RESEARCH ARTICLE

Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing

Bjarne Larsen[1]*, **Kyle Gardner**[2], **Carsten Pedersen**[1], **Marian Ørgaard**[1], **Zoë Migicovsky**[2], **Sean Myles**[2], **Torben Bo Toldam-Andersen**[1]

1 Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg C, Denmark, **2** Department of Plant, Food and Environmental Sciences, Dalhousie University, Faculty of Agriculture, Agricultural Campus, Truro, NS, Canada

* bjl@plen.ku.dk

Abstract

In recent years, new genome-wide marker systems have provided highly informative alternatives to low density marker systems for evaluating plant populations. To date, most apple germplasm collections have been genotyped using low-density markers such as simple sequence repeats (SSRs), whereas only a few have been explored using high-density genome-wide marker information. We explored the genetic diversity of the Pometum gene bank collection (University of Copenhagen, Denmark) of 349 apple accessions using over 15,000 genome-wide single nucleotide polymorphisms (SNPs) and 15 SSR markers, in order to compare the strength of the two approaches for describing population structure. We found that 119 accessions shared a putative clonal relationship with at least one other accession in the collection, resulting in the identification of 272 (78%) unique accessions. Of these unique accessions, over half (52%) share a first-degree relationship with at least one other accession. There is therefore a high degree of clonal and family relatedness in the Danish apple gene bank. We find significant genetic differentiation between *Malus domestica* and its supposed primary wild ancestor, *M. sieversii*, as well as between accessions of Danish origin and all others. Using the GBS approach allowed us to estimate ploidy levels, which were in accordance with flow cytometry results. Overall, we found strong concordance between analyses based on the genome-wide SNPs and the 15 SSR loci. However, we argue that GBS is superior to traditional SSR approaches because it allows detection of a much more detailed population structure and can be further exploited in genome-wide association studies (GWAS). Finally, we compare GBS with SSR for the purpose of identifying clones and pedigree relations in a diverse apple gene bank and discuss the advantages and constraints of the two approaches.

Citation: Larsen B, Gardner K, Pedersen C, Ørgaard M, Migicovsky Z, Myles S, et al. (2018) Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. PLoS ONE 13(8): e0201889. https://doi.org/10.1371/journal.pone.0201889

Editor: Lewis Lukens, University of Guelph, CANADA

Received: November 6, 2017

Accepted: July 24, 2018

Published: August 15, 2018

Data Availability Statement: All data are available via doi:10.5061/dryad.b5s2hn4.

Funding: The authors are deeply thankful to Foreningen PlanDanmark grant no. 13024 and to the Department of Plant and Environmental Sciences, UCPH for funding the project. This work was supported by the National Sciences and Engineering Research Council of Canada (NSERC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

Apple germplasm diversity has been explored and described for decades using low-density markers. Here, simple sequence repeat (SSR) markers have been the preferred approach to characterize and compare the germplasm kept in several European apple collections [1]. The heritable, co-dominant information of SSR markers makes them powerful tools for exploring apple gene bank collections in order to reveal genetic diversity, pedigrees, and mislabelled accessions. However, recently introduced genome-wide marker systems provide alternatives to low density marker systems for genotyping.

High-density marker systems based on single nucleotide polymorphisms (SNPs) are useful because they allow genome-wide comparisons which can reveal small genetic differences between individuals that are otherwise quite similar. In apple, medium density Illumina Infinium arrays containing 8k and 20k SNPs were initially developed [2, 3], followed recently by a high-density 487k SNP Affymetrix Aciom array [4]. SNP arrays allow for the investigation of genetic variation, genome-wide association studies (GWAS), genomic selection [4] and deducing the mosaic founder composition of cultivars through reconstruction of pedigrees [5]. However, they are relatively expensive to use and may result in poor hybridization in diverse perennial crops [6]. The development of new next-generation sequencing techniques, such as genotyping-by-sequencing (GBS) protocols [7], allow the simultaneous discovery and genotyping of markers. In comparison to the development of SNP arrays, GBS offers a reduced cost by enabling marker discovery and genotyping in a single step in high diversity species like apple. Moreover, GBS is even applicable in species for which no reference genome is available. GBS was recently applied to characterize a large apple collection in the USA [8].

Important breeding material is often kept in gene bank collections where a lack of genomic information and scarce documentation of agronomic traits pose serious threats to the potential utilization of germplasm resources. Phenotyping of diverse gene bank material is essential for identifying accessions with superior traits for breeding purposes. Phenotype data collected from gene banks can be paired with genome-wide marker information to facilitate genomics-assisted breeding [8, 9]. Genomics-assisted breeding is especially valuable in tree crops with long juvenile phases, such as apple, where genetic screening at the seedling stage may replace several years of the traditional breeding process [10–12]. Genotyping is also a useful tool for verifying the identity of accessions, especially in old gene bank collections that have been renewed and replanted several times, increasing the risk of curation error and thus misidentification. Finally, genotyping is a valuable tool for identifying clones, since frequent incorrect identification, clonal selections, inaccurate passport information and lack of historical documentation complicates apple classification [13, 14].

Here, we used GBS to genotype a collection of 363 apple accessions, including 14 *Malus sieversii* accessions, belonging to the Pometum gene bank collection (University of Copenhagen, Denmark). This is the most comprehensive collection of local Danish apple cultivars, which we recently studied using SSR markers and flow cytometry [15]. We use GBS to generate further insights into the population structure, relatedness and ploidy levels as well as compare the strengths of this high-density, genome-wide marker information with low density SSR markers.

Materials and methods

Plant material and SNP genotyping

We sampled 349 *Malus domestica* and 14 *M. sieversii* accessions (S1 Table) belonging to the Pometum (University of Copenhagen, Denmark). Young leaves from vigorously growing shoots were sampled. One leaf from each accession was immediately transferred to silica-gel and stored in individual airtight plastic bags. Extraction, quantification and further procedures were performed in 96-well plates. DNA extraction was performed using the DNeasy® 96 Plant Kit (Qiagen®, Hilden, Germany) following the manufacturer's protocol. Total DNA content was quantified with the dsDNA dye (Promega) on the Agilent Mx3005P QPCR System. GBS based on Elshire et al., 2011 [7] using the enzyme *Ape*K1 and 96 samples multiplexed was performed by the Biotechnology Resource Centre at Cornell University, USA using 100 bp long single-end reads on HiSeq2000 (Illumina, San Diego, CA).

