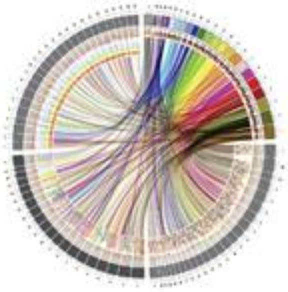
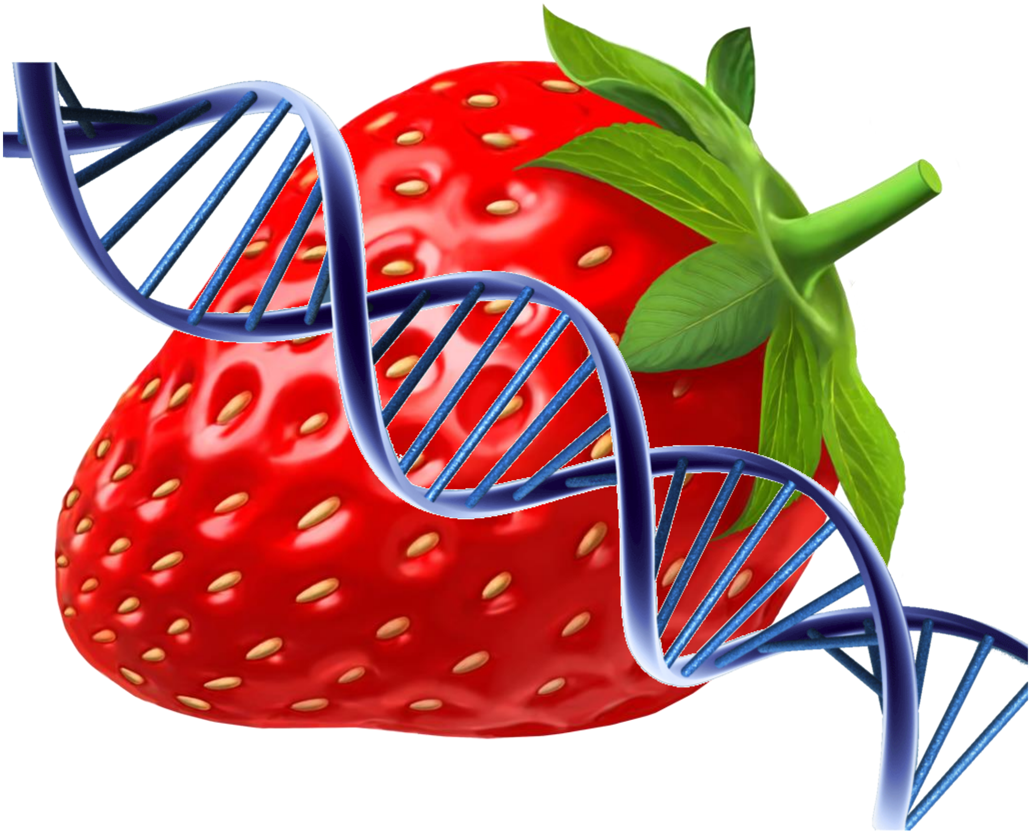


Master’s Thesis

IDENTIFICATION OF HAPLOTYPES IN OCTOPLOID STRAWBERRY (*Fragaria* × *ananassa*) CULTIVARS USING ISTRAW90® SNP ARRAY DATA.

Miriam Payá Milans



Barcelona, July of 2015

AUTONOMOUS UNIVERSITY OF BARCELONA

FACULTY OF BIOSCIENCES

MSc in Bioinformatics

MASTER THESIS

IDENTIFICATION OF HAPLOTYPES IN OCTOPLOID STRAWBERRY (*Fragaria* × *ananassa*) CULTIVARS USING ISTRAW90® SNP ARRAY DATA.

.

Master’s student:
Miriam Payá Milans, PhD

Academic tutor:
Dr. Raquel Egea

Project supervisors:
Konstantinos Alexiou, PhD
Dr. Amparo Monfort

Work presented for completion of the Master's Degree by:

Miriam Payá Milans, PhD

Academic tutor

Dr. Raquel Egea
Associate professor (IBB)

Project supervisors

Konstantinos Alexiou, PhD
Postdoc (IRTA)

Dr. Amparo Monfort
Research Scientist (IRTA)

Barcelona, July of 2015

Abstract

Cultivated strawberry (*Fragaria x ananassa*) is a hybrid octoploid species whose fruit is highly appreciated due to its organoleptic properties and health benefits. Despite its economic importance, the octoploid genome represents a great challenge to study the genome structure and molecular mechanisms related to fruit quality and agronomic traits. Recently, a strawberry SNP genotyping array was developed providing thus a powerful tool for genome-wide scanning. The present study centered on utilizing SNP data from 37 cultivars, bred by the company PLANASA, to determine trait-relevant haplotypes in the cultivars and to improve a strawberry genetic linkage map in development, generated from a F2 population derived from a “Camarosa” x “Dover” cross. We developed a strategy to extend the genetic linkage map by association of high quality SNPs to the previously mapped ones based on genotype similarity with consistent results. The resultant 33 maps, composed of 2,106 markers, were derived from joining 71 fragments in function of their annotated homeolog group; the highlight is on the 633 markers that were newly incorporated. As genetic distance among cultivars responded to both phenotype and genotype properties, a thorough analysis on those linked by a common feature may shed light on the markers involved with it. Functional annotations of SNPs consisted of prediction of effects on genes, with focus on those variants that disrupted gene transcription; a list of 280 markers with predicted high impact on genes is provided. This work may be continued by performing association analysis when cultivars are fully phenotyped in order to define candidate genes related with traits of economic significance.

Index

Figures

Tables

1. Introduction

1.1. Biology and genetics of strawberry

Strawberries are small red fruits very appreciated for their characteristic flavor and aroma, bright red color and sweetness. There are around twenty (Stewart, 2011) wild species of *Fragaria* distributed worldwide (Folta and Davis, 2006). They differ not only on geographical distribution, but in genetic content due to the ability of these plants to naturally polyploidize (Figure 1). The most widely spread species is the wild strawberry *F. vesca*, a diploid with a genome size of 200 Mb organized into 2n = 14. Two relevant natural octoploids are *F. chiloensis* and *F. virginiana*, with 2n = 8x = 56, whose crosses gave rise to the cultivated octoploid hybrid *F.* x *ananassa*, originated in French gardens during the XVIII century (Darrow, 1966). A conventional model assumes that the species most closely related to the natural polyploid species are the diploids *F. vesca* and *F. nubicola* (Potter *et al*., 2000), although a recent model considers that one of the octoploid subgenomes originated with the donor *F. vesca*, one with the diploid *F. iinumae* and two unknown donors close to *F. iinumae* (Tennessen *et al*., 2014).

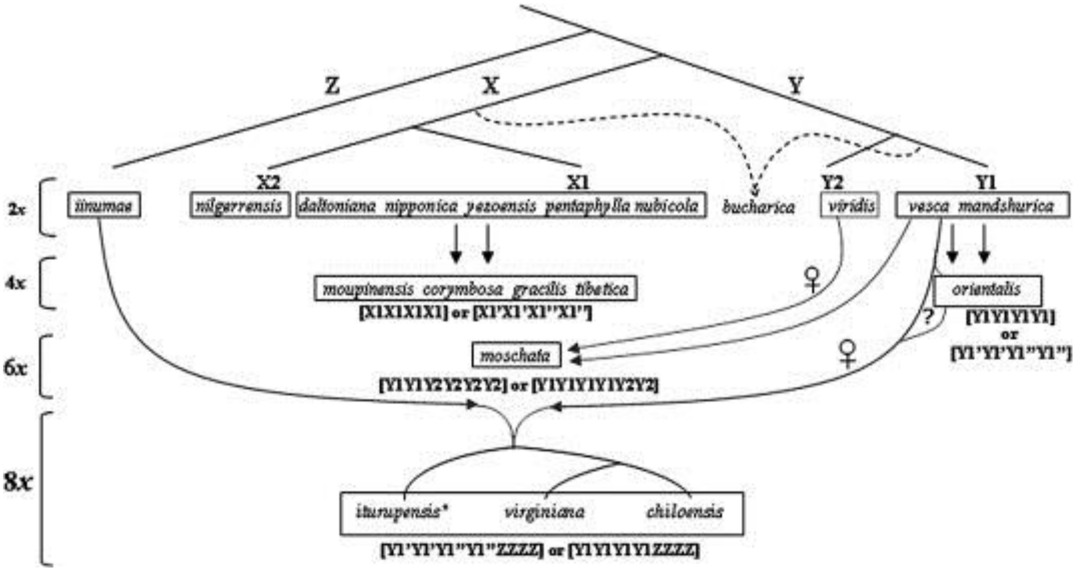

Z
X
Y
X2
X1
Y2
Y1
2x
4x
6x
8x
iinumae
nilgerrensis
daltoniana nipponica yezoensis pentaphylla nubicola
bucharica
viridis
vesca mandshurica
moupinensis corymbosa gracilis tibetica
[X1X1X1X1] or [X1'X1'X1''X1'']
orientalis
?
[Y1Y1Y1Y1]
or
[Y1'Y1'Y1''Y1'']
♀
♀
moschata
[Y1Y1Y2Y2Y2Y2] or [Y1Y1Y1Y1Y2Y2]
iturupensis*
virginiana
chiloensis
[Y1'Y1'Y1''Y1''ZZZZ] or [Y1Y1Y1Y1ZZZZ]

Figure 1. Hypothetical evolution of *Fragaria* spp. (Rousseau-Gueutin *et al*., 2009).

Fragaria belongs to the family *Rosaceae*, which contains many other economically relevant genera. To the subfamily *Maloidae* belong *Prunus* (stone fruits and almonds), *Pyrus* (pears) and *Malus* (apples), and in *Rosoidae* are *Rubus* (blackberries and raspberries), *Fragaria* (strawberries) and ornamental species in the genus *Rosa*. The increasing availability of genomic resources in rosaceous species is promoting evolutionary studies through comparative genomics. In this respect and benefiting from the synteny in the *Rosaceae* species, modes of evolution could be evaluated for *Fragaria, Malus* and *Prunus,* together with the reconstruction of the common ancestor genome (Jung *et al*., 2012). Focusing on the tribe *Potentilleae*, with genera close to *Fragaria*, the similarity in fruits revealed the independent evolution of "strawberries" in different species (Staudt, 1962).

Strawberries are not real fruits, but the berry is formed when the receptacle becomes enlarged and fleshy at maturity (Figure 2). The dark spots in the surface are the real fruits, referred as achenes, which contain the seeds. The offspring derived from the seeds do not usually harbor the desirable parental characteristics, so propagation is performed through "runner plants" that are genetically identical to their "mother plant". After decades of breeding, strawberry cultivars vary widely in size, color, flavor, shape, degree of fertility, season of ripening, liability to disease and constitution of plant. Regarding these characteristics, numerous cultivars have been cultivated for both consumption and ornamental purposes. In base to their phytochemicals (ellagic acid, anthocyanins, quercetin, and catechin) and vitamins (ascorbic acid and folic acid), strawberries have been considered as functional food (Basu *et al*., 2014).

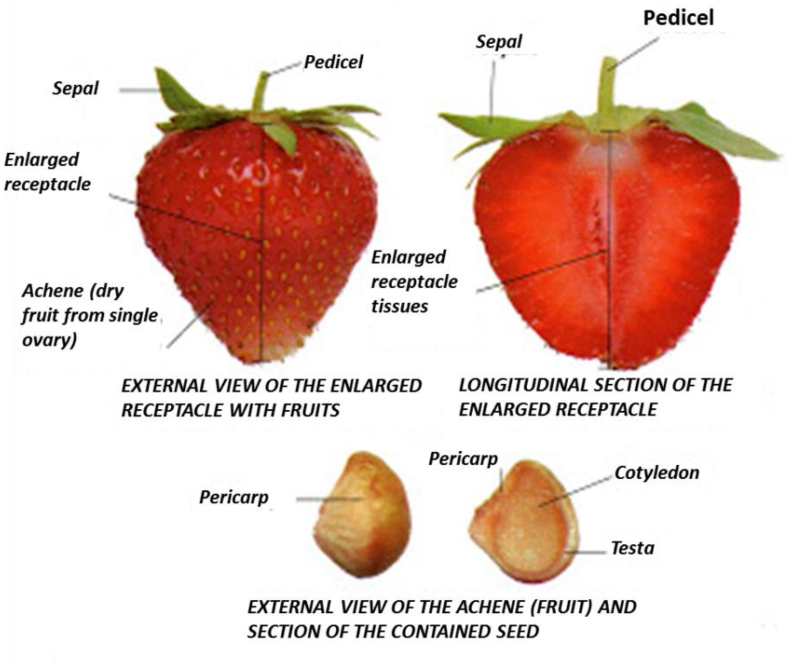

Sepal
Pedicel
Enlarged receptacle
Achene (dry fruit from single ovary)
Enlarged receptacle tissues
Sepal
Pedicel
EXTERNAL VIEW OF THE ENLARGED RECEPTACLE WITH FRUITS
LONGITUDINAL SECTION OF THE ENLARGED RECEPTACLE
Pericarp
Pericarp
Cotyledon
Testa
EXTERNAL VIEW OF THE ACHENE (FRUIT) AND SECTION OF THE CONTAINED SEED

Figure 2. Structure of strawberries and their seeds.

