Evolutionary Applications
Open Access
Evolutionary Applications ISSN 1752-4571

ORIGINAL ARTICLE

Genomic basis of the differences between cider and dessert apple varieties

Diane Leforestier,[1] Elisa Ravon,[2] Hélène Muranty,[2] Amandine Cornille,[3,4] Christophe Lemaire,[1] Tatiana Giraud,[3,4,*] Charles-Eric Durel[2,*] and Antoine Branca[3,4,*]

1 UMR 1345 Institut de Recherche en Horticulture et Semences, Université d'Angers, Angers, France
2 UMR 1345 Institut de Recherche en Horticulture et Semences, INRA, Beaucouzé, France
3 Ecologie, Systématique et Evolution, Université Paris-Sud, Orsay, France
4 Ecologie, Systématique et Evolution, CNRS, Orsay, France

Keywords
BayeScan, F_{ST}, genomewide association, linkage disequilibrium, *Malus domestica*, outlier.

Correspondence
Charles-Eric Durel, UMR 1345 Institut de Recherche en Horticulture et Semences, INRA, Beaucouzé, France.
Tel.: +33-2-41225759;
fax: +33-2-41225755;
e-mail: charles-eric.durel@angers.inra.fr

*Co-senior authors

Received: 7 October 2014
Accepted: 15 April 2015

doi:10.1111/eva.12270

Abstract

Unraveling the genomic processes at play during variety diversification is of fundamental interest for understanding evolution, but also of applied interest in crop science. It can indeed provide knowledge on the genetic bases of traits for crop improvement and germplasm diversity management. Apple is one of the most important fruit crops in temperate regions, having both great economic and cultural values. Sweet dessert apples are used for direct consumption, while bitter cider apples are used to produce cider. Several important traits are known to differentiate the two variety types, in particular fruit size, biennial versus annual fruit bearing, and bitterness, caused by a higher content in polyphenols. Here, we used an Illumina 8k SNP chip on two core collections, of 48 dessert and 48 cider apples, respectively, for identifying genomic regions responsible for the differences between cider and dessert apples. The genome-wide level of genetic differentiation between cider and dessert apples was low, although 17 candidate regions showed signatures of divergent selection, displaying either outlier F_{ST} values or significant association with phenotypic traits (bitter versus sweet fruits). These candidate regions encompassed 420 genes involved in a variety of functions and metabolic pathways, including several colocalizations with QTLs for polyphenol compounds.

Introduction

Domestication and variety diversification have been models for studying the mechanisms underlying adaptation since Darwin (1856), being the result of a strong and recent selection by humans for desired traits in organisms used as food (Meyer et al. 2012; Larson and Burger 2013; McTavish et al. 2013), ornaments (Yuan et al. 2014), pets (Axelsson et al. 2013), or for their metabolic abilities (Douglas and Klaenhammer 2010). Dissecting the genomic changes occurring during domestication and variety diversification has thus a fundamental importance for our understanding of evolutionary processes, in addition to applied interests for improving the desired traits in domesticated organisms and managing the germplasm diversity. Studying the footprints of adaptation in genomes may indeed allow to identify the important traits or metabolic pathways that were under selection during domestication and variety diversification, as well as the genetic bases of these traits (Wang et al. 1999, 2005; Whitt et al. 2002; Palaisa et al. 2003; Gallavotti et al. 2004; Yamasaki et al. 2005; Walsh 2008). Identifying the genomic regions involved may accelerate further improvement of traits controlling agricultural productivity and performance, such as yield, organoleptic or nutritional quality, and resistance to biotic and abiotic stresses, using marker-assisted selection (Soller 1994; Collard and Mackill 2008; Prada 2009). It may also help conservation management programs aiming at maintaining important functional biodiversity in core collections as well as in wild relatives of crop species.

The cultivated apple tree (*Malus domestica* Borkh.) is one of the most important fruit crops in temperate regions, with great economic and cultural values (Juniper and Mabberley 2006). Dessert apples are popular because of their

taste, nutritional properties, storability and convenience of use. The fruits of the specific varieties used to produce cider are smaller and bitter, as are those from crabapples, that is, the fruits of the wild apple species. The bitterness is due to a high content in polyphenols (Sanoner et al. 1999). Not all cider cultivars are, however, extremely bitter (Pereira-Lorenzo et al. 2009). Cider apples are also known for their fibrous structure, which allows longer storage (Lea and Piggott 2003; del Campo et al. 2005). In addition, cider apples more often display biennial bearing (Dapena et al. 2005), that is, with crop occurring only every two years. Finally, cider apples are more susceptible than dessert apples to fire blight, a disease caused by the bacteria *Erwinia amylovora* (Paulin et al. 1988; Lespinasse and Paulin 1990). Thousands of apple cultivars have been documented (Morgan et al. 2002), although only a few now dominate the market. Surprisingly, the history of apple domestication has just begun to be unraveled (Cornille et al. 2014). Genetic analyses have revealed a Central Asian origin of cultivated apple, with an initial divergence from the wild species *Malus sieversii*, together with an unexpectedly large secondary contribution through introgression from the European wild species *Malus sylvestris* (Velasco et al. 2010; Cornille et al. 2012). In contrast to expectations, cider cultivars did not appear the most introgressed by wild species based on microsatellites (Cornille et al. 2012). This suggests either a recent selection in the cider varieties for traits favorable for apple-based beverages from the standing genetic variation in the domesticated gene pool, or the introgression of only few genes from crabapples into the cider varieties. However, cider beverage has been produced for centuries in Western Europe especially by the Celts using native crabapples even before the invasion of the Romans who brought the domesticated apples. Much effort has been devoted since the 17th century in Europe to generate cider apple cultivars with high contents in sugar and polyphenols for producing high-quality cider (Morgan et al. 2002).

Although some *M. sieversii* individuals produce large apples, the variability in fruit size and color is wide. The selection by humans in cultivated apples targeted many phenotypic traits, including among others the number of fruits, their size, color, shape, flavor, taste, texture, storage capacity, harvesting ease, juvenile phase length and disease resistance (Janick 2005). QTL mapping has been used to dissect the genetic architecture of several desired traits, through crosses between cultivars (Calenge et al. 2004; Segura et al. 2008; Celton et al. 2011; Guitton et al. 2012; Longhi et al. 2012; Verdu et al. 2014). However, the footprints of selection have been little studied so far in apples compared to annual crops (Yamasaki et al. 2005; Camus-Kulandaivelu et al. 2008). The recently released 'Golden Delicious' genome sequence (Velasco et al. 2010) and the availability of medium-density genotyping tools (Chagne et al. 2012a) have made it possible to generate population-scale data for investigating genome-wide patterns of selection.

