International Journal
of
Research in Agronomy

E-ISSN: 2618-0618
P-ISSN: 2618-060X
© Agronomy
www.agronomyjournals.com
2020; 3(2): 09-21
Received: 12-05-2020
Accepted: 16-06-2020

Christopher Nyarukowa
Department of Biochemistry, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa

Mari van Reenen
Human Metabolomics, North-West University (Potchefstroom Campus), South Africa

Robert Koech
Kenya Agriculture and Livestock Research Organisation, Tea Research Institute, P.O. Box 820-20200 Kericho, Kenya

Samson Kamunya
Kenya Agriculture and Livestock Research Organisation, Tea Research Institute, P.O. Box 820-20200 Kericho, Kenya

Richard Mose
James Finlay (Kenya) Limited, P.O. Box 223, Kericho 20200, Kenya

Zeno Apostolides
Department of Biochemistry, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa

Corresponding Author:
Zeno Apostolides
Department of Biochemistry, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa

Multivariate models for identification of elite mother bushes with high commercial potential for black tea from mature seedling fields of *Camellia sinensi*

Christopher Nyarukowa, Mari van Reenen, Robert Koech, Samson Kamunya, Richard Mose and Zeno Apostolides

DOI: https://doi.org/10.33545/2618060X.2020.v3.i2a.33

Abstract

Tea producers are in demand of new high yielding cultivars, which produce high quality tea liquors. To breed for these phenotypic traits is challenging due to their polygenic disposition and influence by environment. Two *C. sinensis* populations, namely Comm cultivars from open pollinated field selections, and NComm cultivars from the reciprocal cross of two parents were used. These cultivars were employed to identify the metabolites responsible for distinguishing Comm cultivars, with high yield, high quality and DT from NComm cultivars that did not show these traits. PCA and PLS-DA models were constructed on UPLC/DAD data, which showed clear separation between the Comm and NComm cultivars. CHAID decision trees constructed aimed to classify the 303 genotypes as either Comm or NComm cultivars using subset of compounds. Breeders can predict the quality of new selections from mature seedling fields by employing CHAID decision trees, or the CAF/EC ratio, as predictors.

Keywords: *Camellia sinensis;* catechin; metabolomics; theaflavin

Introduction

Tea (*Camellia sinensis*) is one of the most widely consumed beverages across the world (Hicks, 2009) [14]. The crop which originated in China is grown certain regions of Asia (India, China, Sri Lanka and Japan), Africa (Kenya, Uganda, and Malawi), and Latin America (Argentina). The tea beverage is prepared by brewing or boiling the dried tea leaves in water. Kenya is the world's third largest producer of tea after India and China though it is the leading exporter of black Crush Tear and Curl (CTC) tea (Elbehri *et al.*, 2015) [8]. The tea industry therefore contributes significantly to Kenya's economy by contributing over 26% and 4% of total foreign exchange earnings and Gross Domestic Product (GDP), respectively (Kenya National Bureau of Statistics, 2012) [23]. Tea producers are in demand of new cultivars, which are high yielding, drought tolerant, and produce high quality tea liquors. Tea gets its distinctive astringent and somewhat bitter taste from caffeine (Horie *et al.*, 1997) [17], even though several other metabolites such as the catechins (catechin (CAT), epicatechin (EC), epicatechin gallate (ECg), epigallocatechin (EGC), and epigallocatechin gallate (EGCg)) and all other polyphenols, carbohydrates, and amino acids are influential in its overall taste and aroma (Adkins *et al.*, 2007; Nyarukowa *et al.*, 2016) [1]. The amino acid theanine, which makes up approximately two-thirds of a tea leaf's total free amino acids content, is with other less abundant amino acids, responsible for the sweet and brothy "umami" taste of green tea (Vuong *et al.*, 2011). However, it is noteworthy to indicate that the metabolite composition, which influences tea quality, varies between green and black tea. Unlike green tea, whose quality depends on amino acids, particularly theanine, catechins and caffeine, the quality of black tea depends on theaflavins (theaflavin (TF1), theaflavin-3-gallate (TF2), theaflavin-3'-gallate (TF3), and theaflavin-3,3'-digallate (T4)), thearubigins, catechins and caffeine (Le Gall *et al.*, 2004) [26]. The four TFs are formed during black tea processing by oxidation of green tea catechins in presence of polyphenol oxidase as shown: (1) EC + EGC = TF1; (2) EC + EGCg = TF2; (3) ECg + EGC = TF3; (4) ECg + EGCg = TF4. This therefore indicates that the green leaf catechins are important and thus tea cultivars rich in catechins are likely to produce higher quality teas (Takemoto and Takemoto, 2018) [38].

The employment of seeds obtained from Assam, India, saw the beginning of improvements in Kenya's tea breeding programmes, which brought about the establishment of the initial two polyclonal seed baries at Kangaita and Timbilil (Anon, 1990) [2] following the 1980 formation of Tea Research Foundation of Kenya (TRFK), now known as the Tea Research Institute (TRI). Other large tea producing companies such as James Finlay (Kenya) and George Williamson (Kenya) followed and instituted programmes that saw the establishment of their own improved seed baries.

Traditionally tea breeding, involved selecting of vigorous growing plants, uprooting them from the wild forest, or seedling tea fields and planting in a separate seed garden, called a seed barie, away from slow growing plants. The seeds collected from the seed baries are normally slightly better than seeds collected from seedling gardens with vigorous and slow growing plants. Early studies (Green, 1971) [11] failed to establish reliable correlations between growth and yield properties of mother bushes, and their resultant F_1 progeny clones. In the 1950's vegetative propagation from stem cuttings became possible for tea (Banerjee, 1992) [3]. Subsequent studies (Nyirenda, 1991) [31] have shown adequately strong correlations between the mother bush area, shoot number, and yield of their vegetative propagated clones. A strong positive correlation has also been observed (Shanmugarajah *et al.*, 1991) [37] between clones and their mother bush height, leaf area, stem girth, and stem dry weight in matured seedling fields. All mature seedling tea fields are pruned on a four or five year cycle. The tea breeder normally selects only 100 bushes every year that recover quickly from the prune and meet several criteria e.g. good bush shape, leaf pose, DT and termite resistance, among other traits. These elite mother bushes are believed to be high yielders. Stem cuttings are used to propagate each of the 100 mother bushes into 15-bush observation plots, called clones. The limit of 100 yearly selections is due to the high cost establishing and of maintaining the 15-bush plots. The yield of each clone is measured after five years. Black tea is produced from each of the ten highest yielding clones selected each year, and the tea quality is scored by expert tea tasters. Normally, only one or two of the 100 selected mother bushes produce clones with high yield and good taste. The clones with high yield and good quality are advanced to further field trials and if suitable, are released to the commercial growers. The success rate, from the 100 mother bushes until release to commercial growers is about 1%. Initially, mass selection was employed as tea improvement method, proving a success, to an extent. It however, failed to generate a robust type of tea, possessing satisfactory cup attributes and plant morphological consistency. The developed progenies had not been specifically chosen for their high quality and yielding traits, and as such the resultant seedlings were a mixture of miscellaneous and mediocre genotypes (Wachira, 2001) [41]. Plant breeders have been finding it daunting to develop high yielding tea clones from seedling mother bushes. Our aim is to develop new methods with molecular markers, for selecting mother bushes to increase this success rate.

The effects of global warming, fluctuations in weather patterns are being observed in Kenya, particularly the increased temperatures, leading to prolonged drought spells in the tea growing regions (Elbehri *et al.*, 2015) [8]. Due to these changes in the climate, tea production is likely to be drastically reduced because of a shortage of suitable lands at lower altitudes and the result of this is that farmers have to seek lands at higher, dryer altitudes most of which are occupied by conservation forests. Moreover, evidence has been furnished, over the course of the past 30 years, that temperatures in tea growing regions have been increasing at a rate of 0.2°C per decade (Cheserek *et al.*, 2015) [7]. In addition to this, stresses concomitant with temperature fluctuations in tea producing areas such as Kericho, Kisii, and Nandi, have added to the tea production limitations in Kenya. Tea production is also reliant on well distributed rains; a rise or drop in temperatures as a result of the fluctuations in the rainfall patterns, adversely influences the quantity and quality of tea (Chang, 2015) [5]. The cultivation of tea has also been extended to previously deemed marginal and unsuitable tea growing areas further exacerbating tea quality and tolerance to environmental stresses (Owuor *et al.*, 2010) [33].

The insufficient understanding of the genetics involved when breeding for yield and quality is a problem not only for breeders, but for the tea industry as a whole. Currently, the practice of making field selections based on traits such as recovery from prune and leaf pose have a success rate of about 1% when it comes to identifying elite mother bushes that become commercial successes (Chen *et al.*, 2013) [6]. The tea industry is in need of new methods for field selections to increase this success rate. Metabolomics is one approach than can be broadly applied in screening of elite tea lines, evaluation of quality and physiological changes in tea (Jiang *et al.*, 2019). The key to metabolomics research is the employment of analytic tools to comprehensively analyse metabolites. Holistic metabolic profiles have been obtained from intricate animal and plant samples, using high resolution, information-rich powerful spectrometric techniques. Liquid chromatography coupled with mass spectrometry (LC-MS), due to its advancements within the field, is a central technique in metabolomics research (Khan and Mukhtar, 2007) [24], with it being used predominantly in differential profiling and biomarker identification (Theodoridis *et al.*, 2012) [39]. Metabolomics analyses can either employ a targeted or an untargeted approach. The objective of the targeted approach is the identification and quantification of specific metabolites for which pure standards exist to confirm the identities of the metabolites detected in the samples i.e. the chemical properties of the metabolites under investigation are known. Targeted metabolomics is customarily hypothesis driven, while untargeted metabolomics leads to hypothesis generation, which involves assessing all the metabolites in a biological system (Zhou *et al.*, 2012) [48]. LC-MS has been established as predominant favourite targeted profiling technique especially for plant metabolomics studies (Zhou *et al.*, 2012) [48].