Raw sequence data was first parsed with a custom python program (see [16]) to split the single multiplexed fastq file into 96 separate fastq files indexed by GBS barcode. During the splitting process several quality control procedures were implemented including (1) discarding sequences with ambiguous bases in the barcode or restriction remnant, (2) 3' adapter trimming (i.e. if the genomic fragment was less than ~100 base pairs in length), (3) detection and trimming of chimeric sequence (by examining reads for a second restriction site and discarding any reads where a restriction site was present), and (4) discarding any trimmed sequences less than 30 bp in length. Individual fastq files were then independently aligned to the Malus 1.0p reference genome (www.rosaceae.org) with bwa 0.6.1 [17] using default parameters (eg. allowing a maximum of 4% alignment mismatch). The individual aligned sam files were converted to their binary form (bam), merged, and sorted using Picard tools 1.69 prior to importing into GATK 3.4 [18] for variant calling. We allowed GATK (Unified Genotyper) to call SNPs with minimal filters, including requiring a base quality score of at least 30 (-stand_call_conf 30.0 -stand_emit_conf 10.0), and a prior on heterozygosity of 0.01 (-hets 0.01). Raw variant call files were then filtered with vcftools 0.1.13b [19] to allow bi-allelic SNPs only, a sequence depth of 8 reads (—minDP 8) for a genotype to be called, a minimum distance between neighbouring SNPs of 10 bp (—thin 10), a maximum of 20% missing data per individual sample and locus (—max-missing 0.80). To remove potential paralogous loci, we discarded SNPs having mean read depths above the 90th percentile of the empirical mean read depth distribution across all loci. SNPs with extreme deviations ($p < 0.0001$) from Hardy-Weinberg equilibrium (eg. excessive heterozygosity) were also removed. Filtered vcf files were converted to PLINK ped/map format [20] for downstream analysis.

Identifying clones, polyploids and first degree relatives using GBS

Initially, we performed GBS on 363 accessions which yielded 29,494 SNPs. Next, we restricted our analyses to SNPs with a minor allele frequency (MAF) >0.05, which resulted in 15,802 SNPs. Of these, 14,841 SNPs (93.9%) were mapped to the assembled portion of chromosomes 1–17 of the Golden Delicious genome version 1.0p (www.rosaceae.org). We calculated identity-by-descent (IBD) for all pairs of samples using PLINK. We considered two accessions to be clones of each other when the IBD ($\hat{\pi}$) was >0.85. In theory, IBD = 1 for pairwise clonal relationships. However, two factors can result in IBD < 1 for pairs of accessions that are clonally related. First, reductions in IBD can result from genotyping errors, which likely result primarily from the poor quality of the reference genome: paralogous regions of the genome are collapsed and thus appear as single copy regions in the reference genome so that sequence coverage variation between samples results in different genotype calls between clones. Second, it is possible that somatic mutations between clones exist and these result in a reduction of IBD values. Even with high-quality genotype data from a genotyping microarray, IBD values as low as 0.95 for clonal relationships were previously found in grapes [21]. Considering the uncertainty of the genotype calls with the use of GBS and a relatively poor quality reference genome, we argue that it is reasonable to observe IBD values as low as 0.85 for pairs of accessions that are clonally related. Finally, given the distribution of IBD among all pairwise comparisons, the most parsimonious explanation for the clear bump at the top in the distribution is that these represent clonal relationships.

Next, we used the network package in R to calculate a network adjacency matrix in which pairwise comparisons with $(\hat{\pi}) > 0.85$ were indicated with a '1' and all other comparisons were indicated with a '0'. We visualized clonal relationships using this matrix and the 'plot.network' function in the network package [22]. Before identifying first-degree relationships, we kept only one representative from each clonal group at random. Next, we calculated the observed heterozygosity by individual using the–het function in PLINK and plotted the results, observing a bi-modal distribution which allowed us to easily identify polyploid accessions due to excess heterozosity. Polyploids were excluded from further analysis. Thus, the final data set included 248 unique, diploid genotypes from the original 363 accessions.

After the removal of duplicate clones and polyploids, we repeated the IBD analysis in order to identify first-degree relationships [23, 24]. Accessions with well-known pedigrees such as 'Aroma', 'Discovery', 'Elstar', 'Gloster', 'Ingrid Marie' and 'James Grieve' were used to calibrate the expected range of IBD values for first-degree relationships. Reported first-degree relationships had IBD values ranging from 0.43 to 0.52, and thus, we considered all accessions with pairwise values in this interval to be putative first degree relatives. We used these thresholds to create a network adjacency matrix and visualized the results using the 'plot.network' function in the network package in R [22].

In order to examine the population structure, we initially used PLINK to filter for unique, diploid *Malus domectica* accessions which left us with 234 individuals. After filtering for 5% MAF and pruning for LD (command:—indep-pairwise 10 3 0.5), 10459 SNPs remained for analysis. Using fastSTRUCTURE [25] we tested K = 1 to K = 8 and used the "choosek" function to determine the optimal K value, which we selected as K = 1.

SSR genotyping

SSR genotyping using 15 SSR markers was previously performed on 485 accessions, which included the 363 accessions genotyped using GBS in this study [15]. In the previous work, we identified first-degree relationships using the software CERVUS [26] with a LOD score threshold of 95% [15].

Examining population structure using PCA

A large number of the studied cultivars derive from few major ancestors, which resulted in distinct genetic clustering shown in previous SSR-based study [15]. Therefore, in the principal components analysis (PCA) we decided to include only two offspring from these major ancestors, 'Cox Orange, 'Pigeon blanc' and 'Melonenapfel'. In addition, for the SSR data, accessions with >20% missing data across the 15 SSRs examined were removed from the dataset. Population structure among the remaining 204 accessions was investigated using the adegenet package [27, 28] in R v.3.3.2 [29]. The 'scaleGen' function was used to replace missing data by the mean allele frequencies. PCA was performed using the 'dudi.pca' function, while centering and scaling the data, and accessions were labelled according to species. Subsequently, accessions labelled as *Malus sieversii* were removed from the data set and accessions with >20% missing data were removed, resulting in 190 *M. domestica* accessions. Missing data was replaced by mean allele frequencies and PCA was performed again. *M. domestica* accessions were labelled according to origin and harvest time.