In order to improve marker assisted selection of cultivars in relation to a set of desired characteristics, molecular tools are being developed. At the beginning, multiple genetic markers were used to generate genetic linkage maps. The first reference of *F.* x *ananassa* map was developed by Rousseau-Gueutin *et al*. (2008), consisting of 32 linkage groups, from which 28 were assigned to one of the seven diploid linkage groups of a *F. vesca* reference map. Nowadays, two *F. vesca* reference genomes with gene annotations are available (Shulaev *et al*., 2011; Tennessen *et al*., 2014; Darwish *et al*., 2015). Due to its complexity, the genome of the octoploid is not available, although the development of newer linkage maps are benefiting from these references.

1.2. Strawberry breeding

Strawberry is cultivated all over the world, preferentially in temperate regions (Figure 3, data from FAOSTAT 2015). There are few countries producing relevant quantities of this crop, but most countries are small producers. Spain is the highest exporter (230 k tonnes), followed by the United States of America (USA) (115 k) and Mexico (63 k) (average 2002-2012). USA is also one of the highest importers together with Germany, France and Canada at comparable rates (75-100 k tonnes). Despite being the highest producer, strawberries produced in China are mostly for local consumption, as shown by the low levels of imports and exports for this crop.

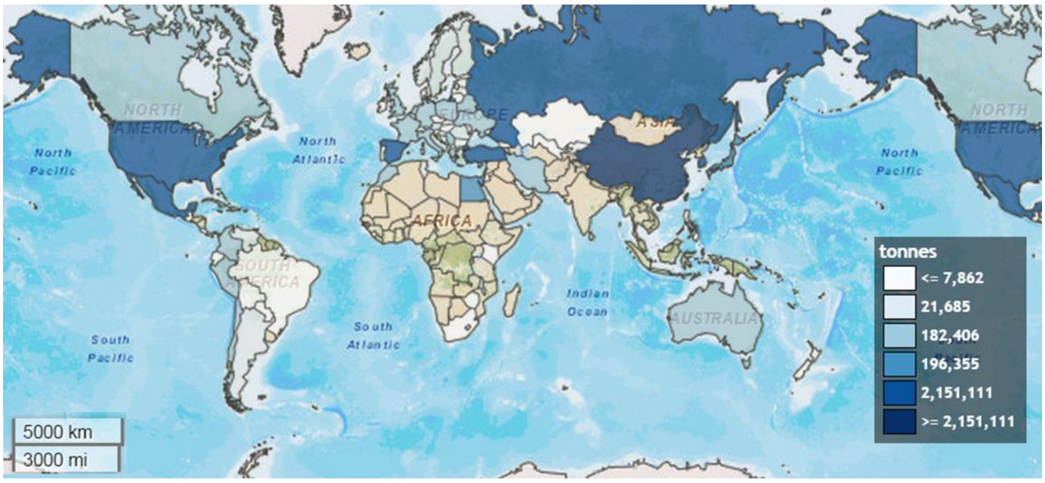


Figure 3. Strawberry production quantities by country, average 2003-2013.

There is a huge step in production among the greatest producers and the rest as seen in Figure 4, where China almost doubles USA's strawberry production, which is at the same time four times higher than the next producer, Spain. The production in these two countries is likely related to their increased land resources for agricultural purposes, ranking first and third respectively in top producer countries in 2013 (FAOSTAT 2015). However, the crop yield is not correlated to the production. In this case, USA leads the ranking with a yield of 535 k Hg/Ha, followed by Morocco with 59 k less and Spain with a yield 100 k even lower (69% of USA's). At this point, yield decreases gradually among countries. Even though China is the highest producer, the yield is half that of USA, 273 k Hg/Ha, indicating huge differences in agricultural systems.

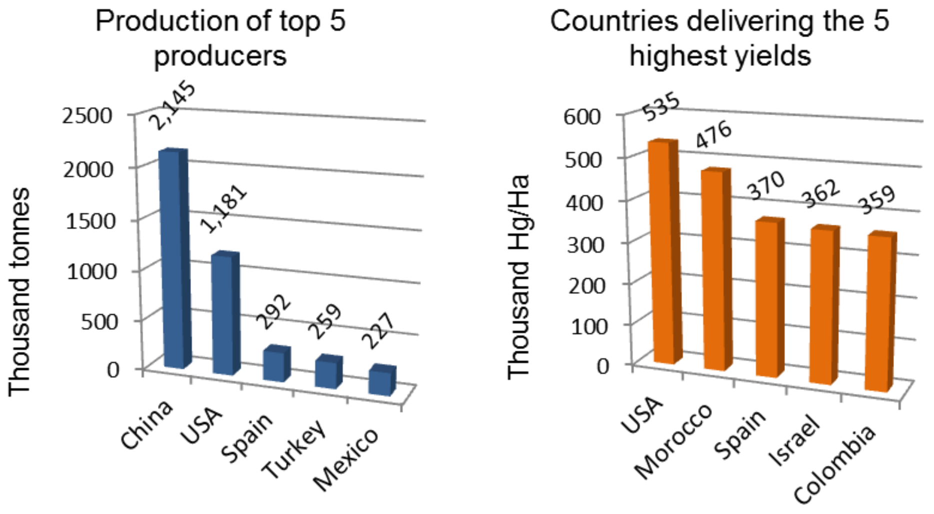


Figure 4. Top 5 producers and countries with highest yields for strawberry, average 2003-2013.

A comparison of world and Spanish strawberry production and yield reflects the apparent low impact of the Spanish stock to the global amount (Figure 5). During the last 20 years, world production has increased from around 3 million tonnes to almost 8 million tonnes, while in Spain production fluctuates around 260-350 k tonnes, most of which is exported. World yield has risen from 130 to 215 Hg/Ha, half that of Spain, which has fluctuated between 285 and 430 Hg/Ha. This reveals a much higher efficiency of the culture in Spain.

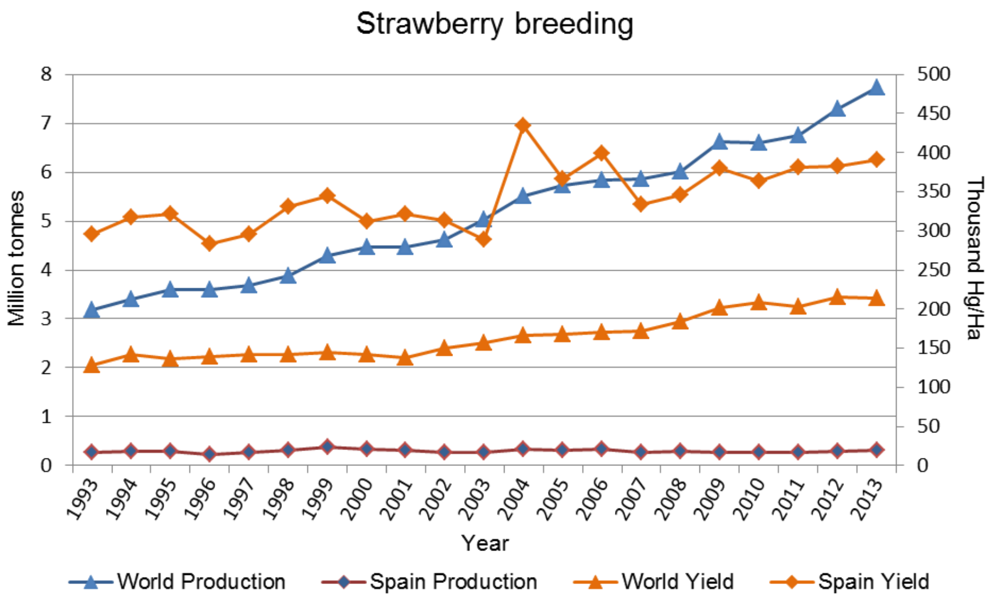


Figure 5. Strawberry production and yield in the world and Spain from 1993 to 2013.

As deduced by the imports/exports data, most of the cultivated strawberries worldwide are for local commerce and still they are in demand. Being the major exporter, strawberry breeding in Spain has a global economic impact. Thus, the study of both genetic and environmental factors affecting their growth, yield and quality is of primary importance.

1.3. Nutrients and phytochemicals

Strawberries have a rich content in nutrients and phytochemicals that have relevant biological effect on humans. These fruits are excellent for their high content in vitamin C, 150% of recommended daily value, as well as low calories and fat, high fiber and manganese and presence of the phytochemicals: anthocyanins and ellagic acid (Table 1). In general, strawberries are a good source of minerals, like manganese, potassium, iodine, copper or iron, vitamins, mainly vitamin C and folate, and other compounds which usually show antioxidant properties (Giampieri *et al*., 2012).

Table 1. Nutrient composition of raw strawberries.Adapted from the U.S. Department of Agriculture, Agriculture Research Service, accessed on June 2015.

Type	Nutrient	Per 100 g	Type	Nutrient	Per 100 g
Proximates	Water (g)	90.95	Minerals	Calcium (mg)	16
	Energy (kcal)	32		Iron (mg)	0.41
	Protein (g)	0.67		Magnesium (mg)	13
	Total lipid (g)	0.3		Phosphorus (mg)	24
	Carbohydrate (g)	7.68		Potassium (mg)	153
	Dietary fiber (g)	2		Sodium (mg)	1
	Sugars (g)	4.89		Zinc (mg)	0.14
	Sucrose (g)	0.47		Copper (mg)	0.048
	Glucose (g)	1.99		Manganese (mg)	0.386
	Fructose (g)	2.44		Selenium (μg)	0.4
Vitamins	Vitamin C (mg)	58.8		Fluoride (μg)	4.4
	Thiamin (mg)	0.024		Flavonoids	
	Riboflavin (mg)	0.022	Anthocyanidins	Delphinidin (mg)	0.3
	Niacin (mg)	0.386		Pelargonidin (mg)	24.8
	Pantothenic acid (mg)	0.125		Cyanidin (mg)	1.7
	Vitamin B6 (mg)	0.047	Flavan-3-ols	(+)-Catechin (mg)	3.1
	Folate (μg)	24		(-)-Epigallocatechin (mg)	0.8
	Choline (mg)	5.7		(-)-Epicatechin (mg)	0.4
	Betaine (mg)	0.2		(-)-Epicatechin 3-gallate (mg)	0.2
	Vitamin A, RAE (μg)	1	Flavonols	Kaempferol (mg)	0.5
	β-carotene (μg)	7		Quercetin (mg)	1.1
	Lutein + zeaxanthin (μg)	26	Proanthocyanidin	Monomers (mg)	3.7
	Vitamin E (mg)	0.29		Dimers (mg)	5.3
	β-tocopherol (mg)	0.01		Trimers (mg)	4.9
	γ-tocopherol (mg)	0.08		4-6mers (mg)	28.1
	δ-tocopherol (mg)	0.01		7-10mers (mg)	23.9
	Vitamin K (μg)	2.2		Polymers (>10mers) (mg)	75.8

Strawberries are rich in phytochemicals, represented by the phenolic compounds flavonoids (anthocyanins, flavonols and flavanols), hydrolysable tannins (ellagitannins and gallotannins), phenolic acids (hydroxybenzoic acids and hydroxycinnamic acids) and condensed tannins (proanthocyanidins), in order of relevance. Anthocyanins, the most abundant class of polyphenolic compounds in strawberry, are responsible of high antioxidant capacity. Among the diversity of pigments found in this fruit, pelargonidin-3-glucoside possesses the major presence (Table 1), while cyanidin-3-glucoside is at a constant low level among varieties (Giampieri *et al.*, 2012). Another class of phenolic compounds is ellagitannins, which liberate ellagic acid after hydrolysis, constituting active principals typical of medicinal plants. Other relevant bioactive molecules are flavanols, reported to possess antioxidant, antimicrobial, antiallergic, and antihypertensive properties. The content of these bioactive compounds not only vary during the stages of fruit maturing, but also on the breeding conditions and genetics (Giampieri *et al*., 2012). During storage, although smell and flavor of the fruits deteriorate, the antioxidant capacity increases due to the complex metabolism taking place at this post-harvest period.

The cultivated strawberry *F.* x *ananassa* present large fruits, bright red color and prolonged shelf life, but at the cost of the intensity and variety of the aroma (Negri *et al*., 2015). The aroma is the result of a complex mixture, regarding that as many as 360 volatile compounds have been identified in ripe strawberries, including esters, aldehydes, ketones, alcohols, terpenes and furanones (Menager *et al*., 2004). Even small amounts of a volatile may make great difference. Due to this complexity, little information is still available about genes associated with aroma biogenesis. The best characterized enzyme related to aroma is a strawberry alcohol acyltransferase (SAAT) gene that plays a crucial role in flavor biogenesis in ripening fruit (Aharoni *et al*., 2000). Another reported gene is omega-6 fatty acid desaturase, which correlates with the presence of the volatile γ-decalactone (Chambers *et al*., 2014). Currently, to assess the improvement of the aroma in cultivated strawberry after breeding programs, comparison of volatile profiles related to aroma in diverse species and cultivars are revealing the most desirable compounds. As an example, one survey reported significant differences between *F. vesca* and *F.* x *ananassa* in the accumulation of individual esters, ketones and terpenoids (Ulrich and Olbricht, 2013).

