Whole-genome analysis of diversity and SNP-major gene association in peach germplasm

Diego Micheletti, Maria Teresa Dettori, Sabrina Micali, Valeria Aramini, Igor Pacheco, Cassia da Silva Linge, Stefano Foschi, Elisa Banchi, Teresa Barreneche, Bénédicte Quilot-Turion, et al.

PLOS | ONE

RESEARCH ARTICLE

Whole-Genome Analysis of Diversity and SNP-Major Gene Association in Peach Germplasm

Diego Micheletti[1¤a], Maria Teresa Dettori[2], Sabrina Micali[2], Valeria Aramini[2], Igor Pacheco[3¤b], Cassia Da Silva Linge[3], Stefano Foschi[4], Elisa Banchi[5], Teresa Barreneche[6,7], Bénédicte Quilot-Turion[8], Patrick Lambert[8], Thierry Pascal[8], Ignasi Iglesias[9], Joaquim Carbó[10], Li-rong Wang[11], Rui-juan Ma[12], Xiong-wei Li[13], Zhong-shan Gao[13], Nelson Nazzicari[14], Michela Troggio[5], Daniele Bassi[3], Laura Rossini[3,14], Ignazio Verde[2], François Laurens[15,16,17], Pere Arús[1], Maria José Aranzana[1]*

1 IRTA, Centre de Recerca en Agrigenòmica CSIC-IRTA-UAB-UB, Campus UAB, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain, **2** Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CRA), Centro di Ricerca per la Frutticoltura, Roma, Italy, **3** Università degli Studi di Milano, DiSAA, Milan, Italy, **4** Centro Ricerche Produzioni Vegetali, Cesena (FC), Italy, **5** Research and Innovation Centre, Fondazione Edmund Mach (FEM), San Michele all'Adige (TN), Italy, **6** INRA, UMR 1332 de Biologie du Fruit et Pathologie, Villenave d'Ornon, France, **7** Univ. Bordeaux, UMR 1332 de Biologie du Fruit et Pathologie, Villenave d'Ornon, France, **8** INRA UR1052 GAFL, Domaine Saint Maurice, Montfavet, France, **9** IRTA, Estació Experimental de Lleida, Parc de Gardeny, Edifici Fruitcentre, Lleida, Spain, **10** IRTA, Estacio Experimental Mas Badia, La Tallada d'Empordà, Girona, Spain, **11** Zhenzhou Fruit Research Institute, CAAS, Zhengzhou, China, **12** Horticultural Institute, Jiangsu Academy of Agricultural Sciences, Nanjing, China, **13** Department of Horticulture, Zhejiang University, Hangzhou, China, **14** Parco Tecnologico Padano, Lodi, Italy, **15** INRA, UMR 1345 Institut de Recherche en Horticulture et Semences, Beaucouzé, France, **16** Université d'Angers, UMR 1345 Institut de Recherche en Horticulture et Semences, SFR 4207 QUASAV, PRES L'UNAM, Angers, France, **17** AgroCampus-Ouest, UMR 1345 Institut de Recherche en Horticulture et Semences, Angers, France

¤a Current address: Research and Innovation Centre, Fondazione Edmund Mach (FEM), San Michele all'Adige (TN), Italy
¤b Current address: INTA, Universidad de Chile, Santiago de Chile, Chile
* mariajose.aranzana@irta.cat

Citation: Micheletti D, Dettori MT, Micali S, Aramini V, Pacheco I, Da Silva Linge C, et al. (2015) Whole-Genome Analysis of Diversity and SNP-Major Gene Association in Peach Germplasm. PLoS ONE 10(9): e0136803. doi:10.1371/journal.pone.0136803

Editor: Yuepeng Han, Wuhan Botanical Garden of Chinese Academy of Sciences, CHINA

Received: March 14, 2015

Accepted: August 7, 2015

Published: September 9, 2015

Data Availability Statement: Genotypic data are available from the GDR database (http://www.rosaceae.org/) under the accession number tfGDR1013 and in the FruitBreedomics database (http://bioinformatics.tecnoparco.org/fruitbreedomics/).

Funding: This work has been funded by the EU seventh Framework Programme (http://ec.europa.eu/research/fp7/index_en.cfm) through the project "FruitBreedomics: Integrated Approach for increasing breeding efficiency in fruit tree crops" (Grant #FP7-265582; http://fruitbreedomics.com/); by the Ministero delle Politiche Agricole Alimentari e Forestali -Italy

Abstract

Peach was domesticated in China more than four millennia ago and from there it spread world-wide. Since the middle of the last century, peach breeding programs have been very dynamic generating hundreds of new commercial varieties, however, in most cases such varieties derive from a limited collection of parental lines (founders). This is one reason for the observed low levels of variability of the commercial gene pool, implying that knowledge of the extent and distribution of genetic variability in peach is critical to allow the choice of adequate parents to confer enhanced productivity, adaptation and quality to improved varieties. With this aim we genotyped 1,580 peach accessions (including a few closely related *Prunus* species) maintained and phenotyped in five germplasm collections (four European and one Chinese) with the International Peach SNP Consortium 9K SNP peach array. The study of population structure revealed the subdivision of the panel in three main populations, one mainly made up of Occidental varieties from breeding programs ($POP1OC_B$), one of Occidental landraces ($POP2OC_T$) and the third of Oriental accessions (POP3OR).

(MiPAAF, http://www.politicheagricole.it) through the project "DRUPOMICS: Sequenziamento del genoma del pesco ed utilizzo della sequenza in programmi di miglioramento della qualita` del frutto del pesco e della resistenza alle malattie" (Grant # DM14999/7303/08) and "ESPLORA: Esplorazione della biodiversità vegetale ed animale alla ricerca di alleli superiori da inserire nei programmi avanzati di miglioramento genetico a sostegno dell'agricoltura nazionale" (Grant #DM 14658/7303/10); by the Spanish Ministry of Science and Innovation (http://www.micinn.es/) through the project AGL2012-40228-C02-01 (uso de la secuencia genomica para la caracterizacion de la variabilidad intraclonal e interespecifica en Melocotonero y almendro) and by the INIA (http://www.inia.es/) through the Project RF2012-00024-C04-04 ("conservación y caracterización de germoplasma introducido"). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Analysis of linkage disequilibrium (LD) identified differential patterns of genome-wide LD blocks in each of the populations. Phenotypic data for seven monogenic traits were integrated in a genome-wide association study (GWAS). The significantly associated SNPs were always in the regions predicted by linkage analysis, forming haplotypes of markers. These diagnostic haplotypes could be used for marker-assisted selection (MAS) in modern breeding programs.

Introduction

Since its domestication in China more than 5,000 years ago, the genetic variability of cultivated peach, *Prunus persica* (L.) Batsch, has been shaped by three major forces [1]. The first is inbreeding, favored by the self-pollinating system of peach, an unusual situation compared to most other *Prunus* species that have an efficient gametophytic self-incompatibility system. Selfing, plus the directional selection of the domestication and the selective processes of plant breeding, have resulted in a dramatic loss of variability when compared to other cultivated species of the genus [2, 3]. The second force is random drift, as a consequence of the small number of parents used in the breeding programs that started in the US in the middle of the 20th century [4, 5]. A limited number of founders may also have been used in Asian breeding programs, as inferred from the work of Li et al [6]. The third force is heterosis that, because peach is easily propagated by grafting, can be exploited in commercial breeding. The usual breeding scheme of peach and most fruit trees is the selection of individuals arising from the first generation of crosses between two accessions that are themselves, and therefore their offspring, partly heterozygous. The consequence of all this is that cultivated peach has a low level of genetic variability, high linkage disequilibrium conservation and, in most modern cultivars, a relatively high level of heterozygosity. This has been demonstrated using sets of SSR markers with good genome coverage [6,7], and, more recently, using resequencing data [8,9].

Due to its economic value (second, after apple, among temperate fruits) and its self-pollinating behavior, peach has been one of the tree species most amenable to genetic studies [10, 11], becoming a model system for genomics research in the Rosaceae. In particular, the genomes of different *Prunus* species are highly conserved [12], allowing many major genes and QTLs from peach and other *Prunus* to be positioned on a single genetic map [13], using a set of high quality and comparable linkage maps, most developed at the end of the last century [14]. Recently, the whole genome sequence of peach has been released, including the resequencing data of several peach accessions [8], opening a new era for the development of genetic studies and their application to modern breeding in this species. It has been possible to estimate the SNP variability of this species [8, 9, 15, 16] and a 9k SNP array v1 [17] has been developed by the International Peach SNP Consortium (IPSC). This SNP array is currently being used for genetic analysis [18–22].

The high abundance and relatively low cost of SNPs compared to SSRs allows for a complementary and deeper study of peach germplasm variability. In this paper, we used the IPSC 9K SNP array [17] to genotype a large collection of peach accessions including an ample set of Occidental and Oriental materials. Our results confirm previous findings on variability parameters and subpopulation structure, provide a finer analysis of linkage disequilibrium (LD), and allow the whole-genome association analysis of a set of morphological characters with Mendelian inheritance.