In this study, we set out to identify genomic regions under divergent selection between cider and dessert apples using two core collections, one of each variety type ($N = 48$ each), and 3704 SNP markers. First, we analyzed the population genetic structure in our sample to assess the differentiation between dessert and cider apple varieties using a much higher number of markers than in a previous study (Cornille et al. 2012). We also investigated the extent of linkage disequilibrium (LD) as a function of genomic distance within the genome to infer the expected maximal distance between the causal variation and the markers displaying association with the phenotype. We then looked at F_{ST} statistics for identifying outlier loci that would differentiate cider and dessert varieties significantly more than the average genomic background. Finally, a genome-wide association analysis was performed, taking into account genetic structure and kinship, contrasting the (i) cider versus dessert variety types or (ii) high versus low bitterness cultivars. Altogether, these analyses aimed at localizing the genomic regions that have been under divergent selection and responsible for the phenotypic differences between cider and dessert apples. We then examined in these regions the putative functions of genes to find candidates that have potentially undergone differential changes during the divergence between cider and dessert apples. Recent selection programs on cider apples aim at improving yield, regularity of production, resistance to pests and pathogens, while maintaining their specific technologic characteristics (e.g., high content in polyphenols). The identification of the genomic regions responsible for the differences between dessert and cider variety types could therefore be of great use for instance in a marker-assisted selection approach trying to select new cider varieties combining a higher content in polyphenols with the agronomic performances of dessert apples such as regular annual bearing, higher yield, and fruit size.

Material and methods

Plant material

The two apple core collections used in this study had been previously constituted by choosing the individuals that maximized the genetic diversity based on a set of 24 microsatellite markers in the INRA Angers germplasm collection of dessert and cider apple cultivars. Shortly, the core collections were built by retaining individuals from larger sets of apple accessions (737 and 188 for dessert and cider apples, respectively) using the 'Maximum Length Subtree' option of the DARwin software (Perrier et al. 2003; Perrier and

Jacquemoud-Collet 2006). The two core collections included 48 dessert and 48 cider apple cultivars, respectively (Supporting information). Reflecting the content of the INRA germplasm collection, both core collections mainly include old (generated before the 1950's) French apple cultivars, some of them being clones of cultivars grown in other European countries under different names. Because Western Europe has been the main place where the selection of dessert and cider apples has taken place (Morgan et al. 2002), the core collections we studied should be quite representative of the selection history of dessert and cider apples.

SNP arrays

Genomic DNA was extracted from leaves of the 96 individuals using the NucleoSpin® Plant II kit (Macherey-Nagel GmbH and Co KG, Düren, Germany). Because apple leaves are full of polysaccharides and phenols that contaminate the extracted DNA and may prevent hybridization on the array, DNA samples were purified as follows: 0.1 volume of sodium acetate (final concentration 0.3 M), 2.5 volumes of cold 100% ethanol, and 1 μL of glycogen were added, the tubes were centrifuged at 13 000 *g* for 30 min, the supernatant was discarded, 200 μL of 70% ethanol was added, the tubes were centrifuged at 13 000 *g* for 10 min, the supernatant was discarded, the tubes were air-dried overnight, and the DNA was resuspended in the appropriate volume of water. DNA samples were then checked for quality using Nanodrop 1000 (Thermo Scientific, Wilmington, DE, USA), quantified using PicoGreen® (Invitrogen, Grand Island, NY, USA), and processed onto the International RosBREED SNP Consortium (IRSC) apple 8k SNP array v1 (Chagne et al. 2012a) following the Illumina® protocol.

SNP filtering

SNPs were filtered using the Genotyping Module (version 1.8.4) of the Illumina® GenomeStudio software (Illumina Inc., San Diego, CA, USA). A visual inspection of each SNP was performed, and SNPs exhibiting a good genotypic clustering in distinct spots were kept. Paralogous SNPs were removed by performing BLAST onto the apple genome and removing probes having two equally good best hits onto the reference genome. This step was necessary for avoiding potential paralogy, due to the whole-genome duplication having occurred in the apple evolutionary history (Velasco et al. 2010). There were a few missing data in the dataset obtained from GenomeStudio, we therefore used fastPHASE 1.2 (Scheet and Stephens 2006) with the default parameters, and we indicated whether an individual belonged to the cider subgroup or to the dessert one to impute the missing data and to phase the SNPs belonging to a given linkage group (LG). Because the core collections were designed to maximize the genetic diversity and because SNPs for the 8k array were chosen among the most polymorphic markers in 27 dessert apple genomes (Chagne et al. 2012a), the allelic frequencies obtained may be biased compared to the full genetic pools of dessert and cider apples. Therefore, we excluded analyses based on the site frequency spectrum and focused only on analyses less sensitive to such biases.

Estimation of linkage disequilibrium

The levels of linkage disequilibrium were estimated using the r^2 parameter between all pairwise comparisons using the Haploview 4.2 software (Barrett et al. 2005) and a minor allele frequency (MAF) cutoff of 0.01. A first analysis was run without taking into account the structure and kinship in the collections; the levels of linkage disequilibrium were then corrected for population structure (see below) and kinship using the R package LDcorSV (Mangin et al. 2012). The kinship matrix, reflecting the degree of genetic covariance among individuals, was calculated with the Cocoa 1.1 software (Maenhout et al. 2009).

Analysis of population structure

The ADMIXTURE 1.23 software (Alexander et al. 2009) was used to investigate the genetic population structure in the dataset. The number of genetic clusters K was assessed using values ranging from 1 to 10, and we chose the number of clusters for which the cross-validation error was the lowest. The cross-validation procedure masks one-fifth of the genotypes (five runs altogether) and calculates estimates for these genotypes. Each genotype is then predicted, and the software calculates a prediction error across all masked genotypes. The Q matrix, that is, the posterior probabilities for each individual to belong to a given cluster, outputted by ADMIXTURE 1.23 was used for the genotype–phenotype association analysis.

Differentiation between cider and dessert apples – Detection of outlier loci

Pairwise single locus F_{ST} between the two core collections was calculated using either GENETIX 4.05 (Belkhir et al. 1996–2004) or BayeScan 2.1 (Foll and Gaggiotti 2008). The Bayesian method implemented in the latter (Beaumont and Balding 2004) was run to detect outlier loci using the following parameters: after 20 pilot runs of 50 000 iterations and an additional burn-in of 500 000 iterations, we used 3 000 000 iterations (thinning interval of 50 and sample size of 50 000).

Phenotype–genotype association

A genome-wide association study (GWAS) was run using the univariate linear mixed model (LMM) implemented in GEMMA (Zhou and Stephens 2012), taking into account the centered kinship matrix (K) calculated in GEMMA and the Q matrix from ADMIXTURE. A first analysis was performed on the two core collections by giving cider cultivars a score of 1 and dessert cultivars a score of 0. However, because not all cider cultivars are bitter, a second analysis was performed, this time not considering the cider versus dessert cultivars classification, but instead the bitterness of the cider apple cultivars, as recorded in the literature (Boré and Fleckinger 1997): bitter cider cultivars were given a score of 1 while sweet cider cultivars and dessert cultivars were given a score of 0. Both binary situations were treated as quantitative traits, as the linear mixed model is recognized as a robust approximation of a generalized linear model (Zhou et al. 2013). Markers were considered significantly associated with the phenotype for P-value $\leq 10^{-3}$. P-values obtained from GEMMA were used in R environment using the qqman package to generate a Manhattan plot (Turner 2014).

