6-27-2019

Estimation of genetic diversity and relatedness in a mango germplasm collection using SNP markers and a simplified visual analysis method

David N. Kuhn

Natalie Dillon

Ian Bally

Amy Groh

Jordon Rahaman

See next page for additional authors

Authors

David N. Kuhn, Natalie Dillon, Ian Bally, Amy Groh, Jordon Rahaman, Emily Warschefsky, Barbie Freeman, David Innes, and Alan H. Chambers

Scientia Horticulturae 252 (2019) 156–168

Contents lists available at ScienceDirect

Scientia Horticulturae

journal homepage: www.elsevier.com/locate/scihorti

Estimation of genetic diversity and relatedness in a mango germplasm collection using SNP markers and a simplified visual analysis method

David N. Kuhn[a,*], Natalie Dillon[b], Ian Bally[b], Amy Groh[c], Jordon Rahaman[c], Emily Warschefsky[c], Barbie Freeman[a], David Innes[b], Alan H. Chambers[d]

[a] Subtropical Horticulture Research Station, USDA-ARS, Miami, FL, USA
[b] Centre for Tropical Agriculture, Horticulture and Forestry Science, Department of Agriculture and Fisheries, QLD, Australia
[c] International Center for Tropical Botany, Florida International University, Miami, FL, USA
[d] Tropical Research and Education Center, Horticultural Sciences, University of Florida, Homestead, FL, USA

ARTICLE INFO

Keywords:
Affinity propagation clustering
Mangifera species
Polyembryony
Self-Incompatibility
SNP genotyping

ABSTRACT

Mango is a globally important tropical fruit but lacks genomic tools to support cultivar identification and to enable breeding efforts. Assessing the genetic diversity and relatedness of mango germplasm is essential for identifying genetically distant parents with favorable agronomic traits to produce hybrid populations enabling selection of improved cultivars. We thus genotyped 1915 mango accessions from the United States, Senegal, Thailand, and Australia with 272 single nucleotide polymorphism (SNP) markers identifying over 520,000 genotypes. These accessions represent the available diversity from both public and private germplasm collections in these countries, as well as accessions from smaller international collections. The study included M*angifera indica*, other *Mangifera* species, and accessions from half sibling populations. Genotype data were analyzed using an affinity propagation method to define 258 groups. Using a simple visual method, no more than 30 SNPs are needed to distinguish a single cultivar of interest from all other cultivars in the dataset enabling the accurate identification of important commercial cultivars. As these SNP markers provided accurate genotype data for accessions from different genera as well as half siblings, the majority of the genetic diversity of the mango germplasm and related species that were genotyped has been captured. The dataset contains a large collection of open-pollinated half siblings from known maternal parents. A simple visual method can also be used to identify self-pollinated individuals among the half siblings of known maternal parents and, in some cases, to infer likely candidates for the paternal parent. Identification of self-pollinated individuals is particularly important in terms of selection of improved cultivars, as due to high levels of heterozygosity, self-pollinated progeny are likely to uncover deleterious recessive alleles. Genotyping of progeny at the seedling stage and removal of self-pollinated progeny can increase the efficiency and decrease the costs of selection of improved cultivars from open-pollinated populations.

1. Introduction

Mango (*Mangifera indica*) is a tropical tree species valued for the nutritional and organoleptic qualities of its fruit. Global mango production reached 46.51 million metric tons in 2016 (FAO, 2016b), and ranked sixth in global fruit production behind watermelons, bananas, apples, grapes, and oranges (FAO, 2016a). Mango has been cultivated for an estimated 4,000 years or more (De Candolle, 1885) with seed propagation followed by grafting of superior types resulting in approximately 1,000 named cultivars (Litz, 2009). Mango is imbedded in the cultural backgrounds of many tropical nations including major producers such as India, Mexico, China, Thailand, Indonesia, and Pakistan (Mahato et al., 2016). The center of origin for mango is believed to be somewhere in Northeastern India, Bangladesh, and Nepal, perhaps stretching into Northern Myanmar and Thailand, though current distributions of wild *M. indica* are not well known (Kostermans and Bompard, 1993; Litz, 2009; Singh et al., 2016). In contrast, Malesia (particularly Sumatra, Borneo, and the Malay Peninsula) is the center of diversity for the genus *Mangifera* (Kostermans and Bompard, 1993). Mangos from South East Asia comprise a tropical, polyembryonic group that contain seeds with both zygotic and nucellar embryos. A second group includes subtropical, Indian types with monoembryonic seeds (Mukherjee and Litz, 2009). Since its original domestication, mango has been dispersed and cultivated throughout the tropical and subtropical

* Corresponding author.
E-mail address: David.Kuhn@ars.usda.gov (D.N. Kuhn).

https://doi.org/10.1016/j.scienta.2019.03.037
Received 28 September 2018; Received in revised form 12 March 2019; Accepted 20 March 2019
Available online 01 April 2019

world and has been associated with the migration of people and trade with and in between regions (Bompard, 2009; Duval et al., 2006; Kostermans and Bompard, 1993). Identification of mango species based on morphological characters has provided estimates of the number of species ranging from 45 (Hou, 1974) to 69 (Kostermans and Bompard, 1993). Highlighting the confusion surrounding the taxonomy of *Mangifera*, the taxonomic database the plant list (theplantlist.org) includes 133 specific epithets for *Mangifera*, only 10 of which are currently accepted, and 116 of which are unresolved. Additional data from genome-wide molecular markers should help resolve some outstanding mango diversity questions and aid future plant improvement work.

The majority of mangoes are consumed domestically (Galán Saúco, 2015). Only a few cultivars are typically preferred in these domestic markets, and even fewer meet the stringent criteria for export markets. Most named varieties (cultivars) originate from chance seedlings or selections that are propagated by cuttings, though formal mango breeding programs exist in Australia, Brazil, China, India, Indonesia, Israel, Mexico, Pakistan, and Thailand (Bally and Dillon, 2018; Kuhn et al., 2017). The primary export mangos are restricted to a few specific cultivars originating from a selection process that began in Florida in the first decades of the 20th century including 'Tommy Atkins', 'Kent', and 'Keitt' (Galán Saúco, 2015; Schnell et al., 2006). Export mangos are mostly large, externally attractive fruits with excellent shipping characteristics. The United States is the largest mango importer consuming around one third (˜436,000 metric tons) of global mango imports in 2013 (Galán Saúco, 2010).

Public and private mango collections have attempted to capture the diversity of cultivated accessions along with some wild *Mangifera* species but are challenged by the biology of this species. Mango seeds are not tolerant to desiccation and cannot be stored for long periods of time. Additionally, the long mango juvenile period would not be amenable to rapid data collection on fruit quality traits if starting from seeds or immature trees. Therefore, mango germplasm collections include long-term maintenance of living trees that are propagated periodically by grafting. This process can lead to unintentional mislabeling and propagation of rootstocks rather than scion material. Large mango germplasm collections have been established at the USDA Subtropical Horticulture Research Station (SHRS) in Miami, FL (the largest collection), and the Department of Agriculture and Fisheries, Mareeba, Queensland, Australia. These collections along with smaller collections are listed in Table 1. The collection at SHRS includes several research populations made from a hybridization polycross experiment where multiple clones of six different mango cultivars were planted in proximity and the seedlings from these open-pollinated trees planted and evaluated. The collection at Mareeba also includes selections and several hybrid populations from a long-term Australian mango breeding program. Ideally, any molecular marker, including SNP markers, would uniquely identify half/full sibling progeny accessions as well as accessions from different genera to allow a single set of markers to be used for all germplasm and research collections.

The mango genome is diploid (n = 20) with an estimated haploid size of ˜439 MB (Arumuganathan and Earle, 1991). There have been many studies of mango germplasm accessions with a wide variety of molecular markers focusing primarily on *M. indica* including previous work using SNPs (Kuhn et al., 2014; Mahato et al., 2016; Sherman et al., 2015). Historical development and use of other types of molecular markers (RAPD, ISSR, SSR, SCoT, AFLP, etc.) are platform dependent, and genotype data cannot be easily transferred between labs. The density of these markers can also be insufficient to support breeding and genetics work in mango. Clustering analyses using these alternative molecular markers has identified differences in mango by source region, but the number of markers and accessions tested only allows for a rough estimate of genetic diversity. Conversely, SNP molecular markers are unambiguous, platform-independent, transferable between labs, and of sufficient quantity to enable genotype databases supporting accession identification and breeding efforts.

SNP discovery in mango and other non-model systems usually relies on next generation sequencing and *de novo* transcriptome analysis of a few accessions to develop markers for genotyping collections. A 'Tommy Atkins' mango transcriptome from multiple plant tissues identified 30,000 transcripts and mapped RNA sequences from 24 genetically diverse mango accessions to identify ˜400,000 SNPs that led to 640 high-quality, well isolated SNPs in protein-coding genes (Kuhn et al., 2014). Another study developed 239 SNPs from mango cultivars 'Keitt' and 'Tommy Atkins' and was used to genotype 74 Israeli accessions (Sherman et al., 2015). Recently, 1,054 single nucleotide polymorphism (SNP) markers were used from the three sources described to create a genetic map for mango (Kuhn et al., 2017). The markers from Kuhn et al were used to genotype seven mapping populations of mango totaling 807 individuals. A consensus genetic map was produced that defined 20 linkage groups for mango as expected. In the present study, 384 SNP markers were used to analyze diversity. The majority of these markers are evenly distributed across the 20 linkage groups to capture diversity across the genome.

The primary objective of this study was to genotype the largest number of mango accessions possible from germplasm collections from the United States, Senegal, Thailand, and Australia with sufficient molecular markers to generate accurate genotypes and determine genetic relationships. The resulting database was then used to 1) estimate genetic diversity in the germplasm dataset, 2) estimate mislabeling/misidentification in the germplasm collections, 3) confirm self-pollination, self-compatibility, and pedigree when possible, 4) test marker association with the polyembryony trait, and 5) genotype accessions from other genera and species related to *M. indica* to assess congruency with classical taxonomy. In using the large SNP database to accomplish these goals, we have developed an intuitive, visual analysis method that allows advanced querying of the database using simple functions of a spreadsheet program. Sorting and counting allows a user to determine the likelihood of self-pollination, infer the paternal parent of open pollinated seedlings of known maternal parentage, and choose subsets of SNP markers to distinguish a cultivar of interest from all other germplasm accessions.

