Assessment of Genome-Wide DArT-seq Markers for Tea *Camellia sinensis* (L.) O. Kuntze Germplasm Analysis

MP Malebe[1], NIK Mphangwe[2], AA Myburg[1] and Z Apostolides[1*]

[1]*Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0002, South Africa*

[2]*Tea Research Foundation of Central Africa, Mulanje, Malawi*

[*]*Corresponding author email:* za@up.ac.za

Abstract

Tea is the world's most consumed beverage, next to water. This makes it an economically important crop. However, the genetic relationship between the cultivars at the Tea Research Foundation of Central Africa (TRFCA) have not been classified. This study assessed 56 tea accessions from the TRFCA. The 56 accessions were split into five groups, namely, Cambod type, China type, TRFCA, Tea Research Institute (TRI), and unknown. A total of 16 382 genome-wide genetic markers were developed using a next-generation sequencing (NGS) platform at Diversity Array Technology (DArT). The current study also explored the usefulness of the DArT-seq markers for tea germplasm analysis. The genetic relationships among the cultivars were analysed using the neighbour-joining method & UPGMA, this was successful in clustering the different cultivars into groups of origin. Principal coordinate analysis (PCoA) showed that 33 TRFCA derived cultivars were distributed in all four quadrants. Analysis of molecular variance (AMOVA) revealed that there was a higher proportion of genetic variation (94%) within the groups than there was among the groups (6%). Nei's unbiased genetic distances among groups suggested that the 33 TRFCA derived cultivars had a low genetic distance from the other four groups. This confirms the PCoA inference, that the 33 TRFCA had a high genetic admixture. The genetic structure was utilised to assign three cultivars of unkown origin, to the Cambod type group. Two cultivars were closely related to the China type group. Our findings are a useful guide for future tea breeding programmes in Southern Africa.

Keywords: germplasm, DArT-seq markers, genetic diversity, *Camellia sinensis*

Introduction

Tea is produced from the leaves of *Camellia sinensis* (L.) O. Kuntze. The tea plant is mostly found in eastern and southern Asia, namely China, Japan and India. There are over 50 countries around the world where tea is cultivated (Wheeler and Wheeler 2004). The main types of processed tea are green tea, oolong tea and black tea. There are three species of tea that are used to produce commercial tea. The Assam type (*Camellia assamica*) is indigenous to Southeast Asia, Assam, Indochina and China. It is a tall tree with large leaves, and it is less resistant to cold. The China type (*Camellia sinensis*) is indigenous to China and is a bush with small leaves and is cold resistant. The Cambod type (*Camellia assamica* subsp. *Lasiocalyx*) has been treated as an intermediate between the China and Assam types of tea in terms of its cold tolerance.

In 1956 a tea improvement programme was started at the TRFCA in Malawi (Ellis and Nyirenda 1995). This programme began with the selection of five trees. These five plants were vegetatively propagated and planted out. A further 16 trees were selected, and all 21 were propagated. The best ten plants were selected, from which two commercial clonal seed gardens were planted in 1962. The ten clones were reduced to six clones that were used to establish more clonal seed gardens. The breeding programme progressed, and clones (with good yield and quality) were released and recommended for commercial use (Ellis and Nyirenda 1995). Over the years cultivars from the Tea Research Institute (TRI) (Kericho, Kenya) were also introduced into the TRFCA programme. The progress in the yield and quality of the newly released cultivars seems to have plateaued (Wium 2009). This could be due to a limited genetic diversity in the parental plants used to develop the cultivars. It would be beneficial to identify genetically distant plants for use as parents in the breeding programme. Information regarding the parents and the origins of some of these cultivars has been misplaced over the years. Clarifying the genetic structure within the germplasm will assist in management and conservation of the tea genetic resources. A successful breeding or genetic conservation programme relies on the understanding of the genetic variation in the gene pool (Paul et al. 1997).

The cultivars released in the past 15 years from the TRFCA have been from a selection of F1 progeny produced from deliberate crossings from the above pool of parents and vegetatively propagated. Good cultivars may be found with fortuitous combination of good traits, this is illustrated by the cultivar Yabukita, widely used in Japan (Tanaka 2012). The classification of tea cultivars is usually based on morphological characterisation (e.g. shape of the leaf as well as the colour of leaf and flower). Previously, phylogenetic dendrograms for tea have been constructed using amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD) and microsatellite markers (Roy and Chakraborty 2009; Fang et al. 2012; Wambulwa et al. 2016; Wambulwa et al. 2016b). More recently, 23 microsatellite markers and three chloroplast DNA regions (intergenic spacer regions identified using Sanger sequencing) were used to investigate the genetic relationships of tea accessions in Africa, 20 of these accessions were from the TRFCA (Wambulwa et al. 2016; Wambulwa et al. 2017). The phylogenetic relationships among 18 tea accessions of *Camellia sinensis* and its wild relatives were analysed using 15 444 SNP markers discovered from restriction site-associated DNA sequencing (RAD-seq) (Yang et al. 2016).

The developments made in NGS have allowed for advances in marker identification. The new technologies that incorporate restriction enzymes and NGS for marker discovery include DArT, RAD-seq and genotyping-by-sequencing (GBS) (Baird et al. 2008; Elshire et al. 2011; Sansaloni et al. 2011). Many of these marker identification methods also rely on restriction enzymes to generate a reduced representation of a genome (Davey et al. 2011). Reduction of the genome complexity is used for species with large genomes to ensure an adequate overlap in sequence coverage (Elshire et al. 2011). The difference between DArT-seq and GBS as well as RAD-seq is in how the complexity reduction is achieved. DArT-seq complexity reduction methods efficiently target low copy sequences corresponding predominantly to active genes in the genome (Sansaloni et al. 2011; Li et al. 2015). The advantages of DArT-seq include scoring thousands of unique genome-wide markers in one single experiment, low-cost and no prior sequence information is required. DArT-seq technology has been applied successfully in diversity analyses and phylogenetic studies on several crop species (Alam et al. 2018; Mogga et al. 2018; Zaitoun et al. 2018). However, the application of DArT-seq markers in analysis of genetic resources in tea has not been explored.

Genetic diversity in an agricultural species such as tea is important as it allows for genetic improvement. Knowledge of genetic diversity among the available tea cultivars is an essential prerequisite for future breeding and crop improvement programs (Balasaravanan et al. 2003). Molecular markers can be utilised for characterising crop plant germplasm and decision making for their conservation (Barcaccia 2009). This study explores the usefulness of DArT-seq markers for tea germplasm analysis. The study also aims to reveal the genetic relationship of the selected cultivars.

Materials and methods

Plant material

Fresh tea leaves were collected from 56 cultivars at the cultivar reserve at the TRFCA in Malawi, packed in plastic bags and transported to South Africa via air transport within 12 hours in cooler boxes. They were kept at 4^{o}C until DNA was isolated within seven days. The 56 cultivars consisted of the following groups: 33 TRFCA derived cultivars, ten of which were of unknown origin, seven were China type, three were Cambod type, and three were TRI full-siblings.