In metabolomics, uni- and multivariate statistical techniques are used in combination to help pinpoint variation (e.g. between classes of interest) in datasets that are often large and high-dimensional. The univariate statistical methods used here was the independent samples t-test and Cohen's d effect size. Three multivariate methods were included, principal component analysis (PCA); partial least squares discriminant analysis (PLS-DA) and Chi-square Automatic Interaction Detection (CHAID) decision trees. PCA and PLS-DA are both multivariate methods that project data onto lower dimensional subspaces by summarising variation, making it possible to graphically present large datasets. PCA models are not provided with group or class membership information, while PLS-DA models, though predictive, are complex and often do not generalise well. During the preceding decade, CHAID decision trees gained popularity, as is documented by the trend in peer-reviewed science journals (Miller *et al.*, 2014) [25]. This increase in popularity is attributed to the realisation by researchers of the benefits associated with making use of advanced statistical software packages to perform comprehensive analyses. Decision trees combine inductive reasoning and supervised learning capable of being used for prediction, regression, estimation, data description, visualisation and dimensionality reduction (Milanović, 2016) [27]. CHAID decision trees were constructed to determine the minimum combination of metabolites that can serve as predictors for separating the Comm cultivars from the NComm cultivars. These CHAID decision trees offer a non-algebraic, data partitioning option, becoming a popular alternative to logistic regression, and discriminant analysis in the past two decades (Wilkinson, 1992). Finally, violin plots, that combine box plots with kernel density plots, were used to show original data for key differentiating metabolites.

The objective of this study was to make use of UPLC/DAD generated data to develop CHAID decision trees, to classify the 303 genotypes as either Comm or NComm cultivars. This may then serve in predicting whether a new field selection is likely to become commercialised due to its similarities with the Comm cultivars. This is the first study to use targeted metabolomics to obtain markers which predict commercial potential in *C. sinensis*.

2. Materials and Methods

2.1. Plant material, and UPLC/DAD sample preparation and analysis

The plant material collection, processing and analyses were performed as described in Nyarukowa *et al.*, (2020). Sixty tea clones used by commercial tea growers near the TRI, were identified and designated the Comm cultivars. A further 247 cultivars from the populations TRFK St.504 and TRFK St. 524 were used and designated the NComm cultivars. Fresh shoots comprising two leaves and a bud were harvested from the 303 cultivars in June 2018. The fresh shoots were placed in appropriately labelled zip-lock plastic bags, and placed on ice blocks to keep cool; these were processed at the TRI miniature tea factory. Five hundred grams of tea leaves were used to make black tea according to Koech *et al.*, (2018) [30]. Briefly, the leaves were withered to a %relative water content of 50–65% over an 18 hour period before being passed through crush, tear and curl (CTC) rollers till maceration was achieved. Following maceration, the resultant dhool was aerated at 22–26°C for 90 min, and at 100% humidity for enzymatic oxidation (fermentation) to occur. A TeaCraft Ltd bench top fluid-bed drier system was employed for firing the tea, starting at 120°C for 25 min, and subsequently lowered to 100°C for 10 min. The black tea samples were then ground using a coffee grinder, placed in sealed in zip-lock plastic bags and stored in 4°C fridge until UPLC analysis.

2.2. Extraction of catechins, caffeine, and theaflavins

Samples were collected, and metabolites extracted from the tea samples according to ISO14502-2 (2005). Briefly, amounts of 0.200 ± 0.001 g of green or black tea samples were weighed out using a Mettler Toledo model MS204TS/00 analytical balance (Microsep, South Africa) and transferred to 20 ml thick walled glass test tubes, following which five ml volumes of 70:30 MeOH (Merck, South Africa): water (v/v) at 70°C was added to each, stoppered and vortex mixed for ± five seconds before being placed into a 70°C set water bath. After five minutes, the extraction mixtures were removed from the water bath and vortex mixed before being returned for an additional five minutes. The mixtures were vortex mixed a second time, cooled and then centrifuged at 3,500 g using a Thermo Scientific Heraeus Labofuge (Sepsci, South Africa) Model 300 centrifuge for ten minutes. The resultant supernatants were decanted into respective ten ml volumetric flasks and the extraction step repeated once more. The two extracts were then pooled, and the volume adjusted to ten ml with cold 70:30 MeOH: water (v/v). A one ml volume of each extract was diluted to five ml using stabilising solution, which constituted 10% (v/v) acetonitrile in water, 500 µg/ml EDTA and 10 mg/ml ascorbic acid, all purchased from Sigma-Aldrich, South Africa. About 100 µl of each resultant dilution was then filtered through a 0.2 µm Minisart®RC4 syringe filter (Sartorius, South Africa) with hydrophilic, solvent-resistant regenerated cellulose membranes and the samples were then analysed using UPLC/DAD.

2.3. UPLC/DAD analyses

The UPLC/DAD analyses were accomplished on a Waters ACQUITY UPLC H-Class system (Waters, Milford, MA, USA) equipped with a binary solvent delivery pump, an autosampler, and a photodiode array detector and controlled by the Empower-3 software. Separation was attained on a Waters Acquity HSS T3 column (1.8 µm, 2.1 × 150 mm), at 40ºC, with the mobile phase constituted of solvent A, which was 2% acetic acid and 9% acetonitrile in deionised double distilled water, at a pH of 2.8, and solvent B comprised of 2% acetic acid and 80% of acetonitrile in deionised double distilled water. The mobile phases were filtered through a 0.2 µm cellulose acetate membrane filter and degassed using a Neuberger Laboport (Labotech, South Africa) vacuum pump. A gradient elution method was employed: 0 min (5% B), 0-21 min (5-20% B), 21-30 min (20-25% B), 30-32 min (25-100% B), 32-39 min (100-100% B), 39-40 min (100-5% B), and 40-45 min (5-5% B). A sample injection volume of five µl and a 0.2 ml/min flow-rate were employed for analyses. Catechins (CAT, EC, ECg, EGC, and EGCg), caffeine and gallic acid (Sigma-Aldrich, South Africa) were used as standards. Tryptamine, sulfanilamide and mycophenolic acid (Sigma-Aldrich, South Africa) were the QC internal standards; identification and quantification were at 278 nm, with the individual catechins and caffeine in the samples being identified on retention times of the standards, and UV/vis spectra matches.

2.4. Data pre-processing and statistical analysis

The data pre-processing and statistical analyses were performed as described in Nyarukowa *et al.*, (2020). Briefly, variables with over 50% missing values, in both classes, were eliminated. Since missing values were deemed below the quantification threshold of the instrument, the remaining missing values were imputed with random numbers below the minimum observed. Outliers were removed based on PCA scores plots with 95% CIs, after data transformation and scaling.

PCA plots were included as supportive evidence, along with other validation statistics generated by the PLS-DA model, namely predictive accuracy considering unseen cases. In the current context, both methods were used to visualise the data rather than predict group membership. However, VIP (variable importance in projection) values were generated to rank metabolites according to their predictive ability. Metabolites with VIP values greater or equal to 1 are generally considered strong predictors. Univariate statistics were generated to support and supplement multivariate findings. The independent sample t-test was used to assess the statistical significance of differences between group means, after correcting for multiple testing by controlling the false discovery rate using Benjamini & Hochberg's approach as coded by Groppe *et al.*, (2011) [12]. The practical relevance of differences were quantified using Cohen's d-value. Data pre-processing, PCA, PLS-DA and univariate statistics were performed using MATLAB with Statistics Toolbox (2019), version 9.5.0 (R2018b) software (Natick, Massachusetts: The MathWorks Inc) in conjunction with the PLS_Toolbox (2019), version 8.7 software (Wenatchee, WA: Eigenvector Research Inc. Software available at http://www.eigenvector.com). Chi-square Automatic Interaction Detector (CHAID) trees were constructed here using IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp. The dataset was randomly split into training and test sets. The training set was used to construct CHAID trees, while the test set was used to validate the trees' performance. Lastly, are the violin plot, which were created using JMP Pro 15 statistical software, contain similar information as found in a box plot, but have the indisputable advantage over the box plot because they show the entire data distribution, which is beneficial when working with multimodal data i.e. distribution with several peaks was used (Hintze and Nelson, 1998) [16].

3. Results and Discussion

3.1. Violin plots for UPLC/DAD

To visually represent the abundance of metabolites retained after zero-filtering, violin plots were constructed. A good separation was attained between the Comm and NComm cultivars by CAF, CAT, EC, and TF2-TF4 (Figure. 1).