Results

IPSC 9K SNP array performance

We genotyped 1,580 *Prunus* accessions (1,576 *P. persica* plus four *Prunus*-related species) using the IPSC peach 9K SNP array v1 [17]. The average call rate (number of calls divided by the number of assayed SNPs) was 0.839 and, after exclusion of 40 samples with the poorest quality (call rate lower than the average call rate minus 0.1), the mean call rate increased to 0.854. The list of varieties with genotypic data is presented in S1 Table.

The 8,144 SNPs assayed were divided into five classes from A to E (see methods) depending on their performance. The SNPs classified as A (no-Call < 5% and all three possible genotypes, AA, AB, and BB), represented 53.2% of the total (4,330). After removing those with minor allele frequency (MAF) lower than 0.05 and those with poor clustering in GenomeStudio Data Analysis software (Illumina Inc.), the final set of class A SNPs was 4,271 (52.4%). The percentage of polymorphic SNPs increased to 65.5% when the SNPs in class B (null allele or preferential annealing) and C (duplicated sequences/genes) were included (Table 1). Only the 4,271 class A SNPs were used for subsequent analysis, corresponding to an average of 18.6 SNPs/Mb (considering the total peach genome size of 230 Mbp) and approximately equivalent to 8.2 SNPs/cM of the reference *Prunus* map (519 cM as in Dirlewanger et al. [12]).

Germplasm characterization

We compared all pairs of accessions to exclude known and unknown sport mutations or synonyms from the analysis, and detected 173 groups encompassing 473 cultivars with two or more cultivars with identical genotype (more than 98% of their SNP genotypes identical). Consequently, 300 cultivars were removed, retaining only one accession for each unique genotype. The largest group of clones contained 11 genotypes, two had nine, eight, seven and six genotypes, respectively, and the rest were distributed in four groups of five, 17 of four, 28 of three and 115 of two genotypes. Most (76 out of 83) of the samples with the same name had the same genotype, including all those introduced as controls ('Rich Lady' and 'Diamond Ray' with five replicates each, and 'Maycrest', 'Big Top', 'Elegant Lady' and 'Jing Yu' with four replicates, each from a different germplasm bank). Five pairs of accessions with the same name from two different germplasm collections ('Akatsuki' 'Siberian C', 'Suncrest', 'Fei Cheng Bali' and, 'Wu Yue Xian'), expected to be identical, were different. Additionally one replicate out of three of 'Armking' and one out of four of 'Rubirich' didn't present the expected genotype.

Some known sports were included in the panel. In all cases they grouped together. The largest group of identical genotypes included all known sports of 'Springcrest': 'Maycrest' (four replicates, each from a different repository), 'Early Maycrest', 'Queencrest', 'Springbelle', 'Springold' and 'Springlady', plus two cultivars of unknown origin, 'Harken' and 'San Giovanni'. Their genotypes were identical for all markers analyzed. Another group included

Table 1. Number of SNPs per class of the IPSC 9K array in a collection of 1,580 *Prunus* accessions.

Class	Class description	# SNPs
A	No Call <5%, three genotypes	4,330
B	Null allele or preferential annealing	230
C	Duplicated sequences/genes	778
D	False SNPs	772
E	Failed	2,034
Total		8,144

doi:10.1371/journal.pone.0136803.t001

'O'Henry' and its sports 'John Henry' and 'Summer Lady'. In this case, differences were identified in 11 SNPs, with 'O'Henry' differing from 'John Henry' by nine SNPs and 11 from 'Summer Lady', with 'John Henry' and 'Summer Lady' only differing by one SNP. All these SNPs were located at a region of 5.4 Mb at the top of chromosome 6 (between markers SNP_IGA_604703 and SNP_IGA_621593), where there was a large amount of missing data for many other SNPs in the derived sports 'John Henry' (41 missing data out of 92 SNPs) and 'Summer Lady' (38 out of 92). A similar pattern was observed for the same genomic region in 'Redskin' and its sport 'Redkist' (13 differences; 44 missing data out of 92 SNPs in 'Red Kist') and in the two groups of cultivars classified as having the same genotype and of unknown relationship: 'Vittorio Emanuele' and 'Paola Matteucci' (differing by 8 SNPs in the same region and with 'Paola Matteucci' having 60 missing data out of 92 SNPs) and four Italian accessions with different names 'Pieri 81', 'Bella Lucia', 'Vaccaro Roccalmunto' and 'Zingara Nera', where the former three were identical and the latter differed by four SNPs in this region and had 46 out of 92 SNPs with missing data. In all cases, the differences in SNPs occurred at loci that were heterozygous in the complete accessions and homozygous in those having missing data. A similar case but involving a different region was observed between cultivars 'Springtime' and 'Starcrest'. These were identical in all markers except for eight SNPs located in the upper part of chromosome 4, within a region of 17.9 Mb where there was also a large amount of missing data (179 in 'Springtime', 231 in 'Starcrest', 178 of them in common with the total 581 markers in the region). As in the previous cases, all distinctive SNPs were homozygous in one variety ('Starcrest') and heterozygous in the other.

Cultivar variability and population structure

A total of 1,240 unique accessions were genotyped with the 4,271 class A SNPs. The observed mean heterozygosity (H_o) per individual was 0.286, ranging from 0.003 in 'Burrona di Rosano' to 0.680 in 'Zheng huang 4'. H_o per SNP ranged from 0.046 in SNP_IGA_762094 to 0.781 in SNP_IGA_550718. Chromosome 6 (Ho = 0.28) was the least heterozygous while chromosome 1 had more loci in heterozygosis (Ho = 0.33). The mean expected heterozygosity (H_e) was 0.39, ranging from 0.055 to 0.500. The mean average inbreeding coefficient ($F = (H_o—H_e) / H_o$) was 0.27, ranging from -0.705 to 0.9922. The vast majority of the markers (98.5%) showed significant ($p < 0.001$) deviation from the Hardy-Weinberg Equilibrium (HWE).

A phylogenetic dendrogram with the 1,240 unique accessions (Fig 1) clearly revealed two main groups, one with Oriental accessions and the other including those from Occidental origin. Two clusters could be distinguished within the Occidental group, where the traditional/non-breeding accessions were differentiated from those derived from breeding programs. The position of each accession in the dendrogram is provided in the S1 Table.

A first approximation to the study of population structure was obtained using principal component analysis (PCA) for the complete set of SNPs, which separated the Oriental and Occidental accessions (Fig 2). The Occidental accessions were clearly separated into two major groups, one including the traditional/non-breeding accessions and the other with varieties used and obtained in modern breeding programs. The separation in three clusters was not absolute and a discrete number of accessions occupied a centric position. The majority of these accessions are known to come from crosses between accessions in different clusters.

The software STRUCTURE [23] was used to obtain a more detailed picture of the stratification in the panel. In agreement with the results of the PCA, the most probable number of subpopulations inferred by this method (K) was three. Considering only the accessions with a subpopulation membership coefficient higher than 0.8, the three subpopulations are: population 1, the Occidental materials used in modern breeding programs ($P1OC_{B,}$ with 352

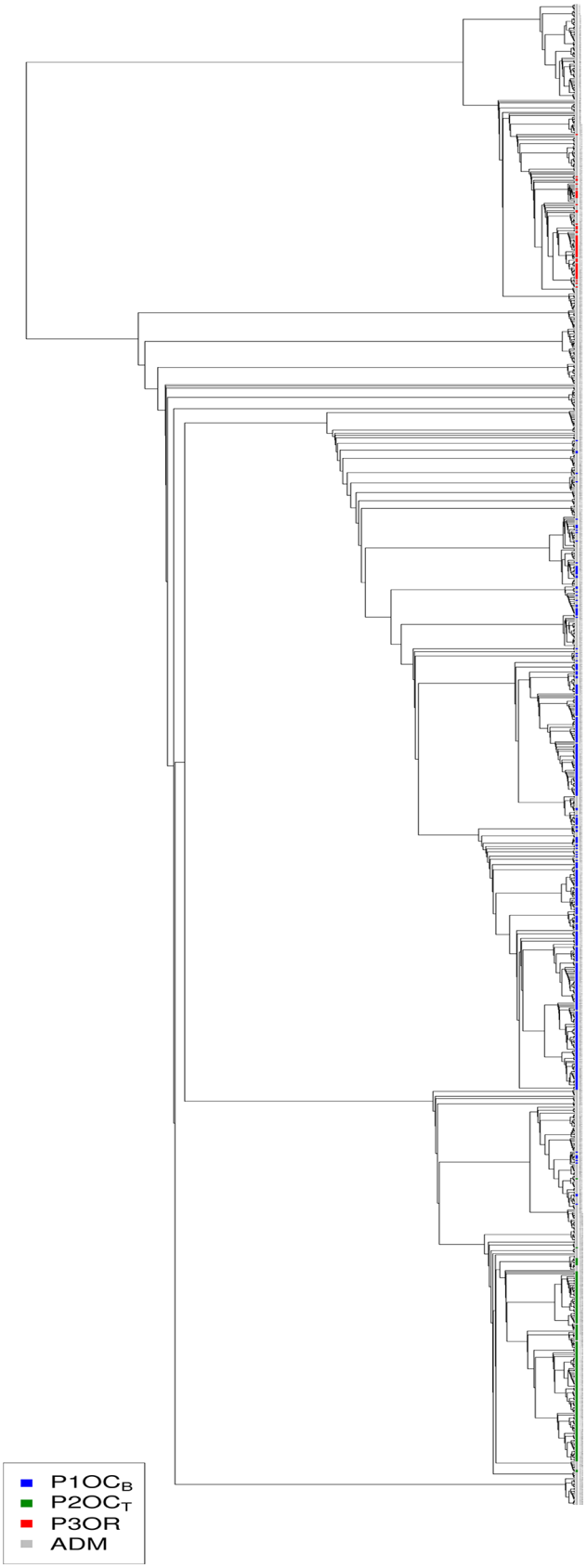
P1OC$_B$
P2OC$_T$
P3OR
ADM

Fig 1. UPGMA tree for the 1,240 accessions. Colors indicate the assignment into populations obtained by STRUCTURE analysis at K = 3. Blue: Occidental/Breeding ($P1OC_B$); Green: Occidental/Traditional ($P2OC_T$); Red: Oriental (P3OR); Gray: Admixture (ADM). Accession names have been replaced by numbers; correspondence between number and accession is provided in S1 Table.

