

**Metabolomic markers for agronomic traits and their possible
biochemical mechanisms in black tea *Camellia sinensis* (L.) O.
Kuntze**

By
Christopher Tavengwa Nyarukowa

Submitted in partial fulfilment of the requirements for the degree


Philosophiae Doctor
(Specialisation in Biochemistry)

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

25 August 2020

SUBMISSION DECLARATION

I, Christopher Tavengwa Nyarukowa declare that the thesis, which I hereby submit for the degree at *Philosophiae Doctor* (Specialisation in Biochemistry) the University of Pretoria, is my work and has not previously been submitted by me for the degree at this or any other tertiary institution.

A handwritten signature in black ink, appearing to read 'Christopher Tavengwa Nyarukowa', is enclosed within a light gray rectangular box.

Signature:

Date: 25 August 2020

PLAGIARISM DECLARATION

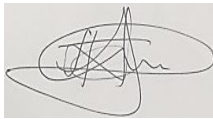
Full name: Christopher Tavengwa Nyarukowa

Student number: u11205670

Title of the work: Metabolomic markers for agronomic traits and their possible biochemical mechanisms in black tea *Camellia sinensis* (L.) O. Kuntze

Declaration

1. I understand what plagiarism entails and am aware of the University's policy in this regard.
2. I declare that this proposal is my own original work. Where someone else's work was used (either from a printed source, internet or any other source), due acknowledgement was given, and reference was made according to departmental requirements.
3. I did not make use of another student's previous work and submit it as my own.
4. I did not allow and will not allow anyone to copy my work with the intention of presenting it as his or her own work.



Signature:

Date: 25 August 2020

ACKNOWLEDGEMENTS

I would like to thank the Lord Almighty, for without Him, this achievement would not be possible. Thank you for giving me the strength to keep going.

I would like to express the utmost gratitude to my supervisor, Prof. Zeno Apostolides, for his invaluable contribution, guidance, constructive criticism and constant support and encouragement. Thank you for the opportunity and overall positive working environment, which enabled me to enjoy the PhD study. I genuinely could not have asked for a more supportive supervisor.

I am forever indebted to my family, the General, Mimaimi, Gwis, The Mountain, Peezy and Mbalisi for the support and encouragement. Thank you always being there during the good as well as the troubling times. You have ALL been a strong foundation and support structure; proud to call you family.

I would like to appreciate the assistance of the University of Pretoria and the Department of Biochemistry, Genetics and Microbiology for providing me with resources, and the staff for their unwavering support throughout the duration of this study. I would also like to thank my CAM research group colleagues for their support and friendship.

I would like to acknowledge the financial support to conduct this research from James Finlay (Kenya) Ltd, George Williamson (Kenya) Ltd, Sotik Tea Company (Kenya) Ltd, Mcleod Russell (Uganda) Ltd, and the Tea Research Institute of Kenya. The *C. sinensis* cultivars used in this study were provided by the Tea Research Institute 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).

DEDICATION

*I dedicate this thesis to
the 5*General and Mimaimi,
for the constant support and unconditional love
this one is for you.*

ABSTRACT

Climate change is causing droughts, which are affecting crop production globally, and disrupting plant metabolism. Due to the unpredictable natural droughts that occur, causing tea farmers significant losses in tea estates, a Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) method for distinguishing between drought tolerant (DT) and drought susceptible (DS) *Camellia sinensis* cultivars was developed based on cultivars from the Tea Research Foundation for Central Africa in Malawi, and validated on 400 samples from the Tea Research Institute in Kenya. From the results, a sample size of 20 tea trees was deemed sufficient to accurately determine the drought susceptibility of a large tea field of approximately 5 - 20 hectares, containing 50 000 - 200 000 tea trees, where the difference between their mean values is approximately 6%. Tea production and subsequently its quality rely on evenly distributed rainfall. Tea consumers concern themselves with the quality of tea, in particular its flavour and aroma. To breed for these phenotypic traits is challenging due to these being qualitative traits inherited from parents, and influenced by environment. Two *C. sinensis* populations, 60 Commercial cultivars and 250 NonCommercial cultivars (TRFK St. 504 and TRFK St. 524) were employed in a part of this study to identify the Quantitative Trait Loci (QTL) responsible for yield, drought tolerance and quality centred on a genetic map constructed using the DArTseq platform. The map comprised 15 linkage groups analogous to chromosome haploid number of tea plant ($2n = 2x = 30$) and spanned 1260.1 cM with a mean interval of 1.1 cM between markers. Sixteen phenotypic traits were evaluated in both populations, and three, 11 and 46 putative QTLs were discovered after mapping on the 15 linkage groups, associated with tea quality from Gas Chromatography-Mass Spectrometry (GC-MS), Nuclear Magnetic Resonance ($^1\text{H-NMR}$) and Ultra-Performance Liquid Chromatography (UPLC) data respectively. The variance explained by the QTLs varied from 4.6 to 96.3%, with an average of 28%. Using the KEGG database, the putative QTLs linked to yield, drought tolerance and quality were secondary metabolites associated with tea phenolic biomolecules and abiotic stress. Principal Component Analysis was performed on the GC-MS, $^1\text{H-NMR}$ and UPLC data, and from these, the UPLC data showed clearer separation and clustering between the Commercial and NonCommercial cultivars. With focus on the UPLC data, it was narrowed down to the five catechins, four theaflavins and caffeine; these were used to develop several logistic regression models. The model based on only the fresh leaf catechins classified over 90% of the 310 genotypes as either Commercial or NonCommercial cultivars. This model may be useful in predicting the

suitability for commercialization of promising selections from mature seedling fields, based on the analysis of their dried green leaves. Last, 20 Commercial and 20 NonCommercial cultivars were analysed using UPLC-MS. New metabolites were identified as contributing to drought tolerance, yield and higher quality of the Commercial as compared to the NonCommercial cultivars.

TABLE OF CONTENTS

SUBMISSION DECLARATION	i
PLAGIARISM DECLARATION.....	ii
ACKNOWLEDGEMENTS.....	iii
DEDICATION.....	iv
ABSTRACT	v
TABLE OF CONTENTS	vii
LIST OF FIGURES	xii
LIST OF TABLES	xvi
LIST OF APPENDICES	xvii
LIST OF ABBREVIATIONS.....	xviii
CHAPTER 1	1
LITERATURE REVIEW	1
1.1 Camellia sinensis	1
1.2 Flavonoids in Tea	2
1.3 Tea polyphenols	4
1.4 Health benefits of tea.....	4
1.5 Black tea.....	8
1.6 Tea quality indicators	9
1.7 Principle metabolites found in tea and their metabolic attributes	10
1.8 The tea breeding history in Africa.....	14
1.9 Improvements in Kenyan tea breeding programmes	15
1.10 Breeding for high yield and environmental stress resistance.....	15
1.11 Current breeding strategies to obtain desirable characteristics	17
1.12 Marker assisted selection	18
1.13 Role of genetic markers in tea breeding improvement	19

1.14 Challenges faced in genomic selection.....	22
1.15 Plant metabolomics.....	22
1.16 Statistical analysis.....	27
1.17 PROBLEM STATEMENT.....	29
1.18 RESEARCH OBJECTIVE.....	30
1.19 RESEARCH OUTPUTS.....	31
1.20 REFERENCES.....	32
CHAPTER 2.....	47
PRIORITISING THE REPLANTING SCHEDULE OF SEEDLING TEA FIELDS ON TEA ESTATES FOR DROUGHT SUSCEPTIBILITY MEASURED BY THE SWAPDT METHOD IN THE ABSENCE OF HISTORICAL IN-FILLING RECORDS.....	47
ABSTRACT.....	47
2.1 INTRODUCTION.....	48
2.2 STATISTICAL ANALYSIS.....	49
2.2.1 Spatial regression Modelling.....	49
2.2.2 Contour Maps: construction and application.....	50
2.2.3 Moran I Test Statistics.....	51
2.2.4 Mann-Whitney test.....	53
2.3 RESEARCH OBJECTIVE.....	55
2.4 HYPOTHESES.....	55
2.5 MATERIALS AND METHODS.....	56
2.5.1 Sample Collection.....	56
2.5.2 Statistical Analysis.....	57
2.6 RESULTS.....	58
2.6.1 %drought score contour RWC plots based on SWAPDT method.....	58
2.7 DISCUSSION AND CONCLUSION.....	67
2.8 REFERENCES.....	69

Appendix 2.1: Peer-reviewed scientific article based on results from Chapter 2.....	72
Appendix 2.2: The meaning of the mean diamonds and x-axis proportional lines used in Chapter 2 statistical analysis	81
CHAPTER 3.....	83
IDENTIFICATION OF QTL'S RESPONSIBLE FOR YIELD, DROUGHT TOLERANCE AND QUALITY TRAITS IN CAMELLIA SINENSIS USING GC-MS, ¹ H-NMR AND UPLC.....	83
ABSTRACT.....	83
3.1 INTRODUCTION.....	84
3.2 RESEARCH OBJECTIVE	87
3.3 HYPOTHESIS	87
3.4 MATERIALS AND METHODS	88
3.4.1 Plant material.....	88
3.4.2 Sample collection and processing.....	88
3.5 GC-MS sample preparation and analysis	89
3.5.1 Sample preparation	89
3.5.2 GC-MS analyses.....	89
3.6 ¹ H-NMR sample preparation and analysis.....	90
3.6.1 ¹ H-NMR buffer solution	90
3.6.2 ¹ H-NMR sample preparation.....	90
3.7 UPLC sample preparation and analysis.....	91
3.7.1 Extraction of catechins, caffeine and theaflavins	91
3.7.2 UPLC analyses	91
3.8 Metabolite identification.....	92
3.9 Determination of tea quality	93
3.9.1 DNA extraction and quantification.....	93
3.9.2 DArTseq assay.....	93
3.9.3 Construction of linkage map	93
3.9.4 QTL analyses.....	93

3.9.5 Fast Adaptive Shrinkage Thresholding Algorithm (FASTA) files preparation	94
3.9.6 Basic Local Alignment Search Tool (BLAST) search	94
3.9.7 Functional annotation and pathway assignment	94
3.10 RESULTS	95
3.10.1 Genetic map construction.....	95
3.10.2 Phenotype segregation and QTL mapping	95
3.11 DISCUSSION AND CONCLUSION	117
3.12 REFERENCES.....	127
Appendix 3.1: DArTseq marker ID with the commensuating separate marker sequence .	137
CHAPTER 4.....	139
MODELS FOR IDENTIFICATION OF ELITE MOTHER BUSHES WITH HIGH BLACK TEA COMMERCIAL POTENTIAL FROM MATURE SEEDLING FIELDS OF CAMELLIA SINENSIS	139
ABSTRACT.....	139
4.1 INTRODUCTION.....	140
4.2 RESEARCH OBJECTIVES	146
4.3 HYPOTHESIS	146
4.4 MATERIALS AND METHODS	147
4.4.1 Plant material.....	147
4.4.2 Sample collection and processing.....	147
4.4.3. GC-MS sample preparation and analysis.....	148
4.4.4 ¹ H-NMR sample preparation and analysis	148
4.4.5 UPLC-DAD and UPLC-MS sample preparation and analysis	149
4.4.6 Metabolite identification	151
4.4.7 Data pre-processing	152
4.4.8 Multivariate statistical analysis	153
4.4.9 Logistic regression analysis	154
4.5. RESULTS	156

4.5.1 List of tables showing detected metabolites using GC-MS, ¹ H-NMR, UPLC-DAD and UPLC-MS.....	156
4.5.2 Violin plots for GC-MS, ¹ H-NMR, UPLC-DAD and UPLC-MS.....	162
4.5.3 GC-MS, ¹ H-NMR and UPLC-DAD PCA plots.....	177
4.5.4 UPLC-MS positive and negative ionisation mode PCA, PLS-DA and S-plots ...	179
4.5.5 LR analysis.....	189
4.6 DISCUSSION AND CONCLUSION	210
4.7 REFERENCES.....	225
Appendix 4.1: Peer-reviewed scientific article based on results from Chapter 4.....	239
Appendix 4.2: Cohen’s d effect size definition, calculation and interpretation used in the statistical analysis of Chapter 4 data	252
CHAPTER 5.....	254
CONCLUDING DISCUSSION AND RECOMMENDATION	254
5.1 Concluding discussion.....	254
5.2 Recommendations	259

LIST OF FIGURES

Figure 1.1: Biosynthesis of flavan-3-ols and their derivatives in <i>C. sinensis</i> leaves and roots.	2
Figure 1.2: Structural configuration of flavonoid molecule.	3
Figure 1.3: Structural configuration of the major flavonoid categories.	4
Figure 1.4: The antioxidant mechanism of action for catechin	6
Figure 1.5: Groups conferring antioxidant properties on catechin	6
Figure 1.6: Catechins predominant in green tea.	7
Figure 1.7: Structures of the major theaflavins.	8
Figure 1.8: Flavonoid metabolism in <i>C. sinensis</i>	11
Figure 1.9: Theanine metabolism in <i>C. sinensis</i>	12
Figure 1.10: Caffeine metabolism in <i>C. sinensis</i>	13
Figure 1.11: Schematic representation of tea breeding program.....	18
Figure 1.12: Diagrammatic representation of the DArT principle	21
Figure 1.13: A metabolomics workflow	24
Figure 1.14: Number of publications on metabolomics.....	26
Figure 2.1: Tea growing areas in Kenya	49
Figure 2.2: Two distributions with different medians but similar shape and spread.	53
Figure 2.3: Two distributions with differing median values	54
Figure 2.4: %RWC drought score contour plots.....	60
Figure 2.5: Mean distribution curves for the fields 12A, 12B, 13A and 13B	62
Figure 2.6: Ideal theoretical vs practical sample size required.....	63
Figure 2.7: Oneway analysis of the % RWC against the two good and two poor fields.	64
Figure 2.8: Power curves.....	66
Figure 3.1: Proposed identification confidence levels in high resolution MS analysis.	92
Figure 3.2: Genetic map of <i>C. sinensis</i> , displaying GC-MS QTL locations	98

Figure 3.3: Genetic map of <i>C. sinensis</i> , displaying ¹ H-NMR QTL locations	101
Figure 3.4: Genetic map of <i>C. sinensis</i> , displaying UPLC QTL locations	108
Figure 4.1: Violin plots showing separation between the Comm and NComm cultivars based on GC-MS metabolites.....	165
Figure 4.2: Violin plots showing separation between the Comm and NComm cultivars based on ¹ H-NMR metabolites.....	169
Figure 4.3: Violin plots showing separation between the Comm and NComm cultivars based on detected UPLC-DAD metabolites.....	172
Figure 4.4: Violin plots showing separation between the Comm and NComm cultivars based on UPLC-MS metabolites.....	176
Figure 4.5: The 3D PCA scores plots for PCs one, two and three, for GC-MS (A) ¹ H-NMR (B) and UPLC-DAD (C)	177
Figure 4.6: The 3D PLS-DA scores plots for LVs one, two and three, for GC-MS (A), ¹ H-NMR (B) and UPLC-DAD (C).....	178
Figure 4.7: (A) The PCA scores plot; (B) the PLS-DA plot, and (C) the s-plot, showing good separation between the Comm and the NComm cultivars in positive ionisation mode.....	180
Figure 4.8: (A) The PCA scores plot; (B) the PLS-DA plot, and (C) the s-plot, showing good separation between the Comm and the NComm cultivars in negative ionisation mode.....	181
Figure 4.9: (A) Pure caffeine and (B) ECg standards identified through their fragments and confirmed by their hit and score values	188
Figure 4.10: Nominal LR using all the detected GC-MS variables.....	189
Figure 4.11: Nominal LR using the seven statistically significant variables from the total detected variables.....	189
Figure 4.12: (A) shows the LR plot using Acetoacetic acid/Psicose as a variable. (B) shows the confusion matrix obtained from the LR analysis.....	190

Figure 4.13: Decision tree based on the GC-MS metabolites	191
Figure 4.14: Nominal LR using all the detected ¹ H-NMR variables.	192
Figure 4.15: Nominal LR model developed on amino acid variables.	192
Figure 4.16: Nominal LR using all CAF, and all five catechin variables.	193
Figure 4.17: (A) shows the LR plot using CAF/CAT as a variable. (B) shows the confusion matrix obtained from the LR analysis.	193
Figure 4.18: Decision tree based on the ¹ H-NMR CAF and all five catechin variables.....	194
Figure 4.19: (A) shows the LR plot using CAF/EGC as a variable, and (B) its confusion matrix.	194
Figure 4.20: (A) shows the LR plot using CAT/EC as a variable. (B) shows the confusion matrix obtained from the LR analysis.	195
Figure 4.21: (A) shows the LR plot using CAF/EC as a variable. (B) shows the confusion matrix obtained from the LR analysis.	195
Figure 4.22: Nominal LR using all ten UPLC-DAD variables.	196
Figure 4.23: Nominal LR using only the four theaflavin variables.	196
Figure 4.24: (A) shows the LR plot using TF4 as a variable. (B) shows the confusion matrix obtained from the LR analysis.	197
Figure 4.25: Nominal LR using CAF, and all five catechin variables.	198
Figure 4.26: Decision tree based on CAF and all five catechin variables	199
Figure 4.27: (A) shows the LR plot using CAT/EC as a variable. (B) shows the confusion matrix obtained from the LR analysis.	200
Figure 4.28: (A) Superimposed green tea UPLC/DAD chromatograms of one Comm and one NComm cultivar.....	201
Figure 4.29: (A) Superimposed black tea UPLC/DAD chromatograms of one Comm and one NComm cultivar.....	202

Figure 4.30: Decision tree based on CAF and the catechin variables, excluding CAT 203

Figure 4.31: (A) shows the LR model for the CAF and catechins, excluding CAT. (B) shows the LR plot based on the CAF/EC ratio. (C) shows the confusion matrix for the CAF/EC ratio. 204

Figure 4.32: (A) shows the LR model for CAT+EC+ECg. (B) shows the confusion matrix for CAT+EC+ECg..... 205

Figure 4.33: (A) shows the LR model for Simple/Complex catechins. (B) shows the confusion matrix for the Simple: Complex catechin model. 206

LIST OF TABLES

Table 3.1: DArT markers distribution among the linkage groups.....	96
Table 3.2: GC-MS QTLs.....	97
Table 3.3: NMR QTLs	99
Table 3.4: UPLC QTLs	102
Table 3.5: The differences between ¹ H-NMR and UPLC QTL markers.	109
Table 3.6: Functional annotation of putative candidate genes in GC-MS related linkage groups of <i>C. sinensis</i> on reference tea genome.....	110
Table 3.7: Functional annotation of putative candidate genes in ¹ H-NMR related linkage groups of <i>C. sinensis</i> on reference tea genome.....	111
Table 3.8: Functional annotation of putative candidate genes in UPLC-DAD related linkage groups of <i>C. sinensis</i> on reference tea genome.....	113
Table 4.1: The list of tentatively identified metabolites detected by GC-MS, expressed in arbitrary units.....	156
Table 4.2: The list of metabolites detected by ¹ H-NMR, expressed in mg/g.....	157
Table 4.3: The list of metabolites detected by the UPLC-DAD, expressed in mg/g.....	158
Table 4.4: The list of metabolites detected by the UPLC-MS, expressed in arbitrary units.	159
Table 4.5: The list of identified metabolites to be distinguishing markers between the Comm and NComm cultivars.....	161
Table 4.6: List of LR models developed showing the %specificity and %sensitivity of each. Prob(Comm) = 1 / (1 + Exp(-(Lin[Comm]))).....	207

LIST OF APPENDICES

Appendix 2.1: Peer-reviewed scientific article based on results from Chapter 2.....	72
Appendix 2.2: The meaning of the mean diamonds and x-axis proportional lines used in Chapter 2 statistical analysis.....	81
Appendix 3.1: DArTseq marker ID with the commensuating separate marker sequence	137
Appendix 4.1: Peer-reviewed scientific article based on results from Chapter 4.....	239
Appendix 4.2: Cohen's d effect size definition, calculation and interpretation used in the statistical analysis of Chapter 4 data	252

LIST OF ABBREVIATIONS

4CL	4-coumarate–CoA
ABA	Abscisic Acid
AFLP	Amplified Fragment Length Polymorphism
ANR	Anthocyanidin reductase
ANS	Anthocyanidin synthase
BEH	Bridged ethylsiloxane hybrid
BLAST	Basic Local Alignment Search Tool
BLASTN	Basic Local Alignment Search Tool (Nucleotide)
BLASTX	Basic Local Alignment Search Tool (Protein)
BRK	Briskiness
BRT	Brightness
BSTFA	Bis (trimethylsilyl)trifluoroacetamide
C4H	Cinnamate 4-hydroxylase
CA	Caffeic acid
CAF	Caffeine
CAT	Catechin
CG	Catechin gallate
CHI	Chalcone isomerase
CHS	Chalcone synthase
CI	Chemically ionised
Comm	Commercial
CV	Coefficient of variation
D ₂ O	Deuterium oxide
DAD	Diode array detector
DArT	Diversity Array Technology
DFR	Dihydroflavonol-4-reductase
DIMS	Direct injection mass spectrometry
DNA	Deoxyribonucleic acid

DT	Drought tolerant
DS	Drought susceptible
EC	Epicatechin
ECg	Epicatechin gallate
EDTA	Ethylenediaminetetraacetic acid
EGC	Epigallocatechin
EGCg	Epigallocatechin gallate
EI	Electron impact ionisation
EV	Electron volts
F3H	Flavanone 3-hydroxylase
F3'H	Flavonoid 3'-monooxygenase
F3'5'H	Flavonoid 3', 5'-hydroxylase
FAO	Food and Agricultural Organisation
FASTA	Fast Adaptive Shrinkage Thresholding Algorithm
FLS	Flavonol synthase
FN	False negative
FP	False positive
FTIR	Fourier transform infrared spectroscopy
GC	Gas chromatography
GCG	Gallocatechin gallate
GDH	Glutamate dehydrogenase
GC	Glutamate synthase
GxE	Genotype x environment
HIV	Human immunodeficiency virus
HPLC	High performance liquid chromatography
IMPDH	Inosine 5'monophosphate dehydrogenase
ISO	International Organisation for Standardisation
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAR	Leucoanthocyanidin reductase
LC	Liquid chromatography

LG	Linkage group
LOD	Odds of logarithm
LOO-CV	Leave-one-out cross-validation
LR	Logistic regression
LV	Latent variable
MAS	Marker assisted selection
MATE	Multidrug and Toxic Compound Extrusion
mQTL	Metabolite quantitative trait loci
MS	Mass spectrometry
NComm	NonCommercial
NIST	National Institute of Standards and Technology
NMR	Nuclear Magnetic Resonance
PAL	Phenylalanine ammonia-lyase
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLS-DA	Partial least squares discriminant analysis
PVE	Phenotypic variation explained
QC	Quality control
QTL	Quantitative Trait Loci
RAPD	Random Amplification of Polymorphic DNA
RT	Retention time
RWC	Relative water content
SABINA	Southern African Biochemistry and Informatics for Natural Products
SAM	S-adenosyl-L-methionone synthase
SASBMB	South African Society of Biochemistry and Molecular Biology
SD	Standard deviation
SEM	Standard error of the mean
SNP	Single Nucleotide Polymorphism
SRM	Selected reaction monitoring
SSR	Simple Sequence Repeat

SWAPDT	Short-time Withering Assessment of Probability for Drought Tolerance
TF1	Theaflavin
TF2	Theaflavin-3-gallate
TF3	Theaflavin-3'-gallate
TF4	Theaflavin-3, 3'-digallate
ToF	Time of flight
TN	True negative
TP	True positive
TR	Thearubigin
TRFCA	Tea Research Foundation of Central Africa
TRFK	Tea Research Foundation of Kenya
TRI	Tea Research Institute
TS	Theanine synthase
UPLC	Ultra performance liquid chromatography
UV	Ultra violet
VIP	Variable importance in projection
WHO	World Health Organisation

CHAPTER 1

LITERATURE REVIEW

1.1 Camellia sinensis

In-depth studies have been conducted on *Camellia sinensis* due to its precise flavonoid profile, responsible for tea health properties. Green tea, rich in catechins, serves as a traditional herbal remedy in China to prevent cardiovascular diseases amongst other chronic diseases. Tea prepared from the leaves of *C. sinensis* has been consumed either as green or black tea, by 70% of the human global population, since time immemorial, owing to its richness in polyphenolic compounds, which are associated with copious health promoting, therapeutic attributes (Tong *et al.*, 2014); tea arrived in Europe only in 1636. Tea producers are in demand of new cultivars, which are high yielding, are drought tolerant, and produce high quality tea liquors. The main types of tea consumed worldwide are green, oolong and black tea, each being determined by the concentration their respective flavan-3-ols (Wambulwa *et al.*, 2017). Over 52 countries worldwide cultivate *C. sinensis*, being consumed either as black (78%), green (20%) or oolong (2%) tea. Green tea is however predominantly favoured in North African and Middle Eastern countries, while black tea is customarily consumed in Western countries (Cooper *et al.*, 2005). Green tea quality evaluation has conventionally been based on the appearance i.e. colour and its intensity, aroma i.e. sweet, floral, grassy, etc., and lastly, its taste i.e. astringency, bitterness, and sweetness. Tea gets its distinctive astringent and somewhat bitter taste from caffeine, even though several other metabolites such as the catechins and other polyphenols, carbohydrates, and amino acids are influential in its overall taste and flavour (Adkins *et al.*, 2007; Nyarukowa *et al.*, 2016). The amino acid theanine, which makes up approximately two-thirds of a tea leaf's total amino acids content, is with other less abundant amino acids, responsible for the sweet and brothy taste of tea. However, it is noteworthy to indicate that the metabolite composition, which influences tea quality, varies between green and black tea. Unlike with green tea, whose quality depends on amino acids, particularly theanine, catechins and caffeine, the quality of black tea relies on theaflavin, thearubigen, catechin and caffeine (Le Gall *et al.*, 2004). According to the World Health Organisation (WHO), about 3.92 million metric tonnes of tea is annually produced, and of this, black tea constitutes 60%, while green tea represents 30% (Meeting and Organization, 2010). Green tea production is however predicted to have

considerably increased compared to black tea due to its immense increased consumption in China (Shen *et al.*, 2018).

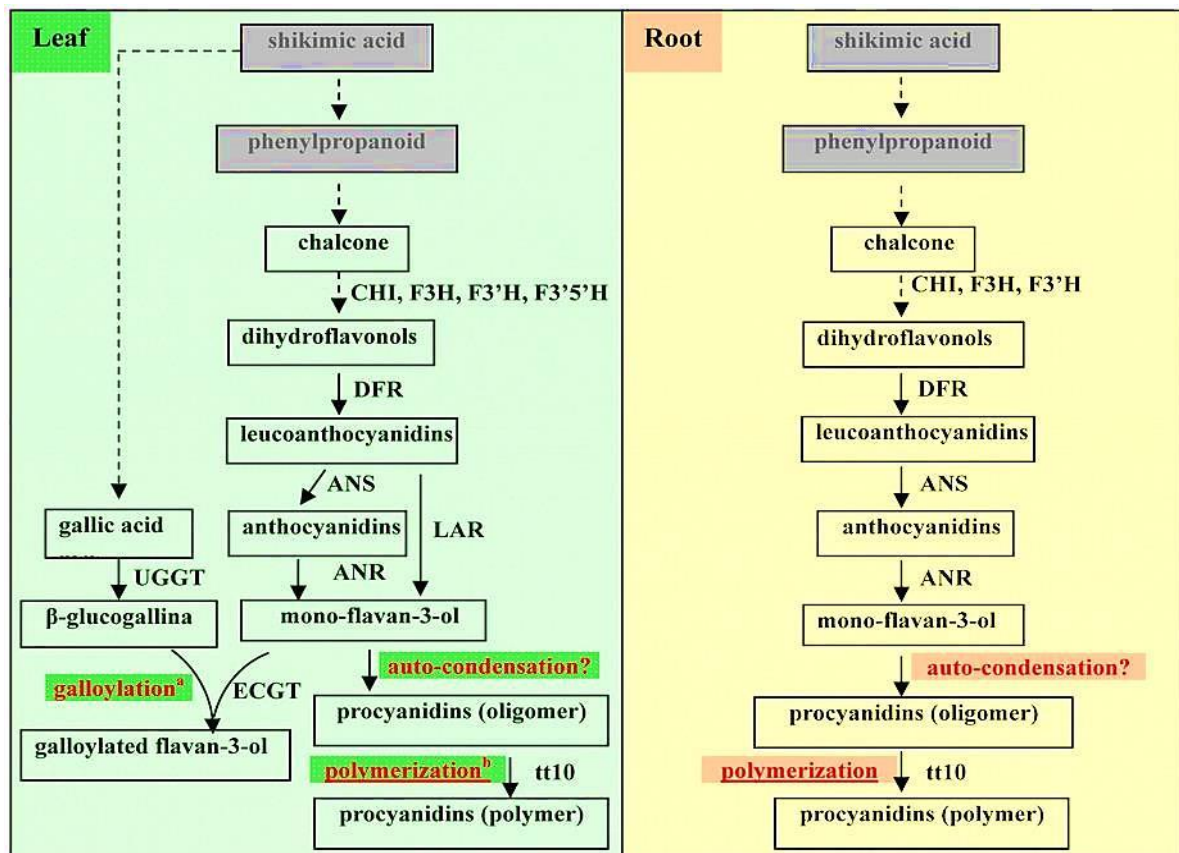


Figure 1.1: Biosynthesis of flavan-3-ols and their derivatives in *C. sinensis* leaves and roots (Jiang *et al.*, 2015).

1.2 Flavonoids in Tea

The dry weight of young *C. sinensis* leaves consists of 25 – 30% flavan-3-ols (Singh *et al.*, 1999). Flavonoids are a sundry class of plant metabolites biosynthesised from phenylpropanoids and derived acetates from carbohydrate metabolism as indicated in Figure 1.1. More than a few disparate types of flavonoids exist, with the most significant being dietary flavonoids, which fall into seven major categories, namely anthocyanidins, chalcones, flavanols, flavones, flavonones, flavonols, and isoflavonoids (Yilmaz, 2006). Flavonoids are essential metabolites required by plants for their growth and development. They protect the plant against microbes and pests by interfering with their interactions with the plant and are involved in the manufacture of phytoalexins, which are insect repellents (Lattanzio *et al.*, 2006). *C. sinensis* makes use of carbon obtained from the metabolism of amino acids tryptophan, tyrosine and phenylalanine for the biosynthesis of 15C flavonoids with a C6-C3-C6 configuration, through the condensation and decarboxylation of phenylpropanoid

derivatives (Cuendet *et al.*, 2001), with chorismic acid being the end product (Jiang *et al.*, 2019). *C. sinensis* then associates carbohydrate metabolism to the shikimate pathway via the pentose pathway, resulting in phenylpropanoid biosynthesis. The flavonoid structure is made up of two aromatic rings A and B, linked by a 3C bridge, in a heterocyclic ring, as shown in Figure 1.2 below.

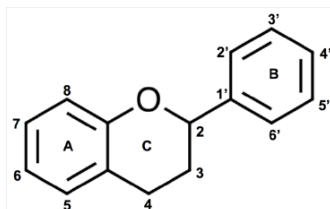
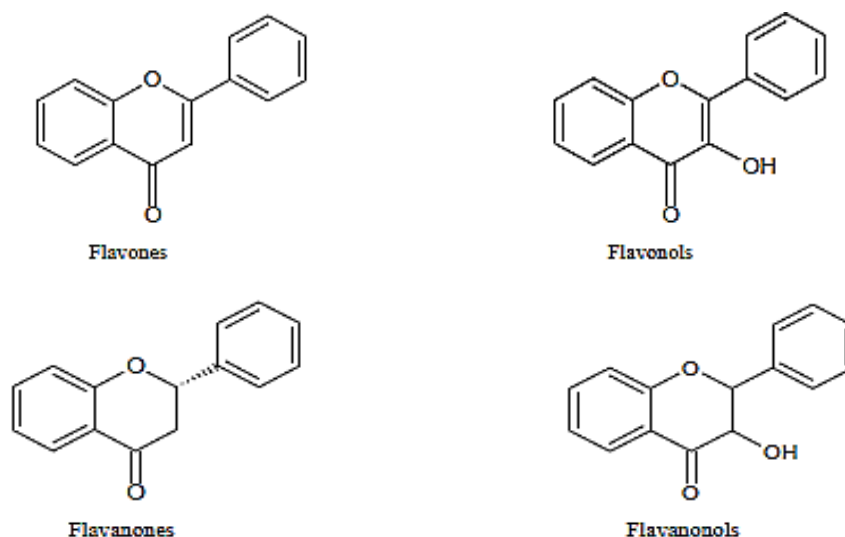


Figure 1.2: Structural configuration of flavonoid molecule.

The A ring is derivative of the acetate pathway, while the B ring is a derivative of phenylalanine via the shikimate pathway. Varying the replacements, through e.g. acylation, alkylation, glycosylation, oxygenation, and sulfation of the A and B rings will result in the formation of varying compounds within each flavonoid class. It is these variations resulting from the substitutions to the C ring, which generate the major flavonoid categories, namely anthocyanidins, chalcones, flavanols, flavones, flavonones, flavonols, and isoflavonoids, as shown in Figure 1.3. In a Kenyan study in 2009, tea cultivars with purple coloured leaves were developed (Kamunya *et al.*, 2009), and a follow up study in 2013 found that said purple leaves were rich in anthocyanins and anthocyanidins, with malvidin being the predominant anthocyanidin responsible for causing the colour (Kerio *et al.*, 2013).



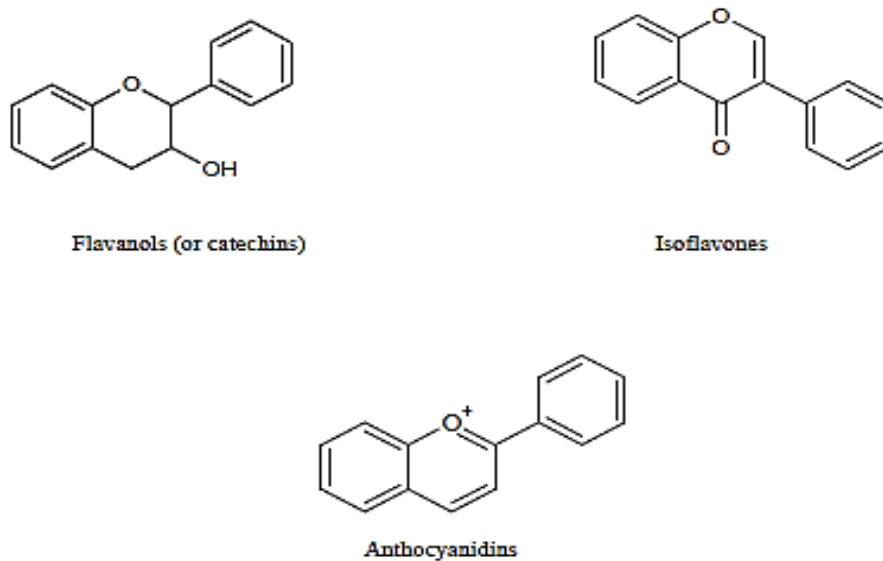


Figure 1.3: Structural configuration of the major flavonoid categories.

1.3 Tea polyphenols

The majority of osmolytes found in plants are secondary metabolites; tea osmolytes are predominantly comprised of polyphenolic metabolites (Cheruiyot *et al.*, 2007). Polyphenols are metabolites consisting of integrated benzene rings, each comprising of numerous hydroxyl groups, and ranging from simple phenolic, to complex polymerised compounds (Balasundram *et al.*, 2006). Although 90% of polyphenols found in tea are flavonoids, other classes of polyphenols exist i.e. phenolic acids and tannins (Sumpio *et al.*, 2006), with tannins being the most active (Bravo, 1998). Polyphenols, being secondary metabolites, are derivatives of condensation reactions between cinnamic acid and three malonyl-CoA groups. These compounds exist as conjugates of mono and polysaccharides connected to phenolic groups (Balasundram *et al.*, 2006).

1.4 Health benefits of tea

The efficacious health promoting properties of green and black tea have been extensively documented in literature, particularly regarding protection against cardiovascular diseases and cancer (Bahorun *et al.*, 2012), obesity and diabetes (Uchiyama *et al.*, 2011), and several metabolic ailments (Thielecke and Boschmann, 2009). Epicatechin (EC), epicatechin gallate (ECg), epigallocatechin (EGC), and epigallocatechin gallate (EGCg) are the major catechins found in tea, with EGCg being the most copious, making up 50 - 80% of the total catechin content (Sang *et al.*, 2011). Structures for the catechin found particularly in green tea are

shown in Figure 1.6. Literature has documented the antioxidant, anti-hypertension, anti-inflammatory, anti-mutagenic, and anti-tumorigenic properties of tea stemming from the catechins, as well as the induced relaxation and enhanced cognitive function stemming from theanine (Lisman *et al.*, 2008; Haskell *et al.*, 2008). Additional research has revealed that the co-administration of drugs and e.g. teas rich in catechins EC and EGCg bring about the inhibition of glucuronidation and sulfation reactions of oral drugs, increasing their bioavailability within the body (Suganuma *et al.*, 2011). The theanine in tea induces relaxation; the reason for this is theanine translocates the blood brain barrier within 30 minutes, in a dose-dependent fashion, following its consumption (Terashima *et al.*, 1999). Furthermore, the relaxing effects of theanine have been documented in physiology studies focused on stress and anxiety (Kimura *et al.*, 2007), while also acting as an antagonist against the stimulatory effects of caffeine (Rogers *et al.*, 2008), and possessing exhibiting anti-hypertensive qualities (Yokogoshi and Kobayashi, 1998). In addition to this, theanine also effectively prevents liver damage resulting from an excessive intake of alcohol (Sadzuka *et al.*, 2005). Due to these documented pharmacological properties possessed by polyphenols, numerous tests on tea extracts have been conducted with these extracts serving as prophylactics e.g. the preclinical trials of polyphenon E in the chemoprevention of lung cancer (Lambert *et al.*, 2005), and EGCg as a human immunodeficiency virus reverse transcriptase antagonist (Nance and Shearer, 2003). Moreover, studies have also shown that theaflavins suppress HIV transcription (Gramza *et al.*, 2005). Lastly, catechins have extensively been documented to possess antioxidative attributes due to a high affinity for scavenging reactive oxygen and nitrogen species as illustrated in Figure 1.4 below. Catechins have a 3,4,5-trihydroxyl group on its B ring, with a gallate group esterified at the third position on its C ring, with positions 5 and 7 of its A ring having hydroxyl groups (Figure 1.5) (Frei and Higdon, 2003); the more hydroxyl groups located on each flavonoid ring, the more antioxidant properties the molecule has (Lien *et al.*, 1999). Catechins possess anti-cancer properties by suppressing the growth of cancerous cells through the blockage of angiogenesis. Catechins prevent diabetes by inhibiting the sodium transporter SGLT1's mechanism of action, therefore suppressing the uptake of glucose, and in turn lowering blood glucose levels (Khan and Mukhtar, 2007). As such, tea is no longer consumed just for the enjoyment of its gratifying taste and aroma, but also for its therapeutic properties (Zhang *et al.*, 2012). Figure 1.6 shows a list of the predominant catechins in green tea.

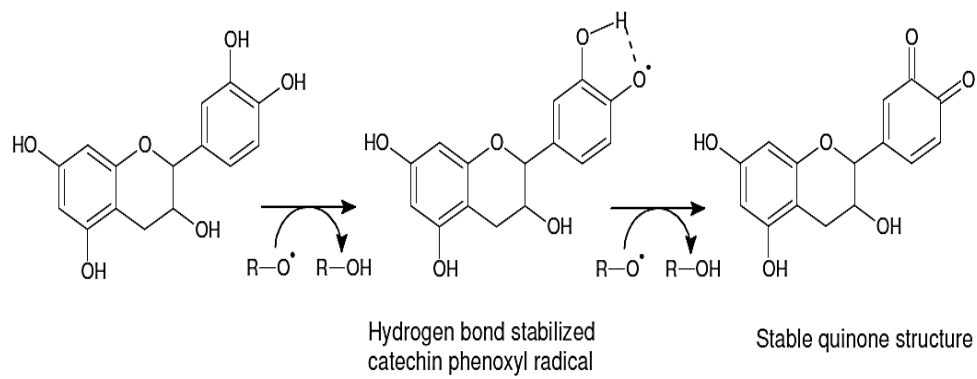


Figure 1.4: The antioxidant mechanism of action for catechin (Amic *et al.*, 2007).

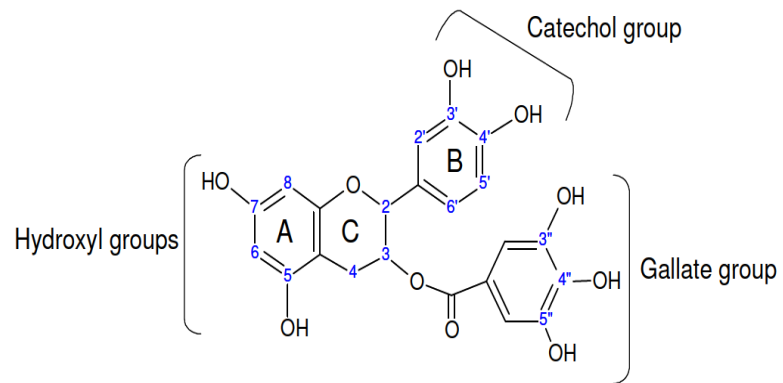


Figure 1.5: Groups conferring antioxidant properties on catechin (Frei and Higdon, 2003).

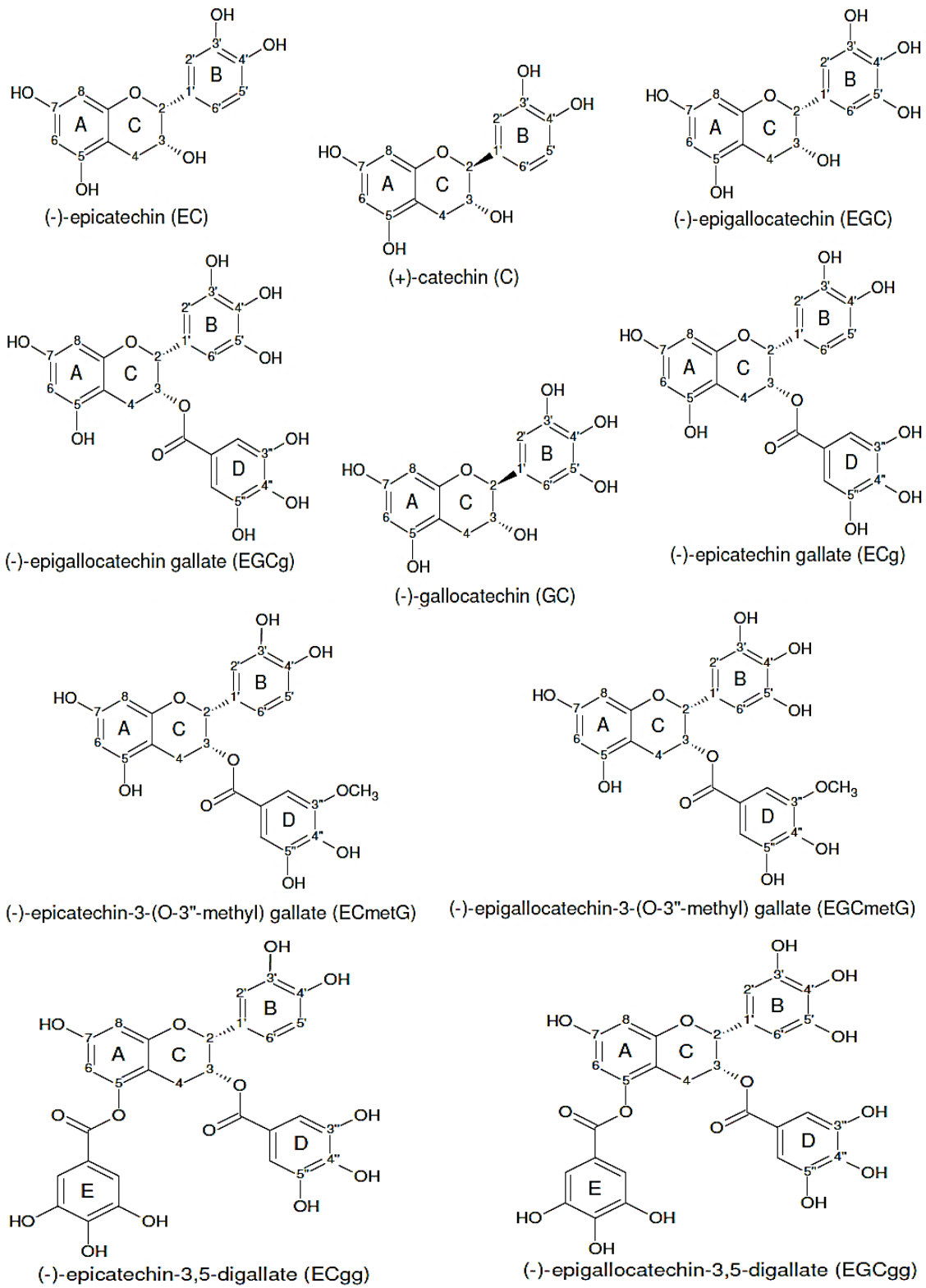


Figure 1.6: Catechins predominant in green tea.

1.5 Black tea

Unlike green tea, which is predominantly consumed in North Africa and the Middle East, black tea is primarily consumed in Western countries as well as in certain Asian countries e.g. India and Sri-Lanka, and throughout East African countries. Kenya, for example, specialises in the production of black tea and as of August 2018, has been declared the world's largest black tea producer and exporter, with a 23% market share, followed by China with 18% and thirdly Sri Lanka with 15% (Kariuki, 2018). In contrast to green tea, black tea predominantly consists of theaflavins, namely theaflavin (TF1), theaflavin-3- gallate (TF2), theaflavin-3'-gallate (TF3) and theaflavin-3, 3'-digallate (TF4). These are shown in Figure 1.7 below. These are formed through the oxidation of catechins, and constitute black tea's primary polyphenols (Sang *et al.*, 2011). In addition to catechins, metabolites such as caffeine, kaempferol, myricetin, theobromine, theophylline and quercetin can be found in black tea in minute quantities (Balentine *et al.*, 1997); these influence the quality of tea liquor.

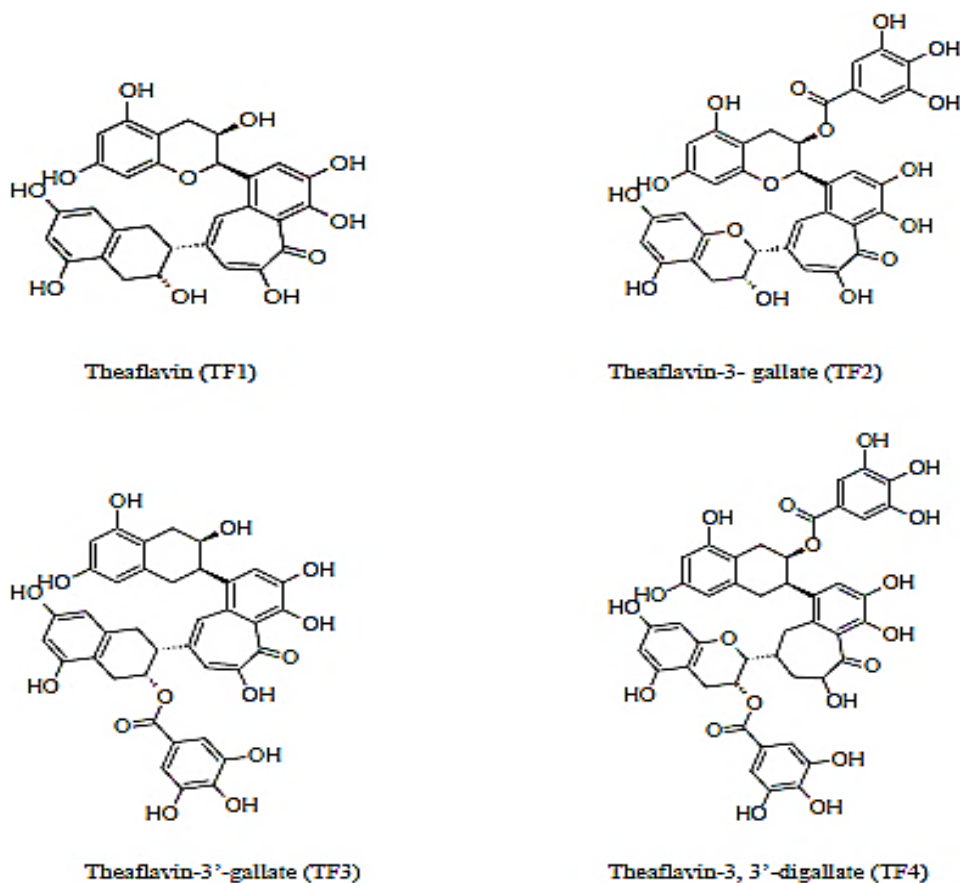


Figure 1.7: Structures of the major theaflavins.

1.6 Tea quality indicators

The polyphenolic metabolites have served as the principal tea quality indicators for a very long time. The organoleptic evaluation of black tea by tea tasters to ascertain tea liquor quality is through sight, smell and or taste of the tea liquor (Kumar *et al.*, 2011). Several studies have documented the correlation which exists between the fresh green leaf's phytochemical composition, the processing parameters employed to get black tea and the sensory quality attributes of the resultant black tea liquor (Obanda *et al.*, 2001). During organoleptic evaluation of black tea liquor, the tea tasters base their decision on five parameters, namely aroma, astringency, brightness, briskness, and colour (Hilton and Ellis, 1972). Caffeine, theaflavins, thearubigins, and several minor flavour related volatile compounds are all contributory to the resultant quality of tea liquor obtained (Owuor *et al.*, 2006a). Furthermore, higher concentrations of carotenoids and chlorophyll in green tea have been submitted as possible black tea quality indicators of the resultant liquor produced from these cultivar clones (Taylor *et al.*, 1992). Hilton and Ellis, (1972) documented the statistically significant linear correlation, which exists between theaflavin content and the market value of black teas from various tea producing countries globally. It is noteworthy to mention that though the Kenyan cultivars showed a positive correlation between theaflavin content and tea price, said correlation was not statistically significant, while cultivars obtained from Central Africa demonstrated a significant correlation between theaflavin content and tea price (Owuor *et al.*, 2006). These observed differences in theaflavin content could emanate from the influential differences of the geographical regions of production (McDowell *et al.*, 1995), or it could be due to the genetic variations in the cultivars i.e. there is inter and intra clonal variation when it comes to the formation of theaflavins via aeration (Magoma *et al.*, 2000). It is for this reason that theaflavin concentration in the tea liquor does not, on its own, always accurately represent black tea quality; the different rates of theaflavin formation due to varying levels of aeration. The astringency of black tea differs with the total composition of theaflavins; the reason being that individual theaflavins possess varying astringencies. The astringency obtained when all four theaflavins i.e. TF1 - TF4 combined, can serve as a black tea quality indicator (Owuor *et al.*, 2006).

1.7 Principle metabolites found in tea and their metabolic attributes

1.7.1 Flavonoids

Flavonoids, as abovementioned, are classified as anthocyanins, flavan-3-ols, flavanones, flavonols, flavones, and isoflavones (Del Rio *et al.*, 2013). Polyphenols are the predominant class of flavonoids, with catechins constituting more than half of them; the major catechins are catechin, EC, ECg, EGC, EGCg, catechin gallate (CG), and gallocatechin gallate (GCG) (Chanphai and Tajmir-Riahi, 2019), with EGCg accounting for the highest quantity. Moreover, methylated catechins and digallic acid bound catechins may also be located in *C. sinensis* (Zhang *et al.*, 2017). These catechin derivatives, in addition to their unique pharmacological effects, also interact with DNA by means of hydrophilic and hydrophobic interactions (Chanphai and Tajmir-Riahi, 2019). O-methylated-EGCg for example has been shown to alleviate Japanese cedar pollinosis (Masuda *et al.*, 2014). Catechins as mentioned earlier provide *C. sinensis* with a chemical defence against pathogens and herbivores (Masuda *et al.*, 2014). Figure 1.8 below shows the schematic representation of the flavonoid biosynthetic pathway.

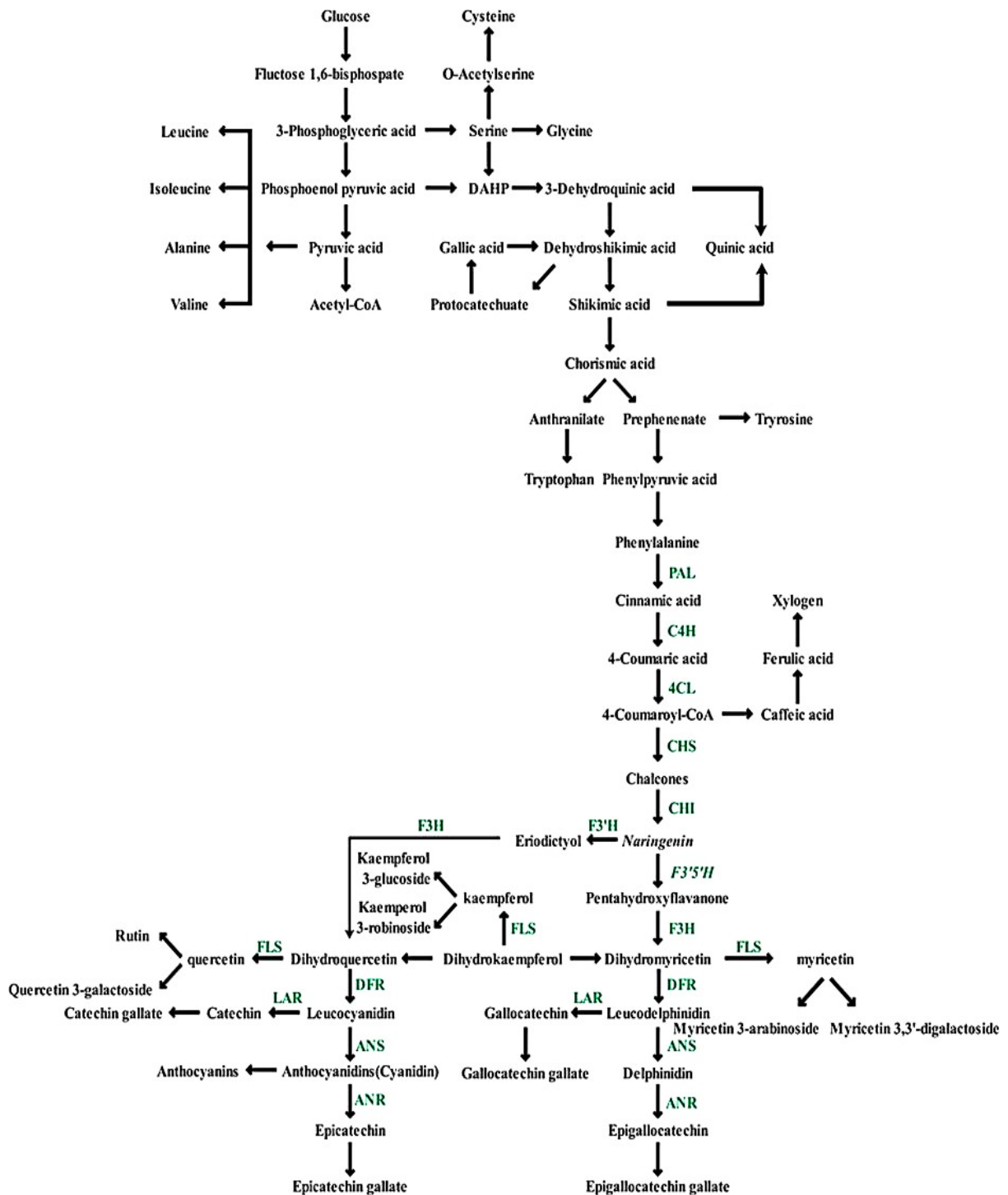


Figure 1.8: Flavonoid metabolism in *C. sinensis*.

Abbreviations: PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate-CoA ligase; CHI, chalcone isomerase; CHS, chalcone synthase; F3'5'H, flavonoid 3', 5'-hydroxylase; F3H, flavanone 3-hydroxylase; F3'H, flavonoid 3'-monooxygenase; FLS, flavonol synthase; DFR, dihydroflavonol-4-reductase; ANS, anthocyanidin synthase; ANR, anthocyanidin reductase; LAR, leucoanthocyanidin reductase (Jiang *et al.*, 2019).

1.7.2 Theanine

According to literature, 26 amino acids have been documented to be found in tea; these include 20 protein, and six non-protein amino acids. Theanine synthase, found in the roots, catalyses glutamic acid into theanine, which falls under the non-protein amino acids, and accounts for approximately 70% (w/w) of the free amino acids found in a leaf's dry weight; root theanine biosynthesis is a result of nitrogen absorption in the root of ammonia and nitrates (Deng *et al.*, 2010). Theanine functions, not only as a nitrogen reservoir, but also as a carbon backbone initiator for synthesis during germination (Sharma *et al.*, 2018). Theanine hydrolase, in the presence of sunlight, hydrolyses glutamic acid and ethylamine to theanine in the leaves. Ammonia oxidase induces the conversion of ethylamine into acetaldehyde, which is a catechin precursor (Kito *et al.*, 1968). Figure 1.9 below shows the metabolism of theanine in *C. sinensis*.

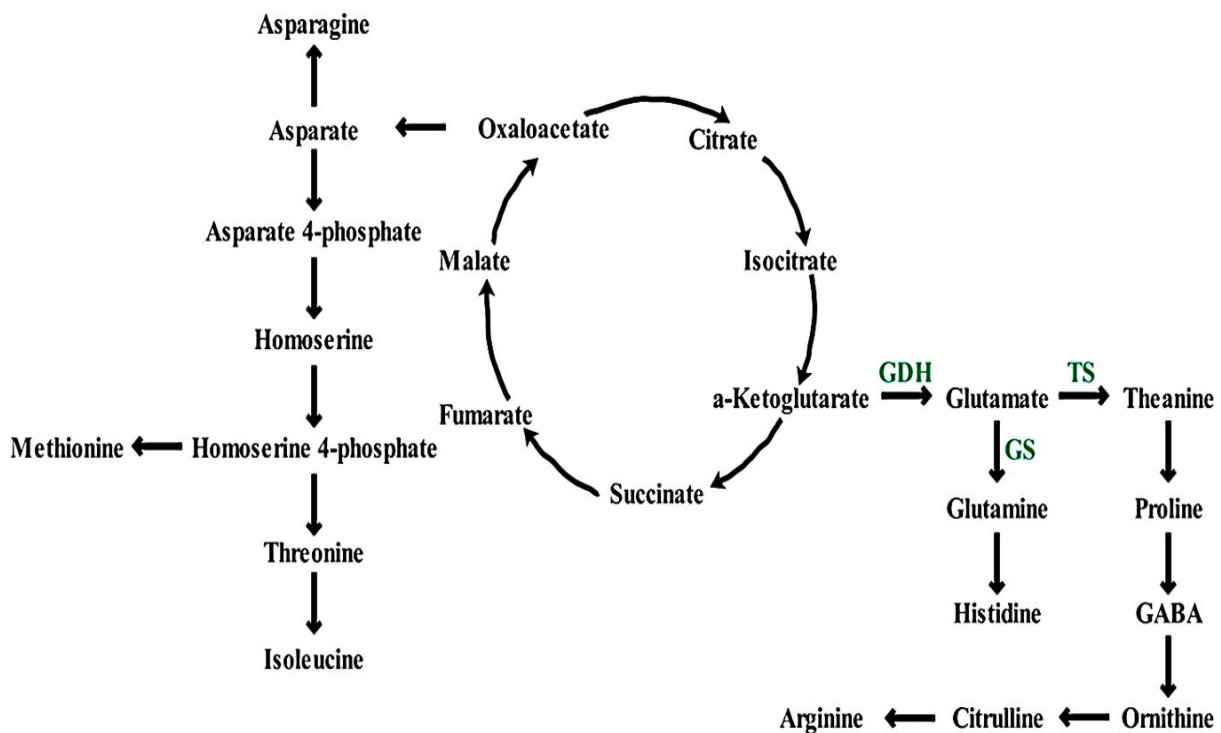


Figure 1.9: Theanine metabolism in *C. sinensis*. GDH represents glutamate dehydrogenase; TS represents theanine synthase, and GS represents glutamine synthase (Jiang *et al.*, 2019).

1.7.3 Caffeine

Caffeine is biosynthesised in the tender leaves of *C. sinensis* from both the *de novo*, and salvage pathways (Ashihara *et al.*, 2013), which involves a four step sequence involving one nucleosidase and three methylation reactions (Mohanpuria *et al.*, 2010). Caffeine's purine ring is formed from *de novo* biosynthesis from CO₂, formate, glutamine, and glycine, with adenine functioning as the predominant source. The alkaloid theobromine is a caffeine precursor. Approximately 99% of the formed caffeine is located in the young, fresh leaves; minute amounts of caffeine are also manufactured in the flowers, fruit, and roots of the tea plant, with most of it being bound to chlorogenic acid in the vacuole (Waldhauser and Baumann, 1996). Figure 1.10 below shows a schematic representation of caffeine's biosynthetic pathway.

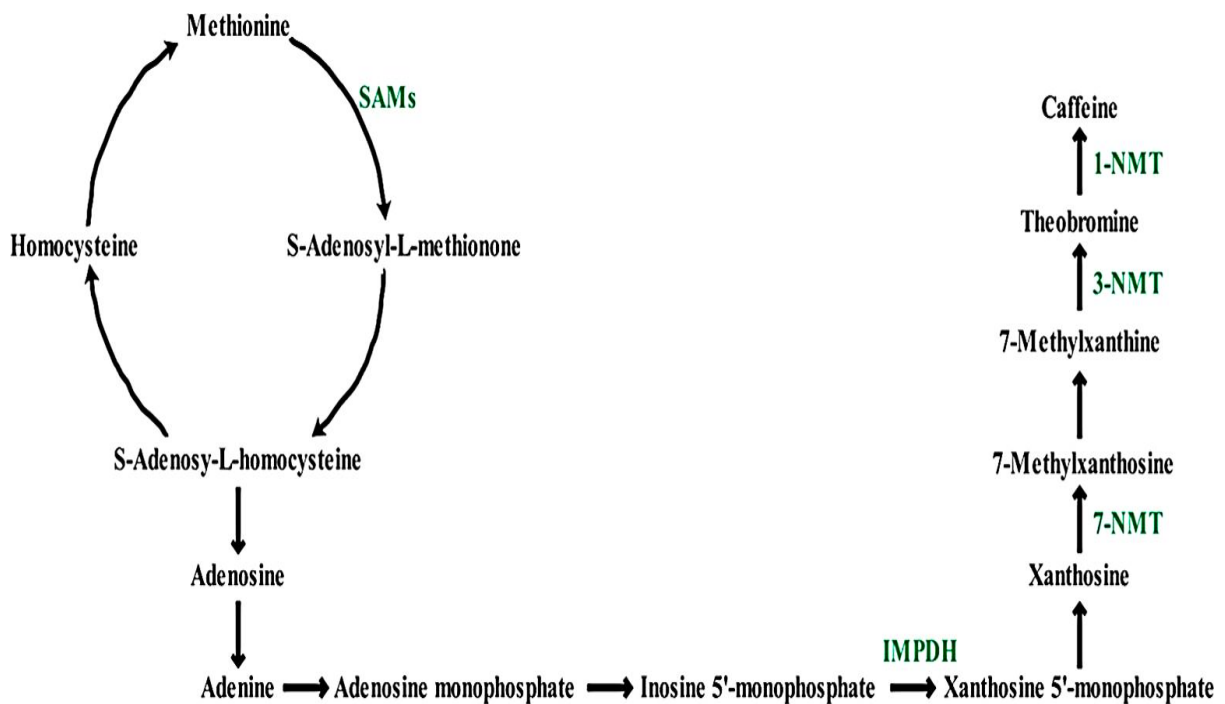


Figure 1.10: Caffeine metabolism in *C. sinensis*. SAMs represents S-adenosyl-L-methionone synthase, and IMPDH represents inosine 5' monophosphate dehydrogenase (Jiang *et al.*, 2019).

1.8 The tea breeding history in Africa

C. sinensis is an evergreen perennial tree distinguished from others, by a large 3.02 Gb diploid genome, with a $2n=2x=30$ chromosome number (Xia *et al.*, 2017). *C. sinensis* is out-crossing and highly self-incompatible (Wachira and Kamunya, 2005), therefore making it heterogeneous and heterozygous (Muoki *et al.*, 2007). It is postulated that the *Camellia* genus encompasses over 300 species; *C. sinensis* (L.) O. Kuntze has been documented to be the foremost prominent agronomic specie (Mondal *et al.*, 2004). The 2012 discovery of *C. cherryana* (Orel and Wilson, 2012), made evident how unstable and highly out-crossed the *Camellia* genus is. The *C. sinensis* chromosomes are small and have median centromeres, an indication of how primitive they are, limiting the advancement of polyploidy screening in tea (Singh *et al.*, 2013b). The first breeding stratagems targeted artificial pollination between cultivars differing in specific morphological attributes in a bid to produce superior cultivars (Willson and Clifford, 2012). The resultant seeds produced from the controlled pollination between the selected cultivar pairs were then replicated on test plots, with a well performing commercial cultivar serving as an observation control for attributes such as growth vigour, leaf quality and yield output. The two cultivar bushes with seedling progeny presenting significant promise were proliferated through vegetative propagation, planted as propagative clones in alternating rows in an isolated seedling fields to yield naturally cross-pollinated seeds. The superlative fields then served as checks for high yielding and better liquor quality producing clones (Green, 1966). The early African tea estates were planted using open pollinated seeds (Cannell *et al.*, 1977). In Assam and several north-east Indian regions, the prominence was placed on mass selection, involving randomised crosses between cultivars, which varied in leaf size and shape, growth rate, and texture (Wight, 1956). The issue with mass selection is that it regularly failed to yield high quality teas and failed to generate cultivars of constant morphological characteristics necessary for improved quality and yields (Richards, 1966). The upside of this is that numerous seed varieties were subsequently developed at Assam, which possess superior quality and yield traits compared to those randomly planted earlier. Sri-Lankan breeders constrained mass selection to selecting the superlative cultivars capable of vegetative propagation to give high yielding, identical progeny (Visser and Kehil, 1958).

1.9 Improvements in Kenyan tea breeding programmes

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) 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 suit and instituted programmes that saw the establishment of their own improved seed baries. 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 morphological consistency. Moreover, 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). Despite the abovementioned challenges, the tea breeding taking place at the TRI has resulted in the generation of new biclonal seed baries, while concurrently expanding the prevailing polyclonal seed baries. As of 2006, approximately 60% of clones concomitant with TRFK 6/8 have been commercialised, stemming from the Timbilil tea estate's breeding programme. Furthermore, 24 out of the 45 developed clones have found success in industry, amongst which are the elite Cambod varieties, TRFK 301/4 and TRFK 301/5. In addition to these, are the clones TRFK 430/90 and TRFK 371/3, which in addition to them having high yield and improved black tea quality, these new cultivars possess biotic and abiotic stress tolerance properties (Kamunya and Wachira, 2006). Breeders have used the TRFCA SFS 150 clone from Malawi and the TRFK 303/577 to produce varieties that are drought tolerant, such as the EPK TN 14-3, and have crossed the TRFCA SFS 150 and EPK TN 14-3 to produce F1 progeny tolerant to cold (Kamunya *et al.*, 2010); only the superior clones are being employed by farmers.

1.10 Breeding for high yield and environmental stress resistance

1.10.1 High yield

Plant breeders have been finding it daunting to develop high yielding clones from seedling mother bushes. Earlier studies (Green, 1971) failed to establish reliable correlations between growth and yield properties of mother bushes, and their resultant F1 progeny clones. Subsequent studies (Nyirenda, 1991) has shown adequately strong correlations between the tea bush area, shoot number, and yield of tea mother bushes and those of their clones. A

strong positive correlation between growth traits and yield in matured tea fields was observed (Shanmugarajah *et al.*, 1991).

1.10.2 Environmental stress

Due to 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 region (Elbehri *et al.*, 2015). Due to these changes in the climate, tea production is drastically being 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 altitudes. 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). In addition to this, stresses concomitant with temperature fluctuations in tea producing areas such as Kericho, Kisii, and Nandi, have added to the grave tea production limitations in Kenya. Tea production is also reliant on well scattered 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). The farming of tea has now been extended to areas previously deemed marginal and unsuitable for growing tea (Owuor *et al.*, 2010).

1.11 Current breeding strategies to obtain desirable characteristics

The objectives of tea breeding programmes vary from one geographical area to another; however due to the abovementioned effects of climate change, breeders are seeking varieties with improved resistance to the environmental stresses, while maintaining high quality. Tea breeders are concentrating on selecting and breeding populations rich in e.g. alkaloids such as caffeine, theobromine and theophylline; amino acids, namely theanine; carbohydrates like fructose and mannose; polyphenols, namely catechins, and proteins (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). In the past, countries such as Kenya, India, and Sri Lanka, which are high black tea producing tend to breed cultivars which will produce black tea rich in theaflavins as they have been documented to be high yielding and high black tea quality clones. Efforts have been made to combine these two qualities into an F1 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). A concerted effort is required to fully grasp and comprehend genetic variability found in tea, to aid in the development of the desired genotypes with the desired attributes (Kaundun and Matsumoto, 2003a). Studies have been conducted on linkage maps, MAS, molecular markers, mutation breeding, and QTLs to enhance breeding strategies (Chen *et al.*, 2013). Figure 1.11 shows a schematic representation of current tea breeding strategy.

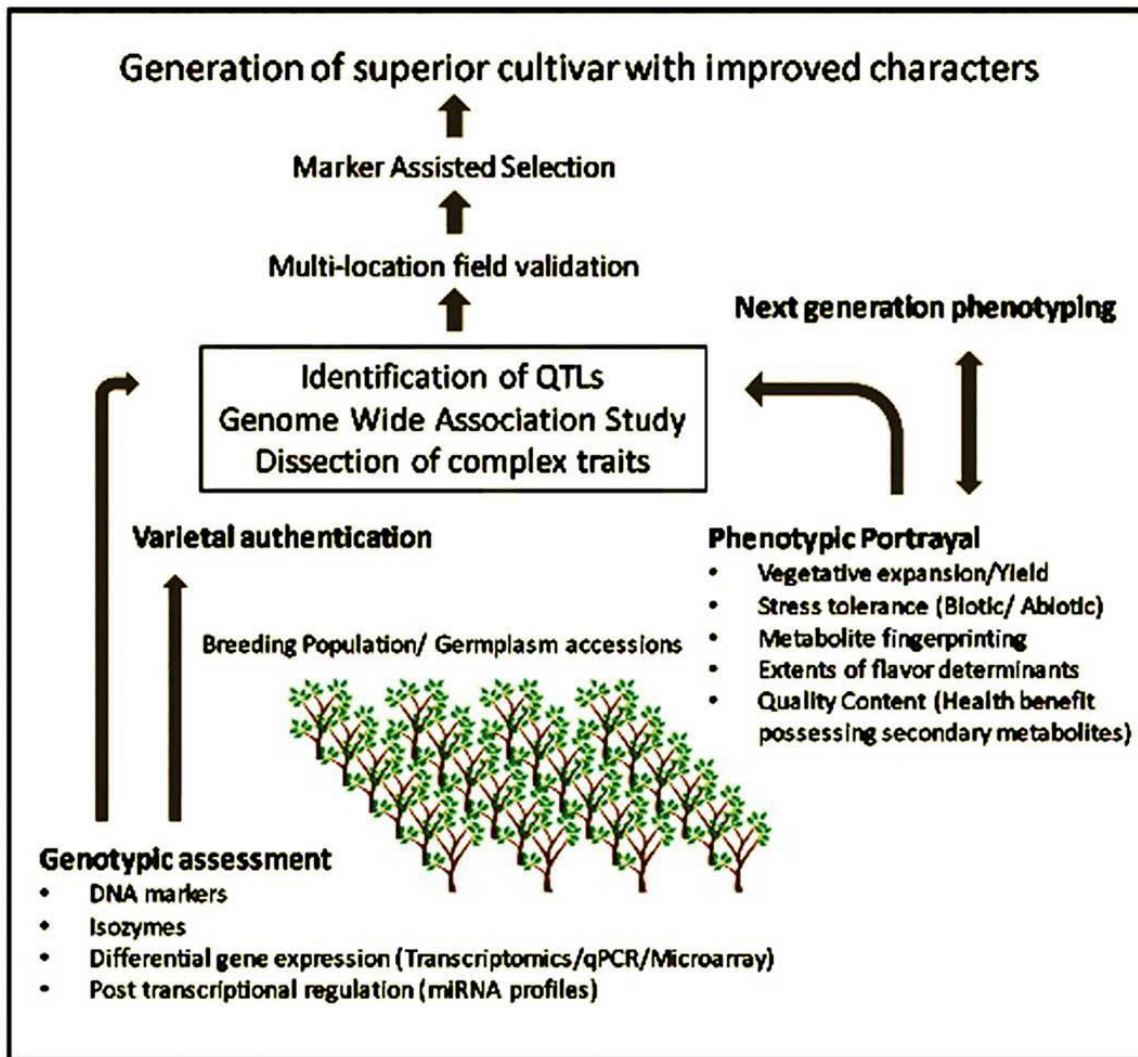


Figure 1.11: Schematic representation of tea breeding program (Hazra *et al.*, 2018).

1.12 Marker assisted selection

MAS is where the plants possessing the genes responsible for expressing the attributes the breeders are interested in, are selected through these molecular markers. Through the advances in, and convenience of molecular markers and genetic maps, the employment of MAS has become possible in many crops, as has been observed with the use of rice molecular markers to discover novel markers in other crops such as barley and wheat (Ellis and Nyirenda, 1995), and has been employed in the improvement of millet (Obanda *et al.*, 2001). Although *C. sinensis* genetic maps have been documented by (Hackett *et al.*, 2000; Adkins *et al.*, 2007; Wright *et al.*, 2002), and studies have been conducted on QTL mapping (Wright *et al.*, 2002; Koech *et al.*, 2018), which have identified markers and linkage groups

that are responsible for the expression of quantitative traits such as caffeine content, making the development of saturated linkage maps a necessary tool for MAS in the breeding of tea.

1.13 Role of genetic markers in tea breeding improvement

Genetic markers are specified chromosomal positions, functioning as genome analysis landmarks. These markers consist of biochemical, DNA, and morphological markers. Several converging biochemical and physiological characteristics are employed in conservative plant breeding to ascertain the cultivars' genetic multiplicity. As a result of its outbreeding nature, *C. sinensis* is decidedly heterogeneous. Studies to assess the genetic diversity of *C. sinensis* through the incorporation of biochemical (metabolite) markers (Das *et al.*, 2013), morphological markers (Chen *et al.*, 2007), and molecular markers (Kaundun and Park, 2002) have been conducted. A lot of these characteristics have, however, been documented to possess continuous variation, with the environment impacting them predominantly, making it difficult to identify robust markers required for genotyping (Ariyaratna and Gunasekare, 2007).

1.13.1 Biochemical/metabolite markers

The existing multiplicity found in *C. sinensis* cultivar varieties has efficiently been ascertained through the use of metabolite markers, such as anthocyanins, caffeine, catechins, and theanine (Li *et al.*, 2013). These biochemical markers are easily influenced by environmental factors, with particular interest being on the plant's stage of development at the time of exposure to these environmental factors (Das *et al.*, 2013).

1.13.2 Morphological markers

Due to the fact that morphological traits relate on a one to one basis with the genes regulating them, these markers can serve as reliable gene indicators. Literature has documented that a majority of morphological markers are a result of mutations (Waycott *et al.*, 1999). Due to the restricted availability of genetic mutants, the use of morphological markers in breeding has not been extensive (Worland *et al.*, 1987). Morphological markers have been documented to be impacted by environmental stresses, and have shown a continuous variation, making it challenging to ascertain distinct taxonomic clusters based on morphological markers (Koech *et al.*, 2018).

1.13.3 Molecular markers

The employment of molecular markers in crop breeding for favourable agronomic characteristics presents breeders with improved opportunities to attain even those traits considered problematic to assess through the use of biochemical, and morphological markers.

Molecular markers, unlike the biochemical or morphological, are plenty, and are least impacted by environmental stresses (Singh *et al.*, 2013a). Furthermore, molecular markers offer an effective, ancillary option for differentiating intra- and or inter-specific germplasm differences, and in so doing, serving as an indispensable tea breeding instrument (Ni *et al.*, 2008). Based on how they are detected, molecular markers can be categorised into three categories, namely DNA sequence based, hybridisation based, and polymerase chain reaction (PCR) based (Angaji, 2011). In a study by Kamunya *et al.*, (2009), *C. sinensis* DNA molecular markers were investigated in cultivars that are known to have a tolerance for drought, cold temperatures, and diseases. In a follow up study, it was ascertained that these markers also serve as indicators for cultivars that will be high yielding and will produce high quality tea liquor (Kamunya *et al.*, 2010). Investigations have also been conducted on Chinary cultivars, by studying the catechin regulatory DNA molecular markers in green tea for MAS breeding (Ma *et al.*, 2014). The use of molecular markers has been documented as a valuable and adequate tool for characterising and discriminating between *C. sinensis* varieties (Kaundun and Matsumoto, 2003b).

1.13.4 Diversity array technology

Diversity Array Technology (DArT) has, according to (Jaccoud *et al.*, 2001), been defined as “*a high throughput microarray hybridisation based method involving the isolation and cloning of randomised DNA fragments from complexity-reduced DNA sample.*” DArT provides a homogeneous, high throughput genotyping, which allows for thousands of molecular markers from thousands of samples to be concurrently assayed, without any preceding sequence information (Wittenberg *et al.*, 2005). The use of DArT has yielded outstanding results in diversity and phylogenetic studies (Steane *et al.*, 2011), and genomic linkage mapping and selection (Poland *et al.*, 2012; Schouten *et al.*, 2012). Studies on crops with multifaceted genomes i.e. apple (Larsen *et al.*, 2018) lemons and oranges (Sagawa *et al.*, 2018), strawberry (Vallarino *et al.*, 2019), sugarcane (Grzebelus, 2015), wheat (Baloch *et al.*, 2017), have documented the successful application of DArT markers. These DArT markers serve as improved alternatives to the presently employed techniques, which include restriction fragment length polymorphism, amplified fragment length polymorphism, simple sequence repeats, and single nucleotide polymorphisms, especially regarding costs and the promptness of marker discovery and whole-genome fingerprint analysis. This technique is cost efficient, non-gel dependent technology that is acquiescent with high throughput mechanisation and the detection of superior markers in a single assay (Figure 1.12).

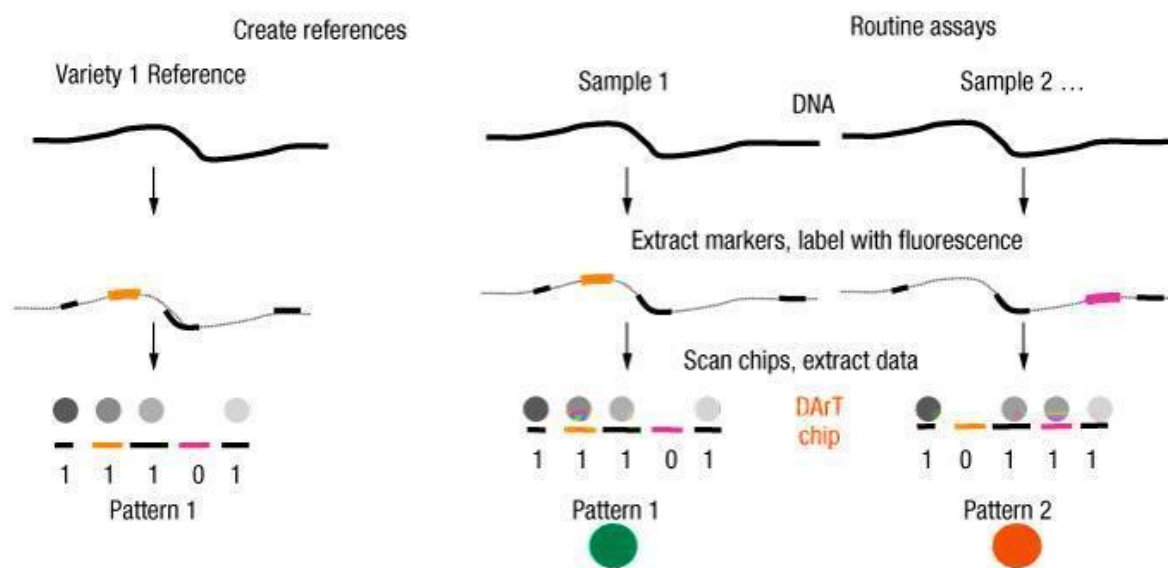


Figure 1.12: Diagrammatic representation of the DArT principle (Koech *et al.*, 2018).

1.13.5 Quantitative trait loci (QTL)

Polygenes are the genes responsible for controlling numerous significant crop traits, including abiotic and biotic stress tolerance, quality, and yield. Analysing QTLs involves detecting linkages between the genotype markers and the resultant phenotype of the progeny assessed for the trait interested in (Collard *et al.*, 2005). The relationship between the marker and the QTL is indicated by the difference (significant) between average values of the traits of interest and the genotype of their markers (Miles and Wayne, 2008). The analysis of QTLs furnishes information regarding the nature of each QTL, its position on the chromosome and its function. Molecular markers have been efficaciously employed in researching quantitative inheritance in crops (Agarwal *et al.*, 2008), such as in tomato, where QTL research resulted in the development of cold tolerant, and insect resistant progeny (Grandillo and Cammareri, 2016), and in wild barley (Elberse *et al.*, 2004), while (Ma *et al.*, 2014) documented markers significantly connected with *C. sinensis*. Metabolomics allows for functional gene mining in metabolic networks integrated with both genomics and transcriptomics. Metabolite quantitative trait loci (mQTLs) furnish information regarding genotypic and phenotypic associations, for mining and validating possible functional genes. Through the employment of this stratagem by (Kato *et al.*, 2000), the genes which encode for the enzyme caffeine synthase, were cloned, bringing with it an opportunity for tea breeders to develop naturally low caffeine content cultivars. In another study, three genes responsible for nitrogen utilisation regulation were identified in two *C. sinensis* cultivars. These genes were also

found to encode for the enzymes which control the caffeine, catechin metabolic pathways in the *C. sinensis* plant. It was, however, unfortunate that the majority of these were candidate genes, which lacked validation by the transcriptome instead of the genome. Furthermore, the absence of established transgenesis and tissue culture limits the transformation system in *C. sinensis*. This is, however, a way of quickly identifying other unknown functional genes linked to metabolism, by combining metabolomics with molecular markers (Li *et al.*, 2017).