2. Materials and methods

2.1. Germplasm accessions

Mango accessions from the United States, Senegal, Thailand, and Australia were included in this study. The sources of all 2232 individuals that were genotyped are summarized in Table 1. Designators were assigned for genotyping and original names appended to designators. Individuals named in the dataset represent a combination of the actual accession name and a unique identifier assigned to distinguish it in the analysis. Combined names may have unusual joining characters such as two colons (::) that have been added to simplify conversion of names from the analysis programs to human friendly names. In some cases, revision of the names is designated by the addition of the revision in parentheses. For Australian germplasm accessions, the addition of (Kensington) to a name denotes that it is a farmer's selection of 'Kensington Pride'.

2.2. Isolation of DNA

DNA was isolated as described in Kuhn et al (Kuhn et al., 2017). Briefly, 3 mm leaf disks were punched from leaves (˜50 mg per sample), disrupted by shaking with 1/52″ stainless steel beads, and extracted using a Mag-Bind Plant DNA DS 96 Kit from Omega BioTek (M1130-01) with automated steps run on a Hamilton Starlet liquid handling robot. DNA was quantified by fluorometry and all DNA samples adjusted to a concentration of 10–20 ng/uL using a Hamilton liquid handling robot.

Table 1
Mango germplasm sources, locations, designators for names, total number of individuals genotyped, and genotyped individuals with less than 5% missing data used in this study.

Population	Station	Location	Designator	Total # of Individuals Genotyped	Final # of Individuals with less than 5% missing data
Germplasm	SHRS ARS	Miami, FL	Mi	308	259
Polycross seedlings	SHRS ARS	Miami, FL	PC named cultivar is maternal parent	386	364
Clones	SHRS ARS	Miami, FL	WA, WB, WF, N2, Ha, TA Specific individual is a clone of named cultivar	59	57
Germplasm	Zill private collection	Boynton Beach, FL	ZLp	48	43
Open pollinated seedlings	Zill private collection	Boynton Beach, FL	ZL named cultivar is maternal parent	56	54
Germplasm	Fruit and Spice Park	Homestead, FL	FSP	172	166
Germplasm	SRS and WRS	Mareeba, Australia	AuMG	634	547
Mangifera laurina hybrids	SRS and WRS	Mareeba, Australia	AuMH named cultivar is maternal parent, M. laurina is paternal parent	76	73
parents for populations	SRS and WRS	Mareeba, Australia	Au, AuMLP, Specific individual is a clone of named cultivar	16	15
Mango hybrid mapping populations between two known parents	SRS and WRS	Mareeba, Australia	AuMP first named cultivar is maternal parent, second is paternal parent	90	75
Germplasm		Senegal	M	64	51
Germplasm		Thailand	M	34	34
Haden X Tommy Atkins	Embrapa	Petrolina, Brazil	HT Haden is maternal parent and Tommy Atkins is paternal parent	27	12
Keitt X Tommy Atkins	Embrapa	Petrolina, Brazil	KT Keitt is maternal parent and Tommy Atkins is paternal parent	81	53
Landrace X Tommy Atkins	Embrapa	Petrolina, Brazil	TD Landrace is maternal parent and Tommy Atkins is paternal parent	96	60
Germplasm and other species	Florida International University (Emily Warschefsky)	Miami, FL	ew	85	52
			Total	2232	1915

2.3. SNP genotyping of germplasm accessions

Each mango accession was genotyped with 384 SNP markers designed as assays to be run on the Fluidigm EP-1 platform (Fluidigm) from a larger set of SNP markers as previously described (Kuhn et al., 2017). The sequences of the 399 and subset of 272 SNP assays used in this study are in Supplemental Table 1 with the associated linkage group, map position, and annotation where available. Assays were performed on a 96 × 96 Fluidigm chip with 91 sample DNAs, five controls and 96 SNP assays. Genotype information in a flat file format was grouped and reformatted using Perl scripts for analysis. Perl scripts are available upon request.

2.4. Analysis of genotype data

Genotype data was encoded in four categories: homozygous allele 1 (1), homozygous allele 2 (2), heterozygous (3) or missing data (0) rather than nucleotide data to allow an unbiased analysis as to the relations of the individuals.

2.4.1. Calculating pairwise distances

The following custom distance function was used to generate pairwise distances:

$$\text{distance}(x, y) = \frac{\sum_{i=1}^{272} comp(x_i, y_i)}{272 - (md(x) + md(y)}$$

where comp (x_i, y_i) is the SNP state comparison scoring function for a given marker i for samples x and y and i) comp (1, 2) or comp (2,1) is equal to 1, ii) comp (b, 3) or comp (3, b), where b is 1 or 2, is equal to 0.5, and iii) comp (a, a), comp (a, 0), comp (0, a), where a is any of the four possible states, is equal to 0. The missing data function, md(x), counts the number of missing data points for a given sample x. The value 272 is the number of markers compared between samples after data was curated to remove markers with greater than 5% missing data and individuals with greater than 5% missing data in a recursive fashion, resulting in a dataset of 272 markers for 1915 individuals.

2.4.2. Affinity propagation analysis

Affinity propagation has been used previously to analyze SNP data (Bryant et al., 2013; Pers et al., 2015), and employs an algorithm that resamples the data space until the group identities and exemplars converge and do not change under further resampling. Affinity propagation, as implemented by the Python library Scikit-learn (Pedregosa et al., 2011), was run with 0.99 damping, 1.0E6 maximum iterations, 5.0E5 convergence iterations, default median preference, and affinity set to "precomputed". The input must be a square pairwise sample similarity matrix which, in this case, was produced by inverting the distance matrix by subtracting each distance from 1. By selecting a convergence iteration value half that of the max iterations, a 50 percent consensus rule was enforced; if the analysis stopped before maximum iterations were reached, then the same groups were resolved at least 50 percent of the time.

2.4.3. Silhouette scores for membership in affinity groups

Per sample and average silhouette scores (Rousseeuw, 1987) were calculated using Scikit-learn (Pedregosa et al., 2011) and the per sample scores are reported in Supplemental Table 2. The inputs are the clusters found by affinity propagation and the pairwise sample distance matrix. Silhouette scores are normalized between 1 and -1. Per sample silhouette scores indicate how well a given sample is matched to its cluster, with 1 being a perfect match and -1 a total mismatch. The average silhouette score reflects how well the dataset has been clustered, with scores closer to 1 indicative of increasingly resolved clustering structure (Rousseeuw, 1987).

2.5. Visualization of genotypic data

Genotype data files were imported into Microsoft Excel for visualization of this large dataset enabling validation of queries without requiring specialized analytical programming scripts in Perl, Python, or R. Grouping information from the affinity propagation analysis, exemplars and silhouette scores were added as columns and the dataset was sorted by affinity group. Cells were colored using conditional formatting (0, grey, missing data; 1, blue, homozygous allele 1; 2, orange, homozygous allele 2; 3, green, heterozygous) which allowed sorting without regard to the actual nucleotide data. Numbers of 0, 1, 2, 3 were calculated for each row and column using the countif function in Excel. Some of the simple analyses performed on the dataset were done using the sorting function in Excel. Columns (markers) were sorted by an individual's genotype across all markers (0–3), number of 1 s, smallest to largest, by heterozygosity and by marker position in the genome, among other methods. Rows (individuals) were sorted by groups, names, homozygosity and heterozygosity, among other methods. Colored data file with affinity propagation groups, silhouette scores and related data are in Supplemental Table 2.

In an affinity group, if two or more accessions share the same or similar silhouette score, they are likely to be genetically identical within the limits of machine genotype error. They will also have similar or identical numbers of homozygous allele 1, homozygous allele 2 and heterozygous states. Using supplemental Table 2, identity can be visually determined by hiding rows in the cluster with dissimilar silhouette scores and scanning across the columns looking for different colored cells among accessions believed to be identical by naming convention.

In addition, to identify a subset of markers that can be used to distinguish a single accession from all other non-identical accessions, supplemental Table 2 can be sorted by the row with the accession of interest. The ranges of 1 s and 2 s can be used in countif formulas to determine the number of mismatches with all other accessions in the database, giving a score of 1 for a 1:2 or 2:1 mismatch.

2.6. Association of polyembryony with SNP markers for germplasm collection

A chi-square test for independence (degrees of freedom = 1106) was used to establish whether a relationship existed between the embryony trait and SNPs at significance threshold p = .001; 369 markers were used across 199 individuals and the derived contingency table of expected values contained ≥ 80% of cells with ≥ 5 counts and no cells with 0 counts.

Association mapping for the polyembryony trait was done in R using post hoc Fisher's Exact Tests with a calculated Bonferroni multiple test corrected significance threshold of p = 2.71e-6. The fisher. Test function was used to test 369 of the 384 markers and 199 individuals with a two-sided alternative hypothesis and a workspace of 2e8The markers with the lowest p-values were compared to the linkage map and QTL mapping from a previous study (Kuhn et al., 2017).

3. Results

3.1. Estimation of genetic diversity

3.1.1. Genotype statistics

Biallelic SNP markers (384) were used to genotype 2,232 germplasm accessions. Data were curated to remove markers with greater than 5% missing data and individuals with greater than 5% missing data in a recursive fashion, resulting in a dataset of genotypes of 272 markers for 1915 individuals (Supplementary Table 2). The individuals included accessions from different genera (*Bouea* and *Anacardium*) of the Anacardiaceae, different *Mangifera* species, germplasm accessions, hybrids of known cultivars and half-sibs from open pollination of

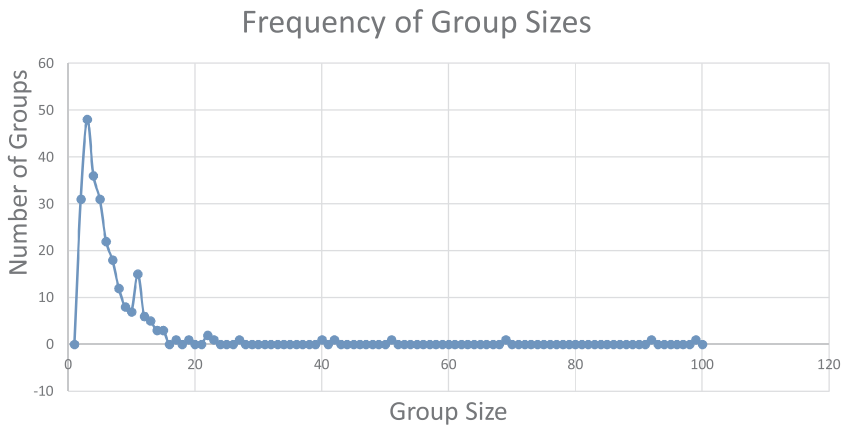


Fig. 1. The distribution of the number of affinity groups having a specific number of individuals. The x-axis represents the number of individuals in an affinity group and the y-axis is the number of affinity groups that have that number of individuals.

selected maternal parents. Missing data per individual varied from 0 to 14 markers of the 272 and missing data for markers varied from 2 to 102 for the 1915 individuals.

