BMC Genomics

RESEARCH ARTICLE Open Access

Integration of Infinium and Axiom SNP array data in the outcrossing species *Malus* × *domestica* and causes for seemingly incompatible calls

Nicholas P. Howard[1,2], Michela Troggio[3], Charles-Eric Durel[4], Hélène Muranty[4], Caroline Denancé[4], Luca Bianco[3], John Tillman[2] and Eric van de Weg[5*]

Abstract

Background: Single nucleotide polymorphism (SNP) array technology has been increasingly used to generate large quantities of SNP data for use in genetic studies. As new arrays are developed to take advantage of new technology and of improved probe design using new genome sequence and panel data, a need to integrate data from different arrays and array platforms has arisen. This study was undertaken in view of our need for an integrated high-quality dataset of Illumina Infinium® 20 K and Affymetrix Axiom® 480 K SNP array data in apple (*Malus* × *domestica*). In this study, we qualify and quantify the compatibility of SNP calling, defined as SNP calls that are both accurate and concordant, across both arrays by two approaches. First, the concordance of SNP calls was evaluated using a set of 417 duplicate individuals genotyped on both arrays starting from a set of 10,295 robust SNPs on the Infinium array. Next, the accuracy of the SNP calls was evaluated on additional germplasm ($n = 3141$) from both arrays using Mendelian inconsistent and consistent errors across thousands of pedigree links. While performing this work, we took the opportunity to evaluate reasons for probe failure and observed discordant SNP calls.

Results: Concordance among the duplicate individuals was on average of 97.1% across 10,295 SNPs. Of these SNPs, 35% had discordant call(s) that were further curated, leading to a final set of 8412 (81.7%) SNPs that were deemed compatible. Compatibility was highly influenced by the presence of alternate probe binding locations and secondary polymorphisms. The impact of the latter was highly influenced by their number and proximity to the 3′ end of the probe.

Conclusions: The Infinium and Axiom SNP array data were mostly compatible. However, data integration required intense data filtering and curation. This work resulted in a workflow and information that may be of use in other data integration efforts. Such an in-depth analysis of array concordance and accuracy as ours has not been previously described in the literature and will be useful in future work on SNP array data integration and interpretation, and in probe/platform development.

Keywords: SNP Array, Single nucleotide polymorphism, Genotyping, *Malus*

* Correspondence: eric.vandeweg@wur.nl
[5]Department of Plant Breeding, Wageningen University and Research, Wageningen, The Netherlands
Full list of author information is available at the end of the article

Background

Single nucleotide polymorphism (SNP) array technology has been increasingly used to generate large quantities of SNP data for use in genetic studies. Over time, next generation arrays are developed that use new sequence data and/or new genome drafts to either refine or expand upon the set of SNPs used in previous arrays. Additionally, different SNP array technologies have been developed, resulting in different array platforms, creating a need for data harmonization and integration [1].

This need has been faced in apple (*Malus* × *domestica*), where a large amount of SNP array data has been generated using the Infinium® IRSC 8 K [2] and 20 K apple SNP arrays [3] on thousands of accessions through over thirty published, as well as ongoing, studies on pedigree reconstruction, genetic linkage map construction, identification of polyploids and aneuploids, quantitative trait loci identification, genome-wide association, and genomic selection; this data has also been used in downstream research like de novo genome assemblies and methodology development for the calling of SNPs [2, 4–33]. These previous and on-going studies have relied on a single SNP array platform, however a recent study provided whole genome SNP data on over 1400 mostly old, unique apple cultivars [32] using the Affymetrix Axiom® Apple 480 K SNP array [34]. Hence, ongoing and future studies on genetic relationships among apple cultivars could benefit from the integration of data across these platforms.

When newer arrays are simply updated arrays with additional SNPs that utilize the same platform, an evaluation of concordance of data among common accessions is straightforward and concordance is often high, such as with the BovineLD BeadChip [35] and the barley 50 K iSelect SNP array [36]. Concordance between the Infinium 8 K and 20 K apple SNP arrays has not been reported, but integration of SNP data across these arrays was seamless in Vanderzande et al. [29]. However, when SNP calls are compared across different platforms that use different technology, such as between the Infinium 20 K and the Axiom 480 K apple SNP arrays, concordance rates may be more variable. This variability is likely due in large part to differences in the chemistry, probe lengths, probe densities used across these platforms, and/or differences in the genotyped germplasm. Concordance rates between the Illumina Infinium 20 K and Affymetrix Axiom Apple 480 K SNP array data were reported as 96 to 98%, based on 53 common individuals [34]. This high rate is promising and is in line with those found in other organisms: an average concordance of 96 to 98.8% was reported in human [37], sheep [38], and swine [39]. However, levels of concordance were not documented at the individual SNP level, as would be needed for accurate data integration. Additionally, none of these studies reported on the technical or biological reasons for the observed SNP call discordances.

In comparative work prior to this, compatibility between array platforms has been determined by evaluating the concordance of SNP calls of genetically duplicate samples genotyped on both platforms, which is usually limited to few individuals. Such evaluations would be made more useful by considering SNP call accuracy via assessments of Mendelian inconsistent and Mendelian consistent errors [40] across direct parent-offspring relationships. This approach could increase the number of informative comparisons and could also expand the data chain to multiple successive generations. Moreover, the use of inheritance patterns surpasses analyses on duplicate genotypes in determining the precise genotype of an individual, allowing, for instance, the revealing of null alleles. The power of compatibility studies may increase even further by an integrated Mendelian error analyses on a mixed data set, rather than within array analyses. The identification and troubleshooting of Mendelian inconsistent and consistent errors have been previously described using Infinium SNP array data in apple, cherry, and peach [29]. Extensive pedigree information exists for apple from breeding records (e. g. from websites such as https://hort.purdue.edu/newcrop/pri/), pomological textbooks (e. g [41].), historic pedigree reconstruction studies [6, 20, 27, 29, 31–33, 42–45], and may also be revealed by the available SNP data.

Mendelian inconsistent and consistent errors in SNP array data can result from the presence of secondary polymorphism(s) on probe sites and/or the presence of duplicated or paralogous sequences. Secondary polymorphisms are sequence differences between the probe and intended target genomic sequence. They may impact the affinity by which a probe binds to a genomic sequence, resulting in distinct signal intensities for the same marker allele (e.g., *B* and *b* for high and low intensity respectively). Thus, individuals with an alternate allele(s) at these secondary polymorphisms have distinct clustering patterns (e.g., *Ab* in addition to *AB).* As individuals may differ for their secondary polymorphism, this gives raise to multiple sub-clusters for the heterozygous genotypes. The location of the *Ab* and *aB* cluster moves towards the *AA* and *BB* homozygous clusters, respectively, with decreasing intensity of the *a* and *b* alleles, which may impact cluster separation and calling accuracy. Secondary polymorphisms may also lead to so called null alleles, where probes completely or nearly completely fail to bind to some of the target genomic sequences [46, 47]. When not accounted for, null alleles may lead to unexpected genotype classes in segregating progenies and thereby in detected, but false, Mendelian inconsistent errors [6, 9, 29]. Duplicated and paralogous sequences may affect cluster separation too because they may also bind to probes, often reducing and compacting the effective cluster space for the target polymorphism [48].

Platforms may differ in their sensitivity to secondary polymorphism and duplicated sequences due to differences in chemistry by each platform [49–51]. The resulting probe hybridization data may also be interpreted differently due to differences in allele calling algorithms used in different genotyping software. How these details might altogether affect calling concordance and SNP call accuracy has yet to be revealed.

This study was undertaken in view of our need for an integrated high-quality dataset of Illumina Infinium® 20 K and Affymetrix Axiom® 480 K SNP array data in apple (*Malus* × *domestica*). While creating the integrated and highly curated dataset, we took the opportunity to thoroughly evaluate observed discordances and inaccurate SNP calls. Thus, the goals of this study were i) to qualify and quantify the compatibility, defined as SNP calls that are both accurate and concordant, of SNP calls across both arrays and ii) to evaluate reasons for observed probe failures and discordant SNP calls in order to improve SNP array data interpretation and probe/platform development. Towards these goals, this study included classical concordance evaluations across individuals genotyped on both platforms, as well as accuracy evaluations by detecting and evaluating Mendelian inconsistent and consistent errors across pedigrees on a mixed dataset. We hereby updated the apple integrated genetic linkage (iGL) map [15] and used a subset of SNPs that showed high performance on the Illumina platform as defined in this paper.

Results

Genetic map and Infinium data curation

There were 10,295 SNPs that passed the Infinium SNP data curation steps and thus were included in the genetic map. Of these, 94.1% (9685) were SNPs retained from the iGL map (Table 1; Additional file 1). For 12.4% (1206) of the SNPs retained from the iGL map, new positions were assigned, and these new positions were all within their respective genetic bins on the iGL, and also within a single centimorgan (cM) of their original position except for six SNPs (Additional file 2), and except for the first 5-6 Mb of LG1, where SNPs were ordered according to other ongoing studies on the Golden Delicious doubled haploid v1.1 (GDDH13) and the anther-derived trihaploid Hanfu line (HFTH1) whole genome sequences (WGSs) (Van de Weg, personal communication).

The level by which co-segregation patterns could be examined varied per SNP and some SNPs were only polymorphic in a small number of individuals. For example, numerous SNPs were only polymorphic in *Malus floribunda* 821 and a small number of its descendants, particularly its grandchild F2–26829–2-2, which was included in the discovery panel used to create the Illumina Infinium 20 K SNP array [3] and which served as a bottleneck in the introgression of the *Rvi6* gene for scab resistance from *Malus floribunda* 821 [52]. Information about minor allele frequencies (MAF) for the 10,295 SNPs included in this study across all individuals except genetic duplicates, can be found in Additional file 1.