PCA was also performed using the SNP genotypes. First, the 204 accessions included in the SSR analysis were extracted from the genotype table using PLINK [23, 24]. Missing data was imputed using LinkImpute (parameters: $k = 3$, $l = 18$) and the resulting imputation accuracy was 93.7% [30]. SNPs with a minor allele frequency (MAF) <0.01 were removed, reducing the SNP set from 24,533 SNPs to 23,446. SNPs were then pruned for linkage disequilibrium using PLINK (—indep-pairwise 10 3 0.5) [23, 24], reducing the number of markers from 23,460 to 17,737 for PCA. The resulting SNP genotype data was analyzed using the same method as the SSRs: by centering and scaling the data using the 'dudi.pca' function in the adegenet package [27, 28] in R v.3.3.2 [29]. We divided accessions based on species and used a Mann-Whitney U test to estimate if species differed along the SSR and SNP PC1 and PC2.

Next, we extracted the 190 *M*. *domestica* accessions from the imputed SNP dataset and repeated the MAF filter of 0.01 and LD-pruning using PLINK [23, 24]. The number of SNPs was reduced from 24,533 SNPs to 17,700, after which PCA analysis was repeated with accessions labelled according to origin and harvest time. We divided accessions based on origin and used a Mann-Whitney U test to estimate if accessions differed along SNP PC2 based on origin. Finally, we tested the correlation between SSR PCs 1 to 5 and SNP PCs 1 to 5, as calculated using all 204 accessions. We used a Pearson's correlation and all p-values were Bonferroni-corrected (multiplied by 25) for multiple comparisons. All PCA results were visualized using the ggplot2 package in R [31].

Results

GBS yielded on average 2.5 million sequence reads per sample for the 363 accessions, with a coefficient of variation of 17%. The accessions genotyped using GBS reflected a subset of accessions previously genotyped using SSR markers. Ploidy levels, determined using flow cytometry, were also available for accessions included in the previous work [15]. This enabled us to compare the strength of GBS, SSRs, and flow cytometry for identifying clones, ploidy levels, establishing first-degree relationships and revealing the underlying genetic structure of the accessions.

GBS reveals triploid accessions

Using genome-wide SNP data, we calculated total heterozygosity by individual, which separated accessions into two groups with heterozygosity ≤ 0.335 or ≥ 0.345 (Fig 1). We compared the accessions in each of the two groups with ploidy levels previously established by flow cytometry [15] and found that accessions with heterozygosity ≤ 0.335 were diploid according to flow cytometry data and that accessions with heterozygosity ≥ 0.345 were triploid. Ploidy levels revealed by both GBS and flow cytometry are given in S1 Table.

Relationships and population structure

We found 230 accessions without any clonal relationships and 119 accessions with at least one putative clonal relationship, resulting in a total of 272 unique genotypes. For some cultivars, somatic mutations have resulted in several clones, such as colour sports, that have been maintained through grafting. We identified 42 putative clonal groups, of which the majority (31) consists of two clonal accessions. The highest number of accessions within a clonal group was 15, which were identified for 'Gravensteiner' (Fig 2).

Analysis of first-degree relationships revealed 142 (52%) accessions with at least one first-degree relative in the collection (Fig 3 and S2 Table). 106 (30%) accessions form a single network that is inter-connected through a series of first-degree relationships. Of the 154 first-degree relationships identified, the majority (96) were discovered using both SSR and SNP markers. 31 first-degree relationships were identified using SSR markers but not SNP markers, whereas 27 were revealed by SNP markers and not by SSR markers (S2 Table).

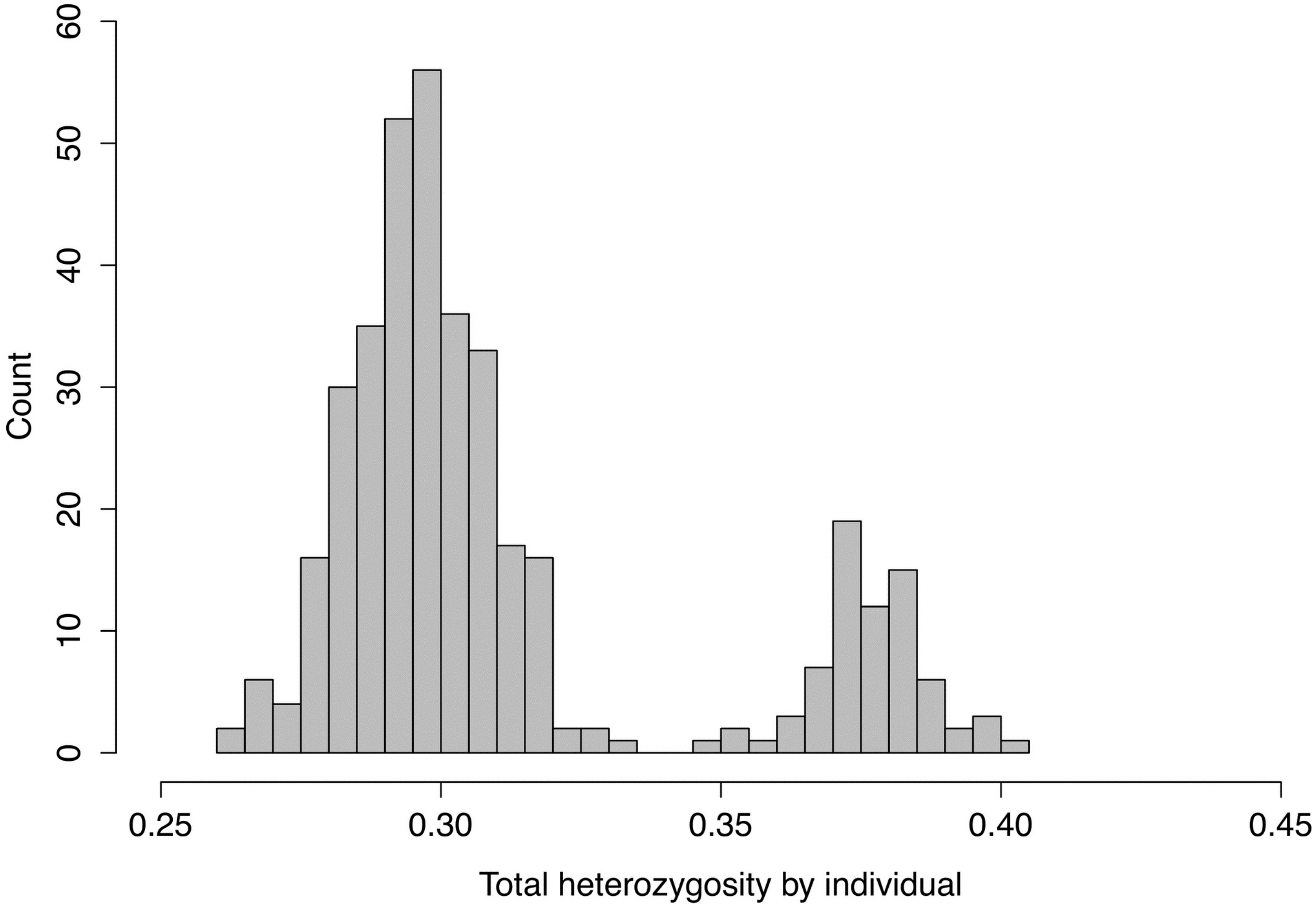

Fig 1. Bar plot of heterozygosity by individual among 15,802 SNPs generated by GBS. The first cluster (heterozygosity ≤ 0.335) contains all diploid accessions whereas the other cluster (heterozygosity ≥ 0.345) comprises triploid accessions according flow cytometry analysis.

