Genetic diversity, population structure and clonal verification in South African avocado cultivars using single nucleotide polymorphism (SNP) markers

Raven Wienk [1,2,3], Marja Mostert-O'Neill[2,3], Nilwala Abeysekara[4], Patricia Manosalva[5], Barbie Freeman[6] and Noëlani van den Berg[1,2,3]

1 Hans Merensky Chair in Avocado Research, University of Pretoria, Pretoria, South Africa

2 Department of Biochemistry, Genetics and Microbiology, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa

3 Faculty of Natural and Agricultural Sciences, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, South Africa

4 Dominican University of California, San Rafael, CA, USA

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

6 Subtropical Horticulture Research Station, USDA-ARS, Miami, FL, USA

Corresponding author: Raven Wienk - raven.wienk@up.ac.za

ORCID Numbers:
Raven Wienk: 0000-0002-4453-5280
Marja Mostert-O'Neill: 0000-0002-6318-3508

ABSTRACT

Since the start of avocado cultivation in South Africa, superior rootstocks and fruit cultivars have been selected based on morphological traits, which is time-consuming and expensive. Technological advances, such as the development of a single nucleotide polymorphism (SNP) genotyping platform for avocado, may reduce these limitations. Therefore, the aim of this study was to implement molecular marker technologies for the validation of clonal material, verification of horticultural varieties and determining the genetic diversity and population structure of an avocado cultivar germplasm in South Africa. An avocado cultivar breeding population, containing 375 individuals, was genotyped using 384 SNP markers. Our affinity propagation analysis indicated a 10.74% mislabelling in the germplasm. The principal

component analysis and discriminate analysis of principal components suggested that the germplasm was admixed in relation to the three known avocado varieties, Guatemalan, Mexican, and West-Indian. Additionally, the ancestral origins were determined for 27 individuals with unknown ancestry. Furthermore, the population diversity was assessed and revealed moderate levels of differentiation in the germplasm, suggesting a high level of gene flow between the different populations. This research highlights the value of clonal verification and horticultural variety identification – for the reliable propagation of material with desired traits. The accurate propagation of material and clonal identity could aid avocado growers to link morphological characters and stress tolerance to accurate genetic backgrounds, which could improve the selection of avocados for current and future environmental stressors, especially as Africa is set to be significantly impacted by climate change.

Keywords: Avocado, SNPs, Population structure, Genetic admixture, Breeding

INTRODUCTION

Avocado (*Persea americana*) comes mainly from three 'varieties', *P. americana* var. *americana* Mill. ('West Indian'), *P. americana* var. *guatemalensis* Williams. ('Guatemalan') and *P. americana* var. *drymifolia* Blake. ('Mexican') (Lahav & Lavi, 2002, Wolstenholme, 2003). Intraspecies reproduction between varieties has led to extensive hybridisation with varying agronomical traits (Davis *et al.*, 1998, Ashworth & Clegg, 2003). The South African avocado industry relies heavily on superior rootstocks and cultivars, with important morphological traits such as improved fruit yield, better fruit quality, and resistance/tolerance to abiotic and biotic factors, which are usually hybrids (Popenoe & Williams, 1947). These hybrids typically show remarkable morphological similarities, making selection and verification of propagated material difficult (Popenoe & Williams, 1947). These morphological trait assessments, selection and development of new cultivars and rootstocks can extend over 20 years - which is a laborious, resource intensive and time-consuming process (Köhne, 2005, Schaffer *et al.*, 2013).

Advances in technology can now assess an individual on the genotypic level to genetically classify and verify horticultural varieties (Schnell *et al.*, 2003) using molecular markers (Karp *et al.*, 1997). Molecular markers are excellent for genomic and evolutionary studies, clonal verification, identifying cryptic relatedness among individuals, and identifying favourable genotypes linked to phenotypic performances in certain environmental conditions (Batley, 2015). Consequently, these technologies have the potential to advance and

improve genomic selection, by reducing the time and costs involved in phenotyping large numbers of individuals, which is vital to the avocado industry (Clegg, 2004).

The most current and popular molecular markers that have been used to study avocado are microsatellites and single nucleotide polymorphisms (SNPs). These markers have been used to investigate sequence nucleotide diversity (Chen *et al.*, 2008), population structure (Chen *et al.*, 2009, Ge *et al.*, 2019a, Ge *et al.*, 2019b, Juma *et al.*, 2020), horticultural variety assignment (Chen *et al.*, 2009), determine genetic diversity (Rubinstein *et al.*, 2019, Ge *et al.*, 2019b, Juma *et al.*, 2020), clarify phylogenetic relationships (Ge *et al.*, 2019a), provide clonal and cultivar verification (Kuhn *et al.*, 2019c) and create linkage maps (Kuhn *et al.*, 2019b).

No research has been conducted on the genetic diversity and population structure of avocados in South Africa. The aim of the present study was to address the lack of genetic diversity and population structure knowledge by sampling and SNP genotyping individuals from an avocado fruit cultivar population by using the 384 SNP platform developed by Kuhn *et al.* (2019c). An affinity propagation analysis (APA) (Frey & Dueck, 2007) was used for clonal verification and identification of mislabelled individuals. Furthermore, the population structure and genetic diversity were investigated using principal component analysis (PCA) and discriminant analysis of principal components (DAPC) (Jombart *et al.*, 2010). These results will be valuable in the establishment of molecular tools that can be used for the effective execution of conservation and breeding practices in the avocado industry.

MATERIALS AND METHODS

Biological material - Germplasm accessions

This study used an avocado breeding population from Tzaneen, Limpopo (South Africa) which was selected and provided by Allesbeste™. It consisted of 375 fruiting cultivar individuals, of which 108 individuals were genetically unique. As sample collection could be error prone, some trees were sampled in duplicate. Each accession had a unique "accession ID", thus, individuals with identical "accession IDs" that were sampled from different trees, were presumed to be genetically identical "clonal/clones". These were sampled to confirm clonal identity and determine the technical error rate. Whereas, individuals with identical "accession IDs" that were sampled from the same tree were classified as "duplicates", these were sampled to determine the machine error rate. Allesbeste™ provided proprietary material for this study and as such, all accessions have been anonymised. All individuals that were genotyped are summarized in **Supplementary File 1 (**10.25403/UPresearchdata.19145087**)**.

SNP genotyping

Ten green, fleshy leaves at intermediate expansion with minimal damage were sampled from each tree. DNA was isolated from the leaf material by the USDA-ARS using the method described by Kuhn *et al.* (2017). Each avocado individual was genotyped with 384 bi-allelic SNP markers run on the Fluidigm EP1™ system with the 96.96 IFC (Fluidigm, San Francisco, CA, USA), with 91 DNA samples and five controls, as previously described by Kuhn *et al.* (2019c). Samples were SNP genotyped by the USDA-ARS. Additional published SNP genotypic data (Kuhn *et al.*, 2019c) was incorporated for population structure analyses and horticultural variety verification, these individuals were from three horticultural varieties, believed to be ancestral to the South African germplasm. These individuals were labelled as "UCR", that included two Guatemalan (G), six Mexican (M), and four West-Indian (WI) individuals, provided by the University of California, Riverside, USA (Kuhn *et al.*, 2019c).

