).

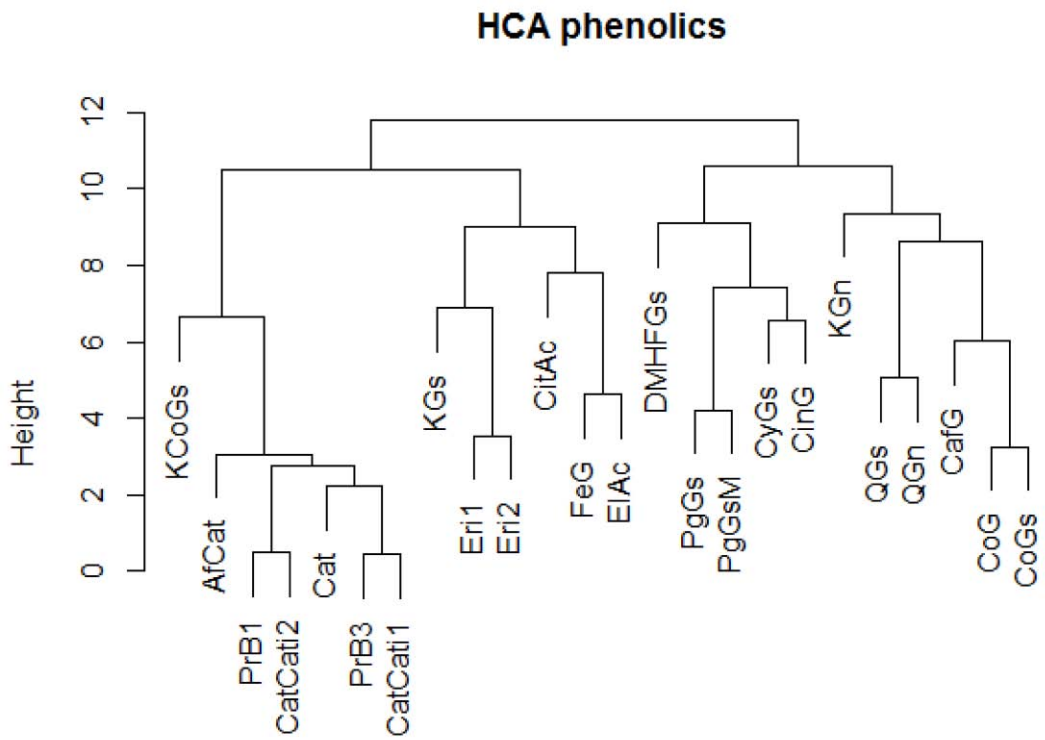
Figure CII. 4 Cluster network analysis (CNA). Graphical representation of the correlation between phenolic compounds. Metabolites are represented as nodes, coloured according to chemical families as specified by the legend. Positive (green) and negative (red) correlations with absolute values >|0.3| are represented as links between the nodes. Links representing absolute correlations >|0.6| are wider the stronger they are and have the maximum colour saturation. Absolute correlation <|0.6| are lighter the weaker they are and have the narrowest width.



According to hierarchical clustering analysis (HCA), metabolites grouped in pairs in four main clusters. The closest compounds forming couples were those that shared high positive correlations, clustering similarly to principal component analysis (Figure CII. 5). In the first branch all the flavan-3-ols clustered together with the KCoGs. The second branch clustered both eriodictyols together with KGs and FeG with ellagic and citric acids. The third branch clustered anthocyanins, cinnamoyl-glucose ester and DMHF-glucoside. In the fourth branch the HCAD coumaroyls and caffeoyl-glucose ester clustered together with quercetin derivatives

and with KGn (Figure CII. 5). These clusters reflect that seen with the PCA, CNA and support mQTL results.

Figure CII. 5 Hierarchical clustering (HCA). Phenolic compounds grouped by accumulation similarity (distance matrix = Euclidean, hierarchical clustering = complete linkage).



Fine genotyping of NILs by IStraw90® Axiom Array

Genotyping results with the IStraw90® (Bassil & Davis *et al.* 2015) revealed that the vast majority of SNPs (79922 from the initial 95063 SNPs, 84.07%) were homozygous in all analyzed lines and therefore not suitable for genotyping the NIL collection. From the remaining 15141 SNPs, 3327 (3.50%) did not surpass quality criteria (<95% call rate) and were discarded. Additionally 5747 (6.04%) were classified as off target variants and probably represent allelic variants not considered by the SNP probes and 3079 (3.24%) did not present both homozygous alleles and were not considered for the analysis. Finally non consistent duplicates were removed and the final set of SNPs was reduced to 1510 SNPs.

Genotyping results allowed a more precise drawing of the introgression's boundaries (Supplementary Table CII.5) and revealed that most NIL were "clean" and did not harbour (or harbour only short) undesired introgressions from *F. bucharica* in the genetic background of *F. vesca*. Several discrepancies between the expected introgressions according to the genetic map (Urrutia *et al.* 2015) and the obtained genotypes (Supplementary Table CII.5) suggested

that there could be some miss-assemblies in the *F. vesca* reference genome v1.1, but comparing with the new v.2.0 assembly (Tennessen *et al.* 2014) discrepancies disappear.

Metabolic QTL analysis

Phenolic metabolites accumulation segregates among the NIL collection. The profile of every NIL showed a different accumulation pattern compared with the RV profile (Figure CII. 3, Supplemental Figure CII. 2). Differences in the introgressed regions are presumably caused by the *F. bucharica* alleles, indicating that phenolic accumulation is to some extent genetically controlled. This variability is reported in relative terms (Figure CII. 1) and in absolute terms (Supplemental Table CII.1). Significant differences in phenolic compounds accumulation between any of the NILs and the control RV were evaluated by independent t-test corrected for false discovery rate (FDR). We accepted these significant differences as mQTL for a specific metabolite when all the studied lines harbouring the same introgression had a significant effect (p-value after correction <0.05) on the accumulation of the metabolite in the same direction. Additionally, to confirm the effect of the QTL and the percentage variability explained by it, we performed a QTL interval mapping analysis for both harvests independently (Table CII. 5). In total, we mapped 100 mQTL for the 24 different compounds, the five chemical families and total polyphenols distributed among the seven linkage groups. In 39 of the mQTL the alleles of the *F. bucharica* introgressions caused over-accumulation of the compounds compared with RV, and in the other 61 mQTL the exotic introgressed regions led to a reduction of the level of the metabolites. Among all the mQTL mapped, 76 were stable in two harvests and 24 only appear in one harvest. There were 49 strong mQTLs (LOD score >1.8, observed variability>20%) that have been graphically represented in Figure CII. 6, 25 of which were stable in both harvests and cited as major QTL.

Anthocyanins were the most abundant phenolics in absolute values, and pelargonidin-3-glucoside the most abundant anthocyanin in almost all the genotypes of both years, followed by cyanidin-3-glucoside. Pelargonidin-3-glucoside-malonate was the least abundant. Lines with introgressions in LG2 (Fb**2**:0-45, Fb**2**:0-63), LG3 (Fb**3**:0-8, Fb**3**:0-15) and LG6 (Fb**6**:84-101, Fb**6**:101-101) had a different balance between anthocyanins, accumulating similar levels of both pelargonidin-3-glucoside and cyanidin-3-glucoside, while the line Fb**2**:0-73 preferentially accumulated cyanidin-3-glucoside (Supplemental Table CII. 1). QTL mapping revealed a single major mQTL over-accumulating cyanidin-3-glucoside at region LG2:45-63 cM that explained 25-69% of the variability (Table CII. 5), and another major mQTL for the under-accumulation of pelargonidin-3-glucoside-malonate at LG5:41-50, that explained 9-27% of the variability. There were also two strong positive mQTLs explaining 28 and 52% of the variability for pelargonidin-3-glucoside and pelargonidin-3-glucoside-malonate, respectively, at LG4:9-20 but this appeared only in the 2012 harvest. Other three minor mQTL were mapped, two at region LG2:63-73 cM under-accumulating pelargonidin-3-glucoside and -3-glucoside-malonate, and one mQTL at region LG6:101-101 cM for the under-accumulation of pelargonidin-3-glucoside malonate (Table CII.5). Total anthocyanins were mapped to three different one-year mQTL, one overlapping with the mQTLs over-accumulating pelargonidins in 2012 at region LG4:9-20, one at the region LG5:50-76 next to the mQTL for under-accumulation of pelargonidin-3-

glucoside-malonate, and one that did not co-localize with mQTLs for individual anthocyanins at LG3:54-94. This is probably due to the fact that there was a slight increase in all individual anthocyanins in LG3:54-94 (Figure CII. 1), not sufficient to map a significant mQTL but the addition of these small increments led to a mQTL for total anthocyanins.

Distribution of total flavonols was fairly homogeneous among the NIL collection, except for a significant decline in NIL Fb**1**:26-61. Regarding specific flavonols, all genotypes had high relative values of kaempferol-glucuronide (the most abundant flavonol in absolute terms) and low relative values of kaempferol-glucoside, in contrast to the behaviour of NIL Fb**1**:26-61 (Figure CII. 1). Absolute flavonol values confirmed this trend, as Fb**1**:26-61 had significantly low flavonol accumulation (1.01 and 0.66 mg/10g of fw in 2012 and 2013, respectively) compared with the control mean of RV (2.14‰ and 1.58 mg/10g of fw in 2012 and 2013, respectively). This is mainly due to a significant decrease in kaempferol-glucuronide concentration that was not compensated by the increased concentration of kaempferol-glucoside (Supplemental Table CII.1). This resulted in a different ratio of metabolites for this specific genotype. According to significance tests, region LG1:26-61 cM gathered the major stable mQTL for the over-accumulation of kaempferol-glucoside (23-41% of explained variability), the under-accumulation of kaempferol-glucuronide (41-50% of explained variability) and total flavonols (16-34% of explained variability) (Figure CII. 6). Positive major mQTL were also mapped for quercetins at LG2:63-73. These mQTLs explained between 5 to 50% of the observed variability for quercetin-glucoside and 23-35% for quercetin-glucuronide. In addition, a major mQTL (18-38% of the explained variability) was mapped for the over-accumulation of kaempferol-coumaryl-glucoside at region LG7:43-59. Other minor mQTLs were mapped for individual flavonols (Table CII.5).

All identified flavan-3-ols were significantly under-accumulated in NILs with introgressions at the beginning of LG5 (Fb**5**:0-11, Fb**5**:0-35, Fb**5**:0-35-56, Fb**5**:0-76) compared with RV in both years. In addition, NIL Fb**1**:26-61 and lines harbouring introgressions in the central region of LG7 (LG7:10-26 cM) accumulated higher concentrations of most of the flavan-3-ols (Supplemental Table CII. 1), but this increase was significant with respect to RV only in 2012 (Table CII.5). Therefore, major mQTL for total and specific flavan-3-ols were mapped to region LG5:0-11 cM. This mQTL explained between 12 to 32% of the variability of total flavan-3-ols. One-year mQTL were mapped at regions LG1:26-61 and LG7:10-26. Additionally, 8-22% of the observed variability of afzelechin-catechin dimers was explained by a mQTL located at LG2:0-30 cM.

Flavanones (eriodictyol isomers) where under-accumulated in NILs Fb**3**:0-8, Fb**3**:0-15, Fb**6**:84-101 and Fb**6**:101-101. These last two lines, with introgressions in LG6, accumulated significantly lower levels of eriodictyol isomer 1 resulting in a different ratio of isomers (this was more evident in 2012) (Supplemental Table CII. 1). Considering the significance tests, we mapped minor mQTLs (<20% of explained variance) for under-accumulation of total flavanones and eriodictyol isomer 2 to region LG3:0-8 cM and eriodictiyol isomer 1 to region LG6:101-101 cM. Strong mQTLs were stable in only one harvest for eriodictyol isomer 2 at region LG1:26-61 and LG4:9-20 (Figure CII. 6).

Relative accumulation of hydroxycinnamic acid derivatives was higher in lines with introgressions in LG2 and LG4 (Figure CII. 1). Lines with introgressions in LG2 presented significantly higher relative and absolute concentrations of cinnamoyl-glucose ester, the most important contributor to total hydroxycinnamic acid derivatives (Supplemental Figure CII. 2). Consequently, we could map positive major stable QTLs for cinnamoyl-glucose ester (37-52% of variability explained) and total HCAD (40-58% of variability explained) to region LG2:0-30 cM (Figure CII. 6). Lines harbouring introgressions at the beginning of LG4 accumulated higher levels of p-coumaroyl-glucose ester, p-coumaroyl-glucoside and feruloyl-glucoside, and major QTL were mapped for these compounds at region LG4:9-20 cM (Figure CII. 6). In addition, a strong one-year QTL was mapped for caffeoyl-glucoside in the same region. Despite these compounds being minor contributors to the total hydroxycinnamic acid-derivatives accumulation, the sum of the small increments was a significantly higher overall concentration of HCAD mapped as a minor mQTL to region LG4:9-20 cM (Table CII.5). The opposite behaviour was shown by lines with introgressions in LG6 (Fb**6**:84-101 and Fb**6**:101-101) where the significant decrease in cinnamoyl-glucose ester accumulation led to a significant under-accumulation of the total HCAD (Table CII.5) although the rest of the HCAD retained average values.

The distribution of ellagic acid was fairly homogeneous along the NIL collection, except for significant increases in lines harbouring introgressions at LG1 and LG4 (Figure CII. 1, Supplemental Table CII.1). These led to the mapping of two positive major mQTLs at regions LG1:26-61 cM and LG4:9-20 cM that contributed to 29-50% and 15-22% of the observed variability, respectively.

Finally, stable QTLs were not mapped for total (poly)-phenols accumulation but one-year mQTLs co-located with mQTLs detected for anthocyanins accumulation. This was expected considering that anthocyanins account for around 80% of the total (poly)-phenols.

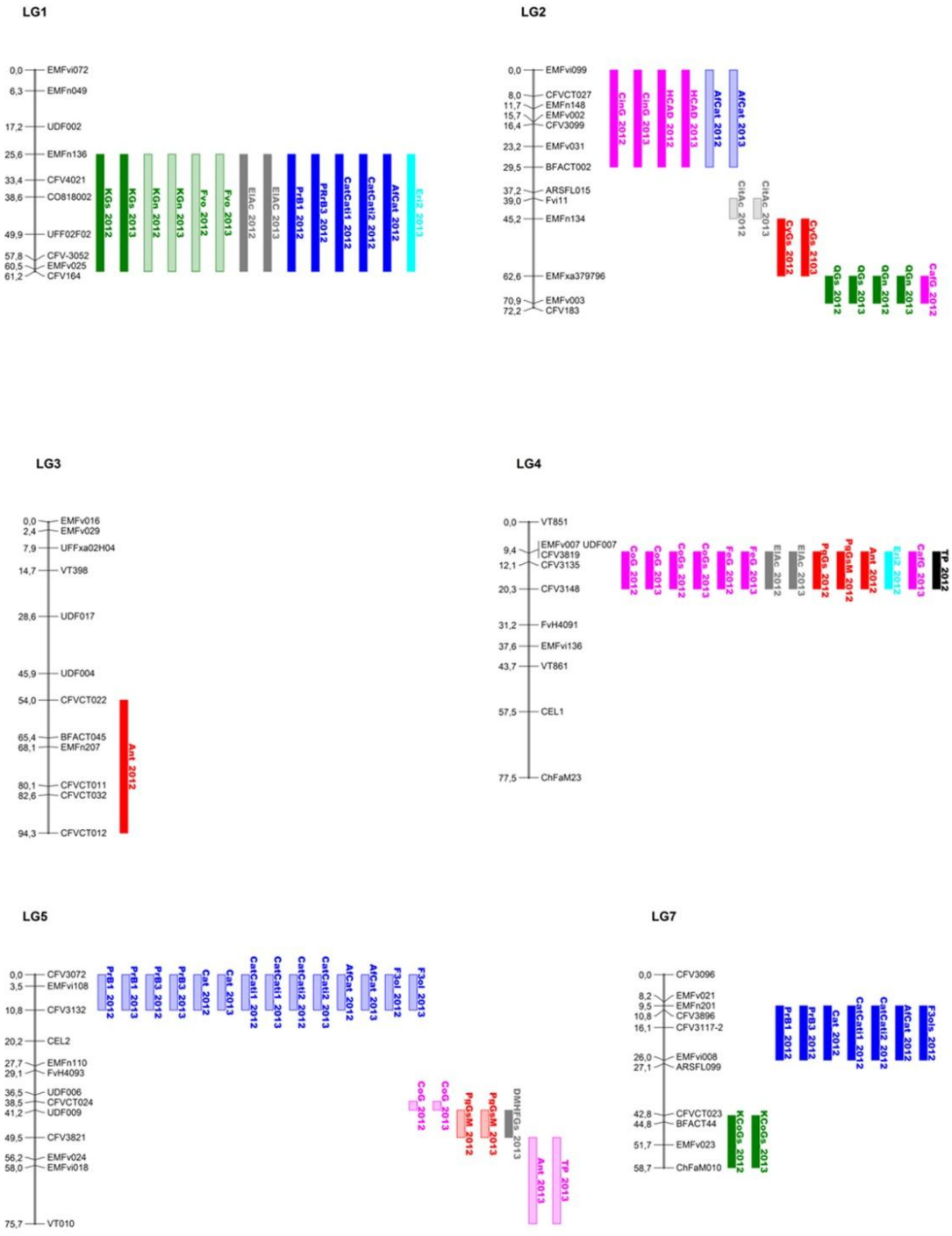
Table CII.5 mQTL for phenolic compounds in *F. vesca*. All detected mQTL are shown. The under or over accumulation effect of the mQTL is indicated by (-) or (+) symbol respectively. The genetic position of the mQTL expressed in cM and the NIL with the shorter introgression harboring the mQTL is given. Finally, the significance test results (t-test p-value after correction and LOD score), the % of the variability explained by the QTL and the stability (1 or 2 harvests) are indicated.