Phenotypic data measurement

Data collection on yield in form of plucked two or three leaves and a bud from established tea was carried out at intervals of 10/11 days or 14 days depending on time of the year. The yield from individual plots were extrapolated to green leaf yield (Kg) per which was converted to made tea per hectare (mt/ha) by a conversion factor of 0.225 prior to statistical analyses. The mean yields provided are for five seasons after the plants were fully established in the field.

The rate of leaf fermentability was determined using the Chloroform test (Sanderson, 1963). This was carried out on harvested two leaves and a bud sampled from five randomly selected bushes per plot to determine the fermentability of the clones. SFS 204 and SFS 150 were used as standards for quick and slow fermentation, respectively. Fermentability was scored based on the change in colour of the shoots after 40, 80 and 120 minutes using a 4-point scale as: 1 - bright red brown (fast fermenter); 2- dull brown (moderate fermenter); 3- greenish tinge (poor fermenter); 4-green (non-fermenter).

Drought tolerance was scored during the dry periods of the year, usually between September and November in field plots by visual assessment of level of wilting using a 3-point scale as: 1. no visual sign of wilting and drooping of leaves 2. mild wilting and drooping of leaves; 3. severe wilting and drooping of leaves with some leaf fall.

Isolation of genomic DNA

DNA extraction was carried out using the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA). The DNA concentration was determined with the use of a UV spectrophotometer, Gene-Quant Pro UV/Vis Spectrophotometer (Amersham Biosciences, Uppsala, Sweden).

Sequencing

Sequencing was carried out at DArT Pty Ltd (Canberra, Australia). DNA quality was controlled prior to shipment by testing the digestibility of the DNA using restriction enzyme *Pst*I (Fermentas, Burlington, Canada) and *Eco*RI (Promega, Madison, USA). Each restriction digestion reaction was carried out as described by the manufacturer; this was then run on 1% agarose gel. The NGS procedure was carried out as described by Sansaloni et al. (2011) using *Pst*I and *Mse*I restriction enzymes.

Sequencing data cleaning and analysis

The data was received in Excel format, 1 or 0 was used to record the presence or absence of each marker in each cultivar, respectively. The markers underwent quality control. The non-polymorphic markers, as well as the markers that had a missing proportion of more than 5%, were removed. The polymorphic information content (PIC) was calculated using PowerMarker 3.25 (Liu and Muse 2005). Markers that had a PIC value lower than 0.18 were removed. SAS® Enterprise Guide® 4.3. (SAS Institute Inc 2010. Administering SAS® Enterprise Guide® 4.3. Cary, NC: SAS Institute Inc.) was used to perform correlation analysis. Duplicates (i.e. markers that associated statistically to one another and resulted in a correlation value of one) were removed from this study. JMP 10 (JMP, Version 10. SAS Institute Inc., Cary, NC, 1989-2007) was used to perform One-way analysis of variance (ANOVA) for the distance between two markers versus the number of markers used to calculate this distance.

Data analysis

The genetic relationships between the cultivars were illustrated based on the neighbour-joining method using PowerMarker 3.25 and MEGA version 5 (Lui and Muse 2005; Tamura et al. 2011). The genetic distance between two cultivars was calculated using the shared allele. The characteristic properties of DArT-seq markers was assessed by calculating the PIC. A Student's t-test was used to determine if there was a significant difference in the distance calculated when using 5 DArT-seq markers and when using 5 000 markers. The Student's t-test was calculated using JMP 10 (JMP, Version 10. SAS Institute Inc., Cary, NC, 1989-2007).

GenAlEx 6.5 was used to analyse the genetic diversity of the germplasm groups (Peakall and Smouse 2006; Peakall and Smouse 2012). AMOVA was used to assess the amount of variation among groups within the germplasm. PCoA was performed to visualise the patterns of the genetic relationship. Pairwise estimates of Nei's unbiased genetic distance between the groups were determined using PowerMarker 3.25 (Nei 1973; Lui and Muse 2005).

The genetic population structure analysis was assessed using the Bayesian admixture procedure implemented in STRUCTURE 2.3.4 (Pritchard et al. 2000). The software was programmed to run using the admixture model and correlated allele frequency. The number of assumed populations (K) was estimated for K ranging from 1-8 using three independent runs of 100 000 burn-in periods and 500 000 Markov Chain Monte Carlo (MCMC) repetitions after burn-in. The most probable number of populations was determined using Structure Harvester (Earl 2012). Structure Harvester followed the recommendation of (Evanno et al. 2005).

Results and discussion

A total of 88 745 markers were developed using DArT-seq. The non-polymorphic markers were removed from this list, and 68 317 polymorphic markers remained. Markers that had a missing proportion of more than 5% were eliminated, 27 354 markers remained. The PIC values for the DArT-seq markers ranged from 0.04-0.38 (Suppl. Table 1). Markers that had a PIC value lower than 0.18 and duplicates (i.e. markers that associated statistically to one another and resulted in a correlation value of one) were removed from this study, 16 499 DArT-seq markers remained. Genetic distance between cultivars was calculated using 16 382 DArT-seq markers out of the total 16 499, as Excel has a limitation of 16 384 columns. The development of this large number of high-throughput markers provides an opportunity to anchor the markers on a tea reference genome, which is yet to be completed (Xia et al. 2017). The DArT-seq markers had an average PIC value of 0.29. The average PIC values of 0.29 in tea was similar to values identified in DArT-seq markers of macademia (0.29) and sugar beet (0.28) (Alam et al. 2018; Simko et al. 2012).

The dendrogram was successful in clustering the different cultivars into groups of origin (e.g. Cambodian type and China type) (Fig. 1 and Suppl. Fig. 1). This dendrogram also clustered siblings together (e.g. the TRI progeny of the cross PC1 X TN14-13). RC13 and MT12 (of unknown origin) clustered with the Cambod type.

One-way ANOVA for the distance between two markers by number of markers was used to calculate this distance. This revealed that there was a high standard error of the mean when five markers are used to calculate the distance between two cultivars (Fig. 2 and Fig. 3). An increase in the number of markers used, resulted in a decrease in the standard error of the mean. PC 113 and PC118 are full siblings. The use of fewer markers results in an incorrect estimation of the genetic relationship which appears to be more distant (Fig. 2). In Fig. 3, the distance between CL12 and PC213 is calculated to be smaller when fewer markers are used. This resulted in an incorrect estimation of the genetic relationship (CL12 and PC213 do not share a common parent). The increase in DArT-seq markers resulted in a more accurate dendrogram. The distance was within 5% of the final distance when ≥2 500 markers were used when calculating the genetic distance between the full siblings, PC113 and PC118. The distance was within 5% of the final distance when using ≥1 000 markers when calculating the genetic distance between the cultivars that did not share parents, CL12 and PC213.