Affinity propagation analysis and visualisation of genotypic data

The data was reformatted to proceed with downstream processes using a custom Perl (Version 5.28.1) script to extract and reformat the genotype information into four categories, as previously described by (Kuhn *et al.*, 2019c). Markers and individuals with greater than 5% missing data were removed in a recursive fashion, as previously described by (Kuhn *et al.*, 2019c). Consequently, 61 individuals and eight markers were removed and excluded from further analysis. Therefore, 326 individuals (including the 12 references & 107 unique accession IDs) and 376 markers remained from the cultivar data. This dataset was named "APA Dataset", as seen in **Table 1**.

Custom Python distance and similarity scripts were used to generate pairwise distances (Python - Version 3.8.6), as described by Kuhn *et al.* (2019c). The similarity matrix was used to perform an Affinity Propagation Analysis (APA) that generated clusters and aided in the identification of mislabelled individuals and confirmation of clonal material (Frey & Dueck, 2007, Bodenhofer *et al.*, 2011, Pedregosa *et al.*, 2011, Kuhn *et al.*, 2019a). Additionally, individuals were assigned silhouette scores (Rousseeuw, 1987), as described by Kuhn *et al.* (2019a). Genotype statistics were obtained by the visualisation and sorting of the genotypic data by accession IDs, affinity groups, silhouette scores, and genotypic profiles in Microsoft Excel (2019), as described by Kuhn *et al.* (2019c).

Table 1 The number of individuals genotyped and the germplasm sources used in this study, including the published dataset – 12 references

	Population		
Germplasm	Cultivar Germplasm	Published Horticultural References [c]	
Source	Allesbeste™	University of California, Riverside	
Location	Tzaneen, Limpopo, South Africa	Various Locations	**Total**
Number of individuals sampled (Original Dataset)	375	12	387
Number of individuals retained for APA analysis [a] (APA Dataset)	314	12	326
Number of individuals retained for Population Analysis [a, b] (Population Analysis Dataset)	147	12	159

[a] Individuals with greater than 5% missing data were removed
[b] Clonal or duplicate individuals were removed
[c] Published horticultural reference individuals (Kuhn *et al.*, 2019c)

The number of SNP differences, machine genotyping error, and technical error were calculated for each "clonal" and "duplicate" set of individuals. The machine genotyping error was calculated using the "duplicate" individuals – individuals sampled from the same tree multiple times. The technical error was calculated using "clonal" individuals – identical "accession ID" individuals sampled from different trees. Mislabelled individuals were identified in two ways; firstly, individuals with identical "accession IDs", but had different genotypic SNP profiles beyond machine genotyping error, were classified as mislabelled type 1. Secondly, individuals with different "accession IDs", but had similar genotypic SNP profiles within machine genotyping error rate, were classified as mislabelled type 2. Mislabelled individuals were highlighted in red in **Supplementary File 1 (**10.25403/UPresearchdata.19145087**)**.

Phylogenetic analysis

The APA Dataset was used to perform a hierarchical cluster analysis to study the individuals with similar genetic characteristics and aid in the identification of mislabelled accessions. The dataset was aligned using MUSCLE (Edgar, 2004) and subsequently used to construct a condensed unweighted pair group method with arithmetic mean (UPGMA) tree (Sneath & Sokal, 1973) using the maximum composite likelihood method, with the confidence examined using bootstrap values calculated for 1000 replicates in MEGA X (Kumar *et al.*, 2018). The dendrogram was exported in the Newick format to be visualised and customized in Interactive tree of life (iTOL) v6 (Letunic & Bork, 2019).

Population structure analysis

After the identification of the mislabelled individuals through the APA and phylogenetic analysis, individuals with the least missing data for each "clone" and "duplicate" within machine genotyping error rate were retained. Consequently, 167 individuals were removed and excluded from further analysis, these were highlighted in yellow in **Supplementary File 1 (**10.25403/UPresearchdata.19145087**).** Thus, 159 cultivar individuals, of which 12 were published horticultural references, were retained and this second dataset was named "Population Analysis Dataset", as seen in **Table 1**, and was used to perform the principal components analysis (PCA), discriminant analysis of principal components (DAPC), structure and diversity analysis. Additionally, one non-polymorphic marker was detected during this analysis and removed – marker SHRSPaS006061, thus, 375 markers remained for the structure analysis. The reduced dataset was reformatted in Microsoft Excel (2019) into a four-bit binary code with A as (1), C as (2), G as (3), and T as (4).

The PCA (Patterson *et al.*, 2006, Reich *et al.*, 2008), DAPC, and allele composition analysis was performed using the Adegenet package (Jombart, 2008, Jombart *et al.*, 2010), whereas the genetic diversity was determined using the MMOD 1.3.3 package (Winter, 2012). These analyses were all performed in RStudio, version 1.3.1093 (RStudio Team, 2016) using R version 4.0.3 (R Development Core Team, 2020).

The PCA was performed to display the genetic relationships among individuals, genetically classify and verify the horticultural variety of individuals and detect structure within the germplasm. The germplasm was analysed in relation to published SNP genotypic data (Kuhn *et al.*, 2019c), which represented the three horticultural varieties (Guatemalan - G, Mexican - M, West-Indian - WI). The number of principal components (PCs) retained was based on preserving majority of the variance while retaining the fewest

PCs (Jombart, 2008). The variance explained by each PC was calculated as the ratio of each eigenvalue to the sum of all calculated eigenvalues.

DAPC was performed to determine the genetic differentiation between different clusters of individuals using the find.clusters() function to determine the number of groups (K) *de novo*, with the optimal K selected using the diffNgroup method (Jombart, 2008). The number of PCs to retain was determined using the optim.a.score() function (Jombart, 2008). The clusters are considered as populations, as it may indicate the individual's horticultural variety. The allele composition analysis/membership probabilities were displayed using the compoplot() function (Jombart, 2008). PCA and DAPC data were imported and visualised using the Plotly R Chart Studio (Plotly Technologies Inc, 2015).

Measures of genetic diversity were evaluated with several "F_{ST} analogues", specifically, Nei's G_{ST} (Nei, 1973, Nei & Chesser, 1983), Hedrick's G_{ST} (Hedrick, 2005), and Jost's D (Jost, 2008) and estimators for Hs and Ht using the diff_stats function (Meirmans & Hedrick, 2011, Winter, 2012). Hs and Ht are estimates of the heterozygosity expected for this population with and without sub-populations, respectively. Population divergence was estimated between all combinations of population clusters nested within varieties, using the pairwise_G_{ST}_Nei, pairwise_G_{ST}_Hendrick and pairwise_D functions, furthermore, the chao_bootstrap function was applied to the populations to determine the robustness of the analysis (Winter, 2012). The 12 reference individuals were removed from the "Population Analysis Dataset", to prevent the reference samples from skewing the analysis.

RESULTS

SNP genotyping statistics and affinity propagation analysis

After removing individuals and markers with more than 5% missing data, the cultivar population contained 326 individuals (including 12 reference individuals and 107 unique accession IDs) and 376 markers. Missing data per individual varied from 0 to 17 markers of the 376 markers, thus, the average missing data from individuals was 3.0 or 0.8%. Missing data for markers varied from 0 to 15 for the 326 individuals, thus the average of missing data of all markers was 2.6 or 0.7%. The heterozygous allele calls for individuals ranged from 2.4% (9/370, six missing data) for accession "UCR524 (WI)" to 75% (282/376) for accession "AB042", and the heterozygous allele calls for markers ranged from 0% (0/324, two missing data) for SNP marker "SHRSPaS006061" to 79.3% (253/319, seven missing data) for SNP marker "SHRSPaS002697". Average allele frequency over all markers for allele 1 was 33.8% and allele 2 was 33.7%.

