Sede Amministrativa: Università degli studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN: Bioscienze e Biotecnologie
INDIRIZZO: Biotecnologie

CICLO: XXVII

Identification of selective sweeps in domesticated apple (*Malus* × *domestica* Borkh.)

Direttore della Scuola: Ch.mo Prof. Giuseppe Zanotti
Coordinatore d'indirizzo: Ch.mo Prof. Fiorella Lo Schiavo
Supervisore: Ch.mo Prof. Gorgio Valle

Dottorando : Emanuela Kerschbamer

The apple is so deeply rooted in the culture of human populations from temperate regions that it is often not recognized as an exotic plant of unclear origin.

Cornille *et al.* 2012

Contents

1

Riassunto

Il melo domestico (*Malus* × *domestica*) è una delle piante più coltivate al mondo ed è tra le specie agricole geneticamente più polimorfiche. Studiare la diversità genetica in melo può dare importanti suggerimenti sul processo di domesticazione e valide risorse per creare mappe genetiche ad alta risoluzione, per analisi di QTL e nei programmi di breeding. I miglioramenti nelle tecnologie di sequenziamento del DNA, dette NGS, hanno ridotto di molto i costi del sequenziamento al punto che i risequenziamenti completi di genomi sono ora fattibili anche per specie ad alta diversità genetica e dal genoma molto grande.

Lo scopo di questo lavoro è l'analisi della variabilità genetica dell'intero genoma di melo e l'identificazione di regioni genomiche sottoposte a selezione durante il processo di domesticazione.

A tale scopo 63 cultivar di melo, rappresentanti l'intera diversità del germoplasma europeo, sono state sequenziate con teconolgia Illumina. Dalle sequenze sono stati predetti oltre 15 milioni di SNP che sono stati filtrati eliminare le predizioni scadenti o legate a regioni ripetute e paraloghe. Ulteriori filtri (minor allele frequency e Hardy-Weinberg equilibrium) sono stati applicati per eliminare gli SNP derivati da errori di genotipizzazione. Il numero finale degli SNP filtrati è risultato di 426'321.

Gli SNP rimasti dopo i filtri di qualità sono stati usati per studiare la struttura di popolazione e la diversità genetica. Dall'analisi delle componenti principali e da un metodo di clusterizzazione implementato in fastStructure, è emersa una debole stratificazione della popolazione analizzata. Questa analisi ha mostrato

la presenza di tre sottopopolazioni con un alto livello di admixture. L'F_{ST} tra ogni coppia di sottopopolazioni è risultato di 0,055, 0,083 and 0,096 indicando un livello di differenziazione moderato.

Due diversi approcci sono stati usati per identificare 'selective sweep'. Il primo è basato sulle frequenze alleliche e sul 'site frequency spectrum' (SFS) ed è implementato nel software SweeD. Il secondo è basato sui pattern di 'linkage disequilibrium' e la statistica ω ed è implementato nel software OmegaPlus. Le regioni del genoma che sono state identificate da entrambi i software sono state usate come regioni candidate sotto selezione positiva. In tutto il genoma le regioni sotto selezione sono risultate 1'194.

In totale 153 predizioni geniche sono state estratte dalle regioni candidate e annotate usando i termini della Gene Ontology e con i pathway metabolici descritti nel database KEGG. Ricerche di similarità in database di piante sono state fatte per trovare geni ortologhi e per capire meglio la funzione dei geni candidati. L'annotazione ha rivelato che i geni sotto selezione positiva sono coinvolti in vari processi quali la fotosintesi, l'ubiquitinazione di proteine, la trasduzione del segnale ormonale delle piante o il metobolismo di amidi e zuccheri. In particolare, per la trasduzione del segnale, sono stati identificati l'importatore dell'auxina e una proteina della famiglia SAUR che agiscono sull'aumento della dimensione cellulare e sulla crescita della pianta e la proteina 2 insensibile all'etilene che porta alla maturazione del frutto e alla senescenza. Le annotazioni funzionali disponibili ascrivono i geni identificati a ruoli fisiologici coerenti con i tratti fenotipici attesi per un processo di domesticazione. Per esempio i tratti legati al miglioramento delle caratterisitche del frutto come la dimensione, il gusto e la dolcezza.

2

Abstract

The domesticated apple (*Malus* × *domestica*) is one of the most cultivated plant over the world and is one of the most genetically polymorphic agricultural species. Studying the genetic diversity of the apple germplasm could provide important hints about the domestication process as giving a valuable resource for high resolution genetic mapping, QTL analysis and breeding programs. Advances in next generation sequencing technologies have driven the costs of DNA sequencing down to the point that whole genome re-sequencing (WGS) is now feasible for high diversity, large genome species. The aim of this work is to gain information on genome-wide genetic variability patterns in apple and to identify regions of the genome that may have been selected during the process of plant domestication.

SNPs were called from Illumina short reads for 63 apple cultivars representative of European germplasm diversity. The identified SNPs (over 15 millions) were filtered for quality and to avoid repeated and paralogous regions. Additional filters (minor allele frequency and Hardy-Weinberg equilibrium) were applied to discard variants derived from genotyping errors resulting in a final number of 426,321 SNPs . The SNPs kept after the quality filters were used to study the population structure and the genetic diversity. A weak stratification of the analyzed population emerged both from the principal component analysis (PCA) and a model based clustering approach performed using fastStructure. This analysis showed the presence of three subpopulations with a high level of admixture. F_{ST} between each couple of sub-groups was 0.055, 0.083 and 0.096 that indicate a moderate differentiation.

Two different approaches were used to identify selective sweeps. The first is based on allelic frequencies and the site frequency spectrum (SFS) and it is implemented in the software SweeD. The second is based on linkage disequilibrium patterns and the omega statistic and it is implemented in the software OmegaPlus. Regions that were identified by both softwares were merged and used as candidate regions for positive selection resulting in 1,194 sweeps on the whole genome.

A total of 153 gene predictions were extracted from these candidate regions and annotated using Gene Ontology terms and mapping on the KEGG pathway database. Similarity searches were also performed against plant databases to find gene orthologs and to better understand the function of candidates. The annotation revealed that genes under positive selection are involved in pathways like photosynthesis, protein ubiquitination, plant hormone signal transduction and starch and sucrose metabolism. In particular for the plant hormone signal transduction, were identified the auxin influx carrier and a SAUR family protein that lead to cell enlargement and plant growth and the ethylene insensitive protein 2 that leads to fruit ripening and senescence. The genes identified in regions under positive selection that were functionally annotated are consistent with the domestication traits for a better fruit: bigger, tastier and sweeter.

3

Introduction

3.1 Apple

The domesticated apple (*Malus* × *domestica* Borkh.) is the most common and culturally important fruit crop of temperate areas. World apple production in the year 2012-2013 was estimated around 76 million tons. China, U.S.A., Turkey, Poland, Italy, and France are the leading producers (FAOSTAT, 2014) and about 65% of the Italian production is concentrated in Trentino - Alto Adige (`www.trentinoagricoltura.it/Trentino-Agricoltura/Settori2/Ortofrutticoltura/Mela`). The uses of this crop are mainly connected to the fruits that are consumed fresh, dried or processed into juice, preserves, cider. Several other related species or hybrids are of use as ornamental trees, rootstocks or for the production of cider.

The fruit production is currently dominated by strains of just few cultivars because at the start of commercial apple production, only a few cultivars seemed to meet the criteria for large scale production [1]. In the in newly established orchards less than ten polyclonal varieties dominate unchallenged. These cultivars belong essentially to the polyclonal groups 'Fuji', 'Gala', 'Golden Delicious', 'Red Delicious' and 'Jonagold', which offer a broad range of highly colored mutants and even some spurs such as those derived from 'Red Delicious' (e.g. 'Red Chief®') that have displaced the original varieties. They are followed by other, smaller groups of cultivars with heterogeneous traits whose importance is increasing, including 'Braeburn', 'Pink Lady®' and 'Pinova'; on the other hand

the traditional 'Elstar', 'Granny Smith', 'Cox's Orange Pippin', 'Renettes' and 'Idared', and their mutants, are gradually losing industry favour [2].

The apple tree reaches a height of three to 12 metres and it is deciduous. When propagated by seeds following standard conditions the apple goes through a long juvenile period of several years (four to nine years) during which it cannot flower. The flower is epigynous, the ovary being enclosed by non-ovarian tissue (fused base of sepals, petals and stamens or cortex of stem, depending on morphology espoused) that remains attached to the ovary at harvest and forms a false fruit or pome. Apple flowers are generally produced in five-flowered inflorescences, known as corymbs. The single flower is pentamerous, bisexual and radially symmetrical. A normal fruit can contains up to ten seeds [1]. The majority of the apple cultivars require cross-pollination and parthenocarpic cultivars are generally of little economic value. Obligatory cross-pollination is a consequence of a strong self-incompatibility based on an S-RNase based gametophytic system [3]. Cultivars differ widely in time of ripening, varying from less than 60 days after full bloom to more than 180. Several factor can affect the maturation time, e.g. rootstock, heavy cropping or low spring and summer temperatures. Apple is a climacteric fruit, this means that a burst of carbon dioxide and ethylene production occurs in the fruits as they ripen. The majority of domesticated apple cultivars is diploid with haploid chromosome number x = 17, but there are some triploids (e.g. 'Reinette du Canada') and tetraploids [4]. A draft genome sequence for domesticated apple is available: the genome sequence for the cultivar Golden Delicious was published in 2010 [5]. The genome size is around 750 Mbp (bp = base pairs) and 66,000 contigs have been mapped on 17 linkage groups in the last update, available on the Genome Database for Rosaceae (http://www.rosaceae.org/). More than 40% of the genome is represented by transposable elements. To date, apple is the the species with the highest number of predicted gene (about 57,000). The sequenced cultivar is highly heterozygous and showed an average of 4 SNPs every 1000 bp [5]. The SNPs identified in 'Golden Delicious' will be fundamental for developing high-throughput genotyping tools for molecular genetics and breeding applications of the near future.

The cultivated apple (*M.* × *domestica*) belongs to the genus *Malus*, and is usually placed in the subtribe Pyrinae, tribe Pyreae, supertribe Pyrodae, fam-

ily Rosaceae. The circumscription of these taxonomic groups lacks agreement among systematics, depending on whether classification schemes emphasize morphological traits, chromosome number intergeneric crossability or molecular polymorphisms. The tribe Pyreae, which includes the genus *Malus* along with other important genera such as *Pyrus*, *Sorbus*, *Cydonia*, *Mespylus* and others, has a high haploid base chromosome number of x = 17 and is generally considered to be monophyletic [6]. A number of models have been proposed to explain the uniquely high number of chromosomes in Pyreae, the most popular being the wide-hybridization hypothesis based on an allopolyploidization event between spireoid (x = 9) and amygdaloid (x = 8) ancestors ([7], [8]). More recent molecular phylogeny studies point to the possibility that Pyreae originated by autopolyploidization or by hybridization between two sister taxa with x = 9 (similar to actual Gillenia), followed by diploidization and aneuploidization to x = 17 [9]. This last hypothesis takes into account that Gillenia and related taxa are New World species and that the earliest fossil evidence of specimens belonging to extant genera of Pyreae are from North America. Furthermore, the autopolyploidization hypothesis was also supported by the observation that the apple genome derives from a relatively recent genomic duplication [5] and the timing of such genome wide duplication, as estimated from the genomic data, agrees with archaeobotanical dates of 48 50 Mya [10]. In addition molecular phylogeny of *Wx* genes in the apple genome confirms the close relationship of Gillenia (x = 9) with the Pyreae (x = 17) lineage. The rate of chromosome rearrangements after polyploidization (12 chromosome events in 60 My) is similar to that for poplar (16 events in 60 My) [11] and lower than in maize (at least 17 chromosome fusion events in 5 My) [12] or in artificial neopolyploids [13]. The genus *Malus* has, according to most authorities, 25 to 30 species and several subspecies of so-called crabapples, many of which are cultivated as ornamental trees for their profuse blossom and attractive fruits. Most the species intercross and, since self incompatibility is common, trees produced from seeds obtained from a botanical garden or arboretum where collection of *Malus* are grown are almost always interspecific or intercultivar hybrids. It is therefore very difficult to be certain of the authenticity of specific names [14]. The 34 primary species of the genus *Malus*

are divided in five section (Malus, Sorbomalus, Eriolobus, Chloromeles and Docyniopsis). The section Malus is further divided in three subsections, Pumilae, Sieboldanae and Kansuenses. The subsection Pumilae contains the domesticated apple and the more closely related s ecies, and it is divided in two series. The first is the series Pumilae and contains: *M. asiatica*, *M.* × *domestica*, *M. micromalus*, *M. orientalis*, *M. prunifolia*, *M. pumila*, *M. sieversii*, *M. spectabilis* and *M. sylvestris*. The second is the series Baccatae and contains: *M. baccata*, *M. floribunda*, *M. halliana*, *M. hupensis*, *M. mandshurica* and *M.sikkimensis*.

3.2 Plant domestication

The technology needed for the cultivation of fruit trees is very different from grain agriculture, as for example fruit plants are perennial and attain full productivity several years after plantation. Horticulture therefore indicates a fully sedentary way of life, not required from annual crops, that allow shifting cultivation. Horticulture started relatively late in the history of food production of the Old World and like happened for grains crops, the first fruit trees seem to begin being cultivated in the Near East core area [15]. The first definite evidence of fruit-tree cultivation appeared in the 4 millennium B.C. in the Chalcolithic contexts. Genetically, domestication of fruit trees means the shifting from sexual reproduction (in the wild) to vegetative propagation (under cultivation). Normally cultivated varieties of fruit trees are maintained vegetatively by cuttings, rooting of twigs, suckers or grafting. Growers, by discarding sexual reproduction and inventing clonal reproduction, were able to select individuals with desirable traits and to duplicate the chosen types to obtain genetically identical saplings. The archaeological excavation indicate two different waves of domestication of fruit trees. The first comprises all the fruit trees multipliable simply by cuttings (in grapevine, fig, pomegranate and sycamore fig), rooting of basal knobs (in olive) or by transplanting offshoots (in date palm). These early domesticated trees were preadapted since they lent themselves easily to vegetative manipulation. Several other fruit trees, such as apple, pear, plum, and cherry were involved in the second wave of domestication, that taken place much later because their culture is based almost entirely on grafting [16]. When and where detached scion

grafting was invented is not yet clear. Very probably the initiation of grafting was outside the area of Mediterranean horticulture and this technology was introduced into this region from east. The earliest documentation of grafting comes from China in connection with citrus cultivation [17]. The adoption of clonal cultivation imply that most fruit trees, in the last five or six millennia since the introduction into cultivation, have undergone very few sexual cycles. Due to the reduced number of generations under domestication is expected that the cultivars have not diverged considerably from their wild relatives. The cultivated varieties of fruit trees can be regarded as exceptional individuals of their species, that excel primarily in fruit size and quality. The absence of profound genetic change in the fruit trees under domestication is also apparent in their ecology. Also the climatic requirements of the cultivars closely resemble those of their progenitors. In hermaphroditic, self-incompatible species such as olive, apple, or pear, the early planters very likely realized that to obtain satisfactory fruit set it is necessary to plant together at least two synchronously flowering clones in order to bring about pollinations between different genotypes.

3.2.1 Apple domestication

The cultivated varieties of apple are most closely related to a variable group of wild and feral crab apples (series *Pumila* in section *Malus* in the genus *Malus*) which are widely distributed in the temperate areas of Asia and Europe. They show a marked differentiation into ecogeographical races which have been ranked as independent species. The European crab apples are usually referred as *M. sylvestris* (L.) Mill. whereas the form growing in Anatolia, north Iran, and Caucasus as reported as a different species *M. orientalis* Uglitzkich. Further east *M. sylvestris* is replaced by *M. sieversii* (Lodeb.) that is split in several additional local species (e.g. *M. kirghisorum*) by some Russian botanists. Furthermore, *M. sieversii* is the only wild species that produces fruit comparable in size and taste to the cultivate apple, *M.* $\times$ *domestica*. Variation in apples is further complicated by the fact that cultivation is practiced in areas supporting wild Malus populations. In recent times, variation patterns in apples have been additionally complicated by crossing European apples with wild and cultivated form from

other geographical origins [15]. Small apples collected from wild and in cut into halves for parching were found in Neolithic and Bronze Age site all over Europe. Small dried apples were found also in a tomb in the Royal Cemetery at Ur, lower Mesopotamia, dated late 3rd millennium B.C. [18], and in site dated tenth century B.C. in the kadesh Barnea oasis, on the border between Negev and Sinai. In spite of such findings, it is generally accepted that apple did not evolve into the a major fruit crop until the classical times, when grafting was introduced. Very little is known about the time and place of apple domestication except that in classical times apples were already extensively grown in Old World. It is therefore futile to try to delimit the area of initial domestication on the basis of evidence available from the living plants. Apples could have been brought into cultivation anywhere in the temperate areas of Europe and western and central Asia. Exceptional Malus individuals may have been picked up not once and in a single place, but many times and in several areas. Furthermore, many cultivars are hybridization product combining genes from several distinct geographical sources (Zohary2000). From a molecular point of view the origin of the domesticated apple, *M.* × *domestica*, has been widely studied and debated. Borkhausen, first describing *M.* × *domestica*in 1803, believed it originated as a hybrid derived from *M. sylvestris*, *M.pumila*, *M.sylvestris* var. praecox [19]. Several authors suggested that *M. sieversii* is the main ancestor of *M.* × *domestica*([14], [20], [21], [22]), based on morphological and molecular data. Wild *M. sieversii* occurs from the Tien Shan mountain range on the border between China and Kazakhstan, Kyrgyzstan and Tajikistan, to the edge of the Caspian Sea [23]. As *M. sieversii* is most diverse in Kazakhstan, this region is considered to be the centre of origin of cultivated apple [14]. On the other hand, Coart *et al.* [24], based on chloroplast DNA information, demonstrate that *M. sylvestris* is probably more closely related to *M.* × *domestica*than was commonly thought. Recent molecular investigations confirmed *M. sieversii*, the wild Central Asian species as the main progenitor with the contribution of other species from Asia to Western Europe along the Silk Route (e.g. *Malus baccata* (L.) Borkh. in Siberia, *Malus orientalis* Uglitz. in the Caucasus, and *Malus sylvestris* Mill. in Europe) [25]. In particular, a significant genetic contribution to *M.* × *domestica*comes from

M. sylvestris. Microsatellite markers studies demonstrated also that dessert apple cultivars are actually genetically closer to *M. sylvestris* than cider cultivars which usually produce smaller fruits that are more bitter than those of dessert cultivars [26].

3.3 Genetic diversity

Genetic diversity describes the differences between genomes of different individuals. It is studied for many different purposes :

- to estimate the level of variation in a population;
- to examine and understand the patterns in which several kinds of polymorphisms are present in the genome;
- to understand the evolutionary mechanisms by which it is maintained;
- to monitor the biodiversity of key indicator species in an threatened environment;
- to uniquely identify individuals in a population;
- to use polymorphisms as genetic markers for pedigree analyses and in breeding programs;
- to identify wild ancestors, understand the origin of actual breeds and the practices of artificial selection;
- to study the phylogeny of different species.

The level of genetic diversity in a population can be estimated using different tests and parameters:

- The nucleotide diversity (π) represents the number of nucleotide differences per site in two sequences in the population
- The observed heterozygosity (H_O) is frequency of heterozygous individuals averaged by the number of sampled loci

- The fixation index (F_{ST}) is a measure of the probability of identity by descent

In population genetics the term population is not used to indicate an entire species but a group of individuals of the same species that live in the same restricted geographical area and can mate with any other member of the opposite sex in the group. A population is said to be stratified or structured when there is a systematic difference in allele frequencies between different subgroups.

Considering the domesticated crops the genetic diversity is the basic ingredient for breeding. Without genetic variability, breeding efforts to develop new, distinct and improved varieties are futile. Therefore, exploring and describing new genetic diversity in old or obsolete varieties, landraces or closely related species, is of the utmost importance to achieve continuous progress in crop breeding. An in-depth knowledge about the genetic relationships among accessions can be helpful to decide the genotypes to cross, to predict the outcome of these crosses and to identify heterotic groups (groups of genotypes that produce high yielding progeny after crossing). Clearly, the genetic diversity present in breeding germplasm can be better exploited if the genetic structure and genetic relationships are better understood.

3.3.1 Single Nucleotide Polymorphism

Nowadays, the most used type of polymorphism is the single nucleotide polymorphism (SNP). They are informative and codominant respect to old-generation types of genetic markers like amplified fragment length polymorphisms (AFLPs) and random amplified polymorphic DNA (RAPDs). A number of different methods based on different principles have been developed to discover and genotype the SNPs [27], [28]. The SNP is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of a species or paired chromosomes in an individual. SNPs can be divided in two different classes: transitions in which the nucleotide is replaced by a nucleotide of the same class (C $\rightarrow$ T and A $\rightarrow$ G) and transversions in which the nucleotide is replaced by a nucleotide of a different class (A $\rightarrow$ C, C $\rightarrow$ G, T $\rightarrow$ A and T $\rightarrow$ G). SNPs are the most frequent polymorphism in the genome, and, for example, in the human

genome 2.8 million of SNPs were identified from the comparison of the reference genome and an individual genome [29], in *Arabidopsis thaliana* at least 500,000 SNPs were identified [30] and more than four million were identified in the *Vitis vinifera* cv. Pinot Noir genome [31]. In population genetics and more in general in genetics and molecular biology, in recent years, SNPs have largely substituted other kind of marker because of their abundance in the genome, the possibility to screen in single assay up to millions of SNPs and the high level of automation. Obviously, there is a difference between the technique useful for the discovery and the genotyping. The direct sequencing has become the favoured approach for the SNPs discovery after the development of the next generation sequencing techniques that brought to a rapid decrease of the sequencing costs [32], [33]. Several different techniques, with different throughput, have been developed for SNP genotyping. The choice of the technique depends on different factors, mainly the number of individuals and SNPs to analyse. The advent of next-generation sequencing (NGS) has revolutionized genomic and transcriptomic approaches to biology. These new sequencing tools are also valuable for the discovery, validation and assessment of genetic markers in populations. Genotyping-by-sequencing (GBS) is now feasible for high diversity, large genome species. This approach uses restriction enzymes to construct libraries reducing the complexity of the genome especially for species with a high genetic diversity that complicates the mapping step or species for which a reference genome is not yet available [34]. Whole genome sequencing (WGS) is becoming more common with the reduction of sequencing costs [35]. This method allows to analyse the whole genome not just a part of it. In this way also low frequency - large effect variants are detected [36].