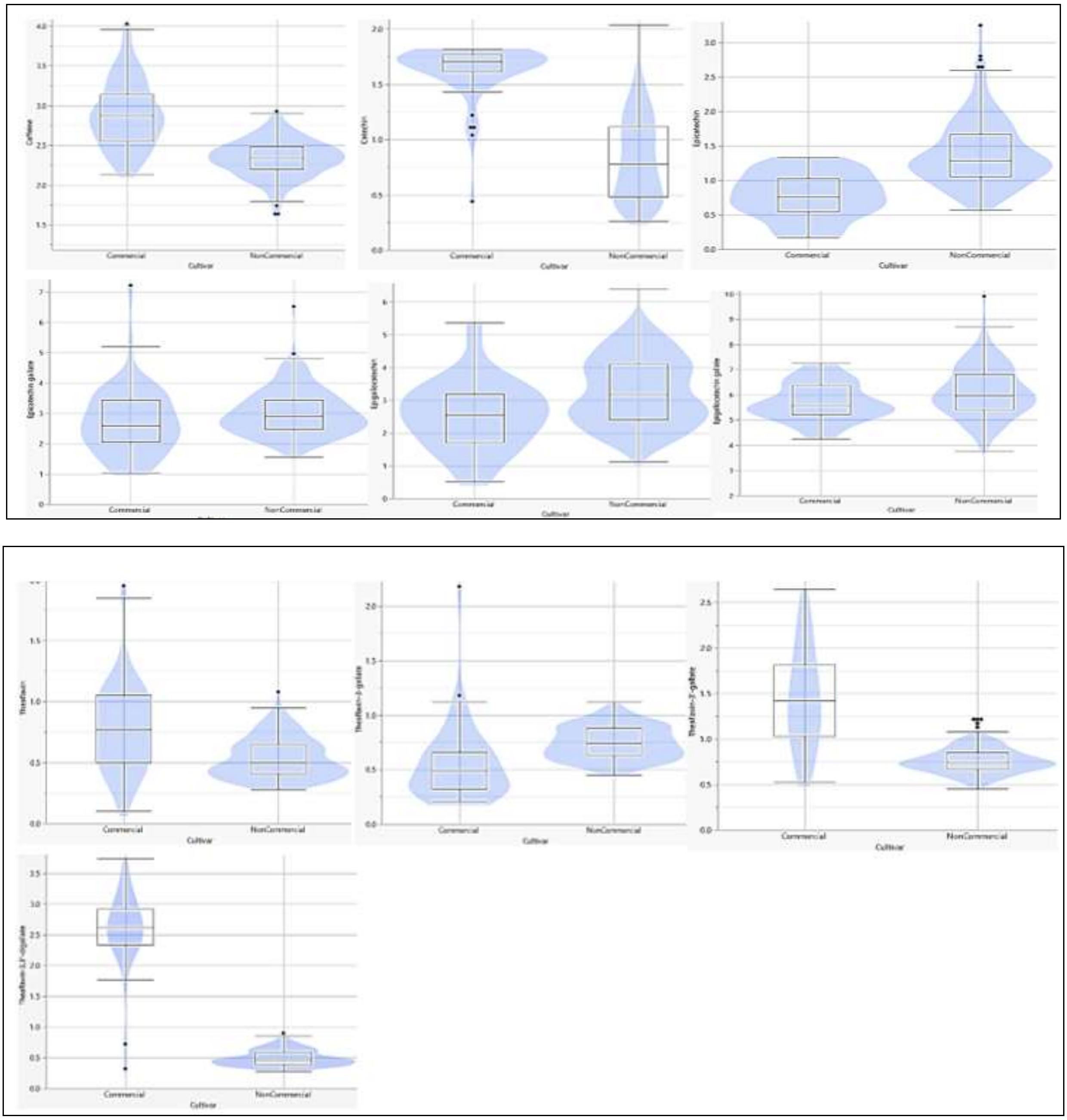


Fig 1: Violin plots showing separation between the Comm and NComm cultivars based on detected metabolites. The y-axis units for the CAF, and the catechins are %w/w dry weight; TF1-TF4 in black tea samples were quantified as EGCg equivalents, based on the EGCg response factor. The black dots represent outliers, which are observations 1.5 x interquartile range (IQR) greater than the 75th quantile or 1.5 x IQR less than the 25th quantile.

3.2. Overview of predictive potential in UPLC/DAD metabolites

PCA and PLS-DA models were used to summarise the variation in the metabolites retained after pre-processing. Scores plots, where each point on the graph represents a sample as projected onto the new lower-dimensional space, were scrutinised to determine the variation between the two groups that can be explained by the measured variables. The PCA plot (Figure 2) indicates that the dominating source of variation can be attributed to the classes in the data. The PLS-DA plot (Figure 3) shows the combined ability of the metabolites to differentiate between classes, thus justifying further investigation for predictive models.

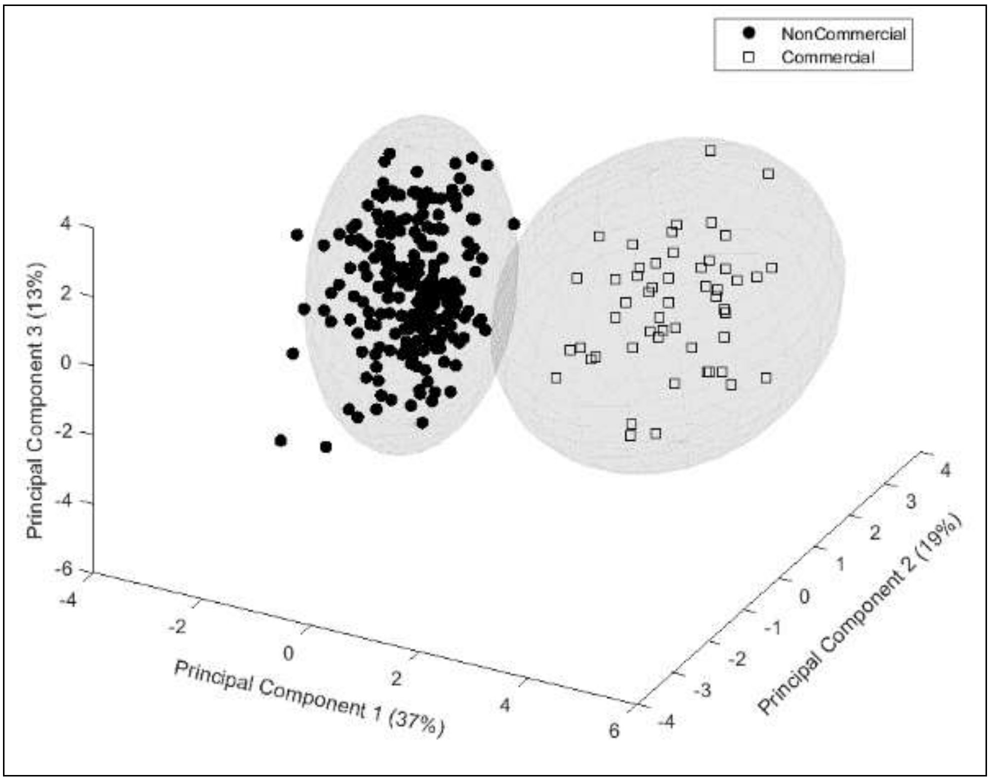


Fig 2: The PCA scores plot for the first three principal components. The plot shows good separation and explaining 69% of the variation observed between the Comm and the NComm cultivars. Ellipsoids represent 95% CI of score centroids of each class. The percentage of the overall variation explained by each component is indicated along each axis.

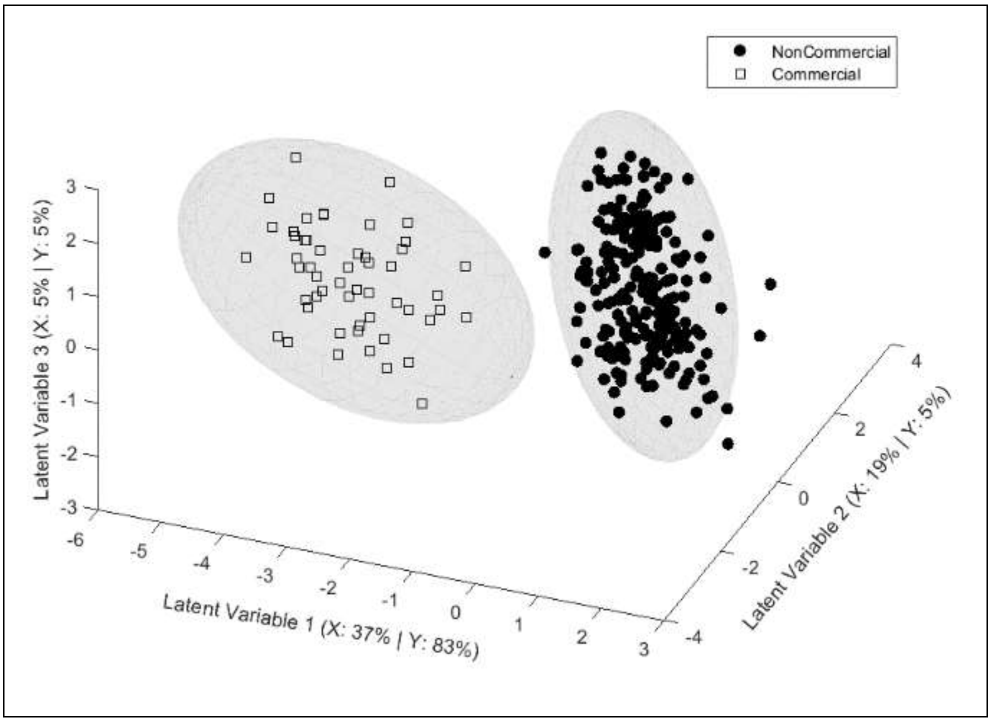


Fig 3: The PLS-DA scores plot for the first three latent variables. The plot shows clear separation between the Comm and the NComm cultivars. The goodness-of-fit values achieved for the UPLC/DAD model was deemed reliable with predictive accuracy R^2=94% and leave-one-out crossvalidated predictive accuracy Q^2=93%. Ellipsoids represent 95% CI of score centroids of each class.