The APA generated 64 cultivar groups for 326 individuals and groups varied from one to 43 individuals. The machine genotyping error ranged from 0% to 2.02% for accession "AB006" with 38 SNP differences. The cultivar germplasm technical genotyping error ranged from 0% to 1.46% for accessions "AB035 & AB266" with 11 SNP differences. The cultivar germplasm contained 35 individuals which were mislabelled, thus, indicating that approximately 10.74% mislabelling is present in the cultivar germplasm (21 individuals were type 1, four individuals were type 2 and 10 individuals were both type 1 & 2). Formatted data with affinity propagation groups, silhouette scores, and genotype data are recorded in **Supplementary File 1 (**10.25403/UPresearchdata.19145087**)**.

Phylogenetic analysis

The genetic distance matrix of the 326 avocado individuals were used to study the genetic relationships in the population through hierarchical clustering. The phylogenetic analysis indicated that the germplasm was divided into three main populations, based on the reference individuals which are highlighted in the darker shade of the respective colours, as seen in **Fig 1**. Based on breeding records and suspected horticultural variety provided by industry, the individuals were coloured accordingly. The UPGMA-based dendrogram produced three major groups, some containing individuals from different horticultural varieties, pointing at genetic admixture between varieties, as seen in **Fig 2**Error! Reference source not found.. Majority of the individuals from the phylogenetic analysis corresponded with the APA, with the exception of some of the mislabelled individuals.

Principal component analysis & population structure analysis

The PCA was used to study the genetic relationships in the cultivar germplasm. The first three eigenvalues were 114.15, 51.22, and 14.73, respectively. The variance explained by the first three PCs were 30.6%, 13.7%, and 4.0%, respectively, hence, the overall variation was 48.3%. The eigenvalues of the analysis showed that majority of the genetic variance was captured by the first three PCs, as seen from the PCA eigenvalue bar graph in **Fig 3 a** and the a-score optimisation in **Fig 3 b**. The scatterplot of the first three PCs for the cultivar germplasm indicated that the reference individuals for Guatemalan (G - blue), Mexican (M - red), and West-Indian (WI - green) had well-defined clusters. The Allesbeste™ cultivar germplasm (orange) appeared to show a cline between the reference individuals, indicating possible genetic admixture, as seen in **Fig 3 c**. The WI cluster separated from the G and M clusters along the first PC, whereas the M cluster separated from the G cluster along the second PC. The Allesbeste™ cultivar germplasm clustered mainly between the G and M clusters, with most individuals grouping closer to the G cluster, as seen in **Fig 3 c**.

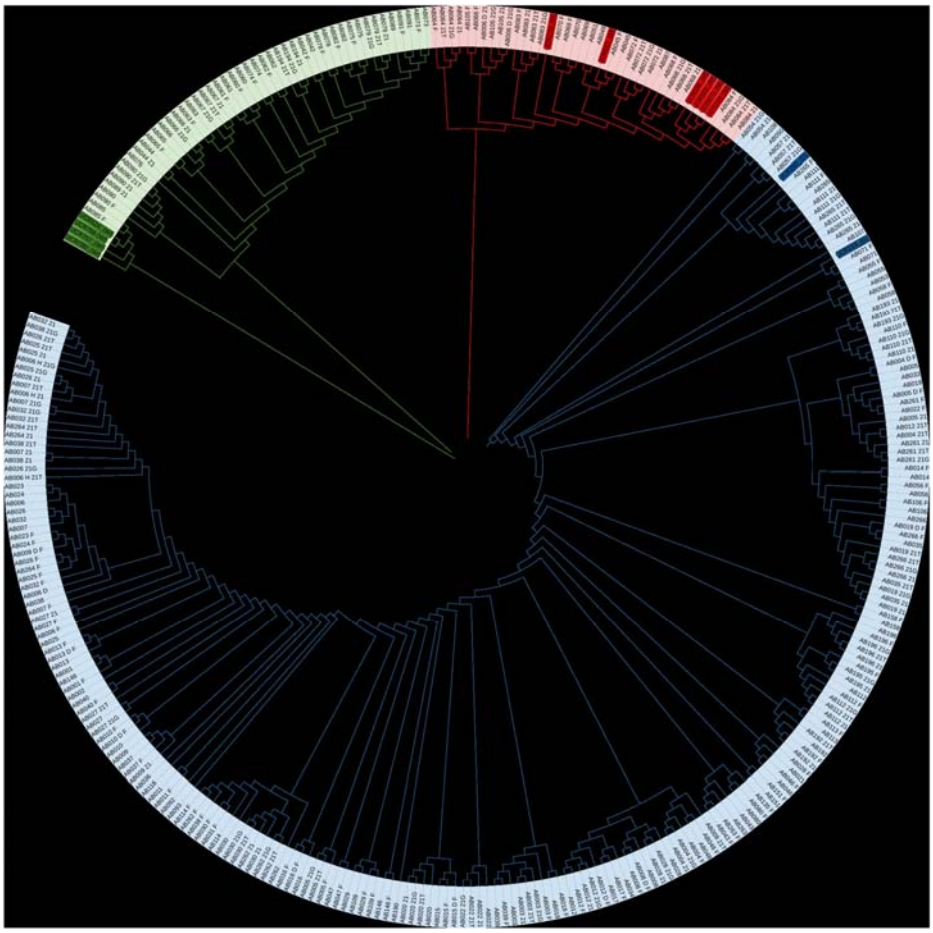


Fig 1 Dendrogram of the 326 avocado trees constructed with UPGMA showing genetic relationships between the analysed samples. Leaves and branches were coloured according to the horticultural variety based on the reference individuals. G: Guatemalan (blue), M: Mexican (red), WI: West Indian (green)