3.3.2 Linkage disequilibrium

Polymorphic loci are considered in linkage equilibrium without selection, mutations or migration. They are in linkage disequilibrium if there is a non-random association of alleles at different loci located on the same chromosome. To estimate the LD several statistics were developed [37] and, in plant, the most widely used are D' [38] and r^2 [39].

Considering a pair of loci with the allele 'A' and 'a' at the first locus, and 'B' and 'b' at the second locus, with allelic frequencies of π_A, π_a, π_B and π_b and with haplotypic frequencies of π_{AB}, π_Ab, π_{aB} e π_{ab}, the difference between observed and expected haplotypic frequencies is $D_{ab} = (\pi_{AB} - \pi_A\pi_B)$. D_{ab} is the fundamental parameter for each statistical methods to estimate the LD. r^2, also described in the literature as Δ^2, is calculated as:

$$r^2 = \frac{(D_{ab})^2}{\pi_A\pi_a\pi_B\pi_b}$$

It is convenient to consider r^2 as the square of the correlation coefficient between the two loci [40]. However, unless the two loci have identical allele frequencies, a value of 1 is not possible. Statistical significance (P-value) for LD is usually calculated using either Fisher's exact test [41], to compare sites with two alleles at each locus, or multifactorial permutation analysis, to compare sites with more than two alleles at either or both loci. D', is calculated as:

$$|D'| = \frac{(D^{ab})^2}{min(\pi_A\pi_b, \pi_a\pi_B)} for D_{ab} < 0$$

$$|D'| = \frac{(D^{ab})^2}{max(\pi_A\pi_B, \pi_a\pi_b)} for D_{ab} > 0$$

D' is scaled based on the observed allele frequencies, so it will range between 0 and 1 even if allele frequencies differ between the loci. D'will only be less than 1 if all four possible haplotypes are observed; hence, a presumed recombination event has occurred between the two loci.

D' and r^2 reflect different aspects of LD and perform differently under various conditions [42]. Although neither r^2 nor D' perform extremely well with small sample sizes and/or low allele frequencies, each has distinct advantages. Whereas r^2 summarizes both recombinational and mutational history, D measures only recombination history and is therefore the more accurate statistic for estimating recombination differences. However, D' is strongly affected by small sample sizes, resulting in highly erratic behavior when comparing loci with low allele frequencies. This is due to the decreased probability of finding all four allelic combinations of low frequency polymorphisms even if the loci are unlinked.

For the purpose of examining the resolution of association studies, r^2 is generally favorable, as it is indicative of how markers might correlate with the QTL of interest. The presence of LD is a prerequisite for association mapping where the LD extent or the physical size of LD blocks, that is chromosomal regions across which all pairs of adjacent loci are in LD [43], determines the marker density required for association mapping. In crop plants, the potential of exploiting LD to detect marker-trait associations was investigated in many crop plants (for example in maize [44], [45] and wheat [46], [47], [48]). The published data suggest that association mapping is a valuable additional tool in the search for the detection of novel genes or QTLs for important agronomic traits. The extensive application of this approach in crop plants is to be expected in the midterm as a result of establishing of the novel high-throughput genotyping and sequencing technologies [49], [50]. In selection studies the extent of LD is determinant to define the resolution at which different statistics are calculated over the genome. Moreover LD levels may vary across genomes due to differences in recombination rates, selective pressures, mating systems and effective population sizes. This leaves a pattern in the genome that can be exploited to identify signatures of positive selection.

3.4 Selection

Population genetics has as a main goal, to determine if the genetic variation in a population is neutral (caused by random genetic drift) or due to selection. It is interesting to study selection to understand the evolutionary past and the basic mechanisms that control molecular evolution. In particular, great interest is put in the study of positive selection, in respect to negative or purifying selection, because it is linked to the arising and evolution of a new function and that provides important information [51]. The identification of genomic regions under positive selection can help to understand the processes that lead to phenotypic differences between and within species [52] .

3.4.1 Selective sweeps

Positive selection acting on a beneficial allele, increases its frequency in the population over time and leaves a pattern in the genome, eliminating the neutral diversity in the flanking regions by genetic hitchhiking (Figure 3.1). The result of this process is called selective sweep [53]. The genomic region subjected to a selective sweep can be recognized by the reduction of genetic diversity and by the pattern of LD and of distribution of alleles at independent sites. Also the haplotype pattern can be useful in the identification of sweeps (Figure 3.2). The analysis of large genomic and SNP datasets should allow the detection of the molecular signatures of new advantageous mutations that have been selected and recently become fixed in the population (frequency = 1). The site frequency spectrum (SFS) is the distribution of the allele frequencies at segregating sites. For each mutation the SFS counts the occurrences of the allele and the sample size (total number of chromosomes, number of individuals multiplied by two in case of diploidy). The SFS is called 'folded' when it is not known which is the ancestral and which is the derived allele. Otherwise in the 'unfolded' case, the occurrences of the derived allele are counted. A selective sweep strongly affects the SFS, leading to an increase of alleles of intermediate frequency while the sweep is acting and to a reduction of these alleles when the sweep is completed. Purifying selection (against deleterious mutations) leaves the same pattern as an ongoing sweep. On the contrary, positive selection will increase the frequency of mutations segregating at high frequencies.

One of the first sweep studies in plants was the investigation of the *Waxy* genomic region in rice (*Oriza sativa*) [54]. This region was chosen because previous studies have shown that a splice donor mutation in intron 1 of the *Waxy* gene is associated with the absence of amylose-characterizing glutinous rice varieties. In [54] paper a 500 Kbp region was sequenced in several rice accessions: *O. nivara*, *O. barthii* and *O. meridionalis* were included as outgroups; *O. sativa* samples included representatives of the five major sub-populations: 21 *indica*, 18 *tropical japonica*, 21 *temperate japonica*, six *aus*, and six *aromatic* landraces from diverse geographical locations in Asia. The authors found a 97% reduction in molecular variation at the *Waxy* gene among individuals that carry the splice

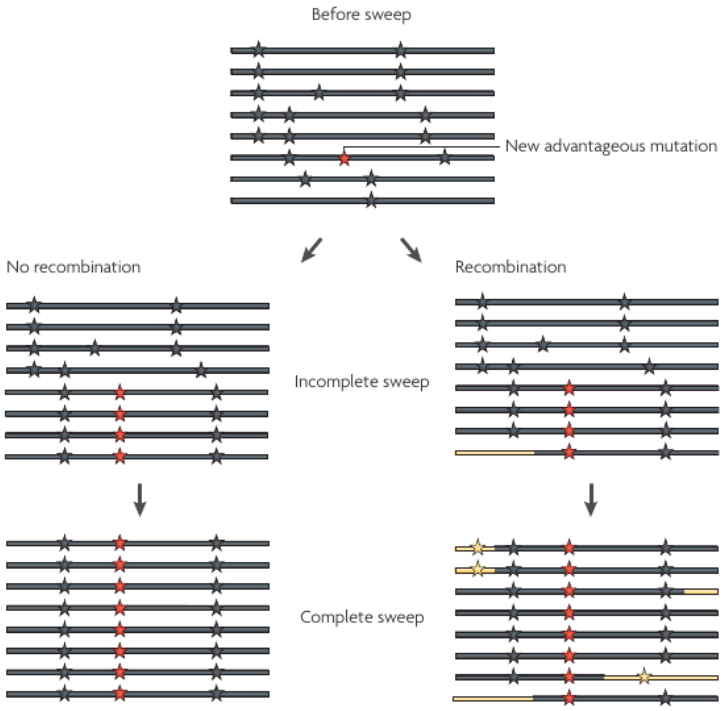


Figure 3.1: Selective sweep. The lines indicate individual DNA sequences or haplotypes, and derived SNP alleles are depicted as stars. A new advantageous mutation (indicated by a red star) appears initially on one haplotype. In the absence of recombination, all neutral SNP alleles on the chromosome in which the advantageous mutation first occurs will also reach a frequency of 100% as the advantageous mutation become fixed in the population. Likewise, SNP-alleles that do not occur on this chromosome will be lost, so that all variability has been eliminated in the region in which the selective sweep occurred. However, new haplotypes can emerge through recombination, allowing some of the neutral mutations that are linked to the advantageous mutation to segregate after a completed selective sweep. As the rate of recombination depends on the physical distance among sites, the effect of a selective sweep on variation in the genomic regions around it diminishes with distance from the site that is under selection. Chromosomal segments that are linked to advantageous mutations through recombination during the selective sweep are coloured yellow. Data that are sampled during the selective sweep at a time point when the new mutation has not yet reached a frequency of 100% represent an incomplete selective sweep. [52]

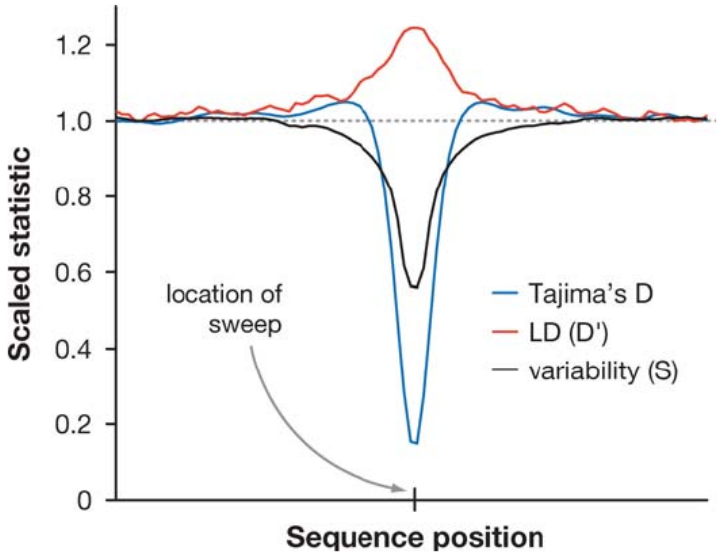


Figure 3.2: The effect of a selective sweep on genetic variation. The figure is based on averaging over 100 simulations of a strong selective sweep. It illustrates how the number of variable sites (variability) is reduced, LD is increased, and the frequency spectrum, as measured by Tajimas D, is skewed, in the region around the selective sweep. All statistics are calculated in a sliding window along the sequence right after the advantageous allele has reached frequency 1 in the population. All statistics are also scaled so that the expected value under neutrality equals one. [51]

donor mutation. This pattern is consistent with positive directional selection for the splice donor mutation. This is an example of *a posteriori* investigation of selective sweep in a candidate region. Whole genome studies of selection patterns in plants are very recent, this is due to the availability of high density SNP data.

Several methods have been developed to detect the described changes in the SFS, usually these methods are neutrality tests. In Tajima's D test the average number of nucleotide differences between pairs of sequences is compared with the total number of segregating sites. If the difference between these two measures of variability is larger than what is expected on the standard neutral model, this model is rejected [55]. When D has positive values it indicates a lack of mutations of intermediate frequency relative to derived mutations that segregate at low or high frequencies, suggesting a sweep has recently completed in the region; when D has negative values it indicates an excess of these mutations with intermediate frequency. Fu and Li extended Tajima's D test to take into account the information on newly arisen alleles at segregating sites by using an evolutionary outgroup [56]. In the presence of purifying or negative selection there will tend to be excess of new mutations because deleterious alleles are present in low frequencies. This will happen also if an advantageous allele has recently become fixed in the population, because then the majority of the mutations in the population are expected to be young. On the contrary, if balancing selection is acting at the locus, then some alleles may be old. Fay and Wus H test is the adaptation of Tajima's D for cases in which very few polymorphisms are present that gives more weight to information from high-frequency derived mutations, to detect unequivocally positive selection [57]. The method of Kim and Stephan uses the pattern of the SFS to identify the location of a selective sweep. It is based on an explicit population genetic model of a selective sweep and in this way they could calculate the expected frequency spectrum in a site as a function of its distance to an advantageous mutation. By fitting the data to the model, they could estimate the location of the selective sweep and the strength of the selective sweep, and perform hypothesis tests regarding the presence of a sweep. This method is particularly useful in that it takes advantage of the spatial pattern left by the sweep along the sequence [58].

Recent positive selection can cause increased levels of population subdivision at certain locations in the genome. When a locus shows extraordinary levels of genetic population differentiation, compared with other loci, this may then be interpreted as evidence for positive selection [59], [60]. Several methods have been proposed for detecting selection based on this idea, for example, that of Akey *et al.* [61], which identifies areas of increased F_{ST}.

An influential approach for detecting recent and strong natural selection is the extended haplotype test [62] and its derivatives [63]. The extended haplotype test relies on the linkage-disequilibrium structure of local regions of the genome. Regions containing a polymorphism under balancing selection will tend to reduce LD if the polymorphism is old, but may increase LD in a transient phase. Selective sweeps also increase levels of LD in a transient phase, although this phase may be relatively short. A haplotype at high frequency with high homozygosity that extends over large regions is a sign of an incomplete selective sweep. The method identifies tracts of homozygosity within a 'core' haplotype, using the 'extended haplotype homozygosity' (EHH) as a statistic. A relative EHH (rEHH) is calculated by comparing the EHH of the core haplotype to the EHH of all other haplotypes in the region. In the version by Voight *et al.* [63], the EHH is summed over all sites away from a core SNP, and compared between the haplotypes that carry the ancestral and the derived allele in the SNP. The statistic (iHS – integrated haplotype score) is then normalized to have a mean of 0 and variance of 1. A related test was proposed by Wang *et al.* [64] , called the linkage disequilibrium decay (LDD) test, which makes use of only homozygous SNP sites and therefore does not require separate phasing of haplotypes. The method of Kim and Stephan [58] was extended by Kim and Nielsen [65] to include pairs of sites to incorporate information regarding linkage disequilibrium.

4

Aims of the project

The study of crop domestication is directed to discover the origin of the wild progenitors of modern crops. In this process the nature of the genetic changes that happen during domestication is investigated, as well as the consequences on the genetic diversity of the domesticated species. An understanding of the domestication process provides insights into the general mechanisms of adaptation and the history of human civilization, but can also guide modern breeding programs aiming to improve crops species [26]. The results presented in this thesis should therefore contribute to better understand the modifications occurred in the genome of the domesticated apple after the domestication process.

More in detail the principal goals of the project were:

- the identification of genomic regions under positive selection that may have been artificially selected during the process of plant domestication in *M.* × *domestica*. Such regions may reveal agronomically interesting traits that could be exploited in marker assisted selection and breeding.
- the whole-genome description of the molecular genetic diversity in *M.* × *domestica*.
- the study of the population structure in *M.* × *domestica*.

All the objectives of the thesis were achieved exploiting the whole-genome resequencing data of a wide collection of apple cultivars representative of the European germplasm collection.

5

Materials and methods

5.1 Samples and data

Sixty-three *M.* × *domestica* cultivars were used in this study (Table 6.1) and two double haploid (DH) accessions, 'X9273' and 'X9748', which were derived from 'Golden Delicious' were included to help identify pseudo-SNPs created from the erroneous assembly of paralogous regions of the apple genome. Thirty-nine of these are from central Europe, four from northern Europe, nine from Russia, five from the U.S., one each from Tunisia, Iran, Canada, Japan, New Zealand and Australia. The cultivated samples were chosen to maximize the genetic diversity in the apple germplasm and they include some of the core European apple breeding founder varieties. These accessions were selected within the frame of the European project FruitBreedomics (www.fruitbreedomics.com) to be resequenced for SNP discovery to develop the 480K Axiom Affymetrix SNP array. This array was produced with the aim of genotyping apple populations for genome-wide association studies. Illumina read data from the resequencing of these 63 accessions were made available for this PhD project.

Table 5.1: Historical origin of the cultivars

Cultivar name	**Country**
Abbondanza	Italy
Ag alma	Russia
Aivaniya	Bulgaria
	Continued on next page

Table 5.1 – continued from previous page

Cultivar Name	Country
Ajmi	Tunisia
Åkerö	Sweden
Alfred Jolibois	France
Amadou	France
Antonovka	Russia
Antonovka Pamtorutka	Russia
Aport Kuba	Russia
Belle et Bonne	Belgium
Borowitsky	Russia
Braeburn	New Zealand
Budimka	Serbia
Busiard	Italy
Cabarette	Belgium
Chodské	Czech Republic
Court-Pendu Henry	Belgium
Cox's Orange Pippin	United Kingdom
De L'Estre	France
Delicious	U.S.
Doctor Oldenburg	Germany
Durello di Forlì	Italy
F2-26829-2-2	U.S.
Filippa	Denemark
Fuji	Japan
Fyriki	Greece
Gelata	Italy/Spain
Godelieve Hegmans	Belgium
Golden Delicious	U.S.
Heta	Sweden
Hetilina	Czech Republic
Ijunskoe ranee	Russia
Jantarnoe	Russia
Jonathan	U.S.
Keswick Codlin	United Kingdom
Košíkové	Czech Republic
Kronprins	Sweden
Lady Williams	Australia
Macoun	U.S.
Maikki	Finland

Continued on next page

Table 5.1 – continued from previous page

Cultivar Name	Country
Malinové holovouské	Czech Republic
McIntosh	Canada
Mela Rosa (PD)	Italy
Mela Rozza	Italy
Ovčí hubička	Russia
Panenské české	Czech Republic
Papirovka	Russia
Patte de Loup	France
Pepino Jaune	France
Precoce de Karaj	Iran
Président Roulin	Belgium
Priscilla-NL	U.S.
Reinette Clochard	France
Reinette Dubois	Belgium
Renetta Grigia di Torriana	Italy
Rosa (FI)	Italy
Skry (Skryzhapel)	Russia
Sonderskow	Czech Republic
Spässerud	Sweden
Worcester Pearmain	United Kingdom
Young America	U.S.

5.2 Alignment and SNP calling

Illumina 100 bp paired-end reads were checked for quality using FastQC (`www.bioinformatics.babraham.ac.uk/projects/fastqc`) and subsequently trimmed and filtered for quality with FastQ clipper from the FastXToolkit (`hannonlab.cshl.edu/fastx_toolkit/index.html`) using a threshold of 26 for minimum quality phred score and a minimum read length after trimming of 80 bp.

BFAST is an algorithm specifically developed to deal with short reads (25-100 bp) from next generation sequencing that supports paired end reads [66]. It exploits whole genome indexes to rapidly mapping reads using a Smith-Waterman method [67]. It was used to map the reads as single ends on the reference genome (cultivar 'Golden Delicious') [5]. BFAST produces as output BAM files that are the binary form of SAM (Sequence Alignment/Map) files [68] (see Appendix B

for details). Aligned reads were filtered following various steps. PCR duplicates were removed with samtools command rmdup [68]. Reads with more than seven mismatches and multiply aligning reads were discarded. A "proper pairing" check implemented in a custom C++ script was applied in order to check if the reads of the pair were on the same chromosome and within a minimum of zero and maximum of 800 bp of distance from each other. A read group tag was added to the header of the BAM files to identify for which sample each read came from before merging all the alignments in a single BAM file. This was done because the SNP calling was performed in pool considering all samples together. Both steps were performed with samtools.

SNPs were called using the command mpileup from samtools and the view command form bcftools (`github.com/samtools/bcftools`) that output a VCF file [69] (see Appendix B for details).

From the called SNPs, those that were heterozygous in the double haploids (DH) were removed. The DHs are *M.* × *domestica* accessions derived from 'Golden Delicious' and are expected to be homozygous at all loci. Any heterozygous calls in these genotypes were those considered evidence of paralogous sequences [70]. Before further filtering a correction was applied to the data, to overcome the fact that the reference genome sequence contains Ns and IUPAC codes that lead to wrong genotype calls. This step simply corrects the genotype call to avoid erroneous calls due to the presence of nucleotide ambiguity codes in the reference genome.

5.3 SNP filtering

SNPs were filtered for quality to retain true SNPs and discard those derived from sequencing errors. Custom scripts written in Python were developed to remove variants with low phred-scaled quality scores (i.e. below 20); a high combined read depth (i.e. higher than 4,000); and a low single-cultivar read depth (i.e. lower than 8) in more than 50% of the cultivars. A minimum phred-scaled quality score of 20 was chosen to ensure only SNPs were selected with a probability of less than 1% for the alternative allele being called wrongly, and the maximum read depth value was used to ensure the removal of SNPs derived from paralogous

regions rather than true heterozygous regions. INDELs and triallelic SNPs were discarded because most population genetics softwares do not deal with them.

Moreover, all SNPs with an allele frequency equal to 1 were discarded since this equated to all the re-sequenced varieties carrying an allele that was different from the 'Golden Delicious' reference genome. Such SNPs were discarded as they were likely to represent potential false SNPs resulting from sequencing errors in the 'Golden Delicious' reference sequence.

Quality filtered variants from the pipeline described above were then processed to meet Affymetrix specific requirements. This step removed variants that were either A/T or C/G transversions, and SNPs in regions where the count of 16-mers was bigger than 300. These filters are required for the development of the SNP array as the probes can account for only two alleles and their hybridization site must be unique.

To define a set of SNPs to be used for population genetic analysis, additional three filters were applied. SNPs were discarded if less than 90% of genotypes were called, the minimum allele frequency (MAF) was lower than 0.1 and the Hardy-Weinberg Fisher's test p-value lower than $1E-7$. The Hardy-Weinberg equilibrium defines the constant frequency of genotypes along generations in a population that respects the main assumptions of: no mutation, no drift, no selection, random mating, large population size. This additional filtering step was applied to discard SNPs with frequencies that deviate too much from the expected ones and could represent genotyping errors. These filters were applied using PLINK version 1.9. Plink is a tool set for whole genome association analysis and it is designed to perform a range of basic, large-scale analyses in a computationally efficient manner [71].

5.4 Population structure

Two methods were used to study the structure of the population.

Principal component analysis. A principal component analysis (PCA) was performed using the R package prcomp (`stat.ethz.ch/R-manual/R-devel/`

`library/stats/html/prcomp.html`). This analysis finds potential patterns in the data by reducing the variables to new ones that explain most of the variability. Data are converted to highlight their similarities and differences, retaining the variance-covariance structure among the samples. The input file was prepared using a custom a python script that starting from a vcf file creates a matrix in which each genotype is codes as 0 (homozygous to the reference), 1 (heterozygous) or 2 (alternative homozygous).