Evaluation of within-platform repeatability

Repeatability of Infinium data was very high, with only 0.0016 and 0.0014% average discordant SNP calls observed per genotyped duplicate when evaluating the 10,295 SNPs included in this study (Subset 1), and when evaluating only 8412 SNPs that were also concordant in Axiom data (Subset 2), respectively (Table 2). For Axiom data, rates of discordant SNP calls varied among SNP subsets between 0.0117 and 0.3199% and were always higher than those for the Infinium data. Logically, discordancy was least for SNP sets that were first filtered for their performance on the Axiom array. Here, discordancy increased with increasing size of the subset (and thus with decreasing filtering intensity) from 0.0117%

Table 1 SNP inclusion/exclusion summary from the Illumina Infinium 20 K array

Classification		n
Included in this study	Retained from original 15,517 SNP iGL map	9685
	Successfully placed via physical information	610
	Total:	10,295
Excluded from this study	Poor clustering	4582
	Overlapping null and homozygous clusters	2604
	Monomorphic	382
	Extra cluster(s) causing illogical segregation	90
	Not included in iGL map, no physical location	60
	Illogical segregation	6
	Total:	7724

Table 2 Frequency of discordant SNP calls across 16 individuals genotyped twice on each array

Array	SNP subset Groups and their percentage (%) discordant*SNP calls					
	1	2	3	4	5	6
Infinium	0.0016	0.0014	–	–	–	–
Axiom	0.3199	0.0706	0.0134	0.0117	0.1516	0.0421

*For accessions that were genotyped more than twice on an array, a single average value was used in the across accession analyses
Subset 1: 10,295 SNPs from the Infinium array that passed the SNP data curation steps in the present study.
Subset 2: 8412 SNPs considered compatible between the array platforms in this study.
Subset 3: 275,223 SNPs in the Axiom array deemed robust [34].
Subset 4: 253,095 SNPs from subset 3 filtered for absence of more than one Mendelian inconsistent error in Muranty et al. [32].
Subset 5: 402,714 SNPs in the Axiom array that were classified as "Poly High Resolution" or "No Minor Homozygous" by Axiom Analysis Suite software.
Subset 6: 320,761 SNPs from subset 5 with SNPs removed that had Mendelian inconsistent errors in two or more parent-offspring relationships, discordant SNP calls in two or more duplicate pairs, or were heterozygous in doubled haploid accessions from Muranty et al. [32]

for the 253,095 SNP of subset 4 to 0.1516% for the 402,714 SNP of subset 6 (Table 2). Discordancy in Axiom data was highest for the SNPs that passed the Infinium data curation steps (Subset 1) (Table 2).

Compatibility of axiom data with included Infinium SNPs

The 417 individuals genotyped on both platforms (Additional file 3) showed an average concordance level of 97.1% across all 10,295 included SNPs, with a minimum of 96.0% and a maximum of 98.1%. Of the 10,295 included SNPs, 65% (6691) had no discordant call(s), while 35% had. Following the SNP data curation process, 8412 (81.7%) SNPs were deemed compatible. These compatible SNPs were classified into nine groups based on the type of adjustment needed to make the SNP compatible (Table 3). Examples of each classification can be found in Additional file 4 and examples of each classification for SNPs deemed discordant in Axiom data can be found in Additional file 5. Classifications per SNP are included in Additional file 1.

Table 3 Distributions of SNPs included in the genetic map study grouped by compatible and incompatible classifications

SNP classification	N		Classification code
Compatible	No adjustment needed	6417	A
	Adjustment needed – type of adjustment		
	Set ambiguous cluster position(s) between cluster groups in Axiom data	821	B
	Heterozygous cluster(s) in Axiom data mistakenly grouped with homozygous cluster	737	C
	Only 1 to 2 discordant SNP calls	223	D
	Malus floribunda 821 specific clustering discordancy between arrays	109	E
	Required adjustment to cluster position in Infinium data	37	F
	Only issue is rare null alleles that are difficult to identify in Axiom data	27	G
	Malus floribunda 821 specific SNP with discordant clustering	26	H
	True allele found for null alleles in Infinium data using Axiom data	11	I
	Set several SNP calls to missing data and unable to call rare null alleles in Axiom data	4	B/G
	Total:	8412	
Incompatible	Poor cluster differentiation with the Axiom array	808	J
	One or more heterozygous cluster overlapping with homozygous cluster in Axiom data	663	K
	All clusters overlap, with no clear cluster differentiation in Axiom data	161	L
	Inconsistent and irresolvable clustering between platforms	86	M
	Extra cluster(s) causing inconsistent clustering and/or illogical co-segregation	78	N
	Missing one cluster in Axiom data	51	O
	Normal clustering but had more than two unresolvable Mendelian inconsistent errors and/or more than 5 Mendelian consistent errors observed in > 5 unrelated individuals in Axiom data	25	P
	Inability to easily identify common null alleles in Axiom data	9	Q
	SNP not in Axiom array	2	R
	Total:	1883	

Of the 8412 compatible SNPs, 6417 (76.3%) required no additional adjustments, whereas 1995 (23.7%) did. The most common adjustments needed were to set errant Axiom SNP calls between clusters to missing data (class B), to reassign clusters in Axiom data (class C), usually heterozygous, and to set to missing data discordant SNP calls when there were only one or two of these (class D). These were followed by four less common classes. Examples of each class and some further descriptions of each can be found in Additional file 4.

Incompatibility between Infinium and axiom platforms

There were 1883 SNPs where Axiom data was deemed incompatible with included Infinium data. They were classified into nine different classes (J-R) (Table 3). The most common issues observed were poor clustering that resulted in an inability to make accurate SNP calls (class J) and overlapping of the heterozygous cluster with one of the homozygous clusters (class K). These were followed by seven less common classes. Examples of each class and some further descriptions of each can be found in Additional file 5.

The effects of paralogous binding sites on SNP exclusion and incompatibility

Probe sequences were retrieved on the expected chromosome from the iGL linkage map for 10,075 (97.9%) of the 10,295 included SNPs and 7175 (92.9%) of the 7724 excluded SNPs (Additional file 1). Of the retrieved probes, 96.0 and 90.8% gave a perfect match (E-value 1.52E-19) with the included and excluded SNPs, respectively. Of the 8412 compatible and 1883 incompatible SNPs, 97.0 and 91.9% showed a full match and 0.08 and 0.63% had less complete matches, with E-values higher than 1E-12, respectively. Hence, included or compatible SNPs had higher sequence similarity, and excluded or incompatible SNPs had lower similarity as estimated by E-values (Fig. 1).

Increasingly lower inclusion rates were correlated with increased numbers of BLAST hits beyond a single BLAST hit (Fig. 2). This was true for each of the three different E-value thresholds used. Inclusion rates decreased from a high of 66% with a single BLAST hit to a low of 6% with more than 10 hits (E-value< 1.0E-16). The results for all three E-value thresholds had this same general trend. Compatibility was slightly sensitive to the number of BLAST hits, with greater than four BLAST hits associated with a reduced compatibility rate (Additional file 6). However, this trend was only observed across a small number of SNPs, as only few SNPs with many BLAST hits passed the Infinium data curation step. Examples of clustering that was likely impaired by paralogous binding sites can be found in J-1 and J-2 of Additional files 5 and 1-1 and 1-2 of Additional file 7.

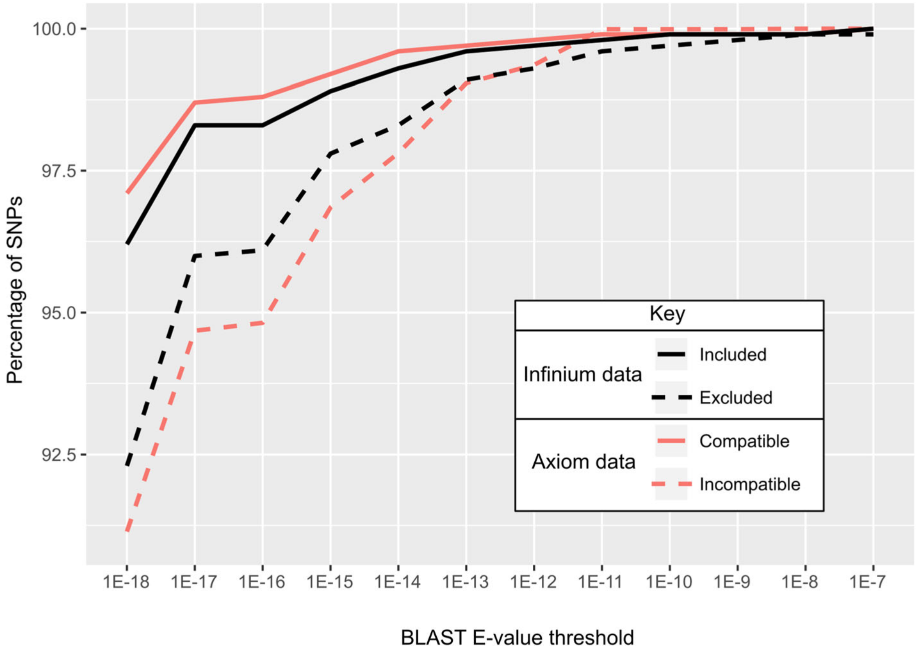
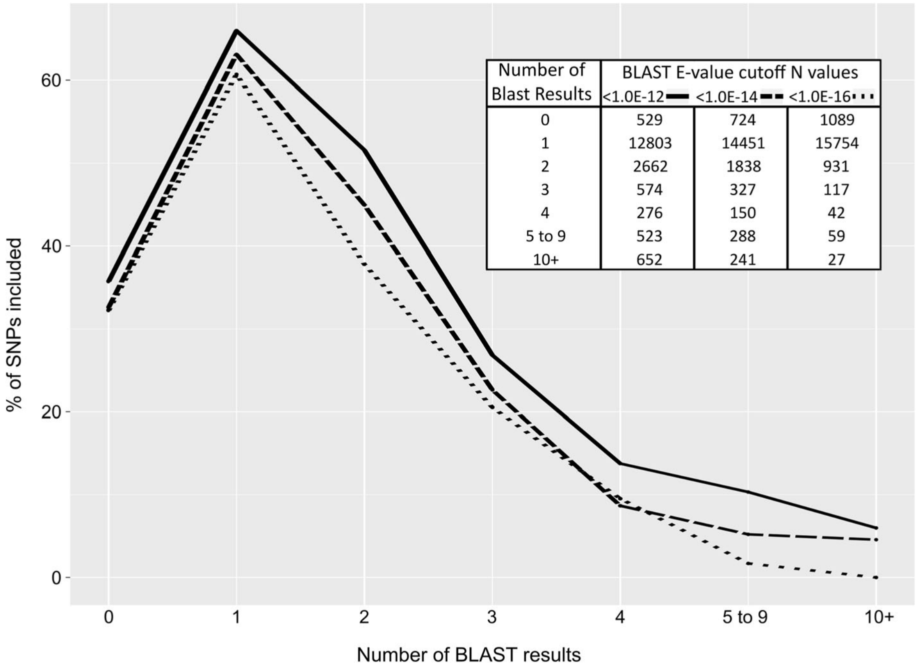
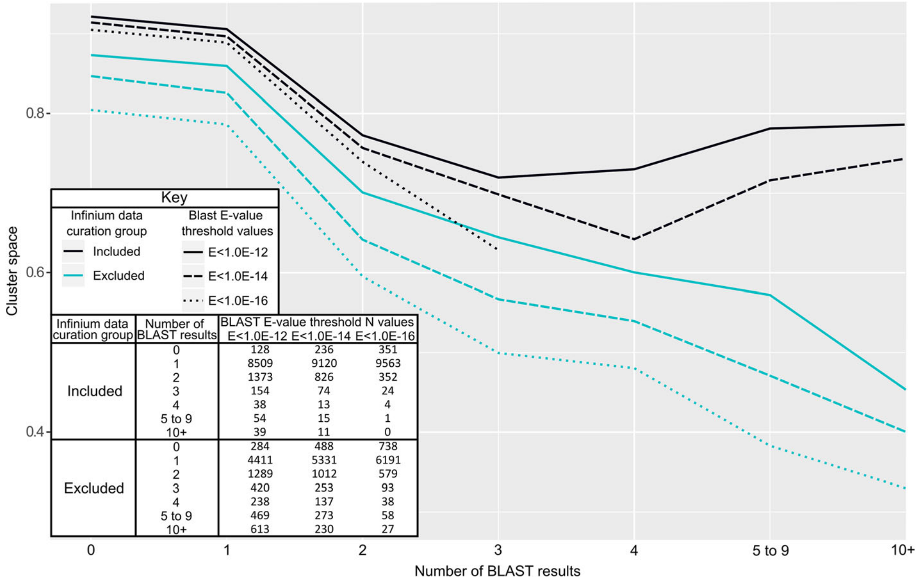


