Tree Genetics & Genomes (2019) 15: 71
https://doi.org/10.1007/s11295-019-1374-1

ORIGINAL ARTICLE

# Creation of an avocado unambiguous genotype SNP database for germplasm curation and as an aid to breeders

David N. Kuhn[1] · Amy Groh[2] · Jordon Rahaman[2] · Barbie Freeman[1] · Mary Lu Arpaia[3] · Noëlani Van den Berg[4] · Nilwala Abeysekara[5] · Patricia Manosalva[5] · Alan H. Chambers[6]

Received: 25 March 2019 / Revised: 2 July 2019 / Accepted: 15 July 2019 / Published online: 27 August 2019


## Abstract

Avocado (*Persea americana*) is an important tropical and subtropical fruit tree crop. Traditional tree breeding programs face the challenges of long generation times and significant expense in land and personnel resources. Avocado selection and breeding can be more efficient and less expensive through the development and application of molecular markers. A total of 1524 individuals were genotyped with 384 SNPs creating the largest SNP genotype database for avocado. These individuals correspond to four extensive germplasm collections including two housed in Florida and two in California. In addition, hybrids and selections from two rootstock breeding programs have been genotyped. Genotype data were analyzed using an affinity propagation method to define 155 groups. The 384 SNP markers provided accurate genotype data for individuals from different *Persea* species as well as half-siblings. Therefore, the majority of the genetic diversity of the avocado germplasm and related species that were genotyped has been captured. 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. Finally, this dataset is unambiguous so breeders can determine the genetic diversity of their breeding stock to optimize avocado breeding and selection programs by identifying outcrossed individuals at the seedling stage, thus increasing the efficiency of avocado genetic improvement.

**Keywords** Avocado germplasm · Affinity propagation clustering · SNP markers · Rootstock breeding · Visual analysis methods for large SNP datasets

Communicated by C. Chen

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s11295-019-1374-1) contains supplementary material, which is available to authorized users.

✉ Alan H. Chambers
achambers@ufl.edu

[1] Subtropical Horticulture Research Station, USDA-ARS, Miami, FL, USA

[2] International Center for Tropical Botany, Florida International University, Miami, FL, USA

[3] Department of Botany and Plant Sciences, University of California Riverside, Riverside, CA, USA

[4] Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, South Africa

[5] Department of Microbiology and Plant Pathology, University of California Riverside, Riverside, CA, USA

[6] Tropical Research and Education Center, Horticultural Sciences, University of Florida, Homestead, FL, USA

## Introduction

Avocado (*Persea americana*) is an economically important tropical and subtropical tree fruit crop. In 2014, worldwide production of avocado was 5.03 million metric tons (MMT), with over 1 MMT produced in Mexico (FAO 2016; Russell et al. 2011). In the USA, which ranks seventh in worldwide avocado production, California is the largest producer with Florida and Hawaii accounting for smaller percentages of the yearly crop. Avocado has three major horticultural groups supported by microsatellite genotyping data (Schnell et al. 2003): *Persea americana var. americana* Mill. 'West Indian' (WI); *var. guatemalensis* Williams, 'Guatemalan' (G); and *var. drymifolia* (Schlecht. & Cham.) Blake, 'Mexican' (M). Like many tree crops, avocado is propagated clonally through grafting to preserve commercially desirable varieties. Almost all the cultivated avocado acreage in California is the cultivar 'Hass', which is currently described as a M × G hybrid (Ashworth and Clegg 2003). Avocado production is the second largest agricultural industry in Florida after citrus and is

based almost exclusively in Miami-Dade County. Growers in southern Florida rely on West Indian (WI) and Guatemalan × West Indian (G × WI) hybrid varieties that are suited to subtropical climates and local soil conditions. Currently, the most widely grown Florida avocados are the varieties 'Simmonds' (WI), 'Donnie' (WI), 'Monroe' (G × WI), and 'Lula' (G × WI) on either open-pollinated 'Lula' or 'Waldin' (WI) rootstock (*personal communication*, Alan Flinn, Florida Avocado Administrative Committee).

Avocado germplasm collections around the world maintain living genetic diversity for this species and develop novel resources for genetics and breeding research. These collections represent avocado diversity both among the commercial cultivars and across the genus. The USDA-ARS Subtropical Horticulture Research Station (SHRS) in Miami, Florida, curates a large avocado germplasm collection with over 1090 trees representing 256 unique cultivars; all avocado germplasm can be viewed and requested at the GRIN-Global website (www.ars-GRIN.gov). The Fairchild Tropical Botanic Garden avocado collection at Williams Farm (Homestead, FL) includes 100 unique cultivars representing diversity among avocado landraces. The University of California Riverside (UCR) maintains germplasm collections at two locations. Material at UC South Coast Research and Extension Center (SCREC) in Irvine includes the following: (a) twelve *Persea* spp. relatives, (b) over 200 cultivars consisting of heritage cultivars and material from previous scion breeding programs (GERMCA), and (c) over 149 unique individuals from the UCR rootstock germplasm collection including hybrids, rootstock advanced selections, and international material imported from several countries including Chile, Guatemala, Mexico, South Africa, Australia, and Israel. The second UCR collection at the Agricultural Operations Station (AgOPs) in Riverside maintains the following: (a) a scion collection at UCR (does not duplicate collection at SCREC) and (b) 10 open-pollinated rootstock breeding blocks.

Avocado breeding requires ~ 15 years before the release of a new cultivar due to the prolonged juvenile period of the species. Most breeding programs in avocado are more accurately represented as selection programs, as hand pollination of maternal parents with specific paternal parents to create full-sib populations is rare. Identifying self-pollinated individuals at the seedling stage in a breeding program is an important screening step before further resources are expended on the individual. In breeding programs, named cultivars are considered genetically identical and fruit can be collected from all clones for selection of progeny with improved phenotypes. A simple molecular marker platform would allow breeders to genotype all clones of a cultivar and remove any mislabeled individuals from further breeding to prevent confounding of phenotypic evaluation.