Model-based approach. The software fastStructure [72] was used to infer population structure. This software uses a model-based approach to assign individuals to a predefined number of subgroups (K) based on their genotypes. Ten independent runs were performed with fastStructure for each K from 1 to 6. The most probable number of subpopulations was estimated by finding the maximum value of marginal likelihood of the data [72].

The repeatability of the performed runs for each K was tested using the software CLUMPP (CLUster Matching and Permutation Program) [73]. The membership coefficients of each individual resulting from each run were permuted with using the greedy algorithm with 1000 repeats. In case multiple solutions are reached, the permutation allows to distinguish the cases of label switching and genuine multimodality and leads to a more truthful interpretation of the results.

5.5 Population genetic analyses

Several statistics were computed to study the genetic variability present in the sample collection considering both whole genome and sliding windows 50,000 bp large with 5,000 bp of step.

Nucleotide diversity. The nucleotide diversity (π) is a measure of genetic variability and represents calculates the number of nucleotide differences per site in two sequences in the population [74]. It was calculated using vcftools (`http://vcftools.sourceforge.net`).

Observed heterozygosity. It was calculated, using a python script having as input the vcf file, as follow:

$$H_o = \frac{\sum_{i=1}^{n} f(Aa)}{n}$$

where n is the number of SNPs in the analyzed window and $f(Aa)$ is the heterozygotes frequency at the *i-th* locus.

Tajima's D. Tajima's D is a statistical test that measures the probability of the sequence to evolve neutrally against the probability of selection by using the difference between the number of segregating sites and the average number of nucleotide differences [55]. This statistic was measured using vcftools.

Fixation index. The fixation index (F_{ST}) is a measure of the probability of identity by descent and it is derived from the Wright's F statistic, that measures the reduction of heterozygosity in a population (Ho) in comparison with Hardy-Weinberg equilibrium heterozygosity (H) [75]. It was measured as follow:

$$F_{ST} = \frac{(H_o - H)}{H_o}$$

The F_{ST} is measured between the different subgroups or subpopulations using the vcftools.

Linkage disequilibrium. The variation of the LD with the distance was analyzed for each pair of identified SNPs. Since the phases between alleles at two heterozygous loci are unknown, for the pairwise comparisons between SNPs at two loci in the same linkage group the composite LD coefficients (Δ_{AB}) reported to have good statistical properties and suggested for routine testing of LD [76]. The software Plink v.1.9 [71] was used to calculate the r^2. A python script was used to parse the plink output and to plot the decay of the r^2 against increasing distance for each pair of SNPs. The LD decay helped to choose the width of the windows at which the selection statistics were later calculated. The average value of r^2 was also calculated for each window in which the genome was divided using a custom python script.

5.6 Identification of regions under selection

Two approaches were used to investigate regions under selection in the genome: one based on the site frequency spectrum (SFS), and the second based on LD.

The first approach computes the composite likelihood ratio (CLR) test and is implemented in the software SweeD (Sweep Detector) [77]. This software calculates the SFS, that is the distribution of derived allele frequencies in the samples, directly from the vcf input file. SweeD is an extension of the SweepFinder algorithm [78] for large scale genomic data. In the context of a selective sweep, the SFS of sites, in proximity of a beneficial mutation, shifts toward high or low frequency derived alleles [57]. If the derived allele can not be distinguished from the ancestral allele, SweeD uses the folded version of the SFS. SweeD was run separately for each chromosome with the vcf as input file and grid as only parameter. The grid parameter was calculated for each chromosome in order to have a measure of the CLR every 5,000 bp (size of the chromosome/5000).

The second approach is based on the recognition of a pattern of linkage disequilibrium which is expected to increase in the regions flanking the selected site, but not across it [51]. The software OmegaPlus is a high performance implementation of the ω statistic [65], that is based on the calculation of the LD between polymorphic sites and gives a measure of the above described pattern. OmegaPlus was run separately for each chromosome with the vcf as input file and grid as option. The grid size was calculated at the same way as for SweeD, to have a measure of ω every 5000 bp. The minimum and the maximum size of the flanking region were fixed to 200 bp (minwin option) and 200,000 bp (maxwin option), respectively.

An R script made available from the authors of SweeD and OmegaPlus (`http://pop-gen.eu/wordpress/selective-sweep-analysis-pipelines/combine-sweed-analysis-with-omegaplus-analysis-to-detect-common-outliers`) was used to select common outliers from both the CLR and the ω statistics. The outliers were 5% of most extreme p-values. Providing an annotation as gff (General Feature Format) file to the script, it finds the genes that are present in region corresponding to the outliers sweeps.

5.7 Candidate genes annotation

Candidate genes for positive selection were annotated with Gene Ontology terms (`http://geneontology.org/`) and InterPro domains (`http://www.ebi.ac.uk/interpro/`) already available from the reference genome sequencing project [5] and accessible at the Genome Database for Rosaceae (`http://www.rosaceae.org/`). A gene set enrichment analysis was performed on GO terms associated to candidate genes, using Bioconductor package topGO (Alexa A. and Rahnenfuhrer J. (2010) topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0). Within topGO a Fisher's exact test was performed on gene counts. The GO term structure was then plotted as a graph. The significant GO terms are drawn as rectangles and the colors represent the relative significance, ranging from dark red (most significant) to light yellow (least significant).

A fasta file containing the candidate genes peptidic sequences was submitted to BlastKOALA (KEGG Orthology And Links Annotation) at KEGG (`http://www.kegg.jp/`). BlastKOALA is a tool that allows the identification of pathways in which the candidate genes are involved. It performs a BLAST of the query against a non-redundant set of KEGG genes. Than it links candidate genes to their KEGG Orthology (KO) identifiers that are categorized under the hierarchy of KEGG pathways. The pathways identified through BlastKOALA help to pinpoint the biological functions in which the candidate genes play a role.

A reciprocal best hit blast was performed to identify candidate genes orthologs in five other species: *Arabidopsis thaliana*, tomato, peach, strawberry and pear. Blastp was performed with standard parameters and the selection of the best hits (e-value $\geq 1E-5$, a minimum of 30% of identity and a minimum of 0.7 of ratio between the length of the query and the length of the hit [79]) was performed with a custom python script.

6

Results

6.1 Alignment, SNP calling and filtering

A total of 5.99 billions of paired-end reads were kept after the removal of poor quality reads (minimum phred score per nucleotide 26) and reads shorter than 80 nt. The reads were divided between the cultivars as shown in table 6.1. The average number of reads for each cultivar was 95.2 millions that correspond to a mean sequencing depth of 25 X. The maximum number of paired reads, 210.4 millions equivalent to a depth of 56 X, belonged to the cultivar Budimka. The minimum, 38.0 millions equivalent to a depth of 10 X, belongs to Maikki. The cleaned reads of each accession were aligned on the reference genome using BFAST [66] and each alignment file were used to perform a SNP calling using samtools [68]. SNPs were called independently for each cultivar and than they were pooled. An average of 3.7 millions variants (SNPs and INDELS) were found in the 63 sequenced accession. The highest, 4.5 millions, and the lowest, 1.6 millions, were found in 'Young America' and 'Golden Delicious', respectively. The minimum number of SNPs was observed in 'Golden Delicious'because the reference genome is 'Golden Delicious'itself and the only SNPs are the heterozygous ones. After the pooling, and the removal of the INDELs, a total of 15,499,525 SNPs were identified. The number of SNPs was reduced of 0.2% after the removing the SNPs that were heterozygous in the double haploids and are likely derived from paralogous regions. Quality filters (SNP quality $>$ 19, read depth $<$ 4001 and

allele frequency < 1 furtherly decreased the number of SNPs to 12,701,549. Additionally, filters specific for the Affymetrix array technology discarded 9,901,376 SNPs. This filters discarded all the INDELs, the not bi-allelic SNPs, the SNPs that present additional SNPs or Ns in the 35 nt flanking the target SNP and the SNPs in which the count of the 16-mer up/down stream the target SNP is higher than 300. Finally the filter for minor allele frequency lower than 0.05 and for Hardy-Weinberg equilibrium distortion (p-value $< 1e-7$) discarded 2,373,852 SNPs. At the end of the filtering process 426,321 SNPs were kept for the further analyses. In Figure 6.1 was reported the filtering steps summary.

Table 6.1: Number of aligned paired end reads for each sample before SNP calling and number of SNPs/INDELs called.

Cultivar name	**Read pairs**	**Number of SNPs/INDELs**
Abbondanza	90,321,460	3,640,158
Ag Alma	127,334,651	4,073,393
Aivaniya	119,632,134	3,856,162
Ajmi	82,950,154	4,063,822
Åkerö	91,666,140	3,664,625
Alfred Jolibois	92,947,232	3,737,886
Amadou	96,043,319	3,899,419
Annurca	86,930,085	3,735,538
Antonovka	122,361,758	4,063,547
Antonovka Pamtorutka	39,392,157	3,303,074
Aport Kuba	71,297,068	3,559,776
Belle et Bonne	113,928,709	3,665,450
Borowitsky	96,122,310	4,117,169
Braeburn	113,833,008	3,601,221
Budimka	210,405,926	3,917,964
Busiard	84,486,599	3,893,658
Cabarette	83,997,287	3,480,301
Chodské	62,687,249	3,558,746
Court-Pendu Henry	99,053,865	3,595,621
Cox's Orange Pippin	92,500,881	3,569,992
De L'Estre	110,615,195	3,456,123

Continued on next page

Table 6.1 – continued from previous page

Cultivar Name	Read pairs	Number of SNPs/INDELs
Delicious	130,861,887	3,441,507
Doctor Oldenburg	125,373,896	4,007,530
Durello di Forlì	46,088,993	3,302,188
F2-26829-2-2	103,188,147	4,173,045
Filippa	90,247,516	3,723,740
Fuji	109,365,440	3,393,911
Fyriki	81,179,122	3,851,739
Gelata	83,186,975	3,856,933
Godelieve Hegmans	98,149,222	3,776,468
Golden Delicious	80,333,013	1,617,646
Heta	44,608,537	3,342,398
Hetlina	88,568,362	3,746,579
Ijunskoe ranee	123,440,174	4,158,540
Jantarnoe	90,682,815	4,047,164
Jonathan	128,267,363	3,415,168
Keswick Codlin	69,130,834	3,466,712
Košíkové	86,098,710	3,688,331
Kronprins	62,159,897	3,497,089
Lady Williams	108,963,567	3,617,078
Macoun	88,835,758	3,608,001
Maikki	38,025,283	3,314,175
Malinové holovouské	81,050,254	3,542,212
McIntosh	126,063,436	3,787,320
Mela Rosa (PD)	95,875,984	3,916,077
Mela Rozza	79,694,154	3,905,327
Ovčí hubička	114,799,463	3,876,049
Panensk české	76,470,501	3,750,462
Papirovka	90,957,520	3,926,332
Patte de Loup	110,531,953	3,627,864
Pepino Jaune	144,185,284	3,749,257
Precoce de Karaj	110,480,537	4,219,837
Président Roulin	106,767,106	3,595,298
Priscilla-NL	117,058,356	3,668,963

Continued on next page

Table 6.1 – continued from previous page

Cultivar Name	Read pairs	Number of SNPs/INDELs
Reinette Clochard	111,511,186	3,645,266
Reinette Dubois	102,336,676	3,518,378
Renetta Grigia di Torriana	64,416,556	3,410,542
Rosa (FI)	70,443,859	3,613,993
Skry (Skryzhapel)	78,536,012	4,008,244
Sonderskow	74,020,116	3,701,811
Spässerud	84,123,712	3,814,082
Worcester Pearmain	96,410,779	3,699,145
Young America	96,202,193	4,514,568

6.2 Population structure

The population structure analyses supported the hypothesis of a unique population in domesticated apple. The principal component analyses does not show a sharp separation in subgroups but indicate the presence of three partially separeted groups. Although the first principal component explained only 8.67% of the variability and the second 3.96% (Figure 6.2).

To have more precise picture of the stratification present in the domesticated apple a model-based clustering analysis was carried out with fastStructure [72]. The number of assumed sub-populations (K) values were ranging from 2 to 6 and 10 independent runs were performed. At K = 4, three solutions resulted from the permutation of the fastStructure runs made using CLUMPP [73]. The most represented solution occurs in the 80% of the runs while each of the remaining two represents the 10%. At all others K the solution was unique. The most probable number of sub-population was identified at K=3 by finding the maximum value of marginal likelihood of the data. Only the accessions with a membership coefficient ≥ 0.70 were assigned to a sub-population while the remaining individuals are cosidered as admixed.

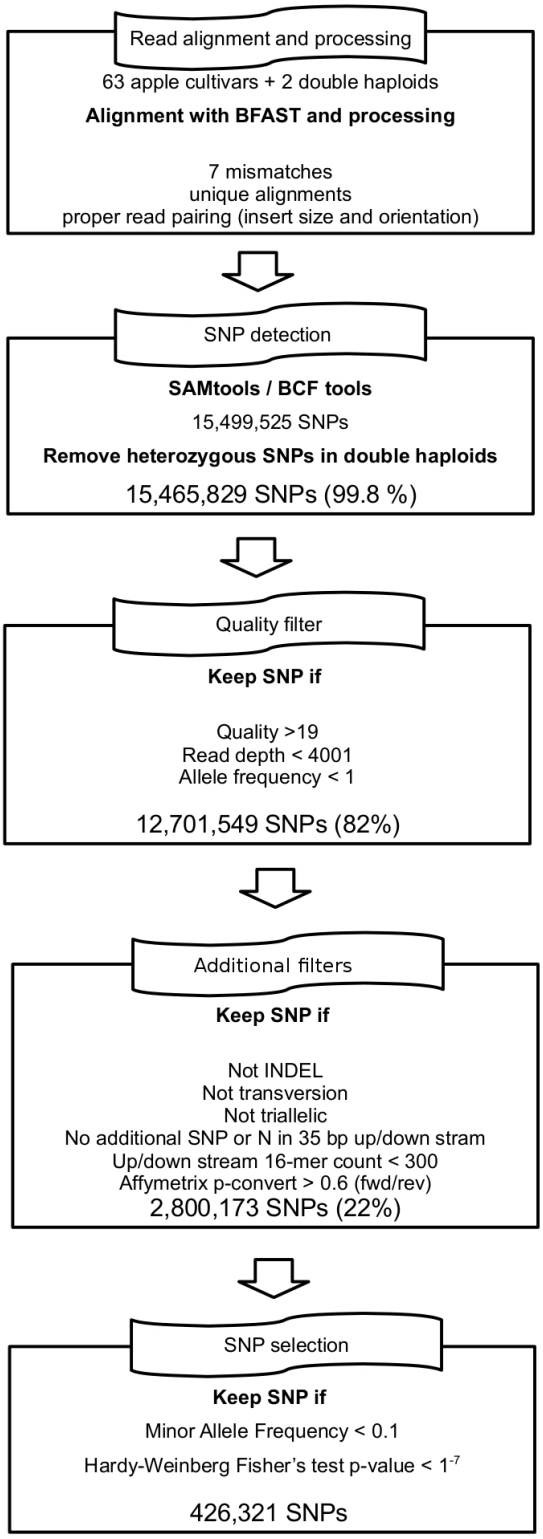


Figure 6.1: Flowchart describing the SNP filtering process.

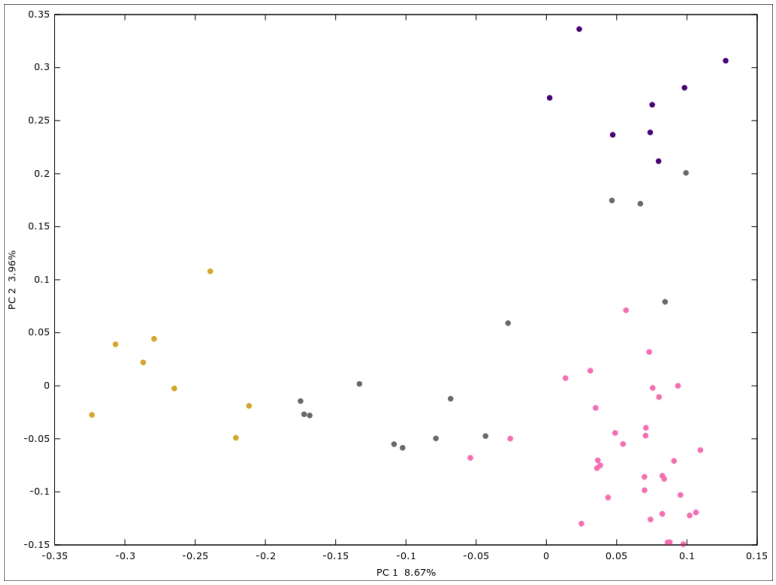


Figure 6.2: Principal component analysis of the cultivars showing first and second principal components. Group 1 in purple, group 2 in yellow, group 3 in pink, admixed individuals in grey.

At K=3 group 1 is composed by eight cultivars: 'Antonovka Pamtorutka', 'Borowitsky', 'Papirovka', 'Spässerud', 'Skry (Skryzhapel)', 'Antonovka', 'Maikki' and 'Åkerö'. Group 2 is composed by eight cultivars: 'Rosa (FI)', 'Fyriki', 'Ijunskoe ranee', 'Aivaniya', 'Ovčí hubička', 'Precoce de Karaj', 'Ajimi' and 'Ag Alma'. Group 3 contains 33 cultivars: 'Doctor Oldenburg', 'Panenské české', 'Kronprins', 'Président Roulin', 'Young America', 'Golden Delicious', 'Court-Pendu Henry', 'Belle et Bonne', 'Reinette Dubois', 'Pepino Jaune', 'Malinov holovouské', 'Chodské', 'Patte de Loup', 'De L'Estre', 'Cabarette', 'Abbondanza', 'Braeburn', 'Cox's Orange Pippin', 'Delicious', 'Fuji', 'Jonathan', 'Lady Williams', 'McIntosh', 'Priscilla-NL', 'Worchester Pearmain', 'F2-26829-2-2', 'Reinette Clochard', 'Alfred Jolibois', 'Keswick Codlin', 'Filippa', 'Košíkové', 'Renetta Grigia di Torriana' and 'Macoun'. The remaining 14 individuals could not be assigned uniquely to one group and were considered as admixed.'Godelieve Hegmans', 'Jantarnoe', 'Heta' and 'Sonderskow' were shared between group 1 and group 3, while 'Aport Kuba', 'Hetlina', 'Mela Rozza', 'Annurca', 'Gelata', 'Busiard', 'Amadou', 'Durello di Forlì', 'Budimka' and 'Mela Rosa (PD)' were shared between group 2 and group 3 (Figure 6.3).

To estimate the genetic differentiation between the subgroups the F_{ST} was calculated as average on the whole genome between each pair of groups excluding the the admixed individuals. The mean F_{ST} was 0.055, 0.083 and 0.096 between group 1 and 2, group 2 and 3, and group 1 and 3, respectively.

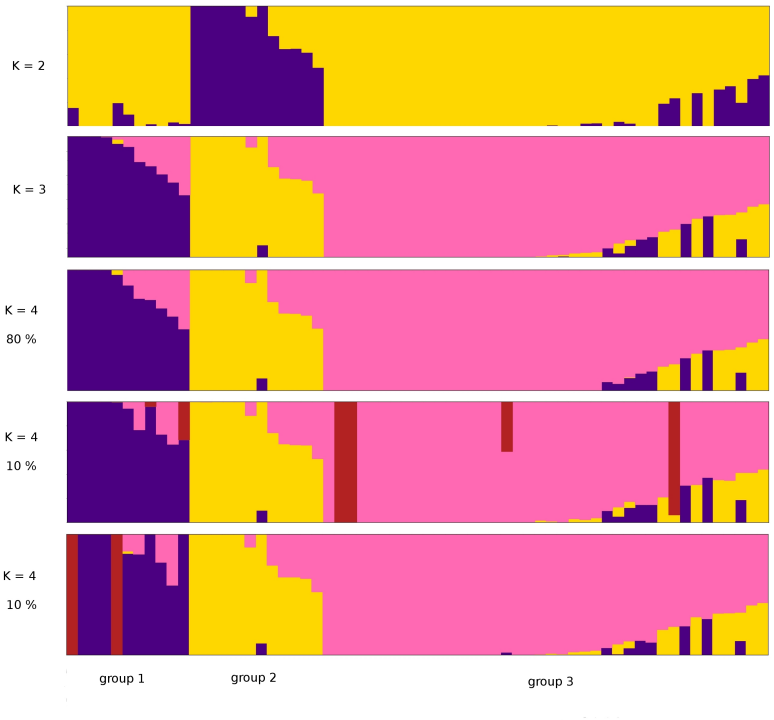


Figure 6.3: Genetic structure of the cultivars from K=2 to K=4. Each individual is represented by a vertical bar, partitioned into K segments representing the amount of assignment of its genome in K clusters identified by different colors. When several clustering solutions were represented within replicate runs, the proportion of simulations represented by each mode is given.

6.3 Genetic variability

The study of the genetic variability was performed through the calculation of the nucleotide diversity, the observed heterozygosity, the Tajimas D and the F_{ST} in sliding windows 50,000 bp wide with 5000 bp step over all the genome.

The average nucleotide diversity (π) was 0.0003 with maximum 0.002683 on chromosome 13 and minimum 0.0000015 on chromosomes 1, 2, 5, 11, 13 and 16.

The average observed heterozygosity in the genome was 0.1939 with minimum 0 in every chromosome and maximum 0.7143 in chromosome 3.

The average Tajima's D was 1.2924 with maximum 5.17005 on chromosome 1 and minimum -1.39945 on chromosome 15. A negative Tajimas D means that in the considered region, there are more alleles at low frequencies than expected, meaning there is a reduction in variability that could be related to a selective sweep or to a recent population expansion after a bottleneck.

The F_{ST} was calculated in sliding windows as average between the F_{ST} measures for each couple of subgroups. The F_{ST} mean in all the genome was 0.0786 with minimum 0 on all chromosomes and maximum 0.4345 on chromosome 10.

The mean linkage disequilibrium measured as r^2 was 0.2322 on the whole genome, the maximum r^2 was 1 in almost all chromosomes and the minimum was 0 on chromosome 15 (Table 6.2).