https://doi.org/10.1371/journal.pone.0201889.g001

Differentiation between *Malus domestica* and *M. sieversii* was found for SSR-based analysis along PC1 ($p = 5.19 \times 10^{-10}$); and based on SNP data along PC1 ($p = 2.83 \times 10^{-8}$) and PC2 ($p = 2.34 \times 10^{-8}$) (Fig 4). The genomic PC positional information for all accessions are listed in S3 Table. Labelling *M. domestica* accessions according their harvest time resulted in no significant separation; whereas accessions of Danish origin vs. other geographical origins differed using SNP-based PCA along PC2 ($p = 0.002$) (S1 Fig).

Discussion

Next generation sequencing combines high-throughput SNP-discovery and genotyping, resulting in high-density SNP-marker data. It is currently replacing traditional genotyping techniques, like SSR-markers, primarily because of its ease and its suitability for GWAS and genomic selection. Genome re-sequencing provides higher resolution, but for many purposes fewer markers are sufficient. Thus, various approaches have been developed to reduce sequencing costs, either by focusing on expressed sequences through RNAseq or on sequences next to restriction enzyme sites. There are several variants of the former approach including GBS [7], RAD-seq [32], SBG [33], and DArTseq [34].

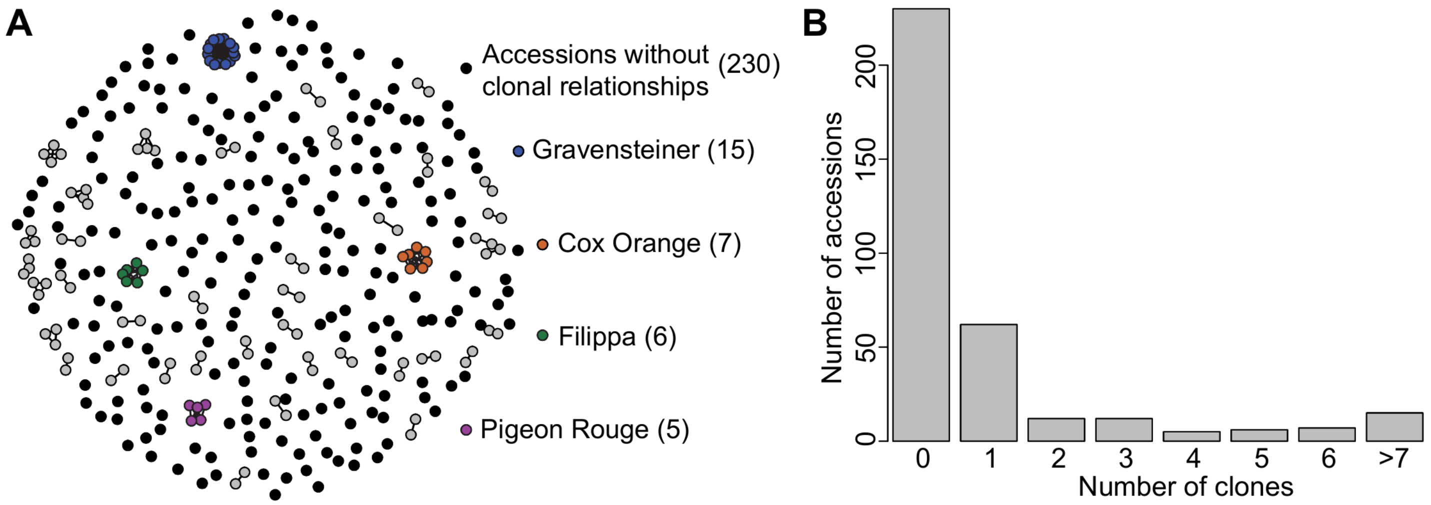

Fig 2. Clonal relationships in the Danish apple germplasm collection. (A) Network of clonal relationships among 349 *Malus domestica* accessions. Each accession is represented by a dot. Accessions without a clonal relationship to other accessions are indicated with a black dot. Accessions with four or fewer clonal relationships are shown in grey together with their clones. Accessions with five or more clonal relationships are indicated by a colour code. (B) For each of the 349 accessions, the number of clonal relationships was evaluated. The majority of accessions (230) are without clonal relationships while 119 (34%) of the accessions have clonal relationships with one or more accessions.

https://doi.org/10.1371/journal.pone.0201889.g002

Here we genotyped 349 apple accessions using the GBS protocol [7] and compared the resulting information with our previous investigations using SSR markers and flow cytometry [15]. When we filtered for less than 20% missing data per individual sample and locus, and a MAF > 5%, 15,802 SNPs remained and 14,841 SNPs (93.9%) of these were mapped to chromosomes 1–17 of the Golden Delicious genome version 1.0p. This is considerably more than the 8,657 SNP-markers obtained with the same GBS-protocol in the USDA apple germplasm collection [8] and also much more than the ~4000 SNPs found among F1 progenies using