The PLS-DA model provides VIP (variable importance in projection) value that ranks metabolites according to their predictive ability. To further supplement this ranking, univariate statistics were derived and all are summarised in Table 1. To demonstrate the potential of specific metabolites for predictive models was explored further in the next section as shown in Table 1.

Table 1: Ranking of metabolites detected by the UPLC/DAD based on their VIP scores.

Variable	Adjusted p-value	Cohen's d-value	VIP
Theaflavin-3,3'-digallate	< 0.0001	5.75	1.8
Theaflavin-3'-gallate	< 0.0001	1.41	1.2
Catechin	< 0.0001	1.90	1.1
Caffeine	< 0.0001	1.42	1.1
Epicatechin	< 0.0001	1.53	1.0
Theaflavin-3-gallate	< 0.0001	0.73	0.7
Epigallocatechin	< 0.0001	0.69	0.7
Theaflavin	< 0.0001	0.73	0.6
Epigallocatechin gallate	0.002	0.41	0.4
Epicatechin gallate	0.013	0.37	0.4

3.3. Predictive modelling based on UPLC/DAD metabolites

Past studies have demonstrated the applicability of theaflavins as markers for black tea quality (Obanda *et al.*, 1997; Wright *et al.*, 2002) [32, 46]. CHAID decision trees were constructed using the four theaflavin (TF1-TF4) variables. Seventy five percent of the 303 genotypes dataset was used to make up the training sample set on which a CHAID decision trees was developed, with the remaining 25% serving as the test sample set, as shown in Figure 4. Cross validation of these CHAID decision trees is important because, as with stepwise regression, prediction errors for any tree applied to new samples may be higher than those of the training samples on which it was constructed. As such cross validation data should be reserved, when possible (Breiman *et al.*, 1984) [4].

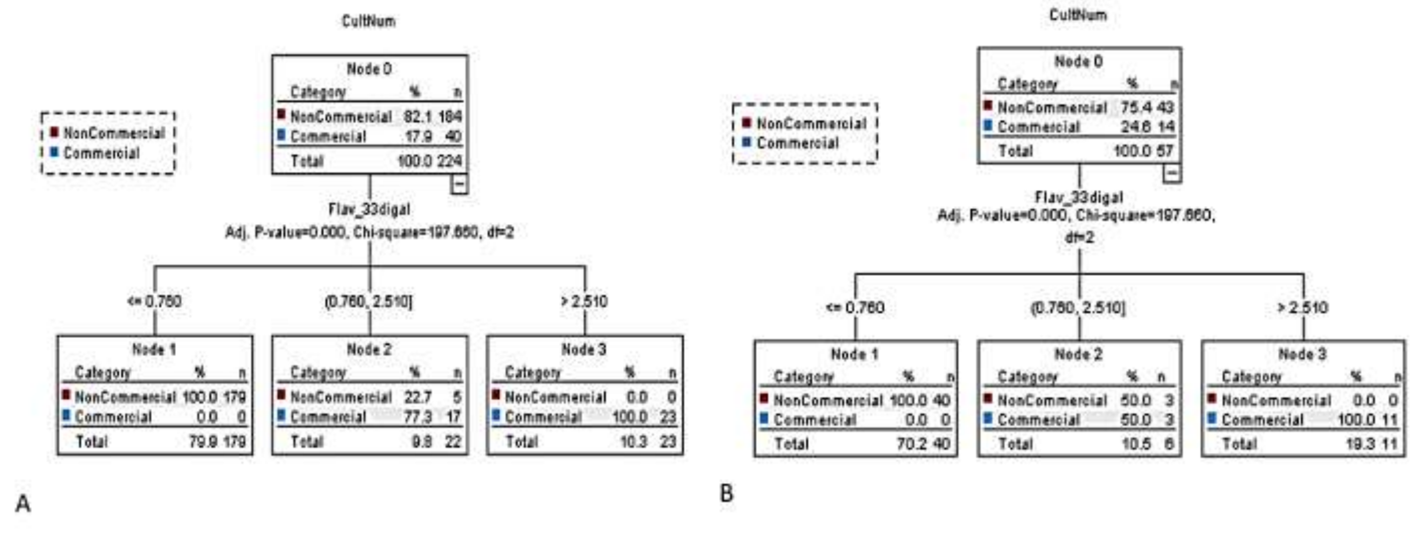


Fig 4: (A) CHAID decision tree – training set, and (B) CHAID decision tree – validation set, based on the four theaflavins variables.

Table 2: Classification accuracy table for CHAID decision tree based on four theaflavins.

Sample	Observed	Predicted		
		Non Commercial	Commercial	Percent Correct
Training	NonCommercial	179	5	97.3
	Commercial	0	40	100.0
	Overall Percentage	79.9	20.1	97.8
Validation	NonCommercial	40	3	93.0
	Commercial	0	14	100.0
	Overall Percentage	70.2	29.8	94.7

Because theaflavins can only be obtained from black tea, which is a laborious and time consuming process, requiring up to five years for a field selection to be propagated from cuttings, grown in a hedge, and produce enough shoots to make black tea, a less laborious solution was sought. CHAID decision trees were constructed from the green leaf analytes. These trees were based on CAF, EC, ECg, EGC and EGCg found in freeze dried green leaf (Figure 5).

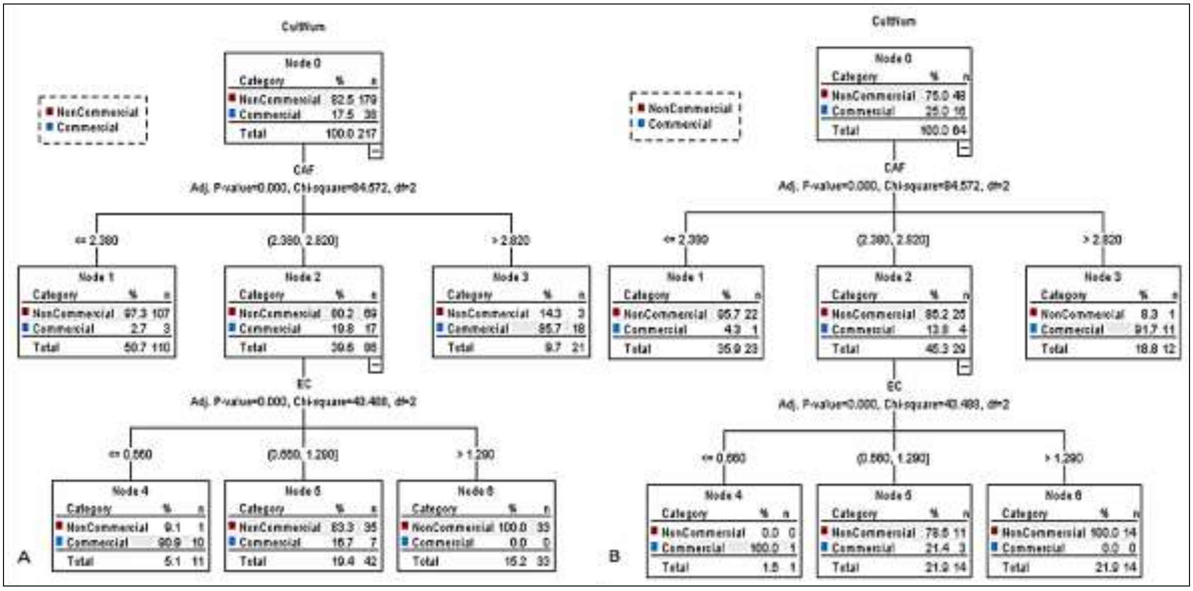


Fig 5: (A) CHAID decision tree – training set, and (B) CHAID decision tree – testing set, based on CAF, EC, ECg, EGC and EGCg.

Table 3: Classification accuracy table for CHAID decision tree based on CAF, EC, ECg, EGC, and EGCg.

Sample	Observed	Predicted		
		Non Commercial	Commercial	Percent Correct
Training	NonCommercial	175	4	97.8
	Commercial	10	28	73.7
	Overall Percentage	85.3	14.7	93.5
Validation	NonCommercial	47	1	97.9
	Commercial	4	12	75.0
	Overall Percentage	79.7	20.3	92.2

Figure 5 shows the CHAID decision tree developed on CAF and the four catechins. The tree, and the accuracy table (Table 3) show that 75% (12/16) of the Comm cultivars were correctly classified in the validation set. Considering that it is a very cumbersome process to manufacture black tea to obtain theaflavins, taking as much as 5 years for the bushes to grow before enough leaves can be harvested, the model making use of the green leaf CAF and four catechins correctly predicted 75% of the Comm cultivars in the validation set as Comm. This saves the tea breeder up to four years, and the labour and resources of cultivating the tea bushes for five years only to learn it is a low yield, drought susceptible and a low quality field selection, and will not be commercialised. From the CHAID tree results, it can be seen that CAF and EC are the important variables that can serve as predictors for distinguishing between Comm and NComm cultivars. A scatter plot of CAF vs EC, the compounds selected by the CHAID decision tree, graphically displays their combined potential to differentiate between Comm and NComm cultivars (Figure 6). The CHAID decision tree in Figure 5 excluded CAT as a variable. The reason for this is that the CAT peak is small and elutes close to two unknown metabolites, which may make it difficult to accurately identify and quantify, especially on columns with lower resolution ability. This can be seen in the chromatogram in Figure 7.