1.14 Challenges faced in genomic selection

When the application of genomic selection is being considered, one important consideration to be taken into account is the interaction between the genotype and the environment (GxE). The application of genomic selection in wheat grain yields totalling 599 lines under four different environments, each analysed by cross-validation, established an accuracy variation of between 0.44 and 0.6 across the environments. Moreover, the study results revealed a significant difference in the accuracy of maize grain yield, ranging between 0.41 and 0.52 under wet and drought conditions (Crossa *et al.*, 2010). In a follow up study with additional environments, it was confirmed that different environments significantly affect the accuracy of the genomic selection results (Crossa *et al.*, 2011). Furthermore, another study employed genomic selection on pines possessing a comparable number of clonal individuals and markers. These clonal individuals were spread across four locations; it was established that the equation generated from results obtained at any one of the locations served as a good predictor, approximating an accuracy of 0.7 within site, and decreasing across the sites (Resende Jr *et al.*, 2012). This indicates that GxE interactions significantly influence the accuracy of genomic selection, even within identical clones exposed to varying environments. As such, it is crucial that the impact of the environment be taken into account when developing genomic prediction models.

1.15 Plant metabolomics

According to Hamanishi *et al.*, (2015) when plants are exposed to abiotic stress, this results in the disruption of the plant's metabolic pathways, to facilitate the plant's survival; they have developed inestimable drought response stratagems (Ogbaga *et al.*, 2014). Crop breeders are continuously seeking knowledge and understanding of the mechanisms employed by crops to survive in drought stricken and salinised environments (Nyarukowa *et al.*, 2016). These adaptive response processes involve regulatory activation of multiple genes, which in turn activates subsequent metabolic pathways. Studies on drought stressed plants have been

conducted, revealing the significance of metabolic regulation, such as a build-up of osmolytes as a response (Slama *et al.*, 2015). Osmolytes such as betaine, mannitol, proline, and trehalose are produced under hyperosmotic stress (Weckwerth *et al.*, 2004). In addition to inducing osmolyte syntheses to maintain turgor through osmotic adjustment (Arbona *et al.*, 2013), these metabolites also serve to stabilise protein conformations, while diminishing protein-solvent interactions, and facilitating the repair of damaged tissues (Ruan and Teixeira da Silva, 2011). Glucosinolates are metabolites rich in nitrogen and sulphur, and are biosynthesised from leucine, methionine, phenylalanine, tryptophan, tyrosine, and or valine (Arbona *et al.*, 2013). Bound to glucosinolate's side chains are a β -thioglucosyl, and a hydroxyaminosulphate residue; these have been associated with the defence mechanisms plants employ against abiotic stresses such as drought. The enzyme myrosinase cleaves the β -thioglycosidic bond on the β -thioglucosyl residue, giving isothiocyanates, nitriles, and thiocyanates (Zandalinas *et al.*, 2012). These are then conjugated with intracellular glutathione` resulting in the glucosinolates' bioactivity (Keum *et al.*, 2005).

Phytometabolomics is the study of plant metabolic profiling, which encompasses the qualitatively and quantitatively analysis of metabolites to better comprehend their metabolic responses under abiotic and biotic stress (Schauer and Fernie, 2006). Comprehending the desiccation response metabolome assists in ascertaining steps involved in the signal transduction pathways (Urano *et al.*, 2009). Metabolic profiling commenced as a diagnostics tool to ascertain herbicide mode of action, and has since grown to include functions such as determining the differences between genetically modified and conservative crops, and genotyping them to discover new genes (Hagel and Facchini, 2008). Figure 1.13 below shows a typical metabolomics workflow:

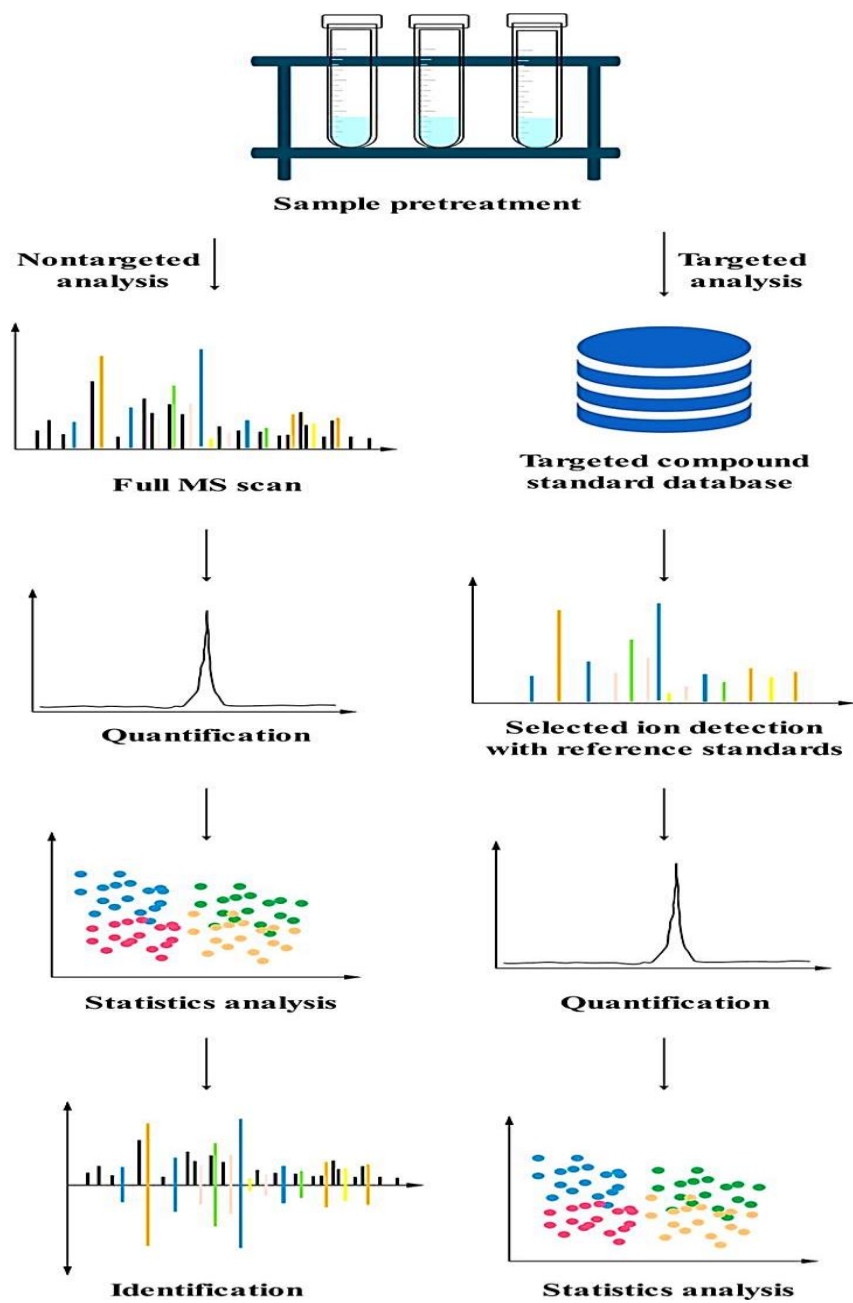


Figure 1.13: A metabolomics workflow documenting steps involved in a metabolomics study (Jiang *et al.*, 2019).

Metabolomics is a post-genomics technique, which explores the link between genes and metabolic networks, by comprehensively investigating the numerous metabolites found in plants (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), with it being used predominantly in differential profiling and biomarker identification (Theodoridis *et al.*, 2012). At present, direct injection mass spectrometry (DIMS), Fourier transform infrared spectroscopy (FTIR), and nuclear magnetic resonance (NMR), are mostly used in plant metabolomics research. Gas chromatography (GC)-MS is another preferred technique employed in plant metabolomics for terpenoid (Chen *et al.*, 2003) and several other volatile analysis (Tikunov *et al.*, 2005). The main disadvantage with GC is its restriction to small volatile biomolecules, meaning it cannot be employed to analyse biomolecules possessing a larger molecular weight i.e. proteins. LC-MS has enjoyed a steady growth over the years, as the metabolomics technique of choice due to its high throughput, soft ionisation, and extensive metabolite coverage (Zhou *et al.*, 2012). 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 the formation of a new hypothesis, which involves assessing all the metabolites in a biological system (Zhou *et al.*, 2012). By combining 1.17, LC-MS, and NMR data, a comprehensive picture of a plant metabolome is obtained. As a result of the increase in plant metabolomics techniques, using non-targeted approaches is now a favourable option for the identification and quantification of hundreds of metabolites, giving an extensive insight into the samples metabolite profiles e.g. tea (Fraser *et al.*, 2012). LC-MS has been established as predominant favourite targeted profiling technique especially for plant metabolomics studies; LC-MS-based metabolomics depends on numerous analytical, computational, and experimental steps (Zhou *et al.*, 2012). When employed as the analytical tool, LC-MS produces mass spectral peak lists, which after being aligned to their respective samples, are related with multivariate statistics, resulting in spectral feature identification.

No single technique adequately detects, identifies, and quantifies all metabolites in animal or plant samples. It is for that reason that the abovementioned metabolomics platforms are to be used in combination to encompass a majority of them. Hybridising GC, and LC-MS with NMR has been documented to result in the most favourable ascertainment of a sample's metabolite profile (Ward *et al.*, 2007). The field of metabolomics research has significantly grown over the last decade from approximately 6 000 publications in the year 2008, to 22 000 in the year 2018. This is shown in Figure 1.14 below.

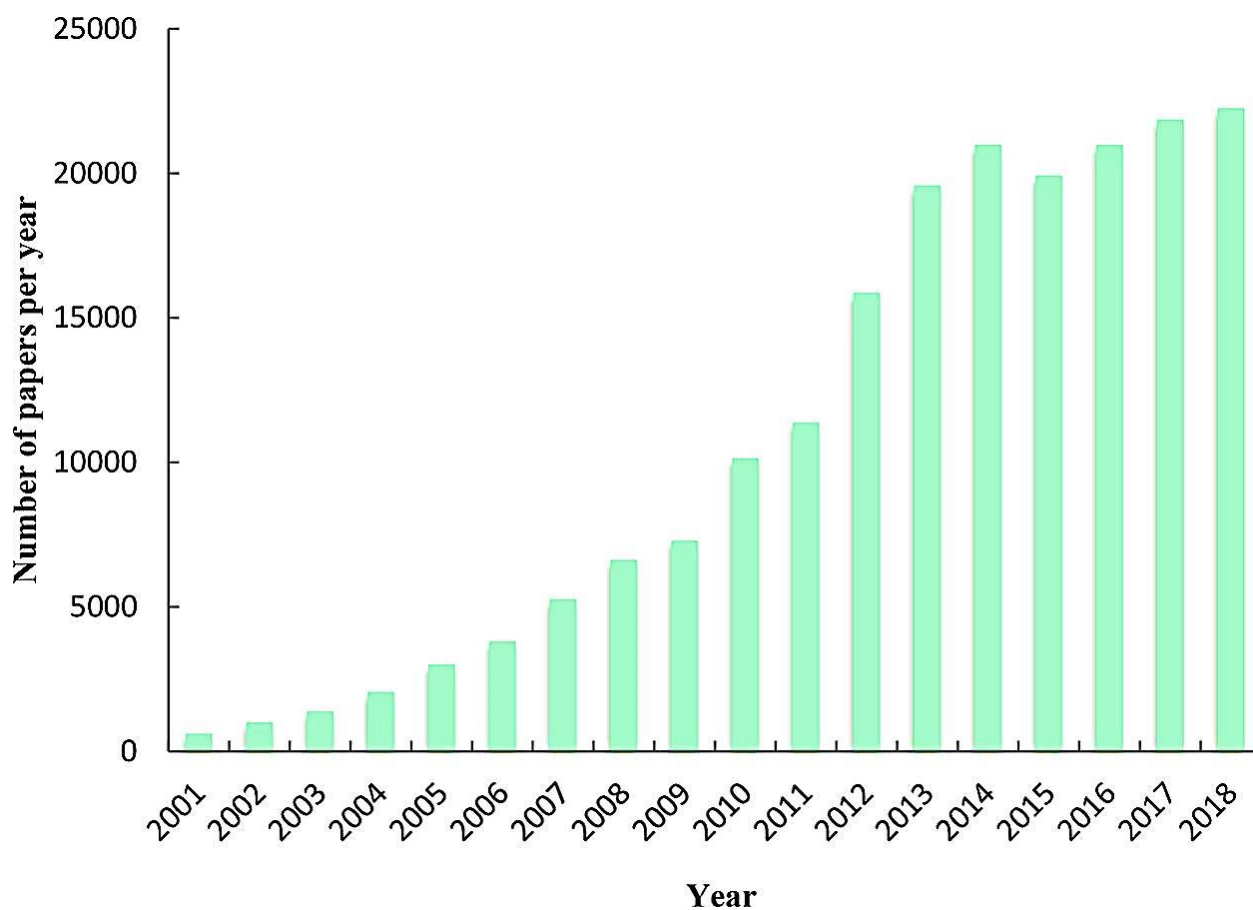


Figure 1.14: Number of publications on metabolomics from 2001 to 2018.

1.16 Statistical analysis

In metabolomics, uni- and multivariate statistical techniques are used in combination to help pinpoint relevant variation (e.g. between groups of interest) in datasets that are often large and high-dimensional. The univariate statistical tool used here was the t-test with resulting p-value and associated effect size. Two multivariate approaches were included, principal component analysis (PCA), and partial least squares discriminant analysis (PLS-DA). PCA is an unsupervised, projection technique that makes it possible to view large datasets by summarising variation through projection onto fewer dimensions. The separation between the classes provides a measure of validation for PLS-DA, which is prone to overfit. PLS-DA is a supervised technique used to identify combinations of variables that can distinguish between classes of samples.

Logistic regression (LR) is a statistical analysis tool generally suitable for testing hypotheses regarding connections between categorical outcome variables and continuous predictor variables. LR solves problems that cannot be solved by simple linear regression, such as any occurring errors that are not normally distributed or are not constant throughout the data range (Peng *et al.*, 2002). Contrasting from discriminant analysis, LR does not make the assumption that the predictor variables possess equal covariance matrices, and that these are normally distributed. It instead makes the assumption that the distributions of any errors equalling the true Y value subtracting the predicted Y value are described by the binomial distribution. This implies an identical probability is maintained across the range of predictor values. This binomial assumption is therefore easily testable using a Z-test (Siegel and Castellan, 1956). LR may be considered robust, provided the samples are random; in so doing this ensures the observations remain independent of one another (Peng *et al.*, 2002). Another useful statistical analysis approach is making use of decision trees, which are created through the use of partition algorithms. These algorithms employ the links between predictors and their corresponding responses, and recursively partition the data, splitting predictors until the desired prediction response is obtained. Through these repeated data partitions, a decision tree is formed. By choosing the best splits from an infinite number of possibilities, the partition algorithm makes the decision trees a powerful modelling tool. Predictors are either continuous or categorical; where continuous, the partitions are a result of a cut off value, with sample values falling above and below this cut off value. If, on the other hand, the predictor is categorical, the samples will be split into two levels (JMP®). The decision tree identifies independent variables with a significant relationship to the dependent variable and evaluates

the continuous variables' interval breaks to identify the most ideal combination. The independent variable possessing the sturdiest relationship with the dependent variable then becomes the decision tree's first branch; each significantly different category, relative to the target variable becomes the leaf. This is continually done to identify each leaf's significant predictor variable until predictors are exhausted (Thomas and Galambos, 2004). Violin plots are a statistical method considered to be a combination of the box plot and a kernel density plot, which are used for plotting numeric data. The violin plot contains the same information as would be found in a box plot, but have the indisputable advantage over the box plot in that they show the entire data distribution, which is beneficial when working with multimodal data i.e. distribution with several peaks (Hintze and Nelson, 1998).

1.17 PROBLEM STATEMENT

The polyphenolic metabolites have served as the principal tea quality indicators for a long time. A positive correlation between theaflavin content and tea price has been reported in Kenyan teas (Obanda *et al.*, 2001), while cultivars obtained from Central Africa demonstrated a significant correlation between theaflavin content and tea price (Hilton and Ellis, 1972). Caffeine, theaflavins, thearubigins, and several minor flavour related volatile compounds are all contributory to the resultant quality of tea liquor obtained (Owuor *et al.*, 2006a), and these are under the control of multiple genes called QTLs. Using conservative approaches for genetic enhancement of the tea plant i.e. increasing caffeine, theaflavins, thearubigins content, has proven laborious and time-consuming, especially since *C. sinensis* is out-crossed and highly self-incompatible (Wachira and Kamunya, 2005), has a low seed production count and the overall lack of available genetic markers. Employing QTLs for certain key agronomic traits can help to select tea cultivars with the desired traits at an early stage of plant growth, especially since tea is a woody plant, and using conservative breeding approaches may bring about delays due to tea having a long juvenile phase of between 22-25 years. When the application of genomic selection is being considered, one important consideration is the interaction between the GxE. GxE interactions have been shown to influence the accuracy of genomic selection, even within identical clones exposed to varying environments. The impact of the environment should be considered when developing genomic prediction models. Phytometabolomics, the study of plant metabolic profiling, which encompasses the qualitatively and quantitatively analysis of metabolites to better comprehend their metabolic responses under abiotic and biotic stress may explore the link between genes and metabolic networks, by comprehensively investigating the numerous metabolites found in plants. Tea is the most consumed beverage, second to water, and is therefore an important commodity. With the environmental changes being brought about by global warming, it has become the most important objective to tea breeders to develop tea cultivars with the potential of producing high yields, which are drought tolerant and essentially produce high quality green and or black teas. Caffeine, catechins, theaflavins, and thearubigins are major determining factors of tea quality, and are yet to be subjected to QTL analysis. This study seeks to, using metabolomics, identify and map out markers associated with QTLs for quality determinants, and to use these to develop predictive models to assist tea breeders in new tea cultivar selections.

1.18 RESEARCH OBJECTIVE

The main objective of this study was to identify and validate the biochemical and molecular markers linked to yield, drought tolerance, and quality traits of black tea.

First, this study prioritised the replanting schedule of seedling tea fields on estates commonly subjected to drought by developing a sampling method and estimating drought susceptibility using the SWAPDT method, this being done in the absence of historical in-filling records.

Second, this study identified putative QTLs associated with amino acids, caffeine, catechins, organoleptic evaluation and %RWC, using DArTseq markers, GC-MS, ¹H-NMR and UPLC platforms, to construct genetic linkage maps for MAS in tea breeding.

Third, this study made use of metabolomics generated data to identify differently expressed metabolites in the two groups of cultivars, following which these metabolites would be used to develop predictive models to classify the 310 genotypes as either Commercial or NonCommercial cultivars. The best model would be used in new field selections.

The thesis is organised into the following chapters:

Chapter 1: Literature review

Chapter 2: Prioritising the replanting schedule of seedling tea fields on tea estates for drought susceptibility measured by the SWAPDT method in the absence of historical in-filling records

Due to the unpredictable natural droughts that occur, causing tea farmers significant losses in tea estates, a two-day method for distinguishing drought tolerant (DT) from drought susceptible (DS) *Camellia sinensis* cultivars was developed. The findings suggest that where historical in-filling records are not available this method may be used to prioritise fields for replanting.

Chapter 3: Identification of QTL's responsible for yield, drought tolerance and quality traits in *Camellia sinensis* using GC-MS, ¹H-NMR and UPLC

Tea consumers concern themselves with the quality of tea. The breeding for these high yielding, DT, high quality phenotypic traits is challenging due to the fact that these are qualitative traits inherited from parents, and influenced by environment. Through the use of molecular markers to identify gene regions linked with the phenotypic traits of interest, marker-assisted selection may be employed to select high yielding, DT and high quality tea cultivars. The putative QTLs involved in identifying these traits of interest in tea based on DArTseq markers reveals the proteins, and possible enzymes linked to the traits of interest.

Chapter 4: Models for identification of elite mother bushes with high black tea commercial potential from mature seedling fields of *Camellia sinensis*

The quality of tea is undeniably affected by variations in its metabolite composition. Tea producers are in demand of new cultivars, which produce high quality tea liquors. This chapter involved the careful study of metabolites influencing tea quality using metabolomics. GC-MS, LC-MS, ¹H-NMR, and UPLC-DAD platforms were used. Logistic regression models were developed to find variables capable of classifying the 310 genotypes as either Commercial or NonCommercial cultivars.

Chapter 5: Concluding discussion and recommendation.

1.19 RESEARCH OUTPUTS

1. Peer-reviewed paper (Appendix 2.1): Nyarukowa C. T., Koech K. R., Loots T., Hageman J., and Apostolides Z (2018). Prioritising the replanting schedule of seedling tea fields on tea estates for drought susceptibility measured by the SWAPDT method in the absence of historical in-filling records. *Journal of Agricultural Science* 10 (7): 26-34.
2. Peer-reviewed paper (Appendix 4.1): Nyarukowa CT., van Reenen M., Koech RK, Kamunya SM., Mose R., and Apostolides Z. Multivariate models for identification of elite mother bushes with high commercial potential for black tea from mature seedling fields of *Camellia sinensis*. *International Journal of Research in Agronomy* Vol. 4, No. 1 (2020).
3. Conference: Nyarukowa C. T., Koech K. R., Loots T., and Apostolides Z (July 2018). SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. Invited poster presentation. The 26th South African Society of Biochemistry and Molecular Biology (SASBMB) conference held in conjunction with the Federation of African Societies of Biochemistry and Molecular Biology (FASBMB), North-West University (NWU), Potchefstroom, South Africa.
3. Conference: Nyarukowa, C., Koech, R., Loots, T. & Apostolides, Z. (August 2018). SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. Presented at the 15th ACGT Regional Plant Biotechnology Forum on “Plant Genomes: From Plants to Networks”, University of the Witwatersrand, Johannesburg, South Africa.

1.20 REFERENCES

- Adkins, N. L., Hall, J. A. & Georgel, P. T. (2007). The use of quantitative agarose gel electrophoresis for rapid analysis of the integrity of protein–DNA complexes. *Journal of Biochemical and Biophysical Methods* 70(5): 721-726.
- Agarwal, M., Shrivastava, N. & Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Reports* 27(4): 617-631.
- Allwood, J. W., Ellis, D. I. & Goodacre, R. (2008). Metabolomic technologies and their application to the study of plants and plant–host interactions. *Physiologia Plantarum* 132(2): 117-135.
- Allwood, J. W. & Goodacre, R. (2010). An introduction to liquid chromatography–mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis* 21(1): 33-47.
- Amic, D., Davidovic-Amic, D., Beslo, D., Rastija, V., Lucic, B. & Trinajstic, N. (2007). SAR and QSAR of the antioxidant activity of flavonoids. *Current Medicinal Chemistry* 14(7): 827-845.
- Angaji, S. A. (2011). Genomics at a Glance (I): Molecular Markers & Genome Mapping. *LAP LAMBERT Academic Publishing*.
- Anon. (1990). Seed garden (barie). Annual Report., 25: Tea Research Foundation of Kenya, Tea Board of Kenya
- Arbona, V., Manzi, M., Ollas, C. d. & Gómez-Cadenas, A. (2013). Metabolomics as a tool to investigate abiotic stress tolerance in plants. *International Journal of Molecular Sciences* 14(3): 4885-4911.
- Ariyaratna, C. & Gunasekare, K. (2007). Genetic base of tea (*Camellia sinensis* L.) cultivars in Sri Lanka as revealed by pedigree analysis. *Journal of Applied Genetics* 48(2): 125-128.
- Ashihara, H., Yokota, T. & Crozier, A. (2013). Biosynthesis and catabolism of purine alkaloids. *In Advances in Botanical Research* 68: 111-138.
- Bahorun, T., Luximon-Ramma, A., Neergheen-Bhujun, V. S., Gunness, T. K., Googoolye, K., Auger, C., Crozier, A. & Aruoma, O. I. (2012). The effect of black tea on risk factors of cardiovascular disease in a normal population. *Preventive Medicine* 54: S98-S102.

- Balasundram, N., Sundram, K. & Samman, S. (2006). Phenolic compounds in plants and agri-industrial by-products: Antioxidant activity, occurrence, and potential uses. *Food Chemistry* 99(1): 191-203.
- Balentine, D. A., Wiseman, S. A. & Bouwens, L. C. (1997). The chemistry of tea flavonoids. *Critical Reviews in Food Science & Nutrition* 37(8): 693-704.
- Baloch, F. S., Alsaleh, A., Shahid, M. Q., Çiftçi, V., de Miera, L. E. S., Aasim, M., Nadeem, M. A., Aktaş, H., Özkan, H. & Hatipoğlu, R. (2017). A whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. *PloS One* 12(1): e0167821.
- Bravo, L. (1998). Polyphenols: chemistry, dietary sources, metabolism, and nutritional significance. *Nutrition Reviews* 56(11): 317-333.
- Cannell, M., Njuguna, C., Ford, E., Smith, R. & Ross-Parker, H. (1977). Variation in yield among competing individuals within mixed genotype stands of tea: a selection problem. *Journal of Applied Ecology*: 969-985.
- Chang, K. (2015). World tea production and trade Current and future development, Food and Agricultural Organization of the United Nations.
- Chanphai, P. & Tajmir-Riahi, H. (2019). Structural dynamics of DNA binding to tea catechins. *International Journal of Biological Macromolecules* 125: 238-243.
- Chen, F., Duran, A. L., Blount, J. W., Sumner, L. W. & Dixon, R. A. (2003). Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. *Phytochemistry* 64(5): 1013-1021.
- Chen, L., Apostolides, Z. & Chen, Z.-M. (2012). Global tea breeding: achievements, challenges and perspectives. Springer Science & Business Media.
- Chen, L., Zhou, Z.-X. & Yang, Y.-J. (2007). Genetic improvement and breeding of tea plant (*Camellia sinensis*) in China: from individual selection to hybridization and molecular breeding. *Euphytica* 154(1-2): 239-248.
- Cheruiyot, E. K., Mumera, L. M., NG'ETICH, W. K., Hassanali, A. & Wachira, F. (2007). Polyphenols as potential indicators for drought tolerance in tea (*Camellia sinensis* L.). *Bioscience, Biotechnology, and Biochemistry* 71(9): 2190-2197.
- Cheserek, B. C., Elbehri, A. & Bore, J. (2015). Analysis of links between climate variables and tea production in the recent past in Kenya. *Donnish Journal of Research in Environmental Studies* 2(2): 5-17.

- Collard, B. C., Jahufer, M., Brouwer, J. & Pang, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* 142(1-2): 169-196.
- Cooper, R., Morré, D. J. & Morré, D. M. (2005). Medicinal benefits of green tea: Part I. Review of noncancer health benefits. *Journal of Alternative & Complementary Medicine* 11(3): 521-528.
- Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S. & Yan, J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186(2): 713-724.
- Crossa, J., Pérez, P., de los Campos, G., Mahuku, G., Dreisigacker, S. & Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. *Journal of Crop Improvement* 25(3): 239-261.
- Cuendet, M., Potterat, O. & Hostettmann, K. (2001). Flavonoids and phenylpropanoid derivatives from *Campanula barbata*. *Phytochemistry* 56(6): 631-636.
- Das, S. K., Sabhapondit, S., Ahmed, G. & Das, S. (2013). Biochemical evaluation of triploid progenies of diploid× tetraploid breeding populations of *Camellia* for genotypes rich in catechin and caffeine. *Biochemical Genetics* 51(5-6): 358-376.
- Deborde, C., Moing, A., Roch, L., Jacob, D., Rolin, D. & Giraudeau, P. (2017). Plant metabolism as studied by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* 102: 61-97.
- Del Rio, D., Rodriguez-Mateos, A., Spencer, J. P., Tognolini, M., Borges, G. & Crozier, A. (2013). Dietary (poly) phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases. *Antioxidants & Redox Signaling* 18(14): 1818-1892.
- Deng, W.-W., Ogita, S. & Ashihara, H. (2010). Distribution and biosynthesis of theanine in Theaceae plants. *Plant Physiology and Biochemistry* 48(1): 70-72.
- Eghbalnia, H. R., Romero, P. R., Westler, W. M., Baskaran, K., Ulrich, E. L. & Markley, J. L. (2017). Increasing rigor in NMR-based metabolomics through validated and open source tools. *Current Opinion in Biotechnology* 43: 56-61.
- Elbehri, A., Azapagic, A., Cheserek, B., Raes, D., Kiprono, P. & Ambasa, C. (2015). Kenya's tea sector under climate change: an impact assessment and formulation of a climate smart strategy. FAO report. FAO, Rome, Italy.

- Elberse, I., Vanhala, T., Turin, J., Stam, P., Van Damme, J. & Van Tienderen, P. (2004). Quantitative trait loci affecting growth-related traits in wild barley (*Hordeum spontaneum*) grown under different levels of nutrient supply. *Heredity* 93(1): 22.
- Ellis, R. & Nyirenda, H. (1995). A successful plant improvement programme on tea (*Camellia sinensis*). *Experimental Agriculture* 31(3): 307-323.
- FAO (Food and Agricultural Organisation). (2018). Retrieved from <http://www.fao.org>.
- Fraser, K., Harrison, S. J., Lane, G. A., Otter, D. E., Hemar, Y., Quek, S.-Y. & Rasmussen, S. (2012). Non-targeted analysis of tea by hydrophilic interaction liquid chromatography and high resolution mass spectrometry. *Food Chemistry* 134(3): 1616-1623.
- Frei, B. & Higdon, J. V. (2003). Antioxidant activity of tea polyphenols in vivo: evidence from animal studies. *The Journal of Nutrition* 133(10): 3275S-3284S.
- Gramza, A., Korczak, J. & Amarowicz, R. (2005). Tea polyphenols-their antioxidant properties and biological activity-a review. *Polish Journal of Food and Nutrition Sciences* 14(3): 219.
- Grandillo, S. & Cammareri, M. (2016). Molecular mapping of quantitative trait loci in tomato. *In The Tomato Genome* 39-73: Springer.
- Green, M. (1966). Clonal selection in seedling stump nurseries. *Tea in East Africa* 6(4): 11-12.
- Green, M. (1971). An evaluation of some criteria used in selecting large-yielding tea clones. *The Journal of Agricultural Science* 76(1): 143-156.
- Griffiths, W. J., Koal, T., Wang, Y., Kohl, M., Enot, D. P. & Deigner, H. P. (2010). Targeted metabolomics for biomarker discovery. *Angewandte Chemie International Edition* 49(32): 5426-5445.
- Grzebelus, D. (2015). Diversity Arrays Technology (DArT) Markers for Genetic Diversity. *In Genetic Diversity and Erosion in Plants* 295-309: Springer.
- Hackett, C. A., Wachira, F. N., Paul, S., Powell, W. & Waugh, R. (2000). Construction of a genetic linkage map for *Camellia sinensis* (tea). *Heredity* 85(4): 346.
- Hagel, J. M. & Facchini, P. J. (2008). Plant metabolomics: analytical platforms and integration with functional genomics. *Phytochemistry Reviews* 7(3): 479-497.
- Halket, J. M., Waterman, D., Przyborowska, A. M., Patel, R. K., Fraser, P. D. & Bramley, P. M. (2005). Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *Journal of Experimental Botany* 56(410): 219-243.

- Hamanishi, E. T., Barchet, G. L., Dauwe, R., Mansfield, S. D. & Campbell, M. M. (2015). Poplar trees reconfigure the transcriptome and metabolome in response to drought in a genotype-and time-of-day-dependent manner. *BMC Genomics* 16(1): 329.
- Haskell, C. F., Kennedy, D. O., Milne, A. L., Wesnes, K. A. & Scholey, A. B. (2008). The effects of L-theanine, caffeine and their combination on cognition and mood. *Biological Psychology* 77(2): 113-122.
- Hazra, A., Dasgupta, N., Sengupta, C. & Das, S. (2018). Next generation crop improvement program: Progress and prospect in tea (*Camellia sinensis* (L.) O. Kuntze). *Annals of Agrarian Science* 16(2): 128-135.
- Hernández, F., Sancho, J., Ibáñez, M., Abad, E., Portolés, T. & Mattioli, L. (2012). Current use of high-resolution mass spectrometry in the environmental sciences. *Analytical and Bioanalytical Chemistry* 403(5): 1251-1264.
- Hilton, P. & Ellis, R. (1972). Estimation of the market value of Central African tea by theaflavin analysis. *Journal of the Science of Food and Agriculture* 23(2): 227-232.
- Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician* 52(2): 181-184.
- Hopfgartner, G., Varesio, E., Tschäppät, V., Grivet, C., Bourgoigne, E. & Leuthold, L. A. (2004). Triple quadrupole linear ion trap mass spectrometer for the analysis of small molecules and macromolecules. *Journal of Mass Spectrometry* 39(8): 845-855.
- Jaccoud, D., Peng, K., Feinstein, D. & Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research* 29(4): e25-e25.
- Jiang, C., Ma, J.-Q., Apostolides, Z. & Chen, L. (2019). Metabolomics for a millenniums-old crop-tea plant (*Camellia sinensis*). *Journal of Agricultural and Food Chemistry* 67(3): 6445-6457.
- Jiang, X., Liu, Y., Wu, Y., Tan, H., Meng, F., sheng Wang, Y., Li, M., Zhao, L., Liu, L. & Qian, Y. (2015). Analysis of accumulation patterns and preliminary study on the condensation mechanism of proanthocyanidins in the tea plant [*Camellia sinensis*]. *Scientific Reports* 5: 8742.
- JMP®, V. SAS Institute Inc., Cary, NC, 1989–2007.
- Kamunya, S. & Wachira, F. (2006). Two new clones (TRFK 371/3 and TRFK 430/90) released for commercial use. *Tea* 27(1/2): 3-14.

- Kamunya, S., Wachira, F., Pathak, R., Korir, R., Sharma, V., Kumar, R., Bhardwaj, P., Chalo, R., Ahuja, P. & Sharma, R. (2010). Genomic mapping and testing for quantitative trait loci in tea (*Camellia sinensis* (L.) O. Kuntze). *Tree Genetics & Genomes* 6(6): 915-929.
- Kamunya, S., Wachira, F., Pathak, R., Muoki, R., Wanyoko, J., Ronno, W. & Sharma, R. (2009). Quantitative genetic parameters in tea (*Camellia sinensis* (L.) O. Kuntze): I. combining abilities for yield, drought tolerance and quality traits. *African Journal of Plant Science* 3(5): 093-101.
- Karori, S., Wachira, F., Ngure, R. & Mireji, P. (2014). Polyphenolic composition and antioxidant activity of Kenyan tea cultivars. *Journal of Pharmacognosy and Phytochemistry* 3(4): 105-116.
- Kato, M., Mizuno, K., Crozier, A., Fujimura, T. & Ashihara, H. (2000). Plant biotechnology: Caffeine synthase gene from tea leaves. *Nature* 406(6799): 956.
- Kaundun, S. & Matsumoto, S. (2003a). Development of CAPS markers based on three key genes of the phenylpropanoid pathway in tea, *Camellia sinensis* (L.) O. Kuntze, and differentiation between assamica and sinensis varieties. *Theoretical and Applied Genetics* 106(3): 375-383.
- Kaundun, S. S. & Matsumoto, S. (2003b). Identification of Processed Japanese Green Tea Based on Polymorphisms Generated by STS–RFLP Analysis. *Journal of Agricultural and Food Chemistry* 51(7): 1765-1770.
- Kaundun, S. S. & Park, Y.-G. (2002). Genetic structure of six Korean tea populations as revealed by RAPD-PCR markers. *Crop Science* 42(2): 594-601.
- Kerio, L., Wachira, F., Wanyoko, J. & Rotich, M. (2013). Total polyphenols, catechin profiles and antioxidant activity of tea products from purple leaf coloured tea cultivars. *Food Chemistry* 136(3-4): 1405-1413.
- Keum, Y. S., Jeong, W. S. & Kong, A. (2005). Chemopreventive functions of isothiocyanates. *Drug News Perspect* 18(7): 445-451.
- Khan, N. & Mukhtar, H. (2007). Tea polyphenols for health promotion. *Life Sciences* 81(7): 519-533.
- Kimura, K., Ozeki, M., Juneja, L. R. & Ohira, H. (2007). L-Theanine reduces psychological and physiological stress responses. *Biological Psychology* 74(1): 39-45.
- Kind, T. & Fiehn, O. (2010). Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews* 2(1-4): 23-60.

- Kito, M., Kokura, H., Izaki, J. & Sasaoka, K. (1968). Theanine, a precursor of the phloroglucinol nucleus of catechins in tea plants. *Phytochemistry* 7(4): 599-603.
- Koech, R. K., Malebe, P. M., Nyarukowa, C., Mose, R., Kamunya, S. M. & Apostolides, Z. (2018). Identification of novel QTL for black tea quality traits and drought tolerance in tea plants (*Camellia sinensis*). *Tree Genetics & Genomes* 14(1): 9.
- Kumar, A., Saini, G. A. U. T. A. M., Nair, A., & Sharma, R. (2012). UPLC: a preeminent technique in pharmaceutical analysis. *Acta Pol Pharm* 69(3): 371-380.
- Kumar, R. S. S., Muraleedharan, N. N., Murugesan, S., Kottur, G., Anand, M. P. & Nishadh, A. (2011). Biochemical quality characteristics of CTC black teas of south India and their relation to organoleptic evaluation. *Food Chemistry* 129(1): 117-124.
- Lambert, J. D., Hong, J., Yang, G.-y., Liao, J. & Yang, C. S. (2005). Inhibition of carcinogenesis by polyphenols: evidence from laboratory investigations. *The American Journal of Clinical Nutrition* 81(1): 284S-291S.
- Larsen, B., Gardner, K., Pedersen, C., Ørgaard, M., Migicovsky, Z., Myles, S. & Toldam-Andersen, T. B. (2018). Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PloS one* 13(8): e0201889.
- Lattanzio, V., Lattanzio, V. M. & Cardinali, A. (2006). Role of phenolics in the resistance mechanisms of plants against fungal pathogens and insects. *Phytochemistry: Advances in Research* 661: 23-67.
- Le Gall, G., Colquhoun, I. J. & Defernez, M. (2004). Metabolite profiling using ¹H NMR spectroscopy for quality assessment of green tea, *Camellia sinensis* (L.). *Journal of Agricultural and Food Chemistry* 52(4): 692-700.
- Li, J.-B., Hashimoto, F., Shimizu, K. & Sakata, Y. (2013). Chemical taxonomy of red-flowered wild *Camellia* species based on floral anthocyanins. *Phytochemistry* 85: 99-106.
- Li, W., Xiang, F., Zhong, M., Zhou, L., Liu, H., Li, S. & Wang, X. (2017). Transcriptome and metabolite analysis identifies nitrogen utilization genes in tea plant (*Camellia sinensis*). *Scientific Reports* 7(1): 1693.
- Lien, E. J., Ren, S., Bui, H.-H. & Wang, R. (1999). Quantitative structure-activity relationship analysis of phenolic antioxidants. *Free Radical Biology and Medicine* 26(3): 285-294.
- Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. & Fernie, A. R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols* 1(1): 387.

- Lisman, J. E., Coyle, J. T., Green, R. W., Javitt, D. C., Benes, F. M., Heckers, S. & Grace, A. A. (2008). Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. *Trends in Neurosciences* 31(5): 234-242.
- Ma, J. Q., Yao, M. Z., Ma, C. L., Wang, X. C., Jin, J. Q., Wang, X. M. & Chen, L. (2014). Construction of a SSR-based genetic map and identification of QTLs for catechins content in tea plant (*Camellia sinensis*). *PloS one* 9(3): e93131.
- Magoma, G., Wachira, F., Obanda, M., Imbuga, M. & Agong, S. (2000). The use of catechins as biochemical markers in diversity studies of tea (*Camellia sinensis*). *Genetic Resources and Crop Evolution* 47(2): 107-114.
- Masuda, S., Maeda-Yamamoto, M., Usui, S. & Fujisawa, T. (2014). 'Benifuuki'green tea containing o-methylated catechin reduces symptoms of Japanese cedar pollinosis: a randomized, double-blind, placebo-controlled trial. *Allergology International* 63(2): 211-217.
- McDowell, I., Taylor, S. & Gay, C. (1995). The phenolic pigment composition of black tea liquors—part I: Predicting quality. *Journal of the Science of Food and Agriculture* 69(4): 467-474.
- McLafferty, F. & Turecek, F. (1997). Interpretation of mass spectra, 1993. Mill Valley: University Science Books.
- Meeting, J. F. W. E. C. o. F. A. & Organization, W. H. (2010). Evaluation of Certain Food Additives: Seventy-first Report of the Joint FAO/WHO Expert Committee on Food Additives. World Health Organization.
- Miles, C. & Wayne, M. (2008). Quantitative trait locus (QTL) analysis.
- Mohanpuria, P., Kumar, V. & Yadav, S. K. (2010). Tea caffeine: metabolism, functions, and reduction strategies. *Food Science and Biotechnology* 19(2): 275-287.
- Molnár-Perl, I. (1999). Simultaneous quantitation of acids and sugars by chromatography: gas or high-performance liquid chromatography? *Journal of Chromatography A* 845(1): 181-195.
- Mondal, T. K., Bhattacharya, A., Laxmikumaran, M. & Ahuja, P. S. (2004). Recent advances of tea (*Camellia sinensis*) biotechnology. *Plant Cell, Tissue and Organ Culture* 76(3): 195-254.
- Munson, M. & Field, F.-H. (1966). Chemical ionization mass spectrometry. I. General introduction. *Journal of the American Chemical Society* 88(12): 2621-2630.

- Muoki, R., Wachira, F., Pathak, R. & Kamunya, S. (2007). Assessment of the mating system of *Camellia sinensis* in biclonal seed orchards based on PCR markers. *The Journal of Horticultural Science and Biotechnology* 82(5): 733-738.
- Nance, C. L. & Shearer, W. T. (2003). Is green tea good for HIV-1 infection? *Journal of Allergy and Clinical Immunology* 112(5): 851-853.
- Ni, S., Yao, M., Chen, L., Zhao, L. & Wang, X. (2008). Germplasm and breeding research of tea plant based on DNA marker approaches. *Frontiers of Agriculture in China* 2(2): 200.
- Niessen, W. M. (2001). Current practice of gas chromatography-mass spectrometry. CRC Press.
- Nyarukowa, C., Koech, R., Loots, T. & Apostolides, Z. (2016). SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. *Journal of Plant Physiology* 198: 39-48.
- Nyirenda, H. (1991). Use of growth measurements and foliar nutrient content as criteria for clonal selection in tea (*Camellia sinensis*). *Experimental Agriculture* 27(1): 47-52.
- Obanda, M., Owuor, P. O. & Mang'oka, R. (2001). Changes in the chemical and sensory quality parameters of black tea due to variations of fermentation time and temperature. *Food Chemistry* 75(4): 395-404.
- Ogbaga, C. C., Stepien, P. & Johnson, G. N. (2014). Sorghum (*Sorghum bicolor*) varieties adopt strongly contrasting strategies in response to drought. *Physiologia Plantarum* 152(2): 389-401.
- Orel, G. & Wilson, P. G. (2012). *Camellia cherryana* (Theaceae), a new species from China. *In Annales Botanici Fennici* 49: 248-255.
- Owuor, P., Obanda, M., Apostolides, Z., Wright, L., Nyirenda, H. & Mphangwe, N. (2006). The relationship between the chemical plain black tea quality parameters and black tea colour, brightness and sensory evaluation. *Food Chemistry* 97: 644-653.
- Owuor, P. O., Wachira, F. N. & Ng'etich, W. K. (2010). Influence of region of production on relative clonal plain tea quality parameters in Kenya. *Food Chemistry* 119(3): 1168-1174.
- Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research* 96(1): 3-14.

- Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H. & Sorrells, M. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome* 5(3): 103-113.
- Proestos, C., Sereli, D. & Komaitis, M. (2006). Determination of phenolic compounds in aromatic plants by RP-HPLC and GC-MS. *Food Chemistry* 95(1): 44-52.
- Resende Jr, M., Munoz, P., Acosta, J., Peter, G., Davis, J., Grattapaglia, D., Resende, M. & Kirst, M. (2012). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytologist* 193(3): 617-624.
- Richards, A. (1966). The breeding, selection and propagation of tea.
- Rogers, P. J., Smith, J. E., Heatherley, S. V. & Pleydell-Pearce, C. (2008). Time for tea: mood, blood pressure and cognitive performance effects of caffeine and theanine administered alone and together. *Psychopharmacology* 195(4): 569-577.
- Rolin, D., Deborde, C., Maucourt, M., Cabasson, C., Fauvelle, F., Jacob, D., Canlet, C. & Moing, A. (2013). High-resolution 1H-NMR spectroscopy and beyond to explore plant metabolome. *In Advances in Botanical Research* 67: 1-66.
- Ruan, C. J. & Teixeira da Silva, J. A. (2011). Metabolomics: creating new potentials for unraveling the mechanisms in response to salt and drought stress and for the biotechnological improvement of xero-halophytes. *Critical Reviews in Biotechnology* 31(2): 153-169.
- Sadzuka, Y., Inoue, C., Hirooka, S., Sugiyama, T., Umegaki, K. & Sonobe, T. (2005). Effects of theanine on alcohol metabolism and hepatic toxicity. *Biological and Pharmaceutical Bulletin* 28(9): 1702-1706.
- Sagawa, C., Cristofani-Yaly, M., Novelli, V., Bastianel, M. & Machado, M. (2018). Assessing genetic diversity of Citrus by DArT_seq™ genotyping. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology* 152(4): 593-598.
- Sang, S., Lambert, J. D., Ho, C. T. & Yang, C. S. (2011). The chemistry and biotransformation of tea constituents. *Pharmacological Research* 64(2): 87-99.
- Schauer, N. & Fernie, A. R. (2006). Plant metabolomics: towards biological function and mechanism. *Trends in Plant Science* 11(10): 508-516.
- Schouten, H. J., van de Weg, W. E., Carling, J., Khan, S. A., McKay, S. J., van Kaauwen, M. P., Wittenberg, A. H., Koehorst-van Putten, H. J., Noordijk, Y. & Gao, Z. (2012).

- Diversity arrays technology (DArT) markers in apple for genetic linkage maps. *Molecular Breeding* 29(3): 645-660.
- Shanmugarajah, V., Kulasegaram, S. & Senanayake, Y. (1991). Nursery plant attributes as criteria for selection of new tea clones. *SLJ Tea Sci* 61: 76-86.
- Sharma, E., Joshi, R. & Gulati, A. (2018). l-Theanine: An astounding sui generis integrant in tea. *Food Chemistry* 242: 601-610.
- Shen, Q., Yu, C., Guo, Y., Bian, Z., Zhu, N., Yang, L., Chen, Y., Luo, G., Li, J. & Qin, Y. (2018). Habitual tea consumption and risk of fracture in 0.5 million Chinese adults: a prospective cohort study. *Nutrients* 10(11): 1633.
- Siegel, S. & Castellan, N. J. (1956). Nonparametric statistics for the behavioral sciences. McGraw-hill New York.
- Singh, A., Rai, V., Chand, R., Singh, R. & Singh, M. (2013a). Genetic diversity studies and identification of SSR markers associated with Fusarium wilt (*Fusarium udum*) resistance in cultivated pigeonpea (*Cajanus cajan*). *Journal of Genetics* 92(2): 273-280.
- Singh, H. P., Ravindranath, S. & Singh, C. (1999). Analysis of tea shoot catechins: Spectrophotometric quantitation and selective visualization on two-dimensional paper chromatograms using diazotized sulfanilamide. *Journal of Agricultural and Food Chemistry* 47(3): 1041-1045.
- Singh, S., Kumar, A., Karthigeyan, S. & Ahuja, P. (2013b). GENETIC IMPROVEMENT OF TEA. *Science Of Tea Technology*: 22.
- Slama, I., Abdelly, C., Bouchereau, A., Flowers, T. & Savouré, A. (2015). Diversity, distribution and roles of osmoprotective compounds accumulated in halophytes under abiotic stress. *Annals of Botany*: mcu239.
- Steane, D. A., Nicolle, D., Sansaloni, C. P., Petrolì, C. D., Carling, J., Kilian, A., Myburg, A. A., Grattapaglia, D. & Vaillancourt, R. E. (2011). Population genetic analysis and phylogeny reconstruction in Eucalyptus (*Myrtaceae*) using high-throughput, genome-wide genotyping. *Molecular Phylogenetics and Evolution* 59(1): 206-224.
- Steinmann, D. & Ganzera, M. (2011). Recent advances on HPLC/MS in medicinal plant analysis. *Journal of Pharmaceutical and Biomedical Analysis* 55(4): 744-757.
- Suganuma, M., Saha, A. & Fujiki, H. (2011). New cancer treatment strategy using combination of green tea catechins and anticancer drugs. *Cancer Science* 102(2): 317-323.

- Sumpio, B. E., Cordova, A. C., Berke-Schlessel, D. W., Qin, F. & Chen, Q. H. (2006). Green tea, the “Asian paradox,” and cardiovascular disease. *Journal of the American College of Surgeons* 202(5): 813-825.
- Taylor, S., Baker, D., Owuor, P., Orchard, J., Othieno, C. & Gay, C. (1992). A model for predicting black tea quality from the carotenoid and chlorophyll composition of fresh green tea leaf. *Journal of the Science of Food and Agriculture* 58(2): 185-191.
- Terashima, T., Takido, J. & Yokogoshi, H. (1999). Time-dependent changes of amino acids in the serum, liver, brain and urine of rats administered with theanine. *Bioscience, Biotechnology, and Biochemistry* 63(4): 615-618.
- Theodoridis, G. A., Gika, H. G., Want, E. J. & Wilson, I. D. (2012). Liquid chromatography–mass spectrometry based global metabolite profiling: a review. *Analytica Chimica Acta* 711: 7-16.
- Thiagarajan-Rosenkranz, P., Draney, A. W. & Lorieu, J. L. (2017). Hybrid NMR: A Union of Solution-and Solid-State NMR. *Journal of the American Chemical Society* 139(13): 4715-4723.
- Thielecke, F. & Boschmann, M. (2009). The potential role of green tea catechins in the prevention of the metabolic syndrome—a review. *Phytochemistry* 70(1): 11-24.
- Thomas, E. H. & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education* 45(3): 251-269.
- Tikunov, Y., Lommen, A., de Vos, C. R., Verhoeven, H. A., Bino, R. J., Hall, R. D. & Bovy, A. G. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology* 139(3): 1125-1137.
- Tong, X., Taylor, A. W., Giles, L., Wittert, G. A. & Shi, Z. (2014). Tea consumption is inversely related to 5-year blood pressure change among adults in Jiangsu, China: a cross-sectional study. *Nutrition Journal* 13(1): 98.
- Uchiyama, S., Taniguchi, Y., Saka, A., Yoshida, A. & Yajima, H. (2011). Prevention of diet-induced obesity by dietary black tea polyphenols extract in vitro and in vivo. *Nutrition* 27(3): 287-292.
- Unger, K. K., Skudas, R. & Schulte, M. M. (2008). Particle packed columns and monolithic columns in high-performance liquid chromatography-comparison and critical appraisal. *Journal of Chromatography A* 1184(1): 393-415.

- Urano, K., Maruyama, K., Ogata, Y., Morishita, Y., Takeda, M., Sakurai, N., Suzuki, H., Saito, K., Shibata, D. & Kobayashi, M. (2009). Characterization of the ABA regulated global responses to dehydration in Arabidopsis by metabolomics. *The Plant Journal* 57(6): 1065-1078.
- Vallarino, J. G., Pott, D. M., Cruz-Rus, E., Miranda, L., Medina-Minguez, J. J., Valpuesta, V., Fernie, A. R., Sánchez-Sevilla, J. F., Osorio, S. & Amaya, I. (2019). Identification of quantitative trait loci and candidate genes for primary metabolite content in strawberry fruit. *Horticulture Research* 6(1): 4.
- Visser, T. & Kehil, F. (1958). Selection and vegetative propagation of tea.
- Wachira, F. (2001). 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: 12-14.
- Wachira, F. N. & Kamunya, S. (2005). Kenyan teas are rich in antioxidants. *Tea* 26(2): 81-89.
- Waldhauser, S. S. M. & Baumann, T. W. (1996). Compartmentation of caffeine and related purine alkaloids depends exclusively on the physical chemistry of their vacuolar complex formation with chlorogenic acids. *Phytochemistry* 42(4): 985-996.
- Wambulwa, M. C., Meegahakumbura, M. K., Kamunya, S., Muchugi, A., Möller, M., Liu, J., Xu, J.-C., Li, D.-Z. & Gao, L.-M. (2017). Multiple origins and a narrow genepool characterise the African tea germplasm: concordant patterns revealed by nuclear and plastid DNA markers. *Scientific Reports* 7(1): 4053.
- Ward, J. L., Baker, J. M. & Beale, M. H. (2007). Recent applications of NMR spectroscopy in plant metabolomics. *FEBS Journal* 274(5): 1126-1131.
- Waycott, W., Fort, S., Ryder, E. & Michelmore, R. W. (1999). Mapping morphological genes relative to molecular markers in lettuce (*Lactuca sativa L.*). *Heredity* 82(3): 245.
- Weckwerth, W., Wenzel, K. & Fiehn, O. (2004). Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 4(1): 78-83.
- Wen, B., Ma, L., Nelson, S. D. & Zhu, M. (2008). High-throughput screening and characterization of reactive metabolites using polarity switching of hybrid triple quadrupole linear ion trap mass spectrometry. *Analytical Chemistry* 80(5): 1788-1799.
- Wight, W. (1956). Commercial selection and breeding of tea in India. *World Crops* 8: 263-268.

- Willson, K. C. & Clifford, M. N. (2012). Tea: Cultivation to consumption. Springer Science & Business Media.
- Wittenberg, A. H., Van Der Lee, T., Cayla, C., Kilian, A., Visser, R. G. & Schouten, H. J. (2005). Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. *Molecular Genetics and Genomics* 274(1): 30-39.
- Wilson, I. D., Nicholson, J. K., Castro-Perez, J., Granger, J. H., Johnson, K. A., Smith, B. W., & Plumb, R. S. (2005). High resolution “ultra performance” liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *Journal of Proteome Research* 4(2): 591-598.
- Worland, A., Gale, M. & Law, C. (1987). Wheat genetics. *In Wheat breeding* 129-171: Springer.
- Wright, L. P., Mphangwe, N. I. K., Nyirenda, H. E. & Apostolides, Z. (2002). 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* 82(5): 517-525.
- Xia, E. H., Zhang, H. B., Sheng, J., Li, K., Zhang, Q. J., Kim, C., Zhang, Y., Liu, Y., Zhu, T. & Li, W. (2017). The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Molecular Plant* 10(6): 866-877.
- Yilmaz, Y. (2006). Novel uses of catechins in foods. *Trends in Food Science & Technology* 17(2): 64-71.
- Yokogoshi, H. & Kobayashi, M. (1998). Hypotensive effect of γ -glutamylmethanamide in spontaneously hypertensive rats. *Life Sciences* 62(12): 1065-1068.
- Zandalinas, S. I., Vives-Peris, V., Gómez-Cadenas, A. & Arbona, V. (2012). A fast and precise method to identify indolic glucosinolates and camalexin in plants by combining mass spectrometric and biological information. *Journal of Agricultural and Food Chemistry* 60(35): 8648-8658.
- Zhao, Y. Y. (2013). Metabolomics in chronic kidney disease. *Clinica Chimica Acta* 422: 59-69.
- Zhang, A., Sun, H., Wang, P., Han, Y. & Wang, X. (2012). Modern analytical techniques in metabolomics analysis. *Analyst* 137(2): 293-300.
- Zhang, Q., Liu, M. & Ruan, J. (2017). Metabolomics analysis reveals the metabolic and functional roles of flavonoids in light-sensitive tea leaves. *BMC Plant Biology* 17(1): 64.

Zhou, B., Xiao, J. F., Tuli, L. & Resson, H. W. (2012). LC-MS-based metabolomics.
Molecular BioSystems 8(2): 470-481.

CHAPTER 2

PRIORITISING THE REPLANTING SCHEDULE OF SEEDLING TEA FIELDS ON TEA ESTATES FOR DROUGHT SUSCEPTIBILITY MEASURED BY THE SWAPDT METHOD IN THE ABSENCE OF HISTORICAL IN-FILLING RECORDS

ABSTRACT

Due to the unpredictable natural droughts that occur, causing tea farmers significant losses in tea estates, a two-day method for distinguishing between drought tolerant (DT) and drought susceptible (DS) *Camellia sinensis* cultivars was developed. This work was based on known cultivars developed at the Tea Research Institute in Kenya and the Tea Research Foundation for Central Africa in Malawi. This paper contains an in-depth description of the application of the Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) method on four 60-year old, *C. sinensis* seedling fields in Kenya. The in-filling history of the four fields due to drought-related deaths was obtained from historical records. The SWAPDT method scores correlated well with the historical records. It has been indicated, from the results obtained in this study, that a sample size of 20 tea trees is sufficient to accurately determine the drought susceptibility of a large tea field of approximately 5 - 20 hectares, containing 50 000 - 200 000 tea trees, where the difference between the fields' mean values, as measured by the SWAPDT method, is approximately 10%.

2.1 INTRODUCTION

Tea made from the leaves of *C. sinensis*, as green or black tea, has been drunk as a mild stimulant due to the caffeine content, since time immemorial (Ellis and Nyirenda, 1995). Tea consumption has increased in recent years, due to the health-promoting effects associated with its high polyphenol content (Preedy, 2012). *C. sinensis* is cultivated in over 52 countries around the world. Global world trade is approximately 78% by value in the form of black, 20% as green and 2% as oolong tea (Nyarukowa *et al.*, 2016). It is an important cash crop for countries such as India and China; in Africa alone, several countries produce tea, namely in Kenya, which is currently ranked third behind Sri Lanka and India with regards to annual production and export of black tea (Chang, 2015), Malawi, Uganda, Tanzania, Zimbabwe, Rwanda, South Africa, Burundi and Mauritius. *C. sinensis* tea estates need to be replanted every 20 - 90 years to maintain high yields. Tea estates are planted in sample blocks of about 5 - 20 hectares, at 10 000 trees per hectare, with seeds from the same batch. Most tea estates in Africa are planted with tea seeds procured from tea-baries (orchards) in India or Sri Lanka. The seed selection criteria employed focused on yield and neglected to consider drought tolerance (Murakami *et al.*, 1999). During severe natural droughts, some of the trees (5 – 15%) die and are replaced with new trees. Tea planters refer to this process as “in-filling”. Most estates keep good records of the in-filling, and hence good and poor fields are easily identified. However, sometimes these records are missing, and a new method is required to determine the drought tolerance of a tea field that might be 20 - 90 years old (Willson and Clifford, 2012).

Tea producers demand new cultivars which are DT, to reduce crop losses. In the coffee industry, farmers are faced with the problem of dealing with coffee rust. How they deal with this is by assessing and analysing the risk of an epidemic by considering the region's characteristics such as climate, soils, crop management patterns, namely shade management, etc. (DaMatta, 2004). This assessment approach has been adopted from studies conducted in West Africa on groundnuts (Avelino *et al.*, 2004) as well as on work conducted by (Savary *et al.*, 2000) on tropical Asian rice. As a result of this, tea farmers are also looking for an inexpensive yet effective method of determining which sample blocks of tea have a high percentage of DS plants so that these sample blocks may be prioritised for replanting. The samples in this study were collected from the James Finlay's estate in Kericho, Kenya which together with surrounding estates (Figure 2.1) produces 23 million kilograms of tea annually.

This part of Kenya enjoys deep rich loam soils, which are high in organic content and combined with the perfect climate and environment are ideal for high yields of good quality tea (<http://www.finlays.net>). In Nyarukowa *et al.*, (2016), a novel logistics probability formula was developed, which can be used to calculate a new cultivar’s probability to be DT after employing the SWAPDT method. The aim of this study was to determine how many tea trees are needed per field to obtain a representative sample of the tea field so that tea fields can be prioritised for replanting.

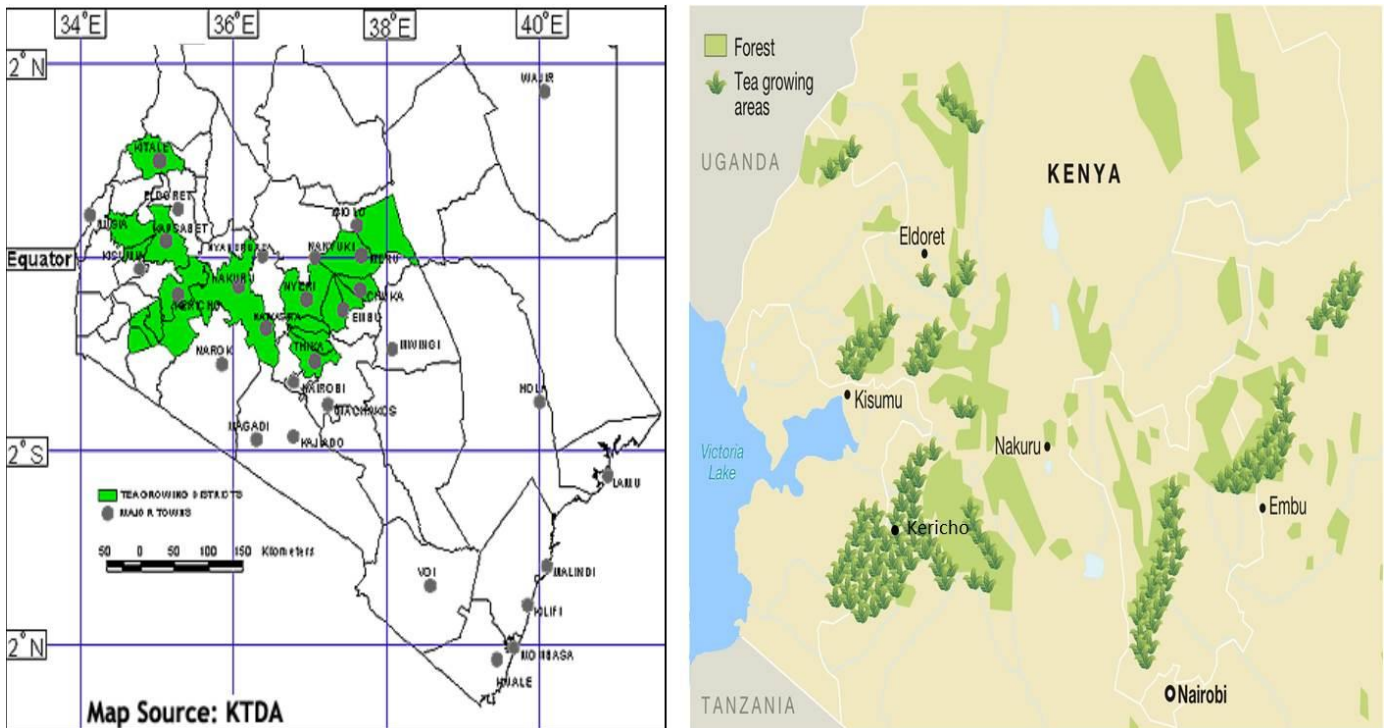


Figure 2.1: Tea growing areas in Kenya (Kenya Tea Development Agency).

<http://blog.dominiontea.com/2014/03/27/kenyan-tea-industry/>

2.2 STATISTICAL ANALYSIS

2.2.1 Spatial regression Modelling

Spatial regression models have been widely employed in biological science disciplines. Central to spatial analysis and quantitative geography, has been the study of methods that specify and fit spatial regression models (Bivand *et al.*, 2013). Spatial autocorrelation is the similarity or dissimilarity of two values of a feature spatially near one another. As such, a positive spatial autocorrelation value signifies that the values of a particular feature cluster spatially while a negative spatial autocorrelation signifies that the locations of said feature are encompassed by neighbours with varying values (Anselin, 1990). According to (Cressie,

1993), spatial data analysis may be classified as point data analysis, lattice data analysis, or geo-statistics; each consists of its own unique objectives and approaches. Point data analysis focuses on ascertaining spatial patterns involved in e.g. cluster formation as aberrations from complete randomness. Lattice data analysis on the other hand looks at the spatial pattern of a particular feature on a regular or irregular spatial lattice perceived at the grid points, the purpose of which is to calculate the spatial pattern by using a pre-determined “neighbourhood structure” to observe the associations between the feature of interest and those expository variables whilst factoring for spatial effects. Lastly, geo-statistical data refers to spatial data collected at points continuous in space. Geo-statistics shares similarities with lattice data analysis, differing only in that geo-statistics has the further objective of predicting values of the feature of interest at locations not yet sampled (Chi *et al.*, 2008). Furthermore, geo-statistics is distinguished from lattice data analysis in that geo-statistics employs distance based functions instead of “neighbourhood structures” (as is the case with lattice data analysis) to denote spatial autocorrelation (Bailey and Gatrell, 1995).

2.2.2 Contour Maps: construction and application

Contour mapping can be used for in-network aggregation schemes, in the presentation of holistic networking of temporal and spatial domains while functioning as a diagnostic tool for the detection of faulty sensors. Contours may be representative of several occurrences e.g. temperature, altitude, etc. Furthermore, altitude and temperature contours may overlap within a single map. Though the contour maps concept is both simple and well documented, the construction of these maps is however quite exigent, particularly in a sensor network environment. Spatial suppression serves to take advantage of the correlation between neighbouring sensors whenever a particular occurrence takes place. When there is e.g. an unexpected temperature change, each sensor (μ) in the shared vicinity of this occurrence determines whether or not it should transmit its reading to the data collection centre, the sink. In instances where the magnitude difference between μ and the reading from another sensor is less than a pre-determined threshold value i.e. β , μ suppresses its report; sensor readings are not reported if the observed values changes over time are not significant. However, to permit outlier detection, sensors detecting such disparities, transmit this data to the sink, even if they ordinarily would not. This then aids the sink in accurately detecting outliers. The sinks then make use of this data of the terrain to precisely recreate contour maps taking into

account the sensor field as well as the positioning of these sensors, ensuring accurate detection and interpolation of outliers (Meng *et al.*, 2006).

2.2.3 Moran I Test Statistics

The most popular spatial correlation test is based on the Moran I test statistic. This test is in a normalised quadratic form, with regards to the variables being tested for spatial correlation (Kelejian and Prucha, 2010). It measures the degree of linear association between the feature of interest (y) at a specific position and the average of the feature at its neighbouring positions (Wy); it can be construed as the regression slope of (y) on (Wy) (Pacheco and Tyrrell, 2002). The Moran I test statistic functions to test that the spatial autocorrelation of a variable in the null hypothesis is zero. Upon the rejection of the null hypothesis, the variable will be deemed to be spatially auto-correlated (Ord and Getis, 1995). The Monte-Carlo test can be employed to solve spatial distribution problems involving spatial point pattern, pattern similarity, space-time interaction and scales of pattern. The Monte Carlo test can also be employed when dealing with a null hypothesis, H_0 , and a corresponding dataset, in which the value u_1 of a selected test statistic, u , is ranked amongst a set of analogous values generated by randomly sampling from the null distribution of u . The rank of the test statistic u_1 , when the u distribution is continuous amongst the set of values $\{u_i: i = 1, \dots, m\}$ enables the determination of the exact significance level for the test (Besag and Diggle, 1977). The ordinary linear regression model is amongst the most useful statistical methods employed in spatial analysis to identify relationships between variables is (Mei *et al.*, 2004). In this technique, the dependent variable, y , is modelled as a linear function of a set of independent variables x_1, x_2, \dots, x_p . Based on n observations ($y_i; x_{i1}, x_{i2}, \dots, x_{ip}$), ($i = 1, 2, \dots, n$), from a study region, the model can be expressed as:

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ik} + \epsilon_i, \quad (1)$$

where $\beta_0, \beta_1, \dots, \beta_p$ are parameters, and $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are error terms assumed to be normally distributed, random, independent variables (Mei *et al.*, 2004).

Determining a study's sample size requires a compromise, which balances power, economy, and timeliness (Dupont and Plummer Jr, 1990). Researchers must define the sample size, power, and hypotheses for their study and to effectively do this, they require software programs that can calculate the missing parameter when given any two of the preceding three parameters. These programs compute the sample size required to identify, using a specified

power, e.g. the difference in efficacy of a particular treatment, the power this difference is detectable with given sample size, and the detectable difference with particular power and sample size. The link between the power and its corresponding sample size is best seen by plotting the power curve as a function of the parameter of interest (Kelsey *et al.*, 1986).

The Student's t-test was used to calculate the probability (p) that the two samples belong to the same population. When $p < 0.05$, there is a 95% certainty that the two samples belong to different populations. The Standard Error of the Mean (SEM) is an inferential statistic that can be used to draw error bars on histograms to visually estimate the p-value. When the sample size (n) > 10 , and the gap between the SEM error bars $> SEM1 + SEM2$, we can be 99% confident that the samples are from two different populations (Cumming *et al.*, 2007). The MANOVA method may also be used to compare multiple fields with each other, and the three methods are expected to produce similar results (Keselman *et al.*, 1998). Oneway analysis using JMP Pro 13 generates “mean diamonds”, which illustrate both the sample mean and confidence interval. The top and bottom of each diamond represent the $(1 - \alpha) \times 100$ confidence interval for each group. The confidence interval computation assumes that the variances are equal across observations, and as such the height of each diamond is proportional to the reciprocal of the square root of the number of observations within the group. The mean line across the middle of each diamond represents the group mean, while the overlap marks appear as lines above and below the group mean. In instances where groups have equal sample sizes, these overlapping marks indicate that the two group means are not significantly different at the given confidence level. Where the mean in one diamond is between the overlap marks of another diamond, this indicates that these two groups are not significantly different at that confidence level (JMP®).

2.2.4 Mann-Whitney test

When a study, such as this one, possesses two groups originating from normally distributed populations, it must be ascertained whether or not these groups are from the same populations i.e. whether there is a significant difference between the mean values of each group. Where no prior knowledge of the distribution exists, the Mann-Whitney test is useful especially in instances where behavioural effects are being observed (Baldino *et al.*, 1979). In medical research, it is commonly used to contrast the outcomes in patient treatments, in non-randomised groups, where the data is continuously distributed and skewed (Fagerland and Sandvik, 2009). This test is often employed as a *t*-test alternative; it assumes that the data consists of randomised, independent samples from two populations possessing similar shape, as shown in Figure 2.2.

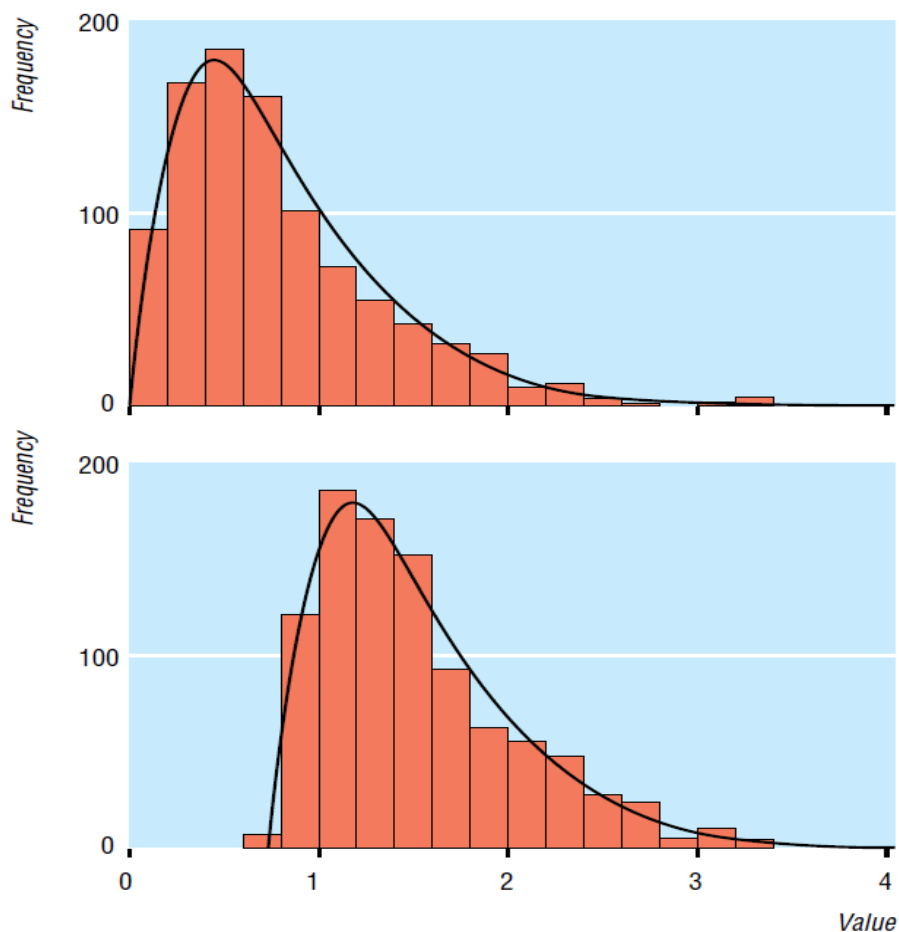


Figure 2.2: Two distributions with different medians but similar shape and spread. The top figure has a distribution skewed 0.75 units to the right, meaning the medians will be different by 0.75 units, whilst both figures maintain identical shapes.

In cases where the sample size is large, the Mann-Whitney test is capable of detecting differences in the sample spread, regardless of the similarity of the medians. The dissimilarities observed in population medians tend to be accompanied by further shape and spread differences; the median differences are however not always the most important. Figure 2.3 elaborates on this.



Figure 2.3: Two distributions with differing median values (0.65 and 1.14 units) and shapes. The distribution showing the larger median also possesses the bigger spread.

2.3 RESEARCH OBJECTIVE

The objective of this study was to prioritise the replanting schedule of seedling tea fields on estates commonly subjected to drought by developing a sampling method and ascertaining drought susceptibility using the SWAPDT method, all of this being done in the absence of historical in-filling records.

2.4 HYPOTHESES

Alternative hypothesis (H_1): There will be a statistically significant difference in the %relative water content (RWC), between the historically “good” and “poor” fields, at the 95% level of confidence.

Null hypothesis (H_0): There will be no statistically significant difference in the %RWC, between the historically “good” and “poor” fields, at the 95% level of confidence.

2.5 MATERIALS AND METHODS

2.5.1 Sample Collection

The field work was conducted on the Kaproret tea estate in Kericho, Kenya, with latitude: 0° 22' 3.86" N and longitude: 35° 16' 59.30" E, during January and February of 2017. Based on historical in-filling records, two good fields, fields 12A and 12B, with fewer in-fillings due-to-deaths-by-drought and two poor fields, 13A and 13B, with higher in-fillings due-to-deaths-by-drought were selected. The good fields were those fields with more drought tolerant plants/ less deaths due to drought, while the poor fields were those fields with more deaths due to drought. These fields were approximately 1200 m apart. The fields were planted from different batches of seeds obtained from Assam, in 1954 and 1956 respectively. They were in the same prune year, receiving the same fertiliser regimen, under rainfed conditions. The longitude and latitude coordinates for field 12A and 12B are 35° 14.75' E, 0° 26' S and for field 13A and 13B are 35° 15.05' E, 0° 26.6' S and at an altitude of 2180 m above mean sea level. These fields were located in regions that have high humidity, fair temperatures and acidic soils. Samples were collected using a “point-intercept within a quadrat method”, in which a 100 m X 100 m quadrat was set up in the middle of each field. The starting point of the sampling was noted as point (0,0). Each quadrat consisted of intersecting lines along ten meter by ten meter pre-determined points on the transect line. This essentially gave ten rows along the “x-axis” and ten rows along the “y-axis”. At each intersecting point, three shoots of two leaves and a bud were harvested from each tree and placed in zip-lock plastic bags. A total of 400 samples were collected, 100 from each of the two good and two poor fields. Following sample collection, the leaves were transported to the Tea Research Institute laboratory, with the zip-locked plastic bags placed in an insulated box on ice. The samples were then subjected to the SWAPDT method as discussed in Nyarukowa *et al.*, (2016) to determine the RWC of the leaves from each field. The SWAPDT method is an inexpensive and practical method developed for the prediction of DT tea cultivars. The rate of RWC loss between the DT and DS cultivars was evaluated by immersing three shoots with two leaves and a bud from a single bush under investigation in 20 ml of distilled water at room temperature and weighed after 24 hours. These turgid leaves were then blot dried and weighed before being oven dried at 37°C and weighed after five hours when their RWC is between 40 - 80%. The leaves were again placed in water for 24 hours, and then weighed, and oven dried at 105°C for 24 hours to obtain each leaf's dry weight.

2.5.2 Statistical Analysis

The data collected on the four fields, were tested for spatial homogeneity. In the case of field 13B, spatial dependence was detected, as provided by (Bivand *et al.*, 2011). A spatial simultaneous autoregressive error model was fitted to the data of field 13B, to predict and consequently remove the spatial signal. This was done using the “spdep” package. Using the one-sided Mann-Whitney test, it was established that Fields 12A and 12B, and 13A and 13B were respectively similar, but that the two groups differ significantly from one another. A Monte-Carlo permutation test approach was then employed, using the p-value of the one-sided Mann-Whitney test as test statistic, applied to comparisons between Fields 12A and 13A, 12A and 13B, 12B and 13A, and 12B and 13B. One thousand repetitions of this test were performed at decreasing sample sizes, to construct an empirical distribution of the p-value for each sample size. The empirical quantile at which 0.05 and 0.01 was observed was recorded. As the sample sizes decreased, the stability of the Mann-Whitney test decreased, as could be seen by the quantiles which usually lead to the rejection of the null hypothesis now moving away from the right-tail of the empirical distribution. The minimum required sample size was set at the level just before the empirical quantile dropped below the required significance level. This test procedure was also repeated while controlling for the mean difference in fields, and by combining the data for Fields 12A and 12B, and 13A and 13B into Fields 12 and 13, respectively.

Power curves for the tea data set were calculated using the package “pwr”. In Figure 2.23, four effect sizes have been defined, which correspond to a difference of 3%, 5%, 10% and 15%. The 3% and 5% represent small effect sizes, with the 10% representing a medium effect size and the 15% a large effect size. The difference found in the pilot data set (35%) would be considered a gigantic effect size. The curves were created for the different amount of fields included in the experiment; these curves serve to determine a suitable number of replicates required for each field for example:

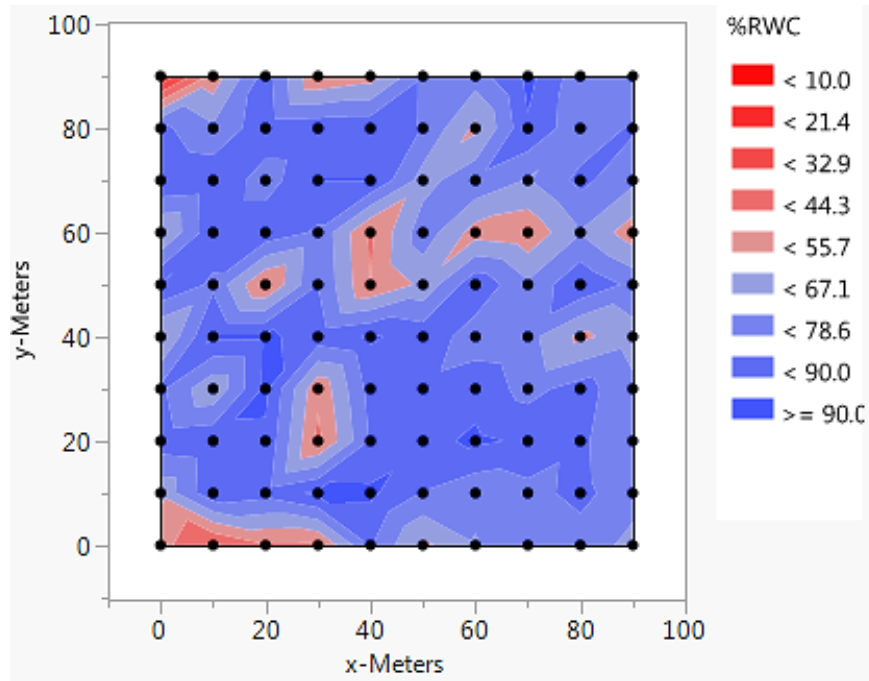
When the desired power is 0.90 and the smallest relevant effect size is 5% and if 100 fields are used in the experiment, the minimum number of necessary replicates is estimated at approximately 22. When 50 fields are used, the minimum number of replicates would be about 33. The sample size calculations are done after eliminating any other “field effects” i.e. using oneway ANOVA. Where “field effects” are not eliminated, a mixed model approach is employed by adding random indicators for fields to the model. This addition results in an

increase in the power. Oneway ANOVA was used to compare the two good and the two poor fields and compute their means, standard deviations, and the Student's t-test, and contour plots of each field were prepared by JMP Pro (ver 13). Excel was used to calculate the SEM, from the same Standard Deviation (SD) at different sample sizes and to obtain the equation of the curve.

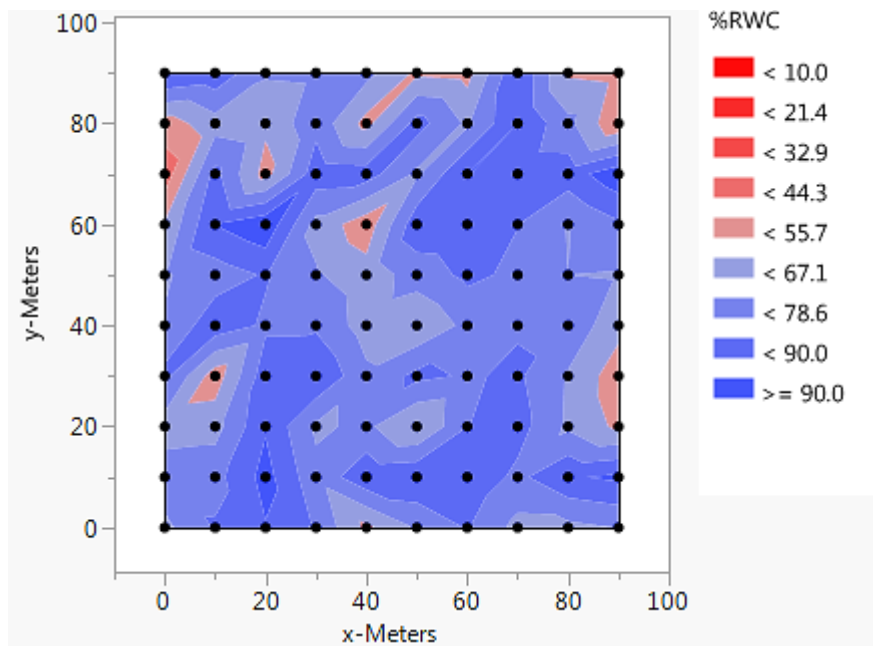
2.6 RESULTS

2.6.1 %drought score contour RWC plots based on SWAPDT method

The contour plots for the four fields are shown in Figure 2.4. These are indicative of the %RWC profiles, which were flat; this eliminates any possible bias due to underground rivers or rocky outcrops. The ANOVA comparison between the two good and the two poor fields shows clear differences, with the mean %RWC of 72.2 for the two good fields, and 35.0 for the two poor fields ($p < 0.0001$). The SD ranged from 13.3 - 20 units, indicating a large variation within each field. This is supported by the large coefficient of variation (CV) values, shown in Figure 2.5.



12A.



12B.