Fig. 1 Cumulative distribution plot demonstrating probe sequences for included/compatible SNPs have lower BLAST E-values. Only SNPs from the Illumina Infinium 20 K Apple SNP array with at least one significant BLAST hit from the 50 nt Infinium probe sequences vs. the GDDH13v1.1 whole genome sequence on the expected chromosome were considered (*N* = 17,250). SNPs with accurate or inaccurate Infinium data were classified as included or excluded, respectively. SNPs with accurate or inaccurate Axiom data were deemed compatible or incompatible (with Infinium data), respectively



Number of Blast Results	BLAST E-value cutoff N values <1.0E-12 —	<1.0E-14 – –	<1.0E-16 ···
0	529	724	1089
1	12803	14451	15754
2	2662	1838	931
3	574	327	117
4	276	150	42
5 to 9	523	288	59
10+	652	241	27

Fig. 2 negative correlation between positive number of BLAST hits and SNP inclusion/compatibility. All 18,019 SNPs from the Illumina Infinium 20 K Apple Infinium array were considered. Three different stringency thresholds for a successful BLAST hit from the 50 nt Infinium probe sequences vs. the GDDH13v1.1 whole genome sequence were considered: 1E-12, 1E-14, and 1E-16. The numbers of SNPs within each group are listed in the included table. Higher numbers of BLAST hits were grouped together because of the diminishing number of SNPs that had higher numbers of BLAST hits. SNPs with accurate Infinium data were classified as included



Infinium data curation group	Number of BLAST results	BLAST E-value threshold N values E<1.0E-12	E<1.0E-14	E<1.0E-16
Included	0	128	236	351
	1	8509	9120	9563
	2	1373	826	352
	3	154	74	24
	4	38	13	4
	5 to 9	54	15	1
	10+	39	11	0
Excluded	0	284	488	738
	1	4411	5331	6191
	2	1289	1012	579
	3	420	253	93
	4	238	137	38
	5 to 9	469	273	58
	10+	613	230	27

Fig. 3 Additional BLAST hits result in lower average cluster space in Infinium SNP array data. Three different stringency thresholds for a successful BLAST hit from the 50 nt Infinium probe sequences vs. the GDDH13v1.1 genome were considered: E < 1E-12 (solid line), E < 1E-14 (dashed line), and E < 1E-16 (dotted line). Data points were excluded from the figure if they were comprised of fewer than 10 SNPs. Cluster space was calculated for each SNP by the difference between 5 and 95% quantiles of observed Theta values from Infinium cluster plot data. SNPs with accurate or inaccurate Infinium data were classified as included or excluded, respectively

However, possible paralogous binding sites were not always associated with problematic clustering (ex. 1–3 in Additional file 7).

Included SNPs had average higher available cluster space than excluded SNPs, regardless of the number of BLAST hits per SNP (Fig. 3). Available cluster space also decreased with increasing numbers of BLAST hits and decreasing BLAST E-value thresholds. This decrease was least among included SNPs and strongest with the more stringent threshold with both included and excluded markers. Some SNPs were still included in the presence of three BLAST hits with the lowest E-value threshold but never were successful in the presence of four such hits.

The effects of secondary polymorphisms on SNP exclusion and incompatibility

The presence of secondary polymorphism(s) at probe sites negatively impacted SNP inclusion of Infinium data and SNP call compatibility of included SNPs in Axiom data. Increasing numbers of secondary polymorphisms were correlated with reduced SNP inclusion and compatibility rates (Additional file 8). We could not effectively compare simultaneously the effects of multiple secondary polymorphisms and their variable positions on SNP inclusion and compatibility rates could not be effectively compared due to low sample sizes for each case. Instead, SNPs that had only a single secondary polymorphism to the intended target genomic sequences were further examined. The closer a single secondary polymorphism was to the target SNP, the more likely this SNP was to be excluded during the SNP curation process in Infinium data or to be deemed discordant (Class C in Table 3) (Fig. 4). Probes with secondary polymorphisms within the first three positions from the target SNP were mostly excluded, due to which too few of them remained to effectively examine their compatibility with Axiom data. Because of this, we also examined Axiom cluster plots for these SNPs and they too had poor clustering. Among included SNPs, the presence of secondary polymorphisms also frequently resulted in the presence of additional heterozygous cluster(s) that were mistakenly called as homozygous in Axiom data (Class C, Table 3), requiring manual cluster adjustment to achieve compatibility (Fig. 4). This effect gradually diminished with increasing distance between the secondary polymorphisms and the 3′-ends of the probes (Fig. 4).

Automated clustering of Infinium data in GenomeStudio was assessed for 187 included SNPs that had a single secondary polymorphism and that required manual cluster adjustment in Axiom data to determine whether

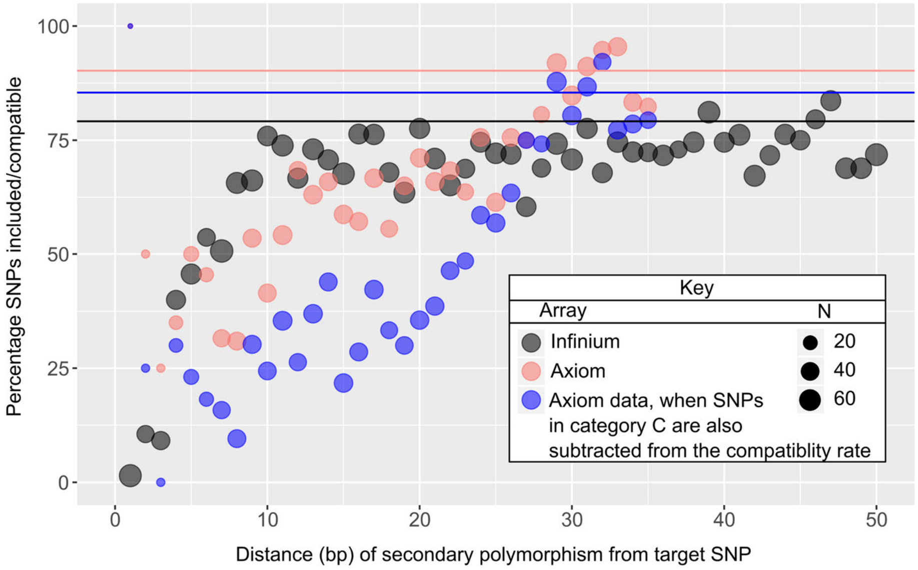


Fig. 4 Closer proximity between secondary polymorphisms and target SNPs result in decreased SNP inclusion and compatibility rates. Secondary polymorphisms and their positions were identified via sequence alignment of 53 cultivars to the GDDH13v1.1 genome. SNPs with accurate Infinium data were classified as included and SNPs with accurate Axiom data were deemed compatible (with Infinium data). The inclusion rate of Infinium data is represented by black. The compatibility of these included SNPs with Axiom data with and without class C SNPs (those with additional heterozygous cluster(s) in Axiom cluster plots requiring manual adjustment to make compatible) being classified as compatible are represented by pink and blue, respectively. The horizontal lines represent the inclusion and compatibility rates for SNPs with no identified secondary polymorphisms at their probe site for the three respective data sources that sized 6632 (black), 6011 (pink), and 6011 (blue) SNPs. SNPs included in this analysis had their alternate allele present in at least 10% of the sequenced individuals, had no more than 25% missing data across the sequenced individuals, and had probe sequence with a single BLAST hit on the GDDH13 WGS with an E-value <1E-12

these SNPs had also required, or would benefit from, manual cluster adjustment in Infinium data as well. The presence of additional heterozygous clusters was often still observed in Infinium cluster plots (ex. C-3 in Additional file 4), but these additional clusters were mostly correctly called as heterozygous. There were only five cases (2.7%) where the additional heterozygous cluster(s) were set mostly or completely as missing data and only a single case (0.5%) where manual cluster adjustment had been also necessary to make SNP data accurate in Infinium data (2–1 and 2–2 in Additional file 7).

Some secondary polymorphisms were identified as likely causes for discordant SNP calls that could be resolved through manual adjustments to Axiom cluster plots (Table 3 – class C; ex. C-1 – C-4 in Additional file 4), or to Infinium data (Table 3 – Class I; ex. I in Additional file 4), or to both platforms' cluster plots (2–1 and 2–2 in Additional file 7). Others were identified as the likely cause of SNP discordancy (ex. J, K, M, N, O, and Q in Additional file 5), or causing probe failures or poor clustering in both arrays (3 in Additional file 7). There were also occasional instances where a secondary polymorphism possibly caused probe failure in Infinium data but not in Axiom data (ex. 4 in Additional file 7). Finally, some instances were identified where prevalent secondary polymorphisms were close to the target SNPs that lacked problematic clustering (ex. 5–1 and 5–2 in Additional file 7).