doi:10.1371/journal.pone.0136803.g001

accessions), population 2, the traditional/non-breeding occidental accessions ($P2OC_T$, 165 accessions) and population 3, the oriental accessions (P3OR, 58 accessions). At K = 2, the division between the Oriental and Occidental cultivars was already evident. Increasing the number of subpopulations to four the Occidental breeding accessions were divided into two subgroups: the first comprising the vast majority of nectarines and the second the vast majority of peaches. Further increments of K maintained these four subpopulations, with additional ones empty or

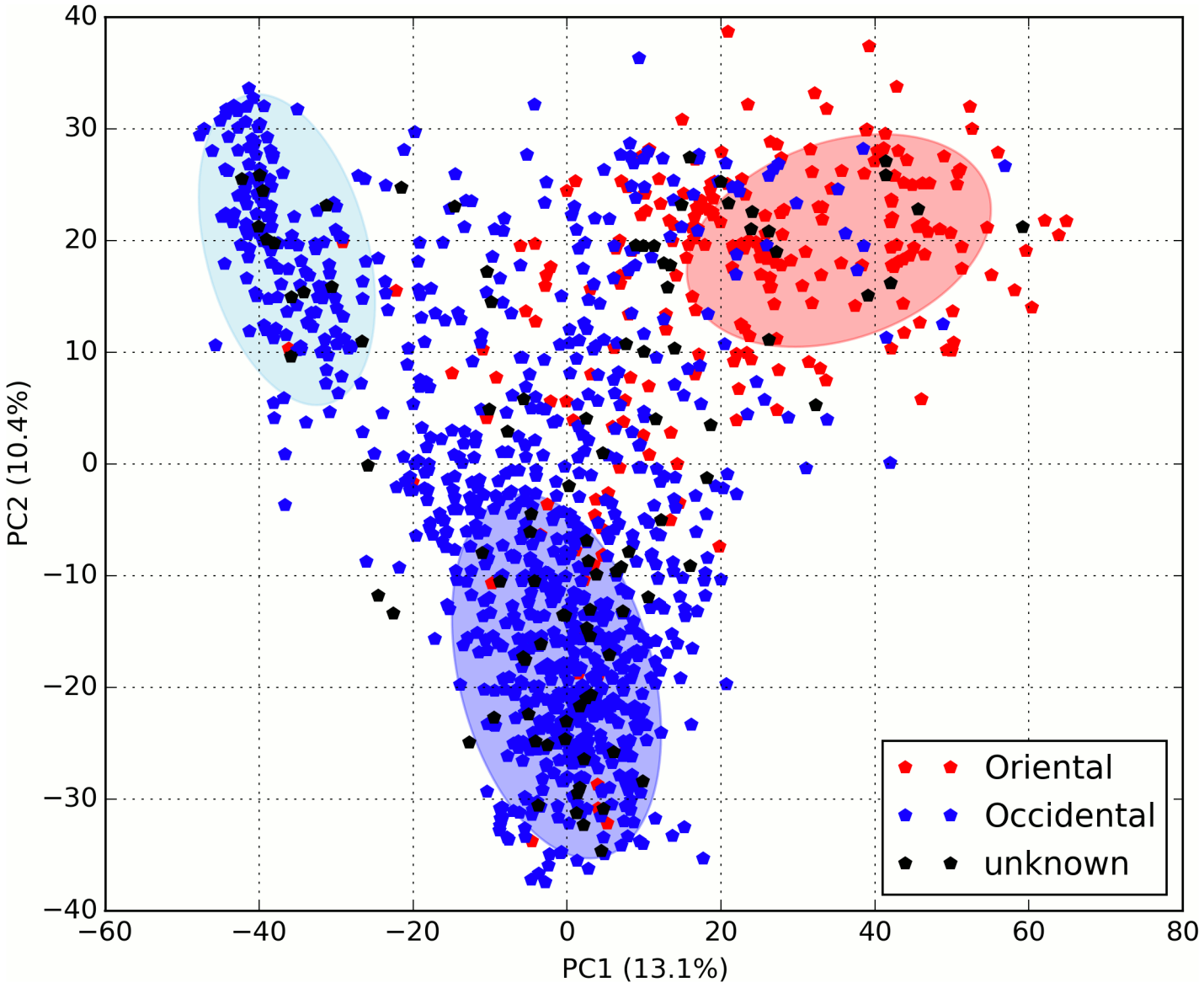

Fig 2. Principal component analysis of the 4,271 positively genotyped SNPs in the 1,240 unique peach accessions. Occidental and Oriental accessions are represented with blue and red colors, respectively. Black symbols indicate accessions with unknown origin.

doi:10.1371/journal.pone.0136803.g002

including only a few individuals with a membership coefficient ≥ 0.8. At all K values, more than half of the individuals (665) were not assigned to a single population and were considered as admixed (ADM) (S1 Fig). The position in the UPGMA tree of the accessions assigned into each subpopulation at K = 3 is presented in Fig 1, where accessions belonging to $P1OC_B$, $P2OC_T$ and P3OR are colored in blue, green and red, respectively. Gray color indicates accessions in the ADM subpopulation.

Distance between populations was calculated through the Fst statistic. Genetic differentiation between $P2OC_T$ and P3OR and between $P1OC_B$ and P3OR were moderate (Fst = 0.14 and Fst = 0.13, respectively), while $P1OC_B$ and $P2OC_T$ showed a greater differentiation (Fst = 0.18)

Linkage disequilibrium (LD)

The LD, measured as r^2, was calculated in the three subpopulations and in the admixed accessions separately, as well as in all 1,240 accessions together using a modified r^2 that corrects for population structure and relatedness [24].

Analyzing the SNP data in each subpopulation, 188, 949 and 353 out of the 4,271 SNPs in $P1OC_B$, $P2OC_T$ and P3OR, respectively, failed the frequency test (MAF < 0.01) and/or the non-calling cut-off of 5% and were discarded. The average value of intra-chromosomal r^2 was 0.096 in $P1OC_B$ and $P2OC_T$, 0.133 in P3OR, and 0.082 in the admixed individuals. As expected, the average value of inter-chromosomal r^2 was smaller than the intra-chromosomal and was 0.016 in $P1OC_B$ and admixed individuals, 0.018 in $P2OC_T$, and 0.042 in P3OR. The LD decayed with distance between markers in all subpopulations. The average value of r^2 dropped below 0.2 at 1.4, 1.0, 1.8, and 0.8 Mb in $P1OC_B$, $P2OC_T$, P3OR and ADM, respectively (Fig 3).

Chromosomal heat-maps showed the presence of several blocks with a high level of LD spread in almost all the chromosomes, generally spanning hundreds of kilobases (S2–S6 Figs). Three of these blocks were common to all subpopulations, two at the top end of chromosome 4, of approximately 1.2 and 2 Mb, and one in chromosome 5 spanning about 2Mb. The organization of the $P1OC_B$ genome in LD blocks was highest, especially chromosomes 1, 2, 3, 4 and 5. The largest block of SNPs in high LD ($r^2 > 0.5$) was observed in LG1, starting at 8 Mb from the top of the chromosome and spanning 12 Mb. This block was followed in length by a 6 Mb block in the same chromosome at 32 Mb from the top. In $P2OC_T$ and P3OR, the SNPs at the bottom of chromosome 6, spanning 2 Mb, were in strong LD. The chromosomal heat-maps of the ADM population revealed only a few blocks of SNPs in high LD, most disappearing when the whole collection of 1,240 cultivars were analyzed with an LD statistic that corrects for population structure and relatedness, only the three common to all four subpopulations remaining.