Identification of candidate genes

The online apple genome browser hosted on http://www.rosaceae.org/, containing the gene model predictions made on the apple genome sequence, was used to investigate the putative functions of genes present in the genomic regions detected in the tests above. The Blast2Go 3.0 software (Conesa et al. 2005) was used to perform BLASTX on these sequences with a maximum Blast ExpectValue of 10^{-3}. After gene ontology (GO) functional annotation, the KEGG tools were used to visualize the corresponding metabolic pathways. A BLASTN was run, and its results were used as inputs in Blast2GO 3.0 to retrieve GO annotations for the entire gene set of the apple genome. The regions of interest were then tested for enrichment of particular gene functions.

Results

SNP genotyping

After visually screening the 7867 SNPs of the IRSC apple 8K SNP array v1 on GenomeStudio, a set of 4234 polymorphic SNPs evenly spread across the apple genome was obtained; after removing potential paralogous SNPs, the number of markers was reduced down to 3704. The number of markers per linkage group was approximately proportional to their length. The average distance between two adjacent SNPs was 140 kb, with the maximum distance separating markers ranging from 1.26 Mb on LG17 to 4.25 Mb on LG15. The distribution of the SNP minor allele frequencies (MAF) was quite uniform across the different possible MAF values (Fig. 1) whether considering the cider or the dessert cultivars. Overall, few data were missing in the dataset, with 2981 markers having no missing data at all and the maximum percentage of missing data being 5.2% and 10.2% per marker and per individual, respectively. This made the inferences using fastPHASE 1.2 highly reliable.

Estimation of linkage disequilibrium

The nonlinear regression model used to analyze the decay of linkage disequilibrium (LD) with the physical distance showed that the squared allele correlation parameter r^2 decayed below 0.2 within 100 kb (Fig. 2). When analyzed separately, the cider and the dessert core collections showed very similar behaviors. The results obtained on the whole dataset when taking into account kinship or/and population structure were very similar too. We therefore assumed that loci distant from more than 100 kb were not in LD and considered windows of 100 kb on both sides of outlier SNPs for finding candidate genes possibly evolving under divergent selection.

Differentiation between cider and dessert apples – Analysis of population structure

Pairwise F_{ST} between cider and dessert apples ranged from 0 to 0.24, with a mean value of 0.014, confirming the weak differentiation between the two core collections. ADMIXTURE analyses revealed a minimum value of the cross-validation error for $K = 2$. Only a quarter of the individuals actually showed a clear assignment (membership probability >0.9) to any cluster, supporting the lack of further structure in the dataset. The Q matrix for $K = 2$ (Fig. 3) confirms the lack of strong differentiation according to the cider/dessert classification, even using genome-wide markers.

Detection of F_{ST} outlier loci and genotype–phenotype associations

Of the 3704 SNPs tested for their probability to have been under divergent selection using Bayescan 2.1, five exhibited significant genetic differentiation. These five outlier SNPs were located as follows: one SNP on LG08 at 11.32 Mb, two SNPs on LG15 at 26.38 Mb and 29.20 Mb, and two SNPs on LG17 at 10.30 Mb (Table 1). The GWAS testing SNP association with cider/dessert variety types revealed six SNPs with significant P-values (i.e., $-$Log10 P-value ≥ 3). These markers were located as follows: one SNP on LG05 at 19.23 Mb, two SNPs on LG08 at 13.05 Mb and

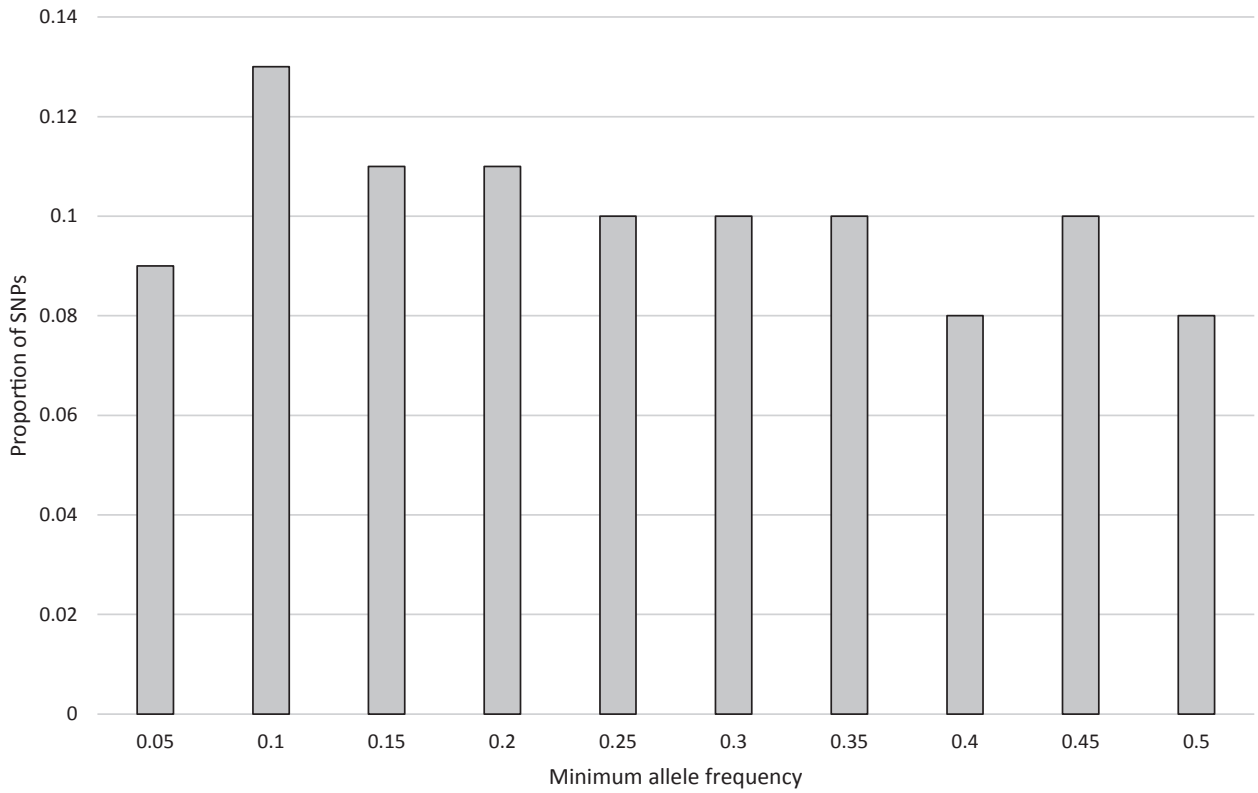

Figure 1 Allele frequency spectrum of the 3704 SNP markers of the array.