Avocado is diploid, with a haploid genome size approximately $8.83 \times 10^8$ bp (Arumuganathan and Earle 1991) and a haploid chromosome number of 12 (Darlington and Wylie 1945). To improve the efficiency of tree breeding, thousands of genetic markers distributed throughout the genome are needed. Single-nucleotide polymorphisms (SNPs) are suitable for this purpose. SNP genotype data is easy to collect in large amounts due to frequency of SNPs and high-throughput SNP assay platforms. SNP markers can be easily designed from available transcriptome or genome assemblies. Sufficient DNA from a single leaf of a plant can be obtained to allow genotyping with hundreds of SNP markers (Barabaschi et al. 2016; Mammadov et al. 2012). Previous studies have developed molecular markers for avocado including SNPs from targeted resequencing (Chen et al. 2008), and microsatellite markers (Alcaraz and Hormaza 2007; Ashworth et al. 2004; Borrone et al. 2007; Ge et al. 2019; Gross-German and Viruel, 2013; Lavi et al. 1994; Mhameed et al. 1997; Sharon et al. 1997). These markers have been combined with phenotypic data including heritability of nutritional components (Calderon-Vazquez et al. 2013) and have been used to identify marker trait associations in avocado (Mhameed et al. 1995; Sharon et al. 1998). A recent study identified > 250,000 polymorphic SNPs from 22 avocado individuals (Ge et al. 2019). Genome-wide SNPs are required to construct a high-density linkage map, identify marker-trait associations to implement MAS, and identify suitable parents from germplasm collections for breeding purposes. Recently, we have developed 5050 SNPs to create a high-density linkage map (Kuhn et al. 2019b). In this study, we selected a subset of 384 SNP markers that are evenly distributed across the 12 linkage groups to genotype avocado germplasm, hybrids, and selections from open-pollinated maternal trees to create an unambiguous dataset for use by avocado breeders worldwide (Borrone et al. 2009; Kuhn et al. 2019b). The SNP markers used in this study were originally developed from the parents ('Simmonds', 'Tonnage', 'Hass', and 'Bacon') of our mapping populations ('Simmonds' × 'Tonnage' and reciprocal, 'Hass' × 'Bacon' and reciprocal) (Olano et al. 2007). The markers were identified from RNA-Seq transcriptome data from each parent and calling SNPs based on comparison to a 'Hass' consensus transcriptome.

The objective of this study was to create an unambiguous database of avocado genotypes by SNP genotyping of the greatest number of the individuals possible in the available avocado germplasm collections, rootstock breeding programs, and commercial clonal material in the USA and South Africa. This database will allow the estimation of genetic diversity in germplasm collections, identification of mislabeled/misidentified individuals, identity of landraces, distinction of self-pollinated progeny from hybrid (outcrossed) progeny of a known maternal parent, and potential paternal parents of progeny from known maternal parents. Breeders will be able to identify useful subsets of SNP markers specific to the particular individuals or cultivars used in their breeding programs.

## Materials and methods

### Germplasm individuals

Germplasm sources of all 1524 avocado individuals that were genotyped are described in Table 1. Genotypes were kept anonymous if requested by the collaborator. Seven different avocado species (*P. americana*, *P. palustris*, *P. caerulea*, *P. cinarescens*, *P. donnell-smithii*, *P. indica*, and *P. schiedeana*) were represented by 15 individuals. Hybrids and selections from two rootstock breeding programs (the University of California Riverside and the Westfalia Technological Services), individuals of commercially produced clonal rootstocks of 'Toro Canyon', and a large collection of open-pollinated half-siblings from known maternal parents were also genotyped.

### Isolation of DNA

DNA was isolated as previously described (Kuhn et al. 2017). Briefly, 3-mm leaf disks totaling approximately 50 mg per sample were punched from leaves, disrupted by shaking with 1/52″ metallic 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 were adjusted to a concentration of 10–20 ng/μl using a Hamilton liquid handling robot.

### SNP genotyping of germplasm individuals

Each avocado individual was genotyped with 384 SNP markers designed as assays to be run on the Fluidigm EP1™ system with the 96.96 IFC (Fluidigm, San Francisco, CA, USA). In addition, 435 individuals with prefixes GERM or GERMCA were previously genotyped with 5050 SNPs that included the 384 used in this study on an Illumina SNP chip (Illumina, San Diego, CA, USA) (see Table 1), and the data for the 384 SNP markers was added to the Fluidigm SNP marker dataset. The sequences of the 384 SNP assays are in Supplemental Table 1 with the associated linkage group (LG), map position in centimorgans (cM), and annotation data provided 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 (available by request) for analysis.

### Analysis of genotype data

#### Data encoding

Genotype data was encoded in four categories as homozygous allele 1 (1), homozygous allele 2 (2), heterozygous (3), or missing data (0) rather than nucleotide data to allow an unbiased analysis without genetic assumptions as to the relations of the individuals.

**Table 1** Sources of germplasm and research individuals genotyped in this study. The information in this table was provided by the curators of the various collections

| Population | Source | Location | No. of individuals sampled[a] | No. of individuals in dataset[b] |
|---|---|---|---|---|
| Germplasm | SHRS ARS USDA | Miami, FL | 360 | 355 |
| Open-pollinated seedlings | SHRS ARS USDA | USHRL ARS Ft. Pierce, FL | 138 | 133 |
| Controls | SHRS ARS USDA | Miami, FL | 64 | 49 |
| Germplasm | University of California, Riverside | SCREC (Irvine, CA) and AgOps (Riverside, CA) | 255 | 241 |
| Germplasm | PBARC ARS USDA | Hilo, HI | 51 | 51 |
| Germplasm | Williams Farm, FTBG | Homestead, FL | 100 | 99 |
| Clonal rootstock | TREC | Homestead, FL | 32 | 32 |
| Germplasm | Westfalia Technological Services | Tzaneen, Limpopo, SA | 27 | 20 |
| Germplasm and research individuals | Anonymous | | 516 | 481 |
| | | | Total: 1524 | Total: 1461 |

[a] Total individuals genotyped using 384 SNPs

[b] Number of individuals with less than 5% missing data. Total 377 SNPs for 1461 individuals

*SHRS* Subtropical Horticultural Research Station, *PBARC* Pacific Basin Agricultural Research Center, *ARS* Agricultural Research Service, *USDA* United States Department of Agriculture, *FTBG* Fairchild Tropical Botanic Garden, *TREC* Tropical Research and Education Center, *USHRL* US Horticultural Research Laboratory, *SCREC* South Coast Research and Extension Center, *AgOps* Agricultural Operations

### Calculating pairwise distances

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

$$\text{distance}\ (x,y) = \frac{\sum_{i=1}^{377}\text{comp}(x_i, y_i)}{377-(\text{md}(x)+\text{md}(y))}$$

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), or 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 377 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 377 markers for 1461 individuals.