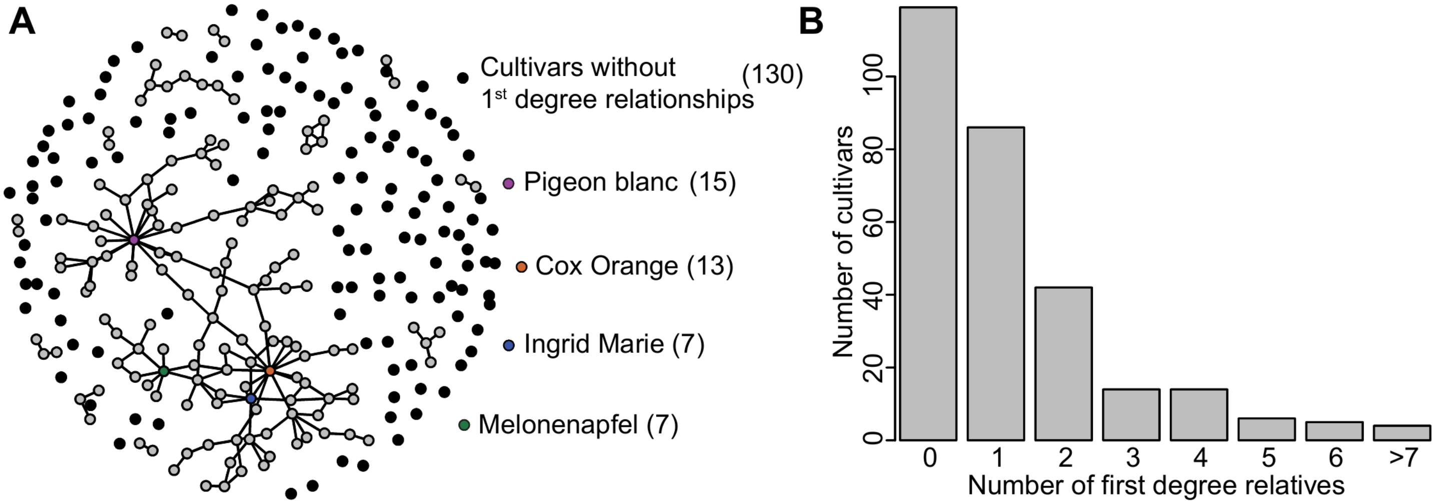

Fig 3. Pedigree structure in the Danish apple germplasm collection. (A) Network of first-degree relationships for each of the 272 unique *Malus domestica* cultivars. Each unique apple cultivar is represented by a dot and edges in the network represent first-degree relationships. Cultivars without first-degree relationships in the collection are indicated by a lone, black dot. In total, 142 cultivars have at least one first-degree relationship. The largest interconnected network includes 106 (39%) of the unique cultivars that are connected through a series of first-degree relatives. (B) The number of first-degree relationships for each of the 272 cultivars. While 130 (48%) cultivars are without first-degree relationships, 52% of the cultivars have a first-degree relationship with at least one other accession.

https://doi.org/10.1371/journal.pone.0201889.g003

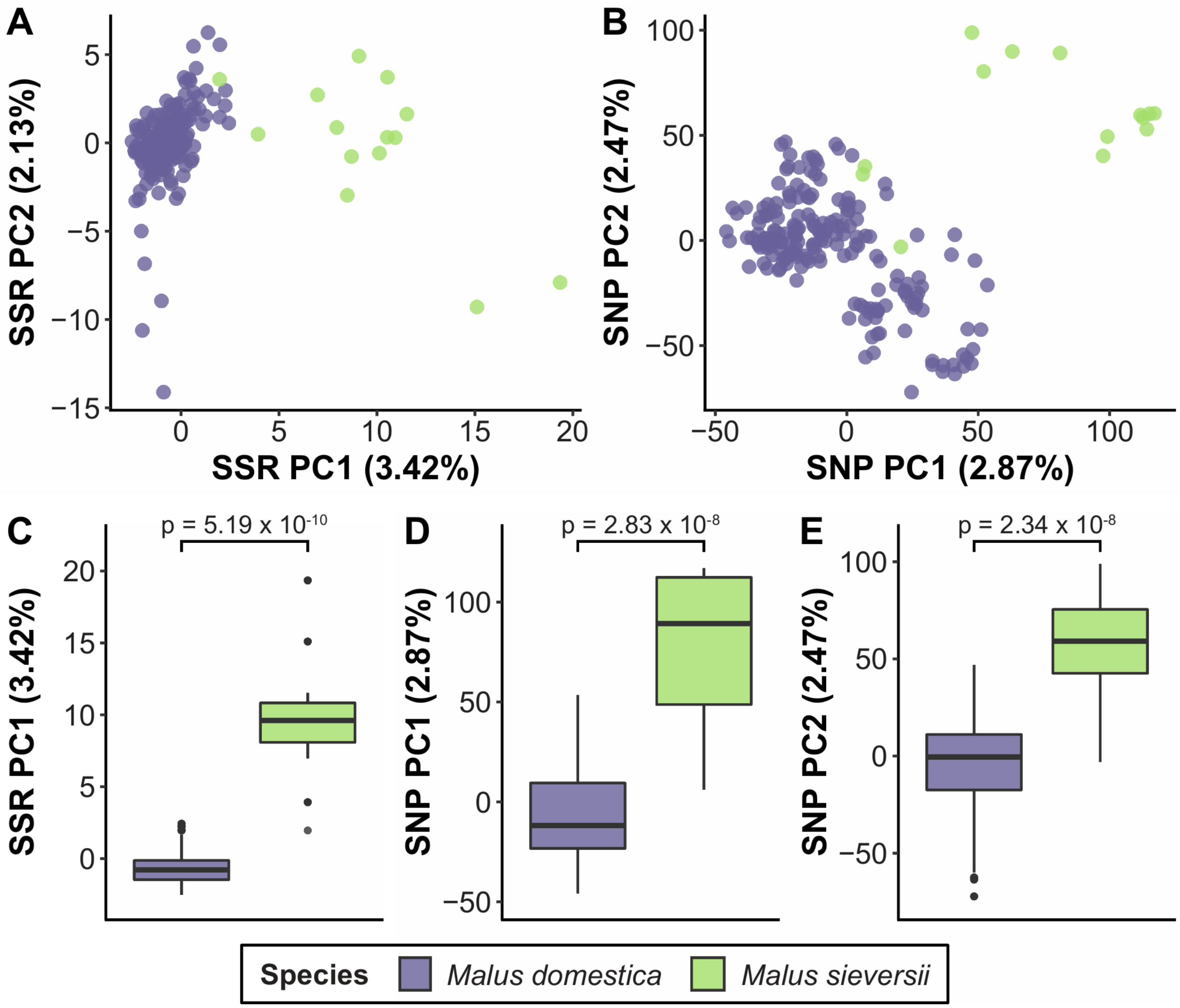

Fig 4. Principal components analysis (PCA) based on (A) SSR and (B) SNP data for 204 accessions. Accessions are labelled based on species, and boxplots of species distribution along (C) SSR PC1, (D) SNP PC1 and (E) SNP PC2 are included. The percentage of variance explained by each PC is indicated in parentheses. Results from Mann-Whitney U-tests between *Malus domestica* and *M. sieversii* are also reported.

https://doi.org/10.1371/journal.pone.0201889.g004

RADseq [35] or GBS [16]. The reason for this high number of SNPs is unlikely to be that the Danish apple collection contains higher levels of genetic diversity. A more likely explanation lies in technical improvements. We carefully equilibrated DNA concentrations before multiplexing which likely resulted in more even coverage across samples and hence fewer SNPs removed due to missing data. Future improvements to genomic resources (e.g. the apple reference genome assembly and genotype imputation algorithms [30, 36]) will allow for higher numbers of SNPs, as evidenced by recent work in apple which identified 122,000 SNPs using GBS [37].