Genome-wide association and haplotypes for major genes

Between 832 and 1,071 of the 1,240 genotyped plants were phenotyped for the following traits, known to be determined by single Mendelian genes [10]: fruit pubescence (*G/g*), fruit shape (*S/s*), fruit flesh color (*Y/y*), non-melting/melting flesh (*M/m*), titrable acidity (*D/d*), leaf gland type (*E/e*), and showy/non-showy flower type (*Sh/sh*) (see Methods for details). For all except fruit shape, the two phenotypic classes were homogeneously distributed. For the trait of fruit shape, the flat individuals (57) were underrepresented compared with round ones (1,003). Association between single SNPs and traits was identified, taking into account kinship and structure. The average value of kinship was 0.58, on a matrix scaled at 2, corresponding to about one quarter of the genome equivalent for example to a grandparent-grandchild, or a half

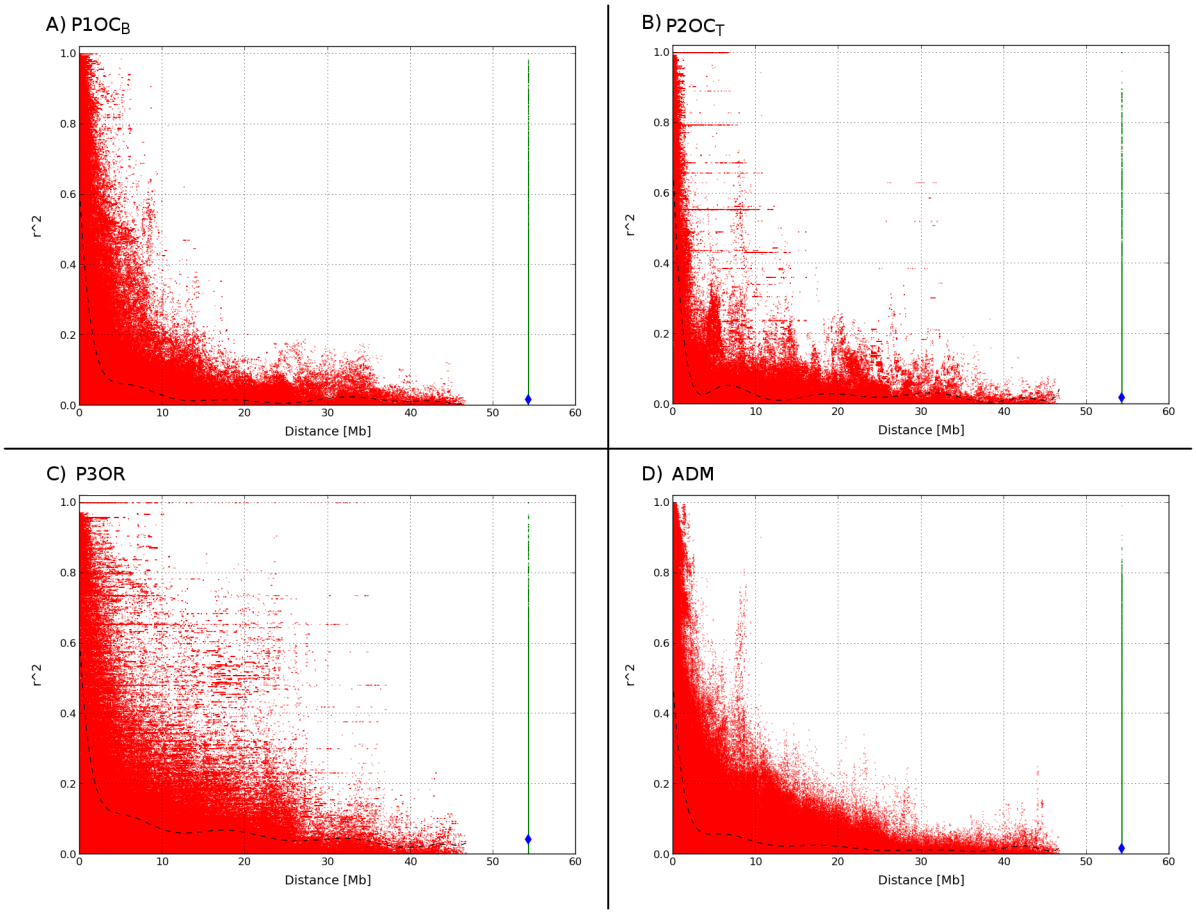

Fig 3. Decay of linkage disequilibrium in a collection of 1,240 peach accessions. The LD (r^2) was calculated for each intra-chromosomal pair of SNPs (red) and for all the inter-chromosomal (green) pairs of SNPs. The blue dotted line indicates the polynomial fitting of the intra-chromosomal comparisons and the average inter-chromosomal comparisons.

doi:10.1371/journal.pone.0136803.g003

sib relationship. We found significant association for all the traits ($-\log_{10}(p) > 5.63$) (Fig 4, S2 Table).

Titratable acidity (TA). The acid/sub-acid fruit trait has been mapped at the beginning of LG5 [25], where the sub-acid allele behaves as dominant. A total of 405 and 427 accessions were classified as acid (TA>8.2 mEq/100 ml. [26]) and sub-acid, respectively. Twenty-three SNPs at the beginning of chromosome 5 were associated with fruit acidity (Fig 4A). The associated SNPs cover a region of about 1.8 Mb (scaffold_5: 467,067..2,270,122) in agreement with the reported position for the D locus [25]. SNP_IGA_544640 (scaffold_5: 629,641) showed the strongest association with the trait, with a p-value of 1.5E-25 and MAF of 0.21. The most frequent allele was observed in 71.4% of sub-acid varieties, in either homozygosity or heterozygosity, while this allele was only present in 3.2% of the acid varieties. The genetic variance (heritability) explained by the model adopted for the association was 0.66. In total, the 23 SNPs produced 99 haplotypes, imputed using fastPHASE [27], with an average heterozygosity of 17.13%. Four haplotypes showed a frequency higher than 5%, two being strongly associated with the sub-acid phenotype (chi-squared p-value < 2.2E-16).

Fruit flesh color. At least three different recessive alleles of the CCD4 gene (ppa006109; scaffold_1:25,639.445..25,641.500) are responsible for the yellow flesh in peach, one produced by a mutation in a microsatellite at the first exon of the gene, one by an SNP in the second exon

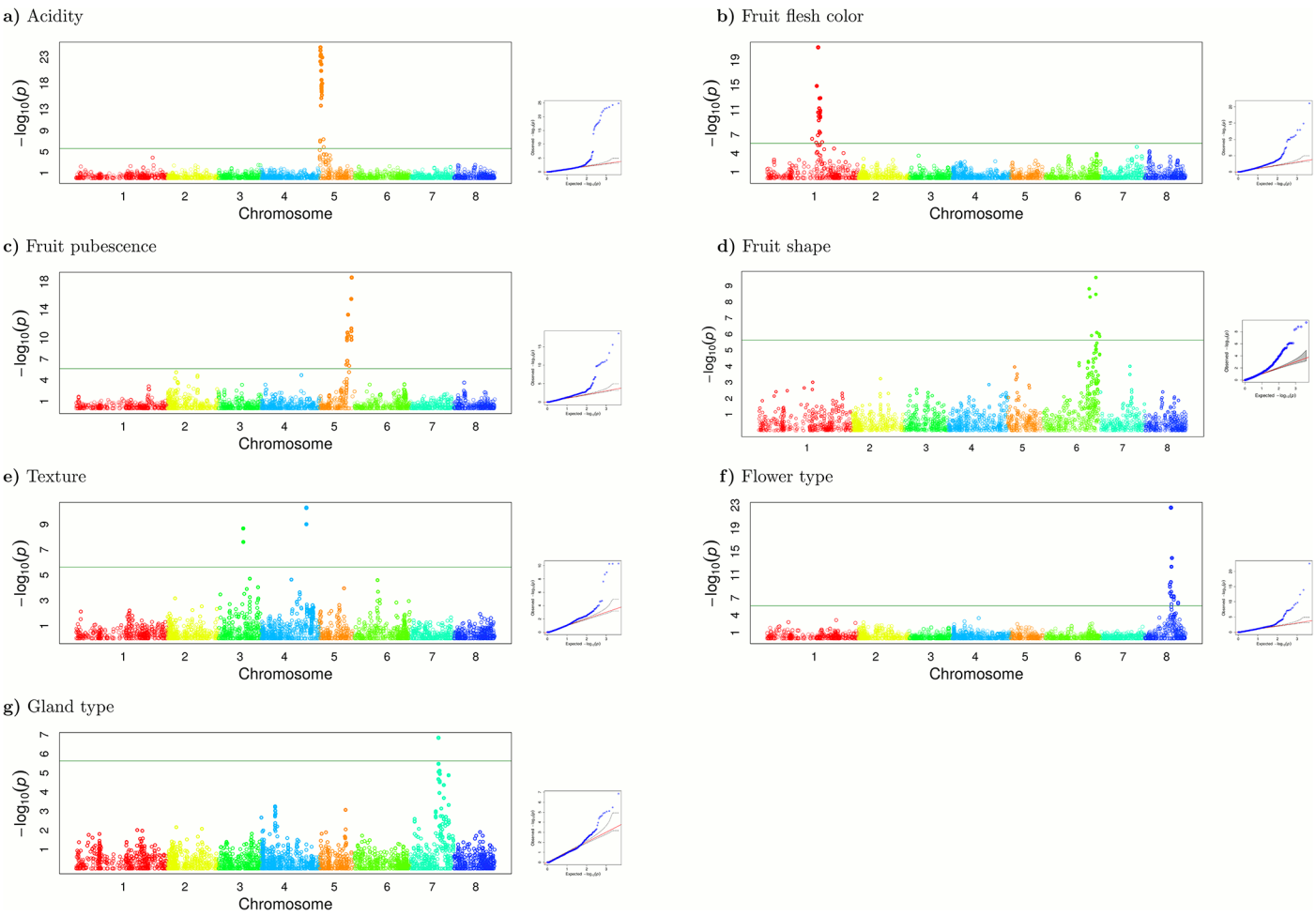

Fig 4. Genome wide association results for seven qualitatively inherited characters in peach. Each Manhattan plot represents one phenotypic trait. Chromosomes are marked with a different color on the horizontal axis. The horizontal green line represents the significance threshold for the association of each character.