1.4. Genetic studies in F. x *ananassa*

1.4.1. Classical genetics

To improve marker-assisted selection of cultivars it is necessary to identify specific markers related to variation among cultivars. Classical breeding would only dispose of morphological and phenotypic information. The development of molecular tools has proved very valuable in the characterization of plant varieties. As explained in paragraph 1.1, strawberries are reproduced by micropropagation, reducing genetic variability in cultivars. This issue made the use of RAPD (*Random Amplified Polymorphic DNA*) very useful at early investigation of the genome during 1990-2000 (Congiu *et al*., 2000). From 2001 and on, studies about genetic diversity incorporated

AFLP (*Amplified Fragment Length Polymorphism*) analysis (Degani *et al*., 2001). Development of new molecular markers consisted of the identification of single sequence repeats (SSR), also known as microsatellites, in a strawberry unigene data set (Folta *et al*., 2005). An improved library of these SSR markers, together with single-dose (SD) markers, which are useful for constructing maps in polyploids (Da Silva and Sobral, 1996), allowed the construction of the first genetic map for octoploid strawberry with 32 linkage groups (Rousseau-Gueutin *et al*., 2008). Since then, many more maps have been developed. In 2009, Sargent *et al*. constructed a genetic linkage map based on microsatellites, gene-specific markers and AFLP and RAPD markers, recovering 69 linkage group fragments, more than the 56 expected in strawberry. Later on 2014, van Dijk *et al*. obtained a SSR linkage map representing each of the 28 chromosome pairs of octoploid strawberry on the base for quantitative trait locus (QTL) discovery, providing the first biologically relevant basis for the discernment and notation of sub-genomes.

The study of a polyploid organism has some disadvantages in the use of genetic markers considering the multiplication of variants and *a priori* unknown distribution among homeolog chromosomes. Interesting results were achieved however using the markers indicated on the previous paragraph. Their development constituted a useful tool to successfully differentiate cultivars (Degani *et al*., 2001). Also, the comparison of diploid and polyploid maps constructed using molecular markers revealed that no major changes occurred after polyploidization, suggesting high synteny between *F. vesca* and *F.* x *ananassa* (Rousseau-Gueutin *et al*., 2008). Moreover, mixed behavior between disomy and partial polysomy was revealed in cultivated strawberry when studying the segregation of microsatellites (Rousseau-Gueutin *et al*., 2008).

Strawberry gene databases were first constructed using ESTs (*Expressed Sequence Tags*) from the genus *Fragaria*, assembling them to create longer transcripts and yielding putative unigenes (Folta *et al*., 2005; Bombarely *et al*., 2010). Then, these assemblies were annotated through homology to Swiss-Prot, *Arabidopsis*, and NCBI nr proteins, functionally characterized and genes relevant to specific physiological processes of economic importance identified. Therefore, it was demonstrated how computational tools allowed mining of important information with limited data. Several *Fragaria* Unigene versions are hosted at the Genome Database for *Rosaceae* (GDR). Since the first available version (v2, constructed on 2005-12-01 from ESTs from GenBank on December 1, 2005) to the latest one (v5, constructed on 2012-12-19 from

ESTs downloaded on July 1, 2012), the number of putative unigenes has risen from 5,565 to 18,234. Many sequencing projects around the world are depositing ESTs from the genus *Fragaria* in the NCBI dbEST database, feeding a future improved build.

1.4.2. SNP array

The latest advance in the development of a genotyping tool for commercial strawberry involved the design of a SNP (*Single Nucleotide Polymorphism*) microarray, the iStraw90 Axiom Array commercialized by Affymetrix, with 95,062 marker loci interrogated with 138,099 probesets (Bassil and Davis *et al*., 2015). SNPs are single variants at orthologous sites within or between individuals. Polyploids constitute a challenge in identifying marker SNPs due to homeologous sequence variants (HSVs) occurring between subgenomes, as well as paralogous sequence variants (PSVs), which occur at non-identical reference coordinates. Finally, several types of marker candidates were discovered: di- and multi-allelic SNPs, di-allelic indels and three categories of haploSNPs (Figure 6 A-C). HaploSNPs coupled a marker SNP with a close HSV, either SNP (Figure 6 A) or indel (Figure 6 B-C), providing a destabilization site resulting in a technical reduction of ploidy. An intrinsic type of site-specific reduction of ploidy was due to deletions in one or more subgenomes. Assessing the goodness of this array to study the whole genome, polymorphic SNPs were well distributed with few gaps in a representation according to their relative physical position across the 7 chromosomes of *F. vesca*.

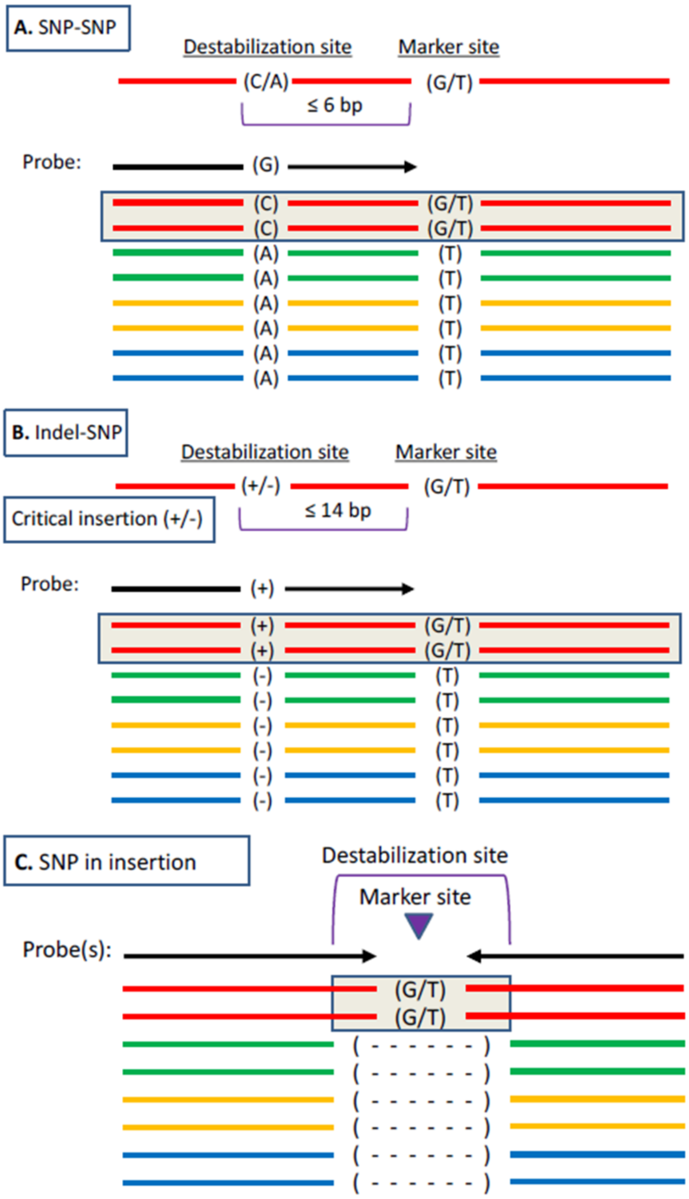


Figure 6. Representation of the three haploSNP categories consisting of SNP-SNP (A), Indel-SNP (B) and SNP-in-Insertion (C). Adapted from Bassil and Davis *et al*. (2015).

After the process of genotyping sample DNA, a list of calls indicating homozygosity (for reference or alternative allele), heterozygosity or no call for each probeset and individual was obtained. The representation of individual calls as function of intensity vs A-B contrast (Figure 7 A) allowed the classification of each probeset in one of six groups. The main parameters analyzed were degree of polymorphism (rate of AA, AB, BB or no call), microarray intensities and call rate (default threshold at 97%). The resultant SNP classes were: (1) "*Poly High Resolution*" (*PHR*), with three well resolved clusters and passing all quality-control (QC) measures; (2) "*No Minor*

Homozygote" (*NMH*), passing all QC but missing one homozygote cluster; (3) "*Off-Target Variant*" (*OTV*), with a low intensity cluster; (4) "*Mono High Resolution*" (*MHR*), showing one monomorphic genotype that passed all QC; (5) "*Call Rate Below Threshold*" (*CRBT*), where genotype call rate was below 97%; and (6) "*Other*", for those SNP patterns not matching any of the previous. The goal of a good genotyping is to identify SNPs with three well separated clusters for two segregating alleles (*Poly High Resolution*).

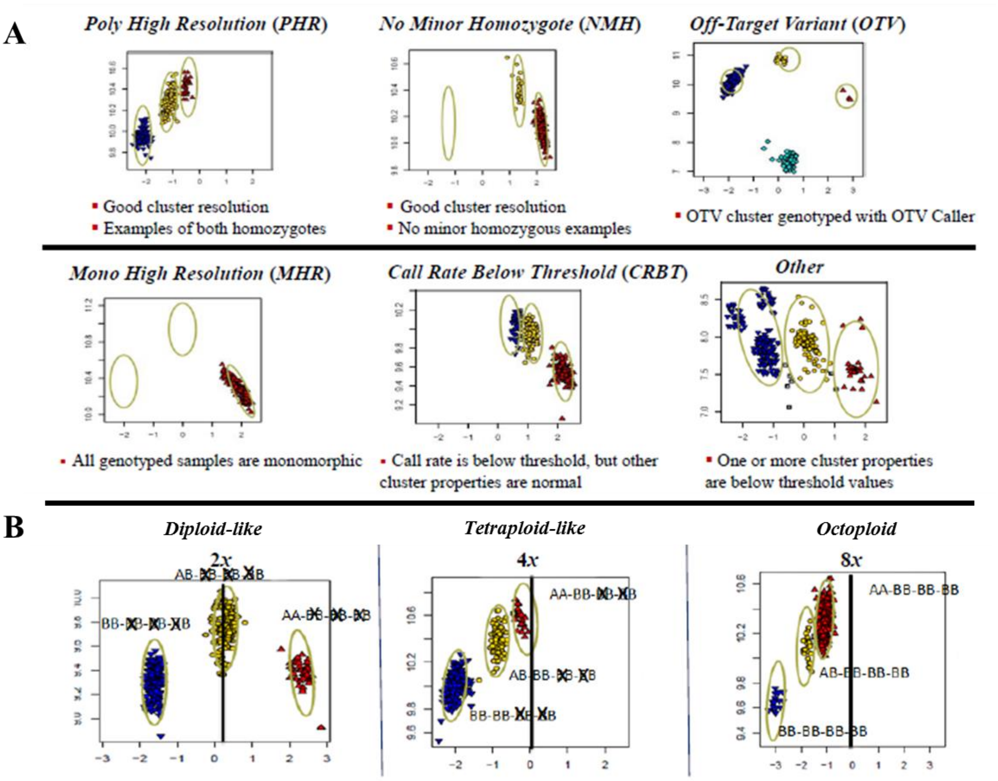


Figure 7. Six SNP quality classes (A) and observed cluster locations in three apparent ploidy levels of a polyploid (B). Adapted from Bassil and Davis *et al*. (2015).

The clustering pattern was affected by the ploidy level (Figure 7 B). A measure that is directly related to ploidy is the homozygosity ratio offset (HomRO), which measures the displacement from zero contrast of the closest homozygous cluster. Based on simulation, a HomRO value ≥ 0.3 was used to classify SNPs as clustering like a diploid, and a HomRO value < 0.3 to classify SNPs clustering like a polyploid. The pattern for polyploid segregation of alleles had the genotype clusters displaced from zero contrast, offset to the positive (subgenomes fixed for the A-allele) or negative (subgenomes fixed for the B-allele) contrast values. Thus, there were both visual and numerical data supporting the relative ploidy of clustering for each SNP.

Genotyping Solution from Affymetrix enables large-scale, high-throughput SNP genotyping for agriculture and human genotyping applications. Plant genomes however present additional challenges due to the complex organization of polyploid genomes. During the last few years, several arrays have been developed for several crops. An example of the evolution of arrays is seen in *Malus*, starting with the development of a 8K SNP array (Chagne *et al*., 2012), which rapidly led to the construction of saturated linkage maps (Antanaviciute *et al*., 2012), but the evaluation of the probes revealed the inaccuracy of the genotyping (Troggio *et al*., 2013) and then a newer 20K array was developed (Bianco *et al*., 2014). SNP genotyping offers a number of advantages over previous marker systems, including abundance of markers, rapid processing of large populations and straightforward allele calling (Thomson, 2014). This technology constitutes a powerful tool to perform fine linkage map construction, genome wide association studies (GWAS), QTL analysis and, additionally, integration of SNP markers into crop and livestock research and breeding.

2. Objectives

The present work was developed as a part of the project called “Development of reliable marker (SNP) sets for high density mapping in octoploid strawberry” in the *Rosaceae* Genetics and Genomics group in CRAG (Barcelona, Spain). In this study, SNP microarrays (IStraw90® Axiom® SNP array, Affymetrix) were used to analyze the genotype of strawberry cultivars from the breeding company PLANASA. The specific objectives of this work are:

I. Determination of SNP haplotypes in strawberry cultivars.

II. Enrichment of the octoploid strawberry genetic map with new polymorphic SNPs.

III. Discovery of candidate genes.

3. Methodology

3.1. Data

The aim of this project is to analyze the genetic differences among different strawberry cultivars cultivated by PLANASA (Table 2) and that may be responsible of their specific characteristics. Thus, IStraw90® Axiom® SNP array data (Affymetrix; Bassil and Davis *et al.*, 2015) from 37 cultivars was provided together with data from an octoploid strawberry F2 population (formed by self-fertilization of F1 hybrid after Camarosa x Dover cross), which included 117 individuals and parental lines Camarosa, Dover and their respective hybrid. *F. vesca* and *F. bucharica* diploids were included in the genotyping process for their use as outgroup in a cladogram. DNA was extracted with DNeasy® plant mini kit (Qiagen, Hilden, Germany) from 37 lines. 50 µL of DNA from each sample, at a concentration $\geq$ 20 ng/µL, was sent for array hybridization. Samples were hybridized with the IStraw90® Axiom® SNP array at USA.