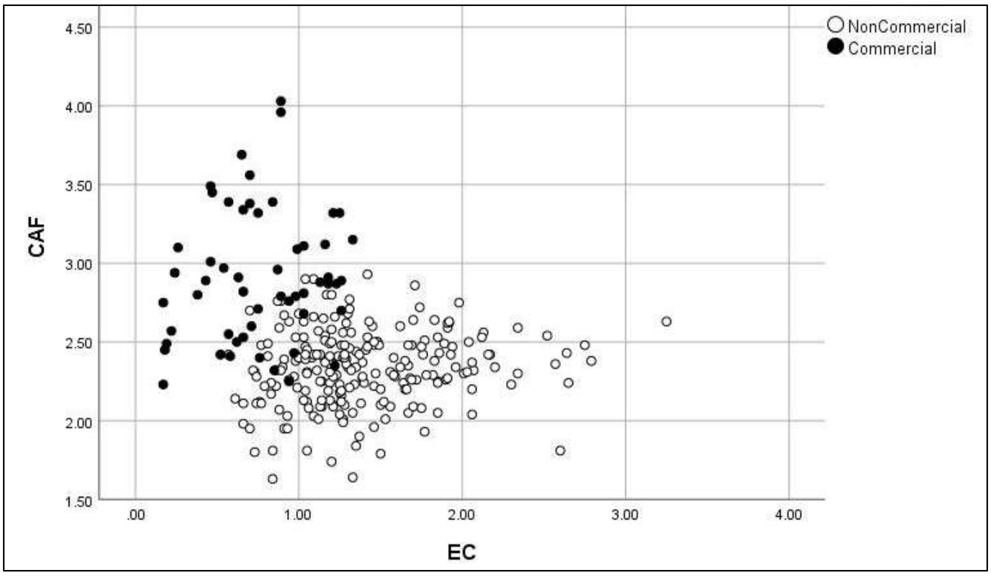
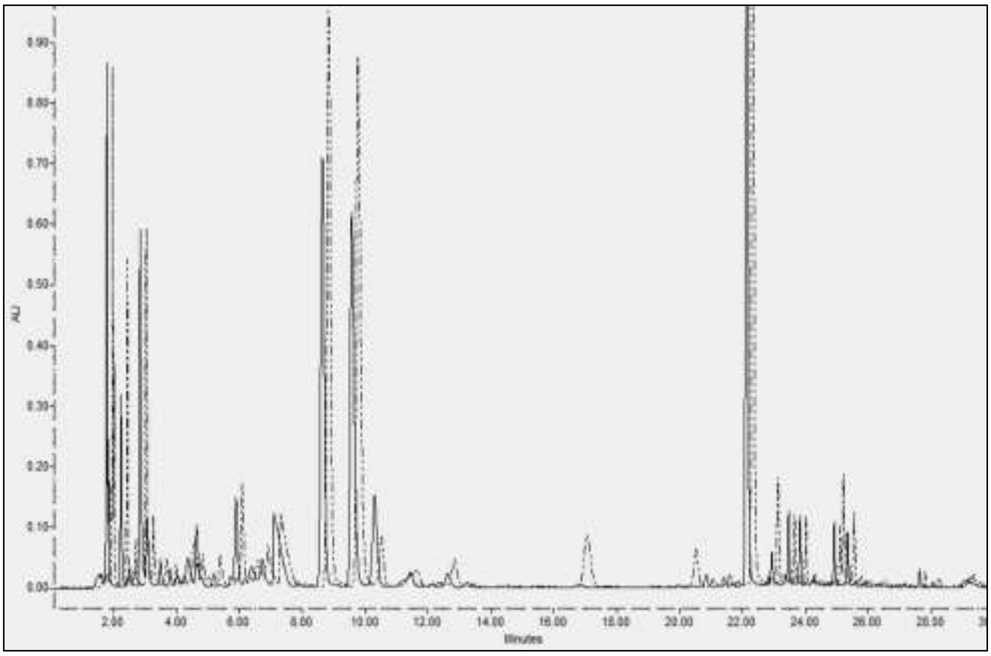
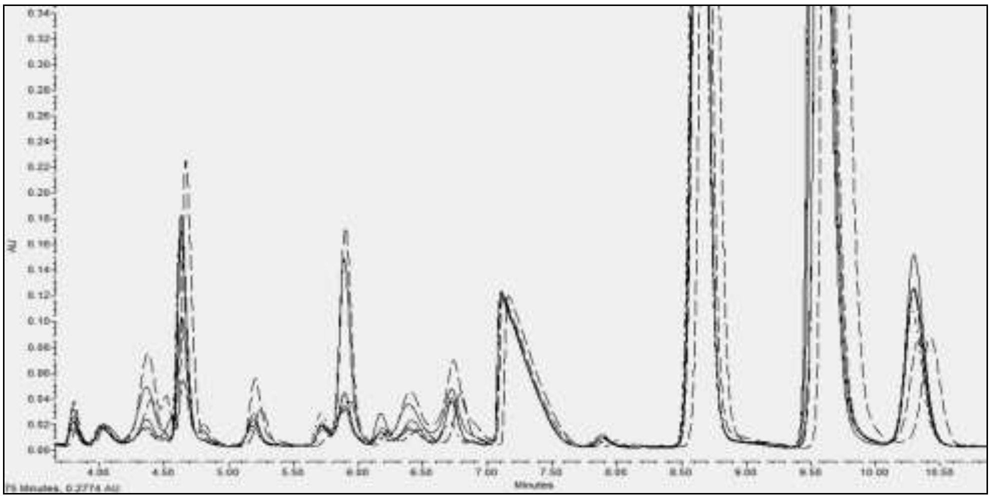


Fig 6: Scatter plot showing the distribution of Comm and NComm cultivars based on %w/w CAF vs EC

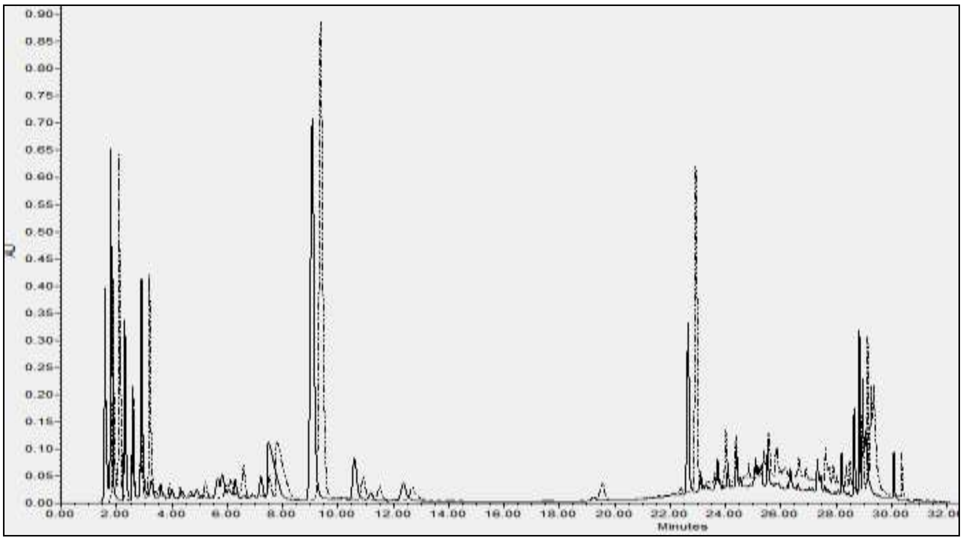


A.

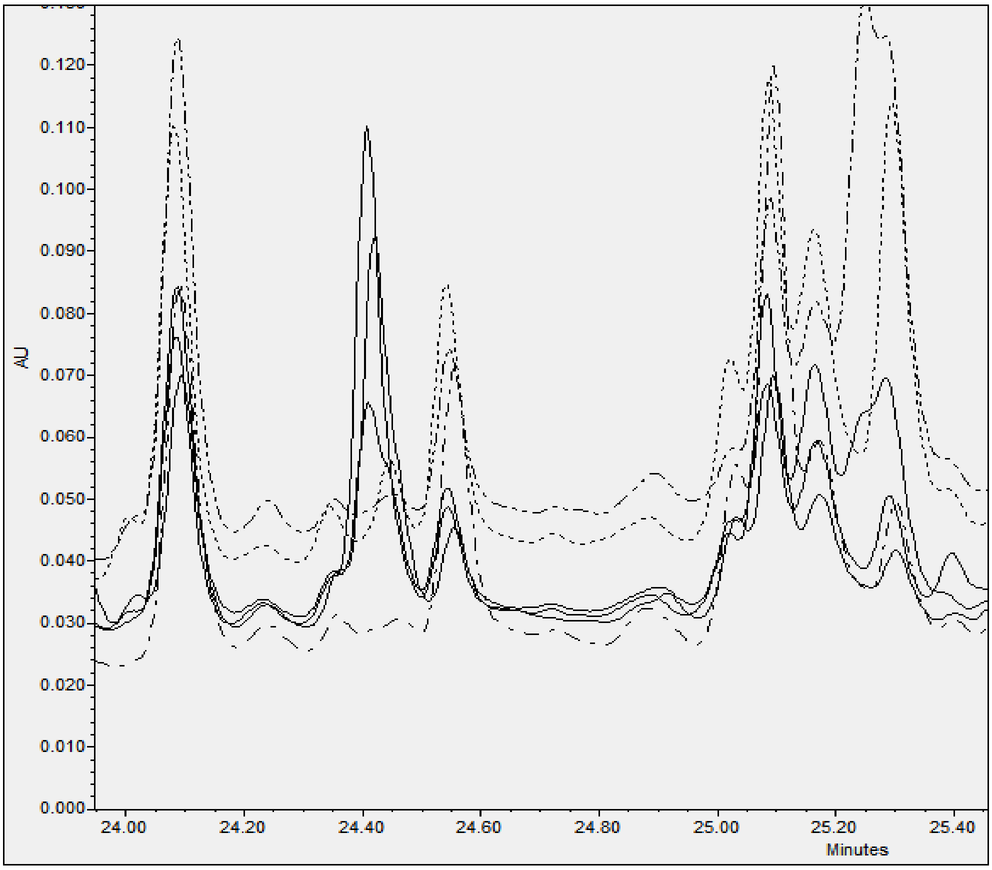


B.