doi:10.1371/journal.pone.0136803.g004

and one by the insertion of a transposable element [28]. In our collection, a total of 525 accessions were yellow-fleshed (*y*) while 499 were white (*Y*). The chromosomal heat-maps showed that the SNPs flanking this gene were in strong LD in a region 2 Mbp long (S1–S5 Figs). This block was shorter in $P2OC_T$ and disappeared in P3OR (where most of the cultivars had white flesh) and ADM. Association analysis identified 21 SNPs located in a 5.2 Mb region in linkage group G1 (scaffold_1:23,352,245..28,551,537) associated to this trait (Fig 4B). The SNP with the strongest association was SNP_IGA_90213 (scaffold_1: 26,479,525), with a p-value of 8.03E-22 and MAF of 0.21. The most frequent variant at this position was strongly linked to white flesh, being present in 61.9% of the white and 6.8% of the yellow varieties. The heritability explained by the allele was 0.65. The 21 associated SNPs were distributed in 152 imputed haplotypes. Six of them showed a frequency higher than 5%, all associated with the trait (Chi-square p-value <0.01), and an average heterozygosity of 92%.

Fruit hairiness. The recessive glabrous fruit trait (nectarine) has recently been associated to a retrotransposon insertion in an MYB gene (ppa026143; scaffold_5:15,897,836..15,899,002) [29]. Of the 1,071 accessions studied, 792 were hairy (peaches, carrying the *G* allele) and 290 glabrous (nectarines, homozygous for the *y* allele). We found 19 SNPs associated with the trait,

mapping in G5 (scaffold_5: 13,952,707..16,774,236). The average MAF of the associated SNPs was 0.31 and the heritability 0.73 (Fig 4C). SNP_IGA_603047 (scaffold_5:16,774,236) showed the highest association (p-value = 2.5E-19, MAF = 0.37 and $R^2 = 0.66$). The most frequent allele was equally present in peaches and nectarines while the less frequent was almost exclusively in peach (98.2% of the varieties with this allele were peaches). PHASE imputed 123 haplotypes using the 19 associated SNPs, five with a frequency > 5% and an average heterozygosity of 8.15%. The most frequent was in high association with the recessive trait: 98% of the accessions with this haplotype in homozygosis had the recessive phenotype.

Flat fruit shape. Flat fruit shape is caused by a dominant allele in obligated heterozygosity, mapped at the end of LG6 [25, 30]. In our panel, only 57 cultivars were flat whereas 1,003 were round. The low number of flat accessions hampered the analysis and it was not possible to identify significant association using the full dataset. This situation could be partially explained considering the relatedness of some flat individuals, for example those from the series UFO (12.3% of flat varieties), and the strong structuring of the trait. In order to solve these drawbacks, we studied 15 independent subsets containing the 57 flat and 100 randomly chosen round cultivars from $P1OC_B$ and admixed accessions. In all the subsets, a single peak of association at the end of chromosome six was evident (scaffold_6:23,101,004..26,601,733). The number of associated SNPs varied between subsets, ranging from five to 22 (Fig 4D). Five were associated with the trait in all runs (SNP_IGA_683904, SNP_IGA_684085, SNP_IGA_685825, SNP_IGA_696241, SNP_IGA_696280), with an average MAF of 0.29 and heritability of 1.0. The first two were in complete LD and showed, on average, higher association with the flat trait. Most of the flat varieties (87.7%) were heterozygous at these two loci while this percentage decreased to 8% in the round varieties, consistent with the genetics of the trait.

Texture. An endoPG gene, mapping at the end of LG4, has been reported to control melting flesh (M) in peach [31]. In our germplasm collection, 213 accessions were classified as non-melting and 746 as melting. We found two SNPs significantly associated in LG3 12 kb apart (scaffold_3:12,836,182..12,878,608) and four in LG4 covering a region of 11.6 kb (scaffold_4:23,497,381..23,508,950) (Fig 4E). The two SNPs associated in LG3 (SNP_IGA_341962; p-value 2.1E-9, and snp_3_12878608; p-value 2.4E-8) were in high LD ($r^2 = 0.909$) and the association was due to an excess of the most frequent allele in homozygosis in the melting varieties. All four SNPs associated in LG4 (SNP_IGA_477941, SNP_IGA_477945, SNP_IGA_477951, SNP_IGA_478039) were in high LD, the first two being in complete LD ($r^2 = 1$) (SNP_IGA_477941 and SNP_IGA_477945; p-value = 5.1E-11). The association was due to an excess of the less frequent allele in homozygosis in the non-melting fruits. Four and six haplotypes were imputed using the two and four associated SNPs of chromosomes 3 and 4, respectively. Two haplotypes had a frequency higher than 5% in both chromosomes. The average heterozygosity was 10.6% in chromosome 3 and 18.9% in chromosome 4.

Flower type (showy/non-showy). The showy/non-showy flower locus (*Sh*) maps in LG8 [32], with the non-showy allele being dominant. Two-hundred and sixty of the accessions analyzed had non-showy flowers while 681 had showy flowers. We found 18 SNPs associated with the trait in a region of 4.3 Mb of chromosome eight (scaffold_8: 13,204,775..17,476,655) (Fig 4F). The average MAF of the associated SNPs was 0.31 and the heritability was 0.62. The most highly associated SNP was SNP_IGA_864149 (scaffold_8: 13,756,987) with a p-value of 2.73E-23, MAF of 0.36 and an r^2 of the model with the SNP of 0.42. The homozygous genotype of the most frequent allele of this SNP was more frequent in the showy (66.2%) than in the non-showy (10%) varieties. A total of 162 haplotypes were imputed using the 18 SNPs associated, six with a frequency higher than 5%.

Leaf gland type. The *E* locus, determining the globular or reniform shape of the leaf glands, maps in LG7 [33]. The globular leaf gland phenotype was present in 190 accessions and

the reniform in 750. A single SNP (SNP_IGA_776161, scaffold_7:14,870,521) was associated to the leaf gland type (Fig 4G). The associated SNP had an MAF = 0.39, R^2 = 0.10 and heritability of 0.31.

Discussion

Analysis of 8,144 SNPs, represented in the IPSC 9k chip, in a diverse sample of 1,580 peach accessions revealed polymorphisms in 5,378 SNPs (66%), while 727 (9%) were false (monomorphic) and 2,037 (25%) failed. These data confirm the accuracy of the validation by Verde et al. [17] using a subset of 96 SNPs of this array, where they found similar proportions (57%, 19% and 24%, respectively). Only 4,271 of the polymorphic SNPs were retained after exclusion of 1,107 that were suspected to contain null/preferential amplification alleles, correspond to duplicate sequences or had MAF <0.05.

Such an array of polymorphic markers allows comparison of genotypes at the whole-genome level, identifying features not possible to discern with the low density marker analysis assays used to date [6, 34]. One of them concerns the differences observed between some of the known and inferred groups of sports analyzed. In four cases, we observed that SNP differences among sports were concentrated at one 5.3 Mb region in the distal end of chromosome 6. These differences involved changes of marker genotype (in 1–11 SNPs) and missing data in many of the 92 markers included in our SNP array in this region (between 38 and 60 SNPs with no data). The pair of cultivars, 'Springtime' and 'Starcrest', presented a similar pattern, although in this case affecting the upper part of chromosome 4. The same region of chromosome 6 is involved in a reciprocal translocation with chromosome 8 in the red-leaved rootstocks 'Nemared' [35], 'Akame' [36] and 'Rubira' [37]. Our results suggest that this region of the genome may be particularly unstable, with major rearrangements occurring frequently. They also suggest that the phenotypic differences between an original cultivar and its sports may be caused by genes located in these unstable DNA fragments.