Using GBS to identify ploidy levels

Triploid apples are relatively common among cultivars: they have a large fruit size and vigorous growth. We identified 19% of the accessions examined as triploids, which are assumed to be the result of fusions between unreduced diploid gametes and normal haploid gametes.

Triploid individuals generally have a 50% higher level of heterozygosity than diploid individuals, assuming Hardy-Weinberg equilibrium conditions, independently of the allele frequency. We found a clear difference in the level of heterozygosity between diploid and triploid accessions (Fig 1). However, it was only about 30% higher in triploids, likely because the studied collection is not a population in Hardy-Weinberg equilibrium. Determination of ploidy level in aspen (*Populus tremuloides* Michx.) was recently described [38] by applying a relatively complex statistical model to GBS data. However, here we show that triploids can be differentiated from diploids simply by the level of heterozygosity in a diverse collection of apple. In other apple germplasm, where triploids are derived from a limited set of diploids, ploidy levels may be indistinguishable. Our use of heterozygosity to distinguish ploidy levels therefore warrants further investigation in other collections of apples and other species. The Infinium SNP-array platform can also be used to determine ploidy level and aneuploidy in apple due to variations in the ratio of signal intensity from the two alleles of each SNP along the chromosomes [39]. However, we do not have information concerning aneuploidy in our collection because it is not easily revealed by flow cytometry.

SSRs vs. SNP-markers for revealing parentages

The number of SSRs and SNP-markers needed to achieve a sufficiently high probability of correct identification of first-degree relationships depends on the level of heterozygosity and the number and frequency of alleles. However, based on theoretical estimations [40, 41] about 5–10 SNPs equals one SSR and about 200 SNP markers are needed to characterize relatedness. The high number of alleles for SSR-markers makes each individual SSR-marker much more informative than bi-allelic SNP genotypes. Identification of clones using SSR markers was performed in a previous study [15] and we found the exact same clonal relationships using SSR markers and SNP markers (S1 Table). Hence, we found the two approaches equally effective for identification of clones. Identifying pedigree relationships, however, is more complicated and we therefore find several reasons to compare the strength of 15 SSR-markers with more than 15,000 SNPs.

Firstly, the use of SSRs is still the gold standard due to the large number of alleles, the high reliability of the resulting genotype calls, the transferability between studies, and the relatively small number of markers required to unambiguously reveal pedigree relatedness. The use of GBS markers for pedigree analysis is, however, still in its infancy and, due to the genotype uncertainty and the relatively low pedigree-relevant informativeness per SNP, there is still uncertainty about where to draw thresholds for pedigree relatedness and what metrics of relatedness (IBD or IBS or other) to use. In addition, pedigree relatedness in a highly heterozygous and clonally propagated organism like apple is severely complicated due to the nature of the pedigree relations: an individual can cross with its own grandparent, cousin etc. and popular cultivars end up being extensively crossed such that their alleles become highly overrepresented in the population overall. This was observed here, were 'Cox's Orange' is a frequent parent to many of the studied cultivars.

Nevertheless, from manually checking the SSR markers and from historical reports on parentages, we conclude that first-degree relationships revealed by both marker systems are parent-offspring relations. Relationships revealed by only one of the two marker systems are primarily first-degree relations with examples of second degree relations. The second degree relations are dominated by half siblings that have 'Cox's Orange' as one parent. The studied accessions includes the most important cultivars that have been grown in Denmark during the past centuries [42–44], such as 'Pigeon blanc', 'Cox's Orange' and 'Ingrid Marie' and these diploid cultivars also have the highest number of first-degree relationships (Fig 3). Our findings are therefore consistent with historical information where the reported place and year of origin [42, 44, 45] has helped us pinpoint the putative parent and offspring in many first-degree relations. We found that 'Melonenapfel', described for the first time in 1788 [46], is the parent of the two important Danish cultivars, 'Filippa' and 'Dronning Louise' (S2 Table).

Gene bank population structure

We found differentiation between accessions of Danish origin and other geographical origins on the basis of the SNP data (S1 Fig). This was not observed for the SSR data. The primary reason for this discrepancy between the SSR and SNP data is likely marker number. The 15 SSR markers provide a far less comprehensive view of the entire genome than the 15k SNPs. With 17 chromosomes, 15 SSRs do not even provide a marker for every chromosome, and given the rapid LD decay and short haplotype blocks observed in diverse collections of apples [8], 15 markers is expected to provide a snapshot of roughly $<0.01\%$ of the segregating haplotypes in the population. Thus, the SNP data are a far more powerful system to detect structure, as described from other studies e.g. [47, 48], and were able to detect even the relatively weak differentiation we observe between Danish and other apple cultivars (S1 Fig).

No well-defined subpopulations were identified within the *Malus domestica* accessions of Danish origin, which is in line with previous findings based on SSR data [15]. Also fastSTRUCTURE analysis suggests that $K = 1$ gives the best description of population structure and thereby support previous findings [15] that population structure is lacking in the *M. domestica* collection. At the species level, *M. domestica* and *M. sieversii* differentiate, which is in accordance with previous findings [8] from the USDA-Germplasm collection. In contrast to this previous work, we did not find that population structure was correlated with harvest time (S1 Fig). This is probably because the accessions studied here do not represent as broad a sample as the USDA collection but rather represent a genetic group of apples adapted to North European costal climate with a short harvest window.

Comparison SSR and SNP marker approaches

An accurate cost comparison between SSR-based marker analysis and GBS-based SNP-marker analysis is difficult because many laboratory-specific conditions will influence the cost, such as the price of labour and availability of instruments. In our study, sequencing costs for the genome-wide SNP data were at least 10 times more expensive than the laboratory costs of obtaining the SSR data. However, obtaining the final SSR genotype data is much more labour-intensive since it requires preparing a large number of PCR reactions, performing allele calling, and many more additional (manual) steps.