The genome wide genetic variability is summarized in Figure 6.3. From the outer to the inner circle nucleotide diversity, observed heterozygosity, Tajima's D and mean F_{ST} are displayed in sliding windows for each chromosome. The decay of the linkage disequilibrium was studied on the whole phased dataset. The r^2 was plotted with the increasing distance between each pair of SNPs on the same chromosome. The correlation coefficient (r^2) decays below the value of 0.2 at SNP distance of about 20 Kbp. The mean r^2 value for 332,773,644 pairwise SNP comparisons within one Mbp was 0.107 (Figure 6.4).

6.4 Selective sweep analysis

The number of genomic regions identified as candidate for positive selection by SweeD alone was 2526, the highest number (210) identified in chromosome 15, the lowest (124) in chromosomes 8 and 16. OmegaPlus alone identified 2354 regions, the highest number (205) identified in chromosome 15, the lowest (97) in chromosome 16. A consensus was build and a total of 1194 sweeps was identified in the whole genome (Table 6.3) by both softwares SweeD and OmegaPlus. Figures 6.5 show the outlier genes identified by the softwares independently and as a consensus. The highest number of sweeps was 106, identified in chromosome 10, the lowest was 41, identified in chromosome 8. As SweeD and OmegaPlus

Table 6.2: Genetic diversity statistics summarized for each chromosome and the whole genome.

Chromosome	Nuc. diversity	Obs. Heterozygosity		Tajima's D			F_{ST}		LD
	Mean	Mean	Max.	Mean	Max.	Min.	Mean	Max.	Mean
1	0.0003	0.1958	0.6437	1.2528	5.1701	-0.7108	0.0779	0.3176	0.2551
2	0.0002	0.1860	0.6190	1.2138	4.3682	-0.5112	0.0556	0.2926	0.2171
3	0.0003	0.1879	0.7143	1.2289	4.1983	-0.6304	0.0643	0.2816	0.2170
4	0.0003	0.1901	0.6032	1.3277	4.5310	-0.7513	0.1266	0.3434	0.2361
5	0.0003	0.1964	0.5555	1.4505	4.4780	-0.6304	0.0865	0.3215	0.2234
6	0.0003	0.1870	0.5714	1.1461	4.1103	-0.6062	0.0704	0.3155	0.2488
7	0.0004	0.2104	0.5714	1.5487	4.6070	-0.6876	0.0715	0.2566	0.2398
8	0.0002	0.1907	0.5397	1.2068	4.3059	-0.4629	0.0778	0.3661	0.2306
9	0.0003	0.2026	0.5397	1.4224	4.8106	-0.5834	0.0902	0.3317	0.2395
10	0.0004	0.1933	0.4906	1.3515	4.2858	-1.1376	0.1054	0.4345	0.2184
11	0.0003	0.1939	0.5238	1.2801	4.0948	-0.4719	0.0567	0.2988	0.2068
12	0.0003	0.1933	0.6191	1.2805	4.4121	-0.6513	0.0896	0.3222	0.2299
13	0.0003	0.1996	0.5556	1.4101	4.1791	-0.7448	0.0731	0.2659	0.2533
14	0.0003	0.1961	0.6349	1.2438	4.4024	-0.6090	0.0613	0.3198	0.2229
15	0.0002	0.1873	0.6058	1.1131	4.5479	-1.3995	0.0703	0.3143	0.2631
16	0.0003	0.2064	0.6349	1.3868	4.2597	-0.6935	0.0881	0.3581	0.2415
17	0.0003	0.1856	0.6190	1.1847	4.2829	-0.5676	0.0884	0.3387	0.2060
Genome mean	0.0003	0.1939	0.7143	1.2924	5.1701	-1.3995	0.0786	0.4345	0.2322

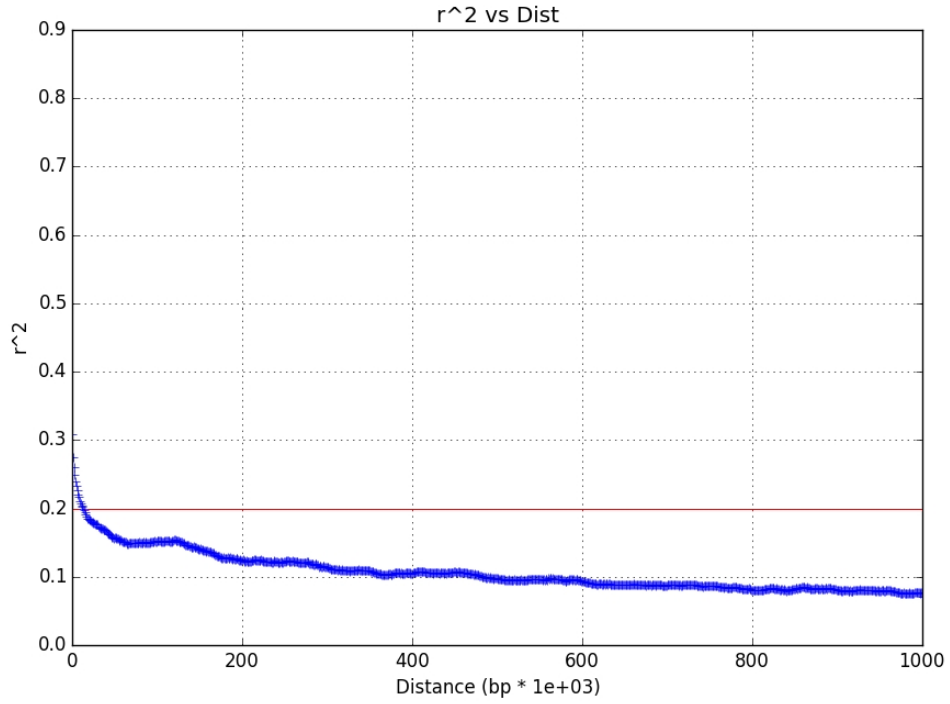


Figure 6.4: Decay of the LD coefficient (r^2) with physical distance. Only pairs of SNPs within 1 Mbp distance were plotted. Each point represents the average r^2 value in 1000 bp bins.

Table 6.3: Summary of selective sweep analysis. For each chromosome the total number of sweeps detected, the number of sweeps per Mbp and the count of genes in sweeps are reported.

Chromosome	# of Sweeps	Sweeps/Mbp	Genes in Sweeps
1	56	1.86	5
2	74	1.69	9
3	69	1.77	12
4	56	1.90	9
5	81	2.06	11
6	68	2.19	5
7	74	2.08	8
8	41	1.16	7
9	77	2.02	9
10	106	2.58	6
11	87	1.98	2
12	65	1.67	11
13	65	1.55	13
14	54	1.59	11
15	76	1.34	12
16	79	2.82	15
17	66	1.83	8
Whole genome	1194	1.86	153
Chr. average	70.2	1.89	9

calculate the CLR and the *omega* statistic in windows, sweeps that were in two consecutive windows were merged in a single region. The minimum length of a sweep was arbitrarily set as the length of the window. The longest sweep region was 274855 bp long, on chromosome 12. The mean length of the sweep was 20864 bp. A total of 153 gene predictions were identified in sweep regions. The highest number of gene predictions (15) was identified in chromosome 16 and the lowest (2) in chromosome 11. On average, 9 gene predictions per chromosome were identified as being under selection (Table 6.3).

6.5 Annotation

The GO enrichment analysis was performed for the ontologies Biological Process and Molecular Function. For the Biological Process ontology, the three most significant terms were transcription from RNA polymerase II promoter (GO:0006366), photosynthesis (GO:0015979) and protein ubiquitination (GO:0016567). In Molecular Function, DNA-directed RNA polymerase activity (GO:0003899), transcription coactivator activity (GO:0003713) and ubiquitin-protein ligase activity (GO:0004842) were the most significant terms. In Table 6.4 and Table 6.5 are listed the 20 most significant terms for both ontologies. They are represented in the ontology tree in Figure 6.6 and Figure 6.7. Thirtyfive of the 153 selected gene predictions were identified in pathways through blastKOALA. Fourteen gene predictions were identified in metabolic pathways (Figure 6.8, eight in the biosynthesis of secondary metabolites (Figure 6.9), four in plant hormone signal transduction (Figure 6.10) and three in purine metabolism (Figure 6.11).

The orthology search through the reciprocal best hit blast revealed that 110 of the 153 candidate gene predictions had an orthologous in at least one of the species investigated: tomato, Arabidopsis thaliana, strawberry, pear and peach. In Figure 6.12 are summarized the number of shared orthologous between apple and one or more other species. Twentyfour gene predictions under selection had orthologous genes in all the five species searched. Seventeen apple gene predictions found orthologs in four of the species analysed: two were shared between tomato, Arabidopsis, strawberry and pear; one was shared between tomato, Arabidopsis, strawberry and peach; two were shared between tomato, Arabidopsis, pear and peach; seven between tomato, strawberry, pear and peach; five between Arabidopsis, strawberry, pear and peach.

Fifteen apple gene predictions were found in three species: one in tomato, Arabidopsis and strawberry; three in tomato, Arabidopsis and pear; one in tomato, Arabidopsis and peach; four in tomato, strawberry and pear; one in tomato, strawberry and peach; two in Arabidopsis, strawberry and peach; one in Arabidopsis, peach and pear; two in strawberry, pear and peach. Twentyfour apple gene predictions were found in two species: two in tomato and strawberry; four in tomato and pear; one in tomato and peach; one in Arabidopsis and strawberry;

Table 6.4: Summary table with the significance of Biological Process GO terms according to different tests.

GO.ID	Term	Annotated	Significant	Expected	Rank in classic	Classic	KS	Weight
GO:0006366	Transcription from RNA polymerase II promoter	32	2	0.01	1	3.1e-05	0.3616	3.1e-05
GO:0015979	Photosynthesis	114	1	0.03	10	0.03038	0.9945	30
GO:0016567	Protein ubiquitination	137	1	0.04	11	0.03641	0.0270	36
GO:0043687	Post-translational protein modification	147	1	0.04	14	0.03902	0.3173	39
GO:0051246	Regulation of protein metabolic process	177	1	0.05	15	0.04682	0.1733	47
GO:0006520	Cellular amino acid metabolic process	877	1	0.24	24	0.21408	0.5670	214
GO:0006508	Proteolysis	1720	1	0.46	38	0.38155	0.0270	382
GO:0006468	Protein phosphorylation	2905	1	0.78	50	0.56478	0.6397	565
GO:0006082	Organic acid metabolic process	1117	1	0.30	31	0.26530	0.7703	1000
GO:0006139	Nucleobase-containing compound metabolic process	6619	2	1.79	51	0.56815	0.6004	1000
GO:0006351	Transcription, DNA-templated	120	2	0.03	2	0.00044	0.4075	1000
GO:0006464	Cellular protein modification process	3609	2	0.97	26	0.25336	0.4685	1000
GO:0006725	Cellular aromatic compound metabolic process	6961	2	1.88	55	0.59994	0.4286	1000
GO:0006793	Phosphorus metabolic process	3757	1	1.01	59	0.66576	0.7784	1000
GO:0006796	Phosphate-containing compound metabolic process	3753	1	1.01	58	0.66534	0.7843	1000
GO:0006807	Nitrogen compound metabolic process	7596	3	2.05	36	0.33702	0.3497	1000
GO:0008150	Biological process	25928	7	7.00	62	1.00000	1.0000	1000
GO:0008152	Metabolic process	19815	7	5.35	21	0.15222	0.0712	1000
GO:0009058	Biosynthetic process	5264	2	1.42	47	0.43160	0.8710	1000
GO:0009059	Macromolecule biosynthetic process	3829	2	1.03	34	0.27693	0.8836	1000

Table 6.5: Summary table with the significance of Molecular Function GO terms according to different tests.

GO.ID	Term	Annotated	Significant	Expected	Rank in classic	Classic	KS	Weight
GO:0003899	DNA-directed RNA polymerase activity	134	2	0.03	1	0.0003	0.4630	0.0003
GO:0003713	Transcription coactivator activity	12	1	0.00	3	0.0024	0.6335	0.0024
GO:0004842	Ubiquitin-protein ligase activity	144	1	0.03	7	0.0284	0.0116	0.0284
GO:0004197	Cysteine-type endopeptidase activity	190	1	0.04	8	0.0373	0.3595	0.0373
GO:0003677	DNA binding	6249	3	1.25	16	0.1123	0.6979	0.1123
GO:0005509	Calcium ion binding	682	1	0.14	17	0.1283	0.0176	0.1283
GO:0004674	Protein serine/threonine kinase activity	2457	1	0.49	24	0.3982	0.5595	0.3982
GO:0004713	Protein tyrosine kinase activity	2467	1	0.49	25	0.3995	0.3748	0.3995
GO:0016491	Oxidoreductase activity	3465	1	0.69	29	0.5168	0.3658	0.5168
GO:0005524	ATP binding	6550	1	1.31	36	0.7643	0.6911	0.7643
GO:0005488	Binding	26243	5	5.23	35	0.7510	0.2941	0.7925
GO:0000166	Nucleotide binding	8051	1	1.61	50	0.8385	0.6122	1.0000
GO:0000988	Protein binding transcription factor act...	41	1	0.01	6	0.0081	0.9040	1.0000
GO:0000989	Transcription factor binding transcripti...	27	1	0.01	4	0.0054	0.8515	1.0000
GO:0001882	Nucleoside binding	6989	1	1.39	40	0.7886	0.6870	1.0000
GO:0001883	Purine nucleoside binding	6989	1	1.39	41	0.7886	0.6870	1.0000
GO:0003674	Molecular function	35113	7	7.00	53	1.0000	1.0000	1.0000
GO:0003676	Nucleic acid binding	10599	3	2.11	23	0.3571	0.6505	1.0000
GO:0003712	Transcription cofactor activity	27	1	0.01	5	0.0054	0.8515	1.0000
GO:0003824	Catalytic activity	18031	6	3.59	13	0.0718	0.8287	1.0000

four in Arabidopsis and pear; four in Arabidopsis and peach; four in strawberry and pear; one in strawberry and peach; three in pear and peach.

Thirty apple gene predictions were found in only one species: five in tomato, five in Arabidopsis, eight in strawberry, eleven in pear and one in peach. The details with gene predictions IDs are shown in Table (Table 6.6).

Table 6.6: Summary of apple candidate genes for positive selection and the identified orthologs in different species.

Apple	Tomato	Arabidopsis	Strawberry	Pear	Peach
MDP0000282147	Solyc06g071850.2.1	AT5G63200.1	gene16780-v1.0-hybrid	PCP003944.1	ppa001586m
MDP0000136041	Solyc12g014540.1.1	AT5G58610.1	gene29090-v1.0-hybrid	PCP000681.1	ppa027200m
MDP0000260203	Solyc01g094970.2.1	AT4G23500.1	gene06153-v1.0-hybrid	PCP022521.1	ppa025007m
MDP0000256881	Solyc10g079860.1.1	AT4G16260.1	gene05145-v1.0-hybrid	PCP007636.1	ppa024457m
MDP0000763939	Solyc11g072210.1.1	AT5G56200.1	gene07740-v1.0-hybrid	PCP044777.1	ppa020870m
MDP0000134878	Solyc05g010780.1.1	AT4G11690.1	gene19555-v1.0-hybrid	PCP045044.1	ppa020757m
MDP0000178257	Solyc02g067100.2.1	AT3G01015.1	gene04312-v1.0-hybrid	PCP030482.1	ppa018071m
MDP0000607972	Solyc08g006180.2.1	AT5G39250.1	gene09531-v1.0-hybrid	PCP011700.1	ppa017095m
MDP0000297452	Solyc05g054670.2.1	AT3G11310.1	gene28775-v1.0-hybrid	PCP036219.1	ppa015680m
MDP0000215587	Solyc10g006640.2.1	AT2G22750.2	gene22289-v1.0-hybrid	PCP019425.1	ppa015634m
MDP0000261622	Solyc10g050690.1.1	AT3G16175.1	gene31888-v1.0-hybrid	PCP043550.1	ppa013597m
MDP0000188674	Solyc07g056470.2.1	AT1G17170.1	gene16882-v1.0-hybrid	PCP039596.1	ppa010831m
MDP0000171795	Solyc10g085380.1.1	AT3G59710.1	gene04519-v1.0-hybrid	PCP030105.1	ppa009526m
MDP0000148885	Solyc01g090550.2.1	AT2G42750.1	gene15792-v1.0-hybrid	PCP015203.1	ppa008584m
MDP0000127659	Solyc10g081220.1.1	AT5G18140.1	gene13651-v1.0-hybrid	PCP032274.1	ppa007978m

Continued on next page

Table 6.6 – continued from previous page

Apple	Tomato	Arabidopsis	Strawberry	Pear	Peach
MDP0000317465	Solyc06g060820.1.1	AT5G56980.1	gene09182-v1.0-hybrid	PCP016414.1	ppa007964m
MDP0000164967	Solyc09g092480.1.1	AT1G50580.1	gene21990-v1.0-hybrid	PCP032181.1	ppa005427m
MDP0000193970	Solyc09g011160.2.1	AT3G15430.1	gene14745-v1.0-hybrid	PCP014077.1	ppa005272m
MDP0000320591	Solyc05g005010.2.1	AT2G44210.2	gene03583-v1.0-hybrid	PCP000380.1	ppa005153m
MDP0000141910	Solyc07g017500.2.1	AT1G29800.1	gene13718-v1.0-hybrid	PCP017807.1	ppa004926m
MDP0000203157	Solyc07g041080.2.1	AT5G08560.1	gene27508-v1.0-hybrid	PCP009203.1	ppa003324m
MDP0000306556	Solyc02g044050.1.1	AT5G60790.1	gene08551-v1.0-hybrid	PCP016496.1	ppa003175m
MDP0000154543	Solyc04g051800.2.1	AT5G09930.1	gene16791-v1.0-hybrid	PCP031156.1	ppa002106m
MDP0000044499	Solyc05g054700.2.1	AT4G01810.1	gene24474-v1.0-hybrid	PCP029086.1	ppa001228m
MDP0000231217	Solyc03g006420.2.1	AT2G01810.1	gene09119-v1.0-hybrid	PCP035713.1	–
MDP0000319369	Solyc09g014300.2.1	AT5G01220.1	gene18817-v1.0-hybrid	PCP014154.1	–
MDP0000246447	Solyc02g055510.1.1	AT3G51700.1	gene33811-v1.0-hybrid	–	ppa009389m
MDP0000214562	Solyc01g111010.2.1	AT5G48590.1	gene13858-v1.0-hybrid	–	–
MDP0000292654	Solyc06g060340.2.1	AT1G44575.1	–	PCP008950.1	ppa009763m
MDP0000168167	Solyc08g077440.2.1	AT2G02480.1	–	PCP002131.1	ppa000379m
MDP0000765473	Solyc04g079830.2.1	AT4G03250.1	–	PCP013723.1	–
MDP0000504077	Solyc08g006150.2.1	AT5G26220.1	–	PCP008932.1	–
MDP0000260339	Solyc01g006940.2.1	AT3G53460.4	–	PCP006056.1	–
MDP0000878006	Solyc08g080580.2.1	AT5G02320.1	–	–	ppa022218m
MDP0000185661	Solyc02g077140.2.1	–	gene31619-v1.0-hybrid	PCP044626.1	ppa025554m
MDP0000202596	Solyc02g090940.2.1	–	gene30400-v1.0-hybrid	PCP029294.1	ppa019013m
MDP0000182101	Solyc11g005230.1.1	–	gene06134-v1.0-hybrid	PCP022209.1	ppa015256m
MDP0000230225	Solyc09g072900.2.1	–	gene06058-v1.0-hybrid	PCP024621.1	ppa010552m
MDP0000893203	Solyc03g121880.2.1	–	gene07500-v1.0-hybrid	PCP019824.1	ppa004252m
MDP0000249364	Solyc03g005090.2.1	–	gene26180-v1.0-hybrid	PCP024857.1	ppa002029m
MDP0000317119	Solyc04g011970.1.1	–	gene03556-v1.0-hybrid	PCP024158.1	ppa001180m
MDP0000122728	Solyc03g123430.2.1	–	gene21524-v1.0-hybrid	PCP043945.1	–
MDP0000934741	Solyc10g079790.1.1	–	gene08835-v1.0-hybrid	PCP026069.1	–
MDP0000747922	Solyc03g044280.1.1	–	gene31665-v1.0-hybrid	PCP024105.1	–
MDP0000212560	Solyc08g081420.2.1	–	gene25929-v1.0-hybrid	PCP006206.1	–
MDP0000165222	Solyc06g050690.2.1	–	gene30513-v1.0-hybrid	–	ppa022399m
MDP0000855311	Solyc02g089560.2.1	–	gene32564-v1.0-hybrid	–	–
MDP0000384437	Solyc12g056420.1.1	–	gene19126-v1.0-hybrid	–	–
MDP0000892561	Solyc11g065810.1.1	–	–	PCP043010.1	–
MDP0000237793	Solyc04g008560.2.1	–	–	PCP042925.1	–
MDP0000132426	Solyc06g036410.1.1	–	–	PCP030167.1	–
MDP0000141543	Solyc09g005460.1.1	–	–	PCP009279.1	–
MDP0000157525	Solyc02g085560.2.1	–	–	–	ppa006614m
MDP0000279619	Solyc12g037970.1.1	–	–	–	–
MDP0000274363	Solyc11g012620.1.1	–	–	–	–
MDP0000392357	Solyc07g043300.1.1	–	–	–	–
MDP0000305623	Solyc05g014130.2.1	–	–	–	–
MDP0000340590	Solyc02g087940.2.1	–	–	–	–
MDP0000264992	–	AT5G56325.1	gene30861-v1.0-hybrid	PCP038345.1	ppa019166m
MDP0000330297	–	AT3G13860.1	gene08410-v1.0-hybrid	PCP024809.1	ppa004110m
MDP0000192598	–	AT5G47740.2	gene04099-v1.0-hybrid	PCP021035.1	ppa016921m
MDP0000249233	–	AT5G42490.1	gene16455-v1.0-hybrid	PCP015706.1	ppa026600m
MDP0000516398	–	AT4G39680.1	gene03084-v1.0-hybrid	PCP008893.1	ppa024614m
MDP0000041327	–	AT4G12120.1	gene19869-v1.0-hybrid	–	ppa021856m
MDP0000222299	–	AT1G15190.1	gene11763-v1.0-hybrid	–	ppa020798m
MDP0000311103	–	AT2G16370.1	gene13946-v1.0-hybrid	–	–
MDP0000203784	–	AT1G73750.1	–	PCP023011.1	ppa005002m
MDP0000240395	–	AT4G25050.2	–	PCP044585.1	–
MDP0000181740	–	AT2G25010.1	–	PCP029781.1	–
MDP0000165501	–	AT4G04180.1	–	PCP013566.1	–
MDP0000275938	–	AT1G10840.1	–	PCP005812.1	–
MDP0000243826	–	AT5G53620.1	–	–	ppa025698m
MDP0000291916	–	AT4G35987.1	–	–	ppa021442m
MDP0000830054	–	AT5G37860.1	–	–	ppa008910m
MDP0000003574	–	ATCG00170.1	–	–	–
MDP0000144321	–	AT4G18240.1	–	–	–
MDP0000232968	–	AT3G51360.1	–	–	–
MDP0000707429	–	AT3G15920.1	–	–	–
MDP0000311242	–	AT1G36340.1	–	–	–
MDP0000144320	–	–	gene29270-v1.0-hybrid	PCP016916.1	ppa017621m
MDP0000180472	–	–	gene06738-v1.0-hybrid	PCP041318.1	ppa011677m
MDP0000186455	–	–	gene13763-v1.0-hybrid	PCP026619.1	–