3.1.2. Affinity propagation analysis

The affinity propagation analysis was done with 1915 individuals and 272 markers as described above and generated 258 affinity groups. The affinity propagation analysis converged after ˜520,000 iterations, well before the maximum number of 1 M iterations, and accession membership in identical affinity groups was observed in greater than 50% of the 520,000 iterations. For each affinity group an exemplar that represents the "center" of the cluster in n-dimensional dataspace is determined by measuring distances of all other cluster members to the exemplar. As the analysis continues, each accession is considered a possible exemplar, distances were calculated and affinity groups determined. Affinity groups varied in size from 2 to 99 individuals. A distribution of the frequency of different sized affinity groups is shown in Fig. 1. The majority of the affinity groups (246 of 258) contained between 2 and 15 individuals.

3.1.3. Silhouette scores

After affinity propagation analysis of the 1915 individuals, a silhouette analysis determined the quality of the membership of a particular accession in an affinity group (Supplemental Table 2). Silhouette scores near or below zero indicate weak evidence for membership in the affinity group, or that the accessions in the group are equally spaced and diffuse in the dataspace such that there is no clear "center" to the cluster. A low silhouette score does not indicate that there is no affinity group, nor that accessions are in a "can't be clustered" bin, nor that individuals have simply been grouped because they are distant from all other clusters, a version of long branch attraction in phylogenetic analyses. With the parameters set for the analysis, convergence of the analysis guarantees that accessions occur more than 50% of the time in the same cluster and clusters are determined by distance to the exemplar and not to other clusters.

3.1.4. Calculation of machine genotyping error

Some cultivars had been genotyped multiple times and allowed the calculation of machine genotyping error. For 26 clones of cultivar 'Haden' in affinity group 235 (Supplementary Table 2) that were labeled as 'Haden' from the Australian, Fruit and Spice Park, SHRS, and Zill collections, there were 35 differences at 16 markers. Seven of the markers showed only one difference among the 26 sequences, with the greatest number of differences for a single marker being seven (Mi_0481). For the 26 clones, there were no "impossible" differences at any marker. For example, an individual that was homozygous for allele 1 and an individual that was homozygous for allele 2 for the same SNP marker (1:2, 2:1) between compared accessions that are putatively identical. Such "impossible" differences are unlikely to occur in the Fluidigm platform used for assay based on how the genotypes are called but have been observed when multiple identical clones are compared. Such differences have not been correlated with specific markers and are assumed to be due to machine error as they occur at less than 0.1% frequncy. All differences observed were between individuals homozygous for the markers and individuals heterozygous at the same marker. Nine of the 26 'Haden' sequences showed no differences over the 272 markers and defined the consensus correct genotype. The greatest number of differences from the consensus genotype was a single individual with seven differences (2.6%), the other clones having one to three differences. Thus, the most conservative estimation of machine genotyping error (7 differences out of 272, 2.6%) was used to determine whether two individuals could be considered identical.

3.1.5. Heterozygosity of individuals and markers

Heterozygosity of individuals ranged from 1% (3/272) for *M. quadrifida* (ew51::SBG_6) to 94% (257/272) for cultivar 'Julie' (ZLp069::0). Low heterozygosity was seen for most individuals labeled as species other than *M. indica*. Heterozygosity of markers ranged from 21.4% (410/1915, Mi_0481, map position LG8 145 cM) to 52.1% (998/1915, mango_rep_c3432, map position LG14 60.4 cM). On average, 17 out of 1915 genotypes were missing across all markers (0.9%), while the average missing data per individual was 2.4 out of 272 markers (0.9%). Average allele frequency over all markers for allele 1 was 51.2% and allele 2 was 48.7%. Allele frequency ranged from allele1:allele2 25:75 for Mi_0358 to 84:16 for SSKP077C2_A650 G.

3.1.6. Estimating mislabeling

To estimate mislabeling, small groups of identical or nearly identical cultivars were used with the assumption that identical genotypes should have identical names. For example, in group 106, there are 99 accessions that fall into three ranges of silhouette scores: -0.3 – 0.17 has 21 accessions, 0.24 – 0.55 has 10 and 0.61 – 0.69 has 68. Accessions in the highest silhouette score range can be considered identical as they have seven or fewer differences to the consensus genotype for 'Kensington Pride'. Many of these are farmer's selections of 'Kensington Pride' (Bally et al., 1996) as indicated by having (Kensington) after the name. Interestingly, not all of the farmers' selections are from clonal 'Kensington Pride', but rather seedlings of 'Kensington Pride', such as AuMG411_Weaver, AuMG352 Kensington Mackay AUS, AuMG250 Bowen Early AUS, AuMG288 Mountain View Mottle HI and Amarto Seedless AUS that have silhouette scores between 0.09 and 0.36. Some accessions are known to be mislabeled such as AuMG249 'Mulgoa Ramasamy' and AuMG465 *M. pajang* that are actually 'Kensington Pride'. Previous names have been kept by the curator to provide historical

consistency. In all, 10 of the 68 (15%) genotypically identical accessions from the Australian germplasm collection (AuMG prefix) may be mislabeled or misidentified. The sample AuMG371 'Carabao Super Manila Harbon' is not a mislabeling at the germplasm level but an error during genotyping, and two samples, AuMG371 and AuMG106 'Kensington Spooner' (AUS) were mislabeled during genotyping. AuMG106 is in group 3 with the other 'Carabao' and AuMG371 is in group 66 with the other 'Kensington Pride'-like accessions. This mislabeling/misidentification estimate is based solely on this group and may not be representative of mislabeling throughout the Australian collection. Two accessions with silhouette scores of 0.68 (AuMG382 Rahder Original (R2E2) and AuMG322 Brown's seedling) are labeled as seedlings, but appear to be 'Kensington Pride' clones. R2E2 is a mid-season variety with large, highly coloured fruit. It was selected in 1982 by Ian Bally, Ross Wright, and Peter Beal as a seedling progeny of the Florida variety 'Kent', and takes its name from the row and position in the field of the original tree at the department´s Bowen Research Facility. Further evidence of mislabeling in germplasm collections comes from identifying groups where two or more differently named accessions are genotypically identical. Only groups that had at least two accessions with the same highest silhouette scores were considered, with the assumption being that these accessions were genetically identical. This occurred in 58 of the 258 groups (22%). Mislabeling was attributed to each of the four largest germplasm collections by the criterion described above and the number of potential mislabeling events were divided by the total number of accessions in the collection. For the Miami, Fruit and Spice Park, Australian, and Senegal collections, the mislabeling was ˜8%. This is likely an underestimate due to the difficulty of identifying mislabeling in groups where multiple genetically identical accessions were all named differently.

3.2. Detection of self-pollination and estimation of self-compatibility

To visually determine the likelihood of self-pollination in hybrids with a known maternal parent, the genotype data in Supplemental Table 2 can be sorted by row using the maternal parent to sort on, followed by counting the number of times that the hybrid and maternal parent differ at homozygous loci. If the maternal parent is correctly identified, differences in the hybrid will only be heterozygotes (3). If there are a number of impossible differences (1:2 or 2:1, maternal parent:hybrid), then the maternal parent has likely been misidentified. An abbreviated example is given in Fig. 2.

There are 41 groups that contain only hybrids, with the total number of hybrids being 418 and the range of hybrids per group as 3–50. Groups that contain hybrids and one of the maternal parents provide an opportunity to determine self-pollination and estimate self-compatibility of the maternal parent. For example, in group 149, 12 of the hybrids that show 'Keitt' as a maternal parent are homozygous for the same alleles at all loci where 'Keitt' is homozygous for that allele. Where 'Keitt' is heterozygous at a locus, different genotypes are seen at these loci in the hybrid progeny. This is consistent with the hybrids being self-pollinated progeny of 'Keitt'. In group 256 which contains 'Tommy Atkins', there are 30 hybrids with 'Tommy Atkins' as the maternal parent. Of these 30, 23 are self-pollinated. The rest of the hybrid analysis is summarized in Table 3. 'Haden' is the maternal parent of 'Tommy Atkins' and both show the highest rate of self-pollination and presumably the greatest amount of self-compatibility.

3.3. Identifying zygotic progeny from polyembryonic parents using silhouette scores

In group 5, AuMG005_Carabao 1_PHIL has a silhouette score of 0.11 while the rest of the group 5 individuals have a silhouette score of 0.85 and are all labeled as Carabao. Using the visual analysis method, AuMG005_Carabao 1_PHIL has no genotypes that would be impossible (1:2, 2:1 genotypes when compared to other group members) if another member of group 5 had been a parent of it. Since AuMG005_Carabao 1_PHIL is not identical to the other 'Carabao' accessions of Group 5, it is likely the zygotic embryo from a maternal 'Carabao' parent. AuMG005 was introduced to Australia in the early 1980′s separate from the other 'Carabao' accessions in the collection. It may have been selected due to favorable trait changes due to its zygotic nature.