	compound	direction	qtl (cM)	smallest NIL	t-test (corrected p-value)	LOD	%expl. Var.	stable
Anthocyanin	Pelargonidin-3-glucoside	-	2:63-73	Fb2.0-73	<0,05	<1,8	3-25%	2
	Pelargonidin-3-glucoside	+	4:9-20	Fb4:0-20	<0,05	2,47	28%	1
	Pelargonidin-3-glucoside-malonate	-	2:63-73	Fb2.0-73	<0,05	<1,8	8-10%	2
	Pelargonidin-3-glucoside-malonate	-	5:41-50	Fb5.41-76	<0,05	2,38	9-27%	2
	Pelargonidin-3-glucoside-malonate	-	6:101-101	Fb6.101-101	<0,05	<1,8	1-10%	2
	Pelargonidin-3-glucoside-malonate	+	4:9-20	Fb4:0-20	<0,05	5,59	52%	1
	Cyanidin-3-glucoside	+	2:45-63	Fb2.39-63	<0,05	7,2	25-69%	2
Flavonol	Kaempferol-glucoside	+	1:26-61	Fb1.26-61	<0,05	3,91	23-41%	2
	Kaempferol-glucoside	+	5:50-76	Fb5.50-76	<0,05	<1,8	3-4%	2
	Kaempferol-glucoside	-	3:0-8	Fb3.0-8	<0,05	<1,8	10-20%	2
	Kaempferol-glucoside	-	6:84-101	Fb6.84-101	<0,05	<1,8	6-8%	2
	Kaempferol-glucuronide	-	1:26-61	Fb1.26-61	<0,05	5,12	41-50%	2
	Kaempferol-glucuronide	-	2:39-45	Fb2.0-45	<0,05	<1,8	3%	2
	Kaempferol-coumaryl-glucoside	+	7:43-59	Fb7:43-59	<0,05	3,57	18-38%	2
	Kaempferol-coumaryl-glucoside	+	5:50-76	Fb5.50-76	<0,05	<1,8	4%	2
	Quercetin-glucoside	+	2:63-73	Fb2.0-73	<0,05	5,45	4-52%	2
	Quercetin-glucoside	-	1:26-61	Fb1.26-61	<0,05	<1,8	1-5%	2
	Quercetin-glucoside	-	6:101-101	Fb6.101-101	<0,05	<1,8	5-14%	2
	Quercetin-glucuronide	+	2:63-73	Fb2:0-73	<0,05	3,22	23-35%	2
Flavan-3-ol	Procyanidin B1	-	5:0-11	Fb5.0-11	<0,05	2,48	9-28%	2
	Procyanidin B1	-	2:0-30	Fb2.0-30	<0,05	<1,8	7-16%	2
	Procyanidin B1	-	5:39-41	Fb5.39-76	<0,05	<1,8	4-6%	2
	Procyanidin B1	-	3:54-94	Fb3.54-94	<0,05	<1,8	1-2%	2
	Procyanidin B1	+	1:26-61	Fb1.26-61	<0,05	2,46	30%	1
	Procyanidin B1	+	7:10-26	Fb7:0-27	<0,05	2,61	30%	1
	Procyanidin B3	-	5:0-11	Fb5.0-11	<0,05	2,82	12-31%	2
	Procyanidin B3	-	2:0-30	Fb2.0-30	<0,05	<1,8	8-14%	2
	Procyanidin B3	-	3:54-94	Fb3.54-94	<0,05	<1,8	2%	2
	Procyanidin B3	+	1:26-61	Fb1:26-61	<0,05	2,37	27%	1
	Procyanidin B3	+	7:10-26	Fb7:0-27	<0,05	1,81	22%	1
	Catechin	-	5:0-11	Fb5.0-11	<0,05	3,05	16-33%	2
	Catechin	-	2:0-30	Fb2.0-30	<0,05	<1,8	9-15%	2
	Catechin	-	3:54-94	Fb3.54-94	<0,05	<1,8	1-3%	2
	Catechin	+	7:10-26	Fb7:0-27	<0,05	2,39	28%	1
	(epi)catechin dimers iso1	-	5:0-11	Fb5.0-11	<0,05	2,83	12-31%	2

	compound	direction	qtl (cM)	smallest NIL	t-test (corrected p-value)	LOD	%expl. Var.	stable
	(epi)catechin dimers iso1	-	2:0-30	Fb2.0-30	<0,05	<1,8	7-13%	2
	(epi)catechin dimers iso1	-	3:54-94	Fb3.54-94	<0,05	<1,8	2%	2
	(epi)catechin dimers iso1	-	7:0-10	Fb7.0-10	<0,05	<1,8	1%	2
	(epi)catechin dimers iso1	+	1:26-61	Fb1:26-61	<0,05	2,59	27%	1
	(epi)catechin dimers iso1	+	7:10-26	Fb7:0-27	<0,05	1,8	21%	1
	(epi)catechin dimers iso2	-	5:0-11	Fb5.0-11	<0,05	2,59	9-29%	2
	(epi)catechin dimers iso2	-	2:30-39	Fb2.0-39	<0,05	<1,8	7-15%	2
	(epi)catechin dimers iso2	-	5:39-41	Fb5.39-76	<0,05	<1,8	4-6%	2
	(epi)catechin dimers iso2	-	3:54-94	Fb3.54-94	<0,05	<1,8	1-2%	2
	(epi)catechin dimers iso2	-	7:0-10	Fb7.0-10	<0,05	<1,8	1%	2
	(epi)catechin dimers iso2	+	1:26-61	Fb1:26-61	<0,05	2,75	31%	1
	(epi)catechin dimers iso2	+	7:10-26	Fb7:0-27	<0,05	2,6	30%	1
	(epi)afzelechin-(epi)catechin dimers	-	2:0-30	Fb2.0-30	<0,05	1,86	8-22%	2
	(epi)afzelechin-(epi)catechin dimers	-	5:0-11	Fb5.0-11	<0,05	1,86	3-22%	2
	(epi)afzelechin-(epi)catechin dimers	-	3:54-94	Fb3.54-94	<0,05	<1,8	1-2%	2
	(epi)afzelechin-(epi)catechin dimers	-	5:39-41	Fb5.39-76	<0,05	<1,8	1-2%	2
	(epi)afzelechin-(epi)catechin dimers	-	7:0-10	Fb7.0-10	<0,05	<1,8	3%	2
	(epi)afzelechin-(epi)catechin dimers	+	1:26-61	Fb1:26-61	<0,05	1,98	22%	1
	(epi)afzelechin-(epi)catechin dimers	+	7:10-26	Fb7:0-27	<0,05	2,01	24%	1
Flavanone	Eriodictyol iso1	-	6:101-101	Fb6.101-101	<0,05	<1,8	11-18%	2
	Eriodictyol iso1	-	5:39-41	Fb5.39-76	<0,05	<1,8	1-10%	2
	Eriodictyol iso1	-	7:0-10	Fb7.0-10	<0,05	<1,8	1-5%	2
	Eriodictyol iso2	-	3:0-8	Fb3.0-8	<0,05	<1,8	5-11%	2
	Eriodictyol iso2	-	7:0-10	Fb7.0-10	<0,05	<1,8	1-5%	2
	Eriodictyol iso2	+	4:9-20	Fb4:0-20	<0,05	3,87	41%	1
	Eriodictyol iso2	+	1:26-61	Fb1:26-61	<0,05	2,44	28%	1
HCAD	p-Coumaroyl-glucose ester	+	4:9-20	Fb4:0-20	<0,05	3,77	28-40%	2
	p-Coumaroyl-glucose ester	-	5:39-41	Fb5.39-76	<0,05	1,86	13-22%	2
	p-Coumaroyl-glucose ester	-	5:11-35	Fb5.0-35	<0,05	<1,8	6-19%	2
	p-Coumaroyl-glucose ester	+	2:0-30	Fb2.0-30	<0,05	<1,8	4-13%	2
	p-Coumaroyl-glucoside ester	+	4:9-20	Fb4:0-20	<0,05	3,31	29-36%	2
	p-Coumaroyl-glucoside ester	+	2:0-30	Fb2.0-30	<0,05	<1,8	6-15%	2
	Cinnamoyl-glucose ester	+	2:0-30	Fb2.0-30	<0,05	5,5	37-52%	2
	Cinnamoyl-glucose ester	-	3:54-94	Fb3.54-94	<0,05	<1,8	2-3%	2
	Cinnamoyl-glucose ester	-	6:101-101	Fb6.101-101	<0,05	<1,8	6-8%	2
	Feruloyl-glucose ester	+	4:9-20	Fb4:0-20	<0,05	9,26	60-71%	2
	Caffeoyl-glucose ester	+	4:9-20	Fb4:0-20	<0,05	9,96	74%	1
	Caffeoyl-glucose ester	+	2:63-73	Fb2:0-73	<0,05	5,5	52%	1
Others	Ellagic acid	+	4:9-20	Fb4:0-20	<0,05	5,11	29-50%	2
	Ellagic acid	+	1:26-61	Fb1.26-61	<0,05	1,87	15-22%	2
	Ellagic acid	-	7:43-59	Fb7.43-59	<0,05	<1,8	1-5%	2
	DMHF-glucoside	+	5:39-50	Fb5:39-76	<0,05	2,67	30%	1

	compound	direction	qtl (cM)	smallest NIL	t-test (corrected p-value)	LOD	%expl. Var.	stable
	Citric acid	-	2:39-45	Fb2.0-45	<0,05	3,09	34%	2
Chemical families	Anthocyanins	+	3:54-94	Fb3:54-94	<0,05	1,99	24%	1
	Anthocyanins	+	4:9-20	Fb4:0-20	<0,05	1,94	23%	1
	Anthocyanins	-	5:50-76	Fb5:50-76	<0,05	1,88	22%	1
	Flavonols	-	1:26-61	Fb1.26-61	<0,05	3,11	16-34%	2
	Flavonols	-	2:39-45	Fb2.0-45	<0,05	<1,8	2-3%	2
	Flavan-3-ols	-	5:0-11	Fb5.0-11	<0,05	2,92	12-32%	2
	Flavan-3-ols	-	2:0-30	Fb2.0-30	<0,05	<1,8	8-16%	2
	Flavan-3-ols	-	3:54-94	Fb3.54-94	<0,05	<1,8	1-2%	2
	Flavan-3-ols	-	7:0-10	Fb7.0-10	<0,05	<1,8	1%	2
	Flavan-3-ols	-	5:39-41	Fb5.39-76	<0,05	<1,8	3-5%	2
	Flavan-3-ols	+	7:10-26	Fb7:0-27	nd	2,32	27%	1
	Flavanones	-	6:101-101	Fb6.101-101	<0,05	<1,8	10-17%	2
	Flavanones	-	3:0-8	Fb3.0-8	<0,05	<1,8	4-16%	2
	Flavanones	-	5:39-41	Fb5.39-76	<0,05	<1,8	1-10%	2
	Flavanones	-	7:0-10	Fb7.0-10	<0,05	<1,8	1-5%	2
	Hydroxycinnamic.ac.deriv	+	2:0-30	Fb2.0-30	<0,05	6,45	40-58%	2
	Hydroxycinnamic.ac.deriv	+	4:0-44	Fb4.0-44	<0,05	<1,8	1-13%	2
	Hydroxycinnamic.ac.deriv	-	5:50-76	Fb5.50-76	<0,05	<1,8	6-8%	2
	Hydroxycinnamic.ac.deriv	-	3:54-94	Fb3.54-94	<0,05	<1,8	2-3%	2
	Hydroxycinnamic.ac.deriv	-	6:84-101	Fb6.84-101	<0,05	<1,8	4-6%	2
	Total polyphenols	+	4:9-20	Fb4:0-20	<0,05	2,27	26%	1
	Total polyphenols	-	5:50-76	Fb5:50-76	<0,05	1,84	22%	1

Figure CII. 6 polyphenolic mQTL. Graphical representation of strong mapped mQTL. Linkage groups (LG) and SSRs (in cM) are depicted in black lines. mQTLs are represented by coloured bars according to chemical families as in Figure 4 (red, anthocyanins; green, flavonols; dark blue, flavan-3-ols; pink, hydroxycinnamic acid derivatives; sky blue, flavanones; black, total (poly)phenols; grey, others). The effect of the introgression on the accumulation of the compound is shown by solid (increased accumulation) or striped (decreased accumulation) bars. The name of the specific metabolite/chemical family is on the right of the bar followed by the harvest year in which it was detected.



Observed candidate genes in mQTL regions

Considering that we were able to map a number of mQTL and that the sequence and annotation of the strawberry genome are publicly available (Shulaev *et al.* 2011), we aimed to identify interesting candidate genes in the mQTL regions described. In order to more precisely define our mQTL genetic intervals, we consider the physical position of introgresions boundaries defined by SNPs (according to *F. vesca* genome reference v1.1) after Axiom® IStraw90® SNPs Array hybridization (Table CII. 6). After a complete genome search, 313 genes were detected on the genetic intervals defined for the mQTLs (Table CII.6). We first checked the list of 99 biochemically characterized *Fragaria* genes (summarized in Supplemental Table 10, Shulaev *et al.* 2011) and found 69 of them located in genetic regions covered by our mQTLs. These included 14 genes that have been described as structural genes directly involved in pigment formation and phenylpropanoid and flavonoid metabolism *(FaCHS, FaCHS2, FaCHS3, FaCHS4 FaDFR, FaDFR1, FaF3H, FaFLS, FaLAR, FaANS, FaANR, FaGT1, FaGT2* and *FaGT7*), one transcription factor (*FaMYB1*) and three other non-structural protein-coding genes (*Fraa1A, Fraa2, Fraa3*) related to flavonoid biosynthesis (Supplemental Table CII.4).

We then checked the list of 32831 annotated predicted genes from the diploid strawberry reference genome for genes annotated *in silico* as enzymes in the phenolic synthesis pathway (*CHS, CHI, FHT, DFR, ANS, LAR, ANR, FLS* and *GT*), and for others that could be involved with substrates and products of our interest (anthocyanidin, flavonoid, flavonol, flavanone, chalcone, cinnamate, coumarate). As a result, we drew up a list of 340 candidate genes (not biochemically characterized) spread over all chromosomes, and some not anchored to the genome. The 255 located in regions covered by our mQTLs were selected (Supplemental Table CII.4).

Finally, we studied the phenotypic changes and the list of candidate genes associated with each QTL interval, case by case, to find any possible link between them. Located in the LG1:26-61 cM region was the biochemically characterized gene14611, known as *FaF3H* or *FaFHT* (Almeida *et al.* 2007), characterized as a flavanone-3-hydroxylase and that could be related to the special balance of flavonols in NILs carrying *F. bucharica* introgression in this region. The *FaCAD1* and *FaCAD2* genes, characterized as cinnamyl alcohol dehydrogenases, were also located in this region. In addition there was one predicted glucosyltransferase (gene14947) whose expression has been inversely related to kaempferol glucuronide accumulation in previous studies with octoploid varies of *Fragaria* (data not shown). There were a further three predicted monooxygenases (gene12513, gene12514 and gene12515) that could be involved with the observed mQTLs.

In section LG2:0-30 cM we found the gene11126, biochemically characterized as flavonol synthase (*FaFLS*) (Almeida *et al.* 2007), and four other genes annotated as FLS but without empirical evidence (gene01063, gene01034, gene23698 and gene11130). These could cause the decrease in flavan-3-ols as they compete for the same substrates (Almeida *et al.* 2007). The functions of other genes, annotated as coumarate—CoA ligase (gene24683), CHS (gene10965, gene10966) and CHI (gene27804), are implicated in the early steps of the phenylpropanoid pathway and so could be related to the increased accumulation of p-coumaroyls and cinnamoyl-glucose.

Table CII.6 mQTL location and predicted genes. Genetic (cM) and physical (bp) mQTL position, the nº of total predicted genes enclosed in the mQTL regions and the nº of genes selected for their putative implication in the phenylpropanoid pathway.

		possition (cM)		possition (bp)			predicted genes	
mQTL	LG	start	end	start	end	size (Mb)	total	phenolic-pathway
1:26-61	1	26	61	3315998	20747404	17,43	2629	29
2:0-30	2	0	30	0	16919611	16,92	2619	32
2:30-39	2	30	39	16919611	18950307	2,03	398	4
2:39-45	2	39	45	18950307	21299464	2,35	488	7
2:45-63	2	45	63	22329904	22663853	0,33	81	3
2:63-73	2	63	73	21489169	22279702	0,79	163	3
3:0-8	3	0	8	315550	1632631	1,32	358	8
3:54-94	3	54	94	5668456	31977233	26,31	3812	33
4:0-9	4	0	9	0	1708726	1,71	235	3
4:9-20	4	9	20	1708726	22629816	20,92	2855	25
4:20-44	4	20	44	22629816	26554022	3,92	607	9
5:0-11	5	0	11	0	2346373	2,35	471	12
5:11-35	5	11	35	2346373	6412800	4,07	741	15
5:39-41	5	39	41	6928955	7246484	0,32	75	1
5:41-50	5	41	50	7246484	11167415	3,92	715	15
5:50-76	5	50	76	11167415	28401116	17,23	2528	31
6:84-101	6	84	101	31889065	32907471	1,02	189	1
6:101-101	6	101	101	32907471	39317498	6,41	1234	36
7:0-10	7	0	10	0	14053681	14,05	2169	38
7:10-26	7	10	26	14053681	17678937	3,63	652	16
7:43-59	7	43	59	19453937	21046215	1,59	325	4

QTLs at region LG2:45-63 cM co-localize with the dihydroflavonol 4-reductases *FaDFR* and *FaDFR1* (gene15174) (Almeida *et al.* 2007), and another gene annotated as DFR (gene15176) could be related to the effect of the mQTL that over-accumulates cyanidin-glucoside. In a neighbouring region (LG2:63-73 cM) there was another mQTL over-accumulating quercetin glucoside (derived from dihydroquercetin as the cyanidin-glucoside) and two mQTL under-accumulating pelargonidin-based anthocyanins (derived from dihydrokaempferol). In this region we found one gene predicted as a leucoanthocyanidin dioxygenaseanthocyanin (LDOX) (gene15078) that may be related with the observed mQTLs. The closeness of these two regions leads us to think that the end of LG2 is important for anthocyanin regulation.

The LG3:54-94 cM genetic interval is very long and harbours 32 candidate genes (none of them biochemically characterized). According to the annotated functions, three genes might be involved in early steps of the pathway (gene27138, gene28898, and gene28093 putatively annotated as chalcone isomerase, coumarate-CoA ligase and trans-cinnamate 4-

monooxygenase, respectively), and the gene 00126, putatively annotated as anthocyanin glucosyltransferase that could be involved in the increased accumulation of anthocyanins.

Region LG4:9-20 cM covers a central region of LG4 and harbours genes such as *Fraa1A*, *Fraa2* and *Fraa3* that have been reported to be important in the control of the flavonoid pathway and pigment accumulation during fruit ripening. It has been seen that expression affects flavonoids and phenylpropanoids accumulation (Muñoz *et al.* 2010). This means that alleles introduced by *F. bucharica* introgression may affect the expression of these genes and favour the accumulation of hydroxycinnamic acid derivatives and anthocyanins.

At the LG5:0-11 cM genetic section, there are nine candidate genes. Among them, *FaANS* (gene32347) and *FaANR* (gene24665) have been functionally characterized as anthocyanin synthase and reductase respectively (Almeida et al 2007). *FaANR* catalyses the transformation of anthocyanidins (pelargonidin and cyanidin) into flavan-3-ols (epicatechin and epiafzelechin), and could therefore be responsible for the under-accumulation of flavan-3-ols in lines carrying a *F. bucharica* introgression in this region.

