

Identification of quantitative trait loci (QTLs) associated with agronomic traits in tea (*Camellia sinensis*) (L.) O. Kuntze

By

Robert Kipyegon Koech

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

14th February 2019

SUBMISSION DECLARATION

I, Robert Kipyegon Koech declare that the thesis/dissertation, 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.

Signature:



Date: 14th February 2019

PLAGIARISM DECLARATION

Full name: Robert Kipyegon Koech

Student number: u14454808

Title of the work: Identification of quantitative trait loci (QTLs) associated with agronomic traits in 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: 14th February 2019

ACKNOWLEDGEMENTS

I would like to express my special and sincere gratitude to my supervisor, Prof. Zeno Apostolides, Head of Complementary and Alternative Medicines (CAM) research group, Biochemistry Division, Department of Biochemistry, Genetics and Microbiology University of Pretoria (UP), South Africa, for giving me the opportunity to do research in his laboratory and providing invaluable guidance throughout my research period. I attribute my successful completion of my PhD degree to his constant encouragement, effort and positive criticism. One could not wish for a friendlier or better supervisor.

My special gratitude to Dr. J.K. Bore the Institute Director, KALRO, Tea Research Institute, Kericho, Kenya who not only permitted me to register for PhD studies but also encouraged me since the commencement of the study. My sincere gratitude also to research collaborators led by Dr. Samson Kamunya, Centre Director, KALRO, Tea Research Institute (TRI), Kericho, and Mr. Richard Mose, Research Manager, Department of Applied Research, James Finlay Kericho (Kenya) Ltd. not only for their many insightful discussions and knowledge about this research study but also their facilitation of KALRO, TRI-Kenya Tea Growers Association (KTGA)-UP collaboration.

I recognise the role played by KALRO-TRI Crop Improvement and Management programme, Tea processing and Value Addition programme and their technicians who assisted me during the collection of phenotypic data, DNA isolation and tea manufacture. I would also like to thank KTGA staff and field technicians for their assistance during my sample collection in their tea estates. I would not forget my colleagues of CAM research group and the staff members, Division of Biochemistry, for their assistance, support and friendship during the entire period of my study.

The research work would not have been possible without the financial support, and I extend my thanks to James Finlay (Kenya) Ltd., George Williamson (Kenya) Ltd., Sotik Tea Company (Kenya) Ltd., Mcleod Russell (Uganda) Ltd., TRI of Kenya and Southern African Biochemistry and Informatics for Natural Products (SABINA), Technology and Human Resources for Industry Programme (THRIP), South Africa, National Research Foundation (NRF) of South Africa, and the University of Pretoria (South Africa) for the funds.

I am grateful to my parents, mother in law and grandparents for their love, prayers and support. I am greatly indebted to my wife Ruth and daughter Abigail for their love, understanding, and prayers and their continued support while I was away from home to complete my PhD studies. Also, I express my sincere thanks to my brothers, sisters in law and brother in laws for their support and valuable prayers.

I thank God, the Almighty, for His showers of blessings during my research period and the successful completion of my PhD studies.

DEDICATION

*I dedicate this thesis to
my family, my wife, Ruth, and my beloved daughter Abigail
for their constant support and unconditional love.*

ABSTRACT

The identification of candidate genes associated with black tea quality and drought tolerance traits in tea (*Camellia sinensis*) plants remains complicated and time-consuming. This study aimed to identify molecular, physiological and biochemical characteristics associated with black tea quality and drought tolerance traits in two tea segregating populations for future marker-assisted selection breeding. The two tea segregating populations TRFK St. 504 and TRFK St. 524 were developed from a reciprocal cross of two heterozygous parental clones, TRFK 303/577 and GW Ejulu. Using Ultra-Performance Liquid Chromatography (UPLC), catechin fractions and theaflavin fractions were analysed in green and black tea, respectively, while caffeine content was analysed in both green and black tea. Black tea samples were further subjected to organoleptic evaluation by professional tea tasters to score liquor characters (colour, brightness, strength and briskness). The percent relative water content (%RWC) using Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) method was used to distinguish between drought susceptible and drought-tolerant tea cultivars. A total of 16 phenotypic data from two segregating tea populations were used to identify the quantitative trait loci (QTLs) influencing tea biochemical, and drought stress traits based on a consensus genetic map constructed using the DArTseq platform. The map consisted of 15 linkage groups and corresponded to the chromosome haploid number of tea plant ($2n = 2x = 30$) and spanned 1 260.1 cM with a mean interval of 1.1 cM between markers. A total of 1 421 DArTseq markers derived from the linkage map identified 53 DArTseq markers to be linked to black tea quality and %RWC. The putative QTLs linked to black tea quality, and drought tolerance traits were submitted to BLAST and assigned functions using Gene Ontology (GO) terms and biosynthetic pathways in the tea genome, gene ontology database and Kyoto Encyclopedia of Genes and Genomes (KEGG). The approach of combining linkage mapping with association mapping allowed identification, and precise authentication of putative QTLs with an additional six more putative QTLs identified. The functional annotations of all the putative QTLs detected were involved in the metabolism of secondary metabolites associated with tea phenolic biomolecules and abiotic stress. The predictive ability of all machine learning models varied across the phenotypic traits. The putative QTLs + annotated proteins + KEGG pathways based prediction approach showed more robustness and usefulness in the prediction of phenotypes than the individual prediction based models. The Extreme Learning Machine (ELM) model had a better

prediction ability for catechin, astringency, brightness, briskness and colour based on putative QTLs + annotated proteins + KEGG pathways approach. The percent variables of importance of putatively annotated proteins and KEGG pathways were associated with the phenotypic traits.

TABLE OF CONTENTS

SUBMISSION DECLARATION	II
PLAGIARISM DECLARATION.....	III
ACKNOWLEDGEMENTS	IV
DEDICATION	VI
ABSTRACT.....	VII
TABLE OF CONTENTS.....	IX
LIST OF FIGURES	XVI
LIST OF TABLES	XVIII
LIST OF APPENDICES.....	XXIII
ABBREVIATIONS	XXIV
CHAPTER 1	1
GENERAL INTRODUCTION.....	1
1.1 BACKGROUND INFORMATION	1
1.2 PROBLEM STATEMENT	5
1.3 RESEARCH JUSTIFICATION	5
1.4 RESEARCH OBJECTIVE	6
1.5 RESEARCH OUTPUTS	8
REFERENCES	9
CHAPTER 2	12
LITERATURE REVIEW	12
2.1 THE TEA PLANT	12
2.1.1 BOTANY OF TEA.....	12
2.1.2 THE CHEMICAL COMPOSITION OF TEA FLUSH	12
2.1.2.1 Tea flavonoids	12
2.1.2.2 Tea Polyphenols.....	14

2.1.2.3 Flavonoids biosynthesis	15
2.1.2.4 Caffeine biosynthesis	15
2.1.2.5 Chemical composition of made tea.....	17
2.1.3 Major tea products	18
2.1.3.1 Green tea	18
2.1.3.2 Oolong tea.....	21
2.1.3.3 Black tea.....	22
2.1.4 Major steps in black tea processing.....	22
2.1.4.1 Withering	22
2.1.4.2 Rolling.....	23
2.1.4.3 Aeration (Fermentation).....	23
2.1.4.4 Drying	24
2.1.4.5 Sorting and grading.....	25
2.1.5 Tea quality indicators	26
2.2 HISTORY OF TEA BREEDING	27
2.2.1 Tea breeding improvements in Kenya.....	28
2.2.2 Selection and breeding of high-quality tea clones	29
2.2.3 Breeding for high yield	31
2.2.4 Breeding for pest and disease tolerance.....	31
2.2.5 Breeding for environmental stress	32
2.2.6 Polyploidy breeding.....	32
2.2.7 Inter-specific hybridisation	33
2.3 FUTURE TEA BREEDING STRATEGIES FOR DESIRABLE ATTRIBUTES	34
2.3.1 Tissue culture.....	35
2.3.2 Hybridisation breeding	35
2.3.3 Mutation breeding.....	36
2.3.4 Breeding for health-promoting effects	36
2.4 Marker-assisted selection	38
2.5 ROLE OF GENETIC MARKERS IN TEA BREEDING IMPROVEMENT	38
2.5.1 Morphological markers.....	39
2.5.2 Biochemical and isoenzyme markers	39

2.5.3 Molecular markers	39
2.6 UTILITY OF MOLECULAR MARKERS IN TARGETING PHENOTYPIC TRAITS	40
2.6.1 DNA-based markers	41
2.6.1.1 Restriction fragment length polymorphism (RFLP)	41
2.6.2 PCR-based methods.....	42
2.6.2.1 Random amplified polymorphic DNA (RAPD)	42
2.6.2.2 Amplified fragment length polymorphism (AFLP)	43
2.6.2.3 Simple sequence repeats (SSRs).....	43
2.6.2.4 Inter-simple sequence repeat (ISSR)	44
2.6.2.5 Sequence-tagged site (STS)	44
2.6.2.6 Cleaved amplified polymorphic sequences (CAPS).....	45
2.6.2.7 SNP	45
2.6.2.8 Diversity array technology (DArTseq)	46
2.7 QTL	47
2.7.1 QTL mapping experiments.....	48
2.7.1.1 Choice of mapping population.....	48
2.7.1.2 Size of mapping populations.....	48
2.7.1.3 Phenotypic data.....	49
2.7.1.4 Statistical methods for QTL mapping.....	49
2.7.2 Linkage mapping	50
2.7.3 Association mapping	51
2.8 ROLE OF MACHINE LEARNING AND GENOMIC-ENABLED PREDICTION IN PLANT BREEDING	52
2.8.1 Factors affecting the accuracy of genomic selection in plant breeding.....	53
2.8.1.1 Effective population size and marker density	53
2.8.1.2 Heritability	54
2.8.1.3 Size and structure of reference population.....	54
2.8.1.4 Statistical models for prediction	55
2.8.2 Genomic selection of outbred species	56
2.8.3 Other challenges in genomic selection	57
REFERENCES	59
CHAPTER 3	87

QUANTITATIVE GENETIC PARAMETERS FOR BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS	87
ABSTRACT.....	87
3.1 INTRODUCTION	88
3.2 RESEARCH OBJECTIVE	89
3.3 HYPOTHESES	89
3.4 MATERIALS AND METHODS	90
3.4.1 Plant material	90
3.4.2 Description of study site	90
3.4.3 Sample collection and processing.....	90
3.4.4 Extraction of catechins, caffeine and theaflavins	91
3.4.5 UPLC analysis of catechins, caffeine and theaflavins.....	91
3.4.6 Determination of plain black tea quality parameters.....	92
3.4.7 Determination of percent relative water content	92
3.4.8 Phenotypic trait analysis	93
3.5 DNA ANALYSIS	93
3.5.1 Extraction, purification and quantification of DNA.....	93
3.5.2 DArTseq assay.....	93
3.5.3 Linkage map construction	94
3.5.4 QTL analyses	94
3.6 RESULTS.....	94
3.6.1 Phenotypic trait analysis	94
3.6.2 Correlation analysis	98
3.6.3 Genetic map construction	108
3.6.4 Phenotype segregation and QTL mapping	108
3.7 DISCUSSION	121
3.8 CONCLUSION	123
3.9 DATA ARCHIVING STATEMENT	124
REFERENCES	125
APPENDICES	128

CHAPTER 4	157
FUNCTIONAL ANNOTATION OF PUTATIVE QTLS ASSOCIATED WITH BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS.....	157
ABSTRACT.....	157
4.1 INTRODUCTION.....	158
4.2 RESEARCH OBJECTIVES.....	159
4.3 HYPOTHESES	159
4.4 MATERIALS AND METHODS	159
4.4.1 Preparation of FASTA files.....	159
4.4.2 BLAST search	160
4.4.3 Functional annotation and pathway assignment.....	160
4.5 RESULTS.....	160
4.5.1 Functional annotation and pathway assignment.....	160
4.5.2 Annotation and mapping of gene ontology	161
4.5.3 KEGG pathway mapping: Flavonoid pathway, caffeine pathway and abiotic stress.	162
4.5.4 Functional annotation of putative candidate genes 15 linkage groups of <i>C. sinensis</i>	163
4.6 DISCUSSION.....	171
4.7 CONCLUSION	176
REFERENCES	178
APPENDICES	183
CHAPTER 5	209
COMBINED LINKAGE AND ASSOCIATION MAPPING OF PUTATIVE QTLS CONTROLLING BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS	209
ABSTRACT.....	209
5.1 INTRODUCTION.....	210
5.2 RESEARCH OBJECTIVE	210
5.3 HYPOTHESES	211
5.4 MATERIAL AND METHODS	211
5.4.1 Population type	211

5.4.2 Phenotypic data.....	211
5.4.3 Population structure.....	211
5.4.4 Linkage mapping, QTL analyses and allelic effects.....	211
5.4.5 Association mapping	212
5.5 RESULTS.....	212
5.5.1 Frequency distribution.....	212
5.5.2 Population structure.....	212
5.5.3 Linkage mapping and association mapping.....	218
5.5.3.1 Interval mapping and allelic effects	218
5.5.3.2 Multiple QTL model mapping and allelic effects.....	218
5.5.3.3 Association mapping.....	219
5.5.4 A comparison of linkage mapping and association mapping.....	227
5.5.5 Functional annotation of putative QTLs in linkage and association mapping	231
5.6 DISCUSSION.....	233
5.7 CONCLUSION	235
REFERENCES	236
APPENDICES	241
CHAPTER 6	257
APPLICATION OF MACHINE LEARNING METHODOLOGIES IN GENOMIC SELECTION OF BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS	257
ABSTRACT.....	257
6.1 INTRODUCTION.....	258
6.2 RESEARCH OBJECTIVE	259
6.3 HYPOTHESES	260
6.4 MATERIALS AND METHODS	260
6.4.1 Genome selection dataset	260
6.4.2 Transforming and ranking phenotypic data.....	260
6.4.3 Partition designs and cross-validation of the dataset.....	260
6.4.4 Prediction models	261

6.4.5 Comparison of performance of different prediction models	261
6.5 RESULTS	261
6.5.1 Percent RWC	262
6.5.2 Caffeine	265
6.5.3 Catechin	269
6.5.4 Tea tasters' scores.....	272
6.5.4.1 Aroma	272
6.5.4.2 Astringency	275
6.5.4.3 Brightness	278
6.5.4.4 Briskness	280
6.5.4.5 Colour	283
6.6 DISCUSSION	286
6.7 CONCLUSION	289
REFERENCES	290
CHAPTER 7	294
CONCLUDING DISCUSSION AND RECOMMENDATION	294
7.1 CONCLUDING DISCUSSION	294
7.2 RECOMMENDATIONS	297

LIST OF FIGURES

Figure 1.1: Map of Kenya showing tea growing areas (Source KTDA, 2006)	2
Figure 2.1: Chemical structure of a flavonoid molecule.....	13
Figure 2.2: Chemical structure of major classes of flavonoids.....	14
Figure 2.3: A schematic diagram of flavonoid biosynthesis pathway.	16
Figure 2.4: A schematic diagram of caffeine biosynthesis pathway.	17
Figure 2.5: Chemical structures of major catechins.....	19
Figure 2.6: Chemical structure of individual theaflavins.....	20
Figure 2.7: Chemical structure of thearubigins.....	21
Figure 2.8: Chemical structures of methylxanthines	21
Figure 2.9: A schematic diagram of major steps in tea manufacture and corresponding types of tea.	25
Figure 2.10: Breeding technology pipeline from the past to the present and to the future.....	42
Figure 2.11: Schematic diagram of the principle of DArTseq.....	47
Figure 3.1: A representative UPLC chromatogram of black tea sample from parental lines TRFK 303/577 (solid line) and GW Ejulu (dotted line) with the three internal standards.	95
Figure 3.2: A representative UPLC chromatogram of green tea sample from parental lines TRFK 303/577 (solid line) and GW Ejulu (dotted line) with the three internal standards.	96
Figure 3.3: Frequency distribution pattern of phenotypic traits of both green and black tea from two parental clones and F ₁ population derived from a reciprocal cross between TRFK 303/577 and GW Ejulu (i.e. St 504 and St 524), respectively. The phenotypic values of the two parental clones are indicated in the graph.....	103
Figure 3.4: DArT-based genetic map of tea plant showing the location of QTL for caffeine, catechins, theaflavins content, liquor brightness and % rwc identified in the two mapping population. Map distance scales in cM are placed at the left margin. Significant QTL for caffeine, individual catechin and individual theaflavin content and liquor brightness is represented by different bars. Bars and lines indicate 1-LOD and 2-LOD support intervals.....	120
Figure 4.1: E-value distribution of the best matches for unigenes (E-value $\leq 1.0e-5$).....	168
Figure 4.2: Similarity distribution of the hit BLAST matches for unigenes.	168
Figure 4.3: Function classifications of GO terms of <i>C. sinensis</i> transcripts based on high-score BLASTX matches in NR plant proteins database, a total of 5,592 unigenes were classified into	

three main GO categories which are biological process (BP), molecular function (MF) and cellular component (CC), respectively.	169
Figure 4.4: KEGG ontology (KO) analysis of differentially expressed genes (DEGs) related to secondary metabolism in <i>C. sinensis</i>	170
Figure 4.5: KO distribution of major enzymes code related to secondary metabolism in <i>C. sinensis</i>	171
Figure 5.1: Delta K values plotted from 1 to 10 for 261 tea accessions panel.	214
Figure 5.2: Population structure of 261 tea accessions panel (K = 2)	215
Figure 5.3: Principal component analysis of 261 tea accessions.	216
Figure 5.4: Phylogenetic tree-based on Nei's genetic distance	217

LIST OF TABLES

Table 3.1: Phenotypic variation of caffeine and individual catechins content in parental lines (TRFK 303/577 and GW Ejulu) and 109 F ₁ progeny St 504.....	97
Table 3.2: Phenotypic variation of caffeine and individual catechins content in parental lines (GW Ejulu and TRFK 303/577) and 152 F ₁ progeny St 524.....	97
Table 3.3: Correlation coefficients (r) between phenotypic traits in the mapping population St 504.....	104
Table 3.4: Correlation coefficients (r) between phenotypic traits in the mapping population St 524.....	106
Table 3.5: DArT markers distribution among the linkage groups	111
Table 3.6: QTL for caffeine, individual catechins and theaflavins content, tea taster's scores and percent relative water content detected in St 504 and St 524 tea plant population	112
Table 4.1: Functional annotation of putative candidate genes in 15 linkage groups of <i>C. sinensis</i> on reference tea genome	165
Table 5.1: Linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St. 504 and St. 524 using interval mapping in MapQTL 6.0	221
Table 5.2: Linkage mapping for putative QTLs in black tea quality and drought tolerance traits in St. 504 and St. 524 using multiple QTL model mapping in MapQTL 6.0	223
Table 5.3: Association mapping results for black tea quality and drought tolerance traits in St. 504 and St.524 using the GLM (Q) and MLM (Q+K) method, respectively (p <0.01)	225
Table 5.4: Association mapping and linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St.504 and St.524 using GLM (Q), MLM (Q + K) methods (p < 0.01) in TASSEL software v5.2.43 and interval mapping in MapQTL 6.0.....	228
Table 5.5: Association mapping and linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St.504 and St.524 using GLM (Q), MLM (Q + K) methods (p < 0.01) in TASSEL software v5.2.43 and multiple QTL model mapping in MapQTL 6.0 ...	229
Table 5.6: Association mapping and linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St.504 and St.524 using GLM (Q), MLM (Q + K) methods (p < 0.01) in TASSEL software v5.2.43 and interval mapping and multiple QTL model mapping in MapQTL 6.0	230

Table 5.7: Association mapping, linkage mapping and functional annotation protein of putative QTLs in black tea quality traits in St.504 and St.524 based on IM and MLM (Q + K) method ($p < 0.01$) in MapQTL 6.0 and TASSEL software v5.2.43, respectively.	232
Table 6.1: The best prediction models based on putative QTLs for %RWC trait + annotated protein + KEGG pathway in both discovery and validation datasets	262
Table 6.2: The best prediction models based on putative QTLs identified for %RWC trait in both discovery and validation dataset	263
Table 6.3: The best prediction models based on annotated protein for putative QTLs identified for %RWC trait in both discovery and validation dataset	264
Table 6.4: The best prediction models based on KEGG pathway for putative QTLs identified for %RWC trait in both discovery and validation dataset	264
Table 6.5: Percent variable of importance of annotated proteins related to %RWC trait in both discovery and validation dataset	265
Table 6.6: Percent variable of importance of KEGG pathways related to %RWC trait in both discovery and validation dataset	265
Table 6.7: The best prediction models based on putative QTLs identified for caffeine trait + annotated protein + KEGG pathway in both discovery and validation dataset	266
Table 6.8: The best prediction models based on putative QTLs identified for caffeine trait in both discovery and validation dataset	266
Table 6.9: The best prediction models based on annotated protein for putative QTLs identified for caffeine trait in both discovery and validation dataset	267
Table 6.10: The best prediction models based on KEGG pathway for putative QTLs identified for caffeine trait in both discovery and validation dataset	268
Table 6.11: Percent variable of importance of annotated proteins related to caffeine trait in both discovery and validation dataset	268
Table 6.12: Percent variable of importance of KEGG pathways related to caffeine trait in both discovery and validation dataset	269
Table 6.13: The best prediction model based on putative QTLs identified for catechin trait + annotated protein + KEGG pathway in both discovery and validation dataset	269
Table 6.14: The best prediction model based on putative QTLs identified for catechin trait in both discovery and validation dataset	270

Table 6.15: The best prediction model based on annotated protein for putative QTLs identified for catechin trait in both discovery and validation dataset	270
Table 6.16: The best prediction models based on KEGG pathway for putative QTLs identified for catechin trait in both discovery and validation dataset	271
Table 6.17: Percent variable of importance of annotated proteins related to catechin trait in both discovery and validation dataset	271
Table 6.18: Percent variable of importance of KEGG pathways related to catechin trait in both discovery and validation dataset	271
Table 6.19: The best prediction model based on putative QTLs identified for tea liquor aroma trait + annotated protein + KEGG pathway in both discovery and validation dataset	272
Table 6.20: The best prediction model based on putative QTLs identified for tea liquor aroma trait in both discovery and validation dataset	273
Table 6.21: The best prediction model based on annotated protein for putative QTLs identified for tea liquor aroma trait in both discovery and validation dataset.....	274
Table 6.22: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor aroma trait in both discovery and validation dataset.....	274
Table 6.23: Percent variable of importance of annotated proteins related to tea liquor aroma trait in both discovery and validation dataset.....	275
Table 6.24: Percent variable of importance of KEGG pathways related to tea liquor aroma trait in both discovery and validation dataset	275
Table 6.25: The best prediction model based on putative QTLs identified for tea liquor astringency trait + annotated protein + KEGG pathway in both discovery and validation dataset	276
Table 6.26: The best prediction model based on putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset.....	276
Table 6.27: The best prediction model based on annotated protein for putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset	277
Table 6.28: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset	277
Table 6.29: Percent variable of importance of annotated proteins related to tea liquor astringency trait in both discovery and validation dataset	278

Table 6.30: Percent variable of importance of KEGG pathways related to tea liquor astringency trait in both discovery and validation dataset	278
Table 6.31: The best prediction model based on putative QTLs identified for tea liquor brightness trait + annotated protein + KEGG pathway in both discovery and validation dataset	278
Table 6.32: The best prediction model based on putative QTLs identified for tea liquor brightness trait in both discovery and validation dataset	279
Table 6.33: The best prediction model based on annotated protein for putative QTLs identified for tea liquor brightness trait in both discovery and validation dataset	279
Table 6.34: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor brightness trait in both discovery and validation dataset	279
Table 6.35: Percent variable of importance of annotated proteins related to tea liquor brightness trait in both discovery and validation dataset	280
Table 6.36: Percent variable of importance of KEGG pathways related to tea liquor brightness trait in both discovery and validation dataset	280
Table 6.37: The best prediction model based on putative QTLs identified for tea liquor briskness trait + annotated protein + KEGG pathway in both discovery and validation dataset	281
Table 6.38: The best prediction model based on putative QTLs identified for tea liquor briskness trait in both discovery and validation dataset	281
Table 6.39: The best prediction model based on annotated protein for putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset	282
Table 6.40: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor briskness trait in both discovery and validation dataset.....	282
Table 6.41: Percent variable of importance of annotated proteins related to tea liquor briskness trait in both discovery and validation dataset	283
Table 6.42: Percent variable of importance of KEGG pathways related to tea liquor briskness trait in both discovery and validation dataset	283
Table 6.43: The best prediction model based on putative QTLs identified for tea liquor colour trait + annotated protein + KEGG pathway in both discovery and validation dataset	284
Table 6.44: The best prediction model based on putative QTLs identified for tea liquor colour trait in both discovery and validation dataset	284

Table 6.45: The best prediction model based on annotated protein for putative QTLs identified for tea liquor colour trait in both discovery and validation dataset	285
Table 6.46: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor colour trait in both discovery and validation dataset	285
Table 6.47: Percent variable of importance of annotated proteins related to tea liquor colour trait in both discovery and validation dataset.....	286
Table 6.48: Percent variable of importance of KEGG pathways related to tea liquor colour trait in both discovery and validation dataset	286

LIST OF APPENDICES

Appendix 3.1: DArTseq marker ID with the corresponding individual marker sequence	128
Appendix 3.2: BioProject PRJNA398959	130
Appendix 3.3: Peer-reviewed scientific article based on results obtained in Chapter 3	138
Appendix 4.1: Similarity distribution of the top BLAST matches for unigenes	183
Appendix 4.2: E-value distribution of the best matches for unigenes ($E\text{-value} \leq 1.0e\text{-}5$).....	183
Appendix 4.4: DArTseq BLAST hit distribution for tea (<i>C. sinensis</i>)	183
Appendix 4.3: DArTseq GO mapping distribution for tea (<i>C. sinensis</i>)	183
Appendix 4.6: DArTseq evidence code distribution for tea (<i>C. sinensis</i>)	184
Appendix 4.5: DArTseq evidence code distribution for BLAST hits (<i>C. sinensis</i>)	184
Appendix 4.8: DArTseq annotation distribution for tea (<i>C. sinensis</i>)	184
Appendix 4.7: DArTseq GO-level distribution for tea (<i>C. sinensis</i>)	184
Appendix 4.9: Gene ontology classification of top 20 sub-groups of biological processes.	185
Appendix 4.10: Gene ontology classification of top 20 sub-groups of molecular functions.	186
Appendix 4.11: Gene ontology classification of top 20 sub-groups of cellular components.	187
Appendix 4.12: List of putative QTLs identified from KEGG pathways with GO terms in <i>C. sinensis</i> transcriptome	188
Appendix 4.13: List of unigenes from primary and secondary metabolic pathways in 15 LGs of <i>C. sinensis</i> transcriptome	195
Appendix 4.14: Peer-reviewed scientific article based on results obtained in Chapter 4.....	1959
Appendix 5.1: Trait-wise association Q-Q plots for GLM. AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness.....	241
Appendix 5.2: Trait-wise association Q-Q plots for GLM (Q). AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness	245
Appendix 5.3: Trait-wise Q-Q association plots for mixed linear model MLM (K). AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness.....	249
Appendix 5.4: Trait-wise Q-Q association plots for mixed linear model MLM (Q+K). AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness	253

ABBREVIATIONS

ABA	Absciscic Acid
AFLP	Amplified Fragment Length Polymorphism
AM	Association Mapping
AMP	Adenosine 5' Monophosphate
AMSL	Above Mean Sea Level
ANOVA	Analysis of Variance
ANR	Anthocyanidin Reductase
ANS	Anthocyanin Synthase
AR	Aroma
AST	Astringency
BLAST	Basic Local Alignment Search Tool
BLASTN	Basic Local Alignment Search Tool (Nucleotide)
BLASTX	Basic Local Alignment Search Tool (Protein)
BP	Biological process
BRK	Briskiness
BRT	Brightness
CAM	Complementary and Alternative Medicines
CAPS	Cleaved Amplified Polymorphic Sequences
CAT	Catechin
CC	Cellular Component
CIM	Composite Interval Mapping
CIMMYT	International Maize and Wheat Improvement Centre
CL	Colour
CLUMPP	Cluster Matching and Permutation Program
CoA	Coenzyme A
CTC	Cut, Tear and Curl
DArT	Diversity Arrays Technology
EC	Evidence Code
EGC	Epigallocatechin
EGCG	Epigallocatechin gallate

ELM	Extreme Learning Machine
EST	Expressed Sequence Tags
FASBMB	Federation of African Societies of Biochemistry and Molecular Biology
FASTA	Fast Adaptive Shrinkage Thresholding Algorithm
GABA	Gamma-Aminobutyric Acid
GBS	Genotyping-by sequencing
GEV	Genomic Estimated Breeding Value
GLM	General Linear Model
GO	Gene Ontology
GRM	Genomic Relationship Matrix
GxE	Genotype x environment
IM	Interval Mapping
ISSR	Inter-Simple Sequence Repeat
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Ontology
KTDA	Kenya Tea Development Agency
KTGA	Kenya Tea Growers Association
LD	Linkage Disequilibrium
LOD	Logarithms of Odds
MAS	Marker-assisted selection
MATE	Multidrug and Toxic Compound Extrusion
MCMC	Markov Chain Monte Carlo
MF	Molecular Function
MIM	Multiple Interval Mapping
MLM	Mixed Linear Model
MQM	Multiple QTL Model
NCBI	National Centre for Biotechnology Information
NN	Neural Networks
NR	Non-Redundant
NRF	National Research Foundation
NWU	North-West University

PAL	Phenylalanine Ammonia Lyase
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PVE	Population Variability Explained
QTL	Quantitative Trait Loci
RAPD	Random Amplification of Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
RKHS	Reproducing Kernel Hilbert Space
RMSE	Root-Mean-Squared-Error
ROS	Reactive Oxygen Species
RWC	Relative Water Content
SABINA	Southern African Biochemistry and Informatics for Natural Products
SAM	S-adenosyl-L-methionine
SASBMB	South African Society of Biochemistry and Molecular Biology
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
STS	Sequence-Tagged Site
SVM	Support Vector Machines
SWAPDT	Short-time Withering Assessment of Probability for Drought Tolerance
TF	Theaflavins
TFDG	Theaflavin digallate
THRIP	Technology and Human Resources for Industry Programme
TR	Thearubigins
TRFK	Tea Research Foundation of Kenya
TRI	Tea Research Institute
UP	University of Pretoria
UPLC	Ultra-Performance Liquid Chromatography
UV	Ultraviolet

CHAPTER 1

GENERAL INTRODUCTION

1.1 Background information

Tea plant (*Camellia sinensis* (L.) O. Kuntze) is an important economic crop grown in 36 tropical and semi-tropical countries around the world for production of tea beverages. In Africa, tea is mainly grown in Kenya, Malawi, Uganda, Tanzania, Zimbabwe, Rwanda, South Africa, Burundi, and Mauritius, although some tea is also grown in Ethiopia, Nigeria and Cameroon (FAO, 2015a). In Kenya, tea was first introduced in 1903 by the Caine brothers who imported dark-leaved “Manipuri” hybrid seed from Assam to establish a plantation at Limuru, Central Kenya (Matheson and Bovill, 1950; Wilson and Co., 1962). Tea cultivation and processing is currently present in 18 of Kenya’s 47 counties with approximately 190,000 ha of land under cultivation (TRFK, 2013) and impacts a large proportion of the population. The tea growing areas are divided into two regions delineated by the Great Rift Valley. The East of Rift block comprises of the Aberdare highlands, Mt. Kenya region, and Nyabene hills while the highlands of Mau escarpment around Kericho, Nandi Hills, Mt. Elgon and the Kisii highlands form the west of Rift block (Figure 1.1). Tea is grown on both sides of Great Rift Valley which is rich in volcanic soil that is found in the highland zones at 1 500-2 700 m above sea level.

Kenya has approximately 190 000 ha of land under tea cultivation, of which 65% is managed by small-scale producers and 35% by large-scale producers (TRFK, 2013). The large-scale sector is owned by 39 large tea companies, while the smallholder sector is operated by half a million farmers who sell their produce to approximately 58 factories (TBK, 2013). The small-scale sector is managed by Kenya Tea Development Agency (KTDA), and the large-scale producers are managed by KTGA. The large-scale tea producers are primarily based west of the Great Rift Valley, while the small-scale tea producers are spread throughout the tea growing region. The large-scale estate sub-sector is composed mainly of transnational companies and some local companies with medium to large plantations of over 20 to 7, 000 hectares. The smallholder sub-sector, on the other hand, consists of individual farmers with an average holding of about 0.1 to 20 hectares (Anon, 2003). Therefore, this makes the tea industry the largest employer and provides income to more than 600,000 smallholder households and more than 150,000 workers at tea estates (FAO, 2015b).

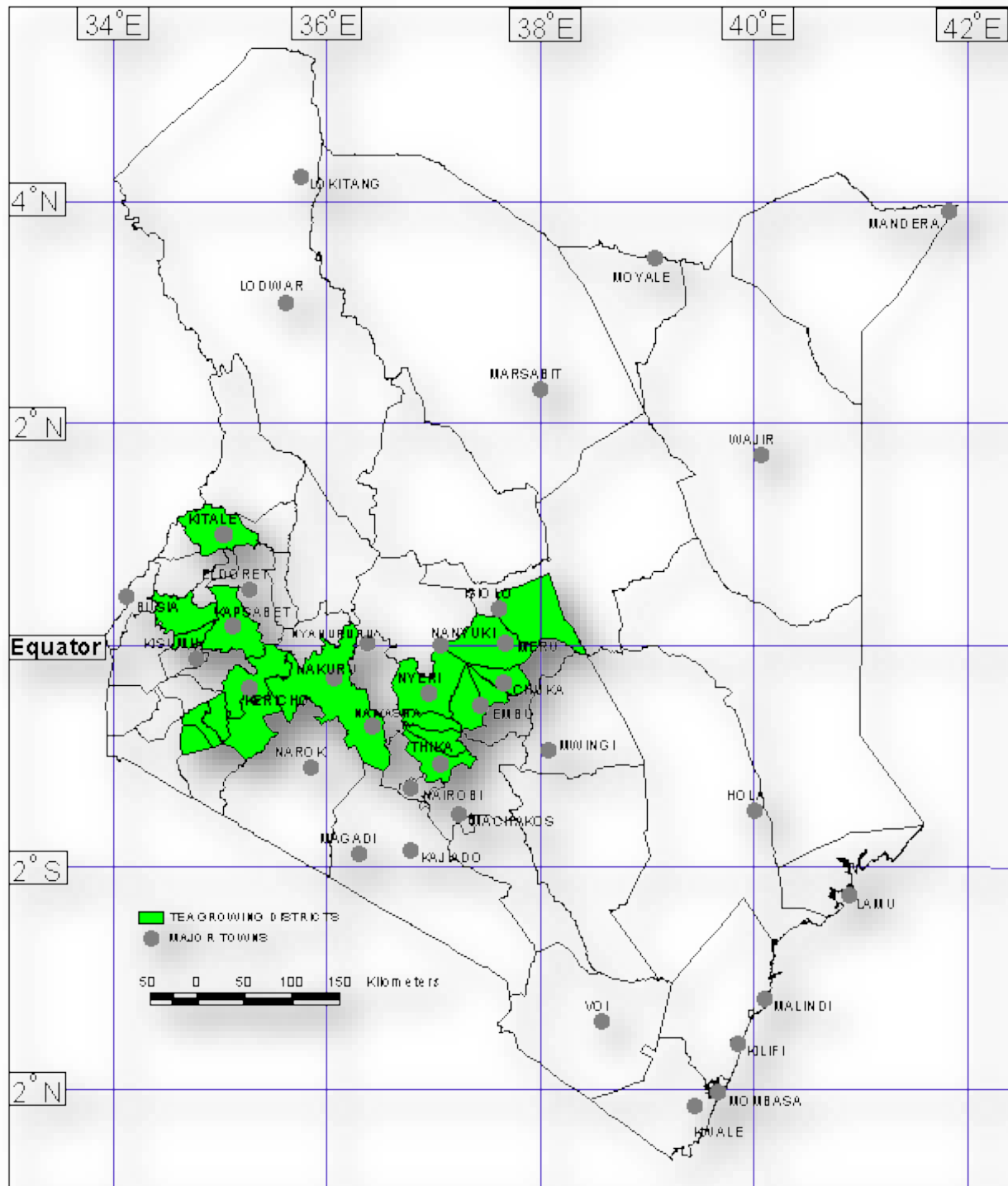


Figure 1.1: Map of Kenya showing tea growing areas (Source KTDA, 2006)

Kenya is ranked third in terms of annual volume production of black tea after China and India but has the highest productivity in yield per hectare as compared to China and India (TBK,

2012). Kenya produces 377 000 tonnes of tea per year (t/yr) which is equivalent to nine percent of world tea production (FAOSTAT, 2013). The majority of the tea produced (343 000 t/yr) is exported and represents approximately 20% of world exports. This therefore makes tea a top foreign exchange earner, and export commodity among agricultural produces with over 26% of all foreign exchange (Kenya National Bureau of Statistics, 2012). Over the last decade, the average price of black tea at the Mombasa auction in Kenya has increased by more than 50% from the 1 990-2 000 average (FAO, 2015a; FAO, 2015b). The rise has been attributed to the introduction of high-yielding plants that are of good quality, tolerant to drought, pests, and diseases (Wachira, 2002; Owuor and Obanda, 2007).

In 2013, the tea industry earned the country Kshs. 114 billion (US\$ 1.33 billion) in export earnings and Kshs. 22 billion (US\$ 0.26) in local sales (AFFA, 2013). But the average international tea price of processed black tea in the world tea market in the same year declined by 2.5% to 2.79 US\$ per kg and declined further by 5.3% in 2014 to 2.65 US\$ per kg (FAO, 2015a; FAO, 2015b). The decline in international tea prices in 2013 and 2014 was due to the weakening of black tea prices and the consistent surplus of black tea supply, which had the effect of depressing auction prices (FAO, 2015a). It was further underpinned by strong growth and demand for other tea products such as green and orthodox teas. Therefore, this calls for encouraged diversification of tea products into other segments such as organic and value-added teas rather than over-reliance on black tea.

Tea production also depends heavily on stable weather conditions, and the effects of climate change are alarming to tea industry stakeholders (FAO, 2015b). One of the major impacts of climate change is the occurrence of severe weather conditions, such as drought, which significantly influence the growth and production of tea (Ng'etich, 1995). Most of the tea growing regions in Kenya experiences a regular three-month drought period between December and March when tea yields drop by about 14 to 20% and climbing up to 30% during severe drought (FAO, 2015b). In recent years, drought affected production and earnings in some tea growing areas in most parts of Kenya. In 1997, 2000 and 2006, drought conditions affected most tea growing areas in Kenya and some factories had shut down or operate below their capacity due to inadequate supply of green tea leaves and the low revenue generated (FAO, 2015b). The current major areas of production in Kenya will become less suitable for tea production by the

year 2020, with many areas unlikely to be tea growing zones by the year 2050 (Cheserek *et al.*, 2015). Therefore, it is critically important for Kenya's tea industry to come up with measures to curb the challenges of climate change in the tea sector.

Tea breeding involves the creation of genetic variability through controlled hybridisation between selected progenitors and evaluation of the newly generated progeny. This is to identify promising individuals with desirable traits which are aimed at improving tea productivity. However, conventional tea breeding requires at least 20-25 years (Chen *et al.*, 2012a), from individual selection to local adaptability testing and to the final release of new tea cultivars to the farmers. Therefore, requires considerable investment in terms of land, labour, time and money. For these reasons, it is more efficient and effective to multiply early-stage populations and select breeding materials by the evaluation of DNA markers for important agronomic traits that are controlled by a few loci before the seedlings are planted in the field (Chen *et al.*, 2012a). The preferable method would be to perform marker-assisted selection (MAS) for many characteristics in many different populations which requires a reference map with many landmark markers. Although these significant advances have been made in tea breeding, to match and satisfy the diverse requirements of domestic and international markets, it is necessary to improve breeding efficiency so as to breed tea cultivars with desirable attributes.

Therefore, precise identification and screening of genetic resources, the development of molecular breeding technologies and the deep understanding of the genetics of important agronomic traits should be given more attention. Currently, the tea breeding program is primarily focused on high-yielding tea, better quality processed tea, genetic resistance to biotic and abiotic stresses prevailing in different tea growing regions and the regional adaptability. However, these agronomic traits in tea plant are quantitative in nature and not amenable to easy manipulations in breeding programs without elaborate and long-term field testing (Wachira, 2001; Kamunya *et al.*, 2004). The majority of developed and released tea cultivars are products of rigorously and highly protracted breeding and clonal selection (Kamunya *et al.*, 2010). Therefore, tea breeding improvement process would be more enhanced with the application of reliable and stable selection tools that are not prone to the continually changing environment.

1.2 Problem statement

Conventional approaches for genetic improvement in tea plant has been extremely difficult and time-consuming, especially for complex traits such as catechins, theaflavins, thearubigins and caffeine content, which are controlled by multiple genes known as QTLs. The application of QTL in mapping in tea plant over the years has been difficult because of its natural characteristics of self-incompatibility, low seed production and lack of available genetic markers. However, a few genetic maps have been reported recently for *C. sinensis* and these are mainly based on dominant markers such as RAPD or AFLP (Hackett *et al.*, 2000; Huang *et al.*, 2006a; Kamunya *et al.*, 2010; Hu *et al.*, 2013) and on co-dominant markers such as SSRs (Ma *et al.*, 2014; Tan *et al.*, 2016). Recently, first SNP-based reference map of tea plant, as well as the most saturated one has been developed (Ma *et al.*, 2015). The use of QTL for important agronomic traits and validation of marker-trait associations can facilitate selection of tea plants with desirable agronomic traits at the early-stage of growth. This is particularly for woody plant such as tea because conventional breeding based on phenotypic selection is often delayed due to its long juvenile stage of about 22-25 years.

Tea being universally the most popular beverage, the need to develop tea cultivars with optimum potential for black or green tea quality has recently become the single most important breeding objective. Currently, Kenya produces cut, tear and curl (CTC) black tea as the only product, for which the prices have declined owing to a global glut. This calls for concerted action, the chief being the diversification and added value of tea products, not only to help Kenya maintain its position as a leading exporter but also to enhance foreign exchange earnings. The catechins, theaflavins (TF), thearubigins (TR) and caffeine though being major determinants of tea quality and essential components for high-quality, diversified tea products, they have not yet been subjected to QTL analysis and the genetic basis of this phenotypes remains poorly understood. The main objective of this study is to determine, identify and map out molecular markers linked to important QTLs for tea quality parameters (catechins, TF, TR and caffeine content) in tea cultivars grown in Kenya.

1.3 Research justification

Genetic maps are essential tools for implementing QTLs analysis and MAS breeding. QTL is a powerful tool for elucidating the molecular basis of complex traits facilitating the identification

of specific genes responsible for particular genetic trait, estimation of gene actions and genetic parameters. The modern molecular techniques based on DNA markers, genetic or linkage maps, and genomic sequencing has enabled breeders to understand genetic structures of agronomic traits and improve crops rapidly and precisely. The integration of these molecular or genetic markers in breeding programmes has been a powerful tool for enhancing the efficiency and effectiveness of identification of QTLs linked to important agronomic traits. The ability to assess complex phenotypes such as quality in tea at the seedling stage using genetic markers would greatly accelerate new cultivar development.

The identification of putative QTLs tightly linked to markers in this study, augmented by known genetic parameters, will demonstrate the great potential of integrating molecular markers in tea breeding. Also, the selection of advantageous traits and markers linked to complex traits could be used to select against negative characteristics and could even be used to select the combination of parents that would give rise to progeny with the desired genotype. In a recently conducted study at TRI, QTL mapping using molecular and phenotypic traits was measured on a pseudo-testcross comprising 42 clonal progeny (TRFK Stock 463) (Kamunya *et al.*, 2009; Kamunya *et al.*, 2010). Over 250 RAPD, 15 SSR, 12 ISSR, and 96 AFLP primers were screened using the bulked segregation analysis technique to identify possible markers associated with quality, yield and stress-related traits in Kenyan tea. The use of QTL for important agronomic traits and validation of marker-trait associations can facilitate selection of tea plants with desired features at early stages of growth. The elucidation of important quantitative traits such as drought tolerance and biochemical components contributing to tea quality properties and its genetic background would be most useful in guiding tea breeding. Such knowledge is currently lacking though the previous studies have attempted to generate the requisite background information.

1.4 Research objective

The main research objective of this study was to identify the physiological, molecular and biochemical characteristics associated with black tea quality traits and drought tolerance trait.

Firstly, this study collected phenotypic data associated with black tea quality traits and drought tolerance trait in two tea segregating populations TRFK St. 504 and TRFK St. 524, which are reciprocal crosses between (TRFK 303/577 × GW Ejulu) and (GW Ejulu × TRFK 303/577), respectively.

Secondly, this study explored the use of novel DArTseq marker technique with an objective of identifying and annotating functions to putative QTLs linked to potentially useful traits associated with black tea quality and %RWC for future MAS breeding in tea.

Thirdly, this study combined linkage mapping and association mapping to precisely authenticate putative QTLs associated with black tea quality and %RWC for future MAS breeding in tea.

Lastly, this study employed machine learning methodologies to build a prediction model that can be used to predict phenotypic traits associated with black tea quality and drought tolerance from the genomic data.

The thesis is organised into the following chapters:

Chapter 2: Literature review

Chapter 3: Quantitative genetic parameters for black tea quality and drought tolerance traits

The selection for high-quality and drought-tolerant tea cultivars requires precise information on careful study of the biochemical molecules which contribute towards black tea quality and drought tolerance. Also, molecular markers can facilitate the indirect selection of high-quality and drought tolerance tea cultivars by identifying genomic regions associated with the phenotypic traits for marker-assisted selection.

Chapter 4: Functional annotation of putative QTLs associated with black tea quality and drought tolerance traits

The information on the genomic and functional annotation of putative QTLs involved in black tea quality and drought tolerance traits in tea based on DArTseq markers will shed more light on the proteins/ enzymes associated with black tea quality and drought tolerance traits.

Chapter 5: Combined linkage and association mapping of putative QTLs controlling black tea quality and drought tolerance traits

A combination of linkage mapping and association mapping was used to provide mutual and precise authentication of QTLs which will enable more reliable results to be obtained.

Chapter 6: Application of machine learning methodologies in genomic selection of black tea quality and drought tolerance traits

The application of machine learning methodologies in plant breeding is capable of linking genomic markers to phenotypic traits using predictive models. This will facilitate identification of genomic regions that are of agronomic value and lead to functional annotation of genomes which will enable real-time high-throughput phenotyping of agronomic traits in tea within a short-time.

Chapter 7: Concluding discussion and recommendation

1.5 Research outputs

- 1. Peer-reviewed publication (Appendix 3.3):** 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.
- 2. Peer-reviewed publication (Appendix 4.14):** Koech, R.K., Malebe, P.M., Nyarukowa, C., Mose, R., Kamunya, S.M., Joubert, F. & Apostolides, Z. (2019). Functional annotation of putative QTL associated with black tea quality and drought tolerance traits. *Scientific Reports* 9(1): 1465.
- 3. Conference proceeding:** Koech, R. K., Malebe, P. M., Nyarukowa, C., Mose, R., Kamunya, S. M. & Apostolides, Z. (July 2018). Novel QTLs for Black Tea Quality Traits and Drought Tolerance in Tea Plants (*Camellia sinensis*). 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.
- 4. Conference proceeding:** Koech, R. K., Malebe, P. M., Nyarukowa, C., Mose, R., Kamunya, S. M. & Apostolides, Z. (June 2017). Detection of Novel QTL for Metabolomic and Agronomic traits in Tea (*Camellia sinensis*). Invited oral presentation. The 13th Annual Conference of the Metabolomics Society, Brisbane, Australia.
- 5. Conference proceeding:** Koech, R. K., Malebe, P. M., Nyarukowa, C., Mose, R., Kamunya, S. M. & Apostolides, Z. (August 2016). Identification of novel QTL for tea quality traits in tea plant. Invited poster presentation. The Southern African Biochemistry and Informatics for Natural Products (SABINA) conference, Mpumalanga, South Africa.

References

- AFFA (2013). Tea market development in Kenya.
- Anon (2003). Kenya Tea Development Authority Annual Report. Nairobi.
- Chen, L., Apostolides, Z. & Chen, Z. M. (2012a). *Global tea breeding: achievements, challenges and perspectives*. Springer Science & Business Media.
- 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.
- FAO (2015a). World tea production and trade current and future development. 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 (2015b). Kenya's tea sector under climate change: An impact assessment and formulation of a climate smart strategy. (Eds A. Elbehri, B. Cheserek, A. Azapagic, D. Raes, M. Mwale, J. Nyengena, P. Kiprono, and C. Ambasa). Rome, Italy.
- FAOSTAT, D. (2013). Food and agriculture organisation of the United Nations. Statistical database.
- 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): 346-355.
- Hu, C. Y., Lee, T. C., Tsai, H. T., Tsai, Y. Z. & Lin, S. F. (2013). Construction of an integrated genetic map based on maternal and paternal lineages of tea (*Camellia sinensis*). *Euphytica* 191(1): 141-152.
- Huang, F., Liang, Y., Lu, J. & Chen, R. (2006). Genetic mapping of first generation of backcross in tea by RAPD and ISSR markers. *Journal of Tea Science* 26(3): 171-176.
- Kamunya, S., Wachira, F. & Muoki, R. (2004). Evaluation of newly developed clones of tea (*Camellia sinensis* (L.) O. Kuntze) for yields, drought tolerance and quality: preliminary indicators. *Tea* 25(1): 12-19.
- 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.
- Kenya National Bureau of Statistics (2012). Kenya facts and figures 2012. Kenya.
- 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.
- Ma, J.Q., Huang, L., Ma, C.L., Jin, J.Q., Li, C.F., Wang, R.K., Zheng, H.K., Yao, M.Z. & Chen, L. (2015). Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). *PloS one* 10(6): e0128798.
- Matheson, J. & Bovill, E. (1950). East African Agriculture. *East African Agriculture*.
- Ng'etich, W. K. (1995). An Evaluation of the responses of tea clones to environment in Kenya., Vol. PhD thesis Silsoe, Bedford: Cranfield University.
- Owuor, P. O. & Obanda, M. (2007). The use of green tea (*Camellia sinensis*) leaf flavan-3-ol composition in predicting plain black tea quality potential. *Food Chemistry* 100(3): 873-884.
- 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 colour, and mature leaf size in tea plant (*Camellia sinensis*). *Tree Genetics & Genomes* 12(3): 1-13.
- TBK (2012). Tea industry performance in 2012. The Tea Board of Kenya.
- TBK (2013). Tea industry performance in 2007. The Tea Board of Kenya.
- TRFK (2013). Annual Technical Report.: Tea Research Foundation of Kenya (TRFK), Tea Board of Kenya.
- Wachira, F. (2002). Detection of genetic diversity and characterisation of Kenyan tea germplasm. *A Tea Genetic Diversity (TGD) Project. TGD Final Project Document, Ed. Tea Research Foundation of Kenya, Kericho, Kenya.*

Wachira, F. N. (2001). Tea improvement in Kenya: An overview of research achievements, prospects and limitations. In *TBK board of directors open day proceedings, 29 January 2001.*, 12-14 Tea Research Foundation of Kenya.

Wilson, S. & Co. (1962). Tea estates in Africa. 7-9 London: Mabey and Filzclarence Ltd.

CHAPTER 2

LITERATURE REVIEW

2.1 The tea plant

2.1.1 Botany of tea

Tea is believed to have originated in Southeast Asia around the intersection of 29 °N latitude and longitude 98 °E (Mondal, 2007). The main cultivated taxa comprise of three natural hybrids. The China type (original *Camellia sinensis* var. *sinensis*) which has small semi-erect leaves and originated from Sichuan Province, south-western China (Van der Vossen and Wessel, 2000). The Assam type (considered a distinct species, *Camellia sinensis* var. *assamica*) which has relatively large, horizontal leaves and is believed to have originated from the forests of Assam in north-eastern India (Van der Vossen and Wessel, 2000). The Cambod type (considered subspecies, *C. assamica* subsp. *lasiocalyx*) which is now considered as a hybrid of *Camellia sinensis* var. *sinensis* and *Camellia sinensis* var. *assamica* (Wambulwa *et al.*, 2016). Hybridisation in tea breeding has been so extreme that it is often debated if archetypal *C. sinensis*, *C. assamica*, or *C. assamica* ssp. *lasiocalyx* still exist (Visser, 1969). *Camellia* taxa hybridise naturally and show a high degree of allogamy (Banerjee, 1992). Studies on tea have reported that at least the three taxa and to some extent, *C. irrawadiensis*, have significantly contributed to the tea gene pool (Banerjee, 1992). Due to this extensive hybridisation, several intergrades, introgressants and putative hybrids have been formed (Mondal, 2002). However, the numerous hybrids currently available are still referred to as China, Assam or Cambod tea depending on morphological proximity to the main taxon (Banerjee, 1992). Tea can be cultivated in regions that have high humidity, fair temperatures and acidic soils, from sea level to high mountains (Willson and Clifford, 2012). The perennial tree crop is grown in many parts of the world for the production of different tea products.

2.1.2 The chemical composition of tea flush

2.1.2.1 Tea flavonoids

Flavonoids are the most widely occurring phytochemicals in many plants and are grouped into several subclasses, including catechins, anthocyanins, flavones, isoflavones, flavonols, and chalcones. Flavonoids compounds play a critical role in growth, development and defence against microorganisms and pests in plants (Lattanzio *et al.*, 2006). In tea, flavonoids are majorly

known to contribute towards teacup quality. These phenolic compounds are water-soluble, most often combined with sugar as glycosides and are located in the vacuole. The natural flavonoids are low molecular weight compounds, consisting of 15 carbon atoms, arranged in a C₆-C₃-C₆ configuration. The basic structure of flavonoids consists of two aromatic rings A and B, joined by a 3-carbon bridge, usually in the form of a heterocyclic ring, C (Figure 2.1).

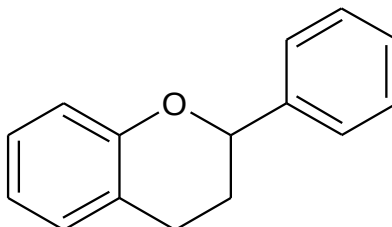
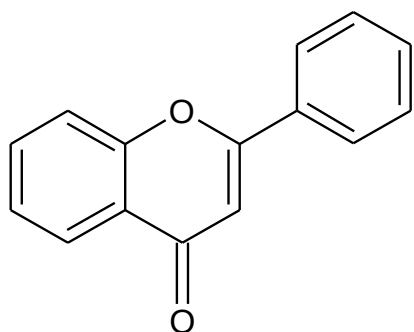
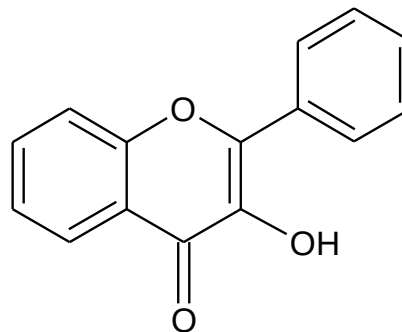


Figure 2.1: Chemical structure of a flavonoid molecule

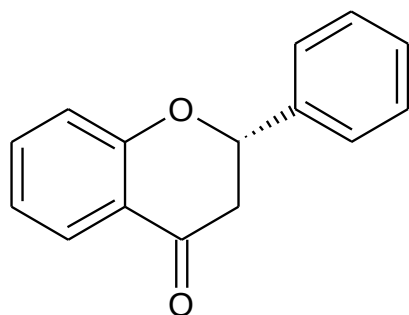
The aromatic ring A is derived from the acetate/malonate pathway, while ring B is derived from phenylalanine through the shikimate pathway. Substitutions to ring A and B give rise to the different compounds within each class of flavonoids. These substitutions may include oxygenation, alkylation, glycosylation, acylation, and sulfation. The variations in substitution patterns to ring C result in major flavonoid classes, including flavones, flavonols, flavanones, flavanonols, flavanols (catechins), isoflavones, and anthocyanidins (Figure 2.2). Purple coloured-leaf tea cultivars have recently been developed in Kenya (Kamunya *et al.*, 2009). The leaves from these cultivars are reported to contain anthocyanins and anthocyanidins (Figure 2.2), with the predominant anthocyanidin being malvidin (Kerio *et al.*, 2013).



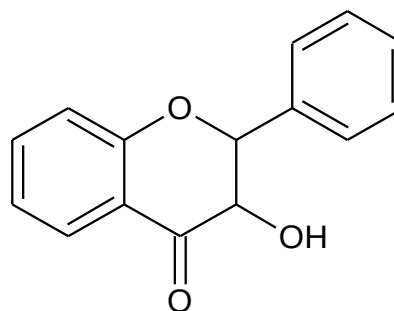
Flavones



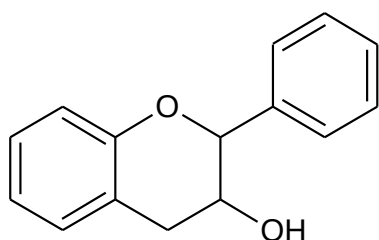
Flavonols



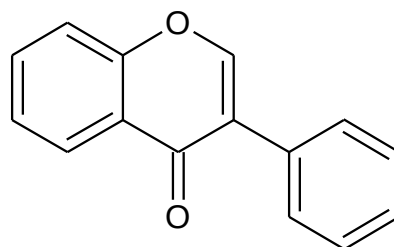
Flavanones



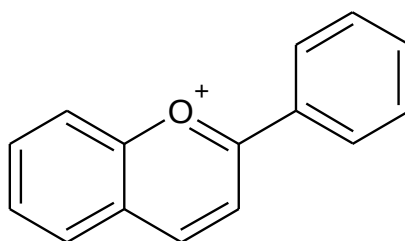
Flavanonols



Flavanols (or catechins)



Isoflavones



Anthocyanidins

Figure 2.2: Chemical structure of major classes of flavonoids

2.1.2.2 Tea Polyphenols

Polyphenols are compounds with bonded benzene rings having one or multiple hydroxyl groups and range from simple phenolic molecules to highly polymerised compounds (Balasundram *et al.*, 2006). Phenolic compounds are derived from the condensation reaction of cinnamic acid with three malonyl-CoA groups. Most naturally-occurring phenolic compounds are present as conjugates with mono- and polysaccharides, linked to one or more of the phenolic groups

(Balasundram *et al.*, 2006). Despite this structural diversity, these wide ranges of phenolic compounds that occur in nature are still referred to as polyphenols. These polyphenols can be categorised into several basic classes, which include: phenolic acids, flavonoids and tannins (Bravo, 1998). Tannins are large polyphenol molecules and form the bulk of the active compound in tea with catechins comprising nearly 90%.

2.1.2.3 Flavonoids biosynthesis

Flavonoids are secondary plant metabolites that are derivatives of shikimate, phenylpropanoid and pentose phosphate pathways (Figure 2.3). The shikimate pathway is responsible for the synthesis of phenylpropanoid which is connected to carbohydrate metabolism through the pentose phosphate pathway. The phenylpropanoid pathway starts with phenylalanine and ends with 4-coumaroyl-Coenzyme A (-CoA) (Weisshaar and Jenkins, 1998). The enzyme phenylalanine ammonia lyase (PAL) catalyses the deamination of phenylalanine to cinnamic acid, which is further modified by hydroxylase and O-methyltransferases (Vankatesh *et al.*, 2007). The enzyme 4-Coumaroyl CoA ligase (4CL) catalyses the formation of Co A esters of cinnamic acid, and these activated intermediates are used in the biosynthesis of diverse compounds via specific branch pathways, such as those leading to either flavonoid or lignin biosynthesis (Winkel-Shirley, 2001). The shikimate pathway is also responsible for formation of gallic acid, which will presumably be attached to the flavan-3-ols with an ester bond in the last stages of the biosynthesis process of the tea flavan-3-ols.

2.1.2.4 Caffeine biosynthesis

Caffeine biosynthetic pathway (Figure 2.4) in tea plant involves four main step sequence consisting of three methylation and one nucleosidase reactions (Mohanpuria *et al.*, 2010; Ashihara *et al.*, 2011). The initial step is the methylation of xanthosine by 7-methylxanthosine synthase (xanthosine 7N-methyltransferase, EC 2.1.1.158). Xanthosine is supplied by four different pathways: de novo purine biosynthesis (de novo route), the degradation pathways of adenine nucleotides (AMP route) and guanine nucleotides (GMP route), and the S-adenosyl-L-methionine (SAM) cycle (SAM route). The second step of caffeine biosynthesis involves N-methylnucleosidase (EC 3.2.2.25) which catalyses the hydrolysis of 7-methylxanthosine. The last two steps of caffeine synthesis are also catalysed by SAM-dependent N-methyltransferase(s), but this enzyme is different from the N-methyltransferase that catalyses the first step in the pathway.

Caffeine synthase (EC 2.1.1.160) enzyme catalyses the last two steps of caffeine biosynthesis, in that, it converts 7- methylxanthine to caffeine via theobromine.

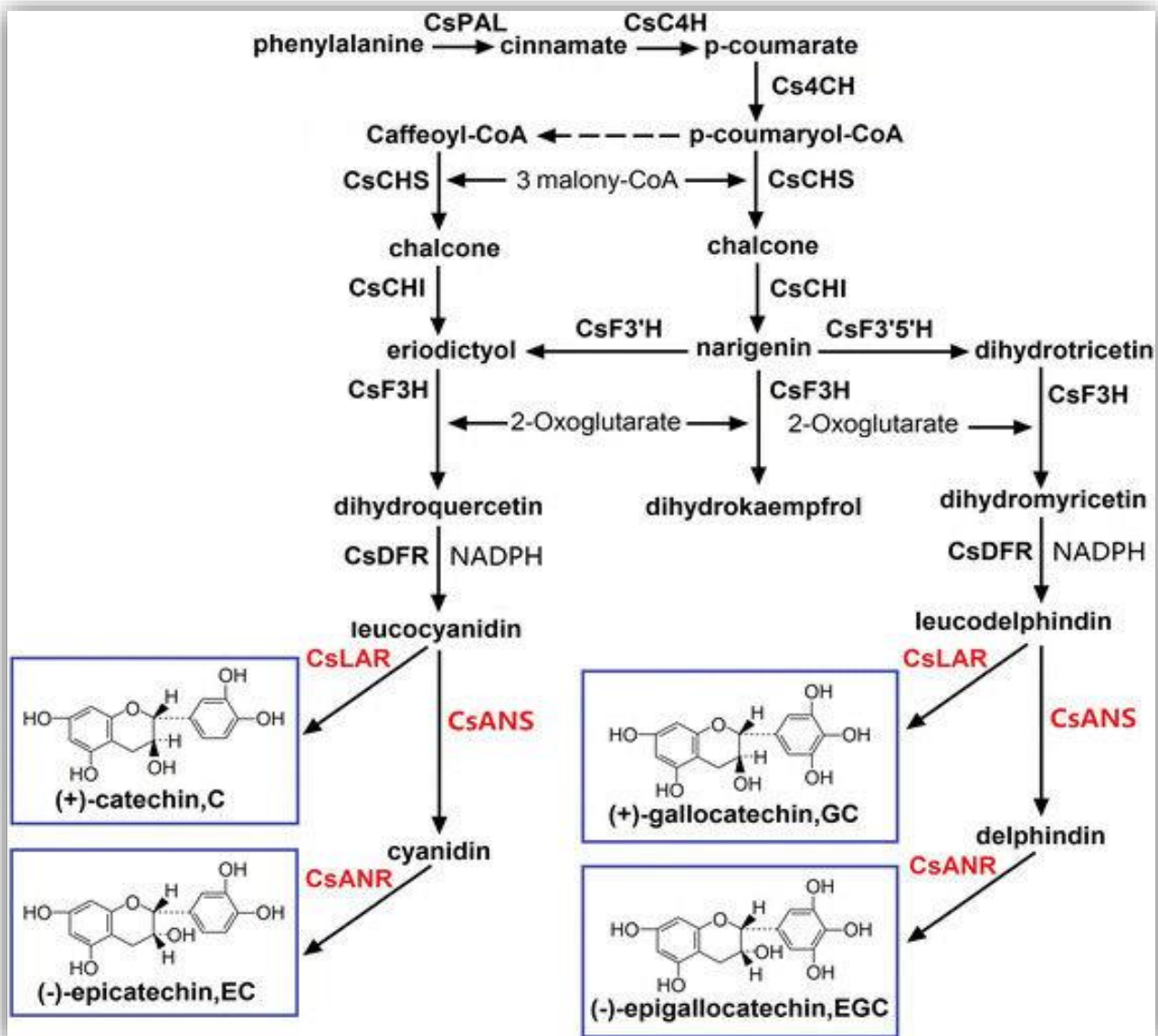


Figure 2.3: A schematic diagram of flavonoid biosynthesis pathway.

Abbreviations: PAL, phenylalanine ammonia lyase; C4H, cinnamic acid 4-hydroxylase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H flavanone 3-hydroxylase; F3'H, flavonoid 3'-hydroxylase; F3'5'H, flavonoid 3',5' -hydroxylase; DFR, dihydroflavanol 4-reductase; LAR, leucoanthocyanidin reductase; ANS, anthocyanin synthase; ANR, anthocyanidin reductase; Source: (Wu *et al.*, 2014).

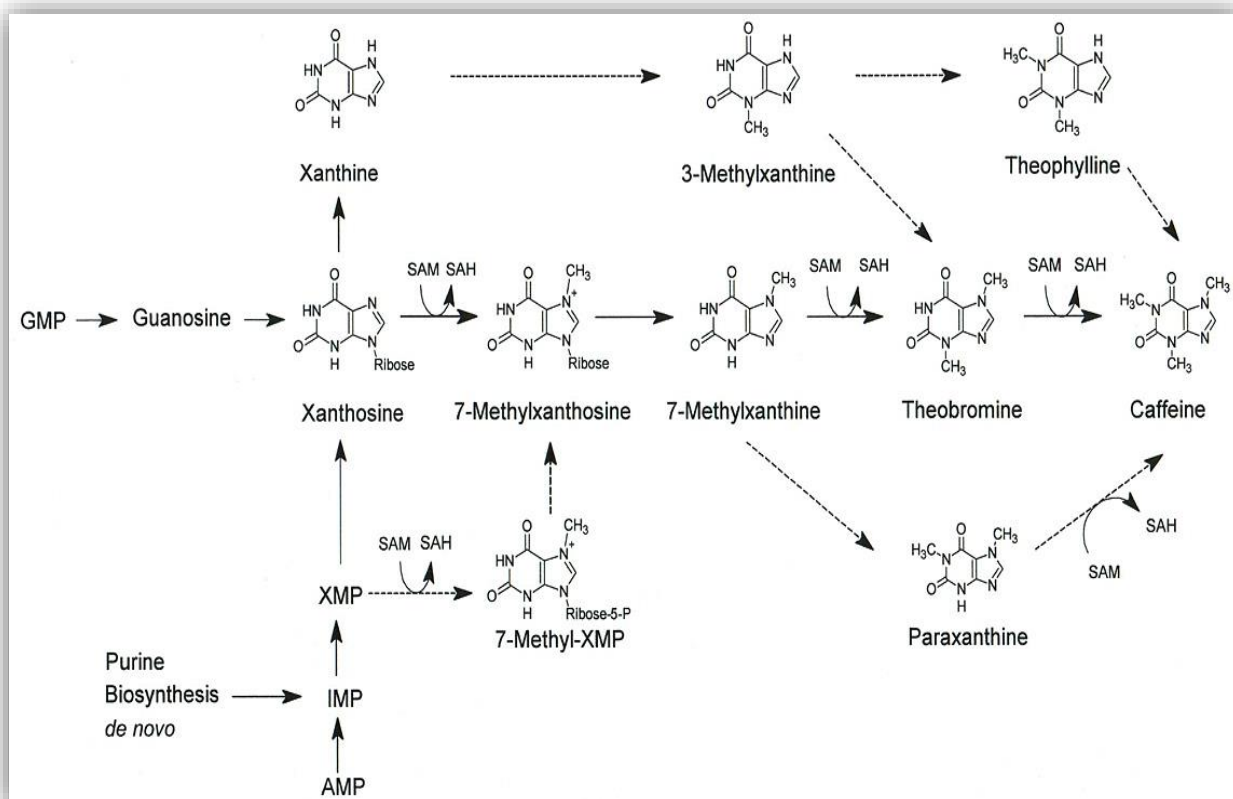


Figure 2.4: A schematic diagram of caffeine biosynthesis pathway.

Abbreviations: GMP, guanosine 5'-monophosphate; XMP, xanthosine 5'-monophosphate; IMP, inosine 5'-monophosphate; AMP, adenosine 5' monophosphate; SAM, S-adenosyl-L-methionine; SAH, S-adenosyl-L-homocysteine. Source: (Anaya *et al.*, 2006).

2.1.2.5 Chemical composition of made tea

The major constituents in tea extracts are catechins, a group of flavonoids, which contribute up to 30% of the dry weight of the tender tea shoots (Pang *et al.*, 2013). Catechins (Figure 2.5) in green tea leaves consist mainly of four flavan-3-ols, which include: (-)-epicatechin (EC), (-)-epicatechin gallate (ECG), (-)-epigallocatechin (EGC) and (-)-epigallocatechin gallate (EGCG) (Ananingsih *et al.*, 2013). Epigallocatechin gallate is the most abundant accounting up to 50-80% of the total catechins (Sang *et al.*, 2011). Theaflavins consist mainly of four major compounds (Figure 2.6), which are termed as theaflavin (TF1), theaflavin-3- gallate (TF2), theaflavin-3'-gallate (TF3) and theaflavin-3, 3'-digallate (TF4). The individual theaflavins and thearubigins (Figure 2.7) which are oxidation products of catechins are the major polyphenolic content

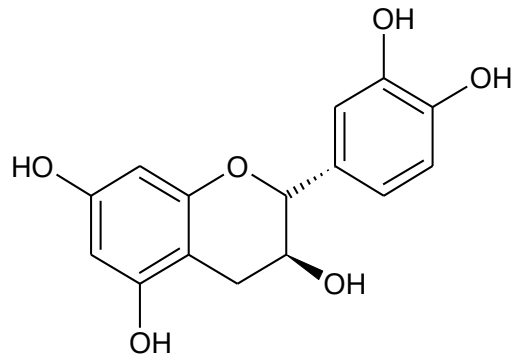
present in black tea (Sang *et al.*, 2011). Besides the catechins, tea also contains kaempferol, quercetin and myricetins glycosides in small amounts (Balentine *et al.*, 1997; Chaturvedula and Prakash, 2011). Caffeine, theobromine, and theophylline (Figure 2.8) are the main alkaloids in tea while phenolic acids and amino acids are present but in small quantities (Balentine *et al.*, 1997). The chemical composition of tea shoots and the reactions that occur during processing are the factors that determine the nature and the quality of the finished tea product.

2.1.3 Major tea products

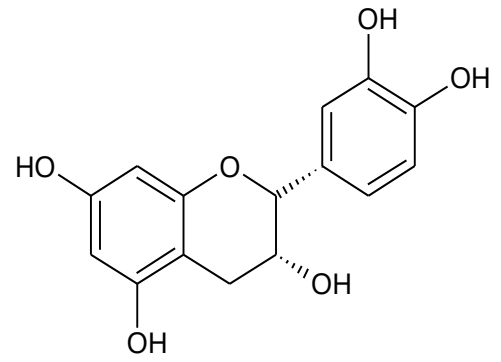
Processed tea is one of the most popular beverages manufactured from the young tender leaves of the plant *C. sinensis* (Cabrera *et al.*, 2003). Tea products can be categorised into six different types of products which include green tea, black tea, oolong tea, white tea, yellow tea, and dark tea according to their different processing procedures. Of the total amount of tea produced and consumed in the world, 78% is black, 20% is green, and less than 2% is oolong tea (Guo, 2013). Though, most of the tea produced in the world can be classified as non-aerated green tea, semi-aerated (oolong) tea and aerated black tea (Tounekti *et al.*, 2013), processing has been diversified into the production of speciality teas such as white, flavoured, organic, decaffeinated, herbal and scented teas. It is the differences between the various processes of tea manufacture that result in differences in the chemical and polyphenol profile between green and oolong black tea.

2.1.3.1 Green tea

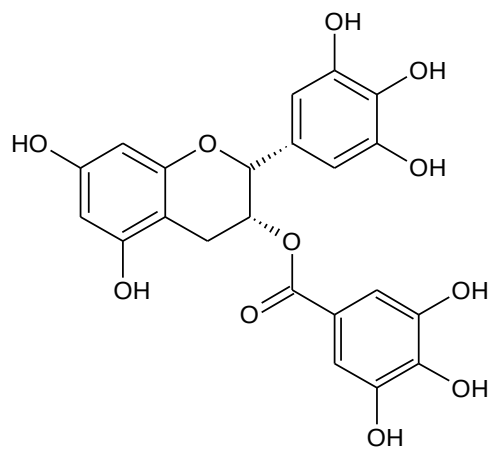
Green tea is mainly consumed in China, Korea, Japan, and a few countries in North Africa and the Middle East. In green tea processing, the leaves are steamed after plucking and then dried relatively rapidly to minimise chemical and enzymatic reactions. Steaming stops the polyphenol oxidase (PPO) enzyme (EC 1.10.31) catalysed oxidation of tea leaf catechins (Willson and Clifford, 2012). Green teas are subject to minimal oxidation to preserve the healthy, natural, and active substances of the fresh tea leaves so that they could be released into the consumer's cup upon infusion.