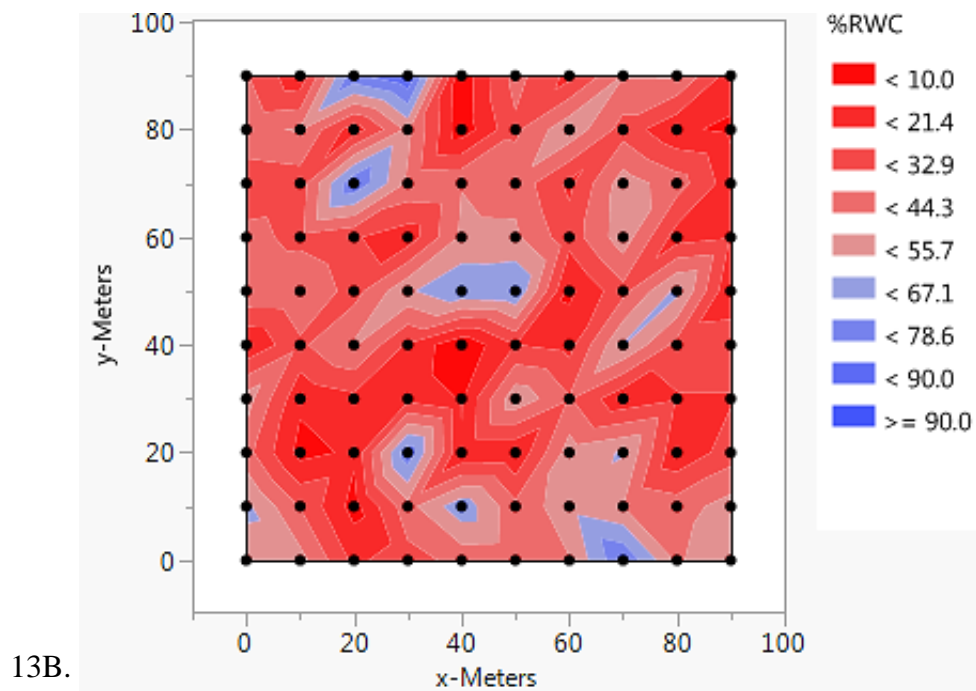
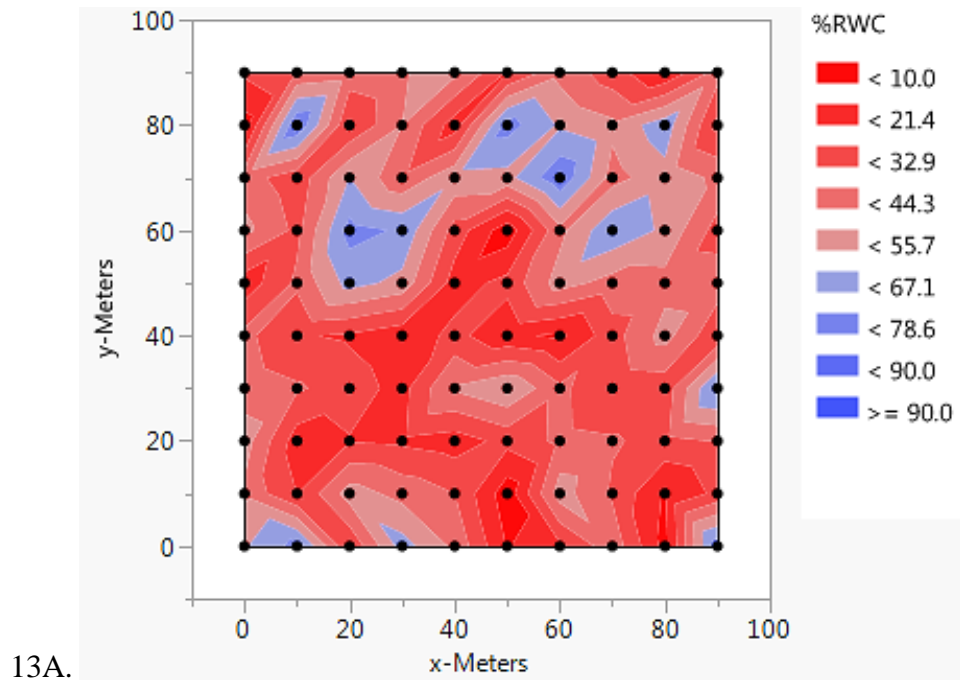
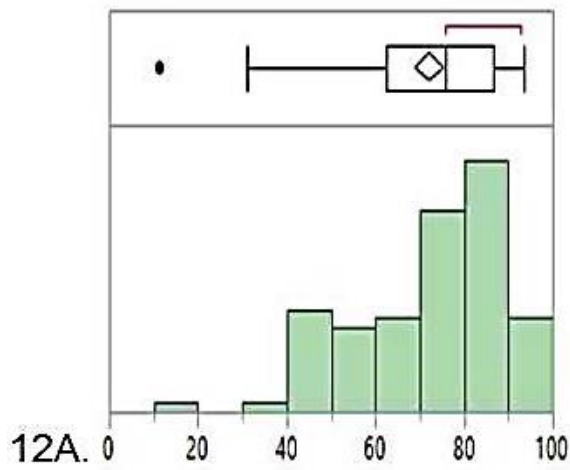
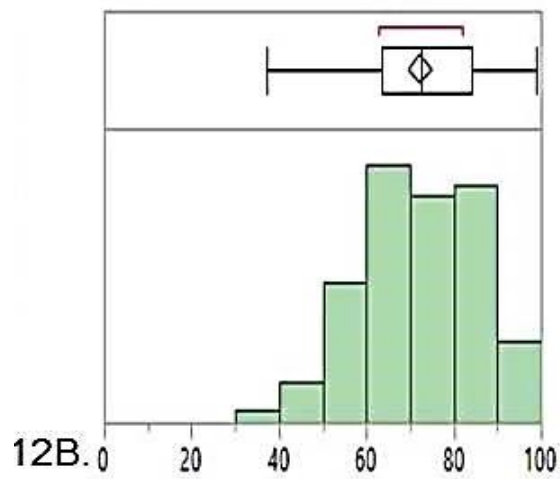


Figure 2.4: %RWC drought score contour plots based on SWAPDT method for the two good fields (12A and 12B) and the two poor fields (13A and 13B).



Summary Statistics	
Mean	72.140969
Std Dev	16.820685
Std Err Mean	1.6820685
Upper 95% Mean	75.478558
Lower 95% Mean	68.80338
N	100
CV	23.316412



Summary Statistics	
Mean	72.293559
Std Dev	13.517742
Std Err Mean	1.3517742
Upper 95% Mean	74.975773
Lower 95% Mean	69.611346
N	100
CV	18.698404

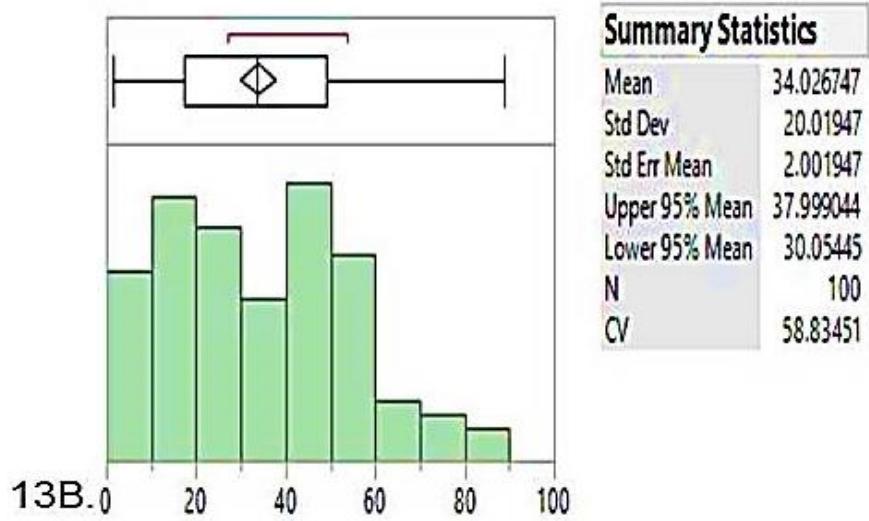
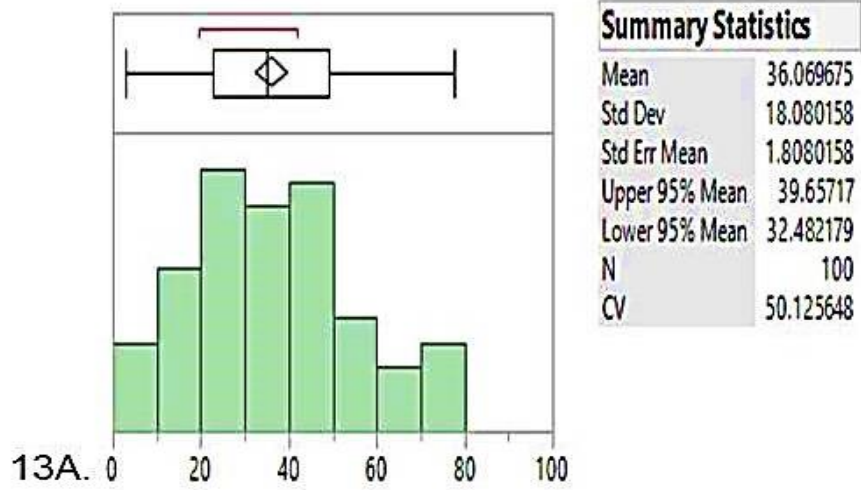


Figure 2.5: Mean distribution curves for the fields 12A, 12B, 13A and 13B. The plots show the Mean, Std Dev, SEM, sample size (N) and the CV for each field.

By approximation, the difference between the two means was expected to be statistically significant provided the difference of the means $>$ the sum of their SEMs or if $\text{Mean1} - \text{Mean2} > \text{SEM1} + \text{SEM2}$. The data for the two good fields were pooled and annotated as Good 1 and Good 2, while that of the poor fields was annotated as Poor 1 and Poor 2. The SEM for different sample sizes ($n = 100, 50, 25, 20, 15, 10$ and 5) for Good 1 and Good 2, and Poor 1 and Poor 2 pools were calculated using the equation shown in Figure 2.6.

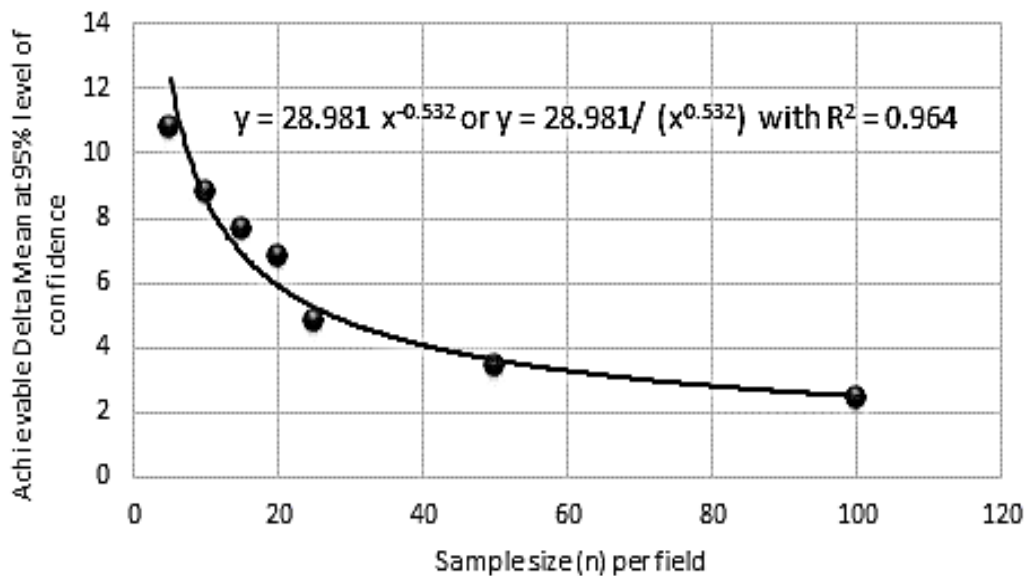
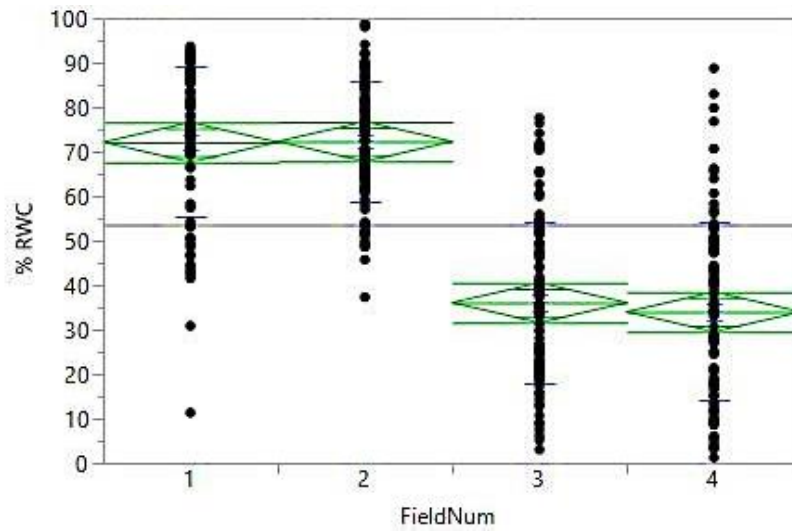


Figure 2.6: Ideal theoretical vs practical sample size required. The plot shows the initially postulated number of samples per field deemed practical versus the actual statistically obtained sample number at the 95% level of confidence.

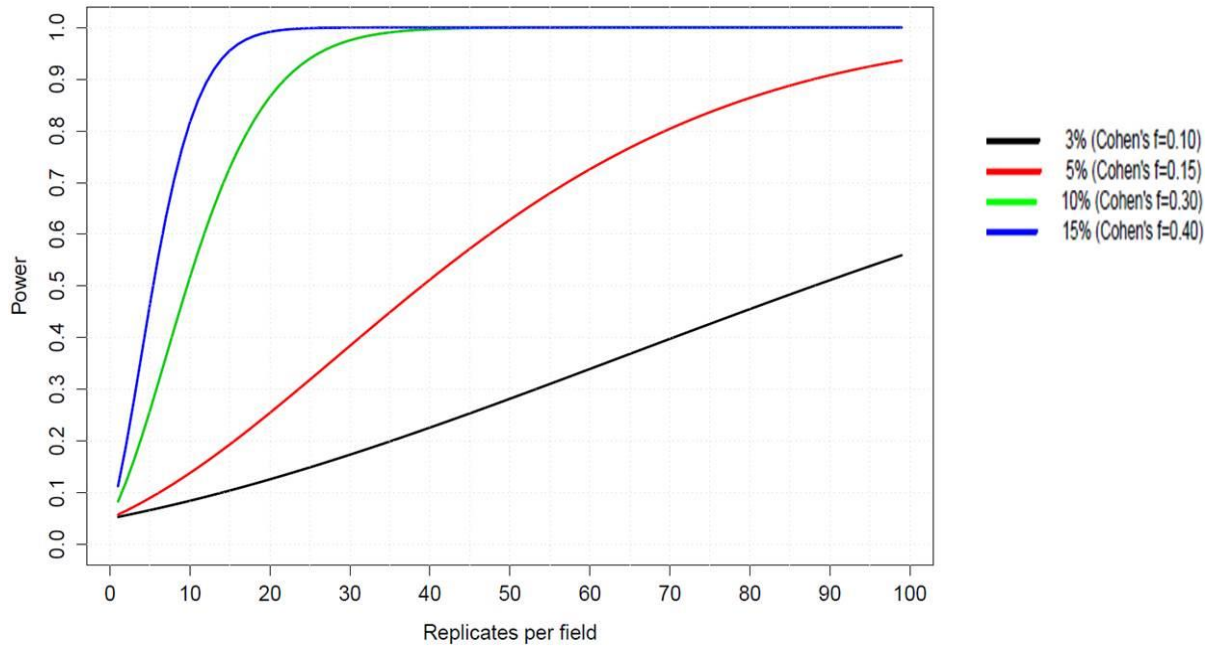
It is important to tea estate management to know the sample size for two fields whose means are close to each other. Figure 2.6 shows the required numbers of samples per field versus the delta mean at the 95% level of confidence. The figure shows that if the delta mean is 8%, the corresponding sample size is approximately 12. Using the SWAPDT method and a sample size of 20, it is possible to distinguish between fields with a delta mean of 6%. The collection of these 20 samples should be in the middle of the field to dispel any possibility of edge effects, and about 10 m apart within rows and 10 m apart between rows. Figure 2.7 below shows the Oneway analysis ANOVA results of the %RWC of the two good and two poor fields.



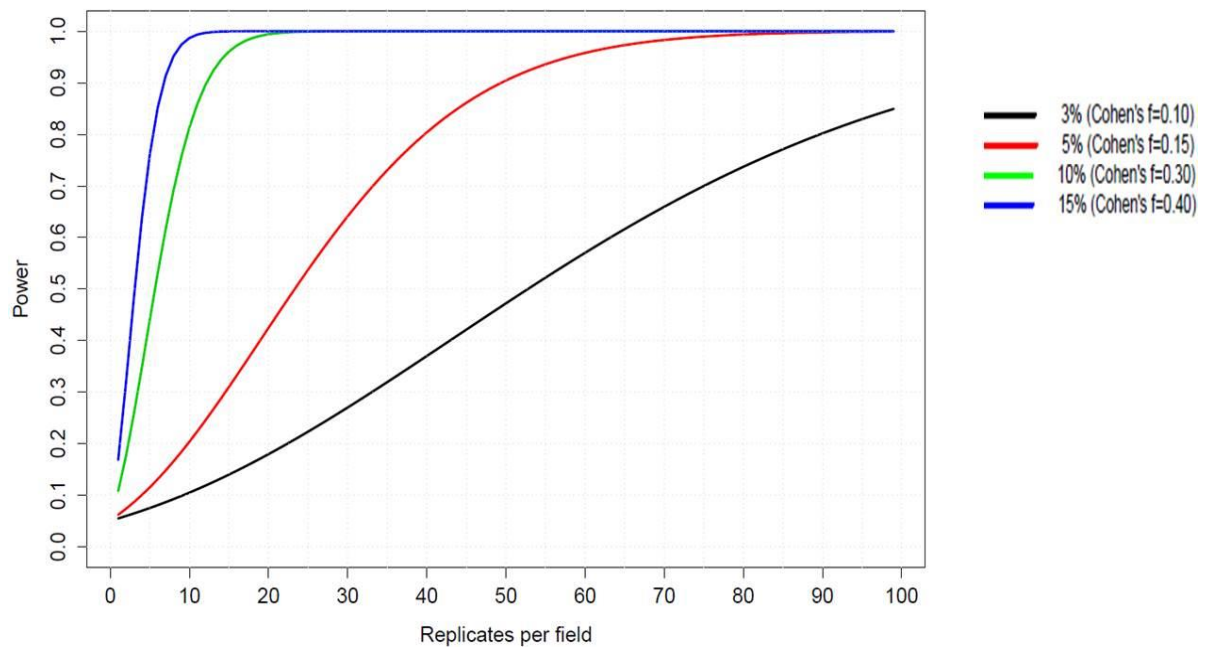
Oneway Anova					
Summary of Fit					
Rsquare		0.539421			
Adj Rsquare		0.535932			
Root Mean Square Error		17.27235			
Mean of Response		53.63274			
Observations (or Sum Wgts)		400			
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
FieldNum	3	138363.70	46121.2	154.5960	<.0001*
Error	396	118140.27	298.3		
C. Total	399	256503.97			

Figure 2.7: Oneway analysis of the % RWC against the two good and two poor fields. Oneway ANOVA was used to calculate the means.

Power curves for the data set were also plotted in Figure 2.8, showing the four effect sizes which correspond with a difference of 3%, 5%, 10% and 15%, as described above in the methods.



A.



B.

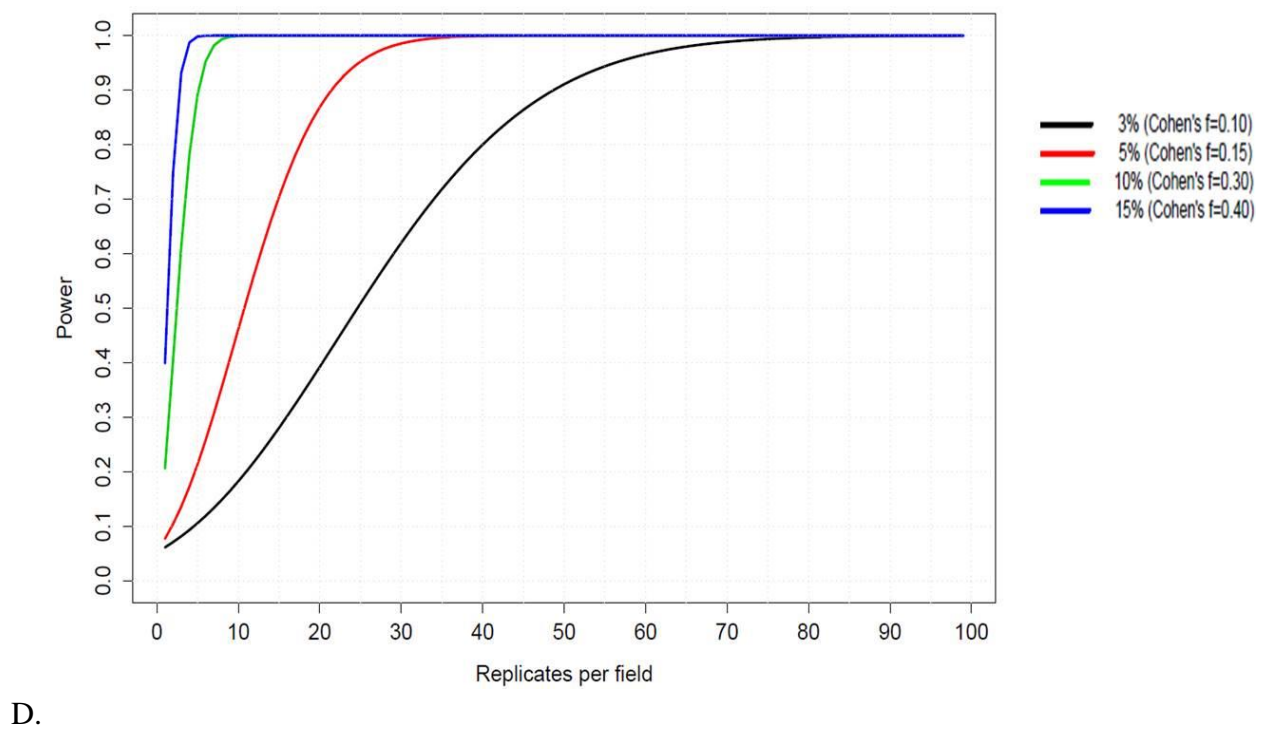
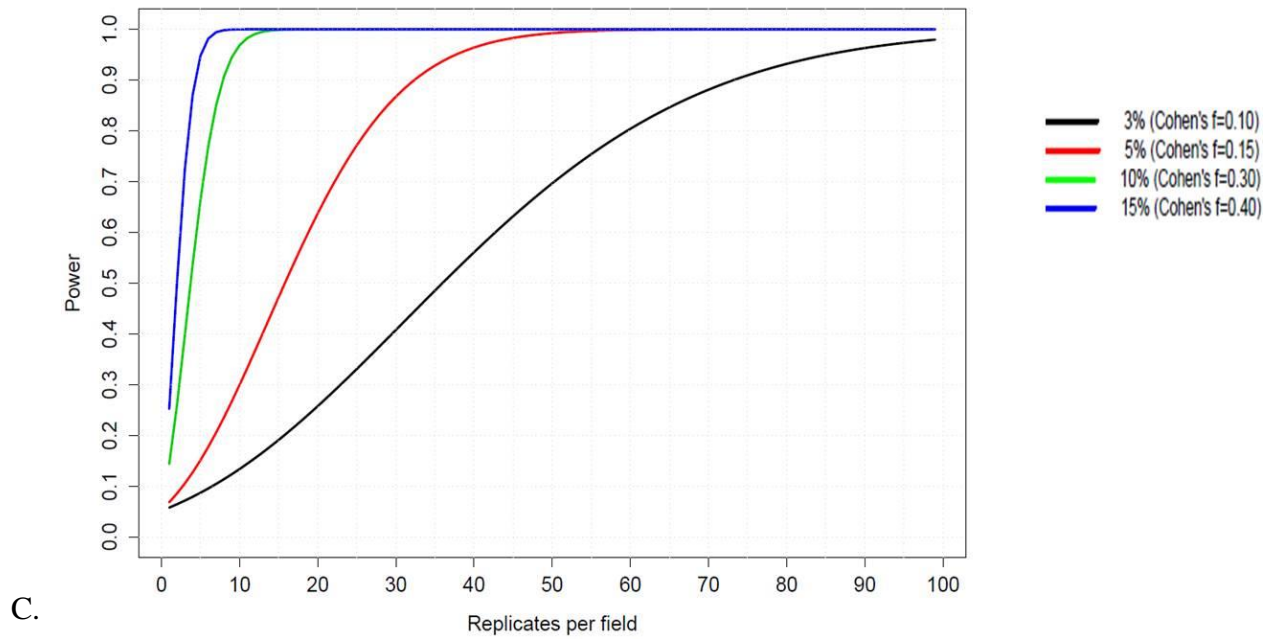


Figure 2.8: Power curves for the data set were plotted, showing four effect sizes which correspond with a difference of 3%, 5%, 10% and 15%. (A) represents the number of replicates for 10 fields; (B) represents the number of replicates for 25 fields. (C) represents the number of replicates for 50 fields, and (D) represents the number of replicates for 100 fields.

2.7 DISCUSSION AND CONCLUSION

The RWC contour maps are almost flat, suggesting that there are no geological features that affect SWAPDT method scores, in these fields. This work sought to find a robust monitoring solution capable of concurrently achieving cost effective, adaptive and timeous sampling; the SWAPDT method was determined to effectively achieve this objective. The contour plots served to show that the high %RWC results of the cultivars from the good fields was not due to e.g. underground streams; likewise to show that the cultivars from the bad fields didn't have e.g. molehills or old hut-sites. Both sets of plots showed that all the fields were normal. The solution attained allowed for improved sampling of the good fields (12A and 12B), and the poor fields (13A and 13B). The algorithm developed performs in-network data suppression for both spatial and temporal dimensions. This algorithm is employable for residual energy monitoring, faulty sensor detection and spatial-temporal event monitoring, with the results confirming the efficiency of the solution's accuracy and cost effectiveness. Figure 2.5 and Figure 2.7 show the mean distribution curves and Oneway ANOVA results, respectively, for the two good fields (12A and 12B) and the two poor fields (13A and 13B). From these results, it is evident that the good fields have a higher %RWC mean, of 72 for both fields 12A and 12B, than the poor field with %RWC means of 34 and 36 for fields 13A and 13B respectively. These results are in agreement with those obtained from the contour plots and what is theoretically expected i.e. that the good fields, with drought tolerant cultivars would have a higher %RWC than the drought susceptible cultivars found on the bad fields. The plot in Figure 2.6 shows the initially postulated number of samples per field deemed practical to observe significant differences between fields versus the actual statistically obtained sample number at the 95% level of confidence. From this figure, it can be seen that the initially postulated 100 samples per field would be required to differentiate between fields with approximately 2% variation between them. As the variation between the fields increases, a smaller sample size is required i.e. when the variation between the fields is more than 6%, as little as 20 samples per field is required. From the means of the good and poor fields shown in Figure 2.5, it is clear that the two sets of fields vary significantly and as such as little as 5 samples per field will be able to distinguish between them on a statistically significant level. The results presented in the ANOVA showed that the SWAPDT method distinguishes good fields from poor fields.

The power of a hypothesis test, according to Wise, (1974), tests the probability that the null hypothesis will correctly be rejected; this test is influenced by sample size. In instances where

a test possesses a low power, an effect may go undetected and it may be wrongly concluded that none exists, leading to the rejection of the H_0 when it, in actuality is true. On the other hand, if the test power is too high, minor effects may be taken to be significant, and this may result in a failure to reject H_0 , when it is actually false. A power value of e.g. 0.9 signifies that if one was to repeat the same experiment several times, taking random samples each time, 90% of the time may result in the correct rejection of the H_0 , and the remaining 10% may result in sampling errors, which result in a failure to reject H_0 . Power curves plot the power of the hypothesis test against the difference between the mean and the target (Ariyavisitakul and Chang, 1993). The sample size is used to ascertain the number of observations required to obtain a certain power value for the testing of the hypothesis at a particular difference. Each power curve represents every combination of power and difference for each sample size when the significance level and the Std Dev are held constant. From Figure 2.8, sample sizes of 10, 25, 50, and 100, and their corresponding power curves are shown. Increasing the sample size employed in the hypothesis test results in a power increase, and this is observed from the power curves. As the sample size increased from 10 to 100, the power values increased from less than 0.1 to 1.0 between fields which vary by as little as e.g. 3%. It is therefore necessary to have enough sample observations to achieve adequate power, without having a large sample size, which leads to a waste of time and money on redundant sampling (Wise, 1974). The findings of the two good and two poor fields correspond with the historical in-filling records that were available for these fields. This finding suggests that where historical in-filling records are not available, the SWAPDT method may be used to prioritize fields for replanting. These results also show that the SWAPDT method developed on tea cultivars from the TRFCA in Malawi can be applied to the seedling tea fields in Kenya, suggesting that the SWAPDT method may apply to other tea growing regions of the world. The sample size of 20 tea trees per field is sufficient to distinguish between fields that vary by 6% or more in their mean SWAPDT score. This sample size will need to be increased if the CV within a field is greater than 60%, i.e. if fields were planted with a more heterogeneous source of seeds than used in the four fields reported here.

2.8 REFERENCES

- Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, 30, 185–207.
- Ariyavisitakul, S. & Chang, L. F. (1993). Signal and interference statistics of a CDMA system with feedback power control. *IEEE Transactions on Communications* 41(11): 1626-1634.
- Avelino, J., Willocquet, L. & Savary, S. (2004). Effects of crop management patterns on coffee rust epidemics. *Plant Pathology* 53(5): 541-547.
- Bailey, T. C. & Gatrell, A. C. (1995). Interactive spatial data analysis. Longman Scientific & Technical Essex.
- Baldino, F., Cowan, A., Geller, E. B. & Adler, M. W. (1979). Effects of antipsychotic and antianxiety drugs on the morphine abstinence syndrome in rats. *Journal of Pharmacology and Experimental Therapeutics* 208(1): 63-66.
- Besag, J. & Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 26(3): 327-333.
- Bivand, R., Anselin, L., Berke, O., Bernat, A., Carvalho, M., Chun, Y., Dormann, C., Dray, S., Halbersma, R. & Lewin-Koh, N. (2011).spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-31, URL <http://CRAN.R-project.org/package=spdep>.
- Bivand, R., Hauke, J. & Kossowski, T. (2013). Computing the J acobian in G aussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. *Geographical Analysis* 45(2): 150-179.
- Chi, Guangqing, and Jun Zhu. (2008). Spatial regression models for demographic analysis." *Population Research and Policy Review* 27(1): 17-42.
- Cressie, N. (1993).Statistics for spatial data New York. Wiley-Interscience.
- Cumming, G., Fidler, F. & Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology* 177(1): 7-11.
- DaMatta, F. M. (2004). Ecophysiological constraints on the production of shaded and unshaded coffee: a review. *Field Crops Research* 86(2-3): 99-114.
- Dupont, W. D. & Plummer Jr, W. D. (1990). Power and sample size calculations: a review and computer program. *Controlled Clinical Trials* 11(2): 116-128.
- Ellis, R. & Nyirenda, H. (1995). A successful plant improvement programme on tea (*Camellia sinensis*). *Experimental Agriculture* 31(3): 307-323.

- Fagerland, M. W. & Sandvik, L. (2009). The wilcoxon–mann–whitney test under scrutiny. *Statistics in Medicine* 28(10): 1487-1497.
- <http://www.finlays.net>
- JMP®, V. SAS Institute Inc., Cary, NC, 1989–2007.
- Kelejian, H. H. & Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics* 157(1): 53-67.
- Kelsey, J. L., Thompson, W. D., Evans, A. S., Faden, R. R., Beauchamp, T. L., King, N. M. P.,... and Hermann, B. P. (1986). *Observational Epidemiologic Studies*.
- Keselman, H., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D. & Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research* 68(3): 350-386.
- Mei, C. L., He, S. Y. & Fang, K. T. (2004). A note on the mixed geographically weighted regression model. *Journal of Regional Science* 44(1): 143-157.
- Meng, X., Nandagopal, T., Li, L. & Lu, S. (2006). Contour maps: Monitoring and diagnosis in sensor networks. *Computer Networks* 50(15): 2820-2838.
- Murakami, T., Nakamura, J., Matsuda, H. & YOSHIKAWA, M. (1999). Bioactive saponins and glycosides. XV. Saponin constituents with gastroprotective effect from the seeds of tea plant, *Camellia sinensis* L. var. *assamica* Pierre, cultivated in Sri Lanka: structures of assamsaponins A, B, C, D, and E. *Chemical and Pharmaceutical Bulletin* 47(12): 1759-1764.
- Nyarukowa, C., Koech, R., Loots, T. & Apostolides, Z. (2016). SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. *Journal of Plant Physiology* 198: 39-48.
- Ord, J. K. & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* 27(4): 286-306.
- Pacheco, A. I. & Tyrrell, T. J. (2002). Testing spatial patterns and growth spillover effects in clusters of cities. *Journal of Geographical Systems* 4(3): 275-285.
- Preedy, V. R. (2012). *Tea in health and disease prevention*. Academic Press.
- Savary, S., Willocquet, L., Elazegui, F. A., Teng, P. S., Van Du, P., Zhu, D., Tang, Q., Huang, S., Lin, X. & Singh, H. (2000). Rice pest constraints in tropical Asia:

characterization of injury profiles in relation to production situations. *Plant Disease* 84(3): 341-356.

Willson, K. C. & Clifford, M. N. (2012). Tea: Cultivation to consumption. *Springer Science & Business Media*.

Wise, M.E. (1974). Interpreting both short-and long-term power laws in physiological clearance curves. *Mathematical Biosciences* 20(3): 327-337.

Appendix 2.1: Peer-reviewed scientific article based on results from Chapter 2

Journal of Agricultural Science; Vol. 10, No. 7; 2018
ISSN 1916-9752 E-ISSN 1916-9760
Published by Canadian Center of Science and Education

Prioritising the Replanting Schedule of Seedling Tea Fields on Tea Estates for Drought Susceptibility Measured by the SWAPDT Method in the Absence of Historical In-filling Records

Christopher Nyarukowa¹, Robert Koech^{1,2}, Theodor Loots³, Jos Hageman⁴ & Zeno Apostolides¹

¹ Department of Biochemistry, University of Pretoria, Hatfield, South Africa

² Kenya Agriculture and Livestock Research Organization, Tea Research Institute, Kericho, Kenya

³ Department of Statistics, University of Pretoria, Hatfield, South Africa

⁴ Department of Mathematical and Statistical Methods, Wageningen University and Research, Wageningen, Netherlands

Correspondence: Zeno Apostolides, Department of Biochemistry, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa. Tel: 27(0)12-420-2486. Fax: 27(0)12-362-5302. E-mail: zeno.apostolides@up.ac.za

Received: April 19, 2018

Accepted: May 19, 2018

Online Published: June 15, 2018

doi:10.5539/jas.v10n7p26

URL: <https://doi.org/10.5539/jas.v10n7p26>

Abstract

Due to the unpredictable natural droughts that occur, causing tea farmers significant losses in tea estates, a two-day method for distinguishing between drought tolerant (DT) and drought susceptible (DS) *Camellia sinensis* cultivars was developed. This work was based on known cultivars developed at the Tea Research Institute in Kenya and the Tea Research Foundation for Central Africa in Malawi. This paper contains an in-depth description of the application of the SWAPDT method on four 60-year old, *C. sinensis* seedling fields in Kenya. The in-filling history of the four fields due to drought-related deaths was obtained from historical records. The SWAPDT method scores correlated very well with the historical records. It has been indicated, from the results obtained in this study, that a sample size of 20 tea trees is sufficient to accurately determine the drought susceptibility of a large tea field of approximately 5-20 hectares, containing 50 000-200 000 tea trees, where the difference between their mean values, as measured by the SWAPDT method, is approximately 10%.

Keywords: *Camellia sinensis*, drought tolerance, field comparison, relative water content, SWAPDT, tea estate

1. Introduction

Tea made from the leaves of *Camellia sinensis*, as green or black tea, has been drunk as a mild stimulant due to the caffeine content, since time immemorial (Ellis & Nyirenda, 1995). Tea consumption has increased in recent years, due to the health-promoting effects associated with its high polyphenol content (Preedy, 2012). *C. sinensis* is cultivated in over 52 countries around the world. Global world trade is approximately 78% by value in the form of black, 20% as green and 2% as oolong tea (Nyarukowa, Koech, Loots, & Apostolides, 2016). It is an important cash crop for countries such as India and China; in Africa alone, several countries produce tea, namely in Kenya, which is currently ranked third behind Sri Lanka and India with regards to annual production and export of black tea (FAO, 2015), Malawi, Uganda, Tanzania, Zimbabwe, Rwanda, South Africa, Burundi and Mauritius. *C. sinensis* tea estates need to be replanted every 20-90 years to maintain high yields. Tea estates are planted in sample blocks of about 5-20 hectares, at 10,000 trees per hectare, with seeds from the same batch. Most tea estates in Africa are planted with tea seeds procured from tea-baries (orchards) in India or Sri Lanka. The seed selection criteria employed focused on yield and neglected to consider drought tolerance (Murakami, Nakamura, Matsuda, & Yoshikawa, 1999). During severe natural droughts, some of the trees (5-15%) die and are replaced with new trees. Tea planters refer to this process as "in-filling." Most estates keep good records of the in-filling, and hence good and poor fields are easily identified. However, sometimes these records are missing, and a new method is required to determine the drought tolerance of a tea field that might be 20-90 years old (Willson & Clifford, 2012).

Tea producers demand new cultivars which are DT, to reduce crop losses. In the coffee industry, farmers are faced with the problem of dealing with coffee rust. How they deal with this is by assessing and analysing the risk

of an epidemic by considering the region's characteristics such as climate, soils, crop management patterns, namely shade management, etc. (DaMatta, 2004). This assessment approach has been adopted from studies conducted in West Africa on groundnuts (Avelino, Willocquet, & Savary, 2004) as well as on work conducted by (Savary et al., 2000) on tropical Asian rice. As a result of this, tea farmers are also looking for an inexpensive yet effective method of determining which sample blocks of tea have a high percentage of DS plants so that these sample blocks may be prioritized for replanting. The samples in this study were collected from the James Finlay's estate in Kericho, Kenya which together with surrounding estates (Figure 1) produces 23 million kilograms of tea annually. This part of Kenya enjoys deep rich loam soils, which are high in organic content and combined with the perfect climate and environment are ideal for high yields of good quality tea (http://www.finlays.net). In Nyarukowa et al. (2016), a novel logistics probability formula was developed, which can be used to calculate a new cultivar's probability to be DT after employing the Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) method. The aim of this study was to determine how many tea trees are needed per field to obtain a representative sample of the tea field so that tea fields can be prioritised for replanting.

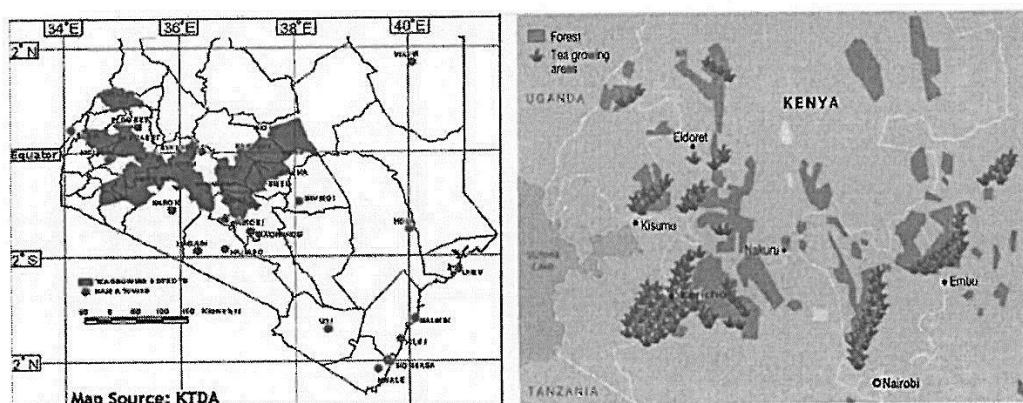


Figure 1. Tea growing areas in Kenya

Source: Kenya Tea Development Agency (<http://blog.dominiontea.com/2014/03/27/kenyan-tea-industry>).

2. Statistical Analysis

Statistical approaches for the analysis of metabolomic data vary. The first step involves the analysis/focus on single trait of interest (univariate analysis), or the effects of multiple metabolites on the outcome being studied (multivariate analysis). Results can then be validated using different approaches and fed into the multivariate analysis. Spatial regression models have been widely employed in biological science disciplines. Central to spatial analysis and quantitative geography, has been the study of methods that specify and fit spatial regression models (Bivand, Hauke, & Kossowski, 2013). The most popular spatial correlation test is based on the Moran I test statistic. This test is in a normalised quadratic form, with regards to the variables being tested for spatial correlation (Kelejian & Prucha, 2010). The Moran I test statistic functions to test that the spatial autocorrelation of a variable in the null hypothesis is zero. Upon the rejection of the null hypothesis, the variable will be deemed to be spatially auto correlated (Ord & Getis, 1995). The Monte Carlo test can also be employed when dealing with a null hypothesis, H_0 , and a corresponding dataset, in which the value u_1 of a selected test statistic, u , is ranked amongst a set of analogous values generated by randomly sampling from the null distribution of u . The rank of the test statistic u_1 , when the u distribution is continuous amongst the set of values $\{u_i; i = 1, \dots, m\}$ enables the determination of the exact significance level for the test (Besag & Diggle, 1977). The ordinary linear regression (OLR) model is amongst the most useful statistical methods employed in spatial analysis to identify relationships between variables is (Mei, He, & Fang, 2004). In this technique, the dependent variable, y , is modelled as a linear function of a set of independent variables x_1, x_2, \dots, x_p . Based on n observations $(y_i; x_{i1}, x_{i2}, \dots, x_{ip})$, ($i = 1, 2, \dots, n$), from a study region, the model can be expressed as:

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ik} + \epsilon_i \quad (1)$$

where, $\beta_0, \beta_1, \dots, \beta_p$ are parameters, and $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are error terms assumed to be normally distributed, random, independent variables (Mei et al., 2004).

Determining a study's sample size requires a compromise, which balances power, economy, and timeliness (Dupont & Plummer, 1990). Researchers must define the sample size, power, and hypotheses for their study and to effectively do this, they require software programs that can calculate the missing parameter when given any two of the preceding three parameters. These programs compute the sample size required to identify, using a specified power, e.g., the difference in efficacy of a particular treatment, the power this difference is detectable with given sample size, and the detectable difference with particular power and sample size. The link between the power and its corresponding sample size is best seen by plotting the power curve as a function of the parameter of interest (Kelsey et al., 1986).

The Student's t-test was used to calculate the probability (p) that the two samples belong to the same population. When $p < 0.05$, there is a 95% certainty that the two samples belong to different populations. The Standard Error of the Mean (SEM) is an inferential statistic that can be used to draw error bars on histograms to visually estimate the p-value. When the sample size (n) > 10, and the gap between the SEM error bars > SEM1 + SEM2, we can be 99% confident that the samples are from two different populations (Cumming, Fidler, & Vaux, 2007). The MANOVA method may also be used to compare multiple fields with each other, and the three methods are expected to produce similar results (Keselman et al., 1998).

Oneway analysis using JMP Pro 13 generates "mean diamonds", which illustrate both the sample mean and confidence interval. The top and bottom of each diamond represent the $(1 - \alpha) \times 100$ confidence interval for each group. The confidence interval computation assumes that the variances are equal across observations, and as such the height of each diamond is proportional to the reciprocal of the square root of the number of observations within the group. The mean line across the middle of each diamond represents the group mean, while the overlap marks appear as lines above and below the group mean. In instances where groups have equal sample sizes, these overlapping marks indicate that the two group means are not significantly different at the given confidence level. Where the mean in one diamond is between the overlap marks of another diamond, this indicates that these two groups are not significantly different at that confidence level (SAS Institute Inc., 1989-2007).

3. Materials and Methods

3.1 Sample Collection

The field work was conducted on the KAPRORET tea estate in Kericho, Kenya, with latitude: $0^{\circ}22'3.86''N$ and longitude: $35^{\circ}16'59.30''E$, during January and February of 2017. Based on historical in-filling records, two good fields, fields 12A and 12B, with fewer in-fillings due-to-deaths-by-drought (planted from a good batch of seeds) and two poor fields, 13A and 13B, with higher in-fillings due-to-deaths-by-drought (planted from a poor batch of seeds) were selected. These fields were approximately 1200 meters apart. The fields were planted from different batches of seeds obtained from Assam, in 1954 and 1956 respectively. They were in the same prune year, receiving the same fertiliser regimen, under rainfed conditions. The longitude and latitude coordinates for field 12A and 12B are $35^{\circ}14.75'E$, $0^{\circ}26'S$ and for field 13A and 13B are $35^{\circ}15.05'E$, $0^{\circ}26.6'S$ and at an altitude of 2180 m above mean sea level. Samples were collected using a "point-intercept within a quadrat method", in which a 100×100 m quadrat was set up in the middle of each field. The starting point of the sampling was noted as point (0,0). Each quadrat consisted of intersecting lines along ten meter by ten meter pre-determined points on the transect line. This essentially gave ten rows along the "x-axis" and ten rows along the "y-axis". At each intersecting point, three shoots of two leaves and a bud were harvested from each tree and placed in zip-lock plastic bags. A total of 400 samples were collected, 100 from each of the two good and two poor fields. Following sample collection, the leaves were transported to the Tea Research Institute laboratory, with the zip-locked plastic bags placed in an insulated box on ice. The samples were then subjected to the SWAPDT method as discussed in Nyarukowa et al. (2016) to determine the relative water content of the leaves from each field. The SWAPDT method is an inexpensive and practical method developed for the prediction of DT tea cultivars. The rate of relative water content (RWC) loss between the DT and DS cultivars was evaluated by immersing three shoots with two leaves and a bud from a single bush under investigation in 20 ml of distilled water at room temperature and weighed after 24 hours. These turgid leaves were then blot dried and weighed before being oven dried at $37^{\circ}C$ and weighed after five hours when their RWC is between 40-80%. The leaves were again placed in water for 24 hours, weighed after and oven dried at $105^{\circ}C$ for 24 hours to obtain each leaf's dry weight.

3.2 Statistical Analysis

The data collected on the four fields, were tested for spatial homogeneity. In the case of field 13B, spatial dependence was detected, as provided by Bivand et al. (2011). A spatial simultaneous autoregressive error model was fitted to the data of field 13B, to predict and consequently remove the spatial signal. This was done using the “spdep” package. Using the one-sided Mann-Whitney test, it was established that Fields 12A and 12B, and 13A and 13B were respectively similar, but that the two groups differ significantly from one another. A Monte-Carlo permutation test approach was then employed, using the p-value of the one-sided Mann-Whitney test as test statistic, applied to comparisons between Fields 12A and 13A, 12A and 13B, 12B and 13A, and 12B and 13B. One thousand repetitions of this test were performed at decreasing sample sizes, to construct an empirical distribution of the p-value for each sample size. The empirical quantile at which 0.05 and 0.01 was observed was recorded. As the sample sizes decreased, the stability of the Mann-Whitney test decreased, as could be seen by the quantiles which usually lead to the rejection of the null hypothesis now moving away from the right-tail of the empirical distribution. The minimum required sample size was set at the level just before the empirical quantile dropped below the required significance level. This test procedure was also repeated while controlling for the mean difference in fields, and by combining the data for Fields 12A and 12B, and 13A and 13B into Fields 12 and 13, respectively.

Power curves for the tea data set were calculated using the package “pwr”. In Figure 4, four effect sizes have been defined, which correspond to a difference of 3%, 5%, 10% and 15%. The 3% and 5% represent small effect sizes, with the 10% representing a medium effect size and the 15% a large effect size. The difference found in the pilot data set (35%) would be considered a gigantic effect size. The curves were created for the different amount of fields included in the experiment; these curves serve to determine a suitable number of replicates required for each field for example:

When the desired power is .90 and the smallest relevant effect size is 5% and if 100 fields are used in the experiment, the minimum number of necessary replicates is estimated at approximately 22. When 50 fields are used, the minimum number of replicates would be about 33. The sample size calculations are done after eliminating any other “field effects” *i.e.* using oneway ANOVA. Where “field effects” are not eliminated, a mixed model approach is employed by adding random indicators for fields to the model. This addition results in an increase in the power. Oneway ANOVA was used to compare the two good and the two poor fields and compute their means, standard deviations, and the Student’s t-test, and contour plots of each field were prepared by JMP Pro (ver 13). Excel was used to calculate the SEM, from the same Standard Deviation (SD) at different sample sizes and to obtain the equation of the curve.

4. Results

4.1 % RWC Drought Score Contour Plots Based on SWAPDT Method

The contour plots for the four fields are shown in Figure 2. These are indicative of the % RWC profiles, which were flat; this eliminates any possible bias due to underground rivers or rocky outcrops. The ANOVA comparison between the two good and the two poor fields shows clear differences, with the mean % RWC of 72.2 for the two good fields, and 35.0 for the two poor fields ($p < 0.0001$). The SD ranged from 13.3-20 units, indicating a large variation within each field. This is supported by the large coefficient of variation (CV) values, shown in Figure 3.

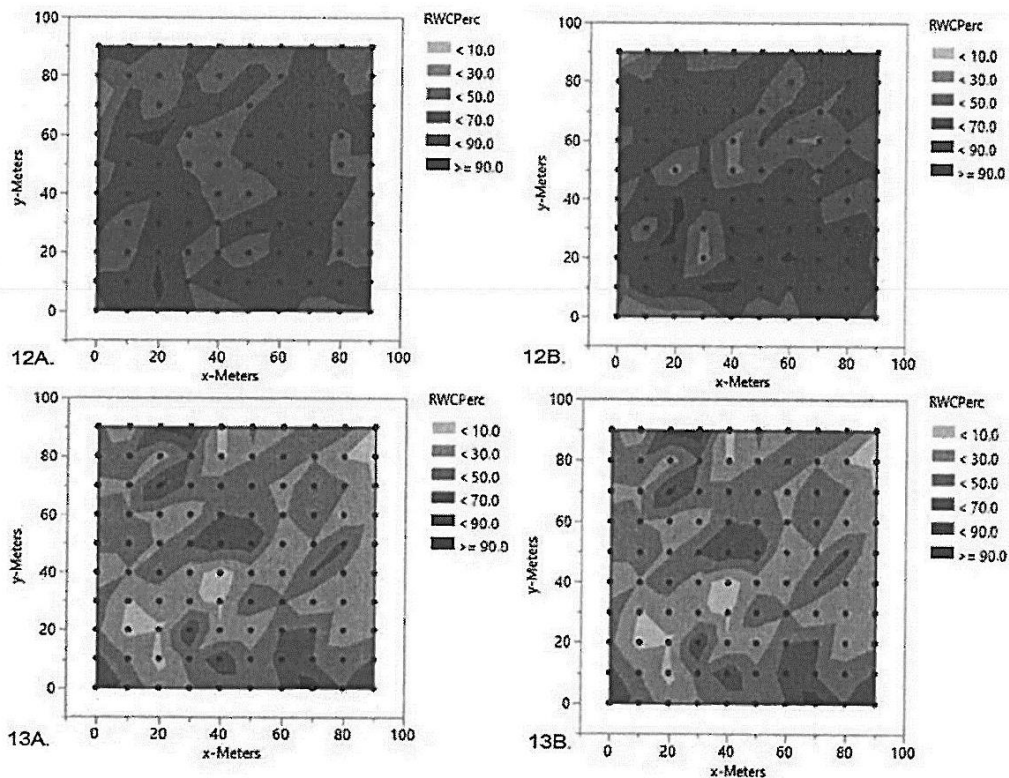


Figure 2. % RWC drought score contour plots based on SWAPDT method for the two good fields (12A and 12B) and the two bad fields (13A and 13B)

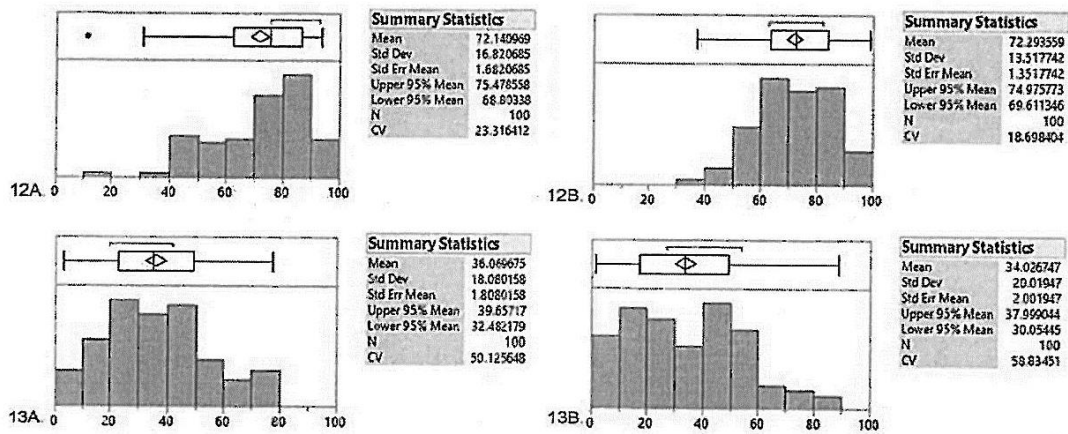


Figure 3. Mean distribution curves for the two good fields (12A and 12B) and the two bad fields (13A and 13B). The plots show the mean, std dev, SEM, sample size and the CV for each field

By approximation, the difference between the two means was expected to be statistically significant provided the difference of the means > the sum of their SEMs or if $Mean1 - Mean2 > SEM1 + SEM2$. The data for the two good fields were pooled and annotated as Good 1 and Good 2, while that of the poor fields was annotated as

Poor 1 and Poor 2. The SEM for different sample sizes (n = 100, 50, 25, 20, 15, 10 and 5) for Good 1 and Good 2, and Poor 1 and Poor 2 pools were calculated using the equation shown in Figure 4.

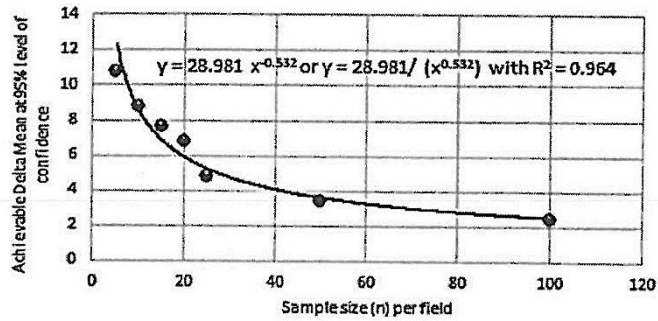


Figure 4. Plot shows the initially postulated number of samples per field deemed practical versus the actual statistically obtained sample number at the 95% level of confidence

It is important to tea estate management to know the sample size for two fields whose means are close to each other. Figure 4 shows the required numbers of samples per field versus the delta mean at the 95% level of confidence. The figure shows that if the delta mean is 8%, the corresponding sample size is approximately 12. Using the SWAPDT method and a sample size of 20, it is possible to distinguish between fields with a delta mean of 6%. The collection of these 20 samples should be in the middle of the field to dispel any possibility of edge effects, and about 10 meters apart within rows and 10 meters apart between rows. Figure 5 below shows the Oneway analysis ANOVA results of the % RWC of the two good and two bad fields.

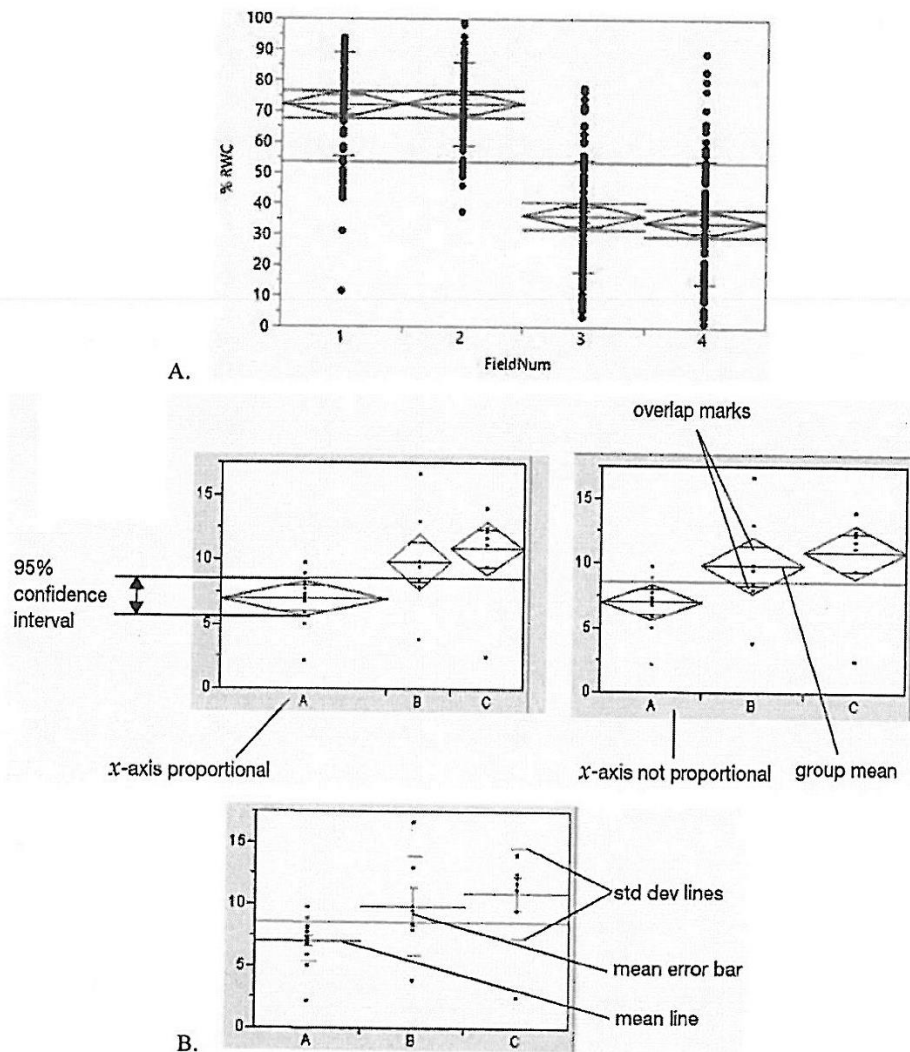


Figure 5. (A). Oneway analysis of the % RWC against the two good and two bad fields. Oneway ANOVA was used to calculate the means. (B). Examples of Mean Diamonds and X-Axis Proportional Options. Mean Lines, Mean Error Bars, and Std Dev Lines. (SAS Institute Inc., 1989-2007)

Power curves for the data set were also plotted in Figure 6, showing the four effect sizes which correspond with a difference of 3%, 5%, 10% and 15%, as described above in the methods.

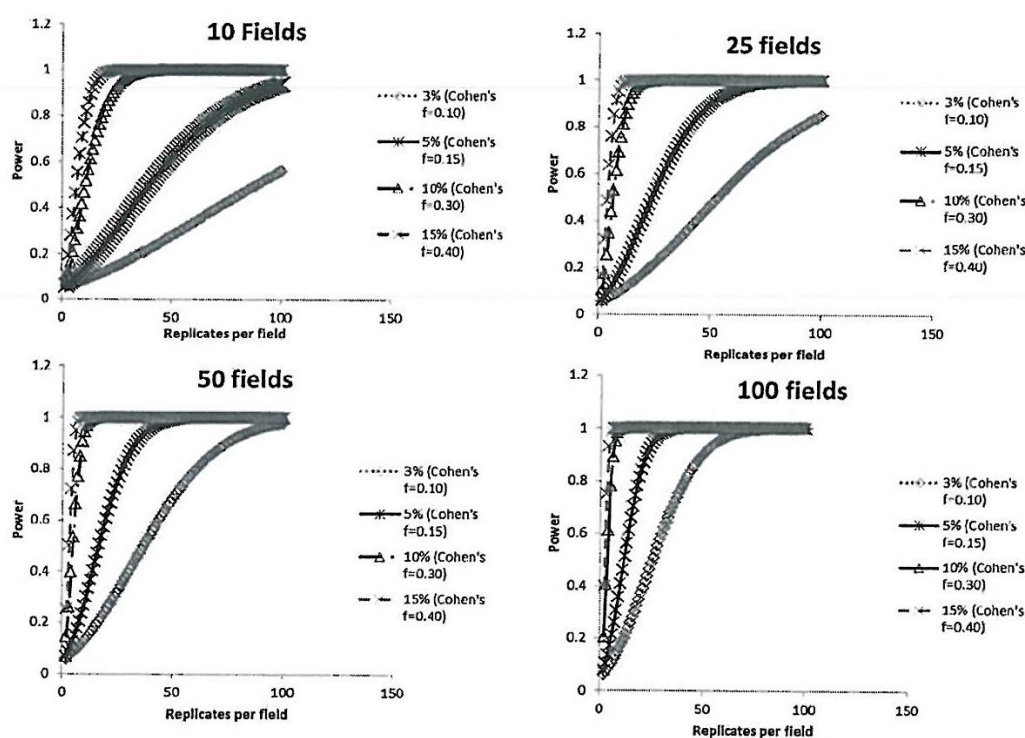


Figure 6. Power curves for the data set were plotted, showing four effect sizes which correspond with a difference of 3%, 5%, 10% and 15%

5. Discussion and Conclusion

The RWC contour maps are almost flat, suggesting that there are no geological features that affect SWAPDT method scores, in these fields.

The results presented in the ANOVA showed that the SWAPDT method distinguishes good fields from poor fields. These findings correspond with the historical records of in-filling that were available for these fields. This finding suggests that where historical in-filling records are not available, the SWAPT method may be used to prioritize fields for replanting. These results also show that the SWAPDT method developed on tea cultivars from the TRFCA in Malawi can be applied to the seedling tea fields in Kenya, suggesting that the SWAPDT method may apply to other tea growing regions of the world. The sample size of 20 tea trees per field is sufficient to distinguish between fields that vary by 6% or more in their mean SWAPDT score. This sample size will need to be increased if the CV within a field is greater than 60%, *i.e.* if fields were planted with a more heterogeneous source of seeds than used in the four fields reported here.

References

- Avelino, J., Willocquet, L., & Savary, S. (2004). Effects of crop management patterns on coffee rust epidemics. *Plant Pathology*, 53(5), 541-547. <https://doi.org/10.1111/j.1365-3059.2004.01067.x>
- Besag, J., & Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Applied statistics* (pp. 327-333). <https://doi.org/10.2307/2346974>
- Bivand, R. S., Hauke, J., & Kossowski, T. (2013). Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis*, 45(2), 150-179. <https://doi.org/10.1111/gean.12008>
- Bivand, R., Anselin, L., Berke, O., Bernat, A., Carvalho, M., Chun, Y., ... Lewin-Koh, N. (2011). *SPDEP: Spatial dependence: Weighting schemes, statistics and models*. CRAN. Retrieved from <https://cran.r-project.org/web/packages/spdep/index.html>

- Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology*, 177(1), 7-11. <https://doi.org/10.1083/jcb.200611141>
- DaMatta, F. M. (2004). Ecophysiological constraints on the production of shaded and unshaded coffee: a review. *Field Crops Research*, 86(2), 99-114. <https://doi.org/10.1016/j.fcr.2003.09.001>
- Dupont, W. D., & Plummer, W. D. (1990). Power and sample size calculations: A review and computer program. *Controlled Clinical Trials*, 11(2), 116-128. [https://doi.org/10.1016/0197-2456\(90\)90005-M](https://doi.org/10.1016/0197-2456(90)90005-M)
- Ellis, R., & Nyirenda, H. E. (1995). A successful plant improvement programme on tea (*Camellia sinensis*). *Experimental Agriculture*, 31(3), 307-323. <https://doi.org/10.1017/S0014479700025485>
- FAO (Food and Agricultural Organisation). (2015). Retrieved from <http://www.fao.org>
- Kelejian, H. H., & Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1), 53-67. <https://doi.org/10.1016/j.jeconom.2009.10.025>
- Kelsey, J. L., Thompson, W. D., Evans, A. S., Faden, R. R., Beauchamp, T. L., King, N. M. P., ... Hermann, B. P. (1986). *Observational Epidemiologic Studies*.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., ... Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350-386. <https://doi.org/10.3102/00346543068003350>
- Mei, C. L., He, S. Y., & Fang, K. T. (2004). A note on the mixed geographically weighted regression model. *Journal of Regional Science*, 44(1), 143-157. <https://doi.org/10.1111/j.1085-9489.2004.00331.x>
- Murakami, T., Nakamura, J., Matsuda, H., & Yoshikawa, M. (1999). Bioactive saponins and glycosides. XV. Saponin constituents with gastroprotective effect from the seeds of tea plant, *Camellia sinensis* L. var. *assamica* Pierre, cultivated in Sri Lanka: Structures of assamsaponins A, B, C, D, and E. *Chemical and Pharmaceutical Bulletin*, 47(12), 1759-1764. <https://doi.org/10.1248/cpb.47.1759>
- Nyarukowa, C., Koech, R., Loots, T., & Apostolides, Z. (2016). SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. *Journal of Plant Physiology*, 198, 39-48. <https://doi.org/10.1016/j.jplph.2016.04.004>
- Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4), 286-306. <https://doi.org/10.1111/j.1538-4632.1995.tb00912.x>
- Preedy, V. R. (2012). *Tea in health and disease prevention*. Academic Press.
- R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org>
- SAS Institute Inc. (1989-2007). *JMP*[®] (Version 13). SAS Institute Inc., Cary, NC.
- Savary, S., Willocquet, L., Elazegui, F. A., Teng, P. S., Van Du, P., Zhu, D., ... Srivastava, R. K. (2000). Rice pest constraints in tropical Asia: Characterization of injury profiles in relation to production situations. *Plant Disease*, 84(3), 341-356. <https://doi.org/10.1094/PDIS.2000.84.3.357>
- Willson, K. C., & Clifford, M. N. (2012). *Tea: Cultivation to consumption*. Springer Science & Business Media.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).

Appendix 2.2: The meaning of the mean diamonds and x-axis proportional lines used in Chapter 2 statistical analysis

The mean diamonds illustrate the sample mean and confidence intervals. Figure 1 shows examples of mean diamonds and x-axis proportional options

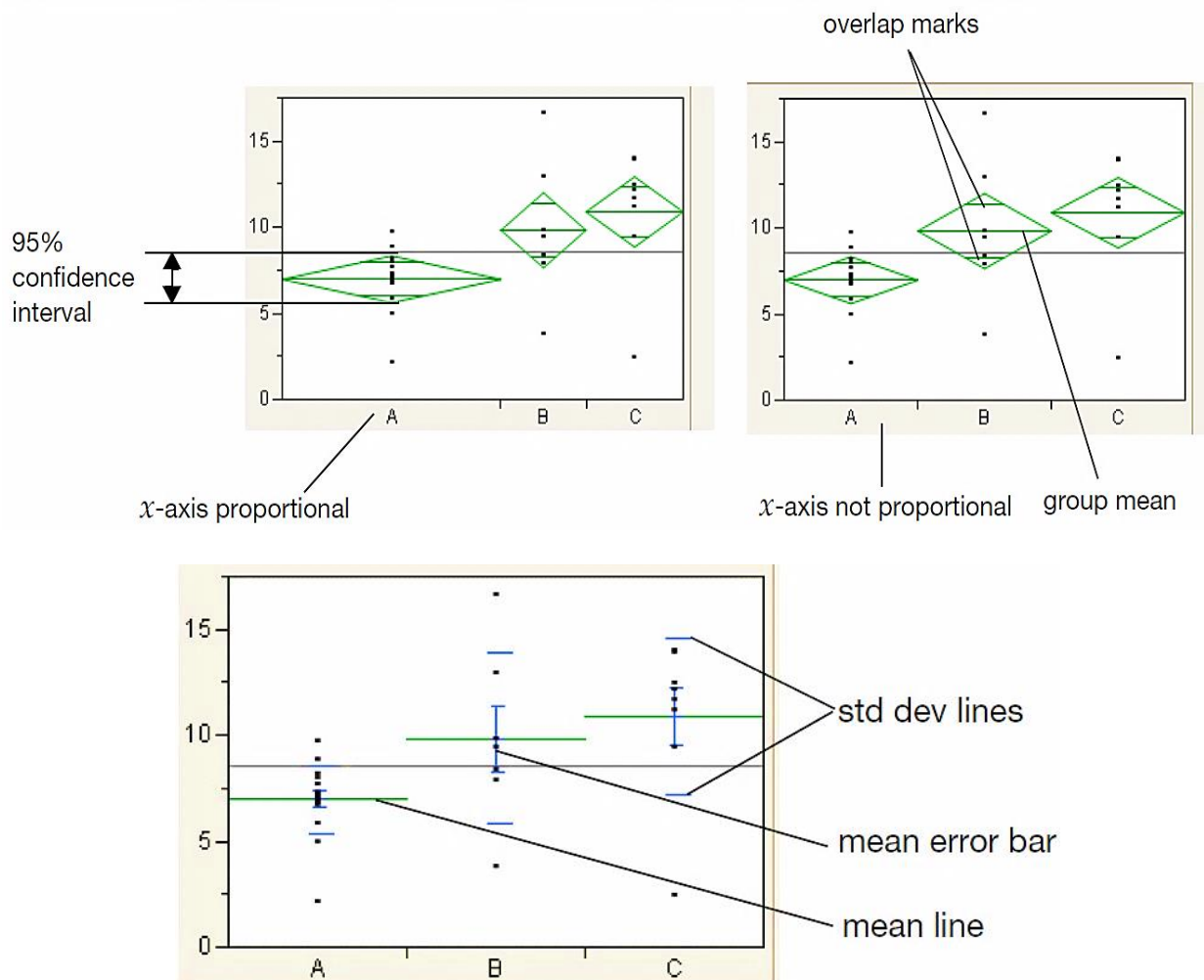


Figure 1: Mean diamonds and x-axis proportional options (JMP®).

The top and bottom of each diamond is representative of each group's $(1-\alpha) \times 100$ CI. The CI calculation/determination assumes the variances to be equal across all the observations, and as such, each diamond's height is proportional to the reciprocal of the square root of each group's total observations. The mean line across the centre of each diamond is representative of each group's mean. In instances where overlaps occur, the lines representing these appear above and below the mean. When groups have equal sample sizes, the overlap marks show that the group means are not significantly different at that particular CI. If one group's

overlap marks are close to another group's mean, this shows that the two groups do not differ at that CI.

CHAPTER 3

IDENTIFICATION OF QTL'S RESPONSIBLE FOR YIELD, DROUGHT TOLERANCE AND QUALITY TRAITS IN CAMELLIA SINENSIS USING GC-MS, ¹H-NMR AND UPLC

ABSTRACT

Camellia sinensis (tea) is one of the most consumed beverages worldwide, and its consumption has, in recent years, increased due to its health-promoting attributes, brought about by its high polyphenol content. In-depth studies have been conducted on *C. sinensis* due to its precise flavonoid profile, responsible for conferring copious therapeutic properties. According to Nyarukowa *et al.*, (2018), approximately 78% of black, 20% of green and 2% of oolong tea by value are globally traded. Green tea, rich in catechins, serves as a traditional herbal remedy in China to prevent cardiovascular diseases among other chronic diseases; this explains why green tea has been substantially studied i.e. because of its antioxidant and anti-carcinogenic properties. Tea production and subsequently its quality are reliant on evenly distributed rainfall. Tea consumers concern themselves with the quality of tea, in particular its flavour and aroma; it is on the basis of these that consumers are willing to pay premium prices for the best quality teas. To breed for these phenotypic traits is however challenging due to the fact that these are qualitative traits inherited from parents, and influenced by environment. Two *C. sinensis* populations (TRFK St. 504 and TRFK St. 524) were employed in this study to identify the Quantitative Trait Loci (QTL) responsible for yield, quality and total polyphenol content centred on a genetic map constructed using the DArTseq platform; populations of 106 TRFK St. 504 and 144 TRFK St. 524 clonal progeny were investigated; map comprised 15 linkage groups analogous to chromosome haploid number of tea plant ($2n = 2x = 30$) and spanned 1260.1 cM with a mean interval of 1.1 cM between markers. Sixteen phenotypic traits were evaluated in both populations. A total of three, 11 and 46 putative QTLs were discovered after mapping on the 15 linkage groups, responsible for tea quality for GC-MS, ¹H-NMR and UPLC data respectively. The variance explained by the QTLs varied from 4.6 to 96.3%, with an average of 28%.

3.1 INTRODUCTION

Tea has been documented to be the most consumed non-alcoholic beverage, worldwide, second to water. According to Dutta *et al.*, (2011), tea production has up scaled from 850 000 tonnes in 2003 to 980 000 tonnes 2007 (Dutta *et al.*, 2011) to 2,414,802 tonnes in 2018 (www.reportlinker.com/tea/reports). Green *C. sinsensis* leaves comprise of predominantly five flavan-3-ols, namely catechin (CAT), epicatechin (EC), epicatechin gallate (ECg), epigallocatechin (EGC), and epigallocatechin gallate (EGCg) (Koech *et al.*, 2018), with EGCg being the principal catechin accounting for approximately 50–80% of the total catechins (Sang *et al.*, 2011). Black tea on the other hand consists of theaflavins and thearubigins as its major polyphenols, resulting from catechin oxidation and polymerisation. Black tea consists of theaflavin (TF1), theaflavin-3-monogallate (TF2), theaflavin-3'-monogallate (TF3), and theaflavin-3,3'- digallate (TF4). Tea's popularity as a beverage is dependent on its flavour, comprising of taste and aroma, with non-volatile organic compounds being responsible for its taste, and the volatile organic compounds responsible for its aroma. Volatile organic compounds in tea fall into one of two groups, with Group I comprising of non-terpenoids such as hexenols, which confer the fresh green flavour, and Group II comprising of terpenoids, responsible for its sweet flowery aroma; high quality teas are rich in Group II compounds and due to their flavoury nature, sell for 4 - 5 times higher (Rawat *et al.*, 2007). According to (Le Gall *et al.*, 2004), the taste of green tea is determined by the type of tea tree, its plucking time, as well as the method of cultivation employed. Amino acids such as theanine, which make up between 60-70% of the amino acids found in tea leaves; these are responsible for tea's brothy taste, while its astringent taste can be attributed to catechin levels and lastly its bitter taste attributed to caffeine (Pongsuwan *et al.*, 2007). Tea quality is undeniably affected by variations in its chemical compositions, which determine its commercial market value (Qin *et al.*, 2013). Sensory evaluation of tea quality by trained specialists, tea tasters, has traditionally been employed to establish its specific aroma profile, which is important in determining the tea quality grade (Schuh and Schieberle, 2006). These trained tasters have developed a language of their own which they use to describe various attributes of a tea infusion; this sensory evaluation has its own deficiencies, such as time consuming, human susceptibility and variability (Group *et al.*, 2011). Due to its increased popularity, tea has enthralled both consumers and researchers, not only because of its taste and aroma but because of medicinal benefits concomitant with tea, owed largely to its bioactive metabolites i.e. flavonoids, with catechins constituting up to 30% of soft-shoot

dry weight in green tea. The catechins and theaflavins in green and black tea respectively have been documented to possess antioxidant, anti-inflammatory, anticancer, as well as cardiovascular disease preventing capabilities (Preedy, 2012).

The profiling of plant metabolites has developed into a major metabolomics field of study, reason being that plants manufacture a wide array of metabolites. Through the use of metabolomics, plant materials have and continue to be classified, with the predominant statistical tool being the principal component analysis (PCA) (Pongsuwan *et al.*, 2007). PCA is the first step in multivariate statistical analysis, which is highly expedient when it comes to e.g. outlier identification, pattern and trend detection. The partial least squares-discriminant analysis (PLS-DA) is another multivariate statistical approach employed in metabolomics data analysis. The PLS-DA is a better suited statistical approach as compared to the PCA, when it comes to differentiation in e.g. the origins of particular samples, especially in instances where the metabolite profiles are influenced/ affected by several factors (Kang *et al.*, 2008). The genetic enhancement of crops is fast becoming a highly employed, continuous practice; this has resulted in an increased demand for plant breeders skilled in metabolomics (Chugh, 2013). When developing novel cultivars, crop breeders encounter a common challenge of ascertaining their selection criteria; *C. sinensis* breeders for example have to select from a list of properties which include, but are not restricted to yield, quality, drought tolerance. Due to the effects of global warming, which are seeing altered precipitation patterns, elevated temperatures and protracted drought spells in the tea growing regions, the Kenyan tea industry has been facing strenuous challenges (FAO 2015). It is for this reason that rigorous breeding programmes need to be developed to produce novel cultivars with better metabolic profiles and improved drought tolerance. A previous study on drought tolerance in *C. sinensis* saw the development of the Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) method which was validated by targeted metabolomics to predict tolerance in tea cultivars by generating metabolic profiles which showed the differences between the drought-tolerant and drought susceptible cultivars under wet conditions; this method employed the %RWC of tea leaves after a five-hour withering period (Nyarukowa *et al.*, 2016). The advancement of genomics-assisted breeding techniques such as Marker Assisted Selection (MAS) has resulted in the significant advancement and refinement of breeding selection precision and efficacy (Chen *et al.*, 2013). The trait selection criterion is a multifaceted one; phenotypic assessments are incapable of denoting the degree of variation with respect to a trait of interest due to environmental effects (Feil and Fraga,

2012). Quantitative genetics, the study of genetic interactions and environment on e.g. *C. sinensis* is capable of empowering breeders with information regarding the quantity of transmissible genetic variation in the available traits for selection (El-Soda *et al.*, 2014). Quantitative genetics also indicates/highlights correlations by furnishing breeders with a comprehension of any genetic relationships existent between the traits directly and or indirectly influencing the crops phenotype. A comprehension of genetic correlations serves to enable the identification of potential markers in instances where the direct measurement of traits is either strenuous or expensive, thus circumventing the selection of unrelated traits (Mackay *et al.*, 2009). According to El-Soda *et al.*, (2014), quantitative genetic analysis also governs the degree to which trait variation and correlations are influenced by environmental factors, ensuring accurate predictions of genetic improvements and enhancing breeding strategy development. Furthermore, in *C. sinensis*, quantitative genetic analysis serves the added advantage of being employed as a means of determining the hereditary potential, in offspring trials, for traits of interest found in either male or female plants e.g. those concomitant with yield, quality, and drought tolerance (Kamunya *et al.*, 2009); this will serve to simplify tea breeding through selection of parents with desirable traits to yield enhanced progeny (Yao *et al.*, 2008).

Genetic linkage maps possess the potential to substantially increase the rate and accuracy of cultivar development strategies for perennial woody tree crops (Hackett *et al.*, 2000). *C. sinensis* is a woody, perennial tree, often having long juvenile periods of 4–5 years, taking approximately 25 years to breed a novel cultivar (Wang *et al.*, 2016). It is characterised by a huge diploid genome of about 3Gb and chromosome number ($2n = 2 \times = 30$) that is self-incompatible, highly heterozygous, and principally allogamous (Koech *et al.*, 2018). It is because of these factors that conventional breeding is both time consuming and laborious, requiring a lot of land for offspring trials (Orel and Wilson, 2012); genetic mapping does away with this, making early selection amenable. This means the selection of parents based on molecular markers of interest linked to Quantitative trait loci (QTL) shortens the breeding cycle. A genetic linkage map was constructed by employing a two-way pseudo-testcross strategy to map the drought-tolerant attribute of tea cultivars originating from (Bali *et al.*, 2015). Despite these advancements, the use of genetic linkage maps is predominantly centred on both prevalent markers like random amplified polymorphic DNA (Hackett *et al.*, 2000) and co-dominant markers like the simple sequence repeats (Tan *et al.*, 2016). Unfortunately these technologies are not extensively pertinent tools capable of correlating genotypes with

phenotypes. There has been an increase, in recent years, in the amount of research conducted on Diversity Arrays Technology (DArT); this technology entails isolating and cloning indiscriminate DNA fragments from complexity-reduced DNA samples (Gupta *et al.*, 2013); because of DArT, standardised genotyping of thousands of markers in parallel in multiple samples without the need for prior sequencing information (Wittenberg *et al.*, 2005) and has been employed in phylogenetic and diversity studies (Steane *et al.*, 2011), genomic selection (Poland *et al.*, 2012), and in the construction of genetic linkage mapping (Koech *et al.*, 2018) in plants such as apples (Schouten *et al.*, 2012), wheat (Zou *et al.*, 2017), and Eucalyptus (Steane *et al.*, 2011). It is on the basis of the increasing research that genetic linkage maps generated from the DArTseq platform can be considered vital in the relation of a particular phenotype to genotype, simplifying the process with which breeders identify and select parents with the sort after traits. In the current study, a high-density linkage map for *C. sinensis* was constructed through the integration of the DArTseq technology, GC-MS, ¹H-NMR, and UPLC techniques for QTLs linked to amino acids, sugars, catechins, caffeine, tea taster scores, and %RWC for future MAS breeding.

3.2 RESEARCH OBJECTIVE

This study's objective was the identification of putative QTLs associated with amino acids, caffeine, catechins, organoleptic evaluation and %RWC DArTseq marker integration with GC-MS, ¹H-NMR and UPLC platforms, constructing genetic linkage maps for MAS in tea breeding.

3.3 HYPOTHESIS

Null hypothesis (H_0): The DArTseq markers are not linked to QTLs for black tea quality parameters and yield.

3.4 MATERIALS AND METHODS

3.4.1 Plant material

All the cultivars used in this study were maintained using uniform agronomic practices in an experimental field site in Kericho (0° 22' S, 35° 17' E), Kenya as described by Koech et al. (2018). Sixty open-pollinated cultivars, pre-selected for their high yield, and good tea liquor since the 1950s formed the Commercial (Comm) group. These cultivars were vegetatively propagated by stem cuttings from elite mother bushes. Each Comm cultivar is cultivated in over 10 Hectares with 13,448 bushes per Ha. The NonCommercial (NComm) group of 250 cultivars were the F₁ progeny of a reciprocal cross between two heterozygous parental clones TRFK 303/577 and GW Ejulu. The NComm cultivars comprise two populations of TRFK St. 504 (TRFK 303/577 x GW Ejulu) with 106 progeny, and the TRFK St. 524 (GW Ejulu x TRFK 303/577) with 144 progeny, which were bred at the Tea Research Institute (TRI) of Kenya (Koech et al., 2018). The GW Ejulu is a commercial cultivar that produces high-quality black tea, with high total catechins and moderate caffeine content; it is, however, a low-yielding and drought-susceptible clone. Cultivar TRFK 303/577, on the other hand, is a high yielding, drought tolerant (DT) commercial cultivar, which produces medium-quality black tea, with moderate levels of caffeine and total catechins. All the plants were vegetatively propagated and planted in 15-bush observation plots comprising 3 rows and 5 plants per row spaced at 1.22 m between rows and 0.61 m within rows (i.e. 13,448 plants/ha) in a randomised complete block design with three replicates.