The UPGMA dendrogram separated Oriental accessions from those cultivated in Occidental countries. Breeding accessions were also clearly separated from those cultivated locally. The kinship analysis revealed that accession pairs were related on average at the level of grandparent-grandchild or half-sibs. This was in agreement with the results from PCA and the STRUCTURE software, identifying three subpopulations within the peach germplasm analyzed: $P1OC_B$, containing most accessions from modern breeding programs, essentially the gene pool derived from that used by early US breeders; $P2OC_T$, including the old European accessions, most of them non-melting peaches that were seed-propagated for a long time and as a consequence highly homozygous; and P3OR that encompasses the majority of Oriental accessions. Previous studies with SSRs [6, 34, 38] also showed that *P. persica* has a strong population stratification and identified the same groups. A recent study [9] identified a different stratification in a panel of 84 wild and cultivated peach accessions: i) wild relatives (species), ii) edible and iii) ornamental accessions. However, only a few Western and breeding accessions were included in the panel. The subpopulations detected are in general consistent with the known history of peach domestication and dispersal, where P3OR represents the original Chinese gene pool from the domestication process, and the old European materials ($P2OC_T$) are those derived from the migration of the Chinese accessions throughout central Asia towards Western Europe. Some of these materials were taken to America by European settlers and were the founders, with a few accessions coming directly from China, of the US breeding programs that started in the early 20th century using the concepts of modern genetics. These programs produced a new wave of successful cultivars that are currently the basis of European and North-American peach production, constituting subpopulation $P1OC_B$.

The population structure identified in this study is also consistent with the genetic bottlenecks described by Verde et al. [8] using a set of whole genome SNPs in a panel of 12 Occidental and Oriental accessions. The first bottleneck clearly denotes the domestication event and is related to the P3OR subpopulation. The second bottleneck is related to peach dissemination in the West and to the modern breeding activities that began in the US and Europe in the last century. This bottleneck is represented by the $P2OC_T$ and $P1OC_B$ subpopulations. In their study Cao et al. [9] suggested the existence of a single bottleneck related to the domestication event failing in identifying dissemination and breeding bottlenecks. However, they included only a few Occidental and breeding accessions in their panel; this lack likely hampered the detection of Western dissemination and modern breeding bottlenecks. A subdivision within this population that is also evident in the SSR studies [7, 34] separates peaches from nectarines, which emphasizes the fact that breeding of these two fruit types has often been separate. A large fraction of the accessions analyzed (54%) were admixed between the subpopulations identified, suggesting that breeding programs also look for useful variability crossing the boundaries of each subpopulation.

The mean observed heterozygosity was $H_o = 0.28$, a lower value than that usually reported with SSRs: $H_o = 0.34$ in a collection of Occidental cultivars [34] and $H_o = 0.47$ for a larger collection of Oriental and Occidental materials [6]. This was expected considering the higher numbers of alleles per locus of SSRs. The proportion of markers with MAF< 0.2 was low when considering all the genotypes studied in the admixed group (17.3–20.0%), and much higher for the three subpopulations (46.8% for $P1OC_B$, 54.2 for $P2OC_T$ and 61.3% for P3OR). This situation can be explained considering that the SNPs were identified in resequencing data of about 50 peach accessions, mostly from the $P1OC_B$ and the admixed groups [17], and by the larger size of these groups (352 and 665 accessions, respectively) compared to the other subpopulations.

Population structure affects the LD throughout the genome. LD decayed with distance in all subpopulations, being faster in the ADM subpopulation. Accessions in this population contain variability from genetically distinct founding populations with different allele frequencies: thereafter they should retain the LD of the parental populations. Additionally the amount of admixture-generated LD between any two loci is highly dependent on the proportion of each population in the admixture process [39]. Faster decay of LD in ADM could be due to the remaining subpopulation stratification in $P1OC_B$, $P2OC_T$ and P3OR or to a relatively low participation of these three populations in the admixture process (see S1 Table).

Previous studies comparing LD dynamics have detected faster decay in Oriental compared to Occidental germplasm [6]. One reason this was not observed in our sample could be the origin of the accessions of the population; here P3OR only included mostly Chinese breed varieties, and no wild germplasm or landraces. This indicates that breeding efforts in China and Occidental countries have resulted in similar LD patterns and, according to Xie et al.[40], in a similar reduction of variability. As expected, the LD pattern was not homogeneous across the chromosome, containing areas with extensive linkage disequilibrium. These areas may correspond to recombination cold-spots and so should be observed in all populations as well as in linkage maps. Two LD blocks in LG4, one at the top and one at the bottom, and one in LG5, were observed in all populations. These regions were also observed by Verde et al.[8], who also plotted the relation between genetic and physical distance in peach chromosomes. A low ratio between genetic and physical distance, indicating a low recombination rate, was only observed for the block at the end of LG4. Selection may explain the LD blocks of the other two regions. Artificial selection has a strong effect on LD. Mosaics of large LD blocks have been found in regions carrying agronomic-related genes [41]. Another example of LD generated by selection detected in our study is the long LD block around the *Y* gene in $P1OC_B$, with most of the

varieties having the recessive yellow flesh phenotype. This disappears in P3OR, where most varieties had white flesh.

Association mapping exploits natural diversity to search for functional variation. This method presents several advantages over linkage mapping but usually they are used together to validate or discard spurious associations. Here we applied association mapping to validate its use in peach. For some of the traits evaluated the genes have been cloned (as is the case of the *Y*, *G* and *F* loci) or mapped in a short interval. In all cases, the SNPs found to be associated with the traits were located in the known region. For the known genes ppa006109 for the *Y* locus, ppa023143 for *G* and ppa006839 for *F*, the most associated SNPs were 838, 875 and 849 Kb upstream of the 5'-UTR of the gene, which corresponds approximately to 1.9 cM (considering the *Prunus* genome of 519 cM and 230 Mb).

We found SNPs tightly linked to the traits considered, but their use in marker assisted selection (MAS) is not straightforward and still needs to be validated. For most of the traits we found haplotypes that explain a large fraction of the phenotypes, however the association is not complete. For example three causal alleles have been reported for the yellow flesh trait [28]. All three are loss-of-function alleles from different mutation events and, consequently, have different SNPs (or haplotypes of SNPs) associated to each allele. Here we showed the considerable extent of LD in peach and that, in populations, SNPs are structured into haplotypes which may be better at predicting the phenotype than single SNPs. Our results do strongly suggest that the use of haplotypes is more powerful than single tightly-linked SNPs in MAS. However, in most cases, these haplotypes will be population-dependent, i.e. they have to be chosen based on the frequency of the haplotypes in the parental lines used in the breeding program.

Conclusions

A collection of 1,580 Occidental and Oriental peach accessions was analyzed with an SNP array covering the whole peach genome. Results showed that the levels and organization of variability are consistent with the known major events in the history of this species after domestication, and generally agree with published information obtained with SSR markers and SNPs.

Given the high levels of LD conservation of the peach genome, whole-genome association analysis with a relatively small number of loci (4,271) showed that SNPs significantly associated with seven major genes were always in the regions predicted by classical linkage analysis. For the genes still not cloned, our data provide sets of tightly linked markers that can be the starting point to identify candidate genes and eventually used for their cloning and molecular characterization. The results also suggest that similar analysis with characters of complex inheritance and high value, such as those related with fruit organoleptic quality and extended post harvest life, will be able to identify genomic regions containing the most important genes responsible for variation of the traits.

The information has other applications for peach breeding and germplasm management, as many commercial cultivars from private and public breeding programs have been characterized genetically to an unprecedented level of detail, enabling the study of the variability within each program and comparison of the variability between breeding programs on a whole-genome basis. This information is useful for parent selection and to search for variability at specific genomic regions not present in the set of parents used in a specific breeding program, but present in the global peach gene pool. Moreover, our results reveal that, for certain characters, the MAS strategy not only requires one or several markers sufficiently close to the gene of interest, but also identification of the different haplotypes for various markers in the region, selecting those that are diagnostic for the character based on the haplotypes of the set of parents used by each breeding program.

Materials and Methods

Plant materials and phenotyping

A set of 1,580 accessions of *Prunus* were selected to be representative of five germplasm collections in four different countries. Of these accessions 1576 were *P. persica* and four were of other related *Prunus*. In detail, 420 accessions were from the peach germplasm collection of Centro di Ricerca per la Frutticoltura (CRA-FRU, Roma), 112 and 239 were provided by the National Institute of Agronomic Research (INRA) of Avignon (France) and Bordeaux (maintained in the Prunus Genetic Resources Center, France), respectively, 367 were from the peach germplasm collection of IRTA at the experimental stations of La Tallada d'Empordà and Gimenells (Spain), 160 were obtained from the collection of University of Milan (Italy), and 282 from the peach collection of Zhejiang University (Hangzhou, China). The complete list of the accessions is reported in S1 Table.

Between 50 and 100 mg of young leaves from each accession were freeze-dried and used for genomic DNA extraction with the DNeasy 96 Plant Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol.

The phenotype of a variable number of plants was available for seven monogenic traits: 1,071 for fruit pubescence (peach/nectarine), 1,060 for fruit shape (round/flat), 1,024 for fruit flesh color (yellow/white), 959 for fruit texture (melting/non-melting), 832 fruit acidity (acid/sub-acid), 940 for gland type (round/reniform), and 741 for flower type (showy/non-showy). All the traits were coded as qualitative.