Prediction of ploidy levels based on the degree of heterozygosity of the SNPs, as observed here, may not account for all collections of apple and thus, needs confirmation by flow cytometry. However, it may well be the case that the degree of heterozygosity clusters in two clearly separated groups. We therefore hypothesize that the use of genome-wide SNP data using next-generation sequencing will be more desirable and efficient for future characterization of germplasm collections. Finally, even though both SSR- and SNP-markers are powerful tools for exploring genetic diversity, only GBS or high-density SNP-arrays provide enough SNP-markers for GWAS, which can then enable marker-assisted breeding. Therefore, despite uncovering similar results in our work using both SSRs and SNPs, most future work will benefit from using SNPs, which allow for both GWAS and detailed population studies to be performed.

Supporting information

S1 Table. Accession list including clonal groups and ploidy levels.
(XLSX)

S2 Table. First-degree relations revealed by SSR and SNP markers.
(XLSX)

S3 Table. Positional information for SSR and SNP PCs visualized in Fig 4.
(XLSX)

S1 Fig. PCA analysis for *Malus domestica* accessions. PCA plot made on basis of SSR-based analysis (A) and SNP-based analysis (B) which enabled to distinguish between accessions of Danish origin and other geographical origins (C).
(PDF)

S2 Fig. Comparisons of SSR and SNPs using PCs with Bonferroni correction.
(PDF)

Acknowledgments

We are deeply thankful to *Foreningen PlanDanmark* and to the Department of Plant and Environmental Sciences, UCPH for funding the project. This work was supported by the National Sciences and Engineering Research Council of Canada (NSERC).

Author Contributions

Conceptualization: Bjarne Larsen, Carsten Pedersen, Marian Ørgaard, Sean Myles, Torben Bo Toldam-Andersen.

Data curation: Bjarne Larsen, Kyle Gardner, Carsten Pedersen, Sean Myles, Torben Bo Toldam-Andersen.

Formal analysis: Bjarne Larsen, Kyle Gardner, Carsten Pedersen, Zoë Migicovsky.

Funding acquisition: Marian Ørgaard, Sean Myles.

Methodology: Bjarne Larsen, Kyle Gardner, Zoë Migicovsky, Sean Myles.

Project administration: Bjarne Larsen, Carsten Pedersen, Marian Ørgaard, Sean Myles, Torben Bo Toldam-Andersen.

Resources: Bjarne Larsen, Sean Myles, Torben Bo Toldam-Andersen.

Software: Bjarne Larsen, Zoë Migicovsky.

Supervision: Kyle Gardner, Carsten Pedersen, Marian Ørgaard, Sean Myles, Torben Bo Toldam-Andersen.

Validation: Bjarne Larsen, Carsten Pedersen.

Visualization: Sean Myles.

Writing – original draft: Bjarne Larsen, Kyle Gardner, Zoë Migicovsky.

Writing – review & editing: Bjarne Larsen, Carsten Pedersen, Marian Ørgaard, Zoë Migicovsky, Sean Myles, Torben Bo Toldam-Andersen.

References

1. Urrestarazu J, Denance C, Ravon E, Guyader A, Guisnel R, Feugey L, et al. Analysis of the genetic diversity and structure across a wide range of germplasm reveals prominent gene flow in apple at the European level. BMC plant biology. 2016; 16(1):130. Epub 2016/06/10. https://doi.org/10.1186/s12870-016-0818-0 PMID: 27277533; PubMed Central PMCID: PMCPmc4898379.

2. Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C. Genome-wide SNP detection, validation, and development of an 8 K SNP array for apple. PLoS One. 2012; 7. https://doi.org/10.1371/journal.pone.0031745 PMID: 22363718

3. Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M, et al. Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (Malus × domestica Borkh). PLoS ONE. 2014; 9(10):e110377. https://doi.org/10.1371/journal.pone.0110377 PMID: 25303088

4. Bianco L, Cestaro A, Linsmith G, Muranty H, Denancé C, Théron A, et al. Development and validation of the Axiom®Apple480K SNP genotyping array. The Plant Journal. 2016; 86(1):62–74. https://doi.org/10.1111/tpj.13145 PMID: 26919684

5. Howard NP, van de Weg E, Bedford DS, Peace CP, Vanderzande S, Clark MD, et al. Elucidation of the 'Honeycrisp' pedigree through haplotype analysis with a multi-family integrated SNP linkage map and a large apple (Malusxdomestica) pedigree-connected SNP data set. Hortic Res. 2017; 4:17003. Epub 2017/03/01. https://doi.org/10.1038/hortres.2017.3 PMID: 28243452; PubMed Central PMCID: PMCPMC5321071 apple cultivar. JJL and DSB, and the University of Minnesota have a royalty interest in this cultivar. These relationships have been reviewed and managed by the University of Minnesota in accordance with its Conflict of Interest policies. The remaining authors declare no conflict of interest.

6. Myles S, Mahanil S, Harriman J, Gardner KM, Franklin JL, Reisch BI, et al. Genetic mapping in grapevine using SNP microarray intensity values. Molecular Breeding. 2015; 35(88). https://doi.org/10.1007/s11032-015-0288-3

7. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one. 2011; 6(5):e19379. https://doi.org/10.1371/journal.pone.0019379 PMID: 21573248

8. Migicovsky Z, Gardner KM, Money D, Sawler J, Bloom JS, Moffett P, et al. Genome to Phenome Mapping in Apple Using Historical Data. The Plant Genome. 2016.

9. Migicovsky Z, Sawler J, Gardner KM, Aradhya MK, Prins BH, Schwaninger HR, et al. Patterns of genomic and phenomic diversity in wine and table grapes. Horticulture Research. 2017; 4. https://doi.org/10.1038/hortres.2017.35 PMID: 28791127

10. Peace CP. DNA-informed breeding of rosaceous crops: promises, progress and prospects. Horticulture Research. 2017; 4:17006. https://doi.org/10.1038/hortres.2017.6 PMID: 28326185

11. Ru S, Main D, Evans K, Peace C. Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. Tree Genetics & Genomes. 2015; 11(1). https://doi.org/10.1007/s11295-015-0834-5

12. Migicovsky Z, Myles S. Exploiting Wild Relatives for Genomics-assisted Breeding of Perennial Crops. Frontiers in Plant Science. 2017; 8(460). https://doi.org/10.3389/fpls.2017.00460 PMID: 28421095

13. Myles S. Improving fruit and wine: what does genomics have to offer? Trends Genet. 2013; 29(4):190–6. Epub 2013/02/23. https://doi.org/10.1016/j.tig.2013.01.006 PMID: 23428114.