3.4.2 Sample collection and processing

Fresh shoots comprising two leaves and a bud were randomly harvested from the respective tea bushes and placed in zip-lock plastic bags, appropriately labelled (Nyarukowa *et al.*, 2018) to be processed at the TRI factory. Half the shoots of each sample were freeze-dried and ground using a coffee grinder, sieved using a 355 µm sieve, sealed in zip-lock plastic bags and stored at 4°C in a fridge until analysis. The other half was used to make black tea according to Koech *et al.*, (2018). Briefly, the leaves were withered to a %relative water content of 50–65% over an 18 hour period before being passed through CTC rollers till maceration was achieved. Following maceration, the resultant dhoor 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.

3.5 GC-MS sample preparation and analysis

3.5.1 Sample preparation

A 70% MeOH solution was used for extraction. For all samples of approximately 150 mg, 1.5 mL extraction solution was added. The samples were vortex mixed and incubated for 10 minutes at 70°C. The samples were vortex mixed halfway through the incubation period as well as at the end. After cooling, the samples were centrifuged for 5 minutes at 6000 g and the 1 mL supernatant transferred to GC vials before drying under nitrogen. The dried samples were derivatised by adding 120 µl methoxyamine (10 mg/mL in pyridine) and incubated for 1 hour at 60°C; followed by the addition of 80 µl BSTFA (containing 1% trimethylchlorosilane) and incubated for another hour at 60°C. Samples were transferred to inserts before GC-MS analysis. Pooled quality control (QC) samples were prepared, and these underwent the same extraction and derivatisation procedures as the samples. Samples were randomly injected with QC samples analysed after every 10th sample i.e. QC1 followed by samples 1 to 10 then QC2 followed by samples 11 to 20 etc. Two additional QCs were analysed in the beginning of each batch to condition the new liner. These QCs were not used in data processing.

3.5.2 GC-MS analyses

Analyses were performed on a GC-TOF-MS system, comprising of an Agilent 7890A GC front-end system with an Agilent 7693 autosampler and a Leco Pegasus HT TOFMS. Hydrogen was used as carrier gas at a flow-rate of 1.8 mL/min; 0.2 µl sample was injected in splitless mode (allowing 30s purge delay). The inlet temperature was kept at 250°C. Compounds were separated on a Restek RX-1MS column (20 m x 180 µm x 0.18 µm). The transfer line and source temperatures were 250 and 200°C, respectively. Solvent delays of 200 s were allowed where after masses (50 – 800 m/z) were recorded at 20 spectra/sec. Universal EI settings were used for ionisation while the detector was operated at 50 V above tune voltage.

3.6 ¹H-NMR sample preparation and analysis

3.6.1 ¹H-NMR buffer solution

A 1.5 M KH₂PO₄ buffer solution was prepared by dissolving 20.4 g of KH₂PO₄ in 80 mL of deuterium oxide (D₂O). Next, 13 mg of sodium azide and 100 mg of trimethylsilyl-2,2,3,3-tetradeuteriopropionic acid (TSP) were dissolved in 10 mL of D₂O and added to KH₂PO₄ solution. The combined solution was mixed well under sonication before adjusting the pH to 7.4 using potassium hydroxide in H₂O. The final solution was then transferred to a 100 mL volumetric flask and the volume topped up to the mark using D₂O.

3.6.2 ¹H-NMR sample preparation

Freeze-dried samples were sent in individual plastic bags of 50 mg weight to the ¹H-NMR lab. A pooled QC sample was created by collecting 5 mg from each of n=294 samples. Samples were prepared by adding 4.5 mL ddH₂O to each 45 mg weight of the dry sample to create a 10 mg/mL concentration. Each sample was vortexed at 0, 20 and 40 minutes. At 60 minutes, a volume of 540 μL of the sample was collected in a microcentrifuge tube, with 60 μL ¹H-NMR buffer solution. The sample was mixed under vortex and centrifuged at 12 000 g for 5 minutes to sediment any particulates. A final volume of 540 μL of supernatant was carefully transferred to a 5 mm ¹H-NMR glass tube and loaded onto an autosampler for ¹H-NMR analysis.

3.6.3 ¹H-NMR analyses

The samples were measured at 500 MHz on a Bruker Avance III HD NMR spectrometer equipped with a triple-resonance inverse (TXI) ¹H{¹⁵N, ¹³C} probe head and x, y, z gradient coils. ¹H spectra were acquired as 128 transients in 64 K data points with a receiver gain of 64 and a spectral width of 10 000 Hz. The sample temperature was maintained at 300K and the H₂O resonance was presaturated by single-frequency irradiation during a relaxation delay of 4 s, with a 90° excitation pulse of 8 μs. Shimming of the sample was performed automatically on the deuterium signal. The resonance line widths for TSP and metabolites were <1 Hz. Fourier transformation and phase and baseline correction were done automatically. Software used for ¹H-NMR processing was Bruker Topspin (V3.5). Bruker AMIX (V3.9.14) was used for metabolite identification and quantification. (Ellinger et al., 2013).

3.7 UPLC sample preparation and analysis

3.7.1 Extraction of catechins, caffeine and theaflavins

Samples were collected as documented in 3.4.2. The International Organisation for Standardisation (ISO) extraction procedure described in document ISO14502-2 (2005) was employed for the extraction of metabolites from the tea samples. Concisely, amounts of 0.200 ± 0.001 g of green and black tea samples were weighed out using a Mettler Toledo model XS205DU 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 were added to each, stoppered and vortex mixed for \pm 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 2000 x g using Thermo Scientific Heraeus Labofuge (Sepsci, South Africa) 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, 500 mg/mL EDTA and 10 mg/ml ascorbic acid, all purchased from Sigma-Aldrich, South Africa. Each resultant dilution was then filtered through a 0.2 μ m Minisart®RC4 syringe filter (Sartorius, South Africa) with hydrophilic, solvent-resistant regenerated cellulose membranes and the samples were then analysed using UPLC-DAD.

3.7.2 UPLC analyses

The UPLC 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 μ m, 2.1 \times 150 mm), 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. 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% B). The mobile phases were filtered through a 0.2 μ m cellulose acetate membrane filter and degassed using a Neuberger

Laboport (Labotech, South Africa) vacuum pump. A sample injection volume of five μL , and a 0.2 mL/min flow-rate were employed for analyses. Catechins (catechin, epicatechin, epicatechin gallate, epigallocatechin and epigallocatechin gallate), caffeine and gallic acid (Sigma-Aldrich, South Africa) were used as standards. Tryptamine, sulfanilamide and mycophenolic acid (Sigma-Aldrich, South Africa) were used as the QC internal standards; identification and quantification was at 278 nm, with the individual catechins and caffeine in the samples being identified on retention times of the standards.

3.8 Metabolite identification

Spectral matching to the NIST11 commercial library (for GC-MS metabolites) and Bruker BBIORFCODE (pH 7.0) and in-house pure compound spectral libraries (pH 7.4) (for ^1H -NMR metabolites) were used to identify the compounds. A level 2 identity was awarded when a spectral match of 80% similarity was achieved. A level 1 identity was awarded when the retention time or retention index of the GC-MS information matched that of standards (Schymanski et al., 2014) or 2D ^1H -NMR information confirmed 1D ^1H -NMR spectral identifications.

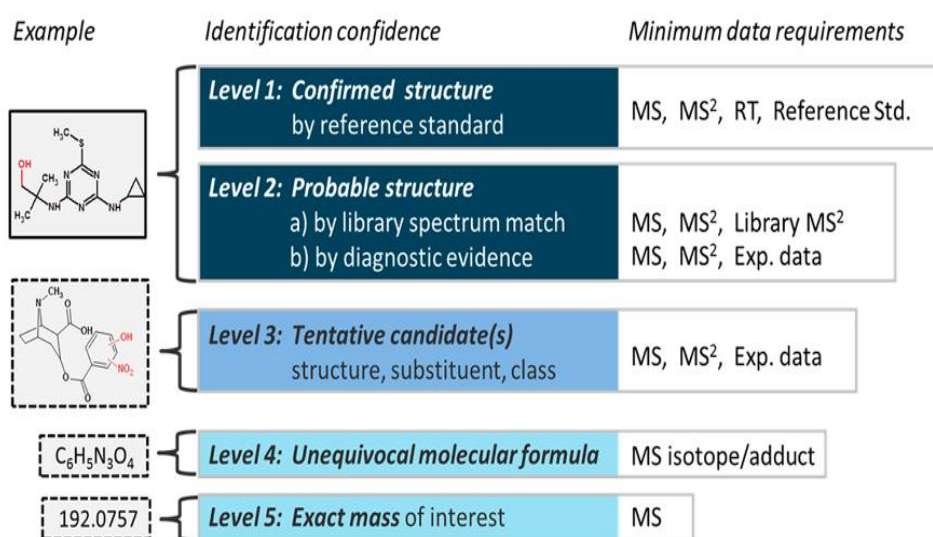


Figure 3.1: Proposed identification confidence levels in high resolution MS analysis. MS² is intended to also represent any form of MS fragmentation (e.g. MS^e, MSⁿ) (Schymanski *et al.*, 2014).

3.9 Determination of tea quality

3.9.1 DNA extraction and quantification

Fresh shoots comprising two leaves and a bud were randomly harvested from the parents, the Comm and NComm cultivars, and placed in zip-lock plastic bags containing dry silica gel, which served to absorb any surface moisture on the leaves; the silica gel had been oven dried for 48 hours at 70°C. Following this, the leaf samples were stored at -20°C prior to DNA extraction; an adapted (Gawel, 1991) method was employed. To quantify the amount of DNA in ng/μl for each sample, the NanoDrop spectrophotometer (NanoDropTechnologies, South Africa) was employed; the DNA integrity was ascertained using agarose gel electrophoresis (Adkins *et al.*, 2007).

3.9.2 DArTseq assay

The DNA samples were sent to the Diversity Array Technology Pty Ltd. in Canberra, Australia where DArTseq analysis was performed; for DNA quality and digestibility analysis, restriction enzyme PstI were obtained from Fermentas, Burlington, Canada and EcoRI from Promega, Madison, USA. The DArTseq technique was conducted as documented in (Sansaloni *et al.*, 2010) using PstI and MseI restriction enzymes. The markers were scored either 0 or 1, representing either the absence or the presence of a polymorphism in each samples genomic representation.

3.9.3 Construction of linkage map

A total of 6 588 DArTseq markers were obtained from the *C. sinensis* sequences, and employing the genomic DNA from TRFK 303/577 and GW Ejulu, and the 250 NComm cultivars, these were tested for segregation. JoinMap 4.0 software was then employed to analyse the derived genotyping data (Van Ooijen, 2006); an odds of logarithm (LOD) of 3–12 and a recombination frequency of 0.4 was used to group the markers, with the distances between markers being ascertained using the Kosambi mapping function. To establish the LOD thresholds at the genome-wide level, 1000 permutations were run, with a p-value of $p < 0.05$.

3.9.4 QTL analyses

To map QTL's, MapQTL 6.0 software was employed with the first set of markers being chosen as cofactors from interval mapping results, while noteworthy markers were selected through backward exclusion. According to (Churchill and Doerge, 1994), 1000 permutations

are required to obtain a significant threshold to accurately accept the existence of a prospective QTL. The maximum LOD values score served to ascertain the position of QTL's on each linkage group. The MapChart software was employed to indicate the location of each QTL for each phenotypic trait.

3.9.5 Fast Adaptive Shrinkage Thresholding Algorithm (FASTA) files preparation

The DArTseq marker sequences were derived from tag sequences linked to each DArTseq marker as documented by Koech *et al.*, (2018). These were produced by Diversity Arrays Technology Pty. Ltd. (Canberra, Australia). FASTA files were then prepared by beginning with single line marker sequence descriptions, which were then followed by the sequence data. Each single line description was differentiated from sequence data through the use of ">" before the description.

3.9.6 Basic Local Alignment Search Tool (BLAST) search

All 1 421 DArTseq markers on LGs 1 to 15 were ran through a BLAST search, with marker tag sequences obtained from the DArTseq map being searched against the tea genome, using the BLASTN program. The %identity and E-value was used to select the best hit. The E-value functioned as a determinant for the number of possible hits one may expect to obtain when searching a database with small E-values designating homology. The %identity indicated the proportion of residues, which were identical between the query and database hit sequences, with lengthy expanses of homology indicating a genuine match.

3.9.7 Functional annotation and pathway assignment

A functional annotation of the 1 421 DArTseq markers was conducted using the BLASTX search against the GenBank protein sequence database, with the E-value threshold set to 10^{-6} (Conesa and Götze, 2008). The Blast2GO program version 3.2 was used for annotating and mapping GO terms. The GO terms linked to each BLAST hit were retrieved, and GO annotation assignment to the query sequences were conducted using the subsequent annotation score parameters; E-value Hit Filter (default=1.0E-6), Annotation Cut-Off (90), GO-Weight (default=5), Hsp-Hit Coverage Cut-Off (default=0). InterProScan, Blast2GO program online sequence search plugin, was used to query the contig sequences for conserved domains/motifs; all 13 applications were selected prior to the run. The resultant GO terms were then combined with the GO terms obtained from the Blast2GO annotation step. Lastly, KEGG mapping was used to ascertain metabolic pathways, with the Blast2GO sequences with matching evidence code (EC) numbers being mapped to the KEGG database.

3.10 RESULTS

3.10.1 Genetic map construction

The regression-based integrated map, which was generated, gave 15 linkage groups, with Figures 3.2, 3.3 and 3.4 showing the important groups; Tables 3.2, 3.3, 3.4 and 3.5 show corresponding information. The linkage map spanned a genetic distance of 1 260.1 cM, averaging a genetic marker locus distance of 1.1 cM. Each linkage group differed in size from the rest, with the smallest being LG 2 spanning 64.8 cM, and the largest, LG 9, spanning 160.1 cM. The number of markers on each LG ranged from 50 to 219 on LGs 1 and 9 respectively, as shown in Table 3.1, with LG 9 having the highest marker density, averaging a locus distance of 1.4 cM; the rest of the LGs ranked lower with a marker density of 0.5-1.7 cM. DArT markers associated with the phenotypic traits of interest were detected on LG 14 generated from the GC-MS, LGs 1, 2, 5, 7, 11, 13 and 14 generated from the ¹H-NMR data, and LGs 1, 2, 4, 6, 7, 8, 9, 10, 12, 13, 14 and 15 generated from the UPLC data. The markers were clustered differently, with some LGs significantly less densely populated with markers than others.

3.10.2 Phenotype segregation and QTL mapping

The analysis of the GC-MS data revealed three putative QTLs; the ¹H-NMR data revealed 11 putative QTLs, and the UPLC data revealed 46 putative QTLs, within the 15 LGs linked with tea liquor quality through the use of Interval Mapping and Multiple QTL Model at a genome-wide significance confidence level threshold of 5%. The logarithm of odds ratio (LOD) scores varied from 3.0 to 3.1 on the GC-MS data; 3.0 to 3.5 for the ¹H-NMR QTLs, and 3.0 to 4.1 on the UPLC data (Tables 3.2, 3.3 and 3.4). The phenotypic variation explained (PVE) expressed as a percentage by all QTLs for the traits of interest ranged between 5.1(catechin) to 96.3% (isoleucine) for the ¹H-NMR QTLs; 6.1 (colour) to 97% (EGC) for the UPLC QTLs, and 4.6 (phloroglucinol) to 7.5% (xylic acid) for the GC-MS QTL's. Four metabolites, namely caffeine, catechin, EC and EGC were detected by both ¹H-NMR and UPLC platforms. These were successfully mapped; it was however noted that they appeared on different LG, at different position and their %PVE differed. This is shown in Table 3.5.

Table 3.1: DArT markers distribution among the linkage groups.

Linkage group	Number of markers	Total length covered (cM)	Average distance between markers (cM)
LG1	50	97.3	0.5
LG2	107	64.8	1.7
LG3	61	82.4	0.7
LG4	98	70.1	1.4
LG5	77	77.0	1.0
LG6	65	82.6	0.8
LG7	139	83.2	1.7
LG8	100	68.3	1.5
LG9	219	160.1	1.4
LG10	114	81.9	1.4
LG11	67	80.7	0.8
LG12	83	77.5	1.1
LG13	68	78.5	0.9
LG14	83	79.2	1.0
LG15	90	76.5	1.2
Total	1421		1260.1
Average		84	1.1

Table 3.2: GC-MS QTLs for arabinose, phloroglucinol, and xylonic acid identified in TRFK St 504 and TRFK St 524 tea cultivar samples.

Traits	QTL	LG	Position (cM)	LOD threshold	PVE (%)	Marker
Arabinose	qArabinose	14	9.083 (9.083 - 9.367)	3.1	5.7	5125565
Phloroglucinol	qPhloroglucinol	14	30.666	3.1	4.6	5136609
Xylonic Acid	qXylonicAcid	14	64.378 (64.378 – 65.383)	3.0	7.5	5075568

LG – Linkage group

LOD – Logarithm of odds ratio

PVE – Phenotypic variation explained

QTL – Quantitative trait loci

The LOD thresholds are determined by $P < 0.05$, and the basis of which, was permutation testing with $n = 1\ 000$.

LG 14

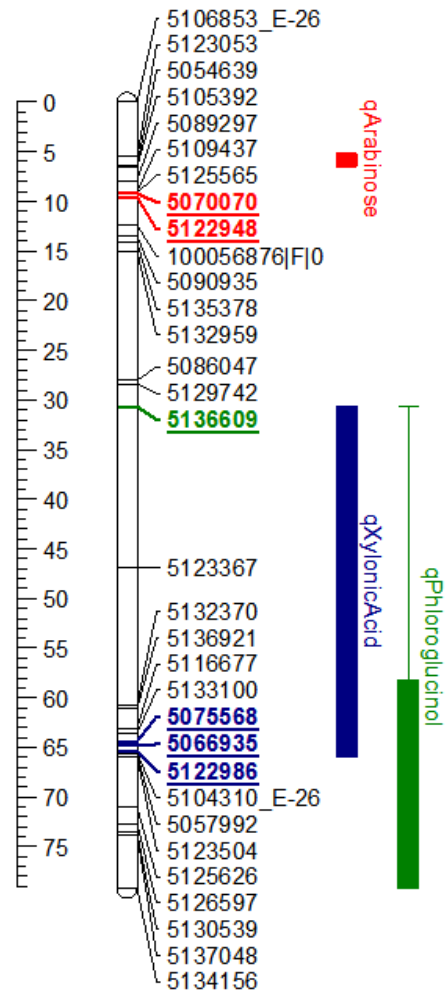


Figure 3.2: Genetic map of *C. sinensis*, displaying GC-MS QTL locations for arabinose, phloroglucinol, and xylonic acid. The map ruler is scaled in cM. Each detected QTL e.g. caffeine, catechins are represented by different coloured bars and lines, which are indicative of 1-LOD and 2-LOD support intervals.

Table 3.3: NMR QTLs for caffeine, catechins, epicatechin, valine, isoleucine, chlorogenic acid, and acetic acid identified in TRFK St 504 and TRFK St 524 tea cultivar samples.

Traits	QTL	LG	Position (cM)	LOD threshold	PVE (%)	Marker
Acetic Acid	qAcetic Acid	14	8.993 (8.993 - 16.426)	3.1	25.7	5109437
Caffeine	qCaffeine	1	37.158 (35.049- 37.158)	3.1	6.6	5109590
Catechin	qCatechin	1	0 (0 - 7.000)	3.0	5.1	5115373
Chlorogenic Acid	qChlorogenic Acid	11	16.971	3.3	6.3	5085772
Epicatechin	qEpicatechin	5	18.151 (11.749 - 18.151)	3.0	18.5	5132307
Epigallocatechin	qEpigallocatechin	13	64.77 (64.77 - 73.899)	3.1	5.4	5133837
Isoleucine	qIsoleucine	7	62.78	3.5	7.5	5120311
		13	29.748 (29.748 - 36.683)	3.3	66.3	5070055
Valine	qValine	2	8.309 (6.818 - 8.309)	3.2	42.6	5123739
		13	19.195 (19.195 - 36.538)	3.1	8.1	5016516
		14	0 (0 - 5.465)	3.0	14.4	5106853_E-26

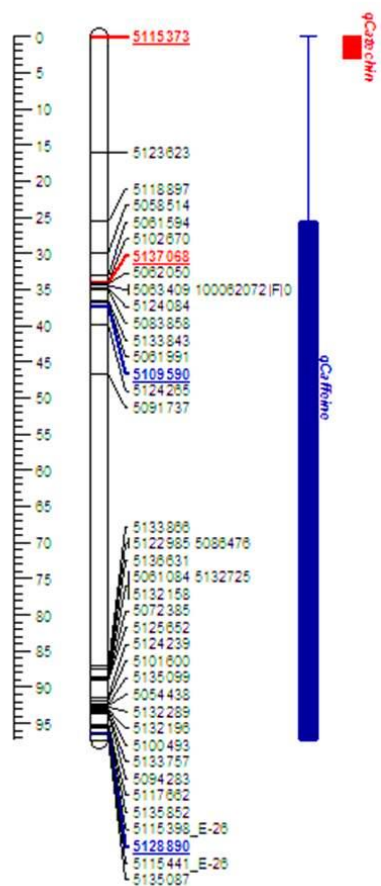
LG – Linkage group

LOD – Logarithm of odds ratio

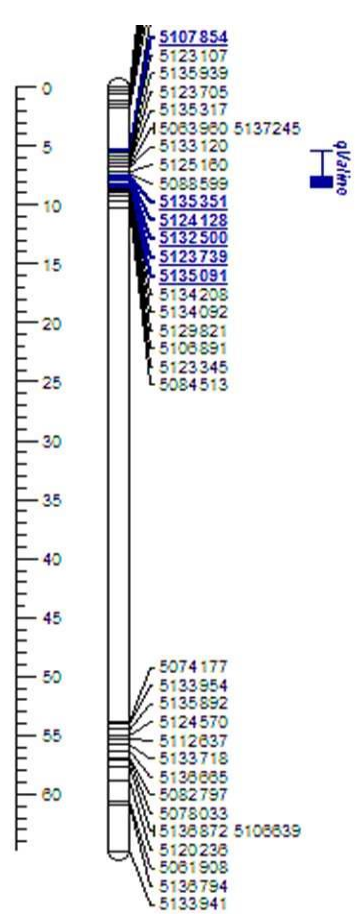
PVE – Phenotypic variation explained

QTL – Quantitative trait loci

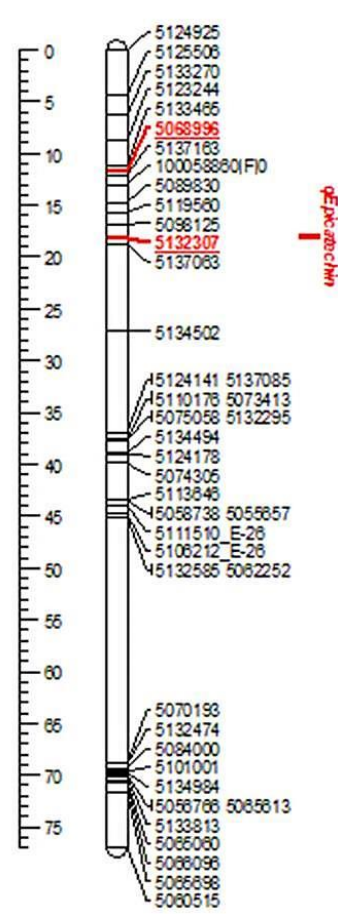
The LOD thresholds are determined by $P < 0.05$, and the basis of which, was permutation testing with $n = 1\ 000$.



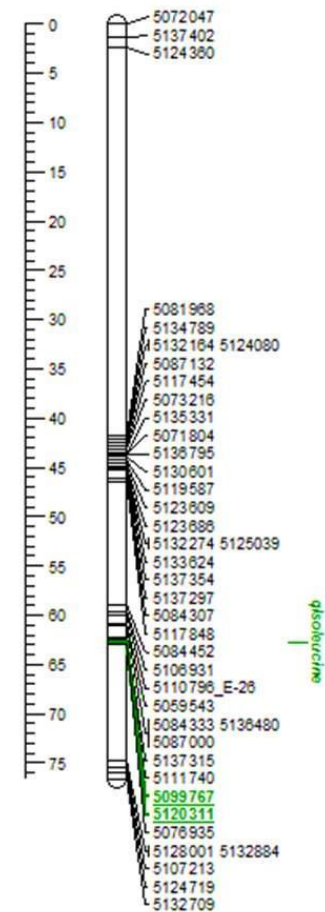
LG 1



LG 2



LG 5



LG 7

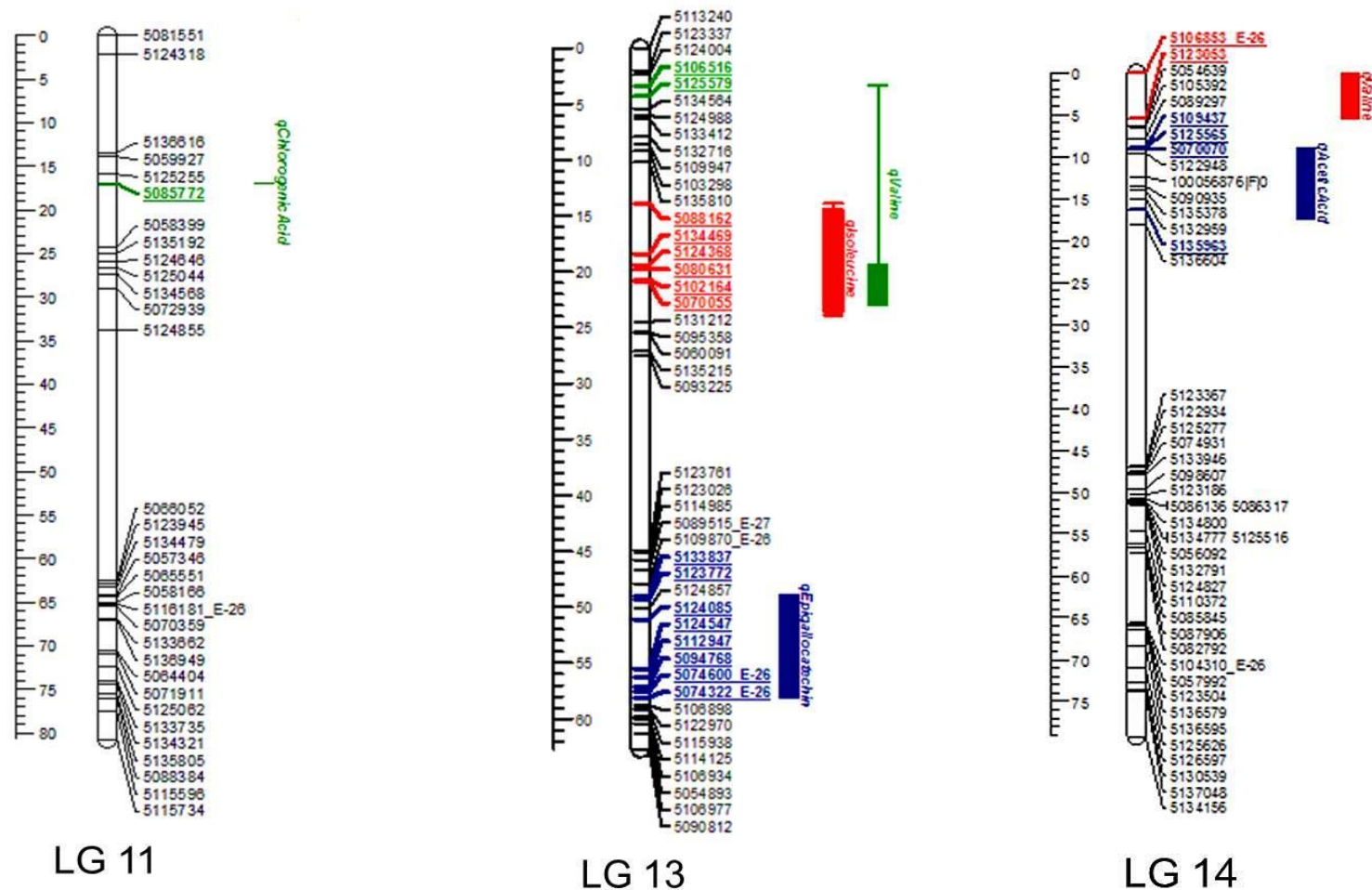


Figure 3.3: Genetic map of *C. sinensis*, displaying ¹H-NMR QTL locations for Acetic acid, Caffeine, Catechin, Chlorogenic acid, Epicatechin, Epigallocatechin, Isoleucine and Valine. The map ruler is scaled in cM. Each detected QTL e.g. caffeine, valine are represented by different coloured bars and lines, which are indicative of 1-LOD and 2-LOD support intervals.

Table 3.4: UPLC QTLs for caffeine, catechins, epicatechin, epicatechin gallate, epigallocatechin, epigallocatechin gallate, theaflavins, tea taster's scores and the % RWC identified in TRFK St 504 and TRFK St 524 tea cultivar samples.

Traits	QTL	LG	Position (cM)	LOD threshold	PVE (%)	Marker
Caffeine	qCaffeine	2	50.9 (50.2 - 51.5)	3.3	6.0	5064585
		4	68.6 (63.1 - 70.1)	3.2	6.7	5112599
		7	48.1 (34.7 - 50.4)	3.2	6.6	5064391
		8	18.8 (16.8 - 26.0)	3.3	6.7	5134558
		13	29.7 (23.7 - 36.5)	3.1	8.1	5088162
		14	5.465 (0 - 5.465)	3.3	6.4	5123053
Catechin	qCAT	2	0 (0 - 2.4)	3.2	6.4	5135436
		4	30.3 (29.9 - 30.3)	3.3	55.6	5063001
		8	12.5 (12.5 - 12.9)	3.3	5.6	5130194
		12	42.9 (36.0 - 54.1)	3.1	10.3	5123751
		13	50.6 (48.6 - 58.7)	3.0	6.9	5111268
		14	60.7 (47.7 - 61.1)	3.2	6.9	5132370
Epicatechin	qEC	2	2.2 (0 - 8.3)	3.3	7.0	5072338
		15	25.4 (20.6 - 25.4)	3.3	7.6	5085963
Epicatechin gallate	qECg	1	96.4 (96.4 - 97.3)	3.0	11.7	5128890
		4	17.7 (13.9 - 37.2)	3.4	8.0	5087113
		6	56.8 (56.5 - 56.9)	3.1	6.7	5098382
		10	20.6 (20.1 - 25.5)	3.3	8.0	5136108
		12	50.4 (50.1 - 50.4)	2.9	23.1	5136790
		13	50.6 (48.6 - 58.7)	3.1	6.4	5088162
		15	75.1 (75.1 - 76.5)	3.3	7.2	5111164
Epigallocatechin	qEGC	1	87.1 (84.4 - 88.6)	3.0	5.8	5133866
		2	7.8 (7.8 - 7.9)	3.2	5.5	5124128
		4	27.2 (27.2 - 28.9)	3.3	56.6	5123475
		6	66.4 (56.5 - 72.4)	3.1	7.2	5073424
		12	42.9 (42.9 - 54.1)	3.0	8.6	5123751
		13	50.6 (48.6 - 58.7)	3.1	6.4	5136623
		14	60.7 (60.7 - 63.1)	3.1	6.9	5132791
Epigallocatechin gallate	qEGCg	4	37.2 (28.9 - 38.6)	3.4	24.1	5087017
		12	48.2 (36.9 - 49.7)	3.0	7.2	5104630
		15	32.1 (17.1 - 33.6)	3.3	6.5	5114089

Theaflavin 1	qTF1	2	4.5 (4.5 - 5.4)	3.1	5.8	5084595
		6	69.6 (69.6 - 75.1)	4.1	7.9	5136045
Theaflavin 2	qTF2	2	4.5 (1.2 - 5.4)	3.1	7.0	5084595
Colour	qCL	7	72.7 (72.7 - 72.7)	3.0	6.5	5132432
Brightness	qBRT	14	71.046 (65.383- 71.046)	3.4	6.8	5125626
Astringency	qAST	1	96.7	3.0	9.1	5135087
		9	(88.8 - 97.3) 87.6 (87.6)	3.2	6.6	5123950
Briskness	qBRK	1	96.7	3.0	7.3	5115441_E-26
		9	(89.2 - 97.3)	3.6	7.4	5123950
		13	87.6 (87.6) 61.582 (61.582)	3.0	7.3	5114985
Aroma	qAR	4	68.6	3.0	6.4	5112599
		10	(68.6) 28.7 (26.9 - 29.0)	3.2	7.0	5128967
%RWC	qRWC	2	60.9	3.2	7.3	5136794
		6	(56.9 - 60.8)	3.3	5.7	5082606
		9	66.2 (66.2 - 66.2) 6.7 (6.2 - 6.7)	3.9	6.9	5130531
Average				3.2	9.9	

%RWC – percent relative water content

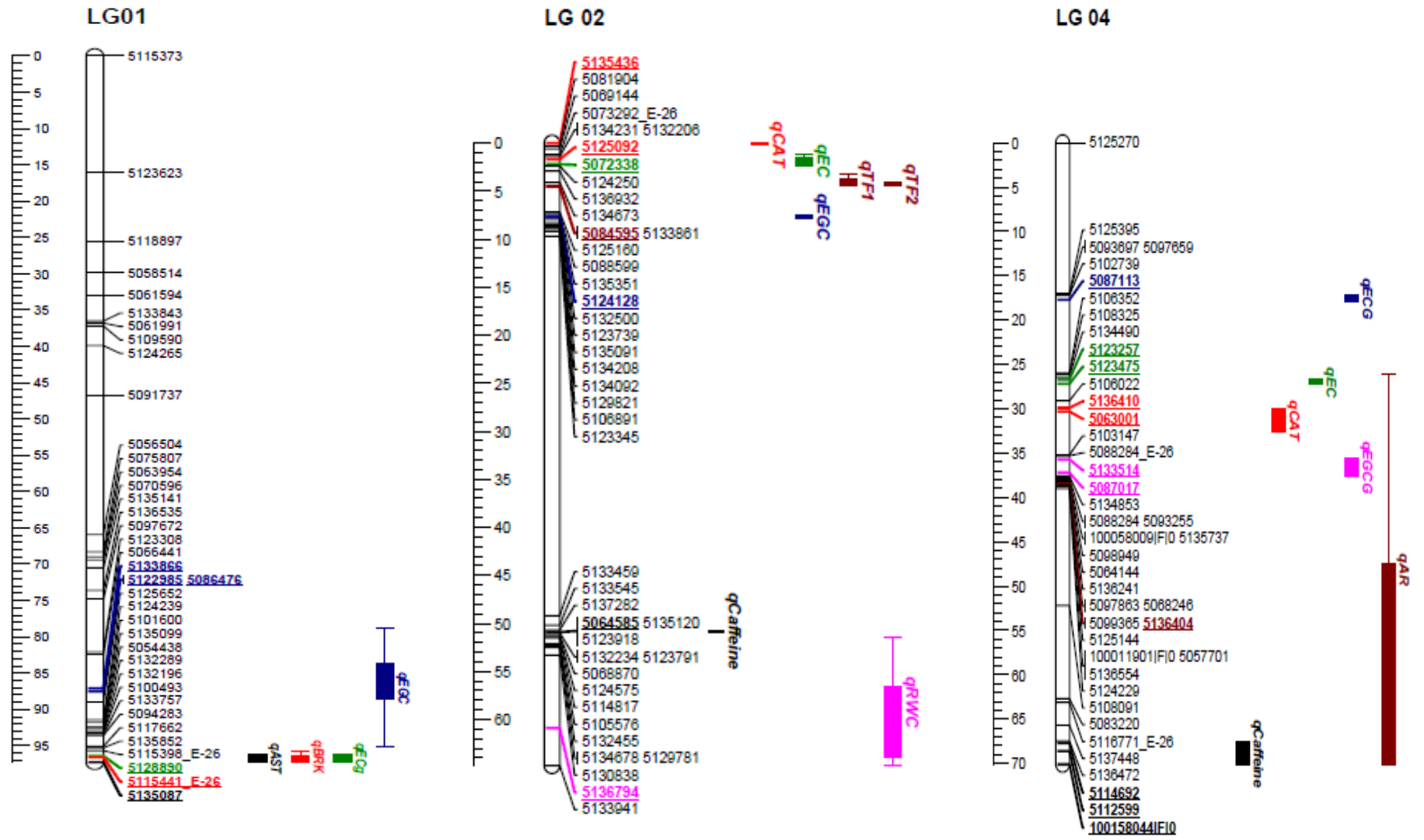
LG – Linkage group

LOD – Logarithm of odds ratio

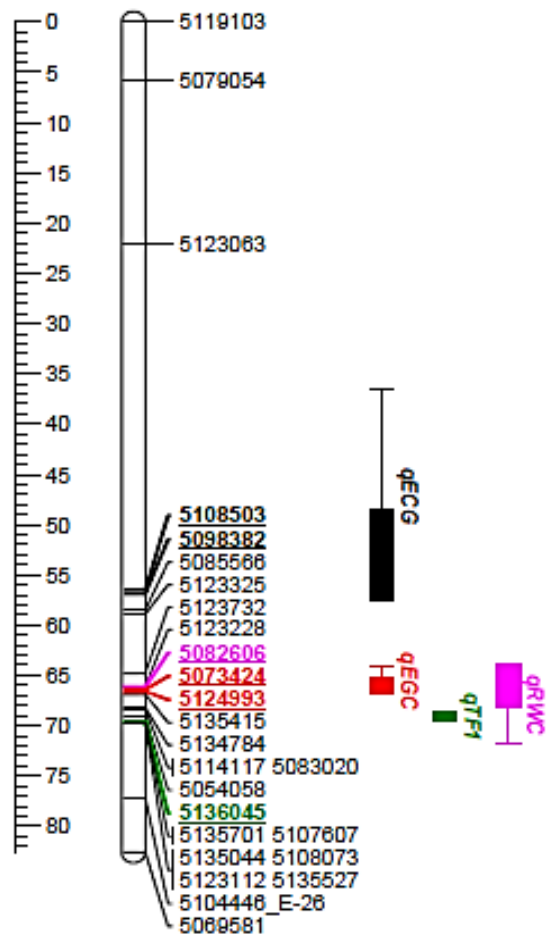
PVE – Phenotypic variation explained

QTL – Quantitative trait loci

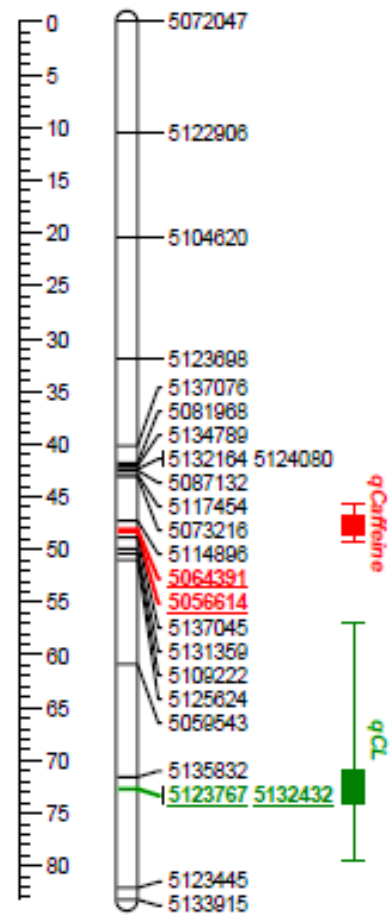
The LOD thresholds are determined by $P < 0.05$, and the basis of which, was permutation testing with $n = 1\ 000$.



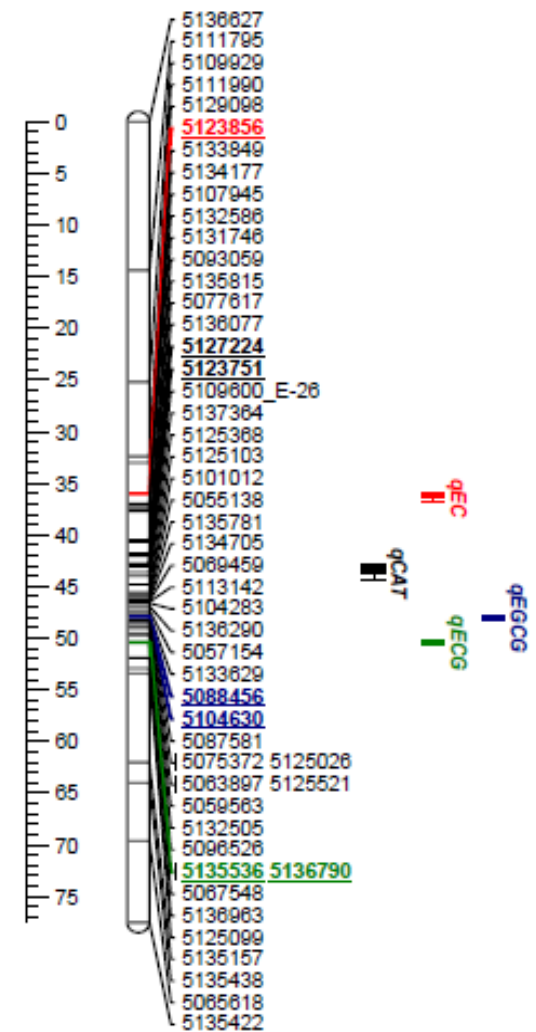
LG 06



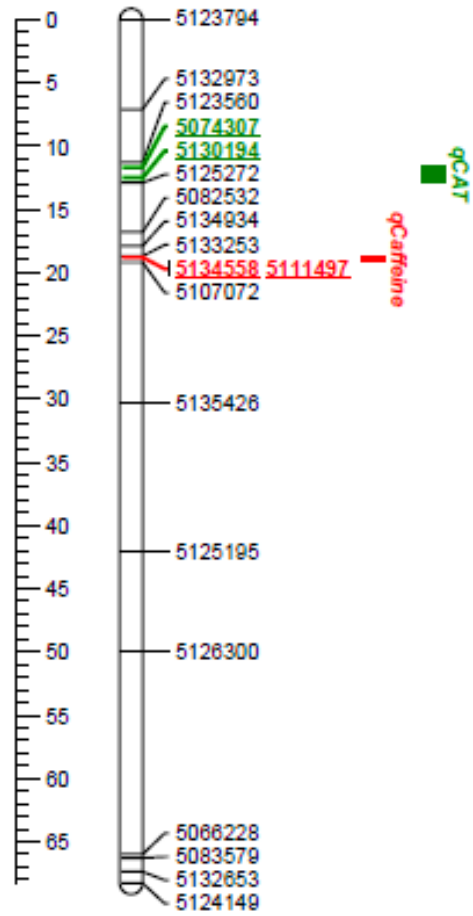
LG 07



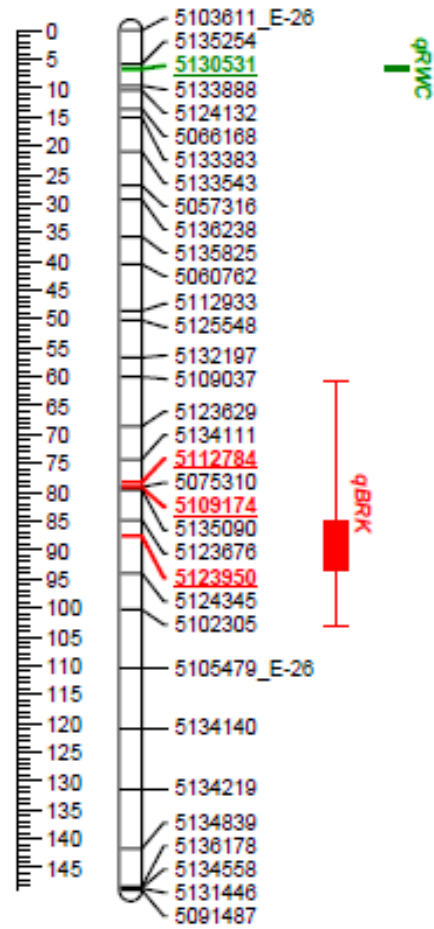
LG 12



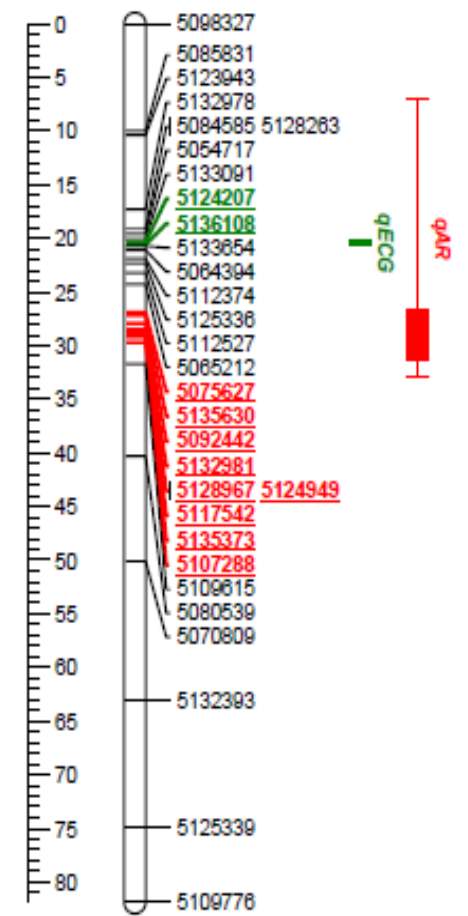
LG 08



LG 09



LG 10



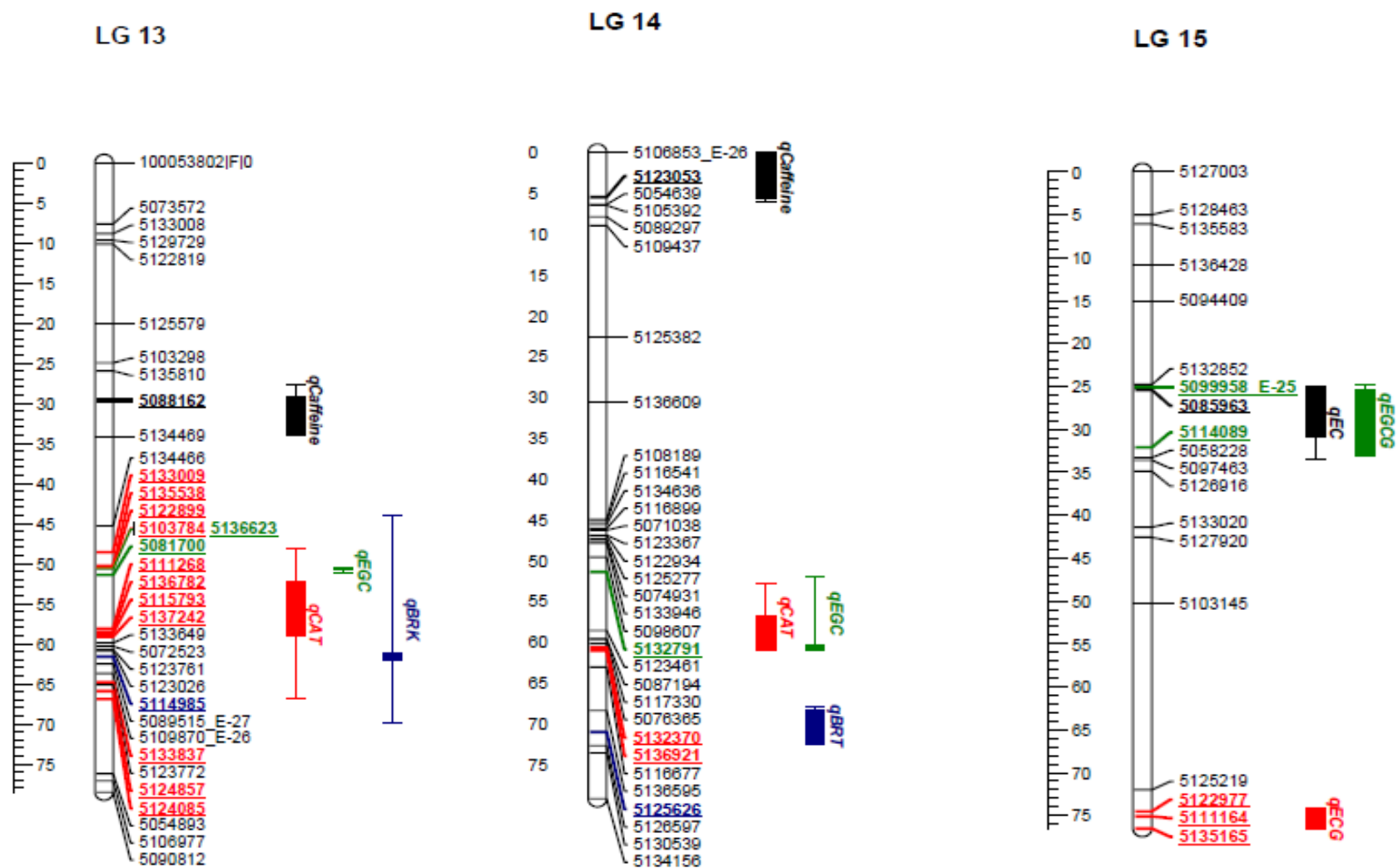


Figure 3.4: Genetic map of *C. sinensis*, displaying UPLC QTL locations for Caffeine, Catechins, Theaflavins, and tea taster's scores. The map ruler is scaled in cM. Each detected QTL e.g. caffeine, catechins are represented by different coloured bars and lines, which are indicative of 1-LOD and 2-LOD support intervals.

Table 3.5: The differences between ¹H-NMR and UPLC QTL markers.

¹ H-NMR QTL	LG	Position	PVE (%)		UPLC QTL	LG	Position	PVE (%)
Caffeine	1	37.1	6.6		Caffeine	2	50.9 (50.2 - 51.5)	6.0
						4	68.6 (63.1 - 70.1)	6.7
						7	48.1 (34.7 - 50.4)	6.6
						8	18.8 (16.8 - 26.0)	6.7
						13	29.7 (23.7 - 36.5)	8.1
						14	5.465 (0 - 5.465)	6.4
Catechin	1	7.1	5.1		Catechin	2	0 (0 - 2.4)	6.4
						4	30.3(29.9 - 30.3)	55.6
						8	12.5 (12.5 - 12.9)	5.6
						12	42.9 (36.0 - 54.1)	10.3
						13	50.6 (48.6 - 58.7)	6.9
						14	60.7 (47.7 - 61.1)	6.9
Epicatechin	5	18.5	18.5		Epicatechin	2	2.2 (0 - 8.3)	7.0
						15	25.4 (20.6 - 25.4)	7.6
Epigallocatechin	13	64.7	5.4		Epigallocatechin	1	87.1 (84.4 - 88.6)	5.8
						2	7.8 (7.8 - 7.9)	5.5
						4	27.2 (27.2 - 28.9)	56.6
						6	66.4 (56.5 - 72.4)	7.2
						12	42.9 (42.9 - 54.1)	8.6
						13	50.6 (48.6 - 58.7)	6.4
14	60.7 (60.7 - 63.1)	6.9						

LG – Linkage group

LOD – Logarithm of odds ratio

PVE – Phenotypic variation explained

QTL – Quantitative trait loci

Table 3.6: Functional annotation of putative candidate genes in GC-MS related linkage groups of *C. sinensis* on reference tea genome.

Nr	QTL	Parent	DArTseq Marker	LG	Position (cM)	LOD	PVE (%)	E-value	Annotated protein	Function
1	qArabinose	TRFK 303/577, GW Ejulu	5125565	14	9.1	3.1	5.7	4.0E-23	['PB1 domain', 'RWP-RK domain']	Abiotic stress
2	qPhloroglucinol	TRFK 303/577, GW Ejulu	5136609	14	30.7	3.1	4.6	5.0E-13	['PB1 domain', 'RWP-RK domain']	Abiotic stress
3	qXylonicAcid	GW Ejulu	5075568	14	64.4	3.0	7.5	2.0E-12	['Pectinesterase']	Drought response

Table 3.7: Functional annotation of putative candidate genes in ¹H-NMR related linkage groups of *C. sinensis* on reference tea genome.

Nr	QTL	Parent	DArTseq Marker	LG	Position (cM)	LOD	PVE (%)	E-value	Annotated protein	Function
1	qAcetic Acid	TRFK 303/577, GW Ejulu	5109437	14	8.9	3.1	25.7	2.0E-25	['Protein of unknown function (DUF677)']	-
2	qCaffeine	TRFK 303/577, GW Ejulu	5109590	1	37.2	3.1	6.6	2.0E-25	['Peptidase C65 Otubain']	Modification of cellular proteins
3	qCatechin	GW Ejulu	5115373	1	7.0	3.0	5.1	1.0E-19	['Histone acetyltransferase subunit NuA4']	Drought response
4	qChlorogenic Acid	TRFK 303/577, GW Ejulu	5085772	11	16.9	3.3	6.3	6.0E-25	['Peptidase C65 Otubain']	Modification of cellular proteins
5	qEpicatechin	TRFK 303/577, GW Ejulu	5132307	5	18.2	3.0	18.5	3.0E-08	['Peptidase family M3']	Abiotic stress
6	qEpigallocatechin	TRFK 303/577	5133837	13	64.8	3.1	5.4	9.0E-18	['PB1 domain', 'RWP-RK domain']	Abiotic stress
7	qIsoleucine	TRFK 303/577, GW Ejulu	5120311	7	62.8	3.5	7.5	5.0E-16	['ABC transporter transmembrane region 2', 'ABC transporter']	Transport protein
		TRFK 303/577	5070055	13	29.7	3.3	66.3	4.0E-23	['Alcohol dehydrogenase GroES-like domain', 'Zinc-binding	Carbohydrate metabolism

									dehydrogenase']	
8	qValine	TRFK 303/577, GW Ejulu	5123739	2	8.3	3.2	42.6	2.0E-6	['C1-like domain']	-
		TRFK 303/577, GW Ejulu	5016516	13	19.2	3.1	8.1	3.0E-17	['Amino acid kinase family']	Nitrogen Assimilation
		TRFK 303/577, GW Ejulu	5106853_E-26	14	5.5	3.0	14.4	9.0E-21	[' Domain of unknown function (DUF4217)']	-

Table 3.8: Functional annotation of putative candidate genes in UPLC-DAD related linkage groups of *C. sinensis* on reference tea genome.

Nr	QTL	Parent	DArTseq Marker	LG	Position (cM)	LOD	PVE (%)	E-value	Annotated protein	Function
1	qECg ^a	TRFK 303/577, GW Ejulu	5128890	1	96.4	6.8	11.7	2.0E-25	['Actin']	Abiotic stress
2	qEC ^a	TRFK 303/577, GW Ejulu	5072338	2	2.2	4.1	5.6	2.0E-25	['Peptidase family M3']	Abiotic stress
3	qEGC ^a	GW Ejulu	5124128	2	7.7	3.2	6.8	2.0E-18	['Kinesin motor domain']	Transport protein
4	qCaffeine ^b	TRFK 303/577, GW Ejulu	5064585	2	50.9	3.4	5.8	6.0E-25	['Peptidase C65 Otubain']	Modification of cellular proteins
5	qECg ^a	TRFK 303/577, GW Ejulu	5097659	4	17.1	4.6	7.8	1.0E-07	['Rpp14/Pop5 family']	-
6	qECg ^a	TRFK 303/577	5087113	4	17.7	4.7	22.9	6.0E-22	['impB/mucB/samB family']	UV protection through DNA repair
7	qEC ^a	GW Ejulu	5134490	4	26.4	12.6	43.7	3.0E-08	['Aminotransferase class I and II']	Phenylalanine, tyrosine and tryptophan biosynthesis
8	qEGCg ^a	GW Ejulu	5134853	4	37.6	14.9	45.1	2.0E-06	['Diaclyglycerol kinase catalytic domain']	Abiotic stress
9	qTF1 ^a	TRFK 303/577, GW Ejulu	5106352	4	26.0	14.7	45.6	2.0E-06	['Thiolase, C-terminal domain']	Benzoic acid biosynthesis
10	qEC ^a	GW Ejulu	5123475	4	27.2	14	51.5	1.0E-10	CSA016461	-
11	qEGC ^a	GW Ejulu	5123475	4	27.2	3.7	51.5	1.0E-10	CSA016461	-
12	qEC ^a	TRFK 303/577	5119221	4	32.7	3.1	53.8	1.0E-19	['Histone acetyltransferase subunit NuA4']	Drought response
13	qEGCg ^a	TRFK 303/577	5119221	4	20.6	3.3	53.8	1.0E-19	['Histone acetyltransferase subunit NuA4']	Drought response
14	qEGC ^a	TRFK 303/577, GW Ejulu	5136058	4	27.6	46.1	54.1	1.0E-07	['Autophagy-related protein 11']	Abiotic stress
15	qTF4 ^a	TRFK 303/577, GW Ejulu	5136058	4	27.6	14.8	54.1	1.0E-07	CSA024230	-
16	qCaffeine ^b	TRFK 303/577, GW Ejulu	5114692	4	68.6	3.8	6.1	4.0E-23	['BT1 family']	Transport protein
17	qCAT ^b	TRFK 303/577	5119221	4	32.7	3.1	2.3	1.0E-19	['Histone acetyltransferase subunit NuA4']	Drought response

18	qECg ^a	TRFK 303/577	5136985	6	26.4	5.1	5.3	6.0E-28	['KOW motif']	
19	qTF1 ^a	TRFK 303/577	5136045	6	69.6	4.5	6.9	6.0E-19	['Catalase']	Abiotic stress
20	qECg ^b	TRFK 303/577, GW Ejulu	5108503	6	56.5	4.8	8.2	2.0E-27	['DnaJ domain']	Drought response
21	qECg ^b	GW Ejulu	5098382	6	56.9	4	8.7	6.0E-19	['Asparagine synthase, Glutamine amidotransferase domain']	Nitrogen mobilization
22	qRWC ^b	TRFK 303/577	5082606	6	66.2	3.3	5.7	9.0E-21	['Alpha adaptin AP2']	Abiotic stress
23	qCaffeine ^b	GW Ejulu	5064391	7	48.1	3.7	6	2.0E-27	['Lipase (class 3)']	Lipid degradation, esterification and transesterification processes
24	qCaffeine ^b	TRFK 303/577, GW Ejulu	5134558	8	18.8	3.9	7.5	9.0E-24	['Nitronate monooxygenase']	Catabolic or anabolic pathways
25	qRWC ^b	TRFK 303/577, GW Ejulu	5130531	9	6.7	4	7	7.0E-06	['MatE']	Drought response, Sequestration of proanthocyanidins
26	qTF2 ^a	TRFK 303/577, GW Ejulu	5128967	10	28.7	3.5	7	2.0E-25	['Acyl-CoA oxidase']	Lipid catabolism and plant hormone biosynthesis
27	qECg ^a	GW Ejulu	5072021	10	25.5	4.3	7.5	8.0E-09	['ATPase family associated with various cellular activities (AAA)']	Heat stress response
28	qECg ^b	GW Ejulu	5136108	10	20.6	4.8	5.7	1.0E-10	['Protein kinase domain']	Abiotic stress
29	qECg ^b	TRFK 303/577, GW Ejulu	5124207	10	20.4	3.1	5.3	8.0E-12	['Acyltransferase']	Phenylpropanoid and Shikimate pathway
30	qEGCg ^a	TRFK 303/577	5088456	12	47.9	3.9	9.8	4.0E-23	['Protein kinase domain']	Abiotic stress
31	qCAT ^a	GW Ejulu	5136077	12	42.8	6.6	11.2	5.0E-13	CSA026168	-

32	qEGC ^a	GW Ejulu	5136077	12	42.8	5.6	11.2	5.0E-13	CSA026168	-
33	qCAT ^a	TRFK 303/577, GW Ejulu	5123751	12	43.0	6.1	11.5	2.0E-18	['Bromodomain']	Scaffolding proteins
34	qEGC ^a	TRFK 303/577, GW Ejulu	5123751	12	43.0	14.9	11.5	2.0E-18	['Bromodomain']	Scaffolding proteins
35	qCaffeine ^b	TRFK 303/577, GW Ejulu	5088162	13	29.7	4.7	5.4	6.0E-28	['PA domain']	-
36	qCAT ^b	TRFK 303/577	5103784	13	50.6	3.8	6.1	8.0E-12	CSA003424	-
37	qCAT ^b	TRFK 303/577	5122899	13	50.5	3.8	6	6.0E-19	CSA033214	-
38	qCAT ^b	TRFK 303/577, GW Ejulu	5133009	13	48.6	3.6	5.8	1.0E-16	['Adaptor complexes medium subunit family']	-
39	qCAT ^b	TRFK 303/577	5122899	13	50.5	3.8	6	6.0E-22	['Protein kinase domain']	Abiotic stress
40	qCAT ^b	TRFK 303/577	5122899	13	50.5	3.8	6	2.0E-21	['14-3-3 protein']	Abiotic stress
41	qCAT ^b	TRFK 303/577	5122899	13	50.5	3.8	6	2.0E-21	['NB-ARC domain']	Disease resistance
42	qCAT ^b	TRFK 303/577	5122899	13	50.5	3.8	6	2.0E-21	['Pectinesterase']	Drought response
43	qCAT ^b	TRFK 303/577	5122899	13	50.5	3.8	6	1.0E-16	['2OG-Fe(II) oxygenase superfamily']	Abiotic stress
44	qCAT ^b	TRFK 303/577, GW Ejulu	5111268	13	50.6	4.1	7.2	6.0E-28	['WD domain']	Transport protein

45	qCAT ^b	TRFK 303/577, GW Ejulu	5123751	13	58.1	4.1	6.5	4.0E-23	['Transmembrane amino acid transporter protein']	Abiotic stress
46	qECg ^b	TRFK 303/577, GW Ejulu	5088162	13	50.6	3.7	5.4	6.0E-28	['PA domain']	-
47	qEGC ^b	TRFK 303/577, GW Ejulu	5123751	13	50.6	3.7	6.5	4.0E-23	['Transmembrane amino acid transporter protein']	Abiotic stress
48	qEGC ^a	GW Ejulu	5116677	14	63.1	4.1	5.2	6.0E-28	CSA002263	-
49	qEGC ^a	GW Ejulu	5116677	14	63.1	4.1	5.2	3.0E-11	['Armadillo/beta-catenin-like repeat']	Heat stress response
50	qBRT ^a	TRFK 303/577	5122986	14	65.4	3.7	7.5	7.0E-06	['Glycosyl hydrolase family 9']	Phenylpropanoid pathway
51	qCAT ^b	-	5132370	14	60.7	4.1	2.5	1.0E-10	['Glutaminyl-tRNA synthetase']	Chlorophyll biosynthesis
52	qECg ^b	GW Ejulu	5111164	15	75.1	4.2	7.2	1.0E-07	['Isocitrate/isopropylmalate dehydrogenase']	Abiotic stress
53	qEGCg ^b	TRFK 303/577, GW Ejulu	5114089	15	32.1	3.8	6.8	8.0E-12	['Cytochrome P450']	Biotic and abiotic stresses

^a Putative QTL identified based on Interval Mapping with LOD>3.0

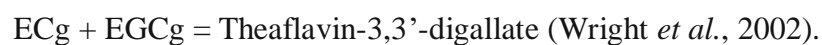
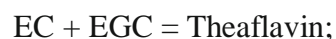
^b Putative QTL identified based on Multiple QTL Model Mapping with LOD>3.0

QTL - Quantitative trait loci; LG - Linkage group; LOD - Logarithm of odd ration; cM - Centi-morgan, CAF - Caffeine, CAT - Catechin, EC - Epicatechin, ECg - Epicatechingallate, EGC - Epigallocatechin, EGCg - Epigallocatechin gallate, TF1 - Theaflavin, TF4 - Theaflavin-3,3'-digallate, BRT - Brightness,%RWC - Percent relative water content

3.11 DISCUSSION AND CONCLUSION

Metabolomics can be employed as a “competent tool” by tea breeders to aid them in the selection and improvement process for good tea cultivars. Yield and quality are some of the traits of interest that have been documented in literature to be significantly influenced by several QTLs (Kamunya *et al.*, 2010). It is also these traits that farmers are keen on, as they get paid more for high quality, high yield teas; an enhancement in these attributes increase/improve their competitiveness not only in the local, but also in the global market. As mentioned earlier, this present study served to identify molecular markers capable of being used in MAS, and in so doing, shorten both the development and commercialisation stages of choice varieties, instead of waiting for periods of up to 20 years to develop new cultivars through conventional breeding and clone selection. Furthermore, this study, being the first of its kind, presents opportunities for the exploration of further QTL analysis of different tea populations. As mentioned, the metabolites from the 310 samples of green tea obtained from the TRI in Kenya were ascertained using GC-MS, ¹H-NMR and UPLC metabolomics platforms; 60 were the Comm cultivars and 250 of these were the NComm cultivars. The ¹H-NMR results indicate that levels of caffeine, catechin, EC and EGC were higher in the Comm cultivars as compared to the NComm cultivars. The ¹H-NMR further detected the amino acids valine and isoleucine. These were found to be higher in the cultivars that were also established to be DT, and were rich in catechins. Cultivars rich in catechins have been shown to be of a higher quality, as was documented in a study on the *Arabidopsis* plants and their metabolic responses to drought stress (Arbona *et al.*, 2003). A study on Poplar trees further showed that isoleucine levels were up-regulated in the DT Poplar trees than in the DS under drought stress (Hamanishi *et al.*, 2015). This is in agreement with the results obtained in this study. Lastly, the UPLC results show that caffeine, catechin, EC, theaflavin, theaflavin-3'-gallate and theflavin-3,3'-digallate were variable importance in projection (VIP) metabolites responsible for distinguishing between the Comm and NComm cultivars. Correlations have been documented in literature between the umami taste found in tea and the metabolites aspartic acid, asparagine, and theanine, while further correlations have been seen between the traits bitterness and astringency with arabinofuranose (Wei *et al.*, 2014), EC, ECg, EGC, EGCg and gallic acid (Robichaud and Noble, 1990), mannose (He *et al.*, 2015) and theobromine (Bonvehi and Coll, 1997). According to literature, fresh green tea leaves contain trace amounts of gallic acid, which then accumulates during auto-oxidation in the

manufacturing process of black tea as a result of galloyl ester breakdown from the catechins and theaflavin gallates. The high levels of gallic acid in some cultivars has been attributed to corresponding high levels of gallated catechins, which result in the generation and consequent degradation of the theaflavins. During black tea manufacture, theaflavic acids are formed, and these in turn oxidize the gallocatechins EGC and EGCg, releasing gallic acid (Kerio *et al.*, 2013). Therefore cultivars rich in gallic acid can be postulated to be rich in gallocatechins, responsible for tea liquor taste. It has been documented that EGCg and ECg are principal taste metabolites in tea, which are responsible for tea astringency, while caffeine is responsible for bitterness (Xu *et al.*, 2018). Furthermore theaflavins, in black tea, catechins, caffeine and glycosidic flavonoids have been shown to serve as metabolomic markers which distinguish high quality teas from the low quality teas (Wang and Ruan, 2009), influencing the price of tea at auctions. The gallated catechins EGC and EGCg significantly contribute to the generation of theaflavins in black tea. As such, tea cultivars high in EGC and EGCg concentrations can, through breeding programs, be developed to enhance the quality of resultant black tea. It has been reported that the ratio of di-hydroxyl flavan-3-ols to tri-hydroxyl flavan-3-ols impacted the quality of black tea; high quantities of simple catechins such as catechin, EC and ECg compared to the gallo-catechins EGC and EGCg, results in higher amounts of theaflavin (Ellis and Nyirenda, 1995). Therefore it is evident that the cultivars with higher catechin content produce high quality green and black tea. From the results obtained in the present study, the NComm cultivars, with higher catechin content correlate to the organoleptic results obtained independently from the tea taster, which agrees with literature. Furthermore, astringency, another trait considered and ascertained by the tea taster, is greatly influenced by theaflavin digallate, which is approximated to be 6.4 times more astringent than theaflavin, while also being 2.88 times more astringent than both theaflavin-3-monogallate and theaflavin-3'-monogallate (Obanda *et al.*, 2001). According to Wright *et al.*, (2002), to generate a single molecule of theaflavin necessitates a dihydroxy and a trihydroxy flavan-3-ol as shown below:



The interrelation between catechins and TFs content with respect to tea quality was extensively studied, with focus on aroma, astringency, brightness, briskness, and colour. Although the organoleptic results indicate a positive correlation between caffeine and catechin content, and quality, they were not convincing. This could be a result of a degradation of the aromatic compounds during storage, and prior to the time of organoleptic evaluation, leading to flat tea. Thearubigin, TF/TR ratio and sensory scores have been documented to significantly decrease with sample storage time (Sedaghatthoor *et al.*, 2013). A significant correlation was observed between the individual sensory traits studied i.e. aroma, astringency, brightness, briskness, and colour, comparable to that obtained by (Owuor *et al.*, 2006b), with no noteworthy correlation between non-gallated theaflavins and the sensory traits, with the exception of the tea liquor brightness. These abovementioned significant correlations are indicative of phenotypic traits controlled by linked genes. It is for this reason that the enhancement of one phenotypic trait may result in the enhancement of the rest of the phenotypic traits under investigation.

The construction of genetic linkage maps is an important requisite for QTL identification of agronomically significant genes such as those responsible for yield and quality, which are influential in the development of better-quality cultivars. Similar to the maps obtained by Taniguchi *et al.*, (2012) of 1298 and 1305 cM, and Ma *et al.*, (2014) of 1143.5 cM, the map obtained in this study was a total length of 1260.1 cM with 1421 markers. This study produced 15 linkage groups, an indication of genome saturation, with $n = 15$. The restriction enzymes PstI (CTGCAG) and MseI (CCGG) performed optimally, with 16,382 DArTseq attained; there was however a gap of more than 20 cM between adjacent markers on LG06 and LG15. This may be due to genome regions which correspond to gap regions in the genetic map; further research needs to be conducted to fill in the gaps in the genetic map used in the current study. The markers mapped in this study are spread over the 15 LGs with marker densities extending from 0.5 to 1.7 cM. A recommended marker density of less than 10 cM is required for genome-wide QTL mapping (Doerge, 2002; Taniguchi *et al.*, 2012); the map contrived and used in this study is therefore ideal for QTL identification. In total, for the GC-MS data, one arabinose, one phloroglucinol, and one xylonic acid were derived, with the %PVE ranging from 4.6 to 7.5 (Table 3.2) and averaging 5.9%. One acetic acid, one caffeine, three catechins (one catechin, one EC and one EGC), one chlorogenic acid, four amino acids (one isoleucine and three valines) were detected using the ¹H-NMR derived data, with the %PVE by each QTL varying from 5.1 to 96.3%, with an average of 34.4% (Table 3.3).

Lastly, six caffeine, 25 catechins, three theaflavins, nine organoleptic scores and three %RWC QTLs were identified, with a %PVE varying between 5.5 to 56.6%, and averaging 9.9% (Table 3.4). The high PVE displayed by the ¹H-NMR QTLs acetic acid, epicatechin, isoleucine and valine, and the UPLC QTLs caffeine, catechins, theaflavins, organoleptic scores, and %RWC suggests that these attributes could possibly be controlled by critical genes. The sample size of 250 employed in this study was comparable to that of 300 used in the study entitled “*construction of a SSR-based genetic map and identification of QTLs for catechins content in tea plant*” (Ma *et al.*, 2014). In addition to the QTLs for catechins obtained across the ¹H-NMR and UPLC platforms, the current study also incorporated QTLs for acetic acid, caffeine, chlorogenic acid, isoleucine and valine from ¹H-NMR, and arabinose, phloroglucinol and xylonic acid from GC-MS, which influence the quality of tea. As shown above in e.g. Figure 3.3, some linkage groups such as LG 13 have several QTLs i.e. QTLs for EGC, isoleucine, and valine. This is indicative that the regions of the chromosomes contain multifunctional genes concomitant with amino acid and catechin production and accretion; this is worthy of further investigation. Moreover, it was interesting to note that the QTLs associated with caffeine, catechin, EC and ECg from both ¹H-NMR and UPLC were located on different LGs, and at different positions on the chromosome, with different %PVE (Table 3.5). This clearly indicates that the genes concomitant with the manufacture and accretion of these metabolites are sparsely situated in different chromosomal regions. The GC-MS and ¹H-NMR results were obtained from an untargeted approach, whereas the UPLC results were obtained from a targeted approach. It has been documented extensively in metabolomics literature that a targeted approach always yields better results than an untargeted one. In the case of a targeted approach, the peaks were confidently identified based on the matching of their retention times and UV spectra to that of pure standards. In the case of the GC-MS and ¹H-NMR, peaks may have been incorrectly identified, even if their identities were based on in-house and online metabolite databases; without the pure standards to corroborate, a likelihood of misidentification exists. This could be the reason why some metabolites e.g. caffeine are found on different LGs. In future works, pure standards for the identified metabolites from an untargeted approach will be purchased and run to confirm the identity. It must also be noted that the sample size of 310 employed in the mapping population employed in the study for the ¹H-NMR was too small, as compared to the sample sizes employed in other similar studies, which had sample sizes of up to 3861 and 4630 respectively (Raffler *et al.*, 2015; Son *et al.*, 2008). Further work will have to be conducted to

confirm these QTLs for MAS to avoid any possible overestimation of the QTL effects, decreasing the statistical power for detection of QTLs possessing lesser effects. ¹H-NMR is a semi-quantitative technique and as such a large sample size is required to minimise obtained variation within samples. Thus for future works, an increased sample size for population mapping is required, to ensure better estimations of QTL loci and effect. Furthermore, the differences obtained could be a result of complications faced during the sample preparation and analysis. As indicated in the materials and methods section, the samples were placed in zip-locked bags before storing at 4°C; during transfer of the samples into test tubes for analyses, different amounts of samples were transferred due to static experienced between the samples and the plastic zip-locked bags. As such further variation was added, which could have affected the final concentrations obtained and subsequently the detected QTL positions per LG. In future studies, glass polytop vials will be used to avoid this variation being added. Literature has documented the significant impact the environment has on the accuracy of QTL detection. This is due to the fact that certain environment-specific QTLs tend to express differently when in different environments; this thus makes such QTLs problematic to employ when breeding for the development of functional traits (Ma *et al.*, 2014).

The GC-MS putative QTLs, qArabinose, qPhloroglucinol, and qXylonic acid, and the ¹H-NMR qEpigallocatechin, in the present study were annotated RWP-RK protein domain, which function in aiding the tea plant against abiotic stress, particularly drought stress (Table 3.6). The RWP-RK protein family are transcription factors which mediate DNA binding. This family's functional analysis revealed several RWP-RK proteins to possess key roles in regulating nitrogen availability in e.g. *Arabidopsis* during stress conditions (Chardin *et al.*, 2014). In the present study, QTLs qCaffeine and qChlorogenic acid, were annotated peptidase C65 Otubain proteins. This family of proteins has been reported to be a very precise ubiquitin iso-peptidase, functioning to remove ubiquitin from proteins. The ubiquitin protein modification is a significant event that causes/ increases protein stability and function in eukaryotic cells; the process is dynamic and reversible (Balakirev *et al.*, 2003). These proteins therefore modify cellular proteins in response to abiotic and biotic stress, aiding the tea plant to cope, and survive. The ¹H-NMR putative QTL, qValine, was annotated an amino acid kinase protein. The synthesis of essential amino acids such as lysine and threonine in plants, is primarily controlled by the feedback inhibition of the kinase proteins aspartate kinase and dihydrodipicolinate synthase. Furthermore, the control of carbon fixation and nitrogen assimilation, and the regulation of carbon and nitrogen into amino acids

during stress conditions is also regulated by this family of proteins. The metabolic regulation of the aspartate kinase gene expression in *Arabidopsis* was studied by Kochhar *et al.*, (1998) who revealed that aspartate conversion into storage amino acid asparagine was subject to reciprocal metabolic control, and that this branch point was a part of a greater nitrogen and carbon regulatory mechanism to enable the plant to store and utilise these during stress. The putative QTL, qIsoleucine, was annotated as GroES-like zinc-binding alcohol dehydrogenase family protein (Table 3.7). These have been shown to function in carbohydrate metabolism. Nyarukowa *et al.*, (2016) reported that drought tolerant tea cultivars had higher levels of carbohydrates than their drought susceptible counterparts. The tolerant cultivars were able to effectively metabolise carbohydrates, providing the plants with energy to “combat” the drought stress conditions. This could explain why this protein was identified in the drought tolerant TRFK 303/577 parental clone.

According to Punyasiri *et al.*, (2004) and Vankatesh *et al.*, (2007), the biosynthetic pathway for *C. sinensis* flavonoids begins with the deamination of phenylalanine to produce trans-cinnamic acid. Trans-cinnamic acid is in turn oxidised to give p-coumaric acid, which then forms p-coumaroyl-CoA. The pivotal step in the flavonoid biosynthetic pathway involves the enzyme chalcone synthase, which catalyses the condensation reaction of p-coumaroyl-CoA and malonyl-CoA to produce chalcone; chalcone isomerisation produces (2S)-flavonones. The identification of the CAT putative QTL for 2-ODDs superfamily protein in both the ¹H-NMR and UPLC/DAD results further validates previously reported literature findings about the 2-ODDs function in flavonoid biosynthesis. The 2-ODDs are non-heme proteins involved in reactions such as C-C bond cleavage, epimerisation, fragmentation, hydroxylation, and ring formation. These 2-ODDs catalyse the formation of flavonoid subclasses, namely (2S)-flavonones which, through a hydroxylation process, are transformed to dihydroflavonols. The formed dihydroflavonols then serves as a substrate for flavonol synthase, which competes with dihydroflavonol 4-reductase to produce anthocyanidins, flavonols, and procyanidins (Table 3.8).

Histone acetyltransferases were identified in the results of this study (¹H-NMR and UPLC). These enzymes are critical in the histone acetylation of chromatin in plants, which is vital in the epigenetic control of gene expression. Acetyl group transfer to core histone tails by histone acetyltransferases, facilitates the transcription of key genes involved in plant drought response and abscisic acid signalling e.g. in *Arabidopsis thaliana* (Kim *et al.*, 2008) and rice (Fang *et al.*, 2014). In a study by Kim *et al.*, (2015), it was reported that the changes in

histone modification could be correlated with the up-regulation of genes involved in drought stress-response. The putative QTLs for catechin (qCatechin in ¹H-NMR and qCAT in UPLC), qEC, and qEGCg in the present study were annotated histone acetyltransferase subunit NuA4 proteins. This corroborates the findings of Jeyaramraja *et al.*, (2003), and Cheruiyot *et al.*, (2008), which reported catechins as possible drought tolerance markers in *C. sinensis*.

Glycosylation, an important process regulated by glycoside hydrolases, involving plant polyphenols has been shown to increase hydrophobic flavonoid solubility and stability (Xu *et al.*, 2016). This process involves glycosidic bond hydrolysis and rearrangement. It is our postulation that the UPLC qBRT putative QTL, associated with the glycosyl hydrolase family 9, functions in synthesising and sequestering phenylpropanoids (Jones *et al.*, 2003). Galloylated catechins found in tea, the result of UDP-glucosyltransferase glucosylation activity, have been reported in literature to be responsible for the astringency and bitterness of teas. Flavonol 3-O-glycosides on the other hand have been reported as being responsible for the dry mouth, velvety mouth-coating sensations experienced as a result of tea consumption (Cui *et al.*, 2016).

Acyltransferases, which are involved in several metabolic pathways such as the biosynthesis of anthocyanidin, and the transfer of acyl groups, using acyl-CoA as the donor, to anthocyanin sugar moieties (Mizutani *et al.*, 2006), were putatively identified in the present study. The putative QTL, qCaffeine, was annotated as BT1 family protein. This protein has been reported by Haferkamp, (2007) as being responsible for exporting adenine and guanosine nucleotides, precursors for caffeine biosynthesis, synthesised entirely in plants plastids (Negishi *et al.*, 1992). The present study putatively annotated QTL qCAT as 14-3-3 protein, a protein involved in abiotic stress response in tea. This finding is in agreement with the findings of Cheruiyot *et al.*, (2008b) where individual catechins were reported to be potential drought tolerance predictors in tea. These 14-3-3 proteins bind to other proteins, inducing specific target-site modification and rearrangement, essential in signal transduction pathways. *Arabidopsis* studies have associated 14-3-3 proteins with ABA signalling, whose primary function includes regulating plant response to abiotic or biotic stress. The 14-3-3 proteins have also been implicated as being in key in physiological stress responses such as carbon and nitrogen metabolism, and various plant growth and development aspects (Koech *et al.*, 2019).

Drought stress has been shown to affect photosynthesis, which in turn affects the plant nutrient availability resulting in e.g. ion intoxication. To combat this, plants rely on reversible

protein phosphorylation by protein kinases at the beginning and later stages of signalling pathways, as a response to abiotic stress. Furthermore, plants regulate the expression of certain protein kinase genes, in order to preserve osmotic homeostasis during drought stress. Putative QTL, qCAT, was also annotated for a protein kinase domain in the present study. Jeyaramraja *et al.*, (2003) reported catechins to be higher in the drought tolerant cultivars as compared to drought susceptible cultivars. This serves as corroboration for the results obtained in the present study, as qCAT was annotated as a protein kinase domain in the TRFK 303/577, the drought tolerant parent.

Amino acids have been reported in literature as not only being essential protein synthesis, but for also being important signalling molecules. In *Arabidopsis* for example, drought and saline stress bring about an upregulation in the expression of proline transporters 1 and 2, while downregulating the expression of amino acids permease 4 and 6 (Hua *et al.*, 2017). It is for this reason that transmembrane amino acid transporter proteins are important; they facilitate the transport of amino acids across biological membranes. Proline has been reported to function as an osmo-protectant, which confers drought tolerance to tea plants subjected to drought stress (Upadhyaya and Panda, 2013). In a study by Nyarukowa *et al.*, (2016) it was reported that during drought stress, drought tolerant tea cultivars convert phenylalanine into proline, which is an osmo-protectant and confers tolerance on the tea cultivars. The present study identified putative QTLs qCAT and qEGC, annotated as putative candidate genes for transmembrane amino acid proteins. It can be postulated that said protein functions as a transmembrane amino acid transporter for proline during times of drought stress. Moreover, these transmembrane amino acid transporter proteins may be involved in transporting phenylalanine, tyrosine and tryptophan, which are precursors for several tea secondary metabolites, importantly polyphenols. These are involved in phenylalanine and shikimate pathways.