Continued on next page

Table 6.6 – continued from previous page

Apple	Tomato	Arabidopsis	Strawberry	Pear	Peach
MDP0000457301	–	–	gene13271-v1.0-hybrid	PCP020021.1	–
MDP0000138601	–	–	gene12621-v1.0-hybrid	PCP010994.1	–
MDP0000259837	–	–	gene05668-v1.0-hybrid	PCP004746.1	–
MDP0000259471	–	–	gene03605-v1.0-hybrid	–	ppa010196m
MDP0000224750	–	–	gene29373-v1.0-hybrid	–	–
MDP0000125711	–	–	gene25264-v1.0-hybrid	–	–
MDP0000138677	–	–	gene24009-v1.0-hybrid	–	–
MDP0000287117	–	–	gene19075-v1.0-hybrid	–	–
MDP0000302549	–	–	gene18716-v1.0-hybrid	–	–
MDP0000269484	–	–	gene07756-v1.0-hybrid	–	–
MDP0000219351	–	–	gene05197-v1.0-hybrid	–	–
MDP0000917312	–	–	gene02373-v1.0-hybrid	–	–
MDP0000314372	–	–	–	PCP032499.1	ppa024752m
MDP0000213893	–	–	–	PCP024059.1	ppa012617m
MDP0000691789	–	–	–	PCP035409.1	ppa001870m
MDP0000200603	–	–	–	PCP044138.1	–
MDP0000304996	–	–	–	PCP035092.1	–
MDP0000136627	–	–	–	PCP029650.1	–
MDP0000171292	–	–	–	PCP028640.1	–
MDP0000276894	–	–	–	PCP022665.1	–
MDP0000320442	–	–	–	PCP017830.1	–
MDP0000301931	–	–	–	PCP017393.1	–
MDP0000275389	–	–	–	PCP011501.1	–
MDP0000320612	–	–	–	PCP010378.1	–
MDP0000437671	–	–	–	PCP004857.1	–
MDP0000010844	–	–	–	PCP002692.1	–
MDP0000198290	–	–	–	–	ppa001677m

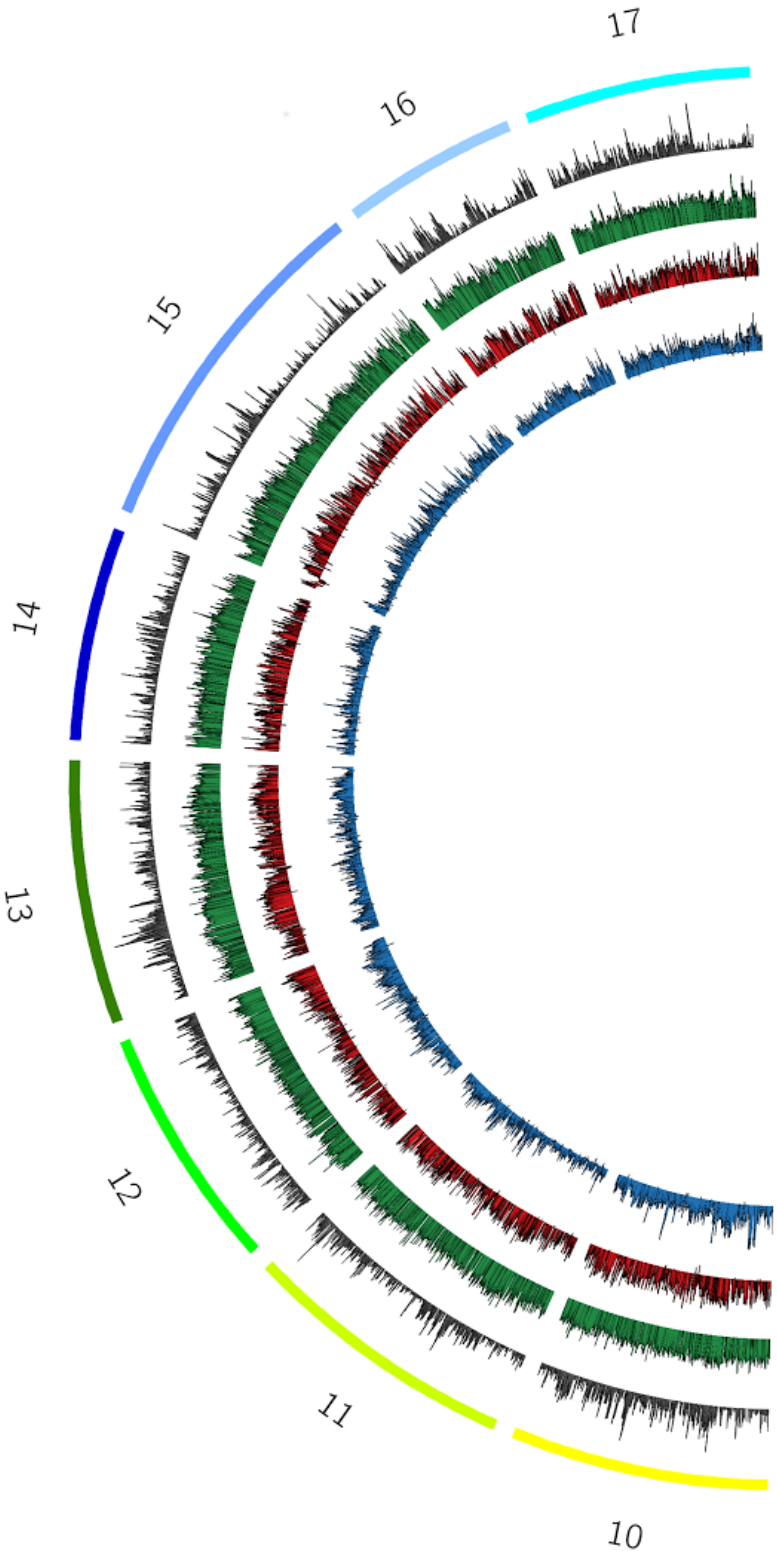
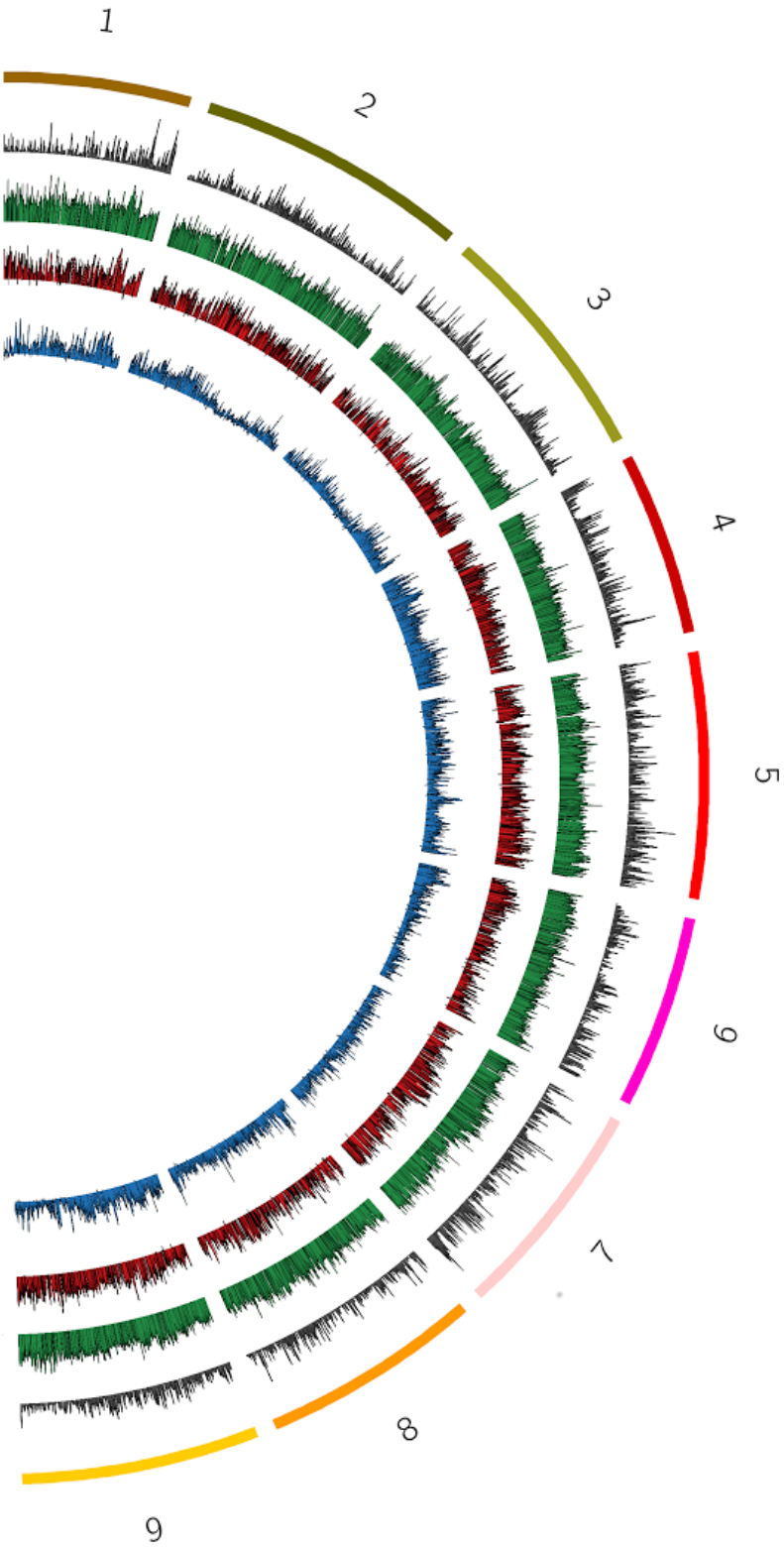
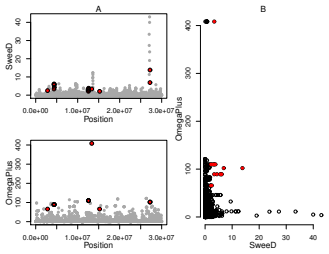


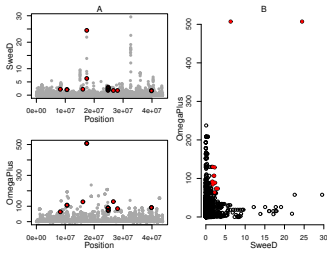
Figure 6.5: Genetic variability summary. For each chromosome from the outer to the inner circle nucleotide diversity, observed heterozygosity, Tajima's D and mean F_{ST} are displayed in sliding windows.


1
2
3
4
5
6
7
8
9

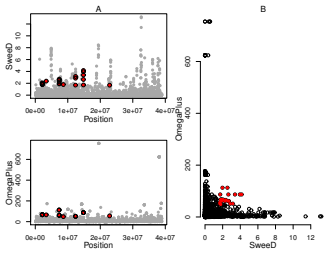
Chromosome 1



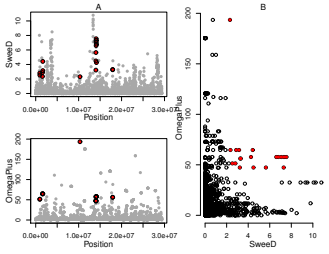
Chromosome 2



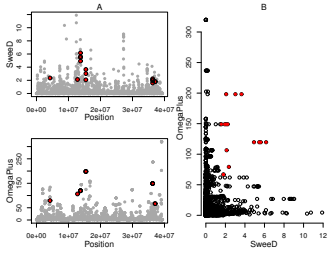
Chromosome 3



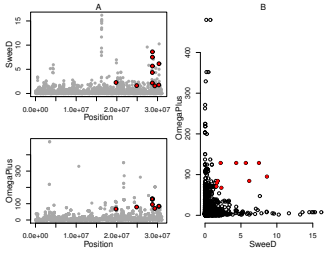
Chromosome 4



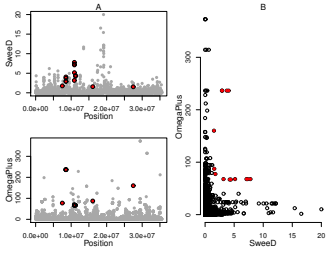
Chromosome 5



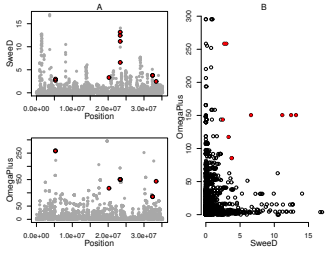
Chromosome 6



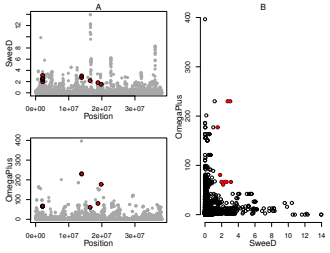
Chromosome 7



Chromosome 8



Chromosome 9



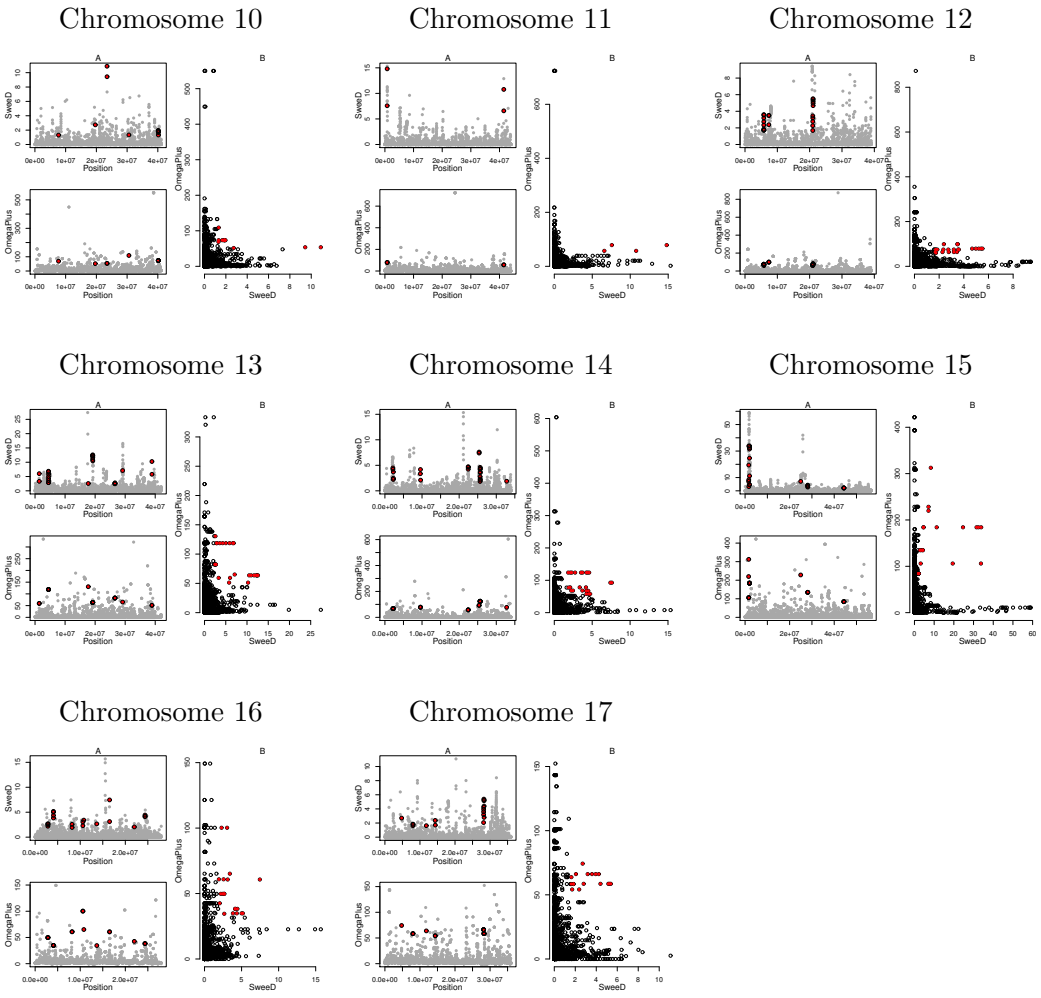


Figure 6.5: Selective sweep analysis for each chromosome. (A) The x axis denotes the position on the chromosome, and the y axis shows the CLR evaluated by SweeD (upper panel) and the ω statistic (bottom panel) evaluated by OmegaPlus. (B) The joint plot for SweeD and OmegaPlus. Red dots denote outliers at a significance level of 5%.

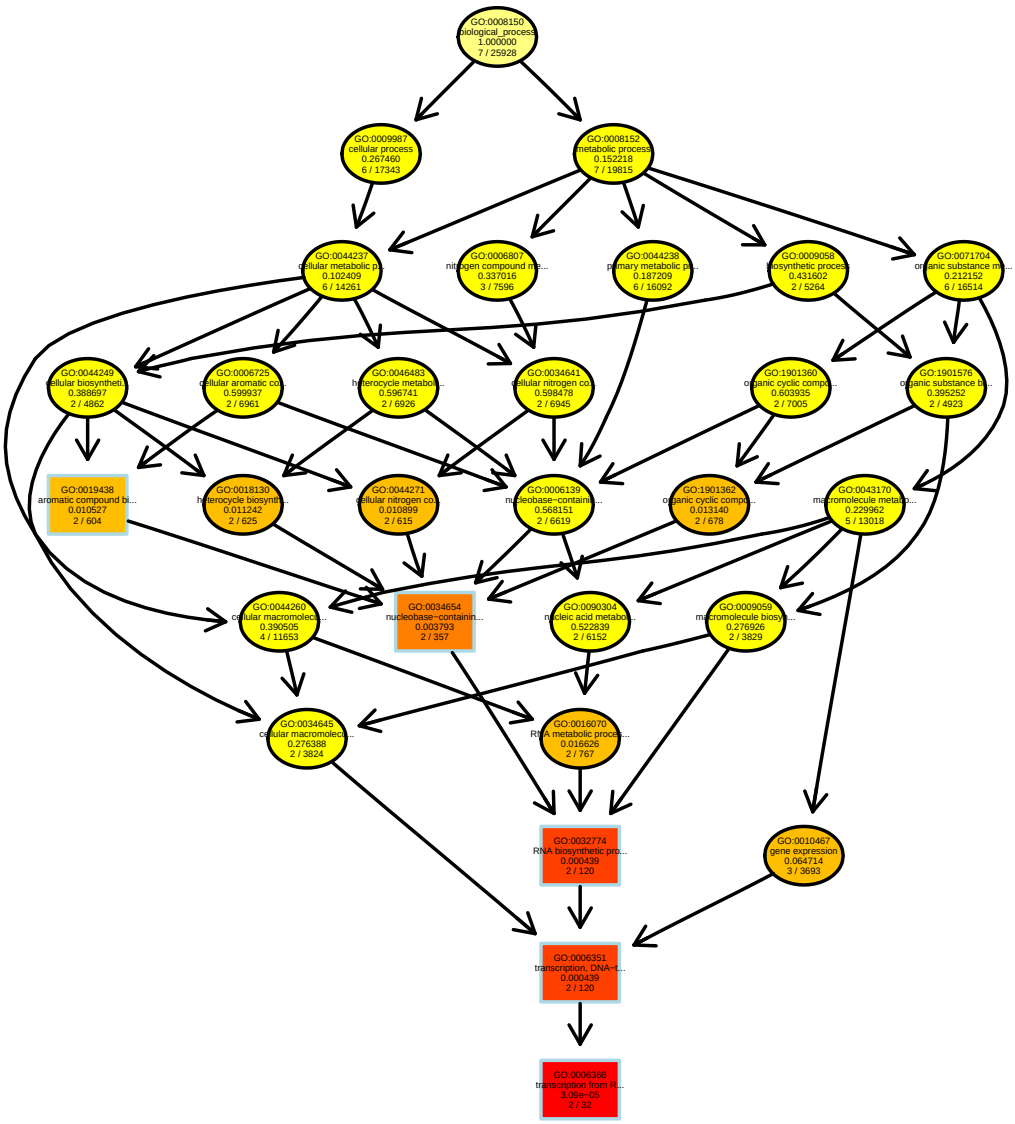


Figure 6.6: The subgraph of the top five GO terms identified from the GO term enrichment in Biological Process. Rectangles indicate the five most significant terms. Rectangle color represents the relative significance, ranging from dark red (most significant) to bright yellow (least significant). For each node, some basic information is displayed. The first two lines show the GO identifier and a trimmed GO name. In the third line the raw p-value is shown. The forth line is showing the number of significant genes and the total number of genes annotated to the respective GO term.