Table 2. List of the cultivars analyzed with their respective geographical location and principal characteristics.

Cultivar	Origin	Characteristics	Cultivar	Origin	Characteristics
1	Huelva	Flavor	D-58	LeBarp	Yield
2	Huelva	Precocity	D-62	LeBarp	Precocity, flavor
3	Huelva	Precocity	D-71	LeBarp	Yield
4	Huelva	Flavor	D-74	LeBarp	Yield, resistance
5	Huelva	Flavor	D-78	LeBarp	Late, yield
6	Huelva	Flavor	D-81	LeBarp	Medium-late, orange
7	Huelva	Flavor	D-85	LeBarp	Medium-late, hardness
8	Huelva	Precocity	Charlotte	Huelva	Wild strawberry flavor
9	Huelva	Precocity	P22	Huelva	Flavor
10	Huelva	Flavor	P23	Huelva	Precocity
11	Huelva	Flavor	P24	Huelva	Flavor, precocity
12	Huelva	Sensitivity	P25	Huelva	Remontant
13	Huelva	Flavor	P26	Huelva	Remontant
14	Huelva	Resistance	D54	Huelva	Precocity
15	Huelva	From S. Miguel	D59	Huelva	Aroma
16	Huelva	Subacid	D73	Huelva	Flavor, remontant
17	Huelva	From Chilo	Camarosa	Huelva	Yield
18	Huelva	Wild strawberry flavor	Dover	Huelva	Resistance
D-45	LeBarp	Flavor, round, orange	DxC Hybrid	Huelva	Precocity
D-55.1	LeBarp	Yield	DxC 21AF-34	Huelva	-
D-57	LeBarp	Sugar, hardness			

3.2. Workflow

A general workflow is shown in Figure 8. Third-party software will be described at their corresponding paragraphs. Custom scripts in Python 3.4 and Perl, as well as a pipeline, were uploaded to GitHub. Detailed functions and commands are described in pipeline.txt on GitHub.

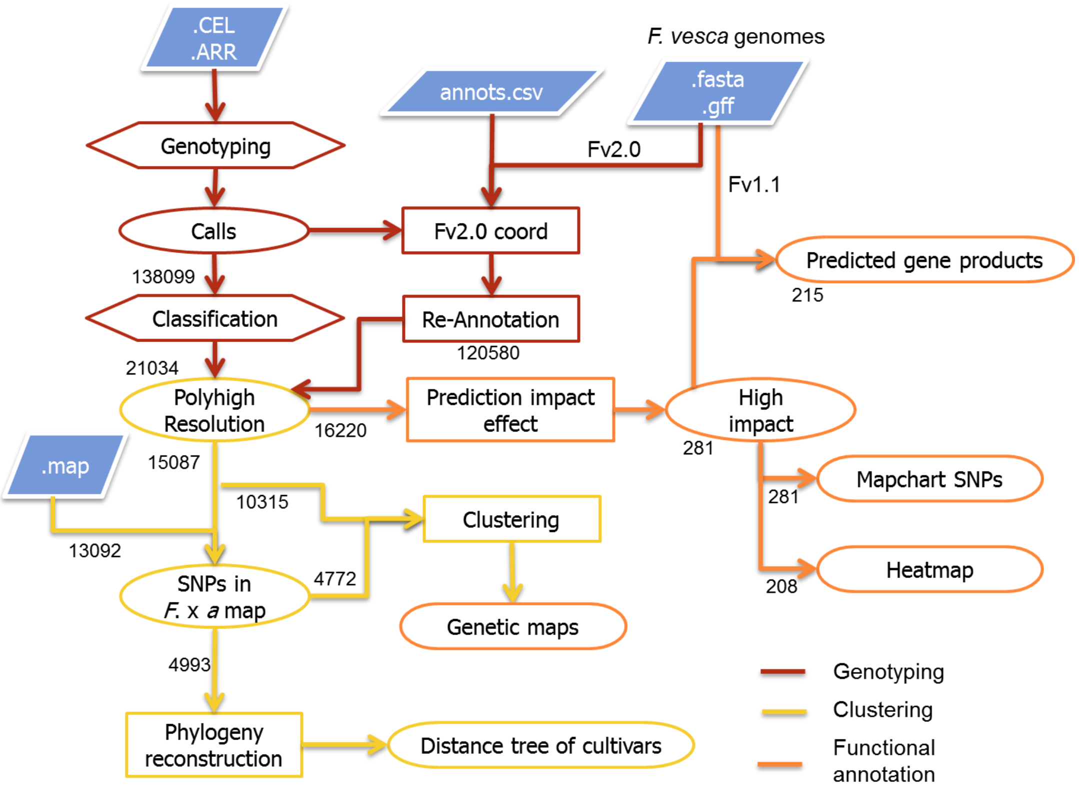


Figure 8. General workflow of the analysis. The numbers correspond to the number of probesets used at each step.

3.3. Genotyping and classification

The first step taken in analyzing the data was generating the genotype calls from CEL intensity files and grouping SNPs into categories according to cluster quality. In this context, SNP calling is the result of processing all the probes related to one SNP at a time.

Genotyping was performed with the Genotyping Console™ (GTC) Software 4.0 (Affymetrix) using modified conditions of the recommended workflow. Firstly, pre-genotyping quality control analysis (QC) was performed. According to "Affymetrix's

Best Practices”, all samples passed the default QC thresholds. Afterwards, probesets were subjected to two genotyping steps. The evaluation of the first genotyping step showed that call rate metrics of essential samples were below the recommended threshold of 97%. Only non-essential samples were excluded before proceeding to the second genotyping step. Finally, the software generated multiple output files including genotyping calls, QC metrics and posteriors (genotype clusters).

A post-process analysis was carried out using SNPolisher v1.5.2 (Affymetrix), an R package designed to classify SNPs into categories using the GTC standard output files. The result was the assignment of SNPs into one of six quality classes according to various quality measures as described at Section 1.4.2. Functions were prepared and run in R based on the user guide of the package, generating probeset QC metrics file and lists of marker loci in each of the previously described classes.

3.4. Re-annotation of probesets

IStraw90® Axiom® SNP array annotations for the 138,099 probesets, including homologous coordinates in *F. vesca* v1.0 draft genome (Shulaev *et al*., 2011) and SNP flanking sequences, were provided by Affymetrix. Before performing downstream analysis on the SNP dataset obtained by SNPolisher, we decided to obtain the coordinates of the SNP probesets in the most recent version of the strawberry genome (version 2.0; Tennessen *et al*., 2014). FASTA sequences corresponding to 127,541 SNP probesets (insertions and deletions were excluded) were aligned to the strawberry genome (version 2.0) using nucleotide blast search (Altschul *et al*., 1990). Ninety nine percent of the probesets (127,308) gave at least one hit. Probes that contained hits with alignment identity > 97.5% were considered as having a coordinate in v2.0 genome, taking into account that probes with hits below that percent of identity were prone to contain gaps in the location of the SNP. Finally, the custom annotation contained 120,580 probesets.

3.5. Functional annotation

Prediction on the effects of SNPs in relation with alterations on gene products was performed using snpEff 4.1 (Cingolani *et al*., 2012), which supports building a custom database with own genome and annotations. The most update high-quality annotation of the strawberry genome was recently published (Darwish *et al*., 2015), but it is based on the v1.1 of the genome. The genome sequence and annotation were downloaded from the GDR website. Thus, a snpEff database containing *F. vesca* v1.1 genome and annotations from Darwish *et al.* (2015) was constructed. A variant call format (vcf) file was generated for all SNP probesets from their sequences, provided by Affymetrix and indicating reference and alternative alleles, using custom script (basic.py; available on GitHub). Additionally, transcript sequences for all genes contained in Darwish annotation were extracted from the genome using samtools (Handsaker *et al*., 2009) and the gffread utility from Cufflinks (Trapnell *et al*., 2010).

snpEff was run using default parameters, reporting the predicted impact effects of SNPs. FASTA sequences of transcripts containing high-impact SNPs were subjected to blastx search against swissprot protein database (release 2015_06 of 27-May-2015), carried out on CLC Genomics Workbench 8.0 (CLC bio, QIAGEN). Blast results were outputted as txt table to search whether hit proteins were mapped in KEGG pathways using custom script (postsnpEff.py; available on GitHub).

Physical homologous coordinates in *F. vesca* for those high-impact SNPs were represented using MapChart 2.2 (Voorrips, 2002). In addition, calls for SNPs containing the two homozygous and hybrid states and predicted to have high impact effect were used to construct a heatmap on MeV: MultiExperiment Viewer 4.9 (TM4, Saeed *et al*., 2003) using HCL (*Hierarchical Clustering*; Eisen *et al*., 1998) and Pearson correlation.

3.6. Clustering and mapping

One of the aims of the study was to enrich the current octoploid genetic map with new polymorphic SNPs that were segregating in the 37 cultivars. In this regard, a genetic map of a *F. x ananassa* population (F2 from Camarosa x Dover cross; provided by the PhD student José Manuel Hidalgo), containing 13,092 loci markers (both SSR and SNPs), was used as a guide in the linkage group analysis (see below). The set of common SNPs present in this map and classified as *PHR* are referred from here on as “FxaRefMap”. An additional filtering of the FxaRefMap was applied, selecting SNPs that belonged to the same chromosome in both versions of *F. vesca* genomes (v1.0 and v2.0); this second group will be referred as “FxaRefMap2”.

Then, *PHR* calls were assigned to one of two sets of v2.0 chromosome files depending on whether they were in FxaRefMap2, using custom Python script (call_manip.py). Then, chromosome by chromosome, *PHR* SNPs were associated to those in FxaRefMap2 at a 90% identity using a modified Perl script from JM Hidalgo

(SNPC.pl, available on GitHub). Finally, the clustered SNPs were prepared for their analysis in JoinMap 4.1 (Stam, 1993; Van Ooijen, 2011) using custom Python script (write_calls4joinmap.py).

JoinMap is software used for the generation of genetic linkage maps in experimental segregating populations of diploid species. In our case, the species of analysis is octoploid. Thus, it would be expected to find, in the best case, a map with 28 linkage groups (7 LG x 4 homeolog groups/LG), or more if regions are split. Also, the analysis was performed using cultivars instead of an F2 population, so instead of assigning certain call to one parent, it was searched a similar call distribution among cultivars, either in phase or counter-phase (concerning homozygous calls). Samples were handled as F2 populations. The results of this program were fragments of maps where recombination between calls was minimized.

Mapping of *PHR* SNPs in JoinMap was performed in two steps. The first step consisted of performing maximum likelihood (ML) mapping with one optimization round on groups selected from the groupings tree with a minimum of 20 SNPs and associated to a same homeolog group. The resultant fragments were visually evaluated and subjected to re-phasing of homozygous calls, preparing them for the second mapping step where either ML or regression mapping algorithm were used depending on their performance.

3.7. Tree of distances

A cladogram was generated for the 37 cultivars, 5 reference varieties and 2 diploid *Fragaria* spp. Calls in FxaRefMap were loaded as polymorphic data into Tassel 4.0 (Bradbury *et al*., 2007). Then, a cladogram based on neighbor-joining clustering method was created using a distance matrix generated by the program. A tree representation was obtained with Mega 6.0 (Tamura *et al*., 2013).

4. Results

4.1. SNP calling

IStraw90 SNP array data from 37 cultivars and 7 additional samples were processed in the context of an octoploid population. Affymetrix Genotyping Console pipeline recommends performing quality control (QC) check followed by two genotyping steps to filter samples (see Section 3.3). In our analysis, all probesets passed QC metrics. After the first genotyping step, 10 samples of interest had call rate below 97% and only one below 95%. These samples were on the focus of the analysis and could not be discarded, so all of them were subjected to the second genotyping step, resulting to 16 samples being below call rate 97%, 5 of which were even below 95.5%. After consulting Affymetrix's recommendations and discussing with experienced users of the software, we decided to use all the cultivars independently of their call rate. Default files with calls and posteriors (see Section 3.3) for all 138,099 probesets in the array were used for classification in quality classes using SNPolisher. The distribution of SNPs into these six classes is described in Table 3.

Table 3. Distribution of genotyped markers and probes by quality class.

Conversion type	Marker Loci	%	Probesets	%
Poly High Resolution	21,034	22	24,489	18
No Minor Homozygote	39,646	42	56,683	41
Off-Target Variant	1,028	1	1,505	1
Mono High Resolution	21,964	23	36,594	26
Call Rate Below Threshold	4,448	5	6,704	5
Other	6,942	7	12,124	9
Total	**95,062**	**100**	**138,099**	**100**

Most of the markers were annotated as high resolution, represented by the classes *PHR*, *NMH* and *OTV*. For downstream analysis, only *PHR* marker loci were used, on the one hand to see only SNPs that distinguish among the three genotypic clusters for each cultivar (two homozygous and one heterozygous) and, on the other hand, to avoid an overlap of probesets targeting the same SNP. At this stage, the number of *PHR* was inflated due to the presence of replicated samples in the additional lines, and also the inclusion of the two diploid *Fragaria* species increased even more the diversity. Then, *PHR* calls were filtered to select those that were segregating in the cultivars, as explained in the following paragraph.