A Student's t-test was used to determine if there was a significant difference in the distance calculated using 5 DArT-seq markers and with the use of 5 000 markers (Table 1 and 2). The null hypothesis was rejected as there was a statistically significant difference between the genetic distance calculated using 5 DArT-seq markers and the genetic distance calculated using 5 000 DArT-seq markers at a 95% level of confidence.

The PCoA analysis (Fig. 4) confirmed the clusters obtained from neighbour-joining method dendrogram seen in Fig. 1. The cultivars sharing CL12 as a common parent clustered together in the upper left quadrant. Those sharing SFS150 as a common parent clustered together in the lower quadrant. The Cambod type clustered in the upper quadrants with MT12 and RC13 (of unknown origin). The full-siblings from TRI group clustered together in the lower right quadrant. The China type group clustered in the left quadrants. The TRFCA group's cultivars were scattered in all four quadrants. This indicates that the TRFCA cultivars are a product of various crosses obtained by the breeders as they had a high proportion of genetic admixture.

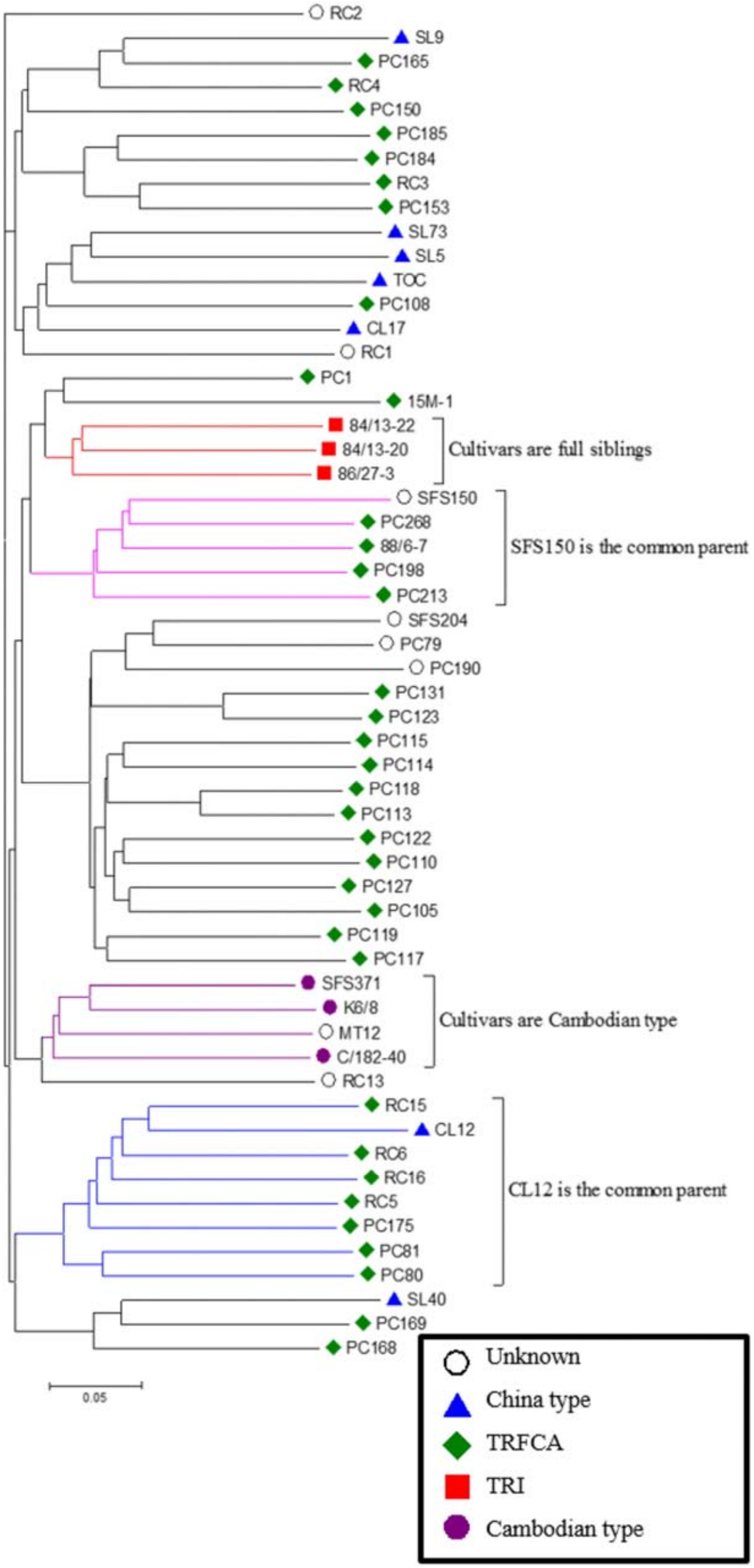


Fig 1. The genetic relationship amongst 56 cultivars using the neighbour-joining method. The dendrogram was generated using 16,382 DArT-seq markers

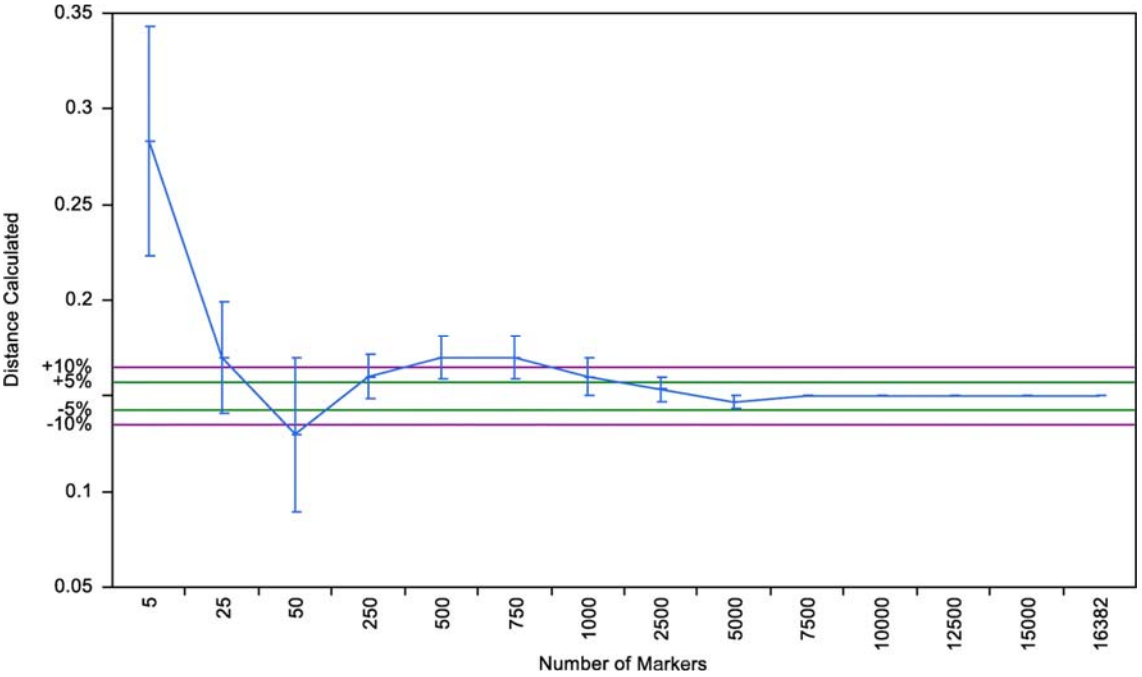


Fig. 2. One-way ANOVA for genetic distance calculated by number of markers for the two siblings PC113 and PC118. Error bars display standard error of mean ($n = 3$ for number of markers ranging from 5 to 5000, $n = 1$ for number of markers > 5000). The purple reference lines mark ± 10% of the final distance. The green reference lines mark ± 5% of the final distance