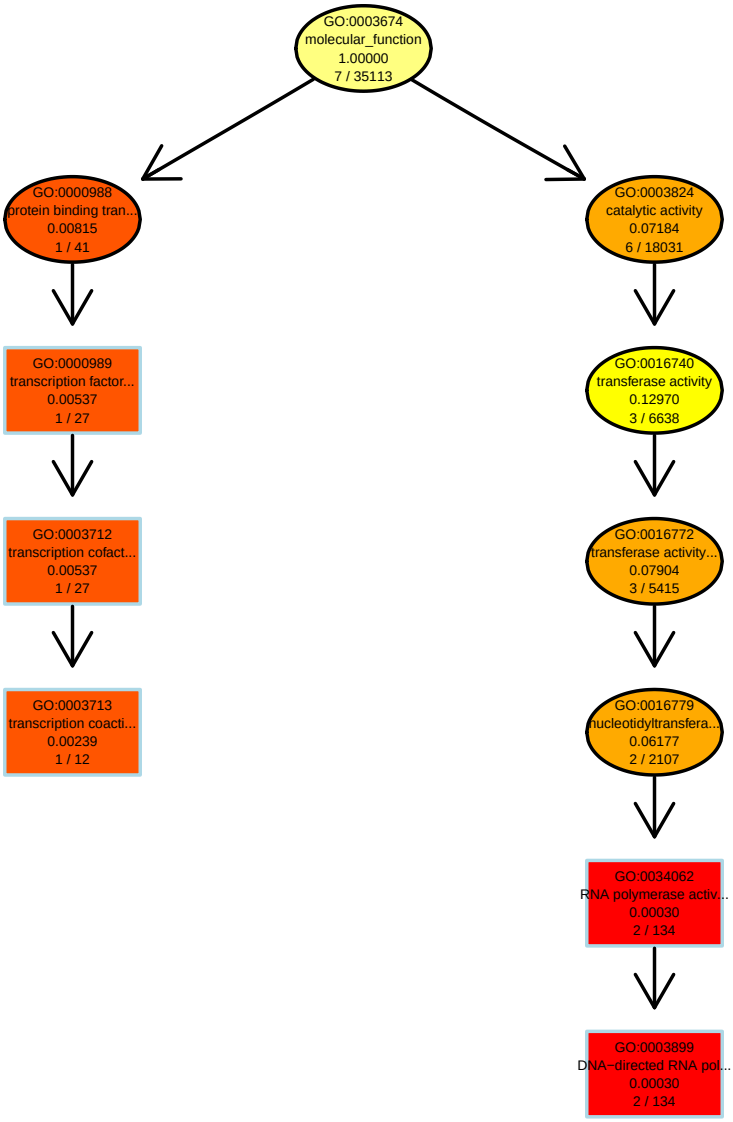


Figure 6.7: The subgraph of the top five GO terms identified from the GO term enrichment in Molecular Function. Rectangles indicate the five most significant terms. Rectangle color represents the relative significance, ranging from dark red (most significant) to bright yellow (least significant). For each node, some basic information is displayed. The first two lines show the GO identifier and a trimmed GO name. In the third line the raw p-value is shown. The forth line is showing the number of significant genes and the total number of genes annotated to the respective GO term.

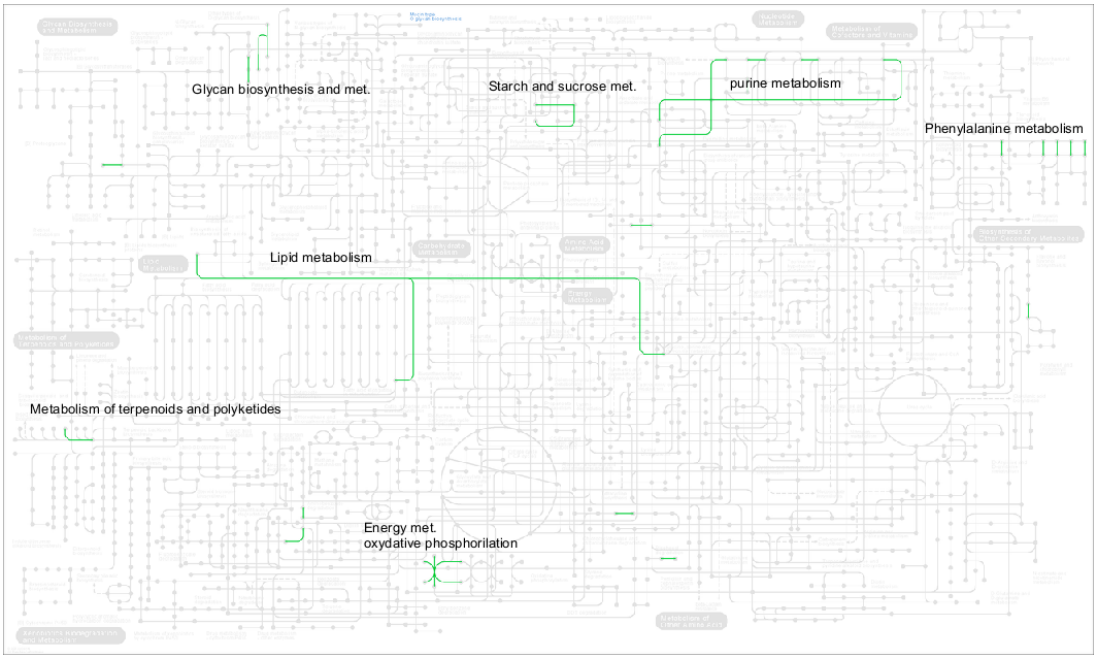


Figure 6.8: Schematic representation of the metabolic pathways. The positively selected genes identified in sweeps are highlighted in green.

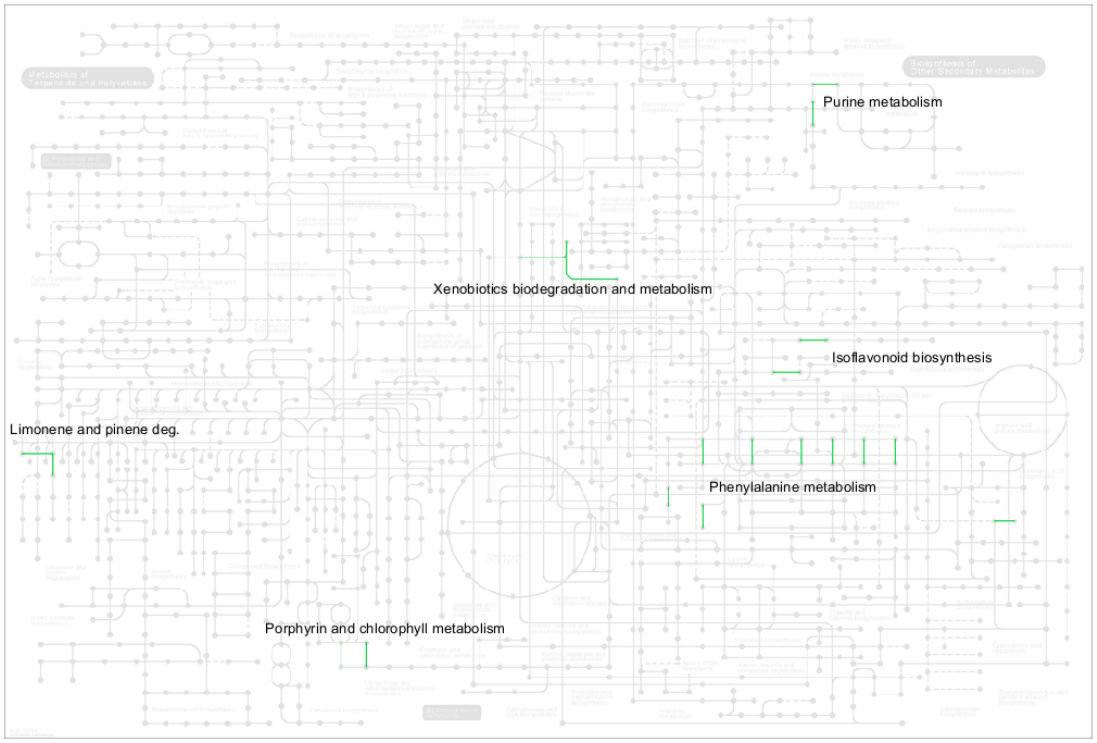


Figure 6.9: Schematic representation of the biosynthesis of secondary metabolite pathways. The positively selected genes identified in sweeps are highlighted in green.

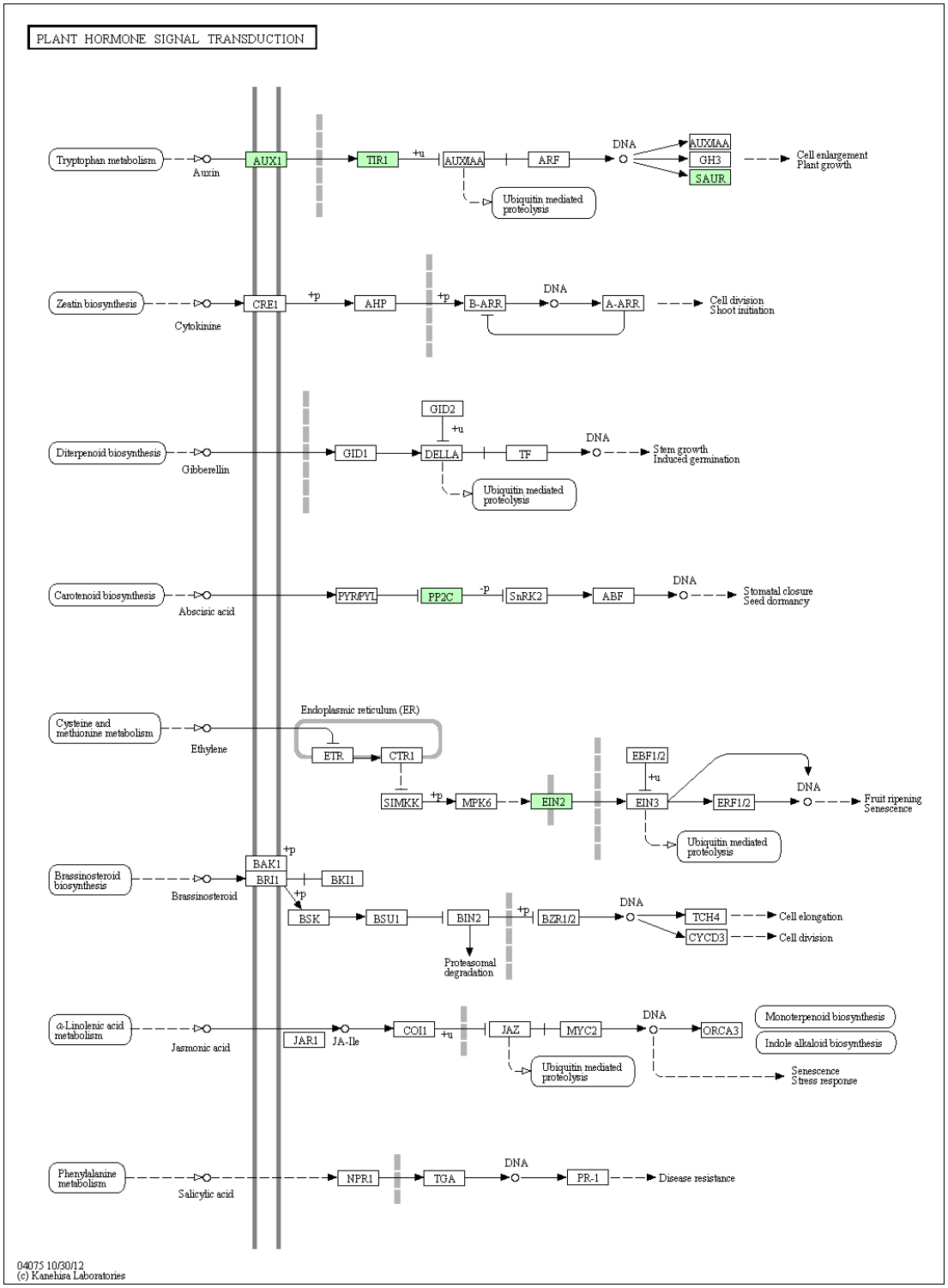


Figure 6.10: Schematic representation of the plant hormone signal transduction pathway. The positively selected genes identified in sweeps are highlighted in green.

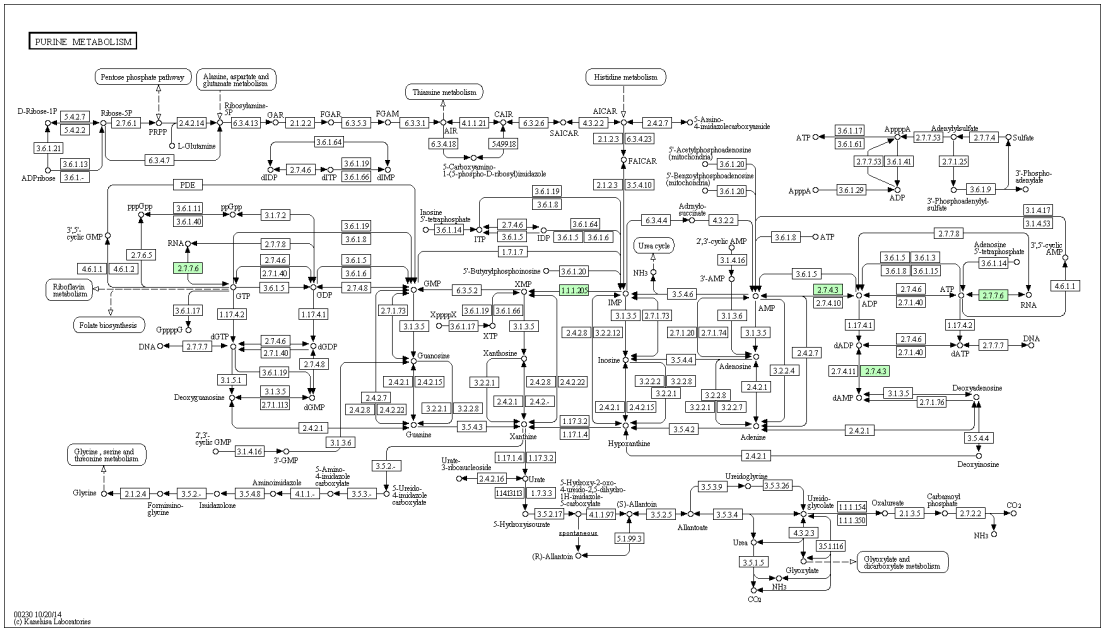


Figure 6.11: Schematic representation of the purine metabolism pathway. The positively selected genes identified in sweeps are highlighted in green.

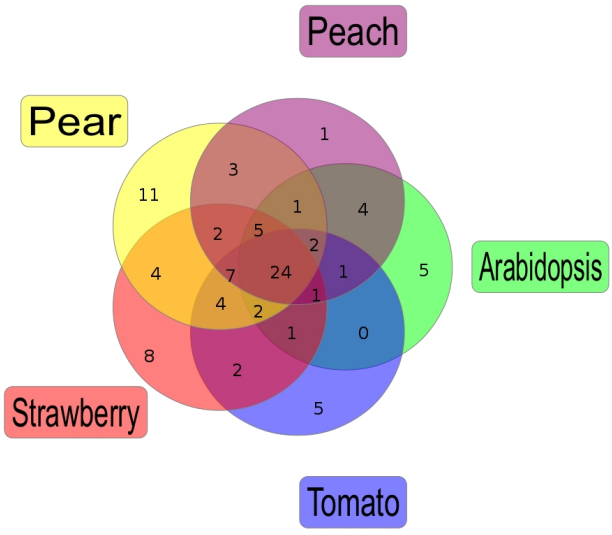


Figure 6.12: Diagram showing the number of orthologous gene candidates for positive selection shared between different species.

7

Discussion

7.1 SNP calling and filtering

The SNP calling and filtering pipeline used in this work was developed by Bianco and colleagues in [70]. This new pipeline successfully avoided the problem of extensive genome paralogy in apple. A more stringent SNP filtering was achieved by including two apple DH accessions in the resequencing samples and removing heterozygous SNPs identified in either of those genotypes during the SNP selection process. Some filtering steps of the pipeline are specifically designed for the experimental technique used (SNPChip array). The SNPchip accuracy designed by means this pipeline is around 90% (Bianco L., personal communication). In order to keep a comparable accuracy the pipeline was applied as whole even with those steps that were not strictly necessary. The high coverage resequencing strategy applied in this work allowed genotypes to be called for most samples, and a minimum of ten reads per variant was imposed to successfully call a genotype. The comparison of sequenced reads for a potential variant site across multiple samples has the potential to differentiate systematic sequencing errors from real SNPs. Other methods have been recently developed for high-throughput genotyping of eukaryotic genomes using short-read sequencing technologies ([80], [34], [81]). These techniques have been employed in the development of high-density linkage maps and the identification of markers linked to agronomic traits in plants. Such studies include, for instance, genotyping by sequencing (GBS) based genetic maps of *Rubus idaeus* [82], *Hordeum vulgare* [83], [84], and restriction-site associated

DNA (RAD) sequencing in *Lolium perenne* [81], *Hordeum vulgare* [85] and *Lupinus angustifolius* [86]. These techniques employ reduced genome representation achieved through restriction enzyme digestion and subsequent PCR analysis from adapted linker sequences, and do not require *a priori* knowledge of the SNPs being interrogated, making them useful for genetic analysis in species where no reference sequence is currently available. On the other hand, sampling chromosomes based on restriction digestion may introduce a bias in allele frequency estimation due to polymorphisms in restriction sites [87]. For the purpose of this thesis to study selection in diverse apple accessions, the WGS approach allows a comprehensive study of the genome avoiding the sampling of only specific regions, and it is feasible on heterozygous species such as apple.

7.2 Population structure

The population structure analysis performed in this work on the domesticated apple cultivars analyzed revealed a unique population. A very weak structure with three poorly differentiated sub-groups was detected by measuring F_{ST} (0.055, 0.083 and 0.096). Fourteen over 63 individuals resulted to be admixed (membership coefficient lower than 0.7), corresponding to 22.2%. Some hypotheses were formulated to explain the assignation of the individuals to the sub-groups. The members of group 1 are mainly cultivars from the North (i.e. Finland, Sweden, Russia). Group 2 is composed by cultivars that may be more related to wild apple species. Group 3 contains the core European apple breeding founder varieties. These results are consistent with other works that studied the population stratification of domesticated apple in local germplasm collections. In [88] 183 apple accessions from the Northeast of Spain were studied and four sub-groups were found. The membership coefficient was higher than 0.80 for 87%, 68%, 79% and 79% of the individuals in each sub-group. The clustering of the Spanish accessions in the sub-groups did not correlate with geographical origin, but in one were found all the old Spanish accessions and in another all the modern commercial cultivars. In [89] the population structure of an Italian germplasm of 383 accessions was studied. The membership coefficient was high for 66.7% of the individuals in one group and for 41.3% of individuals in the other group.

The assignment of each accession to the sub-groups was not clearly linked to any of the phenotypic traits that have been considered (i.e. ripening date, fruit size, shape, color, russet, acidity and sweetness).

7.3 Genetic variability

The knowledge of the existing genetic variability is of fundamental importance in crop species for the breeding and the management of genetic resources [90]. SNP frequency has been extensively investigated on the basis of large resequencing efforts in annual crops, where it appears to vary considerably between species (e.g. from 1 SNP/61 bp to 1 SNP/5,700 bp for maize and rice *Oryza sativa* spp. japonica cultivars, respectively [91], [92]). Many studies have also been reported for forest tree species where high SNP frequencies have been observed (e.g. 1 SNP/63 bp and 1 SNP/130 bp in *Pinus taeda* and *Populus trichocarpa*, respectively [93], [94]). The high outcrossing rates common in most forest tree species is an important factor contributing to the high level of genetic variability in most trees as compared to annual plants. SNP frequencies have been estimated for perennial fruit crop species such as *Citrus* (1 SNP/164 bp, based on database ESTs; [95]), almond (1 SNP/114 bp based onEST-based amplicon resequencing [96]), or cacao (1 SNP/71-223 bp, based on database ESTs [97]). Grape SNP frequency has been estimated in various studies as 1 SNP/117 bp (based on the Pinot Noir genome sequence [31]) or 1 SNP/64-104 bp (based on multi-locus analysis of cultivars; [98], [99]).

With respect to *M.* × *domestica*, previous works have estimated 1 SNP/149 (based on database ESTs; [100]), 1 SNP/225 bp (based on BAC end sequences [101]) and 1 SNP/227 bp (based on the genome sequence of 'Golden Delicious' [5]). The findings of this work of an average of 1 SNP/50 bp across all the cultivars confirm that apple is one of the most genetically polymorphic agricultural species analysed so far and is more than one order of magnitude more polymorphic than human (1 SNP/215 bp; [102]). The estimate of $\pi = 0.0003$ for *M. domestica* cultivars confirms this relatively high polymorphism, which is lower than the range of values recorded for maize ($\pi = 0.0096$ or 0.0036, according to [103] or [104], respectively) and grape ($\pi = 0.0051$ [99]), is in the range of values

measured for barley (π = 0.003-0.004; [104]), and rice (π = 0.0010-0.0027; [105]). Various aspects have contributed to the level of polymorphism in cultivated apple. Both apple and peach [106] are clonally propagated so they bear a reduced number of generations since domestication but apple has a higher level of heterozygosity because it is outcrossing while peach is self-crossing.

7.4 Selective sweep analysis

The increased availability of whole-genome molecular population data has led to the application of various methods to the investigation of selection events at the population level. These methods exploit different features of the data, they often do not have the same efficacy and are prone to errors depending on the biological background of the population (effective population size change, population structure). When the evolutionary history of the population of study is at least partially known, it is good practice to select the method and the tool that perform better in the scenario under investigation. Recently the performance of the most used tools for selection tests has been evaluated on a simulated dataset by Crisci *et al.* [107]. The selection tests SweepFinder [51], SweeD [77], OmegaPlus [108], and iHS [63] were evaluated under different combinations of the models: neutrality, bottlenecks, recurrent hitchhiking and single hitchhiking. Crisci *et al.* found that the standard implementation of SweepFinder and iHS had the highest rate of false negative errors. For both, this was probably due to their reliance on SNP density, where a high SNP density provides more power to the statistic. This means that they suffer from a lack of power in models that reduce diversity, like positive selection and population bottlenecks. Compared to SweepFinder, SweeD had a slightly improved performance, but this was mainly due to a reduced sensitivity to SNP density caused by the inclusion of monomorphic sites. In particular the weakness of iHS is the dependence on empirical outliers. It assumes that positive selection has occurred in the dataset and that these selected loci will have higher frequencies but both of these factors happen for the high proportion of neutral loci identified under bottleneck models, so it is unable to distinguish selective effects from those of a variety of population bottlenecks. OmegaPlus was found to be the most sensitive to the various model parameters and the most

powerful in detecting selection, but it is still prone to high false-positive rates under certain neutral non-equilibrium models. As many natural populations are characterized by a non-equilibrium status, statistics that are more accurate in the identification of selective events, and capable of dealing with the more biologically relevant models are desirable.