### Affinity propagation and silhouette analysis

Affinity propagation analysis of the 1461 individuals and 377 SNP dataset was performed as previously reported (Bryant et al. 2013; Kuhn et al. 2019a; Pedregosa et al. 2011; Pers et al. 2015; Rousseeuw 1987). Silhouette analysis of the affinity groups generated was performed also as previously reported (Kuhn et al. 2019a).

### Calculating landrace identical and impossible genotypes

The numbers of identical and "impossible" genotypes within the landrace subsets and between the landrace subsets were calculated. Identical genotypes within the landrace were calculated by first designating an individual believed to be the best representative of the landrace based on the historical record as the landrace reference individual. The landrace reference individual's genotype was used to sort the entire dataset by row (0, 1, 2, 3). The range of 1s and 2s (homozygous genotypes) was recorded for that individual. The COUNTIF function in Excel was used to count the number of identical genotypes (1-1, 2-2) or impossible genotypes (1-2, 2-1) between an individual in the landrace and the landrace reference individuals. The number of identical and impossible genotypes of each individual in the landrace subset when compared with the landrace reference individual was then averaged over all the individuals in the landrace. This same procedure was used to determine the genotype differences between the landrace reference individual and individuals in the other two landraces. The numbers of identical genotypes and impossible genotypes were averaged over all the individuals in the landrace. The calculations of identical genotypes and impossible genotypes between landraces are based on sorting the entire dataset on three different landrace reference individuals. Therefore, the values for $G$ versus $M$ and $M$ versus $G$ are not symmetrical.

For each landrace subset, the numbers of homozygous allele 1 genotypes, homozygous allele 2 genotypes, and heterozygous genotypes were calculated using the COUNTIF function in Excel without a comparison with the landrace reference individual. These numbers were then averaged over all individuals in the landrace.

To calculate differences between 'Hass' or 'Fuerte' and the landrace subsets, the data were first sorted based on 'Hass' or 'Fuerte'; identical genotypes and impossible genotypes calculated as above and the numbers of identical and impossible genotypes for each individual in the landrace were averaged.

## 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, gray, 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, and 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 to 3), number of 1s, smallest to largest, by average heterozygous allele calls, and by marker position in the genome, among other methods. Rows (individuals) were sorted by groups, names, homozygous and heterozygous SNP calls, among other methods. Colored data file with affinity propagation groups, silhouette scores, and related data are in Supplemental Table 2.

# Results

## Estimation of genetic diversity

### Genotype statistics

Biallelic SNP markers (384) were used to genotype 1524 germplasm individuals. Data were curated to remove SNP 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 377 markers for 1461 individuals (Supplemental Table 2). Missing data per

individual varied from 0 to 20 markers of the 377 individuals, and missing data for markers varied from 0 to 48 for the 1461 individuals. The average of missing data of all markers was 4.8%. Average missing data from individuals was 1.2 from 377 markers or 0.3%.

Some individuals had been genotyped multiple times and allowed the calculation of machine genotyping error. There were 11 samples of 'Tonnage' (4147 genotypes) with only 21 differences among them for a genotyping error of 0.5%. Eight 'Simmonds' samples and a mislabeled 'Lula' sample had no differences in 3393 genotypes for a genotyping error of 0%. Eleven 'Hass' samples, five 'Carmen' samples, 3 HX670 samples, two 'Andes' samples, and 2 unknown samples (anonymous) showed 15 differences out of a possible 8671 genotypes with 11 of differences coming from two SNP markers for a genotyping error of 0.2%. In this group, the reference (correct) genotype was 'Hass'; thus, 43% of the 23 samples were mislabeled or differed in genotype at less than 1%. There were no impossible differences at any marker, i.e., an individual that was homozygous allele 1 and an individual that was homozygous allele 2 at the same locus. Of 32 commercially supplied clones of 'Toro Canyon' received from the University of Florida's Tropical Research and Education Center (Homestead, FL), we observed three differences in 12,064 genotypes for a machine error rate of 0.02%. Interestingly, three individuals from the UCR collection labeled 'Toro Canyon' showed no differences over 377 SNP markers among themselves, but differed from the commercially available 'Toro Canyon' at 17 loci (4.52%). At all 17 loci, the UCR 'Toro Canyon' were scored as homozygous and the commercially available 'Toro Canyon' were scored as heterozygous. Thus, although the most conservative estimation of machine genotyping error (0.3%) would usually be used to determine if two individuals could be considered identical, the 'Toro Canyon' results suggest that individuals that are 96% identical may also be considered identical.

Heterozygous allele calls for individuals ranged from 0.8% (3/374, 3 missing data) for Anonymous091 to 85.7% (323/377) for 'Tonnage'. Heterozygous allele calls for other *Persea* species ranged from ~9 to 15%. Heterozygous allele calls of two individuals of interest were 36.6% for 'Hass' and 2.3% for VC75. Heterozygous allele calls of markers ranged from 6.5% (95/1459) for SHRSPaS006061 (map position LG10 55.5cM) to 54.1% (789/1458) for SHRSPaS002718 (map position LG6, 116.3cM). Average allele frequency over all markers for allele 1 was 49.9% and allele 2 was 49.8%. Allele frequency ranged from allele 1:allele 2 25:75 for Mi_0358 to 84:16 for SSKP077C2_A650G.

### Affinity propagation

Affinity propagation generated 155 groups of the 1461 individuals. Groups varied in size from 2 to 56 individuals, with most of the groups having between 2 and 11 individuals (Fig. 1). In an affinity group, if two or more individuals 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 individuals believed to be identical by naming convention. For example, in group 1 (Supplemental Table 2), 'Alicia Cordero' (row 21), 'Booth 8' (row 22), 'Nabal' seedling (row 29), and 'Sharwil' (row 30) all have the same 0.15 silhouette score. By hiding rows 23–28, all four individuals can easily be seen to be genotypically identical by scrolling from left to right.

In addition, to identify a subset of markers that can be used to distinguish a single accession from all other non-identical individuals, Supplemental Table 2 can be sorted by the row with the accession of interest. The COUNTIF function in Excel can then be used to calculate the ranges of 1s and 2s to determine the number of mismatches with all other individuals in the database, giving a score of 1 for a 1:2 or 2:1 mismatch. An example with a reduced dataset is given in Fig. 2 or can be generated by sorting data in Supplemental Table 2.