Fig 7: (A) Superimposed green tea UPLC/DAD chromatograms of one Comm and one NComm cultivar, offset by 0.25 min for easy identification. The internal standards used were sulphanilamide (1.8 min), Tryptamine (7.3 min) and mycophenolic acid (27.9 min). (B) shows the zoomed in chromatograms of three Comm and three NComm cultivars, showing the position of CAT (5.75 min); CAF (9.60) and EC (10.30 min). In both plots, the dotted line represents the Comm cultivars, and the solid line represents the NComm cultivars.



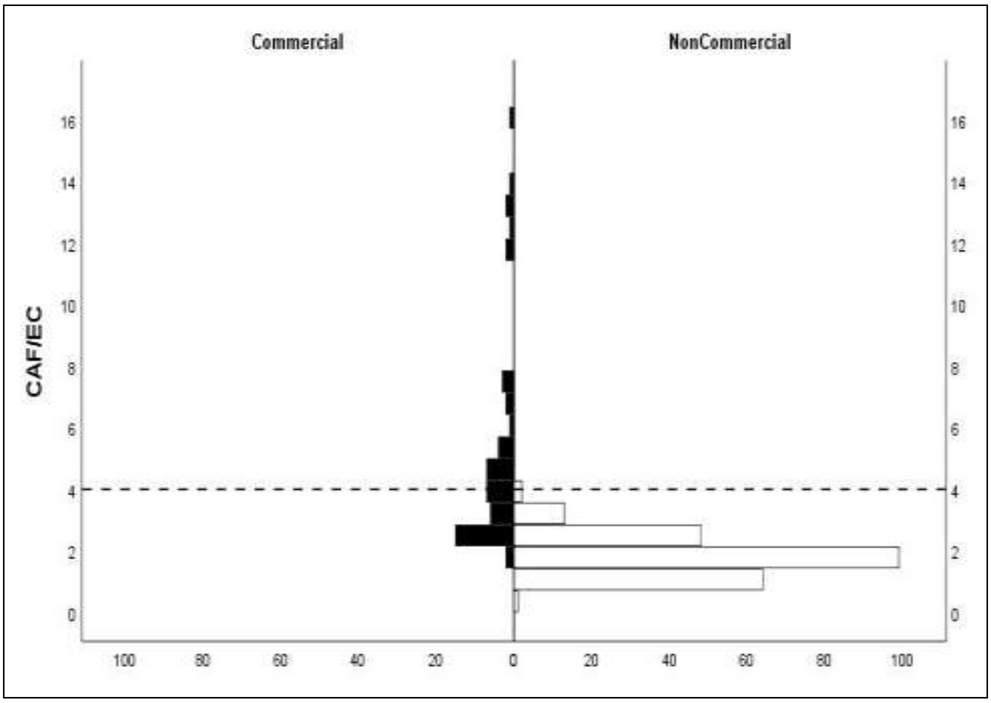
A.



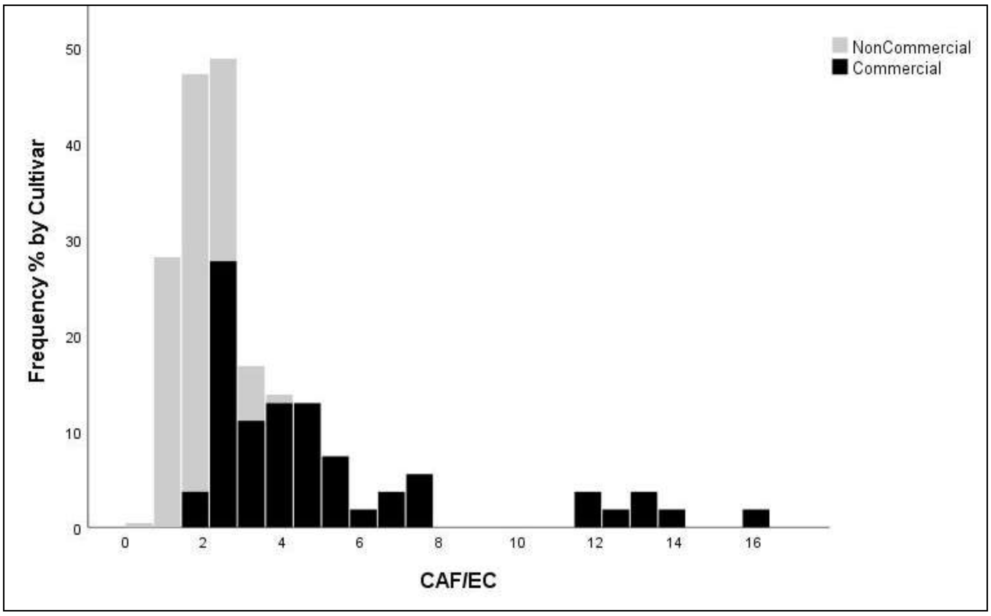
B.

Fig 8: (A) Superimposed black tea UPLC/DAD chromatograms of one Comm and one NComm cultivar, offset and standards as in Figure 7. (B) shows the expanded chromatograms of three Comm and three NComm cultivars, showing the position of TF1 (24.05 min), TF2 (24.40 min), TF3 (24.55 min) and TF4 (25.10 min). In both plots, the dotted line represents the Comm cultivars, and the solid line represents the NComm cultivars. From the (B) figure, it can be seen that TF1, TF3 and TF4 are higher in the Comm cultivars.

as compared to the NComm cultivars Next the ratio of CAF/EC was considered given the inverse relationship observed in Figure 6. The ratio is easier to implement as a distinguishing variable between the Comm and NComm cultivars i.e. the higher the ratio, the higher the likelihood that a cultivar would be Comm, as confirmed in Figure 9.



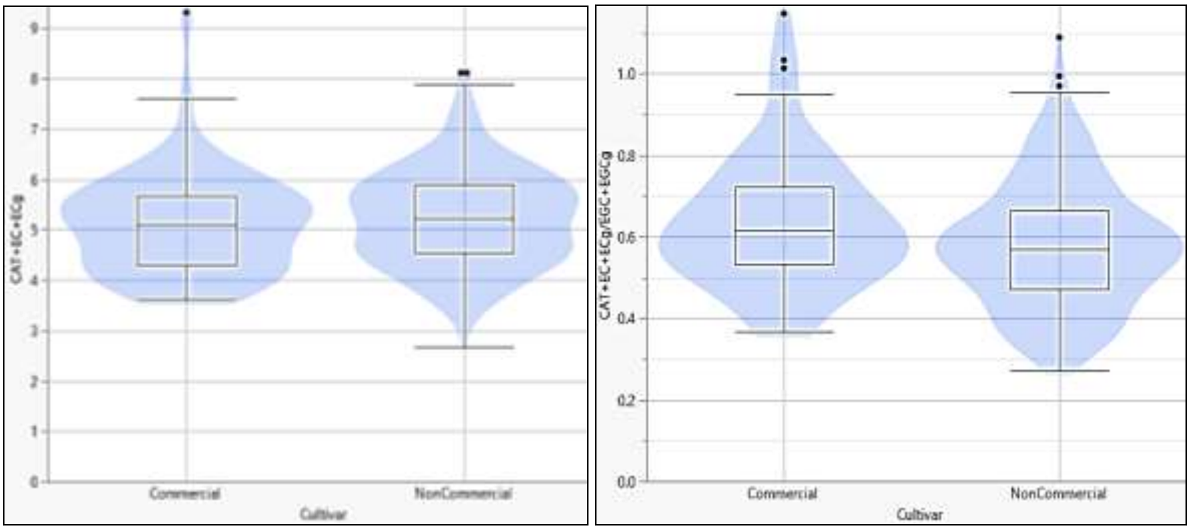
A.



B.

Fig 9: (A) Distribution of CAF/EC ratio by cultivar. (B). Stacked histogram with the CAF/EC ratio as a variable.

In a study by Wright *et al.*, (2000) [25], 20 high, and 20 low quality tea clones were used to investigate the correlations between the catechin profiles of the green tea leaves, and the quality of the resultant black tea produced from them. The results obtained in their study confirmed the findings by Robertson (1983) [35], which showed that the high and low quality tea cultivars differed significantly in CAT, EC, and ECg. The study showed a higher correlation between EC and quality, as compared to ECg, due to the lack of the gallic acid in EC, which has been reported to increase the astringency of tea (Xu *et al.*, 2018) [47]. The high and low quality cultivars thus differed by considering CAT+EC+ECg. Another study by Ellis and Nyirenda, (1995) put forward that the ratio of simple: complex catechins could be a distinguisher between high and low quality teas. These findings do not agree with those of the present study as Table 1 shows that ECg, EGC, and EGCg have VIP scores lower than 1, and as such are not good distinguishers for separating the Comm cultivars from the NComm cultivars. Violin plots were constructed to visualise the possible application of the CAT+EC+ECg and simple: complex ratio in our set of samples, for differentiating between the Comm and NComm cultivars (Figure 10).