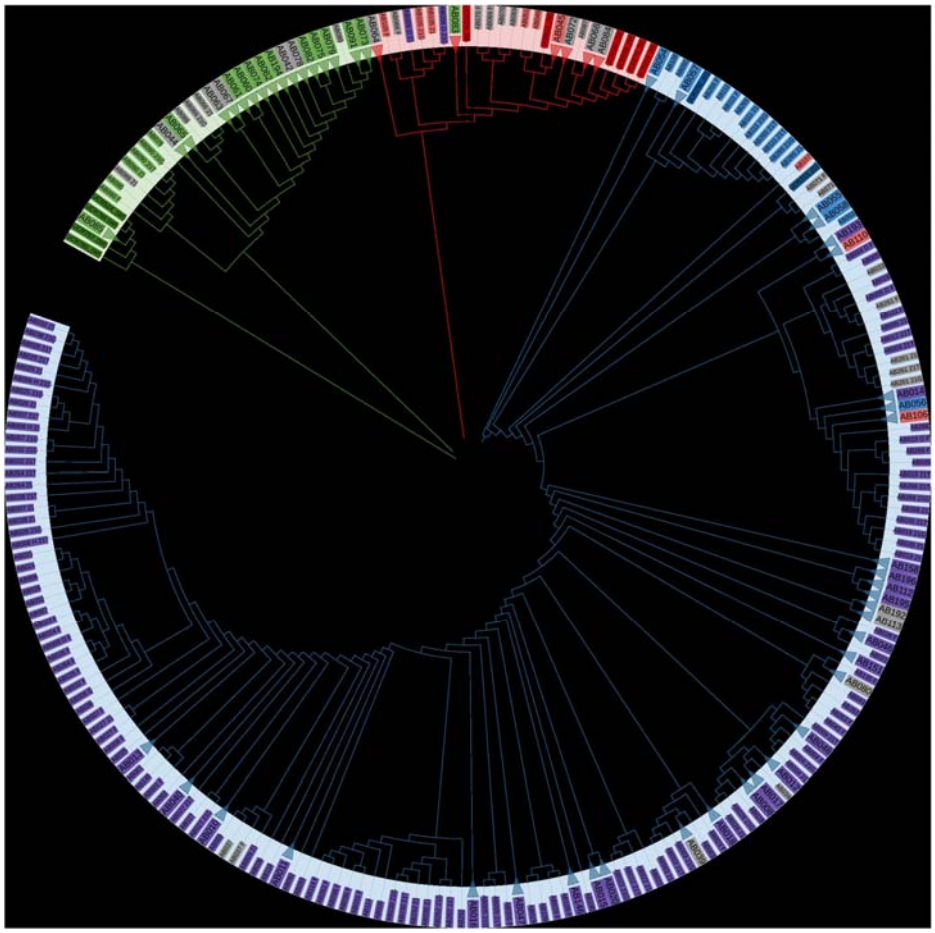
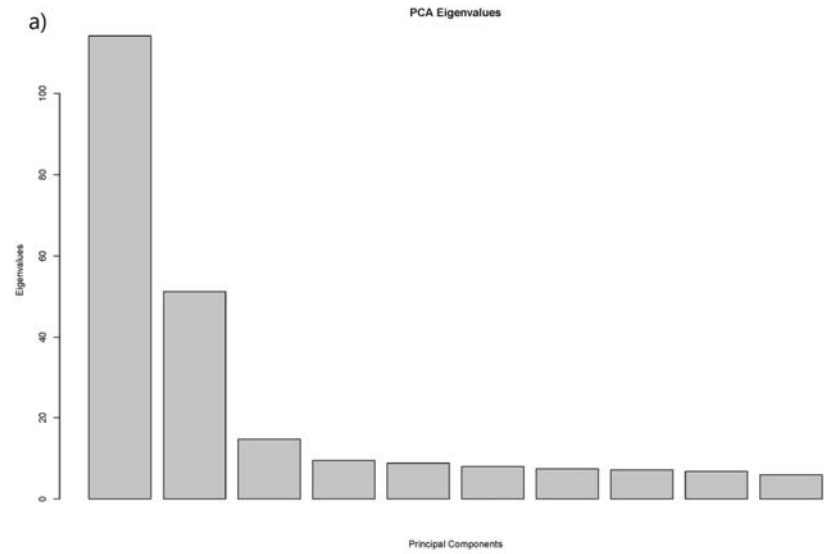
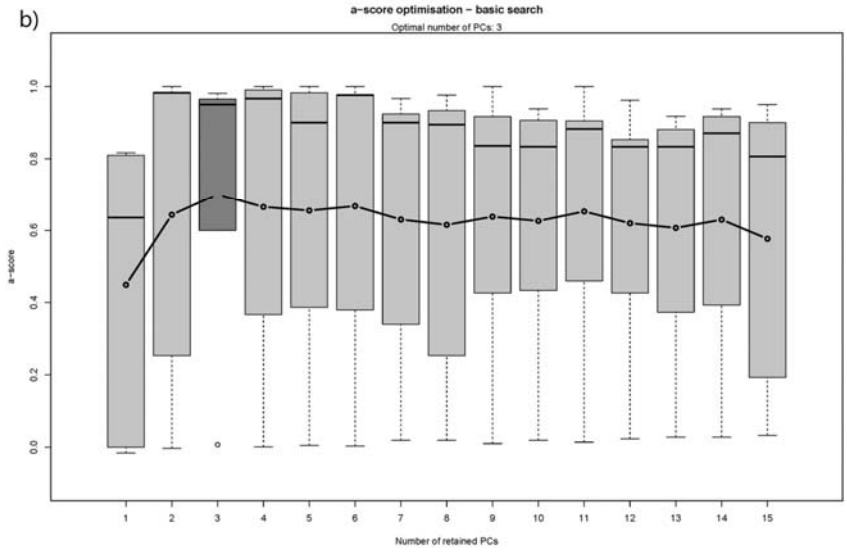


Fig 2 Dendrogram of the 326 avocado trees constructed with UPGMA showing genetic relationships between the analysed samples. Leaves with identical accessions were collapsed into nodes. The reference individuals were coloured according to the horticultural variety (darker shades), G: Guatemalan (blue), M: Mexican (red), WI: West Indian (green). The leaves/nodes representing the individuals were coloured according to horticultural variety information provided by Allesbeste™. G: Guatemalan (blue), M: Mexican (red), WI: West Indian (green), Hybrids (purple), and Unknown (grey)


a)
PCA Eigenvalues
Eigenvalues
Principal Components


b)
a-score optimisation - basic search
Optimal number of PCs: 3
a-score
Number of retained PCs

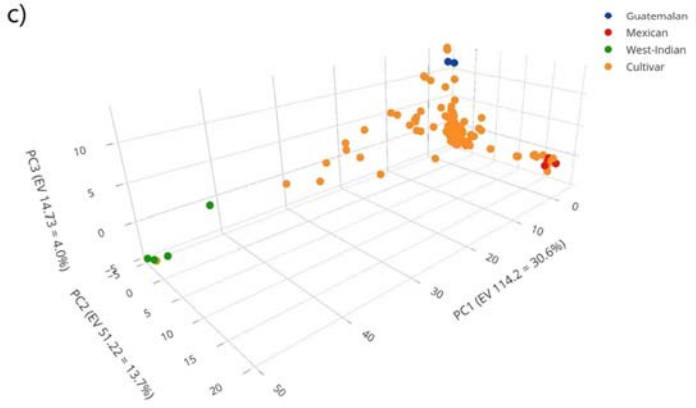

c)
Guatemalan
Mexican
West-Indian
Cultivar
PC3 (EV 14.73 = 4.0%)
PC2 (EV 51.22 = 13.7%)
PC1 (EV 114.2 = 30.6%)

Fig 3 Principal Component Analysis (PCA) of the 159 avocado cultivar germplasm using 375 single nucleotide polymorphisms (SNPs). (a) PCA eigenvalues of the analysis (b) The optimal a-score indicating the number of principal components (PCs) to retain for analysis, indicating three PCs. (c) The eigenvalues and variance of each PC are found within parentheses on each axis. Individuals are represented as dots and the reference varieties are represented by G: Guatemalan (blue), M: Mexican (red), WI: West Indian (green) and the Allesbeste™ cultivar germplasm is represented in orange

Further analysis of the cultivar germplasm based on the above PCA, with the individuals now coloured according to the suspected horticultural variety provided by Allesbeste™ based on breeding records, revealed that the majority of the individuals were G and or GxM hybrids. The cultivar germplasm had 27 individuals with unknown horticultural variety, as seen in **Fig 4**, which were resolved with 16 individuals assigned as G, eight as M, and three as WI. Additionally, there were 17 misclassified individuals, which were reassigned.