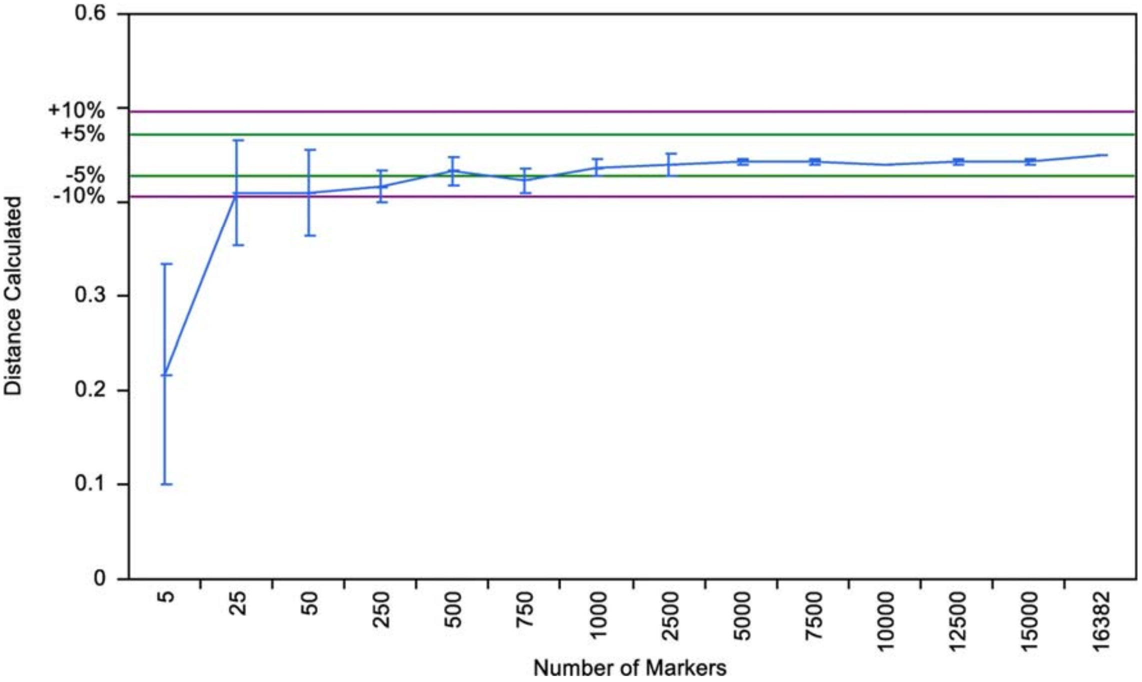


Fig 3. One-way ANOVA for distance calculated by number of markers for cultivars that do not share a parent: CL12 and PC213. Error bars display standard error of mean ($n = 3$ for number of markers ranging from 5 to 5000). The dashed purple reference lines mark ± 10% of the final distance. The solid green reference lines indicate ± 5% of the final distance

Table 1. Student's t-test for distance calculated between full-siblings, PC113 and PC118, P=0,0001, mean values with different group letters are significantly different at the 95% level of confidence

Number of markers	Distance calculated mean	N	Grouping
5	0.28	3	A
5 000	0.15	3	B

Table 2. Student's t-test for distance calculated between cultivars that do not share a parent: CL12 and PC213, P=0,0002. Mean values with different group letters are significantly different at the 95% level of confidence

Number of markers	Distance calculated mean	N	Grouping
5	0.22	3	A
5 000	0.44	3	B

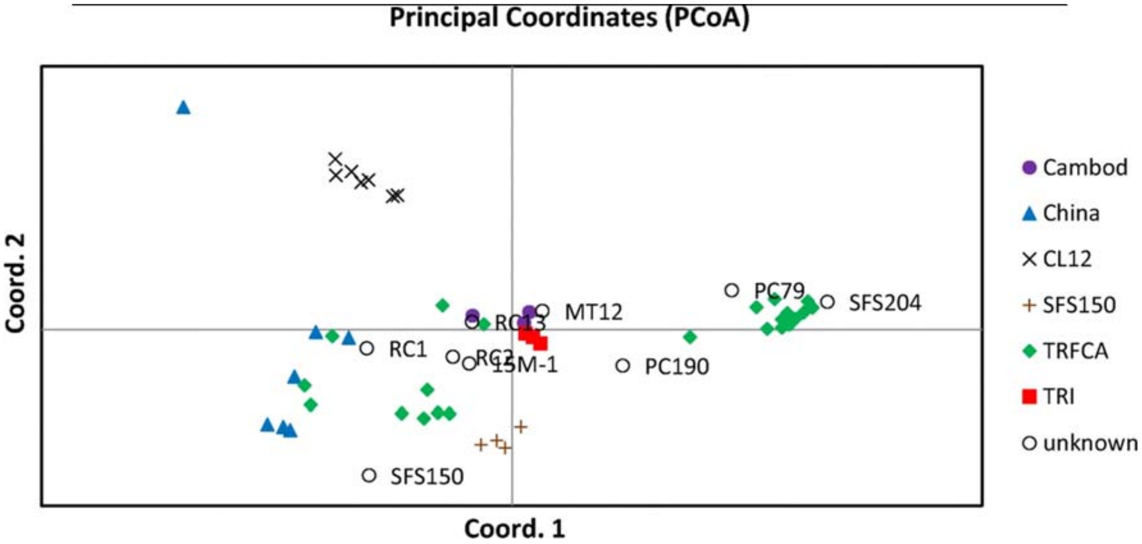


Fig 4. Principal coordinate analysis of 56 tea accessions based on 16,382 DArT markers. Accessions were coloured by origin and parents (The progeny of CL12 and the progeny of SFS150 are indicated as × and +, respectively). Cultivar names were added for cultivars of unknown origin

Table 3 Pairwise estimates of the genetic distance between TRFCA germplasm groups

	TRFCA	China	Unknown	Cambod	TRI
TRFCA	0.00				
China	0.06	0.00			
Unknown	0.07	0.13	0.00		
Cambod	0.08	0.15	0.13	0.00	
TRI	0.09	0.15	0.13	0.14	0.00

The AMOVA revealed that there was a higher proportion of genetic variation (94%) within the groups than among the groups (6%). Nei's unbiased genetic distances among groups are illustrated in Table 3. The TRFCA had a low genetic distance from the other four groups (Cambod type, China type, TRI and unknown). This confirms the

PCoA inference, that the TRFCA has a high genetic admixture of Cambod type, China type and Assam type. The remaining four groups showed a relatively high genetic distance between each other.