Gene25801 has been characterized as a flavonoid 3'hydroxylase, adding hydroxyl groups in position 3' to dihydrokaempferol, kaempferol, apigenin and naringenin, and has been shown to be related to anthocyanin compounds balance (Thill *et al.* 2013). It co-locates with the QTL described in LG5:41-50 cM, so could be involved in the decreased accumulation of pelargonidin-glucoside-malonate. In addition, the transcription factor *FaMYB1*, related to the production of flavonols and anthocyanins during fruit ripening, has been localised in this same region (Almeida *et al.* 2007).

The genetic region in LG5:50-76 cM gathers 29 genes annotated as being related to the phenylpropanoid and flavonoid pathway, but none have been biochemically characterized. Among them, there are four genes predicted as DFR (gene29344, gene31464, gene31465, gene02203), and one as flavonol glucosyltransferase (gene13530) that could be involved in the increased accumulation of kaempferol-glucosides. Gene22073, annotated as a putative leucoanthocyanidin dioxygenase that could be related to the observed QTLs involving total anthocyanin and total (poly)-phenols accumulation, has also been located in this region.

In the LG6:101-101 cM interval, there were 33 candidate genes. Among them, 25 are annotated as probable glucosyltransferases and two have been functionally characterized: *FaGT2* (gene26265) (Lunkenbein *et al.* 2006) and *FaGT7* (gene26344) (Griesser *et al.* 2008). They could be related with the general decrease in glucosides. Other genes in this region, gene 24019 (GT), gene 28428 (DFR) and gene 26403 (GT), are presumably related to pelargonidin-3-glucoside-malonate and kaempferol-coumaryl-glucoside accumulation in octoploid varieties (data not shown) and could here be related with the mQTL for PgGsM and KGs mapped in this region. There were also three predicted cinnamyl alcohol dehydrogenase (gene24025, gene26301 and gene26302), that could be related to the decrease in accumulation of cinnamoyl-glucose.

Localised at region LG7:0-10 cM are four biochemically characterized chalcone synthases: *FaCHS*, *FaCHS2*, *FaCHS3* and *FaCHS4*, the first three corresponding to gene26825 and the last to gene26826 (Almeida *et al.* 2007). These enzymes have been shown to play a central role in

the flavonoid metabolic pathway, so could be involved in the observed under-accumulation of flavanones and flavan-3-ols. Other interesting annotated genes (without empirical evidence) in this region include two probable LAR and four DFR that could be related to the lower accumulation of flavan-3-ols, and one flavonoid monooxygenase and two flavanone glucosyltransferases that could be involved in the decrease of eriodictyol isomers.

Within the boundaries of the region LG7:10-26 cM there are 16 candidate genes but none have been biochemically characterized. Functions predicted for these genes are as leucoanthocyanidin dioxigenase (gene 20980) and chalcone isomerase (gene 23397) that could be related to the general increased accumulation of flavan-3-ols.

Region LG7:43-59 cM includes one functionally characterized anthocyanidin glucosyltransferase (gene12591 named *FaFGT* after Almeida *et al*. (2007) or *FaGT1* after Griesser *et al*. (2008)) and three other genes annotated as flavonoid glucosyltransferases.

Discussion

F. vesca, a source of variability for poly-phenol content in strawberry breeding programmes

There are many factors contributing to (poly)-phenol compounds concentration in strawberry fruits, such as genotype, environmental conditions and growing practices, the degree of maturity at harvest, the procedures and solvents used for extraction, the use of fresh or dry material and the quantification method. So many variable parameters make comparisons between studies difficult.

To our knowledge, no previous publication analyses the (poly)-phenolic content of a mapping population in diploid strawberry (*F. vesca*). One previous study reported on the phenolic content of a single variety of *F. vesca* (Muñoz *et al.* 2011) but quantification was using dry material (DW) and a different internal standard, so absolute values are not comparable with our data, however the proportions between metabolites reported in this work are similar to theirs. For instance, they reported similar average concentrations of pelargonidin-3-glucoside and cyanidin-3-glucoside in contrast to studies on octoploid varieties that found much higher average concentrations of pelargonidin-3-glucoside. They also report lower values of quercetin-3-glucuronide than the other flavonols detected, contrasting with studies on octoploid material where quercetin-3-glucuronide is the major flavonol (Määttä-Riihinen *et al*. 2004; Buendía *et al.* 2010; Aaby *et al.* 2012; Ring *et al.* 2013). This suggests that the relative accumulation of (poly)-phenolic compounds is species-specific.

Comparing our results with studies using fresh material from octoploid varieties, we found that the average concentration for flavonoids was similar to (ellagic acid, flavonols – except for kaempferol-glucuronide) or higher than (anthocyanins, kaempferol-glucuronide) that reported for octoploid varieties. But even more interesting is that the range of concentrations was much broader within our NIL collection than the variability found between octoploid varieties (Aaby *et al.* 2012, Buendía *et al.* 2010) Fewer differences were found between the average concentrations of hydroxycinnamic acid derivatives. The concentration data we report here are similar to those reported by other studies with octoploid varieties (Määttä-Riihinen *et al.* 2004; Aaby *et al.* 2007; Buendia *et al.* 2010). As a result, we consider that *F. vesca* is a good source of variability in terms of quantity and balance, and could be used to improve the (poly)-phenolic content and composition of *Fragaria x ananassa* in breeding programmes. It has been shown that *F. vesca* accumulates higher average concentrations of several (poly)-phenols, mainly anthocyanins but also ellagic acid, and still has great potential for improvement as the range of variation is wide within our population

A NIL collection as a tool for detecting variability in poly-phenol compounds

Introgression line collections have previously been used to successfully map metabolic QTL in tomato (Rousseaux *et al.* 2005; Schauer *et al.* 2006; Alseekh *et al.* 2015), melon (Obando-Ulloa *et al.* 2009) and Arabidopsis (Lisec *et al.* 2008) The great degree of variability observed in the (poly)-phenolic profiling among the *F. vesca* NIL collection and the 72 mQTLs mapped reveal that introgressions from the exotic species *F. bucharica* have a significant effect in the accumulation of (poly)-phenols in the *F. vesca* genetic background. This, together with a

previous study with introgression lines in tomato (introgression from *S. pennellii* in the genetic background of *S. lycopersicum*) that allowed identification of candidate genes for phenolics accumulation (Di Matteo *et al.* 2013), leads us to consider the *F. vesca* NIL collection an appropriate tool for genetic dissection and candidate gene identification in biochemical pathways controlling (poly)-phenols synthesis.

The global view of the mapped mQTLs suggests that some introgressed regions could be related to a specific biological function. The exotic introgression at region LG2:45-63 cM increased the accumulation of cyanidin-3-glucoside and explains by 25-69% of the variance. As cyanidin-3-glucoside represents around the 40% of anthocyanins, by far the most abundant family of (poly)-phenols, an increase in cyanindin-3-glucoside could result in an enhanced total (poly)-phenol content (Supplemental Figure CII. 2). Line Fb**2**:39-63, harbouring the LG2:45-63 cM mQTL, may therefore represent a direct improvement in nutritional quality over *F. vesca* RV.

Other genetic regions also had an interesting set of mQTLs. For example, mQTLs observed in LG1:26-61 cM suggest that alleles from the exotic introgression in this region favour kaempferol-glucoside accumulation to the detriment of other flavonols (KGn, QGs). NIL with exotic introgressions in the central region of LG4 (LG4:9-20) have an interesting (poly)-phenolic phenotype, rich in hydroxycinnamic acid derivatives and ellagic acid, and, under specific environmental conditions, can over-accumulate total (poly)-phenols. Lines harbouring introgressions at region LG6:101-101 cM had an under-accumulation of several glucosides, with 25 of the candidate genes in this region being glucosyltransferases (including two biochemically characterized by Lunkenbein *et al*. (2006) and Griesser *et al*. (2008). This finding points to an effect of the alleles from *F. bucharica* on glucosyltransferase activity. Another interesting set of mQTLs mapped in LG7:0-10 cM, where the under-accumulation of both flavanones and flavan-3-ols suggests that the exotic introgressed alleles could hinder the transformation from naringenin to eriodictyol and block the transformation of leucocyanindin and cyanindin-3-glucoside to catechins. However, the regions delimited as mQTL are still too large to identify a single candidate gene, and further studies such as developing sub-NILs collections for positional cloning or expression analysis between NILs with contrasting phenotypes with the recurrent parental (RV), should be carried out in order to characterize candidate genes.

A previous study with this same NIL collection described 2 QTL for total phenol accumulation in LG2:0-30 cM and LG5:50-76 cM, using the Folin-Ciocalteu method (Urrutia *et al.* 2015) This technique is non-specific and all anti-oxidant substrates in the sample (including phenolics not identified by LC-MS but also other non-phenolic compounds) can interfere in the quantification results. In addition, not all phenolics have the same molar absorptivity in the assay and therefore do not contribute proportionally to the total quantification (Singleton *et al.* 1999) Therefore, these results are not comparable with our total (poly)-phenol quantification, as we sum up the identified phenols by their weights (mg/10g) not by their antioxidant capacity. However, total (poly)-phenol content calculated by LC-MS showed genetic variation in the NIL collection (Figure CII. 3, Table CII.4) and although this was not enough to map stable QTLs, two one-year QTL were mapped at regions LG4:9-20 cM and LG5:50-76 cM. The latter QTL co-localised in position and direction of accumulation with the QTL in LG5:50-76 cM described by

Urrutia *et al*. (2015). The positive QTL described in LG2:0-30 cM by the Folin-Ciocalteu method co-localises with positive and stable QTLs for cinnamoyl-glucose and total hydroxycinnamic acid derivatives.

The genotype and environment effects on (poly)-phenols accumulation

The accumulation pattern of (poly)-phenolic compounds in strawberry fruits in a NIL collection revealed the important effect of the genotype and the environmental conditions on their accumulation. A previous study by (Carbone *et al.* 2009) with an *F x ananassa* breeding population segregating for flavonoid production, found significant genotype-dependent differences at each developmental stage for most analyzed flavonoids. The accumulation of flavonols was also most affected by the environment. In our NIL collection we also observed a major effect of genotype on flavonoid and phenylpropanoid accumulation, and the chemical family most affected by the environment was also total flavonols, although other chemical families were also affected. Analysis of data also revealed that the balance between the metabolites of each class was more stable than the absolute quantification for anthocyanins, flavonols, flavan-3-ols and flavanones, suggesting that the relative abundance of compounds is more tightly regulated than their total accumulation. This was not the case for hydroxycinnamic acid derivatives, where we observed an increase in cinnamoyl-glucose ester not coupled with any variation in the accumulation of the other hydroxycinnamic acid derivatives, suggesting these compounds are not co-regulated (Table CII.1).

Besides the general genotype effect on the (poly)-phenolic accumulation, not all were equally affected by G, E or GxE. These parameters should be taken into account for selective (poly)-phenolic accumulation breeding as they allow us to discard those compounds with more unpredictable behaviour and concentrate efforts on those highly influenced by the genotype and little affected by the environment. For instance, accumulation of p-coumaroyl-glucoside was independent of the environment and had a single positive mQTL at LG2:0-30 cM, so it could be a good candidate compound to improve HCAD content. Another good target might be the accumulation of the anthocyanins PgGs and CyGs, that represent a high percentage of total (poly)-phenols, are highly affected by the genotype, and have a single mQTL each.

Introgression of wild species and NIL collections are a simple way to introduce variability in (poly)-phenolic composition and this is the first complete analysis in wild strawberry fruit. The new genetic regions controlling flavonoids accumulation described here open a way for further candidate gene characterization and provide a new powerful tool to increase the antioxidant capacity of fruits and to produce healthier strawberries for consumers.

Chapter III

Mapping the 'Wild strawberry' aroma in a Fragaria vesca NIL collection

Introduction

Strawberry is the most economically important berry crop. It is cultivated and distributed worldwide and it has been traditionally appreciated for its aroma. However, breeding programs in the past decades addressed more efforts to improve its agronomical performance rather than its organoleptic quality, resulting in a set of varieties producing appealing big brilliant red fruits but lacking in flavor and aroma. The latest becoming the main cause of dissatisfaction among consumers (Bruhn *et al.* 1991). It has been seen that aroma or the total volatile content is an important factor contributing to flavor perception and for overall liking of strawberry fruits (Schwiterman *et al.* 2014).

Aroma is a very complex trait. Over 350 volatile compounds have been identified in *Fragaria* sp. comprising esters, aldehydes, ketones, furanones, alcohols and terpenoids (Latrasse 1991) but only few of them have been reported to contribute to strawberry aroma perceived by humans (Schieberle & Hofmann 1997; Ulrich *et al.* 1997; Ulrich *et al.* 2007).

Esters are the most abundant volatile compounds in the strawberry aroma. Among esters methyl-butanoate, ethyl-butanoate, butyl butanoate, methyl-hexanoate, ethyl-hexanoate, butyl acetate and hexyl acetate have been described as important odorants contributing to the fruity scent. Methyl 2-aminobenzoate (also known as methyl anthranilate) has been reported to confer alone the typical "wild strawberry" aroma that is found in woodland strawberry (*F. vesca*) accessions and very rarely in commercial varieties. Methyl cinnamate has been reported to add pleasant spicy notes and myrtenyl acetate to provide herbaceous refreshing notes (Schieberle & Hofmann 1997; Ulrich *et al.* 1997; Jetti *et al.* 2007; Ulrich *et al.* 2007; Olbricht *et al.* 2008; Schwieterman *et al.* 2014). Furans, and specifically furaneol and mesifurane, are considered important contributors for strawberry aroma adding caramel notes (Schieberle & Hoffmann 1997, Ulrich *et al.* 1997; Ulrich *et al.* 2007; Jetti *et al.* 2007). The terpenoids linalool and nerolidol, that add flowery sweet notes, are also considered important contributors to strawberry aroma (Ulrich *et al.* 1997, Olbrich *et al.* 2008, Schwieterman *et al.* 2014), but they have been described mainly in octoploid cultivars (*F. x ananassa*) and not in diploid wild strawberries (*F. vesca*) (Aharoni *et al.* 2004). The known as 'green compounds', (*Z*)-3-hexenal, (*E*)-2-hexenal and (*Z*)-3-hexen-1-ol, have been reported to contribute to the fresh aroma impression that decreases with ripeness (Ulrich *et al.* 1997, Schieberle & Hoffman 1997). Other important compound for the strawberry aroma is the γ-decalactone conferring 'peach-like' notes (Ulrich *et al.* 1997, Jetti *et al*. 2007, Olbrich *et al.* 2008).

It has been seen that wild strawberries (*F. vesca*) have a more intense and fruity aroma than cultivated varieties mainly due to their higher concentration in esters and terpenoids and that also preserves the pleasant and easily identifiable 'wild-strawberry' aroma due to the accumulation of methyl 2-aminobenzoate (Ulrich *et al.* 1997; Ulrich *et al.* 2007).

Efforts have been addressed to characterize the aroma pattern of different octoploid accessions or mapping populations including commercial and wild material (Jetti *et al.* 2007; Olbricht *et al.* 2008; Zorrilla-Fontanesi *et al.* 2012; Schwieterman *et al.* 2014) and also to compare aroma patterns between octoploid and diploid strawberries (Aharoni *et al.* 2004; Ulrich *et al.* 2007; Dong *et al.* 2013) however, despite of its outstanding organoleptic

characteristics the volatile composition of a diploid *F. vesca* mapping population has never been characterized to our knowledge.

F. vesca is a simplified model for the study of the strawberry genetics as previous studies have reported that the degree of synteny between *F. vesca* and the commercial hybrid *F. x ananassa* is very high (Rousseau-Gueutin *et al.* 2008, Tennessen *et al.* 2014), so transferability of the knowledge can be very fast. In addition, the availability of a high quality reference genome sequence (Shulaev *et al.* 2011) and the recently developed near isogenic line (NIL) mapping collection (Urrutia *et al.* 2015) are powerful tools for the study of any genetic trait.

In this study we perform a detailed profiling and a QTL mapping of the volatile composition of a *F. vesca* NIL collection with the objective of defining genetic regions responsible for the accumulation of volatile compounds, especially for those compounds that are more relevant for wild strawberry aroma.

Materials and Methods

Plant material and sample extraction

We studied the volatilome of diploid strawberry ripe fruits using 42 lines from a near isogenic line (NIL) collection in *F. vesca* that had been previously developed (Supplemental Figure CIII. 1). The recurrent parental of this collection is *F. vesca* var. 'Reine des Vallées' and the donor parental is *F. bucharica* 'FDP 601' (Urrutia *et al.* 2015).

We cultivated 6 to 8 plants from each NIL in the collection, both parental, the hybrid (F_1, *F. vesca* x *F. bucharica*) from which the whole collection descends and a yellow-fruited variety of *F. vesca* named 'Yellow Wonder' (YW), which has a very pleasant pineapple-like aroma. Plants were germinated independently from seed in two consecutive years (2012 and 2013) and grown in a shaded greenhouse in Caldes de Montbui (latitude: 41º 36'N, longitude: 2º 10' E, altitude 203m over the sea level, pre-coastal Mediterranean climate) following the usual agronomical practices for this crop.

Fully ripe berries from all NILs and genotypes that set fruit each year in three to five biological replicates were collected. Each biological replicate consisted in a mix of 10 to 15 fruits (> 5 g) and was individually and immediately frozen in liquid nitrogen, grinded to fine powder and stored at -80ºC until gas chromatography-mass spectrometry (GC-MS) analysis.

Volatile compounds analysis

Volatile compounds were determined in a similar way as described in Zorrilla-Fontanesi *et al.* (2012) at Instituto de Biologia Molecular y Celular de Plantas (IBMCP, Valencia, Spain). Each biological replicate was analyzed as an independent sample. Before the volatile compounds analysis an aliquot of 500 mg of frozen fruit powder from each sample was weighed in a 7 mL glass vial and thawed at 30ºC during 5 min. Then 500 µL of a saturated NaCl solution were added and the mixture was homogenized gently. This mixture (500 µL) was transferred to a 10 mL screw cap headspace vial and analyzed immediately.

Volatiles were sampled HS-SPME (headspace solid phase microextraction) with a 65 µm PDMS/DVB (polydimethylsiloxane/divinyl-benzene) fiber (Supelco, PA, USA). First, vials were tempered at 50ºC for 10 min. Then volatiles were extracted by exposing the fiber to the vial headspace for 30 min at 50ºC with agitation. The extracted volatiles were desorbed to the GC injection port at 250ºC for 1 min in splitless mode. Incubation, volatile extraction and desorption were performed by means of a Combi-PAL autosampler (CTC Analytics, Zwingen, Switzerland). GC-MS was performed in a 6890N gas chromatograph coupled to a 5975B mass spectrometer (Agilent Technologies, CA, USA). Gas chromatography conditions were: DB-5ms column (60 m, 0.25 mm, 1 µm) (J&W Scientific, CA, USA), constant helium flow 1.5 mL min^{-1}; oven programming conditions: 35ºC for 2 min, 5ºC min^{-1} ramp until 250ºC, then 5 min at 250ºC. Compounds were monitorized over the mass-charge ratio (m z^{-1}) range of 35-250.

