ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2014) **12**(3); 330–340
doi:10.1017/S1479262114000136

Molecular-level and trait-level differentiation between the cultivated apple (*Malus* × *domestica* Borkh.) and its main progenitor *Malus sieversii*

Satish Kumar[1]*, Pierre Raulier[2], David Chagné[3] and Claire Whitworth[1]

[1]*The New Zealand Institute for Plant and Food Research Limited, Private Bag 1401, Havelock North 4157, New Zealand,* [2]*Université Catholique de Louvain, 1348 Louvain-La-Neuve, Belgium and* [3]*The New Zealand Institute for Plant and Food Research Limited, Private Bag 11600, Palmerston North, New Zealand*

Received 4 December 2013; Accepted 14 February 2014 – First published online 25 March 2014

Abstract

The present study is the first to compare the trait-level differentiation (Q_{st}) and the molecular-level differentiation (F_{st}) between *Malus* × *domestica* and *Malus sieversii*. A set of 115 accessions representing *M.* × *domestica* (99) and *M. sieversii* (16) were genotyped using the International RosBREED SNP Consortium apple 8K SNP array and phenotyped for eight fruit quality traits in a clonally replicated experiment. A set of 3521 single nucleotide polymorphisms (SNPs) with an average call rate of 98% was retained following SNP data quality filters. About 86% of the total SNPs were polymorphic in *M. sieversii*, while all but three SNPs were polymorphic in *M.* × *domestica*. The patterns of linkage disequilibrium were different, especially at the longer distances, between the two species. No differentiation ($F_{st} = 0$) was observed for nearly 23% of the SNPs, but about 20% of the SNPs exhibited a high genetic differentiation ($F_{st} \geq 0.15$). A highly significant ($P < 0.001$) genome-level $F_{st} = 0.12$ was observed between *M.* × *domestica* and *M. sieversii*. The average estimated Q_{st} value was 0.20 (range 0.08–0.40), and for three of the eight studied traits (crispness, flavour intensity and fruit weight), Q_{st} value was more than twice the estimated genome-level F_{st} value. A higher Q_{st} value than F_{st} value for four of the eight fruit quality traits indicated differential (or directional) selection for these traits in *M.* × *domestica*. The average posterior probability of assignment of *M.* × *domestica* accessions to the *M. sieversii* gene pool was 11%, supporting the hypothesis of *M. sieversii* being one of the progenitors of the domesticated apple.

Keywords: admixture; diversity; germplasm; *Malus* × *domestica*; *Malus sieversii*; SNP

Introduction

The cultivated apple, *Malus* × *domestica*, belongs to the Rosaceae family, and the genus *Malus* is reported to consist of 25 to 30 species. Early studies based on morphological characteristics suggested that several *Malus* species, including *M. sylvestris* Miller, *M. prunifolia* (Willd.) Borkh., and *M. baccata* (L.) Borkh., were involved in the origin and/or domestication of the cultivated apple (Brown, 1975). As Asian *M.* × *asiatica*, *M. baccata*, *M. micromalus*, *M. orientalis*, *M. prunifolia* and *M. sieversii* and European *M. sylvestris* are the species taxonomically closest to *M.* × *domestica*, they are considered to have contributed to the domestic gene pool to differing extents (Robinson *et al.*, 2001). A survey of molecular differences at 23 genes across

* Corresponding author. E-mail: satish.kumar@plantandfood.co.nz

the genus *Malus* supported the proposal that the *M.* × *domestica* gene pool was formed mainly from *M. sieversii* (Velasco *et al.*, 2010). However, when Harrison and Harrison (2011) re-analysed some of the polymorphism data from the Velasco *et al.* (2010) study, they concluded that gene flow from *M. sylvestris* to *M.* × *domestica* could not be ruled out, in line with previous suggestions (Harris *et al.*, 2002; Coart *et al.*, 2006). The modern cultivated apple gene pool mimics an admixed population whereby an individual cultivar might have different degrees of genetic heritage from different *Malus* species.

Malus germplasm has been characterized in several ways, including taxonomic, morphological, and agronomic characterization (Chapman, 1989; Hilu, 1989), biochemical analysis (Doebley, 1989; Røen *et al.*, 2009) and DNA marker-based analysis (Clegg, 1990; Dunemann *et al.*, 1994; Hokanson *et al.*, 1998; Coart *et al.*, 2006; Gharghani *et al.*, 2009; Van Treuren *et al.*, 2010). The relationship between molecular and quantitative measures of genetic diversity is not always straightforward (Frankham, 1999). For example, from a survey of 29 species, McKay and Latta (2002) reported that population differentiation based on molecular markers [i.e. F_{st} (Cockerham and Weir, 1993)] is poorly correlated with quantitative variation at the trait level [i.e. Q_{st} (Spitze, 1993)]. No formal comparisons of F_{st} and Q_{st} have ever been reported for *Malus* species.

A recent study (Cornille *et al.*, 2012) using 26 evenly distributed simple sequence repeat (SSR) markers has presented a detailed analysis of population structure in apple germplasm, but it has been suggested that the use of a large number of single nucleotide polymorphisms (SNPs) will outperform that of SSR markers for population structure analysis (Liu *et al.*, 2005; Helyar *et al.*, 2011). We present here an application of the International RosBREED SNP Consortium (IRSC) apple 8K SNP array (Chagné *et al.*, 2012) to investigate population structure, linkage disequilibrium (LD) and genetic differentiation in an apple germplasm collection comprising *M.* × *domestica* and *M. sieversii* individuals. To our knowledge, this is the first attempt made to directly compare trait-level differentiation and marker-level divergence between *Malus* species. We also tested the marker 'neutrality', which is a desirable feature of markers for species (or population) divergence studies, of our SNP array by conducting a genome-wide association analysis for various important apple fruit quality traits.

Materials and methods

Plant material and phenotypes

In an attempt to expand the *Malus* gene pool, open-pollinated (OP) seeds of nearly 500 apple cultivars and some wild species were imported from around the world and planted in New Zealand during the early 1990s (Noiton *et al.*, 1999; Kumar *et al.*, 2010). From about 35,000 OP seedlings, a core subset of 350 seedlings were identified (using data on trees, fruits and disease resistance traits) to represent the diversity found within the entire collection. Apart from a few exceptions, only one seedling from each maternal family was kept for long-term conservation. All the 350 seedlings were then clonally propagated on to 'M.27' rootstock and two copies (or clones) of each genotype were planted in a common-garden experiment in 2003 in a Plant and Food Research orchard (39°C; 39′S 176°53′E) at Hawke's Bay Research Centre, New Zealand. For the present study, 115 of these 350 individuals were used, representing 99 and 16 individuals of *M.* × *domestica* and *M. sieversii*, respectively. The geographical origin of these 115 individuals, along with the name of their maternal parent, is given in Table S1 (available online). OP seeds of *M. sieversii* accessions were mainly collected from Kazakhstan, but those of *M.* × *domestica* were from different geographical regions including France, the Netherlands, Russia, Scandinavia, the UK and the USA (Table S1, available online).