In this study a consensus between the results of a LD-based method (OmegaPlus) and the results of a SFS-based method (SweeD) was chosen to exploit the strengths of both and avoid false positives errors. This approach was also feasible thanks to the flexibility of the two methods to use the same input file and to the availability of a R script that create a consensus for the outputs. SweeD and OmegaPlus identified a comparable number of selective sweeps (2526 and 2354 respectively) but the consensus between the two was 1194 sweeps, corresponding to 47.26% and 50.72% of the sweeps identified individually. The ratio between the number of sweeps identified in each chromosome and the length of the chromosome confirms that the identification of sweeps is not dependant on the length of the sequence. The highest number of sweeps was identified on chromosome 16, relatively to its length. The number of candidate gene predictions for positive selection identified is 12.8% of the number of sweeps. No other works that apply SweeD and OmegaPlus in crop plants have been published until now. This approach was chosen also because it was the most effective with the data available for this study. Methods that identify positive selection based on the reduction of variability or the F_{ST}, require additional molecular data for the wild population(s). A few papers have been published that applied SweeD/SweepFinder statistic, and the composite likelihood ratio (CLR) to crops, but not a combination of CLR and ω-statistics. In rice (*Oryza sativa* L.), F_{ST} and the CLR across populations were used to identify ecotype differentiated genomic regions (EDRs) for upland and irrigated cultivation [109]. With the construction of the HapMap of maize (textitZea mays L.) in 2009, possibly selected regions were identified through the study of observed recombination, the reduction of nucleotide diversity, Tajima's D and F_{ST} [110]. A more recent study in which 282 maize inbred lines were resequenced , the genome was scanned for sweeps using the CLR approach observing a reduction of nucleotide diversity in the targeted regions and negative Tajima's D values [111]. In the same study it was also found

that the genomic regions targeted by selection were not the same in different subpopulations. This could be the result of different stages of maize improvement, the adaptation to local agricultural conditions, or the same phenotypic trait could be obtained by selection on different genomic regions [111]. The investigation of the origin of cultivated rice through the sequencing of 446 wild rice (*Oryza rufipogon*) and 1083 cultivated japonica and indica varieties revealed selected domestication loci through a genome scan of the nucleotide diversity ratio between wild and cultivated rice [112]. In sorghum (*Sorghum bicolor* (L.) Moench.) domestication loci have been studied with the sequencing of 971 worldwide accessions (wild relatives, elite lines and breeding lines) and the genome scan of expected heterozygosity, average nucleotide diversity and Tajima's D. Known domestication genes or orthologs of domestication loci from maize were found in regions with low heterozygosity [113]. The domestication study of African rice (*Oryza glaberrima*) from the wild progenitor *Oryza barthii* to identify regions of artificial selection, measured genome-wide levels of nucleotide diversity (π), and the site frequency spectrum. These values were used to identify homologs to known *O. sativa* domestication genes, and the integrated haplotype score statistic that pointed out recent uncomplete sweeps [114]. The common bean (*Phaseolus vulgaris* L.) dual domestication was studied with the sequencing of 160 wild and landraces from both the Mesoamerican and Andean gene pools. Working separately for two geographically distinct populations, Schmutz *et al.* calculated the nucleotide diversity ratio between wild and landraces and the F_{ST}. Low nucleotide diversity and high differentiation were used to identify candidate genes associated with domestication in both populations. Negative Tajima's D values confirmed the positive selection in candidate genes [115]. The study on selection and adaptation traits of poplar (*Populus trichocarpa*) exploited four different methods: F_{ST}, allele frequency related to mean annual precipitation and temperature measurements, iHS and allele frequency related to the first and second climate principal components axes. These techniques identified different outlier genes (just one gene in common among all of them), highlighting the potential of the applied methods to recognize different selection pressures as hard or soft sweeps influence the diversity pattern in different ways [116]. The data for the study of genome-wide diversity in a diverse collection of watermelon (*Citrullus lanatus* var. *lanatus*) was produced by

genotyping by sequencing. Selective sweeps were identified by pairwise F_{ST} test between geographically distinct groups. The sweep regions were then checked for the presence of LD blocks ([117]). The breeding history of tomato (*Solanum lycopersicum*) was studied using resequencing data from 360 accessions (10 accessions of wild species, 17 modern commercial hybrids and 333 accessions from different species of the red-fruited clade) [118]. Measuring the nucleotide diversity ratio between the three groups to which the accessions were assigned on the basis of passport information, fruit weight and other morphological traits, Lin and colleagues [118] identified selective sweeps for both the domestication and the improvement processes. A comparative population genomics work in peach (*Prunus persica*) analyzed the polymorphism diversity (θ_w) in three groups (wild, edible and ornamental) showing that a bottleneck occurred between wild and edible peach. Regions under artificial selection in edible and ornamental groups were identified by Tajima's D test. The polymorphism-based methodologies for the estimation of population genetic parameters often give not consistent results despite they estimate similar parameters using similar data sets. Biases in selection estimators are introduced by ignoring the demographic history of the population and by the assumption of neutrality.

In this work 153 candidate gene predictions for positive selection were identified in 1194 sweeps. The GO terms enriched in these regions were: transcription from RNA polymerase II promoter, photosynthesis and protein ubiquitination for the biological process ontology, while DNA-directed RNA polymerase activity, transcription coactivator activity and ubiquitin-protein ligase activity were enriched in the molecular function ontology. BlastKOALA mapped some of the candidate gene predictions for positive selection in the KEGG database also in metabolic pathways, in the biosynthesis of secondary metabolites, plant hormone signal transduction and in the purine metabolism.

In the study of Morris and colleagues [113] in which 971 sorghum accessions were genotyped for 265000 SNPs, the selected regions identified contained genes related to the shattering habit, the starch content and colocalized with mapped loci of height and maturity. Also in the work on the African rice genome [114], with the resequencing of 20 *O. glaberrima* and 94 wild accessions, the selected regions identified contained shattering genes. In the work of Lyu *et al.* [109]

that studied the adaptation of rice to different cultivation modes the adaptive traits were enriched for genes of the classes: lipase containing GDSL domain, peroxidases, glutathione-related, and auxin signaling pathway. The auxin signaling pathway was also identified in the present work by BlastKOALA in the plant hormone signal transduction pathway. In particular the auxin influx carrier (AUX1 LAX family) and a SAUR family protein that lead to cell enlargement and plant growth were recognized. The protein phosphatase 2C which leads to stomatal closure and seed dormancy was also identified in the plant hormone signal transduction pathway, together with ethylene-insensitive protein 2 that leads to fruit ripening and senescence. In the domestication study of common bean, domesticated genes were identified in the pathway that controls the principal floral integrator genes SOC1 and FT (i.e. vernalization genes) [115]. A domestication gene was found to be an ortholog of *Arabidopsis thaliana* BB, a component of the ubiquitin ligase degradation pathway that controls flower and stem size. The finding of this research are common to those found with the data presented here, in which genes candidate for selection were enriched in GO terms protein ubiquitination and ubiquitin-protein ligase activity. Multiple candidate genes for domestication in bean were also components of nitrogen metabolism pathways, which directly affect plant size (nitrogen metabolism, nitrogen transporters and asparagine synthase) [115]. In the comparative study of edible and ornamental peach [106], several candidate genes under selection were identified. Among those, five genes of the auxin response factor gene family were also identified. In addition, an enrichment of gene families related to the carbohydrate metabolic process, tricarboxylic acid cycle, and photosynthesis was found. Similarly, in the current work on domesticated apple, the photosynthesis GO term was found to be enriched in the candidate gene predictions. In peach, another important gene family encodes enzymes that participate in carbohydrate metabolic processes to improve fruit aroma. In particular, a gene that encodes a homologous protein to D-sorbitol-6-phosphate dehydrogenase (S6PDH) was identified. Sorbitol is a transporter of photosynthesis product substances in the Rosaceae family. In the same metabolic process, a candidate gene was identified also in apple, the beta-glucosidase, involved in the starch and sucrose metabolism. Genes encoding ubiquitin protein ligase were identified in peach as well. The study of tomato

domestication and breeding based on the resequencing of 360 accessions, [118] identified selected regions in the genome for both the domestication and the improvement processes. These regions were also found associated with many known QTLs for fruit mass and color.

In apple cultivars, some QTLs for fruit size/fruit weight have been identified on chromosomes 8 and 15 [119] regions corresponding to these QTLs were detected as positively selected also in the present work. In a recent study by Khan and colleagues [120] selective sweep regions were identified by F_{ST} analysis between wild and domesticated apples. Among the selected genes with functional annotation, they found some related to the sugar metabolism, acidity, fruit size and a transcription factor involved in fruit development.

8

Conclusion

This thesis represents the first study of genetic variability and selection in domesticated apple that exploits whole genome resequencing data of a considerable number of individuals. The resequencing provided a clear picture of the complete genomic landscape variability, confirming the high level of nucleotide diversity previously reported in domesticated apple (average $\pi = 0.0003$) [5].

The cultivars selected for this work are part of the common breeding material and have been chosen to maximize the variability. Three sub-groups poorly differentiated, and many admixed individuals were identified in the studied population. This is consistent with other findings showing the absence of a strong population structure in domesticated apple [26], [89].

For the first time in crop plants, a combination of the composite likelihood ratio test and the ω statistic were used to identify genomic regions under positive selection. Both methods identified a total of 1194 candidate genomic regions for selective sweeps that contain 153 gene predictions. The annotation of candidate gene predictions for positive selection by gene ontology terms and mapping on KEGG pathways gave a broad idea on the pathways targeted by artificial selection, such as: photosynthesis, protein ubiquitination, plant hormone signal transduction (auxin in particular) and starch and sucrose metabolism. The same pathways were pointed out as positive selection targets in very recent domestication studies on other crops (peach [106] and bean [115]). These traits are consistent with the domestication process which tends to artificially select tastier, sweeter and bigger fruits.

9

Appendix A: Apple Cultivars

9.1

Description of the apple cultivars origin used in this work available in literature [23], [121].

Abbondanza It is a dessert apple, original from Italy, arose near S.Pietro Capofiume (Bologna) and was selected in 1896. It is scarlet flushed. It has a pronounced raspberry tang, becoming strongly vinous. Sweet, firm, white flesh. Formerly was grown in Emilia-Romagna, where was used mainly for cooking. The crop yield is heavy and the fruit is small (diameter below 5 cm). The picking time is mid of October and it can be stored until January-March.

Aivaniya It is a dessert apple, crisp, juicy with tough skin. It is widely grown in central southern Bulgaria in Plovdiv region (the precise origin is not known). It was the main commercial variety in the 1960s, now it has just local importance. The crop yield is heavy. The picking time is middle/late October and it can be stored until December-March.

Åkerö Dessert apple, probably originated in Sweden. Either the seedlings arose in the ker mansion, southern of Stockholm, or they were imported from Holland by the Tessin family, owners of the ker estate. It was brought to notice in 1858 by the pomologist Olof Eneroth. The mother tree still

stands, although only partially alive. It has a beautiful appearance: pink flush, deep red stripes over pale primrose. It is juicy, refreshing, with savory tang. It is a hardy plant, surviving even the coldest winters. It was grown in Northern Europe. The crop yield is good and the picking time is August-early September.

Alfred Jolibois Dessert apple, original from France. It was raised by Alfred Jolibois at Gurgy, Yonne (Burgundy) and was described in 1948. It has a deep red flush. It is crisp with firm flesh and little flavour. The crop yield is heavy. The picking time is middle/late October and it can be stored until December-March.

Annurca Dessert apple, original from Campania, southern Italy. It is a very old cultivar, already described in 1583. Fruits are small, dark red. It is crisp, and fruity. The picking time is late October and it can be stored until November-February.

Antonovka Apple used both fresh and for cooking. It arose in Kursk, Russia and it has been recorded since 1826. The fruit is large and white, crispy, refreshing and juicy. It was also used as rootstock. It is resistant to scab and partially to mildew. The picking time is early September and it can be stored until November-December.

Braeburn Dessert apple that comes from New Zealand. It was discovered on the property of O. Moran, Waiwhero, Upper Moutere, Nelson. It was grown commercially starting from 1952 by William Bros, Braeburn Orchards. At the beginning it was believed to be a Lady Hamilton seedling. In the 1970s, two forms were present in New Zealand, one maturing earlier than the other, the commercial plants now are of the earlier form. It refreshing, crisp, with firm flesh, sometimes perfumed. It is grown in warm regions: Australia, South Africa, South America, France, Germany, Italy and Canada. The picking time is late October and it can be stored until January-March.

Budimka This cultivar comes from Serbia and it is mostly used for culinary purposes. It was collected in 1936. It has fruity and sharp taste, when

cooked keeps the shape and has a light taste. The picking time is mid October and it can be stored until December, sometimes April.

Cox's Orange Pippin Dessert apple, from the United Kingdom. It was raised in early 1825 by Richard Cox at Colnbrook Lawn, Slough, Bucks. It was probably introduced in early 1850 by Smales and Son, Colnbrook. It ripens perfectly. It is sweet with rich, intense, aromatic flavour and has a creamy flesh. It was described as spicy, honeyed, nutty, pear-like. It was voted the best dessert apple of the south at the 1883 Congress. It was first grown commercially probably in 1862 by Thomas Rivers in Herts. Problems with diseases led to its rejection in early 1900s but it regained commercial popularity with the introduction of lime sulphur sprays in the 1920s. Since the 1970s is the main English apple, but it is grown also in Holland, Germany, Belgium, northern France, New Zealand. The picking time is late September / early October and it can be stored until November-December-early January.

De L'Estre Dessert apple, from France. It arose probably in Corrze (Limousin) and was popularised in the late 1700s by M. Turgot of Limoges. It is crisp, with rich and intense flavour. It was valued for the flavour and the long keeping quality, it was used also for cooking and cider. The picking time is late October and it can be stored until January-April.

Delicious Dessert apple from the U.S. It arose around the 1870s in the farm of Jesse Hiatt in Peru, Iowa. It was commercially introduced by Stark Bros., Missouri. It is densely sweet, aromatic, with creamy flesh. It is mostly grown to be eaten fresh, but it is also used for juice. Over 100 sports have been identified in the U.S. The picking time is mid October and it can be stored until December-March.

Durello di Forlì Dessert apple, very old Italian variety. It has a brisk taste and some sweetness, hard flesh. It was formerly grown in Emilia, Lombardia and Veneto. It was used for cooking. The picking time is mid/late October and it can be stored until January-March.

Filippa Dessert apple from Denmark. It was sown in pot around 1880 by Filippa Johannsen in Hundstrup, South Fyn Island. It is very scented, soft, fruity, with white flesh. The picking time is mid September and it can be stored until October-December.

Fuji Dessert apple from Japan. It was raised in 1939 by H. Niitsi, Hort. Res. St., Morioka. Ralls Janet and Delicious are the parents. It is honeyed and sweet, with crisp, firm and juicy flesh, a though skin. It is grown in Japan and China where is the main variety, but also in the U.S., Italy and France. The picking time is late October and it can be stored until December-March.

Golden Delicious Dessert apple from the U.S. It arose in the early 1890s with H. A. Mullis, Clay County, West Virginia. It was introduced in 1916 by Stark Bros, Missouri. It is honeyed with crisp, juicy, almost yellow flesh. In many countries has dual purpose (dessert and cooking) and it is also used for preparation of baby food. It was extensively planted in the U.S. from the 1920s, in Kent from the 1930s but it diffused widely in Europe in the 1960s. The picking time is late October and it can be stored until November-February, if stored cold also until June.

Jonathan Dessert apple from the U.S. It arose on the farm of Philip Rick, Woodstock, Ulster County, New York. It was introduced in England around 1826 probably by Thomas Rivers. It is crisp, juicy, with plenty of refreshing acidity. In the U.S. it is also valued for sauce and pies. It got early popularity thanks to Judge Buel, president of the Albany Horticultural Society. In 1829 he sent samples to the Massachusetts Horticultural Society who declared it the most promising new variety of the year. After that it spread to the South and West of the U.S. as well as to Europe and Australia. It became widely grown throughout the warmer apple producing regions of the world. The picking time is early October and it can be stored until November-January.

Keswick Codlin Cooking apple from the U.K. It was found at Gleaston Castle, near Ulverston, Lane. It was introduced by J. Sanders of Keswick around 1793. When cooked to juicy, cream froth or puree, it hardly needs any sugar.

It is also a refreshing eating apple. It is one of the most popular cooking apples of the 19th century and the trees were also used as decorative. The picking time is mid/late August and it can be stored until late August-September-October.

Lady Williams Dessert apple from Australia. It arose around 1935 with A. R. Williams, Bonomia Farm, Paynedale, Donnybrook, Western Australia. It is a possible Granny Smith with Jonathan or Rokewood cross. It has a bright red flush and firm crisp flesh. In the 1970s the variety was widely grown in Australia. The picking time is November and it can be stored until January-May.

Macoun Dessert apple from the U.S. It was raised by R. Wellington, NYSAES, Geneva. It is a cross between McIntosh and Jersey Black. It was introduced in 1923 and named after Canadian fruit breeder W. T. Macoun. It is very sweet, scented, with a hint of strawberry flavour, it has snow white, juicy flesh and tough skin. The picking time is mid September and it can be stored until October-December.

McIntosh Dessert apple from Canada. It was found in 1796 by J. McIntosh close to Dundela, Ontario, Canada. It was commercially diffused since 1870. It has white firm flesh, juicy, with some acidity and aromatic, slightly sweet taste. The picking time is late August/mid September.

Papirovka It is considered a synonym of White Transparent cultivar except in Russia, where they are regarded as distinct.

Patte de Loup Dessert apple from France. It arose probably in the late 1700s, around Beuprau, Maine-et-Loire (Western Loire). It is sharp, fruity but mellowing. The picking time is late October and it can be stored until January-April.

Pepino Jaune Dessert apple from France. It was received in 1948 from Cotes du Nord, Brittany. It is sharp, astringent. It is possibly a cider variety. The picking time is late September and it can be stored until October-November.

Priscilla-NL Dessert apple from the U.S. It was raised in 1961 by the cooperative breeding program of Purdue, Rutgers and Illinois University. It has a complex parentage that involves *Malus floribunda* (that carries V_f gene for scab resistance). It was introduced in 1971 and named after the wife of prof. F. D. Hovde, president of Purdue University. It is very sweet, scented, crispy and juicy. It is resistant to scab and mildly resistant to mildew and fireblight. The picking time is late mid October and it can be stored until November-December.

Reinette Clochard Dessert and cooking apple from France, synonymous of Clochard. It is known possibly since mid 1800s but it was described in 1948. It has intense, rich, quite aromatic flavour and firm and creamy flesh. It was one of the best known market-apples, widely grown, traditionally in the West of France. It was valued also for cooking, recommended for Crepes a la Normande and Tarte aux Pommes. The picking time is mid October and it can be stored until November-December, sometimes March.

Renetta Grigia di Torriana Dessert apple from Italy. It arose in Torriana di Barge, Cuneo, West Piedmont and was distributed since 1905. It is scented, sweet, with hardly any acidity. This variety was very popular in Torriana in the 1920s but now is not grown anymore. The picking time is late November and it can be stored until November-February/April.

Worcester Pearmain Dessert apple from the United Kingdom. It arose with Mr. Hale, gardener of Swan Pool, St. Johns, near Worcester. It is sweet with intense strawberry flavour, firm and juicy flesh. It was Englands main early autumn main commercial variety since the late 1800, now is only grown on small extent for the market. It is still a very popular garden apple for its distinctive blossom: almond opening to silvery white. The picking time is early/mid September and it can be stored until September-October.

Young America Apple from the U.S. It arose in New York and was introduced in the 1800s. It is ornamental for the flowering and bears small bright red apples. The picking time is late August/early September.

10

Appendix B: Files Format Description

10.1 SAM/BAM Files

A SAM (Sequence Alignment/Map format) file is TAB-delimited and it is organized in a header section and an alignment section.

Header lines are optional and start with '@'. They are formatted as 'tag:value' where 'tag' is a string explaining the content of 'value'. They contain informations about the version of the format, the reference assembly, the software used to perform the alignment and the type of aligned reads. Alignment lines have 11 mandatory fields and they always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The structure of the alignment lines is reported in Table 10.1

10.2 VCF Files

The VCF format is a text file format that contains meta-information lines, a header line, and data lines. Each data line contain information about a variant. The meta-information lines contain information on the format version and the creation of the file. They can also include specifications of entries that are included

Table 10.1: SAM file Alignment lines structure

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

in the data lines, as 'INFO' and 'FORMAT'. The header line has eight fixed, mandatory columns: '#CHROM', 'POS', 'ID', 'REF', 'ALT', 'QUAL', 'FILTER' and 'INFO'. After these columns a 'FORMAT' column is added if genotypes data are present in the file. The data lines contain for every variant (SNPs or INDEL), the informations specified in the header line columns. Custom information can be added in the 'INFO' and 'FORMAT' fields, but there are standard keywords to specify the most common kind of information (e.g. 'DP' for read depth or 'AF' for allele frequency).

Bibliography

[1] Ferree DC, Warrington IJ: *Apples. Botany, Production and Uses.* CAB International 2003.

[2] Sansavini S, Donati F, Costa F, Tartarini S: **Advances in apple breeding for enhanced fruit quality and resistance to biotic stresses: new varieties for the european market**. *Journal of Fruit and Ornamental Plant Research* 2004, **12**.

[3] Van Nerum I, Keulemans J: **Update on and Review of the Incompatibility (S -) Genotypes of Apple Cultivars**. *HortScience* 2004, **39**(5):943–947.

[4] Considine MJ, Wan Y, D'Antuono MF, Zhou Q, Han M, Gao H, Wang M: **Molecular genetic features of polyploidization and aneuploidization reveal unique patterns for genome duplication in diploid Malus.** *PloS one* 2012, **7**:e29449.

[5] Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagné D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett Ja, Proost S, Rouzé P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R: **The genome of the domesticated apple (Malus domestica Borkh.).** *Nature genetics* 2010, **42**(10):833–9.