3.4. Identifying putative paternal parents of hybrids with known maternal parents

Similarly, silhouette scores can be used to identify the putative paternal parents of cultivars from a known maternal line whether poly or monoembryonic. PC307::WB2-07-44::T/A 05-02, an open pollinated progeny of 'Tommy Atkins' occurs in group 21 with a silhouette score of 0.05 while the other group 21 individuals are 'Nam Doc Mai' (polyembryonic) with silhouette scores ranging from 0.76 to 0.84. The maternal parent is known to be 'Tommy Atkins' and there is only one impossible genotype (1:2, 2:1 as above) of 272 SNP loci, which can be attributed to genotyping error as described above, when PC307 is compared to the 'Nam Doc Mai' accessions in Group 21. Thus, the paternal parent is likely the Thai cultivar 'Nam Doc Mai'. Further visual verification by sorting on the 'Nam Doc Mai' parent and copying and pasting the 'Tommy Atkins' genotype into group 21 showed that there were no genotypes other than the one previously noted that were not consistent with 'Nam Doc Mai' as the paternal parent. This data is not shown but can be demonstrated by sorting Supplemental Table 2 as described. Likewise, in group 58, PC315::WB2-07-58::T/A 05-02, an open pollinated progeny of 'Tommy Atkins' has a silhouette score of 0.13 and the other members of group 58 are approximately genotypically identical accessions of the Thai cultivar 'PPK' (a cultivar of variable spellings, e.g. 'Pu Pui Kali', 'Po Pyu Kalay', 'Po Piju Kalay', 'Po Pyo Ka Kal') with silhouette scores of 0.80. PC315::WB2-07-58::T/A 05-02

Groups	Silhouette	Inds	Contig5810	Contig1365	mango_rep_c9129	Contig6253	Mi_0481	mango_rep_c42460	Mi_0494	Contig2080	Contig1778	mango_rep_c51872
256	0.48	Mi1467_TOMMY ATKINS_FL	1	1	1	1	1	2	2	2	2	2
256	0.16	PC198::WB2-04-66::T/A 09-05	1	1	1	1	1	2	2	2	2	2
256	0.14	PC200::WB2-04-68::T/A 01-01	1	1	1	1	1	2	2	2	2	2
256	0.13	PC202::WB2-04-70::T/A 01-01	1	1	1	1	1	2	2	2	2	2
256	0.17	PC203::WB2-04-71::T/A 01-01	1	1	1	1	1	2	2	2	2	2
256	-0.06	PC207::WB2-04-75::T/A 02-03	3	3	1	3	1	2	3	2	3	2
256	-0.11	PC309::WB2-07-46::T/A 05-02	1	3	3	1	1	3	2	3	2	3
256	0.06	PC346::WB2-08-74::T/A 01-01	3	1	3	3	3	3	2	3	2	2
256	0.07	PC351::WB2-09-04::T/A 01-05	1	3	1	3	3	2	3	3	3	3

Fig. 2. Evidence for self-pollinated progeny using genotype data. Data are from an abbreviated version of the genotype dataset in Supplemental Table 2 for selected individuals in affinity group 256. Column 1 is the number of the affinity group. Column 2 is the silhouette score for the individual named in column 3. Row 2 of column 3 identifies the 'Tommy Atkins' maternal parent genotype and the other accessions are hybrids that have 'Tommy Atkins' as the maternal parent. The 10 following columns contain the genotypes at the SNP markers described in the column headings. Genotype cells are colored and coded as follows: 0, grey, missing data; 1, blue, homozygous allele; 2, orange, homozygous allele 2; 3, green, heterozygous. The first four hybrids share the same homozygous genotypes as the maternal parent and are presumed self-pollinated progeny. The next four hybrids differ from the maternal parent genotypes as heterozygotes and are presumed to have a paternal parent other than 'Tommy Atkins' (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Groups	Silhouette	Inds	Mi_0408	SSKP006C1_G845T	Contig2997	Contig3004	mango_rep_c33609	mango_rep_c10854	Contig5810	mango_rep_c15051	Contig_3554_A93C	mango_rep_c6478	mango_rep_c3996	mango_rep_c4496
150	0.35	AuMG088_Kent_FL	1	1	1	1	1	1	2	2	2	2	0	2
150	0.36	FSP_Mango Row_20120016::Palmer	1	1	1	1	1	1	2	2	2	2	2	2
150	0.35	M138 Senegal_27 cv. Kent	1	1	1	1	1	1	2	2	2	2	2	2
150	0.37	Mi1349_Kent	1	1	1	1	1	1	2	2	2	2	2	2
256	0.49	Mi1350_Tommy Atkins	3	2	1	2	2	2	1	1	3	1	2	1
150	0.00	PC240::WB2-05-65::Unknown (Lost Tag)	1	3	1	3	3	3	3	3	2	3	2	3
150	0.08	PC284::WB2-06-68::Tommy Atkins	1	3	1	3	3	3	3	3	3	3	2	3
150	0.05	PC305::WB2-07-42::T/A 05-02	1	3	1	3	3	3	3	3	2	3	2	3
150	0.06	PC366::WB2-09-59::Kent 03-02	3	3	1	3	3	1	3	2	3	2	2	2
150	-0.03	ZL025::Kent sdlg::N-12	3	3	1	1	1	1	2	3	2	3	2	3

Fig. 3. Evidence for the identification of hybrid parentage. Data are from an abbreviated version of the genotype dataset in Supplemental Table 2 for selected individuals in affinity group 150 and the genotype of 'Tommy Atkins' from affinity group 256. Column 1 is the number of the affinity group of the individual. Column 2 is the silhouette score in its affinity group for the individual named in column 3. The 12 following columns contain the genotypes at the SNP markers described in the column headings. Genotype cells are colored and coded as follows: 0, grey, missing data; 1, blue, homozygous allele; 2, orange, homozygous allele 2; 3, green, heterozygous. Rows 2–5 are the genotypes of three accessions of 'Kent' from three different germplasm collections and a putatively mislabeled accession of 'Palmer'. Row 6 is the genotype of 'Tommy Atkins'. Rows 7–9 are genotypes of hybrids with 'Tommy Atkins' as the maternal parent that putatively have 'Kent' as the paternal parent. Rows 10–11 are genotypes of hybrids with 'Kent' as the maternal parent that putatively do not have 'Kent' as the paternal parent (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

has only one inconsistent genotype when compared to the PPK accessions in the group and 'PPK' is the likely paternal parent of this accession. Visual verification as described above supported this conclusion as there were no genotypes inconsistent with 'PPK' as the paternal parent. PC279::WB2-06-59::T/A 05-02 in group 91 has a silhouette score of -0.01 and three identical accessions of 'Rapoza' have silhouette scores of 0.46, suggesting 'Rapoza' is a likely paternal parent of PC279::WB2-06-59::T/A 05-02. In group 117, by similar analysis and visual verification, the open pollinated progeny of 'Him Sagar', 'Tommy Atkins' and 'Keitt', but not 'Haden', likely have 'Glenn' as the paternal parent.

In group 150, there are nine hybrids with attributed maternal parents as follows, 'Keitt' 2, 'Tommy Atkins' 6, and one unknown, that likely have 'Kent' as a paternal parent based on their silhouette scores. Fig. 3 is an abbreviated example of the visual analysis of group 150 where the genotype data has been sorted by row on 'Kent'. The figure shows a mislabeling of FSP016::FSP_Mango Row_20120016::Palmer as 'Palmer' because it is identical in genotype to 'Kent' from the Australian, Senegal and SHRS germplasm collections. The 'Tommy Atkins' genotype from group 256 has been added to show that the three hybrids with 'Tommy Atkins' as the maternal parent have no genotypes that would contradict 'Kent' as the paternal parent. Finally, the two 'Kent' hybrids are shown to not be self-pollinated by the criteria described above. The presence of a hybrid with a low silhouette score in a group with known cultivars with high silhouette scores does not guarantee that the known cultivar is the paternal parent, as is seen in group 176, where the 'Keitt' seedling has too many impossible genotypes to have 'Brooks' as the paternal parent. This is also seen in group 194 where the 'Keitt' seedling has too many impossible genotypes to have 'Ah Ping' as a paternal parent. Visual verification was used to identify these occurrences and should always be used to verify paternal parentage inferences based on silhouette score.

3.5. Determination of a subset of SNP markers for use in identification of particular cultivars

We have genotyped all 1915 accessions with 272 SNP markers to create a useful database for comparing germplasm accessions from multiple collections. Not all 272 SNP markers are necessary to distinguish an accession from all others in the database. Although there is no single core set of SNPs that can reliably differentiate all accessions in the database due to the breadth of the dataset from half-siblings to other genera, individual subsets of SNP markers can be designed to differentiate a particular cultivar of interest from all other accessions in the dataset. As the dataset appears to encompass the known genetic diversity of the genus *Mangifera*, such a subset could be used to differentiate a cultivar of interest from any other mango.

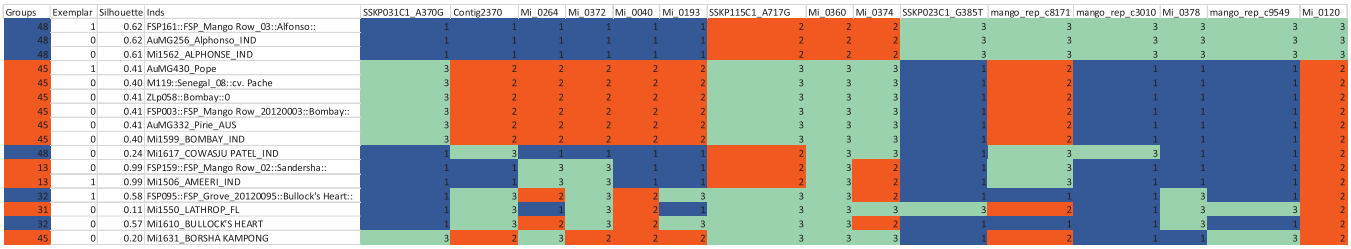
Fig. 4 gives an example of how such a subset for the cultivar 'Alfonso' could be determined. First, the dataset in Supplemental Table 2 is sorted by the 'Alfonso' genotype in row 340 (FSP161 Alfonso), the exemplar in affinity group 48 to order the markers in the increasing sequence 0-3. Next, the range of 1 s and 2 s are determined for 'Alfonso', in this case columns H to DH are 1 s and columns DI-FT are 2 s. Countif formulas are made for these ranges, so that every difference between another accession and 'Alfonso' can be counted. For example: =countif (H2:DH2,2) + countif(DI2:FT2,1) where the values increase (H3:DH3, etc) for each row, will count all the 1:2 and 2:1 differences between all other accessions and 'Alfonso'. Then the dataset can be sorted by column on the number of differences and a subset of SNP markers that will distinguish 'Alfonso' from the genotypically closest accessions can be determined by visual inspection of the sorted dataset. An abbreviated example is given in Fig. 4. Note that even subsets as small as seven markers that show 1:2, 2:1 differences can be reliably used to differentiate 'Alfonso' from its closest accession in the dataset.