4.2. Filtering of probes and annotation to v2.0 genome

The IStraw90 array contained probes not only for SNPs, but also for insertions and deletions. Moreover, it contained references to v1.0 draft genome while a newer version of the genome has become available. Thus, a custom filter of probes was applied to select those probes corresponding to SNPs and map them to the new genome. For this purpose, probeset nucleotide sequences were blasted against v2.0 and those not containing gaps in their alignments were provided with new coordinates (see Section 3.4). This list was used to filter *PHR* marker loci, retrieving 16,220 SNPs with homolog coordinates to v2.0 genome. These SNPs were later used for functional annotation. A final filtering step was applied by selecting only those markers that were segregating in the 37 cultivars, obtaining a final number of 15,087 SNPs that were used for clustering.

4.3. Construction of genotype maps

Maps were constructed using *PHR* SNPs, which were divided based on their chromosome location on v2.0 and analyzed independently. First, SNPs were classified into two groups: one containing those represented in the FxaRefMap2 (a total of 4,772 SNPs; see Section 3.6) and a second containing the rest of them (a total of 10,315 SNPs; Figure 9 A). Afterwards, SNPs on FxaRefMap2 were clustered with *PHR* SNPs that had at least 90% of identity, getting a set of clusters where *PHR* could be associated more than once to the FxaRefMap2 SNPs (clustered, Figure 9 A). After this step, the clusters contained 60-75% of their corresponding initially mapped SNPs and 16-25% of their *PHR* SNPs (Figure 9 B). The latter were checked whether they had been clustered more than once and, if so, assigned to a same homeolog group by comparison with FxaRefMap2. In case they were associated to different homeolog groups, they were discarded (*Dif homeolog*, Figure 9 A). A single file merging both final *PHR*/mapped clustered SNPs was prepared for downstream analysis, containing 11-18% of the initial *PHR*, which represented around half of the mapped SNPs (Figure 9 B). A detailed distribution of clustered SNPs on homeolog groups is provided in Annex I.

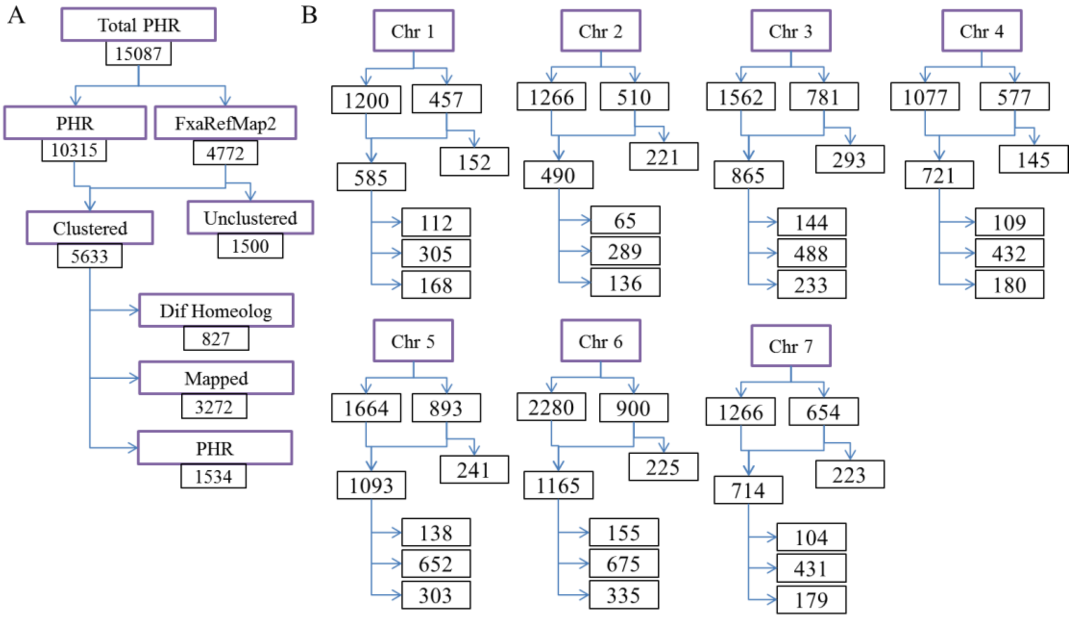


Figure 9. Diagram of SNP clustering. Total number of markers after each step (A) and numbers by chromosome (B). *PHR* SNPs were divided on two non-overlapping groups with respect to being present on FxaRefMap2. *PHR* were compared to each SNP in *FxaRefMap2* and clustered if identity ≥ 90%, yielding a file with *clustered* markers (both types) and a file with *unclustered* mapped SNPs. *PHR* SNPs clustered to more than one homeolog group were discarded (*Dif Homeolog*) and the rest (*Mapped* + *PHR*) formatted for downstream analysis.

Once the SNPs were properly prepared and formatted, they were analyzed in JoinMap, a program that generates maps using SNPs of similar distribution. The program itself generates a groupings tree where groups of SNPs can be selected and analyzed separately. In these groups, SNPs could be inversed regarding homozygous calls. To unify call phases and get longer maps, two mapping rounds were performed. In both rounds, groups were selected with the priority to maximize the number of SNPs assigned to one same homeolog group. In the first round, the main mapping algorithm used was maximum likelihood (ML), which did not have the restriction of a minimum linkage between SNPs, so genotypes in both phases could be seen together in the resultant map fragments. Taking into consideration the relative position on the reference map of the fragments, SNPs were re-phased according to the parental genotypes and subjected to a second mapping round. The final fragments were joined together as shown in Figure 10. The final 33 maps are available online on GitHub (Maps.xlsx), where SNPs are named after their Affymetrix probeset id; numbers on the left indicate the coordinate on the reference genetic map for mapped SNPs; names starting by “A_” indicate SNPs associated to any of the mapped SNPs; the final tag indicates the corresponding homeolog group.

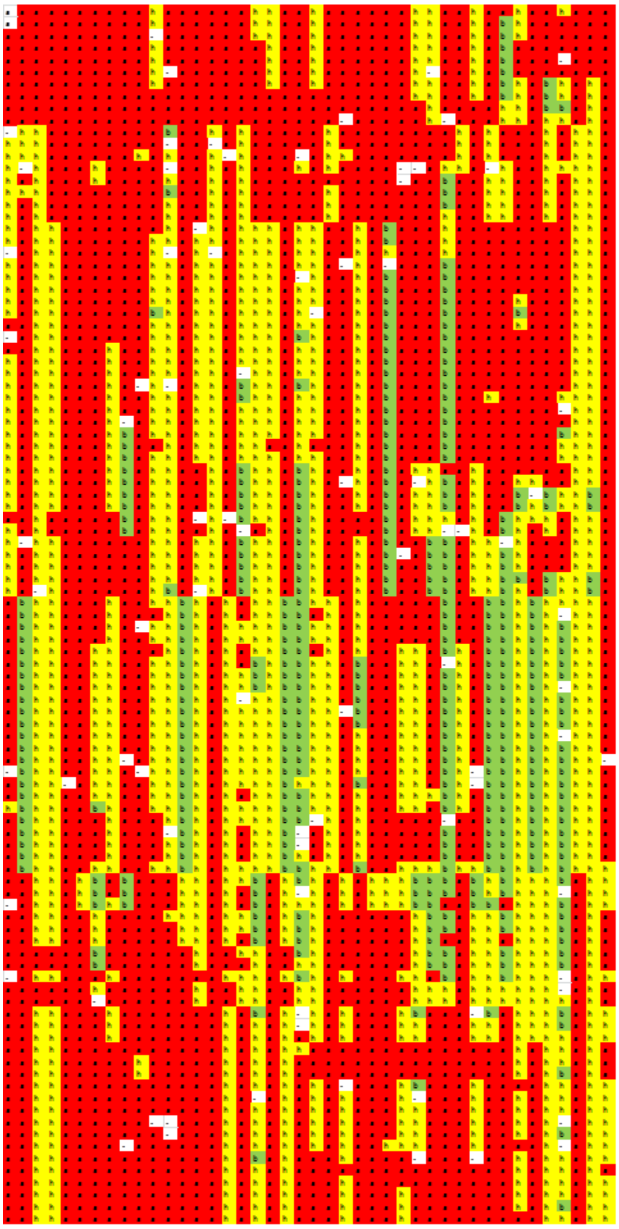


Figure 10. Reconstruction of a genetic map for group LG7-B. JoinMap fragments were ordered in relation to the position of SNPs in the reference map. Cultivars are on columns and SNPs in rows. Genotypes are represented as homozygous (red and green), heterozygous (yellow) and no call (white). Homozygous calls were inversed when appropriate to improve visualization.

When joining fragments together, some of them fitted well with the previous ones. As an example, the map shown in Figure 10 was made out of 3 fragments. Some cultivars maintained homogeneous genotype, while others presented recombination events. Those breaks showing increased recombination would mark the limits between conserved regions. These regions are comparable to haplotypes, genomic regions that are associated and inherited as a block. In this example, it was also noticeable how heterozygosity decreased at the tips of the chromosome suggesting that recombination took place at inner locations.

The distribution of fragments across the seven reference chromosomes was very diverse in number and content of SNPs (Table 4). The total number of fragments ranged from 4 in chromosome 2 and 18 in chromosome 6. Many homeolog groups had only one map, not being able to extend much the current map. On the contrary, some of them had fragments covering most of the current map, either by more small fragments or less big ones. An important observation concerned the number of associated SNPs that were added to the maps. In some homeolog groups, the new information was limited to less than 10 markers. The best cases had 50-80 new markers on their maps, representing a potential improvement to the quality (or resolution) of the current map.

Table 4. Distribution of SNPs in fragments of linkage and homeolog groups generated by JoinMap.

Group	Fragments	Max	Mapped	Assoc	Group	Fragments	Max	Mapped	Assoc
LG1	6	103	190	77	LG5	12	94	226	116
LG1A	2	61	59	17	LG5A*	1	33	25	8
LG1C	3	103	120	54	LG5B	3	63	66	29
LG1D	1	17	11	6	LG5C	5	94	109	63
LG2	4	37	51	36	LG5D	3	17	26	16
LG2A	3	19	32	18	LG6	18	79	391	183
LG2B	1	37	19	18	LG6?	2	55	41	26
LG3	13	48	254	74	LG6A	3	53	66	12
LG3?1	2	21	30	8	LG6B	1	16	10	6
LG3?2	3	41	56	19	LG6C	6	71	125	57
LG3?3	1	48	43	5	LG6C1	1	21	17	4
LG3A	5	44	91	34	LG6D	5	79	132	78
LG3B1	1	22	18	4	LG7	10	39	157	63
LG3B2	1	20	16	4	LG7A	4	29	67	16
LG4	8	80	204	84	LG7B	4	39	64	37
LG4?1	4	23	54	16	LG7C	1	14	13	1
LG4B	3	70	93	45	LG7D	1	22	13	9
LG4D	1	80	57	23					

Looking more closely at fragments, the mean size was similar in all chromosomes except in 1, where instead of being around 20 marker loci long it was near 40 (left boxplot in Figure 11). Some of the fragments did not even have a dozen SNPs, being in some cases low informative. This aspect is detailed in the boxplot on the right side, which indicates the % of associated SNPs in fragments. Here it is seen that some fragments only had FxaRefMap2 SNPs, giving no new information. In most cases, the relation was around one third of new markers in fragments, which is a good contribution. The most noticeable cases provided more than 50% of new markers.

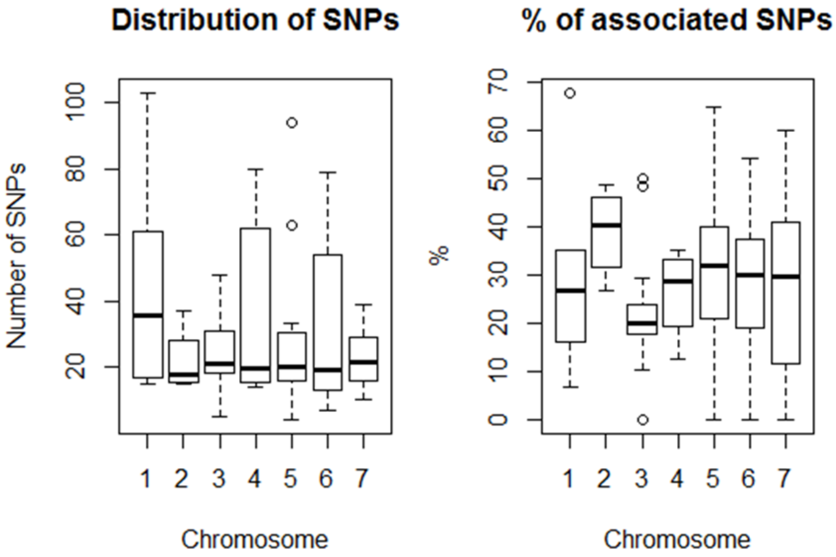


Figure 11. Boxplots of number of SNPs (left) and percentage of associated SNPs (right) across all JoinMap fragments by chromosome.

Due to the steps followed in this process, some information was lost. One clear example is related with the number of homeolog groups remaining after each step (Table 5). When SNPs were clustered, there was around a 40% of SNPs on FxaRefMap2 that did not cluster with the *PHR* (*unclustered*, Figure 9), in some cases representing the whole set of a homeolog group. After that, clustered SNPs that were not part of selectable groups in JoinMap groupings tree were left behind. At this point most of the chromosomes had lost a group. Chromosome 2 was the most affected, getting only maps for 2 of the initially 5 homeolog groups. On the contrary, all initial groups in chromosome 6 were represented in JoinMap fragments.

Table 5. Variation in the number of homeolog groups represented after each step during the clustering and mapping processes.