14. McClure KA, Sawler J, Gardner KM, Money D, Myles S. Genomics: a potential panacea for the perennial problem. American journal of botany. 2014; 101(10):1780–90. https://doi.org/10.3732/ajb.1400143 PMID: 25326620.

15. Larsen B, Toldam-Andersen TB, Pedersen C, Ørgaard M. Unravelling genetic diversity and cultivar parentage in the Danish apple gene bank collection. Tree Genetics & Genomes. 2017; 13(1):14. https://doi.org/10.1007/s11295-016-1087-7

16. Gardner KM, Brown P, Cooke TF, Cann S, Costa F, Bustamante C, et al. Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing. G3: Genes|Genomes|Genetics. 2014; 4(9):1681–7. https://doi.org/10.1534/g3.114.011023 PMID: 25031181

17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. Epub 2009/05/20. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168; PubMed Central PMCID: PMCPmc2705234.

18. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20(9):1297–303. https://doi.org/10.1101/gr.107524.110 PubMed PMID: PMC2928508. PMID: 20644199

19. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–75. Epub 2007/08/19. https://doi.org/10.1086/519795 PMID: 17701901; PubMed Central PMCID: PMCPmc1950838.

21. Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, et al. Genetic structure and domestication history of the grape. Proceedings of the National Academy of Sciences. 2011; 108(9):3530–5. https://doi.org/10.1073/pnas.1009363108 PMID: 21245334

22. Butts CT. network: a Package for Managing Relational Data in R. Journal of Statistical Software. 2008; 24(2):1–36.

23. Purcell S. PLINK v.1.07 2009. Available from: http://pngu.mgh.harvard.edu/purcell/plink/.

24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007; 81(3):559–75. https://doi.org/10.1086/519795 PMID: 17701901

25. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014; 197(2):573–89. Epub 2014/04/05. https://doi.org/10.1534/genetics.114.164350 PMID: 24700103; PubMed Central PMCID: PMCPMC4063916.

26. Kalinowski ST, Taper ML, Marshall TC. Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Molecular Ecology. 2007; 16(5):1099–106. https://doi.org/10.1111/j.1365-294X.2007.03089.x PMID: 17305863

27. Jombart T, Ahmed I. adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011; 27(21):3070–1. https://doi.org/10.1093/bioinformatics/btr521 PMID: 21926124

28. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008; 24(11):1403–5. https://doi.org/10.1093/bioinformatics/btn129 PMID: 18397895

29. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.

30. Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong GY, Myles S. LinkImpute: Fast and Accurate Genotype Imputation for Non-Model Organisms. G3. 2015; 5(11):23383–2390. https://doi.org/10.1534/g3.115.021667 PMID: 26377960.

31. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2009.

32. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008; 3. https://doi.org/10.1371/journal.pone.0003376 PMID: 18852878

33. Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJA, Huvenaars KHJ, et al. Sequence-Based Genotyping for Marker Discovery and Co-Dominant Scoring in Germplasm and Populations. PLOS ONE. 2012; 7(5):e37565. https://doi.org/10.1371/journal.pone.0037565 PMID: 22662172

34. Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, et al. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proceedings. 2011; 5(7):P54. https://doi.org/10.1186/1753-6561-5-s7-p54

35. Sun R, Chang Y, Yang F, Wang Y, Li H, Zhao Y, et al. A dense SNP genetic map constructed using restriction site-associated DNA sequencing enables detection of QTLs controlling apple fruit quality. BMC Genomics. 2015; 16:747. Epub 2015/10/07. https://doi.org/10.1186/s12864-015-1946-x PMID: 26437648; PubMed Central PMCID: PMCPmc4595315.

36. Money D, Migicovsky Z, Gardner K, Myles S. LinkImputeR: user-guided genotype calling and imputation for non-model organisms. BMC Genomics. 2017; 18(1):523. Epub 2017/07/12. https://doi.org/10.1186/s12864-017-3873-5 PMID: 28693460; PubMed Central PMCID: PMCPMC5504746.

37. Migicovsky Z, Li M, Chitwood DH, Myles S. Morphometrics Reveals Complex and Heritable Apple Leaf Shapes. Frontiers in Plant Science. 2018; 8(2185). https://doi.org/10.3389/fpls.2017.02185 PMID: 29354142

38. Gompert Z, Mock KE. Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. Molecular ecology resources. 2017. Epub 2017/02/06. https://doi.org/10.1111/1755-0998.12657 PMID: 28150424.

39. Chagné D, Kirk C, Whitworth C, Erasmuson S, Bicknell R, Sargent DJ, et al. Polyploid and aneuploid detection in apple using a single nucleotide polymorphism array. Tree Genetics & Genomes. 2015; 11 (5):1–6.

40. Ayres KL. The expected performance of single nucleotide polymorphism loci in paternity testing. Forensic Science International. 2005; 154(2–3):167–72. http://dx.doi.org/10.1016/j.forsciint.2004.10.004 PMID: 16182962

41. Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006; 7(10):771–80. http://www.nature.com/nrg/journal/v7/n10/suppinfo/nrg1960_S1.html. https://doi.org/10.1038/nrg1960 PMID: 16983373

42. Bredsted HC. Haandbog i dansk Pomologi, 2. æbler: Hempelske Bog- og Papirhandels Forlag, Odense; 1893.

43. Pedersen A. Danmarks Frugtavl, Beretning fra Fællesudvalget for lokale Iagttagelsesplantninger og Frugtsortundersøgelser: Copenhagen; 1925.

44. Pedersen A. Danmarks Frugtsorter, 1. del. æbler: Alm. Dansk Gartnerforening, Copenhagen; 1950.

45. Matthiessen C. Dansk Frugt: H. Hagerup, Copenhagen; 1913.

46. Hirschfeld CCL. Handbuch der Fruchtbaumzucht: Braunschweig; 1788.

47. Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi RK, et al. Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. PLoS One. 2013; 8(12):e84136. Epub 2013/12/25. https://doi.org/10.1371/journal.pone.0084136 PMID: 24367635; PubMed Central PMCID: PMCPMC3868579.

48. Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, et al. Estimating genomic diversity and population differentiation–an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. BMC Genomics. 2017; 18(1):69. https://doi.org/10.1186/s12864-016-3459-7 PMID: 28077077