3.6. Marker association of embryony type

'Kensington Pride' is a polyembryonic cultivar that is used as the paternal parent in crosses with monoembryonic cultivars in the Australian mango breeding program. In two mapping populations of mango ('Tommy Atkins' x 'Kensington Pride', 'Creeper' x 'Kensington Pride'), the polyembryony trait was associated with SNP markers on Linkage Group 8 (Kuhn et al., 2017). These markers had been included in the 384 SNP markers used to genotype the germplasm collection. For 199 individuals in the Australian germplasm collection embryony and genotype data were used. We found a significant relationship between SNPs and the embryony trait with the chi-squared test for independence ($p = 0.0$). Using post hoc Fisher's Exact Tests, we attempted to associate the embryony trait with a SNP marker across the germplasm collection (Table 4). There were 98 significant (p-value < 2.71e-6) markers out of 369. Mi_0173 which had shown association with the trait in two mapping populations was one of the top 10 markers that showed association with the embryony trait in the germplasm collection. Three of the other top 10 markers were found on LG19 between 0 and 39 cM.

3.7. Species

There were 113 accessions labeled as either genera other than *Mangifera* or species other than *M. indica* within *Mangifera*, with the dataset containing 100 accessions from 23 different *Mangifera* species representing the two subgenera and four sections described by Kostermans and Bompard (Kostermans and Bompard, 1993) (Table 2). In group 0 (Supplementary Table 2), all silhouette scores were low ranging from -0.24 to 0, and accessions represented three genera (*Anacardium*, *Bouea*, and *Mangifera*) in the family *Anacardiaceae* and seven species of *Mangifera* other than *M. indica*, three from subgenus *Mangifera* (*M. gedebe, M. quadrifida, M. casturi*), and the other four from subgenus *Limus (M. macrocarpa, M. foetida, M. caesia, M. pajang)*. The low silhouette scores in this group likely reflect that individuals are



Groups	Exemplar	Silhouette	Inds	SSKP031C1_A370G	Contig2370	Mi_0264	Mi_0372	Mi_0040	Mi_0193	SSKP115C1_A717G	Mi_0360	Mi_0374	SSKP023C1_G385T	mango_rep_c8171	mango_rep_c3010	Mi_0378	mango_rep_c9549	Mi_0120
48	1	0.62	FSP161::FSP_Mango Row_03::Alfonso::	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3
48	0	0.62	AuMG256_Alphonso_IND	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3
48	0	0.61	Mi1562_ALPHONSE_IND	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3
45	1	0.41	AuMG430_Pope	3	2	2	2	2	2	3	3	3	1	2	1	1	1	2
45	0	0.40	M119::Senegal_08::cv. Pache	3	2	2	2	2	2	3	3	3	1	2	1	1	1	2
45	0	0.41	ZLp058::Bombay::0	3	2	2	2	2	2	3	3	3	1	2	1	1	1	2
45	0	0.41	FSP003::FSP_Mango Row_20120003::Bombay::	3	2	2	2	2	2	3	3	3	1	2	1	1	1	2
45	0	0.41	AuMG332_Pirie_AUS	3	2	2	2	2	2	3	3	3	1	2	1	1	1	2
45	0	0.40	Mi1599_BOMBAY_IND	3	2	2	2	2	2	3	3	3	1	2	1	1	1	2
48	0	0.24	Mi1617_COWASJU PATEL_IND	1	3	1	1	1	1	2	3	3	1	3	3	1	1	2
13	0	0.99	FSP159::FSP_Mango Row_02::Sandersha::	1	1	3	3	1	1	2	3	2	1	3	1	1	1	2
13	1	0.99	Mi1506_AMEERI_IND	1	1	3	3	1	1	2	3	2	1	3	1	1	1	2
32	1	0.58	FSP095::FSP_Grove_20120095::Bullock's Heart::	1	3	2	3	2	3	3	3	2	1	1	1	3	1	2
31	0	0.11	Mi1550_LATHROP_FL	1	3	1	3	2	1	3	3	3	3	2	1	3	3	2
32	0	0.57	Mi1610_BULLOCK'S HEART	1	3	2	3	2	3	3	3	2	1	1	1	3	1	2
45	0	0.20	Mi1631_BORSHA KAMPONG	3	2	3	2	2	2	3	3	3	1	2	1	1	3	2

Inds	0	1	2	3	Total differences	1:2, 2:1 differences	1:2, 1:3, 2:1, 2:3 differences
FSP161::FSP_Mango Row_03::Alfonso::	3	105	64	100	0	0	0
AuMG256_Alphonso_IND	10	100	59	103	2	0	2
Mi1562_ALPHONSE_IND	1	105	63	103	2	0	2
AuMG430_Pope	2	88	81	101	88	7	47
M119::Senegal_08::cv. Pache	0	89	81	102	88	7	47
ZLp058::Bombay::0	1	88	81	102	88	7	47
FSP003::FSP_Mango Row_20120003::Bombay::	0	89	82	101	89	7	47
AuMG332_Pirie_AUS	0	88	82	102	89	8	48
Mi1599_BOMBAY_IND	0	90	79	103	90	7	49
Mi1617_COWASJU PATEL_IND	2	106	66	98	97	5	50
FSP159::FSP_Mango Row_02::Sandersha::	2	89	92	89	122	10	61
Mi1506_AMEERI_IND	1	90	92	89	123	10	61
FSP095::FSP_Grove_20120095::Bullock's Heart::	2	89	85	96	121	10	63
Mi1550_LATHROP_FL	1	105	98	68	134	31	66
Mi1610_BULLOCK'S HEART	2	91	83	96	124	11	66
Mi1631_BORSHA KAMPONG	1	84	73	114	110	9	66

Fig. 4. Identifying a subset of SNP markers to differentiate a specific cultivar from all other cultivars in the database using genotype data. An abbreviated version of the genotype dataset in Supplemental Table 2 that has been sorted by row for the cultivar 'Alfonso' (various spellings) and then sorted by column for genotype differences of accessions compared to 'Alfonso'. Column 1 is the number of the affinity group of the individual. Column 2 is the silhouette score in its affinity group for the individual named in column 3. The 15 following columns contain the genotypes at the SNP markers described in the column headings. Genotype cells are colored and coded as follows: 0, grey, missing data; 1, blue, homozygous allele; 2, orange, homozygous allele 2; 3, green, heterozygous. Rows 2–4 are the genotypes of 'Alfonso'. Rows 5–14 are accessions have been compared to 'Alfonso' and sorted by increasing number of differences to 'Alfonso'. The associated table shows the name of the accession from the colored genotype above in Column 1. Columns 2–5 have the number of missing data (0), homozygous allele 1genotypes (1), homozygous allele 2 (2), and heterozygous genotypes (3) for these accessions from the whole genotype dataset, not the abbreviated version. Columns 6–8 have the count of numbers of different genotypes and the type of difference (1:2, 1:3, 2:1,2:3, accession: 'Alfonso') (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

evenly dispersed in this area of the dataspace. Visual inspection of Group 0 in Supplemental Table 2 after zooming out as far as possible reveals and validates the common genotype features of the group, dismissing long branch attraction as a possible explanation of the grouping. Group 1 had much higher silhouette scores, 0.21 to 0.66, and accessions representing 9 *Mangifera* species other than *M. indica*, from both subgenera and two sections of subgenus *Mangifera* (*Mangifera* and *Rawa*). Group 3 had two accessions of *M. altissima* and no other species, but AuMG036 *M. altissima* was genotypically identical to AuMG014 'Pau. Previously these two accessions were found to be genetically identical using 11 SSR markers (Dillon et al., 2013). Kostermans and Bompard (1993) list Pau as a vernacular name for *M. altissima*. Group 8 had five *Mangifera* species including *M. indica*, but seven of the nine accessions of *M. casturi* (subgenus *Mangifera*, section *Mangifera*) were in this group. Five of those *M. casturi* accessions were genotypically identical but also genotypically identical to ew29::KRB_30_M.rufocosta (subgenus *Mangifera*, section *Mangifera*) and ew19::KRP_17_M.macrocarpa (subgenus *Limus*). Group 15 was made up of *M. lalijiwa* and *M. indica* accessions. However, the *M. indica* accession AuMG034_Mangga Madu has previously been suggested to be genetically similar to *M. lalijiwa* (Dillon et al., 2013). Group 20 had 10 of the 17 accessions of *M. odorata* (subgenus *Limus*) that were genotypically identical to two accessions of *M. torquenda* (subgenus *Mangifera*). With limited genetic diversity seen in an intra-species study of *M. odorata* (Yamanaka et al., 2006) and the morphological similarity of the flowers of *M. torquenda* and *M. odorata* (Kostermans and Bompard, 1993) it is assumed that AuMG013_*M. torquenda* and AuMG021_Torquenda Lamantana are mislabeled (Dillon et al., 2013). Group 36 had the only two accessions of *M. pentandra* (subgenus *Mangifera*, section *Euantherae*). Groups 47, 74 and 108 were genotypically distinct but all labeled as *M. laurina*, which seems split into *M. laurina* 'Ipoh' (group 47), *M. laurina* (group 74) and *M. laurina* 'Lombok' (group 108). Group 77 had the only accession of *M. longipes* (not listed by Kostermans and Bompard) which was genotypically identical to ew26::KRB_7_M. caesia. Group 102 had two accessions of *M. applanata* that were genotypically identical to four accessions of *M. rubropetala*, further data are required to determine which species is mislabeled (Dillon et al., 2013).

4. Discussion

The primary objective of this study was to develop a mango SNP database that could be used as a genomics tool assisting curators with germplasm maintenance and plant breeders with the development of new cultivars. This work now represents the largest, most comprehensive genotyping effort in mango to date. SNP markers were selected because they are platform-independent, reproducible across labs, and the resulting databases can be shared globally. SNP genotype data is easy to collect in large amounts due to the high frequency of SNPs, ease of design from transcriptome or genome assemblies, and the availability of high throughput SNP assay platforms. We selected a subset of mango SNP markers that were initially developed to produce a genetic map (Kuhn et al., 2017), and used these markers to genotype as much of the world's mango germplasm as could be obtained for this study. Although we included 384 markers, selecting 50 random subsets of SNPs markers resulted in distance matrices and groupings that were not significantly different from the full dataset (data not shown). This indicates that fewer markers can be used to genotype accessions in the future. The advantage of having a larger set of markers, though, is that it provides users with the ability to select subsets of markers that distinguish particular cultivars of interest.