	Number of homeolog groups			
Chromosome	In Map	In clusters	In fragments	Final %
1	5	4	3	60.0
2	5	4	2	40.0
3	7	7	6	85.7
4	4	4	3	75.0
5	5	5	4	80.0
6	6	6	6	100.0
7	5	4	3	60.0

Around 80% of *PHR* SNPs in each chromosome were not used due to absence of association with SNPs on FxaRefMap2. A test analysis in JoinMap revealed that they actually formed maps (data not shown). These could be representing regions that do not usually segregate in population so markers were not present in the reference genetic map, which mostly benefits from *PHR* SNPs in the population.

4.4. Cladistic representation of cultivars

A tree of distances among cultivars (Figure 12) was constructed using the SNPs included on FxaRefMap (a total of 4, 993 *PHR* SNPs; see Section 3.6). As a result, the selection contained markers that were segregating in both population and cultivars. Also, these markers would be expected to be mostly neutral variations, reducing possible bias of the data. The set of cultivars contained plants from two geographical locations: (1) Huelva (Spain), samples named by a single number, “P” or others; and (2) Le Barp (France), samples starting by “D-” (see Section 3.1). In the cladogram, samples from both locations had tendency to cluster not together but separately of others, indicating that geographical origin did not determine the distribution.

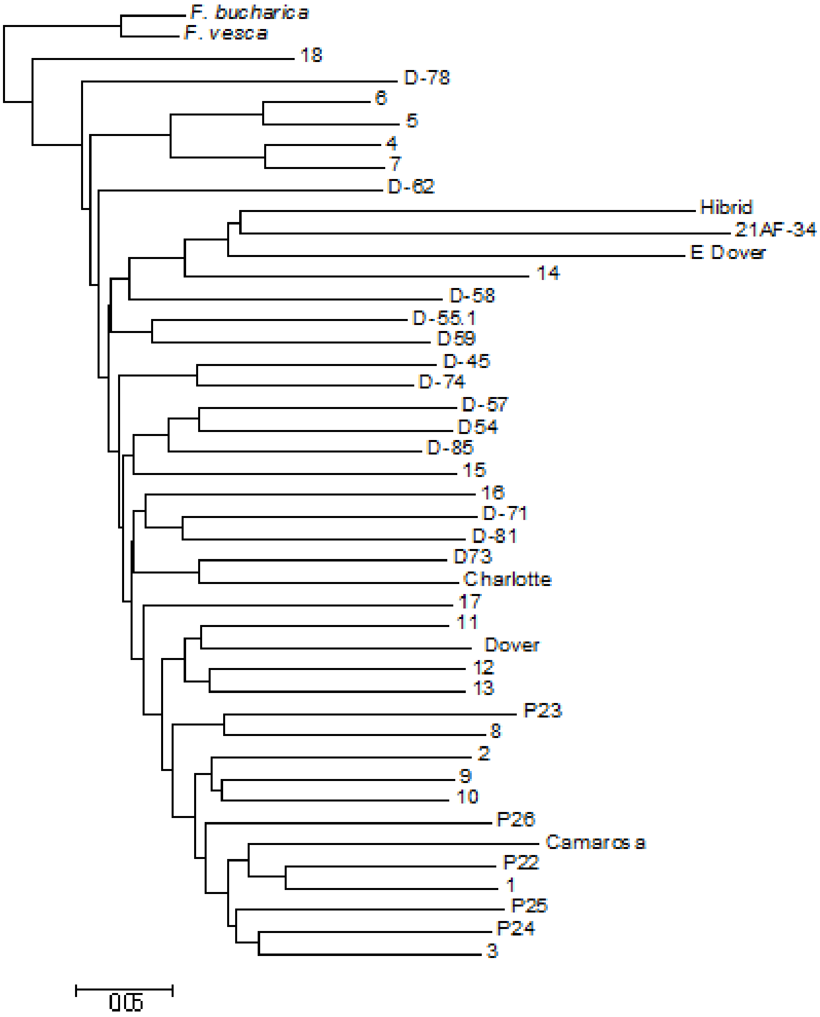


Figure 12. Cladistic representation of genotype distances among 42 cultivars generated by Neighbor-Joining algorithm. The diploids *F. vesca* and *F. bucharica* were used as outgroups.

The breeding company had provided some basic characteristics for most of the cultivars (Table 2), and those from France also had parental information. Looking at the cladogram in relation with some of the given characteristics, more correlation was observed. Sample 18 was close to the diploid species (*F. vesca* and *F. bucharica*) and was described as possessing woodland strawberry flavor, suggesting strong influence of the genotype over this feature in this cultivar. The clade containing samples 4-7 and D-62 represented fruits with improved flavor. Cultivars 2, 8 and 9, which were close in the tree, possessed precocity. In D-45/D-74 clade, the samples were genetically related as D-45 was the female parent of D-74. The limitations in the provided phenotypic data did not allow establishing more accurate relationships, although it was clear that many variables were influencing the clustering.

4.5. Prediction of effects of PHR markers

For functional annotation of *PHR* SNPs, the most recent strawberry gene annotation (Darwish *et al*., 2015) containing 33,073 predicted protein coding genes was used to build a snpEff database. The analysis was performed on a subset of 16,220 *PHR* SNPs that were present in the custom SNP annotation, formatted as vcf. Running snpEff yielded another vcf file with predicted functional annotations, including affected genes, impact of effects, location with respect to the gene features and, when a coding SNP was concerned, an amino acid change. The predicted impact of effect and type of effect are shown on Table 6.

Table 6. Distribution of *PHR* variants and predicted effects as function of functional class type and impact type, respectively.

Impact type	Count	Percent	Functional class type	Count	Percent
High	280	0.65%	Missense	4,125	25.43%
Low	9,987	23.11%	Nonsense	214	1.32%
Moderate	4,084	9.45%	Silent	9,681	59.69%
Modifier	28,864	66.79%	Other	2,200	13.56%
Total Effects	43,215	100%	Total Variants	16,220	100%

There were almost three times more effects than variants (Table 6), indicating that more than one effect were predicted for variants. With respect to impact type, most of the SNPs had a modifier effect in relation with adjacent genes. All silent mutations, composed of *synonymous variant* class, were annotated as “low impact” SNPs and comprised 97% of total effects in this category, together with *splice region* and *intron* localized (2%) and other types of variants. “Moderate impact” effects were due to *missense* (98%) and *splice region* (2%) variants. Finally, mutations with “high impact” mainly implied *gain of stop codon* (74%), but also critical effects were associated to alterations concerning *splice donor/acceptor* sites (14%), *loss of stop codon* (9%) and *loss of start codon* (3%).

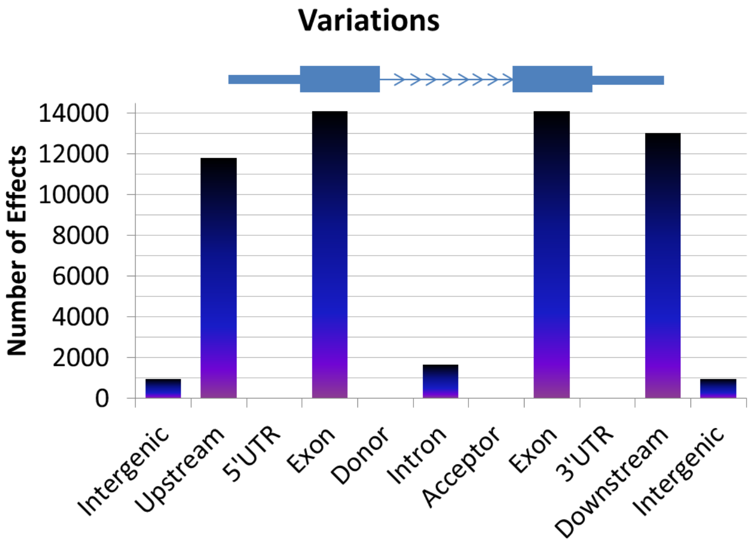


Figure 13. Distribution of effects among genic and intergenic features.

The effects predicted for the given variations were plotted in relation to the position in a hypothetical gene (Figure 13). In the graph, exon and intergenic values in both columns represent the same value. The report indicated that 14,087 variants were localized into exons, representing the 87% of total variants, indicating a strong accumulation of SNPs in genes. Most of these variants were also marked as upstream or downstream variants in relation to the adjacent gene. A small proportion of variation was found in intergenic and intronic regions, where they would be expected to have little effect. Very few SNPs were located on UTR and splice site regions, where they would be likely to affect negatively the transcription or RNA processing.

We also wanted to see how these SNPs with high impact on the genes were distributed on the homolog *F. vesca* chromosomes. For that purpose, their physical coordinates in v1.1 were represented in chromosome charts (Figure 14). As observed, there was a general good coverage of SNPs. The biggest gaps were present at chromosome 1, but the distribution was quite homogeneous. Thus, there was not a region enriched in this type of candidate markers.

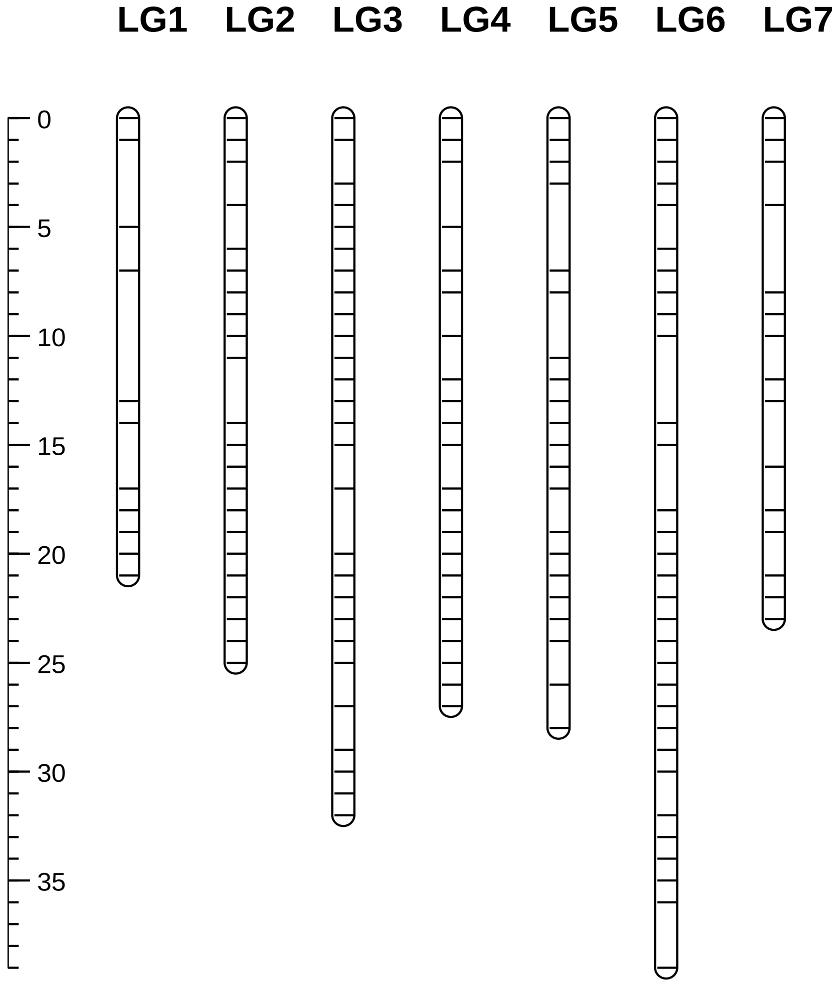


Figure 14. Map chart of the seven linkage groups of *F. vesca* v1.1 draft genome with representation of *PHR* SNPs annotated as "high impact" by snpEff.

The last representation performed consisted of a heatmap (Figure 15). In this case, 208 *PHR* SNPs predicted by snpEff as high impact and possessing both reference and alternative homozygous and heterozygous calls were selected. This representation allowed us to detect possible enrichment of specific SNPs in groups of varieties. For this analysis, reference lines were discarded. The representation was very heterogeneous. The most conserved block was observed on the upper right corner, where the information about the implied cultivars was about precocity (2, 3, 8, 9 and P23) and flavor (1, 10, and P22). The rest of the cultivars showed similar distribution in that set of SNPs, seeing bands on the homozygous calls. Blocks for other clades were not as apparent as the previous one.