The next family of proteins annotated in the present study was MATE proteins. These are active transport proteins utilising membrane electrochemical gradients, whose transport activity is maintained by ATPases. MATE proteins transport secondary metabolites such as anthocyanins, flavon-3-ols, and flavone glycosides into vacuoles, the plant cell's major storage site (Shitan, 2016). The present study saw the putative QTL, qRWC, being annotated for MATE proteins. These proteins are responsible for sequestering vacuole flavonoids in response to water stress (Petruzza *et al.*, 2013). The MATE gene family member, TT12, has been reported to also be involved in the sequestration of proanthocyanidins in seed vacuoles,

resulting in seed coat pigmentation (González *et al.*, 2016), while it transports epicatechin 3'-O-glucoside and cyanidin 3-O glucoside in yeast (Marinova *et al.*, 2007). MATE proteins have also been reported to facilitate and modulate the efflux of ABA and drought-tolerance sensitivity in *Arabidopsis* (Zhang *et al.*, 2014). The fact that the present study found a putative QTL for %RWC annotated as MATE, it thus corroborates the findings of Nyarukowa *et al.*, (2016), which, relying on %RWC, classified tea cultivars as either drought tolerant or susceptible using tea leaf metabolites. In the present study, putative QTL, qECg, was annotated DnaJ protein, which are proteins that are involved in cellular protein homeostasis i.e. protein folding, break down and refolding during plant stress situations (Park and Seo, 2015). They have also been reported to function as chaperones; this can either be alone or by associating with heat-shock protein 70. Wang *et al.*, (2016), in a transcriptomics study on the effect of drought stress on leaf quality of tea, reported an increase in levels of ECg and EGCg during drought stress. Jeyaramraja *et al.*, (2003) also reported on how soil moisture content alterations due to drought affect tea biomolecules in relation to its quality. The results of the present study indicate that ECg may be employed as a possible marker for drought tolerance in *C. sinensis* cultivars.

The qEGCg putative QTL was annotated as diacylglycerol kinase catalytic domain protein. This is postulated to be a drought and cold tolerance marker in *C. sinensis* cultivars. Abiotic stresses such as cold, drought, and salinity have been shown to trigger plants to produce phosphatidic acid (Zhu, 2016). Phosphatidic acid is produced through phosphorylating diacylglycerol, a reaction catalysed by diacylglycerol kinase. Diacylglycerol is essential for the development of, and the response to environmental stimuli. Salinity stress has been shown to increase non-specific phospholipase C activity, promoting diacylglycerol production in *Arabidopsis*. Cold stress has also been reported to induce phosphatidic acid production in suspension-cultured *Arabidopsis* cells (Arisz *et al.*, 2013). Lastly, aminotransferases are enzymes responsible for catalysing the transfer of amino groups from amino donor to acceptor compounds. These proteins are involved in several key metabolic pathways such as amino acid biosynthesis, and secondary metabolites biosynthesis (de la Torre *et al.*, 2014). Prephenate aminotransferase is a class of aminotransferase enzymes responsible for catalysing the final reaction in phenylalanine biosynthesis, a product central to the shikimate pathway; it is a vital precursor for flavonoid synthesis (Ververidis *et al.*, 2007). The putative QTL, qEC, in the present study was annotated as aminotransferase I and

II, leading to the postulation that EC is a marker associated with flavonoid biosynthesis in *C. sinensis*.

In conclusion, as mentioned earlier, this study is the first of its kind, attempting to acquire information pertaining to genomic and functional annotation of proteins responsible for quality, yield and drought tolerance in black tea using DArTseq markers. The results in this study illuminate on the association between proteins and certain metabolites, and how they are responsible for certain traits observed in tea cultivars. The DArTseq markers were able to offer beneficial information, revealing quality and drought tolerance genetic determinants in black tea. The DArTseq markers employed in this current study may serve as valuable markers for constructing linkage maps, and employed to learn new gene functions. The study successfully used SNPs to construct a linkage map. It was however unfeasible to have our map anchored to a previously constructed *C. sinensis* map because of the lack of availability of any anchoring markers. Moreover, earlier studies made use of AFLP, RAPD and SSR markers to construct tea linkage maps, whereas the present study employed DArTseq markers. The information obtained in this study i.e. gene function annotation and DArTseq sequence alignment, compared to recent literature referencing the tea genome has been ground breaking, and has set a platform for further MAS research on tea breeding to be conducted. The results obtained in this work may aid tea breeders select parental clones with desirable DArTseq markers for breeding new tea cultivars with desirable traits.

3.12 REFERENCES

- Adkins, N. L., Hall, J. A. & Georgel, P. T. (2007). The use of quantitative agarose gel electrophoresis for rapid analysis of the integrity of protein–DNA complexes. *Journal of Biochemical and Biophysical Methods* 70(5): 721-726.
- Arbona, V., Flors, V., Jacas, J., García-Agustín, P. & Gómez-Cadenas, A. (2003). Enzymatic and non-enzymatic antioxidant responses of Carrizo citrange, a salt-sensitive citrus rootstock, to different levels of salinity. *Plant and Cell Physiology* 44(4): 388-394.
- Archana, I. & Vijayalakshmi, K. (2018). Antioxidant potential of Phloroglucinol; an in-vitro approach. *International Journal Of Pharmaceutical Sciences and Research* 9(7): 2947-2951.
- Arisz, S.A., van Wijk, R.V., Roels, W., Zhu, J.K., Haring, M.A. & Munnik, T. (2013). Rapid phosphatidic acid accumulation in response to low temperature stress in Arabidopsis is generated through diacylglycerol kinase. *Frontiers in Plant Science*, 4: 1.
- Balakirev, M. Y., Tcherniuk, S. O., Jaquinod, M., & Chroboczek, J. (2003). Otubains: a new family of cysteine proteases in the ubiquitin pathway. *EMBO Reports* 4(5), 517-522.
- Bali, S., Mangain, A., Raina, S. N., Yadava, S. K., Bhat, V., Das, S., Pradhan, A. K. & Goel, S. (2015). Construction of a genetic linkage map and mapping of drought tolerance trait in Indian beverage tea. *Molecular Breeding* 35(5): 112.
- Bonvehi, J. S. & Coll, F. V. (1997). Evaluation of bitterness and astringency of polyphenolic compounds in cocoa powder. *Food Chemistry* 60(3): 365-370.
- Chardin, C., Girin, T., Roudier, F., Meyer, C., & Krapp, A. (2014). The plant RWP-RK transcription factors: key regulators of nitrogen responses and of gametophyte development. *Journal of Experimental Botany* 65(19), 5577-5587.
- Chaturvedula, V. S. P. & Prakash, I. (2011). The aroma, taste, color and bioactive constituents of tea. *Journal of Medicinal Plants Research* 5(11): 2110-2124.
- Chen, L., Apostolides, Z. & Chen, Z.-M. (2012). *Global Tea Breeding: Achievements, Challenges and Perspectives*. Springer Science & Business Media.
- Cheruiyot, E. K., Mumera, L. M., Ng'etich, W. K., Hassanali, A., Wachira, F. & Wanyoko, J. K. (2008b). Shoot epicatechin and epigallocatechin contents respond to water stress in tea [*Camellia sinensis* (L.) O. Kuntze]. *Bioscience, Biotechnology and Biochemistry* 72(5): 1219-1226.
- Chugh, K. (2013). Measuring phenotypic and genetic variances and narrow sense heritability in three populations of annual ryegrass (*Lolium multiflorum* Lam.).

- Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138(3): 963-971.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18): 3674-3676.
- Cui, L., Yao, S., Dai, X., Yin, Q., Liu, Y., Jiang, X., Wu, Y., Qian, Y., Pang, Y. & Gao, L. (2016). Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (*Camellia sinensis*). *Journal of Experimental Botany* 67(8): 2285-2297.
- de la Torre, F., Cañas, R. A., Pascual, M. B., Avila, C., & Cánovas, F. M. (2014). Plastidic aspartate aminotransferases and the biosynthesis of essential amino acids in plants. *Journal of Experimental Botany* 65(19): 5527-5534.
- Doerge, R. W. (2002). Multifactorial genetics: Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3(1): 43.
- Dumas, M.-E., Maibaum, E. C., Teague, C., Ueshima, H., Zhou, B., Lindon, J. C., Nicholson, J. K., Stamler, J., Elliott, P. & Chan, Q. (2006). Assessment of analytical reproducibility of ¹H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. *Analytical Chemistry* 78(7): 2199-2208.
- Dunn, W. B., Goodacre, R., Neyses, L. & Mamas, M. (2011). Integration of metabolomics in heart disease and diabetes research: current achievements and future outlook. *Bioanalysis* 3(19): 2205-2222.
- Dutta, R., Stein, A. & Bhagat, R. (2011). Integrating satellite images and spectroscopy to measuring green and black tea quality. *Food Chemistry* 127(2): 866-874.
- Ebbels, T. M., Lindon, J. C. & Coen, M. (2011). Processing and modeling of nuclear magnetic resonance (NMR) metabolic profiles. In *Metabolic Profiling*, 365-388: Springer.
- El-Soda, M., Malosetti, M., Zwaan, B. J., Koornneef, M. & Aarts, M. G. (2014). Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends in Plant Science* 19(6): 390-398.
- Ellinger, J. J., Chylla, R. A., Ulrich, E. L. & Markley, J. L. (2013). Databases and software for NMR-based metabolomics. *Current Metabolomics* 1(1): 28-40.
- Ellis, R. & Nyirenda, H. (1995). A successful plant improvement programme on tea (*Camellia sinensis*). *Experimental Agriculture* 31(3): 307-323.

- Fang, H., Liu, X., Thorn, G., Duan, J. & Tian, L. (2014). Expression analysis of histone acetyltransferases in rice under drought stress. *Biochemical and Biophysical Research Communications* 443(2): 400-405.
- Feil, R. & Fraga, M. F. (2012). Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics* 13(2): 97.
- Gawel, N. (1991). A modified CTAB DNA extraction procedure for Musa and Ipomea. *Plant Mol. Biol. Report* 9: 292-296.
- Gonzalez, A., Brown, M., Hatlestad, G., Akhavan, N., Smith, T., Hembd, A., Moore, J., Montes, D., Mosley, T. & Resendez, J. (2016). TTG2 controls the developmental regulation of seed coat tannins in Arabidopsis by regulating vacuolar transport steps in the proanthocyanidin pathway. *Developmental Biology* 419(1): 54-63.
- Group, C. P. B., Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.-J., Liu, J.-Q., Chen, Z.-D., Zhou, S.-L. & Chen, S.-L. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences* 108(49): 19641-19646.
- Gupta, P. K., Rustgi, S. & Mir, R. R. (2013). Array-based high-throughput DNA markers and genotyping platforms for cereal genetics and genomics. In *Cereal Genomics II*, 11-55: Springer.
- Hackett, C. A., Wachira, F. N., Paul, S., Powell, W. & Waugh, R. (2000). Construction of a genetic linkage map for *Camellia sinensis* (tea). *Heredity* 85(4): 346.
- Hamanishi, E. T., Barchet, G. L., Dauwe, R., Mansfield, S. D. & Campbell, M. M. (2015). Poplar trees reconfigure the transcriptome and metabolome in response to drought in a genotype- and time-of-day-dependent manner. *BMC Genomics* 16(1): 329.
- He, M., Tian, H., Luo, X., Qi, X. & Chen, X. (2015). Molecular progress in research on fruit astringency. *Molecules* 20(1): 1434-1451.
- Henery, M. L., Moran, G. F., Wallis, I. R. & Foley, W. J. (2007). Identification of quantitative trait loci influencing foliar concentrations of terpenes and formylated phloroglucinol compounds in *Eucalyptus nitens*. *New Phytologist* 176(1): 82-95.
- Hoekstra, F. A., Golovina, E. A. & Buitink, J. (2001). Mechanisms of plant desiccation tolerance. *Trends in Plant Science* 6(9): 431-438.

- Hua, X., Wang, T., Chen, Y., Zhang, M., Chen, J., Liu, J. & Han, H. (2017). Arabidopsis amino acid permease1 contributes to salt stress-induced proline uptake from exogenous sources. *Frontiers in Plant Science* 8: 2182.
- Jeyaramraja, P., Pius, P., Raj Kumar, R. & Jayakumar, D. (2003). Soil moisture stress-induced alterations in bio-constituents determining tea quality. *Journal of the Science of Food and Agriculture* 83(12): 1187-1191.
- Jones, P., Messner, B., Nakajima, J. I., Schäffner, A. R. & Saito, K. (2003). UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in *Arabidopsis thaliana*. *Journal of Biological Chemistry* 278(45): 43910-43918.
- Kamunya, S., Wachira, F., Pathak, R., Korir, R., Sharma, V., Kumar, R., Bhardwaj, P., Chalo, R., Ahuja, P. & Sharma, R. (2010). Genomic mapping and testing for quantitative trait loci in tea (*Camellia sinensis* (L.) O. Kuntze). *Tree Genetics & Genomes* 6(6): 915-929.
- Kamunya, S., Wachira, F., Pathak, R., Muoki, R., Wanyoko, J., Ronno, W. & Sharma, R. (2009). Quantitative genetic parameters in tea (*Camellia sinensis* (L.) O. Kuntze): I. combining abilities for yield, drought tolerance and quality traits. *African Journal of Plant Science* 3(5): 093-101.
- Kang, J., Choi, M.-Y., Kang, S., Kwon, H. N., Wen, H., Lee, C. H., Park, M., Wiklund, S., Kim, H. J. & Kwon, S. W. (2008). Application of a ¹H nuclear magnetic resonance (NMR) metabolomics approach combined with orthogonal projections to latent structure-discriminant analysis as an efficient tool for discriminating between Korean and Chinese herbal medicines. *Journal of Agricultural and Food Chemistry* 56(24): 11589-11595.
- Kaplan, F. & Guy, C. L. (2004). β -Amylase induction and the protective role of maltose during temperature shock. *Plant Physiology* 135(3): 1674-1684.
- Kerchev, P. I., Fenton, B., Foyer, C. H. & Hancock, R. D. (2012). Plant responses to insect herbivory: interactions between photosynthesis, reactive oxygen species and hormonal signalling pathways. *Plant, Cell & Environment* 35(2): 441-453.
- Kerio, L., Wachira, F., Wanyoko, J. & Rotich, M. (2013). Total polyphenols, catechin profiles and antioxidant activity of tea products from purple leaf coloured tea cultivars. *Food Chemistry* 136(3-4): 1405-1413.
- Khan, N. & Mukhtar, H. (2007). Tea polyphenols for health promotion. *Life Sciences* 81(7): 519-533.

- Kim, J. M., Sasaki, T., Ueda, M., Sako, K. & Seki, M. (2015). Chromatin changes in response to drought, salinity, heat, and cold stresses in plants. *Frontiers in Plant Science* 6: 114.
- Kim, J. M., To, T. K., Ishida, J., Morosawa, T., Kawashima, M., Matsui, A., Toyoda, T., Kimura, H., Shinozaki, K. & Seki, M. (2008). Alterations of lysine modifications on the histone H3 N-tail under drought stress conditions in *Arabidopsis thaliana*. *Plant and Cell Physiology* 49(10): 1580-1588.
- Kobayashi-Hattori, K., Mogi, A., Matsumoto, Y. & Takita, T. (2005). Effect of caffeine on the body fat and lipid metabolism of rats fed on a high-fat diet. *Bioscience, Biotechnology, and Biochemistry* 69(11): 2219-2223.
- Kochhar, S., Kochhar, V. K., & Sane, P. V. (1998). Subunit structure of lysine sensitive aspartate kinase from spinach leaves. *IUBMB Life* 44(4), 795-806.
- Koehn, R. K., Malebe, P. M., Nyarukowa, C., Mose, R., Kamunya, S. M. & Apostolides, Z. (2018). Identification of novel QTL for black tea quality traits and drought tolerance in tea plants (*Camellia sinensis*). *Tree Genetics & Genomes* 14(1): 9.
- Koehn, K.R., Malebe, M. P., Nyarukowa, C. T., Mose, R., Kamunya, M. S. & Apostolides, Z. (2019). Functional annotation of putative QTL associated with black tea quality and drought tolerance traits. *Scientific Reports* 9: 1465.
- Kowalsick, A., Kfoury, N., Robbat Jr, A., Ahmed, S., Orians, C., Griffin, T., Cash, S. B. & Stepp, J. R. (2014). Metabolite profiling of *Camellia sinensis* by automated sequential, multidimensional gas chromatography/mass spectrometry reveals strong monsoon effects on tea constituents. *Journal of Chromatography A* 1370: 230-239.
- Lavarack, B., Griffin, G. & Rodman, D. (2002). The acid hydrolysis of sugarcane bagasse hemicellulose to produce xylose, arabinose, glucose and other products. *Biomass and Bioenergy* 23(5): 367-380.
- Le Gall, G., Colquhoun, I. J. & Defernez, M. (2004). Metabolite profiling using ¹H NMR spectroscopy for quality assessment of green tea, *Camellia sinensis* (L.). *Journal of Agricultural and Food Chemistry* 52(4): 692-700.
- Lesack, K. & Naugler, C. (2011). An open-source software program for performing Bonferroni and related corrections for multiple comparisons. *Journal of Pathology Informatics* 2.

- Ludwig, C. & Viant, M. R. (2010). Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques* 21(1): 22-32.
- Ma, J. Q., Yao, M. Z., Ma, C. L., Wang, X. C., Jin, J. Q., Wang, X. M. & Chen, L. (2014). Construction of a SSR-based genetic map and identification of QTLs for catechins content in tea plant (*Camellia sinensis*). *PloS one* 9(3): e93131.
- Mackay, T. F., Stone, E. A. & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10(8): 565.
- Marinova, K., Pourcel, L., Weder, B., Schwarz, M., Barron, D., Routaboul, J. M., Debeaujon, I. & Klein, M. (2007). The Arabidopsis MATE transporter TT12 acts as a vacuolar flavonoid/H⁺-antiporter active in Proanthocyanidin-accumulating cells of the seed coat. *The Plant Cell* 19(6): 2023-2038.
- Mizutani, M., Katsumoto, Y., Fukui, Y., Togami, J., Nakamura, N., Okuhara, H. & Tanaka, Y. (2006). An Acyltransferase involved in biosynthesis of polyacylated anthocyanin.
- Negishi, O., Ozawa, T. & Imagawa, H. (1992). Biosynthesis of Caffeine from Purine Nucleotides in Tea Plant. *Bioscience, Biotechnology and Biochemistry* 56(3): 499-503.
- Nitin Seetohul, L., Islam, M., O'Hare, W. T. & Ali, Z. (2006). Discrimination of teas based on total luminescence spectroscopy and pattern recognition. *Journal of the Science of Food and Agriculture* 86(13): 2092-2098.
- Nyarukowa, C., Koech, R., Loots, T. & Apostolides, Z. (2016). SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. *Journal of Plant Physiology* 198: 39-48.
- Nyarukowa C. T., Koech K. R., Loots T., Hageman J. & Apostolides Z. (2018). Prioritising the replanting schedule of seedling tea fields on tea estates for drought susceptibility measured by the SWAPDT method in the absence of historical in-filling records. *Journal of Agricultural Science* 10(7): 26-34.
- Obanda, M., Owuor, P. O. & Mang'oka, R. (2001). Changes in the chemical and sensory quality parameters of black tea due to variations of fermentation time and temperature. *Food Chemistry* 75(4): 395-404.
- Obanda, M., Owuor, P. O. & Taylor, S. J. (1997). Flavanol composition and caffeine content of green leaf as quality potential indicators of Kenyan black teas. *Journal of the Science of Food and Agriculture* 74(2): 209-215.

- Orel, G. & Wilson, P. G. (2012). *Camellia cherryana* (Theaceae), a new species from China. In *Annales Botanici Fennici*, Vol. 49, 248-255: BioOne.
- Owuor, P. O., Obanda, M., Nyirenda, H. E., Mphangwe, N. I., Wright, L. P. & Apostolides, Z. (2006). The relationship between some chemical parameters and sensory evaluations for plain black tea (*Camellia sinensis*) produced in Kenya and comparison with similar teas from Malawi and South Africa. *Food Chemistry* 97(4): 644-653.
- Park, C. J. & Seo, Y.-S. (2015). Heat shock proteins: a review of the molecular chaperones for plant immunity. *The Plant Pathology Journal* 31(4): 323.
- Paterson, A. H., Damon, S., Hewitt, J. D., Zamir, D., Rabinowitch, H. D., Lincoln, S. E., Lander, E. S. & Tanksley, S. D. (1991). Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics* 127(1): 181-197.
- Petrussa, E., Braidot, E., Zancani, M., Peresson, C., Bertolini, A., Patui, S. & Vianello, A. (2013). Plant flavonoids-biosynthesis, transport and involvement in stress responses. *International Journal of Molecular Sciences* 14(7): 14950-14973.
- Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H. & Sorrells, M. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome* 5(3): 103-113.
- Pongsuwan, W., Fukusaki, E., Bamba, T., Yonetani, T., Yamahara, T. & Kobayashi, A. (2007). Prediction of Japanese green tea ranking by gas chromatography/mass spectrometry-based hydrophilic metabolite fingerprinting. *Journal of Agricultural and Food Chemistry* 55(2): 231-236.
- Punyasiri, P., Abeysinghe, I., Kumar, V., Treutter, D., Duy, D., Gosch, C., Martens, S., Forkmann, G. & Fischer, T. (2004). Flavonoid biosynthesis in the tea plant *Camellia sinensis*: properties of enzymes of the prominent epicatechin and catechin pathways. *Archives of Biochemistry and Biophysics* 431(1): 22-30.
- Preedy, V. R. (2012). *Tea in health and disease prevention*. Academic Press.
- Qin, Z., Pang, X., Chen, D., Cheng, H., Hu, X. & Wu, J. (2013). Evaluation of Chinese tea by the electronic nose and gas chromatography–mass spectrometry: Correlation with sensory properties and classification according to grade level. *Food Research International* 53(2): 864-874.
- Raffler, J., Friedrich, N., Arnold, M., Kacprowski, T., Rueedi, R., Altmaier, E., Bergmann, S., Budde, K., Gieger, C. & Homuth, G. (2015). Genome-wide association study with

- targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genetics* 11(9): e1005487.
- Rawat, R., Gulati, A., Babu, G. K., Acharya, R., Kaul, V. K. & Singh, B. (2007). Characterization of volatile components of Kangra orthodox black tea by gas chromatography-mass spectrometry. *Food Chemistry* 105(1): 229-235.
- Robichaud, J. L. & Noble, A. C. (1990). Astringency and bitterness of selected phenolics in wine. *Journal of the Science of Food and Agriculture* 53(3): 343-353.
- Sang, S., Lambert, J. D., Ho, C.-T. & Yang, C. S. (2011). The chemistry and biotransformation of tea constituents. *Pharmacological Research* 64(2): 87-99.
- Sansaloni, C. P., Petroli, C. D., Carling, J., Hudson, C. J., Steane, D. A., Myburg, A. A., Grattapaglia, D., Vaillancourt, R. E. & Kilian, A. (2010). A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. *Plant Methods* 6(1): 16.
- Schouten, H. J., van de Weg, W. E., Carling, J., Khan, S. A., McKay, S. J., van Kaauwen, M. P., Wittenberg, A. H., Koehorst-van Putten, H. J., Noordijk, Y. & Gao, Z. (2012). Diversity arrays technology (DArT) markers in apple for genetic linkage maps. *Molecular Breeding* 29(3): 645-660.
- Schuh, C. & Schieberle, P. (2006). Characterization of the key aroma compounds in the beverage prepared from Darjeeling black tea: quantitative differences between tea leaves and infusion. *Journal of Agricultural and Food Chemistry* 54(3): 916-924.
- Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P. & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: communicating confidence. ACS Publications.
- Sedaghat, S., Haghghat, S. R. & Shokrgozar, S. (2013). Storage period effects on the qualitative characteristics of scented tea. *International Journal of Bioscience* 3(7): 66-73.
- Shitan, N. (2016). Secondary metabolites in plants: transport and self-tolerance mechanisms. *Bioscience, Biotechnology and Biochemistry* 80(7): 1283-1293.
- Son, H. S., Hwang, G. S., Kim, K. M., Kim, E. Y., van den Berg, F., Park, W. M., Lee, C. H. & Hong, Y. S. (2008). 1H NMR-based metabolomic approach for understanding the fermentation behaviors of wine yeast strains. *Analytical Chemistry* 81(3): 1137-1145.
- Steane, D. A., Nicolle, D., Sansaloni, C. P., Petroli, C. D., Carling, J., Kilian, A., Myburg, A. A., Grattapaglia, D. & Vaillancourt, R. E. (2011). Population genetic analysis and

- phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping. *Molecular Phylogenetics and Evolution* 59(1): 206-224.
- Tan, L. Q., Wang, L. Y., Xu, L. Y., Wu, L. Y., Peng, M., Zhang, C. C., Wei, K., Bai, P. X., Li, H. L. & Cheng, H. (2016). SSR-based genetic mapping and QTL analysis for timing of spring bud flush, young shoot color, and mature leaf size in tea plant (*Camellia sinensis*). *Tree Genetics & Genomes* 12(3): 52.
- Taniguchi, F., Furukawa, K., Ota-Metoku, S., Yamaguchi, N., Ujihara, T., Kono, I., Fukuoka, H. & Tanaka, J. (2012). Construction of a high-density reference linkage map of tea (*Camellia sinensis*). *Breeding Science* 62(3): 263-273.
- Upadhyaya, H. & Panda, S. K. (2013). Abiotic stress responses in tea [*Camellia sinensis* L (O) Kuntze]: an overview. *Reviews in Agricultural Science* 1: 1-10.
- Van Ooijen, J. (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. *Kyazma BV, Wageningen* 33(10.1371).
- Vankatesh, P., Jaiprakash, M., Prasad, P., Pilia, B., Sadhale, P. P. & Sinka, V. (2007). Flavanoid biosynthesis in *Camellia sinensis*.
- Ververidis, F., Trantas, E., Douglas, C., Vollmer, G., Kretschmar, G. & Panopoulos, N. (2007). Biotechnology of flavonoids and other phenylpropanoid-derived natural products. Part I: Chemical diversity, impacts on plant biology and human health. *Biotechnology Journal* 2(10): 1214-1234.
- Wang, K. & Ruan, J. (2009). Analysis of chemical components in green tea in relation with perceived quality, a case study with Longjing teas. *International Journal of Food Science & Technology* 44(12): 2476-2484.
- Wang, W., Xin, H., Wang, M., Ma, Q., Wang, L., Kaleri, N. A., Wang, Y. & Li, X. (2016). Transcriptomic analysis reveals the molecular mechanisms of drought-stress-induced decreases in *Camellia sinensis* leaf quality. *Frontiers in Plant Science* 7: 385.
- Wei, F., Furihata, K., Miyakawa, T. & Tanokura, M. (2014). A pilot study of NMR-based sensory prediction of roasted coffee bean extracts. *Food Chemistry* 152: 363-369.
- Wittenberg, A. H., Van Der Lee, T., Cayla, C., Kilian, A., Visser, R. G. & Schouten, H. J. (2005). Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. *Molecular Genetics and Genomics* 274(1): 30-39.
- Wright, L. P., Mphangwe, N. I. K., Nyirenda, H. E. & Apostolides, Z. (2002). 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* 82(5): 517-525.
- Xu, Y. Q., Zhang, Y. N., Chen, J. X., Wang, F., Du, Q. Z. & Yin, J. F. (2018). Quantitative analyses of the bitterness and astringency of catechins from green tea. *Food Chemistry* 258: 16-24.
- Xu, L., Qi, T., Xu, L., Lu, L. & Xiao, M. (2016). Recent progress in the enzymatic glycosylation of phenolic compounds. *Journal of Carbohydrate Chemistry* 35(1): 1-23.
- Yan, S.-H. (2007). NIR evaluation of the quality of tea and its market price. *Spectroscopy Europe* 19(2): 16-19.
- Yang, C., Hu, Z., Lu, M., Li, P., Tan, J., Chen, M., Lv, H., Zhu, Y., Zhang, Y. & Guo, L. (2018). Application of metabolomics profiling in the analysis of metabolites and taste quality in different subtypes of white tea. *Food Research International* 106: 909-919.
- Yao, M., Chen, L. & Liang, Y. (2008). Genetic diversity among tea cultivars from China, Japan and Kenya revealed by ISSR markers and its implication for parental selection in tea breeding programmes. *Plant Breeding* 127(2): 166-172.
- Yoshida, T. & Sakamoto, T. (2009). Water-stress induced trehalose accumulation and control of trehalase in the cyanobacterium *Nostoc punctiforme* IAM M-15. *The Journal of General and Applied Microbiology* 55(2): 135-145.
- Zhang, J., Wang, X., Yu, O., Tang, J., Gu, X., Wan, X. & Fang, C. (2010). Metabolic profiling of strawberry (*Fragaria* × *ananassa* Duch.) during fruit development and maturation. *Journal of Experimental Botany* 62(3): 1103-1118.
- Zhang, H., Zhu, H., Pan, Y., Yu, Y., Luan, S. & Li, L. (2014). A DTX/MATE-type transporter facilitates abscisic acid efflux and modulates ABA sensitivity and drought tolerance in *Arabidopsis*. *Molecular Plant* 7(10): 1522-1532.
- Zhu, J. K. (2016). Abiotic stress signaling and responses in plants. *Cell* 167(2): 313-324.
- Cheruiyot, E., Mumera, L., Ng'etich, W., Hassanali, A. & Wachira, F. (2008a). Threshold soil water content for growth of tea [*Camellia sinensis* (L.) O. Kuntze]. *Tea* 29(2): 29-38.
- Zou, J., Semagn, K., Iqbal, M., N'Diaye, A., Chen, H., Asif, M., Navabi, A., Perez-Lara, E., Pozniak, C. & Yang, R.-C. (2017). Mapping QTLs controlling agronomic traits in the 'Attila' × 'CDC Go' spring wheat population under organic management using 90K SNP array. *Crop Science* 57(1): 365-377.

Appendix 3.1: DArTseq marker ID with the commensuating separate marker sequence

Marker ID	Marker Sequence (5'-3')
5063001	TGCAGCCTTATTAGTTTTACGTTACGTATGAAAACTATGCAATTGTCGAATTTGTTTAGGAGCTCCCA
5064391	TGCAGCCATAGTTTGAAGAACAGCAGAGTTTTTGACCCAAGAATCATGCTCTTAGTTTACAATCTGTAT
5064585	TGCAGTTCAACATCATCAGCACTTTTAGTTTTAGACAACACATTTATTGACCAAAGCTGAGTGCCATG
5072338	TGCAGCTCTGATTCACTTGTTTCTCAATATTTGAACTTCCACGGTATGAAGGTATATATTTCAAATATG
5073424	TGCAGAGAAATTCCTCCACACACCGAGGCTCGTTTGGGAAGTAACGTTTTTACTAAAAAAAAAATTACAG
5082606	TGCAGCAAGCAAAAAAAGTGGAATCCCTCGGAACACTGGACTCGCCGAGGCTAGTCAGAGCCTCTTTTG
5084595	TGCAGTTGAATCTGTAAGAGTGAGACACCCATTAGGCACCCAATAACTTTAGAAATTCAGAAGAAAAAT
5085963	TGCAGCTTGTGGAGCTGCTTATAACTTTTTATTTTGAGAATCGGTGTCAGAGTGTTTGGTGCATCTTTT
5087017	TGCAGTAACAAACATTATAATTTTTGTGTTGCATTATAAAAAGCAAGACAATACTAATCAAGCCTTGATT
5087113	TGCAGCGGTTGATGCTGTACGTTTACTCACAGCAGCATCTCCTATCCCGACAAGATGGGTTGACTTTTA
5088162	TGCAGATTTTTTAGGACAACAAGATTCTCCAAGGTCTGAATAATTATGAGAACAAAAATATGCATTCAC
5098382	TGCAGGGCATCAAACAGGACATTAGTCACAAGTAATTAGCATAAGCAACGGATACTATAATTACAGATC
5104630	TGCAGCTGTAGATGAGGAGGTACGCTAGTCTAATTTTAGAGGAATAAAGATGGAGAGTTTTACAGATCG
5111164	TGCAGTACCAGATGGACAAGGAGAGAGAAGCGAGTTAGAGGTCTTTACAGATCGGAAGAGCGGTTTCAGC
5111268	TGCAGAGGTAATCTACACCTATCAGTTCACTTCCCATATGTCGTTGAGTACAAACAAGGACACAAAGTT
5112599	TGCAGGAAATCCGGCATTTAGAGGCATTTACAGATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGA
5114089	TGCAGTCCTTTTTGATTGTTTCTGATCCTTCCATCACGAAACATATATTACAGATCGGAAGAGCGGTTT
5114985	TGCAGCAAACACTTGCTAATTTTCACTCTATCTTACCAAATGAGGCATTACAGATCGGAAGAGCGGTT
5123053	TGCAGGTCATTTTATAATGTGACACATCGGAGATTACAGATCGGAAGAGCGGTTTCAGCAGGAATGCCGA

5123475 TGCAGAGCAATTACCACAGTTTATCGCTGCTGGCAACTTACAGATCGGAAGAGCGGTTTCAGCAGGAATG
5123751 TGCAGCTGCCTTTCGTCATGTGGCTTGCCATCAAGAAACCAGAGAAGTTTAGCTTGTCTTGGCTTATTA
5123950 TGCAGCCATTGGTGCCTACCTGGAGTTCATACACATGGAATGAACTTCTGTATGCTCACCTACTGTCA
5124128 TGCAGCTAGAATAATGTCTACCCACAATCAAATATAGATCTTACAGATTATTACAGATCGGAAGAGCGG
5125626 TGCAGTCGGGCCACTTTTTTTTTAGTTGGTGTGAATGTCCCGTAATTTATTACAGATCGGAAGAGCGGTT
5128967 TGCAGAAGAAAGGCACACATGAGCAGCTGTGATGGAATAACCCAGTTGCAATCTACCACATATTACAGA
5130194 TGCAGTGGTGGAAAGCTTGTACATCTGGAGTTTTACAGATCCGAAGAGCGGTTTCAGCAGGAATGCCGAGA
5130531 TGCAGTCAGCATAACCGGTGCAGCTATCTGCCACATCTTCTTTGATTCATCTCGACATTTTCTGTTACA
5132370 TGCAGCTGTCCTTTCACCAAGTAACGCATATAAGTTTGCATCGACGAGTTGCTATTCCAAATAAGAGCA
5132432 TGCAGGTTTCTTCTTTCACCCAAAAAAAAAAGCCCTCATTTTGGTTCAATTTTACAGATCGGAAGAGCGG
5132791 TGCAGCAATTTGCTAATAACCTCCTCATCCTTGCCTTCAAATCAATGCCTTCAAAAACAAATCTCATA
5134558 TGCAGCTGAAGTGGTAAAGAGGCTTGTTGAAGGGGCTCAACTCCTCATCCATCAAAGATTTACAGATCG
5135436 TGCAGGAAAGGCAAGGGAAATAACAACAAATAATTACAGATCGGAAGAGCGGTTTCAGCAGGAATGCCGA
513647 TGCAGAGCAATTACCTGGTGTGAATGTCCCGTAATTTTAGAGGCATTTACAGATCGGAAGAGCGGGCAG

CHAPTER 4

MODELS FOR IDENTIFICATION OF ELITE MOTHER BUSHES WITH HIGH BLACK TEA COMMERCIAL POTENTIAL FROM MATURE SEEDLING FIELDS OF *CAMELLIA SINENSIS*

ABSTRACT

Tea (*Camellia sinensis*) has enthralled both consumers and researchers, and its popularity has increased, due to its taste, aroma and its medicinal attributes, owed largely to its metabolites. The catechins and theaflavins in green and black tea respectively have been documented to possess antioxidant, anti-inflammatory, anticancer, and cardiovascular disease preventing properties (Preedy, 2012). Tea consumers concern themselves with the quality of tea its taste and aroma; it is based on these that consumers will pay premium prices for the best quality teas. The quality of tea is undeniably affected by variations in its metabolite composition (Qin *et al.*, 2013). Tea producers are in demand of new high yielding, DT cultivars, which produce high quality tea liquors. To breed for these phenotypic traits is challenging due to these being quantitative traits, controlled by many genes, inherited from parents, and influenced by environment. In this study, two groups of black tea cultivars, one commercial (Comm) and the second non-commercial (NComm), were compared using a metabolomics approach. Data were generated via untargeted GC-MS and LC-MS; semi-targeted ¹H-NMR, and targeted UPLC-DAD. PCA and PLS-DA were performed on the metabolomics data, which showed clear separation and clustering between the Comm and NComm cultivars. Several logistic regression models were developed and it was found that the model based on the UPLC-DAD theaflavins worked best, with the model based on UPLC-DAD caffeine and the five catechins classifying the 303 genotypes as either Comm or NComm cultivars, worked equally well.

4.1 INTRODUCTION

Tea has been documented to be the most consumed non-alcoholic beverage worldwide, second only to water. Tea quality is undeniably affected by variations in its metabolite composition, which determine its commercial market value. According to Le Gall *et al.* (2004), the taste of green tea is determined by the cultivar of tea tree, the season of plucking, as well as the method of cultivation employed. Tea producers are in demand of new cultivars, which are high yielding, are drought tolerant, and produce high quality tea liquors. Tea gets its distinctive astringent and somewhat bitter taste from caffeine, 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). The amino acid theanine, which makes up approximately two-thirds of a tea leaf's total free amino acids content, is together with other less abundant amino acids, responsible for the sweet and brothy "umami" taste of green tea. 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 is dependent on amino acids, particularly theanine; catechins and caffeine, the quality of black tea is dependent on theaflavins and thearubigins, its major phenolics, which result from the dimerisation and polymerisation of the catechins. The major theaflavins are theaflavin (TF1), theaflavin-3-gallate (TF2), theaflavin-3'-gallate (TF3), and theaflavin-3,3-digallate (TF4) (Le Gall *et al.*, 2004). The four TFs are synthesised by polyphenol oxidase in *C. sinensis* leaves from the combination of green tea catechins, 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 *C. sinensis* cultivars rich in catechins are likely to produce higher quality teas (Takemoto and Takemoto, 2018).

Tea's popularity as a beverage is dependent on its flavour, comprising of taste and aroma. Non-volatile organic compounds are responsible for its taste, while volatile organic compounds are responsible for its aroma. Volatile organic compounds in tea fall into one of two groups, with Group I comprising of non-terpenoids, such as hexanol which confers the fresh green aroma; and Group II comprising of terpenoids, responsible for its sweet flowery aroma (Chaturvedula and Prakash, 2011). High-quality black teas are rich in Group II compounds and due to their flowery nature, achieve significantly higher prices; e.g. high-

quality varieties of Longjing tea (pan-roasted green tea) sell for USD 16.5, while the low-quality varieties sell for USD 1 per Kg (Yu *et al.*, 2008).

It has been reported that areas with higher rainfall tend to produce teas of inferior quality and that black teas from South India have elevated concentrations of aroma causing metabolites during the dry season as compared to the rainy season. In these studies, 40 or fewer compounds were used to classify tea quality. The profiling of plant metabolites has developed into a major metabolomics field of study, the reason being that plants manufacture a wide array of metabolites. The genetic improvement of crops with metabolomics is fast becoming a popular method; this has resulted in an increased demand for plant breeders skilled in the field of metabolomics. When developing new cultivars, crop breeders encounter a common challenge of identifying important selection criteria. Tea breeders criteria of selection include but not limited to yield, quality, and drought tolerance. *C. sinensis* is an important cash crop for many countries with China, India, Kenya and Sri Lanka being leading world producers and exporters of black tea (ITC, 2019). According to the Kenya National Bureau of Statistics (2019), tea is the largest agribusiness in Kenya, with the total export volume of January 2019 being significantly higher, at 47.92 Million Kg compared to the January 2018 total export volume of 31.94 Million Kg. It is for this reason that tea quality is an important selection criterion from an economic perspective as it is the major determinant of market price. Tea quality, whether for black, white, purple or green tea, is governed by the metabolic profile/composition of the tea leaves, influencing its aroma, briskness, brightness and taste (Dutta *et al.*, 2011); the agronomic traits i.e. yield and quality are dependent on leaf physiognomies. As previously noted, due to the effects of global warming, specifically altered precipitation patterns, elevated temperatures and protracted drought spells in the tea growing regions, the Kenyan tea industry has been facing challenges. It is for this reason that rigorous breeding programmes need to be developed to produce new cultivars with better metabolic profiles and improved drought tolerance.

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) following the 1980 formation of Tea Research Foundation of Kenya (TRFK), now known as the TRI. Other large tea producing companies such as James Finlay (Kenya) and George Williamson (Kenya) followed suit and instituted programmes that saw the establishment of their own improved

seed baries. 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 morphological consistency. Moreover, the developed progenies hadn't 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). As of 2006, approximately 60% of clones associated with TRFK 6/8 have been commercialised, stemming from the Timbilil tea estate's breeding programme. Furthermore, 24 out of the 45 developed clones have found success in industry, amongst which are the elite Cambod varieties, TRFK 301/4 and TRFK 301/5. In addition to these, are the clones TRFK 430/90 and TRFK 371/3, which in addition to them having high yield and improved black tea quality, these new cultivars possess biotic and abiotic stress tolerance properties (Kamunya and Wachira, 2006). Breeders have used the TRFCA SFS 150 clone from Malawi and the TRFK 303/577 to produce varieties that are drought tolerant, such as the EPK TN 14-3, and have crossed the TRFCA SFS 150 and EPK TN 14-3 to produce F1 progeny tolerant to cold (Kamunya *et al.*, 2010). Plant breeders have been finding it daunting to develop high yielding clones from seedling mother bushes. Earlier studies (Green, 1971) failed to establish reliable correlations between growth and yield properties of mother bushes, and their resultant F1 progeny clones. Subsequent studies (Nyirenda, 1991) have shown adequately strong correlations between the tea bush area, shoot number, and yield of tea mother bushes and those of their clones. A strong positive correlation between seedling height, leaf area, stem girth, stem dry weights and yield in matured tea fields was observed (Shanmugarajah *et al.*, 1991). Due to 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 region (Elbehri *et al.*, 2015). Due to these changes in the climate, tea production is drastically being 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. 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). 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). The cultivation of tea has now been extended to previously deemed marginal and unsuitable tea growing areas (Owuor *et al.*, 2010).

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 poise have a success rate of about 1% when it comes to identifying elite mother bushes that become commercial successes. The tea industry is in need of new methods for field selections to increase this success rate. Metabolomics has been defined as “*the study of the quantitative measurement of the dynamic multi-parametric metabolic response of biological system and changes in metabolite concentrations or fluxes related to genetic or environmental perturbations*”. It is a discipline which assesses, classifies and quantifies endogenous and exogenous metabolites in a variety of biological samples. The information obtained from the study of metabolites is crucial in that it informs scientists about a biological system’s functional state, explaining the organism’s phenotypic traits (Schauer and Fernie, 2006). Comprehending the desiccation response metabolome assists in ascertaining steps involved in the signal transduction pathways (Urano *et al.*, 2009). Metabolic profiling commenced as a diagnostics tool to ascertain herbicide mode of action, and has since grown to include functions such as determining the differences between genetically modified and conservative crops, and genotyping them to discover new genes (Hagel and Facchini, 2008). 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. Nuclear magnetic resonance ($^1\text{H-NMR}$) spectroscopy and gas chromatography mass spectrometry (GC-MS) were two of the analytical metabolomics platforms used in this study. GC-MS, though tedious in the sample preparation stage, has a higher sensitivity as compared to $^1\text{H-NMR}$, capable of detecting metabolites with concentrations lower than the limit of detection of $^1\text{H-NMR}$. $^1\text{H-NMR}$ has the advantage of having simple sample preparation, as well as being semi-quantitative and non-destructive), but limited by resolution and the availability of plant metabolites in compound databases. Ultra-performance liquid chromatography coupled with a diode array detector (DAD) and mass spectrometry (LC-MS) were other platforms used in the study. Due to its advancements within the field, UPLC is a central technique in metabolomics research (Khan and Mukhtar, 2007), with it being used predominantly in

differential profiling and biomarker identification (Theodoridis *et al.*, 2012). 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). LC-MS has become a method of choice for profiling metabolites in complex biological samples, i.e. plant metabolomics samples (Zhou *et al.*, 2012).

In metabolomics, uni- and multivariate statistical techniques are used in combination to help pinpoint relevant variation (e.g. between groups 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. 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). 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). Decision trees are created through the use of partition algorithms. These algorithms employ the links between predictors and their corresponding responses, and recursively partition the data, splitting predictors until the desired prediction response is obtained. Through these repeated

data partitions, a decision tree is formed. By choosing the best splits from an infinite number of possibilities, the partition algorithm makes the decision trees a powerful modelling tool. Predictors are either continuous or categorical; where continuous, the partitions are a result of a cut off value, with sample values falling above and below this cut off value. If, on the other hand, the predictor is categorical, the samples will be split into two levels (JMP®). The decision tree identifies independent variables with a significant relationship to the dependent variable and evaluates the continuous variables' interval breaks to identify the most ideal combination. The independent variable possessing the sturdiest relationship with the dependent variable then becomes the decision tree's first branch; each significantly different category, relative to the target variable becomes the leaf. This is continually done to identify each leaf's significant predictor variable until predictors are exhausted (Thomas and Galambos, 2004).

Logistic regression (LR) is a statistical analysis tool generally suitable for testing hypotheses regarding connections between categorical outcome variables and continuous predictor variables. LR solves problems that cannot be solved by simple linear regression, such as any occurring errors that are not normally distributed or are not constant throughout the data range (Peng *et al.*, 2002). Contrasting from discriminant analysis, LR does not make the assumption that the predictor variables possess equal covariance matrices, and that these are normally distributed. It instead makes the assumption that the distributions of any errors equalling the true Y value subtracting the predicted Y value are described by the binomial distribution. This implies an identical probability is maintained across the range of predictor values. This binomial assumption is therefore easily testable using a Z-test (Siegel and Castellan, 1956). LR may be considered robust, provided the samples are random; in so doing this ensures the observations remain independent of one another (Peng *et al.*, 2002).

Another useful statistical analysis approach employed in this study were violin plots, which are a statistical method considered to be a combination of the box plot and kernel density plot, which are used for plotting numeric data. The violin plot contains the same information as would be found in a box plot, but have the indisputable advantage over the box plot in that they show the entire data distribution, which is beneficial when working with multimodal data i.e. distribution with several peaks (Hintze and Nelson, 1998).

4.2 RESEARCH OBJECTIVES

1. To use data generated through untargeted GC-MS, and semi-targeted ¹H-NMR metabolomics platforms to identify metabolites, which were expressed differently in the Comm and NComm cultivars.
2. To make use of UPLC-DAD generated targeted metabolomics data to develop LR models and CHAID decision trees, to classify the 303 genotypes as either Comm or NComm cultivars. The best model may then serve in predicting whether a new field selection is likely to become commercialised due to its similarities with the Comm cultivars.
3. To use untargeted LC-MS data to identify any additional metabolites not detected by platforms mentioned in objectives 1 and 2, to distinguish between the Comm and NComm cultivars.

4.3 HYPOTHESIS

Null hypothesis (H_0): There will be no statistically significant difference between the metabolite profiles and metabolite concentrations detected by all the metabolomics platforms employed between the Comm and NComm cultivars, at the 95% confidence interval.

4.4 MATERIALS AND METHODS

4.4.1 Plant material

All these NComm materials were vegetatively propagated and planted in two sites (Kericho (0.3689° S, 35.2863° E) and Kirinyaga (0.6591° S, 37.3827° E), Kenya, in 15-bush plots at 1.22 m and 0.62 m between and within plots, respectively, and replicated thrice following a complete randomised block design (Koech *et al.*, 2018). Sixty open-pollinated cultivars, pre-selected for their high yield, and good tea liquor since the 1950s formed the Comm group. These cultivars were vegetatively propagated by stem cuttings from elite mother bushes. Each Comm cultivar was cultivated in over 10 Hectares with about 10 000 trees per Ha. The NComm group of 250 cultivars were the F₁ progeny of a reciprocal cross between two heterozygous parental clones TRFK 303/577 and GW Ejulu. The NComm cultivars were various clonal materials code-named TRFK St. 504 (TRFK303/577 (♂), GW Ejulu (♀)) with 106 progeny and TRFK St. 524 (GW Ejulu (♀), TRFK 303/577(♂)) with 144 progeny, which were developed at the TRI of Kenya (Koech *et al.*, 2018). The GW Ejulu clone produces high-quality black tea, with high total catechins and moderate caffeine content; it is, however, a low-yielding and drought-susceptible clone. TRFK 303/577, on the other hand, is a high yielding, drought tolerant (DT) clone, which produces medium-quality black tea, with moderate levels of caffeine and total catechins. Each NComm cultivar was previously tested for yield and quality but was found unworthy of commercialisation. From the 303 cultivars, eight Comm and eight NComm cultivars were selected at random, and prepared for UPLC-MS analysis.

4.4.2 Sample collection and processing

About 500 grams fresh shoots comprising two leaves and a bud were harvested from the respective tea bushes, between September 2013 and February 2014. The fresh shoots were placed in appropriately labelled zip-lock plastic bags (Nyarukowa *et al.*, 2018). The plastic bags were placed on ice blocks to keep cool until processing at the TRI mini-factory within 24 hours. Half the shoots of each sample were freeze-dried and ground using a coffee grinder, sieved using a 355 µm sieve, sealed in zip-lock plastic bags and stored at 4°C in a fridge until analysis. The dried green tea samples are referred to as “green tea” in this thesis. The other half was used to make black tea according to Koech *et al.*, (2018). Briefly, the leaves were withered to a % relative water content of 68 - 72% over an 18 hour period before being passed through 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 zip-lock plastic bags and stored in 4°C fridge until UPLC analysis.

4.4.3. GC-MS sample preparation and analysis

4.4.3.1 Sample preparation

A 70% MeOH solution was used for extraction. For all samples of approximately 150 mg, 1.5 mL extraction solution was added. The samples were vortex mixed and incubated for 10 minutes at 70°C. The samples were vortex mixed halfway through the incubation period as well as at the end. After cooling, the samples were centrifuged for 5 minutes at 6000 g and one mL supernatant transferred to GC vials before drying under nitrogen. The dried samples were derivatised by adding 120 µl methoxyamine (10 mg/mL in pyridine) and incubated for 1 hour at 60°C; followed by the addition of 80 µl BSTFA (containing 1% trimethylchlorosilane) and incubated for another hour at 60°C. Samples were transferred to inserts before GC-MS analysis. Pooled quality control (QC) samples were prepared, and these underwent the same extraction and derivatisation procedures as the samples.

4.4.3.2 GC-MS analyses

Analyses were performed on a GC-TOF-MS system, comprising of an Agilent 7890A GC front-end system with an Agilent 7693 autosampler and a Leco Pegasus HT TOFMS. Hydrogen was used as carrier gas at a flow-rate of 1.8 mL/min; 0.2 µl sample was injected in splitless mode (allowing 30s purge delay). The inlet temperature was kept at 250°C. Compounds were separated on a Restek RX-1MS column (20 m x 180 µm x 0.18 µm). The transfer line and source temperatures were 250 and 200°C, respectively. Solvent delays of 200 s were allowed where after masses (50 – 800 m/z) were recorded at 20 spectra/sec. Universal EI settings were used for ionisation while the detector was operated at 50 V above tune voltage.

4.4.4 ¹H-NMR sample preparation and analysis

4.4.4.1 ¹H-NMR buffer solution

A 1.5 M KH₂PO₄ buffer solution was prepared by dissolving 20.4 g of KH₂PO₄ in 80 mL of deuterium oxide (D₂O). Next, 13 mg of sodium azide and 100 mg of trimethylsilyl-2,2,3,3-tetradeuteriopropionic acid (TSP) were dissolved in 10 mL of D₂O and added to KH₂PO₄

solution. The combined solution was mixed well under sonication before adjusting the pH to 7.4 using potassium hydroxide in H₂O. The final solution was then transferred to a 100 mL volumetric flask and the volume topped up to the mark using D₂O.

4.4.4.2 ¹H-NMR sample preparation

Freeze-dried samples were sent in individual plastic bags of 50 mg weight to the ¹H-NMR lab. A pooled QC sample was created by collecting 5 mg from each of n=294 samples. Samples were prepared by adding 4.5 mL ddH₂O to each 45 mg weight of the dry sample to create a 10 mg/mL concentration. Each sample was vortexed at 0, 20 and 40 minutes. At 60 minutes, a volume of 540 µL of the sample was collected in a microcentrifuge tube, with 60 µL ¹H-NMR buffer solution. The sample was mixed under vortex and centrifuged at 12 000 g for 5 minutes to sediment any particulates. A final volume of 540 µL of supernatant was carefully transferred to a 5 mm ¹H-NMR glass tube and loaded onto an autosampler for ¹H-NMR analysis.

4.4.4.3 ¹H-NMR analyses

The samples were measured at 500 MHz on a Bruker Avance III HD NMR spectrometer equipped with a triple-resonance inverse (TXI) ¹H{¹⁵N, ¹³C} probe head and x, y, z gradient coils. ¹H spectra were acquired as 128 transients in 64 K data points with a receiver gain of 64 and a spectral width of 10 000 Hz. The sample temperature was maintained at 300K and the H₂O resonance was presaturated by single-frequency irradiation during a relaxation delay of 4 s, with a 90° excitation pulse of 8 µs. Shimming of the sample was performed automatically on the deuterium signal. The resonance line widths for TSP and metabolites were <1 Hz. Fourier transformation and phase and baseline correction were done automatically. Software used for ¹H-NMR processing was Bruker Topspin (V3.5). Bruker AMIX (V3.9.14) was used for metabolite identification and quantification.

4.4.5 UPLC-DAD and UPLC-MS sample preparation and analysis

4.4.5.1 Extraction of catechins, caffeine, and theaflavins

Samples were collected, and metabolites extracted from the tea samples as documented in the International Organisation for Standardisation (ISO) extraction procedure, described in document ISO14502-2 (2005). Briefly, amounts of 0.200 ± 0.001 g of green and black tea samples were weighed out using a Mettler Toledo model XS205DU 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 2000 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, 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 and UPLC-MS. Twenty Comm and 20 NComm cultivars were randomly selected and analysed using UPLC-MS to identify additional metabolites not detected by the other metabolomics platforms.

4.4.5.2 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 40°C, 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% 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 used as 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. The internal QC standards were also identified based on their retention times and UV/vis spectra.

4.4.5.3 UPLC-MS analyses

The high-resolution UPLC-MS analyses were performed using a Waters Synapt G2 Quadrupole time-of-flight (QTOF) MS connected to a Waters Acquity UPLC (Waters, Milford, MA, USA). Electrospray ionisation was employed in negative mode, with cone voltage set at 15 V, and the desolvation temperature set at 275°C. The desolvation gas was set at 650 L/h, with all other MS settings optimised to obtain the best resolution and sensitivity. The data was acquired by scanning from 150 m/z to 1500 m/z in both resolution mode and MS^E mode. The ESI capillary voltage was set at 3.1 kV. Two MS data channels were acquired in MS^E mode, the first at a low collision energy (4 V) and the second at a ramped collision energy (40–100 V) allowing for the collection of fragmentation data. Leucine enkephalin was employed as the lock mass (reference mass) for accurately determining the masses. Instrument calibration was performed using sodium formate. A Waters HSS T3, 2.1 × 100 mm, 1.7 µm column was used for the chromatographic separation. The mobile phases consisting of deionised double distilled H₂O with 0.1% formic acid (solvent A) and acetonitrile containing 0.1% formic acid (solvent B). The gradient used started off with 0% B for 1 min and increased to 28% B over 20 min, before increasing to 40% B in 1 min then finally to 100% B over 2 min, where it was held isocratic for 1.5 min, followed by re-equilibration to initial conditions for 4 min. A flow rate of 0.3 mL/min, injection volume of 2 µL, and a column temperature of 55°C were used.

4.4.6 Metabolite identification

Spectral matching to the NIST11 commercial library (for GC-MS metabolites) and Bruker BBIORFCODE (pH 7.0) and in-house pure compound spectral libraries (pH 7.4) (for ¹H-NMR metabolites) were used to identify the compounds. The UPLC-DAD metabolites were identified using pure standards and an in house library based on retention times. A level 2 identity was awarded when a spectral match of 80% similarity was achieved. A level 1 identity was awarded when the retention time or retention index of the GC-MS, and UPLC-DAD information matched that of standards or 2D ¹H-NMR information confirmed 1D ¹H-NMR spectral identifications. The potential UPLC-MS biomarkers were identified and confirmed by comparing their mass spectra and retention times against the reference standards. A full spectral library, comprising of MS/MS data in both positive and negative ionisation modes, was obtained. The MassFragment™ application manager (Waters MassLynx v4.1, Waters corp., Milford, USA) was employed to facilitate the MS/MS fragment ion analysis process by way of chemically intelligent peak-matching algorithms.

This information was then submitted for database searching, either in-house or using the online MassBank (<http://www.massbank.jp/>) data source.

4.4.7 Data pre-processing

4.4.7.1 GC-MS, ¹H-NMR and UPLC-DAD

ChromaTOF software (Leco) was used to perform data extraction for the GC-MS data, which included baseline removal using the “spanning” tracking method, with an offset of 1. The software performed automatic smoothing. An expected peak width of 3 s and signal-to-noise ratio of 20 was used to detect the peaks with five apexing masses. GC-MS data was normalised using the “total useful signal” correction method. A subset of the data was aligned for exploratory statistical analysis since the add-on function of ChromaTOF (Statistical Compare) is unable to align > 250 samples. The subset consisted of approximately 140 randomly selected Comm and NComm cultivar samples from every batch (including QCs). With the exploratory statistics, a list of compounds that differed between the groups was generated, which was used to create a reference chromatogram within ChromaTOF. The reference was used to extract the peaks of interest from all the samples in a “targeted” manner. The target peaks lists were aligned into a data matrix with MS Excel using the “consolidate” function. The reason for pre-processing was to transform the data to enhance ease and improve the data analysis. ¹H-NMR spectra pre-processed involved binning and scaling, with bins spanning 0.04 – 0.05 ppm. ¹H-NMR variables were scaled relative to the internal standard (TSP) by dividing each bin by the corresponding TSP value for the same sample. Next, the combined GC-MS, ¹H-NMR and UPLC-DAD variables with more than 10% missing values in both groups were eliminated; if two variables had a high correlation, one was removed; outliers were removed. The remaining missing values for each group, deemed to be below the quantification threshold of the instrument, were imputed with random numbers drawn from a uniform distribution between one and two-thirds of the lowest non-zero observations. Imputations were performed for each variable independently.

4.4.7.2 UPLC-MS

The acquired UPLC-MS data was processed using the MarkerLynx™ version 4.1 software. The processing steps included filtering, peak detection, peak alignment and normalisation. The function one data for all the samples was processed; this is the data collected at low energy and as such does not include ion fragmentation data as only the mother ions of each metabolite are detected. MarkerLynx software parameters were set to process the 1–13 min retention time (Rt) range of the chromatograms, mass range 100–1000 Da, mass tolerance

0.01 Da, mass window 0.05 Da and a Rt window of 0.20 min. Only data matrices that had noise levels less than 50% (MarkerLynx cut off) were retained for downstream chemometric and statistical analyses. Mandatory data scrutiny was meticulously done post data pre-processing steps. This included assessment of the number of extracted features (<10,000 features, as a rule of thumb), applying the 80% rule (i.e. features found in less than 20% of the analysed samples were removed). Furthermore, MarkerLynx processing included removal of adducts (an exclusion list was included in the MarkerLynx automated processing: e.g. sodium adducts, formic acid adducts, etc.). The quality of data and stability of the analysis were monitored using QC samples. The generated clean data matrices were then analysed, applying different algorithms and approaches, to extract information that describes the effects of ESI electronic parameters (capillary and cone voltages) on acquired MS signals (number and abundance of features) and downstream overall data structures. Following the processing, a table of detected metabolite markers with their corresponding normalised peak heights across all the samples was generated. PCA and PLS-DA were then performed using the MarkerLynx™ version 4.1 software.

4.4.8 Multivariate statistical analysis

The univariate statistical tool used here was the t-test with resulting p-value and associated effect size. Independent samples t-tests were performed assuming unequal group variance and a 5% significance using MATLAB with Statistics Toolbox (2019), version 9.5.0 (R2018b) software (Natick, Massachusetts: The MathWorks Inc). Effect sizes were incorporated as an indication of the practical relevance of significant differences ($P \leq 0.05$) based on Cohen's d-value and calculated manually as the absolute difference between group means divided by the larger of the two group standard deviations (SD). To control the family wise error rate, p-values were adjusted using the Bonferonni-Holm correction for multiple testing.

PCA scores plots were generated to provide a visual summary of the predominant variation in each dataset and the association with the two experimental groups. This could be achieved as the PCA models constructed here were unsupervised and so received no group information. PLS-DA is another multivariate statistical approach employed in metabolomics data analysis. PLS-DA has been described as a versatile algorithm capable of being used for discriminative variable selection, as well as descriptive and predictive high-dimensional dataset modelling. PLS-DA is a better suited statistical approach, compared to the PCA, when it comes to distinguishing between the two groups of samples as it is a supervised method, especially in

instances where the metabolite profiles are influenced/ affected by several factors. That said, PLS-DA models are prone to overfit and must be validated. A leave-one-out cross-validation (LOO-CV) procedure was followed here to validate the variance explained in the grouping variables. PLS-DA scores plots were generated to assess the predictive ability of the model. Prior to identifying compounds that are largely responsible for any visible separation, the goodness-of-fit statistics (R-squared), as well as the LOO-CV (Q-squared) statistics, were compared to assess model validity. R-squared values above 80% were considered sufficient but conditioned to no dramatic deterioration during LOO-CV, that is Q-squared values above 60% were considered acceptable. PCA and PLS-DA analysis 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>). Prior to statistical analysis, the data were pre-processed to help ensure the accuracy of results. The GC-MS, ¹H-NMR and UPLC-DAD datasets were log transformed (shifted natural log transformation with shift parameter set to 1) to correct for the skewness in distribution known to plaque metabolomics data, and auto-scaled (subtracting the mean and dividing by the SD) so compounds in different abundances receive equal attention during multivariate analysis. Finally, the PCA model based scores plots and Hotelling's T-squared distances were employed to detect outliers within each group given a 95% confidence interval (CI).

4.4.9 Logistic regression analysis

For the LR model development, JMP Pro 15 software was used. Firstly, the 303 cultivars were separated into the Comm and NComm groups i.e. 56 Comm and 247 NComm cultivars. Next, a validation column was created using the predictive modelling function of the software in which 75% of the 303 samples were randomly selected and assigned as the training sample set. The remaining 25% was held out, and was used as the testing sample set to determine the predictive accuracy of the developed LR models. Next, LR was performed on the 75% using different metabolite combinations as predictors. Once the LR models had been developed, the probability formulas for each were saved, and the number of misclassifications was determined manually on the 25% test sample set. The %sensitivity and %selectivity were determined using the following formulas:

Sensitivity = $TP/(TP + FN)$, where TP are the true positives i.e. Comm cultivars correctly classified as Comm; FN are the false negatives i.e. Comm cultivars misclassified as NComm cultivars.

Specificity = $TN/(TN + FP)$, where TN are the true negatives i.e. NComm cultivars correctly classified as NComm; FP are the false positives i.e. NComm cultivars misclassified as Comm cultivars.

These steps were repeated three times per model, which entails that each time the validation column was generated, a random, different sample list within the 303 dataset made up the training and testing samples sets. The mean of each was then used as the %sensitivity and %specificity for each developed model, as reported in Table 4.9.

4.5. RESULTS

4.5.1 List of tables showing detected metabolites using GC-MS, ¹H-NMR, UPLC-DAD and UPLC-MS

The tables below show all the metabolites detected in this study using the various metabolomics platforms.

Table 4.1: The list of tentatively identified metabolites detected by GC-MS, expressed in arbitrary units.

Comm vs NComm variables	Analytical platform	Relative normalised intensity		Fold change of the mean	Cohen's d-value	Reported literature concentration (mg/g)	References
		Comm	NComm				
Acetoacetic acid	GC-MS	0.050	1.2	1.2	0.21	20.02	Naveed et al., 2017
Arabinose	GC-MS	0.011	2.2*	2.2*	0.81	20.03	Naveed et al., 2017
Catechin	GC-MS	0.070	1.3	1.3	0.47	29.18	Gramza et al., 2006
1-Cyclohexenecarboxylic Acid	GC-MS	0.034	0.6*	0.6*	0.42	4.72	Baeza et al., 2016
Gallic acid	GC-MS	0.056	1.3*	1.3*	0.62	5.10	Kaneko et al., 2006
Glycerol	GC-MS	0.006	1.4	1.4	0.01	10.03	Jones et al., 2008
Phloroglucinol	GC-MS	0.003	1.0	1.0	0.13	45.10	Matanjun et al., 2008
Psicose	GC-MS	0.0005	2.0	2.0	0.69	1.31	Mu et al., 2012
Ribitol	GC-MS	0.007	1.4*	1.4*	0.32	20.04	Roser et al., 1992
Sucrose	GC-MS	0.040	1.0	1.0	0.46	30.90	Kumar et al., 2011
Threonic acid	GC-MS	0.006	0.5*	0.5*	0.86	12.20	Naveed et al., 2017
Xylonic acid	GC-MS	0.001	2.3*	2.3*	0.29	4.50	Habibi et al., 2004
Total sweeteners	GC-MS	0.095	3.6*	3.6*	0.83	N.A	N.A

*indicate a statistically significant difference in the mean concentration of the metabolite between the Comm and NComm cultivars at the 95% level of significance after correcting for multiple testing. N.A = not available.

Table 4.2: The list of metabolites detected by ¹H-NMR, expressed in mg/g.

Comm vs NComm variables	Analytical platform	Comm concentration (mg/g)			NComm concentration (mg/g)			Fold change of the mean	Cohen's d-value	Reported literature concentration (mg/g)	References
		Min	Max	Mean	Min	Max	Mean				
Acetic acid	¹ H-NMR	25.01	60.97	34.8	28.80	70.71	42.2	0.8	1.40	40.01	Bandurski and Schulze, 1977
Alanine	¹ H-NMR	0.11	0.32	0.2	0.16	0.40	0.2	1.0	0.31	4.21	Min et al., 2017
Caffeine	¹ H-NMR	6.24	20.02	12.6	6.14	19.12	11.6	1.1*	1.10	24.33	Chin et al., 2008
Catechin	¹ H-NMR	6.17	30.32	15.4	2.66	28.89	14.2	1.1	0.90	29.18	Gramza et al., 2006
Chlorogenic acid	¹ H-NMR	2.65	6.44	4.2	2.51	6.07	4.0	1.0	1.10	6.92	Marks et al., 2007
Epicatechin	¹ H-NMR	7.92	28.16	14.6	8.84	26.26	14.2	1.0	0.51	70.66	Gramza et al., 2006
Epicatechin gallate	¹ H-NMR	4.01	23.03	13.5	4.32	22.13	13.2	1.0	0.80	170.30	Gramza et al., 2006
Epigallocatechin	¹ H-NMR	20.55	117.41	51.7	17.12	113.49	49.6	1.1*	0.40	151.29	Gramza et al., 2006
Epigallocatechin gallate	¹ H-NMR	19.02	66.60	39.0	17.36	57.58	35.6	1.1*	0.40	173.87	Gramza et al., 2006
Formic acid	¹ H-NMR	19.07	46.40	27.2	16.01	31.08	21.0	1.3*	1.70	21.01	Sanhueza, E., and Andreae, 1991
Gallic acid	¹ H-NMR	0.53	2.04	1.1	0.60	2.98	1.0	1.1	2.00	5.10	Kaneko et al., 2006
Glucose	¹ H-NMR	5.38	11.72	7.5	4.97	10.36	7.2	1.0	0.50	6.91	Melgarejo et al., 2000
Isoleucine	¹ H-NMR	0.13	0.41	0.2	0.01	0.27	0.2	1.0	0.10	2.60	Min et al., 2017
Leucine	¹ H-NMR	0.06	0.23	0.2	0.08	0.30	0.1	2.0*	0.40	3.90	Min et al., 2017
Methanol	¹ H-NMR	0.07	0.29	0.2	0.07	0.25	0.1	2.0*	1.71	0.04	Fall and Benson, 1996
Sucrose	¹ H-NMR	5.91	21.54	15.0	6.61	23.14	13.6	1.1*	0.20	30.90	Kumar et al., 2011
Theanine	¹ H-NMR	2.82	22.22	8.6	4.19	13.79	8.0	1.1	1.90	30.00	Vuong et al., 2011
Quinic acid	¹ H-NMR	1.01	2.70	1.9	1.03	3.04	2.0	1.0	0.71	5.04	Rodrigues et al., 2007
Valine	¹ H-NMR	0.11	0.29	1.7	0.09	0.27	1.7	1.0	0.40	3.40	Min et al., 2017
Total amino acid	¹ H-NMR	13.22	33.89	21.8	11.97	32.85	20.9	1.0	1.04	N.A	N.A
Total catechins	¹ H-NMR	55.73	241.18	121.6	51.22	160.30	114.9	1.1*	0.60	N.A	N.A
Total sweeteners	¹ H-NMR	4.37	22.91	9.3	3.51	15.81	8.7	1.1*	0.71	N.A	N.A

*indicate a statistically significant difference in the mean concentration of the metabolite between the Comm and NComm cultivars at the 95% level of significance after correcting for multiple testing. N.A = not available.

Table 4.3: The list of metabolites detected by the UPLC-DAD, expressed in mg/g.

Comm vs NComm variables	Analytical platform	Comm concentration (mg/g)			NComm concentration (mg/g)			Fold change of the mean	Cohen's d-value	Reported literature concentration (mg/g)	References
		Min	Max	Mean	Min	Max	Mean				
Caffeine	UPLC-DAD	21.3	40.3	29.0	16.3	29.3	23.3	1.2*	1.42	24.33	Chin et al., 2008
Catechin	UPLC-DAD	4.4	18.2	16.4	2.6	20.4	8.3	2.0*	1.90	29.18	Gramza et al., 2006
Epicatechin	UPLC-DAD	1.7	13.3	7.7	5.7	32.5	13.9	0.6*	1.53	70.66	Gramza et al., 2006
Epicatechin gallate	UPLC-DAD	10.4	72.3	27.2	15.5	65.3	30.2	0.9	0.37	170.30	Gramza et al., 2006
Epigallocatechin	UPLC-DAD	5.2	53.8	25.5	11.2	63.9	32.8	0.8	0.69	151.29	Gramza et al., 2006
Epigallocatechin gallate	UPLC-DAD	42.6	72.6	57.1	37.6	99.1	61.0	0.9	0.41	173.87	Gramza et al., 2006
Theaflavin	UPLC-DAD	1.1	19.5	7.9	2.8	10.8	5.3	1.5*	0.73	21.01	Ding et al., 1992
Theaflavin-3-gallate	UPLC-DAD	2.1	21.8	5.5	4.5	11.2	7.5	0.7*	0.73	16.10	Ding et al., 1992
Theaflavin-3'-gallate	UPLC-DAD	5.3	26.5	14.6	4.5	12.3	7.6	1.9*	1.41	32.11	Ding et al., 1992
Theaflavin-3,3-digallate	UPLC-DAD	3.2	37.4	26.1	2.8	9.0	5.0	5.2*	5.75	41.10	Ding et al., 1992
Yield	UPLC-DAD	479.7	5560.0	3022.6	366.2	2345.3	1589.8	2.0*	2.00	N.A	N.A
Total catechins	UPLC-DAD	64.3	230.2	133.9	72.6	281.2	146.1	0.9	0.91	N.A	N.A
Total theaflavins	UPLC-DAD	19.3	87.5	54.0	16.8	39.5	25.4	2.1*	2.10	N.A	N.A

*indicate a statistically significant difference in the mean concentration of the metabolite between the Comm and NComm cultivars at the 95% level of significance after correcting for multiple testing. N.A = not available. Yield is expressed in KgMT/Ha/year.

Table 4.4: The list of metabolites detected by the UPLC-MS, expressed in arbitrary units.

Comm vs NComm variables	Analytical platform	Comm concentration			NComm concentration			Fold change of the mean	Cohen's d-value	Reported literature concentration (mg/g)	References
		Min	Max	Mean	Min	Max	Mean				
Argininosuccinate	UPLC-MS	3.5	7.4	5.3	1.4	3.3	2.4	2.2*	3.10	9.40	Naveed et al., 2017
Caffeic acid	UPLC-MS	0.2	0.8	0.5	0.1	0.5	0.3	1.7*	1.67	19.20	Ravn et al., 1994
Caffeine	UPLC-MS	0.1	0.6	0.3	0.1	0.2	0.2	1.5*	1.56	24.33	Chin et al., 2008
Catechin	UPLC-MS	1.7	4.4	3.1	0.7	2.5	1.4	2.2*	2.41	29.18	Gramza et al., 2006
Citric acid	UPLC-MS	0.5	0.9	0.7	0.2	0.7	0.5	1.4*	1.42	2.32	Melgarejo et al., 2000
Epicatechin	UPLC-MS	0.4	1.9	1.5	0.5	1.7	0.9	1.7*	1.15	70.66	Gramza et al., 2006
Epicatechin gallate	UPLC-MS	1.2	2.8	2.1	1.7	2.5	2.2	1.0	0.49	170.30	Gramza et al., 2006
Epigallocatechin gallate	UPLC-MS	0.2	0.7	0.5	0.1	0.6	0.3	1.7*	1.11	173.87	Gramza et al., 2006
Gallic acid	UPLC-MS	0.4	1.3	0.7	0.6	3.3	1.4	0.5*	1.07	5.10	Kaneko et al., 2006
Gallocatechin	UPLC-MS	0.2	0.9	0.6	0.3	0.8	0.5	1.2	0.29	5.76	Blainski et al., 2017
Gluconic acid	UPLC-MS	2.5	5.1	3.5	1.7	4.1	2.6	1.3*	1.14	11.80	Naveed et al., 2017
Glucose	UPLC-MS	3.4	4.8	4.1	2.8	4.0	3.3	1.2*	2.03	6.91	Melgarejo et al., 2000
Glutamic acid	UPLC-MS	0.1	0.3	0.2	0.1	0.3	0.2	1.0	0.41	6.90	Min et al., 2017
Kaempferol 3- <i>O</i> - β -rutinoside	UPLC-MS	0.7	1.9	1.2	0.4	2.8	1.2	1.0	0.06	0.11	Karakaya and El, 1999
Lysine	UPLC-MS	0.4	1.2	0.7	0.3	0.4	0.3	2.3*	0.82	2.30	Min et al., 2017
Maltose	UPLC-MS	0.6	1.1	0.8	0.3	0.9	0.5	1.6*	1.51	7.10	Naveed et al., 2017
Myoinositol	UPLC-MS	4.7	10.1	6.0	4.9	8.9	6.6	0.9	0.39	36.60	Naveed et al., 2017
Quercetin	UPLC-MS	0.6	1.2	0.9	0.9	2.6	1.6	0.6*	1.79	0.04	Karakaya and El, 1999
Rutin	UPLC-MS	2.5	4.9	4.0	2.7	7.1	4.9	0.8	0.83	2.69	Kreft et al., 2006
Theanine	UPLC-MS	0.3	1.1	0.6	0.2	0.4	0.3	2.0*	2.06	30.00	Vuong et al., 2011
Theobromine	UPLC-MS	0.1	0.7	0.4	0.2	0.4	0.3	1.3*	0.98	4.86	Sun et al., 2006
Total amino acid	UPLC-MS	0.7	2.6	1.4	0.7	1.5	1.0	1.4*	1.45	N.A	N.A
Total catechins	UPLC-MS	3.8	10.6	7.6	3.2	8.2	5.4	1.4*	1.12	N.A	N.A
Total sweeteners	UPLC-MS	8.6	16.0	11.0	8.0	13.8	10.4	1.1	0.97/	N.A	N.A

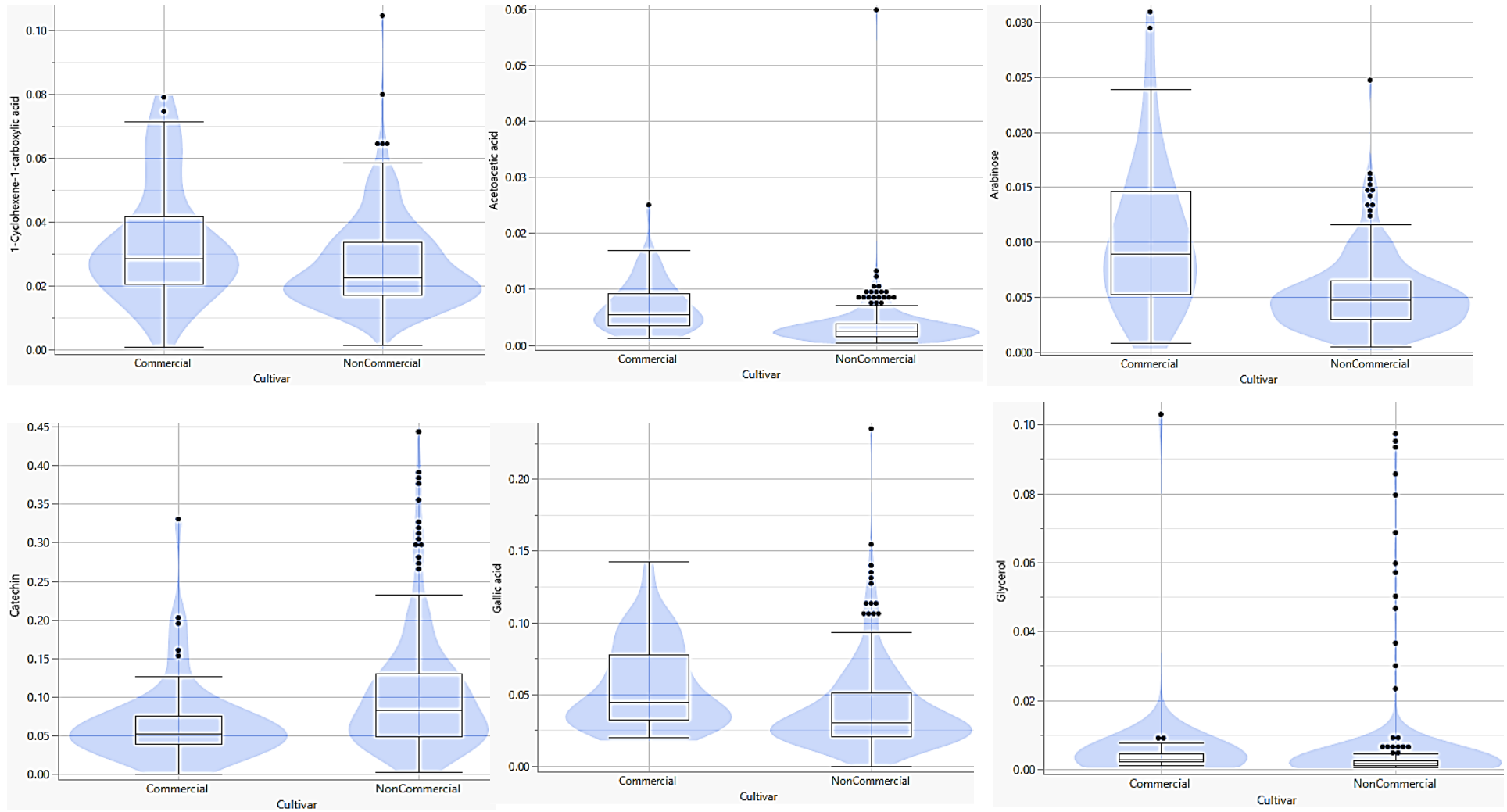
*indicate a statistically significant difference in the mean concentration of the metabolite between the Comm and NComm cultivars at the 95% level of significance after correcting for multiple testing. N.A = not available

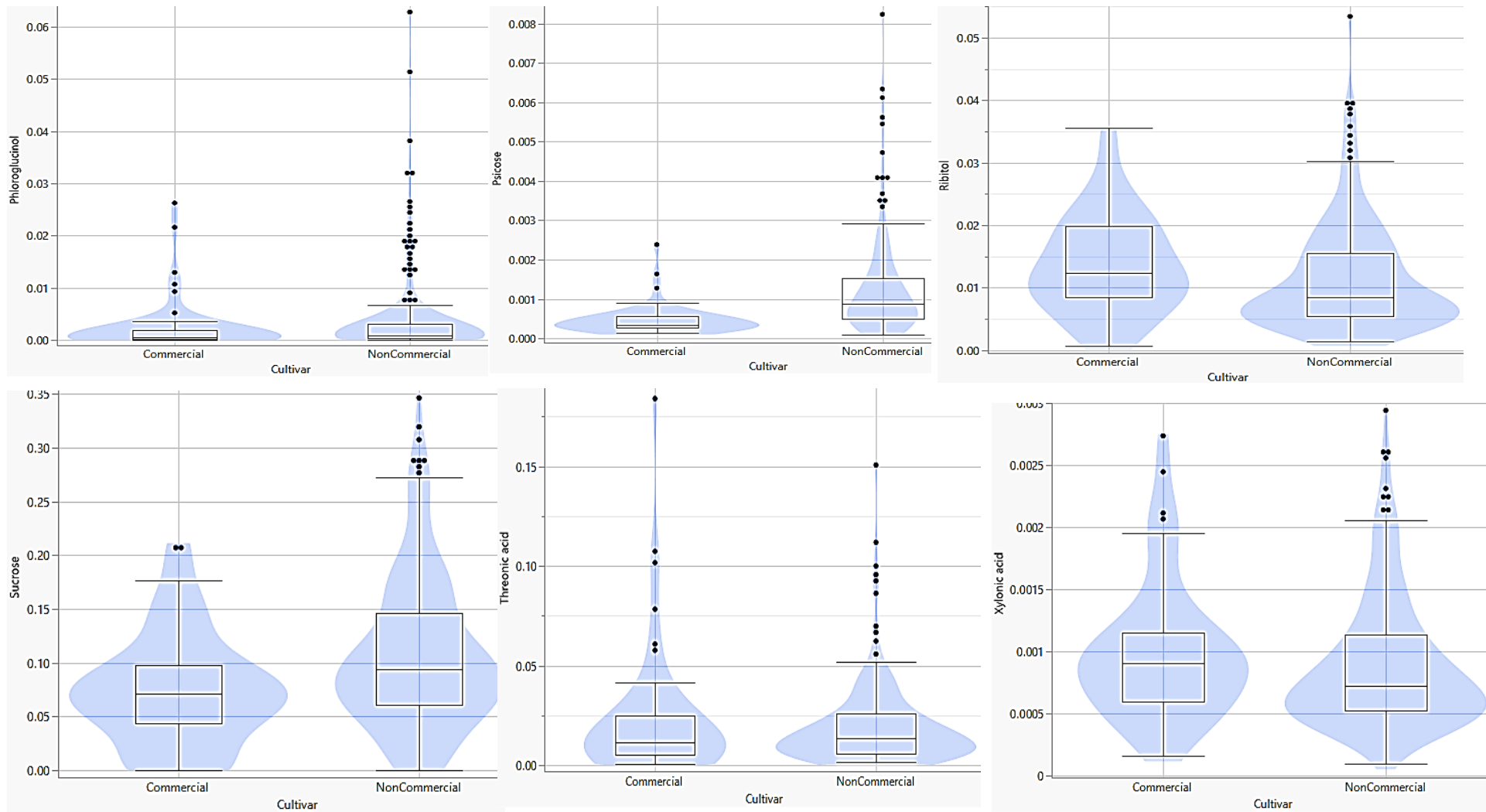
Table 4.5: The list of identified metabolites to be distinguishing markers between the Comm and NComm cultivars. Table shows the ionisation mode under which they were detected, their m/z value and corresponding retention time (RT/min).