The inclusion and concordance of SNPs whose probes had BLAST E-values greater than 1.0E-12 depended on multiple different factors (Additional file 9). The differences between the probe and target sequences for SNPs whose BLAST hits were identical in both the GDDH13 and HFTH1 WGSs were mostly present in the middle or the 5′ end of the probe sequences. However, there were five SNPs (12.5% of those evaluated) with apparent mismatches within the first 10 positions from the target SNP. The observed sequence differences between probes and the GDDH13 WGS seem to be almost exclusively true, as both alternative sequences were mostly confirmed by re-sequencing data from individuals in germplasm group 5, or by the HFTH1 WGS [53]. Only occasional SNPs were identified where a likely error in the GDDH13 WGS was the reason for the low E-value observed in the BLAST data generated. An example of true major sequence differences was probe SNP_FB_0514806 (Infinium name, with the Axiom name AX-115193385), which had a six-nucleotide gap at position 25–31 and a mismatch at position 41 from the target SNP on the GDDH13 WGS but had a full match on the HFTH1 WGS [53]. This SNP was included and compatible without needing any adjustments to the Axiom cluster plot (Classification A), despite the substantial difference between the probe and target genomic position on the GDDH13 WGS.

Discussion

Our results demonstrate a high degree of SNP compatibility (81.7%) between the Infinium® 20 K SNP array and the Affymetrix Axiom® Apple 480 K SNP array across 10,295 SNPs we deemed robust in Infinium data (Table 3). Our initial SNP call concordance rate for duplicate genotypes across the two platforms, 97.1%, was in line with what had been previously reported for these arrays [34]. Our further efforts towards resolving discordances and excluding SNPs with inaccurate SNP calls on either array have now allowed the creation of a combined dataset including SNPs with compatibility clustering and thus will have an exceedingly low SNP call error rate, even for individuals that were only genotyped on one of the two arrays. This highly curated combined dataset will be useful in subsequent studies. Additionally, we were able to identify reasons for both probe failures and discordance through a combined analysis of cluster plot data, SNP co-segregation patterns, and sequence data alongside the GDDH13 WGS [16]. Such an in-depth analysis of array concordance and accuracy has not been previously described in the published literature and may prove useful for future array design and SNP array data interpretation. Of the 8412 SNPs deemed compatible, 2030 are also present on the Illumina Infinium 8 K SNP array (Additional file 1) [2], indicating prospects for data integration with that array as well.

Genetic map and Infinium data curation

A set of 10,295 SNPs passed our rigorous data curation step (Table 1 and Additional file 1). They represent 57% of the initial 18,019 on the Infinium array and exclude 5613 SNPs that were originally included in the iGL map [15]. The set is larger than the 6849 SNPs used in QTL discovery studies on multi-parent populations in apple [14, 17], thanks to our rigorous filtering and curation effort. Many of the excluded SNPs showed *AB* subclusters and/or null alleles, which made them highly informative for the purpose of linkage map creation using full-sib families [15, 54], but which made them less suitable for use in a diversity set of germplasm due to an inability to accurately call null alleles across such germplasm.

Evaluation of within-platform repeatability

Overall, average within-platform repeatability of genetic duplicates was high, depending on the SNP subset used (Table 2). This high repeatability is in line with that observed for other Illumina arrays, such as that for *Zea mays* (99.7076%) [55], *Medicago sativa* (100%) [56], and *Populus nigra* (100%) [57], and for other Axiom arrays, such as that for *Glycine max* (> 99%) [58], *Cicer arietinum* (> 99%) [59], and *Juglans regia* (99.2%) [60]. Although in this study the Infinium platform had a higher repeatability than the Axiom platform, this may not be an equal comparison, as the calculated repeatability rate was highly affected by our intense preliminary filtering process performed on the Infinium data.

It should be noted that a very minor level of discordancy could be due to minute mutation differences between clones held in different collections. However, the chance that a SNP coincides with a mutation between clones is thought to be rare because mutation rates are usually low, and regions of mutation are usually small. Therefore, we assumed discordant SNP calls to have been entirely due to cluster position variance within each platform. Some level of discordancy could be attributed to laboratory or sampling issues. Low quality DNA has been the reason for high levels of discordancy among technical replicates in the Axiom *J. regia* 700 K SNP array [60].

Compatibility of SNP clustering between platforms

The results of this study should not be used to determine whether the two platforms differ in their overall performance, meaning that the results and interpretations are not meant to be an endorsement or a repudiation of either platform. This is because the SNPs on the Infinium 20 K apple array had all been consciously included in the Axiom array to maximize the potential for cross-platform compatibility, even when Axiom's selection criteria for array inclusion or calling performance were violated [34]. The clustering discordances observed across both arrays in this study may be due to the differences in chemistry and probe length between the platforms [49–51]. The Infinium and Axiom SNP platforms make use of a selective, locus specific primer/probe that runs up to the targeted SNP. The platforms differ on how the targeted SNP is captured. The Infinium platform is based on a polymerase executed extension reaction building in fluorescently labelled nucleotides [49] whereas the Axiom platform is based on an end-to-end hybridization between a locus specific and a dye-labelled and allele specific, but otherwise non-selective probe [51]. In addition, the locus specific probes are 50-mers with the Infinium and 35-mers with the Axiom platform. Furthermore, each Infinium probe is bound in multiple copies to 20–30 beads that are in solution [61], while with Axiom the locus specific probes are located at two defined positions on a two-dimensional substrate (while the non-selective allele specific probe is in solution) [51]. While this study gave insight into the possible role of probe size (see discussion section "Effects of secondary polymorphisms on inclusion and compatibility"), how the other differences between the arrays affect calling concordance and SNP call accuracy has yet to be revealed. Despite the differences between the platforms, our results show a high degree of concordance (97.1%) for duplicate genotypes across the 10,295 SNPs deemed to have robust performance on the Infinium array. This rate was in line with the 98% previously observed [34], which evaluated a subset of the duplicate pairs evaluated in this study. This rate is also comparable to the 98.8% for humans [37], the 97.38% observed for sheep [38], and the 98.4% for swine [39].

However, despite this high level of concordance, 1995 SNPs showed significant clustering differences between the platforms, resulting in the need for additional work to maximize compatibility. Of the 10,295 SNPs included from the Infinium array, 19% required adjustments to make them compatible and 18% were deemed incompatible. Researchers using data from both arrays could try to avoid issues by only using the 6417 SNPs that did not require any additional adjustments, though this would result in either a smaller number of SNPs available or a higher level of missing data present in the combined dataset, which this study sought to avoid. However, manual reclustering of the 1995 cluster plots was very time consuming. The most common manual curation performed was to set some percentage of individuals to missing data in the Axiom data due to their errant positions between clusters in cluster plots (B and D in Table 3 and Additional file 4). This issue occurred in 52% of the 1995 SNPs requiring manual curation to make them compatible and could possibly have been greatly reduced or largely eliminated by using more stringent threshold settings in the Axiom Analysis Suite [62]. Though this might have also resulted in additional missing data, it would have been a time saving step that would have possibly still retained the majority of accurate SNP calls for the relevant SNPs. The other SNPs requiring manual curation required more advanced analyses to identify and are not currently as amenable to automated adjustment. These generally required either manual reassignment of sub-clusters (39%; C and F in Table 3 and Additional file 4) or the addressing of issues related to null alleles (2%; G and I in Table 3 and Additional file 4). Sub-clusters and issues relating to null alleles could be targeted in future clustering algorithms to improve SNP call accuracy, possibly by the inclusion of pedigree information in clustering algorithms. However, such automated clustering may still result in removal of SNPs with accurate clustering, as automated clustering used in Bianco et al. [34] excluded 8.4% of the SNPs that required no adjustment to make compatible in this study (noted per SNP in Additional file 1). As such automated clustering solutions are currently unavailable, manual curation methods were used in this study. We considered the effort of manual curation worth our time, as we wished to obtain an integrated SNP array dataset with a maximum number of compatible SNPs to use for downstream analyses. However, this manual curation was time consuming, with diminishing returns for each unit time spent on addressing the increasingly complex cases. Others seeking to create such an integrated SNP array dataset with highly accurate SNP calls may do well to establish a desired balance between time spent and level of curation while considering the requirements of the dataset in the context of their research aims.

Effects of paralogous binding sites on inclusion and compatibility

In our study, the effect of the overall sequence (dis) similarity between probes and intended genomic sequences on SNP inclusion and compatibility was examined. Hereto, all the SNPs' BLAST hits that met the various E-value thresholds were compared together regardless of where sequence differences occurred. This was a conscious decision, as the nature of the differences between intended and non-intended targets was often complex and involved multiple mismatches, making it difficult to quantitatively classify the differences in affinity at a more detailed level.

As much as 94.1% of the probe sequences had a full match to the GDDH13 WGS. This high proportion was not surprising, considering the full genetic relatedness between GDDH13 and 'Golden Delicious' [16], the source for the GDv1 and GDv2 WGSs that were used to design the Infinium and Axiom probes, respectively [3, 34]. The remaining 5.9% of probes may come from incompleteness of the GDDH13 WGS, from local sequence errors in the GDv2 or GDDH13 WGSs, or from sequence or structural differences between the two 'Golden Delicious' homologs. The former two reasons lead to misleading high E-values. The latter issue of structural differences is plausible considering the highly heterozygous nature of apple [63]. These differences could have resulted in true secondary polymorphisms and indels that gave rise to clustering issues. Hence, it is not surprising that full matches were more prevalent in included and in compatible SNPs than in excluded and in-compatible SNPs (Fig. 1).

Another factor that influenced exclusion and incompatibility rates was the number of paralogous binding sites as estimated by the number of BLAST hits. Probes with multiple BLAST hits were more likely to be excluded (Fig. 2) because of a loss in available cluster space (Fig. 3). Many SNPs had two BLAST hits, usually with the non-intended target having a higher E-value and often appearing on the homoeologous chromosomes from the recent whole genome duplication in *Malus* [63]. Other SNPs had more than two significant BLAST hits. These additional BLAST hits could be due to copy number variants. Indeed, physically adjacent significant BLAST hits were occasionally identified. In any case, the binding of a probe to one or more of these non-intended targets would result in additional signal, typically for only one of the two possible marker alleles, whichever was present at the non-intended site(s). This would cause the heterozygous and one of the homozygous clusters to shift towards the allele with extra representation (as in cases 1–1 and 1–2 of Additional file 7). Occasionally, both homozygous clusters may shift towards each other, reducing the effective cluster space from both sides of the cluster plot. This happens in the presence of multiple effective hits that produce additional signal for each of the two possible marker alleles. These observations are in line with Hyten et al. [48], which previously reported a reduction in cluster space due to paralogous binding sites. These results highlight the importance of a high-quality reference genome and adequate SNP filtering for array design.