A model-based clustering method for deducing population structure using DArT-seq markers was implemented (Fig. 5). The best value for K was calculated to be two (Suppl. Fig. 2). Population structure indicating cultivar identity number can be found in Suppl. Fig. 3. The cultivar name and the traits measured can be identified by referring to the identification number in Suppl. Table 2.

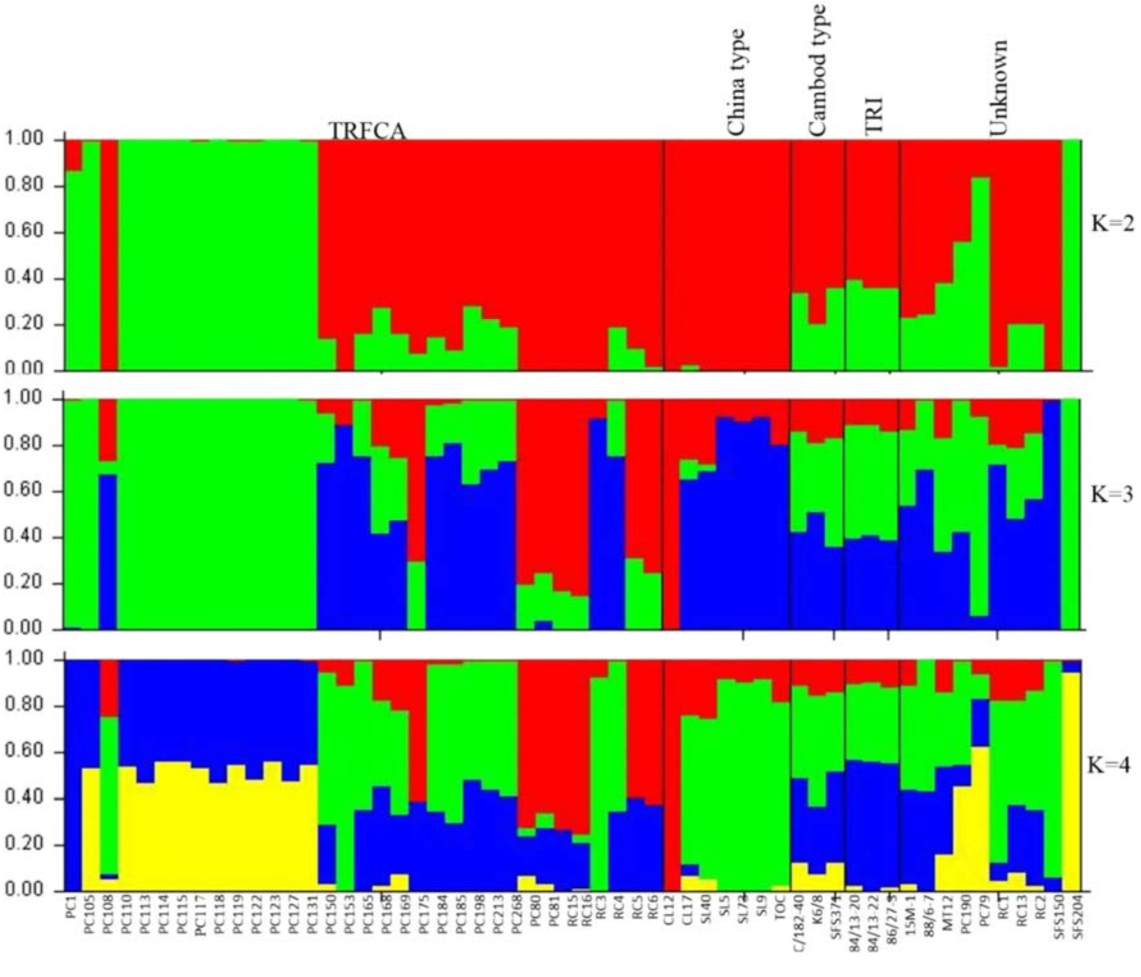


Fig 5. Genetic structure of 56 tea cultivars as inferred by the programme STRUCTURE at $K = 2$, 3 and 4. The accessions are grouped along the horizontal axis

At K=4 CL12 (China type) and PC1 were identified as non-admixed cultivars. These two cultivars were brought into the TRFCA in 1956 when the tea breeding programme was established. CL12 is a known parent of seven cultivars in this TRFCA group. There was no clear observation of the population structure for the TRFCA cultivars, the 33 TRFCA derived cultivars were admixtures. Admixed cultivars were also identified in the ten cultivars from the unknown group. The study confirms that tea is highly heterogeneous (Yang et al. 2016). The following cultivars of unknown origin, were assigned to groups: MT12, RC13 and 15M-1 were closely related to the Cambod type group. Further classification of the following TRFCA cultivars was carried out using the structure: RC3 and PC153 were closely related to the China type group (Fig. 5).

The DArT-seq and RAD-seq technologies have an advantage in saving time when compared to AFLP, RAPD and SSR marker technologies. DArT-seq technology produces a higher number of markers than the RAD-seq (SNP marker) technology (Alam et al. 2018; Yang et al. 2016). The cost of marker development of DArT-seq and RAD-seq technologies markers is similar. Due to the higher number of DArT-seq markers produced, the average cost per data point is less than SNP markers.

Conclusion

Recent studies have looked at identifying more markers in *Camellia sinensis* and characterising the genetic relationship between tea cultivars in Africa (Wambulwa et al., 2016; Wambulwa et al. 2016b). An increased number of markers across the genome would result in a better representation of the genome during germplasm analysis. The genetic relationship among the 56 cultivars was characterised using DArT-seq markers. The present study highlights the advantages of the DArT-seq marker system for diversity analysis in breeding or genebank materials for crop improvement as well as for estimating the diversity of the germplasm collections. The markers that have been used for genetic relationship studies in the past are not only limited in number (hundreds of markers assessed) but are time-consuming, costly and low-throughput markers. DArT-seq technology stands as an appropriate and cost-effective system to discover hundreds of polymorphic genomic loci, scoring thousands of unique genomic-wide DNA fragments in one single experiment, without requiring existing DNA sequence information. This method would be useful in analysing a closely related germplasm where more markers would are required. The genetic relationship between the cultivars that were not classified previously (due to lost records or products of open pollination) has been inferred.

The present results contribute to understanding the existing genetic diversity, relationship and population structure among the TRFCA germplasm cultivars. Such information is crucial to tea breeding programs worldwide, in meeting the future demands of future breeding programs as well as in formulating effective conservation strategies for genetic diversity within cultivars. This method, illustrated with the TRFCA collection, may be applicable to different crop germplasms worldwide.