Table 2
Germplasm accessions in the genotype database and their corresponding position in Kosterman and Bompard's taxonomic categories for the genus Mangifera. The genus, subgenus, section and species are from Kosterman and Bompard's proposed taxonomy of the genus Mangifera. Affinity group of putative accessions of that species. Number of putative accessions of that species in that affinity group and the total number of accessions of any type in that affinity group.

Genus	Subgenus	Section	Species	Affinity Group	Number of accessions of species/total accessions in group
Mangifera	Mangifera	Marchandora	gedebe	0	3/23
Mangifera	Mangifera	Euantherae	caloneura		
Mangifera	Mangifera	Euantherae	cochinchinensis	2	1/10
Mangifera	Mangifera	Euantherae	pentandra	36	2/5
Mangifera	Mangifera	Rawa	parvifolia		
Mangifera	Mangifera	Rawa	paludosa		
Mangifera	Mangifera	Rawa	griffithii	1	2/22
Mangifera	Mangifera	Rawa	gracilipes		
Mangifera	Mangifera	Rawa	merrillii		
Mangifera	Mangifera	Rawa	microphyla		
Mangifera	Mangifera	Rawa	minutifolia		
Mangifera	Mangifera	Rawa	andamanica		
Mangifera	Mangifera	Rawa	nicobarica		
Mangifera	Mangifera	Mangifera	altissima	3	2/4
Mangifera	Mangifera	Mangifera	similis	1	2/22
Mangifera	Mangifera	Mangifera	similis	108	1/11
Mangifera	Mangifera	Mangifera	torquenda	20	2/17
Mangifera	Mangifera	Mangifera	mucronulata		
Mangifera	Mangifera	Mangifera	applanata	102	2/9
Mangifera	Mangifera	Mangifera	longipetiolata		
Mangifera	Mangifera	Mangifera	quadrifida	0	1/23
Mangifera	Mangifera	Mangifera	quadrifida	1	6/22
	Mangifera	Mangifera	quadrifida	2	1/10
	Mangifera	Mangifera	quadrifida	8	1/12
	Mangifera	Mangifera	quadrifida	35	2/5
	Mangifera	Mangifera	quadrifida	36	1/5
	Mangifera	Mangifera	quadrifida	129	1/3
Mangifera	Mangifera	Mangifera	sumbawaensis		
Mangifera	Mangifera	Mangifera	timorensis		
Mangifera	Mangifera	Mangifera	magnifica	1	1/22
Mangifera	Mangifera	Mangifera	linearifolia		
Mangifera	Mangifera	Mangifera	sulavesiana		
Mangifera	Mangifera	Mangifera	swintoniodes		
Mangifera	Mangifera	Mangifera	dewildei		
Mangifera	Mangifera	Mangifera	monandra		
Mangifera	Mangifera	Mangifera	casturi	0	1/23
Mangifera	Mangifera	Mangifera	casturi	1	1/22
	Mangifera	Mangifera	casturi	8	7/12
	Mangifera	Mangifera	casturi	46	1/5
Mangifera	Mangifera	Mangifera	indica		
Mangifera	Mangifera	Mangifera	rubropetala	4	1/2
	Mangifera	Mangifera	rubropetala	102	4/9
Mangifera	Mangifera	Mangifera	rigida		
Mangifera	Mangifera	Mangifera	dongnaiensis		
Mangifera	Mangifera	Mangifera	zeylanica	1	1/22
	Mangifera	Mangifera	zeylanica	147	6/7
Mangifera	Mangifera	Mangifera	oblongifolia	124	1/14
Mangifera	Mangifera	Mangifera	rufocostata	8	1/12
Mangifera	Mangifera	Mangifera	austro-yunnanensis		
Mangifera	Mangifera	Mangifera	collina		
	Mangifera	Mangifera	laurina	47	2/10
	Mangifera	Mangifera	laurina	74	6/12
	Mangifera	Mangifera	laurina	108	5/10
Mangifera	Mangifera	Mangifera	pedicellata		
Mangifera	Mangifera	Mangifera	flava		
Mangifera	Mangifera	Mangifera	austro-indica		
Mangifera	Mangifera	Mangifera	sylvatica		
Mangifera	Mangifera	Mangifera	minor		
Mangifera	Mangifera	Mangifera	lalijiwa	4	4/7
Mangifera	Mangifera	Mangifera	pseudo-indica		
Mangifera	Mangifera	Mangifera	orophila		
Mangifera	Limus		lagenifera		
Mangifera	Limus		decandra		
Mangifera	Limus		superba		
Mangifera	Limus		blommesteinii		
Mangifera	Limus		pajang	0	1/23
Mangifera	Limus		pajang	1	1/22
Mangifera	Limus		caesia	0	1/23
	Limus		caesia	1	1/22
	Limus		caesia	77	1/3
Mangifera	Limus		kemanga		
Mangifera	Limus		macrocarpa	0	1/23

(*continued on next page*)

Table 2 (*continued*)

Genus	Subgenus	Section	Species	Affinity Group	Number of accessions of species/total accessions in group
	Limus		macrocarpa	8	1/12
Mangifera	Limus		foetida	0	7/23
	Limus		foetida	42	1/3
Mangifera	Limus		leschenaultii		
Mangifera	Limus		odorata	20	10/17
	Limus		odorata	148	1/2
	Limus		odorata	244	1/6
	Uncertain		subsessilifolia	1	3/22
					100 individuals

Table 3
Estimation of self-compatibility for maternal parents by analysis of hybrids of known maternal parentage. Column 1 is the putative maternal parent of the hybrid. Column 2 is the number of hybrids with that maternal parent. Column 3 is the number of hybrids that are presumed to be self-pollinated by visual analysis. Column 4 is the percentage of total hybrids of the maternal parent that are self-pollinated as an estimate of self-compatibility.

Maternal parent of hybrid	Number of hybrids	Number of self-pollinated individuals	% self-pollination
Tommy Atkins	104	23	22%
Keitt	207	12	6%
Kent	21	0	0%
Haden	28	6	21%
Him Sagar	26	0	0%

Table 4
Top 10 markers associated with polyembryony in germplasm dataset. Marker names, linkage group from mango genetic map, position in cM on linkage group, and p-value estimation of association of marker with the polyembryony trait are detailed. Map information is from Kuhn et al., 2017.

Marker	Linkage Group	Position on Linkage Group in cM	p-value
Contig2850	19	25.5	2.09E-21
Mi_0173	8	46.1	2.68E-20
Contig560	19	0	7.32E-18
Mi_0426	4	105.6	1.47E-17
SSKP036C1_A393G	Not mapped	Not mapped	1.61E-17
Mi_0227	18	9.8	1.75E-17
Contig2005	19	39.0	8.91E-16
mango_rep_c4227	13	74.1	1.78E-15
Mi_0252	15	32.6	5.96E-15
mango_rep_c8171	17	26.9	8.34E-15

[1,] Results of the permutation test for the mango dataset:
Chi-squared significance (p): 0.0.
Chi-squared Degrees of freedom: 1106 Bonferroni threshold (p) for post hoc tests: 2.7100271002710026e-6.

4.1. SNP markers

The SNP markers were developed from full sibling populations and had no SNPs identified within 100 nucleotides on either side of the variant nucleotide to allow design of SNP assays for the high throughput platforms (Kuhn et al., 2017). In addition, the SNP markers designed at SHRS (Mi_) were developed from RNA sequence data from multiple *M. indica* accessions and a single *M. casturi* accession chosen for their genetic diversity as measured by other methods (Schnell et al., 2006). The high conservation demanded by filtering the flanking sequences and their origin from coding regions suggested that there was a chance that some of the SNP flanking sequences might be conserved in other *Mangifera* species and permit assaying of the variable locus outside of *M. indica*. However, the presence of heterozygosity in the parents of the full sibling population suggests that this locus can be fixed for either allele in other individuals including grandparents or more distant individuals in the lineage of the full siblings. Further, the use of SNP markers for estimating genetic diversity at the population, species, and genus level poses several challenges.

The majority of markers developed (˜70%) were capable of reliably genotyping accessions from other *Mangifera* species and other genera in the *Anacardiaceae*. Since the SNP markers were developed from primarily *M. indica*, it might be expected to see lower variability (heterozygosity) in other species. What was unexpected was that these accessions outside *M. indica* would frequently be homozygous for the identical allele. Until this study, species identification has been based on phenotypic characters.

If individuals from other genera can be accurately genotyped, then these markers should be sufficient to capture all the genetic diversity within the genus *Mangifera*. We successfully used the markers to distinguish accessions from the genus down to the half sibling level. The accessions that have been genotyped do not represent a population in the genetic sense of the term, because of the wide scope of the genotyped accessions, the vegetative reproduction of the named clonal cultivars, and the inclusion of open pollinated progeny of clonally propagated cultivars.

4.2. Affinity propagation

Our initial attempts to analyze the data in this study using the traditional analytical methods (neighbor joining or UPGMA clustering, STRUCTURE, PCA) did not produce results that made biological sense based on our prior knowledge of geographic and pedigree data. A biologically-relevant factor contributing to this complexity includes the often unknown pedigree relationships among parents, hybrids, self-pollinated progeny, and siblings. The biallelic nature of SNP markers also made analysis challenging. SNPs can be homozygous allele 1, homozygous allele 2, or heterozygous. This does two things to the analysis. First, it dramatically increases the identity by chance as there are only three possible states. Second, it essentially erases the importance of “private” or rare alleles which drive most of the genetic diversity estimation for other analysis methods (eg. STRUCTURE).

An alternative analysis method was therefore developed to meet the primary objectives of this research to assist curators and breeders with germplasm identification. As described in Materials and Methods, this method made no genetic assumptions about the data while generating the affinity propagation groups that best matched prior expectations while enabling the inference of unknown accessions. The literature reflects affinity propagation’s relative ubiquity in solving clustering problems with respect to molecular data. Affinity propagation has played a crucial role in GWAS analysis software (e.g. DEPICT) (Pers et al., 2015), subspecies identification (Borile et al., 2011), germplasm evaluation (de Oliveira et al., 2015), and DNA motif discovery (Sun et al., 2015). Affinity propagation cannot reliably determine the relationships of groups or perform a higher order grouping, but there is greater than 50% support for membership in these groups.