Chromatograms and mass spectra were analyzed using the Enhanced ChemStation software (Agilent Technologies, CA, USA). Volatile compounds were unambiguously identified by comparison of both retention time and mass spectra to those of commercial standards (SIGMA-Aldrich, MO, USA) run under the same conditions. When the commercial standards

were not available, compounds were tentatively identified by comparison of their mass spectrum to those in the NIST 05 mass spectral library. These compounds are marked with a "T" after its chemical name (Table CIII. 1). Identified compound areas were normalized by comparison with the peak area of the same compound in a reference sample which was injected regularly each five to six samples. This reference sample consisted on a homogeneous mix of all the samples analyzed each year. This normalization was done in order to correct for variations in sensitivity and fiber aging.

Data and mQTL analysis

Data are expressed as base 2 logarithm of the ratio between each sample and reference sample values. All lines that set fruit each year were processed and analyzed by GC-MS and taken into account for the QTL mapping, however for the exploratory analysis only those genotypes that gave enough fruit production both years were considered.(C.III Supplemental Figure 1). All the statistical analysis and graphical representations were done using the free source software R 2.15.1 (RCoreTeam, 2012) with the Rstudio 0.92.501 interface (Rstudio, 2012) unless otherwise is specified. Pearson's correlation between the metabolites and the genotypes were calculated using the *rcorr* function from Hmisc package (Harrell, 2014). The analysis of variance (ANOVA) fitting the model: G+E+GxE was calculated using the *Anova* function from the car package for R (Fox 2011) , and the omega squared values (ω^2) were obtained from ANOVA parameters according to the following formula: $(SS_i – df_i * MS_{err}) * (MS_t + MS_{err})^{-1}$. Principal Components Analysis (PCA) was calculated with the *prcomp* function and scaled values. The Hierarchical Clustering Analysis (HCA) was calculated considering euclidean distance and complete linkage clustering method. The Cluster Network Analysis (CNA) was calculated with the *qgraph* function from the qgraph R package (Epskamp *et al*. 2012). The significance tests were recursively calculated using the *t.test* function comparing every genotype mean value with the *F. vesca* RV mean value in the same harvests. The obtained p-values were corrected by false discovery rate (FDR) with *p.adjust* function. Significant threshold was set at p-value < 0.05 after correction. QTLs were mapped to a specific genetic region only when all the NILs harboring an exotic introgression in this region had a significant effect and in the same direction for the specific metabolite of study. To confirm these QTL and estimate their effect an interval mapping analysis was performed using MapQTL v.6 (Van Ooijen 2009). Stable regions that explained around 20% or more of the variability and showed LOD scores >1.8 were considered major QTLs. The graphical representation of the mQTLs was done using MapChart 2.2 (Voorrips, 2002).

Results

Profiling and variability of the volatile compounds in the NIL collection

To evaluate the differences in volatile composition by GC-MS and detect genomic regions responsible for its control, we analyzed the volatile metabolic profiles of fully ripe strawberry fruits from a *F. vesca* NIL collection (Urrutia *et al.* 2015), its recurrent parental, *F. vesca* var. Reine des Vallées (RV), used as a control and the white-fruited *F. vesca* var. Yellow wonder (YW) used as an external control or out-group in two consecutive years (2012 and 2013). Metabolite profiling analyses were performed with all the genotypes that set enough fruit each year but for the exploratory statistical analysis we only took into account those 25 genotypes that could be analyzed with at least 3 biological replicates both years (Supplemental Figure CIII. 1).

We unambiguously identified 88 volatile compounds by comparison with commercial standards run under the same conditions and tentatively identified 12 additional compounds marked with a T at the end of the compound name (Table CIII. 1). The identified compounds comprise 11 alcohols, 16 aldehydes, 46 esters, 4 furans, 14 ketones, 8 terpenoids and one lactone and include most of the important contributors to strawberry aroma described in the literature (Schieberle & Hofmann 1997; Ulrich *et al.* 1997; Ulrich *et al.* 2007, Olbrich *et al.* 2008). These considered 'key compounds' are marked with an arrow symbol in Table CIII. 1. Compounds data are expressed as log2 of a ratio (sample/reference sample) and 3 to 5 biological replicates were taken into account for the analysis. The reference sample is a homogenous mix of all samples present in the analysis.

Mean and standard deviation (sd) values for all compounds for both RV and the NIL collection were calculated independently. We could observe that for most compounds RV and average NIL collection show mean log2 ratio values near 0 (what is equivalent for ratio=1) in both harvests (Table CIII. 1). This was expected as the studied population is nearly isogenic, therefore all the analyzed lines (NIL collection and RV) share a common genetic background and the reference sample used for calculating the ratios was a mix of all of them. We also calculated the range (min. and max. log2 ratio) for each compound within the NIL collection each harvests and found that all identified volatile compounds segregated among our NIL collection. Ranges went from fold changes of 0.0005 (log2 ratio = -10.87) for ethyl octanoate in 2012 harvest to fold changes of 120 (log2 ratio = 6.91) for γ-decalactone in 2013. It can be observed that NIL collection ranges have more extreme values for minimal than for maximal ratios. This was expected as *F. vesca* RV, the genetic background of the collection, has a higher fruit quality and aroma than *F. bucharica*, the donor parental, so this result may indicate that *F. bucharica* introgressions tend to decrease volatile accumulation. However not all compounds varied in the same proportions, the less and most variable compound within the collection were decanal (-1.35 to 0.58 in 2012) and γ-decalactone (-6.64 to 6.91 in 2013) respectively. It is also noteworthy mention that nerolidol segregated in a dominant absence-presence manner, thus indicating that *F. vesca* RV is not able to produce nerolidol and the introgressed alleles of *F. bucharica* conferred this ability to the *F. vesca* genetic background (Supplemental Table CIII. 1). This was also confirmed by the F_1 sample (hybrid *F. vesca* RV x *F. bucharica*), present only in 2013 analysis, that also produced nerolidol.

Table CIII. 1. Volatile compounds summary. All identified compounds and their assigned number codes and clusters are presented. Tentatively identified compounds are indicated with a T after the chemical name. A selected set of important compounds contributing to strawberry aroma are indicated with an arrow. Data are expressed as the log2 of the ratio between samples and a reference. Mean ratios and standard deviation (sd) were calculated for each compound in the recurrent parental F.vesca (RV) and in the average NIL collection for 2012 and 2013 harvests. The range of the ratios (min and max values) was calculated for the NIL collection in both harvests.

					Recurrent parental (RV)				NIL collection							
					2012		2013		2012				2013			
	Code	Compound	Cluster	Family	mean	sd	mean	sd	mean	sd	range		mean	sd	range	
	1	1-decanol	A	alcohol	-0,29	1,73	-0,54	1,60	-2,63	2,26	-6,73	2,44	-2,14	2,19	-6,64	1,85
	2	1-hexanol	C	alcohol	0,02	0,38	-0,11	0,71	0,11	0,82	-1,33	2,78	0,13	0,71	-1,56	1,97
	3	1-octanol	D2	alcohol	0,56	0,90	-0,16	0,80	-1,01	1,17	-3,41	1,81	-1,06	1,25	-4,64	1,13
	4	1-penten-3-ol	C	alcohol	0,19	0,68	-0,31	0,65	0,15	0,78	-2,07	2,62	0,09	0,58	-1,94	1,29
	5	2-heptanol	D2	alcohol	0,04	1,19	-0,57	2,07	-1,90	2,23	-7,50	2,18	-2,02	1,80	-6,64	2,05
	6	2-nonanol	D2	alcohol	0,10	0,94	-0,19	1,30	-1,68	2,17	-10,59	2,01	-1,86	1,91	-6,64	0,92
	7	2-tridecanol T	D2	alcohol	-0,19	1,73	-0,68	2,27	-0,95	1,66	-4,39	2,92	-1,61	1,77	-5,64	2,01
	8	2-undecanol T	D2	alcohol	0,06	0,95	-0,40	1,89	-1,24	1,68	-4,55	2,38	-1,62	1,56	-5,06	0,90
	9	(E)-2-hexen-1-ol	B	alcohol	0,03	0,57	-0,45	0,71	-0,53	2,02	-5,78	2,89	-0,90	2,42	-6,64	2,12
	10	Ethanol	A	alcohol	-0,76	1,17	-1,83	1,68	-1,07	1,86	-7,45	2,86	-2,03	2,17	-5,64	2,58
	11	Eugenol	D2	alcohol	-1,31	0,44	-0,97	1,81	-1,46	1,49	-4,24	4,27	-1,79	1,94	-4,64	4,26
	12	3,4-dimethylbenzaldehyde	C	aldehyde	-0,01	0,36	-0,06	0,16	-0,12	0,62	-1,26	1,60	-0,07	0,28	-0,94	1,23
	13	Benzaldehyde	C	aldehyde	0,03	0,35	-0,03	0,39	0,20	0,83	-1,68	2,48	0,06	0,70	-1,69	1,45
	14	Decanal	C	aldehyde	-0,04	0,38	-0,11	0,57	-0,23	0,40	-1,35	0,58	-0,06	0,50	-1,09	0,89
	15	(E)-2-decenal	D2	aldehyde	0,14	0,47	-0,51	0,51	0,10	0,62	-1,92	1,37	-0,33	0,84	-2,94	2,64
	16	(E)-2-heptenal	C	aldehyde	0,32	0,46	0,06	0,70	0,84	1,35	-1,95	4,59	0,90	1,38	-2,40	4,27
→	17	(E)-2-hexenal	C	aldehyde	0,39	0,20	-0,01	0,38	0,03	0,65	-2,04	0,89	-0,04	0,69	-1,69	0,87
	18	(E)-2-nonenal	C	aldehyde	-0,70	0,37	-0,12	0,45	-0,63	0,48	-2,04	0,33	-0,15	0,63	-1,89	1,20
	19	(E)-2-octenal	C	aldehyde	0,67	0,59	0,07	0,44	0,32	0,74	-1,53	1,85	0,13	0,57	-2,00	1,49

					Recurrent parental (RV)				NIL collection							
					2012		2013		2012				2013			
	Code	Compound	Cluster	Family	mean	sd	mean	sd	mean	sd	range		mean	sd	range	
	20	(E)-2-pentenal	C	aldehyde	0,72	0,77	0,27	1,36	0,88	1,03	-2,71	3,11	0,86	0,93	-1,74	2,83
	21	(E,Z)-2,4-heptadienal	C	aldehyde	0,27	0,52	-0,12	0,59	0,03	0,60	-0,95	1,96	0,06	0,55	-1,43	1,95
	22	Heptanal	C	aldehyde	0,25	0,58	0,17	0,78	0,13	0,63	-0,98	1,81	0,38	0,77	-1,25	2,29
	23	Hexanal	C	aldehyde	0,22	0,27	-0,04	0,45	0,29	0,41	-1,14	1,23	0,28	0,44	-1,12	1,23
	24	Nonanal	C	aldehyde	0,24	0,67	0,56	0,61	0,21	0,69	-1,35	2,30	0,14	0,78	-1,22	2,23
	25	Octanal	C	aldehyde	0,93	0,99	1,07	1,54	1,00	1,21	-1,90	4,09	0,89	1,37	-1,69	3,89
	26	Pentanal	C	aldehyde	0,22	0,55	0,40	0,92	0,30	0,78	-2,02	1,71	0,47	0,75	-1,69	2,17
→	27	(Z)-3-hexenal	C	aldehyde	0,26	0,36	0,32	0,68	0,84	1,13	-1,00	3,91	1,15	1,42	-0,94	4,65
	28	1-methylbutyl butanoate	D2	ester	0,04	1,37	0,41	1,36	-1,59	1,63	-3,74	2,89	-1,45	1,38	-2,94	3,61
	29	1-methylethyl butanoate	D2	ester	0,72	0,97	0,36	0,92	-0,02	0,96	-2,02	2,08	-1,11	1,29	-4,64	1,84
	30	1-methylethyl acetate	C	ester	0,17	0,77	-0,95	0,69	0,39	0,74	-1,91	1,83	-0,74	1,01	-3,18	1,36
	31	1-methylhexyl acetate	D2	ester	-0,08	1,04	-0,74	2,09	-1,20	2,11	-7,69	3,12	-1,95	1,73	-6,64	1,90
	32	1-methyloctyl butanoate	D2	ester	-0,62	0,87	0,11	1,23	-1,71	1,42	-4,09	2,25	-1,63	1,59	-4,32	2,62
	33	2,3-butanedioldiacetate T	A	ester	-1,19	1,64	-2,36	2,01	-1,32	2,14	-4,72	3,69	-2,22	2,32	-6,64	3,18
	34	2 -methylbutyl acetate	C	ester	0,22	0,54	-0,53	0,72	0,52	0,78	-1,26	2,40	-0,51	1,00	-2,40	1,66
	35	3-methyl-2-butenyl acetate	C	ester	0,33	0,48	-0,42	0,96	0,29	1,03	-2,46	2,86	-0,21	1,17	-2,56	2,48
	36	3-methylbutyl acetate	C	ester	0,44	0,57	-1,72	0,48	0,04	1,06	-6,00	2,79	-0,85	0,94	-3,32	1,66
	37	Benzyl acetate	D1	ester	1,03	1,03	0,15	1,47	-0,15	1,36	-3,00	2,27	-0,34	1,42	-3,64	1,99
→	38	Butyl acetate	D2	ester	0,01	0,54	-0,18	0,55	-0,03	1,28	-2,94	2,31	-0,50	1,24	-3,64	2,01
→	39	Butyl butanoate	D2	ester	-0,03	0,83	0,50	1,13	-1,42	2,00	-5,86	2,78	-1,57	2,50	-6,64	2,94
	40	Butyl hexanoate	A	ester	-0,17	1,10	-0,06	1,41	-2,06	2,35	-5,97	2,32	-1,82	2,70	-6,64	3,20
	41	Cinnamyl acetate	D1	ester	0,05	0,65	-1,12	3,42	-1,59	2,71	-5,44	4,31	-2,46	2,68	-6,64	1,91
	42	Decyl acetate	A	ester	-0,24	1,14	-0,69	1,67	-2,76	2,49	-7,62	2,18	-2,17	2,34	-6,64	1,51
→	43	(E)-2-hexenyl acetate	B	ester	1,57	0,51	-0,55	0,63	-0,48	2,26	-8,59	2,03	-1,21	2,50	-6,64	2,45

					Recurrent parental (RV)				NIL collection							
					2012		2013		2012				2013			
	Code	Compound	Cluster	Family	mean	sd	mean	sd	mean	sd	range		mean	sd	range	
	44	Ethyl 2-hexenoate	A	ester	-2,09	2,49	-0,92	1,54	-2,80	3,12	-6,06	3,54	-2,42	2,59	-5,64	2,64
	45	Ethyl acetate	A	ester	-0,79	2,11	-1,72	1,78	-2,12	3,19	-6,73	4,15	-2,22	2,45	-6,64	1,57
→	46	Ethyl butanoate	D2	ester	0,33	1,05	-0,51	2,29	-1,30	2,27	-6,84	2,03	-0,99	1,81	-6,64	1,64
	47	Ethyl decanoate	A	ester	-1,70	2,54	-1,97	1,76	-4,65	3,72	-9,38	2,89	-3,58	2,65	-6,64	3,65
	48	Ethyl dodecanoate	A	ester	-1,84	2,56	-2,43	1,88	-3,77	3,16	-6,29	4,09	-4,29	2,87	-6,64	4,53
→	49	Ethyl hexanoate	A	ester	-0,36	1,69	-0,44	2,29	-3,36	3,48	-9,74	1,67	-1,94	2,34	-6,64	1,26
	50	Ethyl methylthioacetate T	D2	ester	0,99	1,25	-0,32	2,20	-1,33	2,36	-5,15	2,53	-0,32	1,96	-5,64	2,38
	51	Ethyl octanoate	A	ester	-0,90	2,32	-1,03	2,08	-4,61	4,09	-10,87	2,24	-2,88	2,57	-6,64	2,77
→	52	Hexyl acetate	D2	ester	0,50	0,43	0,26	0,52	-0,23	0,92	-2,68	1,40	-0,24	0,88	-3,06	1,74
	53	Hexyl butanoate	D2	ester	-0,36	1,17	0,04	1,19	-1,52	1,94	-5,16	2,21	-1,25	2,22	-6,64	3,39
	54	Hexyl hexanoate	D2	ester	-0,35	0,89	-0,17	2,12	-1,72	2,02	-5,88	1,96	-1,13	2,14	-6,64	2,84
→	55	Methyl 2-aminobenzoate T	D2	ester	-0,45	0,50	-0,47	2,97	-0,58	2,19	-6,34	2,56	-1,22	2,39	-6,64	3,70
	56	Methyl 2-hexenoate	D2	ester	0,83	0,85	-0,02	0,95	-0,37	1,67	-3,08	3,96	-1,04	1,54	-3,64	2,30
	57	Methyl 3-hydroxyoctanoate T	D2	ester	0,75	0,78	0,12	2,79	-1,24	2,17	-8,35	2,62	-1,09	1,98	-6,64	3,03
	58	Methyl acetate T	D2	ester	0,26	1,43	-0,67	0,85	0,54	1,01	-2,12	3,48	-0,46	0,97	-3,84	2,14
	59	Methyl benzoate	D1	ester	0,42	1,06	-0,73	1,10	-0,30	1,24	-3,30	2,81	-0,57	1,56	-4,32	2,83
→	60	Methyl butanoate	D2	ester	0,66	1,10	-0,06	1,37	-0,74	1,55	-4,59	2,80	-1,09	1,82	-6,64	1,88
→	61	Methyl cinnamate T	D1	ester	-0,68	0,91	-1,66	1,73	-0,35	2,14	-6,57	3,48	-1,01	2,14	-5,64	4,49
	62	Methyl decanoate	D2	ester	-0,08	0,80	-0,92	1,86	-1,17	1,74	-6,46	1,62	-1,54	1,95	-6,64	2,24
	63	Methyl dodecanoate	A	ester	-0,28	1,46	-1,02	1,62	-1,60	1,81	-3,50	2,72	-2,14	1,88	-4,64	3,29
→	64	Methyl hexanoate	D2	ester	0,02	1,05	-0,28	1,79	-1,80	2,19	-6,52	1,68	-1,11	1,78	-5,64	1,79
	65	Methyl octanoate	D2	ester	0,27	0,58	-1,00	2,02	-0,76	1,44	-4,19	1,49	-1,31	1,72	-6,64	1,63
→	66	Myrtenyl acetate	D2	ester	0,88	0,48	0,41	0,84	-0,02	0,78	-2,13	1,70	-0,36	0,94	-2,56	2,24
	67	Nonyl acetate	D2	ester	0,05	0,58	-0,03	1,19	-0,42	1,06	-2,34	1,56	-0,63	1,13	-2,32	2,29