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

Acknowledgements

The first author acknowledges the financial assistance of the Carnegie Regional Initiative in Science and Education (Carnegie-RISE) through the Southern African Biochemistry and Informatics for Natural Products (SABINA) network. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF, Carnegie-RISE or SABINA.

Data archiving statement

The sequences of the Dart-seq markers have been deposited at GenBank under the accession KCRD00000000. The version described in this paper is the first version, KCRD01000000.

References

Alam M, Neal J, O'Connor K, Kilian A, Topp B (2018) Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. PloS one 13(8):e0203465

Bachmann B, Luke W, Hunsmann G (1990) Improvement of PCR amplified DNA sequencing with the aid of detergents. Nucleic Acids Res 18:1309

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376. doi: 10.1371/journal.pone.0003376 [doi]

Balasaravanan T, Pius P, Kumar RR, Muraleedharan N, Shasany A (2003) Genetic diversity among south Indian tea germplasm (*Camellia sinensis*, *C. assamica* and *C. assamica* spp. *lasiocalyx*) using AFLP markers. Plant Sci 165:365-372

Barcaccia G (2009) Molecular markers for characterizing and conserving crop plant germplasm. In: Jain SM, Brar DS (eds.) Molecular Techniques in Crop Improvement. Springer, Netherlands, pp 231-254

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Rev Genet 12:499-510

Earl DA (2012) Structure Harvester: a website and program for visualizing Structure output and implementing the Evanno method. Conserv Genet Resour 1;4(2):359-61

Ellis R, Nyirenda H (1995) A successful plant improvement programme on tea (*Camellia sinensis*). Exp Agric 31:307-323

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. doi: 10.1371/journal.pone.0019379 [doi]

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of cultivars using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620

Fang W, Cheng H, Duan Y, Jiang X, Li X (2012) Genetic diversity and relationship of clonal tea (*Camellia sinensis*) cultivars in China as revealed by SSR markers. Plant Syst Evol 298:469-483

Freeman S, West J, James C, Lea V, Mayes S (2004) Isolation and characterization of highly polymorphic microsatellites in tea (*Camellia sinensis*). Mol Ecol Notes 4:324-326

Li H, Vikram P, Singh RP, Kilian A, Carling J, Song J, Burgueno-Ferreira JA, Bhavani S, Huerta-Espino J, Payne T, Sehgal D (2015) A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. BMC genomics. 16(1):216

Lui K, Muse S (2005) PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics 21:2128-2129

Mogga M, Sibiya J, Shimelis H, Lamo J, Yao N (2018) Diversity analysis and genome-wide association studies of grain shape and eating quality traits in rice (*Oryza sativa L.*) using DArT markers. PloS one 13(6):e0198012

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A 70:3321-3323

Paul S, Wachira F, Powell W, Waugh R (1997) Diversity and genetic differentiation among populations of Indian and Kenyan tea (Camellia sinensis (L.) O. Kuntze) revealed by AFLP markers. Theor Appl Genet 94:255-263

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288-295

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. Bioinformatics 28:2537-2539. doi: bts460 [pii]

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Roy S, Chakraborty B (2009) Genetic diversity and relationships among tea (*Camellia sinensis*) cultivars as revealed by RAPD and ISSR based fingerprinting. Indian Jl of Biotechnol 8:370-376

Sanderson GW (1963) The chloroform test. A study of its suitability as a means of rapidly evaluating fermenting properties of clones. Tea Quarterly 34:193-196

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. 5:P54

Simko I, Eujayl I, van Hintum TJ (2012) Empirical evaluation of DArT, SNP, and SSR marker-systems for genotyping, clustering, and assigning sugar beet hybrid varieties into populations. Plant Sci 184:54–62 pmid:22284710

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739. doi: 10.1093/molbev/msr121 [doi]

Tanaka J (2012) Japanese tea breeding history and the future perspective. In: Chen L, Apostolides Z and Chen Z (eds) Global Tea Breeding. Springer Heidelberg, New York, pp 227-239

Wambulwa MC, Meegahakumbura MK., Chalo R, Kamunya S, Muchugi A, Xu JC, Liu J, Li D, Gao LM (2016) Nuclear microsatellites reveal the genetic architecture and breeding history of tea germplasm of east Africa. Tree Genet Genomes 12:1-10

Wambulwa MC, Meegahakumbura MK, Kamunya S, Muchugi A, Möller M, Liu J, Xu JC, Ranjitkar S, Li DZ, Gao LM (2016) Insights into the Genetic Relationships and Breeding Patterns of the African Tea Germplasm Based on nSSR Markers and cpDNA Sequences. Front Plant Sci 7

Wambulwa MC, Meegahakumbura MK, Kamunya S, Muchugi A, Möller M, Liu J, Xu JC, Li DZ, Gao LM (2017)Multiple origins and a narrow genepool characterise the African tea germplasm: concordant patterns revealed by nuclear and plastid DNA markers. Sci Rep 7(1):4053

Wheeler DS, Wheeler WJ (2004) The medicinal chemistry of tea. Drug Dev Res 61:45-65