Though SNPs whose probes had more BLAST hits were more likely to be removed during the SNP data curation steps and to be deemed discordant, many were both included and deemed compatible in Infinium and Axiom data, respectively (e.g., 1–3 of Additional file 7). This could be due to i) the presence of just one segregating locus that did not suffer from additional complicating issues like secondary polymorphisms of major effect, due to which clusters remained well separated despite a reduced effective cluster space, ii) paralogous binding site(s) indicated by the additional BLAST hit(s) having no or weak binding affinity to the probes, or iii) part of the BLAST hits being false due to remaining assembly errors in the GDDH13 WGS (see below).

Effects of secondary polymorphisms on inclusion and compatibility

Secondary polymorphisms make a probe differ from its target genomic sequence. The closer a secondary polymorphism was located to the 3′ end of the probe, the more likely the probe was excluded during the Infinium data curation step, was deemed incompatible with Axiom data, or required manual adjustment to Axiom cluster data to make the genotype calls compatible (Fig. 4). Probes with secondary polymorphisms on the first seven positions from the 3′ end had the lowest inclusion rate on Infinium data, and those excluded often had null alleles present. The increased presence of null alleles in probes with secondary polymorphisms within the 3′ portion of probes and their negative effect on concordance and accuracy has also been previously reported with Illumina human SNP arrays [46, 47].

A higher impact of secondary polymorphisms on clustering was observed for Axiom data than for Infinium data (Fig. 4). This is possibly due to the difference in probe size. The longer probe size in the Infinium array could result in a stabilizing effect by the 5’end for binding when the secondary polymorphism(s) exist in the middle or 3′ end of the probe. This stabilizing effect apparently allowed for successful SNP calling even when multiple SNPs or long gaps existed in the middle of the probe sequences, though in some cases this resulted in the need for some manual clustering (Additional file 9). With the smaller 35-mer Axiom probe, this stabilizing effect may be greatly reduced to approximately the final 7 nucleotides. The probable increase of stability with increasing probe size might have been one reason why the initial 30-mer oligonucleotides of the Axiom platform have been extended to the current 35-mers [51]. These results and insights have not been previously reported and could be used to guide SNP filtering for future array construction, including for decisions on probe size.

SNPs were identified with prevalent secondary polymorphisms close to the target SNP that nevertheless lacked problematic clustering (ex. 5–1 and 5–2 in Additional file 7). There were also instances where the exact cause of poor clustering could not be identified (ex. L-2 in Additional file 5). Sequence data and the GDDH13 WGS are not perfect, and it could be that we simply did not have the available information to detect the likely cause for poor clustering in some cases. Likely, there were also other factors unaccounted for in this study, such as the exact nucleotides that were mismatching or the GC content of probes.

Array construction and reference WGSs

Secondary polymorphism and paralogous sequences are known to negatively impact SNP array marker performance [46, 47], the first causes problems with probe hybridization which might affect the probe efficiency while the second might introduce false SNPs obtained from the consensus of sequences coming from each of the two slightly different regions when erroneously merged together. Therefore, measures are undertaken to limit their occurrence in the design of a new SNP array. Following the initial SNP discovery process, candidate SNPs are checked i) for the lack of additional polymorphisms in their flanking sequence and ii) filters on the read depth at the SNP site are used as a proxy for identifying the risk of erroneous calls due to the fusion of reads from paralogous regions. The effectiveness of the additional polymorphisms filtering step is a function of the size and genetic diversity represented by the discovery panel, which consisted of 13 and 63 individuals in for the 20 K Infinium and 480 K Axiom array, respectively. Hence, compatibility issue rates observed for the Infinium array SNPs observed in this work are likely not representative for the other 460 K SNPs included on the Axiom array since the filtering pipeline of the latter had more chances to identify additional polymorphisms in the probe due to the larger discovery panel. The quality of the reference WGS used in the alignment of sequencing data of the discovery panel members is another factor in array performance. To avoid the presence of paralogous sequences, several filters can be applied, such as a check on the read depth at the SNP site and/or a kmer analysis of the probes of selected candidate SNP markers. For example, it is possible to align the entire probe (or multiple subsequences of the probe i.e., kmers) against the reference genome and make sure that these sequences do not appear multiple times across the genome. The efficiency of this approach is a function of the size of the queried segments, the allowed number of mismatches, the sensitivity of the probe for mismatches given the genotyping platform, and the quality of the reference WGS. In the design of the 20 K array, the full probes (50-mers) were aligned to the reference genome allowing for two mismatches, while a kmer analysis was performed on all the 24-mers from the probes to make sure they did not appear multiple times in the genome [3]. Our results indicate this approach to be effective for the Axiom genotyping platform, but not for the Infinium, where probes with more than two polymorphisms in the first 24 positions from the target SNP may still result in relevant marker signal (see examples in Additional file 9). Also, filtering against secondary polymorphisms takes out a major source for null alleles. In contrast, identification of indels through a SNP discovery panel is quite error prone when the resequencing data are of short read length and at moderate read depth.

The quality of the reference WGS used in the design of an array also matters. The WGS used in the design of the 20 K Infinium array was of a complex nature. It consisted of a pseudomolecule for each chromosome in apple plus a series of additional scaffolds, which mostly were highly similar homologous and homoeologous sequences. These scaffolds could not be distinguished for their chromosomal origin, so could not be collapsed into the primary sequence. However, the recent availability of chromosome scale assemblies like the GDDH13 or that of the HFTH1 WGS [53], gives better opportunities to perform the aforementioned filters more effectively.

The GDDH13 whole genome sequence

While the GDDH13 WGS is a great improvement over the previous ‘Golden Delicious’ WGSs, three separate observations in our study revealed that many regions of the GDDH13 WGS are still misassembled (Additional file 1). First, some probes had more than one perfect BLAST hit, while the SNP array cluster plots showed signal for just one locus (1–3 in Additional file 7). Second, we observed mismatches between the GDDH13 physical positions and the genetic positions that were not due to mapping errors (see Additional file 1, column “phys_blast_GDDH13v1.1”). Third, we identified specific Infinium probes that did not give a BLAST hit on the GDDH13 WGS but resulted in regular heterozygous signal in ‘Golden Delicious’, the cultivar that served as source for the GDDH13 individual due to incompleteness in the GDDH13 WGS (Additional file 9). These results confirmed similar conclusions based on the comparison between the GDDH13 WGS and the iGL genetic linkage map [64]. A thorough evaluation of the assembly issues encountered is outside of the scope of this paper but is part of an ongoing study (Van de Weg, personal communication).

Conclusions

We identified 8412 SNPs that could be used to construct an accurate integrated Infinium and Axiom apple SNP array dataset. This specific result will aid future studies involving tracing the inheritance of haplotypes in a combined array dataset. Problematic clustering encountered in our study was primarily due to secondary polymorphisms, alternate probe binding locations interfering with probes’ target sequences, and differences between probes and target genomic sequences. The proximity of the sequence difference(s) to the target SNP between probes and their target and non-target sequences was identified as a major factor in SNP inclusion and compatibility. However, secondary polymorphisms identified roughly on the final 40 (of 50) positions on the Infinium array and the final seven positions on the Axiom array did not impact SNP inclusion and compatibility. The results could be used to help guide further SNP filtering for array construction and SNP array data interpretation.

Methods

Available array data

Illumina Infinium 20 K SNP array data came from previous [11, 15, 20, 27] and ongoing studies. Affymetrix Axiom 480 K SNP array data came from Bianco et al. [34] completed by Muranty et al. [32]. SNP intensity data were made available on request. SNP calling procedures progressively developed during the successive stages of this work as described below.

Germplasm

Each accession used for SNP curation, the analysis of duplicates, and the analysis of SNP call concordance and accuracy in this study is listed in Additional file 3. This group of accessions was filtered for being diploid and for having high SNP call quality using methods outlined in Vanderzande et al. [29]. Each cultivar was assigned a *Malus* UNiQue genotype code (MUNQ) value as described in Muranty et al. [32].

Five general groups of germplasm were distinguished, which partly overlapped. The first group was comprised of 1566 Infinium genotyped cultivars/accessions, some of which were included more than once as different accessions from different institutions. The second group was comprised of 1466 individuals genotyped on the Axiom array. The third group was comprised of 30 full-sib families totalling 2422 seedlings genotyped on the Infinium array and evaluated in various previous genetic linkage mapping and QTL discovery studies [11, 15, 21, 22, 27] (these families are listed in Additional file 10). The fourth group consisted of known pedigree ancestors of the above families (group 3) and individuals of groups 1 that were known to have direct genetic relationships. Finally, a fifth group comprised of individuals whose sequence data was used in the Axiom array SNP discovery panel [34] was used for the identification of secondary polymorphisms. Groups 1 and 2 were used to identify genetic duplicates between the two groups; these duplicates were used to evaluate concordance. Groups 3 and 4 were used for the genetic map revision, curation of Infinium SNP data, and filtering of SNPs. Groups 2, 3, and 4 were used for accuracy evaluations.

Identification of genetic duplicates analysed on both arrays

Germplasm groups 1 and 2 (Infinium and Axiom genotyped cultivars, respectively) were used to identify genetic duplicates. SNP calls for the first group were generated using default settings in GenomeStudio® v2.03 (Illumina Inc. San Diego, CA, USA). SNP calls for the second group were obtained from Muranty et al. [32]. Individuals were deemed genetic duplicates when they shared more than 97% of SNP calls from 7206 selected SNPs that were deemed “robust” in Muranty et al. [32] and, on the Infinium platform, not excluded from Vanderzande et al. [29], free of apparent null alleles in full-sib families from germplasm group 3, and that were included in the iGL map with single locus segregation [15]. This set of SNPs was only used for this purpose; a more thoroughly evaluated set of SNPs was used for subsequent steps. The cutoff value of 97% was established by comparing the SNP calls for the wild genotype *Malus floribunda* 821 from both platforms and rounding down to the closest whole percentage point, as *Malus floribunda* 821 was most prone to atypical clustering with each of the arrays and showed a relatively high level of discordant SNP calls compared to other expected genetic duplicates, due to its wild origin compared to the domesticated pool for which both arrays were initially built.