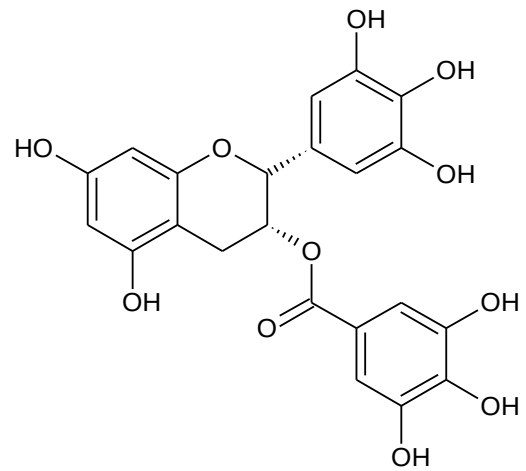
Epicatechin (EC)



Epicatechin gallate (ECG)

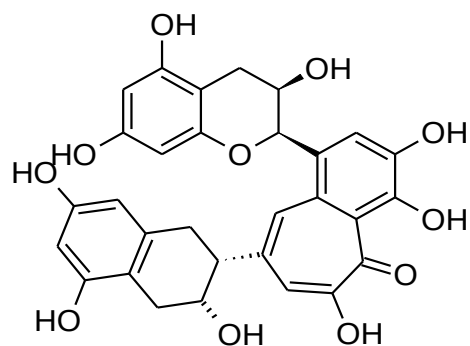


Epigallocatechin (EGC)

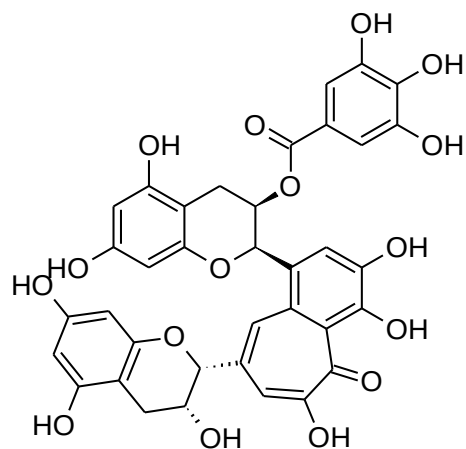


Epigallocatechin gallate (EGCG)

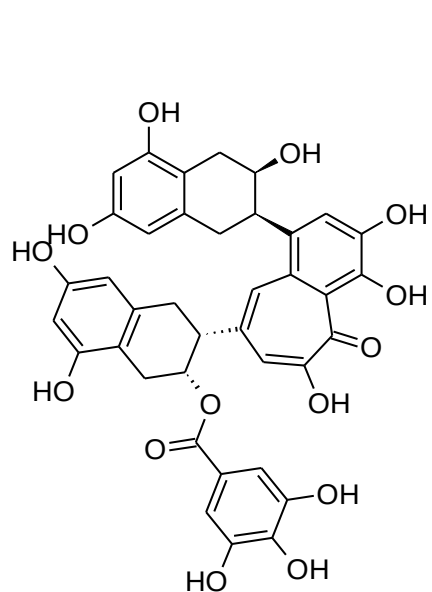
Figure 2.5: Chemical structures of major catechins



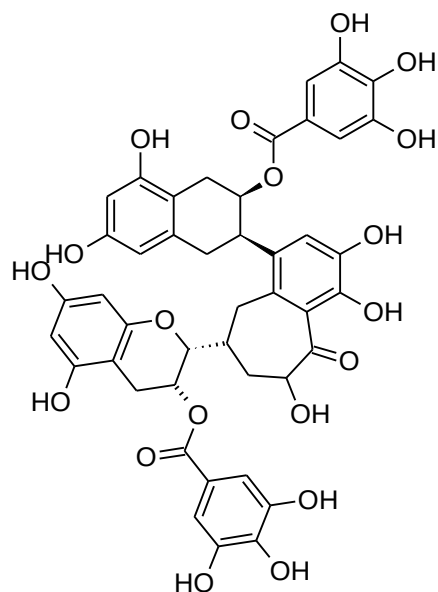
Theaflavin (TF1)



Theaflavin-3-gallate (TF2)



Theaflavin-3'-gallate (TF3)



Theaflavin-3, 3'-digallate (TF4)

Figure 2.6: Chemical structure of individual theaflavins

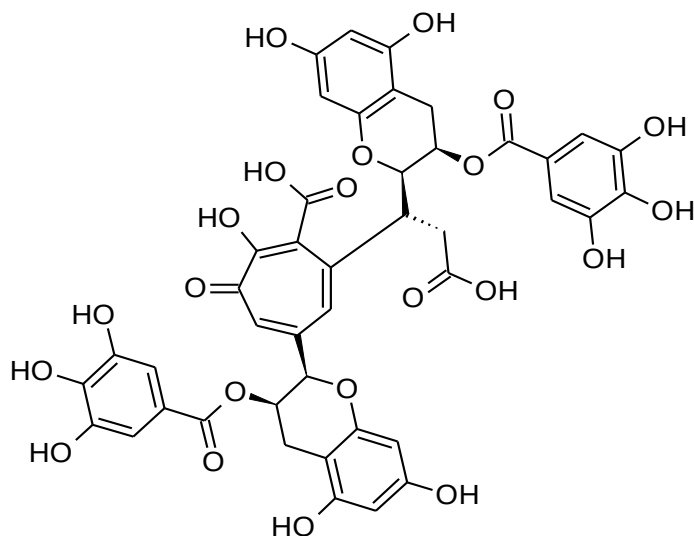
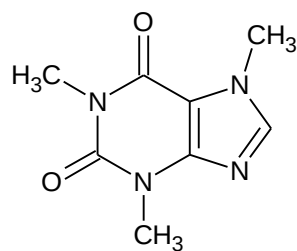
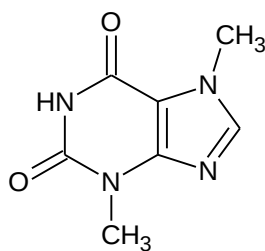


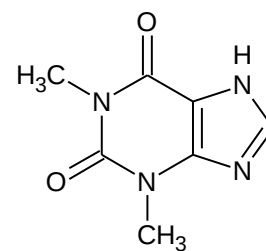
Figure 2.7: Chemical structure of thearubigins



Caffeine



Theobromine



Theophylline

Figure 2.8: Chemical structures of methylxanthines

2.1.3.2 Oolong tea

Oolong tea is a semi-fermented tea mainly consumed in Asia, North Africa, and the United States. Oolong tea is processed from young green tea shoots which are freshly harvested early in the morning. The leaves are then allowed to wither under the sunlight for a few hours prior to the semi-fermentation process, in which tea leaves are oxidised, pan fired at $\sim 200^{\circ}\text{C}$, rolled to form a ball shape, and dried in a specialised oven at various desired temperatures. Experts empirically control the reaction time for the contact between phenolic compounds and oxidative enzymes during this semi-fermentation process, and the final fermentation degree of oolong tea ranges

from 20 to 80%, depending upon the demand of customers. After long-term storage, oolong tea tends to absorb substantial moisture from the air and thus needs to be refined by periodical drying. However, the longer oolong tea is stored and further oxidised gradually, the better it is in terms of taste and beneficial effects on human health (Chen *et al.*, 2013).

2.1.3.3 Black tea

Black tea is consumed primarily in Western countries and in some Asian countries like India, Sri-Lanka, and regions of Africa. Kenya specialises in black tea production and processing and is the third largest producer of black tea after China and India (ITC, 2009). In contrast to green tea, the general steps of processing black tea are; withering of leaves under certain plucking standards, rolling, aerating, drying and sorting (Figure 2.9).

2.1.4 Major steps in black tea processing

2.1.4.1 Withering

Withering is the first step in black tea processing with the aim of reducing the moisture content of the fresh-plucked tea leaves, making them amenable to subsequent processing steps (Figure 2.9). Also, diverse and important biochemical changes occur, which involve changes in proteins, caffeine, sugars, organic acids, polyphenol oxidase activity, chlorophyll, minerals, volatile components, and permeability of cell membranes (Dev and Bajaj, 1980). Withering can be divided into physical and chemical wither because of the physical and chemical changes, respectively. The physical withering is associated with loss of moisture from the shoot which leads to changes in cell membrane permeability. The chemical wither involves the breakdown of proteins into amino acids and other chemical changes that take place. The short chemical withering period has been shown to favour the formation of theaflavins and thus improves the liquor brightness (Obanda *et al.*, 2004). Combined physical and chemical withering produces the highest quality black tea because it results in higher theaflavin content (Owuor *et al.*, 1986b).

The withering process is also important for the formation of volatile compounds during black tea fermentation/ aeration (Takeo, 1984). Thus, appropriate withering is critical to ensure good quality black tea (Tomlins and Mashingaidze, 1997). During withering, the tea leaves are spread evenly in perforated troughs fitted with a powerful exhaust fan underneath to draw the moisture from the tea leaves and carry the humid air out of the withering area. Alternatively, during humid conditions, hot air is blown from underneath the trough through the tea leaves to remove the

moisture. The humid conditions during withering process favour the formation of theaflavins in black tea processing (Obanda *et al.*, 1997). After the withering process, the tea leaves are ready for rolling process.

2.1.4.2 Rolling

During rolling, the leaves are macerated, and the cell structures are disrupted, which brings various enzymes into contact with the polyphenols. The rolling of tea leaves can be achieved by using orthodox rollers such as rotor vanes, Lawrie's tea processor, or CTC machines. Sri-Lanka is the major world producer of orthodox tea while Kenya and India specialise in black CTC tea production (Chen *et al.*, 2012a). There is a major difference between orthodox and black tea in terms of theaflavin content. In general, orthodox teas have low theaflavin contents while CTC teas have high contents of theaflavins and thearubigins (Owuor *et al.*, 1986a). The difference in theaflavin and thearubigin contents is due to the difference in these two processing conditions. A high theaflavin content and high flavour index is associated with the general good quality of tea. However, orthodox process is the method of choice for maximum realisation of the innate potential quality of tea leaves. This explains the high amount of volatile constituents formed during processing of orthodox tea compared to black CTC tea. Other variables such as climatic factors in the tea growing regions and the growth rates of tea bushes have also been found to affect the amounts of theaflavins in the resulting black tea (Owuor *et al.*, 1986a). The chemical and biochemical reactions initiated in the fresh tea leaves during withering and rolling processes are accelerated during aeration or fermentation process.

2.1.4.3 Aeration (Fermentation)

In contrast to green tea, black tea processing undergoes aeration commonly referred to as fermentation, which involves the oxidation of catechins and catechin gallates by polyphenol oxidase (PPO) enzyme (EC 1.10.31), other peroxidase enzymes and non-enzymatic reactions to form the unique character of black tea. There are two important reactions during fermentation/aeration. First, there is the development of colour, strength, and quality of tea brews from the production of non-volatile compounds through the enzymatic oxidation of catechins and their gallates. Secondly, there is production of volatile compounds responsible for the characteristic aroma of black tea.

During aeration, catechins condense with the ortho-quinones arising from the oxidation of the B-ring di- and trihydroxylated catechins to form theaflavins. Previous studies were undertaken to evaluate the quality of black tea by chemical means, have revealed theaflavins and thearubigins to be major components that are largely responsible for briskness, brightness, strength and colour of black tea (Owuor *et al.*, 1986a; Owuor *et al.*, 2006a). Theaflavins which are homogeneous substances, give a yellow colouration in aerated black tea and contribute to the briskness and brightness of the tea liquor (Obanda *et al.*, 2001). Thearubigins are heterogeneous substances responsible for the colour and taste of tea, formed from epitheaflavic acid upon oxidation of gallic acid by theaflavins. The high ratio of simple catechins (EC and EGC) are known to facilitate a high ratio of thearubigins formation from theaflavins. Volatile compounds responsible for characteristic aroma of black tea are also produced during aeration. The larger the flavour index for these volatile compounds, the more the superior the tea flavour (Owuor *et al.*, 1986a).

The rate of aeration has been shown to be profoundly influenced by genetic constitution, seasonal and climatic factors, agronomic and management practices, and tea processing techniques (Cloughley, 1980). Studies on clonal tea revealed that long fermentation under warm conditions enhanced the formation of these flavonoids leading to increased brightness and total colour of black tea liquors (Obanda and Owuor, 1993). However, aeration under warm conditions for too long usually results in more intensely coloured black teas with high contents of thearubigins and low levels of theaflavins (Obanda *et al.*, 2001). The levels of theaflavins in black tea also increase quadratically with aeration time, while the rate of production of individual theaflavins varies between clones as fermentation progresses (Owuor and Obanda, 1993). The estimation of theaflavins has been used to give useful information regarding the optimum aeration time during black tea processing, which is also an indicator of quality and value of plain teas.

2.1.4.4 Drying

Drying or firing of aerated tea leaves (dhool) is primarily intended to cause cessation of enzyme activity and reduce the moisture content to about 3% of the dry mass so that the tea can be stored. Fluidised bed dryer has been used mostly in the final firing of tea which not only reduce the moisture content but also affect levels of volatile compounds, amino acids, carboxylic acids

and Maillard reactions (Hara *et al.*, 1995). Firing at fairly high temperature is necessary for the development of the colour, taste and aroma of black tea.

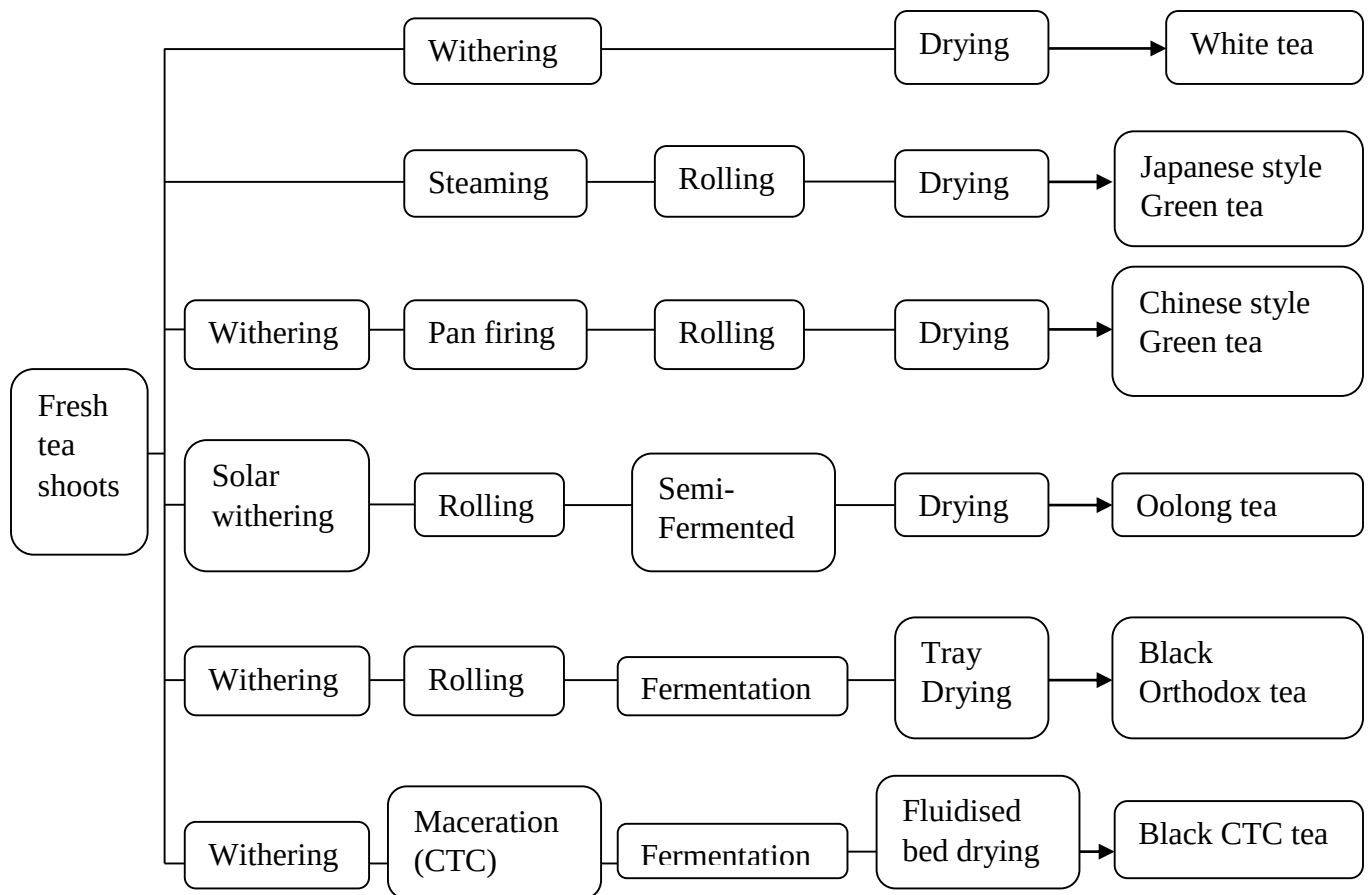


Figure 2.9: A schematic diagram of major steps in tea manufacture and corresponding types of tea. Source: (Hilal and Engelhardt, 2007).

2.1.4.5 Sorting and grading

Sorting of processed tea is mainly done to remove fibre in order to have clean teas. This is usually achieved by passing the teas over a series of screens, which are continually shaken or vibrated, mounted one above the other. This is then followed by grading of teas on particle size, shape, and cleanliness, which is an important stage for the marketing of tea. The major primary grades in Kenya are Pekoe Dust and Dust 1 (D 1), Pekoe Fannings 1 (PF 1), and Broken Pekoe 1 (BP 1). In the tea trade, commonly found, grades include broken orange pekoe, flowery broken orange pekoe, and broken orange pekoe fannings (Hara *et al.*, 1995). The larger particle size grades normally contain larger amounts of aroma compounds than the small particle size grades (Owuor *et al.*, 1987).

2.1.5 Tea quality indicators

Phenolic compounds have long been regarded as one of the principal quality parameters or indicators of tea quality. Organoleptic characteristics of black tea are assessed, using either sight, smell and/or taste of the beverage. Tea tasters, therefore, evaluate the quality of tea using their sense of taste and sight and integrating them to make a valuable judgement. Previous studies have demonstrated a correlation between sensory quality characteristics of resultant black tea, the chemical composition of the green leaf and the black tea processing parameters (Obanda *et al.*, 2001). During the sensory evaluation of black tea, the tasters' judgement is usually based on five parameters, colour, strength, briskness, flavour and overall quality of tea (Roberts and Smith, 1963; Hilton and Ellis, 1972). Theaflavins contribute to all of these tea quality characteristics but some other components such as thearubigins, caffeine, and volatile flavour compounds also have some effects on tea quality (Owuor, 1982). Thus, the total evaluation of black tea appears to be determined by a complex combination of chemical parameters. In addition, the levels of certain pigments have also suggested as a means of predicting black tea quality. These pigments (carotenoid and chlorophyll) in green tea leaf have been used to differentiate teas from various clones (Taylor *et al.*, 1992).

Numerous studies have since displayed a close and statistically significant linear relationship between the content of theaflavins and sensory evaluations of tea tasters or market prices (Hilton and Ellis, 1972) of teas from different countries of origin (McDowell *et al.*, 1991). In Kenya, the correlation between the content of theaflavins and the evaluation of tea tasters has shown a generally positive but statistically non-significant correlation (Owuor *et al.*, 1986a; Owuor *et al.*, 2006b). In Central Africa, significant relationship was demonstrated between total theaflavins and sensory evaluation or tea prices black teas. (Hilton and Ellis, 1972). These differences observed in theaflavins in black teas from the two regions could be due to the differences in the geographical area of production (McDowell *et al.*, 1995) or the genetic differences (Magoma *et al.*, 2000) in the cultivars used in the studies. Further, the rate of formation of theaflavins during aeration varies between clones and within a clone (McDowell *et al.*, 1995). However, the total theaflavins were not considered the critical quality parameters for Kenya's teas (Owuor *et al.*, 2006b). Thus, total theaflavins in the tea liquor may not correctly represent the total astringency and hence the quality of the tea. The individual theaflavins formed at different rates during the aeration result in different levels and ratios of theaflavins in black tea (Obanda *et al.*, 1997).

Black tea astringency varies with composition and total amounts of theaflavins since individual theaflavins have different astringencies. Therefore, astringency with the four theaflavins, simple theaflavin (TF), Theaflavin-3-gallate (TF3G), Theaflavin-3'-gallate (TF3'G) and Theaflavin-3,3'-digallate 3'-digallate 3'-digallate (TFDG) can be better estimators of black tea quality (Owuor *et al.*, 2006a), both in processing and the end product.

2.2 History of tea breeding

The tea plant is a woody long-lived perennial tree characterised by a large diploid genome and recently established to be 3.02 Gb and chromosome number ($2n=2x=30$) (Xia *et al.*, 2017). The tea plant is a highly out-crossing and strongly but not absolutely self-incompatible tree species (Rogers, 1975; Wachira and Kamunya, 2005; Muoki *et al.*, 2007), owing to which it is highly heterogeneous and heterozygous (Banerjee, 1992). Presently, *Camellia* genus is believed to comprise more than 300 species with *Camellia sinensis* (L.) O. Kuntze as the most important agronomic species (Mondal *et al.*, 2004). The recent discovery of *C. cherryana* in 2012 (Orel and Wilson, 2012), indicates the instability and high out-crossing character or property of the genus. Therefore, the allogamous nature of tea coupled with its long generation time, and large genome severely limits the development of true hybrids through conventional breeding. In addition, tea chromosomes are generally small and possess median centromere, indicating their primitiveness and very little evolution in the karyotype (Bezbaruah, 1971) which limits tea improvement to polyploidy screening.

The early breeding strategies were aimed at artificial pollination between plants that differed from one another in some morphological features as a means for producing superior planting materials (Willson and Clifford, 2012). The seeds that were initially produced by controlled hand pollination between selected pairs of bushes and the resulting seedling progenies were replicated in test plots along with a popular commercial jat as control for observations on vigour of growth, morphological uniformity, leaf quality and yield. The two bushes whose seedling progeny showed promise were multiplied by vegetative propagation and planted as generative clones in alternate rows in isolated seed gardens to produce seed by natural cross-pollination. The best seedling fields were then used as checks for high yield and better cup quality for the newly developed outperforming clones (Hainsworth, 1965; Green, 1969; Njuguna, 1985).

In Assam and other parts of north-east India, the emphasis was on mass-selection, which involved random crosses between plants apparently varying in leaf shape, size, texture and growth features (Wight, 1956). Mass-selection, however, often failed not only to produce tea of high-quality but also the uniform morphological attributes essential for high yields and quality (Barua, 1963). It, however, resulted in the development of several seed varieties in Assam that were superior in yield and quality to ‘jats’ that had earlier been randomly planted. In Sri-Lanka, mass-selection was restricted to choosing the outstanding mother bushes which could be propagated vegetatively to produce high-yielding uniform stands (Visser and Kehil, 1958). Unlike in Assam, the emphasis in Sri-Lanka was not so much in the selection of seed bearers (Visser, 1969).

In Africa, the early tea plantations were initially grown from open-pollinated seeds (Cannell *et al.*, 1977) which later formed the base population from which ‘jat’ were selected. Lack of vegetative propagation methods led breeders to establish seed gardens using improved ‘jat’ selections. Mass-selection for ‘jats’ was done where selection was done for general vigour, plucking point density and large shoot size (Green, 1966). The selected ‘jats’ later became progenitors for raising seedling population by open pollination. This method of breeding resulted in slow progress in yield and black tea quality improvement as well as gave rise to non-uniformity of selected unique genotypes. Furthermore, little progress was realised owing to marked environmental heterogeneity (Hasselo, 1964), and continuous exploitation of the same population for further improvement and expansion (Green, 1971).

2.2.1 Tea breeding improvements in Kenya

Tea improvements in Kenya started with the introduction of seeds from Assam, India, which were used to establish the first two polyclonal seed baries at Kangaita (Anon., 1980) and Timbilil (Anon., 1990) after the incorporation of Tea Research Foundation of Kenya (TRFK) in 1980, presently, Tea Research Institute (TRI). Also, the major tea companies, notably James Finlay Kenya Ltd, Brooke Bond (Kenya), Eastern Produce Kenya Ltd and George Williamson (Kenya), initiated their own tea improvement activities by establishing seed baries. Mass-selection was also used as the method for tea improvement, and it was successful, to some extent, in improving the indigenous Assam tea probably because of the presence of desirable genetically linked morphological characters that breed true to type. However, it failed to produce a hardier kind of

tea with acceptable cup characters and morphological uniformity, which could adapt itself to conditions in other regions. Furthermore, since the progenies developed had not been particularly selected for high yield and quality, the resultant seedling populations of mixed genotypes were genetically inferior and diverse (Njuguna, 1985; Wachira, 2001). But above all these challenges, breeding efforts at the TRI, it has led to the expansion of the existing polyclonal seed barie as well as the establishment of new bi-clonal seed baries using elite commercial and promising cultivars.

The selection and breeding process has accelerated the release hundreds of new tea cultivars but, due to the limitations of phenotypic selection; this spurred the search and development of superior but genetically uniform tea clones, which are of better quality and high yields compared to the existing seedling teas (Magoma *et al.*, 2000). Presently, a total of 27 clones (60%) related to TRFK 6/8 have been released for commercial utilisation from this breeding programme at the Timbilil Estate, Kericho (Wachira, 2002; Kamunya, 2003). Also, 45 clones have been developed out of which 24 were selected from seedling populations, including the most popular clone TRFK 6/8 (Kamunya *et al.*, 2010). Among the cultivars developed, few elite cultivars like TRFK 301/4 and TRFK 301/5 (Cambod varieties) were credited with some special characteristics and have been released to the industry for on-farm diversification (Wachira, 1994a; Mamati *et al.*, 2001). Besides high yield and better black tea quality, the two new released clones (TRFK 430/90 and TRFK 371/3), are also tolerant to both abiotic and biotic stress factors (Kamunya and Wachira, 2006). Similarly, clone TRFCA SFS 150 and TRFK 303/577 have been accounted for drought tolerance and EPK TN 14-3 for tolerance to high soil pH, TRFCA SFS 150 and EPK TN 14-3 for cold tolerance, TRFK 12/2 as poor fermenter, 311/287 for tetraploids and TFRK 7/9, TRFK 57/15, AHP SC 31/27, AHP S15/10 for mite tolerant (Seurei, 1996; Kamunya *et al.*, 2010). However, only a few new released clones with superior qualities are being utilised by farmers while the earlier clones are underutilised and rendered irrelevant.

2.2.2 Selection and breeding of high-quality tea clones

Tea quality is an inherent polygenic trait and one of the most important considerations in selecting potential commercial cultivars. It is usually evaluated by both organoleptic and chemical analysis of made tea. Fermentability tests (Sanderson, 1963) and total polyphenol contents (Obanda *et al.*, 1997) are currently being used as potential indicators of black and green

tea quality. However, black tea quality has relied on the relationships between leaf morphology and biochemical parameters in the selection of high-quality tea. It was observed that tea quality significantly correlated with morphological characteristics such as leaf size, colour, and pubescence (Willson and Clifford, 2012). Previous studies have also indicated that the density and length of pubescence would have an influence on the appearance of green tea and on the flavour and aroma of black tea (Wang *et al.*, 2011; Willson and Clifford, 2012). The light leafed cultivars are considered to produce high-quality black tea, while dark-green and pale green cultivars are associated with low quality (Liang *et al.*, 2003). In addition, other than the genetic inheritance, various other factors such as climate or seasonal changes (Agarwal, 1989; Owuor and Othieno, 1991a; Botheju *et al.*, 2000), agricultural practices (Dutta, 1960; Rahman *et al.*, 1978) and processing methods (Takeo, 1984; Agarwal, 1989; Owuor and Orchard, 1990a; Owuor and Orchard, 1990b; Hajra, 2001) were found to play a significant role in determining the quality of the final product. It was also reported that the quality of harvested leaves (plucking standards) affects the quality, and it was reflected in the tea tasters' evaluation (Dutta, 1960; Basu and Choudhury, 1984). Therefore, unlike other traits, the selection of cultivars for better cup quality or made tea quality is complicated at the early stages of cultivar testing and evaluation program. Furthermore, tea plant is an allogamous species, and a large variation of characters occurs from bush to bush in existing seedling tea population and cross-pollinated progenies. Usually, such variation is exploited through the selection programs to develop new tea cultivars with enhanced traits. As the selection program advances from preliminary selection to the final, the variation among the test accessions decreases. Hence, selection of cultivars which also possess traits, which could contribute to better tea quality are eliminated during the early stages of the evaluation programs, unless they possess attributes related to high yield. This is further aggravated by the absence of reliable quality testing methods using green leaf, and the problems encountered during organoleptic analysis. Organoleptic or sensory evaluation is highly subjective and influenced by many factors, which may not have a direct relation to quality (Biswas *et al.*, 1971; Biswas *et al.*, 1973). Although green leaf from a single bush can be processed as black tea (Lamb, 1939), the quantity is not adequate to a tea taster for an accurate estimation. A fair amount of green leaf (Owuor and Obanda, 1998), i.e., about 500 g of green leaf is needed to process green leaf into made tea using miniature processing system and to

subject it for organoleptic analysis. Thus, with the available number of plants per accession at early stages of evaluation trials, this requirement cannot be met.

2.2.3 Breeding for high yield

Developing high-yielding clones from mother bushes of a seedling population has been a challenging task for plant breeders. The early studies (Green, 1971) could not establish consistent correlations between the growth and yield parameters of tea mother bushes, and those of tea clones developed from them. However, subsequent work (Nyirenda, 1989; 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. Further, a strong positive linear correlation was observed (Kulasegeram *et al.*, 1991) between several growth attributes and yield in mature tea grown-fields.

2.2.4 Breeding for pest and disease tolerance

Tea, like any other agricultural crop, is exposed to many pests and diseases. The damage caused by the tea pests has led to a significant impact on quality and yield. In Kenya during heavy infestation, red crevice mite (*Brevipalpus phoenicis*) is estimated to cause yield losses of between 14-30% (Sudo, 1995). Similarly, scale insects (*Aspidiotus* species) can cause yield losses of about 10% (Sudo *et al.*, 1996). However, tea cultivars are known to vary in their morphological and genetic characteristics, which differentially impact on the damage levels caused by pests and diseases (Banerjee, 1987; Hajra, 2001). For example, in Kenya, clone EPK TN14-3 has been noted to be less susceptible to scale insects though also moderately preferred by the red crevice mite (Sudo, 1995). Consequently, polyphenol-rich clones were resistant to mites (Langat *et al.*, 1998), while polyphenol-rich but nitrogen-poor clones were resistant not only to *Brevipalpus phoenicis* Geijske's scarlet mite but also to *Ologonychus coffeae* (Nietner) spider mite (Sudo, 1997). An evaluation of the resistance of tea clones to tea grey blight was also conducted in Kenya. Results showed that the leaf content of total polyphenols inversely correlated to the size of lesions developed following inoculation with *Pestalotiopsis theae* (Sawada) Steyaert pathogen. The tea cultivar with high polyphenol content developed a small lesion on tea leaves (Langat *et al.*, 1998). It was also reported that tea cultivars with high catechins, especially ECG and epigallocatechin gallate (EGCG), were resistant to Kanzawai spider mites (Sudo, 1997). Also, the tea germplasms have exhibited great variation in response

to damage by the root nematodes (*Meloidogyne* spp.), with some tea cultivars being less severely affected than others even in tea fields where they are grown adjacent to one another (Kamunya *et al.*, 2008). Since Kenyan tea is credited for being free of pesticide residues, it is important to use cost-effective and environmentally friendly pest management options such as breeding for pest-resistant or tolerant clones. Therefore, understanding the mechanism of tea plant resistance to pests and diseases is the basis of breeding the resistant cultivar.

2.2.5 Breeding for environmental stress

Changing weather patterns is being experienced in Kenya with increasing temperatures, prolonged drought periods and changing precipitation patterns, especially in the traditional tea growing areas (FAO, 2015). Climate change is drastically reducing tea production with suitable lands being pushed further up the altitude, denting earnings from tea production (Cheserek *et al.*, 2015). Furthermore, there is evidence that over the last 30 years there has been an upward warming trend of 0.2°C/decade in mean temperatures in the tea growing regions (Cheserek *et al.*, 2015). Similarly, stress associated with low temperatures in some parts west of the Rift (Kericho, Kisii, and Nandi), has contributed to serious constraints in tea production. Tea production relies on well-distributed rainfall and increase or decrease in temperatures with changes in rainfall patterns will influence both the quantity and quality of tea (FAO, 2015b). Tea farming is also being extended into the non-traditional tea growing areas that were formerly considered marginal and therefore, unsuitable for the plant. Therefore, the need for consolidation of inherent tolerance to environmental stress in the tea plant is imperative for sustainable tea production (Nagarajah and Ratnasuriya, 1981). In Kenya, the breeding program in TRI, clones with tolerance attributes to various abiotic stress factors, such as clones TRFCA SFS150, TRFK 303/577, EPK TN14-3 and EPK D99/10, have already been identified and introduced into the breeding program.

2.2.6 Polyploidy breeding

Naturally-occurring polyploids in tea may arise either through spontaneous chromosome doubling in somatic tissues or through the occurrence of unreduced gametes. Natural polyploids are more common in Japanese tea varieties than in tea populations from other areas (Banerjee, 1992). Natural polyploids have also been found in tea populations in Kenya (Wachira and Kiplangat, 1991). Naturally-evolved triploids, tetraploids, pentaploids and aneuploids, resulting from open-pollinated progenies, are also found in tea populations in India (HP, 1968).

Polyploidy breeding in tea has been a routine practice for a long period (Gunasekera and Ranatunga, 2003). Polyploidy plays a major role in increasing the amount of genetic material in plants and results in a change in the chromosome number. Most of the cultivated varieties of *Camellia sinensis* and other relatives are diploid ($2n=2x=30$), with a few varieties of triploids ($2n=3x=45$), tetraploids ($2n=4x=60$) and hexaploids ($2n=6x=90$) (Mondal, 2014). Apart from increasing the biomass of crops (Banerjee, 1992), induced polyploids could be used as an additional source of genetic variation to improve the overall performance of existing diploids, or to enhance particular characters such as the shoot size trait (Wachira and Kiplangat, 1991), while retaining most of the characteristics of the diploid progenitor. Also, they could also be used as a source of breeding material for the production of secondary polyploid cultivars by crossing with diploids (Singh, 1980). In tea, various morphological characteristics have been found to be associated with tetraploids (Gunasekera and Ranatunga, 2003). In spite of their high-growth vigour and superior shoot size traits, polyploids have shown to give consistently lower yields than diploids (Wachira, 1994b; Wachira and Ng'etich, 1999). Polyploids tested so far are, however, of restricted genetic background, most having been derived from low to medium yielding progenitors like BBK 2, BBK 5, BBK 7 and BBK 35 (Anon., 1999). Although the value of polyploids in overcoming barriers to inter-specific gene introgression has been recognised in plant species (Gunasekera and Ranatunga, 2003), this has not been fully exploited in perennial crops (Wachira and Ng'etich, 1999). The intra-specific and inter-specific variation of caffeine and catechin content in both wild and domesticated tea genotypes has been studied (Das *et al.*, 2013). However, these early efforts of characterising black tea produced by polyploid genotypes of *Camellia* showed very poor cup characteristics (Gunasekera and Ranatunga, 2003; Das *et al.*, 2012).

2.2.7 Inter-specific hybridisation

Tea breeds freely among the two cultivated species, i.e. *sinensis* and *assamica*, and up to a limited extent with few wild relatives. Earlier, (Wight and Barua, 1957) hybridised *C. irrawadiensis* with *C. sinensis*, and the progenies resembled later but were inferior in tea quality. Later, Bezbaruah and Gogoi (1972) made a successful hybridisation between *C. japonica* and *C. sinensis*, and the progenies were found to be intermediate but produced low yield as well as low quality. The hybrids between *C. sinensis* and *C. japonica* have shown to have a high level of tolerance to diseases such as tea grey bright, tea anthracnose and cold damage during winter

(Takeda, 1990). However, inter-specific hybrids obtained from crossing *C. sinensis* with *C. japonica*, *C. pitardii*, *C. assimilis*, *C. caudata*, *C. salicifolia*, *C. irrawadiensis*, and *C. taliensis* showed very low pollen fertility (Takeda, 1990). In another study, six inter-specific backcross progenies were developed by crossing hybrids of *C. saluenensis* and *C. japonica* (Parks and Kondo, 1974). The segregation data obtained from these six families, three traits were involved in flavonoid constituents, and two were for pubescence characters. (Li *et al.*, 2005) made a cross between high-quality *C. sinensis* and *C. ptilophylla* and out of the 62 progenies, three were identified to be commercially viable by the biochemical parameters or organoleptic test. At the same time, these three progenies were found to be tolerant to cold, a trait which came from *C. ptilophylla*. In a natural cross between *C. taliensis* and *C. sinensis*, two low-caffeine but high-theobromine containing plants were produced (Ogino *et al.*, 2009). Genetic analysis indicated that caffeine-less character might be controlled by one recessive locus. Thus, this caffeine-less plant might be used as a genetic resource for introducing caffeine-less trait in cultivated tea through breeding (Ogino *et al.*, 2009). Indeed, evidence of presence of potential inter-specific hybrids has been obtained by selection in pioneer seedling plantations with morphological traits derived from non-tea *Camellia* species. Such species have been selected in Kenya for further evaluation at the KALRO-TRI, which focuses on the development of diversified tea products from the resulting hybrids (Anon., 1998). Furthermore, numerous intra-specific crosses aimed at diversification of tea products have also been carried out at KALRO-TRI with Japanese Chinary germplasm *C. sinensis* 'Yabukita' and 'Yutakamidori' (Anon, 2001), which are popular commercial cultivars for green tea in Japan. The progenies from these crosses are still young with analytical investigations expected to cast more light on their quality.

2.3 Future tea breeding strategies for desirable attributes

Although, tea breeding objectives vary depending on the geographical region and the local needs. Tea breeding is currently focusing on the selection of populations with high functional components such as polyphenols (catechins), alkaloids (caffeine, theophylline and theobromine), amino acids (theanine), carbohydrates and proteins (Wachira and Kamunya, 2005; Cabrera *et al.*, 2006; Karori *et al.*, 2014). This is because tea is becoming recognised as a health drink rather than a social drink. Previously, especially in black tea producing countries, such as Kenya, India, and Sri-Lanka, tea breeding was more biased towards the development of high-yielding and high black tea quality clones. Therefore, the efforts to combine the two traits in one genotype through

hybridisation breeding have increased significantly over the last decades. But lack of prerequisite knowledge about the inheritance patterns and the combining abilities of desirable traits it has resulted in slow progress in tea improvement (Wachira and Kamunya, 2005). Thus, this calls for concerted efforts on a full understanding of genetic variability in tea because it will help to develop desirable genotypes with desirable traits or qualities (Kaundun and Matsumoto, 2003a; Kaundun and Matsumoto, 2003b). Recently, scientific research studies on mutation breeding, molecular markers, QTL, MAS and genetic or linkage maps (Kamunya *et al.*, 2009; Kamunya *et al.*, 2010; Chen *et al.*, 2012a) have been carried out with the need to improve on tea breeding strategies.

2.3.1 Tissue culture

The techniques of plant cell and tissue culture are also called in vitro techniques (Singh, 1995). This technique can be employed for large-scale propagation of disease-free clones and gene pool conservation. At present, most of the tea plantations in major tea growing countries are planted with clonally propagated cultivars. The development of true hybrids from *C. sinensis* using conventional breeding methods has often been hampered by poor fertility and low viability of hybrid embryos arising from crosses aimed at producing new hybrids (Wachira and Ogada, 1995a). However, micro-propagation by somatic embryogenesis from immature zygotic embryos before abortion provides an embryo rescue technique leading to plant multiplication (Wachira and Ogada, 1995a). Successful in vitro regeneration of *C. sinensis* by somatic embryogenesis has been reported (Wachira and Ogada, 1995a). Tissue culture utilisation in developing pure lines of tea through generation of doubled haploids is still under investigation (Anon, 2001), with useful information to help elucidate tea genetics expected soon.

2.3.2 Hybridisation breeding

Hybridisation is one of the main methods of obtaining genetic variation and breeding of new varieties. Tea plant has been shown to cross with ten different species of the genus *Camellia*, and this has made hybridisation breeding a success (Bezbaruah, 1987). So far, hybridisation breeding has been achieved in *C. irrawadiensis*, *C. taliensis*, *C. japonica* and *C. kissi* species (Ackerman, 1970; Ackerman, 1973; Bezbaruah, 1974; Bezbaruah, 1987). Hybridisation breeding has been used in improving the vegetative and hardiness characters of tea by crossing *C. sinensis* with some closely-related ‘wild’ *Camellia* species (Wachira, 1994c). *C. irrawadiensis* and *C. taliensis*

species have gained special attention as species that lack caffeine, but their liquors lack the quality of tea. However, no new elite cultivars have produced commercially acceptable black tea of good quality though efforts continue to develop elite tea cultivars. Also, numerous crosses aimed at diversification of tea products have also been carried out at TRI with Japanese Chinary germplasm *C. sinensis* 'Yabukita' and 'Yutakamidori' (Anon, 2001) which are popular commercial cultivars for green tea in Japan. Presently, up to 100 clonal hybrids between *C. sinensis* and other *Camellia* species have been developed and are being evaluated at the TRI to shade more light on their quality (Anon, 2007).

2.3.3 Mutation breeding

Mutation breeding is artificially using physical, chemical and biological factors to induce the tea plant to produce genetic variation or to get new valuable genetic materials for further breeding use. Unlike hybridisation breeding, mutation breeding has the advantage of improving a defect in an otherwise elite cultivar, without losing its agronomic and quality characteristics. Early studies on mutation breeding on tea did indicate that the artificial treatments produced inferior mutants that had reduced vigour, stunted growth, few foliage and branches (Singh and Sharma, 1982). However, a recent study on combined effects of gamma-rays and chemical mutagens on biological damage to tea seed has shown that physicochemical mutagen was more significant than that treated with either gamma-ray or chemical mutagens (Chen *et al.*, 2007). A new variety in China which is in its final adaptability test with very early sprouting in the spring, high cup tea quality, resistant to disease and suitable for fine green tea has been developed through mutation breeding (Chen *et al.*, 2007). In TRI, mutagens have been incorporated in vitro cultured materials with the aim of developing new genotypes that can tolerate stress factors and secondary polyploids (Anon, 2001). This is a clear indication that mutation breeding is becoming one of the practical breeding approaches for tea plant.

2.3.4 Breeding for health-promoting effects

Tea being a popular beverage has attracted the attention of both the consumers and the scientific community because of its multiple health-promoting effects (Preedy, 2012). A number of bioactive compounds, such as the flavonoids, caffeine, L-theanine and γ -aminobutyric acid (GABA) are present in tea (Lin *et al.*, 2012; da Silva Pinto, 2013). Tea catechins (green tea) and theaflavins (black tea) have demonstrated to have a wide variety of pharmacological properties,

including antioxidant, anti-inflammatory, anticancer, anti-diabetic and neuroprotective effects (Preedy, 2012). Green tea catechins have been associated with numerous pharmacological properties such as; antimicrobial (Koech *et al.*, 2013), anti-inflammatory (Karori *et al.*, 2008), anti-aging (Esposito *et al.*, 2002), antimalarial (Sannella *et al.*, 2007) and antioxidant activity (Wachira and Kamunya, 2005; Kerio *et al.*, 2013; Karori *et al.*, 2014). Green tea has also been shown to have significant protection against age-associated pathologies and degenerative diseases such as, Parkinson's disease, Alzheimer's disease, and ischemic damage (Mandel and Youdim, 2004). Green tea has also shown anti-diabetic effects in animal models of insulin resistance and humans (Wu *et al.*, 2004; Suzuki *et al.*, 2012; Miyoshi *et al.*, 2015). ECG has been shown to inhibit growth of cancer cells by stimulating apoptosis or programmed cell death both of which are crucial aspects of cancer prevention (Zaveri, 2006; Mao *et al.*, 2010; Mbuthia *et al.*, 2016). EGCG has also been reported to be a powerful antagonist of human immunodeficiency virus reverse transcriptase (HIV-RT) (Xu *et al.*, 2017a). The health benefits associated with green tea are mainly attributed to antioxidant properties of polyphenolic catechins to scavenge reactive oxygen species (Zaveri, 2006). The antioxidant properties are due to the presence of the phenolic hydroxy groups on the B-ring in non-galloylated catechins (EC and EGC) and in the B-ring and D-ring of the galloylated catechins (ECG and EGCG) (Salah *et al.*, 1995). The presence of the 3,4,5-trihydroxy B-ring has been shown to be important for antioxidant properties and radical scavenging activity (Nanjo *et al.*, 1996; Valcic *et al.*, 1999).

Theaflavin-3,3'-digallate 3'-digallate (TF3), a main theaflavin of black tea, has been shown to be effective in anti-obesity (Ko *et al.*, 2015), anti-inflammation (Ko *et al.*, 2014), cancer chemoprevention (Pan *et al.*, 2013; Gao *et al.*, 2016), lowering hepatic lipid accumulation (Jin *et al.*, 2013), and insulin sensitization (Pan *et al.*, 2016). Theaflavins were also found to suppress the transcription of HIV in the cell, and the gallic acid moiety of theaflavins can enhance suppressive activity (Wang and Li, 2006). Theaflavins found in black tea also possess anti-oxidative properties, which are equally as effective as catechins in green tea (Karori *et al.*, 2007; Jochmann *et al.*, 2008) which contribute to health benefits associated with black tea. Caffeine present in tea is not only known to contribute to tea briskness, and other taste characteristics (Obanda *et al.*, 1997; Nitin *et al.*, 2006), but is also known for its stimulatory effect (Sharangi, 2009). Caffeine has also been reported to suppress the body weight gain by suppressing food intake and reducing adipose tissue (Kobayashi-Hattori *et al.*, 2005). These

health benefits associated with tea intake has led to high consumption of tea products, especially regarding tea extracts in pharmaceutical, food and beverage and cosmetic industries. Also, to these health benefits, tea quality is a complex phenomenon and depends mainly on the biochemical components and colour attached to the tea infusions. Therefore, this calls for concerted efforts in tea breeding programs to develop tea cultivars with improved biochemical content. This will not only lead to the development of high-quality diversified, and value-added tea products but will also maintain Kenya's position as a leading tea exporter and increase its foreign exchange earnings.

2.4 Marker-assisted selection

MAS involves the selection of plants carrying genomic regions that are involved in the expression of traits of interest through molecular markers. The development and availability of an array of molecular markers and linkage maps has made MAS possible to be utilised in crop plants. For example, rice markers have been used to identify new syntenic markers in barley (Perovic *et al.*, 2004) and wheat (Liu and Anderson, 2003). Furthermore, this approach has been used to improve neglected crops such as pearl millet (Gale and Devos, 1998). In tea, a linkage map exists (Hackett *et al.*, 2000; Kamunya *et al.*, 2010; Taniguchi *et al.*, 2012; Hu *et al.*, 2013; Ma *et al.*, 2014) with a few studies on QTL mapping (Ma *et al.*, 2015; Tan *et al.*, 2016), but there is no report on comprehensive QTL analysis in the crop to date. In previous studies, (Tanaka, 1996; Wachira, 1996; Huang *et al.*, 2006a; Taniguchi *et al.*, 2012) only identified markers and genomic regions that influenced the expression of several quantitative traits and unsaturated linkage maps. Therefore, the development of saturated linkage maps will provide a greater arsenal of tools for QTL mapping and MAS in tea breeding.

2.5 Role of genetic markers in tea breeding improvement

Genetic markers are specific locations on the chromosome and serve as landmarks for genome analysis. The three basic types of genetic markers include: morphological, biochemical and DNA markers. There are numerous overlapping morphological, biochemical and physiological attributes used in conventional plant breeding to determine the genetic diversity of the cultivars. Tea is a highly heterogeneous plant due to its outbreeding nature. Presently, there are numerous studies that have been conducted to evaluate genetic diversity of tea plant by incorporating various techniques, such as morphological markers (Chen *et al.*, 2005), biochemical and

isoenzyme markers, (Magoma *et al.*, 2000; Magoma *et al.*, 2003; Das and Bhagobaty, 2010) and molecular markers (Wachira *et al.*, 1995; Wachira *et al.*, 1997; Kaundun and Park, 2002b). However, most of these attributes have shown to have continuous variation, a high degree of plasticity and mostly influenced by environmental factors (Banerjee, 1988; Ariyaratna and Gunasekare, 2007). This has led to difficulty in identifying discrete taxonomic groups since more robust molecular markers are needed for genotyping.

2.5.1 Morphological markers

Morphological traits often have one to one correspondence with genes controlling the traits. Due to that, morphological markers can be used as reliable indicators for specific genes and as genetic markers on chromosomes. Most morphological markers have originated from mutation studies (Wettstein-Knowles and Shewry, 1992; Waycott *et al.*, 1999). Morphological markers have not been used extensively in plant breeding because of the limited availability of different mutants, and those available were not neutral in their effect on agronomic phenotype (Worland *et al.*, 1987). Also, morphological markers are influenced by environmental factors, show continuous variation and a high degree of plasticity making it difficult to identify discrete taxonomic groups based on morphological markers.

2.5.2 Biochemical and isoenzyme markers

The variations and diversity between the tea cultivar types has been effectively identified using biochemical markers, such as catechins, caffeine, volatile compounds, theanine and anthocyanins (Magoma *et al.*, 2000; Deng *et al.*, 2010; Li *et al.*, 2013). Isoenzymes, which are the product of genes and give a direct measure of gene homology, and variation has been used to characterise tea germplasms (Magoma *et al.*, 2003; Das and Bhagobaty, 2010). The isoenzymes that have been used include; aspartate aminotransferase (AST), esterase, glutamate dehydrogenase (GDH), malate dehydrogenase (MDH) and peroxidase (PER). However, both biochemical and isoenzymes markers are influenced by environmental factors, the developmental stage of a plant and few biochemical assays are available (Das and Bhagobaty, 2010).

2.5.3 Molecular markers

The use of molecular markers in breeding and selection for desirable agronomic traits in crops offers many enhanced possibilities and opportunities even for traits that are difficult to evaluate using morphological, biochemical or isoenzyme markers. Unlike morphological, biochemical or

isoenzyme markers, molecular markers are many, are not affected by environmental factors and the developmental stage of the plant (Singh *et al.*, 2013). Also, molecular markers provide an efficient, auxiliary means to distinguish intra- or inter-specific variation of germplasms thus, it is an essential tool in tea breeding (Hackett *et al.*, 2000; Ni *et al.*, 2008). Molecular markers have been broadly classified into three classes based on the method of their detection, which include; hybridisation-based, polymerase chain reaction (PCR)-based and DNA sequence-based (Angaji, 2011).

Molecular DNA markers for tea cultivars that can tolerate low and high temperatures, drought, pests, and diseases and produce high yield and quality processed tea has been studied (Kamunya *et al.*, 2009; Kamunya *et al.*, 2010). Molecular DNA markers controlling catechins contents has also been studied in Chinary varieties for green tea processing and for future MAS in tea breeding (Ma *et al.*, 2014). Molecular markers have proven to be useful and sufficient tools to characterise and discriminate tea varieties and commercial teas from the market (Wachira, 2001; Kaundun and Matsumoto, 2003b). Expressed sequence tags (ESTs) markers for blister blight disease resistance (Agarwal and Das, 2009) and insect resistance (Ahmed and Das, 2009) has been studied for MAS to have resistant varieties. Also, drought specific AFLP based markers (Mishra and Sen-Mandi, 2004b) has been studied for future MAS in tea breeding. To achieve the ultimate goal of genetic improvement of tea, further efforts are required to construct a saturated linkage map using molecular markers.

2.6 Utility of molecular markers in targeting phenotypic traits

The advancement in molecular biology in the early 1980s and 1990s enabled the application of molecular markers, which improved the speed and precision of trait mapping (Figure 2.10). The recent developments in molecular tools and novel biotechnology-based approaches offer the promise of providing additional information on the genetic basis of desirable traits, and the nature of linkages with other traits and so may assist major advances in crop improvement in the future (He *et al.*, 2014). The molecular markers can alleviate the problems of working with phenotypic traits by offering easier and manipulation of complex and difficult to measure traits. They can define the presence of ‘non-active’ genes, identify where the traits are controlled by several genes and reduce the need for elaborate and expensive procedures. Molecular markers also potentially allow plant genotyping at almost any stage of development (Gupta *et al.*, 2013).

Various applications of molecular markers include: uncovering phylogeny and population structure of crop germplasm and identifying genetic variation in parental germplasm (Yang *et al.*, 2015; Jannatdoust *et al.*, 2016; Wambulwa *et al.*, 2016), identifying gene functions via QTLs or association mapping, which are implanted as MAS (Moncada *et al.*, 2016; Toppino *et al.*, 2016) and linking genomes of related species via comparative mapping for gene discovery and comparisons of gene functions (Higashi and Saito, 2013; Denoeud *et al.*, 2014; Zhang *et al.*, 2015). Molecular genetic markers with specific emphasis on DArT markers were utilised in this study in identifying qualitative and QTL-associated with tea quality and relative water.

2.6.1 DNA-based markers

2.6.1.1 Restriction fragment length polymorphism (RFLP)

RFLP markers were the first DNA-based genetic markers developed (Botstein *et al.*, 1980). RFLPs are co-dominant markers, and hence can be good estimators of heterozygosity (Miller and Tanksley, 1990). The advantage is that once a library of probes is generated, RFLPs can be generated without any sequence information. However, there are considerable disadvantages with the technology, such as low levels of polymorphism (Jena and Kochert, 1991). For a previously unexplored species, it is necessary to develop a suitable probe library, of either cDNA or genomic DNA. The generation of RFLP data is also time-consuming, particularly with single-copy probes. The assay is one of the most costly to perform as many steps are involved, usually with radioactive labels. Also, large quantities of DNA are also required of 5 to 10 µg per digest and, as a result, whole plants are often required for DNA extractions.

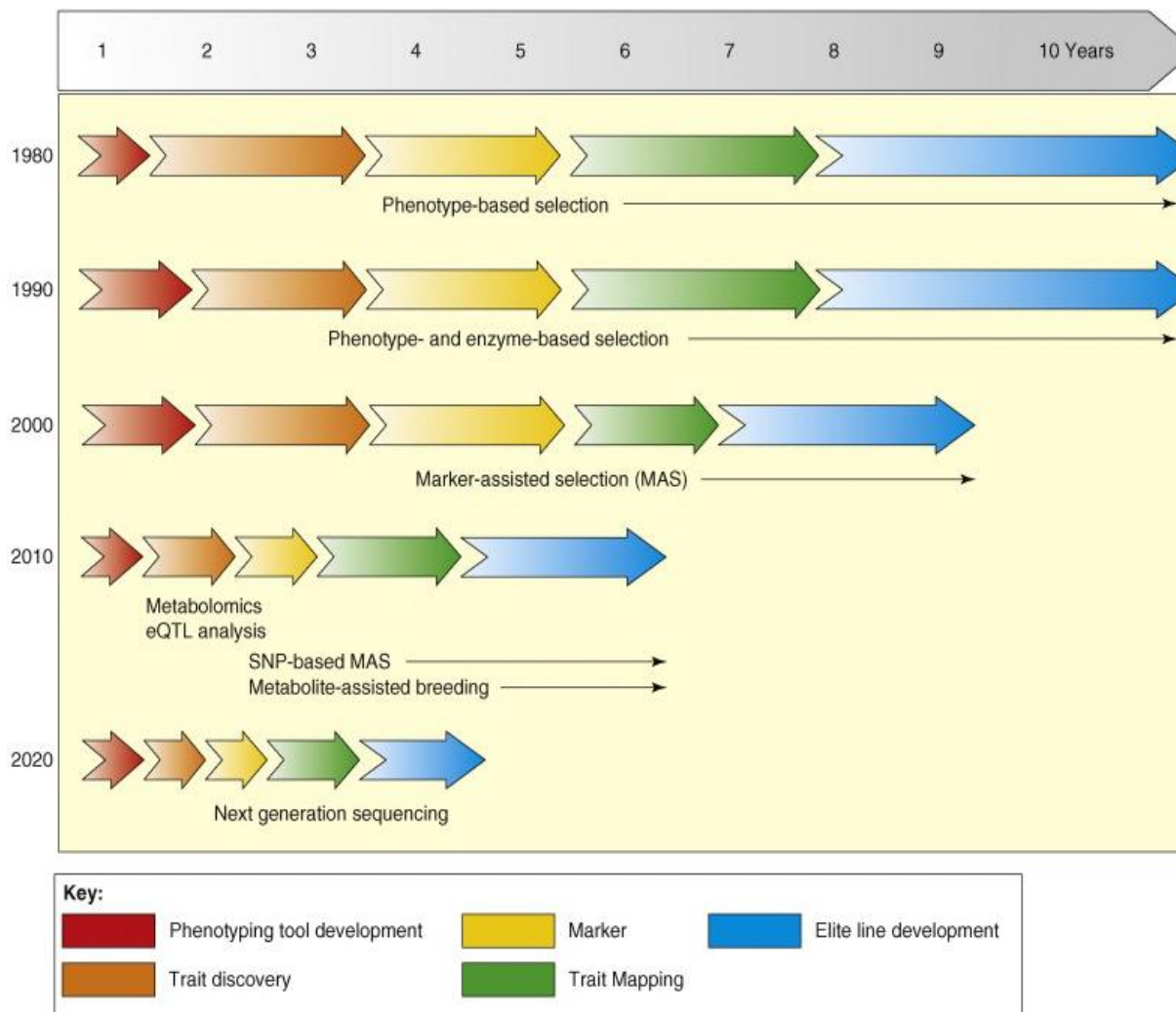


Figure 2.10: Breeding technology pipeline from the past to the present and to the future.

Abbreviations: SNP, single nucleotide polymorphism, QTL, quantitative trait loci. Source: (Fernie and Schauer, 2009).

2.6.2 PCR-based methods

2.6.2.1 Random amplified polymorphic DNA (RAPD)

Random Amplified Polymorphic DNA Random amplified polymorphic DNA (RAPD) are DNA fragments amplified by the PCR using short synthetic primers (generally ten bp) of random sequence (Williams *et al.*, 1990). RAPD markers were the first of the PCR-based markers and were developed independently by (Welsh and McClelland, 1990) and (Williams *et al.*, 1990). Polymorphism between individuals is defined as the presence or absence of a particular RAPD band. RAPD markers have since been used for linkage mapping and marker analyses in many

crop plants (Williams *et al.*, 1990; Carlson *et al.*, 1991; Hackett *et al.*, 2000). Since a single RAPD primer can anneal to many locations in the genome, multiple loci are revealed by a single primer. Therefore, it is possible to obtain a large number of RAPD genetic markers in a short amount of time and at relatively low cost. Although RAPD markers are easy and quick to use, the main drawback is their low reproducibility (Schierwater and Ender, 1993), and hence highly standardised experimental procedures are needed because of their sensitivity to the reaction conditions. RAPD analyses generally require purified, high molecular weight DNA, and precautions are needed to avoid contamination of DNA samples because short random primers are used to amplify DNA fragments.

2.6.2.2 Amplified fragment length polymorphism (AFLP)

AFLP markers are a recent development as compared to RAPD markers (Vos *et al.*, 1995). AFLP involves the restriction of genomic DNA, followed by ligation of adaptors complementary to the restriction sites and selective PCR amplification of a subset of the adapted restriction fragments. They are like RAPDs in that many markers can be assayed quickly using PCR, and they are generally dominant; but, AFLPs appear to be more repeatable than RAPDs. Polymorphisms are detected from the differences in the length of the amplified fragments by polyacrylamide gel electrophoresis (Mitchelson, 2003) or by capillary electrophoresis. The first report of the use of AFLPs in tree plants was by (Cervera *et al.*, 1996) who used this marker system to genetically map a disease resistance gene in *Populus*. Genetic linkage maps based on AFLPs have also been constructed in other plants (Paul *et al.*, 1997; Hackett *et al.*, 2000; Loh *et al.*, 2000; Wachira, 2001). AFLP markers are dominant markers; there is the need to adopt certain strict but subjectively determined criteria for acceptance of bands in the analysis due to the high number and different intensity of bands per primer combination.

2.6.2.3 Simple sequence repeats (SSRs)

Microsatellites, also known as SSRs, short tandem repeats or sequence-tagged microsatellite sites, are tandemly repeated units of short nucleotide motifs that are 1-6 bp long (Tautz and Renz, 1984). SSR loci are individually amplified by PCR using pairs of oligonucleotide primers specific to unique DNA sequences flanking the SSR sequence. Microsatellites may be obtained by screening sequences in databases or by screening libraries of clones. An alternative source of SSRs is to utilise EST and other sequence databases (Kantety *et al.*, 2002). SSR markers are

characterised by their hyper-variability, reproducibility, co-dominant nature, locus specificity and random dispersion throughout most genomes. SSR markers are considered excellent genetic markers for genetic fingerprinting, population genetic studies, particularly for examination of gene flow and molecular breeding (Powell *et al.*, 1996). The advantages of SSRs are that they can be readily analysed by PCR and are easily detected on polyacrylamide gels. Also, SSR genotyping throughput is high, and more variable than RFLPs or RAPDs and can be automated. Also, start-up costs are low for manual assay methods, especially when the markers have been developed and SSR assays require only very small DNA samples (~100 ng per individual). The disadvantages of SSRs are the labour-intensive development process, particularly when this involves screening genomic DNA.

2.6.2.4 Inter-simple sequence repeat (ISSR)

The ISSR are semi-arbitrary markers amplified by PCR in the presence of one primer complementary to a target microsatellite. Each band corresponds to a DNA sequence delimited by two inverted microsatellites (Zietkiewicz *et al.*, 1994; Tsumura *et al.*, 1995; Nagaoka and Ogihara, 1997). ISSRs have been used for genetic diversity studies in different crop plants (Nagaraju *et al.*, 2002; Reddy *et al.*, 2002; Obeed *et al.*, 2008). These markers are also suitable for the identification and DNA fingerprinting (Gupta and Varshney, 2000; Gupta *et al.*, 2002). ISSRs are quick and easy to handle, an inexpensive genotyping technique based on variation in the regions between microsatellites. It is reproducible and does not require genome sequence information; it leads to high multi-locus, polymorphic patterns and produces dominant markers (Mishra *et al.*, 2003). Furthermore, ISSRs is known to be able to discriminate between closely-related genotypes (Fang and Roose, 1997; Hodkinson *et al.*, 2002) and it can detect polymorphisms without any previous knowledge of the crop's DNA sequence.

2.6.2.5 Sequence-tagged site (STS)

STS is a relatively short and single-copy DNA sequence that can be specifically amplified by PCR (Olson *et al.*, 1989). STS markers are co-dominant, locus-specific, and more reproducible. The advantages of using STS markers include; genotypes are easy to score and interpret, and only a small quantity of DNA is needed for one assay. Also the cleaved and un-cleaved amplification products can be adjusted arbitrarily by the appropriate placement of the PCR primers. The procedure is technically simple with robust results because the amplification

product is always obtained (Drenkard *et al.*, 2000). STS markers have been used to assess the relationships among cultivated tea and allied wild species of the genus *Camellia* (Wachira *et al.*, 1997), thus, providing useful information on genetic relationships at the subgenus and sectional level in *Camellia*.

2.6.2.6 Cleaved amplified polymorphic sequences (CAPS)

CAPS are alternative molecular markers, which combine both PCR and RFLP techniques to reveal polymorphism. CAPS markers that have successfully been applied to different plants species (Hu *et al.*, 2014; Shavrukov, 2016). In tea, CAPS markers have been used in genetic diversity of tea by investigating the genes responsible for catechin and tannin biosynthesis (Kaundun and Matsumoto, 2003a). CAPS is the simple and identifications of CAPS results in single or very few fragments after digestion, representing a very positive advantage of this method. However, CAPS markers have clear limitations, in the relatively lengthy step of analysis as well as relatively high-cost of rarer restriction enzymes. Furthermore, CAPS markers are less adaptable for high-throughput automated systems.

2.6.2.7 SNP

An SNP is a single nucleotide base difference between two DNA sequences or individuals. SNPs provide the ultimate or simplest form of molecular markers as a single nucleotide base is the smallest unit of inheritance, and thus they can provide maximum markers (Edwards *et al.*, 2007). SNPs are increasingly becoming the marker of choice in genetic analysis and are used routinely as markers in agricultural breeding programs (Gupta *et al.*, 2001). These applications include; high-resolution genetic map construction (Kota *et al.*, 2001; Morales *et al.*, 2004; Lopez *et al.*, 2005), linkage disequilibrium-based association mapping (Zhu *et al.*, 2003), genetic diagnostics, genetic diversity analysis, cultivar identification, phylogenetic analysis, and characterisation of genetic resources (Gupta *et al.*, 2001; Rafalski, 2002a; Rafalski, 2002b). SNPs are highly stable co-dominant markers and can be identified within or near genes underlying agronomic traits. The abundance of SNPs in plant genomes makes them attractive tools for MAS breeding to identify individual plants containing a combination of alleles of interest from large segregating populations. SNPs can be very easily automated and quickly detected, with high efficiency for detection of polymorphism. However, due to high costs for start-up or marker development,

high-quality DNA required and high technical or equipment demands limit the application of SNPs in some laboratories and practical breeding programs.

2.6.2.8 Diversity array technology (DArTseq)

Diversity Arrays Technology (DArTseq) is a high-throughput microarray hybridisation-based technique which involves isolation and cloning of a random set of DNA fragments from complexity-reduced DNA sample (Figure 2.11). This technique has experienced increased interest in the recent years (Jaccoud *et al.*, 2001; Gupta *et al.*, 2013). DArT has provided a standardised high-throughput genotyping, whereby thousands of markers can be readily assayed in parallel in thousands of samples without prior sequence information (Wittenberg *et al.*, 2005). The DArT array has demonstrated excellent performance for phylogenetic and diversity analyses (Steane *et al.*, 2011), genomic selection (Poland *et al.*, 2012) and genetic linkage mapping (Schouten *et al.*, 2012). DArT has been applied successfully in plants with complex genomes such as sugarcane (Racedo *et al.*, 2016), wheat (Zou *et al.*, 2016), strawberry (Sánchez-Sevilla *et al.*, 2015), apple (Schouten *et al.*, 2012) citrus (Liu *et al.*, 2016b) and trees such as eucalyptus. DArT markers are bi-allelic, behave in a dominant or co-dominant manner and detect single-base changes as well as indels. These markers are better alternative to currently used techniques, including RFLP, AFLP, SSR, and SNP in terms of cost and speed of marker discovery and analysis for whole-genome fingerprinting. It is cost-effective, sequence-independent, a non-gel based technology that is amenable to high-throughput automation and the discovery of hundreds of high-quality markers in a single assay.

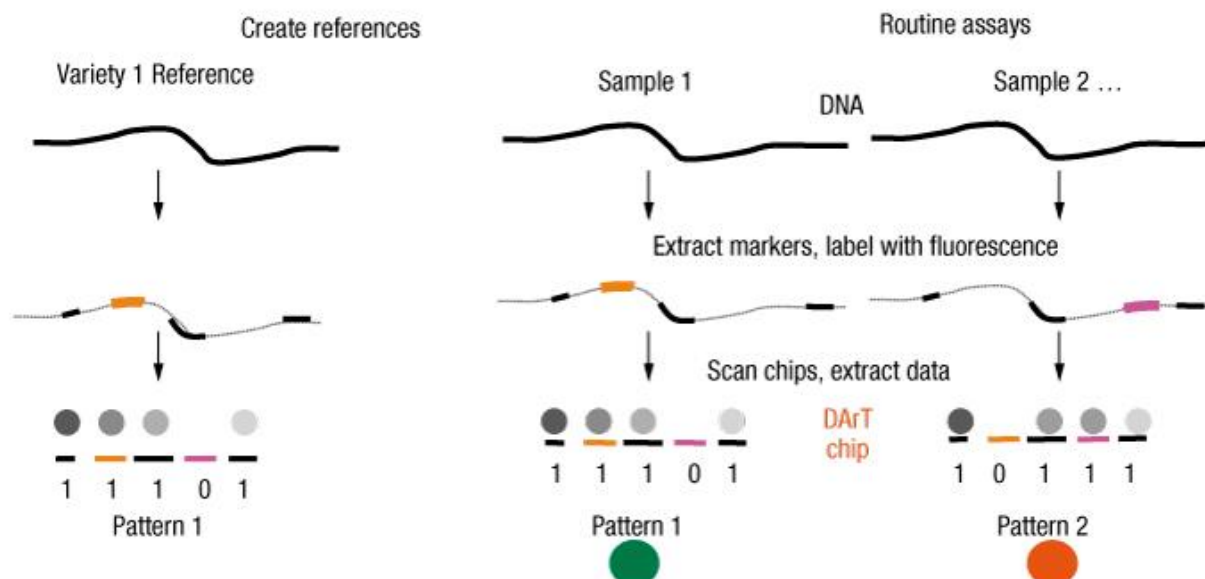


Figure 2.11: Schematic diagram of the principle of DArTseq.

Source: <http://www.diversityarrays.com/>

2.7 QTL

Several important traits in crop plants, such as yield, quality abiotic and biotic stress tolerances are quantitative and are controlled by many genes, known as polygenes. The principle of QTL analysis is the detection of linkage between phenotype and genotype of markers in the progeny evaluated for trait of interest and at regular intervals throughout the genome (Collard *et al.*, 2005). A significant difference between mean trait values of the traits of interest and genotype of markers indicates the association between the marker and the QTL (Miles and Wayne, 2008). The genetic basis of uncovering these traits is of importance for plant breeding, genomics, and germplasm utilisation. QTL analysis provides information about the nature of QTL, their location within the chromosome and what they do and how they act and interact. In several crops, molecular markers have successfully been used for studying quantitative inheritance (Elberse *et al.*, 2004; Agarwal *et al.*, 2008). The linkage between genetic markers and quantitative traits of economic importance has been documented in many plants. For example, in the tomato, selection for specific QTL in segregating progeny has led to the development of insect-resistant lines and cold tolerance (Grandillo and Cammareri, 2016). (Elberse *et al.*, 2004) detected several QTL affecting growth-related traits in wild barley. Studies on cacao detected

stable yield QTL in a study that spanned 15 years (Crouzillat *et al.*, 2000). In tea, (Wachira, 1996; Kamunya *et al.*, 2010; Hu *et al.*, 2013; Ma *et al.*, 2014) studied the existence of different markers that are significantly associated with QTL influencing important traits. The major objectives of QTL mapping in plants are to increase the biological knowledge of the inheritance and genetic architecture of traits and to identify markers that can be used as indirect selection tools in breeding programs.

2.7.1 QTL mapping experiments

2.7.1.1 Choice of mapping population

The choice of mapping population greatly depends upon the intention of the experimenter, trait complexity, the timeframe as well as the resources available for undertaking QTL analysis. The parents selected should have contrasting phenotypes for the traits of interest (e.g. tolerant and susceptible lines). Also, the parents should be genetically divergent to enhance the identification of as many polymorphic markers as possible across the whole-genome (Semagn *et al.*, 2010). Segregating populations can be F₂ or F₂-derived population, backcrosses, recombinant inbred lines, near-isogenic lines, backcross inbred lines and double haploid populations. The size of mapping populations is a key factor affecting the power of QTL detection. Theoretically, using 100 progeny could greatly overestimate phenotypic variances associated with a correctly identified QTL, whereas increasing the population size to 500 would lead to only a slight overestimate. A population size of 1 000 would provide an estimate of the magnitude of the phenotypic effect close to the actual value (Xu, 2003).

2.7.1.2 Size of mapping populations

Large mapping population sizes enable the detection of QTL that have small effects on traits of interest. However, population sizes of greater than 500 individuals are not practicable as they are laborious and costly for phenotyping and genotyping, and are more useful in detecting QTL with major effects (Jiang, 2015). A mapping population size of between 50-300 individuals, combined with reasonably reliable phenotypic data and inter-marker spaces on the linkage map of about 10-15 cM apart, together with an appropriate statistical method, will usually lead to the identification of QTL (Burdon and Wilcox, 2011). However, with only 50 individuals, QTL effects would have to be very large to be detected.

2.7.1.3 Phenotypic data

The phenotypic data required for QTL mapping are estimates of the phenotypic performance of individuals across environments. The power of QTL detection depends upon the sample size, the heritability of the trait, the genetic differences among progenies, the effect of the QTL, and the environment used for phenotypic evaluation (Semagn *et al.*, 2010; Würschum and Kraft, 2014). Although current molecular tools can provide high-throughput with low cost genotyping, phenotyping imposes limits on sample sizes due to its laborious, time-consuming and high-cost nature. High trait heritability is a prerequisite for reliable QTL results and low bias in the estimation of QTL effects. Phenotypic data can be pooled over locations since pooled analysis from genetically similar populations may have a higher power of QTL detection than single-population based analyses (Sun *et al.*, 2010). It is also desirable to have replicated trials across locations for detection of QTL x environment interaction if any.

2.7.1.4 Statistical methods for QTL mapping

The various mathematical methods for QTL mapping can be grouped based on their requirements for genetic maps. First, those methods that do not require prior genetic linkage map construction, e.g. single marker analysis (ANOVA) and partial least squares regression. Secondly, those methods that require a genetic map for the population, e.g. simple interval mapping, composite interval mapping, multiple interval mapping (Collard *et al.*, 2005; Semagn *et al.*, 2006a).

Single marker analysis is the simplest method for QTL detection, in which a t-test, analysis of variance (ANOVA) and linear regression are included but, which does not require a complete linkage map. However, the two major limitations of this statistical method are (i) the reduced chance of detecting a QTL that is further apart from the markers and in identifying associations with other QTL and (ii) the underestimation of QTL effects because of recombination (Collard *et al.*, 2005).

Composite interval mapping (CIM) analysis combines interval mapping with multiple regression analysis. In other words, linked markers treated as cofactors are incorporated into a multiple regression model (Zeng, 1994). Thus, CIM takes into account the genetic variation due to other QTL, resulting in more precise QTL mapping and greater power of QTL detection. CIM is

further extended to another approach called multiple interval mapping (MIM). The MIM uses multiple marker intervals to fit multiple QTL models (Kao *et al.*, 1999).

However, there are always errors in any QTL analytical approach. As with any statistical test, two types of error, Type I and Type II, can occur in QTL mapping (Jansen, 1994). (1) A Type I error is a false-positive, in which a segregating QTL is detected when, in fact, it is not present. (2) A Type II error is a false negative, in which a QTL is not detected when it actually exists. These two types of errors can be controlled by proper critical values (Type I) and determined by the experimental design and the size of the QTL effect (Type II). Statistically, an efficient QTL mapping method should have high detection power and low false discovery rate.

2.7.2 Linkage mapping

Linkage maps provide the position and relative genetic distances between markers on the chromosome. The distance along the map is expressed in centi-Morgan (cM), a unit which is converted from Kosambi recombination fraction (Kosambi, 1943) or Haldane mapping functions (Haldane, 1919). Kosambi's mapping function allows some interference, which is the effect in which the occurrence of crossover in certain region reduces the probability of crossover in the adjacent region (Haldane, 1919). Linkage between markers is usually expressed as the logarithm of the ratio of linkage versus no linkage, called the logarithm of odds (LOD). Normally, a maximum LOD of 3, where linkage is 1 000 times more likely to occur than no linkage is required for linkage assertion. The principle of linkage map construction is the segregation of genes and markers via chromosome recombination during meiosis. Linked markers are grouped into linkage groups, which indicate chromosome segments or entire chromosome. The linkage map serves as a tool for identifying chromosomal regions containing genes and QTL-associated with target traits (Alonso-Blanco and Méndez-Vigo, 2014).

QTL controlling yield, drought tolerance, and quality traits and pubescence on some specific tea germplasms were studied for MAS (Kamunya *et al.*, 2010). Linkage map from the female parent, SFS150, was established with RAPD and AFLP markers (Hackett *et al.*, 2000). The molecular markers related with theanine content, date of bud sprouting, resistance to anthracnose and cold tolerance in tea was detected (Tanaka, 2000). Recently, QTL controlling catechins content has been developed using SSR markers for the MAS in the near future (Ma *et al.*, 2014). As mentioned, tea plant has a long juvenile period, and conventional breeding will require large

investments in land and time for progeny trials. A linkage map could provide a vital tool for relating phenotype to genotype hence facilitates identification and selection of recombinant parents with desirable attributes. This will also increase the speed and precision of tea breeding and improvement programmes.