4.2.1. Silhouette scores

The silhouette scores that measure quality of membership in a group

have allowed easier identification of mislabeled, misidentified, potential self-pollinated individuals, and even, in some cases, the identification of potential paternal parents for open pollinated progeny. In some cases, silhouette scores for all members of a group are close to zero or even negative. Based on how the scores are calculated, the low scores are due to a combination of within group dispersal and the density of the space the entire dataset occupies; i.e. the dispersed group members are close to the outermost borders of other groups. However, visual representation of group members (Supplementary Table 2) makes group relationships clear. In a group where the exemplar and all members have low silhouette scores, the accessions are evenly dispersed through that portion of data space identified by the affinity grouping and so each is an equally poor center (exemplar) for the group. Thus, although high silhouette scores always indicate a high degree of genetic identity, low silhouette scores, especially if low for an entire group, do not mean that the group members are not genetically related, but that they are more dispersed in the volume of data space that they inhabit.

4.3. Estimating genetic diversity in the germplasm dataset

Affinity propagation generated 258 groups of the 1915 accessions, with the most frequent group size having three members and most of the groups having between 2 and 15 members. These groups can be used to assist in the curation of germplasm collections. The 258 groups represent genetic diversity from the genus level down to the half sibling level and, overall, the groups reflect this diversity correctly. Some species are grouped by themselves, while others from groups with multiple species. Half siblings are grouped either by maternal parent or paternal parent. *M. indica* germplasm accessions, the most important group in this study, are found to be grouped by geographic location (e.g. 'Carabao' in the Philippines, 'Nam Doc Mai' in Thailand), or by genetic identity. For example, the 259 clonal accessions in the SHRS germplasm collection appear in 138 groups. A single individual from each of the 138 groups would be sufficient to capture all the genetic diversity in the germplasm collection should it be necessary to choose individuals for a backup collection at another site. The number of individuals representing the genetic diversity in the SHRS collection may decrease once mislabeled/misidentified accessions have been removed. Thus, all the genetic diversity encompassed by the current germplasm collection could be maintained in a collection approximately half the size of the current one and with greater confidence of the identity of the accessions based on comparison with genotypes of accessions from other collections. Another advantage for the SHRS program is that this information will prove useful in determining priority of rescue of trees after a hurricane or in prioritizing grafting of trees for regenerating the collection.

A useful outcome from this study could also be the identification of a subset of genetically identical accessions found in multiple locations around the world. These could form the basis of a study to evaluate their responses to different climatic and soil environments and provide a means to determine the relative genetic by environmental (GxE) influences on traits of interest to growers and breeders.

4.4. Estimating mislabeling/misidentification in the germplasm collection

Correctly identifying accessions in germplasm collections is crucial to their utility and for the distribution of material to requestors. Mislabeling/misidentification is a common problem with all germplasm collections and, in recent years, curators have turned to molecular markers to reduce the amount of mislabeling and misidentification in collections. Genotyping labeled clones from other germplasm collections to verify the identity of the potentially mislabeled clones may also be confounded by the source of the material. Germplasm exchange in the past was quite common, so it should be possible to come to a consensus for labeled clones and specific genotypes.

We have taken two approaches to estimate the prevalence of mislabeling using molecular markers, neither one provides a complete solution. In the Australian germplasm collection, there are a large number of farmers' selections of 'Kensington Pride' that are close to genetically identical and all members of a single affinity group. Within this group are other named accessions that are not related to 'Kensington Pride' but are almost genetically identical to it. Using this single group, we calculated a potential mislabeling of 15% that may represent the upper limit of mislabeling for the entire collection as there were fewer accessions in the group than in the whole collection.

The other estimate of mislabeling occurred when two or more accessions appeared in more than one group. Without regard to which clone was correct, it clearly indicated mislabeling. For all the germplasm collections, this method estimated ˜8% mislabeling, which is an underestimate and, also the lower limit of mislabeling for the collection. The lower estimate was done by counting mislabeling events for each collection and dividing by the total number of accessions from that collection that were genotyped. In any event, mislabeling is a serious problem for all these germplasm collections and the genotype data should be used to reduce this issue either by relabeling based on genotype or removal of accessions from a collection.

4.5. Zygotic and maternal embryos of polyembryonic cultivars

Polyembryonic maternal parents that are well adapted to the environment and soil type of the region are frequently used to generate clonal rootstock based on the belief that such maternal parents "breed true" and the maternal apomictic embryo is selected over the zygotic embryo. At SHRS, we have employed 'Turpentine' as rootstock for over 30 years. In addition, Turpentine was the paternal parent of 'Haden', a Florida cultivar developed in the early 20th century. Our genotyping of all accessions labeled 'Turpentine' at SHRS and comparison to accessions from other germplasm collections demonstrate that our 'Turpentine' accessions fall into five different groups (78, 113, 131, 196, 231). Group 78 is the likely 'Turpentine' as accessions labeled 'Turpentine' from two other germplasm collections are identical to three SHRS accessions and four other SHRS probably mislabeled/misidentified accessions that are likely to be rootstock. Group 131 contains two SHRS accessions labeled 'Turpentine' as does group 231, with groups 113 and 196 containing one labeled 'Turpentine' accession each. None of these accessions can be first generation outcrossed progeny of the accessions in group 78 due to the presence of a large number of impossible genotypes (maternal parent homozygous for allele 1, progeny homozygous for allele 2). Thus, the 'Turpentine' accessions in the groups other than 78 are either two zygotic generations from the original maternal parent or mislabeled. If clonal rootstock is a necessity, for example for clonal trials of scions, rootstock seedlings should be genotyped using a small subset of markers that will distinguish apomictic embryos from zygotic embryos prior to grafting of the scion.

4.6. Self-pollination and self-compatibility

In mango orchards populated with clones derived from perhaps several thousand years of selection and vegetative propagation, self-compatibility may be less frequent as pollen from other clones is more available. Because we had genotypes from more than 400 hybrids where the maternal parent was known to be monoembryonic and many other mango cultivars were in proximity to the maternal parent, determination of the frequency of self-pollination in the progeny would allow a relative estimation of self-compatibility among maternal parents. The visual representation of the data using colors for the genotypes made it relatively easy to determine by inspection if progeny were likely to result from self-pollination. Self-pollinated progeny would share the same homozygous alleles as the maternal parent. By simply counting the number of genotypes for an individual that differed from the maternal parent at those homozygous loci, self-pollinated progeny

could be distinguished from outcrossed progeny. Identifying and removing self-pollinated progeny is important in breeding programs as most maternal parents are heterozygous and the likelihood of generating homozygous recessive allele progeny is increased. By genotyping the progeny at the seedling stage, self-pollinated progeny can be removed from the population that are to be grown up for evaluation and selection, saving the breeder the significant cost per tree to maintain and evaluate experimental plots. Anecdotally, we have observed that self-pollinated Tommy Atkins progeny on our station rarely flower (David Kuhn, *personal observation*).

4.7. Inferring paternal parents of hybrids of known maternal parents

Inferring potential paternal parents of hybrids using affinity propagation analysis and silhouette scores with a simplified visual analysis is a novel approach to analyzing a large dataset of SNP genotypes from a large number of individuals. As in the analysis of zygotic embryos, the key is to look for impossible genotypes between the potential paternal parent and the hybrid. In the case where the hybrids are found in the same group as a potential paternal parent, simply hiding the rows that are not hybrids or the potential parent allows a quick visual scan through the markers to identify hybrids that have one or no impossible genotypes. Such identification of potential paternal parents is particularly important to gain outcrossed material for breeding programs. With the current dataset, genotyping of progeny from maternal trees at the seedling stage and regeneration of the affinity groups can be used to rapidly identify specific paternal parents of hybrids and to select these parents for more detailed phenotypic evaluation.

4.8. Marker association in germplasm collections

Hybrid mapping populations with known maternal and paternal parents that differ significantly for a trait provide a much more sensitive means to associate traits with markers than do germplasm collections. In a mapping population, linkage disequilibrium is high, such that markers every 5 cM are often sufficient to obtain significant associations (J. van Ooijen, JoinMap 4 manual). In germplasm collections, linkage disequilibrium is low as many more meioses have occurred between individuals than in an F1 mapping population.

For the qualitative trait of polyembryony, we were able to find several markers in the same relative positions in the same linkage group of two different mapping populations that shared a single maternal polyembryonic parent ('Kensington Pride'). We attempted to identify similar associations within the germplasm collection that had many polyembryonic accessions. We used the Fisher's exact statistical test to determine association, since polyembryony is a qualitative trait with only two states (polyembryonic and monoembryonic). The marker (Mi_0173) that had shown an association with polyembryony in the mapping populations, also showed the second highest association (lowest p-value score) across the germplasm collection as shown in Table 4. In addition, several other markers found on a different linkage group and within 30 cM of each other were in the top ten markers.

Trait association using germplasm collections can be confounded by mislabeling, low amounts of phenotypic data, and the potential for undetected family structure. We performed a permutation test (chi-squared test for independence) that showed that the polyembryony trait was associated with markers in the germplasm collection genotype data ($p = 0.0$). The Bonferroni correction for post hoc Fisher's Exact Tests was approximately 2.71e-6, with 98 markers significant at $p < 2.71e\text{-}6$. Such a high number of markers significantly associated with the trait may be explained by the dataset representing a collection not a population or family. Family structure, especially for polyembryonic individuals that are clonal, may cause a false association of markers to the polyembryony trait. We cannot discount the possibility that there are more genes that regulate polyembryony despite only finding one in our mapping populations. The two mapping populations that showed a significant association of markers on the same linkage group with polyembryony shared a single polyembryonic paternal parent. Thus, the associations of markers on other linkage groups with polyembryony discovered in the germplasm study may be equally valid as the parents are unknown. The germplasm collection study is likely to lead to the identification of further candidate loci that may be linked to genes regulating polyembryony.