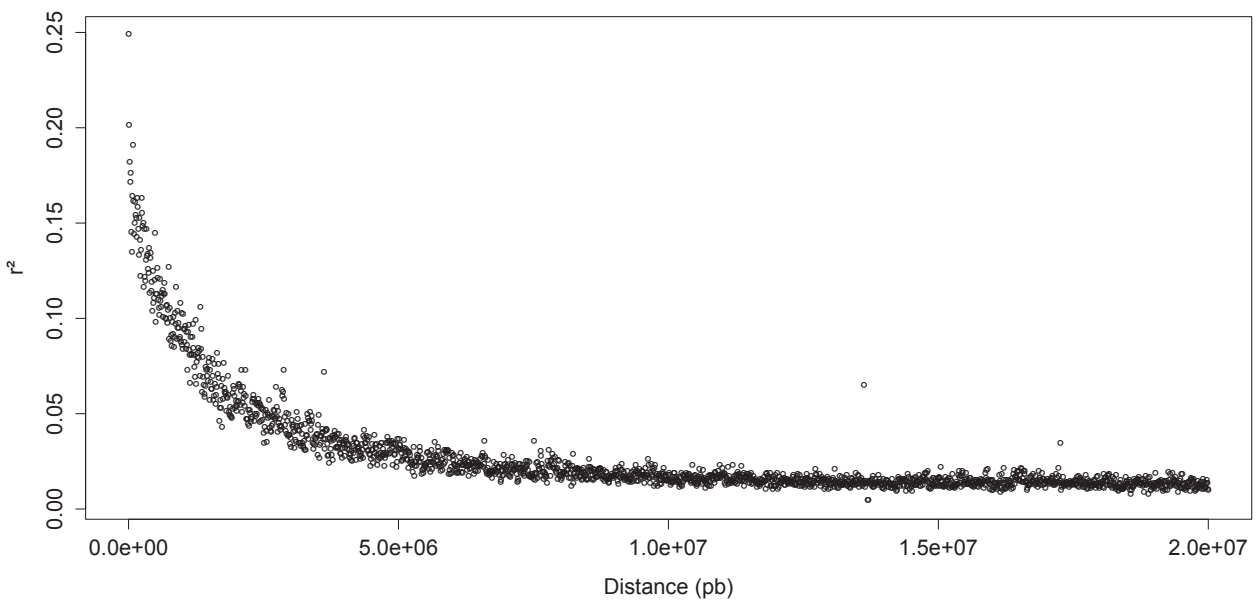

Figure 2 Decay of average linkage disequilibrium (measured as r^2) versus physical distance in increments of 10 000 bp. Both core collections, cider and dessert, were included in the analysis because no difference was observed when correcting for structure and/or kinship.

19.85 Mb, one SNP on LG09 at 29.70 Mb, one SNP on LG12 at 1.03 Mb, and one SNP on LG15 at 23.86 Mb (Table 2 and Fig. 4A). The bitter/sweet trait was found significantly associated with six SNPs located as follows: 2 SNPs on LG01, respectively located at 2.61 Mb and 2.67 Mb, one SNP on LG15 at 23.12 Mb, one SNP on

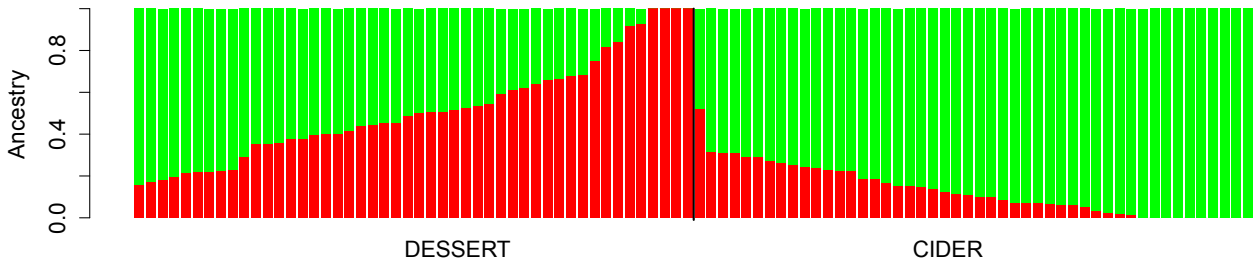

Figure 3 Population structure of 96 apple cultivars from the cider and dessert genetic pools. Membership probabilities were obtained with ADMIXTURE for $K = 2$. The bar plot, generated using the qqman package in R, shows each individual as a vertical bar.

Table 1. SNPs showing significant levels of F_{ST} detected by BayeScan 2.1.

SNP Name	LG	Position	F_{ST}
GDsnp01132	8	11 328 418	0.23
RosBREEDSNP_SNP_CA_29926704_Lg15_RosCOS1232_MAF50_MDP0000283141_exon1	15	26 329 550	0.23
RosBREEDSNP_SNP_AG_33667246_Lg15_01897_MAF40_151341_exon1	15	29 068 827	0.24
RosBREEDSNP_SNP_CT_10901071_Lg17_00918_MAF10_1668766_exon3	17	10 334 128	0.19
RosBREEDSNP_SNP_CT_10898449_Lg17_00918_MAF10_466062_exon6	17	10 336 750	0.19

LG, Linkage Group.

LG16 at 1.45 Mb, and two SNPs on LG17, respectively located at 8.42 Mb and 15.88 Mb (Table 2 and Fig. 4B).

Genes around candidate SNPs associated with phenotypes

We looked at the gene predictions available on the first version of the genome of apple within 200 kb around the seventeen SNPs detected above as putatively under diversifying selection. In the regions containing the five F_{ST} outlier loci detected by BayeScan, 85 predicted genes were found, whose main classes of putative functions are reported in Supporting information. In the 12 regions carrying the markers found to be associated with the cider/dessert or bitter/sweet phenotypes, 179 and 156 predicted genes were found respectively, whose main classes of putative functions are shown in Supporting information. Among these genes, the most represented biological processes were as follows: (i) amino acid metabolism and starch and sugar metabolism for the F_{ST} outliers, (ii) nucleotide metabolism and glycerolipid metabolism for the variety type associated regions, and (iii) purine metabolism and thiamine metabolism for the bitterness associated

Table 2. SNPs showing significant association with the cider/dessert or bitter/sweet phenotypes when taking into account structure and kinship between individuals using GEMMA.

SNP Name	LG	Position	*P*-value
RosBREEDSNP_SNP_CT_22024068_Lg5_RosCOS3072_MAF30_MDP0000753788_exon2*	5	19 238 624	3.83×10^{-4}
RosBREEDSNP_SNP_TC_15251985_Lg8_00354_MAF10_753213_exon1*	8	13 053 086	8.98×10^{-5}
RosBREEDSNP_SNP_TG_23835076_Lg8_RosCOS3331_MAF40_488673_exon1*	8	19 848 379	1.99×10^{-4}
RosBREEDSNP_SNP_GA_33077622_Lg9_01200_MAF20_MDP0000613052_exon1*	9	29 701 351	2.24×10^{-4}
RosBREEDSNP_SNP_GA_1240623_Lg12_RosCOS3293_MAF40_1686868_exon1*	12	1 033 191	4.12×10^{-4}
RosBREEDSNP_SNP_AG_27056933_Lg15_02084_MAF30_1677692_exon1*	15	23 859 694	2.06×10^{-4}
RosBREEDSNP_SNP_AG_32748739_Lg1_RosCOS2753_MAF10_520680_exon1†	1	26 153 648	1.86×10^{-5}
RosBREEDSNP_SNP_AC_33325153_Lg1_01951_MAF10_132337_exon1†	1	26 730 062	2.10×10^{-4}
GDsnp01850†	15	23 124 410	7.20×10^{-4}
RosBREEDSNP_SNP_AC_1452699_Lg16_MDP0000303483_MAF50_MDP0000303483_exon2†	16	1 452 699	2.78×10^{-4}
RosBREEDSNP_SNP_CT_8827345_Lg17_01842_MAF30_MDP0000891106_exon4†	17	8 427 545	7.87×10^{-5}
RosBREEDSNP_SNP_CT_17294445_Lg17_01964_MAF10_1662340_exon9†	17	1 588 1764	7.14×10^{-4}

LG, Linkage Group.
*SNP associated with the cider/dessert phenotype.
†SNP associated with the bitter/sweet phenotype.