### Silhouette analysis

After affinity propagation analysis, a silhouette analysis determined the quality of the membership of a particular accession in a group (Supplementary Table 2). Silhouette scores near or below 0 indicate weak evidence for membership in the group, but no evidence for membership in another group or that a new group would be formed to include this individual. Some groups have low or negative silhouette scores for all members of the group. This indicates that the group is broadly dispersed in the dataspace and that there is no concentrated cluster of individuals in the group.

## 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 1, there are four individuals that have the same silhouette score and are genotypically identical at all 377 SNP loci, but have four different names (GERM_NABAL SEEDLING_WB1-12-04, GERM_SHARWIL_WA2-12-37, GERM_ALICIA CORDERO_WB4-05-03, GERM_BOOTH 8 (replant)_WB4-09-03). As individuals named 'Nabal' seedling and 'Sharwil' occur in other groups (group 2 for 'Nabal' seedling and groups 53 and 98 for 'Sharwil') and an individual named 'Alicia' (group 121) which may be 'Alicia Cordero',

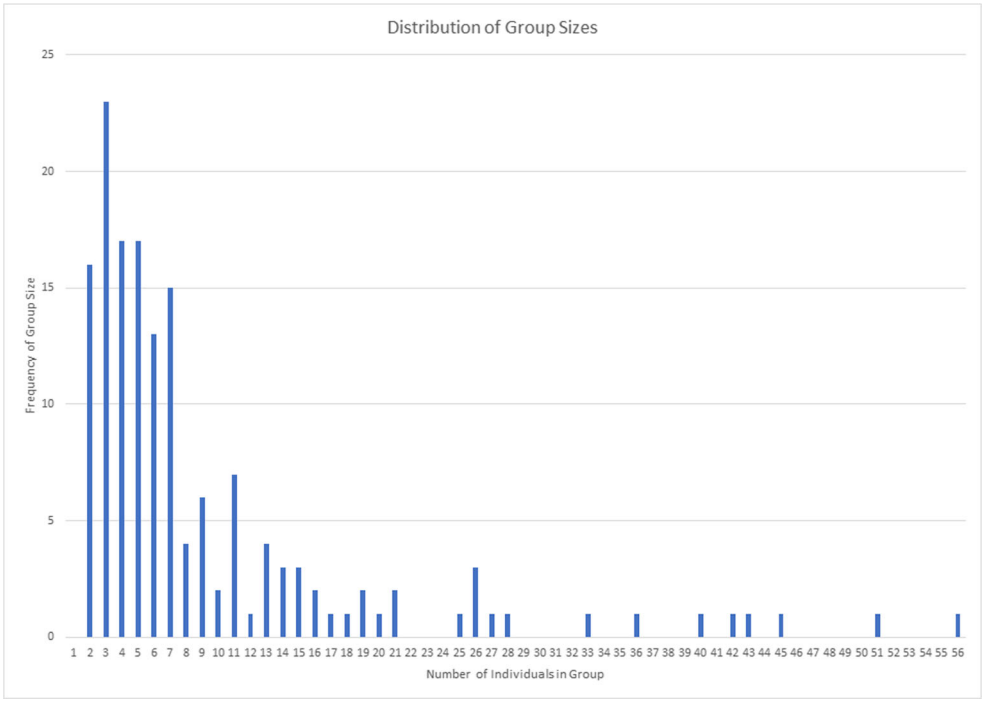


**Fig. 1** Distribution of affinity group size. *X*-axis is the number of individuals in an affinity group. *Y*-axis is the number of groups containing that number of individuals

it is likely that 'Booth 8' is the correct genotype and the other individuals are mislabeled.

Further evidence of mislabeling in germplasm collections comes from identifying groups where two or more differently named individuals are genotypically identical. Only groups that had at least two individuals with the same highest silhouette scores were considered, with the assumption being that these individuals were genetically identical. An estimation of mislabeling for the SHRS germplasm collection (prefix GERM_) was made by identifying groups where two or more SHRS individuals were present and determining if individuals with identical genotypes had different names. This condition held for 23 of the 155 groups and 38 examples of mislabeling were identified. As there are 294 SHRS germplasm individuals, this would represent a minimum of 13% mislabeling in our collection. This is an underestimate, as many groups had SHRS individuals that were genotypically identical to individuals with different names in other germplasm collections. Similar amounts of mislabeling were found in the UCR SCREC (prefix GERMCA).

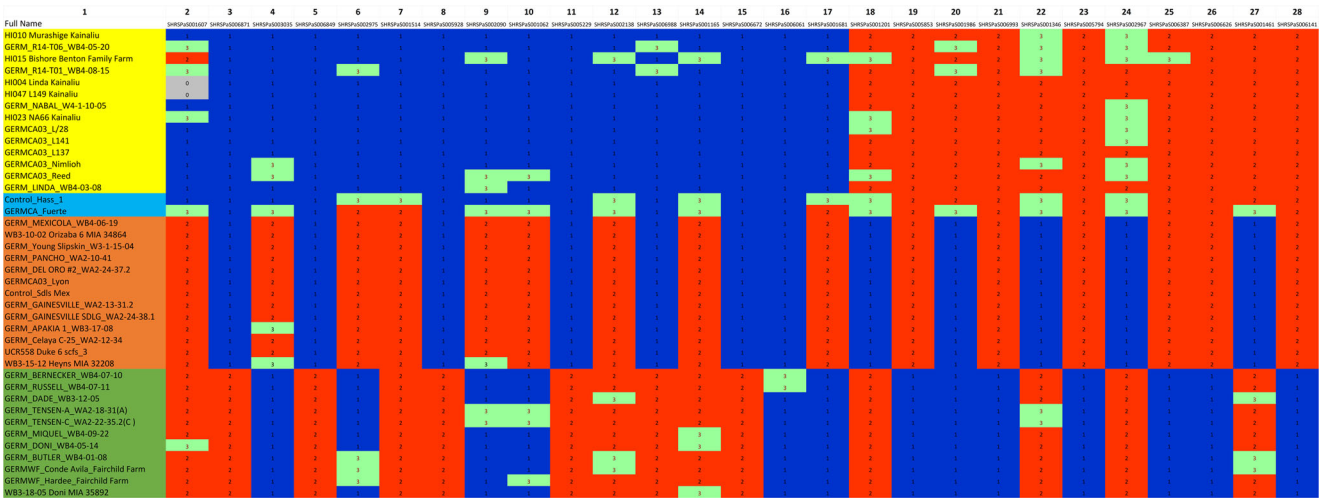


**Fig. 2** Examples of SNP markers that distinguish between landraces. The figure was created from data in Supplemental Table 3. In column 1, Guatemalan cultivars (yellow), 'Hass' and 'Fuerte' genotypes (blue), Mexican cultivars (orange), and West Indian cultivars (green). SNP genotype coloring scheme (columns 2–28), homozygous allele 1 blue, homozygous allele 2 orange, heterozygous light green