Fruit harvesting, which was carried out in the fruiting season (February–May) in 2011, began when fruits were judged to be mature, based on a change in skin background colour from green to yellow, and when the starch pattern index was between 3 and 4 (Brookfield *et al.*, 1997). Fruits were harvested and analysed separately from both trees propagated from each original genotype. Samples of six fruits from each seedling were stored for 42 d at 0.5°C and then for a further 1 d at 20°C before evaluation. Phenotypic information on eight traits describing visual and sensory fruit properties was obtained. Detailed protocols for the assessment of a range of traits have been reported previously (Kumar *et al.*, 2010). Briefly, crispness (CRISP), juiciness (JUICE), flavour intensity (FINT), over-colour amount (OCOL) and russet coverage (RUS) were scored on a scale of 0 (none) to 9 (highest). Fruit weight (WT) was measured as the average weight (g) of six fruits from each seedling. Bulked juice from the cortical flesh of the sample fruit was used to measure titratable acidity (TA) using an automatic acid titrator (Metrohm 716 DMS, Herisau, Switzerland) and the percentage of malic acid in fruit juice was recorded. Soluble solid concentration (SSC) for each fruit was measured using a digital refractometer (Atago PR-32, Tokyo, Japan).

Trait-level differentiation

Estimates of variance among individuals within a species and between the two species were obtained for each trait using the following model:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zp} + \boldsymbol{Wg}/\boldsymbol{p} + \boldsymbol{e}, \quad (1)$$

https://doi.org/10.1017/S1479262114000136 Published online by Cambridge University Press

where $\boldsymbol{y}$ is a vector of observations on each seedling, $\boldsymbol{b}$ is a vector of fixed effects (i.e. overall mean, row effect), $\boldsymbol{p}$ is a vector of random effect of species, $\boldsymbol{g/p}$ is a vector of random effect of genotypes within a species and $\boldsymbol{e}$ is a vector of random residual terms. $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ are the known incidence matrices relating the observations in $\boldsymbol{y}$ to effects in $\boldsymbol{b}$, $\boldsymbol{p}$ and $\boldsymbol{g/p}$, respectively. The variances associated with the random effects $\boldsymbol{p}$, $\boldsymbol{g/p}$ and $\boldsymbol{e}$ were σ^2_B, σ^2_W and, σ^2_e respectively. Equation (1) was implemented in the ASReml software (Gilmour *et al.*, 2006). We calculated estimates of broad-sense heritability (H^2) and Q_{st} (following Spitze, 1993) for each trait from the following:

$$H^2 = \frac{\sigma^2_B + \sigma^2_W}{\sigma^2_B + \sigma^2_W + \sigma^2_e} \quad (2)$$

$$Q_{st} = \frac{\sigma^2_B}{\sigma^2_B + 2\sigma^2_W} \quad (3)$$

The approximate standard errors of H^2 and Q_{st} were obtained using the delta method as implemented in the ASReml software.

SNP genotyping and quality control

Seedlings were genotyped using the IRSC apple 8K array v1 (Chagné *et al.*, 2012), based on the Infinium® II technique. Genomic DNA (gDNA) was extracted from each seedling using the NucleoSpin® Plant II kit (Macherey-Nagel GmbH and Co KG, Düren, Germany) and quantified using the Quant-iT™ PicoGreen® Assay Kit (Invitrogen, Eugene, USA). For this reaction, 200 ng of gDNA were used as a template, following the manufacturer's instructions. SNP genotypes were scored using the Genotyping Module (version 1.8.4) of the Illumina® GenomeStudio software (Illumina, Inc., San Diego, USA). The reliability of each genotype call was measured using the *GenCall* score set at a minimum of 0.15, which is a lower bound for calling genotypes relative to its associated cluster. SNPs were subsequently discarded using a sequence of criteria in the following order: *GenCall* score at the 50% rank (*50% GC*) <0.40, cluster separation (*ClusterSep*) <0.25, more than 5% missing calls, and segregation discrepancy. The BEAGLE 3.1 software (Browning and Browning, 2007) was then used for imputing missing SNP genotypes.

Population structure and LD

Population structure was investigated using the Bayesian clustering method implemented in STRUCTURE (Pritchard *et al.*, 2000), which uses Markov chain Monte Carlo (MCMC) simulations to infer the proportion of ancestry of genotypes in K distinct predefined clusters. Ten independent runs were carried out, and we used 50,000 MCMC iterations after a burn-in of 5000 steps.

We calculated pairwise LD between SNPs to evaluate the extent of LD in apple germplasm. The degree of LD was quantified with the parameter r^2 (Hill and Robertson, 1968) obtained with and without taking into account the population structure and cryptic relatedness following the method of Mangin *et al.* (2012). Estimates of coefficients of relationships (i.e. cryptic relatedness) among all the seedlings were calculated using SNP data following the method of Van Raden (2008). Estimates of r^2, accounting for population structure and cryptic relatedness, were obtained using the R software LDcorSV (Mangin *et al.*, 2012).

Molecular-level differentiation

For evaluating molecular-level differentiation between the two species, F_{st} value at each locus, and also across all the loci, was calculated following the method of Weir and Cockerham (1984). The significance of F_{st} values was assessed using Fisher's exact tests implemented in GENEPOP 4.2 (Rousset, 2008). The significance of the minor allele frequency (MAF) difference between the species was compared using the Mann–Whitney U test.