Genetic map revision, curation of Infinium SNP data, and filtering of SNPs

We used a revised version of the iGL map [15] to ensure adequate evaluation of the accuracy of SNP data at the later stages of this research. We revised the SNP sorting order and cM positions for SNPs in this map using the process described in Vanderzande et al. [29] using physical coordinates obtained by blasting [65] the 50 nt long probe sequences from all SNPs on the Infinium array [3] against the GDDH13 WGS [16]. BLASTn parameters used are as follows: Expected Value Threshold 0.001, Word Size 11, Match/Mismatch Scores 1 and − 2, Gap Costs; existence 5 and extension 2. Parameters were chosen to generate a dataset neutral to bias in detection of orthologs and paralogs [66]. To ensure the accuracy of the physical positions, results from the more extensive, 121 long probe source sequences [3] were used in a few instances where no results were found using only the 50 nt probe sequences. For SNPs with multiple matches, only matches with Expect (E) values < 1.0E-12, approximately equivalent to less than 4 mismatches in a 50 nt sequence, that were not in conflict with coordinates of genetically flanking SNPs were considered. When multiple matches with physical positions between flanking SNPs had E-values < 1.0E-12, the match with the lowest E-value was used. SNPs were rearranged to match these physical coordinates when not resulting in spurious double recombinations in seedlings used to construct the 8 K [20] and 20 K iGL maps [15]. We allowed re-ordering of SNPs up to 2 cM away from their original position. These new positions were accepted when they did not lead to false Mendelian consistent errors.

New cM position estimates were made for SNPs where rearrangements resulted in conflicts with the original cM positions. This was accomplished by using a linear algebra approach using the relative physical distances to the nearest flanking SNPs with concordant positions. Following this step, SNPs that were originally not included in the 20 K iGL map were placed into the genetic map. Their genetic positions were estimated using the same algebraic method as previously described. For rearranged SNPs that required new cM positions that had no blast hits, a cM position was assigned that was equidistant between the genetically flanking SNPs. Occasional SNPs that required rearrangement and new cM positions to address spurious double recombinations (due to incorrect SNP order) had BLAST hits that were inconsistent with flanking SNPs. If one of these adjacent SNPs was inconsistent with the SNP that needed a new cM position and its other flanking SNP, another adjacent SNP that had a consistent BLAST hit was used for cM position imputation using linear algebra. If this was not possible, a cM position that was equidistant between the adjacent SNPs was assigned instead. The ordering of SNPs for the first ~ 6 Mb of LG1, was based on unpublished data made available by Eric van de Weg.

After the construction of this preliminary updated map, all Infinium genotyped individuals that had direct parent-child relationships (listed in Additional file 3) were called for all SNPs in the updated map using the default clustering settings in GenomeStudio. Mendelian consistent and inconsistent errors were identified using GenomeStudio and FlexQTL™ [67] as outlined in Vanderzande et al. [29]. The resulting output on Mendelian errors from FlexQTL™ was used to curate the data as described in Vanderzande et al. [29] and to classify SNPs as either included or excluded. SNPs classified as included were used in later portions of our study and did not have unresolved Mendelian inconsistent errors and were not involved in numerous spurious double recombination events, or false Mendelian consistent errors. SNPs that did not meet these criteria were visually inspected to determine whether they could be addressed to be able to fit the criteria for being included or whether they should instead be excluded. Excluded SNPs generally had poor clustering or problems due to null alleles. Poor clustering was defined as a lack of clarity in differentiating between *AA*, *AB*, and *BB* clusters. SNPs with null alleles had individuals with cluster positions of lower intensity, or in other words, a lower "Norm R" value, in GenomeStudio cluster plots compared to true homozygous clusters. These were identified via the individuals with those lower intensity cluster positions also having a parent or an offspring with a cluster position of lower intensity located below the opposite homozygous cluster. In such a parent-offspring relationship, GenomeStudio assigns one of the individuals in this relationship a SNP call of "*AA*" and the other individual a SNP call of "*BB*", when in reality the two individuals share a third allele, termed "Null" (*N* in additional files). These null alleles may be due to either indels or additional polymorphisms on probe sequences resulting in calling problems obscuring the true alleles at the intended locus [6, 9, 67]. In segregating families, the presence of such alleles could be unequivocally assessed based on atypical segregation ratios. For cultivars with known, genotyped parentages, null alleles may also generate Mendelian consistent and Mendelian inconsistent segregation errors. For cultivars for which parents are unknown or not genotyped and also lack genotyped offspring, calling was solely based on the shape of genotype clusters in GenomeStudio cluster plots. Here, in many cases, heterozygous null genotypes could not be clearly differentiated from truly homozygous genotypes. If the null allele for such SNPs was rare, the SNP was still classified as included. This distinction was made because we wished to avoid excluding SNPs if they were only problematic in a very small number of individuals, particularly those involving genetically unique or obscure individuals. SNPs were also removed if they were found to be monomorphic.

Identification and evaluation of discordant SNPs in duplicated individuals

Only SNPs that passed the previously described curation steps on the Infinium array data were considered for evaluating discordant SNPs in duplicate individuals. SNP calls for Axiom genotyped individuals were obtained using Axiom Analysis Suite software using the default diploid settings. Discordant SNP calls were identified for all pairs of individuals that were genotyped on both arrays. They were further evaluated to determine the nature of the discordancy and whether manual cluster adjustments could be made to improve concordance. In cases of ambiguity regarding which of the discordant calls was accurate, pedigree information was used to evaluate inheritance or co-segregation using FlexQTL™. Confirmed null alleles in Infinium data were not tallied as discordant SNP calls with Axiom data, as they had been hand called during the Infinium data curation step. If manual cluster adjustment of one or both platforms could result in the resolution of the discordance, the SNP was included in the following steps. If not, the SNP was deemed discordant.

Identification of higher order SNP data quality issues affecting concordance and accuracy

Following the previous investigation of discordance between duplicated genotypes, discordance and accuracy were investigated in a single, mixed Infinium-Axiom dataset using all germplasm included in this study. Individuals were filtered to include only those having published parent-offspring relationships (Additional file 3). This dataset was imported into FlexQTL™ to identify Mendelian inconsistent and consistent errors which were then systematically evaluated as described in Vanderzande et al. [29]. Through this evaluation SNPs were identified that had irresolvably inaccurate calls. SNPs with inaccurate Infinium data were reclassified as excluded. Following this reclassification, SNPs with inaccurate Axiom data were deemed incompatible and classified as described in the section "Classification of compatible and incompatible SNP clustering".

Evaluation of within-platform repeatability

Repeatability of SNP calls was evaluated for each platform using individuals that were genotyped twice or more on both platforms from separate DNA extractions (biological replicates) and using different subsets of SNPs. These subsets were 1) the SNPs that passed the Infinium data curation steps in this study, 2) SNPs considered concordant and accurate between both arrays, 3) SNPs in the Axiom array previously deemed robust [34], 4) SNPs from subset 3 that were filtered for absence of more than one pedigree relationship in Muranty et al. [32], 5) SNPs in the Axiom array that were classified as "Poly High Resolution" or "No Minor Homozygous" by Axiom Analysis Suite software, and 6) SNPs from subset 5 with SNPs removed that had Mendelian errors in two or more parent-offspring relationships, discordant SNP calls in two or more duplicate pairs, or were heterozygous in doubled haploid accessions from Muranty et al. [32]. The few available technical replicates were not considered.

Classification of compatible and incompatible SNP clustering

During the above-described identification and evaluation of discordant SNP in duplicated individuals, SNPs deemed compatible, which we defined as having concordant and accurate SNP calls between both platforms, were classified into nine groups according to the type and prevalence of the applied manual cluster adjustments to achieve compatibility (listed in Table 3, with further descriptions and examples of each classification provided in Additional file 4). SNPs deemed incompatible were classified into nine groups and organized by how prevalent each observed issue was (listed in Table 3, with further descriptions and examples of each classification provided in Additional file 5). Incompatible SNPs were defined as those where Axiom SNP calls were either irresolvably discordant with curated Infinium data or SNPs with Axiom SNP calls that were deemed inaccurate. No systematic effort was made to identify the exact cause of each instance of discordancy.

Identification of reasons for SNP exclusion and incompatibility

We explored whether non-intended genomic sites that also hybridize to probes could be causing SNPs to be excluded during the initial Infinium data curation steps. All 18,019 SNPs on the Infinium array were grouped by the number of BLAST hits versus the GDDH13 WGS [16] that each had, using the 50 nt probe sequence from the Infinium array [3]. Next, for each group the percentage of excluded SNPs was determined. The number of BLAST hits was tallied at three different E-value thresholds to demonstrate trends: < 1.0E-12, < 1.0E-14, and < 1.0E-16. A maximum E-value threshold of < 1.0E-12 was chosen because it included intended match for over 99.3% of SNPs from the genetic map revision portion of this work.

Next, we explored whether the presence of additional non-intended genomic sites reduced cluster space available for SNP calling in Infinium cluster plots. Hereto, the effective cluster space was compared between SNPs grouped by inclusion or exclusion and by the number of BLAST hits at the different E-value thresholds. This was done because a reduction in cluster space could result in poor resolution between heterozygous and homozygous clusters, which could reduce SNP call accuracy. Available cluster space was calculated for each SNP by the difference between 5 and 95% quantiles of observed Theta values (or in other words, the space available on the x-axis on the Infinium cluster plots). This metric was used instead of the difference between the minimum and maximum Theta values to account for occasional outliers.

We also assessed how secondary polymorphisms at the probe site were influencing SNP inclusion rates of Infinium data and SNP call compatibility rates of included SNPs in Axiom data. To accomplish this, we first identified secondary polymorphisms among 53 individuals from the 480 K Axiom array SNP discovery panel [34] (Group 5 in Additional file 3) by aligning their sequence data to the GDDH13 WGS [16] using the Burrows-Wheeler Alignment tool [68] in conjunction with SAMtools [69] and identified all variants within the physical coordinates for the 50 nt long Infinium probes, as identified through the previously described BLAST hits. Secondary polymorphisms were considered if they were present in at least 10 % of the sequenced individuals. We used the resulting data to demonstrate how secondary polymorphisms influenced SNP calling quality in two ways. First, for each relevant (in) compatibility classification, we identified one or more representative SNPs. Next, their sequence-based polymorphism data was examined alongside cluster plot data to determine whether the observed clustering problems resulting in inaccurate and thus incompatible SNP calls were likely caused by secondary polymorphisms, as to provide explicit example cases. Second, information on the positions of secondary polymorphisms was correlated to the inclusion/exclusion rate in Infinium data and by compatibility of Axiom data with included Infinium SNPs. Additionally, to determine the effects of major differences between a probe and an intended target sequence, we also evaluated all cases of included SNPs that had probes with E-values greater than 1.0E-12. To minimise the risk of examining artefacts due to high E-values being caused by local sequence errors in the GDDH13 WGS [16], we also considered BLAST hits from the HFTH1 WGS [53] for this step, which became available during this study.