2.7.3 Association mapping

Association mapping (AM) and linkage mapping are two methods widely used for locating genes or QTL in plant breeding (Xu *et al.*, 2017b). In recent years, AM has been applied in *Arabidopsis*, maize, barley, durum wheat, spring wheat, sorghum, sugarcane, sugar beet, soybean, cotton, grape and forest tree species (Abdurakhmonov and Abdugarimov, 2008; Khan and Korban, 2012) as well as rice (Han and Huang, 2013). Recently, AM has been used to study the genetic relationship between tea caffeine synthase gene (TCS1) and the caffeine content in tea plant and its related species (Jin *et al.*, 2016). AM which identifies QTL by examining the trait-marker associations has been used to overcome the limitations of linkage mapping (Kraakman *et al.*, 2004). In linkage mapping, only two alleles at any given locus can be studied in bi-parental crosses and gives low mapping resolution (Flint-Garcia *et al.*, 2003). Thus, AM generally offers higher resolution mapping due to a greater number of recombination events. However, AM has a lower power to detect the effect of rare alleles compared to linkage analysis which has a higher statistical power due to greater allele replication (Korte and Farlow, 2013). Moreover, population structure may cause false positives in AM, but this has been overcome by using a mixed-model, which take both population structure (Q) and kinship (K) into account to reduce false positives (Yu *et al.*, 2006). A comparisons of different statistical models (Q, and Q+K) have been conducted for *Arabidopsis* (Zhao *et al.*, 2007), maize (Yu *et al.*, 2006), rice (Zhao *et al.*, 2011), cotton (Abdurakhmonov *et al.*, 2008), grape, poplar and pine (Eckert *et al.*, 2010; Khan and Korban, 2012). However, false-positive might not be avoided through the aforementioned models. To avoid false-positive, it is required that the significant associations identified within one population should be validated in another population (Wray *et al.*, 2013). Therefore, a combination of linkage mapping and AM could provide mutual authentication and enable more reliable results to be obtained.

2.8 Role of machine learning and genomic-enabled prediction in plant breeding

In the early 1980s, the development of different molecular marker systems drastically increased the total number of polymorphic markers available to plant breeders. One of the most notable high-throughput genotyping technique is SNPs, which have been used intensively in QTL studies (Crossa *et al.*, 2017). More than 10, 000 QTL using different marker systems have been reported in more than 120 studies covering 12 plant species (Bernardo, 2008) that aimed to improve quantitative traits of agronomic and economic importance. Initially, molecular markers were integrated with phenotypic traits by applying MAS technique. However, attempts to improve complex quantitative traits by using QTL-associated marker detection have been unsuccessful due to the difficulty of finding the same QTL across multiple environments and due to QTL-environment interactions or different genetic backgrounds (Weinig and Schmitt, 2004). Linkage analysis for QTL mapping is mainly done on bi-parental populations but has low power for detecting marker-trait association due to low recombination rates of the chromosomes (Flint-Garcia *et al.*, 2003). In early 2000, AM was started with the objective of overcoming the set back of low power of linkage analysis, thus facilitating the detection of marker-trait associations in bi-parental populations and fine-mapping chromosome segments with high recombination rates. However, the main limitation of fine AM is the low power for detecting rare variants that may be associated with traits of agronomic importance (Korte and Farlow, 2013). Thus, the challenge of AM and QTL detection resides in identifying and quantifying rare QTLs with small effects for important agronomic traits that are highly affected by the environment.

Contrary to QTL and AM, application of machine learning methods in genomic selection uses all molecular markers for genomic-enabled prediction of the performance of the candidates for selection (Crossa *et al.*, 2014). The main advantages of genomic selection over phenotype-based selection in plant breeding are that it reduces the cost per cycle and the time required for variety development (Isik, 2014). Therefore, plants provide an excellent framework for applications of machine related methods in genomic selection because they provide a large number of offspring to be tested, their ability to be cloned and inbred easily, the short life-cycles of annuals, and their potential genomic properties such as high levels of linkage disequilibrium (LD) (Sorkheh *et al.*, 2008). Machine learning approaches provide a larger and more general suite of flexible methods. In plant research, several studies have used genome-enabled prediction methodologies such as Reproducing Kernel Hilbert Spaces Regression (RKHS), Neural Networks (NN), Support Vector

Machines (SVM), Random Forests and Boosting (Piepho *et al.*, 2008; Long *et al.*, 2011; Crossa *et al.*, 2014; Crossa *et al.*, 2017). For example, SVM, NN and kernel methods, have been used to detect various stresses in plants (Singh *et al.*, 2016). In maize and wheat, parametric linear regression models, for example, the Bayesian least absolute shrinkage and selection operator (LASSO) (least absolute shrinkage and selection operator) (BL) and the genomic best linear unbiased prediction (GBLUP), as well as semi-parametric nonlinear models, that is, RKHS and the NN methods have been used to predict genetic values of lines in a breeding population (Crossa *et al.*, 2014; González-Camacho *et al.*, 2016). Also, a number of regression or classification models against some parametric models have been used to predict wheat rust resistance or grey leaf spot resistance in maize (Ornella *et al.*, 2017). The choice and the application of a particular machine learning methodology depends largely on the situation and specific problem with each method having its pros and cons.

2.8.1 Factors affecting the accuracy of genomic selection in plant breeding

2.8.1.1 Effective population size and marker density

Effective population size is the number of randomly mating individuals, is a strong determinant of genomic selection accuracy since it influences the effective number of independent chromosome segments. A small population size inbreeding rate is high, however, in plant species like tea which is self-incompatible, has a large number of effective independent chromosome segments. Thus, a small effective population size is better when applying genomic selection in tea breeding as there are fewer independent chromosome segments to estimate. Moreover, one way in which genomic selection can be exploited is to artificially minimise effective population size by using half-sib or within-family designs. This strategy is potentially a good starting point to apply genomic selection for outbred species such as tea.

The number molecular markers required for genomic selection depends on the extent of linkages disequilibrium, which is determined largely by effective population size and the population structure design. For example, (Kumar *et al.*, 2012) had 2500 markers, where 50% of adjacent markers had an LD >0.2, and reached a genomic selection accuracy of 0.7 in a domesticated apple variety (*Malus domestica* Borkh). This explains why within-family genomic selection designs need fewer markers because markers need to track only the large chromosome segments shared by family members (Hayes *et al.*, 2009). Marker type can also affect the accuracy of

genomic selection. Besides SNPs, some other marker types are available, such as DArT, SSR, RAPD, and RFLP.

In simulated data, (Solberg *et al.*, 2008) compared the genomic selection accuracy achieved when using SNPs and SSRs, and found that it required two to three times more SNP markers to achieve accuracy similar to SSRs, because SSRs have multiple alleles and therefore, contain more information. Furthermore, a study by (Poland *et al.*, 2012) in wheat compared genomic estimated breeding value (GEBV) accuracy when using a similar number of SNPs identified by genotyping-by-sequencing (GBS) and DArTs, and found that using SNPs resulted in higher accuracy. Use of SNPs from GBS is under investigation in many species. GBS may avoid ascertainment bias, leading to a broader allele frequency spectrum across the genome. Also, capturing lower frequency markers could reduce the loss of favourable, low-frequency alleles as a result of implementing genomic selection.

2.8.1.2 Heritability

Heritability is defined as the proportion of the phenotypic variability that is genetic in origin. It has been demonstrated that marker effects can be estimated more accurately when heritability is high in genomic selection (Zhao *et al.*, 2013a). Also, genomic selection is used as an approach to achieve more genetic gains in lowly heritable traits (Iwata and Jannink, 2011). In overall, the heritability of agronomic traits is usually high enough to ensure accurate genomic selection implemented in plants, as long as there are sufficient markers and large-size reference populations.

2.8.1.3 Size and structure of reference population

The reference or training population provides the phenotypic and genotypic information to estimate the marker effects that make up the genomic selection prediction equation. Increasing the number of reference individuals allows for more accurate estimates of marker effects, and consequently, improves the accuracy of GEBV for selected candidates. Daetwyler *et al.* (2013) found that the accuracy of GEBV can be improved by 50% by increasing reference population number from 500 to 2 000.

Previous studies on reference population sizes have reached 4 700 lines in maize (Guo *et al.*, 2012), 1 120 in apples (Kumar *et al.*, 2012) and less than 1 000 in both wheat (Crossa *et al.*,

2010) and loblolly pine (Resende *et al.*, 2012a; Resende *et al.*, 2012c). Whereas genotyping costs have decreased dramatically in recent years, phenotyping costs remain high, particularly due to the requirement of replicates within and across years, and across environments (Jannink *et al.*, 2010). Thus, the collection of phenotypes becomes a limitation to increase the reference population size for genomic selection.

Previous studies have shown the importance of the relationship of validation individuals to the reference population (Daetwyler *et al.*, 2013; Würschum *et al.*, 2013). Increasing the relationships effectively reduces effective population size and the effective number of independent chromosome segments. Thus, generally, the stronger the relationship, the higher is the accuracy that can be achieved. In addition, it is more effective to increase the accuracy of GEBV prediction by increasing the relationship of training and validation populations, rather than by increasing training size with less relatedness to predicted individuals (Riedelsheimer *et al.*, 2013), because it increases the probability that chromosome segments evaluated in training populations also exist in validation sets. Hence, it is necessary for selected candidates to be adequately related to the reference population to maintain accuracy (Daetwyler *et al.*, 2013). In their study, it was found that having diverse reference population ensures that phenotypes related to many different chromosome segments are observed, which facilitates accurate prediction of a wider range of selection candidates. Therefore, it is clear that the design of the reference population affects the predictive ability of genomic selection.

2.8.1.4 Statistical models for prediction

Various statistical models have been developed for genomic selection studies, which can be separated into two categories: parametric and non-parametric models. As a parametric model, ridge regression best linear unbiased prediction (RR-BLUP) was first proposed as a genomic selection model by (Hayes and Goddard, 2001). This model assumes that all marker effects are derived from a normal distribution with a common variance. The GBLUP model is an equivalent model to RR-BLUP (VanRaden, 2008) and makes use of a genomic relationship matrix (GRM) generated by evaluating marker covariance across all individuals (Albrecht *et al.*, 2011). The GRM provides greater resolution in genetic relationships among individuals (Hayes *et al.*, 2009). Thus, genomic selection models that use a GRM often result in higher accuracy than pedigree selection in both animals and plants (Hayes *et al.*, 2009; Crossa *et al.*, 2010). Also, GRM can be

applied even when pedigree records are not available, which is particularly useful in plant species with relatively short commercial breeding history and lack of good pedigree records (Hayes *et al.*, 2013). Partial least squares and principal component regression are also parametric models available for genomic selection (Jannink *et al.*, 2010). These models reduce the number of effects to be estimated by computing latent variables, which are then used for predictions.

Non-parametric models can capture non-additive genetic effects, for example, dominance and epistasis in addition to the additive genetic effects that are evaluated in parametric methods. Thus, they may predict phenotypes better than parametric models, particularly if non-additive effects are important. However, many non-additive effects are not fully passed on to progeny, and thus selection decisions basing on predictions from these models may not lead to an increase in genetic gain in the population. RKHS is a semi-parametric regression approach that adds reproducing kernel functions into the model (Gianola *et al.*, 2006). In a previous study by (Heslot *et al.*, 2012) on the genomic selection of various traits in different plant species with 11 statistical models, such as RR-BLUP, Bayes-LASSO, and RKHS, showed that no single model outperformed the others across all traits. This implies that the effectiveness of the methods may depend on how well their assumptions on marker effects match the genetic architecture of target trait (Clark *et al.*, 2011; Zhao *et al.*, 2013b). Although lack of power due to an insufficient data could also be a reason for not distinguishing between models.

2.8.2 Genomic selection of outbred species

The application of genomic selection in outbred plants requires a family design to artificially reduce effective population size, which achieves a higher accuracy with a smaller reference population. This will require prohibitively large reference populations to attain a meaningful accuracy. In a previous study on two eucalyptus lines with populations size of 11 and 51, respectively, the accuracy was consistent with the higher accuracy being achieved in the lower population size line with 0.80 against 0.65, respectively (Resende *et al.*, 2012b). Increasing marker number is also an effective strategy for improving the accuracy of genomic selection in outbred plants. Crossa *et al.* (2010) used 1 148 SNPs on 300 lines of maize from the same heterotic group and reported an accuracy of 0.5 for grain yield. With the same phenotypes and statistical models, but 55 000 SNPs, the accuracy reached 0.8 for the same trait in (González-Camacho *et al.*, 2012). However, current across-family models remain restricted in outbred

plants by limited marker densities, and moderate accuracy is mainly achieved using within-family or combined-family models. For example, (Resende *et al.*, 2012b) found that, although >70% of SNPs were in common between their two eucalyptus lines (polymorphic SNPs that passed the quality filters in both lines); prediction across lines was still inaccurate. Similarly, limited accuracy was also found for maize in an across-family model with a higher SNP number (~50k) (Guo *et al.*, 2014).

The accuracy of genomic selection is closely-related to trait heritability in outbred species. The accuracy in inbred maize lines varied from 0.26 for stalk lodging ($h^2 = 0.55$) to 0.74 for root lodging ($h^2 = 0.75$) using the same reference population (Lorenzana and Bernardo, 2009). Such correlation between GEBV accuracy and trait heritability also exists in trees. For example, in loblolly pine, accuracy was 0.17 for lignin, and tripled to 0.51 for branch angle mean; heritability was also tripled, with $h^2 = 0.11$ for lignin and 0.33 for branch angle mean (Resende *et al.*, 2012c). Moreover, genomic selection accuracy in within-family designs for outbred species appears to be less sensitive to various statistical models, with comparable accuracies being achieved. For instances, both RR-BLUP and Bayes-LASSO achieved accuracies >0.7 for six traits in a bi-parental design (Kumar *et al.*, 2012), and comparable accuracy for maize grain yield between Bayes-LASSO and RKHS was found by (González-Camacho *et al.*, 2012).

2.8.3 Other challenges in genomic selection

The genotype x environment (GxE) interaction is an important factor that should be taken into account when considering the application of genomic selection. Crossa *et al.* (2010) applied genomic selection for wheat grain yield with a total of 599 lines across four environments. They analysed the prediction ability by cross-validation for each environment, respectively, and found that accuracy could vary significantly (0.44-0.6) under different environments. Similarly, their results also revealed a large difference in accuracy for maize grain yield between well-watered and severe drought condition (0.52 and 0.41). A further study with more environments (Crossa *et al.*, 2011) confirmed that accuracy varied greatly in different environments. Moreover, two studies applied genomic selection on loblolly pines with a similar number of individuals and markers (Resende *et al.*, 2012a; Resende *et al.*, 2012c). Resende *et al.* (2012a) allocated clonal individuals into four locations and found that the equation from any one location could provide good prediction within site (~0.7), but the prediction was much lower across the sites. It

demonstrates that GxE interactions can cause significant differences in genomic selection accuracy even for the identical individuals (clones) in different environments. Thus, it is important to take environmental effects into account, which generally are fitted as fixed effects in genomic prediction models.

References

- Abdurakhmonov, I., Kohel, R., Yu, J., Pepper, A., Abdullaev, A., Kushanov, F., Salakhutdinov, I., Buriev, Z., Saha, S. & Scheffler, B. (2008). Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics* 92(6): 478-487.
- Abdurakhmonov, I. Y. & Abdukarimov, A. (2008). Application of association mapping to understanding the genetic diversity of plant germplasm resources. *International Journal of Plant Genomics* 2008.
- Ackerman, W. L. (1970). Interspecific hybridization of *Camellia*. *Amer. Camellia Yearb* 1970: 65-79.
- Ackerman, W. L. (1973). Species compatibility relationships within the genus *Camellia*. *Journal of Heredity* 64(6): 356-358.
- Agarwal, B. (1989). Factors affecting quality of tea during processing.
- 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.
- Agarwal, N. & Das, S. (2009). Dissecting molecular mechanisms related to blister blight disease in tea an est approach. In *Proceedings of the INdo-US workshop on Epigenetic Regulation and Genome Control*, 99-99 Hyderabad, India.
- Ahmed, P. & Das, S. (2009). A functional genomic approach to analyze related transcripts in tea. In *Proceedings of the INdo-US workshop on Epigenetic Regulation and Genome Control*, 100 Hyderabad, India.
- Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M., Simianer, H. & Schön, C.-C. (2011). Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics* 123(2): 339.
- Alonso-Blanco, C. & Méndez-Vigo, B. (2014). Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. *Current Opinion in Plant Biology* 18: 37-43.
- Ananingsih, V. K., Sharma, A. & Zhou, W. (2013). Green tea catechins during food processing and storage: a review on stability and detection. *Food Research International* 50(2): 469-479.

- Anaya, A. L., Cruz-Ortega, R. & Waller, G. R. (2006). Metabolism and ecology of purine alkaloids. *Front in Bioscience* 11(Suppl 1): 2354.
- Angaji, S. A. (2011). *Genomics at a glance (I): Molecular markers and genome mapping*. Lambert Academic Publishing.
- Anon. (2001). Annual Report., 22-25: Tea Research Foundation of Kenya, Tea Board of Kenya.
- Anon. (2007). Annual Technical Report. Tea Research Foundation of Kenya. Tea Board of Kenya.
- Anon. (1980). Breeding. Annual Report., 30: Tea Research Foundation of Kenya, Tea Board of Kenya.
- Anon. (1990). Seed garden (barie). Annual Report., 25: Tea Research Foundation of Kenya, Tea Board of Kenya.
- Anon. (1998). Annual Report., 38-41.: Tea Research Foundation of Kenya, Tea Board of Kenya.
- Anon. (1999). Annual Report., 41: Tea Research Foundation of Kenya, Tea Board of Kenya.
- 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., Kato, M. & Crozier, A. (2011). Distribution, biosynthesis and catabolism of methylxanthines in plants. In *Methylxanthines*, 11-31: Springer.
- 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.
- Banerjee, B. (1987). Can Leaf Aspect Affect Herbivory? A Case Study with Tea. *Ecology* 68(4): 839-843.
- Banerjee, B. (1988). Cup that cheers. *Nature* 332: 580.
- Banerjee, B. (1992). Botanical classification of tea. In *Tea*, 25-51: Springer.
- Barua, D. N. (1963). Selection of vegetative clone. In *Tea Encyclopaedia* Vol. 163, 32-88.
- Basu, R. & Choudhury, M. (1984). How plucking & pruning affect quality of plains tea. *Two & a Bud* 31: 19-21.
- Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Science* 48(5): 1649-1664.

- Bezbaruah, H. (1971). Cytological studies on the Thea and related *Camellias*. *Unpublished Ph. D. Thesis, Gauhati*.
- Bezbaruah, H. (1974). Tea breeding a review. *Indian Journal of Genetics and Plant Breeding* 34: 89-100.
- Bezbaruah, H. (1987). Use of interspecific hybrids in tea breeding. *Two and a Bud* 34: 1-4.
- Bezbaruah, H. & Gogoi, S. (1972). An interspecific hybrid between tea (*Camellia sinensis* L.) and *C. Japonica* L. *Proceedings of the Indian Academy of Science -Section B* 76(5): 219-220.
- Biswas, A. K., Biswas, A. K. & Sarkar, A. (1971). Biological and chemical factors affecting the valuations of northeast Indian plains teas: II. Statistical evaluation of the biochemical constituents and their effects on briskness, quality and cash valuations of black teas. *Journal of the Science of Food and Agriculture* 22(4): 196-204.
- Biswas, A. K., Sarkar, A. R. & Biswas, A. K. (1973). Biological and chemical factors affecting the valuation of North East Indian plains teas. III. Statistical evaluation of the biochemical constituents and their effects on color, brightness and strength of black teas. *Journal of the Science of Food and Agriculture* 24(12): 1457-1477.
- Botheju, W., Abeysinghe, I. & Herath, N. (2000). Effect of standard of plucking on quality and profitability of made tea produced in the upcountry Uva region during the non flavour season. *Sri Lanka Journal of Tea Science* 66(1/2): 36-41.
- Bravo, L. (1998). Polyphenols: chemistry, dietary sources, metabolism, and nutritional significance. *Nutrition Reviews* 56(11): 317-333.
- Burdon, R. D. & Wilcox, P. L. (2011). Integration of molecular markers in breeding. *Genetics, genomics and breeding of conifers*: 276-322.
- 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.
- Cabrera, C., Giménez, R. & López, M. C. (2003). Determination of tea components with antioxidant activity. *Journal of Agricultural and Food Chemistry* 51(15): 4427-4435.
- 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.

- Carlson, J., Tulsieram, L., Glaubitz, J., Luk, V., Kauffeldt, C. & Rutledge, R. (1991). Segregation of random amplified DNA markers in F₁ progeny of conifers. *TAG Theoretical and Applied Genetics* 83(2): 194-200.
- Cervera, M. T., Gusmão, J., Steenackers, M., Peleman, J., Storme, V., Broeck, A. V., Van Montagu, M. & Boerjan, W. (1996). Identification of AFLP molecular markers for resistance against *Melampsora larici-populina* in *Populus*. *TAG Theoretical and Applied Genetics* 93(5-6): 733-737.
- Chaturvedula, V. S. P. & Prakash, I. (2011). The aroma, taste, colour and bioactive constituents of tea. *Journal of Medicinal Plants Research* 5(11): 2110-2124.
- Chen, J., Wang, P., Xia, Y., Xu, M. & Pei, S. (2005). Genetic diversity and differentiation of *Camellia sinensis* L.(cultivated tea) and its wild relatives in Yunnan province of China, revealed by morphology, biochemistry and allozyme studies. *Genetic Resources and Crop Evolution* 52(1): 41-52.
- Chen, L., Apostolides, Z. & Chen, Z. M. (2012a). *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.
- Chen, Y. J., Kuo, P. C., Yang, M. L., Li, F. Y. & Tzen, J. T. (2013b). Effects of baking and aging on the changes of phenolic and volatile compounds in the preparation of old Tieguanyin oolong teas. *Food Research International* 53(2): 732-743.
- 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.
- Clark, S. A., Hickey, J. M. & Van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution* 43(1): 18.
- Cloughley, J. B. (1980). The effect of fermentation temperature on the quality parameters and price evaluation of Central African black teas. *Journal of the Science of Food and Agriculture* 31(9): 911-919.

- Collard, B., 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.
- Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R., Dreisigacker, S. & Yan, J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*.
- 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.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., Camacho-González, J. M., Pérez-Elizalde, S. & Beyene, Y. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*.
- Crossa, J., Perez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R. & Li, Y. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112(1): 48.
- Crouzillat, D., Ménard, B., Mora, A., Phillips, W. & Pétiard, V. (2000). Quantitative trait analysis in Theobroma cacao using molecular markers. *Euphytica* 114(1): 13-23.
- da Silva Pinto, M. (2013). Tea: A new perspective on health benefits. *Food Research International* 53(2): 558-567.
- Daetwyler, H., Hayden, M., Bansal, U., Bariana, H. & Hayes, B. (2013). Genomic selection for disease and morphological traits in diverse wheat landraces. *Plant and Animal Genomes* 21: 11-16.
- Das, R. & Bhagobaty, R. (2010). Assessment of the genetic variability present in the tea germplasm of a tea garden in Assam, India using morphological and biochemical markers. *Tea Leaf*, 36(3): 133-139.
- Das, S. C., Das, S. & Hazarika, M. (2012). Breeding of the tea plant (*Camellia sinensis*) in india. In *Global Tea Breeding*, 69-124: Springer.
- Das, S. K., Sabhapondit, S., Ahmed, G. & Das, S. (2013). Biochemical evaluation of triploid progenies of diploid x tetraploid breeding populations of *Camellia* for genotypes rich in catechin and caffeine. *Biochemical Genetics* 51(5-6): 358-376.

- Deng, W.-W., Ogita, S. & Ashihara, H. (2010). Distribution and biosynthesis of theanine in Theaceae plants. *Plant Physiology and Biochemistry* 48(1): 70-72.
- Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F. & Aprea, G. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345(6201): 1181-1184.
- Dev, C. & Bajaj, K. (1980). Biochemical changes during withering of tea shoots. *Two and a Bud* 27(1): 13-16.
- Drenkard, E., Richter, B. G., Rozen, S., Stutius, L. M., Angell, N. A., Mindrinos, M., Cho, R. J., Oefner, P. J., Davis, R. W. & Ausubel, F. M. (2000). A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in *Arabidopsis*. *Plant Physiology* 124(4): 1483-1492.
- Dutta, S. (1960). Experiments on pruning and tipping in relation to plucking and yield. *Two and a Bud* 7(1): 9-21.
- Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C. & Neale, D. B. (2010). Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., *Pinaceae*). *Genetics* 185(3): 969-982.
- Edwards, D., Forster, J. W., Chagne, D. & Batley, J. (2007). What Are SNPs? *Association mapping in plants*: 41-52.
- 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): 22-33.
- Esposito, E., Rotilio, D., Di Matteo, V., Di Giulio, C., Cacchio, M. & Algeri, S. (2002). A review of specific dietary antioxidants and the effects on biochemical mechanisms related to neurodegenerative processes. *Neurobiology of Aging* 23(5): 719-735.
- Fang, D. & Roose, M. (1997). Identification of closely related citrus cultivars with inter-simple sequence repeat markers. *TAG Theoretical and Applied Genetics* 95(3): 408-417.
- FAO (2015a). World tea production and trade current and future development. 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 (2015b). Kenya's tea sector under climate change: An impact assessment and formulation of a climate smart strategy. (Eds A. Elbehri, B. Cheserek, A. Azapagic, D. Raes, M. Mwale, J. Nyengena, P. Kiprono and C. Ambasa). Rome, Italy.
- Fernie, A. R. & Schauer, N. (2009). Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics* 25(1): 39-48.
- Flint-Garcia, S. A., Thornsberry, J. M. & Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54(1): 357-374.
- Gale, M. & Devos, K. (1998). Plant comparative genetics after 10 years. *Science* 282(5389): 656-659.
- Gao, Y., Rankin, G. O., Tu, Y. & Chen, Y. C. (2016). Theaflavin-3, 3'-digallate decreases human ovarian carcinoma OVCAR-3 cell-induced angiogenesis via Akt and Notch-1 pathways, not via MAPK pathways. *International Journal of Oncology* 48(1): 281-292.
- Gianola, D., Fernando, R. L. & Stella, A. (2006). Genomic assisted prediction of genetic value with semi-parametric procedures. *Genetics*.
- González-Camacho, J., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J., Mahuku, G., Babu, R. & Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *TAG Theoretical and Applied Genetics* 125(4): 759-771.
- González-Camacho, J. M., Crossa, J., Pérez-Rodríguez, P., Ornella, L. & Gianola, D. (2016). Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genomics* 17(1): 208.
- Grandillo, S. & Cammareri, M. (2016). Molecular mapping of quantitative trait loci in tomato. In *The Tomato Genome*, 39-73: Springer.
- Green, M. (1971). An evaluation of some criteria used in selecting large-yielding tea clones. *The Journal of Agricultural Science* 76(1): 143-156.
- Green, M. J. (1966). Clonal selection in seedling stump nurseries. *Tea in East Africa* 6(4): 11-12.
- Green, M. J. (1969). New Clonal Release from the TRI. *Tea* 10(1): 15-18.
- Gunasekera, M. & Ranatunga, M. (2003). Polyploidy in tea (*Camellia sinensis* L.) and its application in tea breeding. a review.
- Guo, M. (2013). *Functional foods: principles and technology*. Elsevier.

- Guo, Z., Tucker, D. M., Lu, J., Kishore, V. & Gay, G. (2012). Evaluation of genome-wide selection efficiency in maize nested association mapping populations. *TAG Theoretical and Applied Genetics* 124(2): 261-275.
- Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D. & Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. *TAG Theoretical and Applied Genetics* 127(3): 749-762.
- Gupta, P., Roy, J. & Prasad, M. (2001). Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science* 80(4): 524-535.
- Gupta, P. & Varshney, R. (2000). The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113(3): 163-185.
- 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.
- Gupta, P. K., Varshney, R. K. & Prasad, M. (2002). Molecular markers: principles and methodology. In *Molecular techniques in crop improvement*, 9-54: 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): 346-355.
- Hainsworth, E. (1965). The 1965 clonal release. *Tea* 6(1): 15.
- Hajra, N. G. (2001). *Tea cultivation: comprehensive treatise*. International Book Distributing Co.
- Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* 8(29): 299-309.
- Han, B. & Huang, X. (2013). Sequencing-based genome-wide association study in rice. *Current Opinion In Plant Biology* 16(2): 133-138.
- Hara, Y., Luo, S., Wickremashinghe, R. & Yamanishi, T. (1995). Botany (of tea). *Food Reviews International* 11: 371-374.
- Hasselo, H. (1964). Productivity gradients on sloping tea land in Ceylon.
- Hayes, B. & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4): 1819-1829.

- Hayes, B. J., Cogan, N. O., Pembleton, L. W., Goddard, M. E., Wang, J., Spangenberg, G. C. & Forster, J. W. (2013). Prospects for genomic selection in forage plant species. *Plant Breeding* 132(2): 133-143.
- Hayes, B. J., Visscher, P. M. & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research* 91(1): 47-60.
- He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H. & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science* 5: 484.
- Heslot, N., Yang, H. P., Sorrells, M. E. & Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science* 52(1): 146-160.
- Higashi, Y. & Saito, K. (2013). Network analysis for gene discovery in plant-specialized metabolism. *Plant, Cell and Environment* 36(9): 1597-1606.
- Hilal, Y. & Engelhardt, U. (2007). Characterisation of white tea-Comparison to green and black tea. *Journal für Verbraucherschutz und Lebensmittelsicherheit* 2(4): 414-421.
- 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.
- Hodkinson, T. R., Chase, M. W. & Renvoize, S. A. (2002). Characterization of a genetic resource collection for *Miscanthus* (*Saccharinae*, *Andropogoneae*, *Poaceae*) using AFLP and ISSR PCR. *Annals of Botany* 89(5): 627-636.
- HP, B. (1968). Genetic improvement of tea in North East India-Its problems and possibilities. *Indian Journal of Genetics and Plant Breeding*: 126-&.
- Hu, C. Y., Lee, T. C., Tsai, H. T., Tsai, Y. Z. & Lin, S. F. (2013). Construction of an integrated genetic map based on maternal and paternal lineages of tea (*Camellia sinensis*). *Euphytica* 191(1): 141-152.
- Hu, C.-Y., Tsai, Y.-Z. & Lin, S.-F. (2014). Development of STS and CAPS markers for variety identification and genetic diversity analysis of tea germplasm in Taiwan. *Botanical Studies* 55(1): 12.
- Huang, F., Liang, Y., Lu, J. & Chen, R. (2006). Genetic mapping of first generation of backcross in tea by RAPD and ISSR markers. *Journal of Tea Science* 26(3): 171-176.

- Isik, F. (2014). Genomic selection in forest tree breeding: the concept and an outlook to the future. *New Forests* 45(3): 379-401.
- ITC (2009). International Tea Committee: Annual Bulletin of Statistics. *International Tea Committee, Ltd., London, United Kingdom*.
- Iwata, H. & Jannink, J.-L. (2011). Accuracy of genomic selection prediction in barley breeding programs: a simulation study based on the real single nucleotide polymorphism data of barley breeding lines. *Crop Science* 51(5): 1915-1927.
- 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.
- Jannink, J. L., Lorenz, A. J. & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* 9(2): 166-177.
- Jannatdoust, M., Darvishzadeh, R., Ziaiefard, R., Ebrahimi, M. A., Maleki, H. H., Gholinezhad, E. & Hatamnia, A. A. (2016). Analysis of genetic diversity and population structure of confectionery sunflower (*Helianthus annuus* L.) native to Iran. *Journal of Crop Science and Biotechnology* 19(1): 37-44.
- Jansen, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics* 138(3): 871-881.
- Jena, K. K. & Kochert, G. (1991). Restriction fragment length polymorphism analysis of CCDD genome species of the genus *Oryza* L. *Plant Molecular Biology* 16(5): 831-839.
- Jin, D., Xu, Y., Mei, X., Meng, Q., Gao, Y., Li, B. & Tu, Y. (2013). Antiobesity and lipid-lowering effects of theaflavins on high-fat diet induced obese rats. *Journal of Functional Foods* 5(3): 1142-1150.
- Jin, J.-Q., Yao, M.-Z., Ma, C.-L., Ma, J.-Q. & Chen, L. (2016). Association mapping of caffeine content with TCS1 in tea plant and its related species. *Plant Physiology and Biochemistry* 105: 251-259.
- Jochmann, N., Lorenz, M., von Krosigk, A. I., Martus, P., Böhm, V., Baumann, G., Stangl, K. & Stangl, V. (2008). The efficacy of black tea in ameliorating endothelial function is equivalent to that of green tea. *British Journal of Nutrition* 99(4): 863-868.
- Kamunya, S. (2003). Tea improvement: strides, constraints & opportunities. *Tea* 24(1): 38-49.
- 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.
- Kamunya, S. M., Wachira, F. N., Lang'at, J., Otieno, W. & Sudoi, V. (2008). Integrated management of root-knot nematode (*Meloidogyne* spp.) in tea (*Camellia sinensis*) in Kenya. *International Journal of Pest Management* 54(2): 129-136.
- Kantety, R. V., La Rota, M., Matthews, D. E. & Sorrells, M. E. (2002). Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Molecular Biology* 48(5): 501-510.
- Kao, C.-H., Zeng, Z.-B. & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* 152(3): 1203-1216.
- Karori, S., Ngure, R., Wachira, F., Wanyoko, J. & Mwangi, J. (2008). Different types of tea products attenuate inflammation induced in *Trypanosoma brucei* infected mice. *Parasitology International* 57(3): 325-333.
- 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.
- Karori, S., Wachira, F., Wanyoko, J. & Ngure, R. (2007). Antioxidant capacity of different types of tea products. *African Journal of Biotechnology* 6(19).
- 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. *TAG 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. (2002b). 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 colored tea cultivars. *Food Chemistry* 136(3): 1405-1413.
- Khan, M. A. & Korban, S. S. (2012). Association mapping in forest trees and fruit crops. *Journal of Experimental Botany* 63(11): 4045-4060.
- Ko, H. J., Lo, C. Y., Wang, B. J., Chiou, R. Y. Y. & Lin, S. M. (2014). Theaflavin-3, 3'-digallate, a black tea polyphenol, attenuates adipocyte-activated inflammatory response of macrophage-associated with the switch of M1/M2-like phenotype. *Journal of Functional Foods* 11: 36-48.
- Ko, H. J., Lo, C. Y., Wang, B. J., Chiou, R. Y. Y. & Lin, S. M. (2015). Theaflavin-3, 3'-digallate, a black tea polyphenol, stimulates lipolysis associated with the induction of mitochondrial uncoupling proteins and AMPK-FoxO3A-MnSOD pathway in 3T3-L1 adipocytes. *Journal of Functional Foods* 17: 271-282.
- 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, K., Wachira, F., Ngure, R., Wanyoko, J., Bii, C. & Karori, S. (2013). Antibacterial and synergistic activity of different tea crude extracts against antibiotic-resistant *S. aureus*, *E. coli* and a clinical isolate of *S. typhi*. *Science Journal of Microbiology* 2013.
- Korte, A. & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9(1): 29.
- Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Annals of Eugenics* 12(1): 172-175.
- Kota, R., Varshney, R. K., Thiel, T., Dehmer, K. J. & Graner, A. (2001). Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum Vulgare* L.). *Hereditas* 135(23): 145-151.
- Kraakman, A. T., Niks, R. E., Van den Berg, P. M., Stam, P. & Van Eeuwijk, F. A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 168(1): 435-446.

- KTDA (2006). ChaiNews. KTDA. Nairobi.
- Kulasegeram, S., Shanmugarajah, V. & Senanayake, Y. (1991). Nursery plant attributes as criteria for selection of new tea clones.
- Kumar, S., Chagné, D., Bink, M. C., Volz, R. K., Whitworth, C. & Carlisle, C. (2012). Genomic selection for fruit quality traits in apple (*Malus domestica* Borkh.). *PloS one* 7(5): e36674.
- Lamb, J. (1939). A note on the manufacture of leaf from selected individual bushes.
- Langat, J., Otieno, W. & Musau, J. (1998). Evaluation of some Kenya tea (*Camellia sinensis*) clones for resistance/susceptibility to *Pestalotiopsis Theae* (Sawada) as influenced by some chemical attributes of mature green leaf. *Tea Board of Kenya (Kenya)*.
- 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.
- 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, X., Ye, T., Huang, Q., Fu, D., Zhang, C. & Zeng, L. (2005). Study on distant hybridization for commercial tea production. In *2005 international symposium on innovation in tea science and sustainable development in tea Industry*, TRA, CAAS, China Tea Science Society, 389-395 Hangzhou, China.
- Liang, Y., Lu, J., Zhang, L., Wu, S. & Wu, Y. (2003). Estimation of black tea quality by analysis of chemical composition and color difference of tea infusions. *Food Chemistry* 80(2): 283-290.
- Lin, S. D., Mau, J. L. & Hsu, C. A. (2012). Bioactive components and antioxidant properties of γ -aminobutyric acid (GABA) tea leaves. *LWT-Food Science and Technology* 46(1): 64-70.
- Liu, S. & Anderson, J. A. (2003). Targeted molecular mapping of a major wheat QTL for *Fusarium* head blight resistance using wheat ESTs and synteny with rice. *Genome* 46(5): 817-823.
- Liu, X., Tang, L., Wu, H., Xi, W., Yu, J. & Zhou, Z. (2016). Development of DArT markers and evaluation of phylogenetic relationship of key Citrus species. *Genetic Resources and Crop Evolution* 63(8): 1307-1318.

- Loh, J. P., Kiew, R., Set, O., Gan, L. H. & Gan, Y. Y. (2000). A study of genetic variation and relationships within the bamboo subtribe *Bambusinae* using amplified fragment length polymorphism. *Annals of Botany* 85(5): 607-612.
- Long, N., Gianola, D., Rosa, G. J. & Weigel, K. A. (2011). Marker-assisted prediction of non-additive genetic values. *Genetica* 139(7): 843-854.
- Lopez, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J. & Verdier, V. R. (2005). Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). *TAG Theoretical and Applied Genetics* 110(3): 425-431.
- Lorenzana, R. E. & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *TAG Theoretical and Applied Genetics* 120(1): 151-161.
- 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 QTL for catechins content in tea plant (*Camellia sinensis*). *PloS one* 9(3): e93131.
- Ma, J.Q., Huang, L., Ma, C.L., Jin, J.Q., Li, C.F., Wang, R.K., Zheng, H.K., Yao, M.Z. & Chen, L. (2015). Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). *PloS one* 10(6): e0128798.
- Magoma, G., Wachira, F., Imbuga, M. & Agong, S. (2003). Biochemical differentiation in *Camellia sinensis* and its wild relatives as revealed by isozyme and catechin patterns. *Biochemical Systematics and Ecology* 31(9): 995-1010.
- 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.
- Mamati, G. E., Wachira, F. N. & Njuguna, C. K. (2001). 2001 TRFK released clones. *Tea* 22(1): 6-7.
- Mandel, S. & Youdim, M. B. (2004). Catechin polyphenols: neurodegeneration and neuroprotection in neurodegenerative diseases. *Free Radical Biology and Medicine* 37(3): 304-317.

- Mao, J. T., Nie, W. X., Tsu, I. H., Jin, Y. S., Rao, J. Y., Lu, Q. Y., Zhang, Z. F., Go, V. L. W. & Serio, K. J. (2010). White tea extract induces apoptosis in non-small cell lung cancer cells: the role of peroxisome proliferator-activated receptor- γ and 15-lipoxygenases. *Cancer Prevention Research* 3(9): 1132-1140.
- Mbuthia, K., Wachira, F., Ngure, R., Mireji, P. & Wachira, S. (2016). Antimutagenic effect of Kenyan tea cultivars in a bacterial test system. *The Pharma Innovation* 3(8, Part A).
- McDowell, I., Feakes, J. & Gay, C. (1991). Phenolic composition of black tea liquors as a means of predicting price and country of origin. *Journal of the Science of Food and Agriculture* 55(4): 627-641.
- 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.
- Miles, C. & Wayne, M. (2008). Quantitative trait locus (QTL) analysis. *Nature Education* 1(1): 208.
- Miller, J. & Tanksley, S. (1990). RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. *TAG Theoretical and Applied Genetics* 80(4): 437-448.
- Mishra, P. K., Fox, R. T. & Culham, A. (2003). Inter-simple sequence repeat and aggressiveness analyses revealed high genetic diversity, recombination and long-range dispersal in *Fusarium culmorum*. *Annals of Applied Biology* 143(3): 291-301.
- Mishra, R. K. & Sen-Mandi, S. (2004b). Genetic diversity estimates for Darjeeling tea clones based on amplified fragment length polymorphism markers. *Journal of Tea Science* 24(2): 86-92.
- Mitchelson, K. R. (2003). The use of capillary electrophoresis for DNA polymorphism analysis. *Molecular Biotechnology* 24(1): 41-68.
- Miyoshi, N., Pervin, M., Suzuki, T., Unno, K., Isemura, M. & Nakamura, Y. (2015). Green tea catechins for well-being and therapy: prospects and opportunities. *Botanics: Targets Therapy* 5: 85-96.
- Mohanpuria, P., Kumar, V. & Yadav, S. K. (2010). Tea caffeine: metabolism, functions, and reduction strategies. *Food Science and Biotechnology* 19(2): 275-287.

- Moncada, M. D. P., Tovar, E., Montoya, J. C., González, A., Spindel, J. & McCouch, S. (2016). A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. *Tree Genetics & Genomes* 12(1): 5.
- Mondal, T. K. (2002). Assessment of genetic diversity of tea (*Camellia sinensis* (L.) O. Kuntze) by inter-simple sequence repeat polymerase chain reaction. *Euphytica* 128(3): 307-315.
- Mondal, T. K. (2007). Tea. *Compendium of Transgenic Crop Plants*.
- Mondal, T. K. (2014). *Breeding and Biotechnology of Tea and Its Wild Species*. Springer Science & Business Media.
- 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.
- Morales, M., Roig, E., Monforte, A., Arus, P. & Garcia-Mas, J. (2004). Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). *Genome* 47(2): 352-360.
- 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.
- Nagaoka, T. & Ogihara, Y. (1997). Applicability of inter-simple sequence repeat polymorphisms in wheat for use as DNA markers in comparison to RFLP and RAPD markers. *TAG Theoretical and Applied Genetics* 94(5): 597-602.
- Nagarajah, S. & Ratnasuriya, G. (1981). Clonal variability in root growth and drought resistance in tea (*Camellia sinensis*). *Plant and Soil* 60(1): 153-155.
- Nagaraju, J., Kathirvel, M., Kumar, R. R., Siddiq, E. & Hasnain, S. E. (2002). Genetic analysis of traditional and evolved Basmati and non-Basmati rice varieties by using fluorescence-based ISSR-PCR and SSR markers. *Proceedings of the National Academy of Sciences* 99(9): 5836-5841.
- Nanjo, F., Goto, K., Seto, R., Suzuki, M., Sakai, M. & Hara, Y. (1996). Scavenging effects of tea catechins and their derivatives on 1, 1-diphenyl-2-picrylhydrazyl radical. *Free Radical Biology and Medicine* 21(6): 895-902.
- 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): 200-207.

- Nitin, S., 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.
- Njuguna, C. K. (1985). 1986 TRFK Released clones. *Tea* 6(2): 4-5.
- Nyirenda, H. (1989). Effectiveness of assessment of vigor and productivity in young vs old bushes and mature clones of tea (*Camellia sinensis*). *Annals of Applied Biology* 115(2): 327-332.
- 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. (1993). Changes in black tea quality due to different fermentation temperature regimes and time during manufacture of slow fermenting Clone S15/10 in a commercial factory. *Tea* 14(2): 123-135.
- 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., Mang'oka, R. & Kavoi, M. M. (2004). Changes in thearubigin fractions and theaflavin levels due to variations in processing conditions and their influence on black tea liquor brightness and total color. *Food Chemistry* 85(2): 163-173.
- 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.
- Obeed, R., Harhash, M. & Abdel-Mawgood, A. (2008). Fruit properties and genetic diversity of five ber (*Ziziphus mauritiana* Lamk.) cultivars. *Pakistan Journal of Biological Sciences* 11: 888-893.
- Ogino, A., Tanaka, J., Taniguchi, F., Yamamoto, M. P. & Yamada, K. (2009). Detection and characterization of caffeine-less tea plants originated from interspecific hybridization. *Breeding Science* 59(3): 277-283.
- Olson, M., Hood, L., Cantor, C. & Botstein, D. (1989). A common language for physical mapping of the human genome. *Science* 245(4925): 1434-1435.
- Orel, G. & Wilson, P. G. (2012). *Camelliacherryana* (Theaceae), a new species from China. In *Annales Botanici Fennici*, Vol. 49, 248-254: BioOne.

- Ornella, L., González-Camacho, J. M., Dreisigacker, S. & Crossa, J. (2017). Applications of Genomic Selection in Breeding Wheat for Rust Resistance. In *Wheat Rust Diseases*, 173-182: Springer.
- Owuor, P. (1982). Can theaflavins content alone be adequate parameter in black tea quality estimation? A review. *Tea* 3: 36-40.
- Owuor, P., Mutea, M., Obanda, A. & Reeves, S. (1986b). Effects of withering on some quality parameters of black tea. *Preliminary results. Tea 1986b* 7: 13-17.
- Owuor, P. & Obanda, M. (1993). Fermentation temperature and duration effects on black tea quality. *Tea Board of Kenya (Kenya)*.
- Owuor, P. & Obanda, M. (1998). The use of chemical parameters as criteria for selecting for quality in clonal black tea in Kenya. Achievement, problems and prospects. *Tea Board of Kenya (Kenya)*.
- Owuor, P., Obanda, M., Apostolides, Z., Wright, L., Nyirenda, H. & Mphangwe, N. (2006b). The relationship between the chemical plain black tea quality parameters and black tea color, brightness and sensory evaluation. *Food Chemistry* 97: 644-653.
- Owuor, P. & Orchard, J. (1990a). Changes in the plain tea quality parameters, individual theaflavins composition and tasters' evaluation of black tea due to extent of two-stage wither, fermentation time and time of plucking. *Tea* 11(1): 29-33.
- Owuor, P. & Orchard, J. (1990b). Changes in the quality and chemical composition of black tea due to degree of physical wither, condition and duration of fermentation. *Tea* 11(2): 109-117.
- Owuor, P. & Othieno, C. (1991). Response of black tea quality to locality in smallholder tea farms. *Tea* 12: 41-45.
- Owuor, P. O., Obanda, M., Nyirenda, H. E., Mphangwe, N. I., Wright, L. P. & Apostolides, Z. (2006a). 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.
- Owuor, P. O., Reeves, S. G. & Wanyoko, J. K. (1986a). Correlation of theaflavins content and valuations of Kenyan black teas. *Journal of the Science of Food and Agriculture* 37(5): 507-513.

- Owuor, P. O., Takeo, T., Horita, H., Tsushida, T. & Murai, T. (1987). Differentiation of clonal teas by terpene index. *Journal of the Science of Food and Agriculture* 40(4): 341-345.
- Pan, H., Gao, Y. & Tu, Y. (2016). Mechanisms of body weight reduction by black tea polyphenols. *Molecules* 21(12): 1659.
- Pan, M. H., Lai, C. S., Wang, H., Lo, C. Y., Ho, C. T. & Li, S. (2013). Black tea in chemoprevention of cancer and other human diseases. *Food Science and Human Wellness* 2(1): 12-21.
- Pang, Y., Abeysinghe, I. S. B., He, J., He, X., Huhman, D., Mewan, K. M., Sumner, L. W., Yun, J. & Dixon, R. A. (2013). Functional characterization of proanthocyanidin pathway enzymes from tea and their application for metabolic engineering. *Plant Physiology* 161(3): 1103-1116.
- Parks, C. R. & Kondo, K. (1974). Breeding studies in the genus *Camellia* (Theaceae). I. A chemotaxonomic analysis of synthetic hybrids and backcrosses involving *Camellia japonica* and *C. saluenensis*. *Brittonia* 26(4): 321-332.
- Paul, S., Wachira, F., Powell, W. & Waugh, R. (1997). Diversity and genetic differentiation among populations of Indian and Kenyan tea (*Camellia sinensis* (L.) O. Kuntze) revealed by AFLP markers. *TAG Theoretical and Applied Genetics* 94(2): 255-263.
- Perovic, D., Stein, N., Zhang, H., Drescher, A., Prasad, M., Kota, R., Kopahnke, D. & Graner, A. (2004). An integrated approach for comparative mapping in rice and barley with special reference to the Rph16 resistance locus. *Functional and Integrative Genomics* 4(2): 74-83.
- Piepho, H., Möhring, J., Melchinger, A. & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161(1-2): 209-228.
- 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.
- Powell, W., Machray, G. C. & Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* 1(7): 215-222.
- Preedy, V. R. (2012). *Tea in health and disease prevention*. Academic Press.

- Racedo, J., Gutiérrez, L., Perera, M. F., Ostengo, S., Pardo, E. M., Cuenya, M. I., Welin, B. & Castagnaro, A. P. (2016). Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biology* 16(1): 142.
- Rafalski, A. (2002a). Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* 5(2): 94-100.
- Rafalski, J. A. (2002b). Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science* 162(3): 329-333.
- Rahman, F., Bhagavaty, H., Basu, R., Choudhury, R., Das, A., Dutta, S., Gilchrist, R., Sharma, K. & Trinick, J. (1978). Effect of different field management practices on cup characters and valuation of teas [India]. *Two and a Bud*.
- Reddy, M. P., Sarla, N. & Siddiq, E. (2002). Inter-simple sequence repeat (ISSR) polymorphism and its application in plant breeding. *Euphytica* 128(1): 9-17.
- Resende, M., Munoz, P., Acosta, J., Peter, G., Davis, J., Grattapaglia, D., Resende, M. & Kirst, M. (2012a). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytologist* 193(3): 617-624.
- Resende, M. D., Resende, M. F., Sansaloni, C. P., Petrolí, C. D., Missiaggia, A. A., Aguiar, A. M., Abad, J. M., Takahashi, E. K., Rosado, A. M. & Faria, D. A. (2012b). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist* 194(1): 116-128.
- Resende, M. F., Muñoz, P., Resende, M. D., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F. & Kirst, M. (2012c). Accuracy of genomic selection methods in a standard dataset of loblolly pine (*Pinus taeda* L.). *Genetics: genetics*. 111.137026.
- Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J.-L. & Melchinger, A. E. (2013). Genomic predictability of interconnected bi-parental maize populations. *Genetics: genetics*. 113.150227.
- Roberts, E. & Smith, R. (1963). The phenolic substances of manufactured tea. IX.-the spectrophotometric evaluation of tea liquors. *Journal of the Science of Food and Agriculture* 14(10): 689-700.
- Rogers, S. (1975). Preliminary observations on pollen tube incompatibility in some tea clones.

- Salah, N., Miller, N. J., Paganga, G., Tijburg, L., Bolwell, G. P. & Riceevans, C. (1995). Polyphenolic flavanols as scavengers of aqueous phase radicals and as chain-breaking antioxidants. *Archives of Biochemistry and Biophysics* 322(2): 339-346.
- Sánchez-Sevilla, J. F., Horvath, A., Botella, M. A., Gaston, A., Folta, K., Kilian, A., Denoyes, B. & Amaya, I. (2015). Diversity Arrays Technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria* × *ananassa*). *PloS one* 10(12): e0144960.
- Sanderson, G. (1963). The chloroform test-A study of its suitability as a means of rapidly evaluating fermenting properties of clones.
- 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.
- Sannella, A. R., Messori, L., Casini, A., Vincieri, F. F., Bilia, A. R., Majori, G. & Severini, C. (2007). Antimalarial properties of green tea. *Biochemical and Biophysical Research Communications* 353(1): 177-181.
- Schierwater, B. & Ender, A. (1993). Different thermostable DNA polymerases may amplify different RAPD products. *Nucleic Acids Research* 21(19): 4647-4648.
- 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.
- Semagn, K., Bjørnstad, Å. & Ndjiondjop, M. (2006a). An overview of molecular marker methods for plants. *African Journal of Biotechnology* 5(25).
- Semagn, K., Bjørnstad, Å. & Xu, Y. (2010). The genetic dissection of quantitative traits in crops. *Electronic Journal of Biotechnology* 13(5): 16-17.
- Seurei, P. (1996). Tea improvement in Kenya: a review. *Tea Board of Kenya*.
- Sharangi, A. (2009). Medicinal and therapeutic potentialities of tea (*Camellia sinensis* L.)-A review. *Food Research International* 42(5): 529-535.
- Shavrukov, Y. (2016). CAPS markers in plant biology. *Russian Journal of Genetics: Applied Research* 6(3): 279-287.
- Singh, A., Ganapathysubramanian, B., Singh, A. K. & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science* 21(2): 110-124.

- Singh, A., Rai, V., Chand, R., Singh, R. & Singh, M. (2013). Genetic diversity studies and identification of SSR markers associated with Fusarium wilt (*Fusarium udum*) resistance in cultivated pigeon pea (*Cajanus cajan*). *Journal of Genetics* 92(2): 273-280.
- Singh, B. (1995). *Plant Breeding*. New Delhi, India: Kalyani Publishers.
- Singh, I. (1980). Nonconventional approaches in the breeding of tea in North East India. *Two and a Bud* 27(1): 3-6.
- Singh, I. D. & Sharma, P. C. (1982). Studies in radiation breeding in tea plants. In *Proceedings of the 4th Annual Symposium on Plantation Crops*, 1-19.
- Solberg, T., Sonesson, A., Woolliams, J. & Meuwissen, T. (2008). Genomic selection using different marker types and densities. *Journal of animal science* 86(10): 2447-2454.
- Sorkheh, K., Malysheva-Otto, L. V., Wirthensohn, M. G., Tarkesh-Esfahani, S. & Martínez-Gómez, P. (2008). Linkage disequilibrium, genetic association mapping and gene localization in crop plants. *Genetics and Molecular Biology* 31(4): 805-814.
- 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.
- Sudoj, V. (1995). Effect of spraying petroleum white oil. Murphoil on control of scale insects *Aspidiotus* sp. and their effect on natural enemies *Aphytis* sp. *Tea Research Foundation of Kenya (Kenya)*.
- Sudoj, V. (1997). Tea pests with special reference to mites: research achievements and future thrusts. *Tea Board of Kenya (Kenya)*.
- Sudoj, V., Khaemba, B. & Wanjala, F. (1996). Screening of tea clones for their resistance to *Brevipalpus phoenicis* Geikjkes (Acari: *Tenuipalpidae*) attack. *Journal of Plantation Crops* 24: 291-295.
- Sun, Y., Wang, J., Crouch, J. H. & Xu, Y. (2010). Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. *Molecular Breeding* 26(3): 493-511.
- Suzuki, Y., Miyoshi, N. & Isemura, M. (2012). Health-promoting effects of green tea. *Proceedings of the Japan Academy, Series B* 88(3): 88-101.

- Takeda, Y. (1990). Cross compatibility of tea (*Camellia sinensis*) and its allied species in the genus *Camellia*. *JARQ, Japan Agricultural Research Quarterly* 24(2): 111-116.
- Takeo, T. (1984). Effect of the withering process on volatile compound formation during black tea manufacture. *Journal of the Science of Food and Agriculture* 35(1): 84-87.
- 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 colour, and mature leaf size in tea plant (*Camellia sinensis*). *Tree Genetics & Genomes* 12(3): 1-13.
- Tanaka, J. (1996). RAPD linkage map of tea plant and the possibility of application in tea genetics and breeding. *Tea Research Journal* 84(Suppl): 44-45.
- Tanaka, J. (2000). Construction of linkage and QTL analysis of tea plant. In *Proceedings of the International Symposium on Molecular Biology and Tea Breeding*, 23-26.
- 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.
- Tautz, D. & Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* 12(10): 4127-4138.
- 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.
- Tomlins, K. & Mashingaidze, A. (1997). Influence of withering, including leaf handling, on the manufacturing and quality of black teas- A review. *Food Chemistry* 60(4): 573-580.
- Toppino, L., Barchi, L., Scalzo, R. L., Palazzolo, E., Francese, G., Fibiani, M., D'Alessandro, A., Papa, V., Laudicina, V. A. & Sabatino, L. (2016). Mapping quantitative trait loci affecting biochemical and morphological fruit properties in eggplant (*Solanum melongena* L.). *Frontiers in Plant Science* 7.
- Tounekti, T., Joubert, E., Hernández, I. & Munné-Bosch, S. (2013). Improving the polyphenol content of tea. *Critical Reviews in Plant Sciences* 32(3): 192-215.
- Tsumura, Y., Yoshimura, K., Tomaru, N. & Ohba, K. (1995). Molecular phylogeny of conifers using RFLP analysis of PCR-amplified specific chloroplast genes. *TAG Theoretical and Applied Genetics* 91(8): 1222-1236.

- Valcic, S., Muders, A., Jacobsen, N. E., Liebler, D. C. & Timmermann, B. N. (1999). Antioxidant chemistry of green tea catechins. Identification of products of the reaction of (+)-epigallocatechin gallate with peroxy radicals. *Chemical Research in Toxicology* 12(4): 382-386.
- Van der Vossen, H. & Wessel, M. (2000). *Plant resources of South-East Asia. No. 16. Stimulants*. Backhuys Publishers.
- Vankatesh, P., Jaiprakash, M., Prasad, P., Pilia, B., Sadhale, P. P. & Sinka, V. (2007). Flavanoid biosynthesis in *Camellia sinensis*.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy science* 91(11): 4414-4423.
- Visser, T. (1969). Tea (*Camelliasinensis* [L.] O Kuntze). *Miscellaneous Papers. Landbouwhogeschool Wageningen*: 459-493.
- Visser, T. & Kehil, F. (1958). Selection and vegetative propagation of tea.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. V. D., Hornes, M., Friters, A., Pot, J., Paleman, J. & Kuiper, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23(21): 4407-4414.
- Wachira, F. (1994a). Breeding and clonal selection in relation to black tea quality: a review. *Tea Board of Kenya (Kenya)*.
- Wachira, F. (1994c). Triploidy in tea. *C. sinensis*: 53-60.
- Wachira, F. (1996). Development of molecular marker in *Camellia*. PhD Thesis, Dundee University.
- Wachira, F. (2002). Detection of genetic diversity and characterization of Kenyan tea germplasm. *A Tea Genetic Diversity (TGD) Project. TGD Final Project Document, Ed. Tea Research Foundation of Kenya, Kericho, Kenya*.
- Wachira, F. & Kamunya, S. (2005). Kenyan teas are rich in antioxidants. *Tea* 26(2): 81-89.
- Wachira, F. & Kiplangat, J. (1991). Newly identified Kenyan polyploid tea strains. *Tea* 12(1): 10-13.
- Wachira, F. & Ng'etich, W. (1999). Dry-matter production and partition in diploid, triploid and tetraploid tea. *The Journal of Horticultural Science and Biotechnology* 74(4): 507-512.
- Wachira, F. & Ogada, J. (1995a). *In vitro* regeneration of *Camellia sinensis* (L.) O. Kuntze by somatic embryogenesis. *Plant Cell Reports* 14(7): 463-466.

- Wachira, F., Waugh, R., Powell, W. & Hackett, C. (1995). Detection of genetic diversity in tea (*Camellia sinensis*) using RAPD markers. *Genome* 38(2): 201-210.
- Wachira, F. N. (1994b). Triploidy in tea (*Camellia sinensis*): Effect on yield and yield attributes. *Journal of Horticultural Science* 69(1): 53-60.
- Wachira, F. N. (2001). Tea improvement in Kenya: An overview of research achievements, prospects and limitations. In *TBK board of directors open day proceedings, 29 January 2001.*, 12-14 Tea Research Foundation of Kenya.
- Wachira, F. N., Powell, W. & Waugh, R. (1997). An assessment of genetic diversity among *Camellia sinensis* L.(cultivated tea) and its wild relatives based on randomly amplified polymorphic DNA and organelle-specific STS. *Heredity* 78(6).
- Wambulwa, M. C., Meegahakumbura, M. K., Kamunya, S., Muchugi, A., Möller, M., Liu, J., Xu, J. C., Ranjitkar, S., Li, D. Z. & Gao, L. M. (2016). Insights into the genetic relationships and breeding patterns of the African tea germplasm based on nSSR markers and cpDNA sequences. *Frontiers in Plant Science* 7.
- Wang, K., Liu, F., Liu, Z., Huang, J., Xu, Z., Li, Y., Chen, J., Gong, Y. & Yang, X. (2011). Comparison of catechins and volatile compounds among different types of tea using high-performance liquid chromatography and gas chromatography mass spectrometer. *International Journal of Food Science and Technology* 46(7): 1406-1412.
- Waycott, W., Fort, S., Ryder, E. & Michelmore, R. (1999). Mapping morphological genes relative to molecular markers in lettuce (*Lactuca sativa* L.). *Heredity* 82(3): 245-251.
- Weinig, C. & Schmitt, J. (2004). Environmental effects on the expression of quantitative trait loci and implications for phenotypic evolution. *AIBS Bulletin* 54(7): 627-635.
- Weisshaar, B. & Jenkins, G. I. (1998). Phenylpropanoid biosynthesis and its regulation. *Current Opinion in Plant Biology* 1(3): 251-257.
- Welsh, J. & McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research* 18(24): 7213-7218.
- Wettstein-Knowles, P. v. & Shewry, P. (1992). Cloned and mapped genes: current status. *Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology.*: 73-98.
- Wight, W. (1956). Commercial selection and breeding of tea in India. *World Crops* 8: 263-268.
- Wight, W. & Barua, P. (1957). What is tea? *Nature* 179: 506-507.

- Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A. & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* 18(22): 6531-6535.
- Willson, K. C. & Clifford, M. N. (2012). *Tea: Cultivation to consumption*. Springer Science & Business Media.
- Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiology* 126(2): 485-493.
- 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.
- Worland, A., Gale, M. & Law, C. (1987). Wheat genetics. In *Wheat breeding*, 129-171: Springer.
- Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E. & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 14(7): 507.
- Wu, L.Y., Juan, C.C., Hwang, L. S., Hsu, Y.P., Ho, P. H. & Ho, L. T. (2004). Green tea supplementation ameliorates insulin resistance and increases glucose transporter IV content in a fructose-fed rat model. *European Journal of Nutrition* 43(2): 116-124.
- Wu, Z. J., Li, X. H., Liu, Z. W., Xu, Z. S. & Zhuang, J. (2014). De novo assembly and transcriptome characterization: novel insights into catechins biosynthesis in *Camellia sinensis*. *BMC Plant Biology* 14(1): 277.
- Würschum, T. & Kraft, T. (2014). Cross-validation in association mapping and its relevance for the estimation of QTL parameters of complex traits. *Heredity* 112(4): 463-468.
- Würschum, T., Reif, J. C., Kraft, T., Janssen, G. & Zhao, Y. (2013). Genomic selection in sugar beet breeding populations. *BMC Genetics* 14(1): 85.
- 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.
- Xu, J., Xu, Z. & Zheng, W. (2017). A review of the antiviral role of green tea catechins. *Molecules* 22(8): 1337.
- Xu, S. (2003). Theoretical basis of the Beavis effect. *Genetics* 165(4): 2259-2268.
- Xu, Y., Li, P., Yang, Z. & Xu, C. (2017). Genetic mapping of quantitative trait loci in crops. *The Crop Journal* 5(2): 175-184.

- Yang, X., Li, H., Liang, M., Xu, Q., Chai, L. & Deng, X. (2015). Genetic diversity and phylogenetic relationships of citron (*Citrus medica* L.) and its relatives in southwest China. *Tree Genetics and Genomes* 11(6): 129.
- Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M. & Holland, J. B. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38(2): 203.
- Zaveri, N. T. (2006). Green tea and its polyphenolic catechins: medicinal uses in cancer and noncancer applications. *Life Sciences* 78(18): 2073-2080.
- Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136(4): 1457-1468.
- Zhang, H. B., Xia, E. H., Huang, H., Jiang, J. J., Liu, B. Y. & Gao, L. Z. (2015). *De novo* transcriptome assembly of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. *BMC Genomics* 16(1): 298.
- Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C. & Marjoram, P. (2007). An *Arabidopsis* example of association mapping in structured samples. *PLoS Genetics* 3(1): e4.
- Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A. & Mezey, J. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature Communications* 2: 467.
- Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., Ranc, N., Piepho, H. P. & Reif, J. C. (2013a). Choice of shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations. *Plant Breeding* 132(1): 99-106.
- Zhao, Y., Zeng, J., Fernando, R. & Reif, J. C. (2013b). Genomic prediction of hybrid wheat performance. *Crop Science* 53(3): 802-810.
- Zietkiewicz, E., Rafalski, A. & Labuda, D. (1994). Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics* 20(2): 176-183.

Zou, J., Semagn, K., Iqbal, M., N'Diaye, A., Chen, H., Asif, M., Navabi, A., Perez-Lara, E., Pozniak, C. & Yang, R. C. (2016). Mapping QTL Controlling agronomic traits in the 'Attila'×'CDC Go'spring wheat population under organic management using 90K SNP Array. *Crop Science*.

CHAPTER 3

QUANTITATIVE GENETIC PARAMETERS FOR BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS

Abstract

Tea (*Camellia sinensis*) contains polyphenols and caffeine which have been found to be of popular interest in tea quality. Tea production relies on well-distributed rainfall which influences tea quality. Phenotypic data for two segregating tea populations TRFK St. 504 and TRFK St. 524, were collected and used to identify the QTL influencing tea biochemical and drought stress traits based on a consensus genetic map constructed using the DArTseq platform. The populations comprised of 261 F₁ clonal progeny. The map consisted of 15 linkage groups which correspond to chromosome haploid number of tea plant ($2n = 2x = 30$) and spanned 1 260.1 cM with a mean interval of 1.1 cM between markers. A total of 16 phenotypic traits were assessed in the two populations. Both interval and multiple QTL mapping revealed a total of 46 putative QTLs in the 15 linkage groups associated with tea quality and %RWC at a significant genome-wide threshold of 5%. In total, six caffeine QTLs, 25 catechins QTLs, three theaflavins QTLs, nine QTLs for tea taster score and three QTLs for %RWCs were detected. Out of these 46 QTLs, 19 QTLs were identified for ten traits in three main regions on LG01, LG02, LG04, LG12, LG13, and LG14. The QTLs-associated with caffeine, individual catechins, and theaflavins were clustered mostly in LG02 and LG04 but in different regions on the map. The explained variance by each QTL in the population ranged from 5.5 to 56.6%, with an average of 9.9%. Identification of QTLs that are tightly linked to markers associated with black tea quality coupled with UPLC assay may greatly accelerate development of novel tea cultivars owing to its amenability at seedling stage. Also, validated molecular markers will contribute greatly to adoption of marker-assisted selection for drought tolerance and tea quality improvement.

3.1 Introduction

In the development of new crop cultivars, plant breeders are faced with challenges of choosing among many potential selection criteria. In tea plant (*Camellia sinensis*), these selection criteria include yield per hectare, cup quality, and tolerance to both abiotic and biotic factors. Besides the yield, the quality of tea is also important from an economic perspective, as quality is an essential factor influencing the market price (Yan, 2007). The quality of tea, especially, appearance, taste, and smell aspects, is largely determined by the biochemical composition of the fresh tea leaves (Obanda *et al.*, 1997; Dutta *et al.*, 2011). Drought is also a major environmental stress factor that affects the growth and development of tea plants, and the traits associated with drought tolerance are quantitative in nature. Conventional breeding for developing high-quality black tea and drought-tolerant tea varieties is therefore time-consuming and labour-intensive. The %RWC represents an easy measure of water deficit and provides the best sensor for drought stress in a plant.

In a previous study on a method for Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) in *Camellia sinensis* validated by targeted metabolomics (Nyarukowa *et al.*, 2016), was developed for predicting the drought tolerance of tea cultivars. The metabolite profiles obtained using the SWAPDT method showed that drought-tolerant tea cultivars differed from drought susceptible tea cultivars. The SWAPDT method relied on the %RWC of tea leaves after the five-hour withering method. Therefore, the SWAPDT method which was used in the current study provides a basis for selection of new drought-tolerant tea cultivars that may lead to improvement of crop productivity, amidst challenges imposed by drought due to climate change. However, with the development of genomics-assisted breeding (GAB) approaches, such as MAS, can greatly improve the precision and efficiency of selection in crop breeding, including tea (Chen *et al.*, 2012a).

The selection criteria in breeding is complex as many of the traits relevant are quantitative in nature, and are controlled by a large number of genes, each with small effects. Thus, it is generally impossible to determine the specific genotype (or breeding value) of an individual simply by assessing its phenotype (Udall, 2003). Phenotypic assessments also do not indicate how much variation in the trait is the result of environmental influences (van Eeuwijk *et al.*, 2010; Feil and Fraga, 2012). However, quantitative genetics which is the study of the interaction of genetics and environment can inform breeders as to the amount of heritable genetic variation

in traits available for selection (El-Soda *et al.*, 2014). It can also provide an understanding of the genetic relationships between traits that both directly and indirectly affect the phenotype, indicating genetic correlations (Mackay *et al.*, 2009). Understanding the genetic correlations can be used to identify potential indicators where it is difficult or expensive to measure traits directly to avoid selection of unrelated traits (Mackay *et al.*, 2009). Quantitative genetic analysis can also determine the degree to which environmental factors influence trait variation and the correlations between traits (El-Soda *et al.*, 2014). This knowledge is essential for the accurate prediction of genetic gains and the development of breeding strategies. In tea, quantitative genetic analysis of progeny trials has the added benefit of providing a means of assessing the genetic potential for traits of interest that are expressed either in male or female plants. The traits include those that are associated with the yield, cup quality, and abiotic and biotic tolerance. The information gained from quantitative genetic analysis can simplify the tea breeding programs through selection of parents with desirable traits. The selection for high-quality tea requires precise information on the diversity available and also a careful study of the biochemical molecules which contribute towards the quality. Molecular markers can also facilitate the indirect selection for drought tolerance traits that are difficult to score directly making it possible to identify the genomic regions associated with drought tolerance for marker-assisted selection.

3.2 Research objective

The objective of this study was to identify putative QTLs linked to caffeine content, catechins fractions and theaflavins fractions, organoleptic evaluation (tea tasters' score) and percent %RWC by integrating DArTseq markers and UPLC results to construct a high-density genetic linkage map for future molecular assisted selection in tea breeding.

3.3 Hypotheses

Alternative hypothesis (H_1): The DArTseq markers are linked to QTLs for black tea quality parameters and %RWC.

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

3.4 Materials and methods

3.4.1 Plant material

Two pseudo-testcross populations consisting of 109 progeny from TRFK St. 504 (TRFK 303/577 x GW Ejulu) and reciprocal cross consisting of 152 progeny coded TRFK St. 524 (GW Ejulu x TRFK 303/577), were developed by crossing two heterozygous parental clones TRFK 303/577 and GW Ejulu. The two parental clones, TRFK 303/577 and GW Ejulu, are of Assamica and China varieties, respectively. The two clones were chosen on the basis of their contrasting attributes. GW Ejulu is a low-yielding, high black tea quality and moderate levels of caffeine, but high in total catechins and individual catechins contents. Clone TRFK 303/577, which is an open-pollinated progeny of clone TRFK 6/8, is high-yielding, drought-tolerant, medium in black tea quality, caffeine and individual catechins.

3.4.2 Description of study site

The cross comprising all above clonal progenies and their parents were established in 2007 at TRI, in Timbilil Estate (Kenya, Kericho county, latitude 0° 22'S, longitude 35° 21'E, altitude 2180 m AMSL). The trial populations were set up in a complete randomised block design with three replications in plots of 15 plants spaced at 0.61 m within rows and 1.22 m between rows. The populations received 150 Kg N per hectare per year in the form of NPKS 25:5:5:5 compound fertilizer. The trial populations were brought into bearing following the recommended management practices (Anon, 2002; Kamunya *et al.*, 2010).

3.4.3 Sample collection and processing

About 1 Kg of fresh shoots in the form of two leaves and a bud were plucked from heterozygous parents and three independent F₁ progeny in a complete randomised block design between September 2013 and February 2014. The fresh-plucked shoots were placed in appropriately labelled sampling bags for processing at the TRI miniature factory. Five hundred grams of fresh green tea leaves were used to produce green tea by immediately drying using a microwave for 5 minutes, which also deactivated the oxidising enzyme polyphenol oxidase. The remaining 500 g was used to manufacture black tea using a standard black tea manufacturing procedure by physically withering the tea leaves up to a moisture content of 50-65% for 18 hours. The leaf was passed through the CTC rollers four times to achieve maceration equivalent to that obtained using commercial CTC rollers in black tea manufacture (Owuor and Othieno, 1991b). Upon

maceration, the tea (dhool) was then aerated for 90 minutes at ambient temperature (22-26°C) and 100% relative humidity to achieve enzymatic oxidation or so-called “fermentation”. A bench top fluid-bed drier system (TeaCraft Ltd) was utilised in firing the tea, initially at 120°C for 20-25 minutes then brought to 100°C for 10 minutes. Green and black tea samples were grounded separately using a coffee grinder and stored in sealed silver lined aluminium sachets for UPLC analysis and organoleptic assessment for processed black tea.

3.4.4 Extraction of catechins, caffeine and theaflavins

International Organisation for Standardisation (ISO) extraction method was used as described in the ISO14502-2 (2005). Briefly, 0.200 ± 0.001 g of green and black tea samples were weighed out on a Mettler Toledo analytical balance (Microsep, South Africa) and transferred into a 20 ml glass test tube. A five ml volume of 70:30 methanol (Merck, South Africa): water (v/v) at 70°C was added to each extraction tube, stoppered and vortex mixed for approximately five seconds before placing into a water bath set at 70°C. The extraction mixture was vortex mixed after five minutes and after ten minutes. The extracts were cooled and centrifuged at $2000 \times g$ using Thermo Scientific Heraeus Labofuge (Sepsci, South Africa) 300 centrifuge for ten minutes, with the resultant supernatant decanted into a ten ml volumetric flask. The extraction step was repeated twice. Both extracts were pooled, and the volume adjusted to ten ml with cold 70:30 methanol: water (v/v). One ml volume of the extracts was then diluted to five mL with stabilising solution (10% v/v acetonitrile, 500 mg/mL EDTA and 10 mg/ml ascorbic acid, Sigma-Aldrich, South Africa). The dilution was filtered through 0.2 μ m Minisart®RC4 syringe filters (Sartorius, South Africa) with hydrophilic, solvent-resistant regenerated cellulose membranes and ready for UPLC analysis.

3.4.5 UPLC analysis of catechins, caffeine and theaflavins

The UPLC analysis was performed on 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 achieved on Waters Acquity HSS T3 column (1.8 μ m, 2.1 $\times$ 150 mm). The mobile phase was composed of solvent A (2% acetic acid and 9% acetonitrile in water, pH 2.8) and solvent B (2% acetic acid and 80% of acetonitrile in water) with gradient elution: 0 min (5% A; 95% B), 0-21 min (5-20% A; 80-95% B), 21-30 min (20-25% A; 75-80% B), 30-32 min (25-100% A; 0-75% B), 32-39 min

(100-100% A), 39-40 min (100-5% A; 0-95% B), and 40-45 min (5-5% A; 95% B) was used in this study. The mobile phase was filtered through a 0.2 µm cellulose acetate membrane filter and degassed using a Neuberger Laboport (Labotech, South Africa) vacuum pump. The sample injection volume was five µL, and the flow-rate was set at 0.2 mL/min. The standards consisted of catechins (CAT, EC, ECG, EGC and EGCG), caffeine and gallic acid (Sigma-Aldrich, South Africa). Tryptamine, sulfanilamide and mycophenolic acid (Sigma-Aldrich, South Africa) were used as internal standards for quality control. Three mixed standards; mixture A (catechins, caffeine, gallic acid and three internal standards), mixture B and mixture C which both had all the standards except the three internal standards with varying concentrations were used. The identification and quantification of individual catechins, caffeine and individual theaflavins were performed at 278 nm. The individual catechins and caffeine peaks in the tea samples were identified based on retention times of the individual catechins and caffeine standards. The individual theaflavins in black tea samples were quantified as EGCG equivalents, based on the response factor of EGCG.