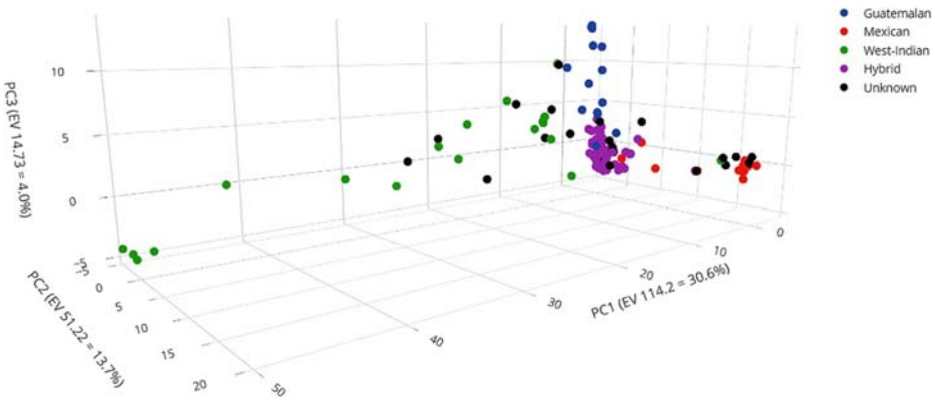


Fig 4 Principal Component Analysis (PCA) of the 159 avocado cultivar germplasm – coloured according to the suspected horticultural varieties based on Allesbeste™ breeding records. The eigenvalues of each principal component are found within parentheses on each axis. Individuals are represented as dots and the horticultural varieties are G: Guatemalan (blue), M: Mexican (red), WI: West Indian (green), GxM: Guatemalan x Mexican hybrid (purple) and Unknown variety (black)

The DAPC was used to investigate the population differentiation between groups of individuals and identify clusters of genetically related individuals. Based on information from literature and industry, hybrids are common between the horticultural varieties, therefore, a DAPC analysis was performed from K = 2 until K = 7 to identify potential hybrids in the germplasm. The eigenvalues of the analysis showed that majority of the genetic variance was again captured by the first three PCs. According to the diffNgroup method, the optimum number of genetic clusters were K = 5, which was best supported and appeared to be the most biologically relevant scatterplot. This scatterplot shows the first two PCs of the DAPC for K = 5,

as seen in **Fig 5 a**. Clusters are shown by different colours and inertia ellipses, while dots represent individuals, indicating the Guatemalan (blue), Mexican (red), and West-Indian (green), Cluster 1 (cyan - possible GxWI hybrids), and Cluster 2 (magenta - possible GxM hybrids). Three groups of genetically closer clusters can be identified, Guatemalan, Cluster 1 and Cluster 2, as seen in **Fig 5 a**. This scatterplot also indicates that majority of the West-Indian accessions are hybrids. Additionally, the scatterplot was shown using the first three PCs of the DAPC of the cultivar germplasm for K = 5, indicating the Guatemalan (blue), Mexican (red), and West-Indian (green), Cluster 1 (cyan), and Cluster 2 (magenta), as seen in **Fig 5 b.** The scatterplot showed a cline, indicating genetic admixture between the genetic clusters. The cultivar germplasm consisted of 9.4% G, 13.2% M, 3.8% WI, 10.7% Cluster 1 (possible GxWI hybrids), and 62.9% Cluster 2 (possible GxM hybrids). Majority of the results from the DAPC matched the suspected horticultural variety provided by Allesbeste™.

The allele composition analysis of the cultivar germplasm indicated the inferred structure and membership probabilities, where each individual is represented by a coloured bar with length proportional to the estimated membership to each cluster (Pritchard *et al.*, 2000), as seen in **Fig 6**Error! Reference source not found.. Majority of the germplasm individuals were composed of the G cluster and Cluster 2 (GxM hybrid), which corresponds to the DAPC results. The reference individuals are located in the enclosed area, from individual 147 to 159 in the genomic composition plot, as seen in **Fig 6**. All genomic composition plots from K=2 until K=7 is recorded in **Supplementary Fig 1**.

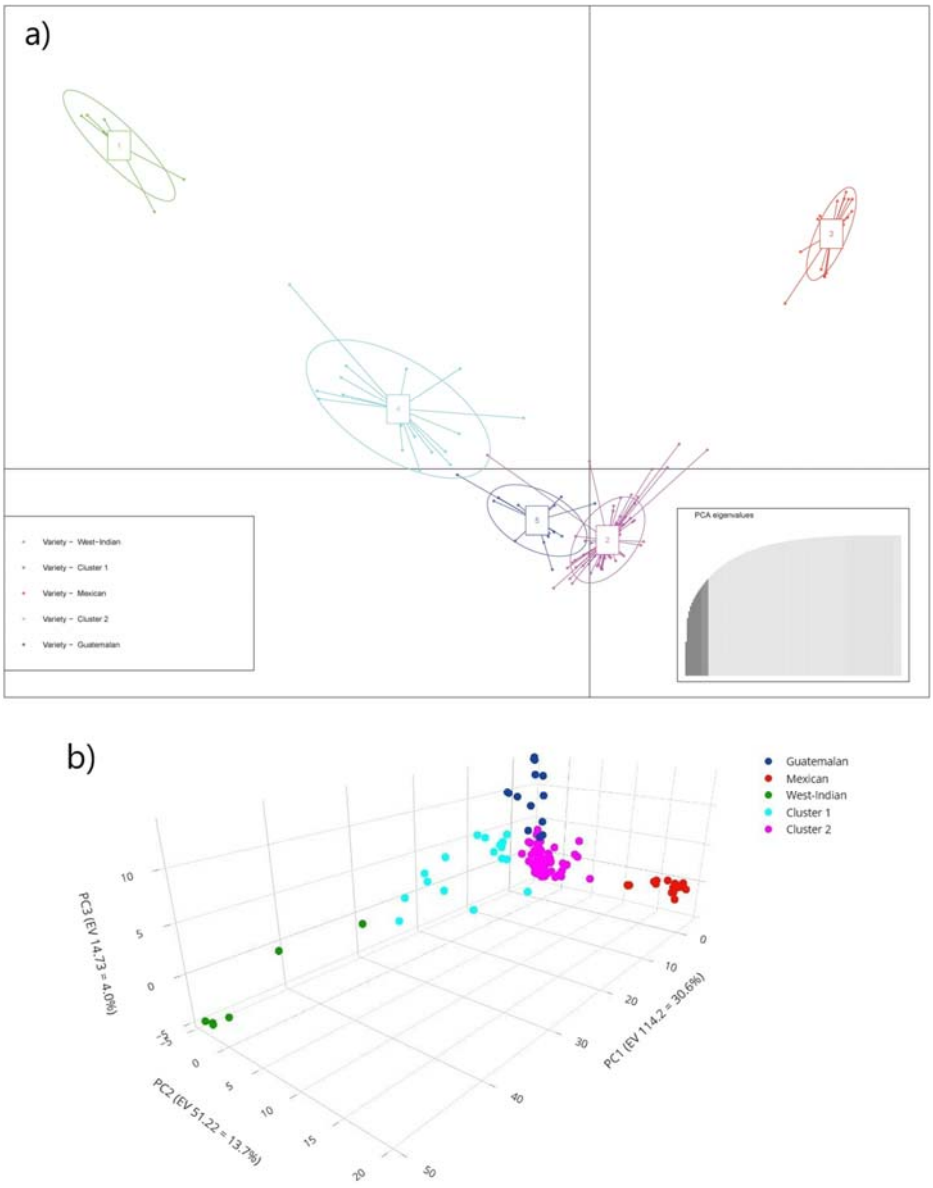


Fig 5 Discriminant Analysis of Principal Components (DAPC) of the 159 avocado cultivar germplasm using 375 single nucleotide polymorphisms (SNPs). (a) Scatterplot shows the first two PCs of the DAPC for K = 5, with clusters shown by different colours and inertia ellipses, while dots represent individuals. The PCA eigenvalue plot is inset on the bottom right (b) Scatterplot shows the first three PCs of the DAPC for K = 5, the eigenvalues and variance of each principal component are found within parentheses on each axis. Individuals are represented as dots and the varieties are represented by G: Guatemalan (blue), M: Mexican (red), WI: West Indian (green), Cluster 1 (cyan) and, Cluster 2 (magenta)