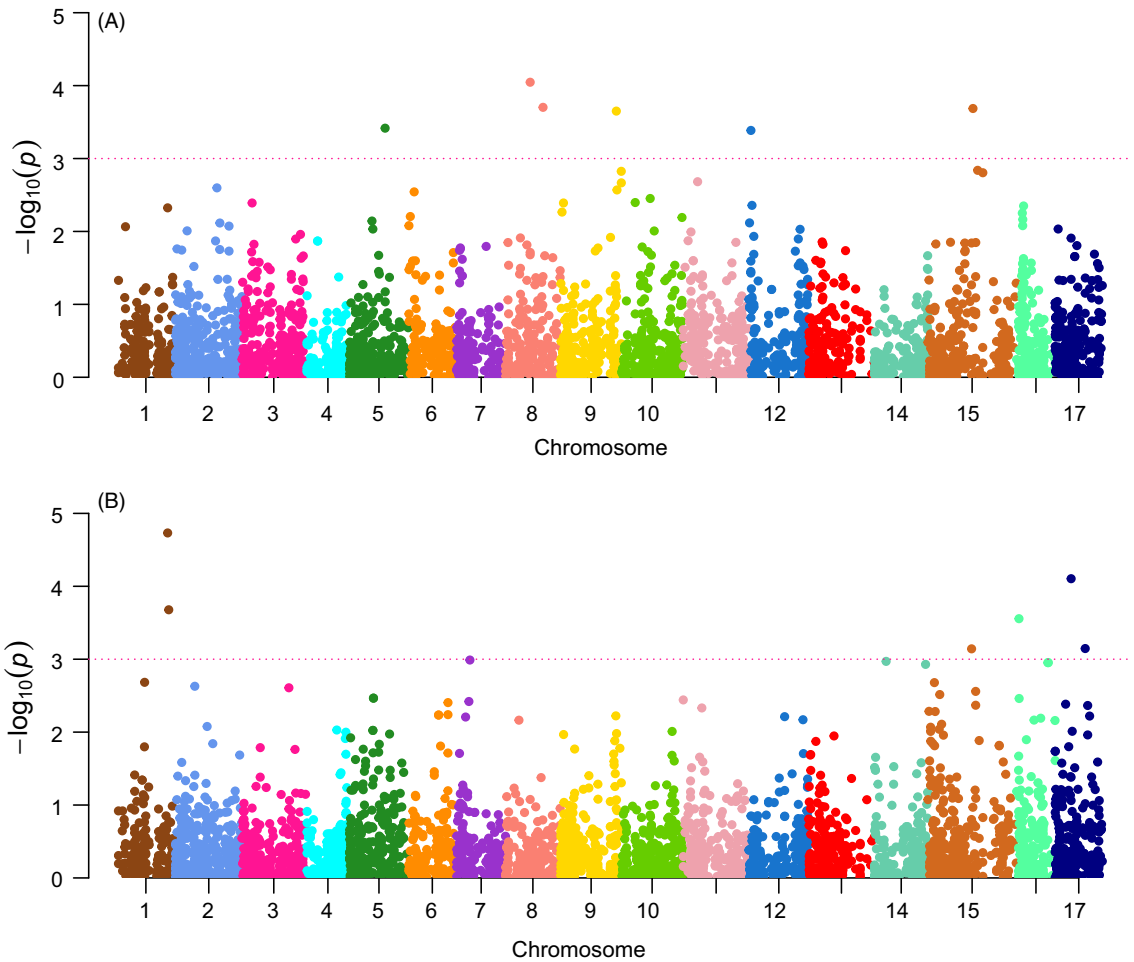

Figure 4 Manhattan plot of the GWAS testing for association between genotypes and the cider/dessert (A) or bitter/sweet phenotype (B). The −log10 of the *P*-value of 3704 SNPs after correction for structure and kinship is plotted against the physical position. SNPs above the blue line are those exhibiting significant *P*-values and thus associated with the cider/dessert or bitter/sweet phenotype.

regions. The enrichment test made using the entire predicted gene set as reference did not yield any significant result.

Discussion

Possible biases due to sample and marker choices

We used in this study core collections that maximize the genetic diversity present in larger initial collections, which may have generated biases in allelic frequencies. However, F_{ST} outliers and GWAS methods should be robust to such biases, and even conservative. Indeed, core collections balance the initial extreme allelic frequencies, so that association should be valid across even more diverse genotypes to be significant in core collections. The use of core collections instead of random sampling may in addition have led to an underestimation of linkage disequilibrium. Indeed, the increased distances between accessions within a core collection reflect an increased number of generations from the most recent common ancestor and thus a higher number of crossing-overs between linked loci (Nordborg and Tavare 2002). The LD values in the core collection are, however, again conservative and are actually the appropriate estimates to consider for the definition of the window size around the significant SNPs in the core collections.

Possible ascertainment biases in the SNP array design result from the choice of the markers among the most polymorphic SNPs based on genome resequencing of 27 dessert apple cultivars (Chagne et al. 2012a). The direct consequence is a more uniform distribution of the MAF spectrum than generally observed for resequencing data (Pe'er et al. 2006). This SNP ascertainment bias most probably led to overestimating the r^2 values (Nielsen and Signorovitch 2003; Nielsen 2004; Lachance and Tishkoff 2013) and thus the LD extent. In the end, the combined impact of the core collection sampling and the SNP ascertainment bias

on the LD estimation is difficult to assess. In addition, SNPs exhibiting contrasted frequency in the dessert and cider apple pools may have been discarded from the 8k apple array, even if they had a higher frequency in the cider apple gene pool, thus restricting the chance of detecting the corresponding genomic regions. These ascertainment biases, however, are again conservative: they may have led us to miss some genomic regions involved in cider versus dessert cultivars, but should not have yielded false positives. The regions detected here should therefore be considered as interesting candidates, but not an exhaustive list.

Low level of genomic differentiation between cider and dessert variety types

A previous study had reported a lack of population genetic structure between cider and dessert apples, using only a couple of dozen of microsatellite markers (Cornille et al. 2012). Our results confirm this result using a much higher number of markers of a different type (i.e., SNPs instead of SSR) along the genome, with no clear assignment of most of the different cultivars to either one or the other of the two inferred clusters according to their variety type. The low mean value of F_{ST} between cider and dessert apples (0.014 in our study) also supports the lack of genome-wide differentiation and is consistent with the mean F_{ST} value of 0.02 found by Cornille et al. (2012). Actually, some cultivars, discarded from the present study, are known to be used for both cider and dessert (e.g., Bagué Petit, Raccroupi, Cazo Jaune), which means the phenotypic classification in cider and dessert apples is not morphologically clear-cut either.