					Recurrent parental (RV)				NIL collection							
					2012		2013		2012				2013			
	Code	Compound	Cluster	Family	mean	sd	mean	sd	mean	sd	range		mean	sd	range	
	68	Octyl acetate	A	ester	0,30	0,87	-0,26	1,43	-2,03	2,17	-6,75	1,72	-1,76	2,05	-6,64	1,31
	69	Octyl butanoate	A	ester	-0,50	1,50	-0,45	1,76	-2,57	1,93	-5,12	2,63	-1,98	2,38	-5,64	3,30
	70	Octyl hexanoate	A	ester	-0,01	2,15	-0,09	2,11	-3,13	2,10	-4,94	2,98	-2,44	2,58	-5,64	3,13
	71	Pentyl acetate	C	ester	0,48	0,47	-0,39	0,31	0,18	0,63	-1,08	1,60	-0,34	0,83	-2,94	1,95
	72	Propyl butanoate	D2	ester	0,17	0,57	-0,27	1,06	-0,67	1,75	-5,29	2,39	-0,90	1,76	-5,64	1,91
→	73	(Z)-3-hexenyl acetate	C	ester	0,13	0,93	-0,69	0,76	-0,21	1,29	-2,57	3,35	0,04	1,51	-3,47	4,46
	74	2,1-pentenylfuran	C	furan	0,63	0,54	-0,03	0,53	0,20	1,07	-3,59	2,03	0,11	0,89	-2,94	1,48
	75	2-pentylfuran	C	furan	0,17	0,30	-0,21	0,32	0,23	0,73	-2,17	1,53	0,11	0,48	-1,09	1,32
→	76	Furaneol	D1	furan	0,43	1,11	-1,94	2,78	-0,18	1,74	-6,00	3,27	-0,92	2,97	-6,64	6,91
→	77	Mesifurane	D1	furan	0,50	0,76	-2,53	1,91	0,05	1,42	-4,29	2,44	-2,46	1,76	-6,64	0,30
	78	1-penten-3-one	C	ketone	0,73	0,47	0,69	0,25	0,63	0,92	-1,68	2,16	0,31	0,65	-1,47	1,68
	79	2-heptanone	D2	ketone	0,25	0,93	0,17	1,50	-1,13	1,84	-5,90	1,87	-0,75	1,67	-6,64	1,35
	80	2-nonanone	D2	ketone	0,16	0,74	0,20	1,29	-1,23	2,07	-10,17	1,76	-1,07	1,99	-6,64	1,01
	81	2-pentadecanone T	A	ketone	0,31	1,56	-0,72	2,79	-1,43	2,18	-5,76	2,66	-2,00	2,39	-6,64	2,34
	82	2-pentanone	D2	ketone	0,17	1,44	0,17	1,97	-1,65	2,34	-7,25	2,75	-1,08	1,96	-6,64	1,86
	83	2-tridecanone T	D2	ketone	0,08	0,98	-0,72	2,19	-1,32	1,86	-6,57	2,02	-1,69	1,89	-6,64	1,39
	84	2-undecanone T	D2	ketone	0,18	0,72	-0,17	1,74	-0,85	1,70	-5,57	2,11	-1,07	1,55	-5,64	1,23
	85	4-tridecanone T	D2	ketone	-0,01	0,96	0,10	1,05	-0,20	0,95	-1,07	3,26	-0,16	0,81	-1,15	2,40
	86	6-methyl-5-hepten-2-one	C	ketone	-0,32	0,54	-0,09	0,54	-0,05	0,62	-1,77	1,30	0,38	0,49	-1,51	1,50
	87	Acetone	C	ketone	0,44	0,56	-0,47	1,35	-0,15	0,91	-2,89	1,78	-0,60	0,95	-3,64	1,01
	88	Acetophenone	D1	ketone	0,65	0,57	0,16	1,09	-0,27	1,23	-3,86	1,74	-0,54	1,34	-3,47	3,42
	89	α-ionone	D2	ketone	0,68	0,43	0,32	1,19	0,00	1,20	-2,61	2,41	-0,06	0,92	-2,25	2,30
	90	β-ionone	D2	ketone	0,48	0,49	-0,39	0,97	0,11	0,83	-2,10	1,79	-0,19	0,73	-2,94	0,99
	91	(Z)-geranyl acetone	D2	ketone	-0,02	0,62	-0,13	0,57	-0,08	0,79	-2,41	1,63	0,00	0,62	-1,79	1,36

					Recurrent parental (RV)				NIL collection							
					2012		2013		2012				2013			
	Code	Compound	Cluster	Family	mean	sd	mean	sd	mean	sd	range		mean	sd	range	
→	92	γ-decalactone	D2	lactone	0,72	1,09	-0,01	3,67	-1,18	1,82	-6,72	3,37	-1,18	2,57	-6,64	6,91
	93	α-farnesene	D2	terpenoid	0,20	1,34	0,32	1,18	-0,40	1,59	-4,81	2,80	-0,45	1,11	-2,84	2,74
	94	α-pinene	C	terpenoid	0,07	0,48	-0,85	0,72	0,32	0,79	-1,65	2,05	-0,31	1,12	-2,84	2,58
	95	Limonene	C	terpenoid	0,23	0,51	-0,42	0,49	-0,17	0,75	-1,49	2,47	-0,30	0,52	-1,56	0,92
→	96	Linalool	C	terpenoid	0,02	0,47	-0,61	0,66	0,01	0,94	-1,98	3,31	0,25	1,09	-1,69	3,54
	97	Myrtenol	C	terpenoid	0,87	0,36	-0,48	0,64	-0,11	0,90	-2,58	1,67	-0,44	1,01	-2,56	2,97
	98	Nerol	C	terpenoid	0,44	0,36	-0,42	0,44	-0,18	1,00	-4,23	1,85	-0,25	1,36	-6,64	2,56
→	99	Nerolidol	C	terpenoid	0,00	0,00	0,00	0,00	0,15	0,52	0,00	2,75	0,14	0,48	0,00	2,54
	100	Terpineol	C	terpenoid	0,27	0,43	-0,03	0,61	0,15	0,51	-1,22	1,22	0,07	0,75	-1,74	1,93

Relations between volatile compounds and NILs

To further study the accumulation pattern of the different volatile compounds within the NIL collection and to establish groups of compounds that were co-regulated we represented the mean log2 ratios for all NIL, RV and YW in both years in a heat map grouped by hierarchical clustering (Figure CIII. 1). Volatile compounds were classified in 4 main groups or clusters named A, B, C and D. D cluster was subdivided in D1 and D2. For clarity, the cluster allocation of each compound together with a nº code is given in Table CIII. 1. Cluster A is composed by 16 volatiles and enriched in long carboxylic acid derived esters and in octyl derived esters, cluster B is formed by only two compounds (*E*)-2-hexenyl acetate and (*E*)-2-hexen-1-ol. Cluster C is integrated by 35 metabolites and it groups all aldehydes (except (E)-2-decenal), all terpenoids (except α-farnesene) and it is enriched in 4C alkyl acetates. Cluster D has 47 compounds, the small subgroup D1 has 7 volatiles and is enriched in aromatic compounds (carrying a benzoic ring) and also includes two furans (mesifurane and furaneol). Subgroup D2 clusters 40 compounds and is enriched in esters derived from butanoic and acetic acids and long chain alcohols and ketones. The heatmap reveals that *F. vesca* RV maintains a nearly average log2 ratio for all compounds without extreme values for any of them. However *F. vesca* YW, shows more extreme accumulation ratios and it is enhanced in volatile compounds from cluster A and D2 (both rich in esters) and shows a decreased profile in compounds from cluster B, C and D1. (Figure CIII. 1). This was expected as the ratio is compared with a mix of all samples that share the genetic background of RV (C.III Supplemental Figure 1) in contrast with YW that is not genetically related with the NIL collection.

NIL show differential accumulation patterns of volatile compounds. Line Fb**1**:26-61 shows an average to low accumulation profile in compounds from all clusters with a slight up-regulation in compounds 72, 60, 53 and 39 from cluster D2, these compounds are down-regulated in line Fb**7**:0-10. NIL harboring introgressions in LG2 also show a general intermediate to low profile for most volatile compounds, however we can find a distinctive accumulation pattern in compounds 44, 33, 45, 81, 10 from cluster A that are over-accumulated in NILs with introgressions at the beginning of LG2 (Fb**2**:0-30 and Fb**2**:0-39) and under-accumulated in NILs with bigger introgressions covering the end of the LG2 (Fb**2**:0-45, Fb**2**:0-63, Fb**2**:073). Lines with *F. bucharica* introgressions at the beginning of LG3 (Fb**3**:0-8, Fb**3**:0-15) are characterized by an over-accumulation of compounds from cluster C, and specially compounds 96 and 99. NIL Fb**4**:0-44 showed a different profile for the two years showing a strong under-accumulation of most volatiles in 2012 harvests and average-mild profile in 2013 so it is not possible to trace an accumulation trend in this genotype. Differences are very evident in lines with introgressions in LG5, as lines harboring introgressions at the beginning of the LG5 (Fb**5**:0-11, Fb**5**:0-35, Fb**5**:0-35-56) under accumulate compounds from clusters A and D2 and show nearly average profiles for compounds in clusters B, C and D1, however lines with introgressions at the end of LG5 (Fb**5**:39-76, Fb**5**:41-76, Fb**5**:50-76) show mostly positive ratios profile, a clear over-accumulation of compounds 27 and 16 (cluster C), and a clear under-accumulation in cluster B and compounds 17 and 18 (cluster C). Line Fb**5**:0-76 that harbors a *F. bucharica* homozygous introgression that covers the whole LG5 presents an intermediate behavior and under-accumulates compounds in cluster A, B, D2 and compounds 17 and 18 from cluster C and over-accumulates compounds 27 and 16. NIL Fb6:0-5 harbors a very small introgression and presents a mild profile for all clusters, similar to RV. Lines with

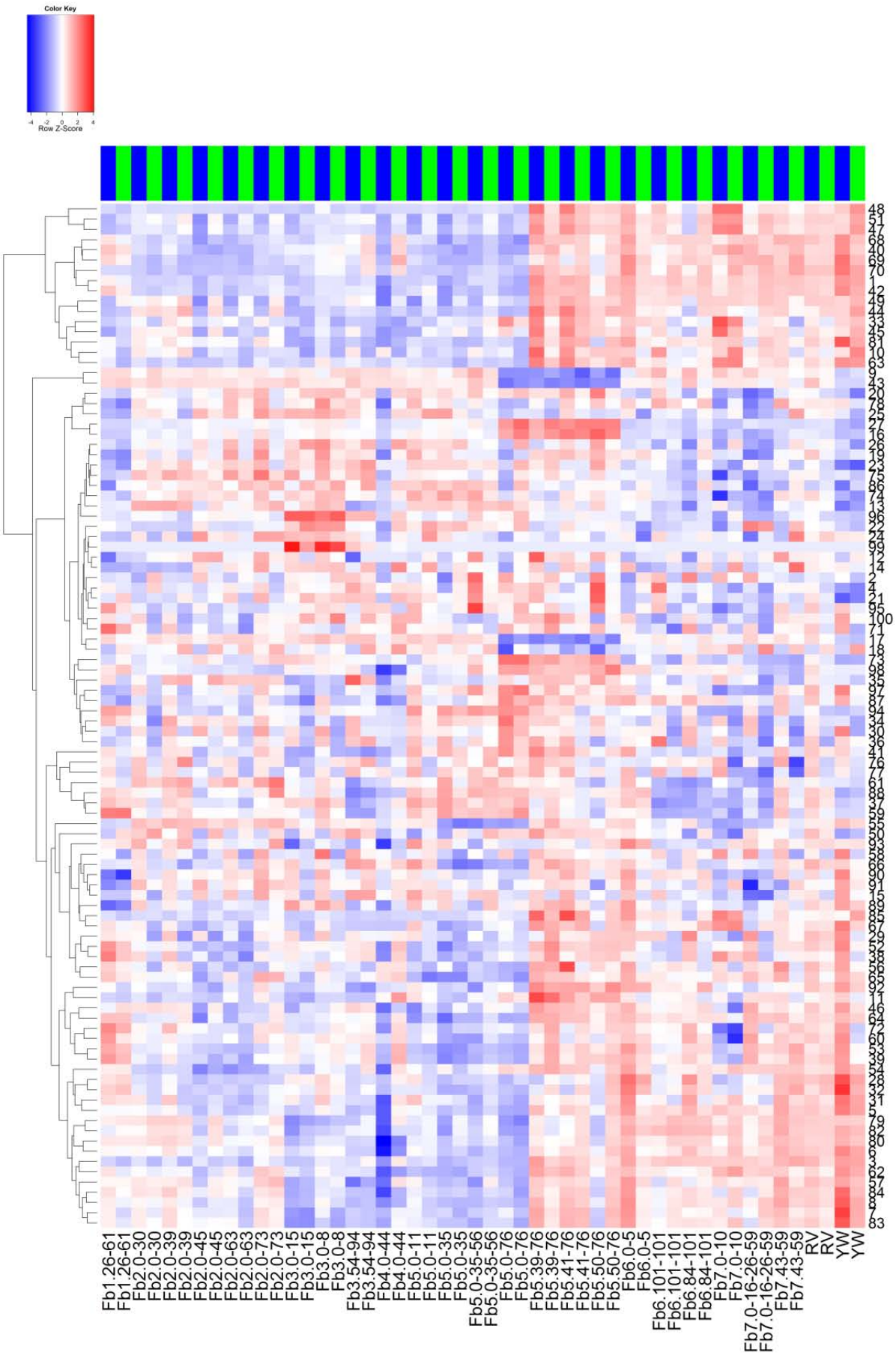

Color Key
-4 -2 0 2 4
Row Z-Score
Fb1.26-61
Fb1.26-61
Fb2.0-30
Fb2.0-30
Fb2.0-39
Fb2.0-39
Fb2.0-45
Fb2.0-45
Fb2.0-63
Fb2.0-63
Fb2.0-73
Fb2.0-73
Fb3.0-15
Fb3.0-15
Fb3.0-8
Fb3.0-8
Fb3.54-94
Fb3.54-94
Fb4.0-44
Fb4.0-44
Fb5.0-11
Fb5.0-11
Fb5.0-35
Fb5.0-35
Fb5.0-35-56
Fb5.0-35-56
Fb5.0-76
Fb5.0-76
Fb5.39-76
Fb5.39-76
Fb5.41-76
Fb5.41-76
Fb5.50-76
Fb5.50-76
Fb6.0-5
Fb6.0-5
Fb6.101-101
Fb6.101-101
Fb6.84-101
Fb6.84-101
Fb7.0-10
Fb7.0-10
Fb7.0-16-26-59
Fb7.0-16-26-59
Fb7.43-59
Fb7.43-59
RV
RV
YW
YW

Figure CIII. 1 Relative volatile compounds profiling in heatmap and hierarchical clustering (HCA) of compounds. Ratio values of all studied volatile compounds per genotype are represented in the heatmap in a blue (negative) to red (positive) scale. Compounds are numerically codified as specified in Table CIII. 1. Genotypes include NILs that were analyzed both years, RV and YW and the top bar identifies sample's harvest year: 2012 (blue) and 2013 (green). The HCA and dendrogram of volatile compounds was done according to metabolites ratio distances (euclidean distance, complete linkage). Clusters are indicated with capital letters in the dendrogram and in Table CIII. 1.

introgressions at the end of LG6 (Fb6:84-101 and Fb6:101-101) show an intermediate profile and a slight under-accumulation in volatiles from clusters C and D1. Lines that harbor introgressions in LG7 show a tendency towards over-accumulation in volatiles from cluster A and an under-accumulation of compounds from cluster C and D1. Fb7:0-10 also shows and under accumulation of several compounds from D2. Numeric mean log2 ratios for all samples analyzed each year are provided in Supplemental Table CIII. 1.

In order to assess the stability of the volatile accumulation pattern within the NIL collection, RV and YW, we calculated the Pearson's correlation of volatile levels between harvests (2012 and 2013). 82 of the 100 compounds showed significant (p-value < 0.05) positive year to year correlation (Table CIII. 2) and importantly, 75 of them correlated with a strict p-value (< 0.01). These very significant correlation indexes ranged from 0.51 (α-ionone) to 0.94 (nerolidol) being the median 0.74, there are three compounds besides nerolidol with correlation values above 0.90 (1-decanol, (*Z*)-3-hexenal and (*E*)-2-hexenyl acetate). Therefore, accumulation pattern of most volatile compounds across the NIL collection was consistent both years thus pointing to an important genotypic effect under their accumulation. All 'key compounds' contributing to strawberry fruit aroma present significant positive year to year correlation except furaneol and butyl acetate. Stability of the lines was demonstrated previously with a (poly)-phenolic profiling of the NIL collection (Urrutia *et al*. accepted) and although the correlation between genotypes attending to volatile profiling is lower, the median of all genotypes is above 0.70. To deeper study the relations between the different clusters of volatile compounds we performed a cluster network analysis (CNA) based on correlations between compounds (Figure CIII. 2). Individual correlation coefficients and significant values are provided in Supplemental Table CIII. 2. CNA reveals that cluster A and D2 are tightly and positively correlated. This can be explained considering that both clusters group most of the identified alcohols and esters and that alcohols are usually transformed to the alkyl group of esters. Compounds in cluster B (9 and 43) correlate positively with (*E*)-2-hexenal (17) and negatively with (Z)-3-hexenal (27) from cluster C. These compounds are derived from hexene and these correlations suggest some kind of *cis-trans* regulations. Grouped in cluster C are also most terpenoids among which the strongest correlation is detected between linalool (96) and nerolidol (99). Cluster D1 is not related with clusters D2 or A, this might be due to the fact that are mainly benzoic-ring derived compounds that may share independent biosynthetic pathways, for instance acetophenone (88) or methyl cinnamate (61) can be transformed to a benzyl alcohol that can be esterified into benzyl acetate (37). Furans, allocated in clusters C and D1 are not highly correlated between them. The alkyl furans in cluster D1, 2,1-

pentenylfuran (74) and 2-pentylfuran (75), are correlated between them and to several 5 to 8 C aldehydes (19,20,23,26). While the link between mesifurane (76) and furaneol (77) (cluster C) is very light but still present. Furaneol is also linked to acetates 35 and 36 from cluster D1.