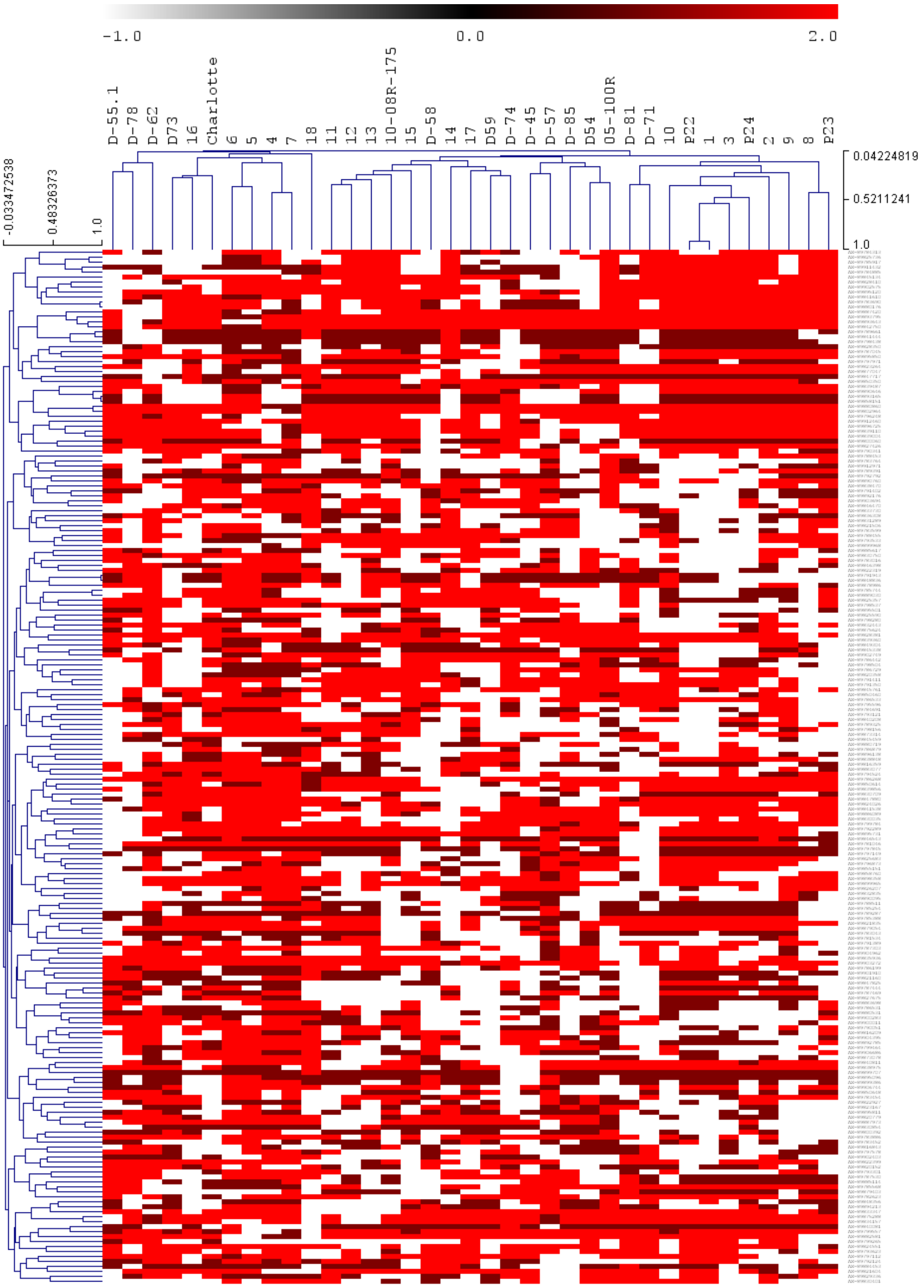


Figure 15. Heatmap representation of *PHR* SNPs classified as high impact. Black and red represent reference and alternative homologous calls, respectively. White represents no calls and heterozygous calls.

4.6. Gene annotation of "high impact" SNPs

All transcripts of predicted genes were obtained from the genome v1.1 using the latest annotation. From those, 276 sequences carried at least one of the 280 SNP markers with predicted high potential to impact gene expression. These putative genes were subjected to blast search against the curated UniProtKB/Swiss-Prot database to identify reliable hits. The results yielded 8 proteins without any hit and, from the rest of the hits, 215 presented an e-value less than 0.1. Hits were also assigned to a KEGG pathway when available, although only 77 of them had a KEGG id, 49 of which had pathway annotation that was reduced to 41 on hits with e-value less than 0.1. A table of gene annotations is available on GitHub (annotation.xlsx).

The number of pathways where these SNPs were involved was large. Some examples involved biosynthesis (gluconeogenesis, fatty acid, phenylpropanoids), metabolism (starch and sucrose, galactose, amino acids), transcription factors, transporters, nucleic acid processes (repair, transport, degradation, splicing) and protein degradation. These results also demonstrate the complexity of a marker-to-phenotype association analysis.

5. Discussion

5.1. Validation of the mapping strategy using genotypes

The absence of a reference genome for the octoploid strawberry sets the focus on building genetic maps. Most of published maps are mainly based on SSR markers (Rousseau-Gueutin *et al*., 2008; Sargent *et al*., 2009; van Dijk *et al*., 2014), allowing their comparison with *F. vesca* references. In the case of using IStraw90 markers, only one genetic map has been published to date based upon the cross “Holiday” x “Korona” (Bassil and Davis *et al*., 2015); our provided reference genetic map benefited from this one. The development of genetic maps based upon crosses between different cultivars will allow obtaining finer maps after their comparison.

We faced the problem of handling a set of cultivars derived from variable parents instead of a F2 population. Thus, we adopted a different strategy based on genetic conservation in close regions; instead of assigning alleles to one parent, markers were grouped by genotype similarity among samples, using SNPs that segregated in the population as seeds for clustering. Traditional genetic maps using populations, in *Fragaria* causes loss of information due to the abundance of monogenic markers (Sargent *et al*., 2009). Here we used cultivars, where higher levels of recombination are promoted, with promising results.

A second problem was the use of 42 individuals for building the maps, a relatively low number. The potential drawback is to cluster together markers that randomly have adopted identical or very similar genotypes. A clustering test performed over all *PHR* SNPs revealed some of them from different LGs with identical genotype distribution. Then, we decided to separate the analysis by LG. Even though this fragmentation avoids wrong clustering of SNPs from diverse LG, some wrong associations still were seen. The self-evident cases, which were easily discarded, implied those markers that associated to more than one homeolog group. Looking at the genotypes of the parental lines along the maps, those markers requiring double recombination processes to fit are possibly wrongly associated.

In our maps, conserved regions with low minor allele frequency were also observed. These regions usually concentrated at the extremes of the maps. In some cases, markers localized at different genetic positions in the reference map are associated together and to more than one haplotype. This is the case for some regions where initial and final SNPs in the reference LG were associated due to the low allele frequency. The interesting cases imply central markers belonging to at least two polymorphic haplotypes. Taking into account that we are mapping genotypes instead of parental markers, this fact suggests co-segregation of markers within regions of low recombination and that may be separated physically. The extent to which these markers are consecutive or interleaved on the genome cannot yet be addressed until the whole octoploid genetic map has been discovered. Testing the reordering of markers based on their physical position revealed a mix of haplotypes along the genomic sequence.

The use of genotypes may be considered poorly reliable at first. Our results indicate consistency of this method considering that markers were grouped usually in relation to the reference genetic map. In this regard, most of the fragments generated by JoinMap contained markers within a narrow genetic distance range. This consistency suggests that the maps yet to be constructed and formed by the non-associated *PHR* SNPs may contain as well markers genetically related. The transfer of markers in these maps to a population analysis may be useful to fill some of the gaps in the actual genetic maps.

5.2. Improvement of the current genetic map

We have produced an extension of the available *F.* x *ananassa* genetic linkage map using SNP marker loci from a set of cultivars, with a total of 71 linkage groups (LGs), made up of 2,106 loci (1,473 of which are shared with the reference genetic map FxaRefMap2). Our contribution is the association of 633 new markers to the map, which are distributed in 73% of the initial reference homeolog groups. This number may rise if we include those SNPs that have identical segregation pattern to the ones selected and mapped by JoinMap but were left apart from the analysis by the program, which only maps non-identical markers.

Since we had sequence information for the SNP markers, they have been associated according to their homologous groups to the latest diploid *Fragaria* reference map (Tennessen *et al*., 2014). Most of the markers on the genetic map associated to one LG (named after genome v1.1) had homologous coordinates on the same chromosome in genome v2.0. But for 5% of the markers, the annotated probeset coordinate on v1.1, the assigned homologous coordinate on v2.0 and the genetic LG were discordant. Three patterns were observed: (1) genetic and v1.1 LGs were equal but not v2.0 LG; (2) v1.1 and v2.0 LGs were equal but not genetic LG; and (3) the three were different. We suggest as possible causes; (i) the reference genome v2.0 has translocated regions or sequences; (ii) the SNPs belong to repeated regions so our homologous coordinate on v2.0 is wrong; (iii) there are translocation events in *F.* x *ananassa* with respect to *F. vesca*; and (iv) the genetic map needs further improvement.

5.3. Identification of potential candidate genes

An ultimate goal towards understanding the variation in the quality traits is to identify the responsible genes. In our work, we have used the most recent strawberry annotation (Darwish *et al.*, 2015) to predict potential effects of polymorphic SNPs on genes, focusing on those changes with the most drastic effects on gene products. From these, the 91 SNPs that are already present in the *F.* x *ananassa* genetic map are putative targets regarding future studies.

We obtained a list of 215 highly affected genes that were segregating in the cultivars. These genes in *F. vesca* may have up to eight repetitions in *F.* x *ananassa*, even though the selected markers in the microarray preferentially showed reduced ploidy. In this case, we do not know whether copies of the genes in other homeolog chromosomes are silenced or more than one homeolog isoform may be active at the same time. Considering that only one of the subgenomes carried the active gene, the potential effect of SNPs would be enhanced in homozygosity. This consideration was taken to build the heatmap (Figure 15), highlighting only homozygous calls.

One work focusing on the isolation of ripening-related genes in strawberry (Manning, 1998), described cDNA clones enhanced by ripening. Among those genes, pyruvate decarboxylase (PDC) is present in our set of candidate genes. PDC showed an expression profile correlated to that of a gene involved in flavor biogenesis, strawberry alcohol acyltransferase (SAAT; Aharoni *et al.*, 2000). Later, the study of strawberry PDC (Moyano *et al.*, 2004), described two isoforms, one of them related to fruit ripening and aroma biogenesis. In the promoter of these genes there were binding sites for transcriptional factors (TF) like myb, myc, auxin response factor (ARF) and low-temperature responsive element. Among our candidate TFs, one myb and one ARF are present, suggesting a role of this pathway on cultivar phenotypic variation at the flavor level.

Other candidate protein mentioned in the bibliography is glutathione *S*-transferase (GST; Manning, 1998; Aharoni *et al*., 2000). GSTs may be involved in the export of anthocyanins from the cytoplasm, their site of synthesis, to the vacuoles for storage (Mueller *et al*., 2000). In strawberry, it may be a possible pigmentation-related gene.

5.4. Multiple sources of variation

As seen on the cladogram (Figure 12) and the heatmap (Figure 15), the variation among cultivars does not relate to specific traits. Phenotype is the manifestation of a complex interaction between the genotype and the environment. Further analysis of markers with alternative alleles in phenotype-genotype blocks could suggest genes implied in that trait. In the example of the two cultivars sharing wild strawberry-like flavor, Charlotte and 18, which are in separate clades, their common genotype may point to flavor related genes. The heatmap failed to produce clear SNP clusters, which would have suggested markers co-segregating in function of a certain characteristic. Also, the relations that grouped the cultivars were variable, including phenotypic and genetic relationship associations. Proposing an additional analysis to perform when characteristics for all cultivars are available, it would be to re do the heatmap based on a selection of cultivars that show being associated by their phenotypes.

5.5. Prospective

The development of a genetic map and, most importantly, the prediction of genes that may be related to quality or agronomical traits, have great potential to impact strawberry breeding. One major goal is determining a small set of genetic markers to facilitate plant selection to breeders. In this regard, we have provided a list of 215 potential marker loci. We have discussed the linked segregation of markers, so it may represent an issue if markers with undesirable effect co-segregate. Another risk derived from planting uniform crops is the reduction of genetic diversity, which may lead to the vulnerability of crop plants to environmental stresses, such as pathogens (Strange and Scott, 2005). A deep knowledge on markers associated with traits such as fruit quality or plant resistance to pathogens will help overcome these problems.

Recent studies are targeting the use of knock out mutants, either in diploid strawberry or *Arabidopsis* when possible, on candidate genes involved in flavor and aroma biogenesis, color or pathogen resistance. Such is the case described in Lin-Wang *et al*. (2014), where they studied the alteration on anthocyanin concentration by regulating the expression of one member of the MYB family. The extrapolation of such results to octoploid strawberry breeding constitutes a potential source not only for quality but also agronomic enhancement.

6. Conclusions

The main conclusions derived from the present work are:

1) IStraw90® Axiom® SNP array in *F.* x *ananassa* yields around 25% SNPs as Poly High Resolution SNPs suitable for constructing high quality genetic maps.

2) Mapping SNP markers based on genotypes is a feasible methodology to improve genetic maps based on a population in nearby regions of low segregation, regarding the consistency of our maps. We provide 33 genotype maps covering 27 of the initial 37 homeolog groups from the reference genetic map.

3) Haplotypes, determined by the co-segregation of markers, are well defined in our maps. They are not consecutive, but interleaved, so larger regions may segregate together.

4) The *F. vesca* latest reference map v2.0 is likely to contain translocations.

5) Genetic distance among cultivars responds to several variables, including phenotype and genetic relationships. Cultivars related by a common characteristic usually share a block of SNPs with identical segregation.

6) Several polymorphic genes are susceptible to loss of function due to the variations caused by their SNPs. Some of these genes are actually studied in relation with quality strawberry traits. Future association studies of phenotype-genotype may reveal candidate genes from this list.

7. Bibliography

Aharoni, A., L. C. Keizer, H. J. Bouwmeester, Z. Sun, M. Alvarez-Huerta, H. A. Verhoeven, J. Blaas, A. M. van Houwelingen, R. C. De Vos, H. van der Voet, R. C. Jansen, M. Guis, J. Mol, R. W. Davis, M. Schena, A. J. van Tunen and A. P. O'Connell (2000). "Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays." Plant Cell **12**(5): 647-662.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Antanaviciute, L., F. Fernandez-Fernandez, J. Jansen, E. Banchi, K. M. Evans, R. Viola, R. Velasco, J. M. Dunwell, M. Troggio and D. J. Sargent (2012). "Development of a dense SNP-based linkage map of an apple rootstock progeny using the *Malus* Infinium whole genome genotyping array." BMC Genomics **13**: 203.