Detected metabolite	Ionisation mode	Measured m/z	RT/min	High energy MS fragment
Caffeic acid	M-H	179.0347	4.9	149, 135
Catechin	M-H	289.0711	4.7	139, 123, 95
Citric acid	M-H	191.0196	1.0	87, 111
Epicatechin	M-H	289.0701	5.7	207, 139, 123, 55
Epicatechin gallate	M-H	441.0807	7.8	273, 153, 139, 123
Epigallocatechin gallate	M-H	457.3720	14.8	441, 289, 153, 139
Gallic acid	M-H	169.0145	1.8	169, 125, 44
Gallocatechin	M-H	305.0657	3.4	223, 195, 163, 139
Gluconic acid	M-H	195.0509	0.7	149, 133
Glucose	M-H	179.0561	0.8	131, 133
Kaempferol 3- <i>O</i> - β -rutinoside	M-H	593.1492	8.7	449, 287, 147, 331
Maltose	M-H	341.1086	0.7	281, 263, 179, 161
Myoinositol	M-H	333.0584	0.6	145, 201, 233
Quercetin	M-H	301.0346	10.3	151, 179, 229, 273
Rutin	M-H	609.1439	7.7	465, 303, 85
Argininosuccinate	M+H	290.1226	1.1	201, 157, 58
Caffeine	M+H	195.0875	8.61	138, 110, 69
Glutamic acid	M+H	148.0602	0.8	130, 102, 84, 56
Lysine	M+H	147.1127	1.0	130, 129, 84, 56
Theanine	M+H	120.0665	1.2	158, 130, 84, 56
Theobromine	M+H	181.0717	3.2	138, 110, 83

4.5.2 Violin plots for GC-MS, ¹H-NMR, UPLC-DAD and UPLC-MS

To visually represent the abundance of metabolites retained after zero-filtering, violin plots were constructed. Each plot has two violin plots, one for the Comm cultivars and another for the NComm cultivars, depicting the concentration/level of each metabolite (indicated on the y-axis). The plots also show box and whisker plots showing the distribution of each metabolite within the Comm and NComm samples, and highlighting those that are outliers. The y-axis represent concentration and the x-axis the cultivar names. From the box plots, it can be seen that some metabolites are statistically significantly different between the Comm and NComm cultivars. The mean line across the centre of each box is representative of each group's mean. In instances where overlaps occur i.e the mean line of one group falls within the lines of the upper and lower quartiles of the other group, this shows that the group means are not significantly different at that particular CI. This therefore means that the concentration of that particular metabolite is not statistically significantly different between the two groups and where the mean line of one group falls outside the box of the other group, that metabolite is statistically significantly different between the two groups at that CI. These are shown in Figure 4.1.





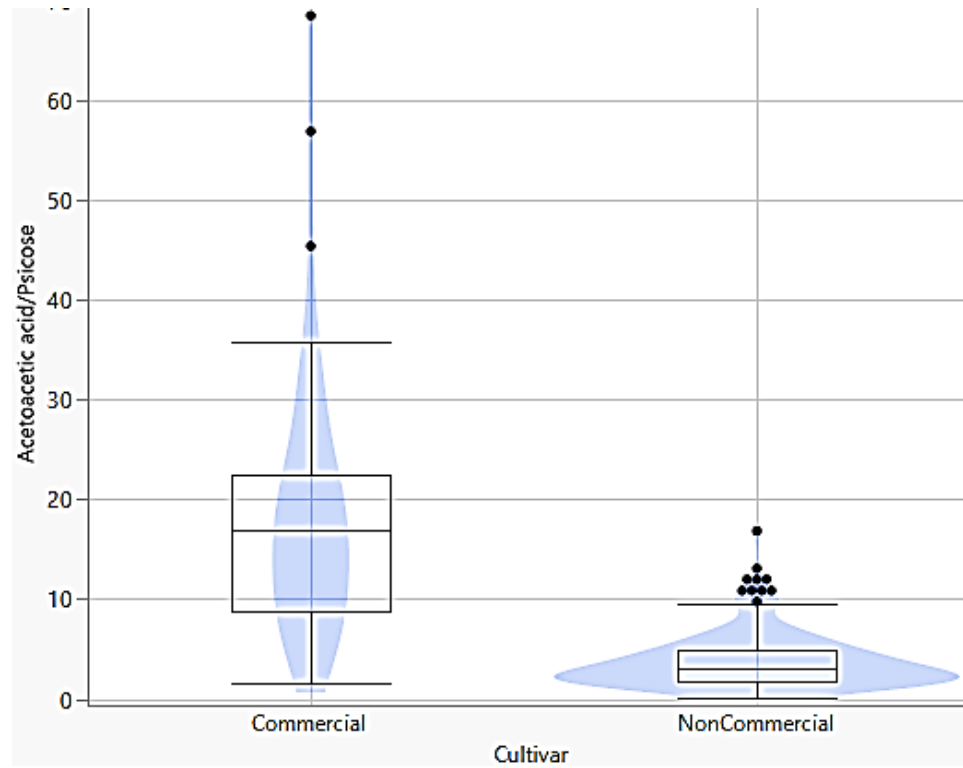
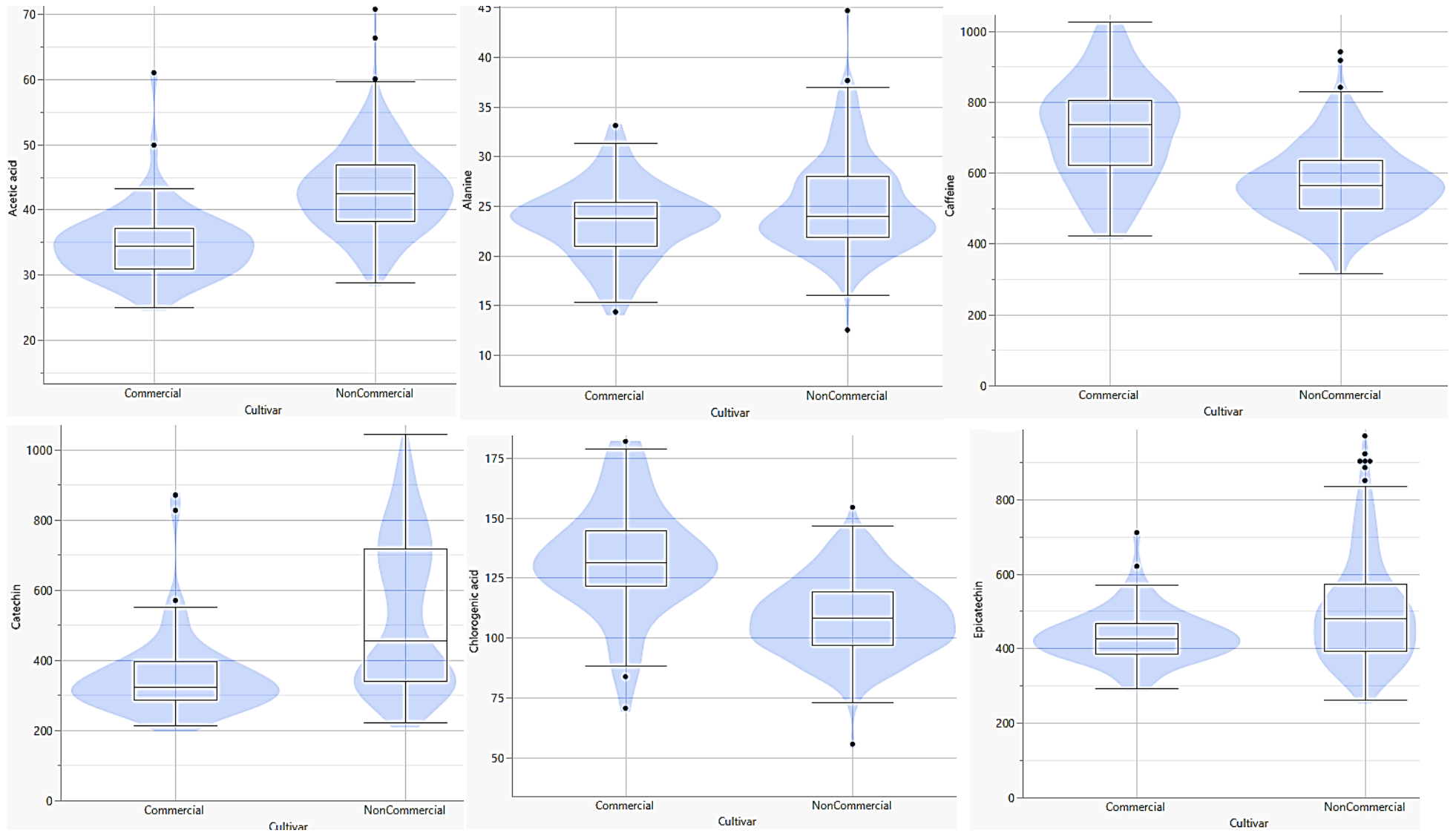
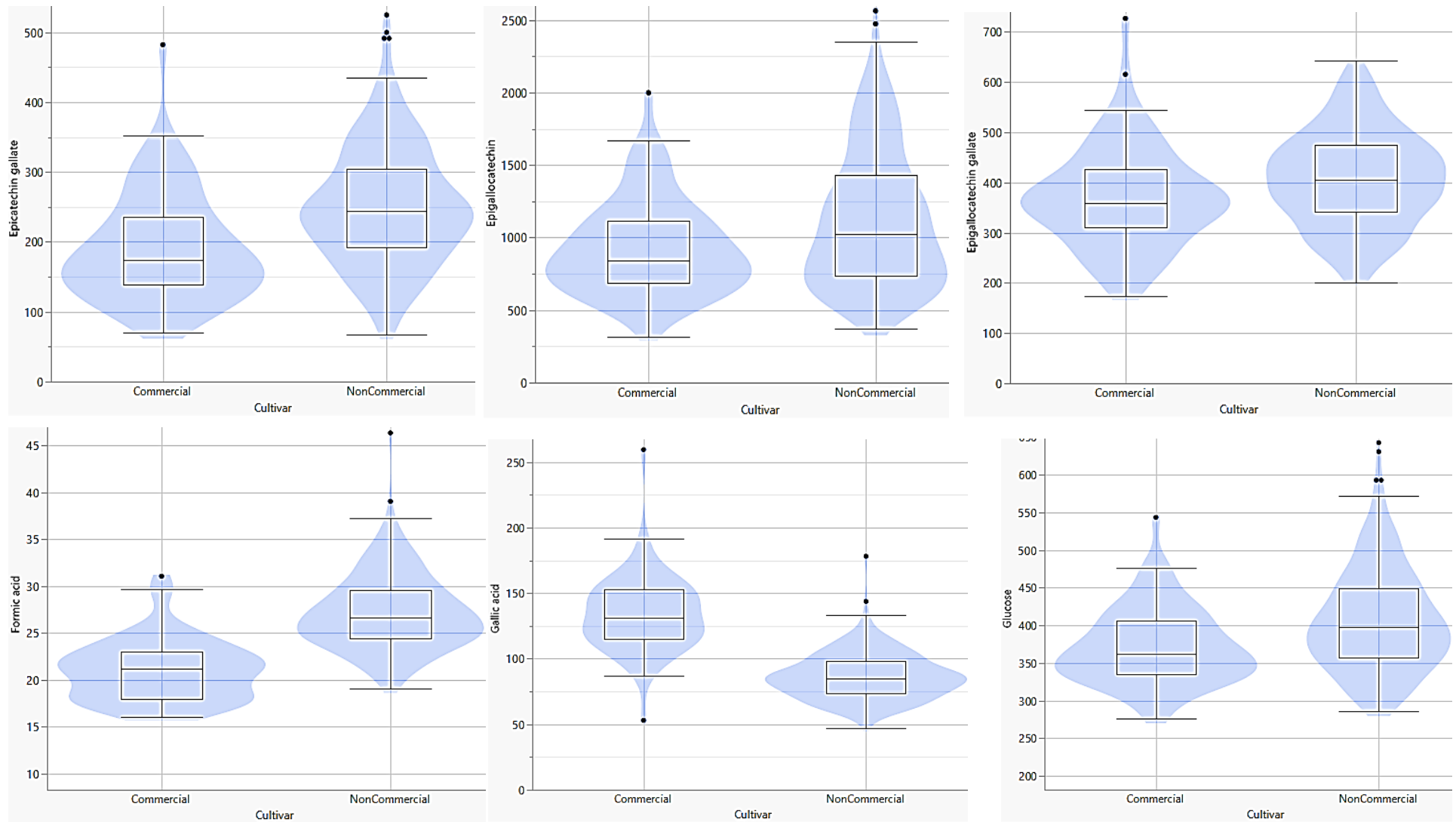
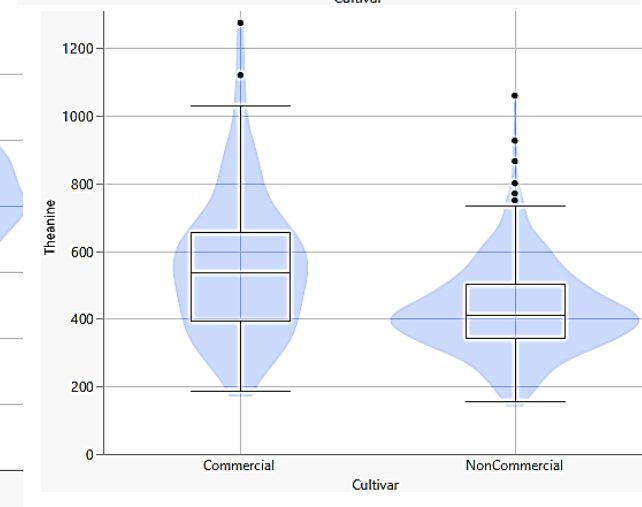
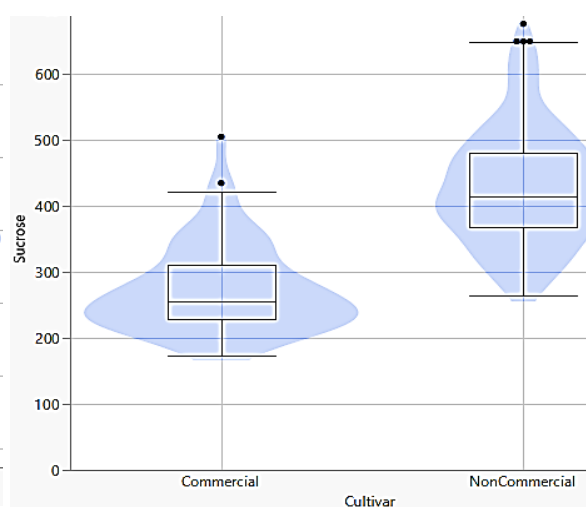
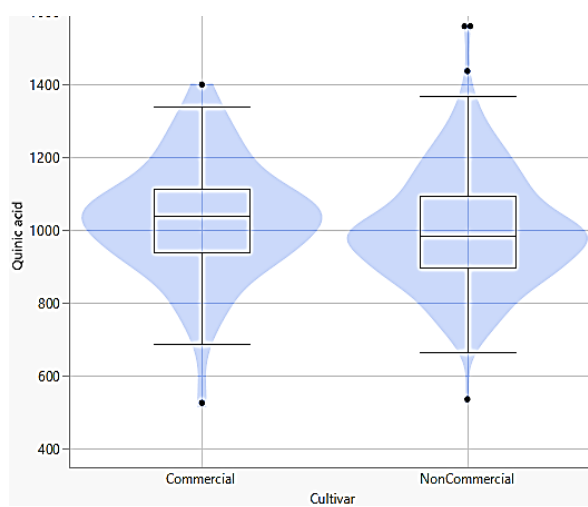
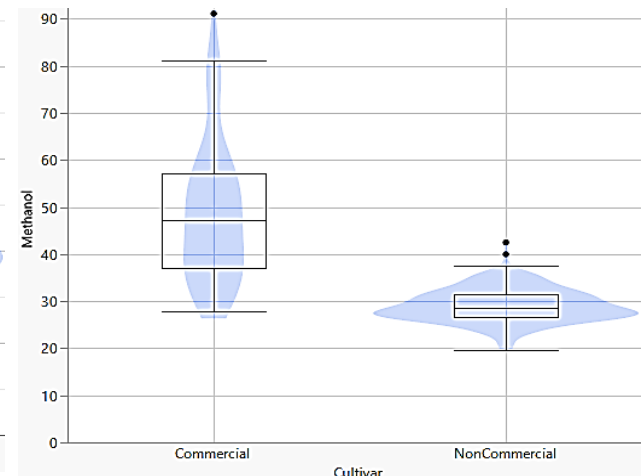
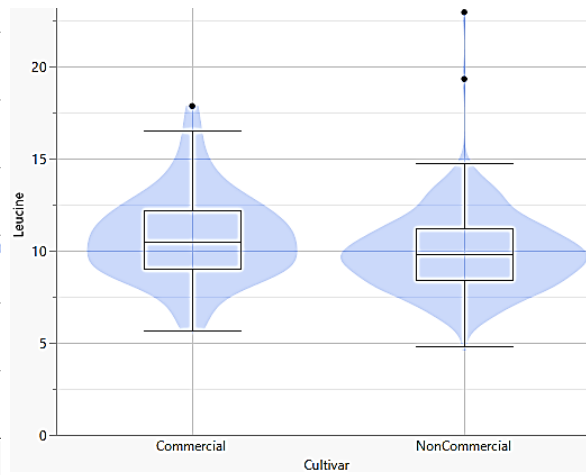
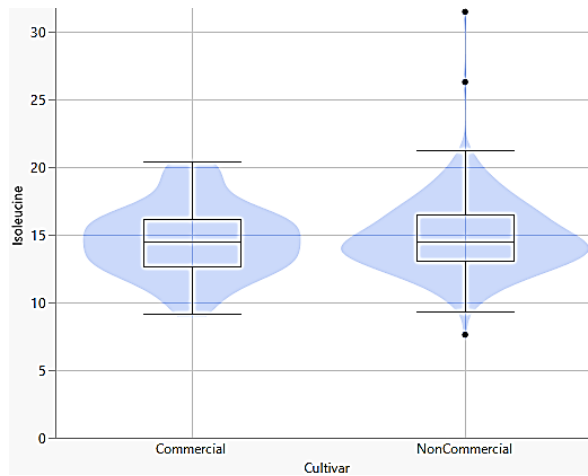


Figure 4.1: Violin plots showing separation between the Comm and NComm cultivars based on GC-MS metabolites. The y-axis units are arbitrary units. The black dots represent outliers, which are observations 1.5 x interquartile range (IQR) greater than the 75th quantile or 1.5 x IQR less than the 25th quantile.







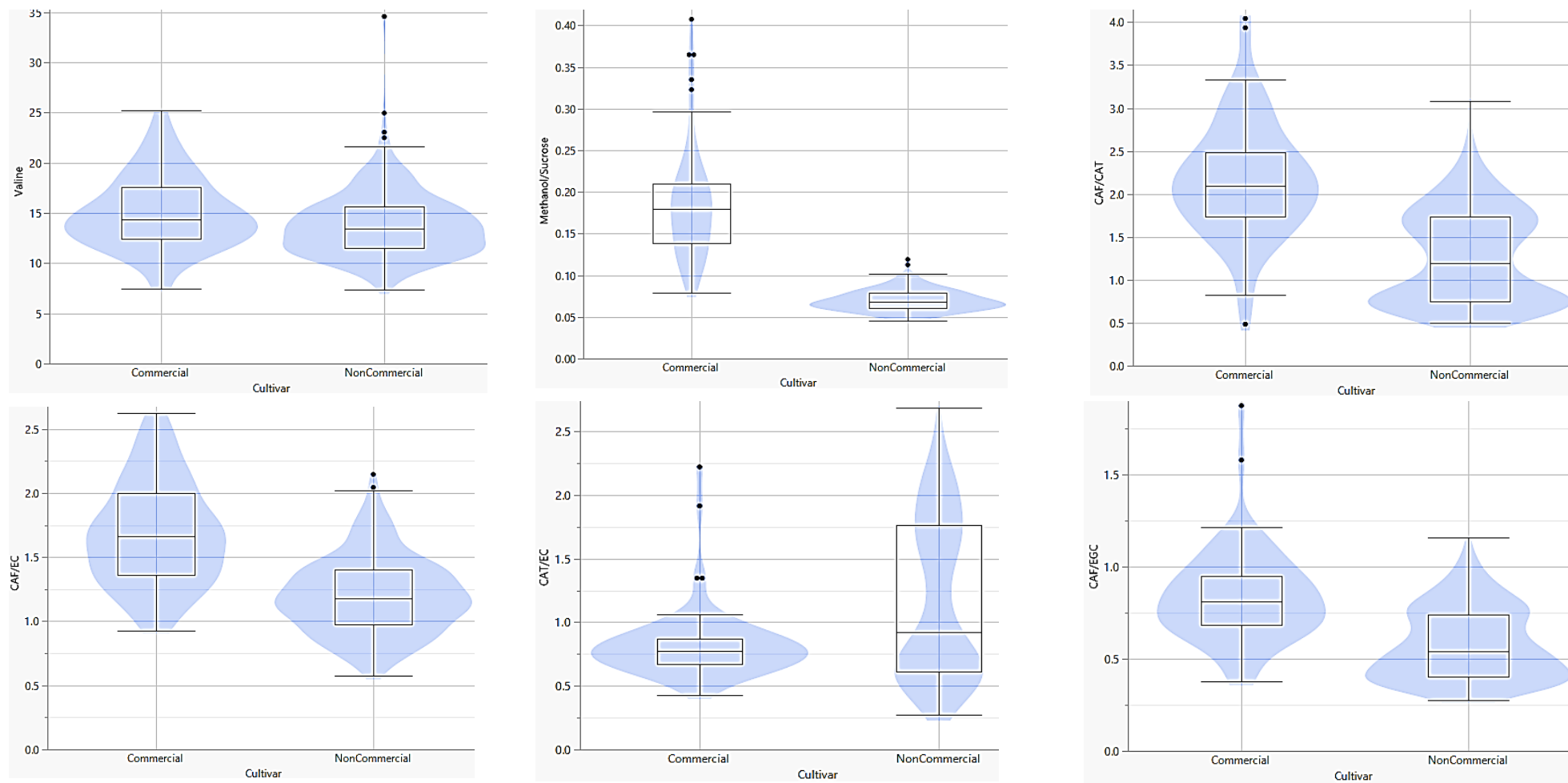
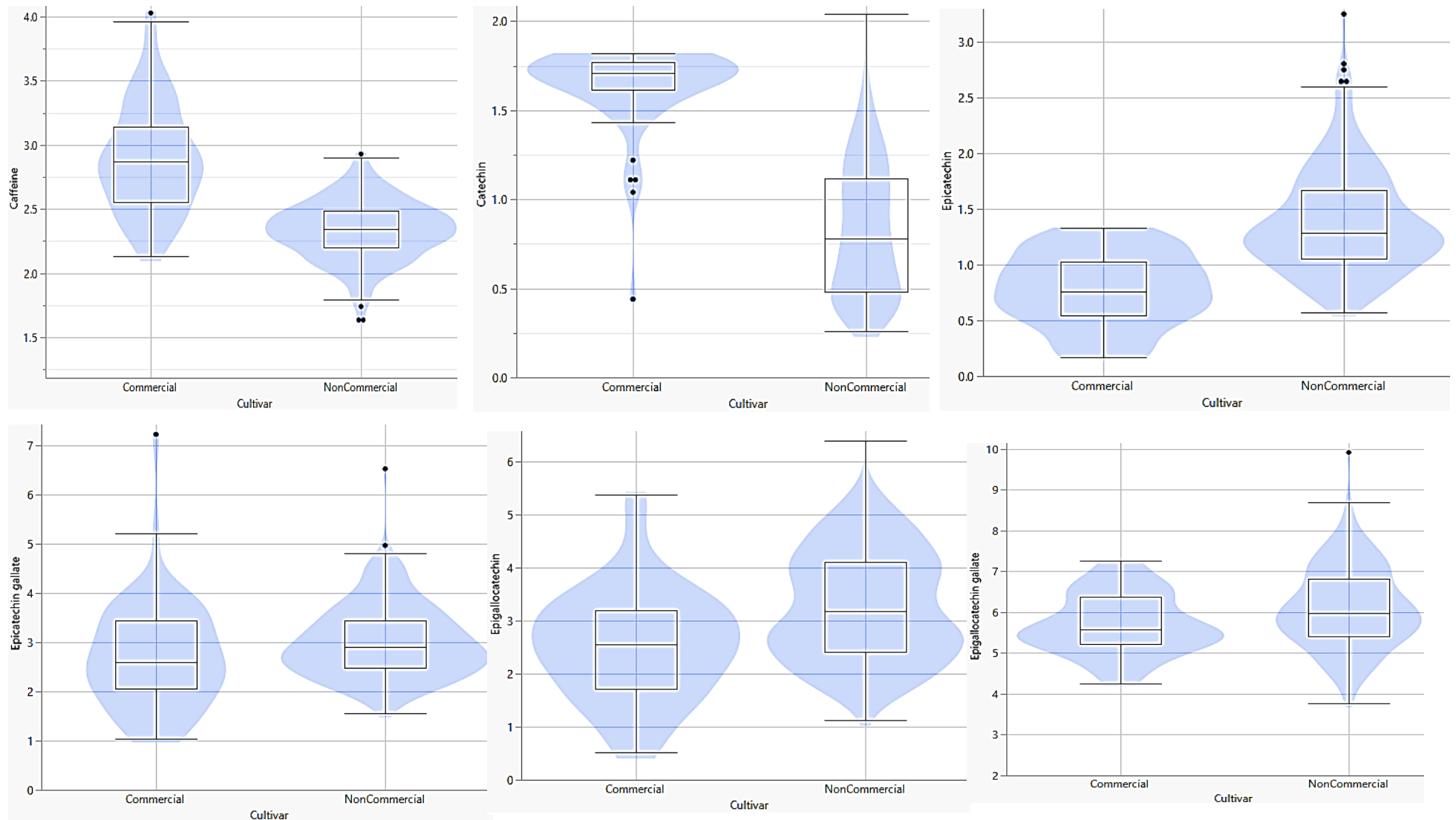
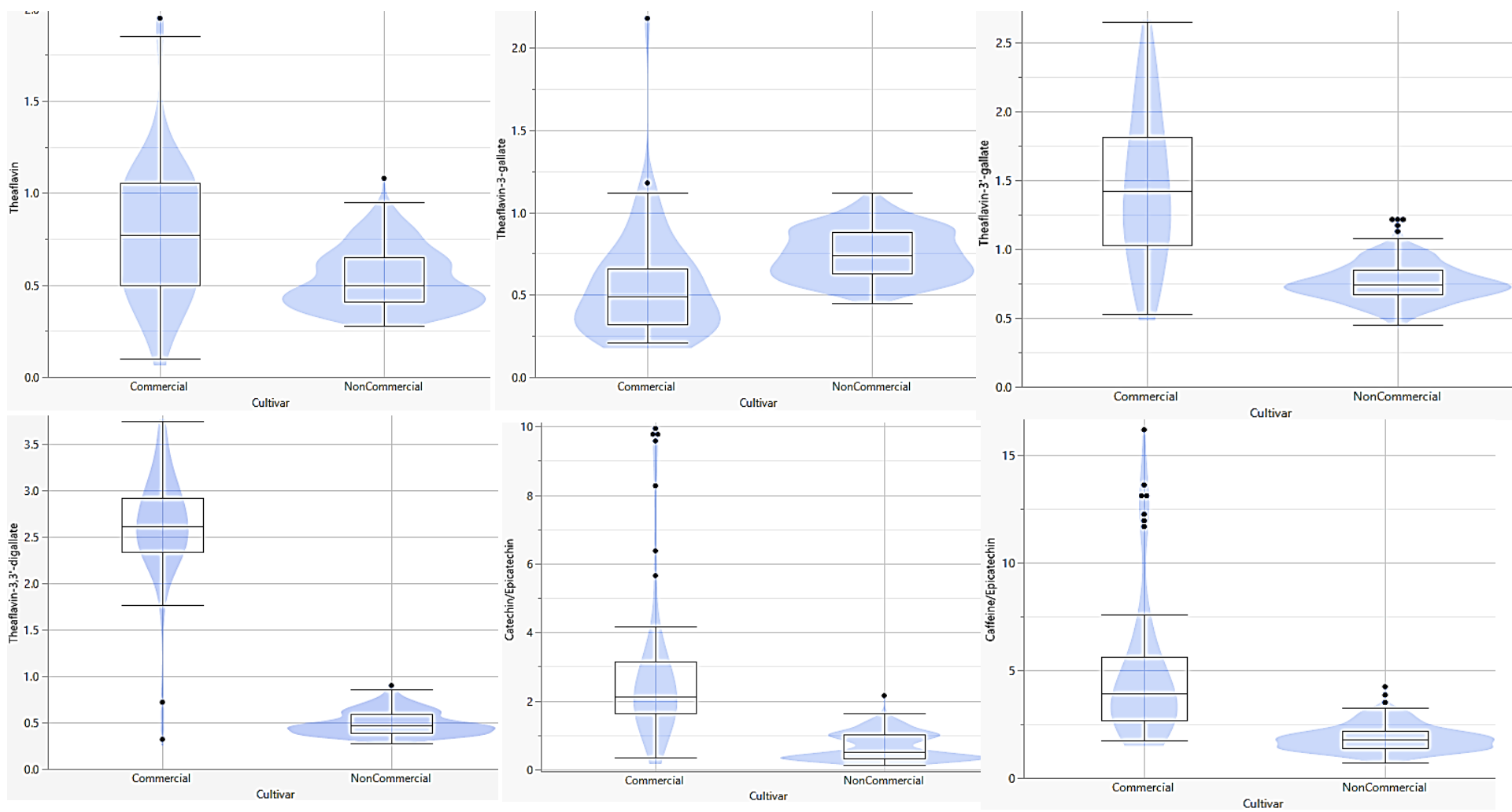


Figure 4.2: Violin plots showing separation between the Comm and NComm cultivars based on ¹H-NMR metabolites. The y-axis units are mg/g dry weights.





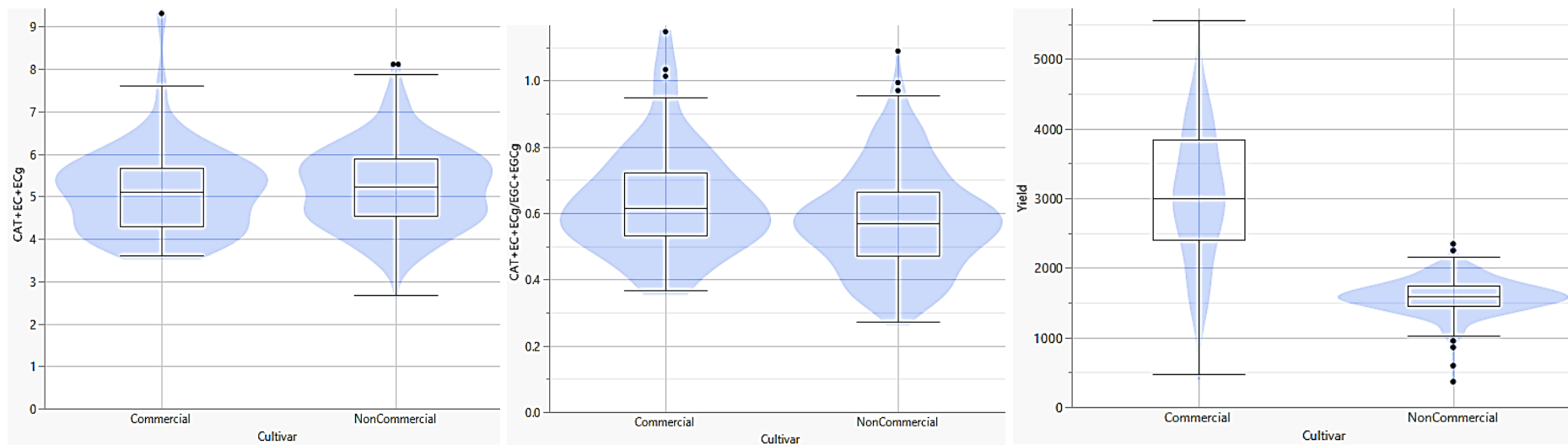
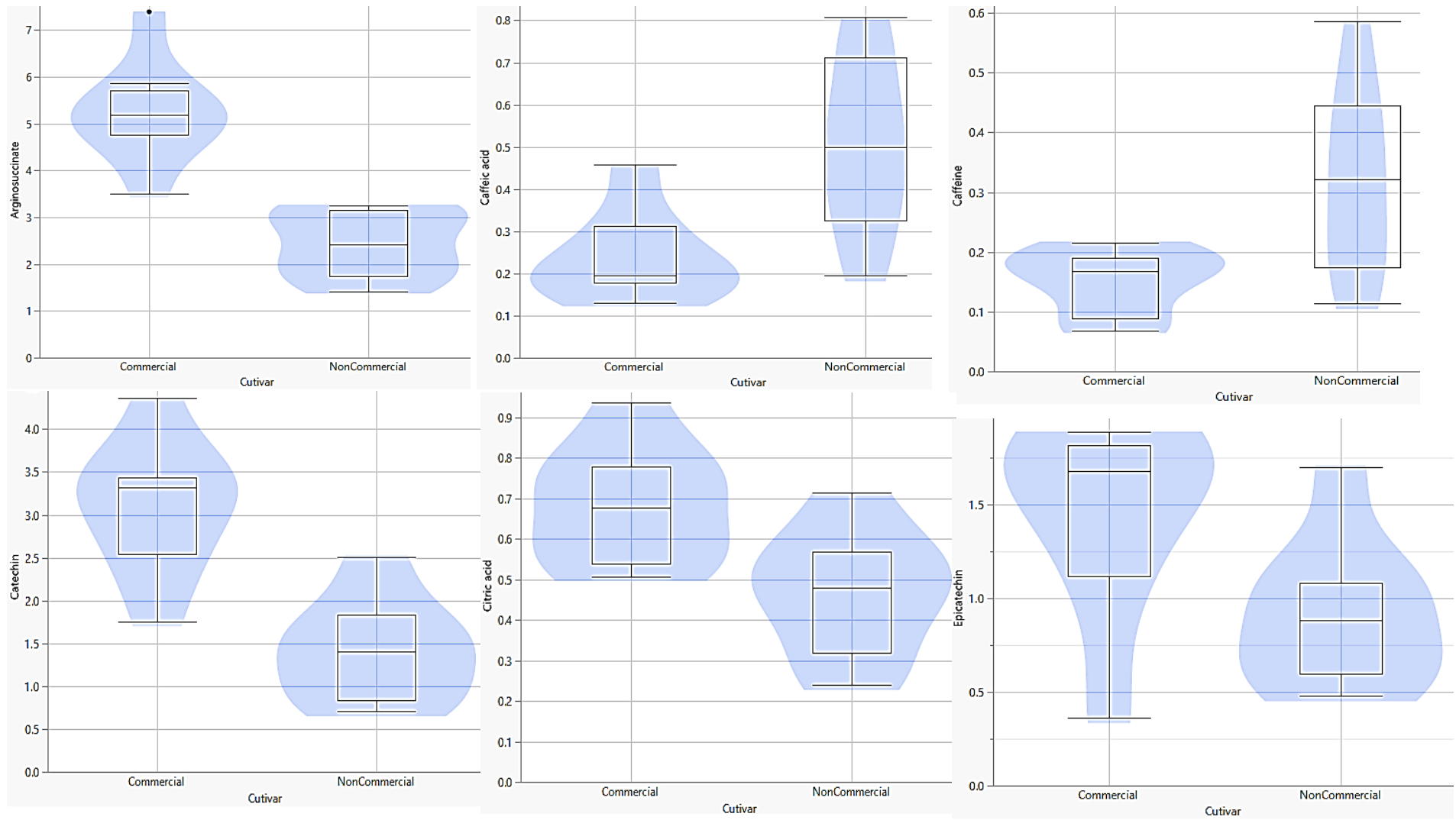
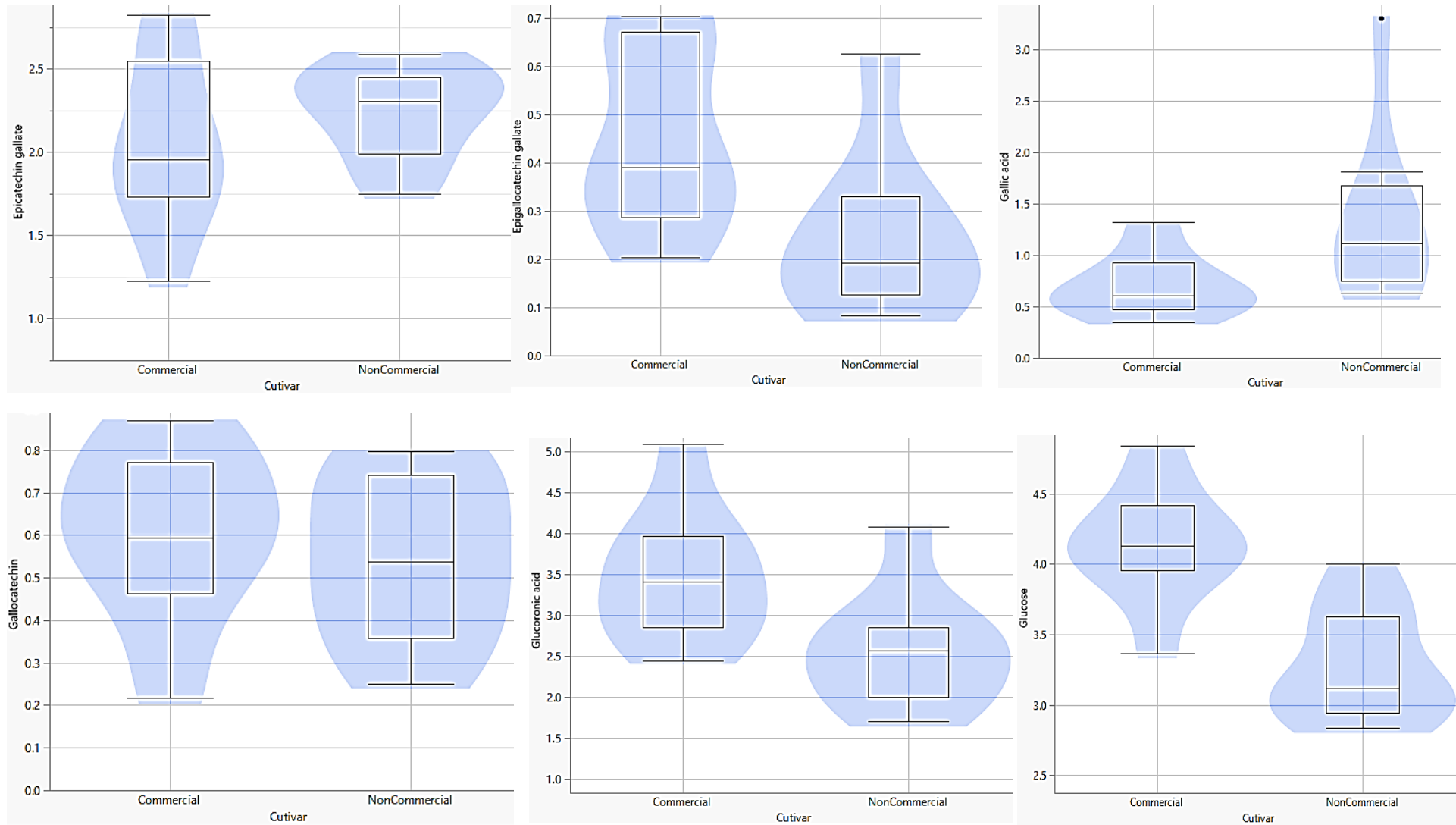
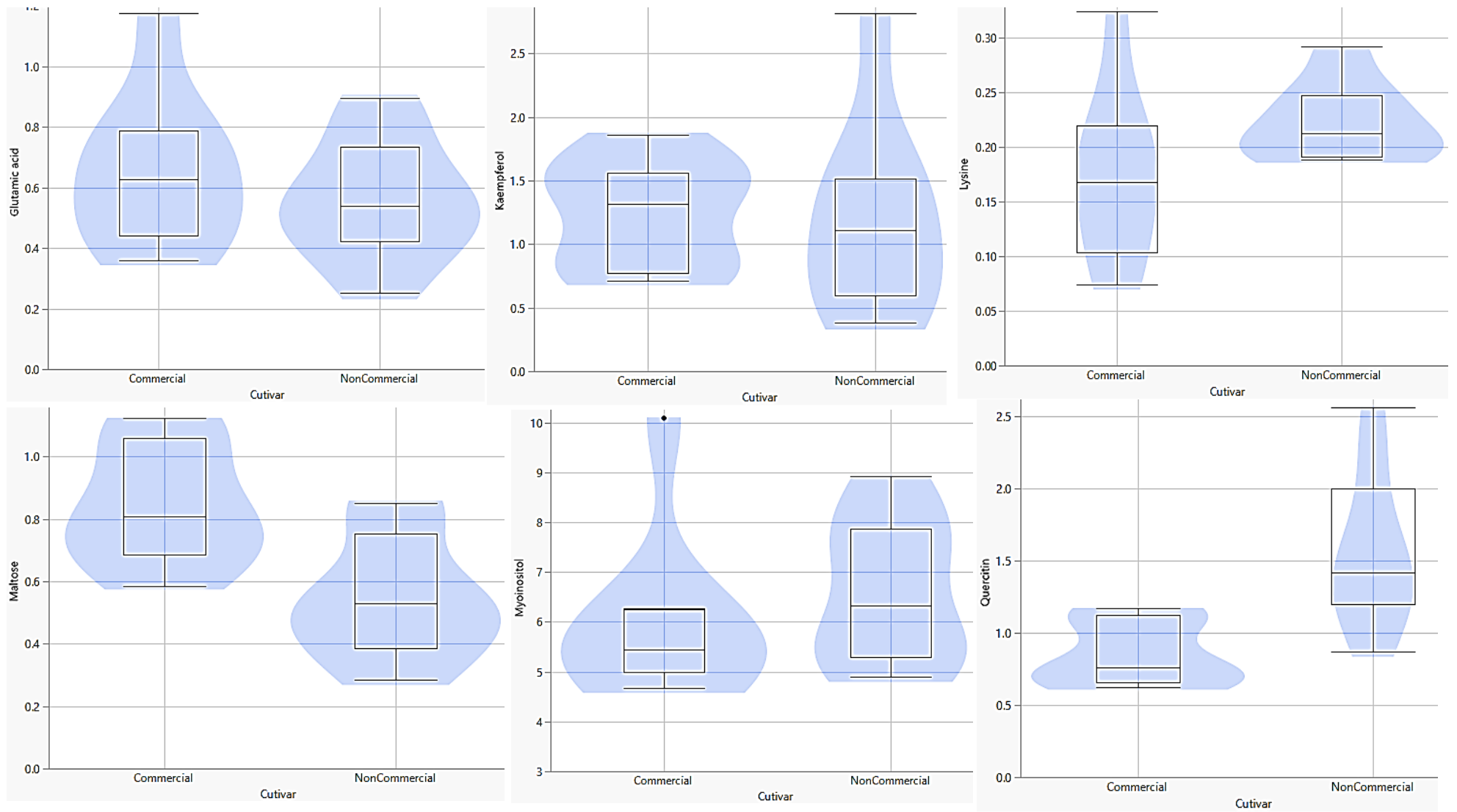


Figure 4.3: Violin plots showing separation between the Comm and NComm cultivars based on detected UPLC-DAD metabolites. The y-axis units for the CAF and catechins are %w/w dry weight; TF1-TF4 in black tea samples were quantified as EGCg equivalents, based on the EGCg response factor; yields are KgMT/Ha/year.







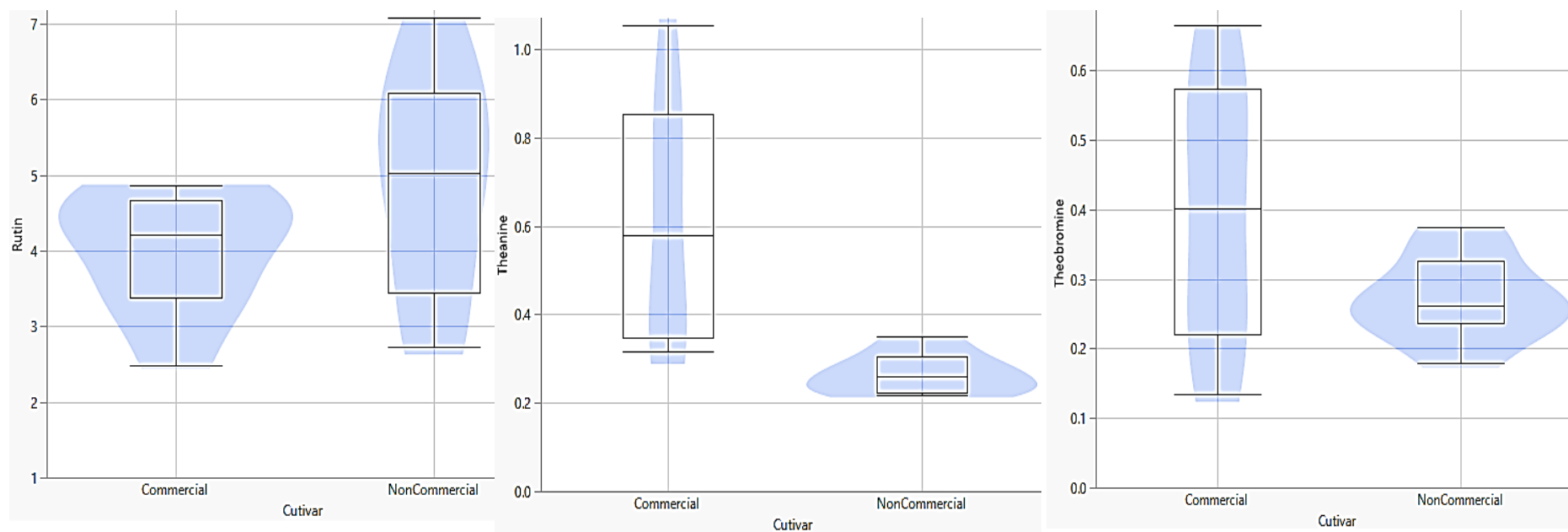


Figure 4.4: Violin plots showing separation between the Comm and NComm cultivars based on UPLC-MS metabolites. The y-axis units are expressed in arbitrary units.

4.5.3 GC-MS, ¹H-NMR and UPLC-DAD PCA plots

Results presented in Figure 4.5 show that the ellipsoids representing 95% CI of score centroids of the Comm and NComm groups separate best by UPLC-DAD (Figure 4.5C) than ¹H-NMR (Figure 4.5B) and GC-MS (Figure 4.5A). The percentage of the overall variation in the measured compounds explained by each principal component (PC) is indicated along the three axes.

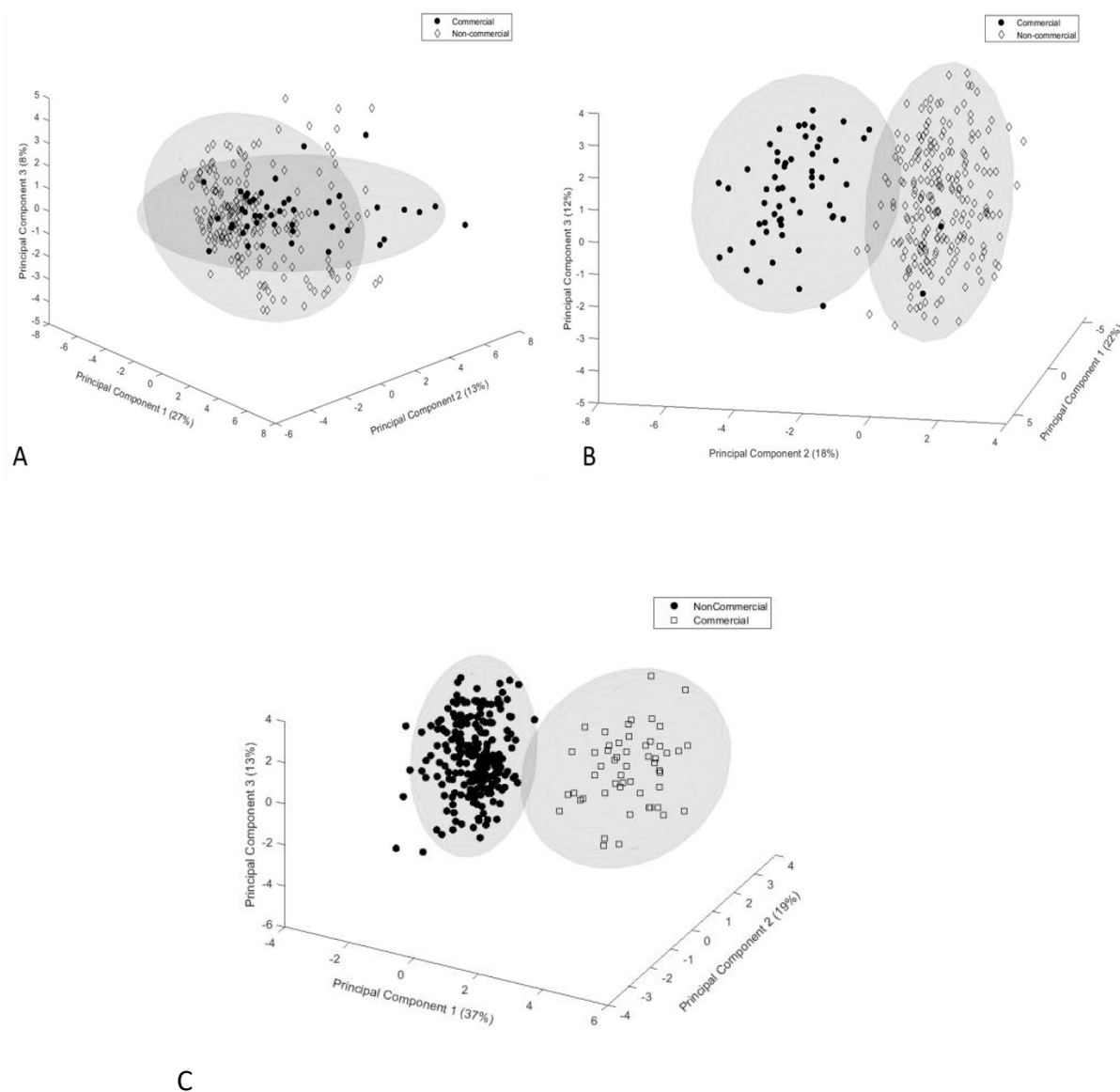


Figure 4.5: The 3D PCA scores plots for PCs one, two and three, for GC-MS (A) ¹H-NMR (B) and UPLC-DAD (C), showing the separation and explaining 48%, 52% and 69% of the variation, respectively from each platform.

4.5.3 GC-MS, ¹H-NMR and UPLC-DAD PLS-DA plots

Similar to PCA plots, ellipsoids represent 95% CI of score centroids of each group. The percentage of the overall variation in the measured compounds (X) and group membership (Y), as explained by each latent variable (LV), which is indicated along each axis, as shown in Figure 4.6.

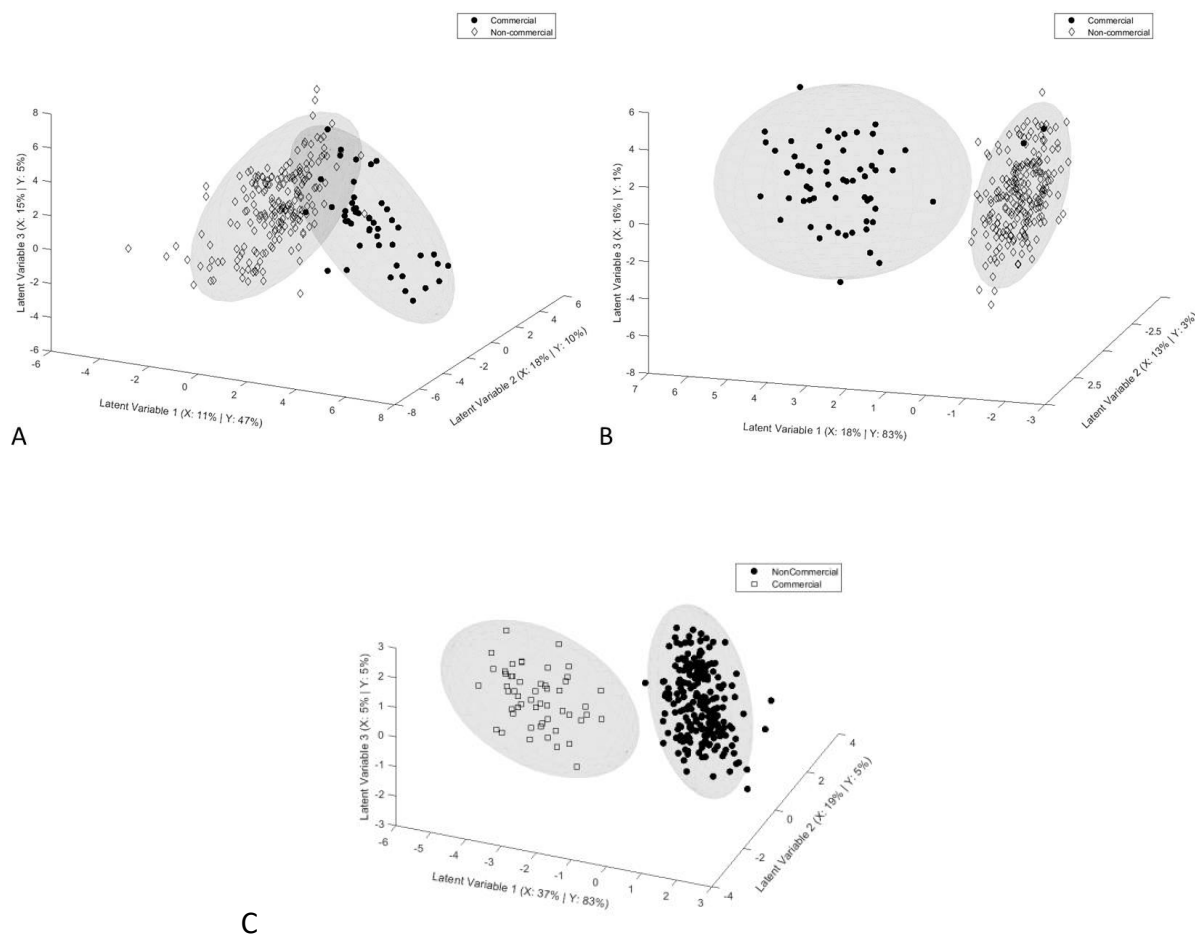


Figure 4.6: The 3D PLS-DA scores plots for LVs one, two and three, for GC-MS (A), ¹H-NMR (B) and UPLC-DAD (C). The goodness-of-fit values achieved for the GC-MS model were $R^2=62\%$ and $Q^2=55\%$ making it unreliable for discriminant identification, when compared to the ¹H-NMR model, which performed better with $R^2=87\%$ and $Q^2=85\%$. The goodness-of-fit values achieved for the UPLC-DAD model were deemed reliable with predictive accuracy $R^2=94\%$ and leave-one-out crossvalidated predictive accuracy $Q^2=93\%$.

4.5.4 UPLC-MS positive and negative ionisation mode PCA, PLS-DA and S-plots

PCA is an unsupervised, projection technique that permits the viewing of large sample datasets by summarising variation through the projection onto fewer dimensions. It is predominantly used multivariate analysis technique for the analysis of metabolomic data. PCA popularity in metabolomics is due to the fact that it is a simple non-parametric method, which permits the viewing of large sample datasets by summarising variation through the projection onto fewer dimensions, revealing inherent data trends. PLS-DA is a supervised technique commonly used for classifying and selecting biomarkers in metabolomics research. It consists of a PLS regression, used to identify combinations of variables that can distinguish between groups of samples. PLS-DA improves the observed PCA separation between the two sample groups, with the percentage of the overall variation explained by each component is indicated along each axis. Each point on the graph represents a sample as projected onto the new lower-dimensional space, with the ellipsoids representing the 95% CI of each group. Figures 4.7 and 4.8 show UPLC-MS PCA, PLS-DA and S-plots under the positive and negative ionisation modes.

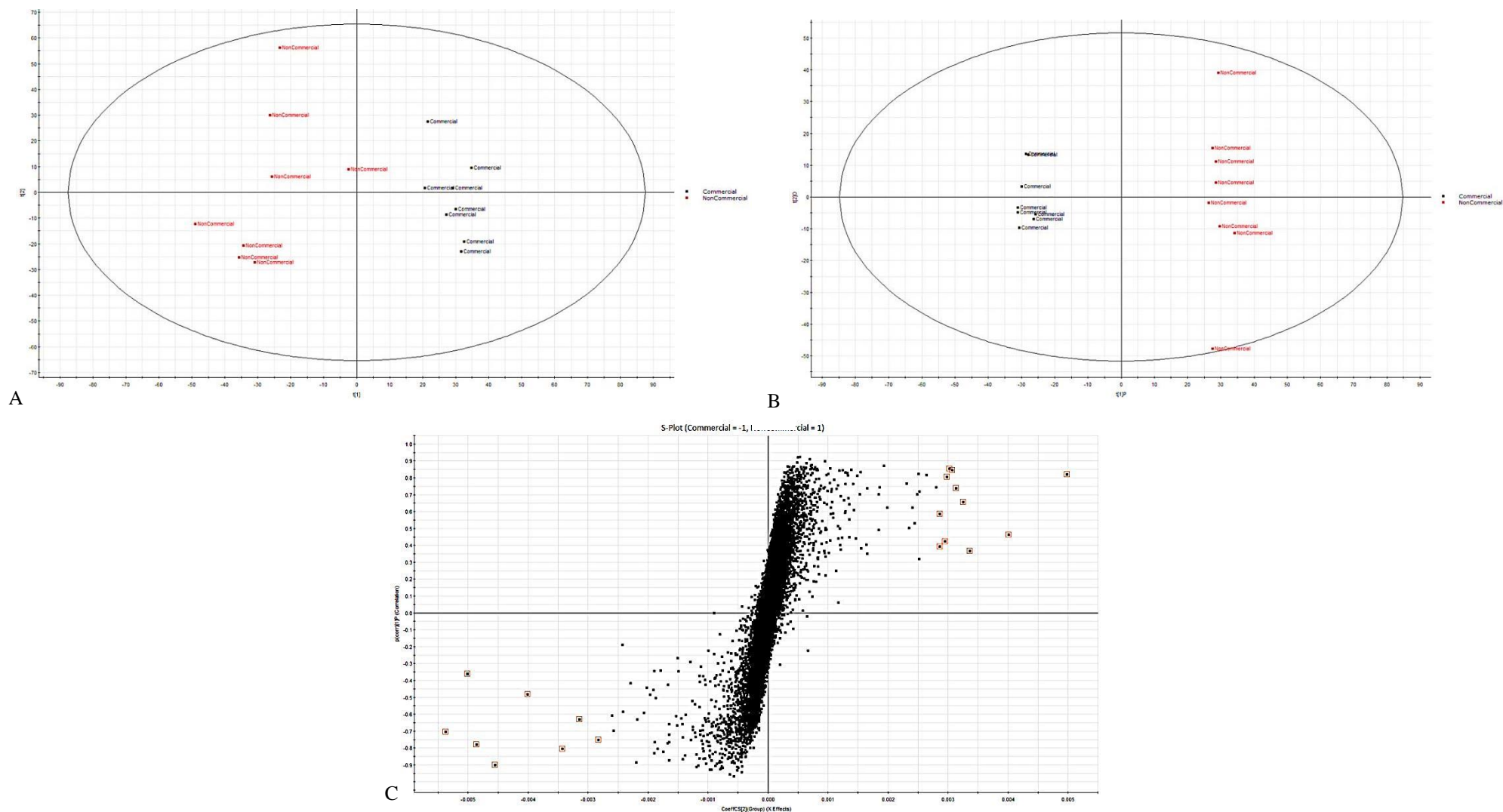


Figure 4.7: (A) The PCA scores plot; (B) the PLS-DA plot, and (C) the s-plot, showing good separation between the Comm and the NComm cultivars in positive ion mode, and the metabolite markers distinguishing both groups. On the s-plot, the markers above the x-axis are the metabolites higher in the Comm cultivars as compared to the NComm, and those below the x-axis are higher in the NComm cultivars as compared to the Comm cultivars.

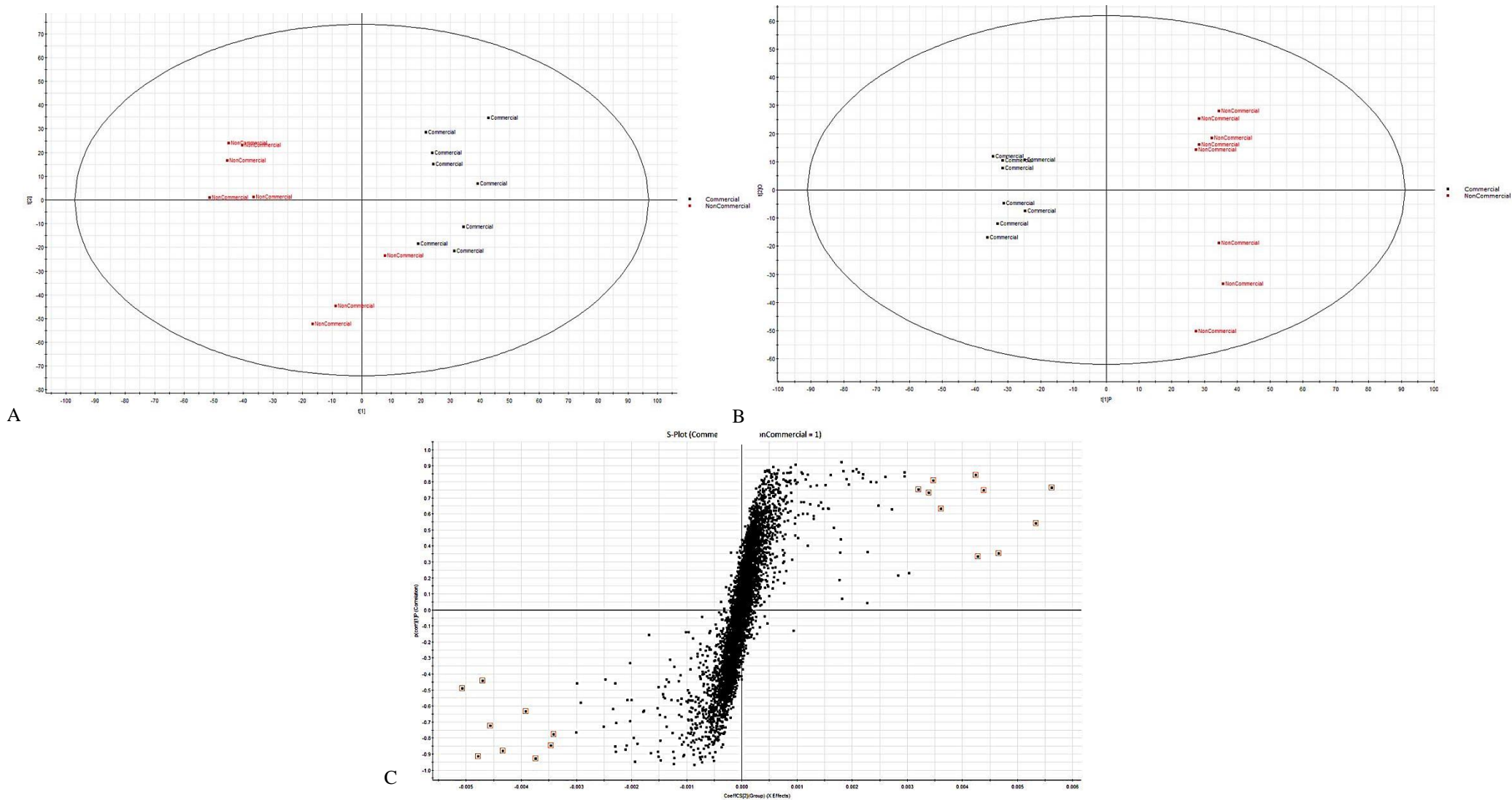
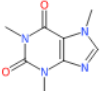
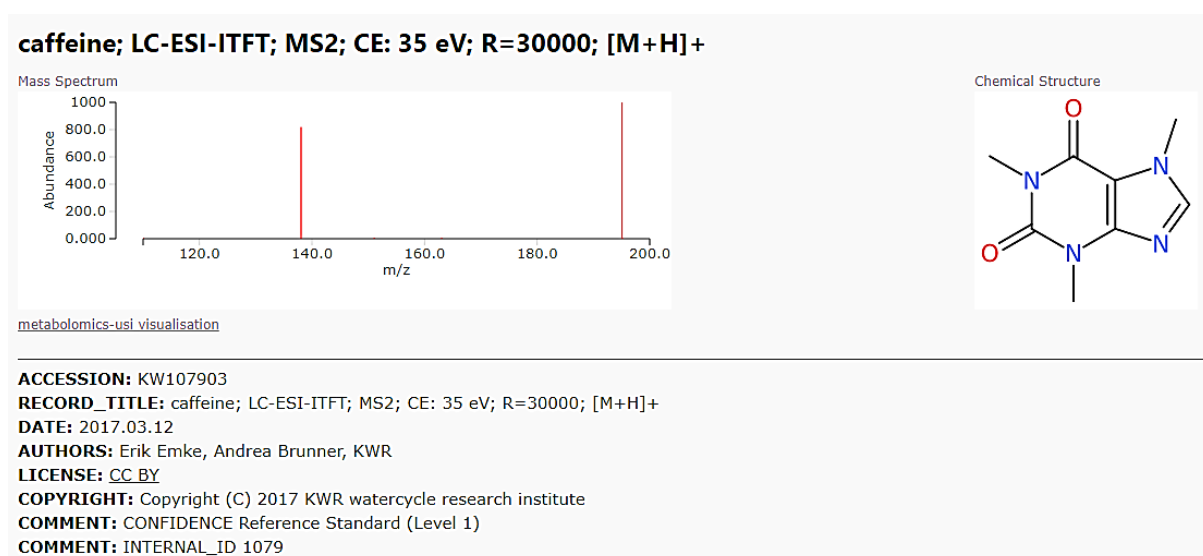


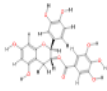
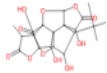
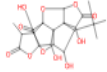
Figure 4.8: (A) The PCA scores plot; (B) the PLS-DA plot, and (C) the s-plot, showing good separation between the Comm and the NComm cultivars in negative ion mode, and the metabolite markers distinguishing both groups. On the s-plot, the markers above the x-axis are higher in the Comm cultivars as compared to the NComm, and those below the x-axis are higher in the NComm cultivars as compared to the Comm cultivar.

As mentioned in 4.4.6, biomarkers were identified and confirmed by comparing their mass spectra and retention times against the reference standards, and the MassFragment application manager was used to facilitate the MS/MS fragment ion analysis process. First, the high energy spectra fragments from the pure standards cocktail for caffeine, in the positive ionisation mode and ECg (annotated catechin gallate since CAT and EC are isomers) in the negative ionisation mode, were uploaded to MassBank online data source where they were identified and confirmed by their hit score values. MassBank uses a database search algorithm to calculate the similarity score between two spectra i.e. database spectrum and query spectrum, based on a modified cosine correlation (Horai *et al.*, 2010). The closer the score is to 1, the higher the certainty of correct identification. Following this confirmation, the peaks in both the negative and positive ionisation mode samples were then identified (Figure 4.9), with Figure 4.9 (E) showing the fragments from the high energy negative ionisation mass spectrum used to identify gallic acid i.e. 44, 125 and mother ion 169.

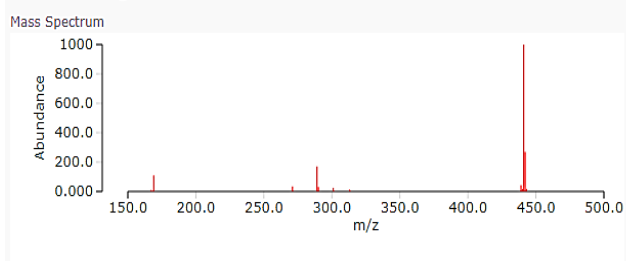
<input type="checkbox"/>	Name	Formula / Structure	Hit	Score
<input type="checkbox"/>	caffeine; LC-ESI-ITFT; MS2; CE: 35 eV; R=30000; [M+H]⁺	C8H10N4O2 	5	0.9983
<input type="checkbox"/>	caffeine; LC-ESI-ITFT; MS2; CE: 35 eV; R=7500; [M+H]⁺	C8H10N4O2 	5	0.9918
<input type="checkbox"/>	Caffeine; LC-ESI-ITFT; MS2; CE: 45%; R=7500; [M+H]⁺	C8H10N4O2 	5	0.9901



(A).

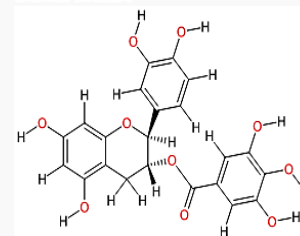
<input type="checkbox"/>	Name	Formula / Structure	Hit	Score
<input type="checkbox"/>	Catechin gallate; LC-ESI-QTOF; MS2; CE:10 eV; [M-H]-	C22H18O10 	8	0.9130
<input type="checkbox"/>	Ginkgolide C; LC-ESI-QTOF; MS2	C20H24O11 	4	0.8569
<input type="checkbox"/>	Ginkgolide C; LC-ESI-QTOF; MS2	C20H24O11 	4	0.8553

Catechin gallate; LC-ESI-QTOF; MS2; CE:10 eV; [M-H]-



[metabolomics-usi visualisation](#)

Chemical Structure



ACCESSION: BS003894

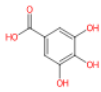
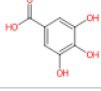
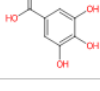
RECORD_TITLE: Catechin gallate; LC-ESI-QTOF; MS2; CE:10 eV; [M-H]-

DATE: 2017.12.01 (Created 2014.08.19)

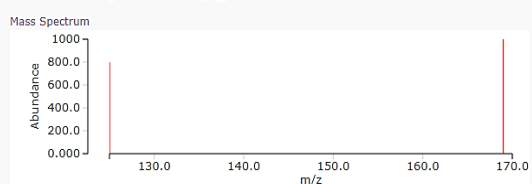
AUTHORS: Plant Biology, The Noble Foundation, Ardmore, OK, US/Dennis Fine, Daniel Wherritt, and Lloyd Sumner

LICENSE: [CC BY-NC-SA 4.0 International](#)

(B).

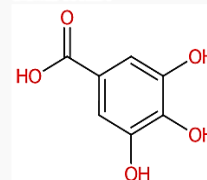
<input type="checkbox"/>	Name	Formula / Structure	Hit	Score
<input type="checkbox"/>	Gallic acid; LC-ESI-QQ; MS2	C7H6O5 	8	0.9989
<input type="checkbox"/>	Gallic acid; LC-ESI-QTOF; MS2	C7H6O5 	8	0.9941
<input type="checkbox"/>	Gallic acid; LC-ESI-QTOF; MS2	C7H6O5 	8	0.9929

Gallic acid; LC-ESI-QQ; MS2



[metabolomics-usi visualisation](#)

Chemical Structure



ACCESSION: PM000401

RECORD_TITLE: Gallic acid; LC-ESI-QQ; MS2

DATE: 2006.04.20

AUTHORS: Sanchez-Rabameda F, et al.

LICENSE: CC-BY-NC

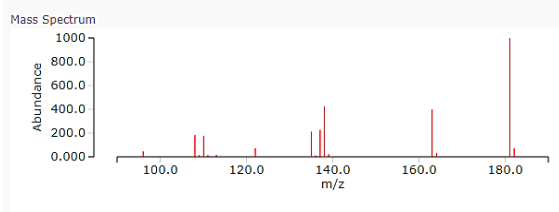
COPYRIGHT: Copyright(C) 2012 Plant Science Center, RIKEN

PUBLICATION: Sánchez-Rabameda, F.; Jáuregui, O.; Casals, I.; Andrés-Lacueva, C.; Izquierdo-Pulido, M.; Lamuela-Raventós, R. M. Liquid Chromatographic/Electrospray Ionization Tandem Mass Spectrometric Study of the Phenolic Composition of Cocoa (*Theobroma Cacao*). *Journal of Mass Spectrometry* 2003, 38 (1), 35–42.
DOI:10.1002/jms.395

(C).

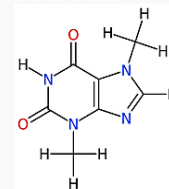
<input type="checkbox"/>	Name	Formula / Structure	Hit	Score
<input type="checkbox"/>	Theobromine; LC-ESI-QTOF; MS2; CE:20 eV; [M+H]⁺	C7H8N4O2 	9	0.9073
<input type="checkbox"/>	DL-4-Hydroxyphenyllactic acid; LC-ESI-QTOF; MS2; CE: 10; R⁻; [M-H]⁻	C9H10O4 	8	0.9051
<input type="checkbox"/>	Theobromine; LC-ESI-QQ; MS2; CE:20 V; [M+H]⁺	C7H8N4O2 	9	0.8961

Theobromine; LC-ESI-QTOF; MS2; CE:20 eV; [M+H]⁺



[metabolomics-usi visualisation](#)

Chemical Structure



ACCESSION: C0000477

RECORD_TITLE: Theobromine; LC-ESI-QTOF; MS2; CE:20 eV; [M+H]⁺

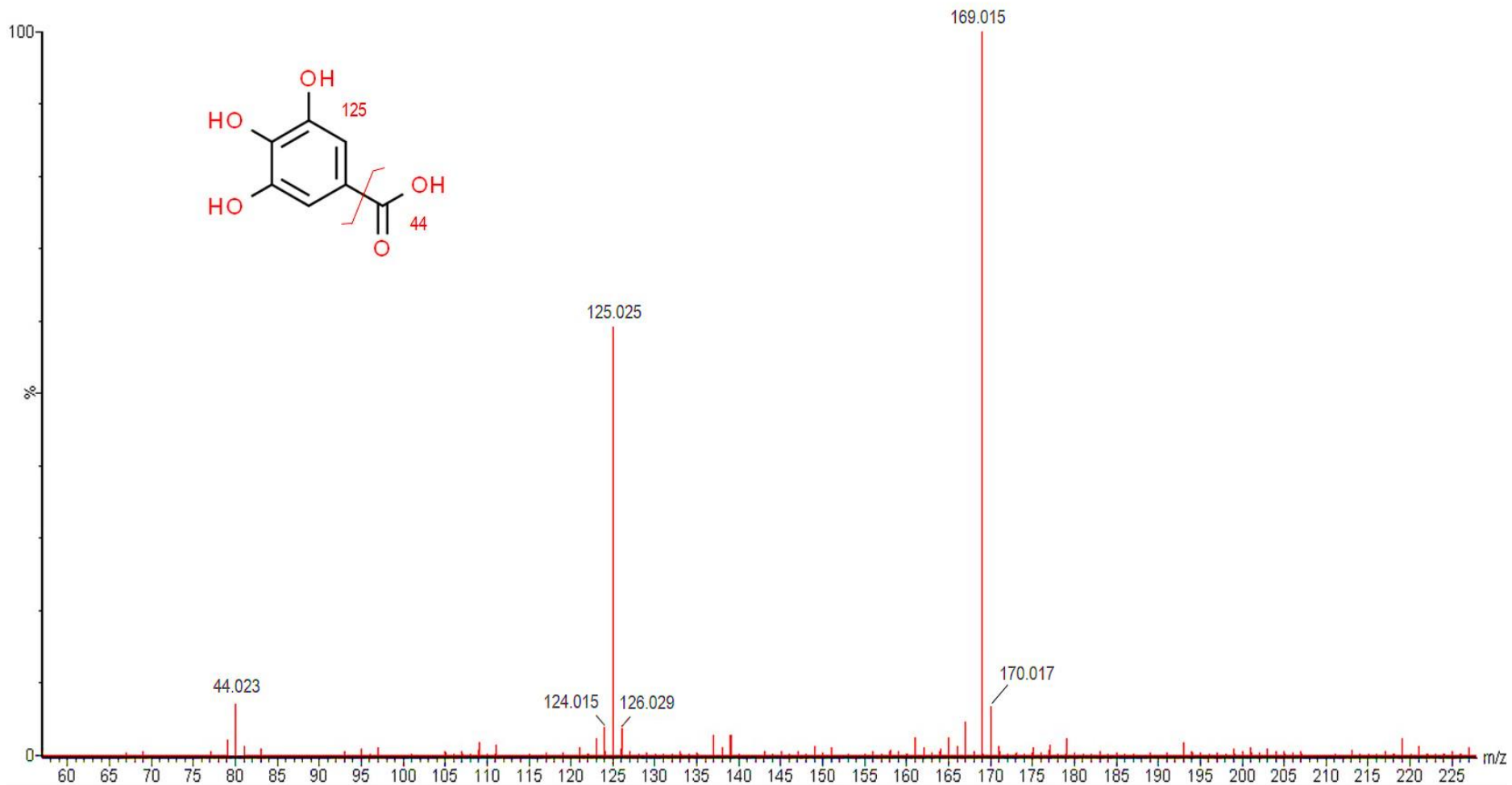
DATE: 2016.01.19 (Created 2008.07.15, modified 2012.11.20)

AUTHORS: Dennis W. Hill, Tzipporah M. Kertesz, Robert Friedman, David F. Grant

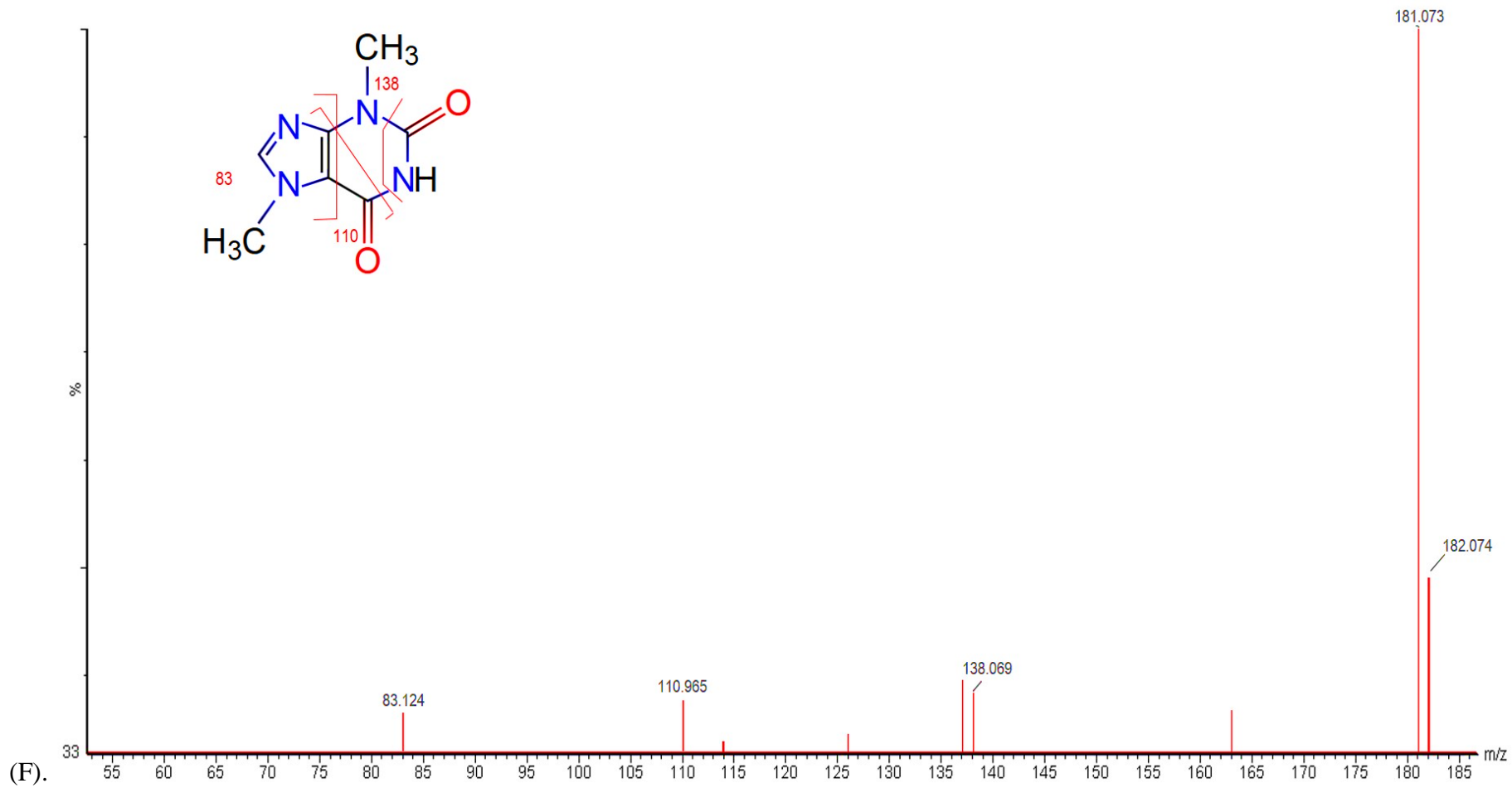
LICENSE: CC BY-SA

PUBLICATION: Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra, <http://pubs.acs.org/doi/abs/10.1021/ac800548g>

(D).



(E).