Long distance LD in the cultivated apple

The r^2 was found here to decrease below 0.2 within 100 kb. In previous studies on apples, r^2 was found decaying below 0.2 within 500 kb in a population of 7 full-sib families genotyped with 2500 SNPs (Kumar et al. 2012) and within 1 cM (corresponding to approximately 500 kb considering that the apple genome is 750 Mb and that the genetic map is 1500 cM long) in a collection of 132 apple cultivars genotyped with 238 SNPs (Micheletti et al. 2008). Such discrepancies with our study may be explained by a sampling of siblings in the former study therefore implying fewer recombination events than in a core collection encompassing a high diversity and several generations between individuals. In the latter study, a fewer number of SNPs, not spanning the entire genome, is also an explanation for a larger range of LD.

Linkage disequilibrium has also been studied in other *Rosaceae* crops such as *Prunus persica*, where r^2 reached 0.1 within 1200 kb in an Oriental peach germplasm (Li et al. 2013), and *Pyrus pyrifolia*, where r^2 fell below 0.2 at approximately 1800 kb in a population of old and modern cultivars, considering the pear genome is 600 Mb and 1100 cM long (Iwata et al. 2013). Studies performed on other allogamous tree species showed lower values of distances above which the LD decayed below 0.2: 200 bp in *Populus tremula* (Ingvarsson 2005) and approximately 2 kb in *Pinus taeda* L. (Brown et al. 2004). These levels of LD appear low compared to our results, probably because the studies were conducted on wild populations of forest trees, in which a much higher number of recombination events probably occurred since the last population bottleneck. In addition, the rather high average distance between our markers may have led to miss some occurrences of short-distance LD.

Differentiated genomic regions between cider and dessert apples

We identified here a total of 17 regions potentially bearing genes responsible for phenotypic differences between cider and dessert apples. Five of these regions harbored F_{ST} outlier loci that exhibited high differentiation levels between cider and dessert cultivars while the other twelve showed significant associations between the genotypic information and the variety type or the bitter trait while accounting for structure and kinship. According to the results on LD decay, 200 kb windows around the significant SNPs were investigated. The enrichment test performed on the three set of genes around significant SNPs, that is, F_{ST} outliers and the two association analyses results, did not detect any particularly overrepresented pathway. No genes known to be involved in the traits differentiating cider and dessert cultivars, such as the polyphenol pathway, were identified around the F_{ST} outliers located on LG08 and LG15. Two genes having high sequence similarity with UDP-glycosyltransferases were found around the two outlier markers on LG17. These genes can play a role in the synthesis pathways of several polyphenol compounds such as flavonoids or anthocyanidins, as exemplified by the *MdPT1* gene (Jugdé et al. 2008) involved in the glycosylation of phloretin into phlorizin, a major dihydrochalcone of apple known to have a bitter taste that may contribute to the peculiar flavor of cider (Whiting and Coggins 1975).

Regarding the results of the association between the genotypic information and the variety type, the two SNPs located on LG08 colocalized with QTLs linked to biennial bearing and yield (Guitton et al. 2012). Cider apples are in fact known to be more subject to biennial bearing than dessert apples (Dapena et al. 2005). However, no gene known to control any traits *a priori* differentiating cider and dessert apples was found within the genomic regions examined around the six SNPs detected as significantly associated

with the variety type. This may be due to lack of knowledge on these genes, and actually 13% of the genes did not have any predicted function. Alternatively, this may be because selection targeted the regulatory elements in the pathways. In fact, several genes coding for transcriptional regulation elements were found in these candidate regions. Finally, the estimation we made on the extent of LD may not reflect reality in these particular regions (as it is a genome-wide mean value we calculated) and could lead the causative factors for our outliers to be located outside of the windows examined.

All the six genomic regions identified when testing the association between the genotypes and the bitter/sweet phenotype were found to colocalize with QTLs responsible for the content of several polyphenolic compounds, either measured in the flesh or measured in the peel of the fruits (Chagne et al. 2012b; Khan et al. 2012b; Kumar et al. 2012; Verdu et al. 2014). The two SNPs located on LG01 colocalized with three QTLs responsible for *p*-coumaroyl quinic acid, hydroxycinnamic acid, and flavonols contents. The SNP located on LG15 colocalized with two QTLs responsible for flavonols and flavonols contents and the two SNPs on LG17, respectively, colocalized with QTLs responsible for quercetin 3-*O*-rutinoside and chlorogenic acid contents. The last area located on LG16 colocalized with a region well known to host several strong effect QTLs responsible for numerous polyphenolic compounds such as catechin, epicatechin, and procyanidins, all belonging to the flavonol class of polyphenols (Chagne et al. 2012b; Khan et al. 2012b). A gene coding for a *LeucoAnthocyanidin Reductase* (*LAR*) was identified underlying this QTL hotspot and is thought to be the gene responsible for the numerous QTLs in this area (Khan et al. 2012a). The *LAR* gene is indeed the one in the polyphenol pathway leading to the formation of the flavonols from leucocyanidin. Interestingly, the SNP significantly associated with the bitter/sweet phenotype and located on LG16 at 1.43 Mb was close to the *LAR* gene (MDP0000376284) located at 1.53 Mb, which makes our result highly consistent with this particular QTL hotspot and the *LAR* candidate gene. Altogether, the six SNPs associated to the bitter/sweet phenotype were located very close (less than 1 Mb on average) to the markers exhibiting the highest LOD score in the QTL analyses.

Applications in cider apple breeding

This study is a first step for the identification of the genetic bases of phenotypic traits that differentiate cider and dessert apple varieties. In addition, our markers will be useful for marker-assisted selection (MAS) for breeding cider varieties carrying both traits already present in cider varieties (such as high polyphenol content) and traits mainly present in dessert apple varieties (such as annual bearing, high yield, or disease resistance). Our markers can indeed guide both the choice of the cider apple progenitors and the selection of seedlings from crosses between dessert and cider varieties and thus segregating for the favorable haplotypes. By genotyping the seedlings of a cross between a cider and a dessert variety type at the loci we identified as linked with traits of interest, one could choose the individuals bearing the favorable alleles and keep the individuals combining traits from the two variety types. Another application of the information we described here could be the inventory of the several traits and genomic regions responsible for them to better manage germplasm diversity in the cultivated apple (Prada 2009).