A. B.

Fig 10: (A) The sum of simple catechins based on Wright *et al.*, 2000 [25]. (B) The ratio of simple to complex catechins based on Ellis and Nyirenda, 1995

The objective of this study was to make use of UPLC/DAD generated data of the metabolites from the 303 Comm and NComm tea cultivars and to classify investigated genotypes as either Comm or NComm cultivars using CHAID decision trees. The best model may then serve as a prediction tool for whether a newly field selected mother bush is likely to become commercialised due to its similarities with the Comm cultivars. This would increase the success rate of field selections from well-established seedling fields. Violin plots serve as a conspicuous means of visualising the differences between classes, carrying substantial statistical information about e.g. medians and outliers. When the mean of one class falls outside the box of the 25th and 75th percentile of the second group, as seen in Figure 1, this indicates that there is a statistically significant difference between these two classes, regarding that metabolite. The metabolites CAF, CAT, EC, and TF2-TF4 in

Figure 1 differentiate the Comm cultivars from the NComm cultivars, making these ideal predictors to be employed in classifying the 303 genotypes into the two classes. Figures 2 and 3 show the PCA and PLS-DA plots based on the UPLC/DAD data. From both plots, a clear separation in the clustering between the two classes is visible, meaning the ten detected metabolites, listed in Table 1, are discriminators between Comm and NComm cultivars. On the basis of these ten the CHAID decision trees in Figures 4 and 5 were developed, with Figure 4 based only on the theaflavins, and Figure 5 only on CAF, EC, ECg, EGC and EGCg.

In this paper, we show that CHAID decision trees can serve as a strong analytical tool for classifying cultivars as either Comm or NComm, based on the black tea theaflavins of the dried green leaf CAF/EC ratio. The dried green leaf CAF/EC ratio can be applied to field selections immediately instead of taking cuttings, growing hedges for 5 years and manufacturing black tea to measure theaflavins. The study also tests other catechin combinations documented in literature to function as predictors for high and low quality teas (Figure 10). The theaflavin variables were separated from the catechin variables, and a CHAID decision tree was developed based on only the theaflavins (Figure 4). The results, and the classification accuracy table (Table 2) show that TF4 was able to correctly classify all Comm cultivars i.e. 100% of the Comm cultivars in the validation set were correctly classified as Comm cultivars. These results corroborate literature findings that report theaflavins as indicators of high tea quality (Wang and Ruan, 2009) [43]. Theaflavins are orangish-brownish pigments, which contribute to the briskness and brightness of black tea (Muthumani and Kumar, 2007) and are the predominant constituents of black tea-cream upon cooling (Roberts, 1963) [34]; it is for this reason they are deemed as an important quality index of black tea. Theaflavin content influences the total colour of tea i.e. teas with higher theaflavins content will have a higher total colour score. Hilton and Ellis (1972), developed several regression formulae, which were used to correlate theaflavin content in Malawian teas, with price. One formula with a highly significant regression coefficient of $p < 0.001$ held:

$$log\text{price} = a\,log\,\text{T.F.} + b\,log\,\text{T.C.} \qquad (1)$$

with a correlation coefficient is 0.82. T.F = theaflavin and T.C = total colour. To validate their findings, they repeated their experiment using tea samples from Malawi, Uganda, Tanzania, Kenya, Assam and New Guinea; similar results were obtained depicting the close correlation between theaflavin content and market price. Our results agree with those of Hilton and Ellis, and show that the Comm cultivars have higher theaflavin content than the NComm cultivars. Their study and its findings however, failed to gain wide acceptance due to the crude extraction method employed. The current study employs UPLC, which allows for the quantitative identification of the individual catechins and theaflavins. Figure 8 shows superimposed black tea Comm and NComm cultivars, and from this figure, it is visible that the Comm cultivars have higher theaflavins content than the NComm cultivars.

Tea breeders are concentrating on selecting and breeding populations rich in e.g. alkaloids such as caffeine, theobromine and theophylline; amino acids, namely theanine, and polyphenols, namely catechins (Karori *et al.*, 2014). The reason for this is that tea liquor has become a renowned healthy drink. Tea consumption has risen annually by 4.5% to 5.5 million tonnes as of 2016, predominantly in China, India and countries with emerging, developing economies; consumption is postulated to increase by another 1.5 million tonnes by 2027 (FAO, 2018). The top three black tea producing countries namely Kenya, India, and Sri Lanka, have bred and selected high yielding or theaflavin rich cultivars. Efforts have been made to combine these two traits into an F_1 progeny via hybridisation breeding, but the lack of requisite knowhow pertaining to inheritance patterns and how to combine desirable attributes into a single progeny has caused sluggish progress in tea breeding (Wachira and Kamunya, 2005) [42]. From Table 1, the predictors CAT, CAF and EC are statistically significant metabolites, capable of classifying the 303 genotypes into the two classes. This implies that tea breeders can now analyse the CAT, CAF and EC content of green leaves from mature seedling field selections and follow decision tree branches, to ascertain whether a new cultivar is likely to be Comm based on their CAT, CAF and EC content. However, the identification and accurate quantification of CAT may be problematic due to its position on the chromatogram, near unknown peaks, and its small peak height (Figure 7), warranting an improvement of the chromatography conditions in the ISO14502-2 (2005) method. Table 1 shows that although CAT has a higher Cohen's d effect size (an effect size used to show the difference between two means) compared to CAF, it has the same VIP score with CAF, making them both equally important variables for distinguishing between Comm and NComm cultivars. The advantage of using CAF instead of CAT is that, unlike CAT, CAF has a large, clean peak at 9.60 min. This peak can be accurately identified and quantified with ease, without possible co-elution faced by CAT. EC is also a large peak with baseline resolution that is easy to quantify.

Figures 6, a scatter plot of CAF vs EC, the metabolites selected by the CHAID decision tree, graphically displays the combined potential of these two metabolites to differentiate between Comm and NComm cultivars. It is however evident from both the tree and scatter plots, that this combination is not a perfect classifier as there are a few misclassifications; and CHAID decision trees cannot be used to rank samples. CAF is higher and EC lower in the Comm cultivars. Hence the ratio of CAF/EC was constructed to increase size of the signal. Figure 9 shows the frequency histogram based on the CAF/EC ratio. This histogram further displays the ranking ability of this ratio i.e. the higher the ratio the higher the likelihood that the sample is of commercial value. In the current sample set, only Comm cultivars have a CAF/EC ratio that exceeds 4. This suggests that the CAT/EC ratio may be useful to identify field selections from mature seedling fields that have a good probability of becoming commercial cultivars.

The present study reported CAT as an important metabolite predictor. This finding is, however, contradictory to the findings of Wright *et al.*, (2000) [25], who showed that CAT correlated least with tea quality. The reason postulated was that CAT is not a precursor of any of the four major theaflavins, and as such was not important as a predictor for high quality cultivars. The research aim of their work was to investigate any correlations between the catechin profiles of the green tea leaves, and the quality of the resultant black tea produced from them. The study involved 20 high, and 20 low quality clones. The results obtained in the Wright study confirmed those obtained by Robertson, (1983) [35], who found that the high and low quality tea cultivars differed significantly in CAT, EC, and ECg. The Wright study also showed a higher correlation between EC and quality, as compared to ECg, due to the lack of the gallic acid in EC. Gallic acid has been shown to increase the astringency of green tea (Xu *et al.*, 2018) [47]. The Wright study concluded that high and low quality cultivars were distinguishable by high sum of simple catechins, namely CAT+EC+ECg, (B-ring di-hydroxy or simple catechins). The results in the current study show lower EC and ECg concentrations in the Comm cultivars, in contrast to Wright's results where EC and ECg were higher in the good cultivars compared to the poor cultivars. This consideration prompted us to construct of a violin plot based on CAT+EC+ECg (Figure 10 A) in the present study. The results however showed no statistically significant difference between the Comm and NComm cultivars, based on the CAT, EC and ECg. In another study by Ellis and Nyirenda, (1995) on simple (CAT, EC and ECg) and complex catechins (EGC and EGCg), they documented that the higher the ratio of simple: complex catechins, the higher the amount of theaflavins produced, which means the higher the quality of the resultant tea liquor. It was therefore concluded that the cultivars with a higher ratio of simple: complex catechins were of higher quality and ought to be selected. In the present study, the ratio of simple: complex catechins were also employed in constructing a violin plot, and there was no statistically significant difference between the Comm and NComm cultivars (Figure 10 B). Our results, however, indicated that the findings of Robertson and Wright were not applicable to the cultivars used in this study. The reason for this could be that the NComm population used in our study was derived from two parents, whereas the cultivars used by Robertson and Wright were open pollinated plants from various parents. Another reason could be that the Robertson and Wright studies employed HPLC, which may have had CAT co-eluting with other compounds, while the co-eluting compounds were separated in our study, with CAT having two shoulders, as is seen in our higher resolution UPLC chromatograms. Lastly, the difference in the results of both studies could be because our study employed a sample size of 303 cultivars whereas Robertson and Wright employed sample sizes of eight and 20 respectively. This difference lends more credibility to our results.

4. Conclusion

The results of this study show that it is now possible for breeders to predict the quality of new selections from mature seedling fields by employing CHAID decision trees, or the CAF/EC, as predictors. By making use of the model based on CAF and the four catechins, breeders will be more successful in identifying and field selections rich in catechins, which as stated in the introduction, will result in teas rich in theaflavins, and higher market price. However, further studies must be done on varieties from other tea producing countries such as Malawi, Sri-Lanka and India, and on populations derived from more parents, to confirm the validity and efficacy of the results obtained. Additionally, chromatographic work must be done to improve on the identification and quantification of CAT, which has been shown to possibly be an important predictor. The method proposed in this study may improve the success of field selections to higher than the current 1%.