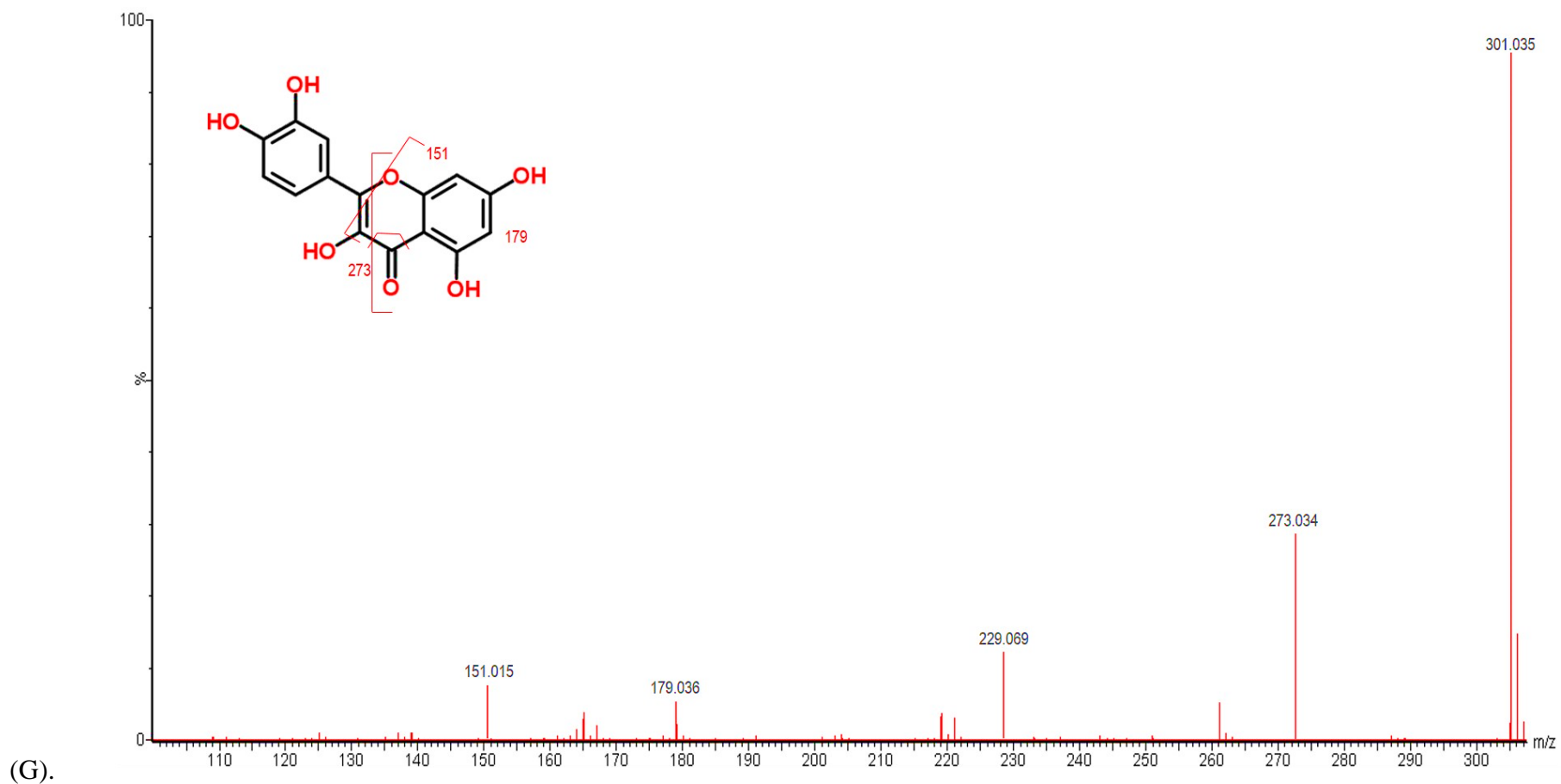


Figure 4.9: (A) Pure caffeine and (B) ECg standards identified through their fragments and confirmed by their hit and score values. A hit of 5 and above, and a score closer to 1 means there is a high certainty of correct identification i.e. caffeine had a hit of 5 and a score of 0.998 and ECg had hit of 8 and score of 0.913. (C) gallic acid and (D) theobromine are the identified peaks in the samples from the negative and positive ionisation modes respectively. (E) high energy negative ionisation fragmentation mass spectrum for gallic acid. (F) high energy positive ionisation fragmentation mass spectrum for theobromine. (G). high energy negative ionisation fragmentation mass spectrum for quercetin.

4.5.5 LR analysis

Nominal LR was performed on the GC-MS, ¹H-NMR, and UPLC-DAD variables, starting with all identified variables and working through the results by taking the most significant i.e. $p < 0.001$, to develop the next model until the best possible model was obtained. In all instances, a confusion matrix was generated to show the number of misclassifications. These results are shown in the figures below.

4.5.5.1 GC-MS LR models

Source	Nparm	DF	ChiSquare	Prob>ChiSq
1-Cyclohexene-1-carboxylic acid	1	1	2.68770517	0.1011
Acetoacetic acid	1	1	31.0642408	<.0001*
Arabinose	1	1	16.6756228	<.0001*
Catechin	1	1	10.130696	0.0015*
Gallic acid	1	1	10.5164949	0.0012*
Glycerol	1	1	0.14945617	0.6991
Phloroglucinol	1	1	12.9818015	0.0003*
Psicose	1	1	46.4458629	<.0001*
Ribitol	1	1	4.05240792	0.0441*
Sucrose	1	1	12.9366782	0.0003*
Threonic acid	1	1	0.01086728	0.9170
Xylonic acid	1	1	1.50817614	0.2194

Confusion Matrix		
Training		
Actual	Predicted Count	
Cultivar	Commercial	NonCommercial
Commercial	44	5
NonCommercial	3	208

Figure 4.10: Nominal LR using all the detected GC-MS variables.

Source	Nparm	DF	ChiSquare	Prob>ChiSq
Acetoacetic acid	1	1	30.171758	<.0001*
Arabinose	1	1	26.1294594	<.0001*
Catechin	1	1	9.44116977	0.0021*
Gallic acid	1	1	13.9005738	0.0002*
Phloroglucinol	1	1	16.3043194	<.0001*
Psicose	1	1	48.4250175	<.0001*
Ribitol	1	1	2.99203747	0.0837
Sucrose	1	1	13.2457502	0.0003*

Confusion Matrix		
Training		
Actual	Predicted Count	
Cultivar	Commercial	NonCommercial
Commercial	43	6
NonCommercial	3	208

Figure 4.11: Nominal LR using the seven statistically significant variables from the total detected variables.

The confusion matrices in Figure 4.10 and 4.11 show five and six Comm cultivars were misclassified as NComm, meaning 90% (44/49) and 88% (43/49) of the genotypes were correctly classified as Comm cultivars, respectively. Psicose and Acetoacetic acid were the most statistically significant variables, and as such the Acetoacetic acid/Psicose ratio was used as a variable to generate a new LR model (Figure 4.12).

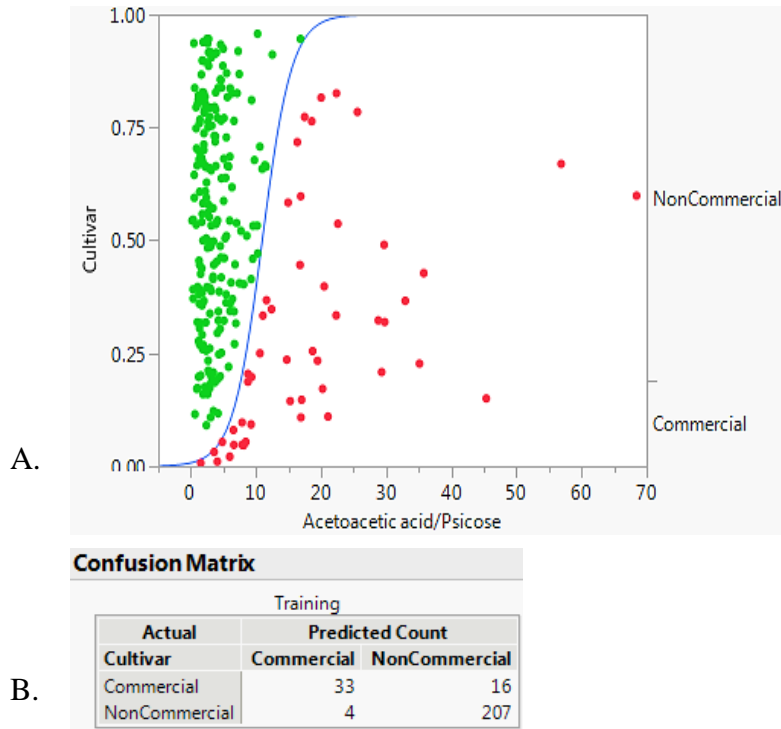


Figure 4.12: (A) shows the LR plot using Acetoacetic acid/Psicose as a variable. (B) shows the confusion matrix obtained from the LR analysis.

4.5.5.2 CHAID decision tree analysis

To confirm the results obtained in the LR model in Figure 4.11, a decision tree was constructed (Figure 4.13).

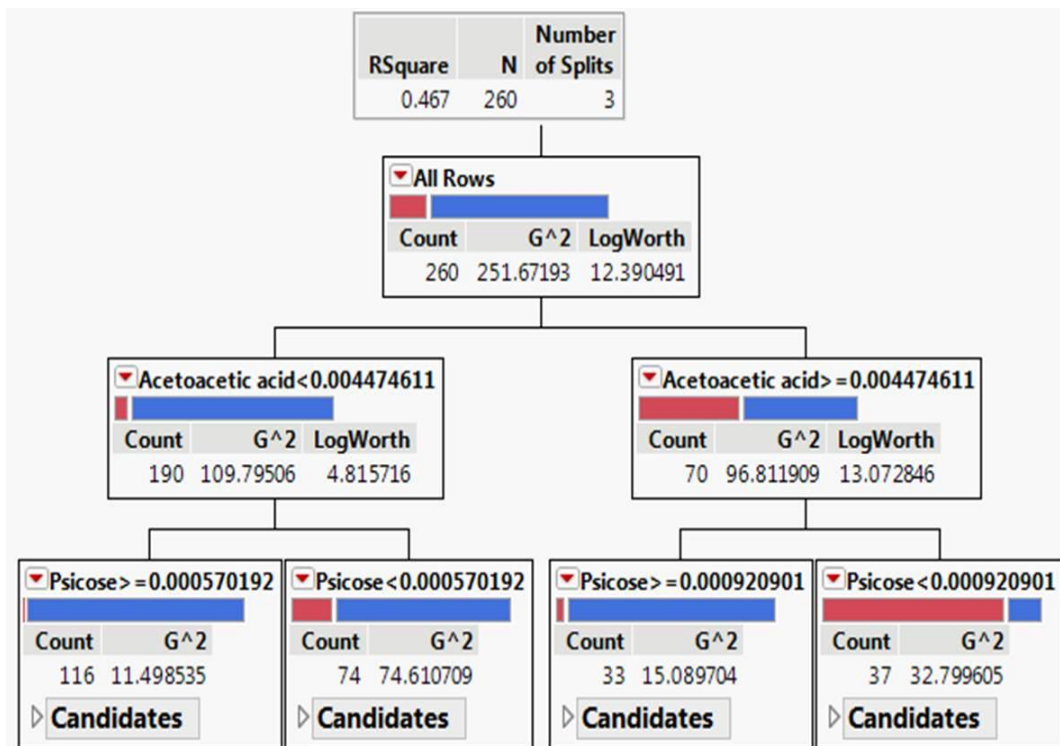


Figure 4.13: Decision tree based on the GC-MS metabolites. The tree shows that Acetoacetic acid and Psicose are predictors for whether a new cultivar will be Comm or NComm.

4.5.5.3 ¹H-NMR LR models

Source	Nparm	DF	L-R	
			ChiSquare	Prob>ChiSq
Acetic acid	1	1	6.0346e-10	1.0000
Alanine	1	1	1.5099e-9	1.0000
CAF	1	1	1.54407e-6	0.9990
CAT	1	1	7.81241e-6	0.9978
Chlorogenic acid	1	1	8.68965e-9	0.9999
EC	1	1	8.9018e-10	1.0000
ECg	1	1	3.5107e-10	1.0000
EGC	1	1	1.96807e-9	1.0000
EGCg	1	1	1.9018e-10	1.0000
Formic acid	1	1	4.84679e-6	0.9982
Gallic acid	1	1	1.4696e-10	1.0000
Glucose	1	1	1.32e-10	1.0000
Isoleucine	1	1	2.97754e-9	1.0000
Leucine	1	1	2.31505e-9	1.0000
Methanol	1	1	191.036729	<.0001*
Quinic acid	1	1	1.0036e-10	1.0000
Sucrose	1	1	1.57977e-8	0.9999
Theanine	1	1	7.4629e-11	1.0000
Valine	1	1	0	1.0000

Confusion Matrix		
Training		
Actual	Predicted Count	
	Commercial	NonCommercial
Commercial	56	0
NonCommercial	0	232

Figure 4.14: Nominal LR using all the detected ¹H-NMR variables.

The confusion matrices in Figures 4.14 show no misclassifications. This means that all 21 ¹H-NMR variables can accurately separate the Comm cultivars from the NComm cultivars. As indicated in the introduction, amino acids have been documented as being important metabolites, which contribute to the quality of tea produced from different tea cultivars. As such, a LR model was developed based on these to see if they could serve as markers to distinguish the Comm from NComm cultivars. The results of this are shown in Figure 4.15 below.

Effect Likelihood Ratio Tests				
Source	Nparm	DF	L-R	
			ChiSquare	Prob>ChiSq
Theanine	1	1	23.4454068	<.0001*
Isoleucine	1	1	14.4877909	0.0001*
Leucine	1	1	11.8225408	0.0006*
Valine	1	1	3.56705993	0.0589
Alanine	1	1	22.0441968	<.0001*

Confusion Matrix		
Training		
Actual Group	Predicted Count	
	Commercial	Non-commercial
Commercial	28	28
Non-commercial	8	224

Figure 4.15: Nominal LR model developed on amino acid variables.

The LR model developed based on the amino acids showed 50% (28/56) Comm cultivars were correctly classified. Next, a LR model based on CAF and the five catechins was developed to see if this model, developed on the ¹H-NMR variables would give the same results as the same model developed on the UPLC-DAD data. This model is shown in Figure 4.16.

Source	Nparm	DF	L-R	
			ChiSquare	Prob>ChiSq
CAF	1	1	86.8761791	<.0001*
CAT	1	1	71.4354484	<.0001*
EC	1	1	0.08390818	0.7721
ECg	1	1	2.19413494	0.1385
EGC	1	1	30.6147274	<.0001*
EGCg	1	1	0.59287867	0.4413

Confusion Matrix		
Training		
Actual	Predicted Count	
	Commercial	NonCommercial
Commercial	51	5
NonCommercial	2	230

Figure 4.16: Nominal LR using all CAF, and all five catechin variables.

The results of this model show that 91% (51/56) of the Comm cultivars were correctly classified, with CAF and CAT being the most significant variables (Figure 4.16), and as such the CAF/CAT ratio was used as a variable to develop a new LR model as seen in Figure 4.17.

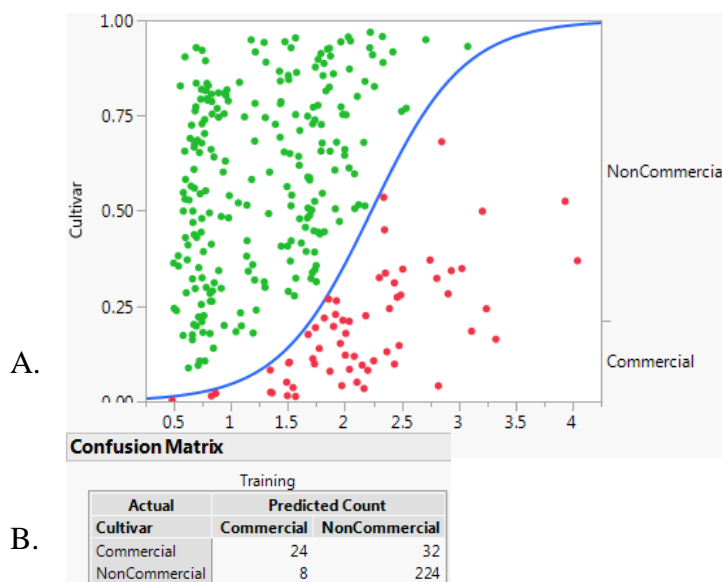


Figure 4.17: (A) shows the LR plot using CAF/CAT as a variable. (B) shows the confusion matrix obtained from the LR analysis.

To confirm the results obtained in the LR model in Figure 4.17, a decision tree was constructed (Figure 4.18).

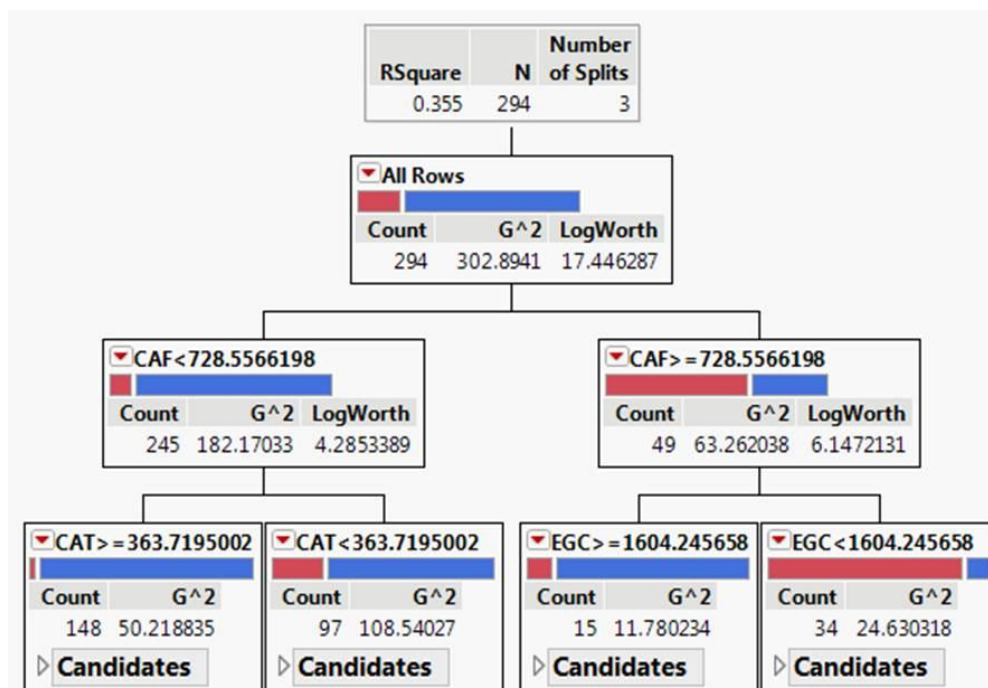


Figure 4.18: Decision tree based on the ¹H-NMR CAF and all five catechin variables. The tree shows that CAF and EGC are predictors for whether a new cultivar will be Comm or NComm.

The decision tree gave rise to a new variable combination i.e. CAF and EGC, which isn't too surprising considering EGC was one of the variables with a $p < 0.0001$ (Figure 4.16). This led to the development of a LR model with CAF/EGC as a variable.

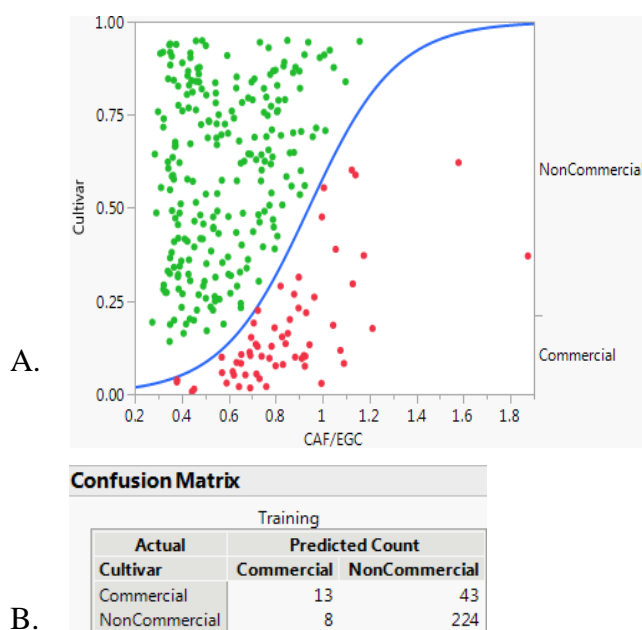


Figure 4.19: (A) shows the LR plot using CAF/EGC as a variable, and (B) its confusion matrix.

Based on the results obtained from the UPLC-DAD data (Figure 4.26), the CAT/EC ratio was shown to work well as a distinguisher between Comm and NComm cultivars. As such, a similar model was developed using the GC-MS data to see whether comparable results would be obtained. This model is shown in Figure 4.20.

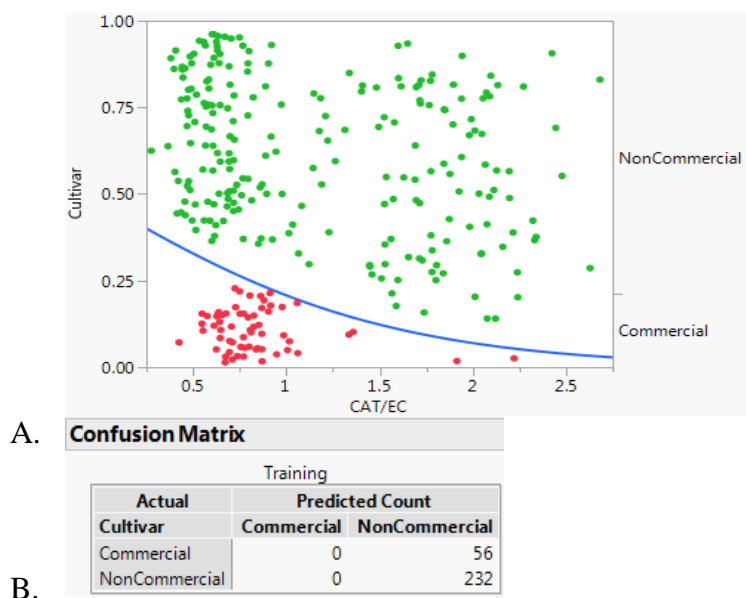


Figure 4.20: (A) shows the LR plot using CAT/EC as a variable. (B) shows the confusion matrix obtained from the LR analysis.

Next, a model based on the CAF/EC ratio, similar to that developed on the UPLC-DAD results was developed, as shown in Figure 4.21.

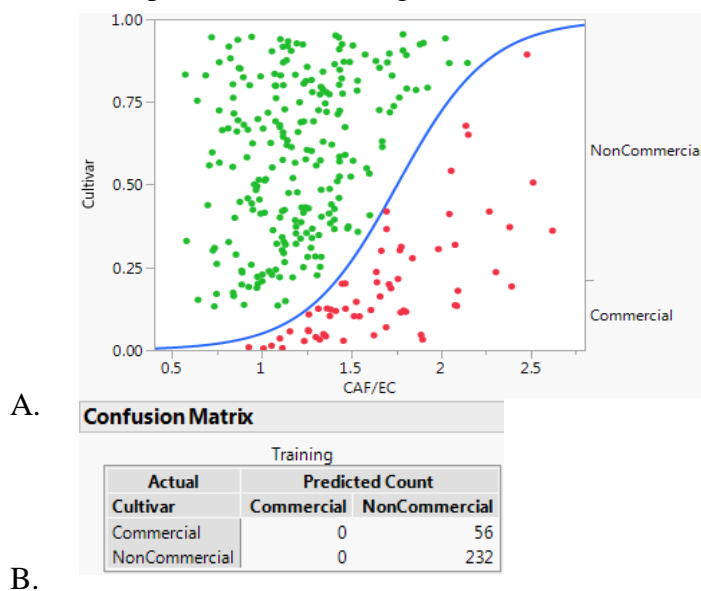


Figure 4.21: (A) shows the LR plot using CAF/EC as a variable. (B) shows the confusion matrix obtained from the LR analysis.

4.5.5.4 UPLC-DAD LR models

Source	Nparm	DF	L-R	
			ChiSquare	Prob>ChiSq
CAF	1	1	4.92929e-6	0.9982
CAT	1	1	5.72304e-7	0.9994
EC	1	1	3.92695e-6	0.9984
ECg	1	1	3.97601e-6	0.9984
EGC	1	1	790.934198	<.0001*
EGCg	1	1	424.393164	<.0001*
Theaflavin	1	1	744.938087	<.0001*
Theaflavin3gallate	1	1	3.19138e-9	1.0000
Theaflavin3'gallate	1	1	13.4755718	0.0002*
Theaflavin3,3'digallate	1	1	24.9898307	<.0001*

Confusion Matrix		
Training		
Actual	Predicted Count	
Cultivar	Commercial	NonCommercial
Commercial	56	0
NonCommercial	0	247

Figure 4.22: Nominal LR using all ten UPLC-DAD variables.

From the results in Figure 4.22, the confusion matrix shows no misclassifications, meaning all ten variables together can accurately separate the Comm cultivars from the NComm cultivars. A LR model was then developed on the four theaflavins, which are markers for tea quality (Obanda *et al.*, 1997; Wright *et al.*, 2002). This model is shown in Figure 4.23.

Source	Nparm	DF	L-R	
			ChiSquare	Prob>ChiSq
Theaflavin	1	1	4.45896e-6	0.9983
Theaflavin3gallate	1	1	0.31955094	0.5719
Theaflavin3'gallate	1	1	0.1818923	0.6698
Theaflavin3,3'digallate	1	1	67.1480193	<.0001*

Confusion Matrix		
Training		
Actual	Predicted Count	
Cultivar	Commercial	NonCommercial
Commercial	54	2
NonCommercial	0	247

Figure 4.14.23: Nominal LR using only the four theaflavin variables.

Figure 4.23 shows that TF4 is the most statistically significantly different metabolite between the Comm and NComm cultivars making it a very important variable. This prompted the development of a LR model based on TF4 (Figure 4.24).

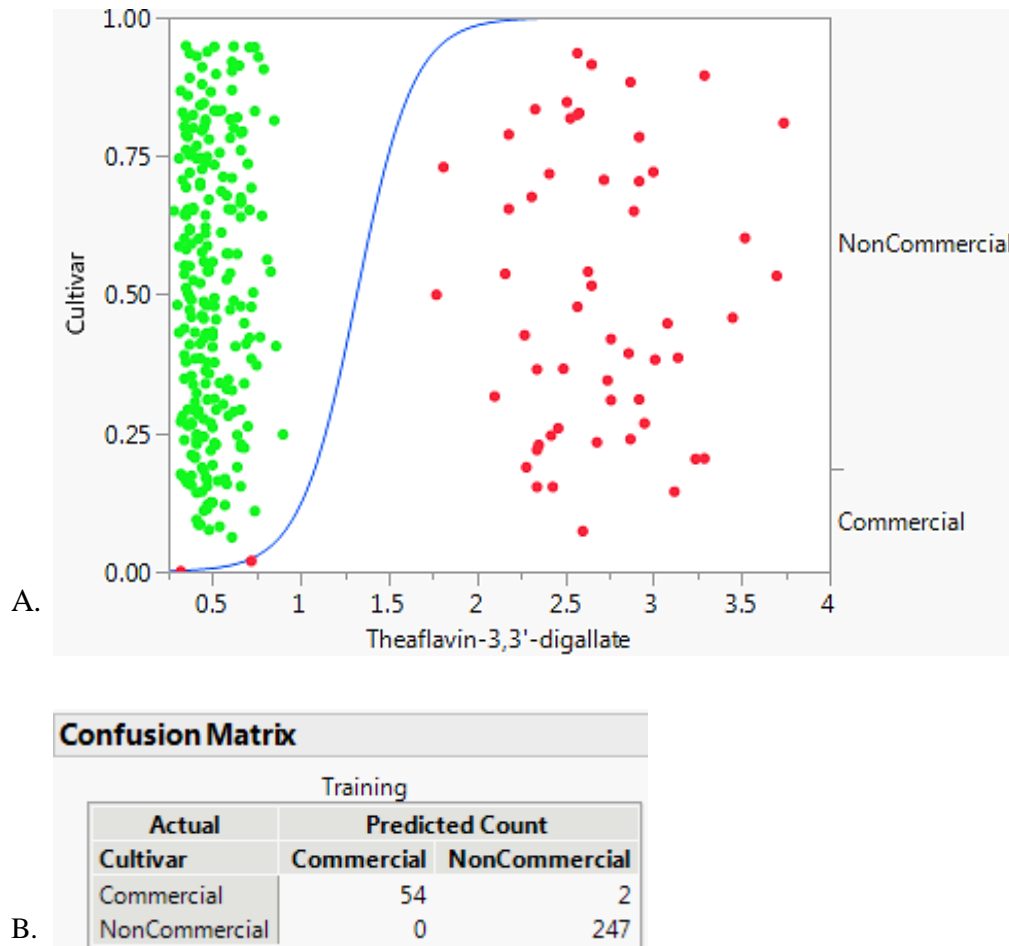


Figure 4.24: (A) shows the LR plot using TF4 as a variable. (B) shows the confusion matrix obtained from the LR analysis.

The confusion matrix in Figure 4.24 shows two Comm cultivars were misclassified as NComm, meaning 96% (54/56) of the genotypes were correctly classified as Comm cultivars, making it comparable to the model developed using all ten variables. In light of the fact that theaflavins are 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, and grown to produce enough shoots to make black tea, a LR model based on CAF, and five catechins of the dried fresh green leaf was developed as shown in Figure 4.25 to see whether these could serve as possible discriminators.

Source	Nparm	DF	L-R	
			ChiSquare	Prob>ChiSq
CAF	1	1	18.2034622	<.0001*
CAT	1	1	40.0911638	<.0001*
EC	1	1	29.1612704	<.0001*
ECg	1	1	6.02876526	0.0141*
EGC	1	1	1.8683257	0.1717
EGCg	1	1	1.02912571	0.3104

Confusion Matrix		
Training		
Actual	Predicted Count	
Cultivar	Commercial	NonCommercial
Commercial	51	5
NonCommercial	3	244

Figure 4.25: Nominal LR using CAF, and all five catechin variables.

Figure 4.25 shows that five Comm cultivars were misclassified as NComm, while three NComm cultivars were misclassified as Comm cultivars. This means 91% (51/56) of the Comm cultivars were correctly classified. To confirm the results obtained in the LR model in Figure 4.25, a decision tree was constructed (Figure 4.26), the results of which coincide with those obtained in the LR model shown in Figure 4.25.

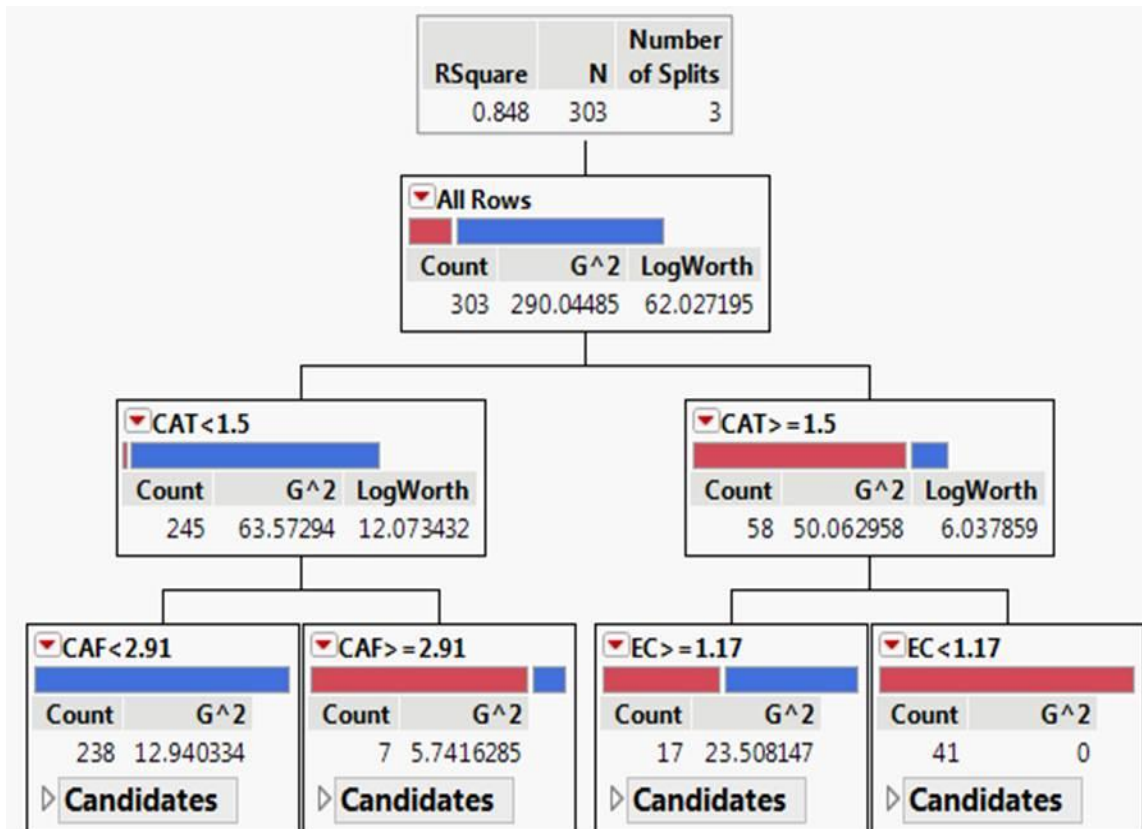


Figure 4.26: Decision tree based on CAF and all five catechin variables. The tree shows that CAT and EC are predictors for whether a new cultivar will be Comm or NComm.

The decision tree shows that CAT and EC are the predictors required to distinguish the Comm cultivars from the NComm cultivars. From the decision tree, 40 of the 56 Comm cultivars or 71%, have a %w/w CAT concentration of > 1.5 and a %w/w EC concentration of < or = 1.13. This therefore means breeders can predict whether a cultivar will be Comm by considering the CAT/EC ratio. As such, a LR model was developed, using the CAT/EC ratio as a predictor, as shown in Figure 4.27.

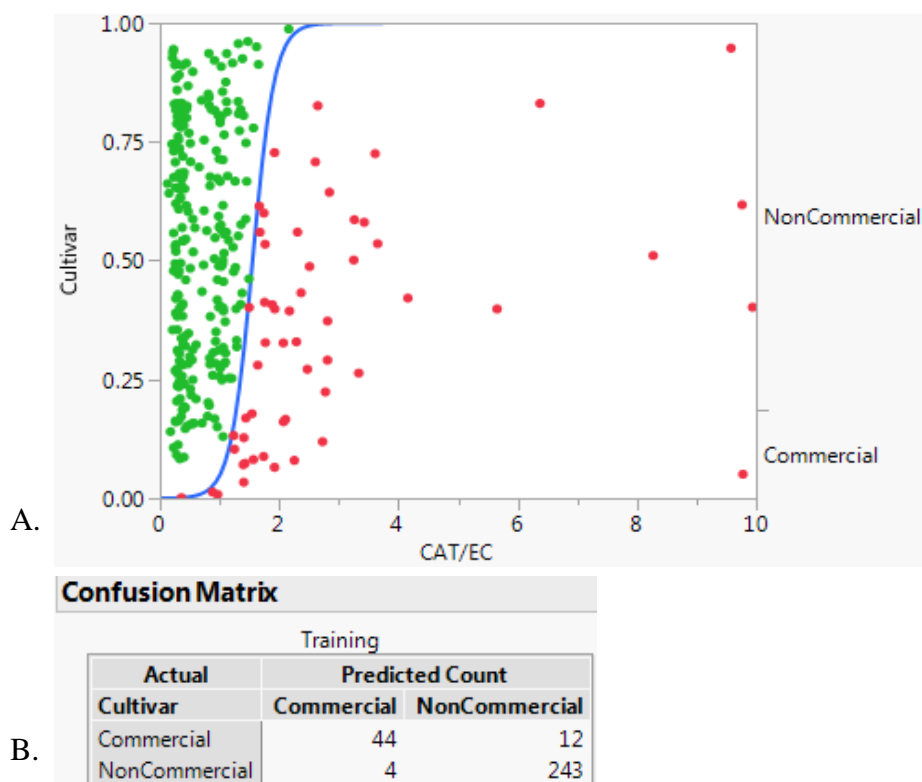


Figure 4.27: (A) shows the LR plot using CAT/EC as a variable. (B) shows the confusion matrix obtained from the LR analysis.

The confusion matrix indicates that 12 Comm cultivars were misclassified as NComm, and four NComm cultivars were misclassified as Comm cultivars. This means the model correctly classified 79% (44/56) of the Comm cultivars. However, from Figure 4.28, it can be seen that the CAT peak is small and elutes very close to an unknown metabolite, which may make it difficult to accurately identify and quantify. This prompted the development, and construction, of another decision tree (Figure 4.30) and LR model (Figure 4.31), in which the CAT variable was excluded, to obtain a new ratio.

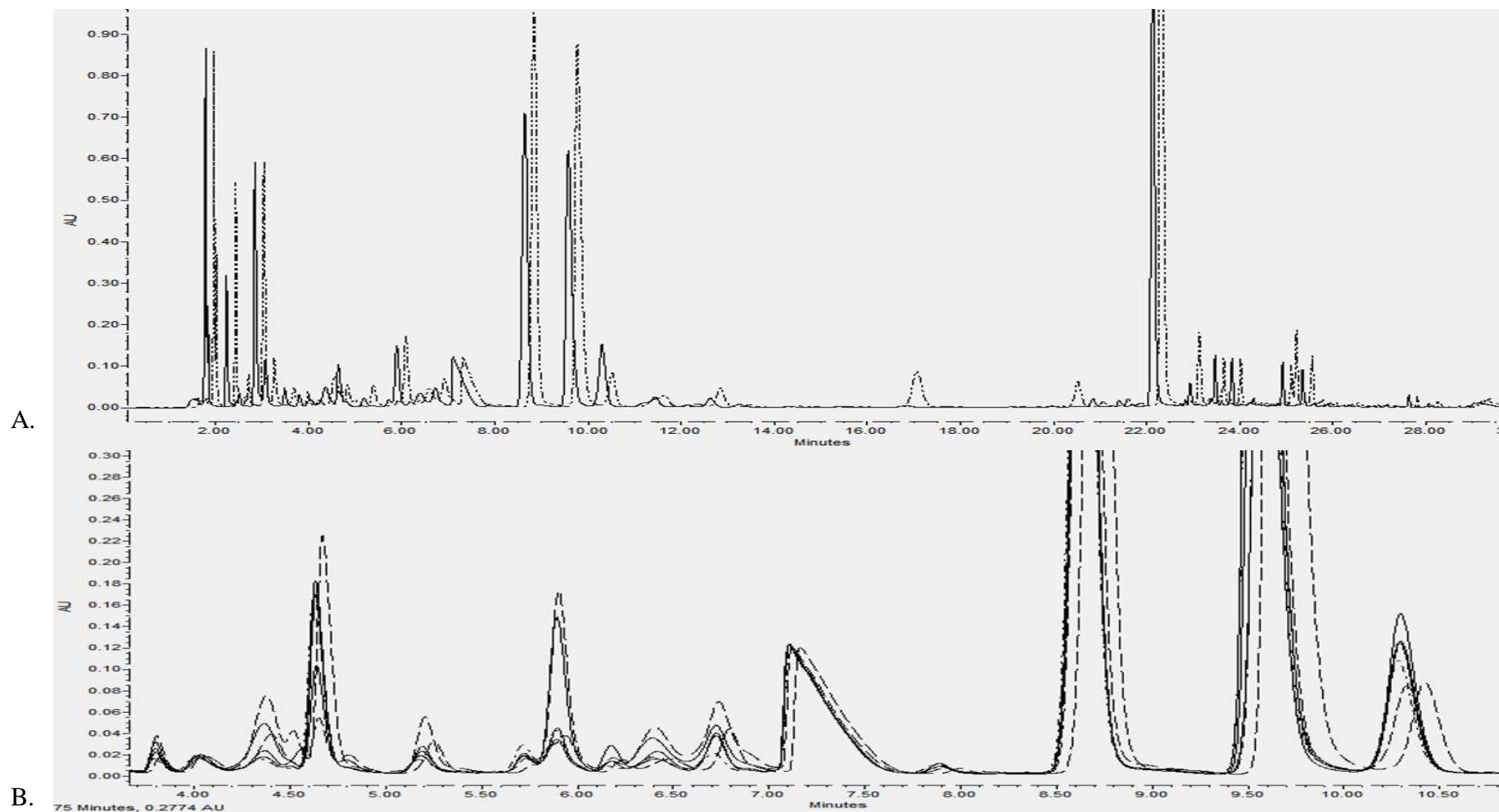


Figure 4.28: (A) Superimposed green tea UPLC/DAD chromatograms of one Comm and one NComm cultivar, offset by 0.25 min for easy identification. The internal standards used were sulphanilamide (1.8 min), Tryptamine (7.3 min) and mycophenolic acid (27.9 min). (B) shows the zoomed in chromatograms of three 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 three dotted lines represents the Comm cultivars, and the three solid lines represents the NComm cultivars.

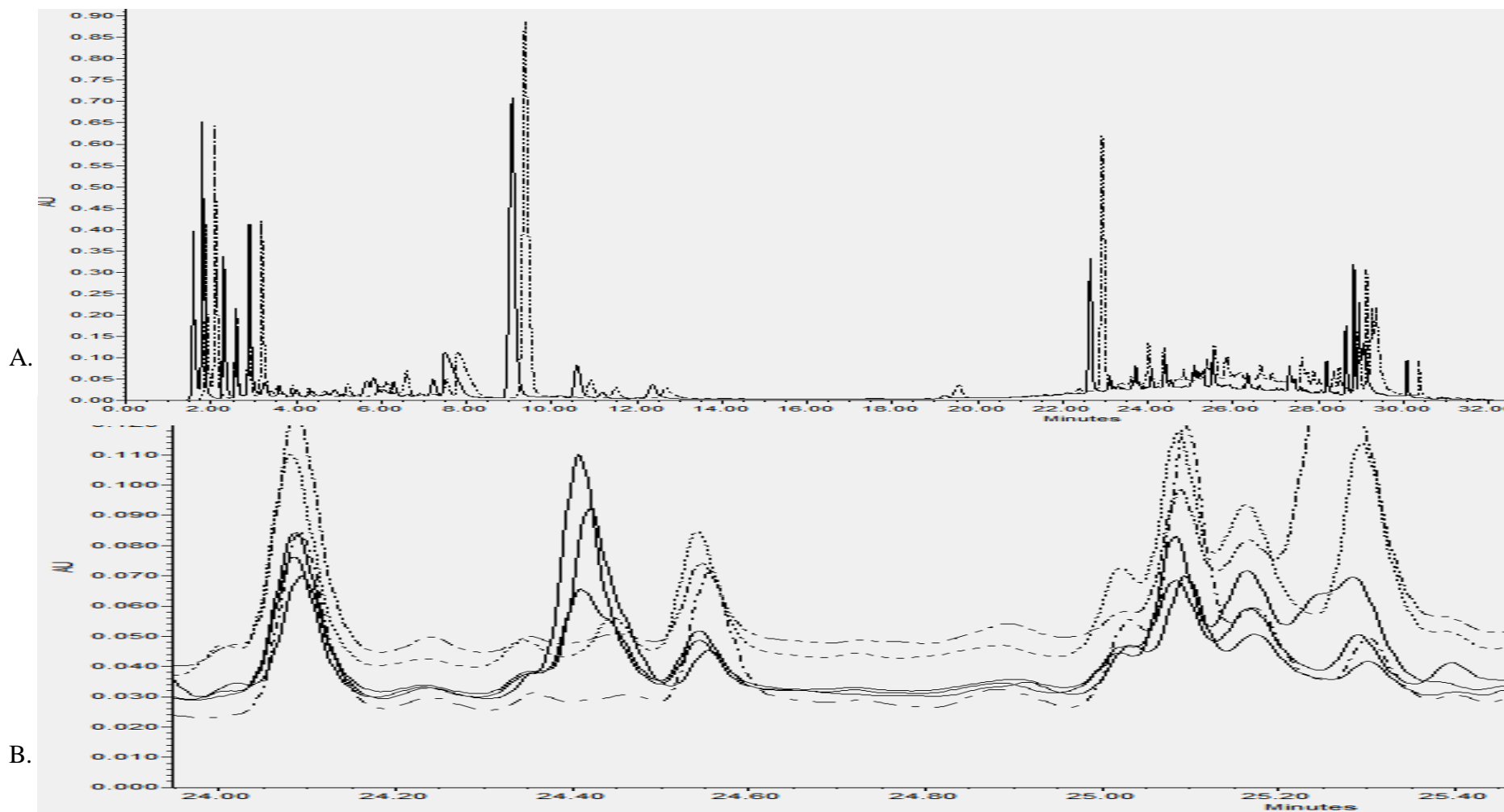


Figure 4.29: (A) Superimposed black tea UPLC/DAD chromatograms of one Comm and one NComm cultivar, offset and standards as in Figure 7. (B) shows the expanded chromatograms of three Comm and three NComm cultivars, showing the position of TF1 (24.05 min), TF2 (24.40 min), TF3 (24.55 min) and TF4 (25.10 min). In both plots, the three dotted lines represents the Comm cultivars, and the three solid lines 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.

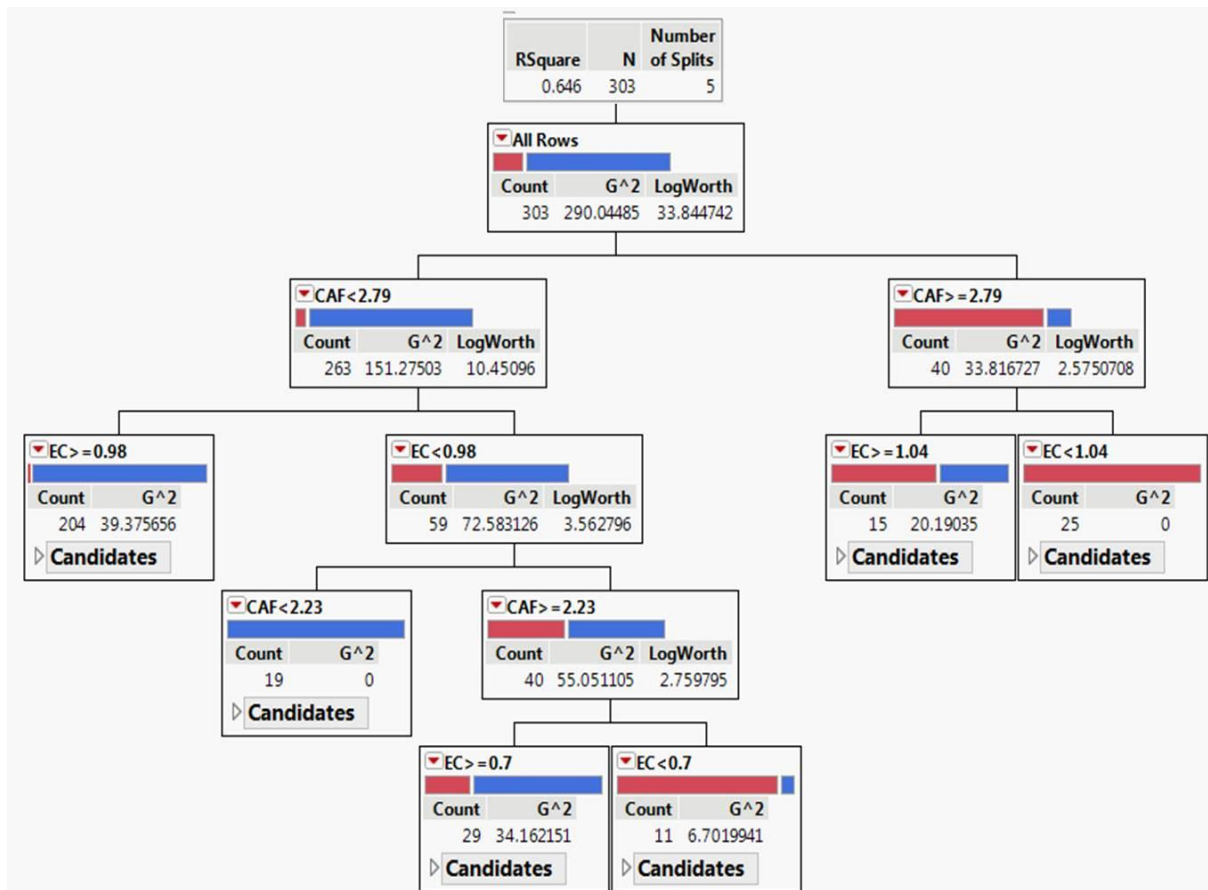


Figure 4.30: Decision tree based on CAF and the catechin variables, excluding CAT. The tree shows that in the absence of CAT, CAF becomes the most significant predictor variable. The CAF/EC ratio serves as a predictor for whether a new cultivar will be Comm or NComm.

A LR model was developed based on the CAF/EC ratio as shown in Figure 4.58 below.

Source	Nparm	DF	L-R	
			ChiSquare	Prob>ChiSq
CAF	1	1	89.6998362	<.0001*
EC	1	1	53.9469241	<.0001*
ECg	1	1	8.9634211	0.0028*
EGC	1	1	1.11649802	0.2907
EGCg	1	1	0.27285565	0.6014

Confusion Matrix		
Training		
Actual	Predicted Count	
Cultivar	Commercial	NonCommercial
Commercial	45	11
NonCommercial	4	243

A.

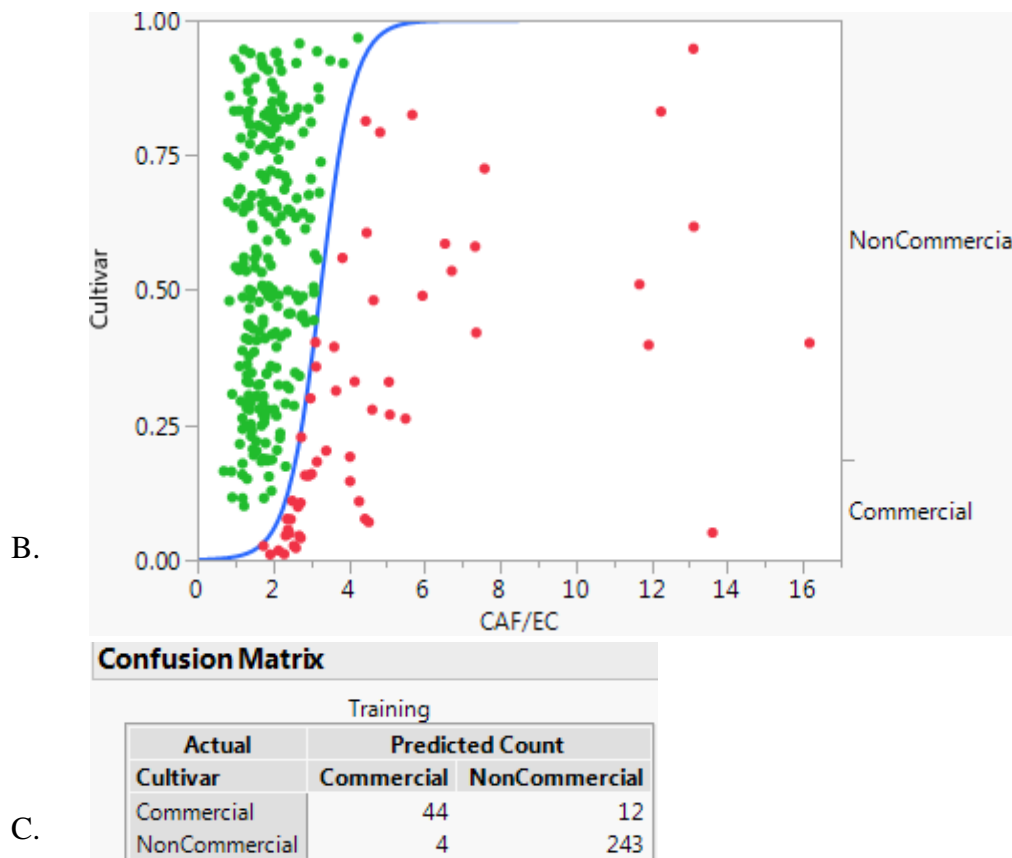


Figure 4.31: (A) shows the LR model for the CAF and catechins, excluding CAT. (B) shows the LR plot based on the CAF/EC ratio. (C) shows the confusion matrix for the CAF/EC ratio.

The LR model based on CAF and the four catechins, with CAT excluded, correctly classified 80% (45/56) of the Comm cultivars. From the confusion matrix of the LR model based on the CAF/EC ratio, it can be seen that the number of misclassifications increases, with 12 Comm cultivars being misclassified as NComm, while four NComm cultivars remain misclassified as Comm cultivars. This therefore means the model correctly classified 79% of the Comm cultivars.

In a study by Wright *et al.*, (2000), 20 high, and 20 low quality tea clones were used 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 results obtained in their study confirmed those of Robertson, (1983), finding that the high and low quality tea cultivars differed significantly in CAT, EC, and ECg. Furthermore, their study showed that CAT correlated least with tea quality, and the reason postulated for this observation was that CAT is not a precursor of any of the four major theaflavins responsible for tea quality. 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). This was also observed with the ungallated EGC highly correlating with quality than

the gallated EGCg. The high and low quality cultivars were thus distinguishable by considering CAT+EC+ECg. A LR model was developed based on CAT+EC+ECg as shown in the Figure 4.32 to see whether or not the findings of Robertson, (1983) and Wright *et al.*, (2000), which were both on Malawian tea cultivars were applicable to the Comm and NComm cultivars used in the present study, which were obtained from Kenya.

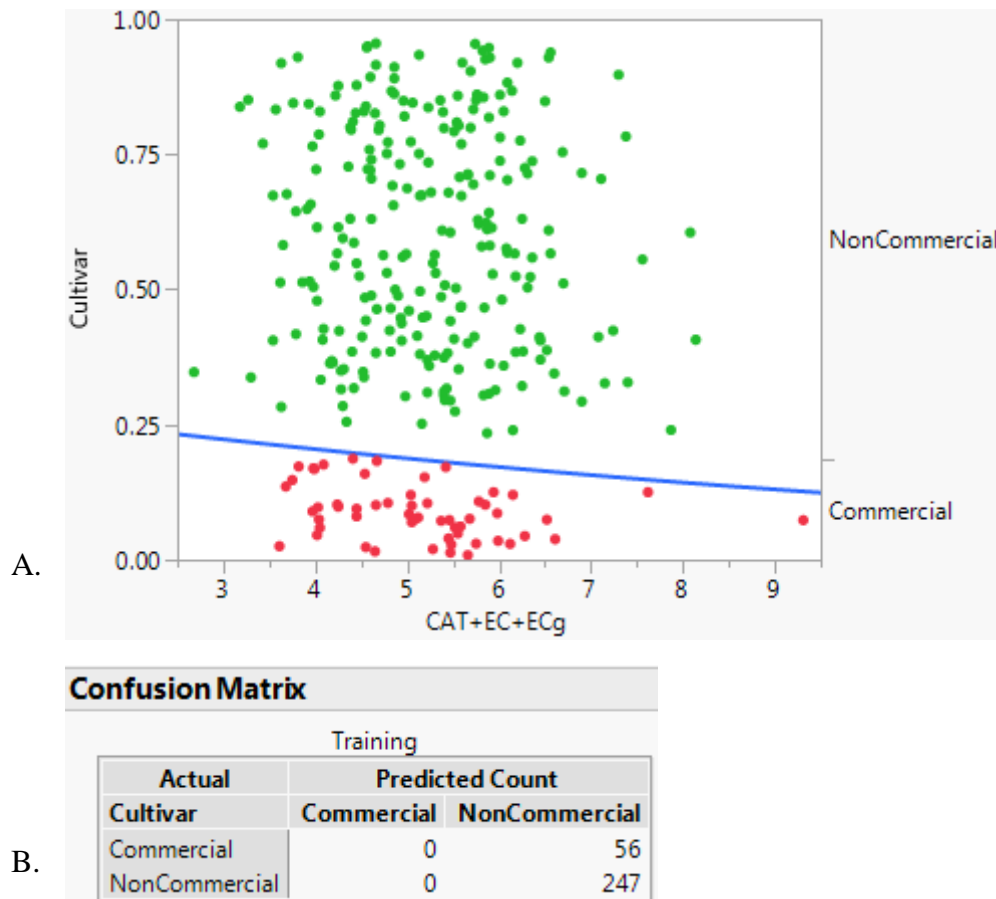


Figure 4.32: (A) shows the LR model for CAT+EC+ECg. (B) shows the confusion matrix for CAT+EC+ECg.

Lastly, the present study also saw the development of a LR model based on the ratio of simple: complex catechins, which was put forward in a study by Ellis and Nyirenda, (1995), as being a distinguisher between high and low quality teas. This LR model is shown in Figure 4.33.

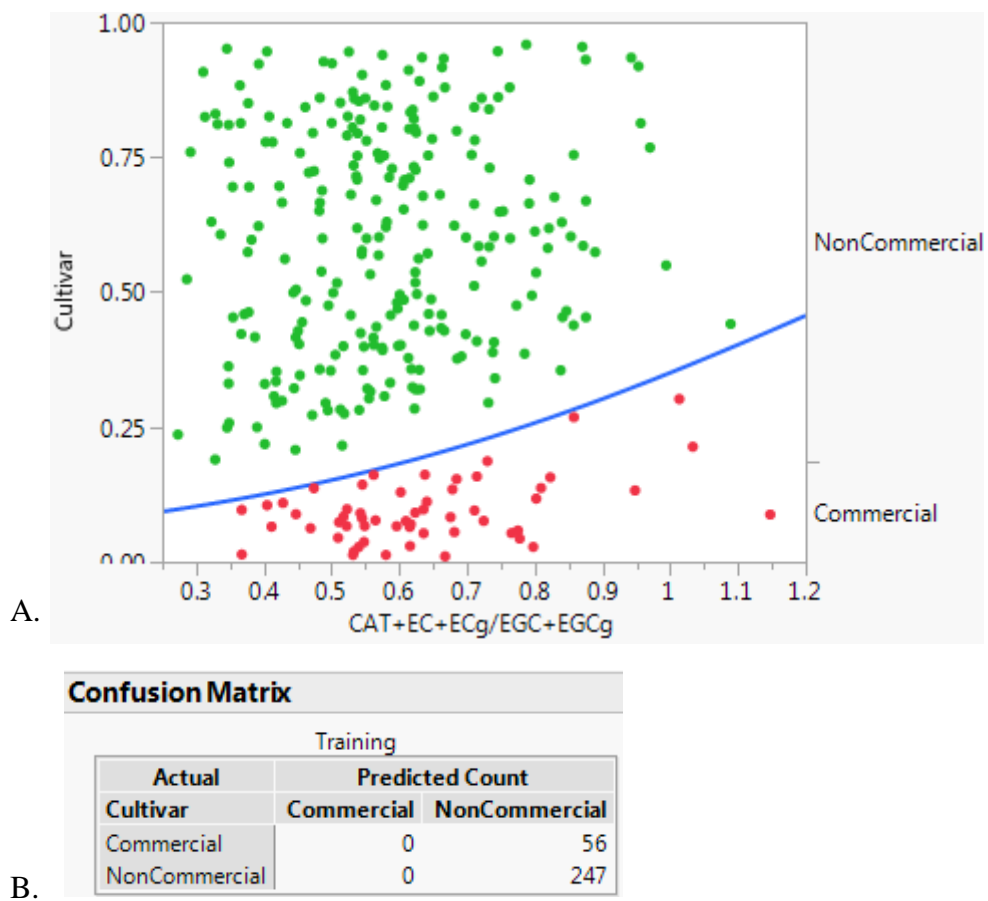


Figure 4.33: (A) shows the LR model for Simple/Complex catechins. (B) shows the confusion matrix for the Simple: Complex catechin model.

Table 4.6: List of LR models developed showing the %specificity and %sensitivity of each. $\text{Prob}(\text{Comm}) = 1 / (1 + \text{Exp}(-\text{Lin}[\text{Comm}])))$.

Model name	Lin (Comm)	%Sensitivity	n=3; (%CV)	%Specificity	n=3; (%CV)
GC-MS					
Model 1 All detected metabolites	$(-1.30) + 102.1 * 1\text{-Cyclohexene-1-carboxylic acid} + 962.2 * \text{Acetoacetic acid} + 762.3 * \text{Arabinose} + (-23.8) * \text{CAT} + 80.1 * \text{Gallic acid} + 7.4 * \text{Glycerol} + (-287.2) * \text{Phloroglucinol} + (-7751.9) * \text{Psicose} + (-137.9) * \text{Ribitol} + (-49.2) * \text{Sucrose} + (-2.7) * \text{Threonic acid} + (-2811.9) * \text{Xylonic acid}$	84	4.0	94	2.3
Model 2 Acetoacetic acid, Arabinose, CAT, Gallic acid, Phloroglucinol, Psicose, Ribitol, Sucrose	$(-1.3) + 865.1 * \text{Acetoacetic acid} + 694.1 * \text{Arabinose} + (-19.9) * \text{CAT} + 78.7 * \text{Gallic acid} + (-280.9) * \text{Phloroglucinol} + (-7755.2) * \text{Psicose} + (-115.5) * \text{Ribitol} + (-39.9) * \text{Sucrose}$	78	3.3	93	1.9
Model 3 Psicose/Acetoacetic acid	$2.2 + (-21.1) * \text{Psicose/Acetoacetic acid}$	67	5.1	77	1.2
¹H-NMR					
Model 1 All detected metabolites	$8.3 + (-0.6) * \text{Acetic acid} + (-0.5) * \text{Alanine} + 0.05 * \text{CAF} + -0.03 * \text{CAT} + 0.2 * \text{Chlorogenic acid} + (-0.03) * \text{EC} + (-0.03) * \text{ECg} + (-0.01) * \text{EGC} + 0.02 * \text{EGCg} + (-1.4) * \text{Formic acid} + 0.05 * \text{Gallic acid} + 0.02 * \text{Glucose} + (-2.7) * \text{Isoleucine} + 2.5 * \text{Leucine} + 2.1 * \text{Methanol} + 0.006 * \text{Quinic acid} + (-0.08) * \text{Sucrose} + (-0.007) * \text{Theanine} + 0.002 * \text{Valine}$	100	3.4	100	1.1

Model 2 Alanine, Isoleucine, Leucine, Theanine, Valine	$1.7 + (-0.3)*\text{Alanine} + (-0.9)*\text{Isoleucine} + 1.1*\text{Leucine} + 0.006*\text{Theanine} + 0.3*\text{Valine}$	50	5.3	78	2.1
Model 3 CAF, CAT, EC, ECg, EGC, EGCg	$(-1.2) + 0.02*\text{CAF} + (-0.01)*\text{CAT} + 0.006*\text{EC} + (-0.02)*\text{ECg} + (-0.01)*\text{EGC} + 0.01*\text{EGCg}$	91	6.2	96	1.8
Model 4 CAF/CAT	$(-5.9) + 2.6*\text{CAF}/\text{CAT}$	43	5.3	75	2.1
Model 5 CAF/EGC	$(-5.06) + 5.4*\text{CAF}/\text{EGC}$	23	4.4	62	1.7
Model 6 CAT/EC	$(-0.09) + (-1.2)*\text{CAT}/\text{EC}$	0	0	100	0
Model 7 CAF/EC	$(-6.8) + 3.9*\text{CAF}/\text{EC}$	0	0	100	0
UPLC-DAD					
Model 1 CAF, CAT, EC, ECg, EGC, EGCg, TF, TF2, TF3, TF4	$(-290.7) + 47.3*\text{CAF} + 35.6*\text{CAT} + (-21.1)*\text{EC} + (-5.1)*\text{ECg} + (-112.8)*\text{EGC} + 54.9*\text{EGCg} + 242.5*\text{TF} + (-6.4)*\text{TF2} + (-461.3)*\text{TF3} + 562.2*\text{TF4}$	86	8.1	100	1.2
Model 2 TF, TF2, TF3, TF4	$(-2.2) + (-0.9)*\text{TF} + (-5.5)*\text{TF2} + (-3.0)*\text{TF3} + 8.4*\text{TF4}$	100	0	100	0
Model 3 TF4	$(-8.2) + 6.2*\text{TF4}$	100	0	100	0
Model 4 CAF, CAT, EC, ECg, EGC,	$(-15.6) + 5.6*\text{CAF} + 6.4*\text{CAT} + (-6.9)*\text{EC} +$	79	22.3	98	1.2

EGCg	$(-1.3)*ECg + (-0.8)*EGC + 0.7*EGCg$				
Model 5 CAT/EC	$(-8.4) + 5.4*CAT/EC$	76	6.1	98	1.6
Model 6 CAF, EC, ECg, EGC, EGCg	$(-15.6) + 5.6*CAF + 6.4* + (-6.9)*EC + (-1.3)*ECg + (-0.8)* EGC + 0.7*EGCg$	86	16.9	97	2.6
Model 7 CAF/EC	$(-7.4) + 2.3*CAF/EC$	60	6.8	98	1.8
Model 8 CAT+EC+ECg	$(-0.9) + (-0.1)*CAT+EC+ECg$	0	0	100	0
Model 9 (CAT+EC+ECg)/(EGC+EGCg)	$(-2.8) + 2.2*((CAT+EC+ECg)/(EGC+EGCg))$	0	0	100	0

Total number of Comm cultivars cultivars used in this study were 49 (GC-MS) and 56 (¹H-NMR and UPLC-DAD). The %Sensitivity and %Specificity were calculated using formulas given in 4.4.9.

4.6 DISCUSSION AND CONCLUSION

Metabolomics statistical data analysis can be employed as a supportive tool to aid breeders in the selection and improvement process for new tea cultivars. As mentioned, the objective of this study was to identify potential classifiers for the 303 genotypes investigated into either of the two groups, Comm or NComm cultivars, with GC-MS, ¹H-NMR, UPLC-DAD and UPLC-MS as the metabolomics platforms. Figures 4.1 to 4.4 show the violin plots for the GC-MS, ¹H-NMR, UPLC-DAD and UPLC-MS data, respectively, showing the differences in each of the detected metabolites in the Comm and the NComm cultivars. Violin plots serve as a conspicuous means of visualising the differences that exist between classes, carrying substantial statistical information about e.g. medians and outliers. When the mean of one class falls outside the 25th and 75th percentile of the second group, as is seen in Figure 4.2, this indicates that there is a statistically significant difference between these two classes, with respect to that particular metabolite. The GC-MS metabolites arabinose, CAT, 1-cyclohexenecarboxylic acid, psicose, ribitol, sucrose, and threonic acid; ¹H-NMR metabolites CAF, EGC, EGCg, Formic acid, Leucine, Methanol, and Sucrose; UPLC-DAD metabolites CAF, CAT, EC, TF1-TF4, and yield; UPLC-MS argininosuccinate, caffeic acid, CAF, CAT, citric acid, EC, EGCg, gallic acid, gluconic acid, glucose, maltose, quercitin and theanine in Figure 4.4 clearly differentiate the Comm cultivars from the NComm cultivars, making these ideal predictors to be employed in classifying the 303 genotypes into the two classes. Furthermore, the PCA and PLS-DA plots in Figures 4.5 and 4.6, respectively, show that the detected metabolites on the ¹H-NMR and UPLC-DAD platforms separate the Comm cultivars from the NComm cultivars, while the GC-MS plots show some overlap between the two groups, signifying that the GC-MS platform metabolites are not capable of separating the Comm from the NComm cultivars. From Figures 4.5 (B) and 4.6 (B), it can be seen that two Comm cultivars cluster with the NComm cultivars. These two Comm cultivars are the parental clones GW Ejulu and TRFK 303/577. The observed clustering is because the metabolite concentrations in both parental clones for the metabolites responsible for separating the two groups are similar to those of the NComm cultivars and as such these parents will cluster with their offspring. This clustering is however not observed in Figures 4.5 (C) and 4.6 (C) because the concentrations of the detected UPLC metabolites in the parental clones are similar to those of the other Comm cultivars and therefore the parental clones have clustered with the Comm cultivars; no clear separation was observed in Figures 4.5 (A) and 4.6 (B) so it can not be determined whether the GC-MS metabolites cluster the

parental clones with the other Comm cultivars, or with the NComm cultivars. Tables 4.1, 4.2 and 4.3 show a total of 12, 19 and 10 metabolites were identified using GC-MS, ¹H-NMR and UPLC-DAD respectively. The GC-MS results showed that arabinose, catechin, gallic acid, glycerol, phloroglucinol, sucrose and xylonic acid, a sugar acid generated through the complete oxidation of xylose, were detectable metabolites, which separated the Comm from the NComm cultivars in terms of arbitrary response units. Arabinose, sucrose and xylonic acid were higher in the Comm cultivars. These three metabolites have been shown to have positive correlations with metabolites such as ribose and sucrose, in other studies. They have been shown to play a role in improving the sweet taste of sugarcane and to be up-regulated during strawberry fruit maturation (Zhang *et al.*, 2010). It can thus be postulated that the tea cultivars with high levels of these sugars will produce sweet-tasting, higher-quality liquor. Arabinose has been reported in a study on serendipity berries from the *Dioscoreophyllum cumminsii* Diels plant, which is indigenous to tropical West Africa, and is grown in Guinea, Cameroon, and in the rain forests of central Africa. Arabinose levels were found to be higher in the fruits of some varieties of these berries and were determined to be the reason for the sweetness in these varieties (Inglett, 2012). In addition, polyols such as arabitol and ribitol, which also enhance the sweetness of fruits, have been reported in the literature at concentrations ranging between 20 and 60 mg/g dry weight (Roser *et al.*, 1992). In a study evaluating the sucrose, and taste-related amino acids content in soybean, a sucrose concentration of 30.9 mg/g dry weight was reported (Kumar *et al.*, 2011). The present study, however, detected average sucrose levels of 15 mg/g and 13.6 mg/g dry weight for the Comm and NComm cultivars respectively. Our results for higher concentrations of arabinose and sucrose in the Comm cultivars, agree with those in the literature where these compounds are higher in sweeter sugarcanes, ripening strawberries and sweeter cultivars of serendipity berries.

Literature has shown that abiotic stress such as drought affects the photosynthetic pathway of plants, and in so doing drastically impacts their primary metabolism, which in turn affects sugars, sugar alcohols, and amino acids. The DT plants, tend to be of a higher quality than the drought susceptible (DS), as they efficiently up-regulate their production of sugars, which they utilise as an energy source during stress (Nyarukowa *et al.*, 2016). This could also explain why the DT cultivars produce better tasting liquor. Further, carbohydrates have also been shown to influence the biosynthesis of other energy-generating metabolites, responsible

for the alteration of gene expression and signal transduction (Hoekstra *et al.*, 2001). Phloroglucinol is a plant polyphenolic compound, which possesses antioxidant properties. It has been compared to ascorbic acid, and has been shown to be more powerful against e.g. DPPH and peroxide radicals; it is considered a natural antioxidant. Phloroglucinol has been described in the literature as sweet, and contributes to the sweet fruity taste of the grapes used to make Pinot noir wines (Cortell *et al.*, 2008). Phloroglucinol was one of the metabolites identified, which was higher in the Comm cultivars as compared to the NComm cultivars. The higher concentration of phloroglucinol in the Comm cultivars agrees with the higher concentration of this compound in sweet fruity Pinot noir grapes. The higher concentration of phloroglucinol in the Comm cultivars can be explained in part by the fact that DT cultivars have been shown to have higher levels of polyphenols, which in turn results in them having a higher levels of antioxidants so they are able to scavenge free radicals better than the DS cultivars and this results in their survival under drought stress (Nyarukowa *et al.*, 2016). Malic acid is a dicarboxylic acid, which was identified as a distinguisher between the Comm and NComm cultivars. This compound has been reported in ripening apples at concentrations of 10 mg/g (Ackermann *et al.*, 1992); it has been reported to be responsible for the sour taste of fruits. In another study on apples by Ma *et al.* (2015), where the sugar and malic acid composition in cultivated vs wild apples was compared, it was found that a significant difference between the malic acid concentrations of the cultivated vs the wild apples existed. Furthermore, the study also showed that malic acid composition highly correlated with that of glucose and sucrose contents, suggesting that the selection of fruit acidity also has a significant effect on the amounts of sugars present in apple fruits. This means sugar metabolism is influenced by malic acid accumulation. The wild apples were shown to be more acidic as compared to the cultivated apples. Apple breeders select for apples richer in malic acid content, due to its strong impact on sugar concentration in apples, resulting in sweeter apples (Ma *et al.*, 2015). The Comm tea cultivar results for malic acid agree with the malic acid results found in literature for commercial apple cultivars (Ma *et al.*, 2015). Psicose was also detected in this study; this metabolite has been documented to confer sweetness to, for example, Worcester sauce and fruit juice. Oshima *et al.*, (2006) reported psicose, a product of fructose breakdown, at concentrations ranging from 0.005 mg/g in coffee to 1.31 mg/g in Worcester sauce. Our study found the arbitrary psicose units to be significantly lower in the Comm cultivars compared to the NComm cultivars, indicating it is a significant distinguisher.

Some key metabolites responsible for the taste of tea are caffeine, which comprises up to 5% of the shoot dry weight; theobromine and theophylline, which are < 3% of the shoot dry weight. Caffeine has, in addition to being a stimulant, been documented to contribute to tea briskness, while theophylline and theobromine have been shown to contribute to the mellowness and sweetness of oolong tea (Chaturvedula and Prakash, 2011). The ¹H-NMR results indicate that levels of caffeine were higher in the Comm cultivars as compared to the NComm cultivars. According to Chin *et al.*, (2008), the average caffeine content in black, and green tea is 7-30 mg/g serving of tea, with the average size of a tea bag being 2 g. The current untargeted study found average green leaf caffeine content of 12.6 and 11.6 mg/g in the Comm and NComm cultivars, respectively. These results agree with those of Mazzafera and Silvarolla (2010), which showed that coffee beans from high-quality cultivars had a higher caffeine concentration of 25 mg/g dry weight compared to the low-quality cultivars with a lower caffeine concentration of 9.64 mg/g.

The ¹H-NMR results show that amino acids valine and isoleucine were detected across all samples. These two amino acids were found to be higher in the Comm cultivars. Moreover, amino acids have been documented to improve taste and aroma in tea infusions, namely alanine, leucine, phenylalanine, tryptophan, tyrosine, and valine. Alanine and phenylalanine have been reported as being responsible for the flowery and rose-like aromas of tea, respectively, whilst leucine has been shown to produce a spicy aroma (Sanderson and Grahamm, 1973).

Chlorogenic acid was one of the metabolites consistently detected in the Comm and NComm cultivars. In a study by Szejtli and Szenté (2005), chlorogenic acid was complexed with the tasteless β -cyclodextrin, to eliminate the undesired bitter taste and resultant sensation in the mouth. In another study on apples, it was shown that from the 20 varieties investigated, the sweeter varieties were those with lower chlorogenic acid concentrations, as low as 1.8 mg/g fresh weight, with the less sweet varieties having as high as 6.9 mg/g fresh weight (Marks *et al.*, 2007). In the present study, chlorogenic acid concentrations were comparable in both groups with the Comm and NComm cultivars having average dry weight concentrations of 4.2 and 4 mg/g, respectively.

Quinic acid was also detected in the Comm and NComm cultivars. Literature has documented quinic acid to be responsible for the astringency taste associated with coffee (Buffo and Cardelli-Freire, 2004). However, there was no statistically significant difference between the Comm and NComm cultivars for quinic acid. The metabolite 1-cyclohexenecarboxylic acid

was also detected in the present study, and was significantly higher in the Comm cultivars than the NComm cultivars. Cinnamic acids are trans-phenyl-3-propenoic acids found in plants as e.g. p-coumaric, caffeic, ferulic, dimethoxycinnamic, and trimethoxycinnamic acids, to name a few. These metabolites conjugate, through their carboxylic groups, with amino acids, polysaccharides, and glycosides. The most predominant of these reactions is the transesterification with quinic acid to form cinnamate esters, which are collectively known as chlorogenic acids. Furthermore, cinnamic acids conjugated with a derivative from quinic acid, shikimic acid, results in 3,4,5-trihydroxy-1-cyclohexenecarboxylic acid, which is also a cinnamate ester. 1-cyclohexenecarboxylic acid, like other chlorogenic acids, is bitter in taste, and has been documented to confer bitterness in green coffee beans (Baeza *et al.*, 2016).

The present study also detected methanol as a potential biomarker, separating the Comm from the NComm cultivars. In a study by Fall and Benson (1996), it was documented that methanol is a natural metabolism product, emitted from plant leaves. Their study showed high concentrations of methanol in the forest air, substantiating the likelihood that the forest plants were producing this compound. Employing GC analysis or direct enzymatic analysis of gas-phase methanol, significant methanol emissions were observed from the leaves of forest plants using leaf and branch enclosure approaches, typically ranging between 0.0003 - 0.017 mg methanol per g dry weight, with young leaves emitting up to 0.04 mg methanol per g dry weight (Nemecek-Marshall *et al.*, 1995). Methanol has been shown to possess a sweet taste and is used in artificial sweeteners (Chattopadhyay *et al.*, 2014). The present study found the concentrations of methanol in the Comm cultivars to be double that of the NCom, at 0.2 and 0.1 mg/g, respectively.

Glycerol is another metabolite that was detected on the GC-MS platform. Literature shows that glycerol enhances the aroma and sweetness levels in wines. In a study investigating the effects of glycerol in red and white wines, it was documented that a glycerol concentration of 10 g/L was sufficient to enhance the aroma, and suppress the bitterness, resulting in these wines being reported as sweet-tasting even when glucose and fructose levels were below the detection threshold reported in other studies (Jones *et al.*, 2008). This study found no significant difference in the concentrations of glycerol in the Comm and NComm cultivars. Since several metabolites described above relate to sweetness, the summation of the metabolites; arabinose, arabitol, glycerol, malic acid, phloroglucinol, psicose, ribitol, sucrose, and xylonic acid for the GC-MS, and glucose, methanol and sucrose for the ¹H-NMR, in each of the 303 cultivars was calculated. This sum of sweeteners was labelled total sweeteners.

Interestingly, the total sweeteners were significantly higher ($P < 0.05$) in Comm cultivars. We report here for the first time in the leaves of black tea cultivars that several metabolites, related to sweetness, which are higher in the Comm than the NComm cultivars. Sweetness may have contributed to these cultivars being selected for commercial production since the 1950s.

As documented below, several metabolites were also detected that are responsible for other taste qualities i.e. bitterness and umami. Plants are rich in bitter-tasting metabolites, which serve to deter herbivores. To reduce bitterness or off-taste, plants produce sweet, acidic, or strong fruity flavoured molecules, which mask the bad tastes (Tripathi *et al.*, 2011). The present study shows that the Comm cultivars have a higher total sweeteners concentration of 9.3 mg/g, which is significantly higher ($P < 0.05$) from the total sweeteners concentration in the NComm of 8.7 mg/g dry weight. Furthermore, the total amino acid concentration, calculated by adding the amino acids alanine, isoleucine, leucine, theanine, and valine, was higher in the Comm than the NCom, at 21.8 and 20.9 mg/g dry weight. These compounds could, therefore, mask the bitterness, resulting from e.g. caffeine and chlorogenic acid. As mentioned in the foregoing, amino acids are responsible for the aroma of tea, therefore, the higher the total amino acids, the more aromatic the tea. The total amino acid and total sweeteners concentrations being higher in the Comm cultivars result in the teas produced from these cultivars having a better taste.

Linolenic acid was another metabolite detected by the GC-MS, which was higher but not statistically significant, in the Comm cultivars. Linolenic acid has been documented as being responsible for the bitter taste observed in soybean lecithins, and linseed oil obtained from different varieties. A study by Toumi *et al.*, (2008), revealed linolenic acid to be the most abundant fatty acid in grapevine leaves, and it was higher in the leaves of DT cultivars, which corroborates our findings. The total lipid membrane composition in plants increases during drought, reducing the amount of water lost by the plant. It is therefore no surprise that DT cultivars have higher linolenic acid content, as it serves as a mechanism for coping with drought stress. Threonic acid is a by-product of ascorbate catabolism, whose production is induced, and regulated by stresses such as light and drought; this metabolite confers protection to the plant. It has been documented to be involved in stomatal closure during drought stress (Renault *et al.*, 2017). The levels of threonic acid found in this study were significantly higher ($P < 0.05$) in the Comm cultivars than the NComm cultivars. This is

expected as threonic acid has been reported to confer drought tolerance properties in plants, and as such it can be postulated that this metabolite contributes to the drought tolerance properties observed in the Comm cultivars.

Correlations have been documented in the literature between the umami taste found in green tea and theanine. Theanine concentrations have been reported to range between 10-50 mg/g dry weight (Vuong *et al.*, 2011) in *C. sinensis* and the higher the concentration, the more the umami taste. The average theanine levels detected in this study were higher, but insignificant, in the Comm (8.6 mg/g) compared to the NComm cultivars (8.0 mg/g) dry weight. Gallic acid was detected to have slightly higher concentrations in the Comm cultivars (1.1 mg/g) than the NComm (1.0 mg/g). In a study by Kaneko *et al.*, (2006), the umami taste intensity of green tea without any additives was evaluated and was scored at intensity of 1.5 out of five. However, following the addition of 5.4 mg/g of natural gallic acid, the intensity of the umami taste increased to a score of 2.4. This higher umami score indicates the significance of gallic acid in the taste of tea. According to literature, fresh green tea leaves contain trace amounts of gallic acid, which then increase during the manufacturing process of black tea as a result of galloyl ester hydrolysis from the catechins and theaflavin gallates. The high levels of gallic acid in some cultivars have been attributed to correspondingly high levels of gallated catechins, which result in the generation and consequent degradation of the theaflavins. It has been documented that EGCg and ECg are principal taste metabolites in tea, which are responsible for tea astringency, while caffeine is responsible for bitterness (Koech *et al.*, 2018). The gallated catechins ECg and EGCg significantly contribute to the generation of theaflavins in black tea. As such, high concentrations of ECg and EGCg may be markers for high-quality black teas. It has been reported that the ratio of di-hydroxyl flavan-3-ols to tri-hydroxyl flavan-3-ols impacted the quality of black tea; high quantities of simple catechins such as catechin, EC and ECg compared to the gallo-catechins EGC and EGCg, results in higher amounts of theaflavins (Kwach *et al.*, 2016). The concentration of EGC in the fresh tea shoots strongly correlates with theaflavins amounts, thus influencing black tea pricing (Yu *et al.*, 2008). The higher EGC found in the Comm cultivars agrees with Yu's findings.

From the results obtained from the ¹H-NMR, the Comm cultivars had a significantly ($P < 0.05$) higher total catechins content of 121.58 mg/g compared to the 114.96 mg/g dry weight of the NComm cultivars. The total catechins were calculated by adding CAT, EC, ECg, EGC, and EGCg. According to Lin *et al.*, (1996), a 200 mL cup of green tea contains 305 mg EGCg,

145 mg EGC, 70 mg ECg, 28 mg EC, and 8 mg CAT. It can, therefore, be concluded that the total catechins may have contributed to the selection of the Comm cultivars since the 1950s. The catechins are the substrates from which theaflavins and thearubigins are produced during the manufacture of black tea. Theaflavins are orangish-brownish pigments, which contribute to the briskness and brightness (Muthumani and Kumar, 2007) and astringency (Obanda *et al.*, 2001) of black tea, all of which are important traits in tea quality determination. The theaflavin digallate, is approximately 6.4 times more astringent than theaflavin, while also being 2.88 times more astringent than both theaflavin-3-monogallate and theaflavin-3'-monogallate (Obanda *et al.*, 2001). They are the predominant constituents of black tea-cream upon cooling (Roberts, 1963); it is for this reason they are deemed as an important quality index of black tea. Furthermore, 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:

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

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. These findings further support those of the current study, which 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-DAD platform, which allows for the quantitative identification of the individual theaflavins. Figure 4.29 shows superimposed black tea Comm and NComm cultivars, and from this figure, it is clearly visible that the Comm cultivars have higher total theaflavins content than the NComm cultivars. This further supports and explains why the Comm cultivars are of higher quality than the NComm cultivars.