SNP genotyping

1,580 *Prunus* accessions were genotyped using the International Peach SNP Consortium (IPSC) peach 9K SNP array v1 [17]. SNP genotypes were scored with the Genotyping Module of the GenomeStudio Data Analysis software (Illumina, Inc.) using the default parameters.

The SNPs were divided into five categories: A, B, C, D, and E. The first three included the polymorphic SNPs and the last two the non-polymorphic. The SNPs with GenTrain higher than 0.4 and GeneCall 10% higher than 0.2 and at least two genotypic classes were classified as polymorphic. The three classes of polymorphic SNPs were defined as follows:

A. SNPs with less than 5% of No Call (failed genotyping) and all three possible genotypes (AA, AB, BB) defined.

B. SNPs with a potential null allele or preferential annealing. Between 5 and 50% of individuals showed a normalized signal intensity value $R < 0.2$ and the remaining individuals represented at least two of the three genotypic classes.

C. Probable presence of duplicated sequences/genes. These SNPs were characterized by the absence of one homozygous cluster with an over-representation of heterozygous individuals, and a percentage of No Call <5%.

The non-polymorphic SNPs were divided as:

D. False SNPs. Characterized by a single genotypic class and a percentage of No Call < 5%.

E. Failed SNPs. Having more than the 50% of No Call and/or with a GenTrain < 0.4 and/or a GeneCall 10% < 0.2.

All further analysis were performed using all Class A SNPs with a minor allele frequency (MAF) higher than 0.05. Genotypic data have been uploaded to GDR database (http://www.

rosaceae.org/) under the accession number tfGDR1013 and in the FruitBreedomics database (http://bioinformatics.tecnoparco.org/fruitbreedomics/).

Phylogenetic tree

The distance matrix was calculated using TASSEL [42], as 1—IBS (identity by state) similarity, being IBS the probability that alleles drawn at random from two individuals at the same locus are the same. For clustering, the distance of an individual from itself was set to 0. The clustering was performed using the R package hclust with the UPGMA method.

Population Structure and Fst

PCA analysis was performed using the "prcomp" R package, freely available at http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html, after genotype numericalization as follows: The following approach was followed for numericalization: the higher value (2) was assigned to the heterozygous state, the lower value (0) was assigned to the less frequent homozygous state, and the intermediate state (1) was assigned to the most frequent homozygous state.

Population structure was studied with the "prcomp" function of R and with the Structure v.2 [23] software. This program uses a clustering method that identifies K subgroups of individuals with distinctive allele frequencies. Individuals can be members of multiple subpopulations with a different coefficient, with the sum of all being equal to 1. To check for population stratification, the program was run under the admixture model assumption with correlated allele frequencies. The run used 100,000 interactions after a burn-in of 10,000 for a value of K ranging from 2 to 20. To avoid bias due to tightly linked markers and to speed up the computation time, the SNPs were pruned based on LD using PLINK [43] to give a final 1,506 SNPs. Pruning was using the "indep" function with a window size of 50 SNPs, five SNPs shift of the window at each step, and a variance inflation factor (VIF) threshold of 10 (equivalent to prune to an LD of 0.75). The VIF is $1/(1-R^2)$ where R^2 is the multiple correlation coefficient for an SNP being regressed on all other SNPs simultaneously. The most probable number of populations was estimated using the method proposed by [44]. The fixation index (Fst) was calculated as in [45] using a custom python script.

HWE and MAF

The Hardy Weinberg equilibrium and the minor allele frequency were calculated for each SNP using PLINK [43]. The SNPs showing severe distortion of the HWE ($p<10e-4$) and/or MAF lower than 0.05 were discarded from further analyses.

Linkage disequilibrium

Given that the phases between alleles at two heterozygous loci are unknown, the linkage disequilibrium (LD) was calculated using the squared correlation based on genotypic allele counts as implemented in PLINK [43]. Additionally, the LD was calculated using the R-package LDcorSV [24], able to correct the classical r^2 measure for population structure and relatedness. To measure LD, SNPs with a minor allele frequency (MAF) lower than 5% were discarded.

The curve representing the observed r^2 values was computed using a polynomial fit as implemented in the "numpy" module (http://docs.scipy.org/doc/) of python v.2.7.3.

Kinship

The kinship matrix was computed using the VanRaden algorithm [46] as implemented in the GAPIT R package [47].

Genome wide associations

Association analysis was using the compressed mixed linear model [48] implemented in the GAPIT R package [47]. Association tests were run using two approaches. One was a global approach where an MLM model for all individuals was run using population structure (k coefficients of ancestry, Q), and a genetic covariance matrix (kinship matrix, K) used as cofactor for SNP effects. The other was a population level approach where each sampling population was run in a separate analysis using the MLM as outlined above. Significant associations were determined using both a B–Y FDR P-value adjusted for multiple testing [49] and a Bonferroni adjustment at the $\alpha = 0.01$ level. Prior to the analyses, SNP markers were filtered for MAF of ≥ 0.05, and a minimum genotyping success of 90%.

The SNPs associated to each trait were phased independently using fastPHASE [27]. For each run of fastPhase 20 random starts and 100 iteration of the EM algorithm were used. To test the association of the inferred haplotypes with the analyzed phenotypes a chi-squared test was performed for each haplotype. The null hypothesis was that the number of accessions showing the two alternative phenotypes was the equal.

Supporting Information

S1 Fig. STRUCTURE output for 1,240 peach accessions with K varying from 2 to 7.
(TIFF)

S2 Fig. Heat map representing the chromosomal level of LD measured as r^2 in $P1OC_B$ (Occidental accessions). The diagonal black line indicates the physical position of the SNPs on the chromosome.
(TIFF)

S3 Fig. Heat map representing the chromosomal level of LD measured as r^2 in $P2OC_T$ (Occidental traditional accessions). The diagonal black line indicates the physical position of the SNPs on the chromosome.
(TIFF)

S4 Fig. Heat map representing the chromosomal level of LD measured as r^2 in P3OR (Oriental accessions). The diagonal black line indicates the physical position of the SNPs on the chromosome.
(TIFF)

S5 Fig. Heat map representing the chromosomal level of LD measured as r^2 in the admixed individuals. The diagonal black line indicates the physical position of the SNPs on the chromosome.
(TIFF)

S6 Fig. Heat map that represents the linkage disequilibrium calculated as r^2 corrected for population structure and relatedness. The diagonal black line indicates the physical position of the SNPs on the chromosome.
(TIFF)

S1 Table. List of varieties analyzed. (1) Asterisks indicate accessions from two different germplasm collections with the same name but with different genotype. (2) n/p = nectarine/peach;

y/w = yellow/white flesh; f/nf = flat/non flat fruit shape; f/c/s = freestone/clingstone/semi-clingstone; m/n = melting/non melting; s/a = sub-acid/acid; r/g = reniform/globose leaf gland; s/ns = showy/non-showy flowers. (3) Numbers indicate the position of the accession in the Fig 1 phylogenetic tree, starting from the top branch.
(XLSX)

S2 Table. SNPs associated with each of the traits studied.
(XLSX)

Acknowledgments

We acknowledge Andrea Caprera (PTP, Lodi, Italy) for his contribution to the organization and standardization of phenotypic datasets; Zhi-jun Shen (Jiangsu Academy of Agricultural Sciences), Ke Cao (Zhengzhou Fruit Research Institute), Hui-juan Jia, Xian-qiao Meng (Zhejiang University), Ming Xie (Zhejiang Academy of Agricultural Sciences) and Jian-bao Tian (Shanxi Academy of Agricultural Sciences) for collecting materials and collecting phenotypic data; Martina Lama (CRPV, Cesena, Italy) for her participation in lab work and the technical staff of the "Biologie du Fruit et Pathologie" unit of the INRA, in particular Hélène Christmann, for her technical help in DNA extraction, and Jean Claude Barbot, Hélène Christmann, Anthony Bernard and Aude Mear for their active collaboration in the phenotyping of the INRA-Bordeaux peach collection. We thank the INRA's "Prunus Genetic Resources Center" for preserving and managing the peach collections and the Experimental Unit of INRA-Toulenne (UEA) for growing the trees. This work has been funded under the UE seventh Framework Programme, the views expressed in this work are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

Author Contributions

Conceived and designed the experiments: IV LR DB BQ-T TB PL TP Z-SG FL PA MJA. Performed the experiments: DM IV MTD SM VA SF CSL LR IP EB TB PL X-WL II JC MJA. Analyzed the data: DM NN MT MTD IV LR PA MJA. Contributed reagents/materials/analysis tools: IV LR DB PL TP MT TB Z-SG L-RW R-JM II JC PA MJA. Wrote the paper: DM PA MJA. Built the bioinformatic infrastructure for data access: NN. Critically revised the manuscript: IV MTD SM LR DB IP MT BQ-T TB PL TP Z-SG FL.