3.4.6 Determination of plain black tea quality parameters

The tea taster's score on black teas was performed on the 261 progenies and their parents. A sample of black tea (2.5 g) was extracted with 150 ml of boiling water for about 5 minutes. The tea sample was then subjected to organoleptic evaluation made by professional tea tasters (WM Cahn, Johannesburg, South Africa) with scores based on liquor colour (CL), liquor brightness (BRT), astringency (AST), liquor briskness (BRK) and liquor aroma (AR). The scores were in the range of 1-5 with the low score representing poor quality and a high score representing high-quality. Black teas that are known to be of high-quality were used as a control to determine the quality of progenies regarding liquor characteristics and taste.

3.4.7 Determination of percent relative water content

The % RWC in the two tea populations was determined using the SWAPDT method as described by Nyarukowa *et al.* (2016). Three shoots with two leaves and a bud collected from heterozygous parents and three independent F₁ progeny in a complete randomised block design were immersed in water at room temperature and weighed after 24 hours. The leaves were then oven dried at 37°C and weighed after five hours. The leaves were again placed in water and

weighed after 24 hours, and then oven dried at 105°C for 24 hours to obtain each leaf's dry weight.

The % RWC was calculated using the formula:

$$\% RWC = [(F_{wt} - D_{wt}) / (FT_{wt} - D_{wt})] \times 100$$

where F_{wt} is the fresh weight, D_{wt} is the dry weight after 24 h in 105°C oven, and FT_{wt} is the weight after 24 h rehydration.

3.4.8 Phenotypic trait analysis

A total of 16 phenotypic traits were assessed for a normal distribution using the JMP 12 (Sall *et al.*, 2012). The mean and the standard deviation for each phenotypic trait in each parent and the progenies were calculated. The significant difference between parental values and progenies was analysed using Student's t-test.

3.5 DNA analysis

3.5.1 Extraction, purification and quantification of DNA

Two tender young leaves and a terminal bud were plucked from a tea bush planted in a complete randomised block design experiments as a representative sample and preserved in zip-lock plastic bags containing dry silica gel (Malebe, 2011). The silica gel was dried in the oven at 70°C for 48 hours to enable it to absorb any surface moisture from the preserved leaves. The green tea leaf samples were stored at -20°C before DNA extraction using a modified method of (Gawal and Jarret, 1991). The NanoDrop spectrophotometer (NanoDrop Technologies, South Africa) which gives the concentration of DNA in ng/μl was used to quantify the DNA. The integrity of DNA was also checked using the agarose gel electrophoresis method (Adkins *et al.*, 2007).

3.5.2 DArTseq assay

The DNA samples were shipped to Diversity Array Technology Pty Ltd (Canberra, Australia) for DArTseq analysis. DNA quality and digestibility were tested using restriction enzyme *Pst*I (Fermentas, Burlington, Canada) and *Eco*RI (Promega, Madison, USA). The DArTseq procedure was carried out as described by (Sansaloni *et al.*, 2010) using *Pst*I and *Mse*I restriction enzymes. Markers were scored 0 or 1 for absence or presence of a polymorphism in the genomic representation of the sample.

3.5.3 Linkage map construction

Six thousand five hundred and eighty-eight DArTseq markers derived from *C. sinensis* sequences were tested for segregation using genomic DNA from the two parents and 261 F₁ individuals. The obtained genotyping data were subjected to linkage analyses using JoinMap 4.0 (Van Ooijen, 2006). Markers were grouped at a logarithm of odds (LOD) of (3-12) and a recombination frequency of 0.4. The genetic distances between markers were calculated with the Kosambi mapping function. The LOD thresholds at the genome-wide level were determined by running 1 000 permutations ($P < 0.05$).

3.5.4 QTL analyses

The QTL mapping was carried out using the MapQTL 6.0 software (Van Ooijen *et al.*, 2000). The first set of markers were selected as cofactors from interval mapping (IM) results, and the significant markers were selected using backwards elimination method. The significant markers at $P < 0.05$ were only used as cofactors in the multiple MQM method to refine QTL detection. The permutation analysis (1,000 permutations) for significance thresholds was used for accepting the presence of a potential QTL (Churchill and Doerge, 1994). The maximum LOD score on a linkage group was used to estimate the positions of QTL. The location and position of the QTL for individual phenotypic traits was indicated using MapChart software (Voorrips, 2002).

3.6 Results

3.6.1 Phenotypic trait analysis

A wide range of variation in individual catechins, theaflavins and caffeine contents was observed in the two parents and F₁ progeny. A chromatogram representing the profiles of black tea and green tea from the two parent lines is presented in Figure 3.1 and Figure 3.2, respectively. Besides the five peaks identified, other several peaks were detected but unidentified. This is an indication that there are other several biochemical compounds present in the tea extracts.

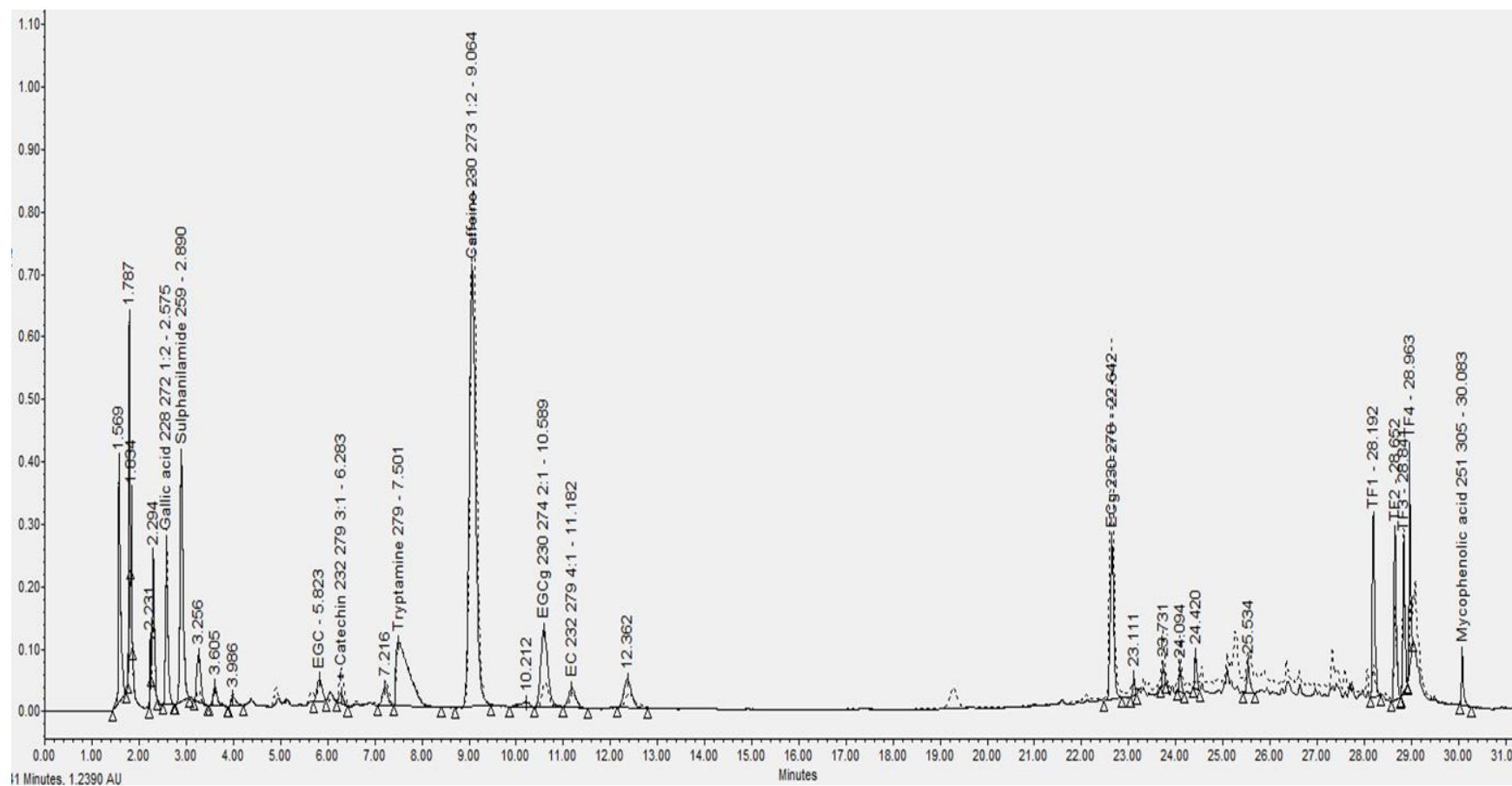


Figure 3.1: A representative UPLC chromatogram of black tea sample from parental lines TRFK 303/577 (solid line) and GW Ejulu (dotted line) with the three internal standards.

KEY: X-axis = Retention time (minutes), Y-axis = Absorbance units (278 nm)

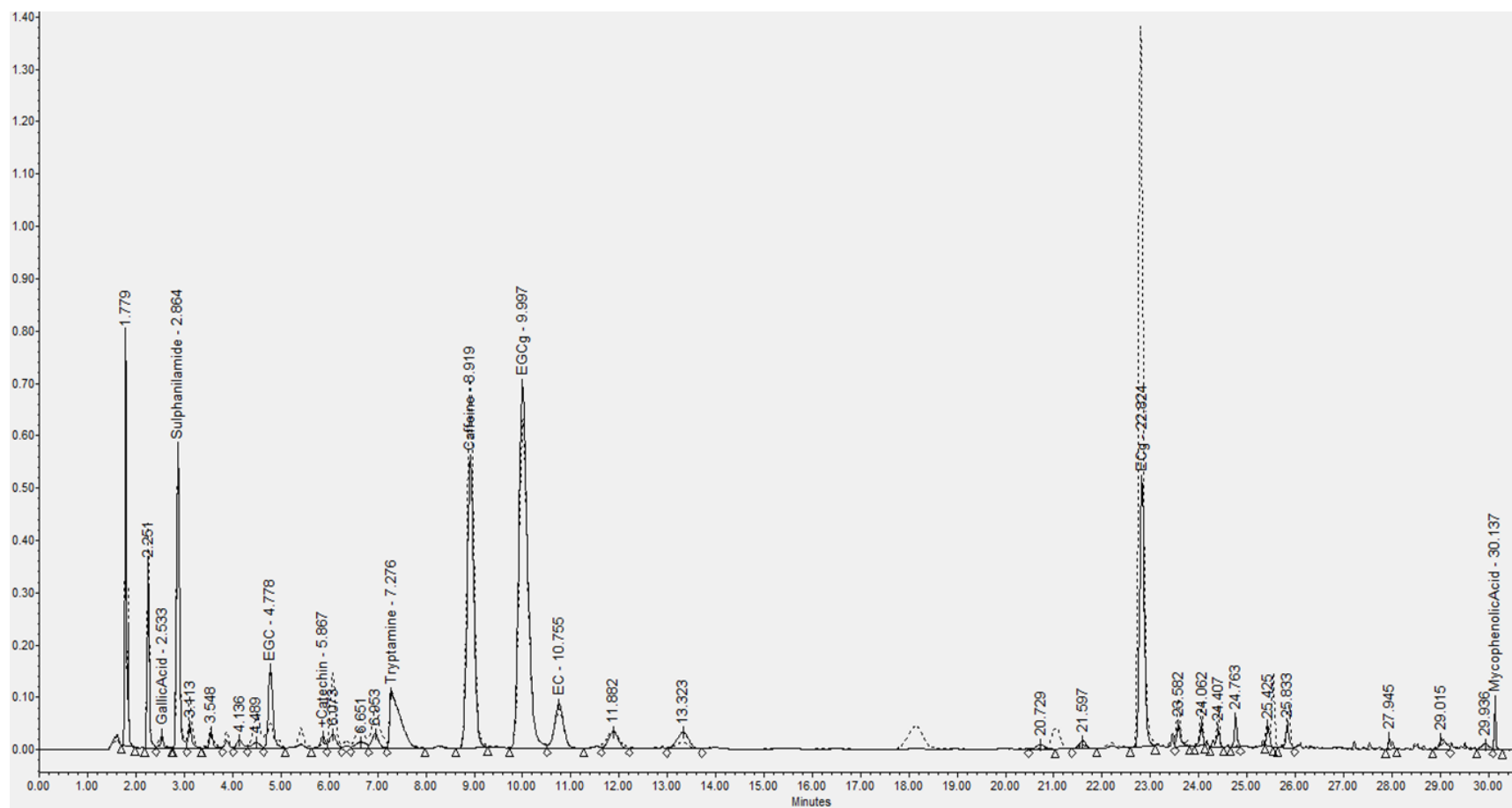


Figure 3.2: A representative UPLC chromatogram of green tea sample from parental lines TRFK 303/577 (solid line) and GW Ejulu (dotted line) with the three internal standards.

KEY: X-axis = Retention time (minutes), Y-axis = Absorbance units (278 nm)

The mean, standard deviation, and coefficient of variation for individual catechins and caffeine content quantified in F₁ progeny and in both parents are shown in Table 3.1 and Table 3.2. Tea liquor characteristics such as colour, brightness, briskness as well as astringency and aroma were also analysed in black tea samples.

Table 3.1: Phenotypic variation of caffeine and individual catechins content in parental lines (TRFK 303/577 and GW Ejulu) and 109 F₁ progeny St. 504

Trait (%w/w)	TRFK 303/577	GW Ejulu	F ₁	F ₁ Range	F ₁ CV (%)
	Mean±SD	Mean±SD	Mean±SD		
Caffeine	2.12±0.11	2.69±0.13	2.25±0.27	1.57-3.02	11.9
CAT	0.44±0.11	1.51±0.17	0.62±0.32	0.17-1.64	53.1
EC	1.22±0.21	0.90±0.07	1.19±0.41	0.51-2.63	34.4
ECG	2.33±0.10	5.21±0.03	3.31±0.83	1.54-6.58	25.6
EGC	3.72±0.72	1.02±0.40	3.25±1.07	0.60-5.79	33.7
EGCG	7.15±0.65	6.50±0.40	6.03±1.13	1.78-12.33	18.9

Table 3.2: Phenotypic variation of caffeine and individual catechins content in parental lines (GW Ejulu and TRFK 303/577) and 152 F₁ progeny St. 524

Trait (%w/w)	GW Ejulu	TRFK 303/577	F ₁	F ₁ Range	F ₁ CV (%)
	Mean±SD	Mean±SD	Mean±SD		
Caffeine	2.69±0.13	2.12±0.11	2.38±0.30	1.53-3.69	12.8
CAT	1.51±0.17	0.44±0.11	0.96±0.36	0.10-2.13	37.8
EC	0.90±0.07	1.22±0.21	1.57±0.51	0.66-3.40	32.6
ECG	5.21±0.03	2.33±0.10	2.82±0.66	1.06-5.24	24.2
EGC	1.02±0.40	3.72±0.72	3.34±1.05	0.60-7.61	32.3
EGCG	6.50±0.40	7.15±0.65	6.14±1.24	1.78-12.33	20.4

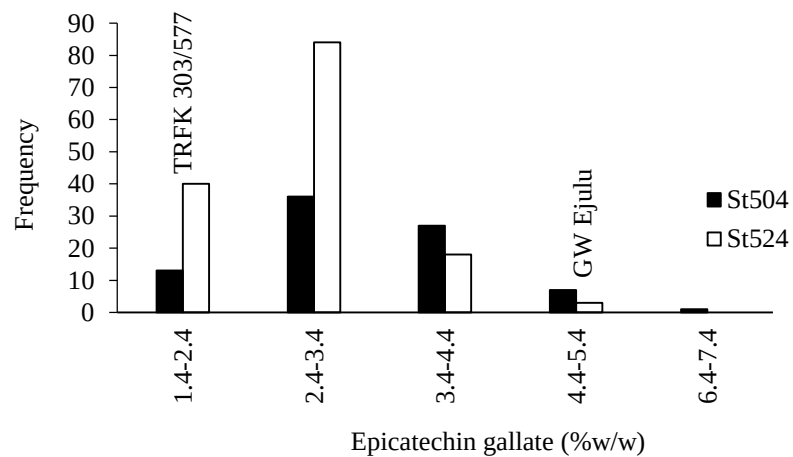
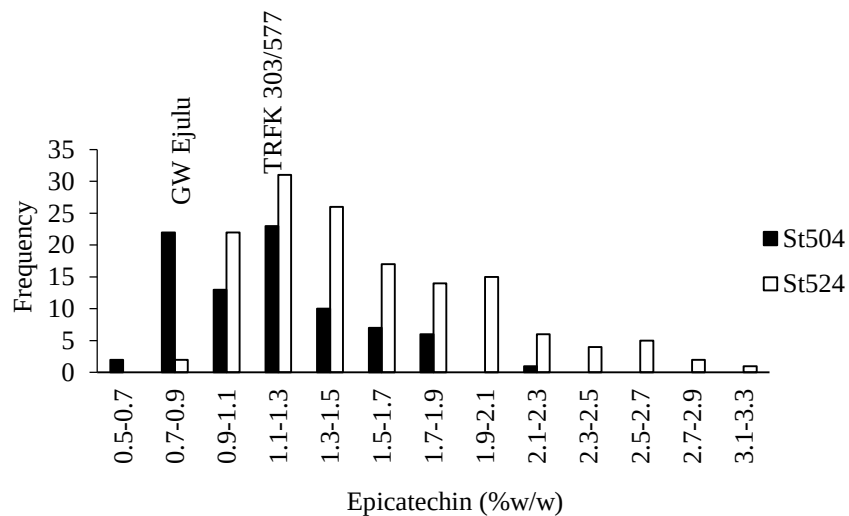
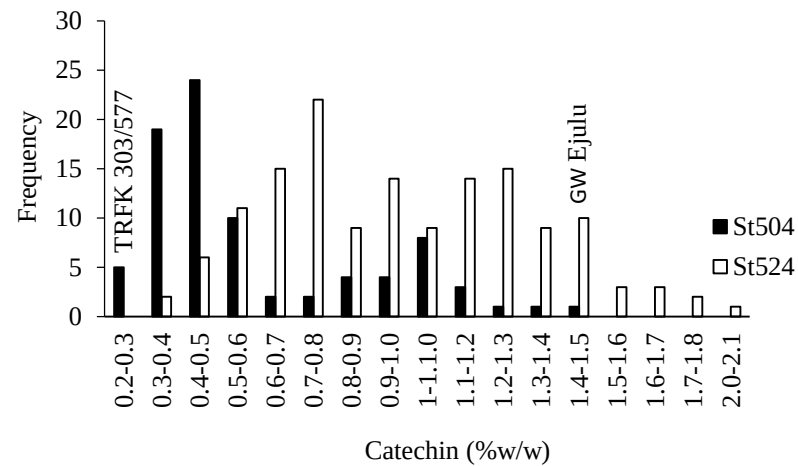
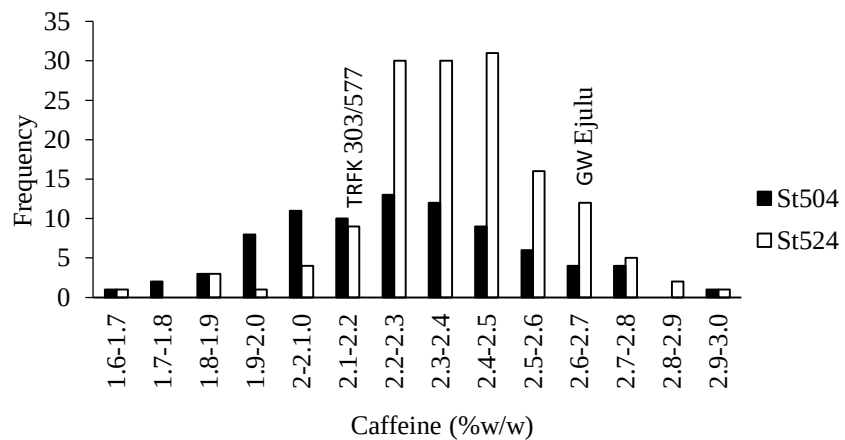
The frequency distributions of phenotype datasets are shown in Figure 3.3. The values of the two parents are indicated in the graphs. The levels of the individual catechins analysed in green teas ranged from 0.1 to 12.3%, with low content of CAT and EC as compared EGCG and EGC, which were the abundant catechins. Caffeine content ranged from 1.5 to 3.7%, while the levels of individual theaflavins in processed black teas ranged from 0.2 to 1.3% with TF3 being the

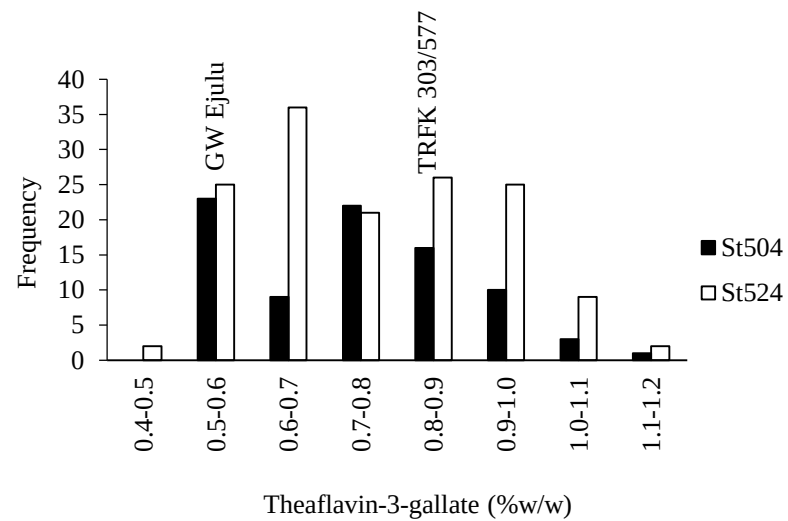
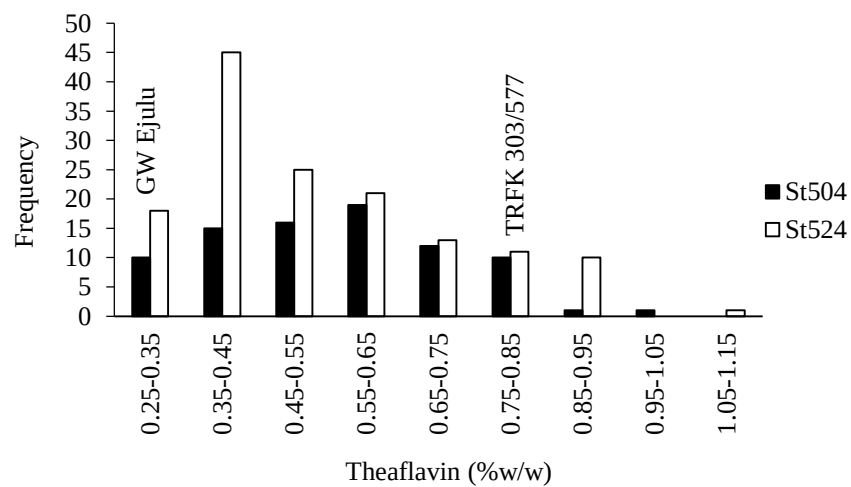
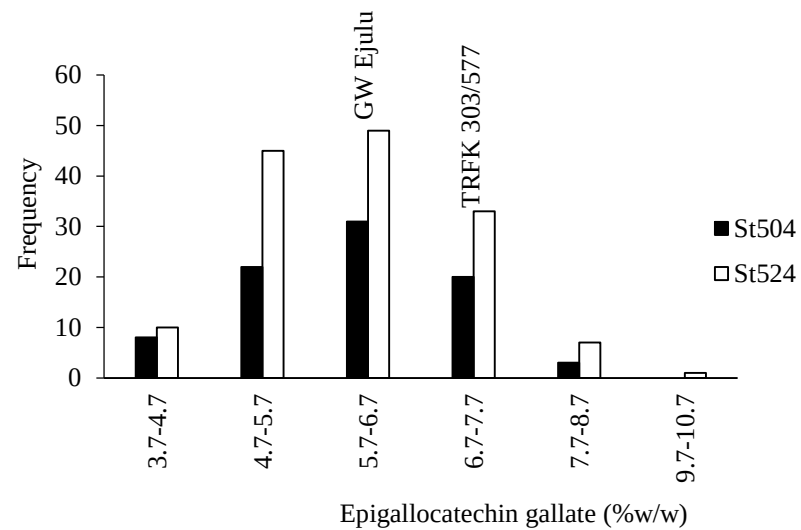
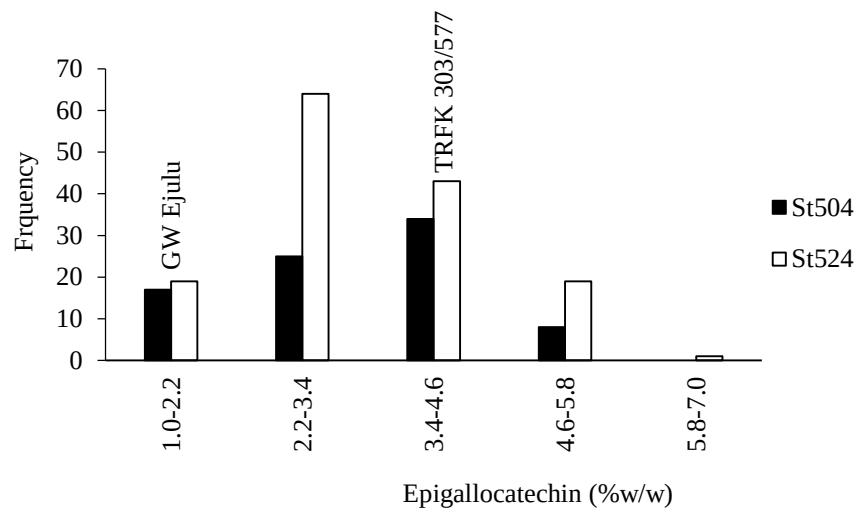
most abundant theaflavin (Figure 3.3). A variation in the levels of individual theaflavins content was observed in parental lines and progenies. Parental clone TRFK 303/577 had high levels of all four theaflavins as compared to parental clone GW Ejulu in both tea populations. A number of traits namely caffeine, EGC, TF1, TF2, TF3 and TF4 score as measured in the two populations (St 504 and St 524) had their F_1 means falling in between the two parental means for respective traits (Figure 3.3). However, the F_1 means for EC, ECG, EGCG, and tea taster scores were marginally lower than the parental means. There were variations in tea taster scores in the parental clones and F_1 progenies. Most of the black tea samples had an average tea quality regarding tea taster's score except for aroma, which scored poorly. This poor aroma score might have been due to the time difference (12 months) between the black tea processing and the tea taster scoring. Tea liquor colour (CL) scored higher than the other tea taster score traits in parental clone GW Ejulu as compared to parental clone TRFK 303/577. There were also variations in the %RWC in F_1 progenies and parental clones in the two tea populations (Figure 3.3). The %RWC ranged from 49 to 99% which show that there were tea plants that were either tolerant or susceptible to drought. Parental clone GW Ejulu which is drought susceptible tea cultivar had low % RWC of 55.1% while parental clone TRFK 303/577 which is drought-tolerant cultivar had high %RWC of 74.1%. In overall, a large number of phenotypic traits for the progeny fell outside to those of the parents indicating transgressive segregation, hence polygenic inheritance of the traits in question (Figure 3.3). Further, the observed normal distribution for the phenotypic traits shows the involvement of several genes in the genetic control of phenotypic traits.

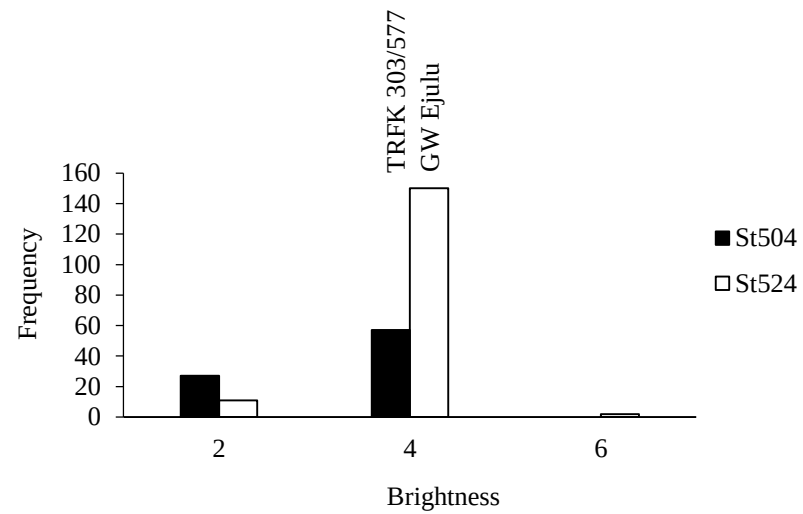
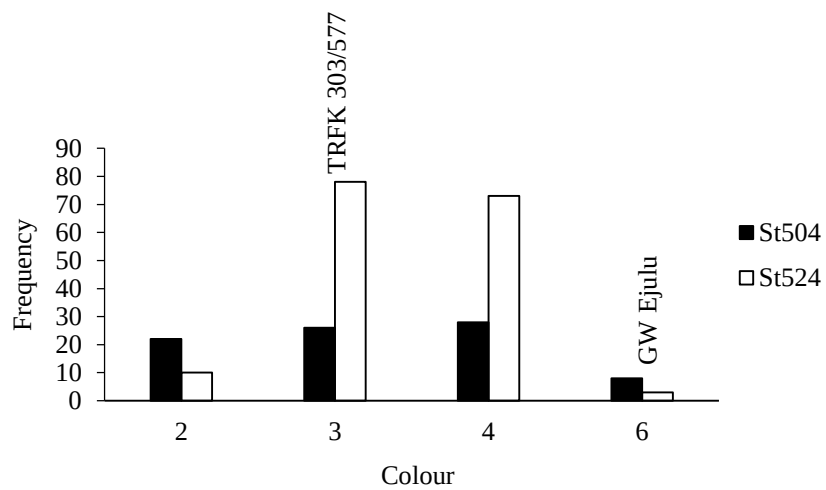
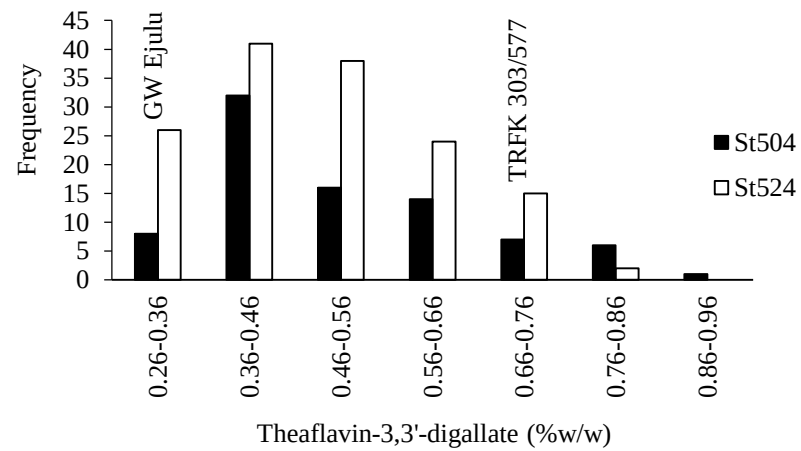
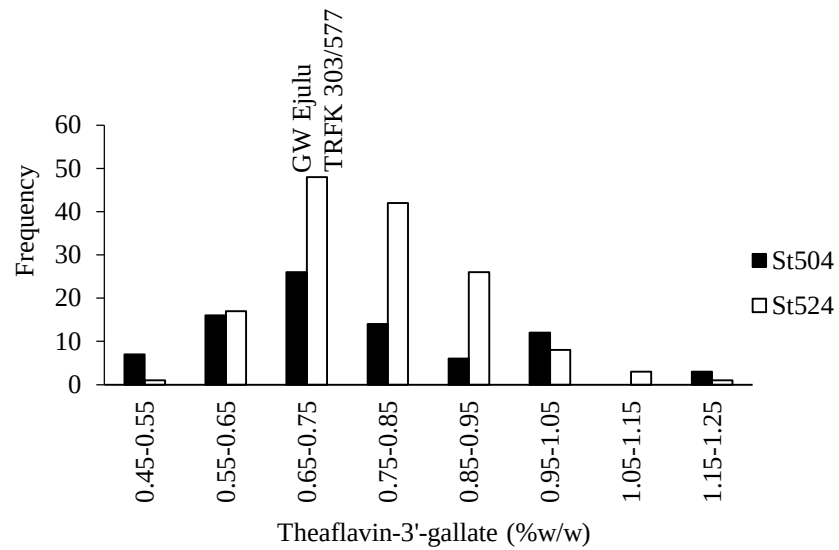
3.6.2 Correlation analysis

Pearson's correlation coefficients between different observations of each trait in the two tea population are shown in Table 3.3 and Table 3.4. The correlations among the phenotypic traits were grouped into catechins (CAT, EC, ECG, EGC, and EGCG), caffeine, theaflavins (TF1, TF2, TF3, and TF4), and tea taster's scores (CL, BRT, AST, BRK and AR) and % RWC. In general, the significant correlation coefficients mostly observed in TRFK St 504 were between EC and EGC ($r=0.62$, $P<0.01$), TF1 and TF2 ($r=0.89$, $P<0.001$), TF1 and TF4 ($r=0.90$, $P<0.001$), TF2 and TF4 ($r=0.89$, $P<0.001$). For tea taster's score, a high correlation was observed between individual AST and BRK ($r=0.96$, $P<0.001$), and it ranged from $r=0.62$ - 0.96 . There were also significant correlations between AR with CL ($r=0.39$ - 0.42 , $P<0.05$), BRT ($r=0.54$ - 0.57 , $P<0.01$),

AST ($r = 0.47-0.51$, $P < 0.01$) and BRT ($r = 0.56-0.58$, $P < 0.01$) (Table 3.3 and 3.4). The correlation coefficient for some of the phenotypic traits in TRFK St. 524, had a similar trend as in TRFK St. 504 for the tea taster's scores except for catechins and theaflavins. The individual catechins correlated significantly with other individual catechins and theaflavins. There was significant correlation between EGC and EGCG ($r = 0.51$, $P < 0.01$), EC and EGC ($r = 0.73$, $P < 0.01$), TF1 and EGC ($r = 0.62$, $P < 0.01$), TF2 and EC ($r = 0.44$, $P < 0.05$), TF1 and EGCG ($r = 0.42$, $P < 0.05$), TF2 and EGC ($r = 0.52$, $P < 0.01$), TF4 and EGC ($r = 0.54$, $P < 0.01$) (Table 3.4). However, there was no correlation between % RWC with all biochemical compounds and tea taster scores in the two tea populations (Table 3.3 and 3.4).







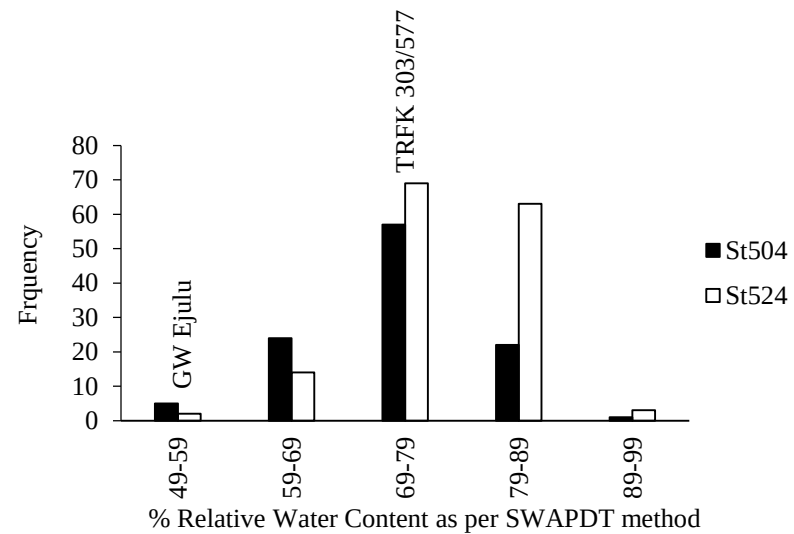
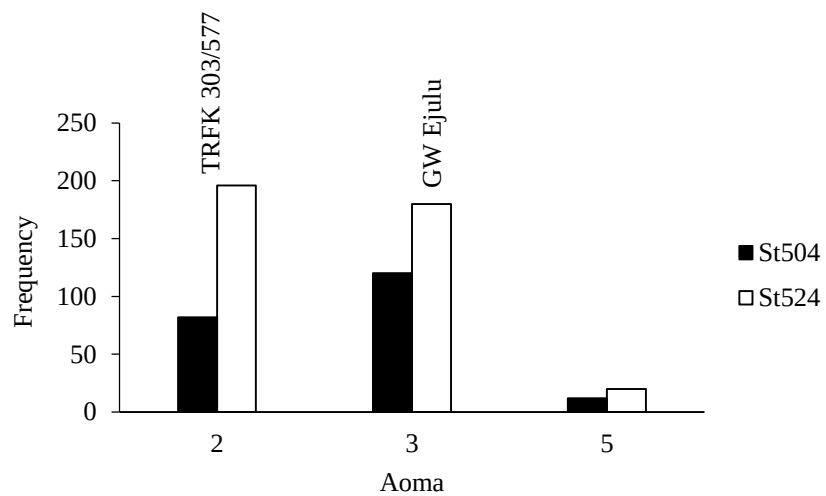
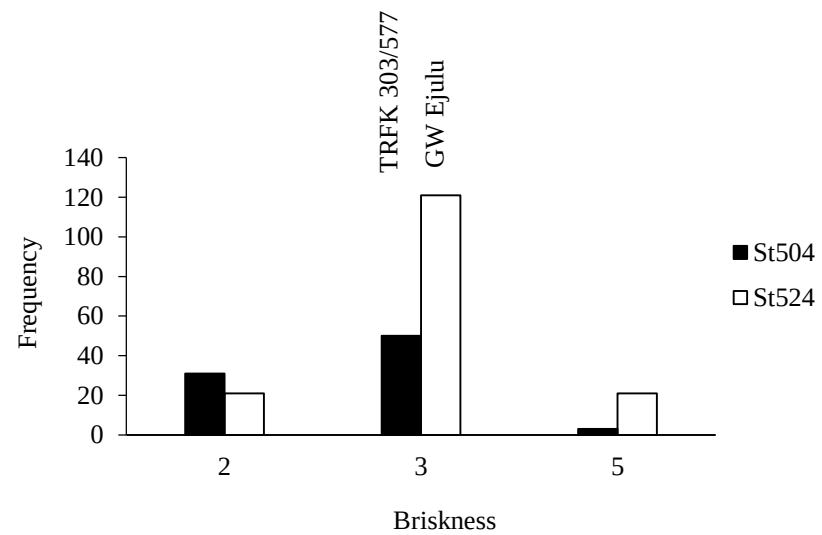
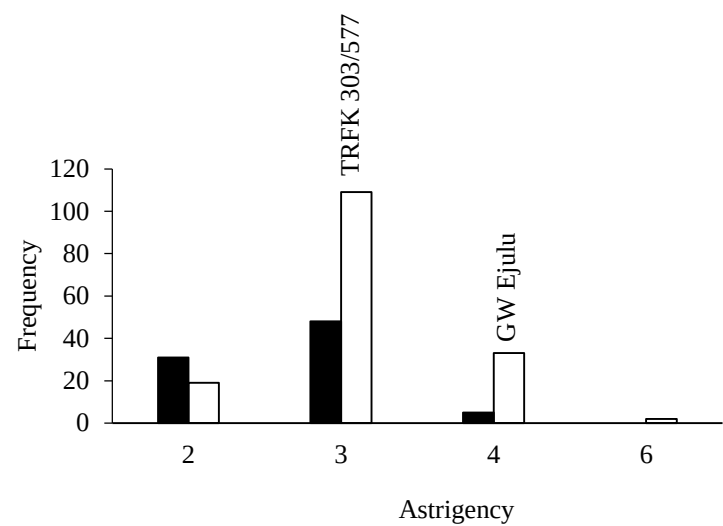


Figure 3.3: Frequency distribution pattern of phenotypic traits of both green and black tea from two parental clones and F₁ population derived from a reciprocal cross between TRFK 303/577 and GW Ejulu (i.e. St 504 and St 524), respectively. The phenotypic values of the two parental clones are indicated in the graph.

Table 3.3: Correlation coefficients (r) between phenotypic traits in the mapping population St 504

	CAFF	CAT	EC	ECG	EGC	EGCG	TF1	TF2	TF3	TF4	CL	BRT	AST	BRK	AR	RWC
CAFF	1															
CAT	-0.06	1														
EC	-0.07	-0.40	1													
ECG	0.44*	0.22	0.26	1												
EGC	-0.13	-0.78**	0.62**	-0.34	1											
EGCG	0.39	-0.57**	0.08	0.12	0.41*	1										
TF1	-0.24	-0.39	0.14	-0.33	0.42*	0.04	1									
TF2	-0.04	-0.36	0.10	-0.09	0.26	0.12	0.89***	1								
TF3	0.03	-0.08	-0.30	-0.01	-0.13	0.11	0.64**	0.83***	1							
TF4	-0.20	-0.28	-0.03	-0.25	0.23	0.08	0.90***	0.89***	0.80***	1						
CL	0.32	-0.08	0.20	0.26	0.12	0.17	-0.03	0.09	-0.13	-0.10	1					
BRT	0.24	-0.18	0.24	0.24	0.16	0.15	0.16	0.32	0.07	0.11	0.73**	1				
AST	0.20	-0.09	0.21	0.24	0.11	0.18	0.15	0.26	0.08	0.16	0.66**	0.83***	1			
BRK	0.22	-0.12	0.23	0.25	0.12	0.16	0.19	0.31	0.13	0.21	0.62**	0.80**	0.96***	1		
AR	0.18	-0.17	0.12	0.19	0.17	0.15	0.18	0.20	0.01	0.15	0.42*	0.54**	0.51**	0.58**	1	
RWC	0.12	0.19	0.16	-0.19	0.07	-0.03	0.06	0.06	0.02	0.06	-0.05	0.05	0.10	0.05	-0.06	1

CAFF- Caffeine; CAT- Catechin; EC- Epicatechin; ECG- Epicatechin gallate; EGC- Epigallocatechin gallate; EGCG- Epigallocatechin gallate; TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate; CL- Colour; BRT- Brightness; AST- Astringency; BRK- Briskness; AR- Aroma; RWC- percent relative water content

**Correlation significant at $p \leq 0.05$ level*

***Correlation significant at $p \leq 0.01$ level*

****Correlation significant at $p \leq 0.001$ level*

Table 3.4: Correlation coefficients (r) between phenotypic traits in the mapping population St 524

	CAFF	CAT	EC	ECG	EGC	EGCG	TF1	TF2	TF3	TF4	CL	BRT	AST	BRK	AR	RWC
CAFF	1															
CAT	0.12	1														
EC	0.07	-0.41*	1													
ECG	0.39	0.10	0.38	1												
EGC	-0.05	-0.74**	0.73**	-0.02	1											
EGCG	0.34	-0.60**	0.19	0.33	0.51**	1										
TF1	-0.01	-0.44*	0.51**	-0.01	0.62**	0.33	1									
TF2	0.13	-0.46*	0.44*	0.11	0.52**	0.42*	0.77**	1								
TF3	0.12	-0.16	-0.14	-0.03	0.02	0.31	0.44*	0.72**	1							
TF4	0.06	-0.51	0.34	-0.08	0.54**	0.42*	0.82**	0.91**	0.71**	1						
CL	0.11	-0.22	0.20	0.26	0.23	0.31	0.24	0.29	0.05	0.20	1					
BRT	0.09	-0.18	0.12	0.05	0.24	0.21	0.29	0.21	0.03	0.26	0.48*	1				
AST	0.09	-0.17	0.16	0.11	0.22	0.24	0.21	0.12	-0.01	0.17	0.57**	0.59**	1			
BRK	0.10	-0.17	0.14	0.10	0.21	0.22	0.19	0.17	0.01	0.25	0.47*	0.69**	0.76**	1		
AR	0.17	-0.15	0.21	0.10	0.26	0.16	0.22	0.21	-0.03	0.21	0.39*	0.57**	0.47*	0.56**	1	
RWC	0.09	0.07	0.06	-0.05	0.06	-0.03	0.12	0.08	0.06	0.12	-0.05	0.00	0.09	0.02	-0.02	1

CAFF- Caffeine; CAT- Catechin; EC- Epicatechin; ECG- Epicatechin gallate; EGC- Epigallocatechin gallate; EGCG- Epigallocatechin gallate; TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate; CL- Colour; BRT- Brightness; AST- Astringency; BRK- Briskness; AR- Aroma; RWC- percent relative water content

**Correlation significant at $p \leq 0.05$ level*

***Correlation significant at $p \leq 0.01$ level*

****Correlation significant at $p \leq 0.001$ level*

3.6.3 Genetic map construction

Regression-based integrated map (RG) resulted into 15 linkage groups (Figure 3.4; Table 3.5). The integrated genetic linkage map covered a total genetic distance of 1 260.1 cM, with an average locus distance of 1.1 cM. The marker density was not always uniform among linkage groups. The individual linkage groups ranged in size from 64.8 cM (LG02) to 160.1 cM (LG09), and the number of markers on each group varied from 50 for LG01 to 219 for LG09 (Table 3.5). Linkage group 9 had the highest marker density with an average locus distance of 1.4 cM, while the other LGs had relatively lower marker density (0.5-1.7 cM/DArT). However, DArT markers linked to phenotypic traits studied were absent on the LG03, LG05 and LG11. The constructed maps contained 15 linkage groups which correspond to haploid chromosome number in *Camellia* spp. ($n = 15$) indicating saturation of the genome. The clustering of markers was observed, and some linkage groups were more densely saturated with markers as compared to others.

3.6.4 Phenotype segregation and QTL mapping

The analysis of 16 phenotypic traits revealed a total of 47 putative QTLs in 15 LGs associated with tea quality and %RWC by IM and MQM at a significance genome-wide threshold of 5% level of confidence. Significant LOD scores ranged from 3.1-47.3 with an average of 6.5, while LOD thresholds of the detected QTLs were between 2.9 and 4.6 with an average of 3.2 (Table 3.6). The interval lengths of the QTLs ranged from 0.14 cM to 23.5 cM. QTLs associated with individual catechins, caffeine and individual theaflavins were mostly clustered in LG02 and LG04 but in different regions on the map (Figure 3.4). Most QTLs for tea taster scores (CL, BRT, AST, BRK, and AR) were detected at a lower significance level of confidence of 90% as compared to traits for biochemical molecules (Table 3.6). The QTL for each trait varied from a total of two (EC) to a total of seven (caffeine) in number. A total of 19 QTLs were identified for ten traits and were clustered in three main regions on LG01, LG02, LG04, LG12, LG13 and LG14 (Table 3.6). The variability explained by each QTL in the population was at average to high levels and ranged from 5.5 to 56.6%, with an average of 9.9%. The population variability explained (PVE) by all QTLs for all the traits ranged between 6.5 (CL) to 97% (EGC).

Seven QTLs for caffeine levels were detected and accounted for 40.5% of the total population explained variability (Table 3.6). Of these, one QTL, qCaffeine, was mapped on LG13 close to marker 5088162, with a LOD score of 4.7 and PVE of 8.1% (Table 3.6). An additional five

putative QTLs were identified on LG02, LG04, LG07, LG08, and LG14 (Table 3.6). The qCaffeine LOD score ranged from 3.2 to 4.7, and the PVE ranged between 6.0% and 8.1%. Six QTLs were detected for CAT content, accounted for a total of 91.1% of the total population variability. The two major QTLs, qCAT, were located on LG04 and LG12 close to marker 5063001 and 5123751, respectively. The two QTLs, had the largest effect on CAT content, accounting for 55.6 and 10.3% of the population variability, respectively (Table 3.6). The QTLs were distributed across the six LGs, with a LOD score ranging from 3.7 to 46.

Two QTLs, qEC, were identified for EC content, which accounted for a total of 14.6% of the population variability (Table 3.6). The qEC was located on LG02 close to marker 5072338 with PVE of 7.0%, and on LG15 close to marker 5085963 with 7.6% of the population variability. Seven QTLs, qECG, were detected for ECG content, which collectively explained 71.1% of the total population variability with LOD score range of 3.7 to 14.9 (Table 3.6). The QTL for ECG, qECG, was located on LG12 close to marker 5136790, had the largest effect on ECG content and accounted for of 23.1% of the population variability. A total of seven QTLs were detected for EGC content, which accounted for a total of 97% the population variability (Table 3.6). Of these, two QTLs, qEGC, were mapped on LG04 and LG12, respectively (Table 3.6). The qEGC on LG04 was located close to marker 5123475 with PVE of 56.6%, while the qEGC on LG12 was located close to marker 5123751 with PVE of 8.6% with LOD score of 47.3 and 14.9, respectively. An additional five putative QTLs were identified on LG1, LG02, LG06, LG13, and LG14 (Table 3.6). The LOD score for the five loci varied from 3.2 to 4.2 and the PVE ranged between 5.5 and 7.2%. The QTL, qEGCG, was identified for EGCG content on LG04, close to marker 5087017, contributing to 24.1% of the population variability (Table 3.6). Also, two putative QTLs were mapped on LG12 and LG15 and explained between 6.5 and 7.2% of the population variability (Table 3.6). The three QTLs detected for EGCG content accounted for a total of 37.8% the population variability (Table 3.6). Two QTLs for theaflavin, qTF1, were identified and it accounted for a total of 13.7% of the population variability (Table 3.6). The qTF1 was located on LG02 and LG06 close to markers 5084595 and 5136045, respectively. The nucleotide sequences for these markers are presented in Appendix 3.1. Only one QTL for qTF2 and another one for qBRT were identified on LG02 and LG14, close to marker 5084595 and 5125626, contributing to 7.0 and 6.8% of the population variability, respectively (Table 3.6).

The QTLs for astringency, qAST, and briskness, qBRK were mapped on two linkage groups (LG01 and LG09). The population variability for the tea taster score varied with aroma, qAR (6.4%) and briskness, qBRK (7.3%), respectively. Tea liquor colour and tea brightness had one QTL each while tea briskness, tea aroma had two QTLs each and tea astringency had three QTLs, respectively. Three QTLs were detected for %RWC, which accounted for a total of 19.9% of the population variability (Table 3.6). The three stable QTLs, qRWC, were mapped on LG02, LG06, and LG09, respectively. Three QTLs were identified for %RWC. The first QTL (qRWC) was located on LG02, the second QTL was located on chromosome LG06, and the third QTL was located on LG09. These QTLs spanned 60.7 cM, 66.2 cM, and 6.7 cM and had peak LOD scores of 4.3, 3.3 and 4.0, respectively (Table 3.6).

Table 3.5: DArT markers distribution among the linkage groups

Linkage group	Number of markers	Total length covered (cM)	Average distance between markers (cM)
LG01	50	97.3	0.5
LG02	107	64.8	1.7
LG03	61	82.4	0.7
LG04	98	70.1	1.4
LG05	77	77.0	1.0
LG06	65	82.6	0.8
LG07	139	83.2	1.7
LG08	100	68.3	1.5
LG09	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 3.6: Putative QTL for caffeine, individual catechins and theaflavins content, tea taster's scores and percent relative water content detected in St 504 and St 524 tea plant population using multiple QTL model mapping

Traits	QTL	LG	Position (cM)	LOD threshold	LOD score	PVE (%)	Nearest Marker
Caffeine	qCaffeine	2	50.905 (50.247-51.548)	3.3	3.4	6.0	5064585
		4	68.574 (63.137-70.142)	3.2	3.8	6.7	5112599
		7	48.071 (34.695-50.409)	3.2	3.7	6.6	5064391
		8	18.762 (16.783-26.008)	3.3	3.9	6.7	5134558
		13	29.748 (23.653-36.538)	3.1	4.7	8.1	5088162
		14	5.465 (0.000-5.465)	3.3	3.7	6.4	5123053
CAT	qCAT	2	0.000 (0.000-2.421)	3.2	3.7	6.4	5135436
		4	30.322(29.886-30.322)	3.3	46.0	55.6	5063001
		8	12.506 (12.506-12.882)	3.3	3.3	5.6	5130194
		12	42.998 (36.019-54.085)	3.1	6.1	10.3	5123751
		13	50.602 (48.559-58.674)	3.0	4.1	6.9	5111268
		14	60.713 (47.731-61.117)	3.2	4.1	6.9	5132370
EC	qEC	2	2.242 (0.000-8.309)	3.3	4.1	7.0	5072338
		15	25.381 (20.639-25.381)	3.3	4.5	7.6	5085963
ECG	qECG	1	96.36 (96.36-97.284)	3.0	6.8	11.7	5128890
		4	17.658 (13.976-37.226)	3.4	4.7	8.0	5087113
		6	56.948 (56.504-56.948)	3.1	4.0	6.7	5098382
		10	20.631 (20.090-25.539)	3.3	4.8	8.0	5136108
		12	50.441 (50.148-50.441)	2.9	14.9	23.1	5136790
		13	50.602 (48.559-58.674)	3.1	3.7	6.4	5088162
		15	75.067 (75.067-76.498)	3.3	4.2	7.2	5111164
EGC	qEGC	1	87.059 (84.375-88.562)	3.0	3.3	5.8	5133866
		2	7.719 (7.719-7.855)	3.2	3.2	5.5	5124128
		4	27.194 (27.194-28.879)	3.3	47.3	56.6	5123475
		6	66.351 (56.504-72.354)	3.1	4.2	7.2	5073424
		12	42.998 (42.839-54.085)	3.0	14.9	8.6	5123751
		13	50.602 (48.559-58.674)	3.1	3.7	6.4	5136623
		14	60.713 (60.713-63.050)	3.1	4.0	6.9	5132791
EGCG	qEGCG	4	37.226 (28.879-38.559)	3.4	15.6	24.1	5087017
		12	48.179 (36.963-49.744)	3.0	4.2	7.2	5104630
		15	32.140 (17.119-33.578)	3.3	3.8	6.5	5114089
TF1	qTF1	2	4.455 (4.455-5.391)	3.1	3.2	5.8	5084595
		6	69.564 (69.564-75.112)	4.1	4.5	7.9	5136045

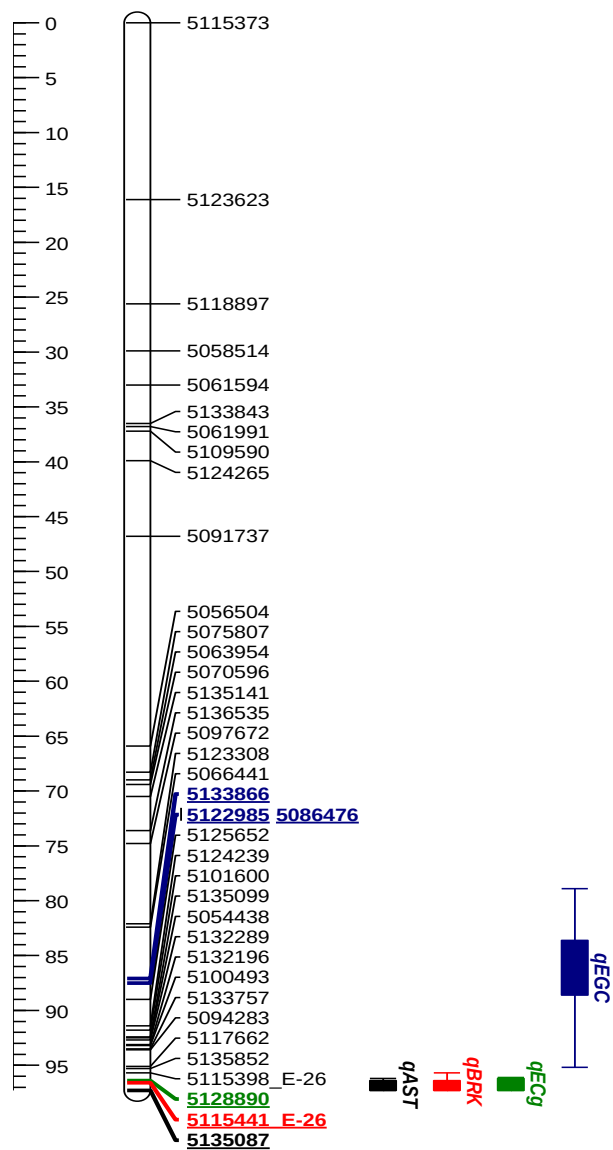
TF2	qTF2	2	4.455 (1.220-5.391)	3.1	4.0	7.0	5084595
CL	qCL	7	72.684 (72.677-72.684)	3.0	3.2	6.5	5132432
BRT	qBRT	14	71.046 (65.383-71.046)	3.4	3.3	6.8	5125626
AST	qAST	1	96.577 (88.83-97.284)	3.0	4.5	9.1	5135087
		9	87.614 (87.614)	3.2	3.2	6.6	5123950
BRK	qBRK	1	96.577 (89.022-97.284)	3.0	3.6	7.3	5115441_E-26
		9	87.614 (87.614)	3.6	3.6	7.4	5123950
		13	61.582 (61.582)	3.0	3.1	6.3	5114985
AR	qAR	4	68.574 (68.574)	3.0	3.1	6.4	5112599
		10	28.656 (26.854-29.038)	3.2	3.5	7.0	5128967
% RWC	qRWC	2	60.868 (56.912-60.868)	3.3	4.3	7.3	5136794
		6	66.217 (66.052-66.217)	3.2	3.3	5.7	5082606
		9	6.684 (6.167-6.684)	3.9	4.0	6.9	5130531
Average				3.2	6.5	9.9	

CAT- Catechin, EC- epicatechin, ECG- epicatechin gallate, EGC- epigallocatechin, EGCG- epigallocatechin gallate, TF1- theaflavin, TF2 - theaflavin-3-gallate, CL- colour, BRT- brightness, AST- astringency, BRK- briskness,% RWC- percent relative water content
LOD- logarithm of odds ratio

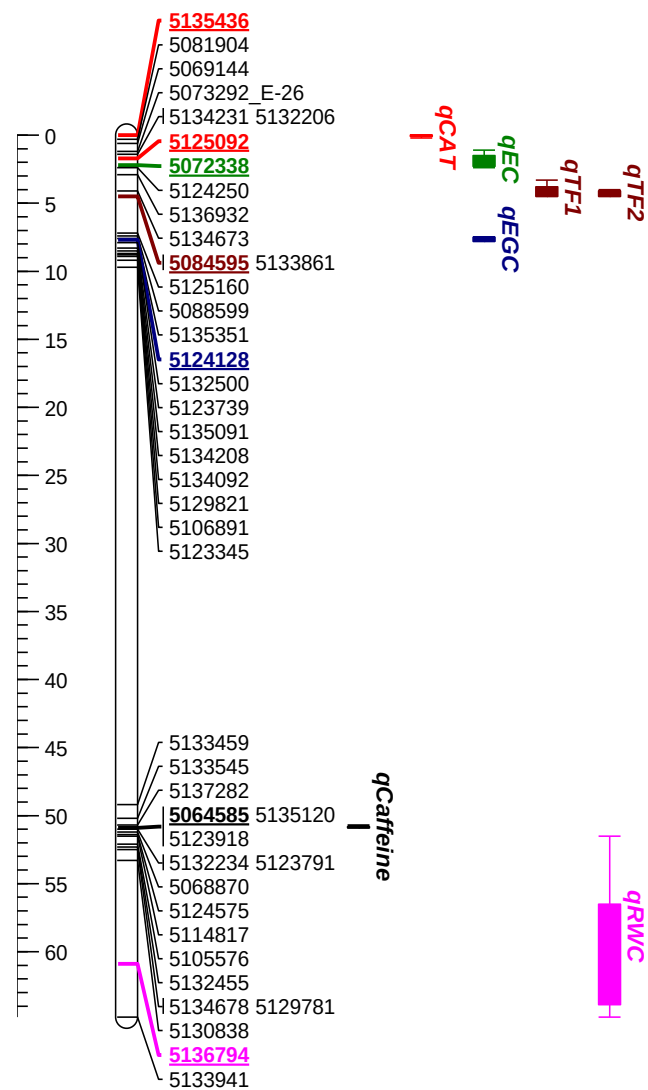
PVE- phenotypic variation explained

The genome-wide LOD significance thresholds ($P < 0.05$) based on permutation testing (n=1 000)

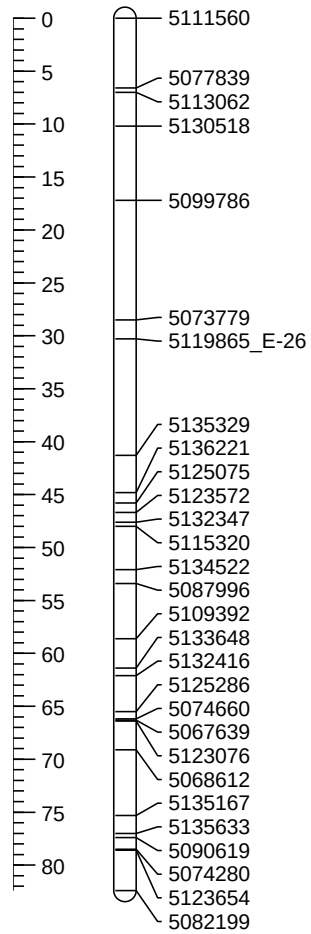
LG01



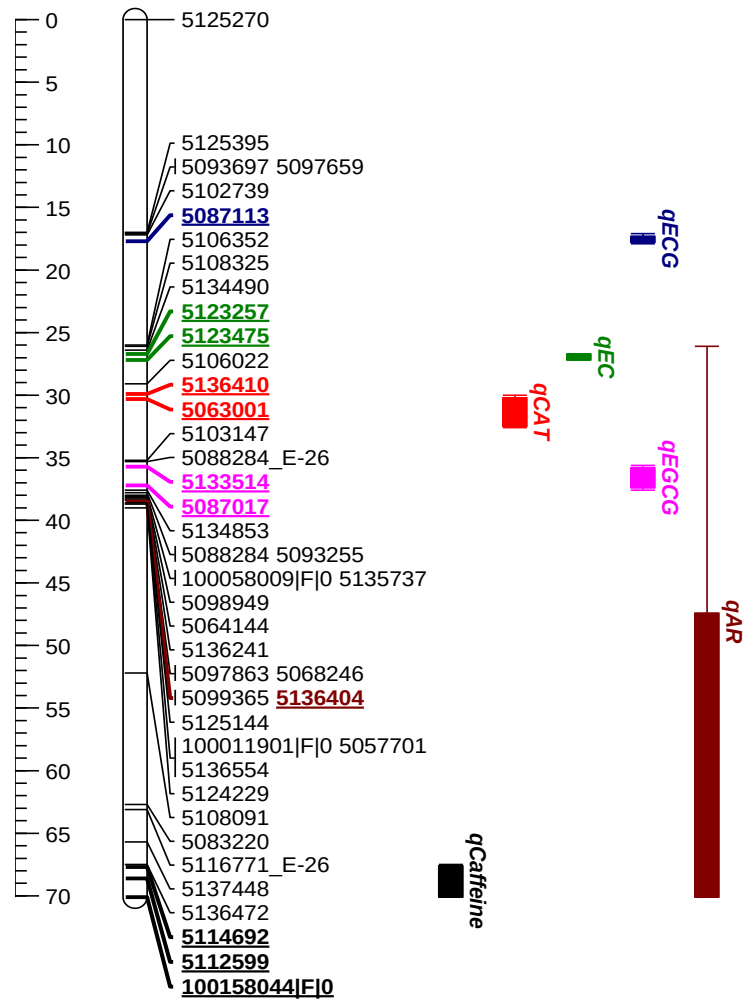
LG 02



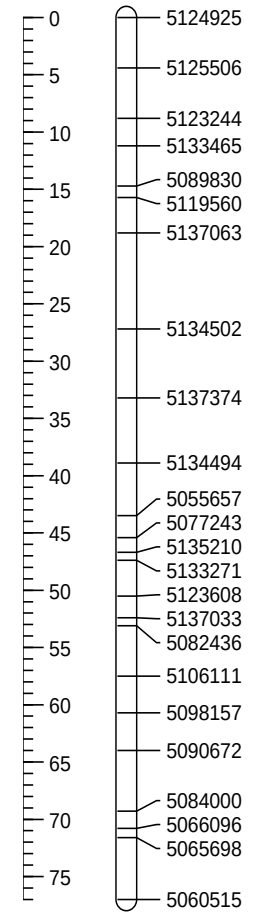
LG 03



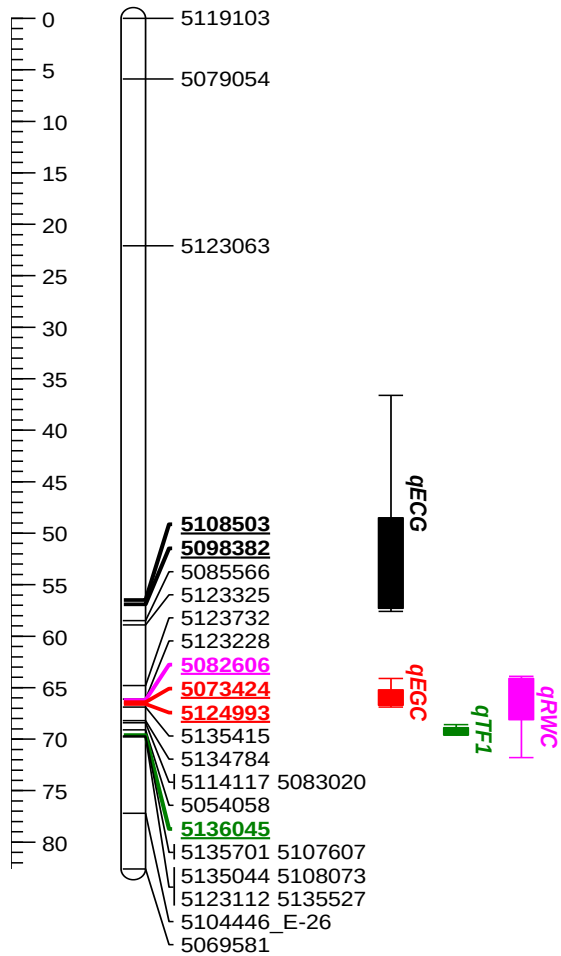
LG 04



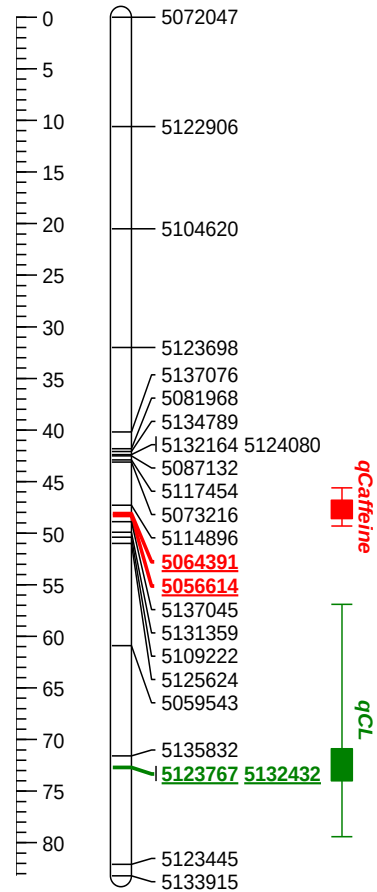
LG 05



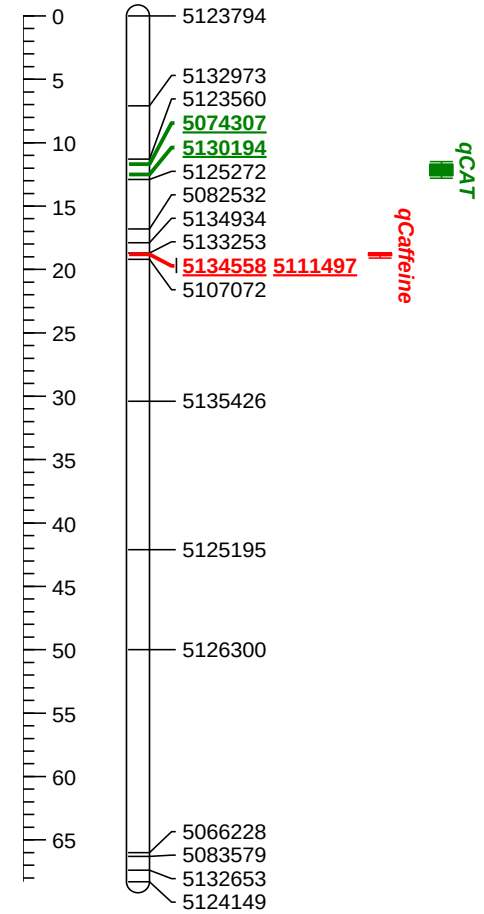
LG 06



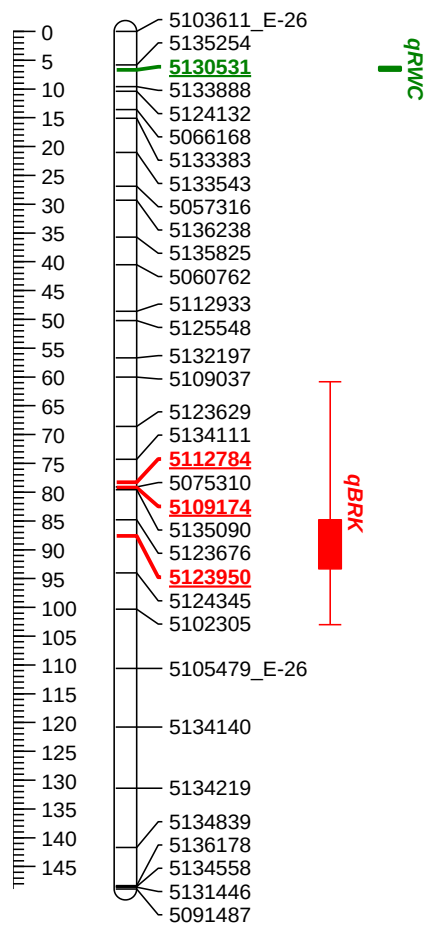
LG 07



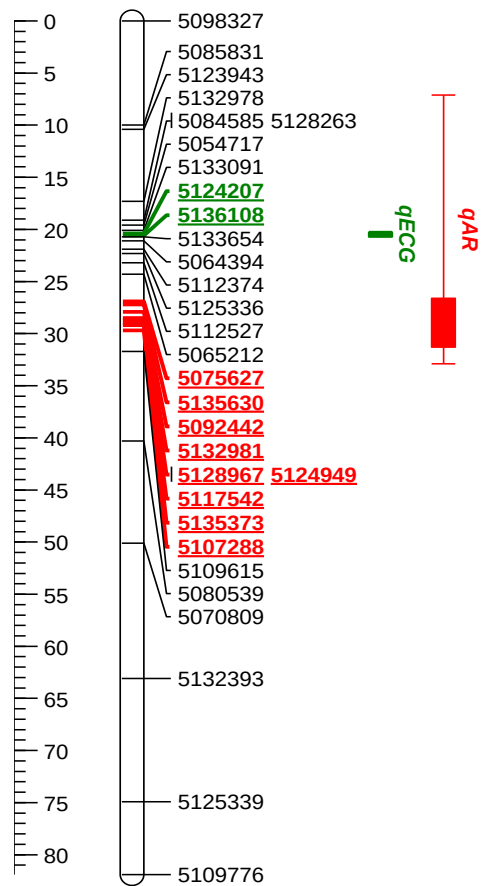
LG 08



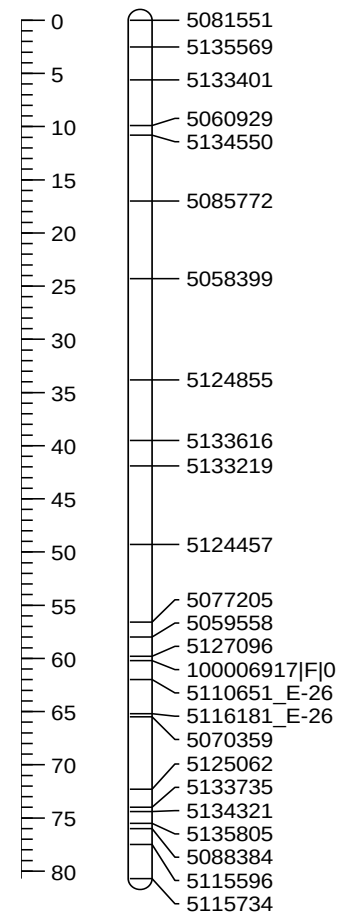
LG 09



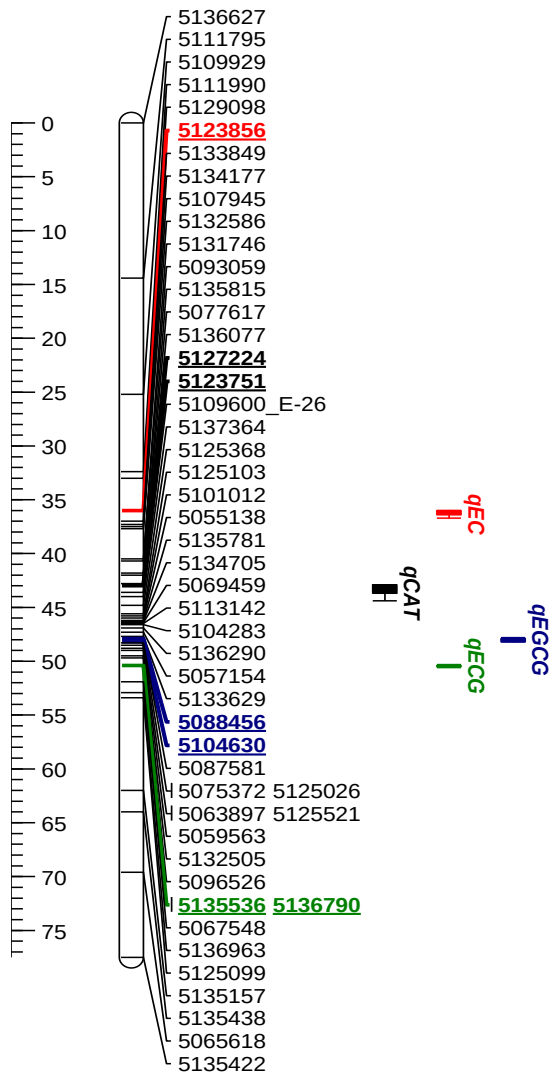
LG 10



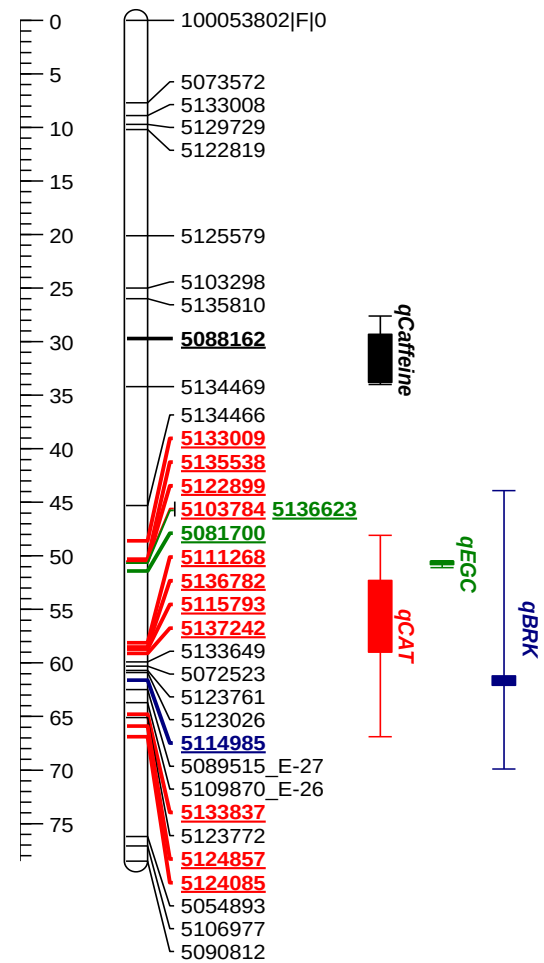
LG 11



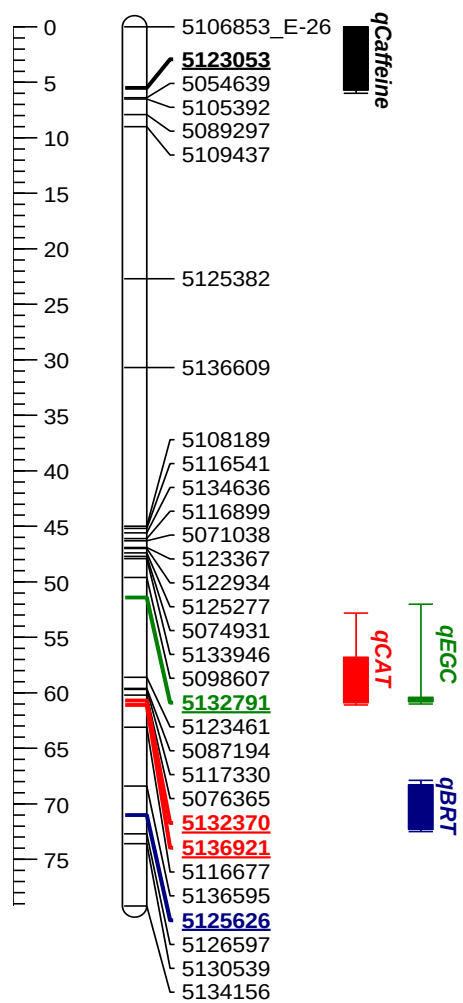
LG 12



LG 13



LG 14



LG 15

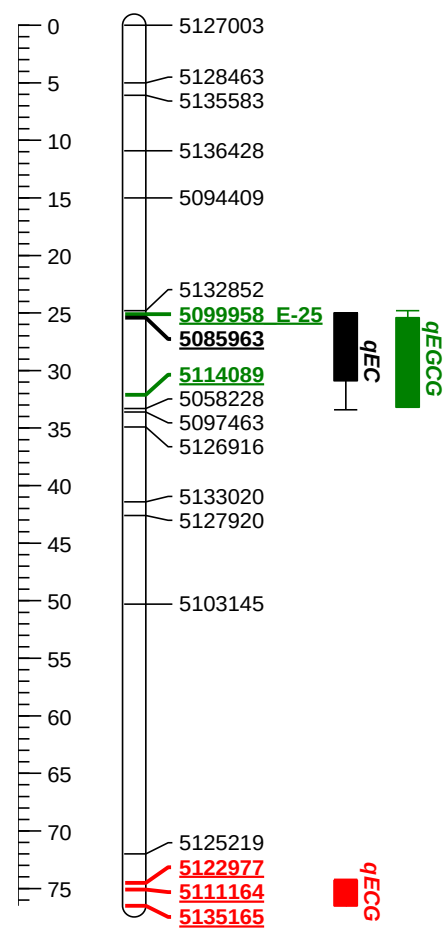


Figure 3.4: DArT-based genetic map of tea plant showing the location of QTL for caffeine, catechins, theaflavins content, liquor brightness and % rwc identified in the two mapping population. Map distance scales in cM are placed at the left margin. Significant QTL for caffeine, individual catechin and individual theaflavin content and liquor brightness is represented by different bars. Bars and lines indicate 1-LOD and 2-LOD support intervals.