Abbreviations

BLAST: Basic local alignment search tool; cM: centimorgan; E-value: Expect values; the number of expected hits of similar quality (score) that could be found just by chance; GDDH13: Golden Delicious Double Haploid 13; HFTH1: Hanfu derived trihaploid line 1; iGL map: Integrated genetic linkage map; SNP: Single nucleotide polymorphism; WGS: Whole genome sequence

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07565-7.

Additional file 1. Updated iGL map for Illumina Infinium 20 K SNP array.

Additional file 2. Changes in position of SNP markers of 9685 SNPs included in this study that were also included in the iGL map [15].

Additional file 3. Germplasm included in study. Additional file 3.1 includes only genetically unique individuals and their associated “analysis names”, pedigrees, and MUNQ values and whose SNP data was used for data curation and/or evaluating concordancy and accuracy. Some individuals genotyped on the Infinium array and whose SNP data was only used for identification of genetic duplicates in Axiom data are not listed in this table and instead are included in Additional file 3.2 contains a listing by accession name and ID and includes all individuals, including duplicates. Additional file 3.3 contains a listing of the institutions who provided germplasm listed in Additional file 3.2.

Additional file 4. Cluster plot examples for classifications of compatible SNPs from Table 3.

Additional file 5. Cluster plot examples for classifications of incompatible SNPs from Table 3.

Additional file 6. Compatibility rate of Axiom data across the 10,295 SNPs that were included from the Illumina curation step grouped by the number of BLAST hits per SNP. Three different BLAST E-value thresholds were considered.

Additional file 7. Additional cluster plot examples.

Additional file 8. Number of secondary polymorphisms verses inclusion of SNPs in Infinium data and compatibility of Infinium data with Axiom data.

Additional file 9. Information for SNPs whose probes have BLAST hit E-values greater than 1.0E-12 that were still included in the Infinium data curation step and possible explanations as to why the apparent differences between their probes and target genomic positions were not sufficient to result in exclusion of the SNP and why these differences may or may not have negatively influenced the compatibility of Axiom clustering. Differences between the probe sequences and either genome sequence are listed in columns I and J as mismatches (M) and/or gaps (G).

Additional file 10. Families used for the improvement of the 20 K integrated genetic linkage map.

Acknowledgements

We would like to thank Diego Micheletti for providing aligned resequencing data and Simone Larger for performing genotyping. We also thank the following institutions for providing germplasm used in this study: Seed Savers Exchange in Decorah, Iowa, USA, Ökowerk in Emden, Germany, and NIAB EMR in Kent, UK.

Authors’ contributions

NPH and EW were the primary designers for the project and MT, LB, CD, and HM contributed to the design of the project. NPH performed SNP data curation, performed analyses, interpreted the data. NPH and EW wrote the manuscript. CE, HM, MT, and LB contributed writing of some portions of the manuscript. JT contributed to the design of and conducted the generation of BLAST data. LB contributed to the design of and conducted the generation of data relating to the identification of secondary polymorphisms. CD curated the germplasm organization and performed MUNQ attribution. MT, EW, CD, HM, and CD contributed to the acquisition of data. All authors provided reviews and revisions for relevant portions of the article. All authors have read and approved of the submission of the manuscript.

Funding

Funding for this research was in part provided by the Niedersächsisches Ministerium für Wissenschaft und Kultur through the EGON project: “Research for a sustainable agricultural production: Development of organically bred fruit cultivars in creative commons initiatives.” Part of the 20 K Infinium SNP data came from the FruitBreedomics project no 265582: “Integrated approach for increasing breeding efficiency in fruit tree crops” (www.fruitbreedomics.com), which was co-funded by the EU seventh Framework Programme.

Availability of data and materials

All accessions included in this study are listed in Additional file 3. The Axiom data used in this study was made available through Muranty et al. [32]. Curated Infinium data for duplicates and pedigree ancestors available at the Genome Database for *Rosaceae* (https://www.rosaceae.org/publication_datasets, accession number tfGDR1050). The highly curated integrated apple SNP array dataset will be shared through following publications on downstream analyses.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institut für Biologie und Umweltwissenschaften, Carl von Ossietzky Univ., Oldenburg, Germany. [2]Department of Horticultural Science, Univ. of Minnesota, St Paul, USA. [3]Fondazione Edmund Mach, San Michele all'Adige, TN, Italy. [4]Université d'Angers, Institut Agro, INRAE, IRHS, SFR 4207 QuaSaV, Beaucouzé, France. [5]Department of Plant Breeding, Wageningen University and Research, Wageningen, The Netherlands.

Received: 4 September 2020 Accepted: 30 March 2021
Published online: 07 April 2021

References

1. Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, et al. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant. 2017;10(8):1047–64. https://doi.org/10.1016/j.molp.2017.06.008.
2. Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, et al. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. Plos One. 2012;7(2):e31745. https://doi.org/10.1371/journal.pone.0031745.
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.
4. Antanaviciute L, Fernández-Fernández F, Jansen J, Banchi E, Evans KM, Viola R, Velasco R, Dunwell JM, Troggio M, Sargent DJ Development of a dense SNP-based linkage map of an apple rootstock progeny using the *Malus* Infinium whole genome genotyping array. BMC Genomics 2012;13(1):1–0. doi: https://doi.org/10.1186/1471-2164-13-203.
5. 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 Genet Genomes. 2015;11(5):94. https://doi.org/10.1007/s11295-015-0920-8.
6. Pikunova A, Madduri M, Sedov E, Noordijk Y, Peil A, Troggio M, et al. 'Schmidt's Antonovka' is identical to 'common Antonovka', an apple cultivar widely used in Russia in breeding for biotic and abiotic stresses. Tree Genet Genomes. 2014;10(2):261–71. https://doi.org/10.1007/s11295-013-0679-8.
7. Kumar S, Chagné D, Bink MC, Volz RK, Whitworth C, Carlisle C. Genomic selection for fruit quality traits in apple (*Malus* × *domestica* Borkh.). Plos One. 2012;7(5):e36674. https://doi.org/10.1371/journal.pone.0036674.
8. Clark MD, Schmitz CA, Rosyara UR, Luby JJ, Bradeen JM. A consensus 'Honeycrisp' apple (*Malus* × *domestica*) genetic linkage map from three full-sib progeny populations. Tree Genet Genomes. 2014;10(3):627–39. https://doi.org/10.1007/s11295-014-0709-1.
9. Di Guardo M, Micheletti D, Bianco L, Koehorst-van Putten HJ, Longhi S, Costa F, et al. ASSIsT: an automatic SNP scoring tool for in-and outbreeding species. Bioinformatics. 2015;31(23):3873–4. https://doi.org/10.1093/bioinformatics/btv446.
10. Di Guardo M, Bink MC, Guerra W, Letschka T, Lozano L, Busatto N, et al. Deciphering the genetic control of fruit texture in apple by multiple family-based analysis and genome-wide association. J Exp Bot. 2017;68(7):1451–66. https://doi.org/10.1093/jxb/erx017.
11. Falginella L, Cipriani G, Monte C, Gregori R, Testolin R, Velasco R, et al. A major QTL controlling apple skin russeting maps on the linkage group 12 of 'Renetta Grigia di Torriana'. BMC Plant Biol. 2015;15(1):150. https://doi.org/10.1186/s12870-015-0507-4.
12. Muranty H, Troggio M, Sadok IB, Al Rifaï M, Auwerkerken A, Banchi E, et al. Accuracy and responses of genomic selection on key traits in apple breeding. Hortic Res. 2015;2(1):1–2. https://doi.org/10.1038/hortres.2015.60.
13. Troggio M, Šurbanovski N, Bianco L, Moretto M, Giongo L, Banchi E, et al. Evaluation of SNP data from the *Malus* Infinium array identifies challenges for genetic analysis of complex genomes of polyploid origin. Plos One. 2013;8(6):e67407. https://doi.org/10.1371/journal.pone.0067407.
14. Allard A, Bink MC, Martinez S, Kelner JJ, Legave JM, Di Guardo M, et al. Detecting QTLs and putative candidate genes involved in budbreak and flowering time in an apple multiparental population. J Exp Bot. 2016;67(9):2875–88. https://doi.org/10.1093/jxb/erw130.
15. Di Pierro EA, Gianfranceschi L, Di Guardo M, Koehorst-van Putten HJ, Kruisselbrink JW, Longhi S, et al. A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. Hortic Res. 2016;3(1):1–3. https://doi.org/10.1038/hortres.2016.57.
16. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nat Genet. 2017;49(7):1099–106. https://doi.org/10.1038/ng.3886.
17. Durand JB, Allard A, Guitton B, Van de Weg E, Bink MC, Costes E. Predicting flowering behavior and exploring its genetic determinism in an apple multi-family population based on statistical indices and simplified phenotyping. Front Plant Sci. 2017;8:858. https://doi.org/10.3389/fpls.2017.00858.
18. Farneti B, Di Guardo M, Khomenko I, Cappellin L, Biasioli F, Velasco R, et al. Genome-wide association study unravels the genetic control of the apple volatilome and its interplay with fruit texture. J Exp Bot. 2017;68(7):1467–78. https://doi.org/10.1093/jxb/erx018.
19. Guan Y, Peace C, Rudell D, Verma S, Evans K. QTLs detected for individual sugars and soluble solids content in apple. Mol Breeding. 2015;35(6):135. https://doi.org/10.1007/s11032-015-0334-1.
20. 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 (*Malus* × *domestica*) pedigree-connected SNP data set. Hortic Res. 2017;4(1):17003. https://doi.org/10.1038/hortres.2017.3.
21. Howard NP, van de Weg E, Tillman J, Tong CB, Silverstein KA, Luby JJ. Two QTL characterized for soft scald and soggy breakdown in apple (*Malus* × *domestica*) through pedigree-based analysis of a large population of interconnected families. Tree Genet Genomes. 2018;14(1):2. https://doi.org/10.1007/s11295-017-1216-y.
22. Howard NP, Tillman J, Vanderzande S, Luby JJ. Genetics of zonal leaf chlorosis and genetic linkage to a major gene regulating skin anthocyanin production (*MdMYB1*) in the apple (*Malus* × *domestica*) cultivar Honeycrisp. Plos One. 2019;14(1):e0210611. https://doi.org/10.1371/journal.pone.0210611.
23. Lemmens E, Alós E, Rymenants M, De Storme N, Keulemans WJ. Dynamics of ascorbic acid content in apple (*Malus* x *domestica*) during fruit development and storage. Plant Physiol Bioch. 2020;151:47–59. https://doi.org/10.1016/j.plaphy.2020.03.006.
24. Migault V, Pallas B, Costes E. Combining genome-wide information with a functional structural plant model to simulate 1-year-old apple tree architecture. Front Plant Sci. 2017;7:2065. https://doi.org/10.3389/fpls.2016.02065.
25. Moriya S, Kunihisa M, Okada K, Iwanami H, Iwata H, Minamikawa M, et al. Identification of QTLs for flesh mealiness in apple (*Malus* × *domestica* Borkh.) Horticult J. 2017:MI-156. doi: https://doi.org/10.2503/hortj
26. Moriya S, Goto S, Kubo T, Kunihisa M, Tazawa J, Kudo T, et al. Identification of *Venturia inaequalis* races in Morioka, Japan and identification of a quantitative trait locus associated with resistance to apple scab in 'Akane' apples. Hortic Res (Japan). 2019;18(4):349–61. https://doi.org/10.2503/hrj.18.349.
27. van de Weg E, Di Guardo M, Jänsch M, Socquet-Juglard D, Costa F, Baumgartner I, et al. Epistatic fire blight resistance QTL alleles in the apple cultivar 'Enterprise' and selection X-6398 discovered and characterized

through pedigree-informed analysis. Mol Breeding. 2018;38(1):5. https://doi.org/10.1007/s11032-017-0755-0.