**Table 2** Average number of identical (1-1, 2-2) and impossible (1-2, 2-1) genotypes within and between landrace subsets, calculated as described in "Materials and methods." Guatemalan (G), Mexican (M), and West Indian (WI) landraces are shown

| | G identical | G impossible | M identical | M impossible | WI identical | WI impossible |
|---|---|---|---|---|---|---|
| Guatemalan (G) | 228.90 | 14.93 | 178.33 | 89.87 | 41.77 | 183.97 |
| Mexican (M) | 173.57 | 99.57 | 325.35 | 4.22 | 88.35 | 181.65 |
| West Indian (WI) | 45.65 | 191.81 | 92.81 | 187.13 | 233.90 | 16.35 |

## Species

In the dataset, there are 15 individuals with names of *Persea* species other than *P. americana*. All individuals have less than 5% missing data indicating that the 377 SNP markers can be used for other species in the genus *Persea*. Of these, *P. nubigena* and *P. floccosa* are identified (http://www.theplantlist.org/browse/A/Lauraceae/Persea/) as *P. nubigena var. guatemalensis*, a synonym for *P. americana var. guatemalensis*, and *P. floccosa* is also listed as *P. americana var. floccosa*. This is borne out by their appearance in group 95 and group 56 respectively. *Persea skutchii* is not recognized as a *Persea* species. *Persea schiedeana* in the SHRS germplasm collection is clearly misnamed and appears in group 65, identical with 'Duke'. The remaining named *Persea* species (*P. palustris*, *P. donnell-smithii*, *P. caerulea*, *P. cinarescens*, and *P. indica*) are all found in group 145 (as is *P. skutchii*). There are four individuals of *P. palustris* and three individuals of *P. caerulea*. The amount of variation within these species is approximately the same as the interspecific variation for all individuals in group 145. The SNPs developed for *P. americana* identify heterozygous alleles in the other species ranging from 7% in *P. palustris* to 15% in *P. cinarescens*.

## Landraces

Avocado has three landraces including *Persea americana var. americana* (West Indian), *var. drymifolia* (Mexican), and *var. guatemalensis* (Guatemalan). Using the http://www.ucavo.ucr.edu/ website for information on avocado varieties as well as GRIN-Global (https://www.ars-grin.gov/) passport data, three subsets representing "pure" individuals of the landraces were defined (Supplemental Table 3). The cultivar that was either the earliest in the historical record to be designated as a particular landrace or a cultivar that was imported as budwood rather than seed was used as the type specimen of the landrace and was chosen for sorting the subsets. The numbers of identical genotypes and impossible genotypes within the landrace subsets and between the landrace subsets were then calculated and averaged over all individuals in the landrace as described in "Materials and methods" and shown in Table 2, as well as the average homozygous and heterozygous SNP calls (Table 3). Visually scanning the genotype data for the sorted subsets showed that markers to distinguish each landrace could be developed (Fig. 2). Markers specific for Guatemalan landrace are seen in columns 2, 7, 12, 14, and 20. Markers specific for Mexican landrace are seen in columns 4, 6, 9, 10, 17, 18, 22, 24, and 27. Markers specific for West Indian landrace are seen in columns 3, 5, 8, 11, 13, 15, 19, 21, 23, 25, 26, and 28. A marker shared by all landraces is shown in column 16 (Fig. 2). Using these markers, hybrids of the landraces can be identified. As examples for the application of the landrace subsets, 'Hass' and 'Fuerte' were added to the subsets and landrace affiliation as well as possible parents of each were estimated (Supplemental Table 3). 'Hass' was most closely allied to the Guatemalan landrace and 'Fuerte' to the Mexican landrace (Table 4). For 'Hass', likely parents were 'Fuerte', 'Julia' (NA 526), and NA 251 (both 'Nabal' seedlings) based on genotype data alone. For 'Fuerte', likely parents were 'Mexicola' and NN63 ('Nabal' seedling) based on genotype data alone. There was no data to support 'Hass' or 'Fuerte' as being self-pollinated progeny of any individuals in the dataset. Based on the average identical genotypes and differences of the landrace subsets, the Mexican and Guatemalan landraces share more common genotypes and are more similar to each other than to the West Indian landrace individuals (Table 2).

## Detection of self-pollination and estimation of self-compatibility

Progeny with a known maternal parent can be screened to determine if they are self-pollinated. In a self-pollinated

**Table 3** Average number of homozygous genotypes and heterozygous genotypes for landrace subsets. Homozygous genotypes for alleles 1 and 2 and heterozygous genotypes were counted for each individual in a landrace subset and averaged as described in "Materials and methods"

| | Homozygous allele 1 | Homozygous allele 2 | Heterozygous |
|---|---|---|---|
| Guatemalan | 145.70 | 140.83 | 89.87 |
| Mexican | 174.74 | 171.17 | 30.87 |
| West Indian | 145.68 | 154.32 | 76.52 |