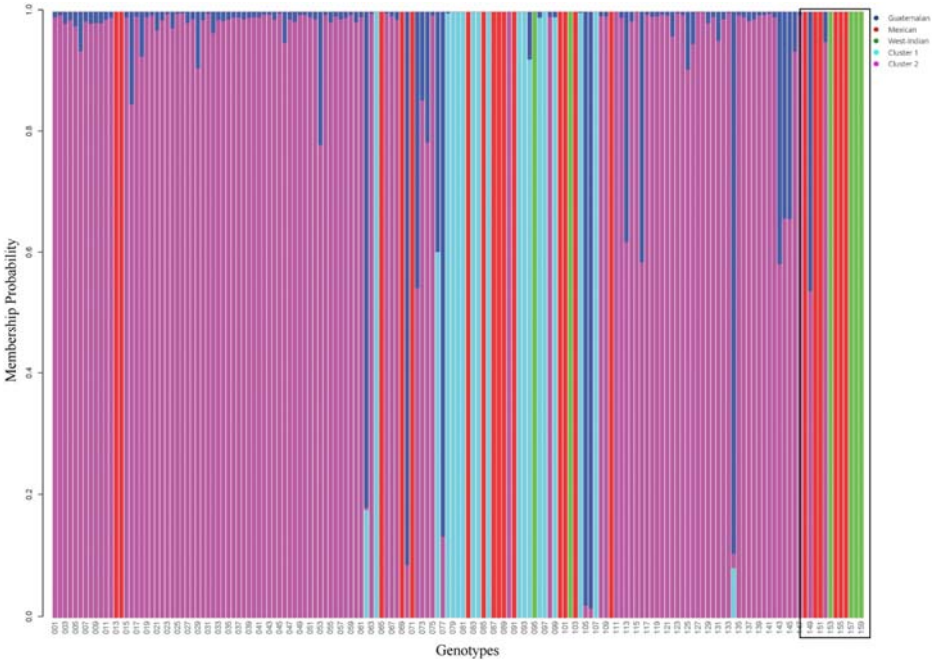


Fig 6 Genomic composition plot of the discriminate analysis of principal components (DAPC) indicating the cluster's composition for the cultivar germplasm for 159 genotypes. Each thin vertical line in the bar plot represents one individual and each colour represents one inferred ancestral population. The length of each colour in a vertical bar represents the proportion of that individual's ancestry that is derived from the inferred ancestral population corresponding to that colour. The same colour in different individuals indicates that they belong to the same cluster, indicating they share the same ancestral population. Clusters: Guatemalan (blue), Mexican (red), West-Indian (green), Cluster 1 (cyan), and Cluster 2 (magenta). Reference individuals are located in the enclosed area

Diversity analysis

The genetic differentiation and diversity present in the germplasm population was evaluated with three "F_{ST} analogues", Nei's G_{ST}, Hedrick's G_{ST}, and Jost's D, additionally Hs and Ht are estimates of the heterozygosity expected for this population with and without the sub-populations defined in the data, respectively. This analysis indicated that the WI vs. Cluster 1 (Nei's G_{ST} = 0.050, Hedrick's G_{ST} = 0.148, Jost's D = 0.058) had the least genetic differentiation, whereas M vs. WI (Nei's G_{ST} = 0.525, Hedrick's G_{ST} = 0.847, Jost's D = 0.509) had the highest genetic differentiation, as seen in **Table 2**.

Table 2 Global population pair-wise F_{ST} comparison among the populations at K=5 identified by the Discriminant Analysis of Principal Components. Hs and Ht are estimates of the heterozygosity expected for this population with and without the sub-populations defined in the data respectively. Clusters: Guatemalan (G), Mexican (M), West-Indian (WI), Cluster 1 (possible GxWI hybrids), Cluster 2 (possible GxM hybrids)

	Hs	**Ht**	**Pair-wise Nei's G_{ST}**	**Pair-wise Hedrick's G_{ST}**	**Pair-wise Jost's D**
G vs. M	0.226	0.313	0.278	0.562	0.225
G vs. WI	0.287	0.438	0.345	0.719	0.423
G vs. Cluster 1	0.375	0.429	0.126	0.359	0.173
G vs. Cluster 2	0.285	0.306	0.069	0.181	0.059
M vs. WI	0.187	0.394	0.525	0.847	0.509
M vs. Cluster 1	0.287	0.399	0.28	0.613	0.313
M vs. Cluster 2	0.199	0.266	0.252	0.502	0.167
WI vs. Cluster 1	0.356	0.374	0.05	0.148	0.058
WI vs. Cluster 2	0.254	0.44	0.422	0.796	0.499
Cluster 1 vs. Cluster 2	0.345	0.424	0.187	0.481	0.242

DISCUSSION

The aim of this study was to use a set of previously developed SNP markers for the validation of clonal material, verification of horticultural variety and determination of the genetic diversity and population structure of an avocado cultivar breeding population in South Africa.

An APA was used to identify mislabelled individuals and confirm clonal material. An APA uses all points simultaneously with no genetic assumptions to determine which individuals would best serve as epitomes and the clustering occurs naturally, thus, decreasing erroneous results (Frey & Dueck, 2007). Previously, Kuhn *et al.* (2019c) used an APA to identify 38 mislabelled individuals in the USDA-ARS Subtropical Horticulture Research Station (SHRS) germplasm collection, thus, indicating 13% mislabelling. Similarly, in this study, the APA identified 35 mislabelled individuals, thus, indicating approximately 10.74% mismatch ratio in the cultivar germplasm. Mislabelling in breeding populations can occur in every phase of avocado production, including incorrect identification in the field, propagation, as well as, during procurement of samples and during genotyping (Kuhn *et al.*, 2019c). It is important to identify mislabelled individuals in germplasms to prevent the propagation of incorrect material, which could be used for budwood purposes. Therefore, identifying these mislabelled individuals may improve breeding efficiency and deployment, while reducing loss of resources and time.

The PCA was used to verify the horticultural variety of individuals in the cultivar germplasm. PCA identifies genetic structures among individuals in the absence of any assumption about the underlying population genetic model (Patterson *et al.*, 2006, Reich *et al.*, 2008), as well as summarizes the overall variability in a population. Based on the PCA, the majority of the South African cultivar germplasm grouped between the Guatemalan and Mexican varieties, and the population appeared to show a cline, rather than well-defined clusters, indicating evidence of genetic admixture. Thus, the cultivar germplasm appeared to consist mainly of GxM hybrids. Furthermore, the cultivar germplasm had 27 individuals of unknown horticultural variety, and 17 individuals with misclassified horticultural variety, which were resolved. Genetic admixture among avocado populations is attributed to the extensive hybridisation between varieties; and this is common as avocado varieties do not have sterility barriers (Davis *et al.*, 1998, Ashworth & Clegg, 2003). Hybrids allow for a desirable blend of important traits in one individual, such as disease resistance and improved yield. Unfortunately, PCA summarises the overall variability in a population and requires an aforementioned definition of clusters to study population structures, thus, these drawbacks warranted further investigation through DAPC.