Abbreviations: qCaffeine- Caffeine; qCAT- Catechin; qEC- Epicatechin; qECG- Epicatechin gallate; qEGC- Epigallocatechin gallate; qEGCG- Epigallocatechin gallate; qTF1- Theaflavin; qTF2- Theaflavin-3-gallate; qCL- Colour; qBRT- Brightness; qAST- Astringency; qBRK- Briskness; qAR- Aroma; qRWC- percent relative water content.

3.7 Discussion

Construction of a genetic linkage map is an important prerequisite for the identification of QTL(s) as well as mapping of agronomically important genes, which will be very useful for developing improved cultivars. Green tea catechins, black tea theaflavins, and caffeine are important biochemical traits that determine tea quality given that the tea price at the auction is based on tea quality. The two principal gallated catechins present were EGCg and EGC which corroborate with results reported by (Karori *et al.*, 2014). The two gallated catechins (EGCg and EGC) have been shown to contribute significantly to the formation of theaflavins. The B-ring in tri-hydroxylated catechins is oxidised at a higher rate than the B-ring in di-hydroxylated catechins, which include: CAT, EC, and EGC because of their lower oxidation ability (Owuor *et al.*, 2006b). Therefore, tea cultivars with high levels of EGCg and EGC can be developed through tea breeding programs to improve black tea quality.

It was demonstrated previously that the ratio of di-hydroxyl flavan-3-ols to tri-hydroxyl flavan-3-ols influenced black tea quality in that, a high proportion of simple catechins (CAT, EC, and EGC) to gallocatechins (EGC and EGCg), lead to a higher proportion of theaflavin (Ellis and Nyirenda, 1995). The strong correlation of the digallate fractions in the present study is an important factor that can be utilised to enhance the quality of black tea. The changes in theaflavins and residual catechin compositions in relation to sensory characteristics of total CL, BRT, and BRK has also been reported in black tea (Obanda *et al.*, 2001). Further, (Obanda *et al.*, 2001), reported that theaflavin digallate is the most astringent, estimated to be 6.4 times more astringent than simple theaflavin, and 2.88 times more astringent than either theaflavin-3-monogallate or theaflavin-3'-monogallate. The formation of a single theaflavin molecule requires a dihydroxy and a trihydroxy flavan-3-ol (Wright *et al.*, 2002), as follows:

Epicatechin (EC) + Epigallocatechin (EGC) = simple theaflavin (TF1);

Epicatechin (EC) + Epigallocatechin gallate (EGCg) = Theaflavin-3-gallate (TF2);

ECG + Epigallocatechin (EGC) = Theaflavin-3'-gallate (TF3);

ECG + Epigallocatechin gallate (EGCg) = Theaflavin-3, 3'-digallate (TF4)

The relationship between catechins and TFs content with the quality of black tea has been investigated with reports showing a significant correlation between sensory parameters with catechins and TFs content. Among all the sensory, parameters investigated such as colour,

brightness, astringency, briskness, and aroma; the results in this present study indicate that neither caffeine, individual catechins nor individual TFs content had any correlation with sensory parameters. This could be due to loss in individual catechins and individual TFs between storage and the time of organoleptic evaluation was done which led to flat tea (Friedman *et al.*, 2009; Sedaghatthoor *et al.*, 2013). Although, there was significant correlation between the individual sensory parameters (CL, BRT, AST, BRK, and AR). These results are similar to those reported by (Owuor *et al.*, 2006b) that there was no significant correlation between simple (non-gallated) theaflavin with sensory parameters except for tea liquor brightness. The significant correlations shows that the phenotypic traits are either controlled by linked genes or genes that have pleiotropic effects on the traits. Hence, improving one phenotypic trait may lead to improvement of other phenotypic traits.

The map spanned a total length of 1 260.1 cM with 1 421 markers, which is close to tea genome obtained by (Ota and Tanaka, 1999) (1 640 cM), (Hackett *et al.*, 2000) (1 349.7 cM), and (Taniguchi *et al.*, 2012) (1 298 and 1 305 cM), (Ma *et al.*, 2014) (1 143.5 cM), but smaller in length than those reported by (Huang *et al.*, 2006a) (2 457.7 and 2 545.3 cM), and (Hu *et al.*, 2013) (4 482.9 cM). The differences in the total map length can result from variation in the number of recombination events in the two parents as well as variations in the numbers and locations of mapped loci. There were 15 linkage groups found in this study, indicating saturation of the genome (n= 15). However, there was a gap greater than 20 cM between the adjacent markers on LG06 and LG15. Even though the combination of *PstI* (CTGCAG) and *MseI* (CCGG) restriction enzymes performed best with 16 382 DArTseq achieved, there was still one gap of larger than 20 cM between adjacent markers on LG06 and LG15. The presence of a large gap may be as a result of genome regions corresponding to the gap regions of the genetic map or could be a result of gaps in both mapping parents which leads no recombination events. Therefore, further studies need to be carried out to fill in the gaps in the current genetic map. However, the mapped markers in the current study were distributed in all 15 LGs with marker density ranging from 0.5 to 1.7 cM. The average interval between two markers was 1.1 cM, which is an improvement to previous genetic maps which ranged from 1.9 to 19.0 cM with the latest by (Ma *et al.*, 2014) spacing 2.9 cM. The marker density for genome-wide QTL mapping which is recommended should be less than 10 cM (Doerge, 2002). The genetic map constructed in the present study is thus suitable for identification of QTL.

In total, six caffeine QTLs, 25 catechins QTLs, three theaflavins QTLs, nine QTLs for tea taster score and three QTLs for %RWC were detected, and the PVE by each QTL varied from 5.5 to 56.6%, with an average of 9.9%. The high PVE exhibited by these QTLs suggests that caffeine, catechins, theaflavins content, tea taster scores, and %RWC may be controlled by few critical genes. However, the small size of 261 mapping population used in this study may lead to overestimation of QTL although the population size is slightly smaller than from the previous study of 300 individuals by (Ma *et al.*, 2014). The individual catechins (ECG, EGC, and EGCG) were mostly detected in LG01, LG04, LG12 and LG13. In addition to QTL for catechins, the present study incorporated QTL for caffeine, black tea theaflavins and tea taster's scores, which are major determinants of tea quality. The major QTL, qCAT, qEC, qEGC, qTF1 and qTF2, detected showed a tendency to co-localize and cluster on LG02 while QTL, qCAT, qECG, qEGC and qEGCG clustered in LG12. The QTL were particularly prominent in the region between 0.00 and 7.72 cM of LG02 and 42.9 to 50.4 cM of LG12, where more than four QTL are located, for each trait.

In LG01, apart from the QTL for individual catechins content, QTL for individual theaflavins, qTF1 and qTF2 were also detected. These regions of the two chromosomes probably contain genes that are multifunctional which are associated with production and accumulation of catechins and theaflavins and deserve further investigation. More interestingly, QTLs associated with caffeine, qCaffeine, and liquor brightness, qBRT, were not clustered with and/or located within the same region on the chromosome with the QTL for catechins and theaflavins in the linkage groups. It is an indication that the genes associated with caffeine production and accumulation as well as liquor brightness are located in different regions on the chromosome from those of catechins and theaflavins. The few loci that were associated with tea catechins and theaflavins will need to be confirmed for MAS. For example, markers 5123475 and 5084595 were consistently associated with individual catechins (EC and EGC) and individual theaflavins (TF1 and TF2), indicated their potential as candidate markers for MAS.

3.8 Conclusion

DARtseq markers are capable of detecting even very small polymorphisms, are cost-effective and are efficient for whole-genome scans and population structure. The main objective of tea breeding is to improve the quality and quantity of the product. The investigation of biochemical

traits such as caffeine, catechins and theaflavins content, which are related to tea quality, presents an alternative approach in tea breeding. The SWAPDT method can also be used since it represents an easy method to measure water deficit in drought stress plants. From the constructed genetic map and QTL analyses, the DArTseq markers were linked to QTLs associated with black tea quality traits and drought tolerance trait. Thus, the results were in agreement with the alternate hypothesis. The use of DArTseq markers in the present study will be useful for tea breeding and will provide valuable information for selecting parents that would give rise to progenies with the desired genotype. Although in the present study, linkage maps might be enhanced by further saturation, they constitute an additional tool that can be used to locate QTLs linked with some important agronomic traits. These will contribute to the development of MAS for tea quality improvement, considering the favourable alleles of the markers involved in the QTL for tea quality. The validated molecular markers will contribute greatly to the adoption of MAS for drought tolerance and tea quality improvement. Also, with the recent availability of a complete genome sequence of the tea plant, this will allow identification genes more precisely in relation to the QTL detected in this study.

3.9 Data archiving statement

The DArT sequences have been submitted to NCBI (<http://www.ncbi.nlm.nih.gov/>). BioProject PRJNA398959, Appendix 2.2.

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.
- Anon (2002). Annual Report. 20-28: Tea Research Foundation of Kenya, Tea Board of Kenya.
- Chen, L., Apostolides, Z. & Chen, Z. M. (2012a). *Global tea breeding: achievements, challenges and perspectives*. Springer Science & Business Media.
- Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3(1): 43-52.
- 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.
- 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.
- Ellis, R. & Nyirenda, H. (1995). A successful plant improvement programme on tea (*Camellia sinensis*). *Experimental Agriculture* 31(03): 307-323.
- Feil, R. & Fraga, M. F. (2012). Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics* 13(2): 97-109.
- Friedman, M., Levin, C.E., Lee, S.U. & Kozukue, N., (2009). Stability of green tea catechins in commercial tea leaves during storage for 6 months. *Journal of Food Science* 74(2), 47-51.
- Gawal, N. & Jarret, R. (1991). A modified CTAB DNA extraction procedure for *Musa* and *Ipomea*. *Plant Molecular Biology Reporter* 9: 262-266.
- 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): 346-355.
- Hu, C. Y., Lee, T. C., Tsai, H. T., Tsai, Y. Z. & Lin, S. F. (2013). Construction of an integrated genetic map based on maternal and paternal lineages of tea (*Camellia sinensis*). *Euphytica* 191(1): 141-152.
- Huang, F., Liang, Y., Lu, J. & Chen, R. (2006). Genetic mapping of first generation of backcross in tea by RAPD and ISSR markers. *Journal Tea Science* 26(3): 171-176.

- ISO14502-2 (2005). Determination of substances characteristic of green and black tea- Part 2: Content of catechins in green tea method using high-performance liquid chromatography. Geneva: International Organization for Standardization.
- 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 and Genomes* 6(6): 915-929.
- 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.
- 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 QTL 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): 565-577.
- Malebe, M. P. (2011). Method for screening plants for drought tolerance. Vol. Msc: University of Pretoria, Pretoria, South Africa.
- 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.
- 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.
- Ota, S. & Tanaka, J. (1999). RAPD-based linkage mapping using F1 segregating populations derived from crossings between tea cultivar 'Sayamakaori' and Strain 'Kana-Ck17'. *Breed Research* 1(Suppl 1): 16.
- Owuor, P., Obanda, M., Apostolides, Z., Wright, L., Nyirenda, H. & Mphangwe, N. (2006b). 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. & Othieno, C. O. (1991). Response of black tea quality to locality in stakeholder tea farms. *Tea* 12: 41-45.
- Sedaghat, S., Haghighat, S.R., & Shokrgozar, S.A. (2013). Storage period effects on the qualitative characteristics of scented tea. *International Journal of Biosciences* 3(7):66-73.
- Sall, J., Lehman, A., Stephens, M. L. & Creighton, L. (2012). *JMP start statistics: a guide to statistics and data analysis using JMP*. SAS Institute.
- 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): 1.
- 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.
- van Eeuwijk, F. A., Bink, M. C., Chenu, K. & Chapman, S. C. (2010). Detection and use of QTL for complex traits in multiple environments. *Current Opinion in Plant Biology* 13(2): 193-205.
- Udall, J. A. (2003). Breeding for quantitative traits in plants. *Crop Science* 43(4): 1578.
- Van Ooijen, J. (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. *Kyazma BV, Wageningen* 33: 10.1371.
- Van Ooijen, J. W., Boer, M., Jansen, R. & Maliepaard, C. (2000). MapQTL 4.0: software for the calculation of QTL positions on genetic maps (user manual). Plant Research International.
- Voorrips, R. (2002). MapChart: software for the graphical presentation of linkage maps and QTL. *Journal of Heredity* 93(1): 77-78.
- 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.
- Yan, S. H. (2007). NIR evaluation of the quality of tea and its market price. *Spectroscopy Europe* 19(2): 16-19.

Appendices

Appendix 3.1: DArTseq marker ID with the corresponding individual marker sequence

Marker ID	Marker Sequence (5'-3')
5063001	TGCAGCCTTATTAGTTTCACGTTACGTATGAAAACTATGCAATTGTCGAATTTGTTTAGGAGCTCCCA
5064391	TGCAGCCATAGTTTGAAGAACAGCAGAGTTTTTGACCCAAGAATCATGCTCTTAGTTTACAATCTGTAT
5064585	TGCAGTTCAACATCATCAGCACTTTTAGTTTGTAGACAACACATTTATTGACCAAAGCTGAGTGCCATG
5072338	TGCAGCTCTGATTCACTTGTCTCAATATTTGAACTTCCACGGTATGAAGGTATATATTTCAAATATG
5073424	TGCAGAGAAATTCCTCCACACACCGAGGCTCGTTTGGGAAGTAACGTTTTTACTAAAAAAATACAG
5082606	TGCAGCAAGCAAAAAAAGTGAATCCCTCGGAACACTGGACTCGCCGAGGCTAGTCAGAGCCTCTTTTG
5084595	TGCAGTTGAATCTGTAAGAGTGAGACACCCATTAGGCACCCAATAACTTTAGAAATTCAGAAGAAAAAT
5085963	TGCAGCTTGTGGAGCTGCTTATAACTTTTTATTTTGAGAATCGGTGTCAGAGTGTGTTGGTGCATCTTTT
5087017	TGCAGTAACAAACATTATAATTTTTGTGTTGCATTATAAAAGCAAGACAATACTAATCAAGCCTTGATT
5087113	TGCAGCGGTTGATGCTGTACGTTTACTCACAGCAGCATCTCCTATCCCGACAAGATGGGTTGACTTTTA
5088162	TGCAGATTTTTTAGGACAACAAGATTCTCCAAGGTCTGAATAATTATGAGAACAAAAATATGCATTAC
5098382	TGCAGGGCATCAAACAGGACATTAGTCACAAGTAATTAGCATAAGCAACGGATACTATAATTACAGATC
5104630	TGCAGCTGTAGATGAGGAGGTACGCTAGTCTAATTTTAGAGGAATAAAGATGGAGAGTTTTACAGATCG
5111164	TGCAGTACCAGATGGACAAGGAGAGAGAAGCGAGTTAGAGGTCTTTACAGATCGGAAGAGCGGTTACAG
5111268	TGCAGAGGTAATCTACACCTATCAGTTCACCTCCCATATGTCGTTGAGTACAAACAAGGACACAAAGTT
5112599	TGCAGGAAATCCGGCATTTAGAGGCATTTACAGATCGGAAGAGCGGTTACAGCAGGAATGCCGAGACCGA
5114089	TGCAGTCCTTTTTGATTGTTTCTGATCCTTCCATCACGAAACATATATTACAGATCGGAAGAGCGGTTT
5114985	TGCAGCAAACACTTGCTAATTTTCACTCTATCTTACCAAAATGAGGCATTACAGATCGGAAGAGCGGTT
5123053	TGCAGGTCATTTTATAATGTGACACATCGGAGATTACAGATCGGAAGAGCGGTTACAGCAGGAATGCCGA
5123475	TGCAGAGCAATTACCACAGTTTATCGCTGCTGGCAACTTACAGATCGGAAGAGCGGTTACAGCAGGAATG
5123751	TGCAGCTGCCTTTCGTCATGTGGCTTGCCATCAAGAAACCAGAGAAGTTTAGCTTGTCTTGGCTTATTA
5123950	TGCAGCCATTGGTGCCTACCTGGAGTTCCATACACATGGAATGAACTTCTGTATGCTCACCTACTGTCA
5124128	TGCAGCTAGAATAATGTCTACCCACAATCAAATATAGATCTTACAGATTATTACAGATCGGAAGAGCGG
5125626	TGCAGTCGGGCCACTTTTTTTTAGTTGGTGTGAATGTCCCGTAATTTATTACAGATCGGAAGAGCGGTT
5128967	TGCAGAAGAAAGGCACACATGAGCAGCTGTGATGGAATAACCCAGTTGCAATCTACCACATATTACAGA
5130194	TGCAGTGGTGGAAGCTTGACATCTGGAGTTTTACAGATCCGAAGAGCGGTTACAGCAGGAATGCCGAGA
5130531	TGCAGTCAGCATAACCGGTGCAGCTATCTGCCACATCTTCTTTGATTCATCTCGACATTTTCTGTTACA
5132370	TGCAGCTGTCCTTTCACCAAGTAACGCATATAAGTTTGCATCGACGAGTTGCTATTCCAAATAAGAGCA
5132432	TGCAGGTTTCTTCTTTCACCCAAAAAAGCCCTCATTTTGGTTCAATTTTACAGATCGGAAGAGCGG
5132791	TGCAGCAATTTGCTAATAACCTCCTCATCCTTGCCTTCAAATCAATGCCTTCAAAAACAAATCTCATA
5134558	TGCAGCTGAAGTGGTAAAGAGGCTTGTTGAAGGGGCTCAACTCCTCATCCATCAAAGATTTACAGATCG
5135436	TGCAGGAAAGGCAAGGGAATAACAACAAATAATTACAGATCGGAAGAGCGGTTACAGCAGGAATGCCGA

5136045	TGCAGCAATAATTTACTTCTTCTATGTTTGGCTATTACAGAAGCCTTGTTCTGTTCAAAC TTCAGCTTT
5136108	TGCAGAACACATGACTCCAATTTGAAGGATACATCCCAAGCACTCCAGATTACAGATCGGAAGAGCGGT
5136623	TGCAGGGACGTTGGCCTATGTTGAGTAGATTACAAATCGGAAGAGCGGTT CAGCAGGAATGCCGAGACC
5136790	TGCAGCTACAAACCAATCCATAGTCACAGCTCAAGGAGTCACTACAGTTACAGACCAAGCCATAAGACC
5136794	TGCAGGTCCAAGCACTTCAACACGTGTGGCTGTTGATTCTGAAGAAGCCATGCCTGATGTTGACAGTGA

Appendix 3.2: BioProject PRJNA398959

Accession	Sample Name	Organism	Tax ID	Cultivar	Isolate	BioProject
SAMN07542559	1	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/1	PRJNA398959
SAMN07542560	2	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/2	PRJNA398959
SAMN07542561	3	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/3	PRJNA398959
SAMN07542562	4	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/4	PRJNA398959
SAMN07542563	5	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/5	PRJNA398959
SAMN07542564	6	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/7	PRJNA398959
SAMN07542565	7	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/8	PRJNA398959
SAMN07542566	8	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/9	PRJNA398959
SAMN07542567	9	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/10	PRJNA398959
SAMN07542568	10	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/11	PRJNA398959
SAMN07542569	11	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/12	PRJNA398959
SAMN07542570	12	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/13	PRJNA398959
SAMN07542571	13	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/15	PRJNA398959
SAMN07542572	14	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/16	PRJNA398959
SAMN07542573	15	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/17	PRJNA398959
SAMN07542574	16	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/18	PRJNA398959
SAMN07542575	17	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/19	PRJNA398959
SAMN07542576	18	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/21	PRJNA398959
SAMN07542577	19	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/22	PRJNA398959
SAMN07542578	20	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/23	PRJNA398959
SAMN07542579	21	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/24	PRJNA398959
SAMN07542580	22	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/27	PRJNA398959
SAMN07542581	23	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/28	PRJNA398959
SAMN07542582	24	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/31	PRJNA398959
SAMN07542583	25	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/32	PRJNA398959
SAMN07542584	26	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/33	PRJNA398959
SAMN07542585	27	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/34	PRJNA398959
SAMN07542586	28	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/35	PRJNA398959
SAMN07542587	29	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/36	PRJNA398959
SAMN07542588	30	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/37	PRJNA398959
SAMN07542589	31	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/38	PRJNA398959
SAMN07542590	32	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/39	PRJNA398959
SAMN07542591	33	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/40	PRJNA398959
SAMN07542592	34	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/41	PRJNA398959

SAMN07542593	35	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/42	PRJNA398959
SAMN07542594	36	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/43	PRJNA398959
SAMN07542595	37	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/44	PRJNA398959
SAMN07542596	38	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/45	PRJNA398959
SAMN07542597	39	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/47	PRJNA398959
SAMN07542598	40	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/48	PRJNA398959
SAMN07542599	41	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/49	PRJNA398959
SAMN07542600	42	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/50	PRJNA398959
SAMN07542601	43	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/51	PRJNA398959
SAMN07542602	44	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/53	PRJNA398959
SAMN07542603	45	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/55	PRJNA398959
SAMN07542604	46	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/57	PRJNA398959
SAMN07542605	47	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/58	PRJNA398959
SAMN07542606	48	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/59	PRJNA398959
SAMN07542607	49	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/60	PRJNA398959
SAMN07542608	50	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/61	PRJNA398959
SAMN07542609	51	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/62	PRJNA398959
SAMN07542610	52	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/63	PRJNA398959
SAMN07542611	53	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/65	PRJNA398959
SAMN07542612	54	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/68	PRJNA398959
SAMN07542613	55	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/69	PRJNA398959
SAMN07542614	56	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/70	PRJNA398959
SAMN07542615	57	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/71	PRJNA398959
SAMN07542616	58	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/73	PRJNA398959
SAMN07542617	59	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/74	PRJNA398959
SAMN07542618	60	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/75	PRJNA398959
SAMN07542619	61	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/77	PRJNA398959
SAMN07542620	62	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/78	PRJNA398959
SAMN07542621	63	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/79	PRJNA398959
SAMN07542622	64	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/80	PRJNA398959
SAMN07542623	65	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/81	PRJNA398959
SAMN07542624	66	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/82	PRJNA398959
SAMN07542625	67	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/83	PRJNA398959
SAMN07542626	68	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/85	PRJNA398959
SAMN07542627	69	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/86	PRJNA398959
SAMN07542628	70	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/88	PRJNA398959

SAMN07542629	71	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/89	PRJNA398959
SAMN07542630	72	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/91	PRJNA398959
SAMN07542631	73	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/92	PRJNA398959
SAMN07542632	74	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/93	PRJNA398959
SAMN07542633	75	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/94	PRJNA398959
SAMN07542634	76	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/95	PRJNA398959
SAMN07542635	77	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/96	PRJNA398959
SAMN07542636	78	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/97	PRJNA398959
SAMN07542637	79	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/98	PRJNA398959
SAMN07542638	80	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/99	PRJNA398959
SAMN07542639	81	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/100	PRJNA398959
SAMN07542640	82	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/101	PRJNA398959
SAMN07542641	83	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/102	PRJNA398959
SAMN07542642	84	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/103	PRJNA398959
SAMN07542643	85	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/104	PRJNA398959
SAMN07542644	86	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/105	PRJNA398959
SAMN07542645	87	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/106	PRJNA398959
SAMN07542646	88	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/107	PRJNA398959
SAMN07542647	89	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/108	PRJNA398959
SAMN07542648	90	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/109	PRJNA398959
SAMN07542649	91	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/110	PRJNA398959
SAMN07542650	92	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/111	PRJNA398959
SAMN07542651	93	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/112	PRJNA398959
SAMN07542652	94	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/113	PRJNA398959
SAMN07542653	95	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/115	PRJNA398959
SAMN07542654	96	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/116	PRJNA398959
SAMN07542655	97	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/118	PRJNA398959
SAMN07542656	98	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/119	PRJNA398959
SAMN07542657	99	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/120	PRJNA398959
SAMN07542658	100	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/121	PRJNA398959
SAMN07542659	101	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/122	PRJNA398959
SAMN07542660	102	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/124	PRJNA398959
SAMN07542661	103	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/125	PRJNA398959
SAMN07542662	104	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/126	PRJNA398959
SAMN07542663	105	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/128	PRJNA398959
SAMN07542664	106	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/129	PRJNA398959

SAMN07542665	107	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/131	PRJNA398959
SAMN07542666	108	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/133	PRJNA398959
SAMN07542667	109	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TFRK 504/134	PRJNA398959
SAMN07542668	110	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/1	PRJNA398959
SAMN07542669	111	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/2	PRJNA398959
SAMN07542670	112	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/3	PRJNA398959
SAMN07542671	113	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/4	PRJNA398959
SAMN07542672	114	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/5	PRJNA398959
SAMN07542673	115	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/6	PRJNA398959
SAMN07542674	116	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/7	PRJNA398959
SAMN07542675	117	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/8	PRJNA398959
SAMN07542676	118	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/9	PRJNA398959
SAMN07542677	119	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/10	PRJNA398959
SAMN07542678	120	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/11	PRJNA398959
SAMN07542679	121	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/13	PRJNA398959
SAMN07542680	122	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/15	PRJNA398959
SAMN07542681	123	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/16	PRJNA398959
SAMN07542682	124	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/17	PRJNA398959
SAMN07542683	125	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/18	PRJNA398959
SAMN07542684	126	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/19	PRJNA398959
SAMN07542685	127	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/22	PRJNA398959
SAMN07542686	128	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/23	PRJNA398959
SAMN07542687	129	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/25	PRJNA398959
SAMN07542688	130	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/26	PRJNA398959
SAMN07542689	131	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/27	PRJNA398959
SAMN07542690	132	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/28	PRJNA398959
SAMN07542691	133	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/29	PRJNA398959
SAMN07542692	134	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/30	PRJNA398959
SAMN07542693	135	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/31	PRJNA398959
SAMN07542694	136	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/32	PRJNA398959
SAMN07542695	137	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/34	PRJNA398959
SAMN07542696	138	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/35	PRJNA398959
SAMN07542697	139	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/36	PRJNA398959
SAMN07542698	140	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/38	PRJNA398959
SAMN07542699	141	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/39	PRJNA398959
SAMN07542700	142	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/40	PRJNA398959

SAMN07542701	143	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/41	PRJNA398959
SAMN07542702	144	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/42	PRJNA398959
SAMN07542703	145	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/43	PRJNA398959
SAMN07542704	146	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/44	PRJNA398959
SAMN07542705	147	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/45	PRJNA398959
SAMN07542706	148	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/48	PRJNA398959
SAMN07542707	149	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/49	PRJNA398959
SAMN07542708	150	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/50	PRJNA398959
SAMN07542709	151	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/51	PRJNA398959
SAMN07542710	152	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/52	PRJNA398959
SAMN07542711	153	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/53	PRJNA398959
SAMN07542712	154	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/55	PRJNA398959
SAMN07542713	155	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/56	PRJNA398959
SAMN07542714	156	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/57	PRJNA398959
SAMN07542715	157	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/58	PRJNA398959
SAMN07542716	158	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/61	PRJNA398959
SAMN07542717	159	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/62	PRJNA398959
SAMN07542718	160	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/63	PRJNA398959
SAMN07542719	161	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/64	PRJNA398959
SAMN07542720	162	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/65	PRJNA398959
SAMN07542721	163	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/66	PRJNA398959
SAMN07542722	164	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/67	PRJNA398959
SAMN07542723	165	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/68	PRJNA398959
SAMN07542724	166	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/69	PRJNA398959
SAMN07542725	167	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/70	PRJNA398959
SAMN07542726	168	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/71	PRJNA398959
SAMN07542727	169	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/72	PRJNA398959
SAMN07542728	170	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/73	PRJNA398959
SAMN07542729	171	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/74	PRJNA398959
SAMN07542730	172	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/75	PRJNA398959
SAMN07542731	173	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/76	PRJNA398959
SAMN07542732	174	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/77	PRJNA398959
SAMN07542733	175	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/78	PRJNA398959
SAMN07542734	176	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/79	PRJNA398959
SAMN07542735	177	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/80	PRJNA398959
SAMN07542736	178	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/81	PRJNA398959

SAMN07542737	179	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/82	PRJNA398959
SAMN07542738	180	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/83	PRJNA398959
SAMN07542739	181	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/84	PRJNA398959
SAMN07542740	182	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/85	PRJNA398959
SAMN07542741	183	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/86	PRJNA398959
SAMN07542742	184	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/87	PRJNA398959
SAMN07542743	185	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/93	PRJNA398959
SAMN07542744	186	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/94	PRJNA398959
SAMN07542745	187	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/96	PRJNA398959
SAMN07542746	188	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/97	PRJNA398959
SAMN07542747	189	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/102	PRJNA398959
SAMN07542748	190	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/104	PRJNA398959
SAMN07542749	191	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/105	PRJNA398959
SAMN07542750	192	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/106	PRJNA398959
SAMN07542751	193	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/108	PRJNA398959
SAMN07542752	194	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/113	PRJNA398959
SAMN07542753	195	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/114	PRJNA398959
SAMN07542754	196	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/115	PRJNA398959
SAMN07542755	197	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/118	PRJNA398959
SAMN07542756	198	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/119	PRJNA398959
SAMN07542757	199	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/121	PRJNA398959
SAMN07542758	200	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/122	PRJNA398959
SAMN07542759	201	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/123	PRJNA398959
SAMN07542760	202	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/124	PRJNA398959
SAMN07542761	203	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/125	PRJNA398959
SAMN07542762	204	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/126	PRJNA398959
SAMN07542763	205	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/127	PRJNA398959
SAMN07542764	206	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/130	PRJNA398959
SAMN07542765	207	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/131	PRJNA398959
SAMN07542766	208	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/132	PRJNA398959
SAMN07542767	209	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/134	PRJNA398959
SAMN07542768	210	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/136	PRJNA398959
SAMN07542769	211	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/137	PRJNA398959
SAMN07542770	212	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/138	PRJNA398959
SAMN07542771	213	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/139	PRJNA398959
SAMN07542772	214	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/140	PRJNA398959

SAMN07542773	215	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/141	PRJNA398959
SAMN07542774	216	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/142	PRJNA398959
SAMN07542775	217	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/143	PRJNA398959
SAMN07542776	218	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/144	PRJNA398959
SAMN07542777	219	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/145	PRJNA398959
SAMN07542778	220	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/146	PRJNA398959
SAMN07542779	221	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/147	PRJNA398959
SAMN07542780	222	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/149	PRJNA398959
SAMN07542781	223	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/150	PRJNA398959
SAMN07542782	224	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/154	PRJNA398959
SAMN07542783	225	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/155	PRJNA398959
SAMN07542784	226	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/156	PRJNA398959
SAMN07542785	227	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/157	PRJNA398959
SAMN07542786	228	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/159	PRJNA398959
SAMN07542787	229	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/160	PRJNA398959
SAMN07542788	230	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/162	PRJNA398959
SAMN07542789	231	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/163	PRJNA398959
SAMN07542790	232	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/165	PRJNA398959
SAMN07542791	233	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/166	PRJNA398959
SAMN07542792	234	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/167	PRJNA398959
SAMN07542793	235	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/168	PRJNA398959
SAMN07542794	236	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/169	PRJNA398959
SAMN07542795	237	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/170	PRJNA398959
SAMN07542796	238	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/172	PRJNA398959
SAMN07542797	239	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/173	PRJNA398959
SAMN07542798	240	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/175	PRJNA398959
SAMN07542799	241	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/176	PRJNA398959
SAMN07542800	242	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/177	PRJNA398959
SAMN07542801	243	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/178	PRJNA398959
SAMN07542802	244	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/179	PRJNA398959
SAMN07542803	245	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/182	PRJNA398959
SAMN07542804	246	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/183	PRJNA398959
SAMN07542805	247	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/184	PRJNA398959
SAMN07542806	248	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/185	PRJNA398959
SAMN07542807	249	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/186	PRJNA398959
SAMN07542808	250	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/187	PRJNA398959

SAMN07542809	251	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/188	PRJNA398959
SAMN07542810	252	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/189	PRJNA398959
SAMN07542811	253	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/190	PRJNA398959
SAMN07542812	254	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/192	PRJNA398959
SAMN07542813	255	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/194	PRJNA398959
SAMN07542814	256	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/195	PRJNA398959
SAMN07542815	257	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/196	PRJNA398959
SAMN07542816	258	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/197	PRJNA398959
SAMN07542817	259	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/198	PRJNA398959
SAMN07542818	260	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/199	PRJNA398959
SAMN07542819	261	<i>Camellia sinensis</i>	4442	<i>Camellia sinensis</i>	TRFK 524/200	PRJNA398959

Appendix 3.3: Peer-reviewed scientific article based on results obtained in Chapter 3

Tree Genetics & Genomes (2018) 14: 9
<https://doi.org/10.1007/s11295-017-1219-8>

ORIGINAL ARTICLE



Identification of novel QTL for black tea quality traits and drought tolerance in tea plants (*Camellia sinensis*)

Robert K. Koech^{1,2} · Pelly M. Malebe¹ · Christopher Nyarukowa¹ · Richard Mose³ · Samson M. Kamunya² · Zeno Apostolides¹

Received: 25 August 2017 / Revised: 6 December 2017 / Accepted: 17 December 2017 / Published online: 8 January 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Tea (*Camellia sinensis*) contains polyphenols and caffeine which have been found to be of popular interest in tea quality. Tea production relies on well-distributed rainfall which influence tea quality. Phenotypic data for two segregating tea populations TRFK St 504 and TRFK St 524 were collected and used to identify the quantitative trait loci (QTL) influencing tea biochemical and drought stress traits based on a consensus genetic map constructed using the DArTseq platform. The populations comprised 261 F₁ clonal progeny. The map consisted of 15 linkage groups which corresponds to chromosome haploid number of tea plant ($2n = 2 \times = 30$) and spanned 1260.1 cM with a mean interval of 1.1 cM between markers. A total of 16 phenotypic traits were assessed in the two populations. Both interval and multiple QTL mapping revealed a total of 47 putative QTL in the 15 LGs associated with tea quality and percent relative water content at a significant genome-wide threshold of 5%. In total, six caffeine QTL, 25 catechins QTL, three theaflavins QTL, nine QTL for tea taster score, and three QTL for percent relative water contents were detected. Out of these 47 QTL, 19 QTL were identified for ten traits in three main regions on LG01, LG02, LG04, LG12, LG13, and LG14. The QTL associated with caffeine, individual catechins, and theaflavins were clustered mostly in LG02 and LG04 but in different regions on the map. The explained variance by each QTL in the population ranged from 5.5 to 56.6%, with an average of 9.9%. Identification of QTL that are tightly linked to markers associated with black tea quality coupled with UPLC assay may greatly accelerate development of novel tea cultivars owing to its amenability at seedling stage. In addition, validated molecular markers will contribute greatly to adoption of marker-assisted selection (MAS) for drought tolerance and tea quality improvement.

Keywords *Camellia sinensis* · Tea quality · Drought tolerance · UPLC · DArTseq · Linkage map

Abbreviations

DArT Diversity Arrays Technology
MAS Marker-assisted selection

QTL Quantitative trait loci
LOD Logarithm of odds
LGs Linkage groups
UPLC Ultra-performance liquid chromatography
CAFF Caffeine
CAT (+)-Catechin
EC (–)-Epicatechin
ECG (–)-Epicatechin gallate
EGC (–)-Epigallocatechin
EGCG (–)-Epigallocatechin gallate
TF1 Simple theaflavin
TF2 Theaflavin -3-monogallate
TF3 Theaflavin-3'-monogallate
TF4 Theaflavin-3, 3'-digallate
CL Liquor color
BRT Liquor brightness
AST Astringency

Communicated by W.-W. Guo

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11295-017-1219-8>) contains supplementary material, which is available to authorized users.

✉ Zeno Apostolides
za@up.ac.za

¹ Department of Biochemistry, University of Pretoria, Pretoria 0002, South Africa

² Kenya Agriculture and Livestock Research Organization, Tea Research Institute, P.O. Box 820, Kericho 20200, Kenya

³ James Finlay (Kenya) Limited, P.O. Box 223, Kericho 20200, Kenya

BRK	Liquor briskness
AR	Liquor aroma
RWC	Relative water content
St	Stock
TRFK	Tea Research Foundation of Kenya

Introduction

Tea plant (*Camellia sinensis* (L.) O. Kuntze) is the second most popular non-alcoholic beverage consumed in the world after water. The main cultivated taxa comprise of three natural hybrids, viz., *C. sinensis* (L.) O. Kuntze (“China” type), *C. assamica* ssp. *assamica* (Masters) Wight (“Assam” jat) and the “Cambod” variety, *C. assamica* ssp. *lasiocalyx* (Planchon ex Watt) Wight (Banerjee 1992). The China type has small semi-erect leaves and the Assam type has relatively large, horizontal leaves. The Cambod type is currently considered as a hybrid of *C. sinensis* var. *sinensis* type and *C. sinensis* var. *assamica* type (Meegahakumbura et al. 2016; Wambulwa et al. 2016). Tea is cultivated widely in many Asian and African countries, and it contributes significantly to the local economy. In Kenya, tea is the single largest agribusiness and contributes over 26% of all foreign exchange earnings and over 4% of the gross domestic product (Kenya National Bureau of Statistics 2012; AFFA 2013).

Tea, being a popular beverage, has attracted the attention of both the consumers and the scientific community because of its multiple health-promoting effects (Preedy 2012). Tea is known to contain a bioactive compounds, such as the flavonoids, caffeine, L-theanine, and γ -aminobutyric acid (GABA) (Lin et al. 2012; da Silva Pinto 2013; Ma et al. 2014). Catechins, a group of flavonoids, are major constituents in tea and contribute up to 30% of the dry weight of the tender tea shoots (Pang et al. 2013; Ma et al. 2014). In green tea leaves, catechins consist mainly of five flavan-3-ols: (+)-catechin (CAT), (–)-epicatechin (EC), (–)-epicatechin gallate (ECG), (–)-epigallocatechin (EGC), and (–)-epigallocatechin gallate (EGCG) (Ananingsih et al. 2013; Ma et al. 2014). The most abundant catechin is the EGCG which accounts for up to 50–80% of the total catechins (Sang et al. 2011). Theaflavins and thearubigins are the major polyphenols found in black tea formed by the enzymatic oxidation and polymerization of catechins during the so-called “fermentation stage” (Owuor et al. 2006b). The major individual theaflavins in black tea include simple theaflavin (TF), theaflavin-3-monogallate (TF-3-g), theaflavin-3'-monogallate (TF-3'-g), and theaflavin-3, 3'-digallate (TFdg). Catechins in green tea and theaflavins in black tea have demonstrated to have a wide variety of pharmacological activities, which include antioxidant, anti-inflammatory, anticancer, antidiabetic, and neuroprotective effects (Preedy 2012).

Tea leaves also contain 2–5% of caffeine and much smaller amounts of theobromine and theophylline (Khan and Mukhtar 2007). Caffeine is not only known for its stimulatory effect (Sharangi 2009), but contributes to tea briskness and other taste characteristics, which is considered an important parameter for the evaluation of tea quality (Obanda et al. 1997; Nitin et al. 2006). Caffeine has also been reported to suppress body weight gain by suppressing food intake and reducing adipose tissue (Kobayashi-Hattori et al. 2005). These health benefits associated with tea intake have led to high consumption of tea products as tea extracts in pharmaceutical, food and beverage, and cosmetic industries. In addition to these health benefits, tea quality is a complex phenomenon and depends mainly on the biochemical components that determine color, aroma, and taste, attached to the tea infusions. As such, tea prices vary greatly depending on the tea quality, which has traditionally been assessed by tea tasters by organoleptic evaluation, to describe various quality attributes of a tea infusion. Furthermore, Kenya tea industry has lately been facing challenges of ever-increasing temperatures, prolonged drought periods, and changing precipitation patterns especially in the tea growing areas (FAO 2015). Therefore, this calls for concerted efforts in tea breeding programs to develop tea cultivars with improved biochemical content and drought tolerance.

In a previous study, a method for Short-time Withering Assessment of Probability for Drought Tolerance (SWAPDT) in *Camellia sinensis* validated by targeted metabolomics (Nyarukowa et al. 2016) was developed for predicting the drought tolerance of tea cultivars. The metabolite profiles obtained using the SWAPDT method showed that drought tolerant tea cultivars differed from drought susceptible tea cultivars. The SWAPDT method relied on the percent relative water content (%RWC) of tea leaves after the five-hour withering method. Therefore, the SWAPDT method which was used in the current study provides a basis for selection of new drought tolerant tea cultivars that may lead to improvement of crop productivity, amidst challenges imposed by drought due to climate change.

The tea plant is a woody long-lived perennial tree characterized by a large diploid genome of approximately 3Gb and chromosome number ($2n = 2 \times = 30$) that is self-incompatible, highly heterozygous, and has a long juvenile period of 4–5 years and about 22–25 years to breed a new cultivar (Chen et al. 2007; Shi et al. 2011). Presently, *Camellia* genus is believed to comprise more than 300 species with *Camellia sinensis* (L.) O. Kuntze as the most important agronomic species (Mondal et al. 2004). The recent discovery of *C. cherryana* in 2012 (Orel and Wilson 2012) indicates the instability and high out-crossing character or property of the genus. These attributes have led to slow progress in tea breeding; however, tremendous progress has been made regarding the availability of good genetic resources for tea breeding.

Recently, Bali et al. (2015) constructed a genetic linkage map using the two-way pseudo-testcross approach for mapping drought-tolerant trait of Indian tea cultivars. Construction of a genetic linkage map for tea using pseudo-testcross approach has been used previously (Wachira et al. 1995; Hackett et al. 2000; Kamunya et al. 2010; Taniguchi et al. 2012; Hu et al. 2013; Ma et al. 2014) due to high out-crossing and heterozygous nature of the tea. In spite of all these advances, these genetic linkage maps are mainly based on dominant markers such as random amplified polymorphic DNA (RAPD) or amplified fragment length polymorphism (AFLP) (Wachira et al. 1995; Hackett et al. 2000; Tanaka 2000; Kamunya et al. 2009, 2010) and co-dominant markers such as simple sequence repeats (SSRs) (Hu et al. 2013; Ma et al. 2014; Tan et al. 2016). Also, the mapping populations used in these previous studies are relatively small (52 to 92 F1 progenies), and a number of gaps that are in between the markers in some of the linkage groups may decrease the accuracy of detected QTL. Thus, these marker technologies have not fully provided a widely applicable tool that can be used to link genotypes to phenotypes. However, SNP discovery and genotyping based on specific-locus amplified fragment sequencing (SLAF-seq) generated a saturated genetic linkage map length of 3965 cM, with an average inter-locus distance of 1.0 cM (Ma et al. 2015).

Diversity Arrays Technology (DArT) which involves the isolation and cloning of a random set of DNA fragments from complexity-reduced DNA sample has experienced increased interest in the recent years (Gupta et al. 2013). DArT has provided a standardized high-throughput genotyping, whereby thousands of markers can be readily assayed in parallel in thousands of samples without prior sequence information (Jaccoud et al. 2001; Wittenberg et al. 2005). The DArT array technology has demonstrated excellent performance for phylogenetic and diversity analyses (Steane et al. 2011), genomic selection (Poland et al. 2012), and genetic linkage mapping (Schouten et al. 2012). DArT platform has been applied successfully to construct more dense genetic linkage maps in plants with more complex genomes such as apple (Schouten et al. 2012), strawberry (Sánchez-Sevilla et al. 2015), sugarcane (Racedo et al. 2016), wheat (Zou et al. 2016), citrus (Liu et al. 2016), and trees such as Eucalyptus (Sansaloni et al. 2010; Steane et al. 2011). Therefore, a genetic linkage map based on DArTseq platform could provide a vital tool for relating phenotype to genotype, hence facilitating identification and selection of recombinant parents with desirable attributes. The genetic linkage map will also increase the speed and precision of tea breeding and improvement programmes. The present study constructed a high-density linkage map for tea plant by integrating DArTseq platform and ultra-performance liquid chromatography (UPLC) technique for QTL linked to catechins, theaflavins, caffeine content, tea taster scores, and percent relative water content for future MAS breeding.

Materials and methods

Plant material

Two pseudo-testcross populations consisting of 109 progeny from TRFK St. 504 (TRFK 303/577 × GW Ejulu) and reciprocal cross consisting of 152 progeny-coded TRFK St. 524 (GW Ejulu × TRFK 303/577), were developed by crossing two heterozygous parental clones TRFK 303/577 and GW Ejulu. The two parental clones, TRFK 303/577 and GW Ejulu, are of Assamica and China varieties, respectively. The two clones were chosen on the basis of their contrasting attributes. GW Ejulu is a low-yielding, high black tea quality and moderate levels of caffeine, but high in total catechins and individual catechin contents. Clone TRFK 303/577, which is an open-pollinated progeny of clone TRFK 6/8, is high yielding, drought tolerant, medium in black tea quality, caffeine, and individual catechins.

Description of study site

The cross comprising all above clonal progenies and their parents was established in 2007 at Tea Research Institute (TRI), in Timbilil Estate (Kenya, Kericho county, latitude 0° 22' S, longitude 35° 21' E, altitude 2180 m AMSL). The trial populations are set up in a complete randomized block design with three replications in plots of 15 plants spaced at 0.61 m within rows and 1.22 m between rows. The populations had been receiving 150 Kg N per hectare per year in the form of NPKS 25:5:5:5 compound fertilizer. The trial populations were brought into bearing following the recommended management practices (Anon, 2002; Kamunya et al. 2010).

Sample collection and processing

About 1 Kg of fresh shoots in the form of two leaves and a bud was plucked from heterozygous parents and three independent F₁ progeny in a complete randomized block design between September 2013 and February 2014. The fresh-plucked shoots were placed in appropriately labeled sampling bags for processing at the TRI miniature factory. Five hundred grams of fresh green tea leaves were used to produce green tea by immediately drying using a microwave for 5 min, which also deactivated the oxidizing enzyme polyphenol oxidase. The remaining 500 g was used to manufacture black tea using a standard black tea manufacturing procedure by physically withering the tea leaves up to a moisture content of 50–65% for 18 h. The leaf was passed through the CTC rollers four times to achieve maceration equivalent to that obtained using commercial CTC rollers in black tea manufacture (Owuor and Othieno 1991). Upon maceration, the tea (dhool) was then aerated for 90 min at ambient temperature (22–26 °C) and 100% relative humidity to achieve enzymatic oxidation or

so-called “fermentation.” A bench top fluid-bed drier system (TeaCraft Ltd) was utilized in firing the tea, initially at 120 °C for 20–25 min then brought to 100 °C for 10 min. Green and black tea samples were grounded separately using a coffee grinder and stored in sealed silver-lined aluminum sachets for UPLC analysis and organoleptic assessment for processed black tea.

Extraction of catechins, caffeine, and theaflavins

International Organization for Standardization (ISO) extraction method was used as is described in the ISO14502-2 2005. Briefly, 0.200 ± 0.001 g of green and black tea samples were weighed out on a Mettler Toledo analytical balance (Microsep, South Africa) and transferred into a 20 ml glass test tube. A 5 ml volume of 70:30 methanol (Merck, South Africa): water (v/v) at 70 °C was added to each extraction tube, stoppered, and vortex mixed for approximately 5 s before placing into a water bath set at 70 °C. The extraction mixture was vortex mixed after 5 min and after 10 min. The extracts were cooled and centrifuged at $2000 \times g$ using Thermo Scientific Heraeus Labofuge (Sepsci, South Africa) 300 centrifuge for 10 min, with the resultant supernatant decanted into a 10-ml volumetric flask. The extraction step was repeated twice. Both extracts were pooled, and the volume adjusted to 10 ml with cold 70:30 methanol: water (v/v). One milliliter volume of the extracts was then diluted to 5 mL with stabilizing solution (10% v/v acetonitrile, 500 mg/mL EDTA, and 10 mg/ml ascorbic acid, Sigma-Aldrich, South Africa). The dilution was filtered through 0.2 μ m Minisart®RC4 syringe filters (Sartorius, South Africa) with hydrophilic, solvent-resistant regenerated cellulose membranes and ready for UPLC analysis.

UPLC analysis of catechins, caffeine, and theaflavins

The UPLC analysis was performed on Waters ACQUITY UPLC H-Class system (Waters, Milford, MA, USA) equipped with a binary solvent delivery pump, an autosampler, and a photodiode array detector (PDA) and controlled by the Empower-3 software. Separation was achieved on Waters Acquity HSS T3 column (1.8 μ m, 2.1×150 mm). The mobile phase was composed of solvent A (2% acetic acid and 9% acetonitrile in water, pH 2.8) and solvent B (2% acetic acid and 80% of acetonitrile in water) with gradient elution: 0 min (5% A), 0–21 min (5–20% A), 21–30 min (20–25% A), 30–32 min (25–100% A), 32–39 min (100–100% A), 39–40 min (100–5% A), and 40–45 min (5–5% A), which was used in this study. The mobile phase was filtered through a 0.2- μ m cellulose acetate membrane filter and degassed using a Neuberger Laboport (Labotech, South Africa) vacuum pump. The sample injection volume was 5 μ L, and the flow-rate was

set at 0.2 mL/min. The standards consisted of catechins (CAT, EC, ECG, EGC and EGCG), caffeine, and gallic acid (Sigma-Aldrich, South Africa). Tryptamine, sulfanilamide, and mycophenolic acid (Sigma-Aldrich, South Africa) were used as internal standards for quality control. Three mixed standards: mixture A (catechins, caffeine, gallic acid and three internal standards), mixture B and mixture C which both had all the standards except the three internal standards, were used. The identification and quantification of individual catechins, caffeine, and individual theaflavins were performed at 278 nm. The individual catechins and caffeine peaks in the tea samples were identified based on retention times of the individual catechins and caffeine standards. The individual theaflavins in black tea samples were quantified as EGCG equivalents, based on the response factor of EGCG.

Determination of plain black tea quality parameters

Tea taster scores

The tea taster's score on black teas was performed on the 261 progenies and their parents. A sample of black tea (2.5 g) was extracted with 150 ml of boiling water for about 5 min. The tea sample was then subjected to organoleptic evaluation made by professional tea tasters (WM Cahn, Johannesburg, South Africa) with scores based on liquor color (CL), liquor brightness (BRT), astringency (AST), liquor briskness (BRK), and liquor aroma (AR). The scores were in the range of 1–5 with the low score representing poor quality and a high score representing high quality. Black teas that are known to be of high quality were used as a control to determine the quality of progenies regarding liquor characteristics and taste.

Determination of percent relative water content

The % RWC in the tea two populations was determined using the SWAPDT method as described by Nyarukowa et al. (2016). Three shoots with two leaves and a bud collected from heterozygous parents and three independent F_1 progeny in a complete randomized block design were immersed in water at room temperature and weighed after 24 h. The leaves were then oven dried at 37 °C and weighed after 5 h. The leaves were again placed in water and weighed after 24 h, and then oven dried at 105 °C for 24 h to obtain each leaf's dry weight.

The % RWC was calculated using the formula:

$$\%RWC = [(F_{wt} - D_{wt}) / (FT_{wt} - D_{wt})] \times 100$$

where F_{wt} is the fresh weight, D_{wt} is the dry weight after 24 h in 105 °C oven, and FT_{wt} is the weight after 24 h rehydration.

Extraction, purification, and quantification of DNA

Two tender young leaves and a terminal bud were plucked from a tea bush planted in a complete randomized block design experiments as a representative sample and preserved in zip-lock plastic bags containing dry silica gel (Malebe 2011). The silica gel was dried in the oven at 70 °C for 48 h to enable it to absorb any surface moisture from the preserved leaves. The green tea leaf samples were stored at −20 °C before DNA extraction using a modified method of (Gawal and Jarret 1991). The NanoDrop spectrophotometer (NanoDrop Technologies, South Africa) which gives the concentration of DNA in nanogram/microliter was used to quantify the DNA. The integrity of DNA was also checked using the agarose gel electrophoresis method (Adkins et al. 2007).

DArTseq assay

The DNA samples were shipped to Diversity Array Technology Pty Ltd. (Canberra, Australia) for DArTseq analysis. DNA quality and digestibility were tested using restriction enzyme *Pst*I (Fermentas, Burlington, Canada) and *Eco*RI (Promega, Madison, USA). The DArTseq procedure was carried out as described by (Sansaloni et al. 2010) using *Pst*I and *Mse*I restriction enzymes. Markers were scored 0 or 1 for absence or presence of a polymorphism in the genomic representation of the sample.

Linkage map construction

Six thousand five hundred and eighty eight DArTseq markers derived from *C. sinensis* sequences were tested for segregation using genomic DNA from the two parents and 261 F₁ individuals. The obtained genotyping data was subjected to linkage analyses using JoinMap 4.0 (Van Ooijen 2006). Markers were grouped at a logarithm of odds (LOD) of (3–12) and a recombination frequency of 0.4. The genetic distances between markers were calculated with the Kosambi mapping function. The LOD thresholds at the genome-wide level were determined by running 1000 permutations ($P < 0.05$).

Phenotypic trait analysis

A total of 16 phenotypic traits were assessed for a normal distribution using the JMP 12 (Sall et al. 2012). The mean and the standard deviation for each phenotypic trait in each parent and the progenies were calculated. The significance difference between parental values and progenies was analyzed using Student's *t* test.

QTL analyses

The QTL mapping was carried out using the MapQTL 6.0 software (Van Ooijen 2006). The first set of markers were selected as cofactors from IM results, and the significant markers were selected using backwards elimination method. The significant markers at $P < 0.05$ were only used as cofactors in the multiple MQM method to refine QTL detection. The permutation analysis (1000 permutations) for significance thresholds was used for accepting the presence of a potential QTL (Churchill and Doerge 1994). The maximum LOD score on a linkage group was used to estimate the positions of QTL. The location and position of the QTL for individual phenotypic traits were indicated using MapChart software (Voorrips 2002).

Results

Phenotypic trait analysis

A wide range of variation in individual catechins, theaflavins, and caffeine contents was observed in the two parents and F₁ progeny. A chromatogram representing the profiles of black tea and green tea from the two parent lines is presented in Supplementary Fig. 1 and Supplementary Fig. 2, respectively. Besides the five peaks identified, there were other several peaks that were detected but unidentified. This is an indication that there are other several biochemical compounds present in the tea extracts. The mean, standard deviation, and coefficient of variation for individual catechins and caffeine content quantified in F₁ progeny and in both parents are shown in Tables 1 and 2. Tea liquor characteristics such as color,

Table 1 Phenotypic variation of caffeine and individual catechins content in parental lines (TRFK 303/577 and GW Ejulu) and 109 F₁ progeny St 504

Trait (%w/w)	TRFK 303/577 Mean ± SD	GW Ejulu Mean ± SD	F ₁ Mean ± SD	F ₁ Range	F ₁ CV (%)
Caffeine	2.12 ± 0.11	2.69 ± 0.13	2.25 ± 0.27	1.57–3.02	11.9
CAT	0.44 ± 0.11	1.51 ± 0.17	0.62 ± 0.32	0.17–1.64	53.1
EC	1.22 ± 0.21	0.90 ± 0.07	1.19 ± 0.41	0.51–2.63	34.4
ECG	2.33 ± 0.10	5.21 ± 0.03	3.31 ± 0.83	1.54–6.58	25.6
EGC	3.72 ± 0.72	1.02 ± 0.40	3.25 ± 1.07	0.60–5.79	33.7
EGCG	7.15 ± 0.65	6.50 ± 0.40	6.03 ± 1.13	1.78–12.33	18.9

Table 2 Phenotypic variation of caffeine and individual catechins content in parental lines (GW Ejulu and TRFK 303/577) and 152 F₁ progeny St 524

Trait (%w/w)	GW Ejulu Mean $\pm$ SD	TRFK 303/577 Mean $\pm$ SD	F ₁ Mean $\pm$ SD	F ₁ range	F ₁ CV (%)
Caffeine	2.69 $\pm$ 0.13	2.12 $\pm$ 0.11	2.38 $\pm$ 0.30	1.53–3.69	12.8
CAT	1.51 $\pm$ 0.17	0.44 $\pm$ 0.11	0.96 $\pm$ 0.36	0.10–2.13	37.8
EC	0.90 $\pm$ 0.07	1.22 $\pm$ 0.21	1.57 $\pm$ 0.51	0.66–3.40	32.6
ECG	5.21 $\pm$ 0.03	2.33 $\pm$ 0.10	2.82 $\pm$ 0.66	1.06–5.24	24.2
EGC	1.02 $\pm$ 0.40	3.72 $\pm$ 0.72	3.34 $\pm$ 1.05	0.60–7.61	32.3
EGCG	6.50 $\pm$ 0.40	7.15 $\pm$ 0.65	6.14 $\pm$ 1.24	1.78–12.33	20.4

brightness, and briskness as well as astringency and aroma were also analyzed in black tea samples as shown in Fig. 1.

The frequency distributions of phenotype datasets are shown in Fig. 1. The values of the two parents are indicated in the graphs. The levels of the individual catechins analyzed in green teas ranged from 0.1 to 12.3%, with low content of CAT and EC as compared to EGCG and EGC, which were the abundant catechins. Caffeine content ranged from 1.5 to 3.7%, while the levels of individual theaflavins in processed black teas ranged from 0.2 to 1.3% with TF3 being the most abundant theaflavin (Fig. 1). A variation in the levels of individual theaflavins content was observed in parental lines and progenies. Parental clone TRFK 303/577 had high levels of all four theaflavins as compared to parental clone GW Ejulu in both tea populations. A number of traits namely caffeine, EGC, TF1, TF2, TF3, and TF4 score as measured in the two populations (St 504 and St 524) had their F₁ means falling in between the two parental means for respective traits (Fig. 1). However, the F₁ means for EC, ECG, EGCG, and tea taster scores were marginally lower than the parental means. There were variations in tea taster scores in the parental clones and F₁ progenies. Most of the black tea samples had an average tea quality regarding tea taster's score except for aroma, which scored poorly. This poor aroma score might have been due to the time difference (12 months) between the black tea processing and the tea taster scoring. Tea liquor color (CL) scored highly than the other tea taster score traits in parental clone GW Ejulu as compared to parental clone TRFK 303/577. There were also variations of the percentage relative water content in F₁ progenies and parental clones in the two tea populations (Fig. 1). The % RWC ranged from 49 to 99% which show that there were tea plants that were either tolerant or susceptible to drought. Parental clone GW Ejulu which is drought susceptible tea cultivar had low % RWC of 55.1% while parental clone TRFK 303/577 which is drought tolerant cultivar had high % RWC of 74.1%. In overall, a large number of phenotypic traits for the progeny fell outside to those of the parents indicating transgressive

segregation, hence polygenic inheritance of the traits in question (Fig. 1). Further, the observed normal distribution for the phenotypic traits shows involvement of several genes in the genetic control of phenotypic traits.

Correlation analysis

Pearson's correlation coefficients between different observations of each trait in the two tea population are shown in Tables 3 and 4. The correlations among the phenotypic traits were grouped into: catechins (CAT, EC, ECG, EGC and EGCG), caffeine, theaflavins (TF1, TF2, TF3, and TF4), and tea taster's scores (CL, BRT, AST, BRK and AR) and % RWC. In general, the significant correlation coefficients mostly observed in St 504 were between EC and EGC ($r = 0.62$, $P < 0.01$), TF1 and TF2 ($r = 0.89$, $P < 0.001$), TF1 and TF4 ($r = 0.90$, $P < 0.001$), and TF2 and TF4 ($r = 0.89$, $P < 0.001$). For tea taster's score, a high correlation was observed between individual AST and BRK ($r = 0.96$, $P < 0.001$), and it ranged from $r = 0.62$ – 0.96 . There were also significant correlations between AR with CL ($r = 0.39$ – 0.42 , $P < 0.05$), BRT ($r = 0.54$ – 0.57 , $P < 0.01$), AST ($r = 0.47$ – 0.51 , $P < 0.01$), and BRT ($r = 0.56$ – 0.58 , $P < 0.01$) (Tables 3 and 4). The correlation coefficient for some of the phenotypic traits in St 524 had a similar trend as in St 504 for the tea taster's scores except for catechins and theaflavins. The individual catechins correlated significantly with other individual catechins and theaflavins. There was significant correlation between EGC and EGCG ($r = 0.51$, $P < 0.01$), EC and EGC ($r = 0.73$, $P < 0.01$), TF1 and EGC ($r = 0.62$, $P < 0.01$), TF2 and EC ($r = 0.44$, $P < 0.05$), TF1 and EGCG ($r = 0.42$, $P < 0.05$), TF2 and EGC ($r = 0.52$, $P < 0.01$), and TF4 and EGC ($r = 0.54$, $P < 0.01$) (Table 4). However, there was no correlation between % RWC with all biochemical compounds and tea taster scores in the two tea populations (Tables 3 and 4).

Genetic map construction

Regression-based integrated map (RG) resulted into 15 linkage groups (Fig. 2; Table 5). The integrated genetic linkage

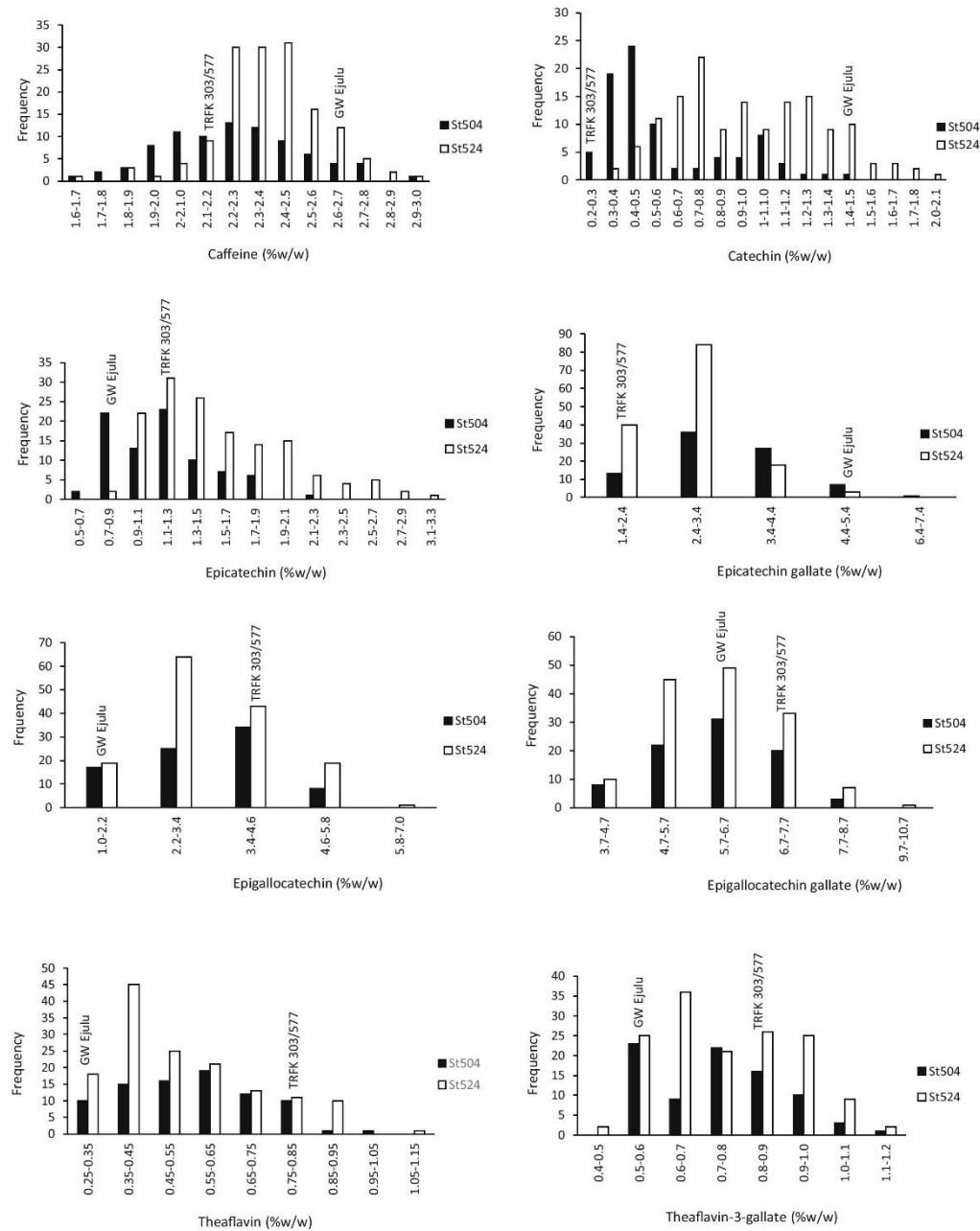


Fig. 1 Frequency distribution pattern of phenotypic traits of both green and black tea from two parental clones and F_1 population derived from a reciprocal cross between TRFK 303/577 and GW Ejulu (i.e. St 504 and St

524), respectively. The phenotypic values of the two parental clones are indicated in the graphs

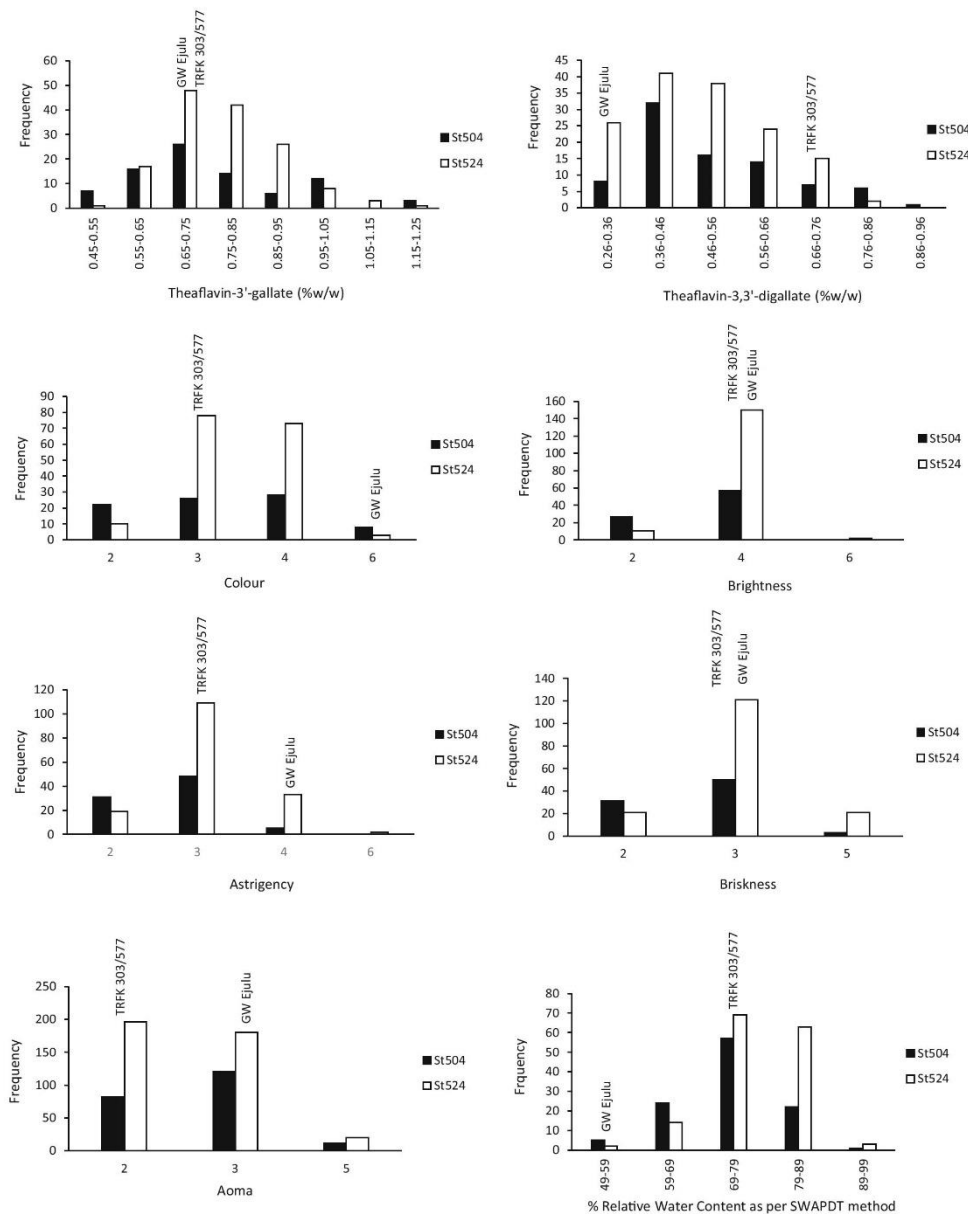


Fig. 1 (continued)

map covered a total genetic distance of 1260.1 cM, with an average locus distance of 1.1 cM. The marker density was not always uniform among linkage groups. The individual linkage groups ranged in size from 64.8 cM (LG02) to 160.1 cM

(LG09), and the number of markers on each group varied from 50 for LG01 to 219 for LG09 (Table 5). Linkage group 9 had the highest marker density with an average locus distance of 1.4 cM, while the other LGs had relatively lower marker

density (0.5–1.7 cM/DaT). However, DaT markers for all the phenotypic traits were absent on the LG03, LG05, and LG11. The constructed maps contained 15 linkage groups which corresponds to haploid chromosome number in *Camellia* spp. ($n = 15$) indicating saturation of the genome. The clustering of markers was observed, and some linkage groups were more densely saturated with markers as compared to others.

Phenotype segregation and QTL mapping

The analysis of 16 phenotypic traits revealed a total of 47 putative QTL in 15 LGs associated with tea quality and % RWC by IM and MQM at a significance genome-wide threshold of 5% level of confidence. Significant LOD scores ranged from 3.1–47.3 with an average of 6.5, while LOD thresholds of the detected QTL were between 2.9 and 4.6 with an average of 3.2 (Table 6). The interval lengths of the QTL ranged from 0.14 to 23.5 cM. QTL associated with individual catechins, caffeine, and individual theaflavins were mostly clustered in LG02 and LG04 but in different regions on the map (Fig. 2). Most QTL for tea taster scores (CL, BRT, AST, BRK, and AR) were detected at a lower significance level of confidence of 90% as compared to traits for biochemical molecules (Table 6). The QTL for each trait varied from a total of two (EC) to a total of seven (caffeine) in number. A total of nineteen QTL were identified for ten traits and were clustered in three main regions on LG01, LG02, LG04, LG12, LG13, and LG14 (Table 6). The variability explained by each QTL in the population was at average to high levels and ranged from 5.5 to 56.6%, with an average of 9.9%. The population variability explained (PVE) by all QTL for all the traits ranged between 6.5 (CL) to 97% (EGC).

Seven QTL for caffeine levels were detected and accounted for 40.5% of the total population explained variability (Table 6). Of these, one QTL, qCaffeine, was mapped on LG13 close to marker 5088162, with an LOD score of 4.7 and phenotypic variation explained (PVE) of 8.1% (Table 6). An additional five putative QTL were identified on LG02, LG04, LG07, LG08, and LG14 (Table 6). The qCaffeine LOD score ranged from 3.2 to 4.7, and the population variability explained ranged between 6.0 and 8.1%. Six QTL were detected for CAT content, accounting for a total of 91.1% of the total population variability. The two major QTL, qCAT, were located on LG04 and LG12 close to marker 5063001 and 5123751, respectively. The two QTL had the largest effect on CAT content, accounting for 55.6 and 10.3% of the population variability, respectively (Table 6). The QTL were distributed across the six LGs, with a LOD score ranging from 3.7 to 46.