28. Vanderzande S, Micheletti D, Troggio M, Davey MW, Keulemans J. Genetic diversity, population structure, and linkage disequilibrium of elite and local apple accessions from Belgium using the IRSC array. Tree Genet Genomes. 2017;13(6):125. https://doi.org/10.1007/s11295-017-1206-0.
29. Vanderzande S, Howard NP, Cai L, Da Silva LC, Antanaviciute L, Bink MC, et al. High-quality, genome-wide SNP genotypic data for pedigreed germplasm of the diploid outbreeding species apple, peach, and sweet cherry through a common workflow. Plos One. 2019;14(6):e0210928. https://doi.org/10.1371/journal.pone.0210928.
30. Verma S, Evans K, Guan Y, Luby JJ, Rosyara UR, Howard NP, et al. Two large-effect QTLs, *Ma* and *Ma3*, determine genetic potential for acidity in apple fruit: breeding insights from a multi-family study. Tree Genet Genomes. 2019;15(2):18. https://doi.org/10.1007/s11295-019-1324-y.
31. Luo F, van de Weg E, Vanderzande S, Norelli JL, Flachowsky H, Hanke V, et al. Elucidating the genetic background of the early-flowering transgenic genetic stock T1190 with a high-density SNP array. Mol Breeding. 2019;39(2):1–5. https://doi.org/10.1007/s11032-019-0929-z.
32. Muranty H, Denancé C, Feugey L, Crépin JL, Barbier Y, Tartarini S, et al. Using whole-genome SNP data to reconstruct a large multi-generation pedigree in apple germplasm. BMC Plant Biol. 2020;20(1):1–8. https://doi.org/10.1186/s12870-019-2171-6.
33. Skytte af Sätra J, Troggio M, Odilbekov F, Sehic J, Mattisson H, Hjalmarsson I, et al. Genetic Status of the Swedish Central collection of heirloom apple cultivars. Sci Hortic-Amsterdam. 2020 Oct 15;272:109599. doi: https://doi.org/10.1016/j.scienta.2020.109599.
34. Bianco L, Cestaro A, Linsmith G, Muranty H, Denance C, Theron A, et al. Development and validation of the axiom® Apple480K SNP genotyping array. Plant J. 2016;86(1):62–74. https://doi.org/10.1111/tpj.13145.
35. Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, et al. Design of a bovine low-density SNP array optimized for imputation. Plos One. 2012;7(3):e34130. https://doi.org/10.1371/journal.pone.0034130.
36. Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, et al. Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci. 2017;8:1792. https://doi.org/10.3389/fpls.2017.01792.
37. Hong H, Xu L, Liu J, Jones WD, Su Z, Ning B, et al. Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. Plos One. 2012;7(9):e44483. https://doi.org/10.1371/journal.pone.0044483.
38. Berry DP, O'brien A, Wall E, McDermott K, Randles S, Flynn P, et al. Inter-and intra-reproducibility of genotypes from sheep technical replicates on Illumina and Affymetrix platforms. Genet Sel Evol. 2016;48(1):86. https://doi.org/10.1186/s12711-016-0267-0.
39. Wijesena HR, Rohrer GA, Nonneman DJ, Keel BN, Petersen JL, Kachman SD, et al. Evaluation of genotype quality parameters for SowPro90, a new genotyping array for swine. J Anim Sci. 2019;97(8):3262–73. https://doi.org/10.1093/jas/skz185.
40. Sobel E, Papp JC, Lange K. Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet. 2002;70(2):496–508. https://doi.org/10.1086/338920.
41. Bussey DJ, Whealy K. The illustrated history of apples in the United States and Canada. Mount Horeb: JAK KAW Press; 2016.
42. Evans KM, Patocchi A, Rezzonico F, Mathis F, Durel CE, Fernandez-Fernandez F, et al. Genotyping of pedigreed apple breeding material with a genome-covering set of SSRs: trueness-to-type of cultivars and their parentages. Mol Breeding. 2011;28(4):535–47. https://doi.org/10.1007/s11032-010-9502-5.
43. Salvi S, Micheletti D, Magnago P, Fontanari M, Viola R, Pindo M, et al. One-step reconstruction of multi-generation pedigree networks in apple (*Malus* × *domestica* Borkh.) and the parentage of Golden delicious. Mol Breeding. 2014;34(2):511–24. https://doi.org/10.1007/s11032-014-0054-y.
44. Lassois L, Denancé C, Ravon E, Guyader A, Guisnel R, Hibrand-Saint-Oyant L, et al. Genetic diversity, population structure, parentage analysis, and construction of core collections in the French apple germplasm based on SSR markers. Plant Mol Biol Rep. 2016;34(4):827–44. https://doi.org/10.1007/s11105-015-0966-7.
45. Larsen B, Toldam-Andersen TB, Pedersen C, Ørgaard M. Unravelling genetic diversity and cultivar parentage in the Danish apple gene bank collection. Tree Genet Genomes. 2017;13(1):14. https://doi.org/10.1007/s11295-016-1087-7.
46. Carlson CS, Smith JD, Stanaway IB, Rieder MJ, Nickerson DA. Direct detection of null alleles in SNP genotyping data. Hum Mol Genet. 2006;15(12):1931–7. https://doi.org/10.1093/hmg/ddl115.
47. Franke L, de Kovel CG, Aulchenko YS, Trynka G, Zhernakova A, Hunt KA, et al. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am J Hum Genet. 2008;82(6):1316–33. https://doi.org/10.1016/j.ajhg.2008.05.008.
48. Hyten, David L., et al. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 2008, 116. Jg., Nr. 7, S. 945–952. doi https://doi.org/10.1007/s00122-008-0726-2.
49. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. Nat Methods. 2006;3(1):31–3. https://doi.org/10.1038/nmeth842.
50. Bellon et al. 2011. SNP genotyping of markers with nearby secondary polymorphisms using Affymetrix' Axiom® Genotyping solution. http://tools.thermofisher.com/content/sfs/brochures/snp_genotyping_sec_polymorphisms_app_brief.pdf
51. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics. 2011;98(2):79–89. https://doi.org/10.1016/j.ygeno.2011.04.005.
52. Koller B, Gianfranceschi L, Seglias N, McDermott J, Gessler C. DNA markers linked to *Malus floribunda* 821 scab resistance. Plant Mol Biol. 1994;26(2):597–602. https://doi.org/10.1007/BF00013746.
53. Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat Commun. 2019;10(1):1–3. https://doi.org/10.1038/s41467-019-09518-x.
54. Bassil NV, Davis TM, Zhang H, Ficklin S, Mittmann M, Webster T, et al. Development and preliminary evaluation of a 90 K axiom® SNP array for the Allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. BMC Genomics. 2015;16(1):155. https://doi.org/10.1186/s12864-015-1310-1.
55. Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, et al. A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. Plos One. 2011;6(12):e28334. https://doi.org/10.1371/journal.pone.0028334.
56. Li X, Han Y, Wei Y, Acharya A, Farmer AD, Ho J, et al. Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. Plos One. 2014;9(1):e84329. https://doi.org/10.1371/journal.pone.0084329.
57. Faivre-Rampant P, Zaina G, Jorge V, Giacomello S, Segura V, Scalabrin S, et al. New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12k Infinium array. Mol Ecol Resour. 2016;16(4):1023–36. https://doi.org/10.1111/1755-0998.12513.
58. Lee YG, Jeong N, Kim JH, Lee K, Kim KH, Pirani A, et al. Development, validation and genetic analysis of a large soybean SNP genotyping array. Plant J. 2015;81(4):625–36. https://doi.org/10.1111/tpj.12755.
59. Roorkiwal M, Jain A, Kale SM, Doddamani D, Chitikineni A, Thudi M, et al. Development and evaluation of high-density axiom® *Cicer*SNP Array for high-resolution genetic mapping and breeding applications in chickpea. Plant Biotechnol J. 2018;16(4):890–901. https://doi.org/10.1111/pbi.12836.
60. Marrano A, Martínez-García PJ, Bianco L, Sideli GM, Di Pierro EA, Leslie CA, et al. A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density axiom™ *J. regia* 700K SNP genotyping array. Plant Biotechnol J. 2019;17(6):1027–36. https://doi.org/10.1111/pbi.13034.
61. Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, et al. Decoding randomly ordered DNA arrays. Genome Res. 2004;14(5):870–7. https://doi.org/10.1101/gr.2255804.
62. Nicolazzi EL, Iamartino D, Williams JL. AffyPipe: an open-source pipeline for Affymetrix axiom genotyping workflow. Bioinformatics. 2014;30(21):3118–9. https://doi.org/10.1093/bioinformatics/btu486.
63. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (*Malus* x *domestica* Borkh.). Nat Genet. 2010;42(10):833–9. https://doi.org/10.1038/ng.654.
64. Peace CP, Bianco L, Troggio M, Van de Weg E, Howard NP, Cornille A, et al. Apple whole genome sequences: recent advances and new prospects. Hortic Res. 2019;6(1):59. https://doi.org/10.1038/s41438-019-0141-7.
65. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1–2):203–14. https://doi.org/10.1089/10665270050081478.
66. Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 2008;24(3):319–24. https://doi.org/10.1093/bioinformatics/btm585.

67. Bink MC, Jansen J, Madduri M, Voorrips RE, Durel CE, Kouassi AB, et al. Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. Theor Appl Genet. 2014;127(5): 1073–90. https://doi.org/10.1007/s00122-014-2281-3.
68. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
69. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93. https://doi.org/10.1093/bioinformatics/btr509.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.