Conclusions

Unraveling the genomic bases of quantitative trait variation is essential for understanding evolution and for accelerating plant breeding (Alonso-Blanco and Méndez-Vigo 2014). Furthermore, the question of sustainable management of germplasm resources is increasingly recognized as a fundamental goal to achieve in many crops (see the DivSeek initiative, http://www.divseek.org/). Recently, it has been suggested that an international consortium for the sustainable management of apple genetic diversity in particular is timely (Volk et al. 2014). Our results on the detection of a few key genomic regions involved in the phenotypic differentiation between cider and dessert apples emerging from an otherwise homogeneous genomic background should be very useful for designing such sustainable apple program. The identified outlier genomic regions will indeed be good targets for screening important genetic variation for conserving both cider and dessert apples specific traits. These programs should also focus on the sustainable conservation of the wild apple gene pools. Wild-to-crop introgressions have indeed been a key driver of the cultivated apple evolution, particularly through introgression from the European crabapple *M. sylvestris* (Cornille et al. 2012). It would be interesting to assess whether the outliers detected here in the cultivated apple have originated from such introgressions from the bitter crabapples. It would feature wild gene pools as sources of key genes for cultivated apple breeding in cider and dessert apples. Overall, our results thus illustrate how genomic can help to feed breeding and conservation programs, and a similar approach could be developed for detecting the genomic basis of other key traits, such as resistance to pathogens or climate adaptation.

Acknowledgements

We thank Philippe Guardiola and Anne Coutolleau of CHU Angers for the genotyping of the individuals using the Inter-

national RosBREED SNP Consortium (IRSC) apple 8K SNP array v1. We thank Laurence Feugey and Arnaud Guyader for providing access to the genetic resources of INRA Angers, UE HORTI for taking care of the plant material and the ANAN platform for DNA quantification. Diane Leforestier also thanks Thibault Leroy for discussions that helped improve this manuscript. TG, AC, and AB thank the BASC labex, the Région Ile de France (PICRI), and the Institut Diversité Ecologie et Evolution du Vivant (IDEEV).

Data archiving statement

The genotypic data have been deposited on http://www.rosaceae.org/search/diversity: Accession Number tfGDR1016.

Literature cited

Alexander, D. H., J. Novembre, and K. Lange 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. **19**:1655–1664.

Alonso-Blanco, C., and B. Méndez-Vigo 2014. Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. Current Opinion in Plant Biology **18**:37–43.

Axelsson, E., A. Ratnakumar, M. L. Arendt, K. Maqbool, M. T. Webster, M. Perloski, O. Liberg et al. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature **495**:360–364.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21**:263–265.

Beaumont, M. A., and D. J. Balding 2004. Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology **13**:969–980.

Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste, and F. Bonhomme 1996–2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier, France.

Boré, J. M., and J. Fleckinger 1997. Pommiers à cidre (variétés de France).

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America **101**:15255–15260.

Calenge, F., A. Faure, M. Goerre, C. Gebhardt, W. E. Van de Weg, L. Parisi, and C. E. Durel 2004. Quantitative Trait Loci (QTL) analysis reveals both broad-spectrum and isolate-specific QTL for scab resistance in an apple progeny challenged with eight isolates of *Venturia inaequalis*. Phytopathology **94**:370–379.

del Campo, G., J. I. Santos, I. Berregi, and A. Munduate 2005. Differentiation of Basque cider apple juices from different cultivars by means of chemometric techniques. Food Control **16**:549–555.

Camus-Kulandaivelu, L., L. M. Chevin, C. Tollon-Cordet, A. Charcosset, D. Manicacci, and M. I. Tenaillon 2008. Patterns of molecular evolution associated with two selective sweeps in the *Tb1–Dwarf8* region in maize. Genetics **180**:1107–1121.

Celton, J. M., S. Martinez, M. J. Jammes, A. Bechti, S. Salvi, J. M. Legave, and E. Costes 2011. Deciphering the genetic determinism of bud phenology in apple progenies: a new insight into chilling and heat requirement effects on flowering dates and positional candidate genes. The New Phytologist **192**:378–392.

Chagne, D., R. N. Crowhurst, M. Troggio, M. W. Davey, B. Gilmore, C. Lawley, S. Vanderzande et al. 2012a. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS ONE **7**:e31745.

Chagne, D., C. Krieger, M. Rassam, M. Sullivan, J. Fraser, C. André, M. Pindo et al. 2012b. QTL and candidate gene mapping for polyphenolic composition in apple fruit. BMC Plant Biology **12**:1–16.

Collard, B. C. Y., and D. J. Mackill 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences **363**:557–572.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**:3674–3676.

Cornille, A., T. Giraud, M. J. Smulders, I. Roldan-Ruiz, and P. Gladieux 2014. The domestication and evolutionary ecology of apples. Trends in Genetics **30**:57–65.

Cornille, A., P. Gladieux, M. J. M. Smulders, I. Roldán-Ruiz, F. Laurens, B. Le Cam, A. Nersesyan et al. 2012. New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genetics **8**:e1002703.

Dapena, E., M. Minarro, and M. D. Blazquez 2005. Organic cider-apple production in Asturias (NW Spain). IOBC wprs Bulletin **28**:161.

Darwin, C. 1987. 1856. Charles darwin's natural selection. In: R. C. Stauffer, ed. Species. Cambridge University Press, Cambridge.

Douglas, G. L., and T. R. Klaenhammer 2010. Genomic evolution of domesticated microorganisms. Annual Review of Food Science and Technology **1**:397–414.

Foll, M., and O. Gaggiotti 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics **180**:977–993.

Gallavotti, A., Q. Zhao, J. Kyozuka, R. B. Meeley, M. K. Ritter, J. F. Doebley, M. E. Pe et al. 2004. The role of *barren stalk1* in the architecture of maize. Nature **432**:630–635.

Guitton, B., J. J. Kelner, R. Velasco, S. E. Gardiner, D. Chagne, and E. Costes 2012. Genetic control of biennial bearing in apple. Journal of Experimental Botany **63**:131–149.

Ingvarsson, P. K. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., *Salicaceae*). Genetics **169**:945–953.

Iwata, H., T. Hayashi, S. Terakami, N. Takada, Y. Sawamura, and T. Yamamoto 2013. Potential assessment of genome-wide association study and genomic selection in Japanese pear *Pyrus pyrifolia*. Breeding Science **63**:125–140.

Janick, J. 2005. The origins of fruits, fruit growing, and fruit breeding. In J. Janick, ed. Plant Breeding Reviews, pp. 255–322. John Wiley & Sons, Inc, Hoboken, NJ.

Jugdé, H., D. Nguy, I. Moller, J. M. Cooney, and R. G. Atkinson 2008. Isolation and characterization of a novel glycosyltransferase that converts phloretin to phlorizin, a potent antioxidant in apple. FEBS journal **275**:3804–3814.

Juniper, B. E., and D. J. Mabberley 2006. The Story of the Apple. Timber Press, Portland, OR.

Khan, S. A., J. Schaart, J. Beekwilder, A. Allan, Y. Tikunov, E. Jacobsen, and H. Schouten 2012a. The mQTL hotspot on linkage group 16 for

phenolic compounds in apple fruits is probably the result of a leucoanthocyanidin reductase gene at that locus. BMC Research Notes **5**:618.

Khan, S. A., P.-Y. Chibon, R. C. H. de Vos, B. A. Schipper, E. Walraven, J. Beekwilder, T. van Dijk et al. 2012b. Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16. Journal of Experimental Botany **63**:2895–2908.