Two QTL, qEC, were identified for EC content, which accounted for a total of 14.6% of the population variability (Table 6). The qEC was located on LG02 close to marker 5072338 with PVE of 7.0%, and on LG15 close to marker 5085963 with 7.6% of the population variability. Seven QTL, qECG, were detected for ECG content, which collectively explained 71.1% of the total population variability with LOD score range of 3.7 to 14.9 (Table 6). The QTL for ECG, qECG, was located on LG12 close to marker 5136790, had the largest effect on ECG content and accounted for of 23.1% of the population variability. A total of seven QTL were detected for EGC content, which accounted for a total of 97% the population variability (Table 6). Of these, two QTL, qEGC, were mapped on LG04 and LG12, respectively (Table 6). The qEGC on LG04 was located close to marker 5123475 with PVE of 56.6%, while the qEGC on LG12 was located close to marker 5123751 with PVE of 8.6% with LOD score of 47.3 and 14.9, respectively. An additional five putative QTL were identified on LG1, LG02, LG06, LG13, and LG14 (Table 6). The LOD score for the five loci varied from 3.2 to 4.2, and the PVE ranged between 5.5 and 7.2%. The QTL, qEGCG, was identified for EGCG content on LG04, close to marker 5087017, contributing to 24.1% of the population variability (Table 6). In addition, two putative QTLs were mapped on LG12 and LG15 and explained between 6.5 and 7.2% of the population variability (Table 6). The three QTL detected for EGCG content accounted for a total of 37.8% the population variability (Table 6). Two QTL for theaflavin, qTF1, were identified and it accounted for a total of 13.7% of the population variability (Table 6). The qTF1 was located on LG02 and LG06 close to markers 5084595 and 5136045, respectively. The nucleotide sequences for these markers are presented in supplementary Table 1. Only one QTL for qTF2 and another one for qBRT were identified on LG02 and LG14, close to marker 5084595 and 5125626, contributing to 7.0 and 6.8% of the population variability, respectively (Table 6).

The QTL for astringency, qAST, briskness, qBRK were mapped on two linkage groups (LG01 and LG09). The population variability for the tea taster score varied with aroma, qAR (6.4%) and briskness, qBRK (7.3%), respectively. Tea liquor color and tea brightness, had one QTL each while tea briskness, tea aroma had two QTL each and tea astringency had three QTL, respectively. Three QTL were detected for % RWC, which accounted for a total of 19.9% of the population variability (Table 6). The three stable QTL, qRWC, were mapped on LG02, LG06, and LG09, respectively. Three QTL were identified for % RWC. The first QTL (qRWC) was located on LG02, the second QTL was located on chromosome LG06, and the third QTL was located on LG09.

Table 3 Correlation coefficients (r) between phenotypic traits in the mapping population St 504

	CAFF	CAT	EC	ECG	EGC	EGCG	TF1	TF2	TF3	TF4	CL	BRT	AST	BRK	AR	RWC
CAFF	1															
CAT	-0.06	1														
EC	-0.07	-0.40	1													
ECG	0.44*	0.22	0.26	1												
EGC	-0.13	-0.78**	0.62**	-0.34	1											
EGCG	0.39	-0.57**	0.08	0.12	0.41*	1										
TF1	-0.24	-0.39	0.14	-0.33	0.42*	0.04	1									
TF2	-0.04	-0.36	0.10	-0.09	0.26	0.12	0.89***	1								
TF3	0.03	-0.08	-0.30	-0.01	-0.13	0.11	0.64**	0.83***	1							
TF4	-0.20	-0.28	-0.03	-0.25	0.23	0.08	0.90***	0.89***	0.80***	1						
CL	0.32	-0.08	0.20	0.26	0.12	0.17	-0.03	0.09	-0.13	-0.10	1					
BRT	0.24	-0.18	0.24	0.24	0.16	0.15	0.16	0.32	0.07	0.11	0.73**	1				
AST	0.20	-0.09	0.21	0.24	0.11	0.18	0.15	0.26	0.08	0.16	0.66**	0.83***	1			
BRK	0.22	-0.12	0.23	0.25	0.12	0.16	0.19	0.31	0.13	0.21	0.62**	0.80**	0.96***	1		
AR	0.18	-0.17	0.12	0.19	0.17	0.15	0.18	0.20	0.01	0.15	0.42*	0.54**	0.51**	0.58**	1	
RWC	0.12	0.19	0.16	-0.19	0.07	-0.03	0.06	0.06	0.02	0.06	-0.05	0.05	0.10	0.05	-0.06	1

CAFF- caffeine, CAT- catechin, EC- epicatechin, ECG- epicatechin gallate, EGC- epigallocatechin gallate, EGCG- epigallocatechin gallate, TF1- theaflavin, TF2- theaflavin-3-gallate, TF3- theaflavin-3-gallate, TF4- theaflavin-3,3'-digallate, CL- color, BRT- brightness, AST- astringency, BRK- briskness, AR- aroma, RWC- percent relative water content

*Correlation significant at $P \leq 0.05$ level

**Correlation significant at $P \leq 0.01$ level

***Correlation significant at $P \leq 0.001$ level

Table 4 Correlation coefficients (*r*) between phenotypic traits in the mapping population St 524

	CAFF	CAT	EC	ECG	EGC	EGCG	TF1	TF2	TF3	TF4	CL	BRT	AST	BRK	AR	RWC
CAFF	1															
CAT	0.12	1														
EC	0.07	-0.41*	1													
ECG	0.39	0.10	0.38	1												
EGC	-0.05	-0.74**	0.73**	-0.02	1											
EGCG	0.34	-0.60**	0.19	0.33	0.51**	1										
TF1	-0.01	-0.44*	0.51**	-0.01	0.62**	0.33	1									
TF2	0.13	-0.46*	0.44*	0.11	0.52**	0.42*	0.77**	1								
TF3	0.12	-0.16	-0.14	-0.03	0.02	0.31	0.44*	0.72**	1							
TF4	0.06	-0.51	0.34	-0.08	0.54**	0.42*	0.82**	0.91**	0.71**	1						
CL	0.11	-0.22	0.20	0.26	0.23	0.31	0.24	0.29	0.05	0.20	1					
BRT	0.09	-0.18	0.12	0.05	0.24	0.21	0.29	0.21	0.03	0.26	0.48*	1				
AST	0.09	-0.17	0.16	0.11	0.22	0.24	0.21	0.12	-0.01	0.17	0.57**	0.59**	1			
BRK	0.10	-0.17	0.14	0.10	0.21	0.22	0.19	0.17	0.01	0.25	0.47*	0.69**	0.76**	1		
AR	0.17	-0.15	0.21	0.10	0.26	0.16	0.22	0.21	-0.03	0.21	0.39*	0.57**	0.47*	0.56**	1	
RWC	0.09	0.07	0.06	-0.05	0.06	-0.03	0.12	0.08	0.06	0.12	-0.05	0.00	0.09	0.02	-0.02	1

CAFF- caffeine, CAT- catechin, EC- epicatechin, ECG- epicatechin gallate, EGC- epigallocatechin gallate, EGCG- epigallocatechin gallate, TF1- theaflavin, TF2- theaflavin-3-gallate, TF3- theaflavin-3'-gallate, TF4- theaflavin-3,3'-digallate, CL- color, BRT- brightness, AST- astringency, BRK- briskness, AR- aroma, RWC- percent relative water content

*Correlation significant at $P \leq 0.05$ level

**Correlation significant at $P \leq 0.01$ level

***Correlation significant at $P \leq 0.001$ level

These QTL spanned 60.7, 66.2, and 6.7 cM and have peak LOD scores of 4.3, 3.3, and 4.0, respectively (Table 6).

Discussion

Construction of a genetic linkage map is an important prerequisite for the identification of QTL(s) as well as mapping of agronomically important genes, which will be very useful for developing improved cultivars. Green tea catechins, black tea theaflavins, and caffeine are important biochemical traits that determine tea quality given that tea price at the auction is based on tea quality. The two principal gallated catechins present were EGCG and EGC which corroborate with results reported by (Karori et al. 2014). The two gallated catechins (EGCG and EGC) have been shown to contribute significantly to the formation of theaflavins. The B-ring in tri-hydroxylated catechins is oxidized high rate than the B-ring in di-hydroxylated catechins, which include CAT, EC, and EGC because of their lower oxidation ability (Owuor et al. 2006a). Therefore, tea cultivars with high levels of EGCG and EGC can be developed through tea breeding programs to improve black tea quality.

It was demonstrated previously that the ratio of di-hydroxyl flavan-3-ols to tri-hydroxyl flavan-3-ols influenced black tea quality in that, high proportion of simple catechins (CAT, EC

and ECG) to gallo catechins (EGC and EGCG), lead to a higher proportion of theaflavin (Ellis and Nyirenda 1995). The strong correlation of the di-gallate fractions in the present study is an important factor that can be utilized to enhance the quality of black tea. The changes in theaflavins and residual catechin compositions in relation to sensory characteristics of total CL, BRT, and BRK have also been reported in black tea (Obanda et al. 2001). Further, Obanda et al. (2001) reported that theaflavin digallate is the most astringent, estimated to be 6.4 times more astringent than simple theaflavin, and 2.88 times more astringent than either theaflavin-3-monogallate or theaflavin-3'-monogallate. The formation of a single theaflavin molecule requires a dihydroxy and a trihydroxy flavan-3-ol (Wright et al. 2002), as follows:

Epicatechin (EC) + Epigallocatechin (EGC) = simple theaflavin (TF1);
 Epicatechin (EC) + Epigallocatechin gallate (EGCg) = Theaflavin-3-gallate (TF2);
 Epicatechin gallate (ECG) + Epigallocatechin (EGC) = Theaflavin-3'-gallate (TF3);
 Epicatechin gallate (ECG) + Epigallocatechin gallate (EGCg) = Theaflavin-3, 3'-digallate (TF4)

The relationship between catechins and TFs content with the quality of black tea has been investigated with reports

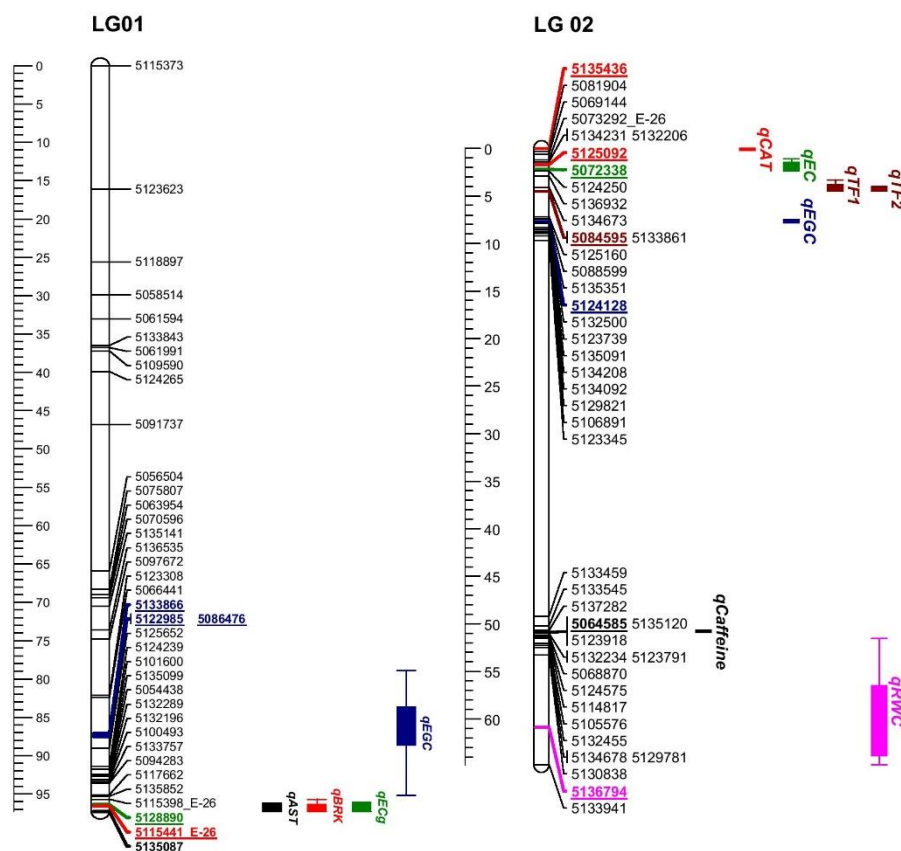


Fig. 2 DArT-based genetic map of tea plant showing location of QTLs for caffeine, catechins, theaflavins content, liquor brightness, and % rwc identified in the two mapping population. Map distance scales in cM are placed at left margin. Significant QTL for caffeine, individual catechin, and individual theaflavin content and liquor brightness is represented by different bars. Bars and lines indicate 1-LOD and 2-LOD support

intervals. qCaffeine- Caffeine; qCAT- Catechin; qEC- Epicatechin; qECG- Epicatechin gallate; qEGC- Epigallocatechin gallate; qEGCG- Epigallocatechin gallate; qTF1- Theaflavin; qTF2- Theaflavin-3-gallate; qCL- Color; qBRT- Brightness; qAST- Astringency; qBRK- Briskness; qAR- Aroma; qRWC- percent relative water content

showing a significant correlation between sensory parameters with catechins and TFs content. Among all the sensory, parameters investigated are color, brightness, astringency, briskness, and aroma; the results in this present study indicate that neither caffeine, individual catechins, nor individual TFs content had any correlation with sensory parameters. This could be due to loss in aromatic compounds between storage, and the time of organoleptic evaluation was done which led to flat tea (Sedaghatthoor et al. 2013). There was a significant correlation between the individual sensory parameters (CL, BRT, AST, BRK and AR), these results are similar to those reported by Owuor et al. (2006b) but there was no significant correlation between

simple (non-gallated) theaflavin with sensory parameters except for tea liquor brightness. The significant correlations show that the phenotypic traits are either controlled by linked genes or genes that have pleiotropic effects on the traits. Hence, improving one phenotypic trait may lead to improvement of other phenotypic traits.

The map spanned a total length of 1260.1 cM with 1421 markers, which is close to tea genome obtained by Ota and Tanaka (1999) (1640 cM), Hackett et al. (2000) (1349.7 cM), and Taniguchi et al. (2012) (1298 and 1305 cM), Ma et al. (2014) (1143.5 cM), but smaller in length than those reported by Huang et al. (2006) (2457.7 and 2545.3 cM), and Hu et al. (2013) (4482.9 cM). The differences in the total map length

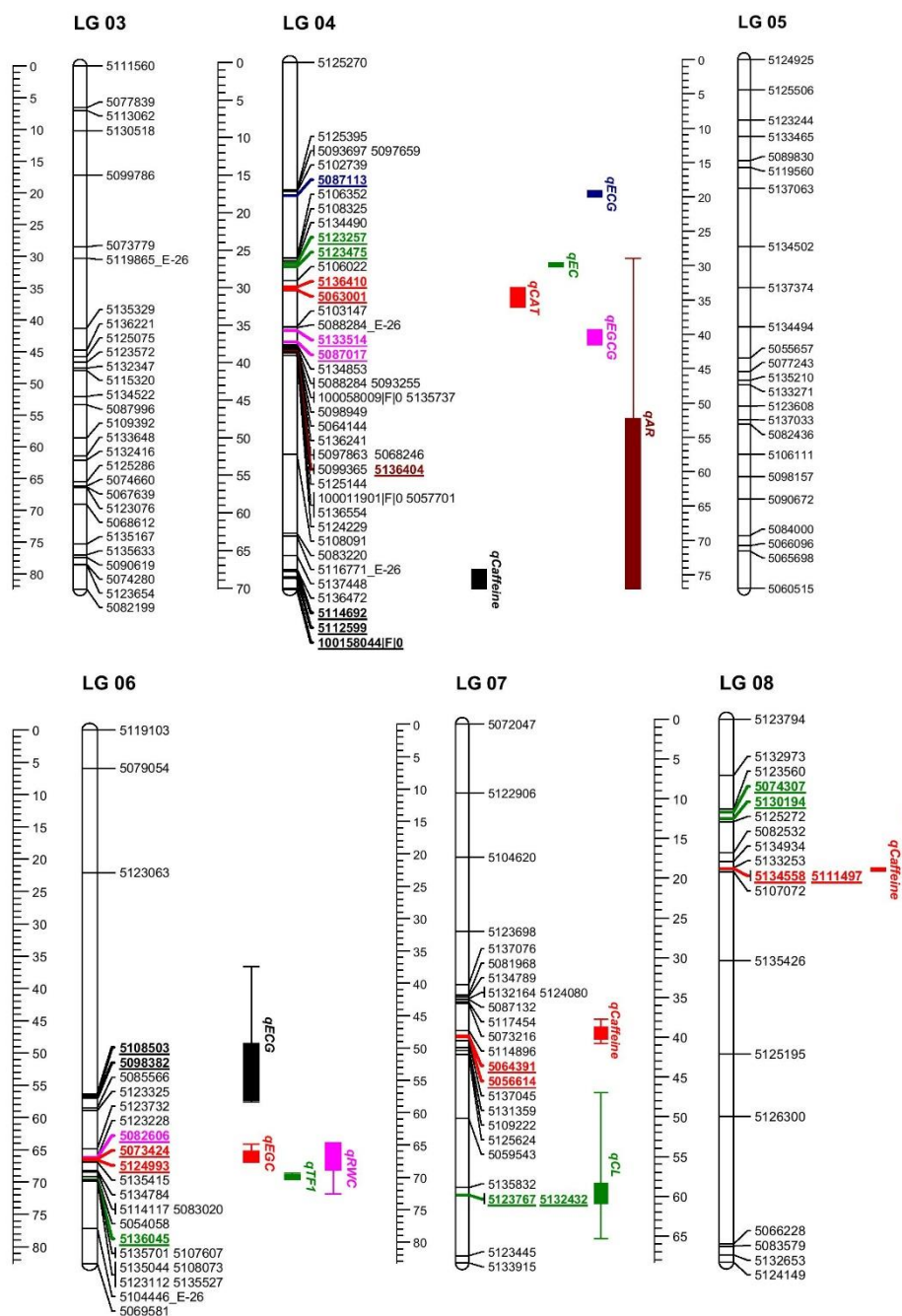


Fig. 2 (continued)

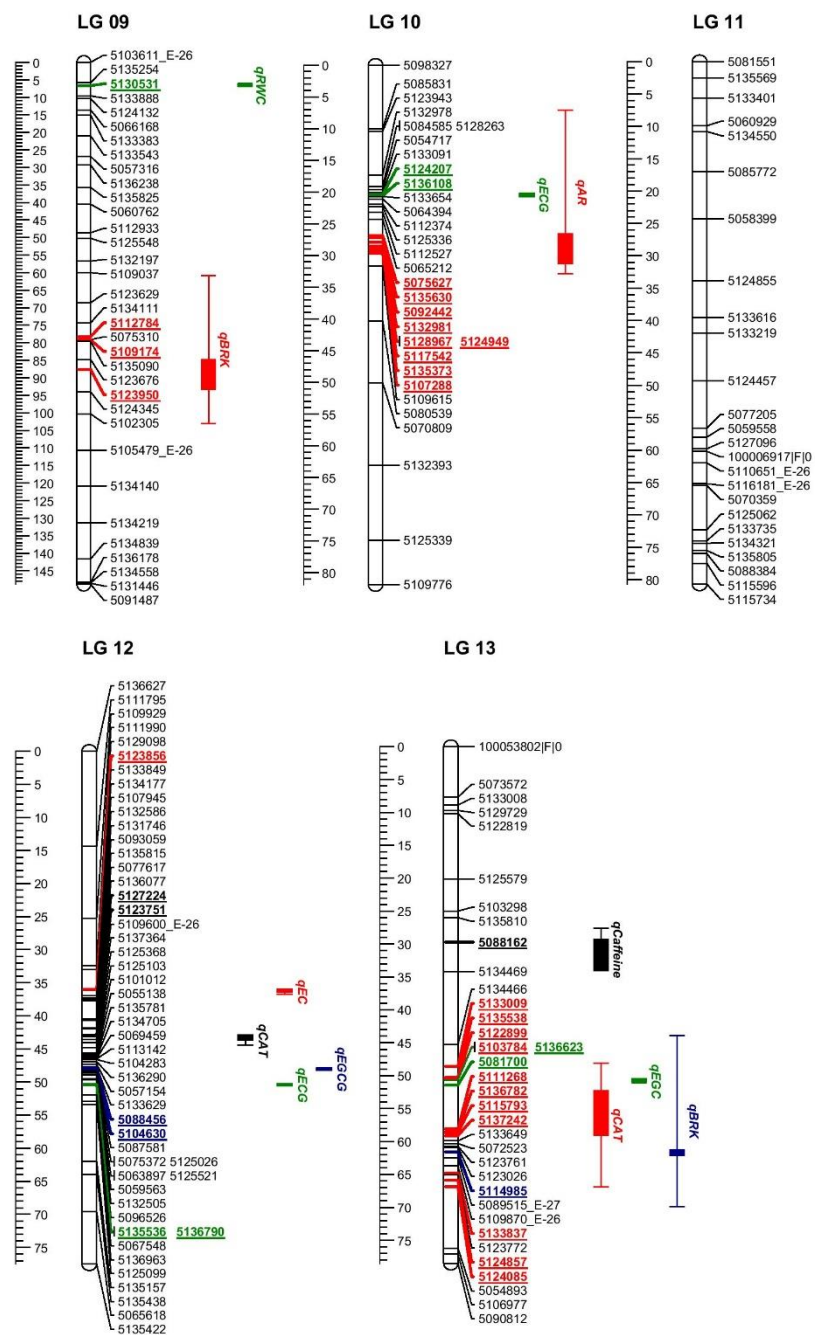


Fig. 2 (continued)

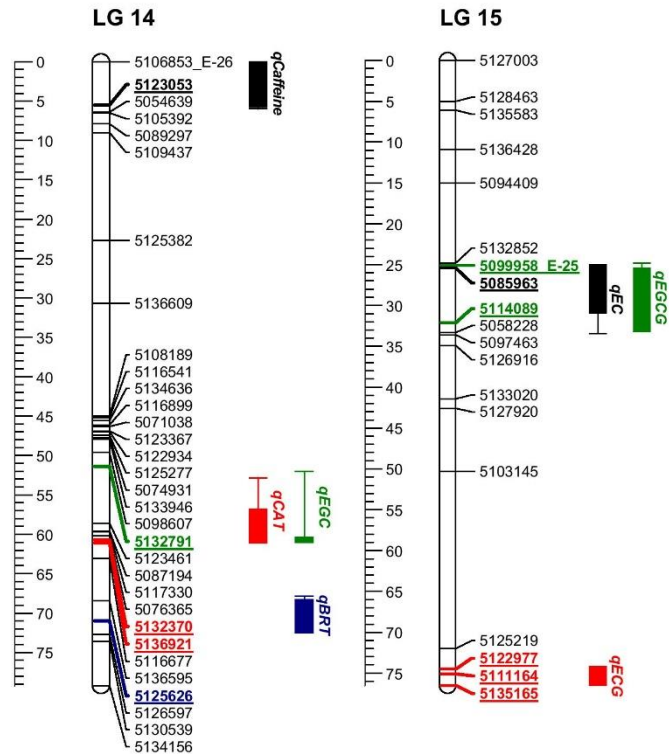


Fig. 2 (continued)

can result from variation in the number of recombination events in the two parents as well as variations in the numbers and locations of mapped loci. There were 15 linkage groups found in this study, indicating saturation of the genome ($n = 15$). However, there was gap greater than 20 cM between the adjacent markers on LG06 and LG15. Even though the

Table 5 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 6 QTLs for caffeine, individual catechins and theaflavins content, tea taster's scores and percent relative water content detected in St 504 and St 524 tea plant population

Traits	QTL	LG	Position (cM)	LOD threshold	LOD score	PVE (%)	Nearest marker
Caffeine	qCaffeine	2	50.905 (50.247–51.548)	3.3	3.4	6.0	5064585
		4	68.574 (63.137–70.142)	3.2	3.8	6.7	5112599
		7	48.071 (34.695–50.409)	3.2	3.7	6.6	5064391
		8	18.762 (16.783–26.008)	3.3	3.9	6.7	5134558
		13	29.748 (23.653–36.538)	3.1	4.7	8.1	5088162
CAT	qCAT	14	5.465 (0.000–5.465)	3.3	3.7	6.4	5123053
		2	0.000 (0.000–2.421)	3.2	3.7	6.4	5135436
		4	30.322 (29.886–30.322)	3.3	46.0	55.6	5063001
		8	12.506 (12.506–12.882)	3.3	3.3	5.6	5130194
		12	42.998 (36.019–54.085)	3.1	6.1	10.3	5123751
EC	qEC	13	50.602 (48.559–58.674)	3.0	4.1	6.9	5111268
		14	60.713 (47.731–61.117)	3.2	4.1	6.9	5132370
		2	2.242 (0.000–8.309)	3.3	4.1	7.0	5072338
ECG	qECG	15	25.381 (20.639–25.381)	3.3	4.5	7.6	5085963
		1	96.36 (96.36–97.284)	3.0	6.8	11.7	5128890
EGC	qEGC	4	17.658 (13.976–37.226)	3.4	4.7	8.0	5087113
		6	56.948 (56.504–56.948)	3.1	4.0	6.7	5098382
		10	20.631 (20.090–25.539)	3.3	4.8	8.0	5136108
		12	50.441 (50.148–50.441)	2.9	14.9	23.1	5136790
		13	50.602 (48.559–58.674)	3.1	3.7	6.4	5088162
EGCG	qEGCG	15	75.067 (75.067–76.498)	3.3	4.2	7.2	5111164
		1	87.059 (84.375–88.562)	3.0	3.3	5.8	5133866
		2	7.719 (7.719–7.855)	3.2	3.2	5.5	5124128
		4	27.194 (27.194–28.879)	3.3	47.3	56.6	5123475
		6	66.351 (56.504–72.354)	3.1	4.2	7.2	5073424
TF1	qTF1	12	42.998 (42.839–54.085)	3.0	14.9	8.6	5123751
		13	50.602 (48.559–58.674)	3.1	3.7	6.4	5136623
		14	60.713 (60.713–63.050)	3.1	4.0	6.9	5132791
		4	37.226 (28.879–38.559)	3.4	15.6	24.1	5087017
		12	48.179 (36.963–49.744)	3.0	4.2	7.2	5104630
TF2	qTF2	15	32.140 (17.119–33.578)	3.3	3.8	6.5	5114089
		2	4.455 (4.455–5.391)	3.1	3.2	5.8	5084595
CL	qCL	6	69.564 (69.564–75.112)	4.1	4.5	7.9	5136045
BRT	qBRT	2	4.455 (1.220–5.391)	3.1	4.0	7.0	5084595
AST	qAST	7	72.684 (72.677–72.684)	3.0	3.2	6.5	5132432
BRK	qBRK	14	71.046 (65.383–71.046)	3.4	3.3	6.8	5125626
AR	qAR	1	96.577 (88.83–97.284)	3.0	4.5	9.1	5135087
		9	87.614 (87.614)	3.2	3.2	6.6	5123950
% RWC	qRWC	1	96.577 (89.022–97.284)	3.0	3.6	7.3	5115441
		9	87.614 (87.614)	3.6	3.6	7.4	5123950
Average		13	61.582 (61.582)	3.0	3.1	6.3	5114985
		4	68.574 (68.574)	3.0	3.1	6.4	5112599
		10	28.656 (26.854–29.038)	3.2	3.5	7.0	5128967
		2	60.868 (56.912–60.868)	3.3	4.3	7.3	5136794
		6	66.217 (66.052–66.217)	3.2	3.3	5.7	5082606
		9	6.684 (6.167–6.684)	3.9	4.0	6.9	5130531
				3.2	6.5	9.9	

CAT catechin, EC epicatechin, ECG epicatechin gallate, EGC epigallocatechin, EGCG epigallocatechin gallate, TF1 theaflavin, TF2 theaflavin-3-gallate, CL color, BRT brightness, AST astringency, BRK briskness, % RWC percent relative water content, LOD- logarithm of odds ratio, PVE- phenotypic variation explained

The genome-wide LOD significance thresholds ($P < 0.05$) based on permutation testing ($n = 1000$)

combination of *PstI* (CTGCAG) and *MseI* (CCGG) restriction enzymes performed best with 16,382 DArTseq achieved, there was still one gap of larger than 20 cM between adjacent markers on LG06 and LG15. The presence of a large gap may be a result of genome regions corresponding to the gap regions of the genetic map or homozygous could be a result of gaps in

both mapping parents which leads no recombination events. Therefore, further studies need to be carried out to fill in the gaps in the current genetic map. However, the mapped markers in the current study were distributed in all 15 LGs with marker density ranging from 0.5 to 1.7 cM. The average interval between two markers was 1.1 cM, which is an

improvement to previous genetic maps which ranged from 1.9 to 19.0 cM with the latest by Ma et al. (Ma et al. 2014), spacing 2.9 cM. The marker density for genome-wide QTL mapping which is recommended should be less than 10 cM (Doerge 2002). The genetic map constructed in the present study is thus suitable for identification of QTL.

In total, six caffeine QTL, 25 catechins QTL, three theaflavins QTL, nine QTL for tea taster score, and three QTL for percent % RWC were detected, and the population variability explained by each QTL varied from 5.5 to 56.6%, with an average of 9.9%. The high PVE exhibited by these QTL suggests that caffeine, catechins, theaflavins content, tea taster scores and % RWC may be controlled by few critical genes. However, the small size of 261 mapping population used in this study may lead to overestimation of QTL although the population size is slightly smaller than from the previous study of 300 individuals by Ma et al. (2014). The individual catechins (ECG, EGC, and EGCG) were mostly detected in LG01, LG04, LG12, and LG13. In addition to QTL for catechins, the present study incorporated QTL for caffeine, black tea theaflavins, and tea taster's scores, which are major determinants of tea quality. The major QTL, qCAT, qEC, qEGC, qTF1, and qTF2 detected showed a tendency to co-localize and cluster on LG02 while QTL, qCAT, qECG, qEGC, and qEGCG clustered in LG12. The QTL were particularly prominent in the region between 0.00 and 7.72 cM of LG02 and 42.9 to 50.4 cM of LG12, where more than four QTL are located, for each trait.

Linkage group one, apart from the QTL for individual catechins content, QTL for individual theaflavins, qTF1, and qTF2 were also detected. These regions of the two chromosomes probably contain genes that are multifunctional which are associated with production and accumulation of catechins and theaflavins and deserve further investigation. More interestingly, QTL associated with caffeine, qCaffeine, and liquor brightness, qBRT, were not clustered with and/or located within the same region on the chromosome with the QTL for catechins and theaflavins in the linkage groups. It is an indication that the genes associated with caffeine production and accumulation as well as liquor brightness are located in different regions on the chromosome from those of catechins and theaflavins. The few loci that were associated with tea catechins and theaflavins will need to be confirmed for MAS. For example, markers 5123475 and 5084595 that were consistently associated with individual catechins (EC and EGC) and individual theaflavins (TF1 and TF2) indicated their potential as candidate markers for MAS.

Conclusion

DARtSeq markers are capable of detecting even very small polymorphisms, are cost-effective, and are efficient for whole-genome scans and population structure. The main objective of tea breeding is to improve the quality and quantity of

the product. The investigation of biochemical traits such as caffeine, catechins, and theaflavins content, which are related to tea quality, presents an alternative approach in tea breeding. The SWAPDT method can also be used since it represents an easy method to measure water deficit in drought stress plants. The use of DARtSeq markers in the present study will be useful for tea breeding and will provide valuable information for selecting parents that would give rise to progenies with the desired genotype. Although, in the present study, linkage maps might be enhanced by further saturation, they constitute an additional tool that can be used to locate QTL linked with some important agronomic traits. These will contribute to the development of MAS for tea quality improvement, considering the favorable alleles of the markers involved in the QTL for tea quality. The validated molecular markers will contribute greatly to adoption of marker-assisted selection (MAS) for drought tolerance and tea quality improvement.

In addition, with recent availability of a complete genome sequence of the tea plant, this will allow identification genes more precisely in relation to the QTL detected in this study.

Acknowledgements The authors acknowledge the financial support to conduct this research, and study grants for RK and PM 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).

Author contributions ZA, SK, and RM were involved with the design of the experiment and plant material used. RK, PM, and CN were involved in the collection of plant material. RK performed the experiments. RK, PM, CN, SK, and ZA analyzed samples and interpreted the data. RK wrote the manuscript and revised by PM, CM, RM, SK, and ZA. The final manuscript was reviewed and approved by all the authors.

Compliance with ethical standards

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

Data archiving statement The DARt sequences have been submitted to NCBI (<http://www.ncbi.nlm.nih.gov/>). BioProject PRJNA398959, Supplementary Table 2.

References

- Adkins NL, Hall JA, Georgel PT (2007) The use of quantitative agarose gel electrophoresis for rapid analysis of the integrity of protein-DNA complexes. *J Biochem Biophys Methods* 70(5):721–726. <https://doi.org/10.1016/j.jbbm.2007.03.006>
- AFFA (2013) Tea market development in Kenya. Nairobi: Agriculture, Fisheries and Food Authority.

- Ananingsih VK, Sharma A, Zhou W (2013) Green tea catechins during food processing and storage: a review on stability and detection. *Food Res Int* 50(2):469–479. <https://doi.org/10.1016/j.foodres.2011.03.004>
- Anon (2002) Tea growers handbook, 5th edn. Tea Research Foundation of Kenya, Kericho
- Bali S, Mamgain A, Raina SN, Yadava SK, Bhat V, Das S, Pradhan AK, Goel S (2015) Construction of a genetic linkage map and mapping of drought tolerance trait in Indian beverage tea. *Mol Breed* 35:1–20
- Banerjee B (1992) Botanical classification of tea. In: Willson KC, Clifford MN (eds) *Tea: cultivation to consumption*. Chapman and Hall, London, pp 25–51. https://doi.org/10.1007/978-94-011-2326-6_2
- Chen L, Zhou ZX, Yang YJ (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. <https://doi.org/10.1007/s10681-006-9292-3>
- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138(3):963–971
- da Silva Pinto M (2013) Tea: a new perspective on health benefits. *Food Res Int* 53:558–567
- Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* 3(1):43–52. <https://doi.org/10.1038/nrg703>
- Ellis R, Nyirenda H (1995) A successful plant improvement programme on tea (*Camellia sinensis*). *J Exp Agric* 31(03):307–323. <https://doi.org/10.1017/S0014479700025485>
- FAO (2015) Kenya's tea sector under climate change: An impact assessment and formulation of a climate smart strategy, by Elbehri A., B. Cheserek, A. Azapagic, D. Raes, M. Mwale, J. Nyengena, P. Kiprono, and C. Ambasa. Rome, Italy
- Gawal N, Jarret R (1991) A modified CTAB DNA extraction procedure for Musa and Ipomea. *Plant Mol Biol Rep* 9(3):262–266. <https://doi.org/10.1007/BF02672076>
- Gupta PK, Rustgi S, Mir RR (2013) Array-based high-throughput DNA markers and genotyping platforms for cereal genetics and genomics. In: Gupta PK and Varshney RK (eds) *Cereal genomics II*. Springer, Netherlands, p 11–55
- Hackett CA, Wachira FN, Paul S, Powell W, Waugh R (2000) Construction of a genetic linkage map for *Camellia sinensis* (tea). *Heredity* 85(4):346–355. <https://doi.org/10.1046/j.1365-2540.2000.00769.x>
- Hu CY, Lee TC, Tsai HT, Tsai YZ, Lin SF (2013) Construction of an integrated genetic map based on maternal and paternal lineages of tea (*Camellia sinensis*). *Euphytica* 191(1):141–152. <https://doi.org/10.1007/s10681-013-0908-0>
- Huang F, Liang Y, Lu J, Chen R (2006) Genetic mapping of first generation of backcross in tea by RAPD and ISSR markers. *J Tea Sci* 26: 171–176
- ISO14502-2 (2005) Determination of substances characteristic of green and black tea-part 2: content of catechins in green tea- method using high-performance liquid chromatography. International Organization for Standardization, Geneva
- Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res* 29(4):e25. <https://doi.org/10.1093/nar/29.4.e25>
- 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. *Afr J Plant Sci* 3:093–101
- 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 Genet Genomes* 6:915–929
- Karori S, Wachira F, Ngure R, Mireji P (2014) Polyphenolic composition and antioxidant activity of Kenyan tea cultivars. *J Pharmacogn Phytochem* 3:105–116
- Kenya National Bureau of Statistics (2012) Kenya facts and Figs 2012. Kenya
- Khan N, Mukhtar H (2007) Tea polyphenols for health promotion. *Life Sci* 81(7):519–533. <https://doi.org/10.1016/j.lfs.2007.06.011>
- 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. *Biosci Biotechnol Biochem* 69(11):2219–2223. <https://doi.org/10.1271/bbb.69.2219>
- Lin SD, Mau JL, Hsu CA (2012) Bioactive components and antioxidant properties of γ -aminobutyric acid (GABA) tea leaves. *LWT-Food Sci Technol* 46(1):64–70. <https://doi.org/10.1016/j.lwt.2011.10.025>
- Liu X, Tang L, Wu H, Xi W, Yu J, Zhou Z (2016) Development of DArT markers and evaluation of phylogenetic relationship of key citrus species. *Genet Resour Crop Evol* 63(8):1307–1318. <https://doi.org/10.1007/s10722-015-0319-2>
- Ma JQ, Yao MZ, Ma CL, Wang XC, Jin JQ, Wang XM, Chen L (2014) Construction of a SSR-based genetic map and identification of QTL for catechins content in tea plant (*Camellia sinensis*). *PLoS One* 9(3):e93131. <https://doi.org/10.1371/journal.pone.0093131>
- Ma JQ, Huang L, Ma CL, Jin JQ, Li CF, Wang RK, Zheng HK, Yao MZ, Chen L (2015) Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). *PLoS One* 10(6): e0128798. <https://doi.org/10.1371/journal.pone.0128798>
- Malebe MP (2011) Method for screening plants for drought tolerance, Vol. Msc, University of Pretoria, Pretoria, South Africa
- Mecgahakumbura MK, Wambulwa MC, Thapa KK, Li MM, Möller M, Xu JC, Yang JB, Liu BY, Ranjitkar S, Liu J, Li DZ (2016) Indications for three independent domestication events for the tea plant (*Camellia sinensis* (L.) O. Kuntze) and new insights into the origin of tea germplasm in China and India revealed by nuclear microsatellites. *PLoS One* 11(5):e0155369. <https://doi.org/10.1371/journal.pone.0155369>
- Mondal TK, Bhattacharya A, Laxmikumar M, Ahuja PS (2004) Recent advances of tea (*Camellia sinensis*) biotechnology. *Plant Cell Tissue Organ Cult* 76(3):195–254. <https://doi.org/10.1023/B:TICU.0000009254.87882.71>
- Nitin SL, Islam M, O'Hare WT, Ali Z (2006) Discrimination of teas based on total luminescence spectroscopy and pattern recognition. *J Sci Food Agr* 86(13):2092–2098. <https://doi.org/10.1002/jsfa.2578>
- 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. *J Plant Physiol* 198:39–48. <https://doi.org/10.1016/j.jplph.2016.04.004>
- Obanda M, Owuor PO, Taylor SJ (1997) Flavanol composition and caffeine content of green leaf as quality potential indicators of Kenyan black teas. *J Sci Food Agr* 74(2):209–215. [https://doi.org/10.1002/\(SICI\)1097-0010\(199706\)74:2<209::AID-JSFA789>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0010(199706)74:2<209::AID-JSFA789>3.0.CO;2-4)
- Obanda M, Owuor PO, Mang'oka R (2001) Changes in the chemical and sensory quality parameters of black tea due to variations of fermentation time and temperature. *Food Chem* 75(4):395–404. [https://doi.org/10.1016/S0308-8146\(01\)00223-0](https://doi.org/10.1016/S0308-8146(01)00223-0)
- Orel G, Wilson PG (2012) *Camellia cheryana* (Theaceae), a new species from China. *Ann Bot Fenn* 49(4):248–254. <https://doi.org/10.5735/085.049.0405>
- Ota S, Tanaka J (1999) RAPD-based linkage mapping using F1 segregating populations derived from crossings between tea cultivar 'Sayamakaori' and Strain 'ana-Ck17'. *Breed Res* 1:16
- Owuor PO, Othieno CO (1991) Response of black tea quality to locality in stakeholder tea farms. *Tea* 12:41–45
- Owuor P, Obanda M, Apostolides Z, Wright L, Nyirenda H, Mphangwe N (2006a) The relationship between the chemical plain black tea

- quality parameters and black tea colour, brightness and sensory evaluation. *Food Chem* 97(4):644–653. <https://doi.org/10.1016/j.foodchem.2005.04.027>
- Owuor PO, Obanda M, Nyirenda HE, Mphangwe NI, Wright LP, Apostolides Z (2006b) 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 Chem* 97(4):644–653. <https://doi.org/10.1016/j.foodchem.2005.04.027>
- Pang Y, Abeysinghe IBS, He J, He X, Huhman D, Mewan KM, Sumner LW, Yun J, Dixon RA (2013) Functional characterization of proanthocyanidin pathway enzymes from tea and their application for metabolic engineering. *Plant Physiol* 161(3):1103–1116. <https://doi.org/10.1104/pp.112.212050>
- 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. *Plant Genome* 5(3):103–113. <https://doi.org/10.3835/plantgenome2012.06.0006>
- Preedy VR (2012) Tea in health and disease prevention. Academic Press, Cambridge
- Racedo J, Gutiérrez L, Perera MF, Ostengo S, Pardo EM, Cuenya MI, Welin B, Castagnaro AP (2016) Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biol* 16(1):142. <https://doi.org/10.1186/s12870-016-0829-x>
- Sall J, Lehman A, Stephens ML, Creighton L (2012) JMP start statistics: a guide to statistics and data analysis using JMP, 6th edn. SAS Institute, Inc., Cary
- Sánchez-Sevilla JF, Horvath A, Botella MA, Gaston A, Folta K, Kilian A, Denoyes B, Amaya I (2015) Diversity arrays technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria × ananassa*). *PLoS One* 10(12):e0144960. <https://doi.org/10.1371/journal.pone.0144960>
- Sang S, Lambert JD, Ho CT, Yang CS (2011) The chemistry and biotransformation of tea constituents. *Pharmacol Res* 64(2):87–99. <https://doi.org/10.1016/j.phrs.2011.02.007>
- Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A (2010) A high-density diversity arrays technology (DArT) microarray for genome-wide genotyping in eucalyptus. *Plant Methods* 6:1
- Schouten HJ, van de Weg WE, Carling J, Khan SA, McKay SJ, van Kaauwen MP, Wittenberg AH, Koehorst-van Putten HJ, Noordijk Y, Gao Z (2012) Diversity arrays technology (DArT) markers in apple for genetic linkage maps. *Mol Breed* 29(3):645–660. <https://doi.org/10.1007/s11032-011-9579-5>
- Sedaghatpour S, Haghighat SR, Shokrgozar SA (2013) Storage period effects on the qualitative characteristics of scented tea. *Int J Biosci* 3(7):66–73
- Sharangi, A (2009) Medicinal and therapeutic potentialities of tea (*Camellia sinensis* L.) - A review. *Food Res Int* 42 (5-6): 529–535
- Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics* 12(1):131. <https://doi.org/10.1186/1471-2164-12-131>
- Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, Myburg AA, Grattapaglia D, Vaillancourt RE (2011) Population genetic analysis and phylogeny reconstruction in eucalyptus (Myrtaceae) using high-throughput, genome-wide genotyping. *Mol Phylogenet Evol* 59(1):206–224. <https://doi.org/10.1016/j.ympev.2011.02.003>
- Tan LQ, Wang LY, Xu LY, Wu LY, Peng M, Zhang CC, Wei K, Bai PX, Li HL, Cheng H, Qi GN (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 Genet Genomes* 12(3):1–13
- Tanaka J (2000) Construction of linkage and QTL analysis of tea plant. *Proceedings of the International Symposium on Molecular Biology and Tea Breeding*, pp 23–26
- 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*). *Breed Sci* 62(3):263–273. <https://doi.org/10.1270/jsbbs.62.263>
- Van Ooijen J (2006) JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B.V., Wageningen
- Voorrips R (2002) MapChart: software for the graphical presentation of linkage maps and QTL. *J Hered* 93(1):77–78. <https://doi.org/10.1093/jhered/93.1.77>
- Wachira FN, Waugh R, Powell W, Hackett C (1995) Detection of genetic diversity in tea (*Camellia sinensis*) using RAPD markers. *Genome* 38(2):201–210. <https://doi.org/10.1139/g95-025>
- Wambulwa MC, Meegahakumbura MK, Kamunya S, Muchugi A, Möller M, Liu J, Xu JC, Ranjitkar S, Li DZ, Gao LM (2016) Insights into the genetic relationships and breeding patterns of the African tea germplasm based on nSSR markers and cpDNA sequences. *Front Plant Sci* 7. <https://doi.org/10.3389/fpls.2016.01244>
- Wittenberg AH, van der Lee T, Cayla C, Kilian A, Visser RG, Schouten HJ (2005) Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. *Mol Gen Genomics* 274(1):30–39. <https://doi.org/10.1007/s00438-005-1145-6>
- Wright LP, Mphangwe NIK, Nyirenda HE, 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. *J Sci Food Agr* 82(5):517–525. <https://doi.org/10.1002/jsfa.1074>
- Zou J, Semagn K, Iqbal M, N'Diaye A, Chen H, Asif M, Navabi A, Perez-Lara E, Pozniak C, Yang RC (2016) Mapping QTL controlling agronomic traits in the 'Attila' × CDC Go'Spring wheat population under organic management using 90K SNP Array. *Crop Sci* 57(1):365–377

CHAPTER 4

FUNCTIONAL ANNOTATION OF PUTATIVE QTLs ASSOCIATED WITH BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS

Abstract

The understanding of black tea quality and %RWC traits in tea by a QTL approach can be useful in the elucidation of the interaction of such complex traits on a molecular basis and in the identification of candidate genes underlying the QTL which is still difficult to do. Nevertheless, the recently published draft tea (*Camellia sinensis*) genome should be able to bridge the gap between the QTL and linkage maps and physical location of candidate genes on the tea genome. The objective of this study was to identify putative QTLs controlling black tea quality and %RWC traits in two tea populations and their F₁ progeny, as well as to identify candidate genes underlying QTLs based on the anchoring of linkage maps on the tea genome and GO information linked to genome annotation. A total of 1 421 DArTseq markers derived from linkage maps identified 53 DArTseq markers linked to black tea quality and %RWC. Of these, 22 DArTseq markers were identified on the tea genome. A total of 5 592 unigenes were assigned GO terms, of these, 56% comprised biological processes, 29% cellular component and 15% molecular functions. A total of 84 unigenes in 15 LGs were assigned to 79 different KEGG database pathways based on categories of secondary metabolite biosynthesis. The three major enzymes identified were transferase (38.9%), hydrolase (29%) and oxidoreductase (18.3%). The 28% of the mapped sequences (405 out of 1 471) were annotated. The putative candidate proteins identified in KEGG database pathways were involved in flavonoid biosynthesis, alkaloid biosynthesis, ATPase family proteins or proteins related to abiotic/ biotic stress response.

4.1 Introduction

The identification of putative candidate genes underlying QTL in plants is often complicated and time-consuming. It is even further complicated by genotype-environment interactions which influence the growth and development of plants. However, using recently unveiled tea genome sequence, it was possible to annotate candidate genes and enzymes involved in the biosynthesis of phenolic compounds and drought tolerance. Phenolic compounds are closely related to the typical flavour of tea infusions and the health benefits ascribed to tea consumption (Preedy, 2012; Wang *et al.*, 2014b). Among these polyphenols, catechins (flavan-3-ols), flavonols, and their derivatives (galloylated catechins and flavonol 3-O-glycosides) are major components with pivotal bioactivities in tea. These compounds have been shown to be the major contributors to the astringent sensation of black tea infusions (Scharbert and Hofmann, 2005). In particular, galloylated catechins have been found to confer astringent and bitter tastes (Rossetti *et al.*, 2009). Flavonol 3-O-glycosides have been found to induce velvety, mouth-drying, and mouth-coating sensations (Cui *et al.*, 2016). Caffeine is important in the cream formation and the briskness of tea and therefore has been used in predicting black tea quality (Obanda *et al.*, 1997; Liang *et al.*, 2002). Polyphenols and caffeine biosynthesis are dependent on structural and regulatory genes; structural genes encode enzymes catalysing the biosynthesis, while regulatory genes control the expression of the genes.

Drought is a major factor which significantly influences the growth and production of tea (Ng'etich *et al.*, 2001; Cheruiyot *et al.*, 2008a). Hence, breeding drought-tolerant tea cultivars that have high recoverability potential is an important target for tea breeders. Although the morphological and physiological mechanisms of drought tolerance in tea plant have been reported, only a few reports are available on the changes of genes related to catechin biosynthesis in response to environmental stresses (Liu *et al.*, 2016a; Tony *et al.*, 2016). This is because drought tolerance is a complex multigenic trait and various regulatory and functional genes are involved. In tea plant, Cheruiyot *et al.* (2007) reported that polyphenols could be used as a potential indicator for drought tolerance. In another study, results suggested epicatechin and epigallocatechin as potential indicators in predicting drought tolerance in tea plant (Cheruiyot *et al.*, 2008b). Recently, transcriptomic analysis of drought susceptible and drought-tolerant tea plants during drought stress has been reported (Maritim *et al.*, 2015; Liu *et al.*, 2016a; Zheng *et al.*, 2016). The molecular responses of plants to drought stress include perception, signal transduction, gene expression, and ultimately

metabolic changes that lead to stress-tolerance. The induced drought stress lead to gene expression which participates in the generation of hormones like abscisic acid (ABA), salicylic acid, and ethylene. For example, ABA has been shown to play crucial roles in regulating the drought response in many plants, and its metabolic pathway involves multiple steps and genes (Zhou *et al.*, 2014). The level of ABA in plants increases rapidly during drought stress, leading to stomatal closure and induction of stress genes to cope with the stress (Sah *et al.*, 2016). Drought stress also results in accumulation of proline, mannitol, sorbitol, formation of radical scavenging compounds (ascorbate, glutathione, tocopherol), and synthesis of protein (Upadhyaya and Panda, 2013; Upadhyaya *et al.*, 2016). Proline acts as an osmoprotectant, and its accumulation is induced as a result of stress due to water deficit. This helps in maintaining water relations, prevents membrane distortion, and also acts as a hydroxyl radical scavenger (Upadhyaya *et al.*, 2016). The increase in generation of reactive oxygen species (ROS) such as superoxide radicals (O_2^-), hydrogen peroxide (H_2O_2), hydroxyl radical ($\cdot OH$), alkoxyl radical ($RO\cdot$), and singlet oxygen ($^1O_2$) within the cell, leads to lipid peroxidation, protein degradation, and enzyme inactivation and affects nucleic acid leading to cell death (Upadhyaya and Panda, 2013; Upadhyaya *et al.*, 2016).

4.2 Research objectives

The objective of this study was to annotate functions and assign biosynthetic pathways to putative QTLs linked to caffeine biosynthesis, catechins biosynthesis, theaflavins fractions, tea tasters' scores, and %RWC.

4.3 Hypotheses

Alternative hypothesis (H_1): Putative QTLs detected are linked to proteins that are associated with caffeine biosynthesis, catechins biosynthesis, theaflavins fractions, tea tasters' score, and %RWC.

Null hypothesis (H_0): Putative QTLs detected are not linked to proteins that are associated with caffeine biosynthesis, catechins biosynthesis, theaflavins fractions, tea tasters' score, and %RWC.

4.4 Materials and methods

4.4.1 Preparation of FASTA files

The sequences of DArTseq- based markers were derived from the tag sequence associated with each DArTseq marker as explained in the previous work (Koech *et al.*, 2018) and were generated by Diversity Arrays Technology Pty. Ltd. (Canberra, Australia). Marker sequences

were arranged in FASTA format by starting with a single-line description of the sequence, followed by sequence data. The single-line description was distinguished from the sequence data by placing a symbol ">" in front of the description

4.4.2 BLAST search

A total of 1 421 DArTseq markers on LG1 to LG15 were subjected to a BLAST search. The marker tag sequences (FASTA format) derived from the DArTseq map were searched using the BLASTN program against the non-redundant assembled draft tea genome. The best hit was then selected based on E-value and %identity. The E-value served as a measure of the number of hits one can "expect" to see by chance when searching a database with small E-value indicating homology. The %identity showed the percentage of identical residues between query sequences and hit sequences from a database, with long stretches of homology more likely to indicate a genuine match.

4.4.3 Functional annotation and pathway assignment

Functional annotation of all 1 421 DArTseq markers was performed on BLASTX search against the non-redundant GenBank protein sequence database with a threshold E-value of 10^{-6} (Conesa *et al.*, 2005; Conesa and Götz, 2008). Annotation and mapping of GO terms were done with the Blast2GO program (Blast2GO v3.2). The GO terms associated with each BLAST hit were retrieved (mapping step) and GO annotation assignment (annotation step) to the query sequences was carried out using the following 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). The contig sequences were also queried for conserved domains/motifs using InterProScan, an on-line sequence search plug-in within the Blast2GO program with all 13 applications selected before the run and the resulting GO terms were merged with the GO term results from the annotation step of Blast2GO. To determine metabolic pathways, KEGG mapping was used. The sequences with corresponding evidence code (EC) numbers obtained from Blast2GO were mapped to the KEGG metabolic pathway database.

4.5 Results

4.5.1 Functional annotation and pathway assignment

The identification of candidate genes for tea was conducted in all the 15 LGs derived from the DArTseq maps. Sequences of each marker on all the LGs were subjected to a BLAST search against tea genome BLAST database to compare the homology and identify the

location of the gene of interest within the tea genome. The results of BLAST searches for markers derived from the DArTseq map are presented in Table 4.1. Of 1 421 DArTseq markers derived from DArTseq map, 53 markers were identified to be linked to black tea quality and %RWC. A total of seven markers were dominant in TRFK 303/577 parent while nine markers were dominant in GW Ejulu parent, respectively. All the 53 markers with unique best hits were identified in the tea genome. The 23 markers showed locations across ten chromosomes in tea, which were LG1 (1), LG3 (1), LG4 (8), LG6 (1), LG9 (1), LG10 (2), LG12 (2), LG13 (4) LG14 (2) and LG15 (1). A total of 14 markers were dominant, eight were dominant in GW Ejulu, and six were dominant in TRFK 303/577. The dominant markers in GW Ejulu were for caffeine, catechin, ECG, EGC, and EGCG. On the other hand, the dominant markers in TRFK 303/577 were for caffeine, ECG, EGC, TF1, and %RWC. All the 13 dominant markers were identified in the tea genome for the two parents used in this study as reciprocal crosses in the two tea populations (Table 4.1).

4.5.2 Annotation and mapping of gene ontology

All unigenes in this study were assigned GO terms based on BLASTX searches against the non-redundant (NR) databases (Figure 4.1 and Figure 4.2; Appendix 4.1 to Appendix 4.13). Our results showed that 86-100% similarity distribution of the top hits in the NR database (Figure 4.1). The E-value distribution of the mapped sequences with high homologies (smaller than $1e-6$) was high as compared to homologous sequences that were greater than $1e-5$ (Figure 4.2). The E-value distribution of the best hits in the NR database showed that 26% annotated sequences had a high identity with their best hits (smaller than between $1e-28$ to $1e-26$), whereas 15% ranged from $1e-21$ to $1e-25$ and another 18% ranged from $1e-15$ to $1e-20$ (Figure 4.1). The similarity distribution revealed that 43% of the query sequences had greater than 91-95% similarity, whereas 21% of the query sequences varied in similarity from 86-90%, and the other 3% varied from 81-85% (Figure 4.2).

A total of 5 592 unigenes were assigned GO terms, of these, biological processes (56%) comprised the largest category, followed by cellular component (29%) and molecular functions (15) (Figure 4.3). In biological process (BP) category, the majority of unigenes were involved in regulation of biological process (35%), cellular process (16%), metabolic process (12%), response to stimulus (11%), developmental process (9%) and localization and signalling (6%) (Figure 4.3). The cellular and metabolic processes were the most highly represented groups in the BP, which suggests that extensive metabolic activities are taking place in the fresh tea flush. The best-represented groups of molecular function (MF) were

binding activity (48%), catalytic activity (27%), signal transducer activity (10%), molecular transducer activity (8%) and transporter activity (7%), respectively (Figure 4.3). Within the MF category, binding and catalytic activities were the most abundant groups. The cellular component (CC) included the cell (21%), cell part (21%), organelle (16%), membrane (11%) organelle part (7%), membrane part (7%), and macromolecular complex (6%), respectively (Figure 4.3). The transcripts that corresponded to the cell, cell parts, and cell organelles were the most abundant groups within the CC category. Among the top 20 functional sub-groups in BP, multicellular organism development, response to stress and response to endogenous stimulus represented the largest group (Appendix 4.9). The top 20 functional sub-groups for MFs, protein binding, receptor binding, and signal transducer activity represented the largest group (Appendix 4.10), while plasma membrane, nucleus and cytoplasm were the largest groups that represented the top 20 sub-groups of CC (Appendix 4.11).

4.5.3 KEGG pathway mapping: Flavonoid pathway, caffeine pathway and abiotic stress

Pathway-based analysis for the tea flush transcriptome is helpful to understand further the biological functions of genes involved in the tea flavonoid biosynthesis pathway, caffeine pathway, and stimulus to response to abiotic stress (drought stress). A total of 53 unigenes linked to putative QTLs were assigned two different KEGG database pathways (Appendix 4.12) while 84 unigenes in 15 linkage groups were assigned to 25 different KEGG database pathways based on categories of secondary metabolite biosynthesis (Figure 3.4, Appendix 4.13). The pathways with most representations were metabolic pathways and biosynthesis of secondary metabolites, which was consistent with the GO categories of the BP (Figure 4.3 and Appendix 4.9). Of these, 52 unigenes were categorised into the metabolism groups, with most of them involved in purine metabolism (11%), cysteine and methionine metabolism (9.6%), pyruvate metabolism (9.6%), alanine, aspartate and glutamate metabolism (7.6%), arginine and proline metabolism (7.6%) and fructose and mannose metabolism (7.6%) and other sub-categories, respectively. The most abundant unigenes were for carbohydrate and amino acids biosynthesis, which are mostly involved in plant hormone signal transduction pathways in response to abiotic stress and biosynthesis of other secondary metabolites such as phenylpropanoid biosynthesis, flavonoid biosynthesis, and alkaloid biosynthesis. The unigenes associated with flavone and flavonol biosynthesis, a key regulatory pathway in the metabolism of the main chemical components in *C. sinensis* leaves were present (Figure 4.4). Unigenes associated with purine metabolism, also a key regulatory pathway which is involved in metabolism of adenine and guanosine nucleotides were present.

Also, major enzymes that are involved in different metabolic pathways were further grouped into the six different categories; oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The distribution of the three abundant enzymes was also consistent with the GO categories of MF (Figure 4.3 and Appendix 4.10). The three major enzymes distribution included; transferase enzymes (38.9%), which are a class of enzymes that catalyse the transfer of specific functional groups (a methyl, or glycosyl group, or galloyl group) from a donor to acceptor molecules, hydrolase enzymes (29%), which catalyse the hydrolysis of compounds and oxidoreductase enzymes (18.3%), which catalyse the oxidation-reduction reactions (Figure 4.5). The three other enzymes that were less abundant in distribution included; lyase enzymes (8.4%), which cleave carbon-carbon, carbon-oxygen, phosphorous-oxygen and carbon-nitrogen bonds without hydrolysis or oxidation reactions, isomerase enzymes (3.1%), convert a molecule from one isomer to another while ligase enzymes (2.3%), catalyse the formation of a new chemical bond by joining of two large molecules (Figure 4.5).

4.5.4 Functional annotation of putative candidate genes 15 linkage groups of *C. sinensis*

The 28% of the mapped sequences (405 out of 1 471) were annotated. Functional annotations of detected QTLs led to the identification of functions of 37 putative candidate genes in parental clones that could be involved in the expression of the targeted traits (Table 4.1). The putative candidate genes involved in secondary metabolism (phenylpropanoid biosynthesis, flavonoid biosynthesis, alkaloid biosynthesis, and volatile compounds biosynthesis) were located in five linkage groups LG04, LG09, LG10 and LG13. The putative candidate genes identified were involved in phenylpropanoid and flavonoid biosynthesis pathways included a group or class of enzymes namely; 2-oxoglutarate/Fe(II)-dependent dioxygenases superfamily (2-ODDs), aminotransferase class I and II, acyltransferase, glycosyl hydrolase family 9. This was in agreement with KEGG pathway mapping where enzymes mapped clustered into six major groups according to different metabolic pathways they are involved with transferases, hydrolases, and oxidoreductases as three abundant enzymes. The multidrug and toxic compound extrusion (MATE) transporter was also putatively identified in LG09. BT1 family protein located in LG04 was putatively annotated for qCAFF. The putative QTLs and their corresponding putative candidate genes associated with phenylpropanoid biosynthesis, and flavonoid biosynthesis were identified in both parental clones.

Proteins related to oxidative stress and ATPase family proteins or proteins related to abiotic stress were found in the confidence intervals of QTL for drought tolerance traits. Among the

53 examined putative QTLs in parental clones (Table 4.1), 18 putative QTLs were related to abiotic stress (ABA signalling, amino acids peptide degradation) and were mostly found on LG04 and LG13. However, the putative QTL related to abiotic stress, including %RWC was also found on LG01, LG06, LG10, LG12, LG14, and LG09, respectively. The putative candidate genes identified included, pectinesterase, 14-3-3 protein, 2OG-Fe (II) oxygenase superfamily, protein kinase domain, transmembrane amino acid transporter protein, peptidase family M3, histone acetyltransferase subunit NuA4, DnaJ domain, and MatE. ATPase family proteins associated with various cellular activities (AAA), and armadillo/beta-catenin-like repeat protein were found to be associated with the heat stress response. The putative QTL in LG13 that was associated with disease stress response was NB-ARC domain protein. Putative candidate genes related with ABA signalling pathway were found on QTL for qCAT, qEC, qEGCG, and qRWC (Table 4.1), respectively. The putative candidate gene related to degradation of peptides by targeting the amino acids residues that are formed during abiotic stress was found on QTL for qEC. Parental clone TRFK 303/577 which is known to be high-yielding and drought-tolerant cultivar, exhibited most putative candidate genes related to drought stress response compared to parental clone GW- Ejulu.

Table 4.1: Functional annotation of putative candidate genes in 15 linkage groups of *C. sinensis* on reference tea genome

Nr	QTL	Parent	DArTseq Marker	LG	Position (cM)	LOD	PEV	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	['Diacylglycerol 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	['Armado/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, CAFF- Caffeine, CAT- Catechin, EC- Epicatechin, ECG- Epicatechin gallate, EGC- Epigallocatechin, EGCG- Epigallocatechin gallate, TF1- Theaflavin, TF4- Theaflavin-3,3'-digallate, BRT- Brightness,% RWC- Percent relative water content

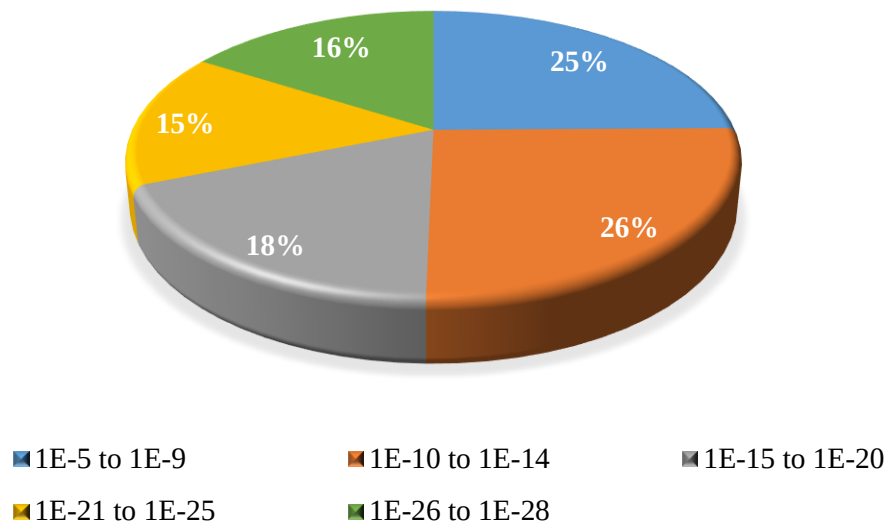


Figure 4.1: E-value distribution of the best matches for unigenes (E-value $\leq 1.0e-5$)

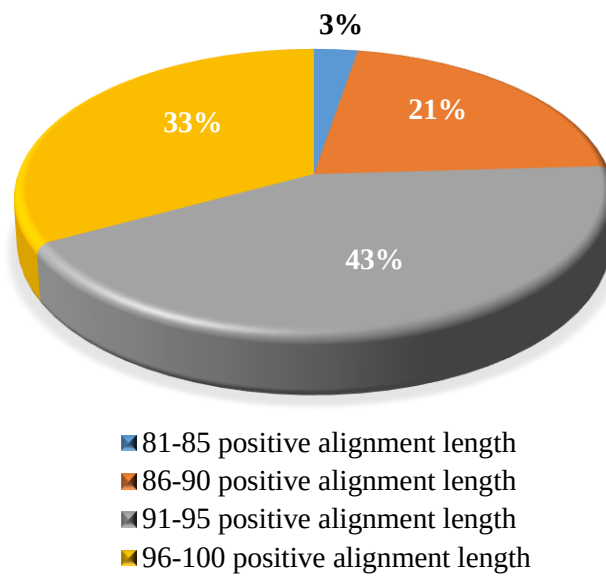


Figure 4.2: Similarity distribution of the hit BLAST matches for unigenes.

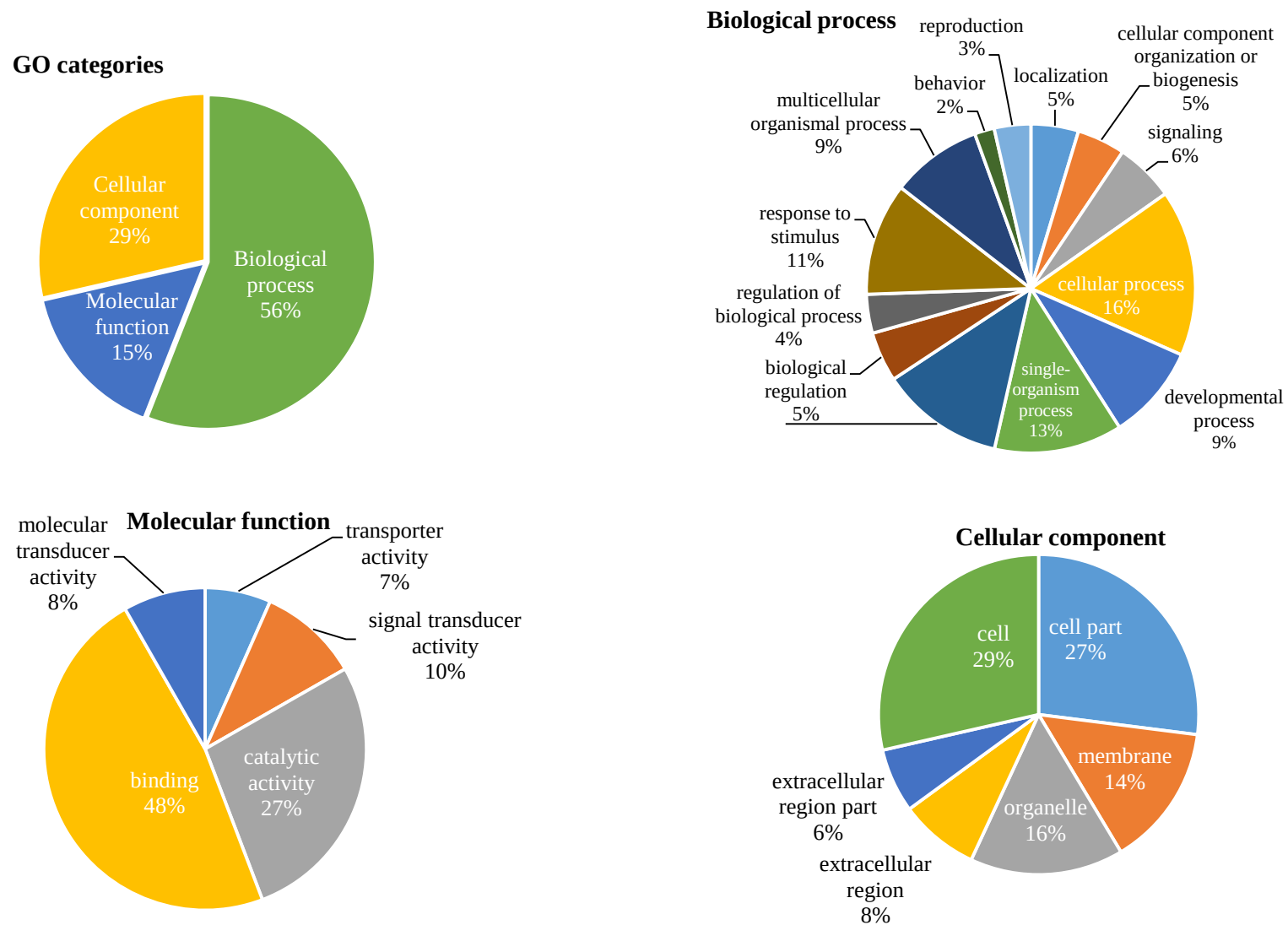


Figure 4.3: Function classifications of GO terms of *C. sinensis* transcripts based on high-score BLASTX matches in NR plant proteins database, a total of 5,592 unigenes were classified into three main GO categories which are biological process (BP), molecular function (MF) and cellular component (CC), respectively.

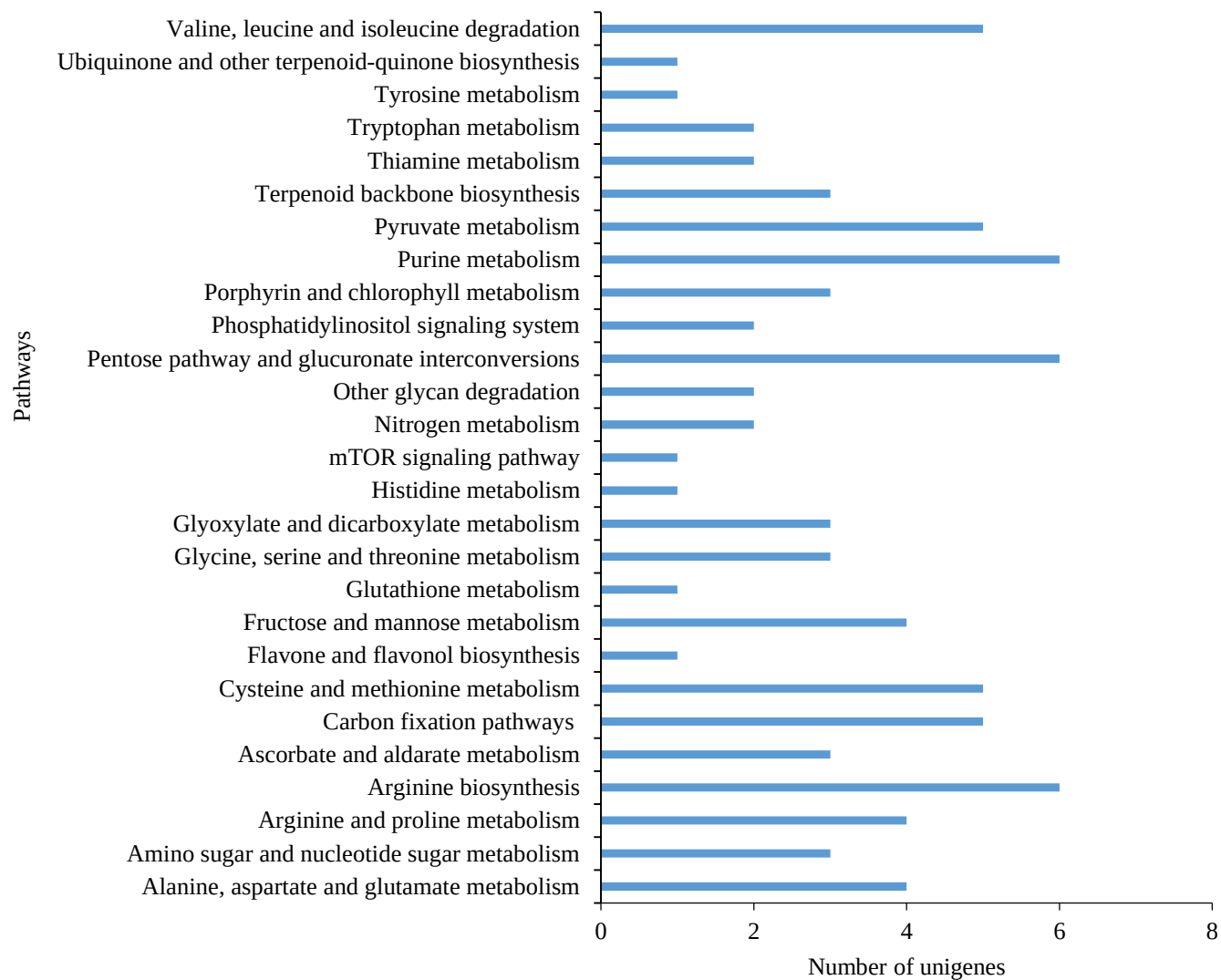


Figure 4.4: KEGG ontology (KO) analysis of differentially expressed genes (DEGs) related to secondary metabolism in *C. sinensis*

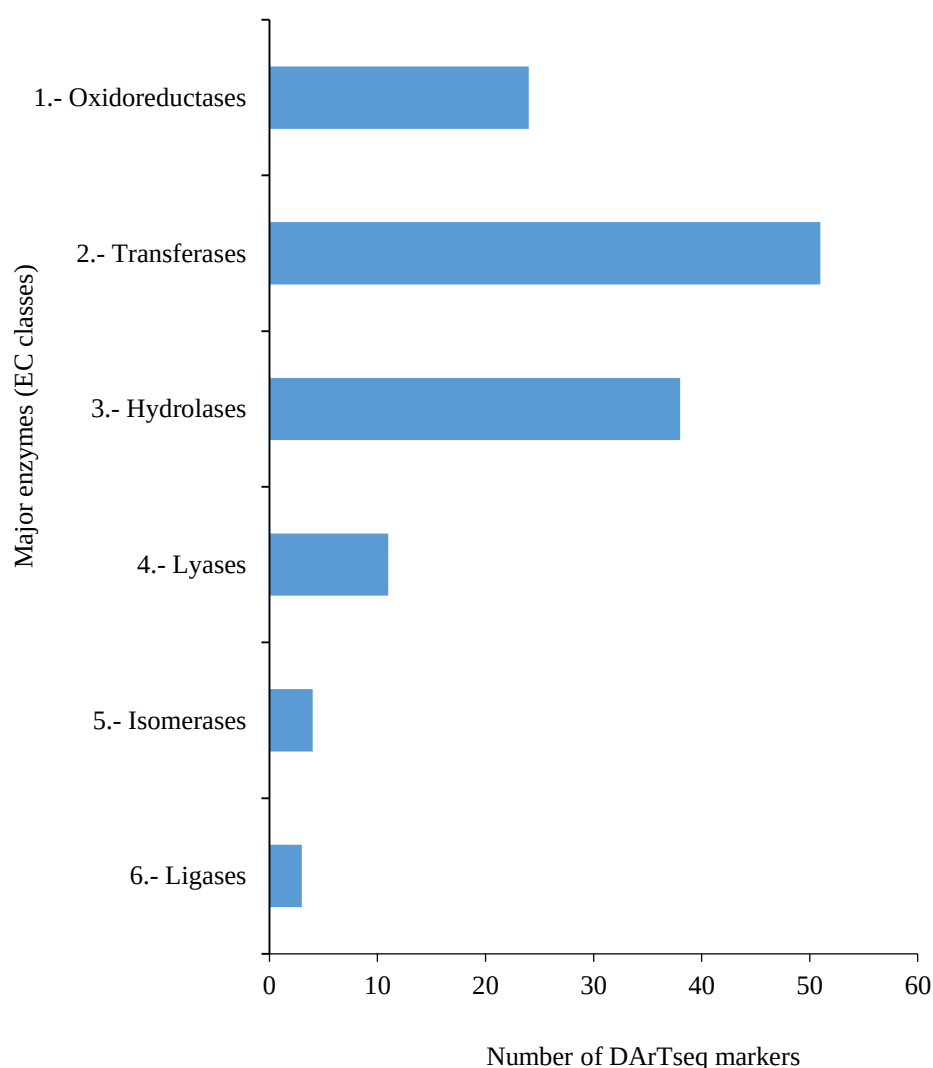


Figure 4.5: KO distribution of major enzymes code related to secondary metabolism in *C. sinensis*

4.6 Discussion

Flavonoid biosynthesis pathway in tea begins with the amino acid phenylalanine, which undergoes deamination to form trans-cinnamic acid (Punyasiri *et al.*, 2004; Vankatesh *et al.*, 2007). The oxidation process of trans-cinnamic acid yields p-coumaric acid, which transforms to form p-coumaroyl-CoA. The committed step in flavonoid biosynthesis is the subsequent condensation of p-coumaroyl-CoA and malonyl-CoA catalysed by chalcone synthase enzyme to form chalcone. The subsequent isomerisation of chalcone forms (2S)-flavonones. The identification of putative QTL, qCAT, for 2-ODDs superfamily protein in this study corroborates with other previous reports on the role of 2-ODDs in flavonoid

biosynthesis. The 2-ODDs which are large superfamily of non-heme proteins, which facilitate numerous oxidative reactions, including hydroxylation, epimerization, C-C bond cleavage, ring formation, and fragmentation. These 2-ODDs superfamily catalyse key oxidative reactions that facilitate the formation of different flavonoid subclasses, for example, the oxidation of (2S)-flavonones. The (2S)-flavonones are converted to dihydroflavonols through hydroxylation process which is then catalysed by another 2-ODD, flavonone 3 β -hydroxylase (F3H). The dihydroflavonols formed is a substrate for flavonol synthase (FNS I), which competes with dihydroflavonol 4-reductase (DR4) to form flavonols, anthocyanidins, and procyanidins.