Marker–trait association analysis

A mixed linear model approach (Yu *et al.*, 2006) that accounts for multiple degrees of relatedness (population structure and cryptic relationships) was used:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{e}, \quad (4)$$

where $\boldsymbol{y}$ is a vector of observations on seedlings, $\boldsymbol{\beta}$ is an unknown vector containing fixed effects, including a genetic marker, population structure ($\boldsymbol{Q}$), and the intercept, and $\boldsymbol{X}$ is the known design matrix related to $\boldsymbol{\beta}$. $\boldsymbol{Z}$ is the known design matrix related to $\boldsymbol{a}$, the unknown vector of random additive genetic effects with $\boldsymbol{a} \sim N(0, \boldsymbol{G}\sigma^2_a)$. The realized relationship coefficient matrix ($\boldsymbol{G}$) was derived using all the available SNPs following the method of Van Raden (2008). The scalar σ^2_a is the additive variance and $\boldsymbol{e}$ is a vector of independent random deviates with variance σ^2_e. Equation (4) was fitted using the GAPIT software (Lipka *et al.*, 2012). To avoid spurious associations that could arise from population structure, we included principal components (PCs) as covariates (i.e. $\boldsymbol{Q}$ matrix). In equation (4), each SNP was tested in turn using a t test (H_0: no additive association between the SNP and the trait) and P values were obtained. A genome-wide significance threshold of

https://doi.org/10.1017/S1479262114000136 Published online by Cambridge University Press

Table 1. Percentage of variance explained by different sources, quantitative trait-level differentiation (Q_{st}) and broad-sense heritability (H^2) for different traits[a]

Traits	Between species	Genotypes within a species	Residual	Q_{st}	H^2
CRISP	0.48	0.36	0.16	0.40 (0.35)	0.84 (0.12)
FINT	0.48	0.38	0.14	0.25 (0.28)	0.86 (0.09)
JUICE	0.21	0.53	0.26	0.14 (0.20)	0.74 (0.09)
OCOL	0.36	0.58	0.05	0.12 (0.18)	0.94 (0.02)
RUS	0.13	0.74	0.13	0.08 (0.14)	0.87 (0.04)
SSC	0.21	0.60	0.19	0.15 (0.18)	0.81 (0.10)
TA	0.30	0.58	0.12	0.19 (0.24)	0.88 (0.06)
WT	0.51	0.44	0.05	0.33 (0.30)	0.95 (0.03)

CRISP, crispness; FINT, flavour intensity; JUICE, juiciness; OCOL, over-colour; RUS, russet; SSC, soluble solid concentration; TA, titratable acidity; WT, fruit weight.
[a] Approximate standard errors of Q_{st} and H^2 estimates are given in parentheses.

$\alpha = 0.05$ (which equates to a marker-wise threshold $p = 1 \times 10^{-5}$ as we tested 3521 SNPs) was used to identify significant marker–trait associations for all traits.

Results

Trait-level differentiation

The percentage of total variance for a trait attributed to species differences varied from 13% for RUS to 51% for WT with an average of 33% (Table 1). Differences among individuals within a species accounted for a larger proportion of variation ranging from 36% for CRISP to 74% for RUS with an average of 53%. The degree of trait-level genetic differentiation (Q_{st}) between the two species ranged widely from 0.08 for RUS to 0.40 for CRISP with an average of 0.21 (Table 1). Juiciness (JUICE) had the lowest (0.73) broad-sense heritability (H^2) and WT had the highest (0.95).

SNP genotyping and molecular-level species differentiation

Following various SNP data quality filters, a set of 3521 SNPs with an average call rate of 98% was retained for further analyses. The MAF at these selected loci ranged from 0.01 to 0.50 with an average of 0.24. The retained SNPs were evenly spread across the apple genome; that is, the proportion (out of 3521) of SNPs on any given linkage group (LG) was generally similar to the relative size of that LG assuming the total genome size as 742 Mb. The accuracy of imputation using the BEAGLE software was high for a majority of SNPs; that is, the accuracy was >0.95 for 95% of the 3521 SNPs.

Of the 3521 SNPs, 3038 (86.3%) and 3518 (99.9%) SNPs were polymorphic in *M. sieversii* and *M.* × *domestica*, respectively. The median of the MAF was significantly lower in *M. sieversii* (0.13) than in *M.* × *domestica* (0.23). The observed and expected heterozygosity was 0.19 and 0.23, respectively, in *M. sieversii* and 0.35 and 0.35,

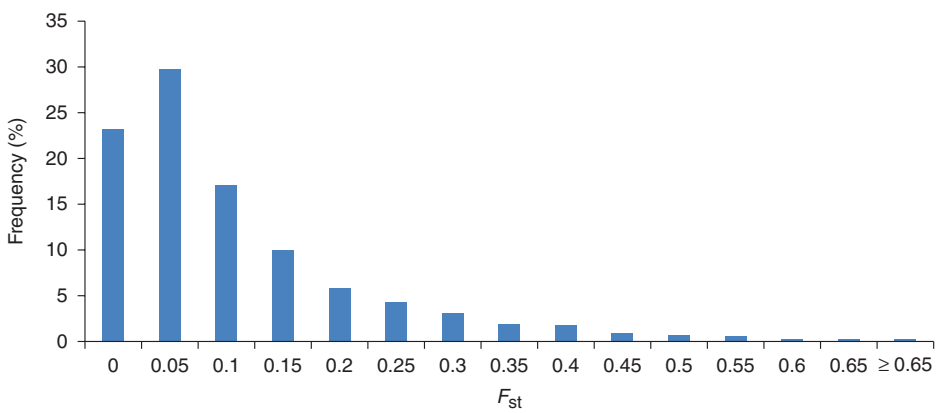

Fig. 1. Genome-wide distribution of F_{st} values.

https://doi.org/10.1017/S1479262114000136 Published online by Cambridge University Press

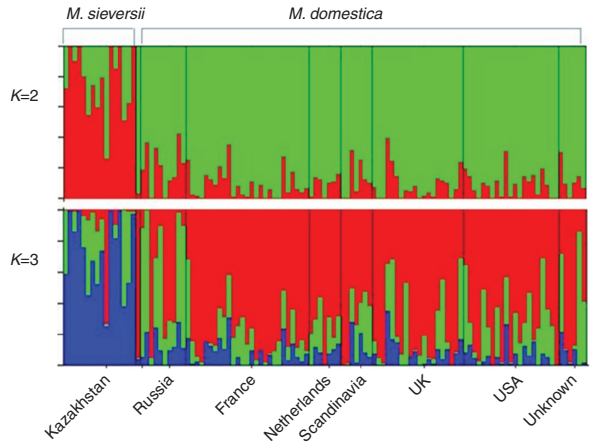

Fig. 2. Model-based Bayesian clustering of 115 accessions of the two *Malus* species using the STRUCTURE software. Each accession's genome is represented by a single vertical line, which is partitioned into coloured segments in proportion to the estimated membership of a species. Results shown are for varying numbers of clusters (i.e. $K = 2$ or 3), so the *y*-axis indicates the posterior probability for assignment to different species or clusters. Black vertical lines separate individuals of different geographical origin or country of origin.

respectively, in *M.* × *domestica*. The maximum expected heterozygosity that can be reached at a biallelic SNP is 0.5. The estimate of F_{IS} value was higher in *M. sieversii* (0.19) than in *M.* × *domestica* (0.10), suggesting somewhat higher inbreeding in the former. Genetic differentiation between the two *Malus* species was determined by calculating pairwise F_{st} at each SNP (Fig. 1). No differentiation ($F_{st} = 0$) was observed for nearly 23% of the SNPs, but about 20% of the SNPs exhibited a high genetic differentiation ($F_{st} \geq 0.15$). The average F_{st} across all the SNPs was 0.12, which was highly significant ($P < 0.001$) as assessed by Fisher's exact test.