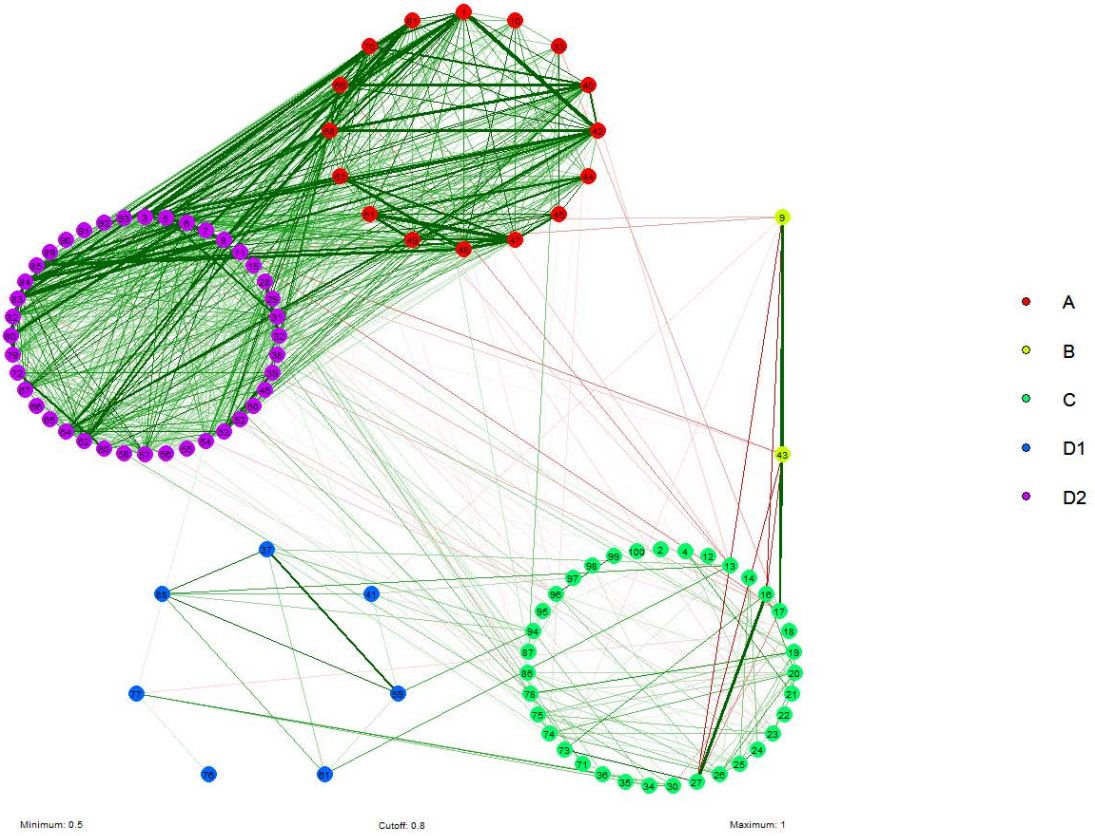


Figure CIII. 2 Cluster network analysis (CNA). Metabolites are represented as nodes colored according to clusters defined by HCA as specified by the legend. Positive (green) and negative (red) correlations with absolute values >|0.5| are represented as links between the nodes. Links representing absolute correlations >|0.8| are wider the stronger they are and have the maximum color saturation. Absolute correlation <|0.8| are vaguer the weaker they are and have the smallest wideness.

With the aim of better understanding volatile variability among the NIL collection, RV and YW, we performed a principal component analysis (PCA) (Figure CIII. 3). Lines were distributed along the first two components (PC1 and PC2) that explained the 31% and the 13% of the observed variability respectively without forming clear groups and remaining near the axis (Figure CIII. 3A). These suggesting that distribution of most of the volatiles is continuous and that differences in the aromatic pattern between the NIL collection are restricted to punctual or small subsets of metabolites rather than to big metabolites clusters. This reinforces the idea that aroma is a complex trait controlled by many locus widespread in all the genome.

From a closer perspective we can observe that NIL tend to separate across PC1 according to their introgressed region (lines with introgressions in LG6, LG7 and the end of LG5, RV and YW

have negative PC1 scores and lines with introgressions in LG2, LG3, LG4 and the beginning of LG5 have positive PC1 scores). While PC2 divides the samples again according to their genotype (lines with introgressions in LG5 have positive PC2 scores while lines with introgressions in LG1, LG6 and LG7 have negative scores) but also according to the harvest year (specially lines with introgressions in LG2 and LG3 but also RV). This suggests that a higher proportion of the observed variability between NIL is due to genotype rather than to the environment and that volatile accumulation in NIL with introgressions in LG2 and LG3 are especially susceptible to environmental conditions.

Attending to loading plots (Figure CIII. 3B), linalool, octanal and 6-methyl-5-hepten-ona together with most esters and alcohols account for variability across PC1. Compounds contributing mostly to variability across PC2 are *cis* isomers derived from hexane ((*E*)-2-hexenal, (*E*)-2-hexenyl acetate and (*E*)-2-hexen-1-ol) and the aldehyde (*E*)-2-nonenal together with (*E*)-2-heptenal, (*Z*)-3-hexenal and myrtenol.

Among all samples, YW is the one that provides more variability to the aroma profile (this was expected considering that is not genetically related with the NIL collection). It is characterized by “pineapple-like” fruity aroma what might be due to a specific combination in esters, γ-decalactone, alcohols and ketones.

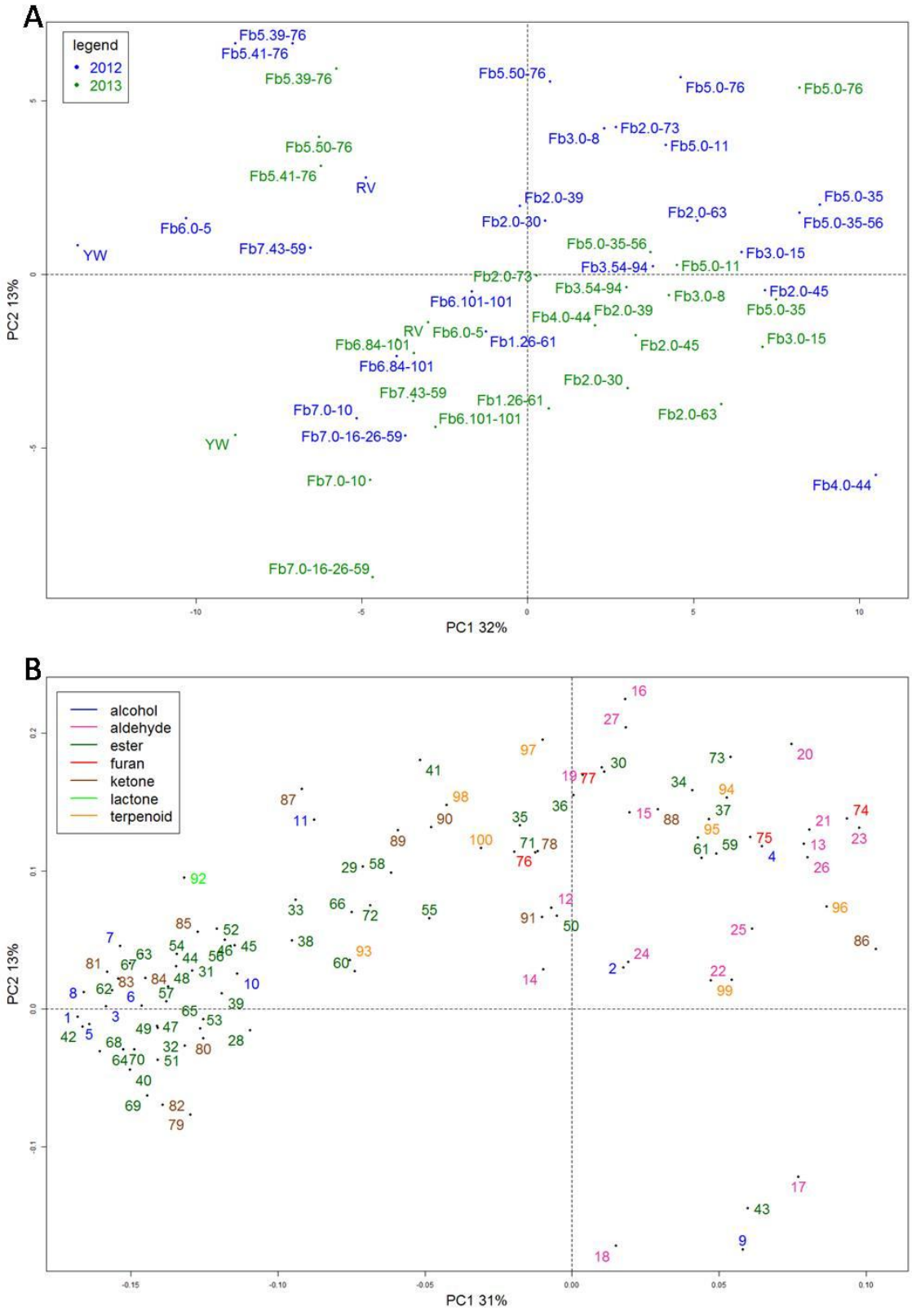


Figure CIII. 3 PCA scores and loading plot. PCA scores plot (A) and loading plot (B) of PC1 vs. PC2. NIL, RV and YW are colored according to the harvest year as specified in the legend **A**. Compounds are coded as specified in Table CIII. 1 and colored according to their chemical family as specified in the legend **B**.

Table CIII. 2. Compounds correlation between harvests. Correlation values were calculated using average genotype values for both harvests. Asterisk after the values indicate significance at different thresholds: p-value <0.05 '*' , p-value <0.01 '**', p-value <0.001 '***'. No significant correlations are indicated by 'ns'.

compound	correlation between harvest	significance
1-decanol	0.90	***
1-hexanol	-0.22	ns
1-octanol	0.78	***
1-penten-3-ol	0.33	ns
2-heptanol	0.72	***
2-nonanol	0.83	***
2-tridecanol T	0.62	***
2-undecanol T	0.78	***
(E)-2-hexen-1-ol	0.82	***
Ethanol	0.41	*
Eugenol	0.81	***
3,4-dimethylbenzaldehyde	0.42	*
Benzaldehyde	0.78	***
Decanal	-0.01	ns
(E)-2-decenal	0.57	**
(E)-2-heptenal	0.9	***
(E)-2-hexenal	0.88	***
(E)-2-nonenal	0.47	*
(E)-2-octenal	0.52	**
(E)-2-pentenal	0.58	**
(E,Z)-2,4-heptadienal	0.34	ns
Heptanal	0.54	**
Hexanal	0.62	**
Nonanal	0.09	ns
Octanal	0.41	*
Pentanal	0.12	ns
(Z)-3-hexenal	0.94	***
1-methylbutyl butanoate	0.76	***
1-methylethyl butanoate	0.16	ns
1-methylethyl acetate	0	ns
1-methylhexyl acetate	0.31	ns
1-methyloctyl butanoate	0.68	***
2,3-butanedioldiacetate T	0.75	***
2 -methylbutyl acetate	0.6	**
3-methyl-2-butenyl acetate	0.62	**
3-methylbutyl acetate	0.27	ns
Benzyl acetate	0.75	***
Butyl acetate	0.25	ns

compound	correlation between harvest	significance
Butyl butanoate	0.63	***
Butyl hexanoate	0.77	***
Cinnamyl acetate	0.67	***
Decyl acetate	0.88	***
(E)-2-hexenyl acetate	0.92	***
Ethyl 2-hexenoate	0.65	***
Ethyl acetate	0.54	**
Ethyl butanoate	0.56	**
Ethyl decanoate	0.77	***
Ethyl dodecanoate	0.79	***
Ethyl hexanoate	0.68	***
Ethyl methylthioacetate T	0.35	ns
Ethyl octanoate	0.76	***
Hexyl acetate	0.52	**
Hexyl butanoate	0.72	***
Hexyl hexanoate	0.63	***
Methyl-2-aminobenzoate T	0.84	***
Methyl-2-hexenoate	0.44	*
Methyl 3-hydroxyoctanoate T	0.54	**
Methyl acetate T	-0.01	ns
Methyl benzoate	0.74	***
Methyl butanoate	0.49	*
Methyl cinnamate T	0.71	***
Methyl decanoate	0.86	***
Methyl dodecanoate	0.87	***
Methyl hexanoate	0.81	***
Methyl octanoate	0.81	***
Myrtenyl acetate	0.74	***
Nonyl acetate	0.59	**
Octyl acetate	0.79	***
Octyl butanoate	0.68	***
Octyl hexanoate	0.77	***
Pentyl acetate	0.3	ns
Propyl butanoate	0.48	*
(Z)-3-hexenyl acetate	0.8	***
2,1-pentenyl_furan	0.72	***
2-pentylfuran	0.61	**
Furaneol	-0.27	ns
Mesifurane	0.69	***
1-penten-3-one	0.54	**
2-heptanone	0.67	***
2-nonanone	0.83	***
2-pentadecanone T	0.71	***

compound	correlation between harvest	significance
2-pentanone	0.76	***
2-tridecanone T	0.64	***
2-undecanone T	0.66	***
4-tridecanone T	0.76	***
6-methyl-5-hepten-2-one	0.63	***
Acetone	0.74	***
Acetophenone	0.76	***
α-ionone	0.51	**
β-ionone	0.55	**
(Z)-geranyl acetone	0.32	ns
γ-decalactone	0.75	***
α-farnesene	0.55	**
α-pinene	0.76	***
Limonene	0.31	ns
Linalool	0.78	***
Myrtenol	0.64	***
Nerol	0.84	***
Nerolidol	0.95	***
Terpineol	0.25	ns

Genotypic and environmental effect over volatile compounds accumulation

After observing a high variability in the accumulation of the volatile compounds in the strawberry ripe fruit among the NIL collection, we performed an analysis of variance (ANOVA). In order to explore to which extent the genotype (G), the environment (E) and their interaction (GxE) are affecting to the observed variability, we fitted the G+E+GxE model considering the harvests in different years as different environments.

The ANOVA test revealed that there were several combinations of factors influencing variability depending on the compound, but it was clear that the G was the most important factor (Figure CIII. 4). Concerning NIL collection, G significantly contributed (p-value<0.05) to variability in 98 out of the 100 studied volatile compounds (Supplemental Table CIII. 3). Among them, there were 33 compounds that were significantly influenced by the three factors G, E and GxE, 16 were influenced by G and E but not by GxE, 33 were influenced by G and GxE but not for E and interestingly there were 17 compounds that were influenced only by the G including the 'key compounds' methyl 2-aminobenzoate, nerolidol, γ-decalactone, ethyl butanoate and (*Z*)-3-hexenal. The two compounds that were not influenced by G were the furaneol that was influenced by E and GxE and the 3,4-dimethylbenzaldehyde that was not influenced by any of the factors.

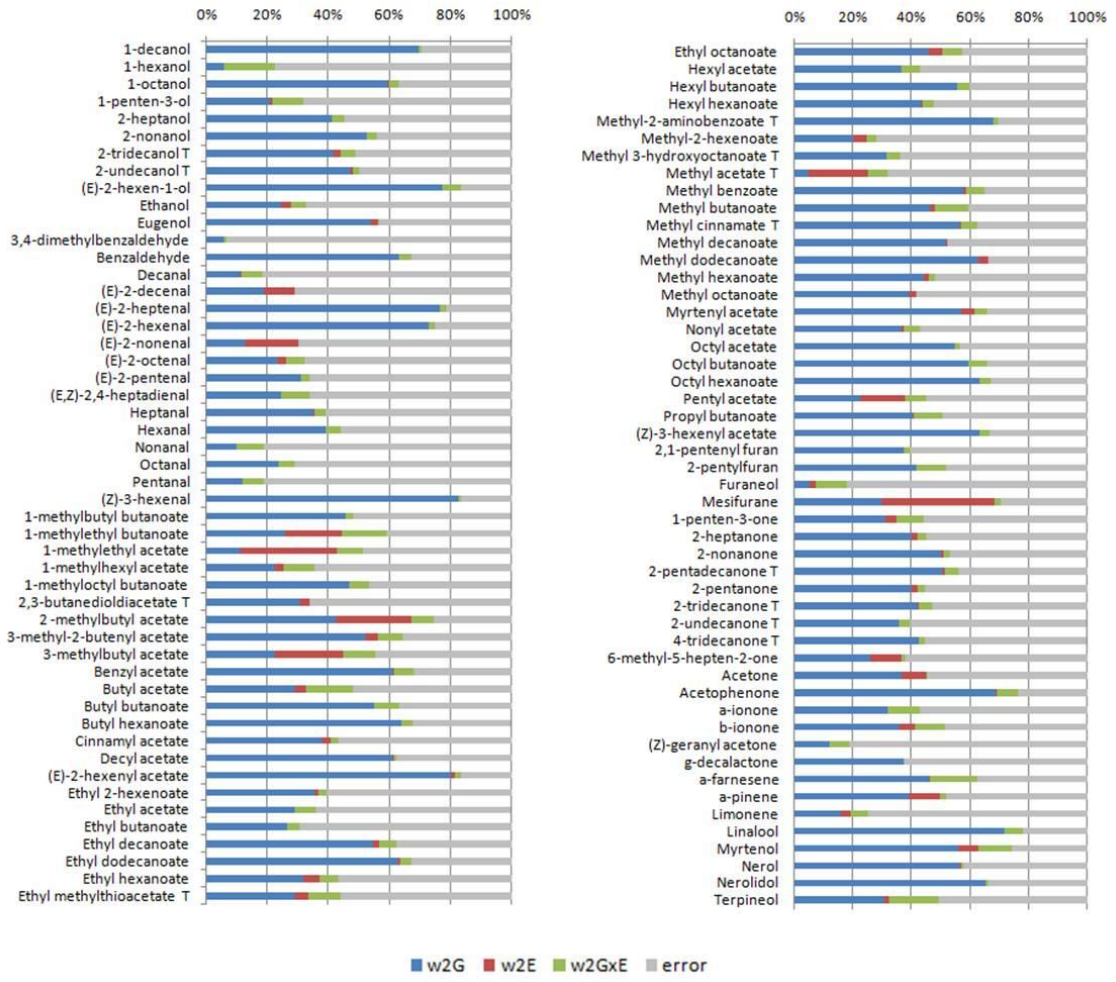


Figure CIII. 4 ω2 values. Percentage of the observed variability attributable to each of the factors: genotype (G), environment (E), their interaction (GxE) or to the error.

Besides the significant influence of each of the factors, they differ in the actual percentage of variability they account for. In general terms, the genotype (with exceptions) has a predominant role over the variability than the environment or the GxE interaction but there is also a high proportion of the observed variability that cannot be attributed to any of the factors and it is what we call error (Figure CIII. 4). See Supplemental Table CIII. 3 for precise quantification. Specifically 35 compounds, including 10 of the 'key compounds' (17, 27, 39, 43 55, 61, 66, 73, 96, 99) due more than 50% of their variability to the genotype and the effect reaches the 70% for 6 of them, including 4 'key compounds' ((*E*)-2-hexenal, (*Z*)-3-hexenal, (*E*)-2-hexenyl acetate and linalool). However there are also some 'key compounds' like butyl acetate, ethyl butanoate, ethyl hexanoate, hexyl acetate, furaneol and γ-decalactone for whom more than 50% of the observed variability can't be assigned to any of the factors (G, E, GxE). The effect of the environment is much less evident and only surpasses 20% of the observed variability for 5 compounds including one of the 'key compounds' (mesifurane).