Glycosylation is an important process for the diverse functions of polyphenolic compounds in plants and is known to be able to increase the solubility and stability of hydrophobic flavonoids (Xu *et al.*, 2016). The glycosylation process is regulated by glycoside hydrolases, which are involved in hydrolysis and/or rearrangement of glycosidic bonds. We hypothesise the putative QTL, qBRT linked to glycosyl hydrolase family 9, a family of glycoside hydrolases in the current study to play a role in the synthesis and sequestration of some phenylpropanoids (Jones *et al.*, 2003). In tea, the galloylated catechins formed through the glucosylation activity of UDP-glucosyltransferase, have been found to confer astringent and bitter tastes, while flavonol 3-O-glycosides have been found to induce velvety, mouth-drying, and mouth-coating sensations (Cui *et al.*, 2016). In legume plant, bur clover (*Medicago truncatula*), epicatechin 3'-O-glucoside, has also been shown to be involved in the biosynthesis of proanthocyanidin in the seed coat (Pang *et al.*, 2007; Zhao and Dixon, 2009). Also, the glycosylated polyphenols also function as acyl donors in biochemical reactions, for example, β -glucogallin, a glucose ester of gallic acid, functions as an acyl transfer donor in the biosynthesis of both galloylated tannins (Niemetz and Gross, 2005) and galloylated catechins (Liu *et al.*, 2012).

Acyltransferases which participates in many secondary metabolism pathways was putatively identified in this current study (qECG) in LG10. Acyltransferases are involved in the biosynthesis of anthocyanidin and other flavonoid groups by transferring acyl groups to the sugar moiety of anthocyanins using acyl-CoA as the donor (Mizutani *et al.*, 2006). The process of modification of anthocyanin and other flavonoid groups with acyl groups is catalysed by acyltransferases and is important in maintaining their stability. In LG04, qCAFF was putatively annotated as BT1 family protein which is has been identified to be involved in export of adenine nucleotides, which are exclusively synthesised in plants plastids, and are

precursors for caffeine biosynthesis (Haferkamp, 2007). Furthermore, nucleotides, adenine, and guanosine are precursors in caffeine biosynthesis in tea plants (Negishi *et al.*, 1992).

The 14-3-3 proteins which bind to other proteins to induce target-site specific alteration and conformation of target proteins are important plant signal transduction pathways. Previous reports implicate plant 14-3-3s in key physiological processes, in particular, abiotic and biotic stress responses, metabolism (especially primary carbon and nitrogen metabolism), as well as various aspects of plant growth and development. In the current study, putative QTL, qCAT in LG13 was putatively annotated 14-3-3 protein, which is abiotic or biotic stress protein in tea. This, therefore, is in agreement with the previous study by Cheruiyot *et al.* (2008b) where individual catechins were demonstrated as potential indicators in predicting drought tolerance in tea plant. Meanwhile, a number of recent studies in *Arabidopsis* have linked 14-3-3s to ABA signalling. Some of the main functions ascribed to ABA include response to abiotic/biotic stress.

Drought stress which usually affects the plant photosynthetic process also interferes with plant nutrient availability leading to ion intoxication. The reversible protein phosphorylation by protein kinases at the early and later stage of the signalling pathways is one of the mechanisms that plants use in response to abiotic stress. Also, to maintain the osmotic homeostasis and tolerance due to drought stress, plants also do regulate gene expression of some specific genes of protein kinases. The presence of a putative QTL, qCAT, which was annotated for protein kinase domain in this study supports the argument of the previous studies on the role of tea catechins as potential indicators of drought tolerance in tea plants. A previous study by Jeyaramraja *et al.* (2003) indicated that catechins were higher in the drought-tolerant cultivars and therefore, corroborated with the presence of qCAT, annotated as protein kinase domain in drought-tolerant parental clone TRFK 303/577.

Transmembrane amino acid transporter protein functions as gateways to permit the transport of amino acids across the biological membrane. Amino acids are not only critical for the synthesis of proteins but other many essential compounds, which act as signal molecules. For example, in *Arabidopsis*, water and salt stress induce a strong expression of proline transporter 1 (ProT1) and proline transporter two while the expression of amino acid permease 4 (AAP4) and amino acid permease 6 (AAP6) is repressed (Hayat *et al.*, 2012; Hua *et al.*, 2017). In tea, proline has been shown to act as an osmo-protectant during drought stress (Upadhyaya and Panda, 2013; Upadhyaya *et al.*, 2016) and the presence of a putative

candidate gene (transmembrane amino acid protein) in this current study, could hypothesise the functions of transmembrane amino acid transporter for proline presence during drought stress. Furthermore, the presence of transmembrane amino acid transporter proteins could be their involvement of transport of aromatic amino acids phenylalanine, tyrosine and tryptophan, which serve as precursors for a wide range of secondary metabolites (polyphenols) in tea. This is because the two identified putative QTL, qCAT and qEGC in LG13 in both parental clones (TRFK 303/577 and GW Ejulu) are products of shikimate and phenylalanine pathways.

MATE proteins are a family of secondary active transporters, which utilise the electrochemical gradient of membrane maintained by ATPases for their transport activity. MATE transporters are involved in the transport of secondary metabolites into vacuoles, which are the major storage site of most of the conjugated flavonoids, which comprise of flavonols, anthocyanins and flavone glycoside (Shitan, 2016). In the current study, putative QTL, qRWC in LG09 was annotated for MATE proteins, which are involved in the sequestration of flavonoids in vacuoles in response to drought stress (Petrussa *et al.*, 2013). Secondary metabolites in plants, especially flavonoids play an important role in response to abiotic stress such as drought stress. A member of the MATE gene family (TT12) has also been elucidated for sequestration of proanthocyanidins in seed vacuoles, which leads to pigmentation of the seed coat (González *et al.*, 2016). In yeast, it was revealed that TT12 proteins could specifically transport glycosidic form like epicatechin 3'-O-glucoside and cyanidin 3-O glucoside (Marinova *et al.*, 2007). The MATE proteins from a legume, bur clover (*Medicago truncatula*) and grapevines (*Vitis*) have been characterised for transport of flavonoids through the sequestration of epicatechin 3-O-glucoside and cyanidin 3-O glucoside (proanthocyanidins) into the vacuole for storage (Zhao and Dixon, 2009). In another study, apart from the role of vacuolar sequestration flavonoids, MATE proteins have been recognised in the transport of alkaloids into vacuoles in tobacco. Furthermore, MATE proteins have facilitated and modulated ABA efflux and sensitivity in drought-tolerant *Arabidopsis* (Zhang *et al.*, 2014). Since a putative QTL for %RWC annotated as MATE in this current study, it, therefore, corroborates with previous reports (Nyarukowa *et al.*, 2016) which relied on the %RWC to classify tea cultivars as either drought-tolerant or drought susceptible based on tea leaves metabolites.

In this current study, putative QTL, qECG in LG06 was putatively annotated DnaJ protein. DnaJ proteins are essential components that contribute to cellular protein homeostasis and complex stabilisation of proteins (protein folding, degradation and refolding) under stress conditions in plants (Park and Seo, 2015). They function as molecular chaperones either alone or in association with heat-shock protein 70 (Hsp70). In a previous study on the transcriptomic analysis of the effect of drought-induced-stress on tea leaf quality, revealed that the levels of ECG and EGCG increased during drought stress (Wang *et al.*, 2016). This was also reported by Jeyaramraja *et al.* (2003) on the effects of soil moisture stress alterations on tea biomolecules in relation to tea quality. Furthermore, in transgenic tobacco, the overexpression of tomato chloroplast-targeted DnaJ enhanced drought tolerance (Wang *et al.*, 2014a) and maintained photosystem II under chilling stress (Kong *et al.*, 2013). Our findings, therefore, suggest that ECG can be used as a potential marker of drought tolerance or cold stress in tea cultivars.

The putative QTL, qEGCG, in LG04 was annotated diacylglycerol kinase catalytic domain protein which we hypothesise as a marker for drought and cold tolerance in tea cultivars. Abiotic stress, including salinity, drought, and osmotic adjustments triggers the production of phosphatidic acid (Zhu, 2016). Phosphatidic acid is formed by phosphorylation of diacylglycerol by diacylglycerol kinase. Thus, phosphatidic acid rather than diacylglycerol has typically been implicated as a major plant secondary messenger (Zhu, 2016). However, the presence of diacylglycerol is necessary for certain developmental processes and the response to particular environmental stimuli. For example, salinity stress in *A. thaliana* increased the activity of non-specific phospholipase C, which promoted the production of diacylglycerol. Also, cold stress has recently been found to stimulate the formation of phosphatidic acid in suspension-cultured *Arabidopsis* cell (Arisz *et al.*, 2013).

Histone acetyltransferases play a critical role in histone acetylation of plant chromatin, which is important in epigenetic control of gene expression. The transfer of acetyl groups to core histone tails by histone acetyltransferases, promote transcription of target genes involved in drought response and abscisic acid signalling in several plants, including *Arabidopsis* (Kim *et al.*, 2008), rice (Fang *et al.*, 2014), and maize (Kapazoglou and Tsaftaris, 2011). Furthermore, previous studies reported up-regulation of drought stress-responsive genes and found to be correlated with changes in histone modification (Kim *et al.*, 2015; Asensi-Fabado *et al.*, 2017). The putative QTL, qCAT, qEC, and qEGCG in this current study were annotated histone acetyltransferase subunit NuA4 protein in drought-tolerant parental clone TRFK

303/577. Therefore, catechins could act as potential indicators or markers of drought tolerance in tea, which corroborates with previous reports by (Jeyaramraja *et al.*, 2003; Cheruiyot *et al.*, 2008b), which investigated the biochemical changes in tea constituents as influenced by varying levels of water stress.

Aminotransferases, also known as transaminases, catalyse amino group transfers from amino donor to amino acceptor compounds. Aminotransferases play a major role in a variety of metabolic pathways, including, amino acid biosynthesis, secondary metabolites biosynthesis, and photorespiration (de la Torre *et al.*, 2014). Prephenate aminotransferase, a class of aminotransferase enzyme, catalyse the final step in the synthesis of phenylalanine, which is the central product of shikimic acid pathway and serves as a precursor for the synthesis of flavonoids (Ververidis *et al.*, 2007). The identification of putative QTL, qEC in the current study annotated as aminotransferase I and II is hypothesised as a marker linked to flavonoids biosynthesis in tea.

4.7 Conclusion

This current study is the first attempt to obtain more information on genomic and functional annotation of proteins/ enzymes involved in black tea quality and drought tolerance traits in tea plant based on DArTseq markers. The results in this study were in agreement with the alternative hypothesis that putative QTL detected were linked to proteins that are associated with caffeine biosynthesis, catechins biosynthesis, theaflavins fractions, tea tasters' score, and %RWC. Therefore, the DArTseq markers used, and putative QTL identified in this current study will shed more light on the proteins/ enzymes associated with the biosynthesis of caffeine, individual catechins in green tea which are converted to individual theaflavins in black tea processing. Also, the DArTseq markers will provide useful information, in particular for studies of the genetic determinants of black tea quality traits and drought tolerance in tea. Also, the DArTseq markers used in the current study can be useful markers for construction of linkage maps or may be used to discover new functions of genes. We successfully constructed a genetic linkage map using SNPs, but it was impossible to anchor it to other previously constructed linkage maps of tea since no anchoring markers were available. Furthermore, it was also difficult to deduce the relative genetic positions of the SNPs on the previous different linkage maps. Indeed, previous studies used RAPD, AFLP and SSR markers to construct tea genetic linkage maps, while in our study, we used DArTseq markers. However, the information obtained through the annotation of putative gene functions and the alignment of DArTseq sequences relative to recently published reference

tea genome has set a milestone for a future and further researches on MAS and breeding of tea plants. Also, this work may help breeders in selection of parents with desirable DArTseq markers for development of new tea cultivars with desirable traits.

References

- 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.
- Asensi-Fabado, M. A., Amtmann, A. & Perrella, G. (2017). Plant responses to abiotic stress: The chromatin context of transcriptional regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1860(1): 106-122.
- 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.
- 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.
- 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.
- Conesa, A. & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 2008.
- 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.
- 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.

- 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.
- Haferkamp, I. (2007). The diverse members of the mitochondrial carrier family in plants. *FEBS Letters* 581(12): 2375-2379.
- Hayat, S., Hayat, Q., Alyemeni, M. N., Wani, A. S., Pichtel, J. & Ahmad, A. (2012). Role of proline under changing environments: a review. *Plant Signalling and Behavior* 7(11): 1456-1466.
- 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.
- Kapazoglou, A. & Tsaftaris, A. (2011). Epigenetic chromatin regulators as mediators of abiotic stress responses in cereals. In *Abiotic Stress in Plants-Mechanisms and Adaptations*: InTech.
- 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.
- Kong, F., Deng, Y., Zhou, B., Wang, G., Wang, Y. & Meng, Q. (2013). A chloroplast-targeted DnaJ protein contributes to maintenance of photosystem II under chilling stress. *Journal of Experimental Botany* 65(1): 143-158.

- Liang, Y., Lu, J. & Zhang, L. (2002). Comparative study of cream in infusions of black tea and green tea [*Camellia sinensis* (L.) O. Kuntze]. *International Journal of Food Science and Technology* 37(6): 627-634.
- Liu, S. C., Jin, J. Q., Ma, J. Q., Yao, M.-Z., Ma, C. L., Li, C. F., Ding, Z. T. & Chen, L. (2016). Transcriptomic analysis of tea plant responding to drought stress and recovery. *PloS one* 11(1): e0147306.
- Liu, Y., Gao, L., Liu, L., Yang, Q., Lu, Z., Nie, Z., Wang, Y. & Xia, T. (2012). Purification and characterization of a novel galloyltransferase involved in catechin galloylation in the tea plant (*Camellia sinensis*). *Journal of Biological Chemistry* 287(53): 44406-44417.
- 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.
- Maritim, T. K., Kamunya, S. M., Mireji, P., Mwendia, C., Muoki, R. C., Cheruiyot, E. K. & Wachira, F. N. (2015). Physiological and biochemical response of tea [*Camellia sinensis* (L.) O. Kuntze] to water-deficit stress. *The Journal of Horticultural Science and Biotechnology* 90(4): 395-400.
- 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.
- Ng'etich, W., Stephens, W. & Othieno, C. (2001). Responses of tea to environment in Kenya. 3. Yield and yield distribution. *Experimental Agriculture* 37(3): 361-372.
- Niemetz, R. & Gross, G. G. (2005). Enzymology of gallotannin and ellagitannin biosynthesis. *Phytochemistry* 66(17): 2001-2011.
- 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.
- 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.

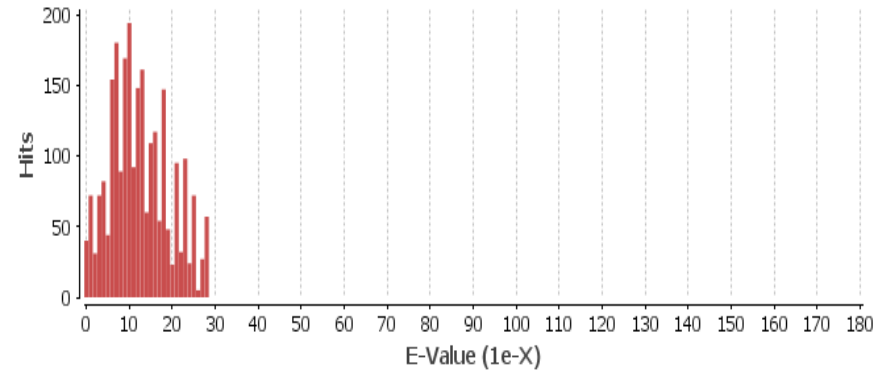
- Pang, Y., Peel, G. J., Wright, E., Wang, Z. & Dixon, R. A. (2007). Early steps in proanthocyanidin biosynthesis in the model legume *Medicago truncatula*. *Plant Physiology* 145(3): 601-615.
- Park, C. J. & Seo, Y.-S. (2015). Heat shock proteins: a review of the molecular chaperones for plant immunity. *The Plant Pathology Journal* 31(4): 323.
- 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.
- 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.
- Rossetti, D., Bongaerts, J., Wantling, E., Stokes, J. & Williamson, A. M. (2009). Astringency of tea catechins: more than an oral lubrication tactile percept. *Food Hydrocolloids* 23(7): 1984-1992.
- Sah, S. K., Reddy, K. R. & Li, J. (2016). Absciscic acid and abiotic stress tolerance in crop plants. *Frontiers in Plant Science* 7: 571.
- 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.
- Shitan, N. (2016). Secondary metabolites in plants: transport and self-tolerance mechanisms. *Bioscience, Biotechnology and Biochemistry* 80(7): 1283-1293.
- Tony, M., Samson, K., Charles, M., Paul, M., Richard, M., Mark, W., Stomeo, F., Sarah, S., Martina, K. & Francis, W. (2016). Transcriptome-based identification of water-deficit stress-responsive genes in the tea plant, *Camellia sinensis*. *Journal of Plant Biotechnology* 43(3): 302-310.
- Upadhyaya, H., B.K., D. & Panda, S. K. (2016). Drought-induced physiological and biochemical changes in leaves of developing seedlings of tea [*Camellia sinensis* (L) O Kuntze] Cultivars *Journal of Tea Science Research* 6(4): 1-11
- 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.

- 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, G., Cai, G., Kong, F., Deng, Y., Ma, N. & Meng, Q. (2014). Overexpression of tomato chloroplast-targeted DnaJ protein enhances tolerance to drought stress and resistance to *Pseudomonas solanacearum* in transgenic tobacco. *Plant Physiology and Biochemistry* 82: 95-104.
- 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.
- 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.
- 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.
- Zhao, J. & Dixon, R. A. (2009). MATE transporters facilitate vacuolar uptake of epicatechin 3'-O-glucoside for proanthocyanidin biosynthesis in *Medicago truncatula* and *Arabidopsis*. *The Plant Cell* 21(8): 2323-2340.
- Zheng, C., Wang, Y., Ding, Z. & Zhao, L. (2016). Global transcriptional analysis reveals the complex relationship between tea quality, leaf senescence and the responses to cold-drought combined stress in *Camellia sinensis*. *Frontiers in Plant Science* 7: 1858.
- Zhou, L., Xu, H., Mischke, S., Meinhardt, L. W., Zhang, D., Zhu, X., Li, X. & Fang, W. (2014). Exogenous abscisic acid significantly affects proteome in tea plant (*Camellia sinensis*) exposed to drought stress. *Horticulture Research* 1: 14029.
- 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.

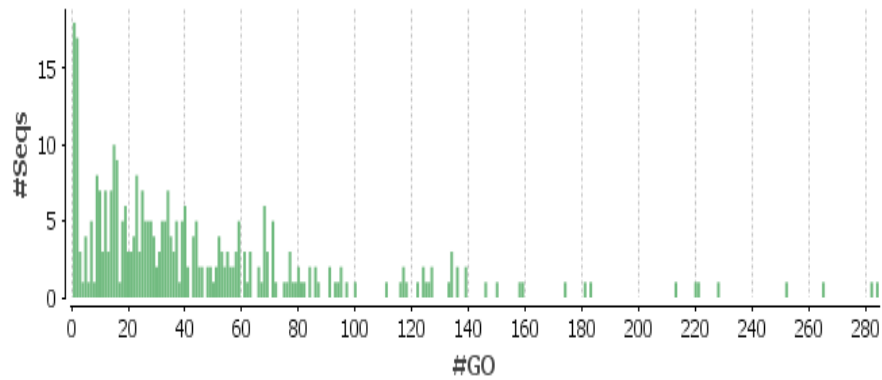
Appendices



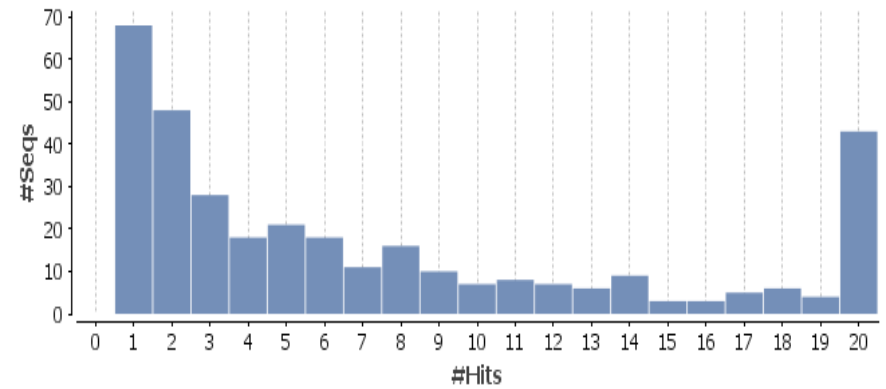
Appendix 4.1: Similarity distribution of the top BLAST matches for unigenes



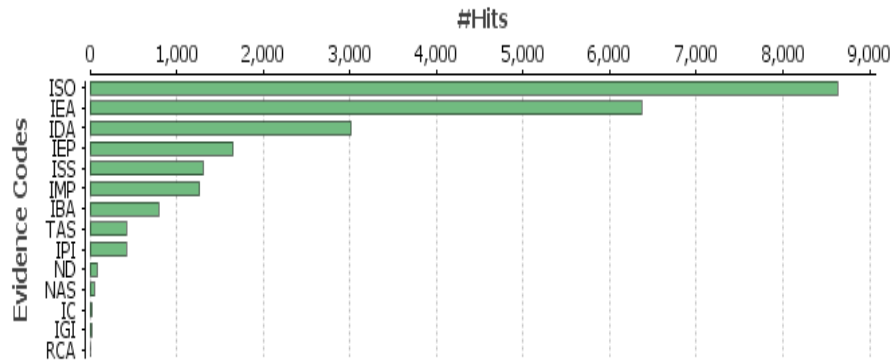
Appendix 4.2: E-value distribution of the best matches for unigenes (E-value $\leq 1.0e-5$)



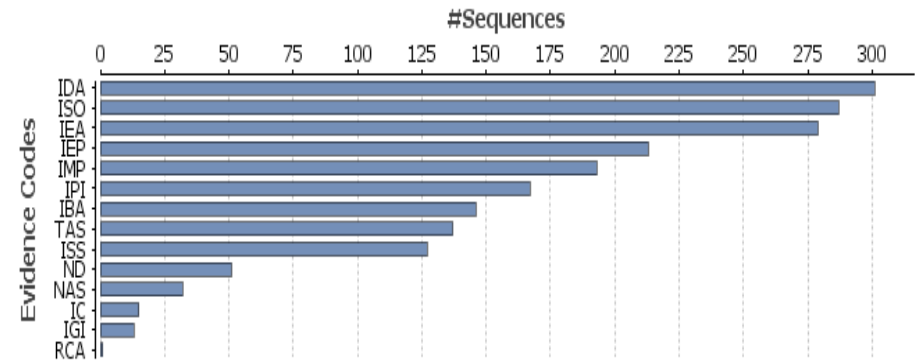
Appendix 4.3: DArTseq GO mapping distribution for tea (*C. sinensis*)



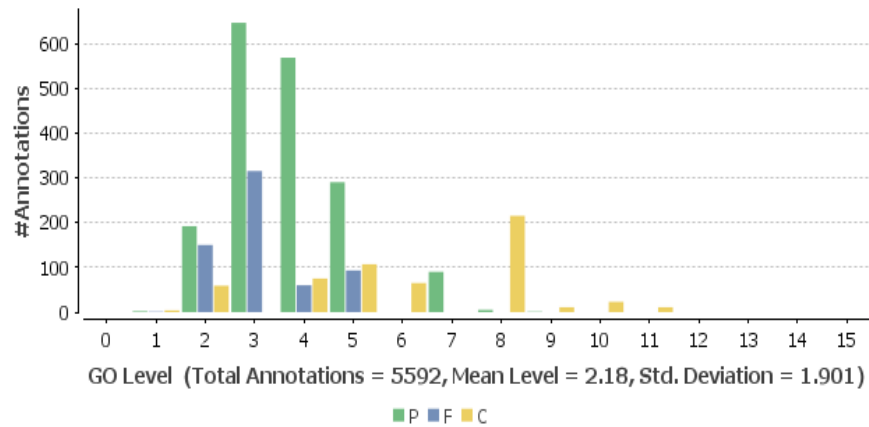
Appendix 4.4: DArTseq BLAST hit distribution for tea (*C. sinensis*)



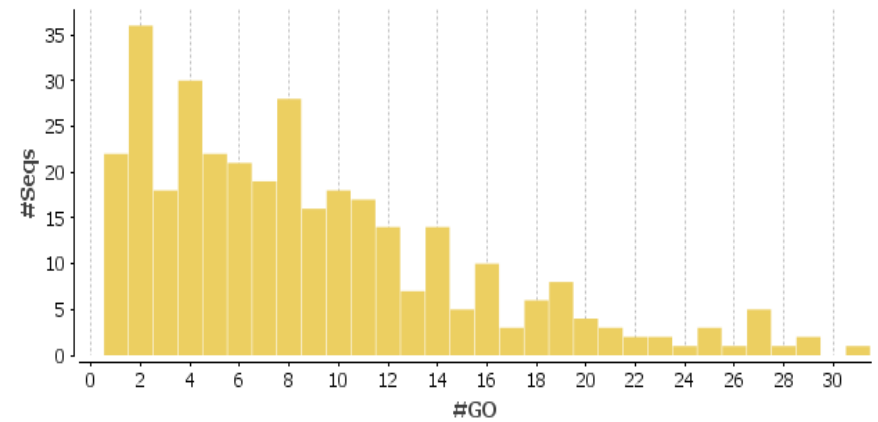
Appendix 4.5: DArTseq evidence code distribution for BLAST hits (*C. sinensis*)



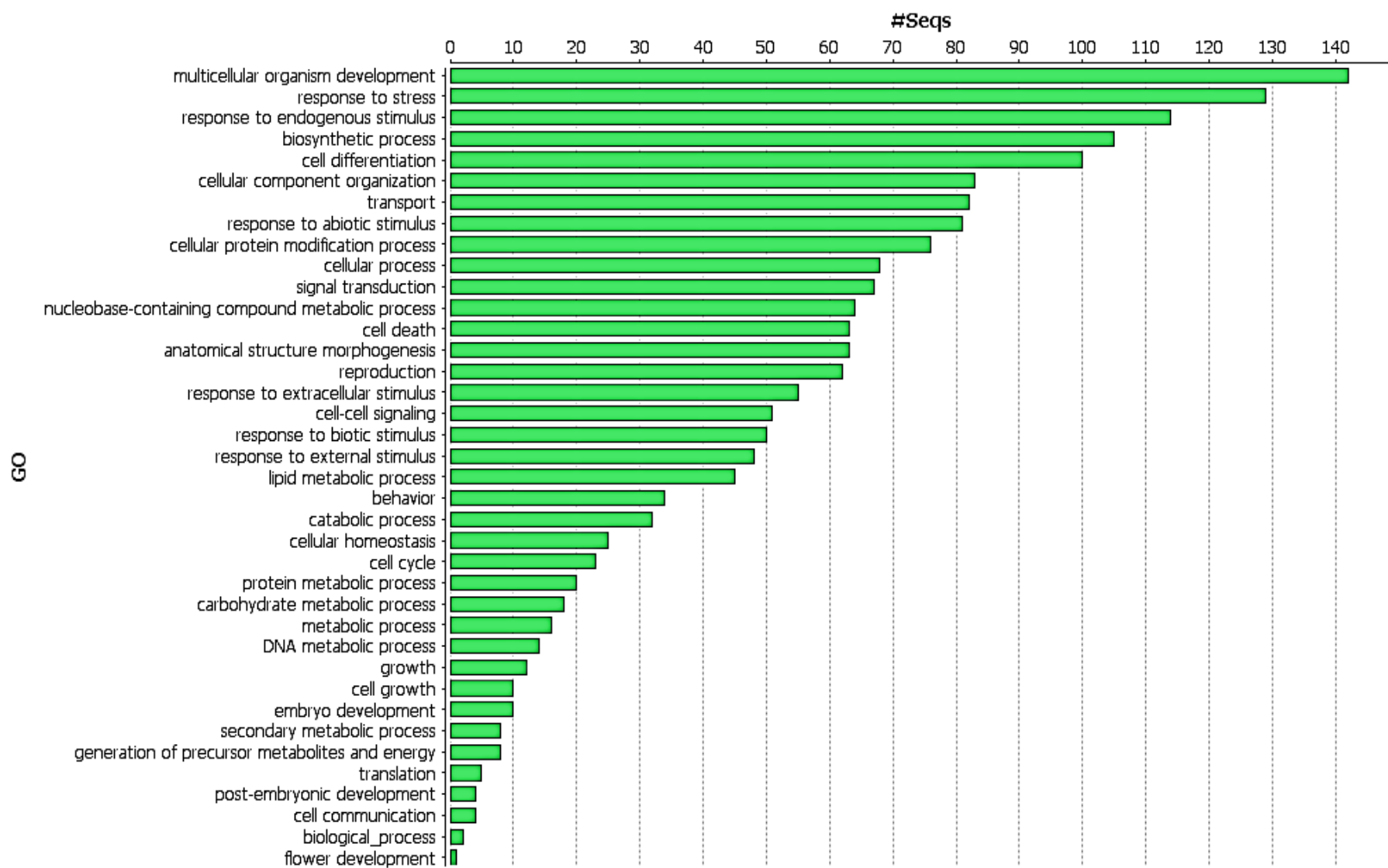
Appendix 4.6: DArTseq evidence code distribution for tea (*C. sinensis*)



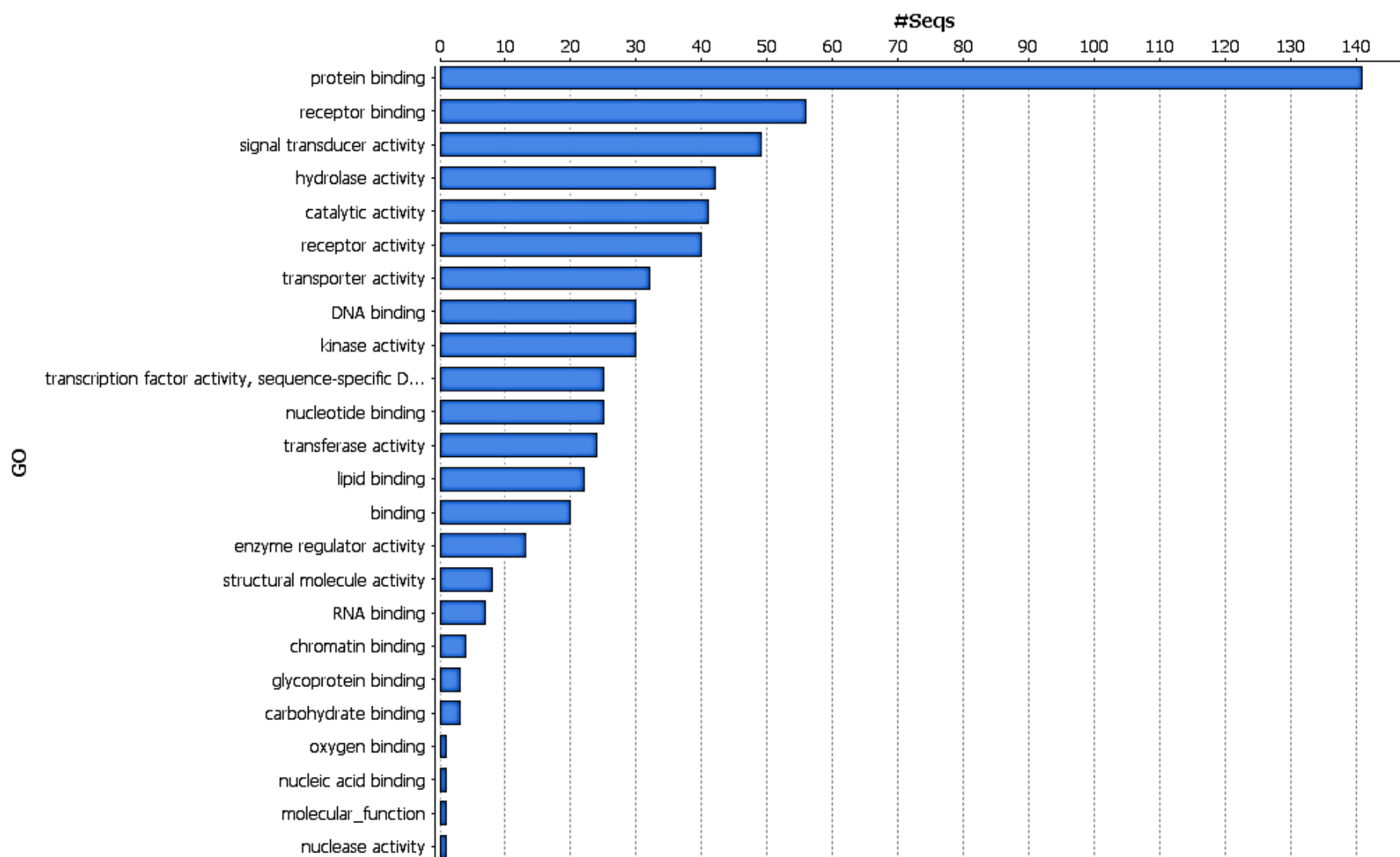
Appendix 4.7: DArTseq GO-level distribution for tea (*C. sinensis*)



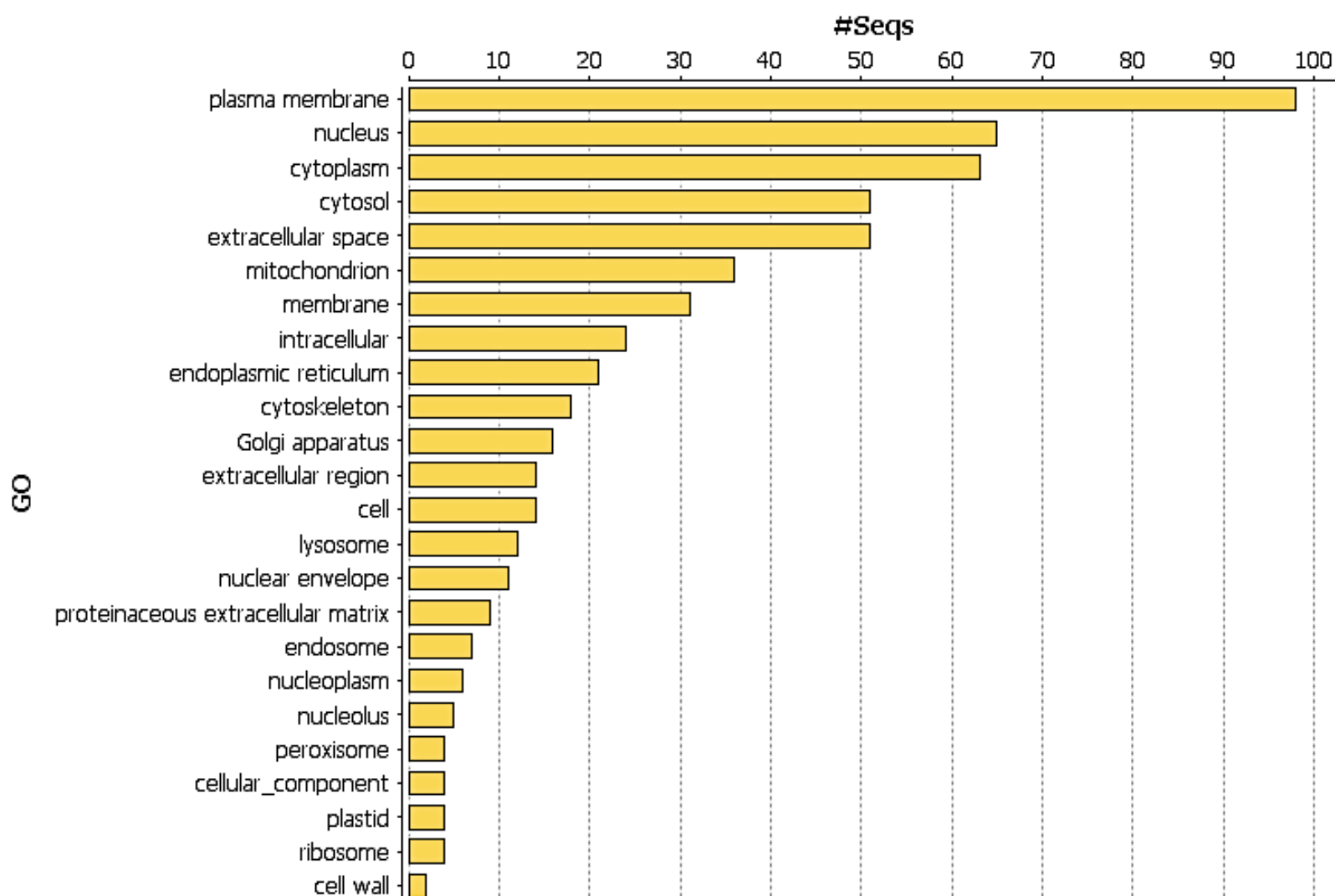
Appendix 4.8: DArTseq annotation distribution for tea (*C. sinensis*)



Appendix 4.9: Gene ontology classification of top 20 sub-groups of biological processes.



Appendix 4.10: Gene ontology classification of top 20 sub-groups of molecular functions.



Appendix 4.11: Gene ontology classification of top 20 sub-groups of cellular components.

Appendix 4.12: List of putative QTLs identified from KEGG pathways with GO terms in *C. sinensis* transcriptome

Nr	Trait	Locus	LG	Pos (cM)	LOD	PEV	E-Value	#GO	GO IDs	GO Name	Pathway	#Enzs in Pathway	Enzyme Codes
1	^a TF1	5084595	2	4.5	3.2	5.7	3.7E-23	9	F:GO:0000166; F:GO:0003824; P:GO:0007275; P:GO:0006629; P:GO:0030154; F:GO:0008289; P:GO:0009058; C:GO:0005622; F:GO:0005215	F:nucleotide binding; F:catalytic activity; P:multicellular organism development; P:lipid metabolic process; P:cell differentiation; F:lipid binding; P:biosynthetic process; C:intracellular; F:transporter activity	Steroid degradation, Steroid hormone biosynthesis	11	EC:1.3.99.5 - 4-dehydrogenase (acceptor), EC:1.1.1.145 - dehydrogenase, EC:1.14.15.4 - 11beta-monooxygenase, EC:1.14.15.5 - 18-monooxygenase, EC:2.8.2.4 - sulfotransferase, EC:2.1.1.6 - O-methyltransferase, EC:1.14.14.1 - monooxygenase, EC:1.1.1.62 - 17-dehydrogenase, EC:1.1.1.239 - dehydrogenase (NAD+), EC:1.1.1.145 - dehydrogenase, EC:2.4.1.17 - 1-naphthol glucuronyltransferase
2	^a TF2	5084595	2	4.5	4.1	7.3	3.7E-23	9	F:GO:0000166; F:GO:0003824; P:GO:0007275; P:GO:0006629; P:GO:0030154; F:GO:0008289; P:GO:0009058; C:GO:0005622; F:GO:0005215	F:nucleotide binding; F:catalytic activity; P:multicellular organism development; P:lipid metabolic process; P:cell differentiation; F:lipid binding; P:biosynthetic process; C:intracellular; F:transporter activity	Steroid degradation, Steroid hormone biosynthesis	11	EC:1.3.99.5 - 4-dehydrogenase (acceptor), EC:1.1.1.145 - dehydrogenase, EC:1.14.15.4 - 11beta-monooxygenase, EC:1.14.15.5 - 18-monooxygenase, EC:2.8.2.4 - sulfotransferase, EC:2.1.1.6 - O-methyltransferase, EC:1.14.14.1 - monooxygenase, EC:1.1.1.62 - 17-dehydrogenase, EC:1.1.1.239 - dehydrogenase (NAD+), EC:1.1.1.145 - dehydrogenase, EC:2.4.1.17 - 1-naphthol glucuronyltransferase
3	^a EGC	5124128	2	7.7	3.3	5.6	2.2E-18	1	F:GO:0005198	F:structural molecule activity	-	-	-
4	^a ECG	5098382	6	57	5.2	8.7	5.5E-19	8	F:GO:0003723; F:GO:0005515; P:GO:0007275; P:GO:0009653; C:GO:0005730; P:GO:0006139; C:GO:0005654; F:GO:0016787	F:RNA binding; F:protein binding; P:multicellular organism development; P:anatomical structure morphogenesis; C:nucleolus; P:nucleobase-containing compound metabolic process; C:nucleoplasm; F:hydrolase activity	Purine metabolism	15	EC:4.6.1.1 - cyclase, EC:4.6.1.2 - cyclase, EC:2.7.4.6 - kinase, EC:2.7.4.3 - kinase, EC:3.6.1.3 - adenylypyrophosphatase, EC:3.1.4.17 - phosphodiesterase, EC:3.5.4.4 - deaminase, EC:2.7.1.4- - kinase, EC:3.5.4.6 - deaminase, EC:3.1.4.35 - phosphodiesterase, EC:2.7.1.2- - kinase,

EC:3.6.1.15 - phosphatase,
 EC:5.4.2.2 - (alpha-D-glucose-1,6-
 bisphosphate-dependent),
 EC:2.7.6.1 - diphosphokinase,
 EC:2.4.2.8 - phosphoribosyltransferase

5	^a TF1	5136045	6	69.6	4.3	7.7	5.5E-19	1	F:GO:0005515	F:protein binding	-	-	-
6	^a CAFF	5111497	8	18.8	4.1	7.1	3.7E-23	4	P:GO:0008150; F:GO:0005515; C:GO:0005737; C:GO:0005576	P:biological_process; F:protein binding; C:cytoplasm; C:extracellular region	-	-	-
7	^a CAT	5103784	13	50.6	3.8	6.1	5.5E-19	14	P:GO:0007275; P:GO:0007165; P:GO:0008152; P:GO:0007267; C:GO:0005886; P:GO:0007049; C:GO:0005635; C:GO:0005856; F:GO:0005515; P:GO:0008219; C:GO:0005737; P:GO:0009719; P:GO:0016043; P:GO:0030154	P:multicellular organism development; P:signal transduction; P:metabolic process; P:cell-cell signalling; C:plasma membrane; P:cell cycle; C:nuclear envelope; C:cytoskeleton; F:protein binding; P:cell death; C:cytoplasm; P:response to endogenous stimulus; P:cellular component organisation; P:cell differentiation	-	-	-
8	^a CAT	5122899	13	50.5	3.8	6	3.7E-23	5	F:GO:0005515; P:GO:0007275; P:GO:0030154; C:GO:0005886; F:GO:0005215	F:protein binding; P:multicellular organism development; P:cell differentiation; C:plasma membrane; F:transporter activity	-	-	-
9	^a ECG	5123761	13	60.7	3.3	5.8	3.6E-20	2	F:GO:0003700; C:GO:0005622	F:transcription factor activity, sequence- specific DNA binding; C:intracellular	-	-	-
10	^a BRT	5135810	13	26	2.8	5.8	1.3E-13	4	C:GO:0005623; P:GO:0019538; F:GO:0005515; F:GO:0016787	C:cell; P:protein metabolic process; F:protein binding; F:hydrolase activity	-	-	-

11	^a CAFF	5054639	14	6.4	3.9	6.9	5.0E-10	12	C:GO:0005615; P:GO:0009628; P:GO:0008219; P:GO:0007275; C:GO:0005737; P:GO:0009719; P:GO:0007267; P:GO:0016043; P:GO:0030154; P:GO:0009653; F:GO:0005102; P:GO:0006950	C:extracellular space; P:response to abiotic stimulus; P:cell death; P:multicellular organism development; C:cytoplasm; P:response to endogenous stimulus; P:cell-cell signalling; P:cellular component organisation; P:cell differentiation; P:anatomical structure morphogenesis; F:receptor binding; P:response to stress	-	-	-
12	^a CAT	5132370	14	60.7	6.5	10.8	1.5E-25	26	P:GO:0007275; P:GO:0009991; P:GO:0006464; F:GO:0008289; P:GO:0009653; P:GO:0009058; P:GO:0009607; P:GO:0009628; F:GO:0003723; P:GO:0008219; P:GO:0007165; P:GO:0006259; F:GO:0005102; P:GO:0006810; P:GO:0006950; P:GO:0000003; P:GO:0007049; P:GO:0040007; C:GO:0005634; F:GO:0003677; C:GO:0005615; C:GO:0005737; P:GO:0009719; P:GO:0016043; P:GO:0030154;	P:multicellular organism development; P:response to extracellular stimulus; P:cellular protein modification process; F:lipid binding; P:anatomical structure morphogenesis; P:biosynthetic process; P:response to biotic stimulus; P:response to abiotic stimulus; F:RNA binding; P:cell death; P:signal transduction; P:DNA metabolic process; F:receptor binding; P:transport; P:response to stress; P:reproduction; P:cell cycle; P:growth;	-	-	-

									F:GO:0016740	C:nucleus; F:DNA binding; C:extracellular space; C:cytoplasm; P:response to endogenous stimulus; P:cellular component organisation; P:cell differentiation; F:transferase activity			
13	^a EGC	5132370	14	60.7	3.7	6.4	1.5E-25	26	P:GO:0007275; P:GO:0009991; P:GO:0006464; F:GO:0008289; P:GO:0009653; P:GO:0009058; P:GO:0009607; P:GO:0009628; F:GO:0003723; P:GO:0008219; P:GO:0007165; P:GO:0006259; F:GO:0005102; P:GO:0006810; P:GO:0006950; P:GO:0000003; P:GO:0007049; P:GO:0040007; C:GO:0005634; F:GO:0003677; C:GO:0005615; C:GO:0005737; P:GO:0009719; P:GO:0016043; P:GO:0030154; F:GO:0016740	P:multicellular organism development; P:response to extracellular stimulus; P:cellular protein modification process; F:lipid binding; P:anatomical structure morphogenesis; P:biosynthetic process; P:response to biotic stimulus; P:response to abiotic stimulus; F:RNA binding; P:cell death; P:signal transduction; P:DNA metabolic process; F:receptor binding; P:transport; P:response to stress; P:reproduction; P:cell cycle; P:growth; C:nucleus; F:DNA binding; C:extracellular space; C:cytoplasm; P:response to endogenous stimulus; P:cellular component	-	-	-

14	^a EGCG	5114089	15	32.1	4	6.8	8.0E-12	5	C:GO:0005634; F:GO:0003700; P:GO:0007275; P:GO:0009058; P:GO:0006139	organisation; P:cell differentiation; F:transferase activity C:nucleus; F:transcription factor activity, sequence- specific DNA binding; P:multicellular organism development; P:biosynthetic process; P:nucleobase-containing compound metabolic process	-	-	-
15	^b TF1	5084595	2	4.5	3.2	5.7	3.7E-23	9	F:GO:0000166; F:GO:0003824; P:GO:0007275; P:GO:0006629; P:GO:0030154; F:GO:0008289; P:GO:0009058; C:GO:0005622; F:GO:0005215	F:nucleotide binding; F:catalytic activity; P:multicellular organism development; P:lipid metabolic process; P:cell differentiation; F:lipid binding; P:biosynthetic process; C:intracellular; F:transporter activity	Steroid degradation, Steroid hormone biosynthesis	11	EC:1.3.99.5 - 4-dehydrogenase (acceptor), EC:1.1.1.145 - dehydrogenase, EC:1.14.15.4 - 11beta-monooxygenase, EC:1.14.15.5 - 18-monooxygenase, EC:2.8.2.4 - sulfotransferase, EC:2.1.1.6 - O-methyltransferase, EC:1.14.14.1 - monooxygenase, EC:1.1.1.62 - 17-dehydrogenase, EC:1.1.1.239 - dehydrogenase (NAD+), EC:1.1.1.145 - dehydrogenase, EC:2.4.1.17 - 1-naphthol glucuronyltransferase
16	^b TF2	5084595	2	4.5	4.1	7.3	3.7E-23	9	F:GO:0000166; F:GO:0003824; P:GO:0007275; P:GO:0006629; P:GO:0030154; F:GO:0008289; P:GO:0009058; C:GO:0005622; F:GO:0005215	F:nucleotide binding; F:catalytic activity; P:multicellular organism development; P:lipid metabolic process; P:cell differentiation; F:lipid binding; P:biosynthetic process; C:intracellular; F:transporter activity	Steroid degradation, Steroid hormone biosynthesis	11	EC:1.3.99.5 - 4-dehydrogenase (acceptor), EC:1.1.1.145 - dehydrogenase, EC:1.14.15.4 - 11beta-monooxygenase, EC:1.14.15.5 - 18-monooxygenase, EC:2.8.2.4 - sulfotransferase, EC:2.1.1.6 - O-methyltransferase, EC:1.14.14.1 - monooxygenase, EC:1.1.1.62 - 17-dehydrogenase, EC:1.1.1.239 - dehydrogenase (NAD+), EC:1.1.1.145 - dehydrogenase, EC:2.4.1.17 - 1-naphthol glucuronyltransferase
17	^b EGC	5124128	2	7.7	3.3	5.6	2.2E-18	1	F:GO:0005198	F:structural molecule activity	-	-	-
18	^b ECG	5098382	6	57	5.2	8.7	5.5E-19	8	F:GO:0003723;	F:RNA binding;	Purine	15	EC:4.6.1.1 - cyclase,

									F:GO:0005515; P:GO:0007275; P:GO:0009653; C:GO:0005730; P:GO:0006139; C:GO:0005654; F:GO:0016787	F:protein binding; P:multicellular organism development; P:anatomical structure morphogenesis; C:nucleolus; P:nucleobase-containing compound metabolic process; C:nucleoplasm; F:hydrolase activity	metabolism		EC:4.6.1.2 - cyclase, EC:2.7.4.6 - kinase, EC:2.7.4.3 - kinase, EC:3.6.1.3 - adenylypyrophosphatase, EC:3.1.4.17 - phosphodiesterase, EC:3.5.4.4 - deaminase, EC:2.7.1.4- - kinase, EC:3.5.4.6 - deaminase, EC:3.1.4.35 - phosphodiesterase, EC:2.7.1.2- - kinase, EC:3.6.1.15 - phosphatase, EC:5.4.2.2 - (alpha-D-glucose-1,6- bisphosphate-dependent), EC:2.7.6.1 - diphosphokinase, EC:2.4.2.8 - phosphoribosyltransferase
19	^b CAT	5132370	14	60.7	6.5	10.8	1.5E-25	26	P:GO:0007275; P:GO:0009991; P:GO:0006464; F:GO:0008289; P:GO:0009653; P:GO:0009058; P:GO:0009607; P:GO:0009628; F:GO:0003723; P:GO:0008219; P:GO:0007165; P:GO:0006259; F:GO:0005102; P:GO:0006810; P:GO:0006950; P:GO:0000003; P:GO:0007049; P:GO:0040007; C:GO:0005634; F:GO:0003677; C:GO:0005615; C:GO:0005737; P:GO:0009719; P:GO:0016043; P:GO:0030154; F:GO:0016740	P:multicellular organism development; P:response to extracellular stimulus; P:cellular protein modification process; F:lipid binding; P:anatomical structure morphogenesis; P:biosynthetic process; P:response to biotic stimulus; P:response to abiotic stimulus; F:RNA binding; P:cell death; P:signal transduction; P:DNA metabolic process; F:receptor binding; P:transport; P:response to stress; P:reproduction; P:cell cycle; P:growth; C:nucleus; F:DNA binding;	-	-	-

20	^b EGCG	5114089	15	32.1	4	6.8	8.0E-12	5	C:GO:0005634; F:GO:0003700; P:GO:0007275; P:GO:0009058; P:GO:0006139	C:extracellular space; C:cytoplasm; P:response to endogenous stimulus; P:cellular component organisation; P:cell differentiation; F:transferase activity C:nucleus; F:transcription factor activity, sequence- specific DNA binding; P:multicellular organism development; P:biosynthetic process; P:nucleobase-containing compound metabolic process	-	-	-

^a Putative QTL identified in Interval Mapping using MapQTL 6.0

^b Putative QTL identified in Multiple QTL Model Mapping using MapQTL 6.0

P (biological process); F (molecular function); C (cellular component)

(-) Not found in KEGG pathway

Appendix 4.13: List of unigenes from primary and secondary metabolic pathways in 15 LGs of *C. sinensis* transcriptome

Nr	Pathway	#Enzs in Pathway	Enzyme	#Unigene	Unigene ID
1	Alanine, aspartate and glutamate metabolism	4	EC:4.1.1.15 - decarboxylase, EC:6.3.4.16 - synthase (ammonia), EC:1.4.1.2 - dehydrogenase, EC:2.6.1.44 - transaminase	7	5134557, 5082783, 5134557, 5082783, 5134840, 5098250, 5125548
2	Amino sugar and nucleotide sugar metabolism	3	EC:5.4.2.2 - (alpha-D-glucose-1,6-bisphosphate-dependent), EC:2.7.1.2 - glucokinase (phosphorylating), EC:2.7.1.1 - hexokinase type IV glucokinase	4	5122964, 5077455, 5134772, 5134772
3	Arginine and proline metabolism	4	EC:4.1.1.50 - decarboxylase, EC:1.14.13.39 - synthase (NADPH), EC:3.5.3.1 - arginine amidinase, EC:1.2.1.3 - dehydrogenase (NAD ⁺)	5	5124570, 5133545, 5124506, 5109744, 5123629
4	Arginine biosynthesis	6	EC:1.14.13.39 - synthase (NADPH), EC:3.5.3.1 - arginine amidinase, EC:6.3.4.16 - synthase (ammonia), EC:2.1.3.3 - carbamoyltransferase, EC:1.4.1.2 - dehydrogenase, EC:3.5.1.14 - acid amidohydrolase	8	5133545, 5124506, 5109744, 5134557, 5082783, 5098949, 5134840, 5080820
5	Ascorbate and aldarate metabolism	2	EC:2.4.1.17 - 1-naphthol glucuronyltransferase, EC:1.2.1.3 - dehydrogenase (NAD ⁺)	3	5075058, 5084636, 5123629
6	Carbon fixation in photosynthetic organisms	2	EC:4.1.2.13 - aldolase, EC:1.1.1.37 - dehydrogenase	5	5106931, 5084452, 5133604, 5124439, 5132470
7	Cysteine and methionine metabolism	5	EC:4.1.1.50 - decarboxylase, EC:1.1.1.37 - dehydrogenase, EC:4.2.1.22 - beta-synthase, EC:1.1.1.27 - dehydrogenase, EC:2.6.1.44 - transaminase	9	5124570, 5133604, 5124439, 5132470, 5112205, 5123470, 5133941, 5098250, 5125548
8	Flavone and flavonol biosynthesis	1	EC:3.2.1.31 - beta-glucuronide glucuronohydrolase glucuronidase	2	5072939, 5068701
9	Fructose and mannose metabolism	4	EC:4.1.2.13 - aldolase, EC:3.2.1.78 - endo-1,4-beta-mannosidase, EC:1.1.1.21 - reductase,	6	5106931, 5084452, 5059558, 5127501, 5123251, 5134772

			EC:2.7.1.1 - hexokinase type IV glucokinase		
10	Glycerolipid metabolism	5	EC:2.7.1.107 - kinase (ATP), EC:1.1.1.2 - dehydrogenase (NADP+), EC:1.1.1.21 - reductase, EC:3.1.1.3 - lipase, EC:1.2.1.3 - dehydrogenase (NAD+)	6	5137030, 5123251, 5123251, 5134984, 5065613, 5123629
11	Glycine, serine and threonine metabolism	3	EC:4.2.1.22 - beta-synthase, EC:2.6.1.51 - transaminase, EC:2.6.1.44 - transaminase	5	5112205, 5098250, 5125548, 5098250, 5125548
12	Glyoxylate and dicarboxylate metabolism	3	EC:6.4.1.3 - carboxylase, EC:2.3.1.9 - C-acetyltransferase, EC:1.1.1.37 - dehydrogenase	5	5122912, 5124564, 5133604, 5124439, 5132470
13	Histidine metabolism	1	EC:1.2.1.3 - dehydrogenase (NAD+)	1	5123629
	mTOR signalling pathway	1	EC:2.7.11.24 - protein kinase	1	5132197
	Nitrogen metabolism	2	EC:6.3.4.16 - synthase (ammonia), EC:1.4.1.2 - dehydrogenase	3	5134557, 5082783, 5134840
14	Other glycan degradation	2	EC:3.2.1.25 - mannanase, EC:3.2.1.18 - neuraminidase	3	5059558, 5127501, 5055477
15	Pentose and glucuronate interconversions	4	EC:1.1.1.2 - dehydrogenase (NADP+), EC:3.2.1.31 - beta-glucuronide glucuronohydrolase glucuronidase, EC:1.1.1.21 - reductase, EC:2.4.1.17 - 1-naphthol glucuronyltransferase	6	5123251, 5072939, 5068701, 5123251, 5075058, 5084636
16	Phosphatidylinositol signalling system	2	EC:2.7.1.107 - kinase (ATP), EC:3.1.4.11 - phospholipase C	1	5137030, 5122970
17	Porphyrin and chlorophyll metabolism	3	EC:4.2.1.24 - synthase, EC:3.2.1.31 - beta-glucuronide glucuronohydrolase glucuronidase, EC:2.4.1.17 - 1-naphthol glucuronyltransferase	3	5132295, 5072939, 5068701, 5075058, 5084636
18	Purine metabolism	6	EC:3.6.1.15 - phosphatase, EC:2.7.4.3 - kinase, EC:3.6.1.3 - adenylypyrophosphatase, EC:5.4.2.2 - (alpha-D-glucose-1,6-bisphosphate-dependent), EC:3.1.4.17 - phosphodiesterase,	22	5123251, 5090210, 5133369, 5125437, 5082614, 5118156, 5105900, 5078215, 5134705, 5137364, 5113142, 5136410, 5125437, 5090210, 5133369, 5105900, 5078215, 5122964, 5077455, 5123174,

			EC:3.5.4.4 - deaminase		5075058, 5098382
19	Pyruvate metabolism	5	EC:2.3.1.9 - C-acetyltransferase, EC:1.1.1.37 - dehydrogenase, EC:4.4.1.5 - lyase, EC:1.1.1.27 - dehydrogenase, EC:1.2.1.3 - dehydrogenase (NAD+)	8	5124564, 5133604, 5124439, 5132470, 5136472, 5123470, 5133941, 5123629
20	Terpenoid backbone biosynthesis	3	EC:2.3.3.10 - synthase, EC:2.3.1.9 - C-acetyltransferase, EC:2.5.1.29 - diphosphate synthase	4	5125140, 5124564, 5096779, 5072048
21	Thiamine metabolism	2	EC:3.6.1.15 - phosphatase, EC:2.7.4.3 - kinase	12	5123251, 5090210, 5133369, 5125437, 5082614, 5118156, 5105900, 5078215, 5134705, 5137364, 5113142, 5136410
22	Tryptophan metabolism	2	EC:2.3.1.9 - C-acetyltransferase, EC:1.2.1.3 - dehydrogenase (NAD+)	2	5124564, 5123629
23	Tyrosine metabolism	1	EC:2.1.1.6 - O-methyltransferase	2	5059558, 5081364
24	Ubiquinone and other terpenoid-quinone biosynthesis	1	EC:1.6.5.2 - dehydrogenase (quinone)	1	5107213
25	Valine, leucine and isoleucine degradation	5	EC:2.3.3.10 - synthase, EC:6.4.1.3 - carboxylase, EC:2.3.1.9 - C-acetyltransferase, EC:2.3.1.16 - C-acyltransferase, EC:1.2.1.3 - dehydrogenase (NAD+)	5	5125140, 5122912, 5124564, 5124564, 5123629

SCIENTIFIC REPORTS

OPEN

Functional annotation of putative QTL associated with black tea quality and drought tolerance traits

Received: 14 June 2018
 Accepted: 12 December 2018
 Published online: 06 February 2019

Robert. K. Koehn^{1,2}, Pelly M. Malebe¹, Christopher Nyarukowa¹, Richard Mose³,
 Samson M. Kamunya², Fourie Joubert¹ & Zeno Apostolides¹

The understanding of black tea quality and percent relative water content (%RWC) traits in tea (*Camellia sinensis*) by a quantitative trait loci (QTL) approach can be useful in elucidation and identification of candidate genes underlying the QTL which has remained to be difficult. The objective of the study was to identify putative QTL controlling black tea quality and percent relative water traits in two tea populations and their F1 progeny. A total of 1,421 DArTseq markers derived from the linkage map identified 53 DArTseq markers to be linked to black tea quality and %RWC. All 53 DArTseq markers with unique best hits were identified in the tea genome. A total of 5,592 unigenes were assigned gene ontology (GO) terms, 56% comprised biological processes, cellular component (29%) and molecular functions (15%), respectively. A total of 84 unigenes in 15 LGs were assigned to 25 different Kyoto Encyclopedia of Genes and Genomes (KEGG) database pathways based on categories of secondary metabolite biosynthesis. The three major enzymes identified were transferases (38.9%), hydrolases (29%) and oxidoreductases (18.3%). The putative candidate proteins identified were involved in flavonoid biosynthesis, alkaloid biosynthesis, ATPase family proteins related to abiotic/biotic stress response. The functional annotation of putative QTL identified in this current study will shed more light on the proteins associated with caffeine and catechins biosynthesis and % RWC. This study may help breeders in selection of parents with desirable DArTseq markers for development of new tea cultivars with desirable traits.

The identification of putative candidate genes underlying QTLs in plants is often complicated and time-consuming. This is even further complicated by genotype-environment interactions, which influence the growth and development of plants. However, using recently unveiled tea genome sequence¹, it was possible to annotate candidate genes and enzymes involved in the biosynthesis of phenolic compounds and drought tolerance. The phenolic compounds in tea are closely associated to the classical flavour of tea infusions and to the health benefits ascribed to tea consumption^{2,3}. Catechins (flavan-3-ols), flavonols, and their derivatives (galloylated catechins and flavanol 3-O-glycosides) are among the major biomolecules in tea with important bioactivities. The polyphenols in tea have been shown to contribute to the astringency in black tea infusions⁴. Especially, galloylated catechins have been found to confer astringent and bitter tastes in tea⁵ while flavanol 3-O-glycosides have been found to confer velvety, mouth-drying, and mouth-coating sensations⁶. During the production of black tea, the enzymatic oxidation of the leaf catechins by polyphenol oxidase enzyme leads to the formation of theaflavins (TF) and thearubigins (TR)⁷. Theaflavins which are responsible for the taste, brightness and contribute to the colour of black teas. Thearubigins are responsible for thickness and colour of both the liquors and infusion⁸. The other plain tea quality parameters or attributes, include black tea brightness, astringency, colour, briskness and aroma⁹. Indeed, theaflavins have become a critical parameter in estimating the quality and the price of black teas¹⁰. Caffeine is important in cream formation and the briskness of tea and therefore, has been used in predicting black tea quality^{11,12}. Polyphenols and caffeine biosynthesis are dependent on structural and regulatory genes; structural genes encode enzymes catalyzing the biosynthesis, while regulatory genes control the expression of the genes.

¹Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, 0002, South Africa. ²Kenya Agriculture and Livestock Research Organization, Tea Research Institute, P.O. Box 820, Kericho, 20200, Kenya. ³James Finlay (Kenya) Limited, P.O. Box 223, Kericho, 20200, Kenya. Correspondence and requests for materials should be addressed to Z.A. (email: za@up.ac.za)

Drought is a major factor which significantly influences the growth and production of tea^{13,14}. Hence, breeding drought-tolerant tea cultivars that have high recoverability potential is an important target for tea breeders. Although, the morphological and physiological mechanisms of drought tolerance in tea plant has been reported, only few reports are available on the changes of genes related to catechin biosynthesis in response to environmental stresses^{15,16}. This is because drought tolerance is a complex multigenic trait, and various regulatory and functional genes are involved. In tea plant¹⁷, reported that polyphenols can be used as a potential indicator for drought tolerance. In another study, results suggested epicatechin and epigallocatechin as potential indicators in predicting drought tolerance in tea plant¹⁸. Recently, transcriptomic analysis of drought susceptible and drought-tolerant tea plants during drought stress has been reported^{15,19,20}. The molecular responses of plants to drought stress include perception, signal transduction, gene expression, and ultimately metabolic changes that lead to stress tolerance. The induced drought stress lead to gene expression which participate in the generation of hormones like abscisic acid (ABA), salicylic acid, and ethylene. For example, ABA has been shown to play crucial roles in regulating the drought response in many plants, and its metabolic pathway involves multiple steps and genes²¹. The level of ABA in plants increases rapidly during drought stress, leading to stomatal closure and induction of stress genes to cope with the stress²². Drought stress also results in accumulation of proline, mannitol, sorbitol, formation of radical scavenging compounds (ascorbate, glutathione, tocopherol), and synthesis of protein^{23,24}. Proline acts as an osmoprotectant and its accumulation is induced to tolerance of tea to water deficit stress, as this helps in maintaining water relations, prevents membrane distortion, and also acts as a hydroxyl radical scavenger²⁵. The increase in generation of reactive oxygen species (ROS) such as superoxide radicals (O₂), hydrogen peroxide (H₂O₂), hydroxyl radical (.OH), alkoxyl radical (.RO), and singlet oxygen (1O₂) within the cell, leads to lipid peroxidation, protein degradation, and enzyme inactivation and affects nucleic acid leading to cell death^{26,27}. The objective of the current study was to annotate and assign biosynthetic pathway functions to putative QTLs linked to caffeine content, catechins fractions, theaflavins fractions, tea taster scores and %RWC.

Materials and Methods

Preparation of FASTA files. The sequences of DArTseq-based markers were derived from the tag sequence associated with each DArTseq marker as explained in the previous work⁹ and were generated by Diversity Arrays Technology Pty. Ltd. (Canberra, Australia). Marker sequences were arranged in FASTA format by starting with a single-line description of the sequence, followed by sequence data. The single-line description was distinguished from the sequence data by placing a symbol ">" in front of the description.

BLAST search. A total of 1,421 DArTseq markers on LG1 to LG15 were subjected to a BLAST search. The marker tag sequences (FASTA format) derived from the DArTseq map were searched using the non-redundant BLASTN program against the assembled draft tea genome (*Camellia sinensis* var *assamica*)¹. The best hit was then selected based on E-value and % identity. The E-value served as a measure of the number of hits one can "expect" to see by chance when searching a database with small E-value indicating homology. The % identity showed the percentage of identical residues between query sequences and hit sequences from a database, with longer stretches of homology more likely to indicate a genuine match.

Functional annotation and pathway assignment. Functional annotation of all 1,421 DArTseq markers was performed on BLASTX search against the non-redundant GenBank protein sequence database with a threshold E-value of 10⁻⁶^{25,26}. Functional annotation and mapping of gene ontology (GO) terms was done with the Blast2GO program (Blast2GO v3.2)²⁷. Each BLAST hit associated with the GO terms were retrieved and annotated using the following annotation score parameters; E-Value Hit Filter (default = 1.0E-6), GO-Weight (default = 5), Hsp-Hit Coverage Cut Off (default = 0), Annotation Cut-Off (90). The contig sequences were also queried for conserved domains/motifs using InterProScan, which is an on-line sequence search plug-in within the Blast2GO program. The Kyoto Encyclopedia of Genes and Genomes (KEGG) mapping was used to determine metabolic pathways. The sequences with corresponding evidence code (EC) numbers obtained from Blast2GO were mapped to the KEGG metabolic pathway database.

Results

Functional annotation and pathway assignment. The identification of candidate genes for tea was conducted in all the 15 LGs derived from the DArTseq maps. Sequences of each marker on all the LGs were subjected to a BLAST search against tea genome BLAST database in an effort to compare the homology and identify the location of the gene of interest within the tea genome. The results of BLAST searches for markers derived from the DArTseq map are presented in Table 1. Of 1,421 DArTseq markers derived from DArTseq map, 53 markers were identified to be linked to black tea quality and %RWC. A total of seven markers were dominant in TRFK 303/577 parent while 9 markers were dominant in GW Ejulu parent, respectively. All the 53 markers with unique best hits were identified in the tea genome. The 23 markers showed locations across ten chromosomes in tea, which were LG1 (1), LG3 (1), LG4 (8), LG6 (1), LG9 (1), LG10 (2), LG12 (2), LG13 (4), LG14 (2) and LG15 (1). A total of 14 markers were dominant, 8 were dominant in GW Ejulu and six were dominant in TRFK 303/577, respectively. The dominant markers in GW Ejulu were for caffeine, catechin, ECG, EGC and EGCg. On the other hand, the dominant markers in TRFK 303/577 were for caffeine, ECG, EGC, TF1 and %RWC. All the 13 dominant markers were identified in the tea genome for the two parents used in this study as reciprocal crosses in the two tea populations (Table 1).

Annotation and mapping of gene ontology. All unigenes in this study were assigned GO terms based on BLASTX searches against the non-redundant (NR) databases (Fig. 1 and 2). Our results showed that 86–100% similarity distribution of the top hits in the non-redundant (NR) database (Fig. 1). The E-value distribution of the mapped sequences with high homologies (smaller than 1e-6) was high as compared to homologous sequences

Nr	QTL	Parent	DA/Tseq Marker	LG	Position (cM)	LOD	PEV	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	qECG ^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	qECG ^a	GW Ejulu	5134853	4	37.6	14.9	45.1	2.0E-06	[Diacylglycerol kinase catalytic domain]	Abiotic stress
9	qTFI ^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	qECG ^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	qECG ^a	TRFK 303/577	5119221	4	20.6	3.3	53.8	1.0E-19	[Histone acetyltransferase subunit NuA4]	Drought response
14	qECG ^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	[BTI 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	qTFI ^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	[Dna] 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	qECG ^a	TRFK 303/577	5088456	12	47.9	3.9	9.8	4.0E-23	[Protein kinase domain]	Abiotic stress
31	qCAT ^b	GW Ejulu	5136077	12	42.8	6.6	11.2	5.0E-13	CSA026168	—
32	qECG ^a	GW Ejulu	5136077	12	42.8	5.6	11.2	5.0E-13	CSA026168	—
33	qCAT ^b	TRFK 303/577, GW Ejulu	5123751	12	43.0	6.1	11.5	2.0E-18	[Bromodomain]	Scaffolding proteins
34	qECG ^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
Continued										

Nr	QTL	Parent	DArTseq Marker	LG	Position (cM)	LOD	PEV	E-value	Annotated protein	Function
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	qEGC ^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	[Glutamyl-tRNA synthetase]	Chlorophyll biosynthesis
52	qEGC ^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

Table 1. Functional annotation of putative candidate genes in 15 linkage groups of *C. sinensis* on reference tea genome. ^aPutative QTL identified based on Interval Mapping with LOD >3.0. ^bPutative QTL identified based on Multiple QTL Model Mapping with LOD >3.0. QTL - Quantitative trait loci; LG - Linkage group; LOD - Logarithm of odd ratio; cM - Centimorgan, CAFF Caffeine, CAT Catechin, EC epicatechin, ECG epicatechin gallate, EGC epigallocatechin, EGCG epigallocatechin gallate, TF1 theaflavin, TF4- Theaflavin-3,3'-digallate, BRT brightness, %RWC percent relative water content.

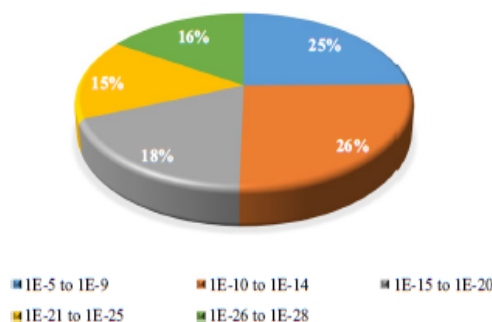


Figure 1. E-value distribution of the best matches for unigenes (E-value $\leq 1.0 \times 10^{-5}$).

that were greater than 1e-5 (Fig. 2). The E-value distribution of the best hits in the NR database showed that 26% annotated sequences had high identity with their best hits (smaller than between 1e-28 to 1e-26), whereas 15% ranged from 1e-21 to 1e-25 and another 18% ranged from 1e-15 to 1e-20 (Fig. 1). The similarity distribution revealed that 43% of the query sequences had greater than 91–95% similarity, whereas 21% of the query sequences varied in similarity from 86–90%, and the other 3% varied from 81–85% (Fig. 2).