5. Acknowledgements

The authors acknowledge the financial support to conduct this research, and study grants for CN from James Finlay (Kenya) Ltd., George Williamson (Kenya) Ltd., Sotik Tea Company (Kenya) Ltd., Mcleod Russell (Uganda) Ltd., the TRI of Kenya, and *Southern African Biochemistry and Informatics for Natural Products (*SABINA). The *C. sinensis* cultivars used in this study were provided by the TRI of Kenya. Supplementary funding was provided by the Technology and Human Resources for Industry Programme (THRIP), an initiative of the Department of Trade and Industries of South Africa (dti), the National Research Foundation (NRF) of South Africa, and the University of Pretoria South Africa.

6. Author contributions

ZA, RM and SK were involved with the experimental design of the research. RK and CN were responsible for plant material collection. CN conducted the experiments. MvR performed statistical analysis. CN wrote the manuscript, which was revised by MvR, RK, RM, SK, and ZA. The manuscript was reviewed and approved by all the authors.

7. Compliance with ethical standards

7.1 Conflict of interest

The authors assert that they have no conflicts of interest.

8. References

1. Adkins NL, Hall JA, Georgel PT. The use of quantitative agarose gel electrophoresis for rapid analysis of the integrity of protein–DNA complexes. Journal of biochemical and biophysical methods 2007; 70(5):721-726.
2. Anon. Seed garden (barie). Annual Report., Tea Research Foundation of Kenya, Tea Board of Kenya, 1990, 25
3. Banerjee B. Selection and breeding of tea. In Tea 1992, 53-86.
4. Breiman LF, Olshen JH, Stone RA. Classification and regression trees. Belmont, Calif.: Wadsworth, 1984.
5. Chang K. World tea production and trade Current and future development, Food and Agricultural Organization of the United Nations, 2015.
6. Chen L, Apostolides Z, Chen ZM. Global tea breeding: achievements, challenges and perspectives. Edn 1, Springer Science & Business Media, 2013.
7. Cheserek BC, Elbehri A, Bore J. Analysis of links between climate variables and tea production in the recent past in Kenya. Donnish Journal of Research in Environmental Studies 2015; 2(2):5-17.
8. Elbehri A, Azapagic A, Cheserek B, Raes D, Kiprono P, Ambasa C. Kenya's tea sector under climate change: an impact assessment and formulation of a climate smart strategy. FAO report. FAO, Rome, Italy, 2015.
9. Ellis R, Nyirenda H. A successful plant improvement programme on tea (*Camellia sinensis*). Experimental Agriculture 1995; 31(3):307-323.
10. FAO. Global tea consumption and production driven by robust demand in China and India. In FAO Intergovernmental group on tea a subsidiary body of the FAO committee on commodity problems (CCP), (Ed K. Chang). Rome, Italy, 2018, 1-13
11. Green M. An evaluation of some criteria used in selecting large-yielding tea clones. The Journal of Agricultural Science 1971; 76(1):143-156.
12. Groppe DM, Urbach TP, Kutas M. Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. Psuchophysiology 2011; 48(12):1726-1737.
13. Hagel JM, Facchini PJ. Plant metabolomics: analytical platforms and integration with functional genomics. Phytochemistry Reviews 2008; 7(3):479-497.
14. Hicks A. Current status and future development of global tea production and tea products. AUJT 2009; 12(4):251-264.
15. Hilton PJ, Ellis RT. Estimation of the Market Value of Central African Tea by Theaflavin Analysis. Journal of the

Science of Food and Agriculture 1972; 23:227-232.

16. Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. The American Statistician 1998; 52(2):181-184.
17. Horie H, Mukai T, Kohata K. Simultaneous determination of qualitatively important components in green tea infusions using capillary electrophoresis. Journal of Chromatography A 1997; 758(2):332-335.
18. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp, 2019.
19. ISO 14502-2: 2005 Determination of substances characteristic of green and black tea. Part *2:* Content of catechins in green tea method using highperformance liquid chromatography, 2006
20. Kamunya S, Wachira F. Two new clones (TRFK 371/3 and TRFK 430/90) released for commercial use. Tea 2006; 27(1, 2):3-14.
21. Kamunya S, Wachira F, Pathak R, Korir R, Sharma V, Kumar R *et al*. Genomic mapping and testing for quantitative trait loci in tea (*Camellia sinensis* (L.) O. Kuntze). Tree genetics & genomes 2010; 6(6):915-929.
22. Karori S, Wachira F, Ngure R, Mireji P. Polyphenolic composition and antioxidant activity of Kenyan tea cultivars. Journal of Pharmacognosy and Phytochemistry 2014; 3(4):105-116.
23. Kenya National Bureau of Statistics (2012). Kenya facts and figures Kenya, 2012.
24. Khan N, Mukhtar H. Tea polyphenols for health promotion. Life sciences. 2007; 81(7):519-533.
25. Koech RK, Malebe PM, Nyarukowa C, Mose R, Kamunya SM, Apostolides Z. Identification of novel QTL for black tea quality traits and drought tolerance in tea plants (*Camellia sinensis*). Tree genetics & genomes 2018; 14(1):9.
26. Le Gall G, Colquhoun IJ, Defernez M. Metabolite profiling using 1H NMR spectroscopy for quality assessment of green tea, *Camellia sinensis* (L.). Journal of agricultural and food chemistry 2004; 52(4):692-700.
27. Milanović M, Stamenković M. CHAID decision tree: Methodological frame and application. Economic Themes 2016; 54(4):563-586.
28. Miller B, Fridline M, Liu PY, Marino D. Use of CHAID decision trees to formulate pathways for the early detection of metabolic syndrome in young adults. Computational and mathematical methods in medicine, 2014.
29. Muthumani T, Kumar RSS. Influence of fermentation time on the development of compounds responsible for quality in black tea. Food Chemistry 2007; 101(1):98-102.
30. Nyarukowa C, Koech R, Loots T, Apostolides Z. SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. Journal of plant physiology 2016; 198:39-48.
31. Nyirenda H. Use of growth measurements and foliar nutrient content as criteria for clonal selection in tea (*Camellia sinensis*). Experimental Agriculture 1991; 27(1):47-52.
32. Obanda M, Owuor PO, Taylor SJ. Flavanol composition and caffeine content of green leaf as quality potential indicators of Kenyan black teas. Journal of the Science of Food and Agriculture 1997; 74(2):209-215.
33. Owuor PO, Wachira FN, Ng'etich WK. Influence of region of production on relative clonal plain tea quality parameters in Kenya. Food chemistry 2010; 119(3):1168-1174.
34. Roberts EAH. The phenolic substances of manufactured tea. X.-the creaming down of tea liquors. Journal of the Science of Food and Agriculture 1963; 14(10):700-705.
35. Robertson A. Effects of physical and chemical conditions on the in vitro oxidation of tea leaf catechins. Phytochemistry 1983; 22(4):889-896.
36. Schauer N, Fernie AR. Plant metabolomics: towards biological function and mechanism. Trends in plant science 2006; 11(10):508-516.
37. Shanmugarajah V, Kulasegeram S, Senanayake Y. Nursery plant attributes as criteria for selection of new tea clones, 1991.
38. Takemoto M, Takemoto H. Synthesis of theaflavins and their functions. Molecules 2018; 23(4):918.
39. Theodoridis GA, Gika HG, Want EJ, Wilson ID. Liquid chromatography–mass spectrometry based global metabolite profiling: a review. Analytica chimica acta. 2012; 711: 7-16.
40. Urano K, Maruyama K, Ogata Y, Morishita Y, Takeda M, Sakurai N *et al*. Characterization of the ABA-regulated global responses to dehydration in Arabidopsis by metabolomics. The plant journal. 2009; 57(6):1065-1078.
41. Wachira F. Tea improvement in Kenya An overview of research achievements. Prospects and limitations in TBK Board of Directors Open day Proceeding, 29, Jan 2001. Tea Research Foundation of Kenya 2001, 12-14.
42. Wachira FN, Kamunya S. Kenyan teas are rich in antioxidants. Tea 2005; 26(2):81-89.
43. Wang K, Ruan J. Analysis of chemical components in green tea in relation with perceived quality, a case study with Longjing teas. International journal of food science & technology 2009; 44(12):2476-2484.
44. Wilkinson L. Tree structured data analysis: AID, CHAID and CART. *Retrieved February, 1, 2008.*
45. Wright LP, Mphangwe NIK, Nyirenda HE, Apostolides Z. Analysis of caffeine and flavan-3-ol composition in the fresh leaf of *Camellia sinesis* for predicting the quality of the black tea produced in Central and Southern Africa. Journal of the Science of Food and Agriculture. 2000; 80(13):1823-1830.
46. Wright LP, Mphangwe NIK, Nyirenda HE, Apostolides Z. Analysis of the theaflavin composition in black tea (*Camellia sinensis*) for predicting the quality of tea produced in Central and Southern Africa. Journal of the Science of Food and Agriculture. 2002; 82(5):517-525.
47. Xu YQ, Zhang YN, Chen JX, Wang F, Du QZ, Yin JF. Quantitative analyses of the bitterness and astringency of catechins from green tea. Food chemistry. 2018; 258:16-24.
48. Zhou B, Xiao JF, Tuli L, Ressom HW. LC-MS-based metabolomics. Molecular BioSystems 2012; 8(2):470-481.