Population structure and LD

We used the 'admixture model' implemented in the STRUCTURE software to infer population structure and introgression (Fig. 2 and Table 2). A pairwise comparison assuming two clusters (i.e. $K = 2$) indicated that the average posterior probability of assignment of *M.* × *domestica* accessions to the *M. sieversii* gene pool was 11%, whereas the average probability for the reverse was 23% (Table 2). *M.* × *domestica* accessions from Russia displayed a relatively high degree of admixture (20%) with *M. sieversii* compared with other seed sources, for which the extent of admixture varied from 9 to 12% (Table 2).

The ΔK statistic, designed to identify the most relevant number of clusters by determining the number of clusters beyond which there is no further increase in likelihood, was greatest for $K = 3$ ($\Delta K = 3897$, $\Pr|\ln L = -386566$). For $K = 3$, the mean estimated membership of *M.* × *domestica* individuals to the three gene pools, namely *M.* × *domestica*, *M. sieversii* and an unknown, was 0.68, 0.06 and 0.26%, respectively (Fig. 2). Assuming that *M.* × *domestica* individuals with membership probability ≥0.20 (an arbitrary threshold; Cornille *et al.*, 2012) of a non-*domestica* gene pool displayed introgression, it was inferred from the results that 10 and 38% of the accessions displayed introgression from *M. sieversii* and an unknown gene pool, respectively. About 9% of the *M.* × *domestica* accessions displayed introgression from *M. sieversii* as well as from an unknown gene pool.

The pattern of LD decay was somewhat different for the two species. Compared with that in *M.* × *domestica*, LD in *M. sieversii* was lower at short distances, but higher at long distances (Fig. 3(a)). The results of the combined analysis revealed a moderate degree of LD at various distances between the markers; for example, the average r^2 for SNPs separated by 0.1 Mb, 0.5 Mb (approximately 1 cM in apple) and 1.0 Mb was 0.13, 0.08 and 0.06, respectively (Fig. 3(b)). LD was slightly lower after adjusting for population structure, but the pattern of LD decay was quite similar to that without accounting for structure.

Genome-wide SNP–trait associations

The numbers of PCs, which constitute the $\boldsymbol{Q}$ matrix to account for population structure, varied between the traits: OCOL = 0; SSC = 1; CRISP = 2; WT and TA = 3; and FINT, JUICE, and RUS = 4. The profiles of *P* values, in terms of $-log_{10}(p)$, for all the tested SNPs for each trait are shown in Fig. 4. Clustering of significant SNPs for OCOL (on LG9) and TA (on LG16) suggested the presence of large-effect quantitative trait locus (QTLs) at these positions. The largest-effect SNP (ss475879551), explaining 23% variation in OCOL, is located at a distance of 0.11 Mb from the *MdMYB10* gene (MDP0000259616)

Table 2. Proportions of estimated membership (using 'admixture' model in STRUCTURE) to each of the two inferred gene pools for accessions of different origins

Country of origin	Gene pool 1 (*Malus sieversii*)	Gene pool 2 (*Malus domestica*)
M. domestica – Russia	0.20	0.80
M. domestica – France	0.09	0.91
M. domestica – Netherlands	0.10	0.90
M. domestica – Scandinavia	0.12	0.88
M. domestica – UK	0.09	0.91
M. domestica – USA	0.09	0.91
Total – *M. domestica*	0.11	0.89
Total – *M. sieversii*	0.77	0.23

https://doi.org/10.1017/S1479262114000136 Published online by Cambridge University Press

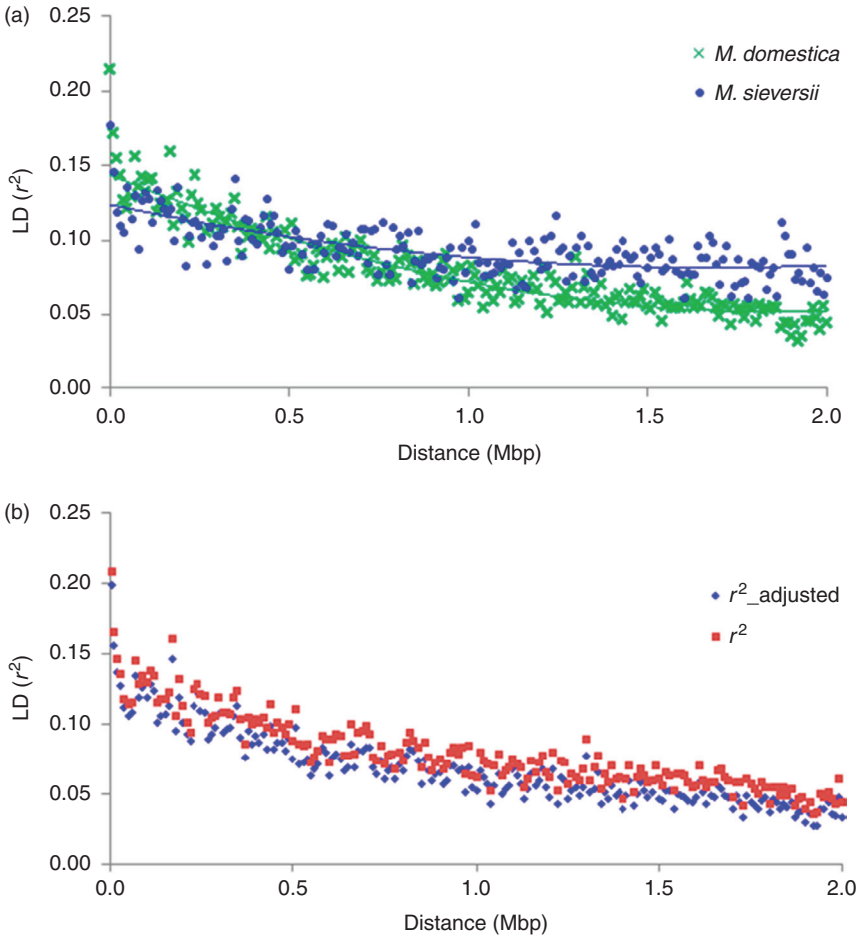

Fig. 3. Patterns of linkage disequilibrium (LD: r^2) in two *Malus* species separately (a) and a combined LD pattern with and without making adjustment for population structure (b).

known for its causal effect on Type 1 red flesh in apple fruits (Chagné *et al.*, 2007; Espley *et al.*, 2007) and is also a key regulator of red skin colour (Lin-Wang *et al.*, 2011). The SNP (ss475881697) with the largest effect (i.e. explaining 8% variation) on TA resides within the leucoanthocyanidin reductase (*LAR1*) gene (MDP0000376284) and the next-best SNPs (ss475881686 and GDsnp01588) are located close to the *Ma* gene (Xu *et al.*, 2012; Khan *et al.*, 2013). The frequency of the minor allele at the largest-effect SNP for OCOL and TA was 0.36 and 0.17, respectively, suggesting that the observed marker–trait association is less likely to be biased by the MAF at these SNP loci.