Kumar, S., D. Chagné, M. C. A. M. Bink, R. K. Volz, C. Whitworth, and C. Carlisle 2012. Genomic selection for fruit quality traits in apple (*Malus*×*domestica* Borkh.). PLoS ONE **7**:e36674.

Lachance, J., and S. A. Tishkoff 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays **35**:780–786.

Larson, G., and J. Burger 2013. A population genetics view of animal domestication. Trends in Genetics **29**:197–205.

Lea, A. G. H., and J. R. Piggott 2003. Fermented Beverage Production, 2nd edn. Kluwer Academic/Plenum Publishers, New York.

Lespinasse, Y., and J. P. Paulin 1990. Apple breeding programme for fire blight resistance: strategy used and first results. Acta Horticulturae (ISHS) **273**:285–296.

Li, X. W., X. Q. Meng, H. J. Jia, M. L. Yu, R. J. Ma, L. R. Wang, K. Cao et al. 2013. Peach genetic resources: diversity, population structure and linkage disequilibrium. BMC Genetics **14**:84.

Longhi, S., M. Moretto, R. Viola, R. Velasco, and F. Costa 2012. Comprehensive QTL mapping survey dissects the complex fruit texture physiology in apple (*Malus* x *domestica* Borkh.). Journal of Experimental Botany **63**:1107–1121.

Maenhout, S., B. De Baets, and G. Haesaert 2009. CoCoa: a software tool for estimating the coefficient of coancestry from multilocus genotype data. Bioinformatics **25**:2753–2754.

Mangin, B., A. Siberchicot, S. Nicolas, A. Doligez, P. This, and C. Cierco-Ayrolles 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity **108**:285–291.

McTavish, E. J., J. E. Decker, R. D. Schnabel, J. F. Taylor, and D. M. Hillis 2013. New World cattle show ancestry from multiple independent domestication events. Proceedings of the National Academy of Sciences of the United States of America **110**:E1398–E1406.

Meyer, R. S., A. E. DuVal, and H. R. Jensen 2012. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. The New Phytologist **196**:29–48.

Micheletti, D., F. Costa, P. Baldi, M. Troggio, M. Pindo, M. Komjanc, M. Malnoy et al. 2008. Linkage disequilibrium analysis to enable more efficient gene and QTL mapping in apple. *RGC4*.

Morgan, J., A. Richards, and E. Dowle 2002. The New Book of Apples: the Definitive Guide to Apples, Including Over 2000 Varieties. Ebury, London.

Nielsen, R. 2004. Population genetic analysis of ascertained SNP data. Human genomics **1**:218–224.

Nielsen, R., and J. Signorovitch 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theoretical Population Biology **63**:245–255.

Nordborg, M., and S. Tavare 2002. Linkage disequilibrium: what history has to tell us. Trends in Genetics **18**:83–90.

Palaisa, K. A., M. Morgante, M. Williams, and A. Rafalski 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. The Plant Cell **15**:1795–1806.

Paulin, J. P., G. Lachaud, R. Chartier, and J. M. Bore 1988. Sensibilité au feu bactérien de variétés de pommiers à cidre. Résultats de 3 années d'expérimentation.35.

Pe'er, I., Y. R. Chretien, P. I. de Bakker, J. C. Barrett, M. J. Daly, and D. M. Altshuler 2006. Biases and reconciliation in estimates of linkage disequilibrium in the human genome. American Journal of Human Genetics **78**:588–603.

Pereira-Lorenzo, S., A. M. Ramos-Cabrer, and M. Fischer 2009. Breeding Apple (*Malus* x *Domestica* Borkh). In: S. M Jain and P. M. Priyadarshan, eds. Breeding Plantation Tree Crops: Temperate Species. pp. 33–81. Springer, New York.

Perrier, X., A. Flori, and F. Bonnot 2003. Data analysis methods. In P. Hamon, M. Seguin, X. Perrier, and J. C. Glaszmann, eds. Genetic Diversity of Cultivated Tropical Plants. pp. 43–76. Science Publishers, Inc., Montpellier, France.

Perrier, X., and J. P. Jacquemoud-Collet 2006. DARwin software v. 6.0.010. http://darwin.cirad.fr/. (accessed on 21 April 2015).

Prada, D. 2009. Molecular population genetics and agronomic alleles in seed banks: searching for a needle in a haystack? Journal of Experimental Botany **60**:2541–2552.

Sanoner, P., S. Guyot, N. Marnet, D. Molle, and J. P. Drilleau 1999. Polyphenol profiles of French cider apple varieties (*Malus domestica* sp.). Journal of Agricultural and Food Chemistry **47**:4847–4853.

Scheet, P., and M. Stephens 2006. 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 **78**:629–644.

Segura, V., C. Cilas, and E. Costes 2008. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects: mixed linear modelling of repeated spatial and temporal measures. The New Phytologist **178**:302–314.

Soller, M. 1994. Marker assisted selection – an overview. Animal Biotechnology **5**:193–207.

Turner, S. D. 2014. qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots (http://biorxiv.org).

Velasco, R., A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana et al. 2010. The genome of the domesticated apple (*Malus* x *domestica* Borkh.). Nature Genetics **42**:833–839.

Verdu, C. F., S. Guyot, N. Childebrand, M. Bahut, J. M. Celton, S. Gaillard, P. Lasserre-Zuber et al. 2014. QTL analysis and candidate gene mapping for the polyphenol content in cider apple. PLoS ONE **9**: e107103.

Volk, G. M., C. T. Chao, J. Norelli, S. K. Brown, G. Fazio, C. Peace, J. McFerson et al. 2014. The vulnerability of US apple (*Malus*) genetic resources. Genetic Resources and Crop Evolution 1–30.

Walsh, B. 2008. Using molecular markers for detecting domestication, improvement, and adaptation genes. Euphytica **161**:1–17.

Wang, H., T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigouroux, M. Faller, K. Bomblies et al. 2005. The origin of the naked grains of maize. Nature **436**:714–719.

Wang, R. L., A. Stec, J. Hey, L. Lukens, and J. F. Doebley 1999. The limits of selection during maize domestication. Nature **398**:236–239.

Whiting, G. C., and R. A. Coggins 1975. Estimation of the monomeric phenolics of ciders. Journal of the Science of Food and Agriculture **26**:1833–1838.

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut, and E. S. Buckler 2002. Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences **99**:12959–12962.

Yamasaki, M., M. I. Tenaillon, I. V. Bi, S. G. Schroeder, H. Sanchez-Villeda, J. F. Doebley, B. S. Gaut et al. 2005. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. The Plant Cell **17**: 2859–2872.

Yuan, J. H., A. Cornille, T. Giraud, F. Y. Cheng, and Y. H. Hu 2014. Independent domestications of cultivated tree peonies from different wild peony species. Molecular Ecology **23**:82–95.

Zhou, X., and M. Stephens 2012. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics **44**:821–824.

Zhou, X., P. Carbonetto, and M. Stephens 2013. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics **9**: e1003264.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Names of the 96 cultivars chosen from the INRA Angers collections of old cider and dessert apple varieties.

Table S2. Results from the Blast2GO software on the genes identified around the significant SNPs.