4.9. Genotyping accessions from other genera and species

With the caveat that the accessions labeled as other species may also have a rate of mislabeling as high as 15% and an unknown rate of misidentification, we were able to distinguish distinct affinity groups that either represented a single species or two labeled species that were genetically identical. However, due to vagaries in the naming of accessions, the grouping of many different species and genera together in a single group, and the small number of species with multiple accessions, it is difficult to assess the value of the genotypic data in clarifying or validating species identification.

The three accessions from *Bouea* (three *B. macrophylla* and one *B. oppositifolia*), the single accession from *Anacardium*, and some accessions of seven *Mangifera* species other than *M. indica* are found in group 0. Silhouette scores are low for all accessions in Group 0 but visual inspection of the data does not suggest that individuals from different genera are identical or that there are no distinct differences that would identify individuals from a particular genus or species. In *Mangifera,* three of the accessions are in subgenus *Mangifera*, but not in the same section and four are in subgenus *Limus*.

Group 1, containing accessions from nine different *Mangifera* species other than *M. indica*, two common name accessions and two accessions labeled *M. species*, highlights the problem with labeling of accessions and species identification. Group 1 has six accessions (AuMG055_12. M. similis, ew14::KRB_2_M. similis, AuMG060_46. M. quadrifida, AuMG532_M. quadrifida (NT DPI), AuMG461_Assam Ramuk, AuMG537_Ramuk) that represent three pairs of labeled accessions that are all genetically identical. It is unclear if 'Ramuk' is the local name for another species or if this would be *M. quadrifida* or *M. similis*. Of the nine accessions in Group 1 labeled as *Mangifera* species, two are in subgenus *Limus*, six in subgenus *Mangifera* and one is not assigned to a subgenus or section (*M. subsessifolia*). Thus, in the groups that contain the greatest number of different species, there is no definitive evidence to support the distinction between subgenera proposed by Kostermans and Bompard (Kostermans and Bompard, 1993).

In cases where there are two or more accessions labeled as a species and either all accessions are in the same affinity group or the majority of accessions are in a single affinity group, the genotype data support the identity of distinct species. From Table 2, there is evidence for *M. gedebe*, *M. pentandra*, *M. griffithii*, *M. altissima*, *M. quadrifida*, *M. casturi*, *M. rubropetala/M. applanata*, *M. zeylanica*, *M. lalijiwa*, *M. foetida*, *M. odorata*, and *M. subsessifolia*. Accessions labeled *M. laurina* occur in three distinct groups that may represent three separate species or subspecies of *M. laurina* (Ipoh, Lombok and Laurina). These subdivisions of *M. laurina* are supported by phenotypic data with Ipoh being phenotypically distinct from the other *M. laurina* species (Ian Bally, *personal communication*). Thus, despite some potential mislabeling/misidentification, there is genotypic support for at least 13 of 23 species represented in the dataset, with the potential for supporting more distinct species identification when more accessions become available for genotyping. At the same time, a clear species identification for *M. indica* becomes less certain, due to accessions labeled as *M. indica* appearing as distinct affinity groups with less in common with other *M. indica* groups as there is between distinct species. One possible explanation is that *M. indica* is really a cultigen with diverse introgression from other species with the confounding factor of selection and vegetative propagation over several thousand years.

5. Conclusions

The generation of the largest mango SNP genotype database derived from diverse germplasm accessions ranging from different genera to half sibling hybrids provides a tool that is useful to mango breeders and researchers worldwide. The employment of a novel grouping method based on affinity propagation and the scoring of the quality of membership in a group by the silhouette analysis has shown that the genotype data can be used successfully to estimate germplasm genetic germplasm and identify mislabeling across the entire range of accessions. In addition, using a simple color coding of the genotypes, we have shown that identification of self-pollinated or outcrossed progeny, estimation of self-compatibility for maternal parents, and identification of likely paternal parents for hybrids of known maternal parents can be easily interrogated by visual inspection. Finally, association of important horticultural qualitative traits with specific SNP markers across the entire germplasm is possible if sufficient reliable phenotypic data is available.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions

DK, IB, ND, BF, EW -mango germplasm collections; DK, BF, JR -SNP markers; DK, AG, JR -data reformatting and analysis; DK, IB, ND, DI, AG, JR, BF, EW, AC conception and design of the work, drafting and revising the manuscript.

Funding

DK, AG, JR, BF were funded by USDA-ARS CRIS #6631-21000 – 022-00D and the National Mango Board NACA#58-6038 – 5-001.

.

Acknowledgments

Thanks to Elaini Oliveira dos Santos Alves (UESC, Bahia, Brazil), Carlos Antonio Fernandes Santos and Francisco Pinheiro Lima Neto (Embrapa Semiarido, Petrolina, Pernambuco, Brazil) for sharing mango germplasm and hybrid accessions. Thanks to Christina Currais (USDA-ARS-SHRS, USA) for outstanding effort in genotyping all the germplasm collections. Special thanks to Leo Ortega and the National Mango Board (USA) for their exceptional support in funding and encouraging this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.scienta.2019.03.037.

References

Arumuganathan, K., Earle, E., 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9, 208–218.
Bally, I.S., Dillon, N.L., 2018. Mango (Mangifera Indica L.) Breeding, Advances in Plant Breeding Strategies: Fruits. Springer, pp. 811–896.
Bally, I., Graham, G.C., Henry, R.J., 1996. Genetic diversity of Kensington mango in Australia. Aust. J. Exp. Agric. 36, 243–247.
Bompard, J., 2009. Taxonomy and Systematics. The Mango: Botany, Production and Uses. CAB International, Wallingford, pp. 19–41.
Borile, C., Labarre, M., Franz, S., Sola, C., Refregier, G., 2011. Using affinity propagation for identifying subspecies among clonal organisms: lessons from M. Tuberculosis. BMC Bioinformatics 12.
Bryant, C., Giovanello, K.S., Ibrahim, J.G., Chang, J., Shen, D.G., Peterson, B.S., Zhu, H.T., Initi, As.D.N., 2013. Mapping the genetic variation of regional brain volumes as explained by all common SNPs from the ADNI study. PLoS One 8.
De Candolle, A., 1885. Origin of Cultivated Plants. D. Appleton.
de Oliveira, E.J., Santana, F.A., de Oliveira, L.A., Santos, V.D., 2015. Genotypic variation of traits related to quality of cassava roots using affinity propagation algorithm. Sci. Agric. 72, 53–61.
Dillon, N.L., Bally, I.S.E., Wright, C.L., Hucks, L., Innes, D.J., Dietzgen, R.G., 2013. Genetic diversity of the Australian national mango genebank. Sci. Hortic. 150, 213–226.
Duval, M.-F., Risterucci, A.-M., Calabre, C., Le Bellec, F., Bunel, J., Sitbon, C., 2006. Genetic diversity of Caribbean mangoes (Mangifera indica L.) using microsatellite markers. VIII International Mango Symposium 820, 183–188.
FAO, 2016a. Global Fruit Production in 2016, by Variety (in Million Metric Tons). Statistica.
FAO, 2016b. Mango Production Worldwide From 2000 to 2016 (in Million Metric Tons). Statistica.
Galán Saúco, V., 2010. Worldwide mango production and market: current situation and future prospects. IX International Mango Symposium 992, 37–48.
Galán Saúco, V., 2015. Trends in world mango production and marketing. XI International Mango Symposium 1183, 351–364.
Hou, D., 1974. Anacardiaceae. Flora malesiana-series 1. Spermatophyta 8, 395–548.
Kostermans, A., Bompard, J.-M., 1993. The Mangoes: Their Botany, Nomenclature. Horticulture and Utilisation. Academic Press, London.
Kuhn, D., Dillon, N., Innes, D., Wu, L.-S., Mockaitis, K., 2014. Development of single nucleotide polymorphism (SNP) markers from the mango (Mangifera indica) transcriptome for mapping and estimation of genetic diversity. XXIX International Horticultural Congress on Horticulture: Sustaining Lives, Livelihoods and Landscapes (IHC2014): IV 1111. pp. 315–322.
Kuhn, D.N., Bally, I.S., Dillon, N.L., Innes, D., Groh, A.M., Rahaman, J., Sherman, A., Cohen, Y., Ophir, R., 2017. Genetic map of mango: a tool for mango breeding. Front. Plant Sci. 8, 577.
Litz, R.E., 2009. The Mango: Botany, Production and Uses. CABI.
Mahato, A.K., Sharma, N., Singh, A., Srivastav, M., Jaiprakash Singh, S.K., Singh, A.K., Sharma, T.R., Singh, N.K., 2016. Leaf transcriptome sequencing for identifying Genic-SSR markers and SNP heterozygosity in crossbred mango variety' Amrapali' (Mangifera indica L.). PLoS One 11.
Mukherjee, S.K., Litz, R.E., 2009. Introduction: botany and importance. The Mango: Botany, Production and Uses. pp. 1–18.
Pedregosa, F., Ga, Varoquaux, Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T., Frayling, T., Speliotes, E.K., Boehnke, M., Raychaudhuri, S., Fehrmann, R.S.N., Hirschhorn, J.N., Franke, L., Trai, G.I.A., 2015. Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. 6.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Schnell, R., Brown, J.S., Olano, C., Meerow, A., Campbell, R., Kuhn, D., 2006. Mango genetic diversity analysis and pedigree inferences for Florida cultivars using microsatellite markers. HortScience 41 993-993.
Sherman, A., Rubinstein, M., Eshed, R., Benita, M., Ish-Shalom, M., Sharabi-Schwager, M., Rozen, A., Saada, D., Cohen, Y., Ophir, R., 2015. Mango (Mangifera indica L.) germplasm diversity based on single nucleotide polymorphisms derived from the transcriptome. BMC Plant Biol. 15.
Singh, N.K., Mahato, A.K., Jayaswal, P.K., 2016. Origin, diversity and genome sequence of mango (Mangifera indica L.). Indian J. Hist. Sci. 51, 355–368.
Sun, C.X., Huo, H.W., Yu, Q., Guo, H.T., Sun, Z.G., 2015. An affinity propagation-based DNA motif discovery algorithm. Biomed. Res. Int.
Yamanaka, N., Hasran, M., Xu, D.H., Tsunematsu, H., Idris, S., Ban, T., 2006. Genetic relationship and diversity of four Mangifera Species revealed through AFLP analysis. Genet. Resour. Crop Evol. 53, 949–954.