[6] Potter D, Eriksson T, Evans RC, Oh S, Smedmark JEE, Morgan DR, Kerr M, Robertson KR, Arsenault M, Dickinson Ta, Campbell CS: **Phylogeny and classification of Rosaceae**. *Plant Systematics and Evolution* 2007, **266**(1-2):5–43.

[7] Campbell CS, Donoghue MJ, Baldwin BG, F WM: **Phylogenetic Relationships in Maloideae (Rosaceae) Evidence from sequences of the internal transcriped spacers.pdf**. *American journal of botany* 1995, **82**(7):903–918.

[8] Shulaev V, Korban SS, Sosinski B, Abbott AG, Aldwinckle HS, Folta KM, Iezzoni A, Main D, Arús P, Dandekar AM, Lewers K, Brown SK, Davis TM, Gardiner SE, Potter D, Veilleux RE: **Multiple models for Rosaceae genomics.** *Plant physiology* 2008, **147**(3):985–1003.

[9] Campbell CS, Evans RC, Morgan DR, Dickinson Ta, Arsenault MP: **Phylogeny of subtribe Pyrinae (formerly the Maloideae, Rosaceae): Limited resolution of a complex evolutionary history**. *Plant Systematics and Evolution* 2007, **266**(1-2):119–145.

[10] Wolfe J, Wehr W: **Rosaceous Chamaebatiaria-like foliage from the Paleogene of western North America**. *Aliso (USA)* 1988, **12**:177–200.

[11] Tuskan G, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao R, Bhalerao R, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**(5793)::1596.

[12] Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C: **Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution.** *The Plant cell* 2008, **20**:11–24.

[13] Doyle J, Flagel L, Paterson A, Rapp R, Soltis D, Soltis P, Wendel J: **Evolutionary genetics of genome merger and doubling in plants.** *Annual review of genetics* 2008, **42**(443):461.

[14] Janick J, Cummins J, Brown S, Hemmat M: *Apples. Fruit breeding.* 1996.

[15] Zohary D, Hopf M: *Domestication of Plants in the Old World.* Oxford University Press, third edition 2000.

[16] White KD: *Roman farming. (Aspects of Greek and Roman life).* Cornell un edition 1970.

[17] Nagy S, Shaw PE, Veldhuis MK: *Citrus Science and Technology.* Avi publis edition 1977.

[18] Ellison R, Renfrew J, Brothwell D, Seeley N: **Some food offerings from Ur, excavated by Sir Leonard Woolley, and previously unpublished**. *Journal of Archaeological Science* 1978, **5**(2):167–177.

[19] Korban S, Skirvin R: **Nomenclature of the cultivated apple**. *HortScience* 1984, **19**(2):177–180.

[20] Hokanson SC, McFerson JR, Forsline PL, Lamboy WF: **Collecting and Managing Wild Malus Germplasm in its Center of Diversity**. *HortScience* 1997, **32**(2):173–176.

[21] Robinson JP, Harris Sa, Juniper BE: **Taxonomy of the genus Malus Mill. (Rosaceae) with emphasis on the cultivated apple, Malus domestica Borkh.** *Plant Systematics and Evolution* 2001, **226**(1-2):35–58.

[22] Harris Sa, Robinson JP, Juniper BE: **Genetic clues to the origin of the apple.** *Trends in genetics : TIG* 2002, **18**(8):426–30.

[23] Morgan J, Richards A: *The New Book Of Apples*. Brogdale Horticultural Trust, Ebury Press 2002.

[24] Coart E, Van Glabeke S, De Loose M, Larsen aS, Roldán-Ruiz I: **Chloroplast diversity in the genus Malus: New insights into the relationship between the European wild apple (Malus sylvestris (L.) Mill.) and the domesticated apple (Malus domestica Borkh.)**. *Molecular Ecology* 2006, **15**:2171–2182.

[25] Cornille A, Giraud T, Smulders MJM, Roldán-Ruiz I, Gladieux P: **The domestication and evolutionary ecology of apples.** *Trends in genetics : TIG* 2013, :1–9.

[26] Cornille A, Gladieux P, Smulders MJM, Roldán-Ruiz I, Laurens F, Le Cam B, Nersesyan A, Clavel J, Olonova M, Feugey L, Gabrielyan I, Zhang XG, Tenaillon MI, Giraud T: **New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties.** *PLoS genetics* 2012, **8**(5):e1002703.

[27] Gupta PK, Roy JK, Prasad M: **Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants**. *Current Science* 2001, **80**(4):524–535.

[28] Ding C, Jin S: **High throughput methods for SNP genotyping**. *Methods in Molecular Biology* 2009, (578):245–25.

[29] Pushkarev D, Neff NF, Quake SR: **Single-molecule sequencing of an individual human genome.** *Nature biotechnology* 2009, **27**(9):847–850.

[30] Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Nordborg M, Ecker JR, Weigel D: **Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana**. *Science* 2007, **317**(July):338–342.

[31] Velasco R, Zharkikh A, Troggio M, Cartwright Da, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett Ja, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety.** *PloS one* 2007, **2**(12):e1326.

[32] Arai-Kichise Y, Shiwa Y, Nagasaki H, Ebana K, Yoshikawa H, Yano M, Wakasa K: **Discovery of Genome-wide DNA Polymorphisms in a Landrace Cultivar of Japonica Rice by Whole-genome Sequencing**. *Plant & cell physiology* 2011, **52**(2):274.

[33] Bräutigam A, Gowik U: **What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research.** *Plant biology* 2010, **12**(6):831–41.

[34] Elshire RJ, Glaubitz JC, Sun Q, Poland Ja, Kawamoto K, Buckler ES, Mitchell SE: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.** *PloS one* 2011, **6**(5):e19379.

[35] Davey JW, Hohenlohe Pa, Etter PD, Boone JQ, Catchen JM, Blaxter ML: **Genome-wide genetic marker discovery and genotyping using next-generation sequencing.** *Nature reviews. Genetics* 2011, **12**(7):499–510.

[36] Wood AR, Tuke MA, Nalls M, Hernandez D, Gibbs JR, Lin H, Xu CS, Li Q, Shen J, Jun G, Almeida M, Tanaka T, Perry JRB, Gaulton K, Rivas M, Pearson R, Curran JE, Johnson MP, Göring HHH, Duggirala R, Blangero J, Mccarthy MI, Bandinelli S, Weedon MN, Singleton A, Melzer D, Ferrucci L, Frayling TM: **Whole-genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes**. *Human molecular genetics* 2014, (November):1–9.

[37] Devlin B, Risch N: **A comparison of linkage disequilibrium measures for fine-scale mapping.** *Genomics* 1995, **29**:311–322.

[38] Lewontin RC: **The interactions of selection and linkage II. Optimus models**. *Genetics* 1964, **50(4)**(1956):49–67.

[39] Weir BS: *Genetic data analysis II: methods for discrete population genetic data.* 1996.

[40] Hill WG, Robertson a: **Linkage disequilibrium in finite populations.** *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 1968, **38**:226–31.

[41] Fisher R: **The Logic of Inductive Inference**. *Journal of the Royal Statistical Society* 1934, **98**:39–82.

[42] Flint-garcia SA, Thornsberry JM, Buckler ESI: **STRUCTURE OF LINKAGE DISEQUILIBRIUM IN PLANTS**. *Annu. Rev. Plant Biol.* 2003, **54**:357–374.

[43] Stich B, Melchinger AE, Piepho HP, Heckenberger M, Maurer HP, Reif JC: **A new test for family-based association mapping with inbred lines from plant breeding programs.** *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 2006, **113**:1121–1130.

[44] Yu J, Buckler ES: **Genetic association mapping and genome organization of maize**. *Current Opinion in Biotechnology* 2006, **17**:155–160.

[45] Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, Tingey S, Rafalski A: **Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize**. *Molecular Genetics and Genomics* 2008, **279**:1–10.

[46] Ravel C, Praud S, Murigneux A, Canaguier A, Sapet F, Samson D, Balfourier F, Dufour P, Chalhoub B, Brunel D, Beckert M, Charmet G: **Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (Triticum aestivum L.)**. *Genome* 2006, **49**:1131–1139.

[47] Rhoné B, Raquin AL, Goldringer I: **Strong linkage disequilibrium near the selected Yr17 resistance gene in a wheat experimental population**. *Theoretical and Applied Genetics* 2007, **114**:787–802.

[48] Tommasini L, Schnurbusch T, Fossati D, Mascher F, Keller B: **Association mapping of Stagonospora nodorum blotch resistance in modern European winter wheat varieties**. *Theoretical and Applied Genetics* 2007, **115**:697–708.

[49] Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES: **Association mapping: critical considerations shift from genotyping to experimental design.** *The Plant cell* 2009, **21**(8):2194–202.

[50] Mackay I, Powell W: **Methods for linkage disequilibrium mapping in crops**. *Trends in Plant Science* 2007, **12**(2):57–63.

[51] Nielsen R: **Molecular signatures of natural selection.** *Annual review of genetics* 2005, **39**:197–218.

[52] Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG: **Recent and ongoing selection in the human genome.** *Nature reviews. Genetics* 2007, **8**(11):857–68.

[53] Maynard Smith J, Haigh J: **The hitch-hiking effect of a favourable gene**. *Genetical Research* 1974, **23**:23–35.

[54] Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD: **Selection under domestication: evidence for a sweep in the rice waxy genomic region.** *Genetics* 2006, **173**(2):975–83.

[55] Tajima F: **Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism**. *Genetics* 1989, **595**(3):585–595.

[56] Fu YX, Li Wh: **Statistical Tests of Neutrality of Mutations**. *Genetics* 1992, (133):693–709.

[57] Fay JC, Wu Ci: **Hitchhiking Under Positive Darwinian Selection**. *Genetics* 2000, (155):1405–1413.

[58] Kim Y, Stephan W: **Detecting a local signature of genetic hitchhiking along a recombining chromosome.** *Genetics* 2002, **160**(2):765–77.

[59] Beaumont Ma, Nichols Ra: **Evaluating Loci for Use in the Genetic Analysis of Population Structure**. *Proceedings of the Royal Society B: Biological Sciences* 1996, **263**:1619–1626.

[60] Majewski J, Cohan FM: **Adapt globally, act locally: The effect of selective sweeps on bacterial sequence diversity**. *Genetics* 1999, **152**(January):1459–1474.

[61] Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a High-Density SNP Map for Signatures of Natural Selection**. *Genome research* 2002, **12**:1805–1814.

[62] Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, Mcdonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: **Detecting recent positive selection in the human genome from haplotype structure**. *Nature* 2002, **419**(October):832–837.

[63] Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS biology* 2006, **4**(3):e72.

[64] Wang ET, Kodama G, Baldi P, Moyzis RK: **Global landscape of recent inferred Darwinian selection for Homo sapiens.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:135–40.

[65] Kim Y, Nielsen R: **Linkage disequilibrium as a signature of selective sweeps.** *Genetics* 2004, **167**(3):1513–24.

[66] Homer N, Merriman B, Nelson SF: **BFAST: an alignment tool for large scale genome resequencing.** *PloS one* 2009, **4**(11):e7767.

[67] Smith T, Waterman M: **Identification of common molecular subsequences**. *Journal of Molecular Biology* 1981, **147**:195–197.

[68] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics (Oxford, England)* 2009, **25**(16):2078–9.

[69] Danecek P, Auton A, Abecasis G, Albers Ca, Banks E, DePristo Ma, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R: **The variant call format and VCFtools.** *Bioinformatics (Oxford, England)* 2011, **27**(15):2156–8.

[70] Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M, Salvi S, Jansen J, Viola R, Gut I, Laurens F, Chagné D, Velasco R, van de Weg E, Troggio M: **Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (Malus domestica Borkh).** *PloS one* 2014, **9**(10):e110377.

[71] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MaR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *American journal of human genetics* 2007, **81**(3):559–75.

[72] Raj A, Stephens M, Pritchard JK: **fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets.** *Genetics* 2014, **197**(2):573–589.

[73] Jakobsson M, Rosenberg Na: **CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure.** *Bioinformatics (Oxford, England)* 2007, **23**(14):1801–6.

[74] Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proceedings of the National Academy of Sciences* 1979, **76**(10):5269–5273.

[75] Wright S: **Genetical structure of populations**. *Nature* 1950, **4215**:247–249.

[76] Schaid DJ: **Linkage Disequilibrium Testing when Linkage Phase is Unknown**. *Genetics* 2004, **166**(January):505–512.

[77] Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N: **SweeD: likelihood-based detection of selective sweeps in thousands of genomes.** *Molecular biology and evolution* 2013, **30**(9):2224–34.

[78] Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data.** *Genome research* 2005, **15**(11):1566–75.

[79] Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215**(3):403–410.

[80] Baird Na, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis Za, Selker EU, Cresko Wa, Johnson Ea: **Rapid SNP discovery and genetic mapping using sequenced RAD markers**. *PLoS ONE* 2008, **3**(10):1–7.

[81] Pfender WF, Saha MC, Johnson Ea, Slabaugh MB: **Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in Lolium perenne**. *Theoretical and Applied Genetics* 2011, **122**:1467–1480.

[82] Ward Ja, Bhangoo J, Fernández-Fernández F, Moore P, Swanson JD, Viola R, Velasco R, Bassil N, Weber Ca, Sargent DJ: **Saturated linkage map construction in Rubus idaeus using genotyping by sequencing and genome-independent imputation**. *BMC genomics* 2013, **14**:2.

[83] Liu H, Bayer M, Druka A, Russell JR, Hackett Ca, Poland J, Ramsay L, Hedley PE, Waugh R: **An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley.** *BMC genomics* 2014, **15**:104.

[84] Poland Ja, Brown PJ, Sorrells ME, Jannink JL: **Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach**. *PLoS ONE* 2012, **7**(2).

[85] Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistué L, Corey A, Filichkina T, Johnson Ea, Hayes PM: **Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley.** *BMC genomics* 2011, **12**:4.

[86] Yang H, Tao Y, Zheng Z, Zhang Q, Zhou G, Sweetingham MW, Howieson JG, Li C: **Draft Genome Sequence, and a Sequence-Defined Genetic Linkage Map of the Legume Crop Species Lupinus angustifolius L**. *PLoS ONE* 2013, **8**(5).

[87] Arnold B, Corbett-Detig RB, Hartl D, Bomblies K: **RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling**. *Molecular Ecology* 2013, **22**:3179–3190.

[88] Pina A, Urrestarazu J, Errea P: **Analysis of the genetic diversity of local apple cultivars from mountainous areas from Aragon (Northeastern Spain)**. *Scientia Horticulturae* 2014, **174**:1–9.

[89] Liang W, Dondini L, De Franceschi P, Paris R, Sansavini S, Tartarini S: **Genetic Diversity , Population Structure and Construction of a Core Collection of Apple Cultivars from Italian Germplasm**. *Plant Molecular Biology Reporter* 2014, :1–16.

[90] Glaszmann JC, Kilian B, Upadhyaya HD, Varshney RK: **Accessing genetic diversity for crop improvement**. *Current Opinion in Plant Biology* 2010, **13**:167–173.

[91] Jones E, Chu WC, Ayele M, Ho J, Bruggeman E, Yourstone K, Rafalski A, Smith OS, McMullen MD, Bezawada C, Warren J, Babayev J, Basu S, Smith S: **Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (Zea mays L.) germplasm**. *Molecular Breeding* 2009, **24**:165–176.

[92] Yamamoto T, Nagasaki H, Yonemaru Ji, Ebana K, Nakajima M, Shibaya T, Yano M: **Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms.** *BMC genomics* 2010, **11**:267.

[93] Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB: **Nucleotide diversity and linkage disequilibrium in loblolly pine.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(42):15255–60.

[94] Gilchrist EJ, Haughn GW, Ying CC, Otto SP, Zhuang J, Cheung D, Hamberger B, Aboutorabi F, Kalynyak T, Johnson L, Bohlmann J, Ellis BE, Douglas CJ, Cronk QCB: **Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of Populus trichocarpa**. *Molecular Ecology* 2006, **15**:1367–1378.

[95] Jiang D, Ye QL, Wang FS, Cao L: **The Mining of Citrus EST-SNP and Its Application in Cultivar Discrimination**. *Agricultural Sciences in China* 2010, **9**(2):179–190.

[96] Wu SB, Wirthensohn MG, Hunt P, Gibson JP, Sedgley M: **High resolution melting analysis of almond SNPs derived from ESTs**. *Theoretical and Applied Genetics* 2008, **118**:1–14.

[97] Lima LS, Gramacho KP, Carels N, Novais R, Gaiotto Fa, Lopes UV, Gesteira aS, Zaidan Ha, Cascardo JCM, Pires JL, Micheli F: **Single nucleotide polymorphisms from Theobroma cacao expressed sequence tags associated with witches' broom disease in cacao**. *Genetics and Molecular Research* 2009, **8**(3):799–808.

[98] Vezzulli S, Micheletti D, Riaz S, Pindo M, Viola R, This P, Walker MA, Troggio M, Velasco R: **A SNP transferability survey within the genus Vitis.** *BMC plant biology* 2008, **8**:128.

[99] Lijavetzky D, Cabezas JA, Ibáñez A, Rodríguez V, Martínez-Zapater JM: **High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology.** *BMC genomics* 2007, **8**:424.

[100] Chagné D, Gasic K, Crowhurst RN, Han Y, Bassett HC, Bowatte DR, Lawrence TJ, Rikkerink EHa, Gardiner SE, Korban SS: **Development of a set of SNP markers present in expressed genes of the apple**. *Genomics* 2008, **92**:353–358.

[101] Han Y, Chagné D, Gasic K, Rikkerink EHa, Beever JE, Gardiner SE, Korban SS: **BAC-end sequence-based SNPs and Bin mapping for rapid integration of physical and genetic maps in apple**. *Genomics* 2009, **93**(3):282–288.

[102] **A map of human genome variation from population-scale sequencing** 2010.

[103] Tenaillon MI, Sawkins MC, Long aD, Gaut RL, Doebley JF, Gaut BS: **Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.).** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(16):9161–9166.

[104] Kota R, Varshney RK, Prasad M, Zhang H, Stein N, Graner a: **EST-derived single nucleotide polymorphism markers for assembling genetic and physical maps of the barley genome**. *Functional and Integrative Genomics* 2008, **8**:223–233.

[105] Zhu Q, Zheng X, Luo J, Gaut BS, Ge S: **Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: Severe bottleneck during domestication of rice**. *Molecular Biology and Evolution* 2007, **24**:875–888.

[106] Cao K, Zheng Z, Wang L, Liu X, Zhu G, Fang W, Cheng S, Zeng P, Chen C, Wang X, Xie M, Zhong X, Wang X, Zhao P, Bian C, Zhu Y, Zhang J, Ma G, Chen C, Li Y, Hao F, Li Y, Huang G, Li Y, Li H, Guo J, Xu X, Wang J: **Comparative population genomics reveals the domestication history of the peach, Prunus persica, and human influences on perennial fruit crops**. *Genome Biology* 2014, **15**(7):415.

[107] Crisci JL, Poh YP, Mahajan S, Jensen JD: **The impact of equilibrium assumptions on tests of selection.** *Frontiers in genetics* 2013, **4**(November):235.

[108] Alachiotis N, Stamatakis a, Pavlidis P: **OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets.** *Bioinformatics (Oxford, England)* 2012, **28**(17):2274–5.

[109] Lyu J, Li B, He W, Zhang S, Gou Z, Zhang J, Meng L, Li X, Tao D, Huang W, Hu F, Wang W: **A genomic perspective on the important genetic mechanisms of upland adaptation of rice.** *BMC plant biology* 2014, **14**:160.

[110] Gore Ma, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer Ja, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES: **A first-generation haplotype map of maize.** *Science (New York, N.Y.)* 2009, **326**(5956):1115–7.

[111] Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, Wang B, Liu Z, Chen J, Li W, Zhang M, Xie S, Lai J: **Genome-wide genetic changes during modern breeding of maize.** *Nature genetics* 2012, **44**(7):812–5.

[112] Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B: **A map of rice genome variation reveals the origin of cultivated rice.** *Nature* 2012, **490**(7421):497–501.

[113] Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S: **Population genomic and genome-wide association studies of agroclimatic traits in sorghum.** *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(2):453–8.

[114] Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, Zuccolo A, Song X, Kudrna D, Ammiraju JSS, Cossu RM, Maldonado C, Chen J, Lee S, Sisneros N, de Baynast K, Golser W, Wissotski M, Kim W, Sanchez P, Ndjiondjop MN, Sanni K, Long M, Carney J, Panaud O, Wicker T, Machado Ca, Chen M, Mayer KFX, Rounsley S, Wing Ra: **The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication**. *Nature Genetics* 2014, **46**(9):982–988.

[115] Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick Ma, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MMS, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea Ca, Wang M, Yu Y, Zhang M, Wing Ra, Cregan PB, Rokhsar DS, Jackson Sa: **A reference genome for common bean and genome-wide analysis of dual domestications.** *Nature genetics* 2014, **46**(7):707–713.

[116] Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen JG, Tuskan Ga, DiFazio SP: **Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations**. *Nature Genetics* 2014, **46**(10):1089–1096.

[117] Nimmakayala P, Levi A, Abburi L, Abburi VL, Tomason YR: **Single nucleotide polymorphisms generated by genotyping by sequencing to characterize genome-wide diversity , linkage disequilibrium , and selective sweeps in cultivated watermelon**. *BMC genomics* 2014, **15**(767).

[118] Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Städler T, Li J, Ye Z, Du Y, Huang S: **Genomic analyses provide insights into the history of tomato breeding.** *Nature genetics* 2014, **46**(11):1220–1226.

[119] Devoghalaere F, Doucen T, Guitton B, Keeling J, Payne W, Ling T, Ross J, Hallett I, Gunaseelan K, Dayatilake G, Diak R, Breen KC, Tustin DS, Costes E, Chagné D, Schaffer R, David K: **A genomics approach to understanding the role of auxin in apple (Malus x domestica) fruit size control**. *BMC Plant Biology* 2012, **12**:7.

[120] Khan MA, Olsen KM, Sovero V, Kushad MM, Korban SS: **Fruit Quality Traits Have Played Critical Roles in Domestication of the Apple**. *The Plant Genome* 2014, **7**(3).

[121] Bergamini A, Faedi W: *Monografia di cultivar di melo*. Roma: Ministero dell'Agricoltura e delle Foreste, Direzione generale della Produzione Agricola, Istituto Sperimentale per la Frutticoltura 1984.