The DAPC was used to determine the population structure of the cultivar germplasm, as it is a multivariate model which assesses the genetic differentiation between different clusters of individuals into groups, while maximizing between-group variability and minimizing within-group variation (Fisher, 1936, Lachenbruch & Goldstein, 1979, Jombart, 2008, Jombart & Ahmed, 2011). DAPC has a few advantages, such as the probabilistic assignment of individuals to groups (like Bayesian approaches) and the visual assessment of structures for different population genetic models (Jombart *et al.*, 2010). In this study, the DAPC allowed for the verification of the horticultural variety of individuals in the breeding population. Based on the DAPC, the cultivar germplasm consisted of 9.4% Guatemalan, 13.2% Mexican, 3.8% West-Indian, 10.7% Cluster 1 (possible GxWI hybrids), and 62.9% Cluster 2 (possible GxM hybrids). The high percentage of Guatemalan, Mexican and possible GxM hybrids in the germplasm is coherent, as the most popular cultivar grown world-wide is Hass (Crane *et al.*, 2013), which is a GxM hybrid (Rendón-Anaya *et al.*, 2019). Furthermore, the Guatemalan variety has high fruit averages and horticultural quality, whereas the Mexican variety has a desirable fruit size (Bergh & Ellstrand, 1986) and has shown some tolerance and resistance to *Phytophthora cinnamomi* (Sánchez-González *et al.*, 2019), which are valuable traits in the industry. Furthermore, Guatemalan and Mexican varieties are typically grown in less tropical areas (Williams, 1977), such as avocado growing regions in South Africa.

Interestingly, majority of the West-Indian accessions in the cultivar germplasm appeared to be GxWI hybrids, even though the industry records indicated these are West-Indian accessions. These GxWI hybrids have been known to have an early harvest period and bridges harvesting gaps (Bergh & Ellstrand, 1986), which could explain the presence of Cluster 1 (possible GxWI hybrids) in the germplasm. However, there does not appear to be any MxWI hybrids within the population. This may be due to lack of sampling or due to the lack of breeding of MxWI hybrids in South Africa. Some West-Indian individuals are more tolerant to salinity and calcareous soils (Ben-Ya'acov & Michelson, 1995), which is not favoured by most avocado cultivars grown in South Africa. Most commercial avocado rootstocks and cultivars are hybrids (Popenoe & Williams, 1947), hence, it is important to correctly identify the horticultural variety of individuals, as this affects the ability of breeding programmes to select accurate and superior individuals. A concern involved in this study is the precise DAPC assignment of individuals, as it may be skewed by the lack of reference samples utilized during analysis (Ottewell *et al.*, 2016). Furthermore, it is important to curate more avocado germplasms in South Africa to include potential MxWI hybrids and improve our understanding of the population. An informative addition to this study would involve linking the genotypic data with phenotypic data to provide a more rounded description of the germplasm at hand.

Genetic diversity was determined with "F_{ST} analogues" that assessed the within and among population variation. MMOD is a package that allows three different "F_{ST} analogues" to be evaluated, Nei G_{ST}, Hendrick's G_{ST}, and Jost's D, which is comparable between studies (Winter, 2012). These "F_{ST} analogues" and their combined use will allow more robust analyses of population structure than what is achievable with only F_{ST} (Meirmans & Hedrick, 2011). Some previously reported F_{ST} values for avocado germplasms among the three varieties were 0.19, 0.22 and 0.25, reported by Boza *et al.* (2018), Guzmán *et al.* (2017) and Gross-German & Viruel (2013), respectively, whereas lower F_{ST} values of 0.061 and 0.05 were reported by Juma *et al.* (2020) and Cañas-Gutiérrez *et al.* (2019) respectively. In this study, the "F_{ST} analogues" indicated that the West-Indian vs. Cluster 1 (possible GxWI hybrids; Nei G_{ST}= 0.050) had the least genetic differentiation, whereas Mexican vs. West-Indian had the highest genetic differentiation (Nei G_{ST} = 0.525).

These studies show the varying levels of diversity in numerous avocado germplasms worldwide. These diversity levels can be affected by the type and number of markers used, the number of individuals and populations assessed, comparable reference samples and different parameters used for the analysis. Genetic diversity allows for a species to adapt to various environmental conditions and stressors (Schleif, 1993), such as climate change and

resistance to new emerging pathogens and pests. The cultivar germplasm analysed in this study contained moderate differentiation between varieties and hybrid clusters. The "F_{ST} analogues" values in this study were similar to other studies, such as Guzmán *et al.* (2017) and Gross-German & Viruel (2013). Moderate levels of differentiation in the germplasm suggests interbreeding between the three varieties, which is seen with Cluster 1 (possible GxWI hybrids) and Cluster 2 (possible MxG hybrids) in this study. Majority of the cultivar germplasm (62.9%) grouped into Cluster 2 (possible GxM hybrids); this would correlate with industry breeding records.

To our knowledge, this study presents the first molecular genetic assessment of an avocado cultivar germplasm in South Africa. In the present study, molecular marker technology was used to identify mislabelled individuals, validate clonal material, verify horticultural variety, determine population structure, and genetic diversity. The results from the study may prevent the future propagation of incorrect material, establish proper management and conservation strategies and lastly, improve cultivar breeding efficiency by aiding in the selection of avocado with the ability to cope with changing environments and emerging pests and pathogens. Molecular markers are a powerful and important tool for avocado breeding programmes.

ACKNOWLEDGMENTS
The authors would like to thank the Hans Merensky Foundation© and Allesbeste™ for funding, as well as, the Forestry and Agricultural Biotechnology Institute (FABI) and the University of Pretoria for the use of their facilities and equipment. Furthermore, the authors would like to thank Dr David Kuhn for the custom affinity propagation scripts. Lastly, I would like to thank Allesbeste™ for providing the plant material.

AUTHOR CONTRIBUTIONS
RW contributed to the study design, experimental design, sample curation, formal analysis, investigation, visualisation, drafting/writing/editing of the manuscript. NVDB contributed to the study conceptualization and design, experimental design, project administration, resources, supervision and funding. MMON and NA were responsible for methodology and technical assistance. PM provided the horticultural reference. BF extracted, processed and performed the SNP genotyping. All co-authors contributed to writing/editing of the manuscript. All authors contributed to and approved the final manuscript.

STATEMENTS AND DECLARATIONS
Financial support was received from the Hans Merensky Foundation© and Allesbeste™. 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. This study received specific approval by the appropriate ethics committee for research involving plants. Permission has been granted for the use of the reference dataset by the relevant co-authors. The figures, tables and text passages have originated from this study and from the author.

REFERENCES

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

Batley J (2015) *Plant Genotyping*. Humana Press, Dordrecht, London.

Ben-Ya'acov A & Michelson E (1995) Avocado rootstocks. Vol. 17 (Janick J, ed.) 381-429. John Wiley and Sons, Inc, New York, NY.

Bergh B & Ellstrand N (1986) Taxonomy of the avocado. *California Avocado Society Yearbook* **70**: 135-146.

Bodenhofer U, Kothmeier A & Hochreiter S (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27**: 2463-2464.

Boza EJ, Tondo CL, Ledesma N, Campbell RJ, Bost J, Schnell RJ*, et al.* (2018) Genetic differentiation, races and interracial admixture in avocado (*Persea americana* Mill.), and *Persea* spp. evaluated using SSR markers. *Genetic Resources and Crop Evolution* **65**: 1195-1215.