Discussion

Q_{st} and F_{st}

Even though molecular marker-based methods have become cheaper and faster to assist rapid decision-making in conservation projects, an understanding of locally adaptive between-population and within-population genetic variation is likely to be of greater importance when choosing populations/genotypes for genetic conservation. Thus, Q_{st} will be particularly relevant to conservation efforts where preserving extant adaptation to local environments is an important goal. Q_{st} varies widely among traits, and the traits that experience the strongest local selection (or adaptation) pressures are expected to be the most divergent (McKay and Latta, 2002). The traits considered in the present study could perhaps be of low relevance from a natural adaptation point of view, but these are certainly desirable for commercial breeding purposes. In this study, Q_{st} was highest for CRISP, followed by WT and FINT, suggesting that these traits could have experienced strongest selection pressure among the studied traits.

The estimated F_{st} value between *M. sieversii* and *M.* × *domestica* in the study was 0.12, which is similar to that reported in an earlier report using SNPs ($F_{st} = 0.14$; Velasco *et al.*, 2010). There are not many SNP-based population genetics studies reported for *Malus* species. SSR-based studies

https://doi.org/10.1017/S1479262114000136 Published online by Cambridge University Press

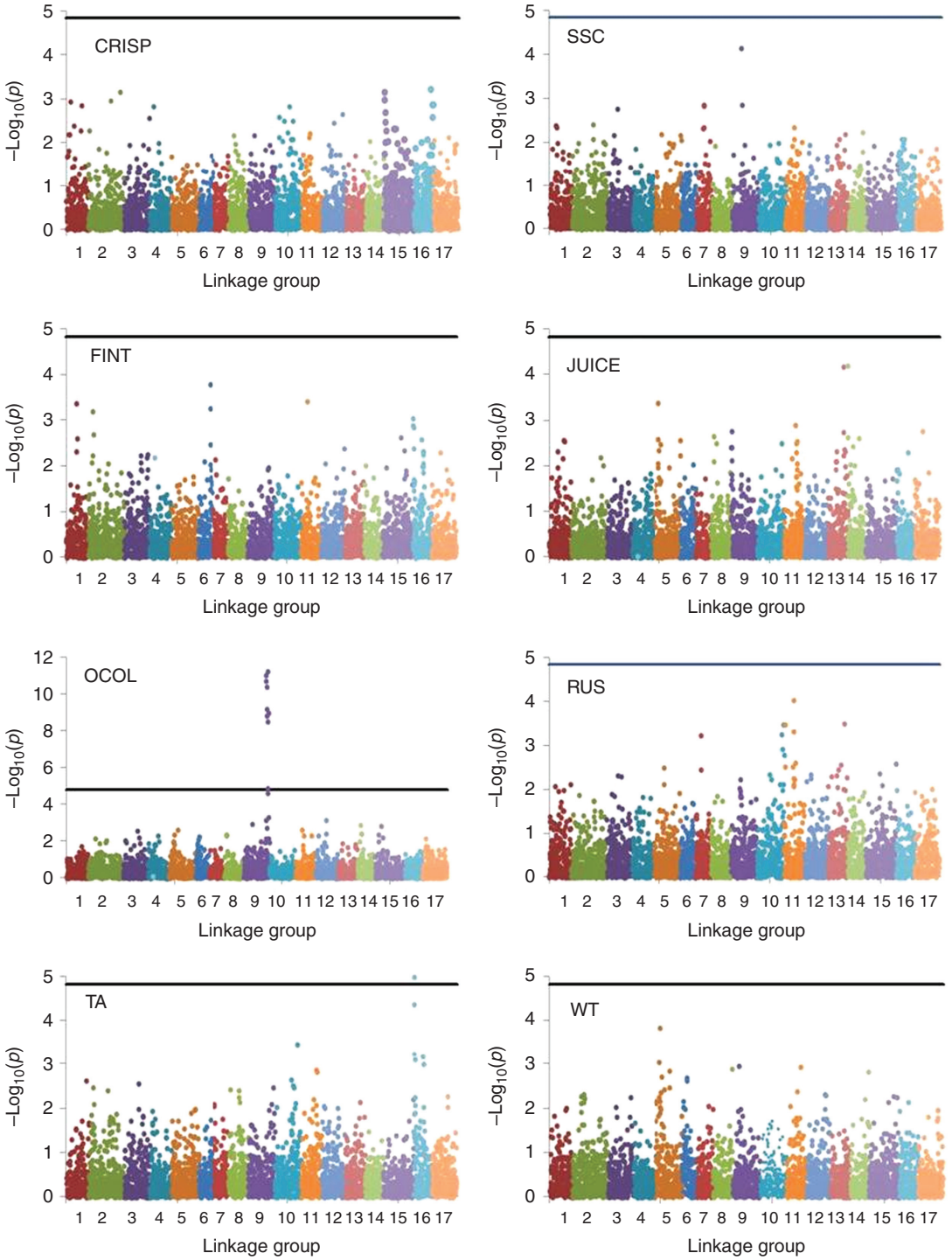

Fig. 4. Manhattan plots of the $-\log_{10}(p)$ values for various apple fruit traits (CRISP, crispness; SSC, soluble solid concentration; FINT, flavour intensity; JUICE, juiciness; OCOL, over-colour; RUS, russet; TA, titratable acidity; and WT, fruit weight) from a genome-wide scan against position on each of 17 linkage groups (represented by different colours). Grey horizontal line indicates the genome-wide significance threshold.

have reported somewhat lower F_{st} values ($F_{st} = 0.09$, Coart *et al.*, 2006; and $F_{st} = 0.06$, Cornille *et al.*, 2012) than SNP-based studies. Differing mutation rates and heterozygosity levels between SSRs and SNPs raise the question of how to compare diversity estimates derived from different marker systems. Nonetheless, studies comparing F_{st} estimates using different marker systems seem to report generally comparable values (Allendorf and Seeb, 2000,

https://doi.org/10.1017/S1479262114000136 Published online by Cambridge University Press

and references therein). Ritland (2000) recommended that F_{st} and Q_{st} comparisons be preferably carried out by employing less mutable loci than microsatellites. Mutation rates are in general considerably lower for SNPs than for microsatellites (Foll and Gaggiotti, 2008), and also SNPs are increasingly preferred over other marker systems (e.g. microsatellites) for population genetics studies for various reasons, including the availability of high numbers of annotated markers, low scoring error rates, and the potential for high-throughput genotyping (Helyar *et al.*, 2011).