Taken together these exploratory results we accept that NIL collection is a good tool for the genetic dissection of the volatile accumulation in strawberry fruit. There is a wide range of

variability within each compound among the NIL collection, that is normally concentrated in a specific set of NIL harboring common introgressed regions (Figure CIII. 1), and a high proportion of this variability seems to be genetically controlled.

QTL analysis

In order to assess which regions of the genome could be responsible for the variations in volatile ratios, a QTL analysis was performed with all samples collected each year independently. For each metabolite we performed individual t-test for NIL ratios against the recurrent parental *F. vesca* RV ratio. Resulting t-test p-values where corrected for multi-testing and finally a corrected p-value<0.05 was accepted as significance threshold. A QTL for a metabolite was assigned to a specific genetic region when all NIL harboring a common *F. bucharica* introgression in that region showed a significant difference in the metabolite ratio compared with *F. vesca* RV in the same direction (Table CIII.3). QTL that were mapped to the same region in 2 harvests were considered stable QTL. Additionally, in order to evaluate the percentage of variability explained by each QTL regarding the whole NIL collection, an interval mapping analysis was conducted. QTL that explained around 20% of the observed variability and had a LOD score >1.80 were considered strong QTL (Figure CIII. 5). Non stable QTL (detected only one harvest) were considered only if they accounted for more than 20% of the observed variability that year.

We were able to map a total of 126 QTL from 81 different compounds the majority of which were esters (40) followed by aldehydes (12), alcohols (11), ketones (8), terpenoids (7) and furans (3). Among the detected QTL, 102 were stable (present in two years) and 24 were present only in one harvests. Besides, 50 of the stable QTL were also strong (explain around 20% of the variability and LOD>1.8) and considered major QTL. The effect of the QTL over the ratio regarding *F. vesca* RV ratio was positive in 30 of them and negative in 96 of them.

From the total of strong and stable QTL, 25 corresponded to compounds that mapped to a single locus accounting for most of the observed variability, including 9 ‘key compounds’ (linalool, nerolidol, mesifurane, methyl hexanoate, methyl cinnamate, (E)-2-hexenal, (E)-2-hexenyl acetate, (Z)-3-hexenal and (Z)-3-hexenyl acetate),meaning that a big proportion of their variability is restricted to a single specific genomic region in the genetic background of *F. vesca* RV. These major QTL confer a good starting point for the study of the accumulation of these compounds. There were also three compounds that mapped two major QTL: the ‘key compound’ methyl 2-aminobenzoate, nerol and 3-methyl-2-butenyl acetate.

The remaining 19 stable major QTL correspond to compounds for which additional minor QTL were mapped (i.e. the ‘key compounds’ methyl butanoate, butyl butanoate and myrtenyl acetate and other compounds such as 1-decanol, benzyl acetate, hexyl butanoate, butyl hexanoate, ethyl octanoate, α-farnesene, and acetophenone among others). Most of the compounds for which strong and stable QTLs were mapped showed a high effect of the genotype and low effect of the environment over their variability (Figure CIII. 3) as expected. However there are some exceptions being the most significant the mesifurane, that although is clearly influenced by the environment (38%), the effect of the genotype (30%) was enough to map a QTL. There were also some compounds, mainly aldehydes (octanal, nonanal, decanal, (E)-2-octenal, (E)-2-nonenal and (E)-2-decenal) but also alcohols and ketones (ethanol, 1-

pentent-3-ol, 1-hexanol, 2-heptanol, acetone, 1-penten-3-one, 2-pentanone and 2-nonanona) that only mapped one-year QTL and not stable ones. Most of these compounds show a low effect of the genotype over their accumulation ratios, and low correlation between harvests resulting in an unpredictable accumulation behavior. Nineteen of the detected volatile compounds were not mapped to any QTL, probably because their correlation between harvests and the genetic effect over their variability was low (i.e. furaneol, (Z)-geranyl acetone, methyl acetate, 1-methylethyl acetate and 1-methylethyl butanoate) or because despite of showing a moderate genotypic effect over their accumulation, their segregation was not sufficient to be statistically significant (i.e. γ-decalactone, 2-heptanone, 2-tridecanone, 4-tridecanone, 2-pentadecanone, hexyl acetate, nonyl acetate and heptanal among others).

There are QTL in all seven linkage groups, however some regions gather several QTL for different compounds that co-localize, this may indicate that those compounds share common pathways or regulatory signals. In many cases, as expected, these co-localizations match with the clusters defined by HCA and also with the compounds that showed to be highly correlated by CNA.

At the end of LG1 (LG1:26-61 cM) (Figure CIII. 5) there are two negative major QTL that explain 16-36% and 18-30% of the observed variability for α- and β-ionone respectively and one positive QTL that explains 8-23% of the variability of pentyl acetate (Table CIII. 3).This region also harbors other minor QTL including one positive QTL for the 'key compound' butyl acetate.

In LG2 there are two regions that gather QTL, the fist one at the beginning of the linkage group (LG2:0-30 cM) harbors five strong stable QTL, two positive-ones for the 'key compound' methyl cinnamate and 2-pentylfurane and three negative-ones for hexyl hexanoate, octyl butanoate and octyl hexanoate. The second region is LG2:39-45 that harbors one stable QTL for the accumulation of 3-methyl-2-butenyl acetate.

Two strong and stable QTL were mapped at LG3:0-8 cM for the 'key' terpenoids nerolidol and linalool. This compounds showed a nearly absence (RV) - presence (NILs harboring introgressions at LG3:0-8 cM) segregating pattern. At the end of LG3 (LG3:54-94 cM) we also found a positive stable QTL for the accumulation of 3-methyl-2-butenyl acetate and one-year QTL for 3-methylbutyl acetate and also a major negative QTL for the accumulation of acetophenone and two minor QTL for cinnamyl acetate and methyl benzoate, from cluster D1.

At LG4 two major stable QTL were mapped for the terpenoids nerol (LG4:9-20 cM) and α-farnesene (LG4:20-44 cM). Besides it harbors five minor stable QTL for 1-decanol, 2-undecanol, 2-undecanone, decyl acetate and acetophenone and 8 one-year QTL for decanal, 2-heptanol, 2-nonanol, 2-pentanone, 2-nonanone, acetone, 1-methylhexyl acetate and methyl decanoate. Mapped QTL reveal a tendency towards long carbon chains volatiles (9-11C) in this region (Figure CIII. 5).The effect of the environment over the NILs with *F. bucharica* introgressions at the beginning of LG4 seems to be especially important.

The majority of the strong and stable QTL are located in LG5. At the beginning of the LG5 (LG5:0-11 cM) there is one positive QTL for the terpenoid α-pinene and four negative QTL for the alcohol 1-decanol and the esters methyl octanoate, methyl decanoate and butyl hexanoate from clusters A and D2. At the central region LG5:11-35 cM strong and stable QTLs with a

negative effect over the ratios of 9 esters, five of them being the 'key compounds' methyl 2-aminobenzoate, myrtenyl acetate, methyl butanoate, butyl butanoate and methyl hexanoate. This region might be fundamental for strawberry aroma as it controls simultaneously the accumulation of several important 'key compounds'. The end of the LG5 (LG5:50-76 cM) also harbors an interesting bunch of QTL. It gathers four QTLs related with the 'key compounds' associated to green-fresh scent. Positive QTL are mapped for (*Z*)-3-hexenal and (*Z*)-3-hexenyl acetate and negative QTL are mapped for their respective trans-2 isomers (*E*)-2-hexenal and (*E*)-2-hexenyl acetate, this reflects the negative correlation between the isomers what might be indicating that *F. bucharica* alleles at this region favor some kind of regulation towards the accumulation of *cis-3* isomers of hexenal and hexenyl acetate. In addition there are three other positive QTL for the terpenoid nerol, the alcohol eugenol and the aldehyde (E)-2-heptenal and one negative QTL for (E)-2-hexen-1-ol.

No strong QTL were mapped in LG6, but region LG6:101-101 cM harbors four minor QTL reflecting a negative effect in aromatic (harboring a benzoic ring) compounds ratio and also a negative effect in myrtenyl acetate accumulation.

LG7, and especially the beginning of the linkage group, seem to be important for volatile accumulation as it accumulates 13 strong and stable QTL, two of them for rare 'key compounds' methyl 2-aminobenzoate (LG7:0-10 cM) and mesifurane (LG7:26-43 cM). Additionally, at region LG7:0-10 cM we could also find four stable and one one-year positive QTL for the accumulation of long carboxylic chain esters from cluster A, and 7 negative stable QTL for terpenoids (limonene and myrtenol), esters, one aldehyde and one furan from clusters C and D2.

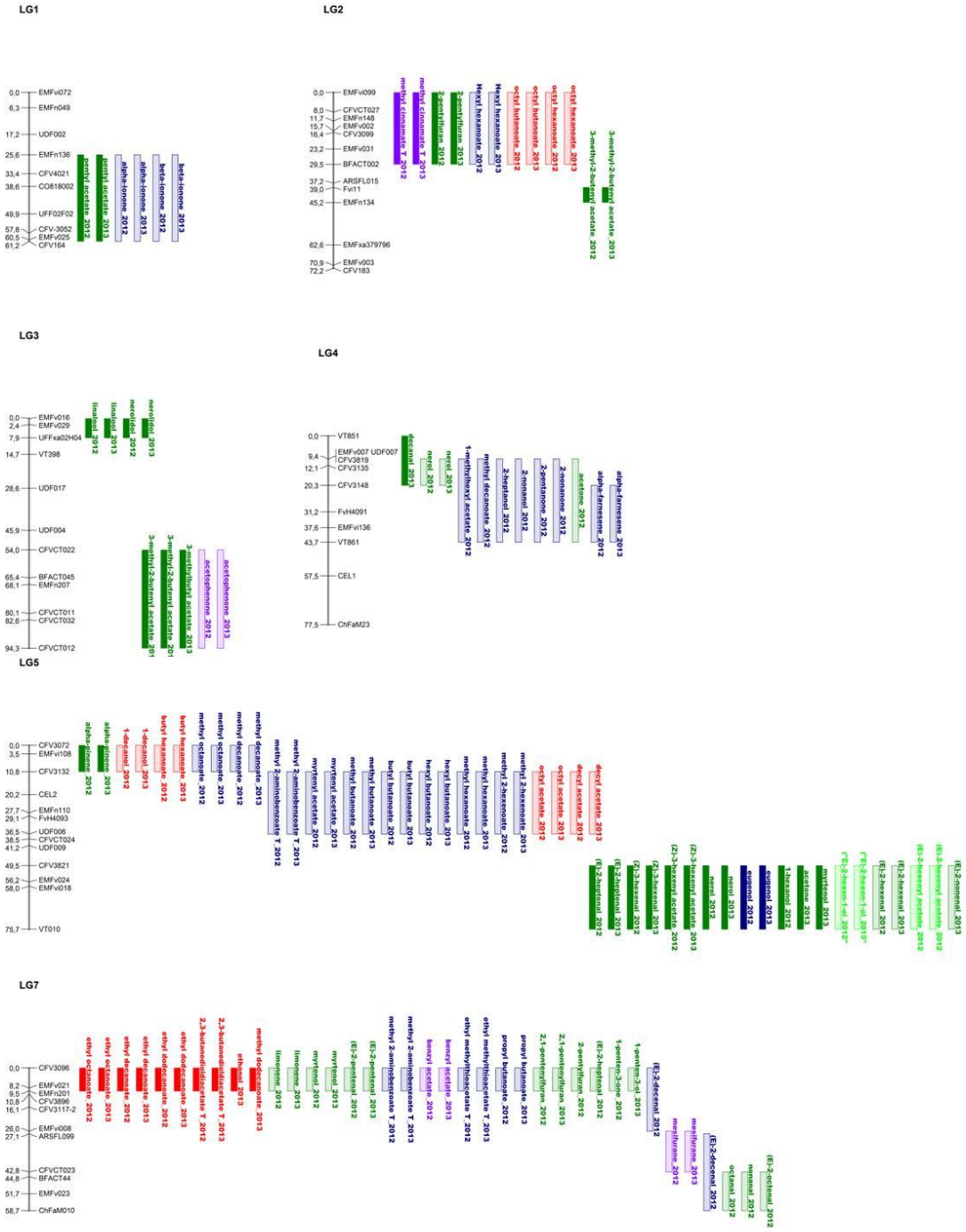


Figure CIII. 5 volatile QTL. Graphical representation of strong QTL mapped. Represented QTL were found to be significantly different (corrected p-value < 0.05) from the recurrent parental (*F. vesca* RV), in the same direction, in at least one harvests for all the NILs harboring the introgressed region and explained around 20% of the variability regarding the NIL collection. QTL names correspond to the volatile compound affected and the year it was detected (2012 or 2013). The colored bars symbolize the cluster to which the compound belongs by HCA as in Figure CIII. 2 (cluster A = red, cluster B = pale green, cluster C= apple green, cluster D1 = blue, cluster D2 =purple) and the positive or negative effect of the QTL over the ratio regarding *F. vesca* RV is represented by the full or light color bars respectively. For locating the QTL, the LG and position in cM of the microsatellites (SSRs) used for genotyping are given.

Table CIII. 3 QTL for volatile compounds detected in a *F. vesca* NIL collection. Detected QTL listed by compound's alphabetical order. The position of the QTL (LG number followed by the start and end position in cM), the positive (up) or negative (down) effect of the QTL over the metabolite's ratio compared with *F. vesca* RV, the NIL harboring the shorter *F. bucharica* introgression (in cM) that includes the QTL, the results of the t-test (corrected p-value) and interval mapping analysis (LOD score), the percentage of variance explained by the QTL regarding the NIL collection and the stability of the QTL (detected in 1 or 2 harvests) are provided.

Table CIII. 3

	Compound	direction	qtl (cM)	shorter NIL	t-test (corrected p.value)	LOD	% explained variance	stable
	(E)-2-decenal	down	LG7:0-26	Fb7:0-27	<0,05	4,64	46%	1
	(E)-2-decenal	down	LG7:27-59	Fb7:0-59	<0,05	6,14	56%	1
	(E)-2-heptenal	down	LG7:0-10	Fb7:0-10	<0,05	2,10	25%	1
	(E)-2-heptenal	up	LG5:50-76	Fb5:50-76	<0,05	15,17	46-87%	2
	(E)-2-hexen-1-ol	down	LG5:50-76	Fb5:50-76	<0,05	15,66	63-88%	2
→	(E)-2-hexenal	down	LG5:50-76	Fb5:50-76	<0,05	16,04	74-88%	2
→	(E)-2-hexenyl acetate	down	LG5:50-76	Fb5:50-76	<0,05	16,75	82-89%	2
	(E)-2-nonenal	down	LG5:50-76	Fb5:50-76	<0,05	4,59	46%	1
	(E)-2-octenal	down	LG7:52-59	Fb7:52-59	<0,05	2,61	30%	1
	(E)-2-pentenal	down	LG7:0-10	Fb7:0-10	<0,05	2,91	30-32%	2
→	(Z)-3-hexenal	up	LG5:50-76	Fb5:50-76	<0,05	14,76	58-86%	2
→	(Z)-3-hexenyl acetate	up	LG5:50-76	Fb5:50-76	<0,05	5,68	44-53%	2
	1-decanol	down	LG5:0-11	Fb5:0-11	<0,05	1,86	16-22%	2
	1-decanol	down	LG3:8-15	Fb3:0-15	<0,05	<1,80	2-5%	2
	1-decanol	down	LG4:20-44	Fb4:0-44	<0,05	<1,80	1-5%	2
	1-hexanol	up	LG5:50-76	Fb5:50-76	<0,05	1,99	23%	1
	1-methylbutyl butanoate	down	LG5:11-35	Fb5:0-35	<0,05	<1,80	10-11%	2
	1-methylbutyl butanoate	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	2-3%	2
	1-methylhexyl acetate	down	LG4:9-44	Fb4:0-44	<0,05	2,68	30%	1
	1-methyloctyl butanoate	down	LG2:0-30	Fb2:0-30	<0,05	<1,80	7-13%	2
	1-methyloctyl butanoate	down	LG5:11-35	Fb5:0-35	<0,05	<1,80	17-21%	2
	1-octanol	down	LG1:26-61	Fb1:26-61	<0,05	<1,80	1-3%	2
	1-octanol	down	LG2:0-30	Fb2:0-30	<0,05	<1,80	10-16%	2
	1-octanol	down	LG5:11-35	Fb5:0-35	<0,05	<1,80	5-17%	2
	1-penten-3-ol	down	LG7:0-10	Fb7:0-10	<0,05	5,25	51%	1
	1-penten-3-one	down	LG7:0-10	Fb7:0-10	<0,05	3,97	42%	1
	2,1-pentenyl furan	down	LG7:0-10	Fb7:0-10	<0,05	4,45	36-45%	2
	2,3-butanedioldiacetate T	up	LG7:0-10	Fb7:0-10	<0,05	2,49	6-28%	2
	2-heptanol	down	LG4:9-44	Fb4:0-44	<0,05	1,94	23%	1
	2-methylbutyl acetate	down	LG7:43-59	Fb7:43-59	<0,05	<1,80	7-8%	2
	2-nonanol	down	LG1:26-61	Fb1:26-61	<0,05	<1,80	1%	2
	2-nonanol	down	LG5:11-35	Fb5:0-35	<0,05	<1,80	13-15%	2