Bassil, N. V., T. M. Davis, H. Zhang, S. Ficklin, M. Mittmann, T. Webster, L. Mahoney, D. Wood, E. S. Alperin, U. R. Rosyara, H. Koehorst-Vanc Putten, A. Monfort, D. J. Sargent, I. Amaya, B. Denoyes, L. Bianco, T. van Dijk, A. Pirani, A. Iezzoni, D. Main, C. Peace, Y. Yang, V. Whitaker, S. Verma, L. Bellon, F. Brew, R. Herrera and E. van de Weg (2015). "Development and preliminary evaluation of a 90 K Axiom(R) SNP array for the allo-octoploid cultivated strawberry *Fragaria* x *ananassa*." BMC Genomics **16**: 155.

Basu, A., A. Nguyen, N. M. Betts and T. J. Lyons (2014). "Strawberry as a functional food: an evidence-based review." Crit Rev Food Sci Nutr **54**(6): 790-806.

Bianco, L., A. Cestaro, D. J. Sargent, E. Banchi, S. Derdak, M. Di Guardo, S. Salvi, J. Jansen, R. Viola, I. Gut, F. Laurens, D. Chagne, R. Velasco, E. van de Weg and M. Troggio (2014). "Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus* x *domestica* Borkh)." PLoS One **9**(10): e110377.

Bombarely, A., C. Merchante, F. Csukasi, E. Cruz-Rus, J. L. Caballero, N. Medina-Escobar, R. Blanco-Portales, M. A. Botella, J. Munoz-Blanco, J. F. Sanchez-Sevilla and V. Valpuesta (2010). "Generation and analysis of ESTs from strawberry (*Fragaria* x *ananassa*) fruits and evaluation of their utility in genetic and molecular studies." BMC Genomics **11**: 503.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss and E. S. Buckler (2007). "TASSEL: software for association mapping of complex traits in diverse samples." Bioinformatics **23**(19): 2633-2635.

Chagne, D., R. N. Crowhurst, M. Troggio, M. W. Davey, B. Gilmore, C. Lawley, S. Vanderzande, R. P. Hellens, S. Kumar, A. Cestaro, R. Velasco, D. Main, J. D. Rees, A. Iezzoni, T. Mockler, L. Wilhelm, E. Van de Weg, S. E. Gardiner, N. Bassil and C.

Peace (2012). "Genome-wide SNP detection, validation, and development of an 8K SNP array for apple." PLoS One **7**(2): e31745.

Chambers, A. H., J. Pillet, A. Plotto, J. Bai, V. M. Whitaker and K. M. Folta (2014). "Identification of a strawberry flavor gene candidate using an integrated genetic-genomic-analytical chemistry approach." BMC Genomics **15**: 217.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3." Fly (Austin) **6**(2): 80-92.

Congiu, L., M. Chicca, R. Cella, R. Rossi and G. Bernacchia (2000). "The use of random amplified polymorphic DNA (RAPD) markers to identify strawberry varieties: a forensic application." Molecular Ecology **9**(2): 229-232.

Da Silva, J. A. and B. W. Sobral (1996). Genetics of polyploids. The impact of plant molecular genetics, Springer**:** 3-37.

Darrow, G. M. (1966). The strawberry; history, breeding, and physiology. New York,, Holt.

Darwish, O., R. Shahan, Z. Liu, J. P. Slovin and N. W. Alkharouf (2015). "Re-annotation of the woodland strawberry (*Fragaria vesca*) genome." BMC Genomics **16**: 29.

Degani, C., L. J. Rowland, J. A. Saunders, S. C. Hokanson, E. L. Ogden, A. Golan-Goldhirsh and G. J. Galletta (2001). "A comparison of genetic relationship measures in strawberry (*Fragaria* x *ananassa* Duch.) based on AFLPs, RAPDs, and pedigree data." Euphytica **117**(1): 1-12.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-14868.

FAOSTAT. (2015). Retrieved Juny 8, 2015: http://faostat3.fao.org/.

Folta, K. M. and T. M. Davis (2006). "Strawberry genes and genomics." Critical Reviews in Plant Sciences **25**(5): 399-415.

Folta, K. M., M. Staton, P. J. Stewart, S. Jung, D. H. Bies, C. Jesdurai and D. Main (2005). "Expressed sequence tags (ESTs) and simple sequence repeat (SSR) markers from octoploid strawberry (*Fragaria* x *ananassa*)." BMC Plant Biol **5**: 12.

GDR. "Genome Database for *Rosaceae*": https://www.rosaceae.org/.

Giampieri, F., S. Tulipani, J. M. Alvarez-Suarez, J. L. Quiles, B. Mezzetti and M. Battino (2012). "The strawberry: composition, nutritional quality, and impact on human health." Nutrition **28**(1): 9-19.

GitHub: https://github.com/eldamarth/MasterThesisBioinf.

Jung, S., A. Cestaro, M. Troggio, D. Main, P. Zheng, I. Cho, K. M. Folta, B. Sosinski, A. Abbott, J. M. Celton, P. Arus, V. Shulaev, I. Verde, M. Morgante, D. Rokhsar, R. Velasco and D. J. Sargent (2012). "Whole genome comparisons of *Fragaria*, *Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies." Bmc Genomics **13**.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Lin-Wang, K., T. K. McGhie, M. Wang, Y. H. Liu, B. Warren, R. Storey, R. V. Espley and A. C. Allan (2014). "Engineering the anthocyanin regulatory complex of strawberry (*Fragaria vesca*)." Frontiers in Plant Science **5**.

Manning, K. (1998). "Isolation of a set of ripening-related genes from strawberry: their identification and possible relationship to fruit quality traits." Planta **205**(4): 622-631.

Menager, I., M. Jost and C. Aubert (2004). "Changes in physicochemical characteristics and volatile constituents of strawberry (Cv. Cigaline) during maturation." J Agric Food Chem **52**(5): 1248-1254.

Moyano, E., S. Encinas-Villarejo, J. A. Lopez-Raez, J. Redondo-Nevado, R. Blanco-Portales, M. L. Bellido, C. Sanz, J. L. Caballero and J. Munoz-Blanco (2004). "Comparative study between two strawberry pyruvate decarboxylase genes along fruit development and ripening, post-harvest and stress conditions." Plant Science **166**(4): 835-845.

Mueller, L. A., C. D. Goodman, R. A. Silady and V. Walbot (2000). "AN9, a petunia glutathione S-transferase required for anthocyanin sequestration, is a flavonoid-binding protein." Plant Physiology **123**(4): 1561-1570.

Negri, A. S., D. Allegra, L. Simoni, F. Rusconi, C. Tonelli, L. Espen and M. Galbiati (2015). "Comparative analysis of fruit aroma patterns in the domesticated wild strawberries "Profumata di Tortona" (*F. moschata*) and "Regina delle Valli" (*F. vesca*)." Front Plant Sci **6**: 56.

Potter, D., J. J. Luby and R. E. Harrison (2000). "Phylogenetic relationships among species of *Fragaria* (Rosaceae) inferred from non-coding nuclear and chloroplast DNA sequences." Systematic Botany **25**(2): 337-348.

Rousseau-Gueutin, M., E. Lerceteau-Kohler, L. Barrot, D. J. Sargent, A. Monfort, D. Simpson, P. Arus, G. Guerin and B. Denoyes-Rothan (2008). "Comparative genetic mapping between octoploid and diploid *Fragaria* species reveals a high level of colinearity between their genomes and the essentially disomic behavior of the cultivated octoploid strawberry." Genetics **179**(4): 2045-2060.

Saeed, A. I., V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush and J. Quackenbush (2003). "TM4: a free, open-source system for microarray data management and analysis." Biotechniques **34**(2): 374-378.

Sargent, D. J., F. Fernandez-Fernandez, J. J. Ruiz-Roja, B. G. Sutherland, A. Passey, A. B. Whitehouse and D. W. Simpson (2009). "A genetic linkage map of the cultivated strawberry (*Fragaria* x *ananassa*) and its comparison to the diploid Fragaria reference map." Molecular Breeding **24**(3): 293-303.

Shulaev, V., D. J. Sargent, R. N. Crowhurst, T. C. Mockler, O. Folkerts, A. L. Delcher, P. Jaiswal, K. Mockaitis, A. Liston, S. P. Mane, P. Burns, T. M. Davis, J. P. Slovin, N. Bassil, R. P. Hellens, C. Evans, T. Harkins, C. Kodira, B. Desany, O. R. Crasta, R. V. Jensen, A. C. Allan, T. P. Michael, J. C. Setubal, J. M. Celton, D. J. Rees, K. P. Williams, S. H. Holt, J. J. Ruiz Rojas, M. Chatterjee, B. Liu, H. Silva, L. Meisel, A. Adato, S. A. Filichkin, M. Troggio, R. Viola, T. L. Ashman, H. Wang, P. Dharmawardhana, J. Elser, R. Raja, H. D. Priest, D. W. Bryant, Jr., S. E. Fox, S. A. Givan, L. J. Wilhelm, S. Naithani, A. Christoffels, D. Y. Salama, J. Carter, E. Lopez Girona, A. Zdepski, W. Wang, R. A. Kerstetter, W. Schwab, S. S. Korban, J. Davik, A. Monfort, B. Denoyes-Rothan, P. Arus, R. Mittler, B. Flinn, A. Aharoni, J. L. Bennetzen, S. L. Salzberg, A. W. Dickerman, R. Velasco, M. Borodovsky, R. E. Veilleux and K. M. Folta (2011). "The genome of woodland strawberry (*Fragaria vesca*)." Nat Genet **43**(2): 109-116.

Stam, P. (1993). "Construction of Integrated Genetic-Linkage Maps by Means of a New Computer Package - Joinmap." Plant Journal **3**(5): 739-744.

Staudt, G. (1962). "Taxonomic studies in the genus *Fragaria* typification of Fragaria species known at the time of Linnaeus." Canadian Journal of Botany **40**(6): 869-886.

Stewart, P. J. (2011). *Fragaria* history and breeding. Genetics, genomics and breeding of berries. K. M. Folta and C. Kole. Enfield, NH

Boca Raton, FL, Science Publishers ;

Marketed and distributed by CRC Press**:** 114-137.

Strange, R. N. and P. R. Scott (2005). "Plant disease: a threat to global food security." Annu Rev Phytopathol **43**: 83-116.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar (2013). "MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0." Molecular Biology and Evolution **30**(12): 2725-2729.

Tennessen, J. A., R. Govindarajulu, T. L. Ashman and A. Liston (2014). "Evolutionary origins and dynamics of octoploid strawberry subgenomes revealed by dense targeted capture linkage maps." Genome Biol Evol **6**(12): 3295-3313.

Thomson, M. J. (2014). "High-throughput SNP genotyping to accelerate crop improvement." Plant Breeding and Biotechnology **2**(3): 195-212.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.

Troggio, M., N. Surbanovski, L. Bianco, M. Moretto, L. Giongo, E. Banchi, R. Viola, F. F. Fernandez, F. Costa, R. Velasco, A. Cestaro and D. J. Sargent (2013). "Evaluation of SNP Data from the *Malus* Infinium Array Identifies Challenges for Genetic Analysis of Complex Genomes of Polyploid Origin." PLoS One **8**(6): e67407.

Ulrich, D. and K. Olbricht (2013). "Diversity of volatile patterns in sixteen *Fragaria vesca* L. accessions in comparison to cultivars of *Fragaria* x*ananassa*." Journal of Applied Botany and Food Quality **86**: 37-46.

van Dijk, T., G. Pagliarani, A. Pikunova, Y. Noordijk, H. Yilmaz-Temel, B. Meulenbroek, R. G. Visser and E. van de Weg (2014). "Genomic rearrangements and signatures of breeding in the allo-octoploid strawberry as revealed through an allele dose based SSR linkage map." BMC plant biology **14**(1): 55.

Van Ooijen, J. W. (2011). "Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species." Genetics Research **93**(5): 343-349.

Voorrips, R. E. (2002). "MapChart: software for the graphical presentation of linkage maps and QTLs." J Hered **93**(1): 77-78.

8. Annexes

Annex I. Number of probesets on FxaRefMap2 and associated by *F*. x *ananassa* homeolog group (HG) after clustering and proportion with respect to the corresponding linkage group (LG).

HG	Mapped	Associated	% of LG
LG1-A	88	27	24.3
LG1-B*	4	10	3
LG1-C	164	108	57.5
LG1-D	49	23	15.2
LG2-?1	21	2	5.4
LG2-A	74	26	23.5
LG2-B	111	78	44.5
LG2-C	83	30	26.6
LG3-?1	61	26	12.1
LG3-?2	70	48	16.4
LG3-?3	83	10	12.9
LG3-A	144	112	35.5
LG3-B1	69	23	12.8
LG3-B2	44	11	7.6
LG3-C2	17	3	2.8
LG4-?1	100	33	21.7
LG4-?2	40	20	9.8
LG4-B	168	78	40.2
LG4-D	124	49	28.3
LG5-?1	29	15	4.6
LG5-A*	54	15	7.2
LG5-B	167	43	22
LG5-C	265	166	45.1
LG5-D	137	64	21
LG6-?	50	26	7.5
LG6-A	100	21	12
LG6-B	36	37	7.2
LG6-C	222	108	32.7
LG6-C1	36	16	5.1
LG6-D	231	127	35.4
LG7-A	119	23	23.3
LG7-B	188	114	49.5
LG7-C	42	7	8
LG7-D	82	35	19.2