Differences in sampling intensity (16 and 99 individuals from *M. sieversii* and *M.* × *domestica*, respectively) could bias comparison between species, so further investigation is needed using larger sample sizes to evaluate SNP-based population differentiation and Q_{st} and F_{st} comparison in *Malus* species. However, it is important to note that the level of polymorphism affects the relationship between sample size and accuracy of parameter estimates (allele frequency, allelic richness, etc.). Estimation is less biased by differences in sample size for loci with low polymorphism than for those with high polymorphism. Thus, unequal sample size will probably bias estimates using SNP markers less than those using microsatellites (Leberg, 2002).

The statistical evidence for selection and local adaptation is provided by comparing the distribution of F_{st} with that of Q_{st}. Under pure neutrality, and if the traits are additive, $Q_{st} = F_{st}$ for any trait, and departures from this neutral expectation are considered to be evidence for selection acting on the quantitative trait under study (Spitze, 1993; Whitlock, 1999; Merilä and Crnokrak, 2001; McKay and Latta, 2002). Traits (e.g. CRISP, WT, FINT and TA) with $Q_{st} > F_{st}$ indicated that directional selection favouring these phenotypes could have been involved in the domesticated apple (*M.* × *domestica*). For some traits (e.g. JUICE, SSC and OCOL) in the present study, Q_{st} values were very similar to F_{st} values, suggesting that the observed degree of differentiation between the two species could have been obtained by genetic drift alone. We observed $Q_{st} < F_{st}$ for only one trait (RUS), which could perhaps suggest that the domestication of apple did not necessarily favour selection on RUS differently from the natural evolution of this trait in *M. sieversii*. Nonetheless, genes, markers and traits will each behave differently in the adaptive divergence of populations, and thus extrapolation from one type of variation to another must be done with caution (McKay and Latta, 2002).

We obtained genetic variation within a species (σ^2_W) by clonally replicating each accession; hence, the estimates are partly confounded by non-additive genetic variances, which could result in conservative estimates of Q_{st} (Lynch and Walsh, 1998; Merilä and Crnokrak, 2001). Thus, our estimates of Q_{st} are likely to represent lower bounds and could be conservative with respect to finding evidence for species differentiation.

Population structure

Our results revealed possible introgression of genetic material into *M. sieversii* from *M.* × *domestica* and *vice versa*. The average posterior probability of assignment of *M. sieversii* accessions to the *M.* × *domestica* gene pool was 23%, which is almost identical to that reported earlier by Cornille *et al.* (2012). The percentage of admixed ancestry of the *M.* × *domestica* gene pool in the *M. sieversii* gene pool was 11%, which supports earlier reports (Coart *et al.*, 2006; Velasco *et al.*, 2010; Cornille *et al.*, 2012) of *M. sieversii* being one of the progenitor species. When three clusters (i.e. *M.* × *domestica*, *M. sieversii* and an 'unknown') were considered, the 'unknown' gene pool was found to contribute about 26% to the genome of the *M.* × *domestica* gene pool. We hypothesized this 'unknown' gene pool to be *M. sylvestris*, but it is difficult to say whether this estimated 26% introgression is indeed from *M. sylvestris* alone or from a mix of some other species, as *M. sylvestris* accessions were not included in the present study. A study carried out by Cornille *et al.* (2012), which included all the three species (*M.* × *domestica*, *M. sieversii* and *M. sylvestris*), reported that the contribution of *M. sylvestris* to the *M.* × *domestica* gene pool was about 16%, which lends support to our hypothesis.

The SNP-derived average pairwise coefficient of relationship (0.55) among the *M. sieversii* accessions was higher than that among the *M.* × *domestica* accessions (0.21). The estimated F_{IS} value, which is a measure of inbreeding, was also higher for *M. sieversii* than for *M.* × *domestica*. *M.* × *domestica* individuals used in this study were derived from OP seeds collected from germplasm repositories. OP seedlings from wider crossing, especially when accessions of different *Malus* species are planted in proximity, could result in higher genetic variation than those from a population comprising named cultivars (Kumar *et al.*, 2010). Each of the 16 *M. sieversii* accessions represented a different maternal family, whereas there were a couple of accessions (out of 99) of *M.* × *domestica* derived from the same maternal family. A likely involvement of common pollen parents (especially in insect-pollinated species) or reciprocal pollination could result in higher observed relatedness than expected. In addition to the smaller sample size (16 *vs.* 99), another reason for the observed relatively low molecular variation in *M. sieversii* accessions could be the ascertainment bias (Helyar *et al.*, 2011), because the panel of 27 accessions used for designing the 8K SNP array involved only one *M. sieversii* accession. Our results indicated that the median MAF was lower in *M. sieversii* than in *M.* × *domestica*, suggesting some ascertainment bias. Still, the likelihood of transferability of *M.* × *domestica* SNPs to *M. sieversii* has been reported to be high (Micheletti *et al.*, 2011). Interestingly, there were

https://doi.org/10.1017/S1479262114000136 Published online by Cambridge University Press

three SNPs, namely ss475883988 (LG6), MDP0000215722 (LG8) and ss475883982 (LG12), found to be polymorphic only in *M. sieversii*, suggesting that alleles at these loci are perhaps fixed in *M.* × *domestica* accessions.

OP seeds of *M. sieversii* were collected from their natural habitat in Kazakhstan. The reported introgression (Coart *et al.*, 2006; Larsen *et al.*, 2006; Cornille *et al.*, 2012) from the domesticated apple to its wild progenitors (*M. sieversii* and *M. sylvestris)*, which is supported by the results of this study, suggests that for gene conservation purposes, it is desirable to utilize molecular and phenotypic data to identify hybrids and/or misclassified individuals (Gross *et al.*, 2012).