**Table 4** Average number of identical (1-1, 2-2) genotypes and impossible (1-2, 2-1) genotypes when all individuals in the Guatemalan (G), Mexican (M), and West Indian (WI) landraces were compared with either 'Hass' or 'Fuerte' as the reference individual. Landrace subset genotype data was sorted on either 'Hass' or 'Fuerte', and identical and impossible genotypes were counted for each individual in the landrace subset and averaged across all individuals in the subset as described in "Materials and methods"

| | G identical | G impossible | M identical | M impossible | WI identical | WI impossible |
|---|---|---|---|---|---|---|
| 'Hass' | 181.73 | 14.33 | 166.17 | 52.38 | 33.68 | 158.13 |
| 'Fuerte' | 154.60 | 22.23 | 195.35 | 7.30 | 35.58 | 143.65 |

individual, all the loci homozygous in the maternal parent will be homozygous for the same allele in the hybrid. For example, group 14 includes the cultivar 'Suardia' (GERM_Suardia_WA2-22-40) and seven open-pollinated seedlings of 'Suardia' out of 11 'Suardia' seedlings in the dataset. If the markers are sorted by the genotype of 'Suardia' (Suardia sdlg), scanning through the genotype data demonstrates immediately that three of the seven seedlings (Pa9853, Pa10248, and Pa9939) are self-pollinated seedlings of 'Suardia', sharing all the homozygous genotypes with the maternal parent. As predicted from self-pollination, these seedlings also have a greater number of homozygote genotypes and fewer heterozygote genotypes than the maternal parent. Thus, an estimate of self-compatibility for 'Suardia' would be 27% (3/11). Group 23 includes 'Melendez' and four 'Melendez' seedlings out of five in the dataset. Sorting the marker genotypes on the parent 'Melendez', visual inspection identifies that none of the four progeny is self-pollinated, as they are not homozygous at all of the loci where 'Melendez' is homozygous. An estimate of self-compatibility for 'Melendez' would be 0%. Finally, group 24 includes 'Bernecker' and four 'Bernecker' seedlings out of six in the dataset. Sorting on 'Bernecker', visual inspection identifies that all four 'Bernecker' seedlings are the result of self-pollination, giving an estimate of self-compatibility of 67% (4/6).

## Discussion

Our goal has been to develop an unambiguous genotype database from a large group of individuals and a large number of SNP markers to aid in the maintenance and application of germplasm collections, to preserve genetic diversity, provide a dataset for breeders that will allow correct identification of individuals in their breeding programs, and allow curators and breeders to estimate the genetic diversity available in their collections and breeding programs. SNP markers were selected as they are platform independent and databases of genotypes can be shared globally. We initially developed avocado SNP markers to produce a genetic map and have used a subset of the markers to genotype all available avocado germplasm individuals as well as research individuals from breeding programs (Kuhn et al. 2019b). The advantage of having a larger set of markers is that it provides users with the ability to select subsets of markers that distinguish the particular cultivars they are interested in.

### SNP markers and analysis of large genotype datasets

There were 377 markers (98%) that showed less than 5% missing data across 1461ndividuals. We successfully used the markers to distinguish individuals from the species down to the half-sib level, suggesting that these markers are an accurate estimate of genetic diversity in the genus *Persea*. Individuals from other genera in the *Lauraceae* were not available to us and therefore have not yet been genotyped. However, due to the high level of identity among the individuals labeled as other species and individuals in the Mexican landrace, these markers are not suitable to reliably identify other species.

The use of SNP markers for estimating genetic diversity at the population and species levels poses several challenges including using the data to determine genetic relatedness and diversity at any level from half-sibs to landraces or across species. SNP markers are biallelic. They can be homozygous allele 1, homozygous allele 2, or heterozygous. This does two things to the analysis. First, it increases the identity by chance dramatically as there are only three possible states. Second, it basically erases the importance of "private alleles" which drive most of the genetic diversity estimation programs currently used (STRUCTURE, Mr. Bayes, etc.).

The analytical method presented here makes no genetic assumptions about the data to generate affinity propagation groups. The silhouette scores that measure quality of membership in a group have allowed easier identification of mislabeling/misidentification, potential self-pollinated individuals in the hybrids, and even, in some cases, potential paternal parents for open-pollinated progeny.

In some cases, silhouette scores for all members of a group are close to 0 or even negative. However, when viewing the group in Excel with conditional formatting of the cells with contrasting colors for the genotypes, it is immediately obvious to the most casual observer that the group members are related. A potential explanation of this contradiction is that unlike

in groups where there are several individuals that are identical or nearly so that provide a strong center to the cluster in the *n*-dimensional data space, individuals in a group with all low silhouette scores are evenly dispersed through that portion of data space and so each is an equally poor center for the group. Nonetheless, these individuals do make up a group and are not just thrown together in some version of long branch attraction. In some ways, the large amount of SNP data available and the sensitivity of the affinity propagation method have identified a group that encompasses a large, but distinct, volume of the data space. It can be easily demonstrated that by judicious selection of the SNP markers for any affinity group, the majority of individuals in the group can be rendered nearly identical, which would raise their silhouette scores but provide no more information than is present in the full dataset. 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 disperse in the volume of data space they inhabit.

## Anonymous individuals

Several contributors of germplasm individuals or breeding selections shared proprietary material for this study, and these were included to strengthen the study results. Affinity propagation analysis works better with more individuals as it can be thought of as filling in some of the less populated portions of the *n*-dimensional dataspace and perhaps providing a new individual to be the nucleus of a new group. The dataset presented here is a snapshot of the process as we continue to add individuals to the dataset. After addition of new genotype data, a new similarity matrix is generated using all the data, a new affinity propagation analysis is done, and new silhouette scores were calculated. As the dataset grows, the number of affinity groups does not increase in a linear fashion which suggests that we have identified the majority of the genetic diversity in the individuals genotyped.

For end users of the dataset in Supplemental Table 2, it is not necessary to redo the affinity analysis if only a simple comparison of a few individuals is desired. Such a comparison can be done by sorting the dataset based on the accession of interest using the COUNTIF function in Excel and identifying which group is most closely allied to the accession. As described here, the end user can use the dataset to determine mislabeling, identify SNP markers that could distinguish the accession of interest, test hypotheses about self-pollination, etc. Further analyses suggest that there is no need to genotype with the full set of 384 SNPs. Subsets of informative SNPs can be used and compared with an edited dataset that only contains genotypes for those SNP loci.

## Intuitive visual analysis of large genotype datasets

The sheer size of the dataset, 377 SNP genotypes for 1461 individuals or 550,797 genotypes, has made analysis challenging. By recoding the dataset and coloring the four states (homozygous allele 1, homozygous allele 2, heterozygous, and missing data) using conditional formatting in Excel and then reducing the visual size of the genotype data to its minimum (10%), it became possible to get a visual impression of the groups, the relatedness of the members of the groups, and the amount and type of homozygous allele calls in groups. Using simple formulas to calculate the number of each genotype state in an individual or for a marker across all individuals followed by sorting also allows a straightforward visual impression of which factors are important in distinguishing groups and members within groups. We have only presented some of the possible ways to sort the data and encourage novel approaches using familiar tools in Excel that do not require competence in other programming languages such as Perl, R, or Python.