Cañas-Gutiérrez GP, Arango-Isaza RE & Saldamando-Benjumea CI (2019) Microsatellites revealed genetic diversity and population structure in Colombian avocado (*Persea americana* Mill.) germplasm collection and its natural populations. *Journal of Plant Breeding and Crop Science* **11**: 106-119.

Chen H, Morrell P, De La Cruz M & Clegg M (2008) Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). *Journal of Heredity* **99**: 382–389.

Chen H, Morell P, Ashworth V, De La Cruz M & Clegg M (2009) Tracing the geographic origins of major avocado cultivars. *Journal of Heredity* **100**: 56–65.

Clegg M (2004) Application of molecular markers to avocado improvement. 24-28. California Avocado Commission, Proceedings of the California Avocado Research Symposium, University of California, Riverside.

Crane J, Douhan G, Faber B, Arpaia M, Bender G, Balerdi C*, et al.* (2013) The avocado botany, production and uses: Cultivars and rootstocks. (Schaffer B, Wolstenholme B & Whiley A, eds.), 200-233. CABI, UK.

Davis J, Henderson D, Kobayashi M & Clegg M (1998) Genealogical relationships among cultivated avocado as revealed through RFLP analyses. *Journal of Heredity* **89**: 319-323.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792-1797.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**: 179-188.

Frey BJ & Dueck D (2007) Clustering by passing messages between data points. *Science* **315**: 972-976.

Ge Y, Tan L, Wu B, Wang T, Zhang T, Chen H*, et al.* (2019b) Transcriptome sequencing of different avocado ecotypes: De novo transcriptome assembly, annotation, identification and validation of EST-SSR markers. *Forests* **10**: 411.

Ge Y, Zhang T, Wu B, Tan L, Ma F, Zou M*, et al.* (2019a) 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**: 215.

Gross-German E & Viruel M (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 Genetics & Genomes* **9**: 539-555.

Guzmán LF, Machida-Hirano R, Borrayo E, Cortés-Cruz M, Espíndola-Barquera MdC & Heredia García E (2017) Genetic structure and selection of a core collection for long term conservation of avocado in Mexico. *Frontiers in Plant Science* **8**: 243.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* **59**: 1633-1638.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403-1405.

Jombart T & Ahmed I (2011) *adegenet* 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**: 3070-3071.

Jombart T, Devillard S & Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**: 1-15.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015-4026.

Juma I, Geleta M, Nyomora A, Saripella GV, Hovmalm HP, Carlsson AS, *et al.* (2020) Genetic diversity of avocado from the southern highlands of Tanzania as revealed by microsatellite markers. *Hereditas* **157**: 1-12.

Karp A, Edwards KJ, Bruford M, Funk S, Vosman B, Morgante M, *et al.* (1997) Molecular technologies for biodiversity evaluation: opportunities and challenges. *Nature Biotechnology* **15**: 625-628.

Köhne S (2005) Selection of avocado scions and breeding of rootstocks in South Africa. New Zealand and Australia Avocado Grower's Conference, Tauranga, New Zealand.

Kuhn D, Livingstone III D, 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. *Scientia Horticulturae* **246**: 1–11.

Kuhn D, Bally I, Dillon N, Innes D, Groh A, Rahaman J, *et al.* (2017) Genetic map of mango: a tool for mango breeding. *Frontiers in Plant Science* **8**: 577.

Kuhn D, Dillon N, Bally I, Groh A, Rahaman J, Warschefsky M, *et al.* (2019a) Estimation of genetic diversity and relatedness in a mango germplasm collection using SNP markers and a simplified visual analysis method. *Scientia Horticulturae* **252**: 156–168.

Kuhn D, Groh A, Rahaman J, Freeman B, Arpaia M, van den Berg N, *et al.* (2019c) Creation of an avocado unambiguous genotype SNP database for germplasm curation and as an aid to breeders. *Tree Genetics & Genomes* **15**: 71.

Kumar S, Stecher G, Li M, Knyaz C & Tamura K (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **35**: 1547–1549.

Lachenbruch P & Goldstein M (1979) Discriminant analysis. *Biometrics* **35**: 69-85.

Lahav E & Lavi U (2002) Genetics and classical breeding. (Whiley A, Schaffer B & Wolstenholme B, eds.), 39-69. CAB International, Wallingford, U.K.

Letunic I & Bork P (2019) Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**: W256–259.

Meirmans PG & Hedrick PW (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources* **11**: 5-18.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**: 3321-3323.

Nei M & Chesser RK (1983) Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**: 253-259.

Ottewell KM, Bickerton DC, Byrne M & Lowe AJ (2016) Bridging the gap: A genetic assessment framework for population-level threatened plant conservation prioritization and decision-making. *Diversity and Distributions* **22**: 174-188.

Patterson N, Price AL & Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* **2**: e190.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O*, et al.* (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**: 2825-2830.

Plotly Technologies Inc (2015) Collaborative data science. (Plotly Technologies Inc, ed.) Montréal, QC.

Popenoe W & Williams L (1947) The expedition to Mexico of October 1947. *California Avocado Society Yearbook* **1947**: 22–28.

Pritchard J, Stephens M & Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**.

R Development Core Team (2020) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Reich D, Price AL & Patterson N (2008) Principal component analysis of genetic data. *Nature Genetics* **40**: 491-492.

Rendón-Anaya M, Ibarra-Laclette E, Méndez-Bravo A, Lan T, Zheng C, Carretero-Paulet L*, et al.* (2019) The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the National Academy of Sciences* **116**: 17081-17089.

Rousseeuw P (1987) Silhouettes - a graphical aid to the interpretation and validation of cluster-analysis. *Journal of Computational and Applied Mathematics* **20**: 53–65.

RStudio Team (2016) RStudio: integrated development for R.

Rubinstein M, Eshed R, Rozen A, Zviran T, Kuhn D, Irihimovitch V, *et al.* (2019) Genetic diversity of avocado (*Persea americana* Mill.) germplasm using pooled sequencing. *BMC Genomics* **20:379**.

Sánchez-González EI, Gutiérrez-Soto JG, Olivares-Sáenz E, Gutiérrez-Díez A, Barrientos-Priego AF & Ochoa-Ascencio S (2019) Screening progenies of mexican race avocado genotypes for resistance to *Phytophthora cinnamomi* Rands. *HortScience* **54**: 809-813.

Schaffer B, Wolstenholme B & Whiley A (2013) *The avocado: botany, production and uses*. CABI, Oxfordshire, UK.

Schleif R (1993) *Genetics and molecular biology*. Johns Hopkins University Press, Baltimore, MD.

Schnell R, Brown J, Olano C, Power E, Krol C, Kuhn D, *et al.* (2003) Evaluation of avocado germplasm using microsatellite markers. *Journal of the American Society for Horticultural Science* **128**: 881–889.

Sneath PH & Sokal RR (1973) *Numerical taxonomy. The principles and practice of numerical classification*.

Williams L (1977) The avocado, a synopsis of the genus *Persea*, subg. *Persea*. *Economic Botany* **31**: 315–320.

Winter DJ (2012) MMOD: an R library for the calculation of population differentiation statistics. *Molecular Ecology Resources* **12**: 1158-1160.

Wolstenholme B (2003) Avocado rootstocks: What do we know; are we doing enough research? *South African Avocado Growers' Association Yearbook* **26**: 106-112.

DATA AVAILABILITY & ARCHIVING STATEMENT

The cultivar germplasm analysed during this study is available in the University of Pretoria Research Repository [Supplementary File 1 - https://doi.org/10.25403/UPresearchdata.19145087].