LD and marker–trait association

A high degree of LD is a desirable feature for reliable marker–trait associations in genome-wide association studies (GWAS) and also for higher accuracy of genomic selection. However, an assumption of no LD between markers is common to various methods of estimating population genetics parameters (e.g. F_{st}). Although the STRUCTURE analysis assumes that SNP loci are independent within the study population, this assumption would not have been grossly violated because the pattern of LD in the present study indicated enough independence across the genome (Fig. 3(a) and (b)). Relatively higher long-distance LD in *M. sieversii* than in *M. domestica* could be due to high genetic relatedness among the *M. sieversii* accessions as discussed earlier. LD structure could vary among different types of plant populations within a species (reviewed by Myles *et al.*, 2009). For a given distance between markers, the extent of LD in our germplasm accessions (Fig. 3) was far less (about one-third) than that reported in advanced-generation crosses (Kumar *et al.*, 2012, 2013).

Inclusion of SNPs that are tightly linked to genes (or regions under selection) could violate the assumption of marker neutrality, hence resulting in misleading inferences about population differentiation and genetic structure (Nielsen *et al.*, 2006). However, a recent review (Kirk and Freeland, 2011) has advocated the use of non-neutral SNPs or causative markers, which could be associated with phenotype, for better understanding of adaptive evolution and population differentiation. A couple of highly significant SNP–trait associations were identified for OCOL and TA in the present study, but the most significant SNPs are probably not the causative ones due to low-to-moderate LD. Moreover, the estimated F_{st} value at two of the most significant SNPs (ss475879551 on LG9 and ss475881697 on LG16) was 0.12 and 0.09, respectively, so these SNPs were not outliers as per the observed F_{st} distribution (Fig. 1). A lack of strong marker–trait associations for most of the studied traits indicated that the SNP array used in this study comprised mainly neutral markers, so our inferences on admixture and population structure of *Malus* accessions would not be biased. The population structure and admixed proportions of *M. sieversii* and *M.* × *domestica* accessions in the present study are quite comparable to those obtained using neutral as well as independent SSR markers (Cornille *et al.*, 2012), suggesting that including a small number of SNPs in LD analysis is not likely to bias the estimates of population differentiation (e.g. Kaeuffer *et al.*, 2007). The *M. sieversii* gene pool has shown significant resistance to apple scab and fire blight (Luby *et al.*, 2002; Forsline and Aldwinckle, 2004) and may well hold genes that allow *M. sieversii* to adapt to cold and dry regions (Yan *et al.*, 2008). Finding marker–trait associations (that can be used in cultivar breeding programmes) in largely unrelated individuals such as germplasm collections will require higher SNP density, compared with that used in this study, due to faster LD decay in such populations.

Conclusions

Trait-level differentiation (Q_{st}) for three of the eight studied traits (crispness, flavour intensity and fruit weight) was more than twice the molecular-level differentiation (F_{st}), suggesting a differential selection and/or adaptation divergence for these traits in *M.* × *domestica*. These differences between Q_{st} and F_{st} suggest that both marker-level and trait-level information should be considered to construct 'core' populations for gene conservation purposes. The observed percentage of admixed ancestry (11%) of the *M.* × *domestica* gene pool in the *M. sieversii* gene pool further supports the hypothesis that the latter is one of the progenitors of the domesticated apple.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1479262114000136

Acknowledgements

This study was partly funded by the New Zealand Ministry of Business, Innovation and Employment (MBIE) through a Core Funding programme. The authors sincerely thank their French colleagues (Francois Laurens and Amandine Cornille) and Plant and Food Research (PFR) colleagues (Richard Volz and Claudia Wiedow) for helpful suggestions and feedback.

References

Allendorf FW and Seeb LW (2000) Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. *Evolution* 54: 640–651.

Brookfield P, Murphy P, Harker R and MacRae E (1997) Starch degradation and starch pattern indices; interpretation and relationship to maturity. *Postharvest Biology and Technology* 11: 23–30.

Brown AG (1975) Apples. In: Janick J and Moore JN (eds) *Advances in Fruit Breeding*. West Lafayette, IN: Purdue University Press, pp. 3–37.

Browning SR and Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics* 81: 1084–1097.

Chagné D, Carlisle C, Blond C, Volz RK, Whitworth C, Oraguzie NZ, Crowhurst RN, Allan AC, Espley RV, Hellens RP and Gardiner SE (2007) Mapping a candidate gene (*MdMYB10*) for red flesh and foliage colour in apple. *BMC Genomics* 8: 212.

Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A, Velasco R, Main D, Rees JD, Iezzoni A, Mockler T, Wilhelm L, van de Weg E, Gardiner SE, Bassil N and Peace C (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS One* 7: e31745.

Chapman C (1989) Principles of germplasm evaluation. In: Stalker HT and Chapman C (eds) IBPGR Training Courses: Lecture Series 2. *Scientific Management of Germplasm: Characterization, Evaluation, and Enhancement*. Rome: International Board for Plant Genetic Resources, pp. 55–64.

Clegg MT (1990) Molecular diversity in plant populations. In: Brown AHD, Clegg MT, Kahler AL and Weir BS (eds) *Plant Population Genetics, Breeding, and Genetic Resources*. Sunderland, MA: Sinauer Associates, Inc., pp. 98–115.

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

Cockerham CC and Weir BS (1993) Estimation of gene flow from *F*-statistics. *Evolution* 47: 855–863.

Cornille A, Gladieux P, Smulders MJM, Roldan-Ruiz I, Laurens F, Le Cam B, Nersesyan A, Clavel J, Olonova M, Feugey L, Gabrielyan I, Zhang X-G, Tenaillon MI and Giraud T (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.

Doebley J (1989) Isozymic evidence and the evolution of crop plants. In: Soltis DE and Soltis PS (eds) *Isozymes in Plant Biology*. Portland, OR: Dioscorides Press, pp. 165–191.

Dunemann F, Kahnau R and Schmidt H (1994) Genetic relationships in *Malus* evaluated by RAPD 'fingerprinting' of cultivars and wild species. *Plant Breeding* 113: 150–159.

Espley RV, Hellens RP, Putterill J, Stevenson DE, Kutty-Amma S and Allan AC (2007) Red colouration in apple fruit is due to the activity of the MYB transcription factor, MdMYB10. *Plant Journal* 49: 414–427.

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

Forsline PL and Aldwinckle HS (2004) Evaluation of *Malus sieversii* seedling populations for disease resistance and horticultural traits. *Acta Horticulturae* 663: 529–534.

Frankham R (1999) Quantitative genetics in conservation biology. *Genetic Resources* 74: 37–244.