References

1. Faust M, Timon B. Origin and Dissemination of Peach. In: Janik J, editor. Horticultural Reviews. New York: John Wiley & Sons, Inc.; 1995. p. 331–79.
2. Byrne DH. Isozyme variability in four diploid stone fruits compared with other woody perennial plants. The Journal of Heredity 1990; 81(1): 68–71.
3. Mnejja M, Garcia-Mas J, Audergon J-M, Arús P. Prunus microsatellite marker transferability across rosaceous crops. Tree Genetics & Genomes 2010; 6(5): 689–700.
4. Scorza R, Mehlenbacher SA, Lightner GW. Inbreeding and coancestry of freestone peach cultivars of the eastern united states and implications for peach germplasm improvement. J Amer Soc Hort Sci 1985; 110(4): 547–52.
5. Hesse CO. Peaches. Advances in Fruit Breeding 1975: 285–335.
6. Li X-w, Meng X-q, Jia H-j, Yu M-l, Ma R-j, Wang L-r, et al. Peach genetic resources: diversity, population structure and linkage disequilibrium. BMC Genetics 2013; 14(1): 84.
7. Aranzana MJ, Carbó J, Arús P. Microsatellite variability in peach [Prunus persica (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. Theor Appl Genet 2003; 106(8): 1341–52. PMID: 12750778

8. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, et al. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 2013; 45(5): 487–94. doi: 10.1038/ng.2586 PMID: 23525075
9. Cao K, Zheng Z, Wang L, Liu X, Zhu G, Fang W, et al. 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. doi: 10.1186/s13059-014-0415-1 PMID: 25079967
10. Monet R, Guye A, Roy M, Dachary N. Peach mendelian genetics: a short review and new results. Agronomie 1996; 16: 321–9.
11. Abbott AG, Georgi L, Yvergniaux D, Wang Y, Blenda A, Reighard G, et al. Peach: the model genome for Rosaceae. Acta Hort 2002; 575: 145–55.
12. Dirlewanger E, Graziano E, Joobeur T, Garriga-Calderé F, Cosson P, Howad W, et al. Comparative mapping and marker-assisted selection in Rosaceae fruit crops. PNAS 2004; 101(23): 9891–6.
13. Abbott AG, Arús P, Scorza R. Genetic engineering and genomics. In: Bassi DRLaD, editor. The peach Botany, production and uses London: CAB International; 2008. pp. 85–105.
14. Arús P, Yamamoto T, Dirlewanger E, Abbott AG. Synteny in the Rosaceae. In: Janik J, editor. Plant Breeding Reviews. Hoboken, NJ: John Wiley & Sons; 2005. pp. 175–211.
15. Aranzana M, Illa E, Howad W, Arus P. A first insight into peach [Prunus persica (L.) Batsch] SNP variability. Tree Genetics & Genomes 2012; 8(6): 1359–69.
16. Ahmad R, Parfitt DE, Fass J, Ogundiwin E, Dhingra A, Gradziel TM, et al. Whole genome sequencing of peach (*Prunus persica* L.) for SNP identification and selection. BMC Genomics 2011; 12: 569. doi: 10.1186/1471-2164-12-569 PMID: 22108025
17. Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, et al. Development and Evaluation of a 9K SNP Array for Peach by Internationally Coordinated SNP Detection and Validation in Breeding Germplasm. PLOS ONE 2012; 7(4): e35668. doi: 10.1371/journal.pone.0035668 PMID: 22536421
18. Eduardo I, Chietera G, Pirona R, Pacheco I, Troggio M, Banchi E, et al. Genetic dissection of aroma volatile compounds from the essential oil of peach fruit: QTL analysis and identification of candidate genes using dense SNP maps. Tree Genetics & Genomes 2013; 9(1): 189–204.
19. Yang N, Reighard G, Ritchie D, Okie W, Gasic K. Mapping quantitative trait loci associated with resistance to bacterial spot (Xanthomonas arboricola pv. pruni) in peach. Tree Genetics & Genomes 2013; 9 (2): 573–86.
20. Frett T, Reighard G, Okie W, Gasic K. Mapping quantitative trait loci associated with blush in peach [Prunus persica (L.) Batsch]. Tree Genetics & Genomes 2014; 10(2): 367–81.
21. Pacheco I, Bassi D, Eduardo I, Ciacciulli A, Pirona R, Rossini L, et al. QTL mapping for brown rot (Monilinia fructigena) resistance in an intraspecific peach (Prunus persica L. Batsch) F1 progeny. Tree Genetics & Genomes 2014; 10(5): 1223–42.
22. Romeu J, Monforte A, Sanchez G, Granell A, Garcia-Brunton J, Badenes M, et al. Quantitative trait loci affecting reproductive phenology in peach. BMC Plant Biology 2014; 14(1): 52.
23. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics2000 June 1, 2000; 155(2): 945–59. PMID: 10835412
24. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 2012; 108(3): 285–91. doi: 10.1038/hdy.2011.73 PMID: 21878986
25. Dirlewanger E, Pronier V, Parvery C, Rothan C, Guye A, Monet R. Genetic linkage map of peach [*Prunus persica* (L.) Batsch] using morphological and molecular markers. Theoretical and Applied Genetics 1998; 97(5–6): 888–95.
26. Eduardo I, López-Girona E, Batlle I, Reig G, Iglesias I, Howad W, et al. Development of diagnostic markers for selection of the subacid trait in peach. Tree Genetics & Genomes 2014; 10(6): 1695–709.
27. Scheet P, Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics 2006; 78(4): 629–44. PMID: 16532393
28. Falchi R, Vendramin E, Zanon L, Scalabrin S, Cipriani G, Verde I, et al. Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. The Plant Journal 2013; 76 (2): 175–87. doi: 10.1111/tpj.12283 PMID: 23855972
29. Vendramin E, Pea G, Dondini L, Pacheco I, Dettori MT, Gazza L, et al. A Unique Mutation in a MYB Gene Cosegregates with the Nectarine Phenotype in Peach. PLOS ONE 2014; 9(3): e90574. doi: 10.1371/journal.pone.0090574 PMID: 24595269

30. Dirlewanger E, Cosson P, Boudehri K, Renaud C, Capdeville G, Tauzin Y, et al. Development of a second-generation genetic linkage map for peach [*Prunus persica* (L.) Batsch] and characterization of morphological traits affecting flower and fruit. Tree Genetics & Genomes 2006; 3(1): 1–13.
31. Peace CP, Crisosto CH, Gradziel TM. Endopolygalacturonase: a Candidate Gene for Freestone and Melting Fleshin Peach. Molecular Breeding 2005; 16(1): 21–31.
32. Ogundiwin E, Peace C, Gradziel T, Parfitt D, Bliss F, Crisosto C. A fruit quality gene map of Prunus. BMC Genomics 2009; 10(1): 587.
33. Dettori MT, Quarta R, Verde I. A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. Genome 2001; 44: 783–90. PMID: 11681601
34. Aranzana M, Abbassi E-K, Howad W, Arus P. Genetic variation, population structure and linkage disequilibrium in peach commercial varieties. BMC Genetics 2010; 11(1): 69.
35. Jáuregui B, de Vicente MC, Messeguer R, Felipe A, Bonnet A, Salesses G, et al. A reciprocal translocation between 'Garfi' almond and 'Nemared' peach. Theor Appl Genet 2001; 102: 1169–76
36. Yamamoto T, Yamaguchi M, Hayashi T. An integrated genetic linkage map of peach by SSR, STS, AFLP and RAPD. J Jpn Soc Hort Sci 2005; 74(3): 204–13.
37. Lambert P, Pascal T. Mapping Rm2 gene conferring resistance to the green peach aphid (Myzus persicae Sulzer) in the peach cultivar 'Rubira'. Tree Genetics & Genomes 2011; 7(5): 1057–68.
38. Cao K, Wang L, Zhu G, Fang W, Chen C, Luo J. Genetic diversity, linkage disequilibrium, and association mapping analyses of peach (Prunus persica) landraces in China. Tree Genetics & Genomes 2012; 8(5): 975–90.
39. Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, et al. Population Structure in Admixed Populations: Effect of Admixture Dynamics on the Pattern of Linkage Disequilibrium. Amer J Hum Genet 2001; 68: 198–207. PMID: 11112661
40. Xie R, Li X, Chai M, Song L, Jia H, Wu D, et al. Evaluation of the genetic diversity of Asian peach accessions using a selected set of SSR markers. Scientia Horticulturae 2010; 125(4): 622–9.
41. Soto-Cerda BJ, Cloutier S. Association mapping in plant genomes. In: Caliskan M Editor. Genetic Diversity in Plants; 2012. pp. 29–54.
42. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 2007; 23: 2633–2635. PMID: 17586829
43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 2007; 81(3): 559–75. PMID: 17701901
44. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 2005; 14: 2611–20. PMID: 15969739
45. 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(12): 1805–1814. PMID: 12466284
46. VanRaden P. Efficient methods to compute genomic predictions. Journal of dairy science 2008; 91(11): 4414–23. doi: 10.3168/jds.2007-0980 PMID: 18946147
47. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics 2012; 28(18): 2397–9. PMID: 22796960
48. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nature genetics 2010; 42(4): 355–60. doi: 10.1038/ng.546 PMID: 20208535
49. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist 2001; 29(4): 1165–88.