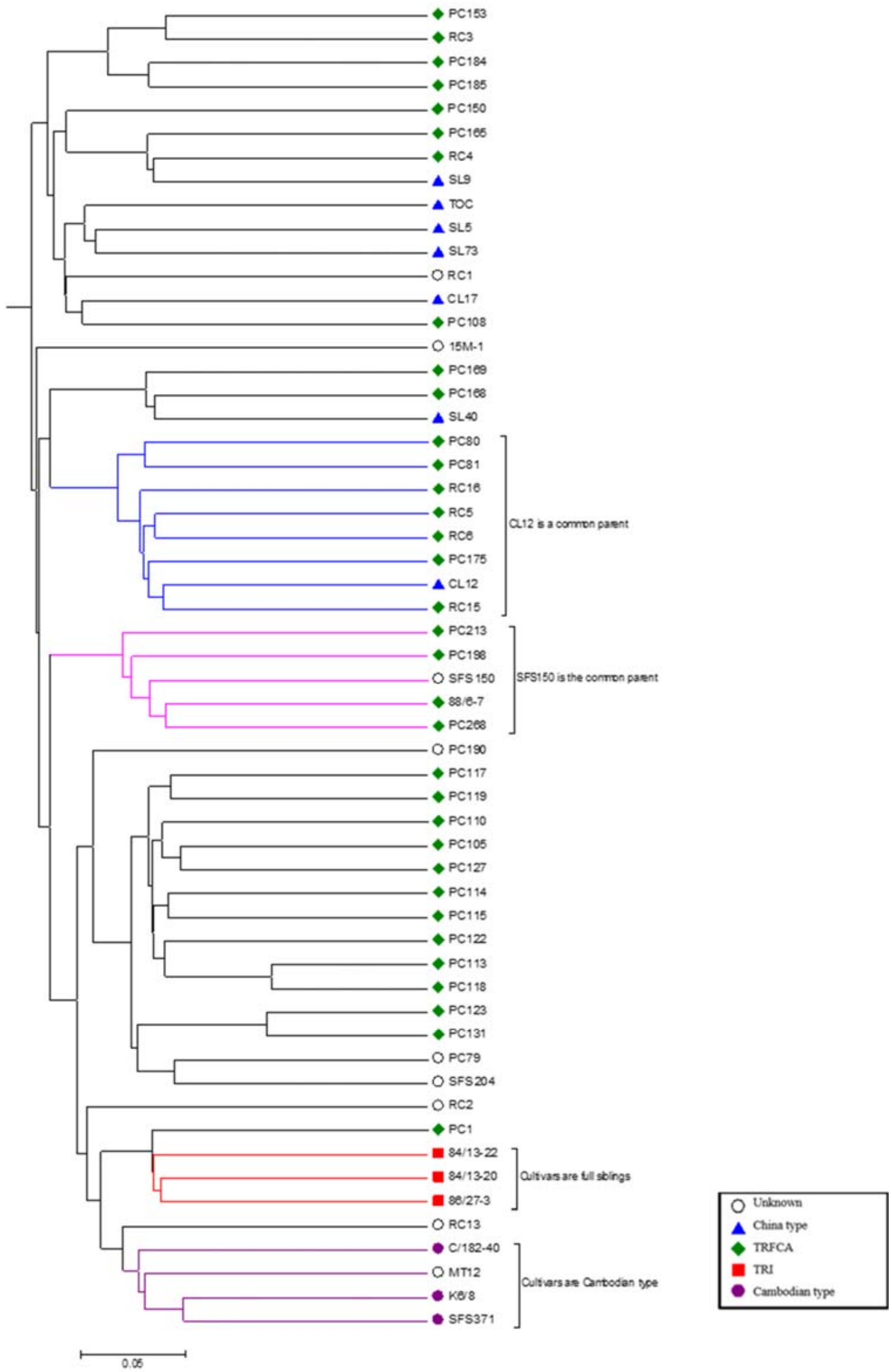
Wium M (2009) Characterization of genetic diversity in selected cultivars and identification of a possible molecular marker for drought tolerance in tea *Camellia sinensis* (L.) *O. Kuntze*. Dissertation, University of Pretoria

Xia EH, Zhang HB, Sheng J, Li K, Zhang QJ, Kim C, Zhang Y, Liu Y, Zhu T, Li W, Huang H (2017) The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Mol Plant 10(6):866-77

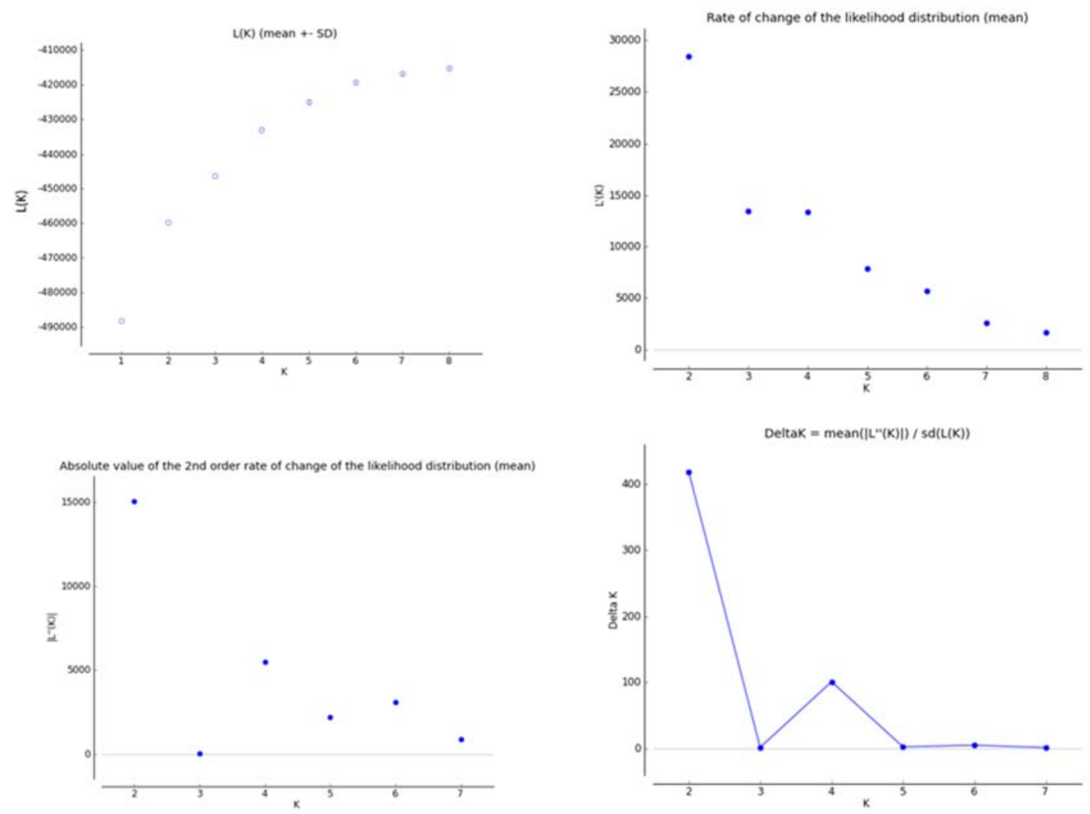
Yang H, Wei CL, Liu HW, Wu JL, Li ZG, Zhang L, Jian JB, Li YY, Tai YL, Zhang J, Zhang ZZ (2016) Genetic divergence between *Camellia sinensis* and its wild relatives revealed via genome-wide SNPs from RAD sequencing. PLoS One 11(3):e0151424

Zaitoun SY, Jamous RM, Shtaya MJ, Mallah OB, Eid IS, Ali-Shtayeh MS (2018) Characterizing Palestinian snake melon (*Cucumis melo* var. *flexuosus*) germplasm diversity and structure using SNP and DArTseq markers. BMC Plant Biol 18(1):246