Gharghani A, Zamani Z, Talaie A, Oraguzie NC, Fatahi R, Hajnajari H, Wiedow C and Gardiner SE (2009) Genetic identity and relationships of Iranian apple (*Malus* × *domestica* Borkh.) cultivars and landraces, wild *Malus* species and representative old apple cultivars based on simple sequence repeat (SSR) marker analysis. *Genetic Resources and Crop Evolution* 56: 829–842.

Gilmour AR, Cullis BR, Harding SA and Thompson R (2006) *ASReml Update: What's New in Release 2.00*. Hemel Hempstead: VSN International Limited.

Gross BL, Henk AD, Forsline PL, Richards CM and Volk GM (2012) Identification of interspecific hybrids among domesticated apple and its wild relatives. *Tree Genetics and Genomes* 8: 1223–1235.

Harris SA, Robinson JP and Juniper BE (2002) Genetic clues to the origin of the apple. *Trends in Genetics* 18: 426–430.

Harrison N and Harrison R (2011) On the evolutionary history of the domesticated apple. *Nature Genetics* 43: 1043–1044.

Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, Cariani A, Maes GE, Diopere E, Carvalho GR and Nielsen EE (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* 11: 123–136.

Hill WG and Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38: 226–231.

Hilu KW (1989) Taxonomy of cultivated plants. In: Stalker HT and Chapman C (eds) IBPGR Training Courses: Lecture Series 2. *Scientific Management of Germplasm: Characterization, Evaluation, and Enhancement*. Rome: International Board for Plant Genetic Resources, pp. 33–40.

Hokanson SC, Szewc-McFadden AK, Lamboy WF and McFerson JR (1998) Microsatellite (SSR) markers reveal genetic identities, genetic diversity and relationships in a *Malus domestica* Borkh. core subset collection. *Theoretical and Applied Genetics* 97: 671–683.

Kaeuffer R, Reale D, Coltman DW and Pontier D (2007) Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity* 99: 374–380.

Khan SA, Beekwilder J, Schaart JG, Mumm R, Soriano JM, Jacobsen E and Schouten HJ (2013) Differences in acidity of apples are probably mainly caused by a malic acid transporter gene on LG16. *Tree Genetics and Genomes* 9: 475–487.

Kirk H and Freeland JR (2011) Applications and implications of neutral *versus* non-neutral markers in molecular ecology. *International Journal of Molecular Sciences* 12: 3966–3988.

Kumar SK, Volz RK, Alspach PA and Bus VGM (2010) Development of a recurrent apple breeding programme in New Zealand: a synthesis of results, and a proposed revised breeding strategy. *Euphytica* 173: 207–222.

Kumar S, Chagné D, Bink MCAM, Volz RK, Whitworth C and Charmaine C (2012) Genomic selection for fruit quality traits in apple (*Malus* × *domestica* Borkh.). *PLoS One* 7: e36674.

Kumar S, Garrick DG, Bink MCAM, Whitworth C, Chagné D and Volz RK (2013) Novel genomic approaches unravel genetic architecture of complex traits in apple. *BMC Genomic* 14: 93.

Larsen AS, Asmussen CB, Coart E, Olrik DC and Kjaer ED (2006) Hybridization and genetic variation in Danish populations of European crab apple (*Malus sylvestris*). *Tree Genetics and Genomes* 2: 86–97.

Leberg PL (2002) Estimating allelic richness: effects of sample size and bottleneck. *Molecular Ecology* 11: 2445–2449.

Lin-Wang K, Micheletti D, Palmer J, Volz R, Lozano L, Espley R, Hellens RP, Chagne D, Rowan DD, Troggio M, Iglesias I and Allan A (2011) High temperature reduces apple fruit colour via modulation of the anthocyanin regulatory complex. *Plant and Cell Environment* 34: 1176–1190.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES and Zhang Z (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28: 2397–2399.

Liu N, Chen L, Wang S, Oh C and Zhao H (2005) Comparison of single nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* 6: S26.

Luby JJ, Alspach PA, Bus VGM and Oraguzie NC (2002) Field resistance to fire blight in a diverse apple (*Malus* sp.) germplasm collection. *Journal of American Society of Horticultural Science* 127: 245–253.

Lynch M and Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

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

McKay JK and Latta RG (2002) Adaptive population divergence: markers, QTL and traits. *Trends in Ecology and Evolution* 17: 285–291.

Merilä J and Crnokrak P (2001) Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology* 14: 92–103.

Micheletti D, Troggio M, Zharkikh A, Costa F, Malnoy M, Velasco R and Salvi S (2011) Genetic diversity of the genus *Malus* and implications for linkage mapping with SNPs. *Tree Genetics and Genomes* 7: 857–868.

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE and Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21: 2194–2202.

Nielsen EE, Hansen MM and Meldrup D (2006) Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): implications for inferring population structure in non-model organisms. *Molecular Ecology* 15: 3219–3229.

Noiton DAM, Hofstee M, Alspach PA, Brewer L and Howard C (1999) Increasing genetic diversity for apple breeding: a preliminary report. *Acta Horticulturae* 484: 105–107.

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* 9: 1195–1204.

Robinson JP, Harris SA and Juniper BE (2001) Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivated apple, *Malus* × *domestica* Borkh. *Plant Systematics and Evolution* 226: 35–58.

Røen D, Ekholm A and Rumpunen K (2009) Estimating useful diversity in the Norwegian core collection of apples. *Acta Horticulturae* 814: 131–136.

Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* 8: 103–106.

Spitze K (1993) Population structure in *Daphnia obtusa*: quantitative genetic and allozyme variation. *Genetics* 135: 367–374.

Van Raden PM (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91: 4414–4423.

Van Treuren R, Kemp H, Ernsting G, Jongejans B, Houtman H and Visser L (2010) Microsatellite genotyping of apple (*Malus* × *domestica* Borkh.) genetic resources in the Netherlands: application in collection management and variety identification. *Genetic Resources and Crop Evolution* 57: 853–865.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stromo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagné D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouzé P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F and Viola R (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nature Genetics* 42: 833–839.

Weir BS and Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution* 38: 1358–1370.

Whitlock MC (1999) Neutral additive genetic variance in a metapopulation. *Genetic Resources* 74: 215–221.

Xu K, Wang A and Brown S (2012) Genetic characterization of the *Ma* locus with pH and titratable acidity in apple. *Molecular Breeding* 30: 899–912.

Yan G, Long H, Song W and Chen R (2008) Genetic polymorphism of *Malus sieversii* populations in Xinjiang, China. *Genetic Resources and Crop Evolution* 55: 171–181.

Yu J, Pressoir G, Briggs WH, Vroh-Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S and Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38: 203–208.