## Estimating genetic diversity in the germplasm dataset

Affinity propagation generated 155 groups which help in curating, maintaining, and selecting individuals for backing up the avocado germplasm collections. The 155 groups represent genetic diversity from the species level down to the half-sib level, and, overall, the groups reflect this diversity correctly. Species are grouped either by themselves or with other species. Half-sibs are frequently grouped with either the maternal or paternal parent. For example, for our SHRS germplasm collection, the 286 individuals appear in 90 groups. A single individual from each of the 90 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. This number may be smaller, based on mislabeling/misidentification of individuals. Thus, all the genetic diversity encompassed by the current germplasm collection could be maintained in a collection approximately one third the size of the current one and with greater confidence of the identity of the individuals based on comparison with genotypes of individuals 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 and in prioritizing grafting of trees for regenerating the collection.

## Estimating mislabeling/misidentification in the germplasm collection

Determining mislabeling in a germplasm collection is difficult if there are only single individuals for each named cultivar. In our genotyping of multiple germplasm collections, we have increased the chance to at least identify the correctly named genotype simply by the majority genotype for that named

cultivar. However, the possibility that the named cultivar in multiple germplasm collections came from a single source also makes determination by simple majority less than certain. Mislabeling can occur at any level, upon addition into the collection, vegetative propagation, and improper identification in the field for application of labels and in the laboratory prior to genotyping.

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 in collections. Even with valid markers, a common problem is, for example, having two identically labeled clones with different genotypes. They cannot both be correct, and, in fact, both may be mislabeled. 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 that other germplasm collections may have acquired one or the other of the clones, or indeed even a third genotype for the identically labeled clone.

We have taken a rather limited approach to estimate mislabeling using molecular markers by using only the SHRS germplasm collection and groups where two or more individuals with identical genotypes, but different names appeared in the same group. Without regard to which clone was correct, it clearly indicated mislabeling as at least one of them was incorrectly labeled. For our germplasm collection, this method estimated ~ 13% mislabeling, which is an underestimate and also the lower limit of mislabeling for the collection. Without regard to the amount of mislabeling, we will act to remove all mislabeled trees from our collection at SHRS to prevent distribution of mislabeled material through the GRIN-Global system. If identification of the correct accession is not possible, both individuals in GRIN-Global will have a note stating that they are genotypically identical and cannot be confidently identified as a specific cultivar.

## Species and landraces

The SNP markers were developed from neutral nucleotide variations in coding regions in *Persea americana*. All 377 **were** genotyped consistently in all the other species in the dataset. We observed a high level of identity among the individuals representing other *Persea* species. Although from 9 to 15% nucleotide variation was found among individuals in the other species, the variable markers could not reliably distinguish one species from another. This may have been due to the low number of individuals genotyped and the relatively high amount of mislabeling/misidentification.

Distinguishing landraces was easily achieved as there were many markers that were landrace specific. Individuals were identified as belonging to a landrace from the earliest historical record, and these subsets were used to identify landrace specific markers. A small number of markers can accurately identify an individual as being in a particular landrace or as a hybrid of two landraces.

Based on the historical record, 'Fuerte' and 'Nabal' seedlings were present in California prior to the selection of 'Hass'. 'Fuerte' itself is purported to be a Guatemalan × Mexican hybrid brought into California in 1910 from Atlixco, Mexico. Thus, it is possible that 'Mexicola' or a similar "pure" Mexican individual and a Guatemalan × Guatemalan hybrid similar to a 'Nabal' seedling could have been the original parents of 'Fuerte'.

## Self-pollination and self-compatibility

Estimating self-compatibility becomes important in choosing maternal parents for selection of progeny with favorable phenotypic characters. For rootstock, screening of seedlings is done at an early stage and usually before genotyping. Thus, choosing a maternal parent with low self-pollination rates increases the chance of outcrossing and of progeny with improved vigor. The current dataset is robust and should give breeders the tools to analyze the identity of new selections, distinguish self-pollinated individuals at the seedling stage, and determine mislabeling of progeny/parental selections with a small number of SNP markers. The ability to estimate self-pollinating tendencies of potential maternal parents could improve the efficiency of creating hybrid populations with fewer selfed progeny. The results presented here could easily be expanded to other avocado breeding programs that rely on specific cultivars to meet their breeding program priorities.

## Inferring potential paternal parents of hybrids using silhouette scores

The affinity propagation analysis and silhouette scores are a novel approach to analyzing a large dataset of SNP genotypes from large numbers of individuals. Using coloring of genotypes has made interacting with the dataset much more intuitive and led to some interesting discoveries. In addition, these methods can be used in some circumstances as a quick means to identify potential paternal parents of hybrids. 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 of importance to gain outcrossed material for breeding programs as it may identify cultivars that are particularly likely to outcross. 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 them for phenotypic evaluation.

## Summary and conclusions

The generation of a large SNP genotype database of diverse germplasm individuals ranging from genera to half-sib hybrids provides a widely useful tool that avocado breeders and researchers worldwide can use as it is based on unambiguous sequence data that can be generated by any platform. Coupled with the novel grouping method of affinity propagation and the scoring of the quality of membership in a group by the silhouette analysis, we have shown that the genotyping data can be used successfully to estimate germplasm genetic diversity and mislabeling across the entire range of individuals. In addition, using a simple color coding of the genotypes, we have shown that identification of self-pollinated or outcrossed hybrids, estimation of self-compatibility for maternal parents, and identification of likely paternal parents for hybrids of known maternal parents can be done by visual inspection.

The selected 384 markers were sufficient to obtain an accurate genotyping of the material, identify offtypes and mislabeling in the collections, and provide genetic evidence to distinguish landraces and distinct individuals both within *P. americana* and for other *Persea* spp. This work creates a foundational database for breeding and germplasm curation in avocado that is extensible as new individual genotypes are added to the dataset. Subsequent submissions to the database will require construction of a new similarity matrix, affinity propagation analysis, and the assignment of silhouette scores. Increasing the size of the dataset, either with known or anonymous samples, increases the utility of the tool without altering its previous effectiveness or contradicting previous results.