Table CIII. 3

	Compound	direction	qtl (cM)	shorter NIL	t-test (corrected p.value)	LOD	% explained variance	stable
	2-nonanol	down	LG4:9-44	Fb4:0-44	<0,05	4,95	48%	1
	2-nonanone	down	LG4:9-44	Fb4:0-44	<0,05	6,48	58%	1
	2-pentanone	down	LG4:9-44	Fb4:0-44	<0,05	2,09	25%	1
	2-pentylfuran	down	LG7:0-10	Fb7:0-10	<0,05	3,16	35%	1
	2-pentylfuran	up	LG2:0-30	Fb2:0-30	<0,05	3,16	21-35%	2
	2-tridecanol T	down	LG3:8-15	Fb3:0-15	<0,05	<1,80	3-17%	2
	2-undecanol T	down	LG4:20-44	Fb4:0-44	<0,05	<1,80	1-15%	2
	2-undecanol T	down	LG5:11-35	Fb5:0-35	<0,05	<1,80	18-20%	2
	2-undecanone T	down	LG4:20-44	Fb4:0-44	<0,05	<1,80	2-31%	2
	3-methyl-2-butenyl acetate	up	LG3:54-94	Fb3:54-94	<0,05	2,07	11-24%	2
	3-methyl-2-butenyl acetate	up	LG2:39-45	Fb2:39-47	<0,05	5,54	5-49%	2
	3-methylbutyl acetate	up	LG3:54-94	Fb3:54-94	<0,05	3,29	36%	1
	acetone	down	LG4:9-44	Fb4:0-44	<0,05	2,24	26%	1
	acetone	up	LG5:50-76	Fb5:50-76	<0,05	2,99	33%	1
	acetophenone	down	LG3:54-94	Fb3:54-94	<0,05	1,80	14-21%	2
	acetophenone	down	LG4:0-20	Fb4:0-20	<0,05	<1,80	6-15%	2
	acetophenone	down	LG6:101-101	Fb6:101-101	<0,05	<1,80	14-20%	2
	acetophenone	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	8-19%	2
	α-farnesene	down	LG4:20-44	Fb4:0-44	<0,05	2,29	10-26%	2
	α-farnesene	down	LG3:8-15	Fb3:0-15	<0,05	<1,80	16%	2
	α-ionone	down	LG1:26-61	Fb1:26-61	<0,05	3,35	16-36%	2
	α-pinene	up	LG5:0-11	Fb5:0-11	<0,05	4,20	35-42%	2
	benzaldehyde	up	LG2:0-30	Fb2:0-30	<0,05	<1,80	13-19%	2
	benzyl acetate	down	LG7:0-10	Fb7:0-10	<0,05	2,26	15-26%	2
	benzyl acetate	down	LG6:101-101	Fb6:101-101	<0,05	<1,80	18-21%	2
	β-ionone	down	LG1:26-61	Fb1:26-61	<0,05	2,58	18-30%	2
→	butyl acetate	up	LG1:26-61	Fb1:26-61	<0,05	<1,80	6-15%	2
→	butyl butanoate	down	LG5:11-35	Fb5:0-35	<0,05	3,51	30-38%	2
→	butyl butanoate	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	1-2%	2
	butyl hexanoate	down	LG5:0-11	Fb5:0-11	<0,05	3,26	30-35%	2
	butyl hexanoate	down	LG2:0-30	Fb2:0-30	<0,05	<1,80	19-20%	2
	butyl hexanoate	up	LG7:0-10	Fb7:0-10	<0,05	<1,80	11-14%	2
	cinnamyl acetate	down	LG3:54-94	Fb3:54-94	<0,05	<1,80	5-8%	2
	cinnamyl acetate	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	4-6%	2
	decanal	up	LG4:0-20	Fb4:0-20	<0,05	2,51	29%	1
	decyl acetate	down	LG5:11-35	Fb5:0-35	<0,05	1,80	20-22%	2
	decyl acetate	down	LG4:20-44	Fb4:0-44	<0,05	<1,80	1-16%	2
	ethanol	up	LG7:0-10	Fb7:0-10	<0,05	3,36	36%	1
	ethyl 2-hexenoate	down	LG5:11-35	Fb5:0-35	<0,05	<1,80	15-18%	2
→	ethyl butanoate	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	3-10%	2
	ethyl decanoate	down	LG1:26-61	Fb1:26-61	<0,05	<1,80	2-4%	2

Table CIII. 3

	Compound	direction	qtl (cM)	shorter NIL	t-test (corrected p.value)	LOD	% explained variance	stable
	ethyl decanoate	up	LG7:0-10	Fb7:0-10	<0,05	3,59	10-38%	2
	ethyl dodecanoate	down	LG2:0-30	Fb2:0-30	<0,05	<1,80	4-15%	2
	ethyl dodecanoate	up	LG7:0-10	Fb7:0-10	<0,05	3,57	5-38%	2
→	ethyl hexanoate	down	LG1:26-61	Fb1:26-61	<0,05	<1,80	1-3%	2
	ethyl methylthioacetate T	down	LG7:0-10	Fb7:0-10	<0,05	1,84	1-22%	2
	ethyl octanoate	down	LG1:26-61	Fb1:26-61	<0,05	<1,80	1-4%	2
	ethyl octanoate	up	LG7:0-10	Fb7:0-10	<0,05	3,19	10-35%	2
	eugenol	up	LG5:50-76	Fb5:50-76	<0,05	4,44	33-45%	2
	hexanal	up	LG3:54-94	Fb3:54-94	<0,05	<1,80	6-7%	2
	hexyl butanoate	down	LG5:11-35	Fb5:0-35	<0,05	4,55	34-46%	2
	hexyl butanoate	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	1%	2
	hexyl hexanoate	down	LG2:0-30	Fb2:0-30	<0,05	3,53	27-38%	2
	limonene	down	LG7:0-10	Fb7:0-10	<0,05	2,03	21-24%	2
→	linalool	up	LG3:0-8	Fb3:0-8	<0,05	6,64	54-59%	2
→	mesifurane	down	LG7:26-43	Fb7:26-45	<0,05	7,27	16-62%	2
→	methyl 2-aminobenzoate T	down	LG7:0-10	Fb7:0-10	<0,05	2,54	8-29%	2
→	methyl 2-aminobenzoate T	down	LG5:11-35	Fb5:0-35	<0,05	6,79	33-60%	2
	methyl 2-hexenoate	down	LG5:11-35	Fb5:0-35	<0,05	3,26	14-35%	2
	methyl 3-hydroxyoctanoate T	down	LG3:8-15	Fb3:0-15	<0,05	<1,80	4-6%	2
	methyl benzoate	down	LG3:54-94	Fb3:54-94	<0,05	<1,80	10-11%	2
	methyl benzoate	down	LG6:101-101	Fb6:101-101	<0,05	<1,80	13-15%	2
	methyl benzoate	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	11-21%	2
	methyl benzoate	up	LG1:26-61	Fb1:26-61	<0,05	<1,80	14-18%	2
→	methyl butanoate	down	LG5:11-35	Fb5:0-35	<0,05	2,79	16-31%	2
→	methyl butanoate	down	LG7:0-10	Fb7:0-10	<0,05	<1,80	1-19%	2
→	methyl cinnamate T	up	LG2:0-30	Fb2:0-30	<0,05	2,81	18-32%	2
	methyl decanoate	down	LG5:0-11	Fb5:0-11	<0,05	2,49	24-28%	2
	methyl decanoate	down	LG4:9-44	Fb4:0-44	<0,05	2,71	31%	1
	methyl dodecanoate	down	LG1:26-61	Fb1:26-61	<0,05	<1,80	1%	2
	methyl dodecanoate	down	LG2:0-30	Fb2:0-30	<0,05	<1,80	8-18%	2
	methyl dodecanoate	down	LG5:11-35	Fb5:0-35	<0,05	<1,80	13-14%	2
	methyl dodecanoate	up	LG7:0-10	Fb7:0-10	<0,05	2,80	31%	1
→	methyl hexanoate	down	LG5:11-35	Fb5:0-35	<0,05	5,54	35-52%	2
	methyl octanoate	down	LG5:0-11	Fb5:0-11	<0,05	4,83	43-48%	2
	myrtenol	down	LG7:0-10	Fb7:0-10	<0,05	4,32	8-44%	2
	myrtenol	down	LG3:8-15	Fb3:0-15	<0,05	<1,80	2-7%	2
	myrtenol	up	LG5:50-76	Fb5:50-76	<0,05	6,71	60%	1
→	myrtenyl acetate	down	LG5:11-35	Fb5:0-35	<0,05	4,67	45-47%	2
→	myrtenyl acetate	down	LG6:101-101	Fb6:101-101	<0,05	<1,80	1%	2
	nerol	down	LG4:9-20	Fb4:0-20	<0,05	4,40	17-45%	2
	nerol	down	LG7:43-59	Fb7:43-59	<0,05	<1,80	7-11%	2

Table CIII. 3

	Compound	direction	qtl (cM)	shorter NIL	t-test (corrected p.value)	LOD	% explained variance	stable
	nerol	up	LG5:50-76	Fb5:50-76	<0,05	4,33	38-44%	2
→	nerolidol	up	LG3:0-8	Fb3:0-8	<0,05	22,38	76-95%	2
	nonanal	down	LG7:43-59	Fb7:43-59	<0,05	5,16	50%	1
	octanal	down	LG7:43-59	Fb7:43-59	<0,05	2,85	32%	1
	octyl acetate	down	LG5:11-35	Fb5:0-35	<0,05	2,40	25-27%	2
	octyl butanoate	down	LG2:0-30	Fb2:0-30	<0,05	2,11	21-25%	2
	octyl hexanoate	down	LG2:0-30	Fb2:0-30	<0,05	1,95	21-23%	2
	octyl hexanoate	down	LG1:26-61	Fb1:26-61	<0,05	<1,80	1-2%	2
	octyl hexanoate	down	LG5:0-11	Fb5:0-11	<0,05	<1,80	13-20%	2
	pentyl acetate	up	LG1:26-61	Fb1:26-61	<0,05	1,99	8-23%	2
	propyl butanoate	down	LG7:0-10	Fb7:0-10	<0,05	2,81	5-31%	2
	propyl butanoate	up	LG1:26-61	Fb1:26-61	<0,05	<1,80	7-13%	2

Discussion

Volatile profiling and particularities of the diploid strawberry

Woodland strawberry (*F. vesca*) aroma is known to present significant qualitative and quantitative differences when compared with commercialized varieties (*F. x ananassa*). *F. vesca* has been reported to have larger concentrations of esters and terpenoids and a more intense aroma besides specific compounds like methyl 2-aminobenzoate (aka methyl anthranilate) that confers its characteristic 'wild strawberry' scent (Ulrich *et al.* 2007). In this study we profile the volatile composition of a NIL collection obtained from an inter-specific cross between *F. vesca* and *F. bucharica (*Urrutia *et al*. 2015*)*. The genetic background of *F. vesca* confers stability and homogenicity to the collection with outstanding organoleptic quality, but the homozygous introgressions of *F. bucharica*, an exotic relative of *F. vesca*, confer enough phenotypic variability among the collection to map QTL for agronomical and metabolic traits (Urrutia *et al*. 2015; Urrutia *et al*. accepted). The alleles of *F. bucharica* had usually the effect of decreasing the accumulation of volatile compounds as most of the mapped QTLs had a negative effect over the ratio of the compounds.

The total number of identified volatile compounds was higher in *F. vesca* NIL collection (100) than in studies done with *F. x ananassa* populations (81 Schwiterman *et al.* (2014) and 87 Zorrilla-Fontansi *et al.* (2012)), although these differences may also arise from the extraction and detection protocol implemented. The *F. vesca* NIL collection aroma profiling revealed a complex composition of at least one hundred compounds being the majority of them esters (46%), followed by aldehydes (16%), ketones (14%), alcohols (11%) and by several terpenoids, furans and lactones (13%), these proportions are in agreement with what has been described in other studies with octoploid strawberries (Schwiterman *et al.* 2014, Zorrilla-Fontanesi *et al.* 2012).All the identified compounds in the *F. vesca* NIL collection have been previously described in strawberry fruit and around 20 of them have been reported to be important for its aroma (Latrasse 1991; Schieberle &Hofmann 1997; Ulrich *et al.* 1997; Ulrich *et al.* 2007).

The identified compounds that were not found in octoploid studies are several esters like methyl 2-aminobenzoate, methyl acetate, methyl cinnamate, methyl 3-hydroxyoctanoate, ethyl methylthioacetate and 2,3-butanedioldiacetate, some terpenoids as α-farnesene and α-pinene and the furanone furaneol (Zorilla-Fontanesi *et al.* 2012,) that might be contributing to the special scent of wild strawberry. In addition we also identified nerolidol and linalool segregating among our collection. These compounds have been reported to be characteristic of octoploid *Fragaria* species and produced by a truncated allele of the *FaNES* gene (Aharoni *et al.* 2004; Chambers *et al.* 2012). However we found a clear QTL at LG3:0-8 cM (Figure CIII. 5) for the accumulation of these two compounds that co-locates with the *FaNES* gene. The *F. vesca* RV parental do not produce any linalool or nerolidol but they were detected in the hybrid (analyzed only in 2013) (Supplemental Table CIII. 1) this suggests that the *F. bucharica* alleles for the *FaNES* gene are producing linalool and nerolidol. Both parental in the NIL collection (*F. vesca* and *F. bucharica*), the F_1 hybrid and the lines in the collection producing linalool and nerolidol (Fb**3**:0-8 and Fb**3**:0-15) together with a *Fragaria x ananassa* as a positive control were genotyped for *FaNES* alleles following the method described by Aharoni *et al.* (2004). The observed results concluded that the truncated *FaNES* allele is not present in our

collection (data not shown). This may suggest that there are several alleles producing linalool in strawberry and that some of them may have arise before octoploidization.

Comparison RV and YW

F. vesca YW is a white fruited strawberry known to have a nice intense fruity aroma with tropical (pineapple-like) notes. It has been used in this study as an out-group of the NIL collection and it has shown a different pattern of volatile accumulation (Figure CIII. 2) enriched in esters from cluster A and D2 and with a higher accumulation ratios than *F. vesca* RV (Figure CIII. 1). A recent study with the white fruited octoploid species *Fragaria chiloensis* (Prat *et al.* 2014), also known for its intense tropical fruity aroma, reported that the characteristic tropical fruit scent came from a set of 6 esters that include ethyl hexanoate (49) and hexyl acetate (52) also detected in our experiment associated to *F. vesca* YW (Figure CIII. 2). The other four compounds (furfuryl acetate, acetyl acetate, 1-methylethyl dodecanoate and ethyl tetradecanoate) were not detected under our experimental conditions this might be due to their absence, as they were detected in a different *Fragaria sp*. or might be due to the different identification method used.

QTLs

Significant year to year correlation was detected for most compounds (82 out of 100) although the correlation index and the significance threshold varied considerably according with the different kinds of inheritance described previously (Olbricht *et al.* 2008). However the correlation parameters that we report here are higher than those reported for volatile compounds in other studies (Eduardo *et al.* 2013). Differences in the relative volatile accumulation pattern in each NIL in the two studied harvests have been seen to be mainly associated to their genotypes (Figure CIII. 3) and in a lower extent to the environment. These contrast with what has been reported in other studies with octoploid strawberry (Forney *et al.* 2000; Zorrilla-Fontanesi *et al.* 2012) and peach (*Prunus persica*) (Eduardo *et al.* 2013; Sanchez *et al.* 2014) were the effect of the environment was more relevant. The special configuration of our mapping population, that is a near isogenic line collection, may be responsible for such stability, avoiding epistatic effects among different QTL. The fact that all lines share a common genetic background, in contrast with other mapping populations where genetic differences among lines is wider, may highlight the effect of the genotype, caused by exotic introgressions, and moderate the effect of the environment over the phenotypic traits as all lines may respond in a similar way.

The QTL mapping revealed 50 major stable QTL that accounted for a high proportion of the variabililty of 47 compounds, including 14 major QTL identified for 13 'key compounds'. Many of the QTL co-locate in few genetic regions according to clusters suggesting that those compounds are co-regulated and controlled by a reduced number of loci. LG5 and LG7 seem the most determinant regions controlling volatile compounds synthesis as they accumulate the largest number of QTL and harbor nine and two major QTL for 'key compounds' respectively. Lines with introgressions at the beginning of LG5 like Fb**5**:0-35 might be interesting for further studies in fruity scent (esters) heritance, lines with introgressions at the end of LG5, like Fb**5**:50-76 could be used to dissect the herbaceous-refreshing aroma and lines with

introgressions in LG7 could be of use for the study of methyl 2-aminobenzoate (Fb**7**:0-10) and mesifurane (Fb**7**:26-45).

Some of the detected QTLs confirm those described by Zorrilla-Fontanesi *et al.* (2012) as they co-locate according to synteny studies (Rousseau-Gueutin *et al.* 2008, Tennessen *et al.* 2014). A QTL for Methyl benzoate was located in LG1:26-61 cM in *F. vesca* and in LGI-F.1: 38 cM in *F. x ananassa.* A QTL for benzyl acetate was located in LG7:0-10 cM in *F. vesca* and in LGVII-F.1c: 9 cM in *F. x ananassa*. A QTL for ethyl decanoate was in LG3:8-15 cM for *F. vesca* and in LGIII-F.1: 4 cM and LGIII-M.1: -8 cM in *F. x ananassa* and a QTL for mesifurane was located in LG7:27-43 cM in *F. vesca* and in LGVII-F.2: 18 cM and LGVII-M.2:65 cM in *F. x ananassa*, this last QTL is associated with the gene *FaOMT* responsible for its accumulation that also co-locates with our QTL (Zorrilla-Fontanesi *et al.* 2012). There were also QTLs that were located in different genetic regions in *F. x ananassa* and *F. vesca* or compounds that showed significant variability in one population and not in the other, thus highlighting that different genetic backgrounds and environments can reveal different genetic traits. For instance we found two QTLs controlling accumulation of methyl 2-aminobenzoate that is characteristic of *F. vesca* aroma and that was not detected in *F. x ananassa*. Previous reports have mapped a QTL for the accumulation of γ-decalactone in the homeolog LGIII-M.2: 50-54 cM (Zorrilla-Fontanesi *et al.* 2012) and a candidate gene *FaFAD1* with an eQTL co-localizes (Sanchez-Sevilla *et al.* 2014), however we did not find any significant QTL for γ-decalactone in our collection although data suggest that there might be an increase in this compound in lines with introgressions at the end of LG5, this increase is not enough to report a significant effect (Supplemental Table CIII. 1 and C.III Supplemental Figure 2). However these may suggest that there are other genetic regions controlling γ-decalactone accumulation in *F. vesca.* Esters derived from the lipoxygenase pathway and C_6 compounds ((*E*)-2-henxe-1-ol, (*E*)-2-hexenal, (*E*)-2-hexenyl acetate, (*Z*)-3-hexenal and (*Z)*-3-hexenyl acetate) are commonly described as 'green compounds' and are usually considered too variable within genotypes or varieties to be used as discriminative compounds (Ulrich *et al.* 1997), however a recent study with an F_1 mapping population in peach (*Prunus. persica*) reported stable QTLs for (*E*)-2-hexenyl acetate and (*Z*)-3-hexenyl acetate (Eduardo *et al.* 2012). Our data revealed a high year to year correlation between these compounds (Table CIII. 2 and C.III Supplemental Figure 2) and QTLs that co-localize for all of them at LG5:50-76 cM, this suggesting that these compounds are stable and co-regulated under our conditions.

A in-depth characterization of the volatiles emitted by the ripe strawberry fruit in a *F. vesca* NIL mapping collection revealed a complex mixture of 100 compounds, that vary in relative abundance across the population presumably because of the effect of *F. bucharica* alleles. The high genetic effect under the heritance of many compounds (35 compounds >50% G effect) allowed the mapping of 50 major QTL including 14 QTL for compounds considered of extreme importance for strawberry scent. Some of them, as methyl 2-aminobenzoate or mesifurane, are rarely found in commercial varieties (*F. x ananassa*) and are of great interest for breeding programs. Therefore, we set up the basis for further studies in the heritance of the woodland strawberry aroma that may lead to improve aroma and marketability of new strawberry varieties.