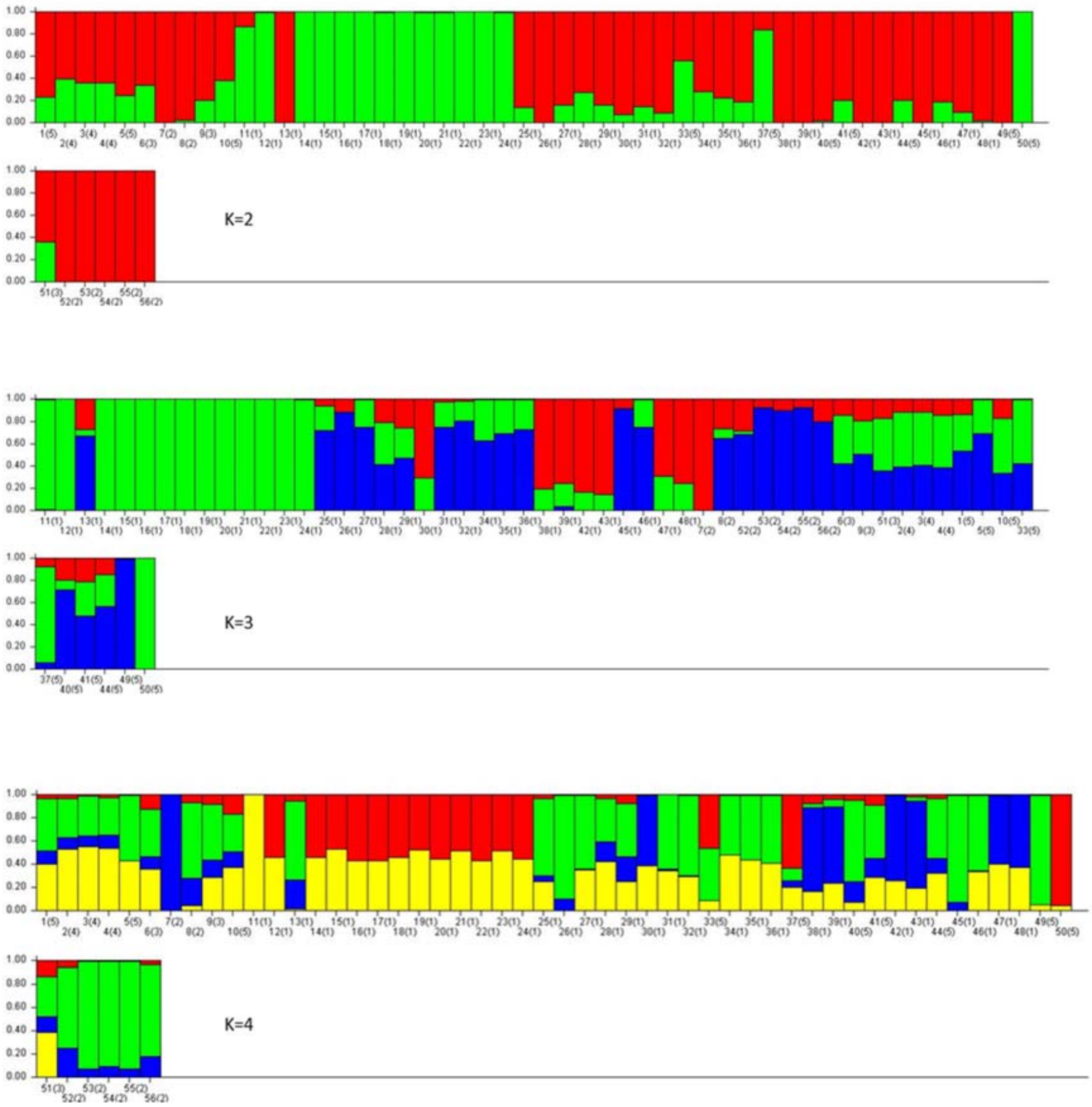
Supplementary Figures and Tables



Suppl. Fig. 1 The genetic relationship among 56 cultivars using the UPGMA method. The dendrogram was generated using 16 382 DArT-seq markers



Suppl. Fig. 2 Estimation of the optimal number of groups K (±SD).



Suppl. Fig. 3 Genetic structure of 56 tea cultivars as inferred by STRUCTRE at K=2, 3 and 4. The accessions are grouped (1-5) along the horizontal axis. Each cultivar is represented by a single bar broken into K coloured segments. The numbers 1-56 represent the individual cultivars. The numbers in brackets correspond to the populations: (1) = TRFCA, (2) = China type, (3) = Cambod type, (4) = Full siblings from TRI and (5) = unknown

Click on this link to be redirected to the Suppl. Table 1 summary statistics table for the 16 382 DArT-seq loci for 56 TRFCA tea samples:

https://www.dropbox.com/s/0ei6hfkkm3s1gg9/Summary%20Statistics%20DArT.xlsx?dl=0

Suppl. Table 2 A list of the individual cultivars, the groups and phenotypic data (mean yield, rate of fermentability and drought tolerance)

Cultivar	Group	Yield	Fermentability	Drought Tolerance
15M-1	5 unknown	3593	1	1
84/13-20	4 TRI full siblings	3176	1	1
84/13-22	4 TRI full siblings		3	1
86/27-3	4 TRI full siblings	2825	2	1
88/6-7	5 unknown	3236	1	2
C/182-40	3 Cambod Type	1239	2	3
CL12	2 China Type	2111	3	2
CL17	2 China Type	1472	3	2
K6/8	3 Cambod Type	1670	1	3
MT12	5 unknown	2562	2	1
PC1	1 TRFCA	2728	1	3
PC105	1 TRFCA	2244	1	3
PC108	1 TRFCA	2825	1	3
PC110	1 TRFCA	2393	1	3
PC113	1 TRFCA	2144	1	3
PC114	1 TRFCA	2547	1	3
PC115	1 TRFCA	2129	1	3
PC117	1 TRFCA	3148	1	3
PC118	1 TRFCA	2424	1	3
PC119	1 TRFCA	1909	1	3
PC122	1 TRFCA	4109	1	1
PC123	1 TRFCA	4560	1	1
PC127	1 TRFCA	1865	2	2
PC131	1 TRFCA	1733	1	3
PC150	1 TRFCA	3503	2	3
PC153	1 TRFCA	2836	2	1
PC165	1 TRFCA	3567	2	1
PC168	1 TRFCA	3773	1	1
PC169	1 TRFCA	2191	3	1
PC175	1 TRFCA	3861	2	1
PC184	1 TRFCA	3556	1	3
PC185	1 TRFCA	4657	2	1
PC190	5 unknown	2835	2	2
PC198	1 TRFCA	3946	2	1
PC213	1 TRFCA	4625	2	1
PC268	1 TRFCA	4200	1	2
PC79	5 unknown	2222	1	3
PC80	1 TRFCA	1881	1	3
PC81	1 TRFCA	2548	2	3
RC1	5 unknown	3845	3	1
RC13	5 unknown	3611	3	1
RC15	1 TRFCA	3166	3	1
RC16	1 TRFCA	3624	3	1

RC2	5 unknown	2801	3	2
RC3	1 TRFCA	3443	3	1
RC4	1 TRFCA	3577	2	1
RC5	1 TRFCA	3251	3	2
RC6	1 TRFCA	3677	3	1
SFS150	5 unknown	4094	2	2
SFS204	5 unknown	2452	1	3
SFS371	3 Cambod Type	1031	1	3
SL40	2 China Type	1163	3	2
SL5	2 China Type		3	2
SL73	2 China Type	2034	3	2
SL9	2 China Type	2272	3	2
TOC	2 China Type	3606	3	2