Figures 4.7 and 4.8 show the positive and negative ionisation mode PCA, PLS-DA and S-plots, respectively. From both the PCA's and PLS-DA's, it can be seen that there is complete separation between the eight Comm and eight NComm cultivars analysed using UPLC-MS. Nine of the 21 detected metabolites were also detected by the other platforms, and these have

been discussed above. The purpose of employing the UPLC-MS platform was to detect and identify additional metabolites, which may not have been detected by the other three platforms and to investigate their influence on tea taste and quality (Figure 4.8). Caffeic acid (CA) was one of the metabolites detected by UPLC-MS. CA has been reported in literature as a potent antioxidant (Gülçin, 2006). In a study by Krishna and Surinder, (2003), it was shown that applying CA to soybean increases its yields. This could explain why the Comm cultivars had a reported higher yield as compared to the NComm cultivars. Furthermore, it has been reported that since the antioxidant enzyme genes responsible for regulating antioxidant enzyme activities in plants and the antioxidant system are interactional (Shin *et al.*, 2014), the pre-treatment with CA affects antioxidant enzyme activation in plants. When these pre-treated plants are subjected to drought stress, the antioxidant enzyme activities are enhanced, thus protecting plants from drought (Wan *et al.*, 2014). Moreover, CA pre-treatment has been reported to enhance drought tolerance in cucumber seedlings through the increased antioxidant enzyme activity; CA also leads to an increase in proline and soluble carbohydrate contents. CA has further been reported to enhance the flavour of potatoes (Thybo *et al.*, 2006), and good quality olive oils (Kiritsakis, 1998). Citric acid was the other metabolite detected. It has been documented to have a bitter taste (Van Der Klaauw and Smith, 1995). Citric acid has been shown to have beneficial effects in the roots of wheat and legumes, as it forms stable molecular complexes with metallic cations, favouring the availability and absorption of water and nutrients, and in so doing, increasing the vigour of the plant (Franco *et al.*, 1992). The role of citric acid in drought stress is well documented in several studies, which have reported that its synthesis and breakdown functions as a pH regulating mechanism in plant cells (Sadak and Orabi, 2015). Glucuronic acid was one of the other metabolites detected. Gluconic acid is an organic acid produced from the oxidation of glucose. As is the case with citric acid, glucuronic acid lowers the pH in the plant cells, enabling them to take up water and survive drought stress (Anastassiadis and Morgunov, 2007). This compound may contribute to the Comm cultivars being drought tolerant. Another important finding in this study was the detection of kaempferol 3-*O*- β -rutinoside. This flavon-3-ol glycosides has been documented as being an important tastant responsible for the velvety astringent taste noted in tea infusions. It has also been reported as an enhancer of the bitterness of caffeine in tea (Scharbert and Hofmann, 2005). Myoinositol was another metabolite also detected in the study. In a study by Rogers *et al.*, (1999), it was shown that myoinositol levels were higher in high quality, compared to the lower quality coffee beans.

Myoinositol is involved in several metabolic pathways in plants, serving as a precursor for e.g. inositol phosphates, phosphoinositides, cell wall polysaccharides through the myoinositol oxidation pathway; it is also involved in signal transduction pathways (Rogers *et al.*, 1999). Quercetin and rutin are flavonol glycosides, which were both detected by UPLC-MS. These two compounds have been reported to being responsible to the bitter taste found in buckwheat plants (Baghel *et al.*, 2012; Suzuki *et al.*, 2015). The amino acids glutamic acid and lysine were also detected. In a study by Solms and Wyler (1979), which sought out to answer the question “Is there a potato taste at all, and what are the corresponding compounds?” This study reported that potatoes have a neutral flavour, yet possess a typical taste and odour. It was found that the amino acids glutamic acid and lysine were responsible for this.

As mentioned, one of the objectives of this study was to make use of UPLC-DAD generated data of the metabolites from the 303 samples of green tea obtained from the TRI in Kenya, to develop LR models, to classify the 303 genotypes investigated as either Comm or NComm cultivars. 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, in turn, increase the success rate of field selections through conventional breeding. Tables 4.1 to 4.3 show the list of 12, 19 and ten metabolites upon which the LR models developed in this study were based. Over the course of the preceding decade, LR has gained popularity, as is documented by the trend in peer-reviewed science journals. 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, including LR (Zeitouni and Chelghoum, 2001). In this study, we show that LR is capable of serving as a strong analytical tool for classifying cultivars as either Comm or NComm, based on the CAT/EC ratio, as well as the CAF/EC ratio from the UPLC-DAD catechins, and that these ratios can be used to predict whether or not new field selections are likely to be commercialised. The study also tests other catechin combinations documented in literature to function as predictors for high and low quality teas (Table 4.6). Multiple regression analysis functions to classify significant variables as genotype class predictors, which elucidate the dependent variable variance. Multiple regression, an extension of simple linear regression, is used when predicting the value of a dependent variable based on the value of two or more independent variables (Statistics, 2013). By contrast, decision

trees identify those variables which would most differentiate Comm and NComm cultivars. Figure 4.10 shows the nominal LR results obtained using all 12 GC-MS variables. Based on these results, 90% (44/49) of the Comm cultivars were correctly classified. Next, the seven statistically significant variables were used to develop another model (Figure 4.11) and this model was found to correctly classify 88% (43/49) of the Comm cultivars. The ChiSquare values in Figure 4.11 show Psicose and Acetoacetic acid to be the most statistically significant variables, and as such the Acetoacetic acid/Psicose ratio was used as a variable to generate a new LR model (Figure 4.12). This model correctly classified 67% (33/49) of the Comm cultivars. This means the GC-MS variables in this study were not the best suited to separate the Comm vs NComm cultivars, prompting the development of LR models on the ¹H-NMR results. The spatial decision tree was developed on the GC-MS data confirmed that Acetoacetic acid and Psicose were indeed important variables in the GC-MS dataset. As with the GC-MS data, the first model developed on the ¹H-NMR data was based on all 19 detected variables (Figure 4.14). From these results, it can be seen that this model correctly classifies all cultivars into the Comm and NComm classes, with the confusion matrix displaying zero misclassifications. However, no further LR models based on the significance of each variable could be developed because from the obtained p-values, all variables except for Quinic acid were significantly different between the Comm and the NComm cultivars. It has been extensively reported in literature that amino acids are good indicators for determining quality of tea liquor. This led to the development of model shown in Figure 4.15 based solely on the detected amino acids. From this model, only 50% of the Comm cultivars were correctly classified. This means that the detected amino acids, on their own, may not be the best variables to separate the cultivars into the two classes. The LR model based on CAF and the five catechins was developed and this model showed 91% of the Comm cultivars were correctly classified, which is identical to the UPLC-DAD model results based on the same variables (Figure 4.25). This shows that CAF and the five catechins are markers for separating the Comm and NComm cultivars. From the ChiSquare values of the model (Figure 4.16), CAF and CAT were found to be the most significant variables, resulting in the CAF/CAT ratio being used as a variable. The LR model based on this ratio correctly classified 43% of the Comm cultivars, showing that this ratio would not work well in separating the Comm from the NComm cultivars. The decision tree developed in Figure 4.17 identified CAF and EGC as important variables, which prompted the development of a LR model based on CAF/EGC. The results of this model show that 23% of the Comm cultivars

were correctly classified, making this ratio unsuitable for separating the two classes of cultivars. Figures 4.19 and 4.20 were developed based on the CAT/EC and CAF/EC ratios, respectively, similar to what was done with the UPLC-DAD results. Both these models correctly classified 0% of the Comm cultivars, failing dismally, by comparison to those developed on the UPLC-DAD data, which correctly classified 79% and 57% of the Comm cultivars, respectively. This could mean that UPLC-DAD detection of the CAT, CAF, and EC metabolites was better than ¹H-NMR and as such the correct quantities were detected, making the ratios suitable predictors in the UPLC-DAD based models than in the ¹H-NMR based models.

Lastly, LR models were developed on the UPLC-DAD metabolites, starting with Figure 4.21, which shows the nominal LR results obtained using all ten UPLC/DAD variables. From these results, it can be seen that this model correctly classifies all 303 genotypes into the Comm and NComm classes, with the confusion matrix displaying zero misclassifications. The theaflavin variables were then separated from the catechin variables, and a LR model was developed based on the four theaflavins (Figure 4.22) and the most significant theaflavin, TF4 (Figure 4.23). From these models, 96% of the Comm cultivars were correctly classified, with only two misclassifications. This shows that theaflavins are as efficient at separating the two groups as all ten metabolites, corroborating literature findings that theaflavins are indicators of high tea quality markers (Wang and Ruan, 2009). Figure 4.24 shows the LR model developed on CAF and the five catechins. The confusion matrix shows five of the 56 Comm cultivars were misclassified and 80% were correctly classified. Considering that it is a very cumbersome process to manufacture black tea to be able 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 catechins would be less cumbersome to ascertain the likelihood that a new field selection will be similar to the Comm cultivars. This saves them up to four years, as well as labour and resources of cultivating the tea bushes for five years only to learn it is a low yield, drought susceptible and low quality field selection, and will not be commercialised.

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). In the past, countries such as Kenya, India, and Sri Lanka, which are high black tea producing breed cultivars rich in theaflavins. Efforts have been made to combine these two qualities into an F₁ 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). From the decision tree given in Figure 4.26, it is evident that the predictors CAT and EC are responsible for classifying the 303 genotypes into the two classes. This implies that tea breeders will now be able to analyse the CAT and EC content of the seedling green leaves and follow the decision tree branches, to ascertain whether a new cultivar is likely to be Comm based on their CAT and EC content. The use of a decision tree for spatial classification is based on the simplicity and effectiveness of this approach. Since Figure 4.26 shows that CAT and EC are the predictors, this means that the ratio of CAT/EC can serve as a predictor. The CAT/EC ratio LR plot (Figure 4.27) shows that to classify each of the 303 genotypes as Comm cultivar, a CAT/EC ratio of 1.5 and above is required, while any ratio below 1.5 indicates the likelihood the cultivar will be NComm. The confusion matrix of this model shows that 12 of the 56 Comm cultivars were misclassified, meaning the model correctly classified 79% of the genotypes. However, as mentioned earlier, the identification and accurate quantification of CAT may be problematic due to its position on the chromatogram, as well as its peak height (Figure 4.28), warranting an improvement of the chromatography conditions in the ISO14502-2 (2005) method. This prompted the development of a decision tree, and LR model which excluded CAT (Figure 4.30 and 4.31 respectively). The decision tree identified CAF and EC as the metabolite predictors, and the LR model developed based on the CAF/EC ratio proved capable of correctly classifying 57% of the genotypes, 22% lower than the CAT/EC ratio based model. This indicates CAT is an important metabolite predictor. This finding is, however, contradictory to the findings of Wright *et al.*, (2000), 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 eight high, and eight low quality clones. The results obtained in the Wright study confirmed those obtained

by Robertson, (1983), 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). The Wright study concluded that high and low quality cultivars were distinguishable by considering CAT+EC+ECg, (B-ring di-hydroxy or simple catechins) prompting the development of a LR model based on CAT+EC+ECg (Figure 4.32) in the present study. The results of this show 100% misclassification of the Comm cultivars as NComm cultivars. In another study by (Ellis and Nyirenda, 1995) on simple (CAT, EC and ECg) and complex catechins (EGC and EGCg) (tri-hydroxy), they documented that the higher the ratio of simple: complex catechins, the higher the amount of theaflavins produced, which in turn 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 developing a LR model, and there was 100% misclassification of the Comm cultivars (Figure 4.33). 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 coeluting with other compounds, while the coeluting 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. The larger sample size lends more credibility to our results. Future work 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. In conclusion, the results of this study show that the UPLC-DAD is the better suited platform for analysing samples in order to generate LR models for prediction purposes. 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. Further 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%.

From the results presented throughout this chapter, it can thus be concluded that objective 1, to use data generated through untargeted GC-MS, and semi-targeted ¹H-NMR metabolomics platforms to identify metabolites, which were expressed differently in the Comm and NComm cultivars was successfully achieved. Objective 2, which sought to make use of UPLC-DAD generated targeted metabolomics data to develop several LR models and decision trees, to classify the 303 genotypes as either Comm or NComm cultivars was successfully achieved. Last, objective 3, which sought to use untargeted LC-MS data to identify any additional metabolites not detected by platforms mentioned in objectives 1 and 2, to distinguish between the Comm and NComm cultivars was also successfully completed. The null hypothesis, which states that there will be no statistically significant difference between the metabolite profiles and metabolite concentrations detected by all the metabolomics platforms employed between the Comm and NComm cultivars, at the 95% confidence interval can therefore be rejected

4.7 REFERENCES

- Ackermann, J., Fischer, M. & Amado, R. (1992). Changes in sugars, acids, and amino acids during ripening and storage of apples (cv. Glockenapfel). *Journal of Agricultural and Food Chemistry* 40(7): 1131-1134.
- Adkins, N. L., Hall, J. A. & Georgel, P. T. (2007). The use of quantitative agarose gel electrophoresis for rapid analysis of the integrity of protein–DNA complexes. *Journal of Biochemical and Biophysical methods* 70(5): 721-726.
- Anastassiadis, S., & Morgunov, I. G. (2007). Gluconic acid production. *Recent patents on biotechnology* 1(2): 167-180.
- Anon. (1990). Seed garden (barie). Annual Report., 25: Tea Research Foundation of Kenya, Tea Board of Kenya.
- Archana, I. & Vijayalakshmi, K. (2018). Antioxidant potential of phloroglucinol; an in-vitro approach. *International Journal of Pharmaceutical Sciences and Research* 9(7): 2947-2951.
- Baeza, G., Sarriá, B., Bravo, L. & Mateos, R. (2016). Exhaustive Qualitative LC-DAD-MS n Analysis of Arabica Green Coffee Beans: Cinnamoyl-glycosides and Cinnamoylshikimic Acids as New Polyphenols in Green Coffee. *Journal of Agricultural and Food Chemistry* 64(51): 9663-9674.
- Baghel, S. S., Shrivastava, N., Baghel, R. S., Agrawal, P., & Rajput, S. (2012). A review of quercetin: antioxidant and anticancer properties. *World J Pharm Pharmaceutical Sci* 1(1): 146-60.
- Bandurski, R. S. & Schulze, A. (1977). Concentration of indole-3-acetic acid and its derivatives in plants. *Plant Physiology* 60(2): 211-213.
- Blainski, A., Antonelli-Ushirobira, T. M., Godoy, G., Leite-Mello, E. V., & Mello, J. C. (2017). Pharmacognostic evaluation, and development and validation of a HPLC-DAD technique for galocatechin and epigallocatechin in rhizomes from *Limonium brasiliense*. *Revista Brasileira de Farmacognosia* 27(2): 162-169.
- Borsani, J., Budde, C. O., Porrini, L., Lauxmann, M. A., Lombardo, V. A., Murray, R., Andreo, C. S., Drincovich, M. F. & Lara, M. V. (2009). Carbon metabolism of peach fruit after harvest: changes in enzymes involved in organic acid and sugar level modifications. *Journal of Experimental Botany* 60(6): 1823-1837.

- Brühl, L., Matthäus, B., Scheipers, A. & Hofmann, T. (2008). Bitter off-taste in stored cold-pressed linseed oil obtained from different varieties. *European Journal of Lipid Science and Technology* 110(7): 625-631.
- Buffo, R. A., & Cardelli-Freire, C. (2004). Coffee flavour: an overview. *Flavour and Fragrance Journal* 19(2): 99-104.
- Cabrera, C., Artacho, R. & Giménez, R. (2006). Beneficial effects of green tea—a review. *Journal of the American College of Nutrition* 25(2): 79-99.
- Chang, K. (2015). World tea production and trade Current and future development, Food and Agricultural Organization of the United Nations.
- Chattopadhyay, S., Raychaudhuri, U., & Chakraborty, R. (2014). Artificial sweeteners—a review. *Journal of Food Science and Technology* 51(4): 611-621.
- Chaturvedula, V. S. P. & Prakash, I. (2011). The aroma, taste, color and bioactive constituents of tea. *Journal of Medicinal Plants Research* 5(11): 2110-2124.
- Cheserek, B. C., Elbehri, A. & Bore, J. (2015). Analysis of links between climate variables and tea production in the recent past in Kenya. *Donnish Journal of Research in Environmental Studies* 2(2): 5-17.
- Chin, J. M., Merves, M. L., Goldberger, B. A., Sampson-Cone, A., & Cone, E. J. (2008). Caffeine content of brewed teas. *Journal of Analytical Toxicology* 32(8): 702-704.
- Chugh, K. (2013). Measuring phenotypic and genetic variances and narrow sense heritability in three populations of annual ryegrass (*Lolium multiflorum Lam.*).
- Clifford, M. N. (2000). Chlorogenic acids and other cinnamates—nature, occurrence, dietary burden, absorption and metabolism. *Journal of Science and Food Agriculture* 80(7): 1033-1043.
- Cortell, J. M., Sivertsen, H. K., Kennedy, J. A., & Heymann, H. (2008). Influence of vine vigor on Pinot noir fruit composition, wine chemical analysis, and wine sensory attributes. *American Journal of Enology and Viticulture* 59(1): 1-10.
- Ding, Z., Kuhr, S., & Engelhardt, U. H. (1992). Influence of catechins and theaflavins on the astringent taste of black tea brews. *Zeitschrift für Lebensmittel-Untersuchung und Forschung*, 195(2), 108-111.
- Dumas, M.-E., Maibaum, E. C., Teague, C., Ueshima, H., Zhou, B., Lindon, J. C., Nicholson, J. K., Stamler, J., Elliott, P. & Chan, Q. (2006). Assessment of analytical reproducibility of ¹H NMR spectroscopy based metabonomics for large-scale

- epidemiological research: the INTERMAP Study. *Analytical Chemistry* 78(7): 2199-2208.
- Dunn, W. B., Goodacre, R., Neyses, L. & Mamas, M. (2011). Integration of metabolomics in heart disease and diabetes research: current achievements and future outlook. *Bioanalysis* 3(19): 2205-2222.
- Dutta, R., Stein, A. & Bhagat, R. (2011). Integrating satellite images and spectroscopy to measuring green and black tea quality. *Food Chemistry* 127(2): 866-874.
- Ebbels, T. M., Lindon, J. C. & Coen, M. (2011). Processing and modeling of nuclear magnetic resonance (NMR) metabolic profiles. *Springer* 365-388.
- Elbehri, A., Azapagic, A., Cheserek, B., Raes, D., Kiprono, P. & Ambasa, C. (2015). Kenya's tea sector under climate change: an impact assessment and formulation of a climate smart strategy. *FAO report. FAO, Rome, Italy.*
- Ellinger, J. J., Chylla, R. A., Ulrich, E. L. & Markley, J. L. (2013). Databases and software for NMR-based metabolomics. *Current Metabolomics* 1(1): 28-40.
- Ellis, R. & Nyirenda, H. (1995). A successful plant improvement programme on tea (*Camellia sinensis*). *Experimental Agriculture* 31(3): 307-323.
- Fall, R., & Benson, A. A. (1996). Leaf methanol—the simplest natural product from plants. *Trends in Plant Science* 1(9): 296-301.
- FAO (2018). 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)*, 1-13 (Ed K. Chang). Rome, Italy.
- FAO (Food and Agricultural Organisation). (2015). Retrieved from <http://www.fao.org>.
- Fehsenfeld, F., Calvert, J., Fall, R., Goldan, P., Guenther, A. B., Hewitt, C. N., ... & Zimmerman, P. (1992). Emissions of volatile organic compounds from vegetation and the implications for atmospheric chemistry. *Global Biogeochemical Cycles* 6(4): 389-430.
- Franco AC, Ball E., & Luttge U. (1992). Differential effects of drought and light levels on accumulation of citric acids during CAM in *Clusia*. *Plant Cell and Environment* 15(1): 821-829.
- Gramza, A., Khokhar, S., Yoko, S., Gliszczynska-Swiglo, A., Hes, M., & Korczak, J. (2006). Antioxidant activity of tea extracts in lipids and correlation with polyphenol content. *European Journal of Lipid Science and Technology* 108(4): 351-362.

- Green, M. (1971). An evaluation of some criteria used in selecting large-yielding tea clones. *The Journal of Agricultural Science* 76(1): 143-156.
- Group, C. P. B., Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., Liu, J. Q., Chen, Z. D., Zhou, S. L. & Chen, S. L. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences* 108(49): 19641-19646.
- Gülçin, İ. (2006). Antioxidant activity of caffeic acid (3, 4-dihydroxycinnamic acid). *Toxicology*, 217(2-3), 213-220.
- Habibi, Y., Mahrouz, M., Marais, M. F., & Vignon, M. R. (2004). An arabinogalactan from the skin of *Opuntia ficus-indica* prickly pear fruits. *Carbohydrate Research* 339(6): 1201-1205.
- Hagel, J. M. & Facchini, P. J. (2008). Plant metabolomics: analytical platforms and integration with functional genomics. *Phytochemistry Reviews* 7(3): 479-497.
- Hamanishi, E. T., Barchet, G. L., Dauwe, R., Mansfield, S. D. & Campbell, M. M. (2015). Poplar trees reconfigure the transcriptome and metabolome in response to drought in a genotype-and time-of-day-dependent manner. *BMC Genomics* 16(1): 329.
- Hilton P. J. & Ellis, R. T. (1972). Estimation of the Market Value of Central African Tea by Theaflavin Analysis. *Journal of the Science of Food and Agriculture* 23: 227-232.
- Hilton, P. J., & Palmer-Jones, R. (1973). Relationship between the flavanol composition of fresh tea shoots and the theaflavin content of manufactured tea. *Journal of the Science of Food and Agriculture* 24(7): 813-818.
- Hinreiner, E., Filipello, F., Berg, H. W., & Webb, A. D. (1955). Evaluation of thresholds and minimum difference concentrations for various constituents of 4 wines. Detectable differences in wine. *Food Technology* 9(10): 489-490.
- Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician* 52(2): 181-184.
- Hoekstra, F. A., Golovina, E. A. & Buitink, J. (2001). Mechanisms of plant desiccation tolerance. *Trends in Plant Science* 6(9): 431-438.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 65-70.
- Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., ... & Oda, Y. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* 45(7): 703-714.

- Iiyama, S., Ezaki, S., Toko, K., Matsuno, T., & Yamafuji, K. (1995). Study of astringency and pungency with multichannel taste sensor made of lipid membranes. *Sensors and Actuators B: Chemical* 24(1-3): 75-79.
- Imamura, K., Tsuchama, Y., Tsunekawa, H., Okamura, K., Okamoto, R. & Harada, K. (1995). Removal of chlorogenic acid from coffee extracts. *Jpn. Kokai JP 07322823*.
- Inglett, G. E. (2012). Intense sweetness of natural origin. *Flavor of Foods and Beverages: Chemistry and Technology* 97.
- International Tea Committee (ITC). Annual Bulletin of Statistics, 2019.
- Jaiswal, R., Sovdat, T., Vivian, F. & Kuhnert, N. (2010). Profiling and characterization by LC-MS n of the chlorogenic acids and hydroxycinnamoylshikimate esters in mate (*Ilex paraguariensis*). *Journal of Agricultural and Food Chemistry* 58(9): 5471-5484.
- JMP®, V. SAS Institute Inc., Cary, NC, 1989–2007.
- Jones, P. R., Gawel, R., Francis, I. L., & Waters, E. J. (2008). The influence of interactions between major white wine components on the aroma, flavour and texture of model white wine. *Food Quality and Preference* 19(6): 596-607.
- Kamunya, S. & Wachira, F. (2006). Two new clones (TRFK 371/3 and TRFK 430/90) released for commercial use. *Tea* 27(1/2): 3-14.
- Kamunya, S., Wachira, F., Pathak, R., Korir, R., Sharma, V., Kumar, R., Bhardwaj, P., Chalo, R., Ahuja, P. & Sharma, R. (2010). Genomic mapping and testing for quantitative trait loci in tea (*Camellia sinensis* (L.) O. Kuntze). *Tree Genetics & Genomes* 6(6): 915-929.
- Kaneko, S., Kumazawa, K., Masuda, H., Henze, A., & Hofmann, T. (2006). Molecular and sensory studies on the umami taste of Japanese green tea. *Journal of Agricultural and Food Chemistry* 54(7): 2688-2694.
- Kang, J., Choi, M.-Y., Kang, S., Kwon, H. N., Wen, H., Lee, C. H., Park, M., Wiklund, S., Kim, H. J. & Kwon, S. W. (2008). Application of a ¹H nuclear magnetic resonance (NMR) metabolomics approach combined with orthogonal projections to latent structure-discriminant analysis as an efficient tool for discriminating between Korean and Chinese herbal medicines. *Journal of Agricultural and Food Chemistry* 56(24): 11589-11595.
- Kaplan, F. & Guy, C. L. (2004). β -Amylase induction and the protective role of maltose during temperature shock. *Plant Physiology* 135(3): 1674-1684.

- Karakaya, S., & EL, S. N. (1999). Quercetin, luteolin, apigenin and kaempferol contents of some foods. *Food Chemistry* 66(3): 289-292.
- Karori, S., Wachira, F., Ngure, R. & Mireji, P. (2014). Polyphenolic composition and antioxidant activity of Kenyan tea cultivars. *Journal of Pharmacognosy and Phytochemistry* 3(4): 105-116.
- Kenya National Bureau of Statistics (2018). Kenya facts and figures 2018. Kenya.
- Kerchev, P. I., Fenton, B., Foyer, C. H. & Hancock, R. D. (2012). Plant responses to insect herbivory: interactions between photosynthesis, reactive oxygen species and hormonal signalling pathways. *Plant, Cell & Environment* 35(2): 441-453.
- Khan, N. & Mukhtar, H. (2007). Tea polyphenols for health promotion. *Life Sciences* 81(7): 519-533.
- Kiritsakis, A. K. (1998). Flavor components of olive oil—A review. *Journal of the American Oil Chemists' Society* 75(6): 673-681.
- Kobayashi-Hattori, K., Mogi, A., Matsumoto, Y. & Takita, T. (2005). Effect of caffeine on the body fat and lipid metabolism of rats fed on a high-fat diet. *Bioscience, Biotechnology, and Biochemistry* 69(11): 2219-2223.
- Koech, R. K., Malebe, P. M., Nyarukowa, C., Mose, R., Kamunya, S. M. & Apostolides, Z. (2018). Identification of novel QTL for black tea quality traits and drought tolerance in tea plants (*Camellia sinensis*). *Tree Genetics & Genomes* 14(1): 9.
- Kowalsick, A., Kfoury, N., Robbat Jr, A., Ahmed, S., Orians, C., Griffin, T., Cash, S. B. & Stepp, J. R. (2014). Metabolite profiling of *Camellia sinensis* by automated sequential, multidimensional gas chromatography/mass spectrometry reveals strong monsoon effects on tea constituents. *Journal of Chromatography A* 1370: 230-239.
- Kreft, I., Fabjan, N., & Yasumoto, K. (2006). Rutin content in buckwheat (*Fagopyrum esculentum* Moench) food materials and products. *Food Chemistry* 98(3): 508-512.
- Krishna, S., & Surinder, K. (2003). Effect of some phenolic compounds and light intensity on nodulation, chlorophyll content, proteins, total free amino acid content and yield of soybean (*Glycine max* (L.) Merrill). *Indian Agriculture*, 47(1/2), 79-84.
- Kumar, V., Rani, A., Goyal, L., Pratap, D., Billore, S. D., & Chauhan, G. S. (2011). Evaluation of vegetable-type soybean for sucrose, taste-related amino acids, and isoflavones contents. *International Journal of Food Properties* 14(5): 1142-1151.
- Kwach, B. O., Owuor, P. O., Kamau, D. M., Msomba, S. W. & Uwimana, M. A. (2006). Variations in the precursors of plain black tea quality parameters due to location of

- production and nitrogen fertilizer rates in Eastern African clonal tea leaves. *Experimental Agriculture* 52(2): 266-278.
- Lavarack, B., Griffin, G. & Rodman, D. (2002). The acid hydrolysis of sugarcane bagasse hemicellulose to produce xylose, arabinose, glucose and other products. *Biomass and Bioenergy* 23(5): 367-380.
- Le Gall, G., Colquhoun, I. J. & Defernez, M. (2004). Metabolite profiling using ¹H NMR spectroscopy for quality assessment of green tea, *Camellia sinensis* (L.). *Journal of Agricultural and Food Chemistry* 52(4): 692-700.
- Ley, J. P. (2008). Masking bitter taste by molecules. *Chemosensory Perception* 1(1): 58-77.
- Lin, Y. L., Juan, I. M., Chen, Y. L., Liang, Y. C., & Lin, J. K. (1996). Composition of polyphenols in fresh tea leaves and associations of their oxygen-radical-absorbing capacity with antiproliferative actions in fibroblast cells. *Journal of Agricultural and Food Chemistry* 44(6): 1387-1394.
- Lorist, M. M., & Tops, M. (2003). Caffeine, fatigue, and cognition. *Brain and Cognition* 53(1): 82-94.
- Ludwig, C. & Viant, M. R. (2010). Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques* 21(1): 22-32.
- Ma, B., Chen, J., Zheng, H., Fang, T., Ogutu, C., Li, S., Han, Y. & Wu, B. (2015). Comparative assessment of sugar and malic acid composition in cultivated and wild apples. *Food Chemistry* 172: 86-91.
- MacDonald, R. C., & Fall, R. (1993). Detection of substantial emissions of methanol from plants to the atmosphere. *Atmospheric Environment. Part A. General Topics* 27(11): 1709-1713.
- Marks, S. C., Mullen, W., & Crozier, A. (2007). Flavonoid and chlorogenic acid profiles of English cider apples. *Journal of the Science of Food and Agriculture* 87(4): 719-728.
- Matanjun, P., Mohamed, S., Mustapha, N. M., Muhammad, K., & Ming, C. H. (2008). Antioxidant activities and phenolics content of eight species of seaweeds from north Borneo. *Journal of Applied Phycology* 20(4): 367.
- Mazzafera, P., & Silvarolla, M. B. (2010). Caffeine content variation in single green Arabica coffee seeds. *Seed Science Research* 20(3): 163-167.

- Melgarejo, P., Salazar, D. M., & Artes, F. (2000). Organic acids and sugars composition of harvested pomegranate fruits. *European Food Research and Technology* 211(3): 185-190.
- Min, L. U., Huaming, A. N., & Daoping, W. A. N. G. (2017). Characterization of Amino Acid Composition in Fruits of Three *Rosa roxburghii* Genotypes. *Horticultural Plant Journal* 3(6): 232-236.
- Mu, W., Zhang, W., Feng, Y., Jiang, B. & Zhou, L. (2019). Recent advances on applications and biotechnological production of D-psicose. *Applied Microbiology and Biotechnology* 94(6): 1461-1467.
- Muthumani, T., & Kumar, R. S. S. (2007). Influence of fermentation time on the development of compounds responsible for quality in black tea. *Food Chemistry* 101(1): 98–102.
- Naveed, M., Brown, L. K., Raffan, A. C., George, T. S., Bengough, A. G., Roose, T. & Hallett, P. D. (2017). Plant exudates may stabilize or weaken soil depending on species, origin and time. *European Journal of Soil Science* 68(6): 806-816.
- Nemecek-Marshall, M., MacDonald, R. C., Franzen, J. J., Wojciechowski, C. L., & Fall, R. (1995). Methanol emission from leaves (enzymatic detection of gas-phase methanol and relation of methanol fluxes to stomatal conductance and leaf development). *Plant Physiology* 108(4): 1359-1368.
- Nitin Seetohul, L., Islam, M., O'Hare, W. T. & Ali, Z. (2006). Discrimination of teas based on total luminescence spectroscopy and pattern recognition. *Journal of the Science of Food and Agriculture* 86(13): 2092-2098.
- Noble, A. C., & Bursick, G. F. (1984). The contribution of glycerol to perceived viscosity and sweetness in white wine. *American Journal of Enology and Viticulture* 35(2): 110-112.
- Nyarukowa C. T., Koech K. R., Loots T., Hageman J. & Apostolides Z. (2018). Prioritising the replanting schedule of seedling tea fields on tea estates for drought susceptibility measured by the SWAPDT method in the absence of historical in-filling records. *Journal of Agricultural Science* 10(7): 26-34.
- Nyarukowa, C., Koech, R., Loots, T. & Apostolides, Z. (2016). SWAPDT: A method for Short-time Withering Assessment of Probability for Drought Tolerance in *Camellia sinensis* validated by targeted metabolomics. *Journal of Plant Physiology* 198: 39-48.

- Nyirenda, H. (1991). Use of growth measurements and foliar nutrient content as criteria for clonal selection in tea (*Camellia sinensis*). *Experimental Agriculture* 27(1): 47-52.
- Obanda, M., Owuor, P. O. & Mang'oka, R. (2001). Changes in the chemical and sensory quality parameters of black tea due to variations of fermentation time and temperature. *Food Chemistry* 75(4): 395-404.
- Obanda, M., Owuor, P. O., & Taylor, S. J. (1997). Flavanol composition and caffeine content of green leaf as quality potential indicators of Kenyan black teas. *Journal of the Science of Food and Agriculture* 74(2); 209-215.
- O'Neill, M., Albersheim, P., & Darvill, A. (1990). The pectic polysaccharides of primary cell walls. In *Methods in plant biochemistry* (Vol. 2, pp. 415-441). Academic Press).
- Oshima, H., Kimura, I. & Izumori, K. (2006). Psicose contents in various food products and its origin. *Food Science and Technology Research* 12(2): 137-143.
- Owuor, P. O. & Obanda, M. (2007). The use of green tea (*Camellia sinensis* (L)) leaf flavan-3-ols composition in predicting plain black tea quality potential. *Food Chemistry* 100(3): 873–884.
- Owuor, P. O., Wachira, F. N. & Ng'etich, W. K. (2010). Influence of region of production on relative clonal plain tea quality parameters in Kenya. *Food Chemistry* 119(3): 1168-1174.
- Peng, C.-Y. J., Lee, K. L. & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research* 96(1): 3-14.
- Pongsuwan, W., Fukusaki, E., Bamba, T., Yonetani, T., Yamahara, T. & Kobayashi, A. (2007). Prediction of Japanese green tea ranking by gas chromatography/mass spectrometry-based hydrophilic metabolite fingerprinting. *Journal of Agricultural and Food Chemistry* 55(2): 231-236.
- Qin, Z., Pang, X., Chen, D., Cheng, H., Hu, X. & Wu, J. (2013). Evaluation of Chinese tea by the electronic nose and gas chromatography–mass spectrometry: Correlation with sensory properties and classification according to grade level. *Food Research International* 53(2): 864-874.
- Ravn, H., Pedersen, M. F., Borum, J., Andary, C., Anthoni, U., Christophersen, C., & Nielsen, P. H. (1994). Seasonal variation and distribution of two phenolic compounds, rosmarinic acid and caffeic acid, in leaves and roots-rhizomes of eelgrass (*Zostera marina* L.). *Ophelia* 40(1): 51-61.

- Rawat, R., Gulati, A., Babu, G. K., Acharya, R., Kaul, V. K. & Singh, B. (2007). Characterization of volatile components of Kangra orthodox black tea by gas chromatography-mass spectrometry. *Food Chemistry* 105(1): 229-235.
- Renault, H., Alber, A., Horst, N. A., Lopes, A. B., Fich, E. A., Kriegshauser, L. & Pineau, A. (2017). A phenol-enriched cuticle is ancestral to lignin evolution in land plants. *Nature Communications* 8: 14713.
- Roberts, E. A. H. (1963). The phenolic substances of manufactured tea. X.—the creaming down of tea liquors. *Journal of the Science of Food and Agriculture* 14(10): 700–705.
- Robertson, A. (1983). Effects of physical and chemical conditions on the in vitro oxidation of tea leaf catechins. *Phytochemistry* 22(4): 889-896.
- Rodrigues, C. I., Marta, L., Maia, R., Miranda, M., Ribeirinho, M., & Máguas, C. (2007). Application of solid-phase extraction to brewed coffee caffeine and organic acid determination by UV/HPLC. *Journal of Food Composition and Analysis* 20(5): 440-448.
- Rogers, W. J., Michaux, S., Bastin, M., & Bucheli, P. (1999). Changes to the content of sugars, sugar alcohols, myo-inositol, carboxylic acids and inorganic anions in developing grains from different varieties of Robusta (*Coffea canephora*) and Arabica (*C. arabica*) coffees. *Plant Science* 149(2): 115-123.
- Roser, D. J., Melick, D. R., Ling, H. U., & Seppelt, R. D. (1992). Polyol and sugar content of terrestrial plants from continental Antarctica. *Antarctic Science* 4(4): 413-420.
- Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* 10(3): 361-374.
- Sadak, M. S., & Orabi, S. A. (2015). Improving thermo tolerance of wheat plant by foliar application of citric acid or oxalic acid. *International Journal of Chemistry and Technology Research* 8(1): 111-123.
- Sanderson, G. W., & Graham, H. N. (1973). Formation of black tea aroma. *Journal of Agricultural and Food Chemistry* 21(4): 576-585.
- Sanhueza, E. & Andreae, M. O. (1991). Emission of formic and acetic acids from tropical savanna soils. *Geophysics Research Letter* 18(9): 1707-1710.
- Scharbert, S., & Hofmann, T. (2005). Molecular definition of black tea taste by means of quantitative studies, taste reconstitution, and omission experiments. *Journal of Agricultural and Food Chemistry* 53(13): 5377–5384.

- Schauer, N. & Fernie, A. R. (2006). Plant metabolomics: towards biological function and mechanism. *Trends in Plant Science* 11(10): 508-516.
- Schuh, C. & Schieberle, P. (2006). Characterization of the key aroma compounds in the beverage prepared from Darjeeling black tea: quantitative differences between tea leaves and infusion. *Journal of Agricultural and Food Chemistry* 54(3): 916-924.
- Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental Science and Technology* 48(4): 2097-2098.
- Shanmugarajah, V., Kulasegeram, S., & Senanayake, Y. (1991). Nursery plant attributes as criteria for selection of new tea clones.
- Shin, S. H., Seo, S. G., Min, S., Yang, H., Lee, E., Son, J. E., ... & Cheng, J. X. (2014). Caffeic acid phenethyl ester, a major component of propolis, suppresses high fat diet-induced obesity through inhibiting adipogenesis at the mitotic clonal expansion stage. *Journal of Agricultural and Food Chemistry* 62(19): 4306-4312.
- Siegel, S. & Castellan, N. J. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-hill New York.
- Sokolowsky, M., & Fischer, U. (2012). Evaluation of bitterness in white wine applying descriptive analysis, time-intensity analysis, and temporal dominance of sensations analysis. *Analytica Chimica Acta* 7(32): 46-52.
- Solms, J., & Wyler, R. (1979). Taste components of potatoes.
- Statistics, L. (2013). Multiple regression analysis using SPSS statistics. *Laerd Research Ltd*.
- Stephan, A. & Steinhart, H. (2000). Bitter taste of unsaturated free fatty acids in emulsions: contribution to the off-flavour of soybean lecithins. *European Food Research and Technology* 212(1): 17-25.
- Sun, H. W., Qiao, F. X., & Liu, G. Y. (2006). Characteristic of theophylline imprinted monolithic column and its application for determination of xanthine derivatives caffeine and theophylline in green tea. *Journal of Chromatography A* 1134(1-2): 194-200.
- Suzuki, T., Morishita, T., Kim, S. J., Park, S. U., Woo, S. H., Noda, T., & Takigawa, S. (2015). Physiological roles of rutin in the buckwheat plant. *Japan Agricultural Research Quarterly* 49(1): 37-43.

- Szejtli, J., & Szente, L. (2005). Elimination of bitter, disgusting tastes of drugs and foods by cyclodextrins. *European Journal of Pharmaceutics and Biopharmaceutics* 61(3): 115-125.
- Takemoto, M., & Takemoto, H. (2018). Synthesis of theaflavins and their functions. *Molecules* 23(4): 918.
- Tan, F., Tan, C., Zhao, A., & Li, M. (2011). Simultaneous determination of free amino acid content in tea infusions by using high-performance liquid chromatography with fluorescence detection coupled with alternating penalty trilinear decomposition algorithm. *Journal of Agricultural and Food Chemistry* 59(20): 10839-10847.
- Taylor, N. W. (1928). A physico-chemical theory of sweet and bitter taste excitation based on the properties of the plasma membrane. *Protoplasma* 4(1): 1-17.
- Theodoridis, G. A., Gika, H. G., Want, E. J. & Wilson, I. D. (2012). Liquid chromatography–mass spectrometry based global metabolite profiling: a review. *Analytica Chimica Acta* 711: 7-16.
- Thomas, E. H. & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education* 45(3): 251-269.
- Thybo, A. K., Christiansen, J., Kaack, K., & Petersen, M. A. (2006). Effect of cultivars, wound healing and storage on sensory quality and chemical components in pre-peeled potatoes. *LWT-Food Science and Technology* 39(2): 166-176.
- Toumi, I., Gargouri, M., Nouairi, I., Moschou, P. N., Salem-Fnayou, A. B. & Ghorbel, A. (2008). Water stress induced changes in the leaf lipid composition of four grapevine genotypes with different drought tolerance. *Biologia Plantarum* 52(1): 161-164.
- Tripathi, A., Parmar, D., Patel, U., Patel, G., Daslaniya, D., & Bhimani, B. (2011). Taste masking: a novel approach for bitter and obnoxious drugs. *JPSBR* 1(3): 36-142.
- Unno, K., Tanida, N., Ishii, N., Yamamoto, H., Iguchi, K., Hoshino, M., ... & Yamada, H. (2013). Anti-stress effect of theanine on students during pharmacy practice: positive correlation among salivary α -amylase activity, trait anxiety and subjective stress. *Pharmacology Biochemistry and Behavior* 111: 128-135.
- Urano, K., Maruyama, K., Ogata, Y., Morishita, Y., Takeda, M., Sakurai, N., Suzuki, H., Saito, K., Shibata, D. & Kobayashi, M. (2009). Characterization of the ABA-regulated global responses to dehydration in Arabidopsis by metabolomics. *The Plant Journal* 57(6): 1065-1078.

- Valpuesta, V. & Botella, M. A. (2004). Biosynthesis of L-ascorbic acid in plants: new pathways for an old antioxidant. *Trends Plant Science* 9(12): 573-577.
- van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7(1): 142.
- Van Der Klaauw, N. J., & Smith, D. V. (1995). Taste quality profiles for fifteen organic and inorganic salts. *Physiology & Behavior* 58(2): 295-306.
- Vuong, Q. V., Bowyer, M. C., & Roach, P. D. (2011). L-Theanine: properties, synthesis and isolation from tea. *Journal of the Science of Food and Agriculture* 91(11): 1931-1939.
- Wachira, F. (2001). 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*: 12-14.
- Wachira, F. N. & Kamunya, S. (2005). Kenyan teas are rich in antioxidants. *Tea* 26(2): 81-89.
- Wan, Y. Y., Chen, S. Y., Huang, Y. W., Li, X., Zhang, Y., Wang, X. J., & Bai, J. G. (2014). Caffeic acid pretreatment enhances dehydration tolerance in cucumber seedlings by increasing antioxidant enzyme activity and proline and soluble sugar contents. *Scientia Horticulturae* 173, 54-64.
- Wang, K. & Ruan, J. (2009). Analysis of chemical components in green tea in relation with perceived quality, a case study with Longjing teas. *International Journal of Food Science & Technology* 44(12): 2476-2484.
- Warrack, B. M., Hnatyshyn, S., Ott, K. H., Reily, M. D., Sanders, M., Zhang, H. & Drexler, D. M. (2009). Normalization strategies for metabolomic analysis of urine samples. *Journal of Chromatography B* 877(5-6): 547-552.
- Welti, R., Li, W., Li, M., Sang, Y., Biesiada, H., Zhou, H. E., ..., & Wang, X. (Profiling membrane lipids in plant stress responses role of phospholipase D α in freezing-induced lipid changes in Arabidopsis. *Journal of Biological Chemistry* 277(35): 31994-32002.
- Worley, B., Halouska, S. & Powers, R. (2013). Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry* 433(2): 102-104.
- Wright, L. P., Mphangwe, N. I. K., Nyirenda, H. E., & Apostolides, Z. (2002). 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* 82(5): 517-525.
- Wright, L. P., Mphangwe, N. I. K., Nyirenda, H. E., & Apostolides, Z. (2000). Analysis of caffeine and flavan-3-ol composition in the fresh leaf of *Camellia sinensis* for predicting the quality of the black tea produced in Central and Southern Africa. *Journal of the Science of Food and Agriculture* 80(13): 1823-1830.
- www.reportlinker.com/tea/reports.
- Xu, Y.-Q., Zhang, Y.-N., Chen, J.-X., Wang, F., Du, Q.-Z. & Yin, J.-F. (2018). Quantitative analyses of the bitterness and astringency of catechins from green tea. *Food Chemistry* 258: 16-24.
- Yan, S.-H. (2007). NIR evaluation of the quality of tea and its market price. *Spectroscopy Europe* 19(2): 16-19.
- Yu, H., Wang, J., Yao, C., Zhang, H., & Yu, Y. (2008). Quality grade identification of green tea using E-nose by CA and ANN. *LWT-Food Science and Technology* 41(7): 1268-1273.
- Zeitouni, K. & Chelghoum, N. (2001). Spatial decision tree-application to traffic risk analysis. In *Proceedings ACS/IEEE International Conference on Computer Systems and Applications* 203-207: IEEE.
- Zhang, J., Wang, X., Yu, O., Tang, J., Gu, X., Wan, X. & Fang, C. (2010). Metabolic profiling of strawberry (*Fragaria × ananassa* Duch.) during fruit development and maturation. *Journal of Experimental Botany* 62(3): 1103-1118.
- Zhou, B., Xiao, J. F., Tuli, L. & Resson, H. W. (2012). LC-MS-based metabolomics. *Molecular BioSystems* 8(2): 470-481.
- Zhou, J., Ho, C. T., Long, P., Meng, Q., Zhang, L. & Wan, X. (2019). Preventive Efficiency of Green Tea and Its Components on Nonalcoholic Fatty Liver Disease. *Journal of Agriculture and Food Chemistry* 67(19): 5306-5317.

Appendix 4.1: Peer-reviewed scientific article based on results from Chapter 4

International Journal of Research in Agronomy 2020; 3(2): 09-21



International Journal of Research in Agronomy

E-ISSN: 2618-0618
P-ISSN: 2618-060X
© Agronomy
www.agronomyjournals.com
2020; 3(2): 09-21
Received: 06-05-2020
Accepted: 08-06-2020

Christopher Nyarukowa
Department of Biochemistry,
University of Pretoria, Private Bag
X20, Hatfield, 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
Kericho, Kenya

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

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

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

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

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

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

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 F₁ 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) ^[23]. 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 \pm 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 (Sepsco, 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 μ g/ml EDTA and 10 mg/ml ascorbic acid, all purchased from Sigma-Aldrich, South Africa. About 100 μ l of each resultant dilution was then filtered through a 0.2 μ m 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 μ m, 2.1 \times 150 mm), at 40°C, 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 μ m 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 μ l 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 Gropppe *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).

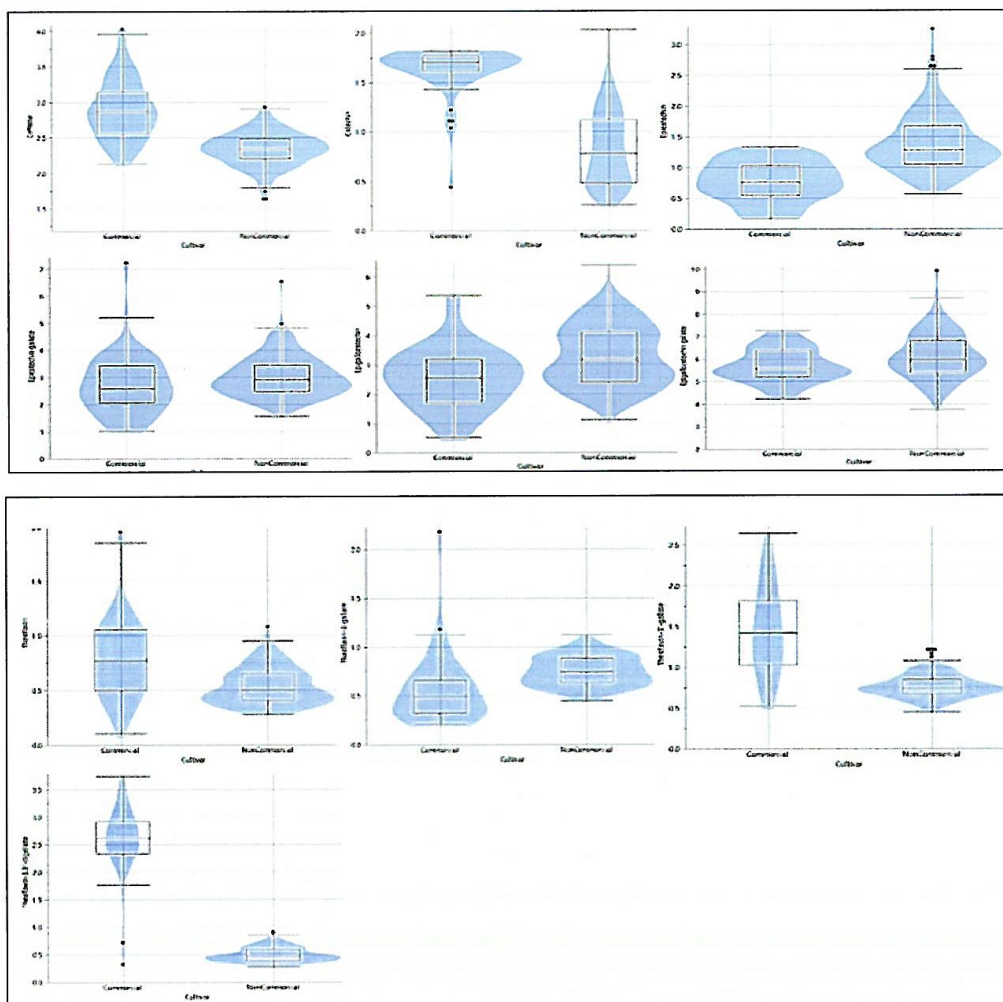


Fig 1: Violin plots showing separation between the Comm and NComm cultivars based on detected metabolites. The y-axis units for the CAF, and the catechins are % w/w dry weight; TF1-TF4 in black tea samples were quantified as EGCg equivalents, based on the EGCg response factor. The black dots represent 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.

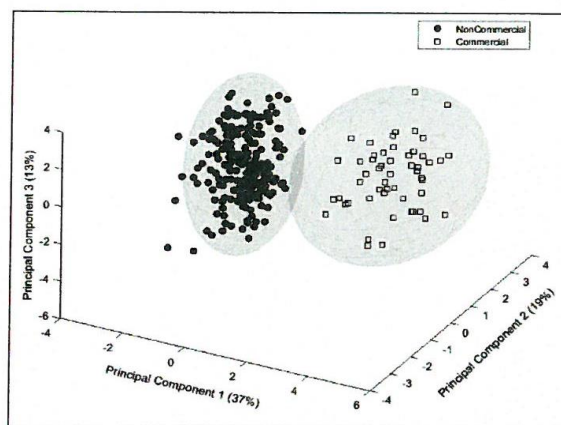


Fig 2: The PCA scores plot for the first three principal components. The plot shows good separation and explaining 69% of the variation observed between the Comm and the NComm cultivars. Ellipsoids represent 95% CI of score centroids of each class. The percentage of the overall variation explained by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

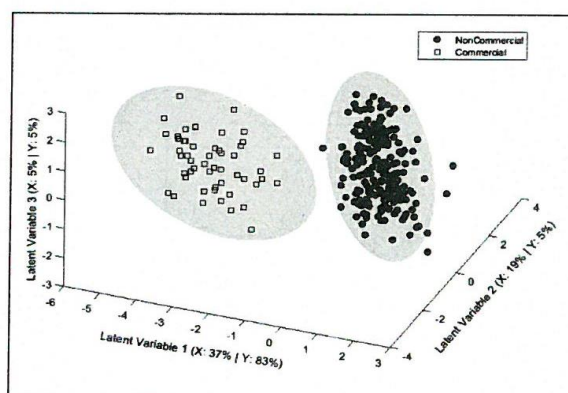


Fig 3: The PLS-DA scores plot for the first three latent variables. The plot shows clear separation between the Comm and the NComm cultivars. The goodness-of-fit values achieved for the UPLC/DAD model was deemed reliable with predictive accuracy $R^2=94\%$ and leave-one-out crossvalidated predictive 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].

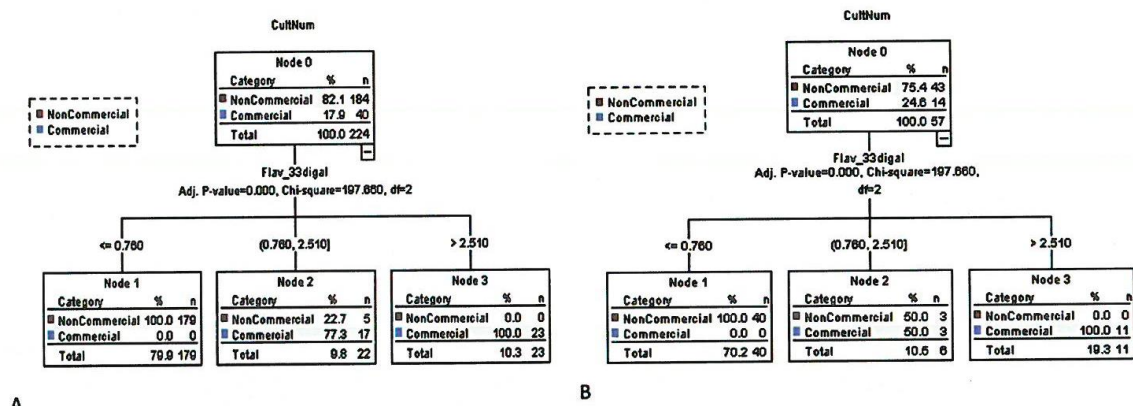


Fig 4: (A) CHAID decision tree – training set, and (B) CHAID decision tree – validation set, based on the four theaflavins variables.

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).

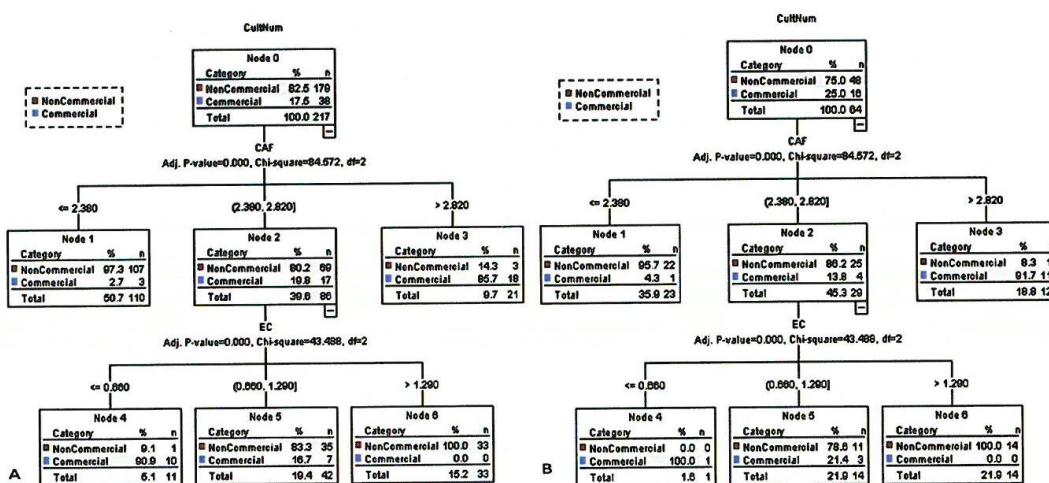


Fig 5: (A) CHAID decision tree – training set, and (B) CHAID decision tree – testing set, based on CAF, EC, ECg, EGC and EGCg.

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.

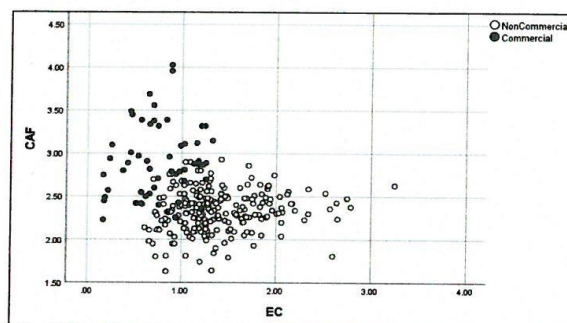
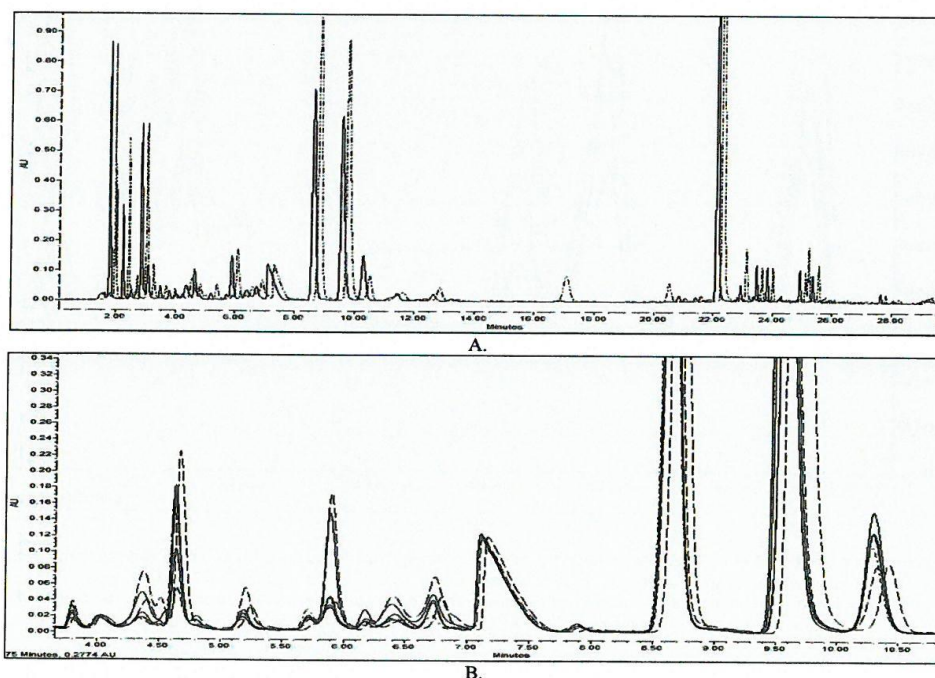
**Fig 6:** Scatter plot showing the distribution of Comm and NComm cultivars based on % w/w CAF vs EC

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 sulphaniamide (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.

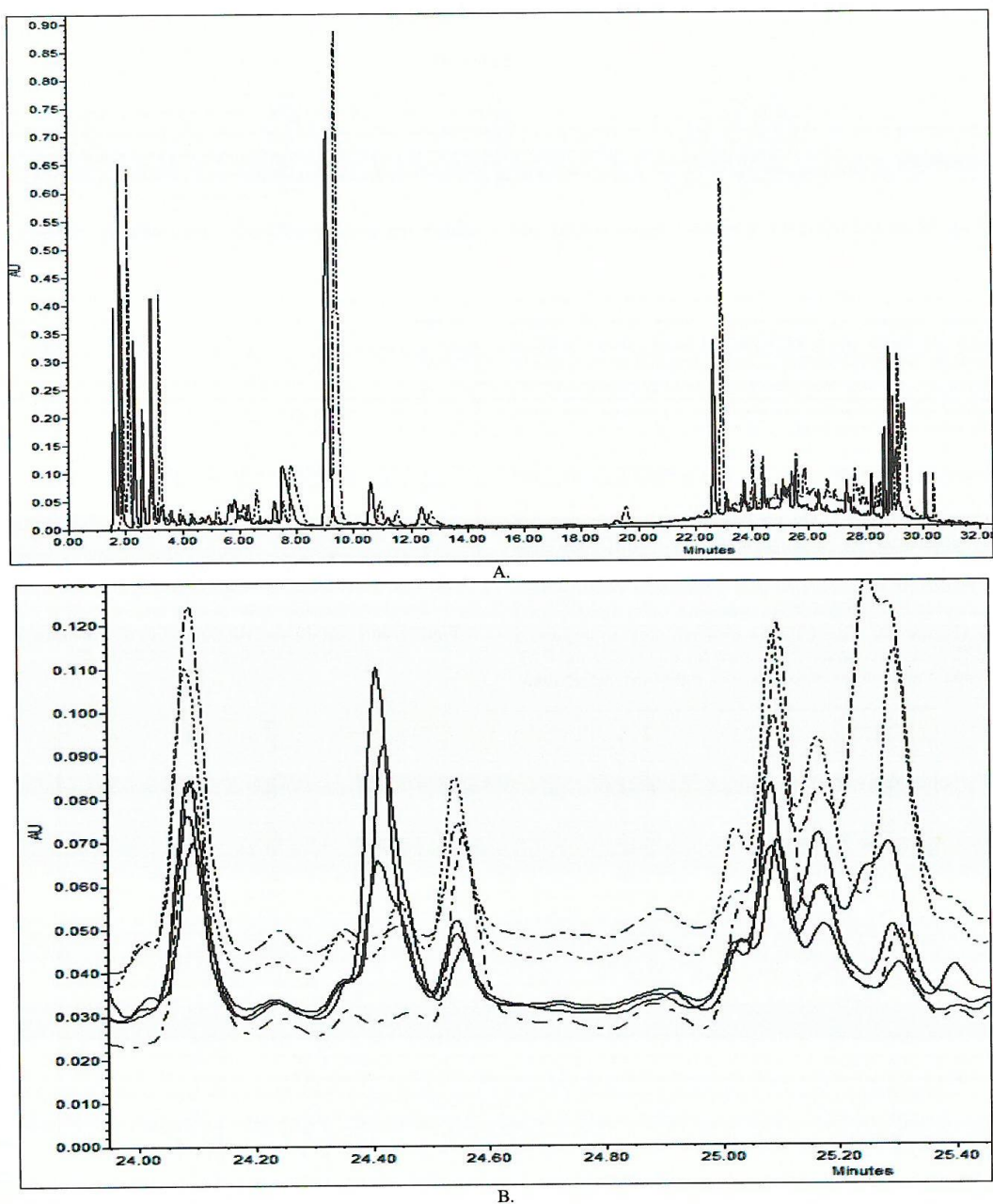


Fig 8: (A) Superimposed black tea UPLC/DAD chromatograms of one Comm and one NComm cultivar, offset and standards as in Figure 7. (B) shows the expanded chromatograms of three Comm and three NComm cultivars, showing the position of TF1 (24.05 min), TF2 (24.40 min), TF3 (24.55 min) and TF4 (25.10 min). 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.

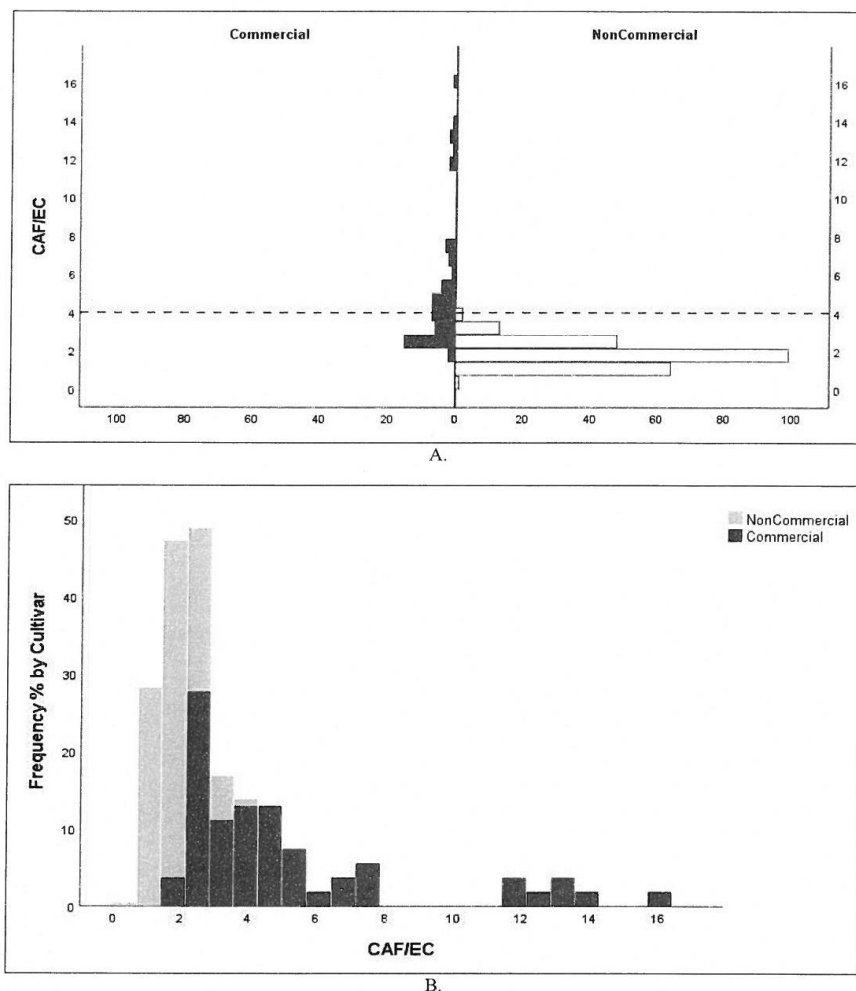


Fig 9: (A) Distribution of CAF/EC ratio by cultivar. (B). Stacked histogram with the CAF/EC ratio as a variable.

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).

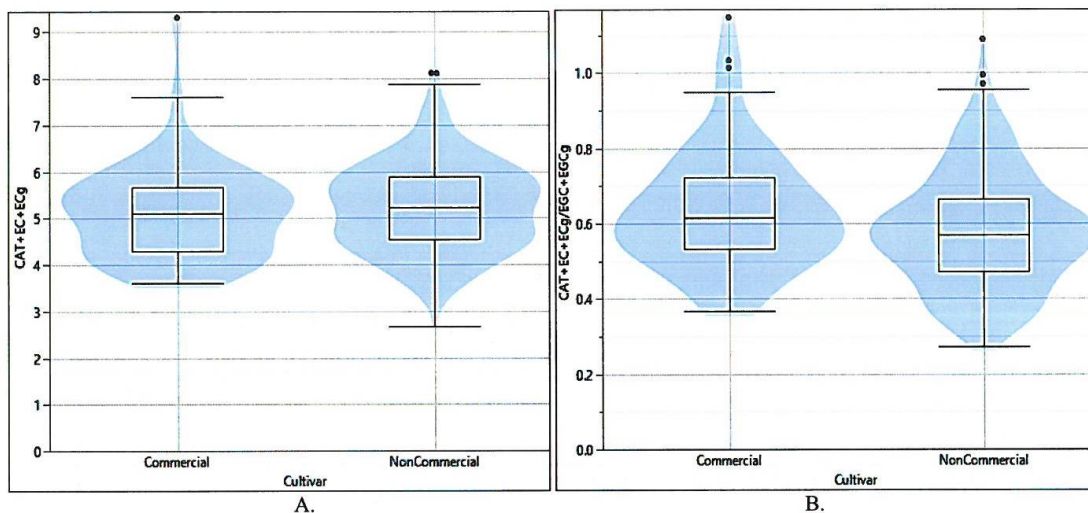


Fig 10: (A) The sum of simple catechins based on Wright *et al.*, 2000 [25]. (B) The ratio of simple to complex catechins based on Ellis and Nyirenda, 1995

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:

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

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

- 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.
- Anon. Seed garden (barie). Annual Report., Tea Research Foundation of Kenya, Tea Board of Kenya, 1990, 25.
- Banerjee B. Selection and breeding of tea. In *Tea*, 1992, 53-86.
- Breiman LF, Olshen JH, Stone RA. Classification and regression trees. Belmont, Calif.: Wadsworth, 1984.
- Chang K. World tea production and trade Current and future development, Food and Agricultural Organization of the United Nations, 2015.
- Chen L, Apostolides Z, Chen ZM. Global tea breeding: achievements, challenges and perspectives. Edn 1, Springer Science & Business Media, 2013.
- Cheserek BC, Elbehri A, Bore J. Analysis of links between climate variables and tea production in the recent past in Kenya. *Domnish Journal of Research in Environmental Studies*. 2015; 2(2):5-17.
- 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.
- Ellis R, Nyirenda H. A successful plant improvement programme on tea (*Camellia sinensis*). *Experimental Agriculture*. 1995; 31(3):307-323.
- 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.
- Green M. An evaluation of some criteria used in selecting large-yielding tea clones. *The Journal of Agricultural Science*. 1971; 76(1):143-156.
- Groppe DM, Urbach TP, Kutas M. Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology*. 2011; 48(12):1726-1737.
- Hagel JM, Facchini PJ. Plant metabolomics: analytical platforms and integration with functional genomics. *Phytochemistry Reviews*. 2008; 7(3):479-497.
- Hicks A. Current status and future development of global tea production and tea products. *AUJT*. 2009; 12(4):251-264.
- 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.
- Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. *The American Statistician*. 1998; 52(2):181-184.
- 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.
- IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp, 2019.
- 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.
- 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.
- 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.
- 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.
- Kenya National Bureau of Statistics. Kenya facts and figures Kenya, 2012.
- Khan N, Mukhtar H. Tea polyphenols for health promotion. *Life sciences*. 2007; 81(7):519-533.
- 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.
- Le Gall G, Colquhoun IJ, Defervez 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.
- Milanović M, Stamenković M. CHAID decision tree: Methodological frame and application. *Economic Themes*. 2016; 54(4):563-586.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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, 2008, 1.
 45. Wright LP, Mphangwe NIK, Nyirenda HE, Apostolides Z. Analysis of caffeine and flavan-3-ol composition in the fresh leaf of *Camellia sinensis* 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, Ransom HW. LC-MS-based metabolomics. *Molecular BioSystems* 2012; 8(2):470-481.

Appendix 4.2: Cohen's d effect size definition, calculation and interpretation used in the statistical analysis of Chapter 4 data

Cohen's d

<p>$d = M_1 - M_2 / \sigma$</p> <p>where</p> $\sigma = \sqrt{[\sum(X - M)^2 / N]}$ <p>where X is the raw score, M is the mean, and N is the number of cases.</p>	<p>Cohen (1988) defined <i>d</i> as the difference between the means, $M_1 - M_2$, divided by standard deviation, σ, of either group. Cohen argued that the standard deviation of either group could be used when the variances of the two groups are homogeneous.</p> <p>In meta-analysis the two groups are considered to be the experimental and control groups. By convention the subtraction, $M_1 - M_2$, is done so that the difference is positive if it is in the direction of <i>improvement</i> or in the predicted direction and negative if in the direction of <i>deterioration</i> or opposite to the predicted direction.</p> <p><i>d</i> is a descriptive measure.</p>
<p>$d = M_1 - M_2 / \sigma_{pooled}$</p> $\sigma_{pooled} = \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]}$	<p>In practice, the pooled standard deviation, σ_{pooled}, is commonly used (Rosnow and Rosenthal, 1996).</p> <p>The pooled standard deviation is found as the root mean square of the two standard deviations (Cohen, 1988, p. 44). That is, the pooled standard deviation is the square root of the average of the squared standard deviations. When the two standard deviations are similar the root mean square will be not differ much from the simple average of the two variances.</p>
<p>$d = 2t / \sqrt{df}$</p> <p>or</p> $d = t(n_1 + n_2) / [\sqrt{df}\sqrt{(n_1 n_2)}]$	<p><i>d</i> can also be computed from the value of the <i>t</i> test of the differences between the two groups (Rosenthal and Rosnow, 1991). In the equation to the left "df" is the degrees of freedom for the <i>t</i> test. The "n's" are the number of cases for each group. The formula without the n's should be used when the n's are equal. The formula with</p>
	<p>separate n's should be used when the n's are not equal.</p>
<p>$d = 2r / \sqrt{1 - r^2}$</p>	<p><i>d</i> can be computed from <i>r</i>, the ES correlation.</p>
<p>$d = g\sqrt{N/df}$</p>	<p><i>d</i> can be computed from Hedges's <i>g</i>.</p>

The interpretation of Cohen's *d*

Cohen's Standard	Effect Size	Percentile Standing	Percent of Nonoverlap
	2.0	97.7	81.1%
	1.9	97.1	79.4%
	1.8	96.4	77.4%
	1.7	95.5	75.4%
	1.6	94.5	73.1%
	1.5	93.3	70.7%
	1.4	91.9	68.1%
	1.3	90	65.3%
	1.2	88	62.2%
	1.1	86	58.9%
	1.0	84	55.4%
LARGE	0.9	82	51.6%
	0.8	79	47.4%
	0.7	76	43.0%
	0.6	73	38.2%
MEDIUM	0.5	69	33.0%
	0.4	66	27.4%
	0.3	62	21.3%
SMALL	0.2	58	14.7%
	0.1	54	7.7%
	0.0	50	0%

Cohen (1988) hesitantly defined effect sizes as "small, $d = .2$," "medium, $d = .5$," and "large, $d = .8$ ", stating that "there is a certain risk in inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" (p. 25).

Effect sizes can also be thought of as the average percentile standing of the average treated (or experimental) participant relative to the average untreated (or control) participant. An ES of 0.0 indicates that the mean of the treated group is at the 50th percentile of the untreated group. An ES of 0.8 indicates that the mean of the treated group is at the 79th percentile of the untreated group. An effect size of 1.7 indicates that the mean of the treated group is at the 95.5 percentile of the untreated group.

Effect sizes can also be interpreted in terms of the percent of nonoverlap of the treated group's scores with those of the untreated group, see Cohen (1988, pp. 21-23) for descriptions of additional measures of nonoverlap.. An ES of 0.0 indicates that the distribution of scores for the treated group overlaps completely with the distribution of scores for the untreated group, there is 0% of nonoverlap. An ES of 0.8 indicates a nonoverlap of 47.4% in the two distributions. An ES of 1.7 indicates a nonoverlap of 75.4% in the two distributions.

Becker, L. A. (2000). Effect size (ES). Retrieved September, 9, 2007.

CHAPTER 5

CONCLUDING DISCUSSION AND RECOMMENDATION

5.1 Concluding discussion

The present study sought to identify metabolomic markers associated with yield, drought tolerance and quality traits and document their possible biochemical mechanisms in black tea *C. sinensis* cultivars. As emphasised throughout the thesis, climate change, due to the effects global warming, is causing droughts, which are affecting crop production. Tea is one of Kenya's key cash crops, providing revenue for up to 3 million individuals, and like the rest of the globe, Kenya is experiencing changes in weather patterns. These include substantial temperatures increases, rainfall decreases and an increase in droughts, frosts and hailstorms. Tea farming depends on a good distribution of rainfall; as such these changes in climate pose a significant threat to its global supply chains; to survive, plants must reconfigure their metabolic pathways. Because of the high occurrence of droughts, the first step of this study saw the validation of the Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) method, developed on four cultivars from the Tea Research Foundation for Central Africa in Malawi in 2016, to distinguish between drought tolerant and drought susceptible cultivars. Method validation was conducted on 400 samples from the Tea Research Institute in Kenya, and has been published in a peer review journal. The obtained results showed that a sample size of 20 tea trees was deemed sufficient to compare the drought susceptibility of large tea fields of approximately 5 - 20 hectares, containing 50 000 - 200 000 tea trees, where the difference between the fields' mean values, as measured by the SWAPDT method, was at least 6%. The SWAPDT scores for each of the 400 samples used correlated with the historical records of the fields from where the samples were taken. With the SWAPDT method validated, this method was then applied to the 310 cultivars (60 open-pollinated cultivars, pre-selected for their high yield, and good tea liquor, which formed the Comm group, and the 250 cultivars which were the F₁ progeny of a reciprocal cross between two heterozygous parental clones TRFK 303/577 and GW Ejulu, which formed the NComm group) used for the genomic and metabolomics studies.

Tea quality relies on the precise metabolite profiles of teas, which are responsible for its flavour and aroma. The present study, employing the two *C. sinensis* populations i.e. 60 Comm cultivars and 250 NComm cultivars (TRFK St. 504 and TRFK St. 524), identified the QTLs responsible for yield, drought tolerance and quality centred on a genetic map constructed using the DArTseq platform. The map

comprised 15 linkage groups analogous to chromosome haploid number of tea plant ($2n = 2x = 30$) and spanned 1260.1 cM with a mean interval of 1.1 cM between markers. A 20 cM gap was noted between LG06 and LG15 adjacent markers, a possible result of gaps in both the parents used for mapping, which could have led to a lack of recombination events. Sixteen phenotypic traits were evaluated in both segregating populations and used to discover QTLs responsible for the traits of interest in both the black and green tea. Three, 11 and 46 putative QTLs were discovered after mapping on the 15 linkage groups, associated with tea quality from GC-MS, $^1\text{H-NMR}$ and UPLC data respectively. Constructing genetic linkage maps is an important requisite for QTL identification of agronomically significant genes such as those responsible for yield and quality, which are influential in developing better-quality cultivars. From the GC-MS data, one arabinose, one phloroglucinol, and one xylic acid QTLs were derived, with the %PVE ranging from 4.6 to 7.5 and averaging 5.9%. From the $^1\text{H-NMR}$ data, one acetic acid, one caffeine, three catechins (one catechin, one EC and one EGC), one chlorogenic acid, five amino acids (two isoleucine and three valines) were detected, with the %PVE by each QTL varying from 5.1 to 96.3%, and averaging 34.4%. Lastly, six caffeine, 25 catechins, three theaflavins, nine organoleptic scores and three %RWC QTLs were identified, with a %PVE varying between 5.5 to 56.6%, and averaging 9.9%. The high PVE displayed by the $^1\text{H-NMR}$ QTLs acetic acid, epicatechin, isoleucine and valine, and the UPLC QTLs caffeine, catechins, theaflavins, organoleptic scores, and %RWC suggests that these attributes could be controlled by critical genes. Besides the QTLs for catechins obtained across the $^1\text{H-NMR}$ and UPLC platforms, the current study also incorporated QTLs for acetic acid, caffeine, chlorogenic acid, isoleucine and valine from $^1\text{H-NMR}$, and arabinose, phloroglucinol and xylic acid from GC-MS, which influence the quality of tea. It was interesting to note that the QTLs associated with caffeine, catechin, EC and ECg from both $^1\text{H-NMR}$ and UPLC were on different LGs, and at different positions on the chromosome, with different %PVE. This clearly indicates that the genes concomitant with the manufacture and accretion of these metabolites are sparsely situated in different chromosomal regions. The variance explained by the QTLs varied from 4.6 to 96.3%, with an average of 28%. The UPLC analysis revealed a wide-ranging variation in the contents of caffeine, the different catechins, and the theaflavins in both the parents and their F1 progeny. As expected, the green tea had high catechins levels while the black tea had high theaflavins content. It was however noteworthy that there was no statistically significant difference in caffeine content between the green and black teas. In addition to the five major peaks i.e. CAF, EC, ECg, EGC and EGCg, several smaller peaks were detected, some of which were identified using LC-MS. This serves as an indication that numerous other metabolites are

present in tea extracts that could be contributing to the resultant quality of the tea, and other traits of interest. Using the KEGG database, the putative QTLs linked to yield, drought tolerance and quality were shown to be secondary metabolites associated with tea phenolic biomolecules and abiotic stress. Sixty seven unigenes associated with detected the putative QTLs were assigned KEGG database pathways based on secondary metabolite biosynthesis categories. The most predominant unigenes involved carbohydrate and amino acids biosynthesis; these are involved in plant hormone signal transduction pathways during abiotic stress and the biosynthesis of flavonoids, phenylalanine, thiamine, tyrosine, and tryptophan. Six enzyme categories were involved in the various metabolic pathways, namely hydrolases, isomerases, ligases, lyases, oxidoreductases, and transferases.

In the present study, PCA and PLS-DA were performed on the GC-MS, ¹H-NMR and UPLC data. The GC-MS results showed that the metabolites arabinose, catechin, gallic acid, glycerol, phloroglucinol, sucrose and xylonic acid were metabolites, which separated the Comm from the NComm cultivars. Literature has shown that abiotic stress such as drought affects the photosynthetic pathway of plants, and drastically affects their primary metabolism, which affects sugars, sugar alcohols, and amino acids. DT plants have been of a higher quality than the DS, as they efficiently up-regulate their production of sugars, which they utilise as an energy source during stress. The DT cultivars produce better quality tea liquor. Carbohydrates such as the detected glucose and sucrose have also been shown to influence the biosynthesis of other energy-generating metabolites, responsible for the alteration of gene expression and signal transduction. Caffeine, another detected metabolite, has besides being a stimulant, been documented to contribute to tea briskness, while theophylline and theobromine have been shown to contribute to the mellowness and sweetness of oolong tea. Several detected metabolites were shown to relate to the sweet taste of tea such as the GC-MS detected arabinose, arabitol, glycerol, malic acid, phloroglucinol, psicose, ribitol, sucrose, and xylonic acid, and the ¹H-NMR detected glucose, methanol and sucrose. The total sweeteners were higher ($P < 0.05$) in Comm cultivars as compared to the NComm cultivars. This study reports for the first time in the leaves of black tea cultivars that several metabolites, related to sweetness, which are higher in the Comm than the NComm cultivars. The present study reported the total amino acid concentration of the detected amino acids alanine, isoleucine, leucine, theanine, and valine, was higher in the Comm (21.8 mg/g) than the NComm (20.9 mg/g) dry weight. These compounds were postulated to possibly mask the bitterness, resulting from e.g. caffeine and chlorogenic acid, and as such contributing to the overall higher quality of the Comm cultivars. Because

256

amino acids are responsible for the aroma of tea, this means the higher the total amino acids, the more aromatic the tea. The total amino acid and total sweeteners concentrations being higher in the Commercial cultivars result in the teas produced from these cultivars having a better taste.

The ¹H-NMR and UPLC detected catechins are the substrates from which theaflavins and thearubigins are produced during manufacturing black tea. Theaflavins contribute to the briskness, brightness and astringency of black tea, which are important traits in tea quality determination. They are the predominant constituents of black tea-cream upon cooling (Roberts, 1963); 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. The GC-MS, ¹H-NMR and UPLC-DAD results were used to generate LR models for prediction. The results also show that it is now easier for breeders to predict the quality of new field selections by employing UPLC-DAD results in one of three, cost effective ways. The first is by making use of the CAF/EC ratio; if the ratio is 3.2 and above, that cultivar is 79% likely to be commercialised. The second way of ascertaining this is by using the decision trees, with CAF and EC as predictors. Third, they can make use of LR. One benefit of using decision trees and the CAF/EC ratio is that these enable the breeders to evaluate their results without the need of consulting a statistician, which is cost effective i.e. no need to pay consultation fee. 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%. By making use of the model LR based on CAF and the five catechins, breeders will be more successful in identifying and field selections rich in catechins, which will cause teas rich in theaflavins, and higher market price. Last, eight Commercial and eight NonCommercial cultivars were analysed using UPLC-MS. Additional metabolites

such as caffeic acid were identified as contributing to drought tolerance, yield and higher quality of the Comm cultivars as compared to the NComm cultivars.

5.2 Recommendations

Based on the findings from the current study, the following recommendations are made for future research:

- To eliminate the possibility of the obtained results being due to the environment, the parental clones and their F₁ progenies employed in the study may be replanted in a different geographical location, such that if the same results obtained in the present study are obtained again, this would eliminate the GxE variable, and lend further credibility to them.
- The ¹H-NMR sample size used for population mapping may have to be increased, to ensure better estimations of QTL loci and effect.
- Further targeted metabolomics studies are warranted on the GC-MS metabolites reported in Table 4.1.
- Improve on the UPLC-DAD chromatography to separate and accurately quantify the CAT peak, as CAT has been shown to be an important metabolite predictor in distinguishing between Comm and NComm cultivars.
- The current study was performed using NComm cultivars from two parents. It is therefore warranted that this part of the study be repeated with open pollinated NComm cultivars, to validate, and strengthening the reliability of the predictive ability of the models obtained in the present study.