**Funding information** This study was funded by Trust Fund Cooperative Agreement 58-6038-7-006 with the California Avocado Commission and USDA-ARS CRIS 58-6038-21000-022-00D.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.



## References

Alcaraz ML, Hormaza JI (2007) Molecular characterization and genetic diversity in an avocado collection of cultivars and local Spanish genotypes using SSRs. Hereditas 144:244–253. https://doi.org/10.1111/j.2007.0018-0661.02019.x

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Ashworth V, Clegg M (2003) Microsatellite markers in avocado (*Persea americana* Mill.): genealogical relationships among cultivated avocado genotypes. J Heredity 94:407–415

Ashworth VETM, Kobayashi MC, De La Cruz M, Clegg MT (2004) Microsatellite markers in avocado (*Persea americana* Mill.): development of dinucleotide and trinucleotide markers. Sci Hortic 101:255–267. https://doi.org/10.1016/j.scienta.2003.11.008

Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Vale G, Cattivelli L (2016) Next generation breeding. Plant Sci 242:3–13

Borrone JW, Schnell RJ, Violi HA, Ploetz RC (2007) Seventy microsatellite markers from *Persea americana* Miller (avocado) expressed sequence tags. Mol Ecol Notes 7:439–444. https://doi.org/10.1111/j.1471-8286.2006.01611.x

Borrone JW, Brown JS, Tondo CL, Mauro-Herrera M, Kuhn DN, Violi HA, Sautter RT, Schnell RJ (2009) An EST-SSR-based linkage map for *Persea americana* Mill. (avocado). Tree Genet Genomes 5:553–560. https://doi.org/10.1007/s11295-009-0208-y

Bryant C, Giovanello KS, Ibrahim JG, Chang J, Shen DG, Peterson BS, Zhu HT (2013) Mapping the genetic variation of regional brain volumes as explained by all common SNPs from the ADNI study. PLoS ONE 8. https://doi.org/10.1371/journal.pone.0071723

Calderon-Vazquez C, Durbin ML, Ashworth VETM, Tommasini L, Meyer KKT, Clegg MT (2013) Quantitative genetic analysis of three important nutritive traits in the fruit of avocado. J Am Hortic Sci 138:283–289

Chen H, Morrell PL, de la Cruz M, Clegg MT (2008) Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). J Hered 99:382–389. https://doi.org/10.1093/jhered/esn016

Darlington CD, Wylie AP (1945) Chromosome atlas of flowering plants. George Allen & Unwin LTD, London, p 16

FAO (2016) Global avocado production in 2013, by country. https://www.statista.com/statistics/577455/world-avocado-production/

Ge Y, Zhang T, Wu B, Tan L, Ma F, Zou M, Chen H, Pei J, Liu Y, Chen Z, Xu Z (2019) Genome-wide assessment of avocado germplasm determined from specific length amplified fragment sequencing and transcriptomes: population structure, genetic diversity, identification, and application of race-specific markers. Genes 10(3):215

Gross-German E, Viruel MA (2013) Molecular characterization of avocado germplasm with a new set of SSR and EST-SSR markers: genetic diversity, population structure, and identification of race-specific markers in a group of cultivated genotypes. Tree Genet Genomes 9(2):539–555

Kuhn DN, Bally ISE, Dillon NL, Innes D, Groh AM, Rahaman J, Ophir R, Cohen Y, Sherman A (2017) Genetic map of mango: a tool for mango breeding. Front Plant Sci 8:577

Kuhn DN, Dillon N, Bally I, Groh A, Rahaman J, Warschefsky M, Freeman B, Innes D, Chambers AH (2019a) Estimation of genetic diversity and relatedness in a mango germplasm collection using SNP markers and a simplified visual analysis method. Sci Hortic 252:156–168

Kuhn D, Livingstone D III, Richards J, Manosalva P, Van den Berg N, Chambers A (2019b) Application of genomic tools to avocado (*Persea americana*) breeding: SNP discovery for genotyping and germplasm characterization. Sci Hortic 246:1–11

Lavi U, Akkaya M, Bhagwat A, Lahav E, Cregan PB (1994) Methodology of generation and characteristics of simple sequence repeat DNA markers in avocado (*Persea americana* M.). Euphytica 80:171–177

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP markers and their impact on plant breeding. Int J Plant Genomics 2012:1–11

Mhameed S, Hillel J, Lahav E, Sharon D, Lavi U (1995) Genetic association between DNA fingerprint fragments and loci controlling

agriculturally important traits in avocado (*Persea americana* Mill). Euphytica 84:81–87. https://doi.org/10.1007/Bf01677560

Mhameed S, Sharon D, Kaufman D, Lahav E, Hillel J, Degani C, Lavi U (1997) Genetic relationships within avocado (Persea americana Mill) cultivars and between Persea species. Theor Appl Genet 94(2):279–286

Olano C, Borrone J, Brown JS, Violi H, Ploetz R, Schnell R (2007) Development of mapping populations for avocado. In: Annu Meet Fla State Hort Soc, pp 26–29

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, Lui JC, Vedantam S, Gustafsson S, Esko T et al (2015) Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun 6:5890. https://doi.org/10.1038/ncomms6890

Rousseeuw PJ (1987) Silhouettes - a graphical aid to the interpretation and validation of cluster-analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Russell JR, Bayer M, Booth C, Cardle L, Hackett CA, Hedley PE, Jorgensen L, Morris JA, Brennan RM (2011) Identification, utilisation and mapping of novel transcriptome-based markers from blackcurrant (*Ribes nigrum*). BMC Plant Biol 11. https://doi.org/10.1186/1471-2229-11-147

Schnell RJ, Brown JS, Olano CT, Power EJ, Krol CA, Kuhn DN, Motamayor JC (2003) Evaluation of avocado germplasm using microsatellite markers. J Am Soc Hortic Sci 128:881–889

Sharon D, Cregan PB, Mhameed S, Kusharska M, Hillel J, Lahav E, Lavi U (1997) An integrated genetic linkage map of avocado. Theor Appl Genet 95:911–921. https://doi.org/10.1007/s001220050642

Sharon D, Hillel J, Mhameed S, Cregan PB, Lahav E, Lavi U (1998) Association between DNA markers and loci controlling avocado traits. J Am Soc Hortic Sci 123:1016–1022