A total of 5,592 unigenes were assigned GO terms, of these, biological processes (56%) comprised the largest category, followed by cellular component (29%) and molecular functions (15%), respectively (Fig. 3). In biological process (BP) category, the majority of unigenes were involved in regulation of biological process (35%), cellular process (16%), metabolic process (12%), response to stimulus (11%), developmental process (9%) and localization and signaling (6%), respectively (Fig. 3). The cellular and metabolic processes were the most highly represented groups in the biological process, which suggests that extensive metabolic activities taking place in the fresh tea flush. The best-represented groups of molecular function (MF) were binding activity (48%), catalytic activity (27%), signal transducer activity (10%), molecular transducer activity (8%) and transporter activity (7%), respectively (Fig. 3). Within the molecular function category, binding and catalytic activities were the most abundant groups. The cellular component (CC) included the cell (21%), cell part (21%), organelle (16%), membrane (11%) organelle part (7%), membrane part (7%), and macromolecular complex (6%), respectively (Fig. 3). The transcripts that corresponded to cell, cell parts, and cell organelles were the most abundant groups within the cellular component category. Among the top 20 functional sub-groups in biological processes, multicellular organism development, response to stress and response to endogenous stimulus represented the largest group (Supplementary Fig. 1). The top 20 functional sub-groups for molecular functions, protein binding, receptor

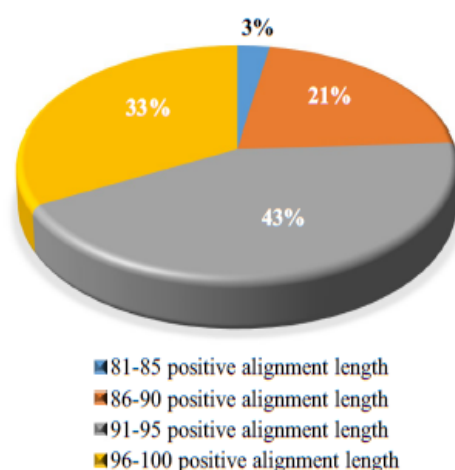


Figure 2. Similarity distribution of the hit BLAST matches for unigenes.

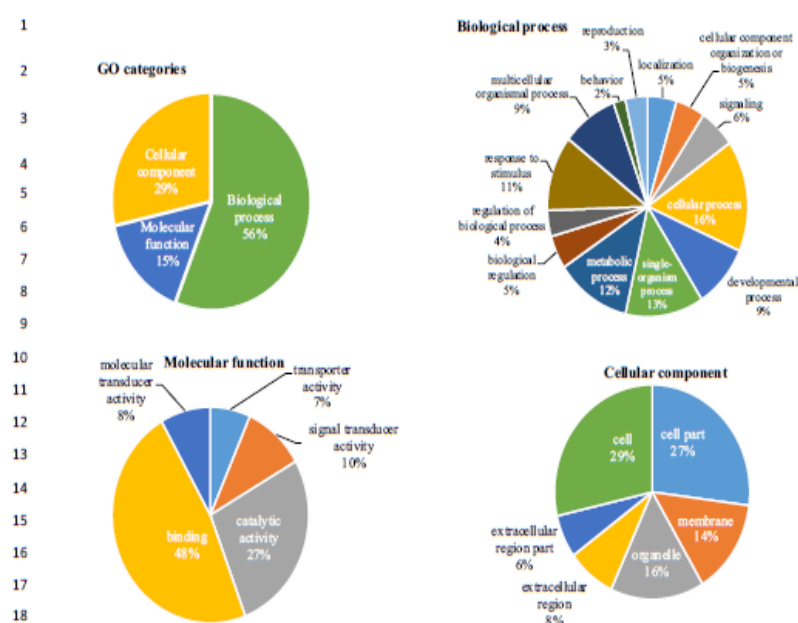


Figure 3. Function classifications of GO terms of *C. sinensis* transcripts based on high-score BLASTX matches in NR plant proteins database, a total of 5,592 unigenes were classified into three main GO categories which are biological process (BP), molecular function (MF) and cellular component (CC), respectively.

binding and signal transducer activity represented the largest group (Supplementary Fig. 2), while plasma membrane, nucleus and cytoplasm were the largest group that represented the top 20 sub-groups of cellular component (Supplementary Fig. 3).

KEGG pathway mapping: Flavonoid pathway, caffeine pathway and abiotic stress. Pathway-based analysis for the tea flush transcriptome is helpful to further understand the biological functions of genes involved in the tea flavonoid biosynthesis pathway, caffeine pathway and stimulus to response to abiotic stress (drought stress). A total of 53 unigenes linked to putative QTL were assigned two different KEGG database pathways (Supplementary Table 1) while 84 unigenes in 15 linkage groups were assigned to 25 different KEGG database

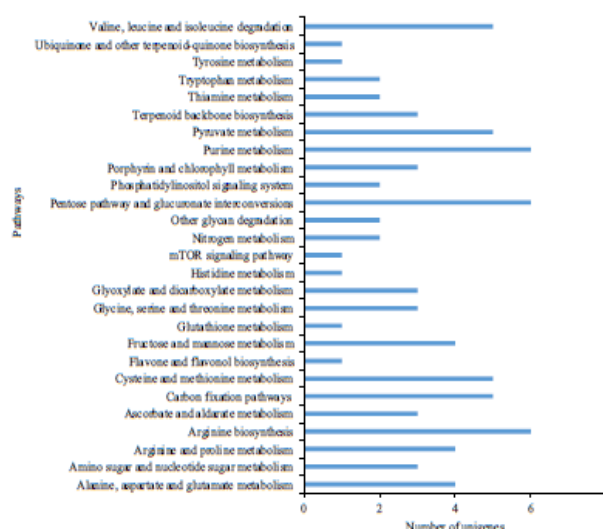


Figure 4. KEGG ontology (KO) analysis of differentially expressed genes (DEGs) related to secondary metabolism in *C. sinensis*.

pathways based on categories of secondary metabolite biosynthesis (Supplementary Table 2, Fig. 4). The pathways with most representations were metabolic pathways and biosynthesis of secondary metabolites, which was consistent with the GO categories of the biological process (Fig. 3 and Supplementary Fig. 1). Of these, 52 unigenes were categorized into the metabolism groups, with most of them involved in purine metabolism (11%), cysteine and methionine metabolism (9.6%), pyruvate metabolism (9.6%), alanine, aspartate and glutamate metabolism (7.6%), arginine and proline metabolism (7.6%) and fructose and mannose metabolism (7.6%) and other sub-categories, respectively. The most abundant unigenes were for carbohydrate and amino acids biosynthesis, which are mostly involved in plant hormone signal transduction pathways in response to abiotic stress and biosynthesis of other secondary metabolites such as phenylpropanoid biosynthesis, flavonoid biosynthesis and alkaloid biosynthesis. The unigenes associated with flavone and flavonol biosynthesis, a key regulatory pathways in the metabolism of the main chemical components in *C. sinensis* leaves were present (Fig. 4). Unigenes associated with purine metabolism, also a key regulatory pathway which is involved in metabolism of adenine and guanosine nucleotides were present.

Also, major enzymes that are involved in different metabolic pathways were further grouped into the six different categories; oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The distribution of the three abundant enzymes was also consistent with the GO categories of molecular function (Fig. 3 and Supplementary Fig. 2). The three major enzymes distribution included; transferase enzymes (38.9%), which are a class of enzymes that catalyze the transfer of specific functional groups (a methyl, or glycosyl group, or galloyl group) from a donor to acceptor molecules, hydrolase enzymes (29%), which catalyze the hydrolysis of compound and oxidoreductase enzymes (18.3%), which catalyze the oxidation-reduction reactions (Fig. 5 and Supplementary Fig. 3). The three other enzymes that were less abundant in distribution included; lyase enzymes (8.4%), which cleave carbon-carbon, carbon-oxygen, phosphorus-oxygen and carbon-nitrogen bonds without hydrolysis or oxidation reactions, isomerase enzymes (3.1%), convert a molecule from one isomer to another while ligase enzymes (2.3%), catalyze the formation of a new chemical bond by joining of two large molecules (Fig. 5).

Functional annotation of putative candidate genes 15 linkage groups of *C. sinensis*. The 28% of the mapped sequences (405 out of 1,471) were annotated. Functional annotations of detected QTL led to the identification of functions of 37 putative candidate genes in parental clones that could be involved in the expression of the targeted traits (Table 1). The putative candidate genes involved in secondary metabolism (phenylpropanoid biosynthesis, flavonoid biosynthesis, alkaloid biosynthesis and volatile compounds biosynthesis) were located in five linkage groups LG04, LG09, LG10 and LG13. The putative candidate genes identified were involved in phenylpropanoid and flavonoid biosynthesis pathways included a group or class of enzymes namely; 2-oxoglutarate/Fe(II)-dependent dioxygenases superfamily (2-ODDs), aminotransferase class I and II, acyltransferase, glycosyl hydrolase family 9. This was in agreement with KEGG pathway mapping where enzymes mapped clustered into six major groups according to different metabolic pathways they are involved with transferases, hydrolases and oxidoreductases as three abundant enzymes. The multidrug and toxic compound extrusion (MATE) transporter was also putatively identified in LG09. BT1 family protein located in LG04 was putatively annotated for qCAFE.

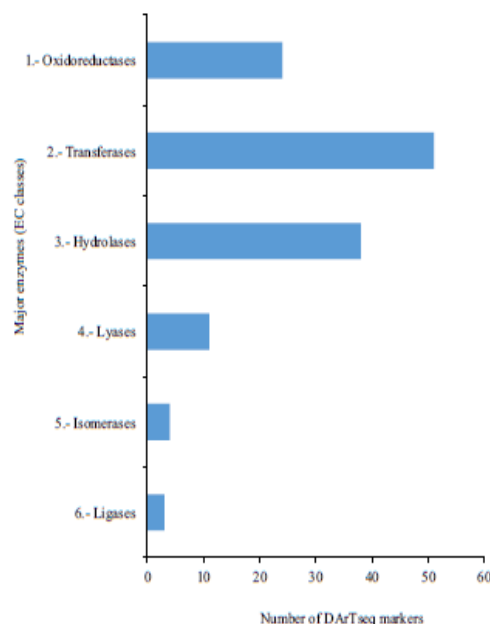


Figure 5. KEGG ontology (KO) distribution of major enzymes code related to secondary metabolism in *C. sinensis*.

The putative QTL and their corresponding putative candidate genes associated with phenylpropanoid biosynthesis, and flavonoid biosynthesis were identified in both parental clones.

Proteins related with oxidative stress and ATPase family proteins or proteins related to abiotic stress were found in the confidence intervals of QTL for drought tolerance traits. Among the 53 examined putative QTL in parental clones (Table 1), 18 putative QTL were related to abiotic stress (ABA signaling, amino acids peptide degradation) and were mostly found on LG04 and LG13. However, the putative QTL related to abiotic stress, including percent RWC was also found on LG01, LG06, LG10, LG12, LG14 and LG09, respectively. The putative candidate genes identified included, pectinesterase, 14-3-3 protein, 2OG-Fe (II) oxygenase superfamily, protein kinase domain, transmembrane amino acid transporter protein, peptidase family M3, histone acetyltransferase subunit NuA4, DnaJ domain and MatE. ATPase family proteins associated with various cellular activities (AAA), and Armadillo/beta-catenin-like repeat protein were found to be associated with the heat stress response. The putative QTL in LG13 that was associated with disease stress response was NB-ARC domain protein. Putative candidate genes related with ABA signaling pathway were found on QTL for qCAT, qEC, qEGCG and qRWC (Table 1), respectively. The putative candidate gene related to degradation of peptides by targeting the amino acids residues that are formed during abiotic stress was found on QTL for qEC. The parental clone TRFK 303/577 which is known to be high yielding and drought-tolerant cultivar, exhibited most putative candidate genes related to drought stress response as compared to parental clone GW-Ejulu.

Discussion

Flavonoid biosynthesis pathway in tea begins with the amino acid phenylalanine, which undergoes deamination to form trans-cinnamic acid^{28,29}. The oxidation process of trans-cinnamic acid yields p-coumaric acid, which undergoes transformation to form p-coumaroyl-CoA. The committed step in flavonoid biosynthesis is the subsequent condensation of p-coumaroyl-CoA and malonyl-CoA catalyzed by chalcone synthase enzyme to form chalcone. The subsequent isomerization of chalcone forms (2S)-flavonones. The identification of putative QTL, qCAT, for 2-ODDs superfamily protein in this study corroborates with other previous reports on the role of 2-ODDs in flavonoid biosynthesis. The 2-ODDs which are large superfamily of non-heme proteins, which facilitate numerous oxidative reactions, including hydroxylation, epimerization, C-C bond cleavage, ring formation and fragmentation. These 2-ODDs superfamily catalyze key oxidative reactions that facilitate the formation of different flavonoid subclasses, for example, the oxidation of (2S)-flavonones. The (2S)-flavonones are converted to dihydroflavonols through hydroxylation process which is then catalyzed by another 2-ODD, flavanone 3β-hydroxylase (F3H). The dihydroflavonols formed is a substrate for flavonol synthase (FNS1), which competes with dihydroflavonol 4-reductase (DR4) to form flavonols, anthocyanidins and procyanidins.

Glycosylation is an important process for the diverse functions of polyphenolic compounds in plants and is known to be able to increase the solubility and stability of hydrophobic flavonoids³⁰. The glycosylation process is regulated by glycoside hydrolases, which are involved in hydrolysis and/or rearrangement of glycosidic bonds. We hypothesize the putative QTL, qBRT linked to glycosyl hydrolase family 9, a family of glycoside hydrolases in

the current study to play a role in the synthesis and sequestration of some phenylpropanoids³¹. In tea, the galloylated catechins formed through the glucosylation activity of UDP-glucosyltransferase, have been found to confer astringent and bitter tastes, while flavanol 3-O-glycosides have been found to induce velvety, mouth-drying, and mouth-coating sensations⁶. In legume plant, bur clover (*Medicago truncatula*), epicatechin 3'-O-glucoside, has also been shown to be involved in the biosynthesis of proanthocyanidin in the seed coat^{32,33}. In addition, the glycosylated polyphenols function also as acyl donors in biochemical reactions, for example, β -glucogallin, a glucose ester of gallic acid, functions as an acyl transfer donor in the biosynthesis of both galloylated tannins³⁴ and galloylated catechins³⁵.

Acyltransferases which participates in many secondary metabolism pathways was putatively identified in this current study (qECG) in LG10. Acyltransferases are involved in the biosynthesis of anthocyanidin and other flavonoid groups by transferring acyl groups to the sugar moiety of anthocyanins using acyl-CoA as the donor³⁶. The process of modification of anthocyanin and other flavonoid groups with acyl groups is catalyzed by acyltransferases and is important in maintaining their stability. In LG04, qCAFF, was putatively annotated as BT1 family protein which has been identified to be involved in export of adenine nucleotides, which are exclusively synthesized in plants plastids, and are precursors for caffeine biosynthesis³⁷. Furthermore, nucleotides, adenine and guanosine are precursors in caffeine biosynthesis in tea plants³⁸.

The 14-3-3 proteins which bind to other proteins to induce target-site specific alteration and conformation of target proteins are important plant signal transduction pathways. Previous reports implicate plant 14-3-3s in key physiological processes, in particular, abiotic and biotic stress responses, metabolism (especially primary carbon and nitrogen metabolism), as well as various aspects of plant growth and development. In the current study, putative QTL, qCAT in LG13 was putatively annotated 14-3-3 protein, which is abiotic or biotic stress protein in tea. This, therefore, is in agreement with previous study by¹⁵ where individual catechins were demonstrated as potential indicators in predicting drought tolerance in tea plant. Meanwhile, a number of recent studies in *Arabidopsis* have linked 14-3-3s to ABA signaling. Some of the main functions ascribed to ABA include response to abiotic/biotic stress.

Drought stress which usually affects the plant photosynthetic process also interferes with plant nutrient availability leading to ion intoxication. The reversible protein phosphorylation by protein kinases at the early and later stage of the signaling pathways is one of the mechanisms that plants use in response to abiotic stress. In addition, to maintain the osmotic homeostasis and tolerance due to drought stress, plants also do regulate gene expression of some specific genes of protein kinases. The presence of a putative QTL, qCAT, which was annotated for protein kinase domain in this study supports the argument of the previous studies on the role of tea catechins as potential indicators of drought tolerance in tea plants. A previous study by³⁹ indicated that catechins were higher in the drought-tolerant cultivars and therefore, corroborated with the presence of qCAT, annotated as protein kinase domain in drought-tolerant parental clone TRFK 303/577.

Transmembrane amino acid transporter protein functions as gateways to permit the transport of amino acids across the biological membrane. Amino acids are not only critical for the synthesis of proteins, but other many essential compounds, which act as signal molecules. For example, in *Arabidopsis*, water and salt stress induce a strong expression of proline transporter 1 (ProT1) and proline transporter two while the expression of amino acid permease 4 (AAP4) and amino acid permease 6 (AAP6) is repressed^{40,41}. In tea, proline has been shown to act as an osmo-protectant during drought stress^{23,24} and the presence of a putative candidate gene (transmembrane amino acid protein) in this current study, could hypothesize the functions of transmembrane amino acid transporter for proline presence during drought stress. Furthermore, the presence of transmembrane amino acid transporter proteins could be their involvement of transport of aromatic amino acids phenylalanine, tyrosine and tryptophan, which serve as precursors for a wide range of secondary metabolites (polyphenols) in tea. This is because the two identified putative QTL, qCAT and qECG in LG13 in both parental clones (TRFK 303/577 and GW Ejulu) are products of shikimate and phenylalanine pathways.

The MATE proteins are a family of secondary active transporters, which utilize electrochemical gradient of membrane maintained by ATPases for their transport activity. MATE transporters are involved in transport of secondary metabolites into vacuoles, which are the major storage site of most of the conjugated flavonoids, which comprise of flavonols, anthocyanins and flavone glycoside⁴². In the current study, putative QTL, qRWC in LG09 was annotated for MATE proteins, which are involved in the sequestration of flavonoids in vacuoles in response to drought stress⁴³. Secondary metabolites in plants, especially flavonoids play an important role in response to abiotic stress such as drought stress. A member of MATE gene family (TT12) has also been elucidated for sequestration of proanthocyanidins in seed vacuoles, which leads to pigmentation of the seed coat⁴⁴. In yeast, it was revealed that TT12 proteins can specifically transport glycosidic form like epicatechin 3'-O-glucoside and cyanidin 3-O-glucoside⁴⁵. The MATE proteins from a legume, bur clover (*Medicago truncatula*) and grapevines (*Vitis*) have been characterized for transport of flavonoids through the sequestration of epicatechin 3-O-glucoside and cyanidin 3-O-glucoside (proanthocyanidins) into the vacuole for storage⁴³. In another study, a part from the role of vacuolar sequestration flavonoids, MATE proteins have been recognized in transport of alkaloids into vacuoles in tobacco. Furthermore, MATE proteins have facilitated and modulated abscisic acid efflux and sensitivity in drought-tolerant *Arabidopsis*⁴⁶. Since a putative QTL for %RWC annotated as MATE in this current study, it therefore, corroborates with previous reports⁴⁷ which relied on the percent %RWC to classify tea cultivars as either drought tolerant or drought susceptible based on tea leaves metabolites.

In this current study, putative QTL, qECG in LG06 was putatively annotated DnaJ protein. DnaJ proteins are essential components that contribute to cellular protein homeostasis and complex stabilization of proteins (protein folding, degradation and refolding) under stress conditions in plants⁴⁸. They function as molecular chaperones either alone or in association with heat-shock protein 70 (Hsp70). In a previous study on the transcriptomic analysis of the effect of drought-induced stress on tea leaf quality, revealed that the levels of ECG and ECGG increased during drought stress⁴⁹. This was also reported by³⁹ on the effects of soil moisture stress alterations

on tea biomolecules in relation to tea quality. Furthermore, in transgenic tobacco, the over expression of tomato chloroplast-targeted DnaJ enhanced drought tolerance⁵⁰ and maintained photosystem II under chilling stress⁵¹. Our findings, therefore, suggest that ECG can be used as a potential marker of drought tolerance or cold stress in tea cultivars.

The putative QTL, qEGCG, in LG04 was annotated diacylglycerol kinase catalytic domain protein which we hypothesize as a marker for drought and cold tolerance in tea cultivars. Abiotic stress, including salinity, drought, and osmotic adjustments triggers the production of phosphatidic acid⁵². Phosphatidic acid is formed by phosphorylation of diacylglycerol by diacylglycerol kinase. Thus, phosphatidic acid rather than diacylglycerol has typically been implicated as a major plant secondary messenger. However, the presence of diacylglycerol is necessary for certain developmental processes and the response to particular environmental stimuli. For example, salinity stress in *A. thaliana* increased the activity of non-specific phospholipase C, which promoted the production of diacylglycerol. In addition, cold stress has recently been found to stimulate the formation of phosphatidic acid in suspension-cultured *Arabidopsis* cell⁵³.

Histone acetyltransferases play a critical role in histone acetylation of plant chromatin, which is important in epigenetic control of gene expression. The transfer of acetyl groups to core histone tails by histone acetyltransferases, promote transcription of target genes involved in drought response and abscisic acid signaling in several plants, including *Arabidopsis*⁵⁴, rice⁵⁵, and maize⁵⁶. Furthermore, previous studies reported up-regulation of drought stress responsive genes and found to be correlated with changes in histone modification^{57,58}. The putative QTL, qCAT, qEC and qEGCG in this current study were annotated histone acetyltransferase subunit NuA4 protein in drought-tolerant parental clone TRFK 303/577. Therefore, catechins could act as potential indicators or markers of drought tolerance in tea, which corroborates with previous reports^{18,39} which investigated the biochemical changes in tea constituents as influenced by varying levels of water stress.

Aminotransferases, also known as transaminases, catalyze amino group transfers from amino donor to amino acceptor compounds. Aminotransferases play a major role in a variety of metabolic pathways, including, amino acid biosynthesis, secondary metabolites biosynthesis and photorespiration. Prephenate aminotransferase, a class of aminotransferase enzyme, catalyze the final step in the synthesis of phenylalanine, which is the central product of shikimic acid pathway and serves as a precursor for the synthesis of flavonoids⁵⁹. The identification of putative QTL, qEC in the current study annotated as aminotransferase I and II is hypothesized as a marker linked to flavonoids biosynthesis in tea.

Conclusion

This current study is the first attempt to obtain more information on genomic and functional annotation of proteins/enzymes involved in black tea quality and drought tolerance traits in tea plant based on DArTseq markers. The DArTseq markers used, and putative QTL identified in this current study will shed more light on the proteins/enzymes associated with the biosynthesis of caffeine, individual catechins in green tea which are converted to individual theaflavins in black tea processing. In addition, the DArTseq markers will provide useful information, in particular for studies of the genetic determinants of black tea quality traits and drought tolerance in tea. Also, the DArTseq markers used in the current study can be useful markers for construction of linkage maps or may be used to discover new functions of genes. We successfully constructed a genetic linkage map using SNPs, but it was impossible to anchor it to other previously constructed linkage maps of tea since no anchoring markers were available. Furthermore, it was also difficult to deduce the relative genetic positions of the SNPs on the previous different linkage maps. Indeed, previous studies used RAPD, AFLP and SSRs markers to construct tea genetic linkage maps, while in our study, we used DArTseq markers. However, the information obtained through the annotation of putative gene functions and the alignment of DArTseq sequences relative to recently published reference tea genome, has set a milestone for a future and further researches on marker-assisted selection and breeding of tea plants. In addition, this work may help breeders in selection of parents with desirable DArTseq markers for development of new tea cultivars with desirable traits.

Data Availability

The DArT sequences have been submitted to NCBI (<http://www.ncbi.nlm.nih.gov/>). BioProject PRJNA398959.

References

- Xia, E. H. *et al.* The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol. Plant*. 10, 866–877 (2017).
- Preedy, V. R. *Tea in health and disease prevention*, (Academic Press, 2012).
- Wang, K. *et al.* Comparison of phenolic compounds and taste of Chinese black tea. *Food Sci. Technol. Res.* 20, 639–646 (2014).
- Scharbert, S. & Hofmann, T. Molecular definition of black tea taste by means of quantitative studies, taste reconstitution, and omission experiments. *J. Agric. Food Chem.* 53, 5377–5384 (2005).
- Rossetti, D., Bongers, J., Wantling, E., Stokes, J. & Williamson, A.-M. Astringency of tea catechins: more than an oral lubrication tactile percept. *Food Hydrocol.* 23, 1984–1992 (2009).
- Cui, L. *et al.* Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (*Camellia sinensis*). *J. Exp. Bot.* 67, 2285–2297 (2016).
- Obanda, M., Owuor, P. O., Mang'ola, R. & Kavoi, M. M. Changes in thearubigin fractions and theaflavin levels due to variations in processing conditions and their influence on black tea liquor brightness and total colour. *Food Chem.* 85, 163–173 (2004).
- Owuor, P. O., Wachira, F. N. & Ng'etich, W. K. Influence of region of production on relative clonal plain tea quality parameters in Kenya. *Food Chem.* 119, 1168–1174 (2010).
- Koech, R. K. *et al.* Identification of novel QTL for black tea quality traits and drought tolerance in tea plants (*Camellia sinensis*). *Tree Genet. Genomes* 14, 9 (2018).
- Owuor, P. O. & Obanda, M. The use of green tea (*Camellia sinensis*) leaf flavan-3-ol composition in predicting plain black tea quality potential. *Food Chem.* 100, 873–884 (2007).
- Obanda, M., Owuor, P. O. & Taylor, S. J. Flavanol composition and caffeine content of green leaf as quality potential indicators of Kenyan black teas. *J. Sci. Food Agric.* 74, 209–215 (1997).

12. Liang, Y., Lu, J. & Zhang, L. Comparative study of cream in infusions of black tea and green tea [*Camellia sinensis* (L.) O. Kuntze]. *Int. J. Food Sci.* **37**, 627–634 (2002).
13. Ng'etich, W., Stephens, W. & Othieno, C. Responses of tea to environment in Kenya. 3. Yield and yield distribution. *Exp. Agr.* **37**, 361–372 (2001).
14. Cheruiyot, E., Mumera, L., Ng'etich, W., Hassanali, A. & Wachira, F. Threshold soil water content for growth of tea [*Camellia sinensis* (L.) O. Kuntze]. *Tea* **29**, 29–38 (2008).
15. Lu, S.-C. *et al.* Transcriptomic analysis of tea plant responding to drought stress and recovery. *PLoS one* **11**, e0147306 (2016).
16. Tony, M. *et al.* Transcriptome-based identification of water-deficit stress responsive genes in the tea plant, *Camellia sinensis* Plant Biotechnol. **43**, 302–310 (2016).
17. Cheruiyot, E. K., Mumera, L. M., Ng'etich, W. K., Hassanali, A. & Wachira, F. Polyphenols as potential indicators for drought tolerance in tea (*Camellia sinensis* L.). *Biosci. Biotechnol. Biochem.* **71**, 2190–2197 (2007).
18. Cheruiyot, E. K. *et al.* Shoot epicatechin and epigallocatechin contents respond to water stress in tea [*Camellia sinensis* (L.) O. Kuntze]. *Biosci. Biotechnol. Biochem.* **72**, 1219–1226 (2008).
19. Maritim, T. K. *et al.* Physiological and biochemical response of tea [*Camellia sinensis* (L.) O. Kuntze] to water-deficit stress. *J. Hort. Sci. Biotechnol.* **90**, 395–400 (2015).
20. Zhang, H. B. *et al.* De novo transcriptome assembly of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. *BMC Genomics* **16**, 298 (2015).
21. Zhou, L. *et al.* Exogenous abscisic acid significantly affects proteome in tea plant (*Camellia sinensis*) exposed to drought stress. *Horticulture, Research* **1**, 14029 (2014).
22. Sah, S. K., Reddy, K. R. & Li, J. Abscisic acid and abiotic stress tolerance in crop plants. *Front. Plant Sci.* **7**, 571 (2016).
23. Upadhyaya, H. & Panda, S. K. Abiotic stress responses in tea [*Camellia sinensis* (L.) O. Kuntze]: an overview. *Reviews in Agricultural Science* **1**, 1–10 (2013).
24. Upadhyaya, H. B. K., D. & Panda, S. K. Drought induced physiological and biochemical changes in leaves of developing seedlings of Tea [*Camellia sinensis* (L.) O. Kuntze] Cultivars. *Journal of Tea Science Research* **6**, 1–11 (2016).
25. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
26. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008** (2008).
27. Blake, J. *et al.* Gene ontology annotations and resources. *Nucleic Acids Res.* **41** (2013).
28. Punyasin, P. *et al.* Flavonoid biosynthesis in the tea plant *Camellia sinensis*: properties of enzymes of the prominent epicatechin and catechin pathways. *Arch. Biochem. Biophys.* **431**, 22–30 (2004).
29. Vanlatesh, P. *et al.* Flavonoid biosynthesis in *Camellia sinensis*. (2007).
30. Xu, L., Qi, T., Xu, L., Lu, L. & Xiao, M. Recent progress in the enzymatic glycosylation of phenolic compounds. *J. Carbohydr Chem.* **35**, 1–23 (2016).
31. Jones, P., Messner, B., Nakajima, I.-I., Schäffner, A. R. & Saito, K. UGT73C6 and UGT78D1, glycosyltransferases involved in flavonoid glycoside biosynthesis in *Arabidopsis thaliana*. *J. Biol Chem.* **278**, 43910–43918 (2003).
32. Pang, Y., Peed, G. J., Wright, E., Wang, Z. & Dixon, R. A. Early steps in proanthocyanidin biosynthesis in the model legume *Medicago truncatula*. *Plant Physiol.* **145**, 601–615 (2007).
33. Zhao, J. & Dixon, R. A. MATE transporters facilitate vacuolar uptake of epicatechin 3'-O-glucoside for proanthocyanidin biosynthesis in *Medicago truncatula* and *Arabidopsis*. *The Plant Cell* **21**, 2323–2340 (2009).
34. Niemetz, R. & Gross, G. G. Enzymology of galotannin and ellagitannin biosynthesis. *Phytochemistry* **66**, 2001–2011 (2005).
35. Liu, Y. *et al.* Purification and characterization of a novel galloyltransferase involved in catechingalloylation in the tea plant (*Camellia sinensis*). *J. Biol Chem.* **287**, 44406–44417 (2012).
36. Mizutani, M. *et al.* An acyltransferase involved in biosynthesis of polyacylated anthocyanin, s198 (2006).
37. Haferkamp, I. The diverse members of the mitochondrial carrier family in plants. *FEBS Lett.* **581**, 2375–2379 (2007).
38. Negishi, O., Ozawa, T. & Imagawa, H. Biosynthesis of Caffeine from Purine Nucleotides in Tea Plant. *Biosci. Biotechnol. Biochem.* **56**, 499–503 (1992).
39. Jeyaraman, P., Pius, P., Raj Kumar, R. & Jayakumar, D. Soil moisture stress-induced alterations in bioconstituents determining tea quality. *J. Sci. Food Agric.* **83**, 1187–1191 (2003).
40. Hayat, S. *et al.* Role of proline under changing environments: a review. *Plant Signal. Behav.* **7**, 1456–1466 (2012).
41. Hua, X. *et al.* *Arabidopsis* AMINO ACID PERMEASE1 contributes to salt stress-induced proline uptake from exogenous sources. *Front. Plant Sci.* **8**, 2182 (2017).
42. Shitan, N. Secondary metabolites in plants: transport and self-tolerance mechanisms. *Biosci. Biotechnol. Biochem.* **80**, 1283–1293 (2016).
43. Petrucci, E. *et al.* Plant flavonoids—biosynthesis, transport and involvement in stress responses. *Int. J. Mol. Sci.* **14**, 14930–14973 (2013).
44. Gonzalez, A. *et al.* TTG2 controls the developmental regulation of seed coat tannins in *Arabidopsis* by regulating vacuolar transport steps in the proanthocyanidin pathway. *Dev. Biol.* **419**, 54–63 (2016).
45. Marinova, K. *et al.* 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**, 2023–2038 (2007).
46. Zhang, H. *et al.* A DTX/MATE-type transporter facilitates abscisic acid efflux and modulates ABA sensitivity and drought tolerance in *Arabidopsis*. *Mol. Plant* **7**, 1522–1532 (2014).
47. Nyanukwira, C., Koeh, 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. *J. Plant Physiol.* **198**, 39–48 (2016).
48. Park, C. I. & Seo, Y. S. Heat shock proteins: a review of the molecular chaperones for plant immunity. *Plant Pathol. J.* **31**, 323 (2015).
49. Wang, W. *et al.* Transcriptomic analysis reveals the molecular mechanisms of drought-stress-induced decreases in *Camellia sinensis* leaf quality. *Front. Plant Sci.* **7**, 385 (2016).
50. Wang, G. *et al.* Overexpression of tomato chloroplast-targeted DnaJ protein enhances tolerance to drought stress and resistance to *Pseudomonas solanacearum* in transgenic tobacco. *Plant Physiol. Biochem.* **82**, 95–104 (2014).
51. Kong, F. *et al.* A chloroplast-targeted DnaJ protein contributes to maintenance of photosystem II under chilling stress. *J. Exp. Bot.* **65**, 143–158 (2013).
52. Zhu, J. K. Abiotic stress signaling and responses in plants. *Cell* **167**, 313–324 (2016).
53. Arisz, S. A. *et al.* Rapid phosphatidic acid accumulation in response to low temperature stress in *Arabidopsis* is generated through diacylglycerol kinase. *Front. Plant Sci.* **4**, 1 (2013).
54. Kim, J. M. *et al.* Alterations of lysine modifications on the histone H3 N-tail under drought stress conditions in *Arabidopsis thaliana*. *Plant Cell Physiol.* **49**, 1580–1588 (2008).
55. Fang, H., Liu, X., Thorn, G., Duan, J. & Tian, L. Expression analysis of histone acetyltransferases in rice under drought stress. *Biochem. Biophys. Res. Commun.* **443**, 400–405 (2014).
56. Kapazoglou, A. & Tsafaris, A. Epigenetic chromatin regulators as mediators of abiotic stress responses in cereals. In *Abiotic Stress in Plants: Mechanisms and Adaptations* (Intech, 2011).
57. Kim, J. M., Sasaki, T., Ueda, M., Sako, K. & Seki, M. Chromatin changes in response to drought, salinity, heat, and cold stresses in plants. *Front. Plant Sci.* **6**, 114 (2015).

58. Asensi-Fabado, M. A., Amtmann, A. & Perrella, G. Plant responses to abiotic stress: The chromatin context of transcriptional regulation. *Biochim Biophys Acta (BBA)* **1860**, 106–122 (2017).
59. Ververidis, F. *et al.* Biotechnology of flavonoids and other phenylpropanoid-derived natural products. Part I: Chemical diversity, impacts on plant biology and human health. *Biotechnol. J.* **2**, 1214–1234 (2007).

Acknowledgements

The authors acknowledge the financial support to conduct this research, and study grants for R.K. and P.M. 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).

Author Contributions

Z.A., S.K. and R.M. were involved with the design of the experiment and plant material used. R.K., P.M. and C.N. were involved in collection of plant material. R.K. performed the experiments. R.K., P.M., C.N., S.K. and Z.A. analyzed samples and interpreted the data. F.J. performed BLASTX and functional annotation. R.K. wrote the manuscript and revised by P.M., C.M., R.M., S.K. and Z.A. The final manuscript was reviewed and approved by all the authors.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-37688-z>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

CHAPTER 5

COMBINED LINKAGE AND ASSOCIATION MAPPING OF PUTATIVE QTLs CONTROLLING BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS

Abstract

The advancements in genotyping have opened new approaches for identification and precise mapping of QTLs in plants, particularly by combining linkage and AM analysis. In this study, a combination of linkage and the AM approach was used to precisely identify and authenticate putative QTLs associated with black tea quality traits and %RWC. The population structure analysis clustered two parents and their respective 261 F₁ progenies from the two reciprocal crosses into two clusters with 141 tea accessions in cluster one and 122 tea accessions in cluster two. The two clusters were of mixed origin with tea accessions in population TRFK St. 504 clustering together with tea accessions in population TRFK St. 524. A total of 70 putative QTLs linked to black tea quality traits and %RWC were detected in IM method. Cofactors obtained in IM were used in multiple QTL model (MQM) mapping where 40 putative QTLs were detected. The phenotypic variance for each QTL ranged from 2.8-23.3% in IM and 4.1-23% in MQM mapping. Using Q-model and Q+K-model in AM, a total of 49 DArTseq markers were associated with 16 phenotypic traits. Significant marker-trait association results obtained using AM were similar to those obtained in IM, and MQM mapping with six more QTLs detected in AM. The functional annotations of six additional putative QTLs detected in AM were involved in purine/thiamine biosynthesis, phenylalanine, tyrosine and tryptophan biosynthesis, carbon fixation and abiotic stress. The combined linkage and AM approach appears to have great potential and could reduce the cost and improve the efficacy and accuracy of selection of desirable traits in tea breeding.

5.1 Introduction

Linkage mapping has been a key tool for identifying the genetic basis of quantitative traits in plants. However, linkage mapping is usually limited by low polymorphism or small population size, which is required for suitable crosses. Also, the mapping resolution is limited by few recombination events, which are considered to estimate the genetic distances between marker loci causative genomic regions for QTLs. AM also known as LD mapping has extensively been used to circumvent the limitations posed by linkage mapping in plants in the last few years (Kraakman *et al.*, 2004). AM which exploits broader genetic diversity offers more advantages over linkage mapping in terms of mapping resolution, allele number, saves time in establishing a marker-trait association and its application in a breeding program (Soto-Cerda and Cloutier, 2012). Although linkage mapping and AM each offer more advantages over the other, they are often applied in conjunction to validate the QTL identified, thus reducing spurious associations.

The integrated approach of linkage-AM has been used in other plants such as *Arabidopsis* (Sterken *et al.*, 2012), wheat (Shi *et al.*, 2017) and maize (Li *et al.*, 2016) to dissect quantitatively inherited traits. In tree species, an integrated method of linkage mapping and AM to decipher the nature of genetic architecture of potential QTLs for growth traits has been reported in poplar hybrids (Du *et al.*, 2016) and maritime pine (Bartholomé *et al.*, 2016). In grapes, Fournier-Level *et al.* (2010) combined linkage mapping and AM to study the genetic patterns of anthocyanin content, a determinant of berry colour, in grapes. In a study to investigate genetic relationship between tea caffeine synthase gene (TCS1) and the caffeine content in tea plant and its related species, AM was used (Jin *et al.*, 2016). In this current study, we combined the high QTL detection power of the linkage mapping with the high-resolution power of AM. The linkage-AM approach will not only accelerate the pace of QTL mapping in tea breeding improvement but will also precisely identify reliable QTLs linked to black tea quality and drought tolerance in tea.

5.2 Research objective

The objective of this study was to combine linkage mapping and AM to precisely identify and authenticate putative QTLs linked to caffeine content, catechins fractions, theaflavins fractions, tea tasters' scores, and %RWC.

5.3 Hypotheses

Alternative hypothesis (H_1): Putative QTLs associated with caffeine content, catechins fractions, theaflavins fractions, tea tasters' scores, and %RWC detected in linkage mapping, and AM are significantly different.

Null hypothesis (H_0): Putative QTLs associated with caffeine content, catechins fractions, theaflavins fractions, tea tasters' scores, and %RWC detected in linkage mapping, and AM are not significantly different.

5.4 Material and methods

5.4.1 Population type

Two pseudo-testcross populations used in this current study consisted of 109 F_1 progeny from TRFK St. 504 (TRFK 303/577 x GW Ejulu) and 152 F_1 progeny from TRFK St. 524 (GW Ejulu x TRFK 303/577).

5.4.2 Phenotypic data

A total of 16 phenotypic traits were used as described in chapter 3.

5.4.3 Population structure

The population structure was inferred from the 1 421 DArT markers data previously used to construct genetic linkage map using the STRUCTURE software version 2.3.4 (Pritchard *et al.*, 2010). Twenty independent runs were carried out using the following parameters: number of populations (K) from 1 to 10, burn-in time and Markov Chain Monte Carlo (MCMC) replication number were both set to 100 000 for model of admixture with correlated allele frequencies. The natural logarithms of the probability data ($\text{LnP}(K)$) and the *ad hoc* delta K statistical were calculated using STRUCTURE Harvester (Earl, 2012), and the optimal K according to the delta K value was then selected (Mezmouk *et al.*, 2011). Finally, the population structure matrix (Q) was obtained by integrating 20 independent replicate runs and applying CLUMPP software (Jakobsson and Rosenberg, 2007). STRUCTURE bar plots were plotted using STRUCTURE Plot v 2.0 (Ramasamy *et al.*, 2014).

5.4.4 Linkage mapping, QTL analyses and allelic effects

Six thousand five hundred and eighty-eight DArTseq markers derived from *C. sinensis* genomic DNA of the two parents and 261 F_1 progeny were subjected to linkage mapping analyses using JoinMap 4.0 (Van Ooijen, 2006) and QTL mapping analyses using MapQTL 6.0 software (Van Ooijen *et al.*, 2000) as described in chapter 2.

The allelic effects were estimated as $A_f = [(\mu_{ac} + \mu_{ad}) - (\mu_{bc} + \mu_{bd})]/4$ for female additivity, $A_m = [(\mu_{ac} + \mu_{bc}) - (\mu_{ad} + \mu_{bd})]/4$ for male additivity and $D = [(\mu_{ac} + \mu_{bd}) - (\mu_{ad} + \mu_{bc})]/4$ for dominance where μ_{ac} , μ_{ad} , μ_{bc} and μ_{bd} are estimated phenotypic means associated to each of the four possible genotypic classes ac, bc, ad and bd, derived from ab×cd cross (Sadok *et al.*, 2013).

5.4.5 Association mapping

Using the Q and K matrices data files from STRUCTURE software as a covariates, 1 421 DArT markers of the 261 F₁ progeny were tested for association with each phenotype using a general linear model (GLM), mixed linear model (MLM) that included kinship, phylogenetic tree in TASSEL (trait analysis by association, evolution and linkage) Version 5.2.43 software (Bradbury *et al.*, 2007). Principal component analysis (PCA) was performed using JMP Pro 14 to visualise the dispersion of the association panel in a graph. The Q and kinship (K) matrices were used to correct the effects of population substructure in the association panel which can cause false-positive associations. The p-value and R² were used to determine whether a QTL is associated with the marker and the magnitude of QTL effects, respectively.

5.5 Results

5.5.1 Frequency distribution

The frequency distribution results for 16 phenotypic traits used in this study are discussed in chapter 3.

5.5.2 Population structure

A total of 261 tea accessions, 109 for population TRFK St. 504 and 152 for population TRFK St.524 were combined to construct a panel to study their genetic structure. The two tea populations, TRFK St. 504 and TRFK St. 524 are a reciprocal cross of two parents TRFK 303/577 and GW Ejulu. A population structure analysis showed that the value of Evanno's ΔK presented a sharp spike at $K = 2$, which suggested that this population panel was clustered into two groups (Figure 5.1). The two groups or clusters are represented by two clades with parental clone TRFK 303/577 clustering with 141 progenies in cluster one (blue) and parental clone GW Ejulu (not shown) clustering with 122 progenies in cluster two (black), respectively (Figure 5.2). The two clusters were accession with mixed origin with tea accessions in TRFK St. 504 population clustering together with tea accessions in TRFK St. 524 population. Furthermore, the results from PCA (Figure 5.3) and a phylogenetic tree-based on Nei's genetic distance was in agreement with the structure analysis results (Figure

5.4). The population structure of the two tea populations was examined with the neighbour-joining algorithm using Euclidian distance on the DArT marker band intensities.

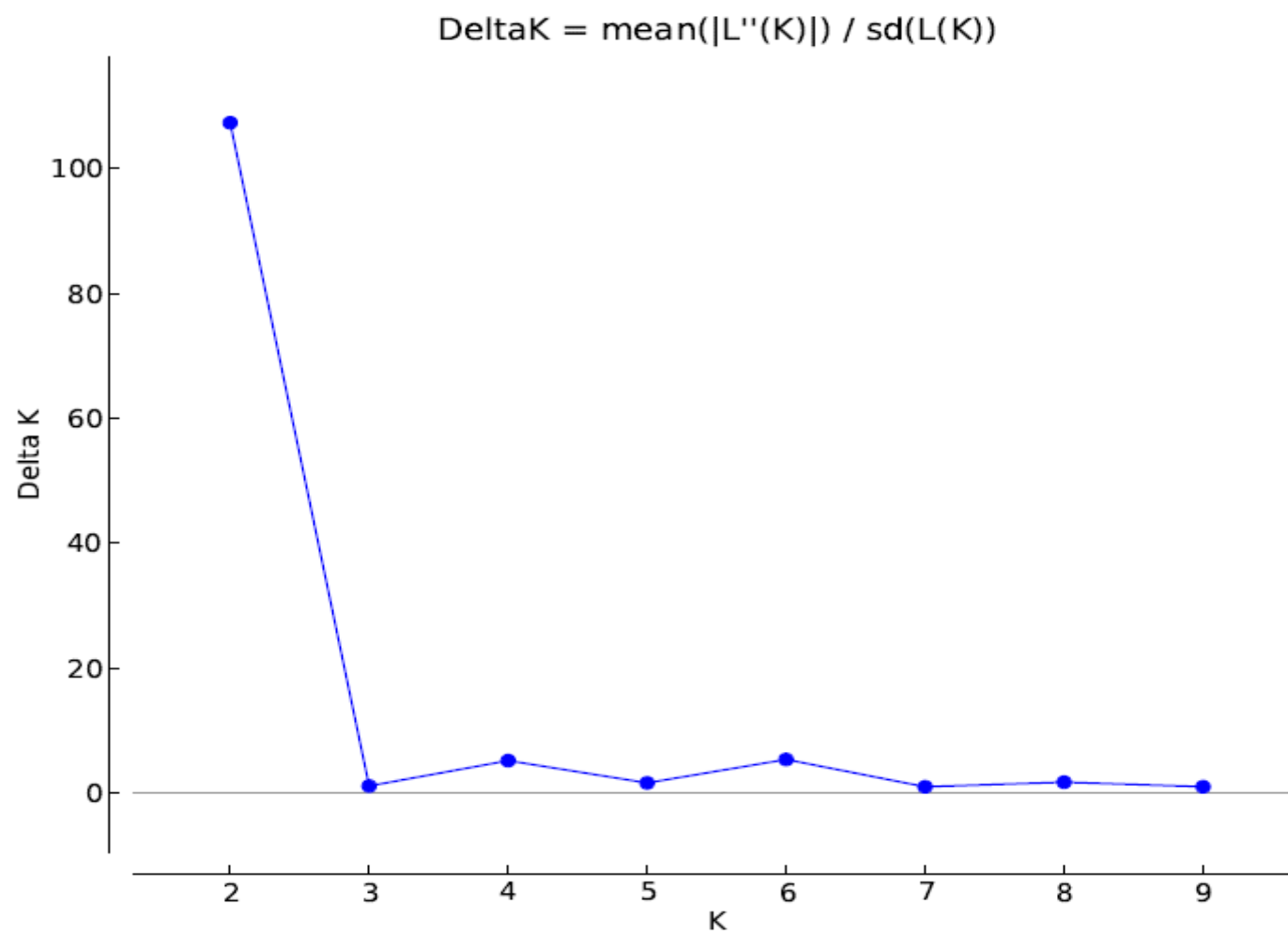


Figure 5.1: Delta K values plotted from 1 to 10 for 261 tea accessions panel.



Figure 5.2: Population structure of 261 tea accessions panel ($K = 2$)

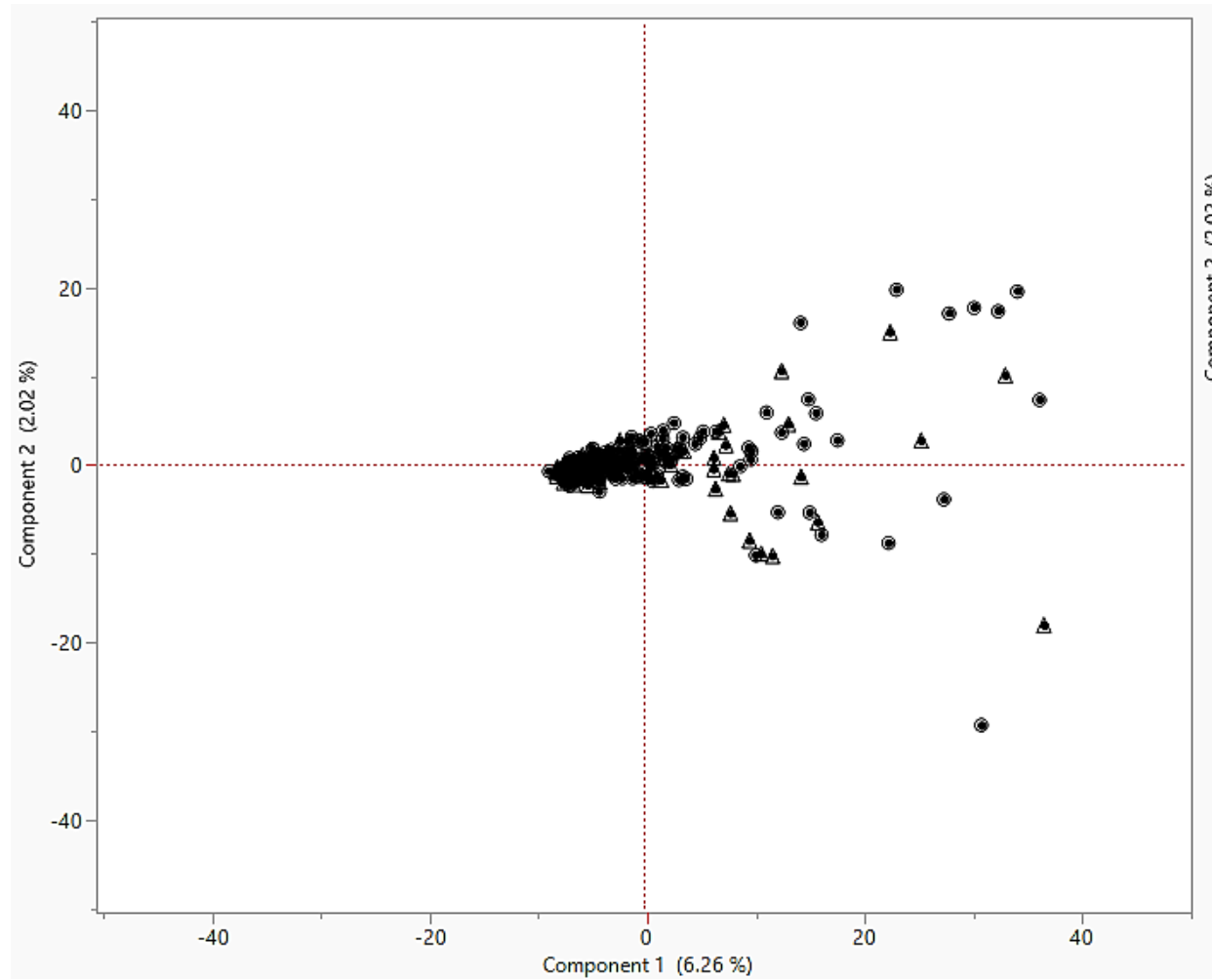


Figure 5.3: Principal component analysis of 261 tea accessions. ● represents cluster 1, ▲ represents cluster 2

5.5.3 Linkage mapping and association mapping

5.5.3.1 Interval mapping and allelic effects

The genetic linkage map was constructed using a total of 1,421 DArT markers as discussed in chapter 3. A total of 71 putative QTLs were identified for the 16 phenotypic traits measured using the interval mapping (Table 5.1). Of all the putative QTLs identified, most QTLs were found in almost all linkage groups, except LG03, LG05 and LG11. For black tea quality traits, 67 putative QTLs were detected and were located on almost all linkage groups, except on LG03, LG05 and LG11. The remaining three putative QTLs which are associated with drought tolerance were identified in LG02, LG06 and LG09. Eight putative QTLs were identified on LG01, seven in LG02 and LG12, 10 in LG04, five in LG06 and LG15, three in LG07, LG08, LG09, two in LG10, 15 in LG13, four in LG15, respectively. The highest number of putative QTLs were associated with catechin (13), EGC (12), tea tasters' score (12), caffeine (10) and EGC (9), respectively. The phenotypic variance for all the identified putative QTLs ranged from 2.8% for EGC to 23.3% for qECG, respectively. The phenotypic variance for %RWC ranged from 5.7 to 7.3%, while for the tea tasters' score, it ranged from 5.8 to 9.1%. Half of the allelic effects identified to be associated with black tea quality and drought tolerance traits had a positive additive effect. There were eight allelic effects associated with female additivity, and 18 were associated with male additive effects, respectively. The highest positive female and male additive effects were qEGC and qECG, respectively. Although, some putative QTLs also exhibited high negative allelic effects, which were from either female or male additive effects or both, such as qEGCG, qCAT, qEC and qCaffeine (Table 5.1). All the putative QTLs identified in IM were used as cofactors in MQM mapping.

5.5.3.2 Multiple QTL model mapping and allelic effects

A total of 46 putative QTLs ($LOD > 3.0$) were identified in 16 phenotypic traits using MQM mapping method from the selected cofactors in IM. Similarly, no putative QTLs in LG03, LG05 and LG11, respectively were associated with any of the phenotypic traits. Most of the putative QTLs previously identified in IM were also identified MQM mapping. Of these 46 putative QTLs identified, 43 putative QTLs that were associated with black tea quality traits while three were associated with qRWC (Table 5.2). The three putative QTLs associated with qRWC were similar to those identified in IM. Four putative QTLs were identified in LG01 and LG12, seven in LG02, five in LG04, three in LG06, LG09, LG13, LG14 and LG15, two in LG07 and LG08 and one in LG01, respectively. A similar trend was observed in the

number of putative QTLs for each trait as in IM with a high putative QTL number for catechin, ECG, EGC and caffeine, respectively. The phenotypic variance for all the identified putative QTL ranged from 4.1% for qCAT to 23% for qECG, respectively. The phenotypic variance for qRWC was similar to IM results, and it ranged from 5.7 to 7.3%, while for the tea tasters' score, it ranged from 6.0 to 9.1%. For allelic effects, similar results to IM with qEGC and qECG showing high positive female additive and male additive effects, respectively, this trend was also observed in MQM mapping. The putative QTLs identified using both IM, and MQM mapping were almost similar although the efficiency and the accuracy of detecting QTL is achieved by employing MQM instead of the single QTL model used in IM.

5.5.3.3 Association mapping

Four models GLM, GLM (Q), mixed linear model with kinship matrix (MLM) (K) and mixed linear model with kinship matrix and population structure matrix (MLM) (Q+K) in TASSEL software v5.2.43 were used to determine AM and the effects of population structure on the AM to reduce the inflation of false-positive associations. The p-values were plotted cumulatively for each model, and the distribution examined. However, no further analysis was done using GLM model since marker-trait association results were characterised by excess small p-values, which indicated an abundance of spurious associations (Appendix 5.1). The association analysis using the Q-model and Q+K-model detected a total of 49 DArT markers associated with 15 different black tea quality traits and one drought-tolerant trait. The GLM (Q) model showed associations between DArT markers and traits ($p < 0.01$) and was confirmed using the MLM (Q+K) model (Table 5.3, Appendix 5.2 and Appendix 5.4). The MLM Q+K-model (MLM with Q-matrix and K-matrix as a correction for population structure) showed a good fit for the p-values ($p < 0.01$), as compared to GLM (Q) models, which were characterised by a few associations with excess of small p-values, which indicates abundance of spurious associations (Table 5.3, Appendix 5.2 and Appendix 5.4). Also, the MLM Q+K-model also showed better small p-values than the K-model (MLM with K-matrix as a correction for population structure) (Appendix 4.3 and Appendix 5.4). The GLM (Q) model may not have accounted for the heterogeneity of the genetic background in some tea cultivars under study, which may have resulted in false-positive associations. Ideally, the distribution of p-values should follow a uniform distribution with less deviation from the expected p-values. However, the two models were used to compare the effect of population matrix and kinship matrix on GLM and MLM, respectively, in associating

mapping. Therefore, taking into account the performance of the two different models, the results from the MLM (Q+K) model appeared to have controlled better population structure and kinship relationships than GLM (Q) model. While it might be tempting to consider p-values that remain extreme after MLM (Q+K) model correction as true associations, we need to keep in mind that minor allele frequency also influences p-values.

A total of 49 DArT markers were associated with 16 different phenotypic traits in six out of 15 linkage groups (Table 5.3). Since, all 1 421 DArT markers in the genotypic dataset had a known map location, the loci associated with particular traits could be allocated to a linkage group. The numbers of marker loci associated with traits together with their location and positions in different linkage groups at a test level of $p < 0.01$ in the two test models are shown in Table 5.3. Choosing a significance level of $p < 0.01$ which involves multiple testing corrections in both GLM (Q) and MLM (Q+K) models, 32 DArT marker loci were detected for all individual catechins (C, EC, ECG, EGC, EGCG), scattered over five linkage groups namely LG01, LG02, LG04, LG12 and LG15 (Table 5.3). For caffeine trait, three DArT markers were only located in LG02 while 11 DArT markers for all individual theaflavins traits (TF1, TF2, TF3, TF4) were detected in LG04 and LG15 (Table 5.3). The trait for tea liquor colour which is an indicator for black tea quality in terms of tea tasters' score appeared to be associated with only one DArT marker locus at the significance level of $p < 0.01$ in LG07 (Table 5.3). Drought tolerance trait (%RWC) was also associated with only one DArT marker locus detected in LG02 (Table 5.3). The majority of the DArT markers loci associated with the traits were found in LG04, LG12, LG15, LG02, and LG01. The three DArT markers associated with catechin trait in LG04, namely, 5136945, 5064764 and 5134490 had the highest proportion of the percent explained phenotypic variation of 41%, 38%, and 25%, respectively. The DArT markers that also showed high proportion of percent explained phenotypic variation were also observed in LG04 for qEGCG, qEGC, and qTF1 traits, respectively. In this study, a pleiotropic locus which is associated with a single locus affecting two or more distinct phenotypic traits was observed. Several pleiotropic loci were identified including DArT markers 5136945, 5112438, 5136058, 5123257, 5106022 and 5064764 which were associated with qC, qEC, qEGC, qTF1, qTF2 and qTF4 phenotypic traits (Table 5.3). All the pleiotropic loci detected and associated with the respective phenotypic traits were located in LG04.

Table 5.1: Linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St. 504 and St. 524 using interval mapping in MapQTL 6.0

Nr	^a Trait	Locus	LG	Position (cM)	LOD	PEV	Af	Am	D	Allelic effect
1	AST	5115441_E-26	1	96.6	4.5	9.1	0.224	-0.061	-0.015	Af
2	AST	5135087	1	97.3	4.5	9.1	0.234	-0.057	-0.014	Af
3	BRK	5115441_E-26	1	96.6	3.6	7.3	0.196	-0.082	-0.020	Af
4	BRK	5135087	1	97.3	3.5	7.2	0.204	-0.078	-0.019	Af
5	ECG	5128890	1	96.4	6.8	11.7	0.326	0.046	0.011	Af
6	ECG	5135087	1	97.3	6.6	11.3	0.347	0.053	0.013	Af
7	ECG	5115441_E-26	1	96.6	6.5	11.1	0.331	0.051	0.013	Af
8	EGC	5133866	1	87.1	3.3	5.8	0.109	0.218	0.054	Am
9	CAFF	5064585	2	50.9	3.3	5.8	5.849	-9.702	-2.425	Af; Am
10	CAT	5135436	2	0.0	3.6	6.1	-0.838	-11.195	-2.799	Am
11	EC	5072338	2	2.2	4.3	7.2	-14.648	-12.555	-3.139	Af; Am
12	EGC	5124128	2	7.7	3.3	5.6	-36.447	1.544	0.386	Af
13	RWC	5136794	2	60.9	4.3	7.3	0.482	1.130	0.283	Am
14	TF1	5084595	2	4.5	3.2	5.7	-0.952	8.158	2.039	Am
15	TF2	5084595	2	4.5	4.1	7.3	2.290	9.140	2.285	Am
16	AR	100158044 F 0	4	70.1	2.9	6	-0.268	0.090	0.023	Af
17	CAFF	5112599	4	68.6	3.8	6.7	2.413	11.852	2.963	Am
18	CAT	5063001	4	30.3	5.3	4.1	-17.604	-53.467	-13.367	Am
19	EC	5123475	4	27.2	14.0	21.9	3.485	44.620	11.155	Am
20	EC	5123257	4	26.7	13.4	21.1	1.703	43.813	10.953	Am
21	EC	5134490	4	26.4	12.4	19.6	0.260	41.569	7.005	Am
22	ECG	5087113	4	17.7	3.8	5.8	-2891.857	3715.230	928.808	Af; Am
23	EGC	5123475	4	27.2	3.7	2.8	7071.043	-6846.394	-1711.598	Af; Am
24	EGCG	5087017	4	37.2	4.9	6	-320.975	-200.380	-50.095	Af; Am
25	ECG	5098382	6	57.0	5.2	8.7	35.650	-32.760	-8.190	Af; Am
26	EGC	5073424	6	66.4	3.9	6.7	-13.905	14.749	3.687	Af; Am
27	EGC	5124993	6	66.6	3.9	6.6	-12.321	15.635	3.909	Af; Am
28	RWC	5082606	6	66.2	3.3	5.7	-2.094	-0.814	-0.203	Af
29	TF1	5136045	6	69.6	4.3	7.7	2.795	3.782	0.946	Af; Am
30	CAFF	5064391	7	48.1	3.4	6	-12.264	-3.397	-0.849	Af
31	CAFF	5056614	7	48.3	3.3	5.9	-12.029	-3.985	-0.996	Af
32	CL	5132432	7	72.7	3.2	6.5	-0.138	-0.118	-0.029	Af; Am
33	CAFF	5134558	8	18.8	4.3	7.5	17.375	-12.332	-3.083	Af; Am
34	CAFF	5111497	8	18.8	4.1	7.1	11.542	-12.927	-3.232	Af; Am
35	CAT	5130194	8	12.5	3.9	6.6	-29.556	-9.720	-2.430	Af

36	AST	5123950	9	87.6	3.2	6.6	-0.036	0.178	0.045	Am
37	BRK	5123950	9	87.6	3.6	7.4	-0.012	0.190	0.047	Am
38	RWC	5130531	9	6.7	4.1	7	1.933	2.876	0.719	Am
39	AR	5075627	10	26.9	3.4	6.9	-0.064	0.110	0.027	Am
40	ECG	5136108	10	20.6	3.3	5.7	4.297	34.819	8.705	Am
41	CAT	5123751	12	43.0	4.2	6.5	-36.478	-3.162	-0.791	Af
42	CAT	5127224	12	42.9	4.2	6.5	-36.467	-3.045	-0.761	Af
43	ECG	5136790	12	50.4	15.0	23.3	-22.365	60.821	15.205	Am
44	ECG	5135536	12	50.4	15.0	23.2	-22.316	60.808	15.202	Am
45	EGC	5123751	12	43.0	5.0	8.5	50.465	-30.542	-7.635	Af; Am
46	EGC	5127224	12	42.9	5.0	8.4	50.241	-30.563	-7.641	Af; Am
47	EGCG	5104630	12	48.2	4.2	7.2	18.115	-52.080	-13.020	Am
48	BRT	5088162	13	29.8	3.0	6.1	0.006	-0.142	-0.035	Am
49	BRT	5135810	13	26.0	2.8	5.8	0.007	-0.137	-0.034	Am
50	CAFF	5088162	13	29.8	3.4	5.4	7.183	-29.136	-7.284	Am
51	CAFF	5080631	13	35.6	2.9	3.1	-14.651	23.420	5.855	Af; Am
52	CAT	5122819	13	10.2	4.3	7.3	-27.341	-14.521	-3.630	Af; Am
53	CAT	5111268	13	58.1	4.6	7.2	14.200	17.804	4.451	Af; Am
54	CAT	5129729	13	9.8	4.2	7.2	-26.752	-14.467	-3.617	Af; Am
55	CAT	5115793	13	58.7	4.3	6.8	12.635	17.248	4.312	Af; Am
56	CAT	5136782	13	58.5	4.2	6.6	12.029	17.237	4.309	Af; Am
57	CAT	5103784	13	50.6	3.8	6.1	4.188	21.966	5.491	Am
58	CAT	5122899	13	50.5	3.8	6	3.656	21.921	5.480	Am
59	ECG	5123761	13	60.7	3.3	5.8	-0.469	-0.112	-0.028	Af
60	ECG	5072523	13	60.3	3.0	5.2	-0.261	-0.231	-0.058	Af; Am
61	ECG	5114985	13	61.6	2.8	5	-0.439	-0.088	-0.022	Af
62	EGC	5136623	13	50.6	3.7	6.4	8.946	-29.764	-7.441	Am
63	BRT	5125626	14	71.1	3.3	6.8	0.000	0.143	0.036	Am
64	CAFF	5123053	14	5.5	4.0	7.1	-7.622	-11.572	-2.893	Af; Am
65	CAFF	5054639	14	6.4	3.9	6.9	-7.260	-11.533	-2.883	Af; Am
66	CAT	5132370	14	60.7	6.5	10.8	-19.945	35.937	8.984	Af; Am
67	EGC	5132370	14	60.7	3.7	6.4	46.492	-7.514	-1.878	Af
68	EC	5085963	15	25.4	4.4	7.5	11.674	-24.472	-6.118	Am
69	EC	5099958_E-25	15	25.1	4.4	7.4	11.686	-24.101	-6.025	Am
70	ECG	5111164	15	75.1	4.2	7.2	22.426	34.202	8.550	Af; Am
71	EGCG	5114089	15	32.1	4.0	6.8	-1.178	56.567	14.142	Am

^a Putative QTL identified in Interval Mapping using MapQTL 6.0. LOD- Logarithm of odds for putative QTL in IM (LOD>2); PEV- percentage phenotypic variation explained by markers in IM; Af- Female additivity; Am- Male additivity; D- Dominance

Table 5.2: Linkage mapping for putative QTLs in black tea quality and drought tolerance traits in St. 504 and St. 524 using multiple QTL model mapping in MapQTL 6.0

Nr	^b Trait	Locus	LG	Position (cM)	LOD	PEV	Af	Am	D	Allelic effect
1	AST	5135087	1	97.3	4.5	9.1	0.234	-0.057	-0.014	Af
2	BRK	5115441_E-26	1	96.6	3.6	7.3	0.196	-0.082	-0.020	Af
3	ECG	5128890	1	96.4	6.8	11.7	0.326	0.046	0.011	Af
4	EGC	5133866	1	87.1	3.3	5.8	0.109	0.218	0.054	Am
5	CAFF	5064585	2	50.9	3.3	5.8	5.849	-9.702	-2.425	Af; Am
6	CAT	5135436	2	0.0	3.6	6.1	-0.838	-11.195	-2.799	Am
7	EC	5072338	2	2.2	4.3	7.2	-14.648	-12.555	-3.139	Af; Am
8	EGC	5124128	2	7.7	3.3	5.6	-36.447	1.544	0.386	Af
9	RWC	5136794	2	60.9	4.3	7.3	0.482	1.130	0.283	Am
10	TF1	5084595	2	4.5	3.2	5.7	-0.952	8.158	2.039	Am
11	TF2	5084595	2	4.5	4.1	7.3	2.290	9.140	2.285	Am
12	AR	100158044 F 0	4	70.1	2.9	6	-0.268	0.090	0.023	Af
13	CAFF	5112599	4	68.6	3.8	6.7	2.413	11.852	2.963	Am
14	CAT	5063001	4	30.3	5.3	4.1	-17.604	-53.467	-13.367	Am
15	EC	5123475	4	27.2	14.0	21.9	3.485	44.620	11.155	Am
16	ECG	5087113	4	17.6	4.7	8.2	-0.133	0.141	-0.090	Af; Am
17	EGC	5123475	4	27.1	3.7	2.8	7071.043	-6846.394	7330.160	Af; Am; D
18	EGCG	5087017	4	37.2	4.9	6	-320.975	-200.380	-50.095	Af; Am
19	ECG	5098382	6	57.0	5.2	8.7	35.650	-32.760	-8.190	Af; Am
20	EGC	5073424	6	66.4	3.9	6.7	-13.905	14.749	3.687	Af; Am
21	TF1	5136045	6	69.6	4.3	7.7	2.795	3.782	6.425	D
22	RWC	5082606	6	66.2	3.3	5.7	-2.094	-0.814	-0.203	Af
23	CAFF	5064391	7	48.1	3.4	6	-12.264	-3.397	-0.849	Af
24	CL	5132432	7	72.7	3.2	6.5	-0.138	-0.118	-0.029	Af; Am
25	CAFF	5134558	8	18.8	4.3	7.5	17.375	-12.332	-3.083	Af; Am
26	CAT	5130194	8	12.5	3.9	6.6	-29.556	-9.720	-2.430	Af
27	AST	5123950	9	87.6	3.2	6.6	-0.036	0.178	0.045	Am
28	BRK	5123950	9	87.6	3.6	7.4	-0.012	0.190	0.047	Am
29	RWC	5130531	9	6.7	4.1	7	1.933	2.876	0.719	Am
30	ECG	5136108	10	20.6	3.3	5.7	4.297	34.819	8.705	Am
31	AR	5075627	10	26.9	3.4	6.9	-0.064	-0.064	-0.064	Am
32	CAT	5123751	12	43.0	4.2	6.5	-36.478	-3.162	-0.791	Af

33	ECG	5136790	12	50.4	15.0	23.3	-22.365	60.821	15.205	Am
34	EGC	5123751	12	43.0	5.0	8.5	50.465	-30.542	-7.635	Af; Am
35	EGCG	5104630	12	48.2	4.2	7.2	18.115	-52.080	-13.020	Am
36	BRK	5114985	13	61.6	3.1	6.3	-0.366	-0.071	0.374	Af; D
37	CAFF	5088162	13	29.8	3.4	5.4	7.183	-29.136	-7.284	Am
38	CAT	5111268	13	58.1	4.6	7.2	14.200	17.804	4.451	Af; Am
39	EGC	5136623	13	50.6	3.7	6.4	8.946	-29.764	-7.441	Am
40	BRT	5125626	14	71.1	3.3	6.8	0.000	0.143	0.036	Am
41	CAFF	5123053	14	5.5	4.0	7.1	-7.622	-11.572	-2.893	Af; m
42	CAT	5132370	14	60.7	6.5	10.8	-19.945	35.937	8.984	Af; Am
43	EGC	5132791	14	51.4	3.0	3.0	-14.572	66.511	-43.428	Am
44	EC	5085963	15	25.4	4.4	7.5	11.674	-24.472	-6.118	Am
45	ECG	5111164	15	75.1	4.2	7.2	22.426	34.202	8.550	Af; Am
46	EGCG	5114089	15	32.1	4.0	6.8	-1.178	56.567	14.142	Am

^b Putative QTL identified in Multiple QTL Model Mapping using MapQTL 6.0

LOD- Logarithm of odds for putative QTL in MQM (LOD>3); PEV- percentage phenotypic variation explained by markers in MQM; Af- Female additivity; Am- Male additivity; D- Dominance

CAFF- Caffeine; CAT- Catechin; EC- Epicatechin; ECG- Epicatechin gallate; EGC- Epigallocatechin gallate; EGCG- Epigallocatechin gallate; TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate; CL- Colour; BRT- Brightness; AST- Astringency; BRK- Briskness; AR- Aroma; RWC- percent relative water content

Table 5.3: Association mapping results for black tea quality and drought tolerance traits in St. 504 and St.524 using the GLM (Q) and MLM (Q+K) method, respectively (p <0.01)

Nr	Trait	Marker	LG	Position (cM)	p-Value	Marker R ² GLM (Q)	p-Value	Marker R ² MLM (Q+K)
1	ECG	5128890	1	96.4	1.18E-03	0.042	5.31E-03	0.031
2	EGC	5086476	1	87.5	8.20E-03	0.031	8.54E-04	0.043
3	EGC	5122985	1	87.5	1.47E-04	0.056	2.41E-03	0.038
4	EGC	5133866	1	87.1	8.94E-03	0.030	6.62E-04	0.046
5	CAFF	5134416	2	44.4	2.46E-04	0.057	5.58E-04	0.050
6	CAFF	5136551	2	43.3	9.60E-04	0.048	7.51E-03	0.030
7	CAFF	5137282	2	50.7	2.27E-03	0.042	9.72E-03	0.026
8	EC	5132500	2	7.9	2.91E-03	0.036	1.37E-04	0.052
9	RWC	5136794	2	60.9	8.58E-03	0.032	3.02E-03	0.033
10	CAT	5063001	4	30.3	2.34E-09	0.098	6.67E-08	0.109
11	CAT	5064764	4	28.8	6.42E-37	0.326	7.80E-22	0.379
12	CAT	5136945	4	28.1	6.39E-40	0.345	1.81E-21	0.406
13	EC	5112438	4	27.4	4.90E-07	0.087	1.50E-05	0.067
14	EC	5123257	4	26.7	3.89E-08	0.101	7.16E-07	0.089
15	EC	5134490	4	26.4	5.15E-19	0.228	6.21E-15	0.246
16	EC	5136058	4	27.6	1.00E-06	0.083	5.31E-06	0.073
17	EC	5136410	4	29.9	2.39E-05	0.065	1.77E-05	0.065
18	ECG	5087113	4	17.7	3.74E-03	0.035	2.92E-04	0.048
19	EGC	5112438	4	27.4	1.50E-11	0.148	1.95E-08	0.134
20	EGC	5123257	4	26.7	9.59E-13	0.162	6.95E-10	0.167
21	EGC	5136058	4	27.6	3.43E-11	0.143	3.60E-09	0.146
22	EGCG	5074553	4	39.5	3.26E-14	0.210	1.86E-12	0.221
23	EGCG	5123463	4	28.9	3.38E-17	0.250	1.09E-13	0.266
24	EGCG	5134853	4	37.6	7.03E-14	0.205	2.00E-09	0.170
25	EGCG	5136554	4	38.7	1.46E-16	0.242	2.08E-11	0.222
26	EGCG	100011901 F 0	4	38.7	1.24E-04	0.066	3.13E-03	0.039
27	TF1	5064764	4	28.8	2.35E-10	0.149	5.78E-06	0.088
28	TF1	5106352	4	26.0	1.95E-04	0.061	8.45E-03	0.032
29	TF1	5123463	4	28.9	1.33E-10	0.153	2.66E-06	0.097
30	TF1	5136945	4	28.1	1.69E-11	0.166	8.20E-07	0.107
31	TF2	5063001	4	30.3	9.76E-04	0.050	4.22E-03	0.035
32	TF2	5064764	4	28.8	1.85E-08	0.122	7.36E-05	0.063

33	TF2	5106022	4	29.1	2.99E-09	0.134	3.06E-04	0.052
34	TF4	5064764	4	28.8	8.64E-06	0.082	1.33E-03	0.042
35	TF4	5106022	4	29.1	1.38E-06	0.094	1.57E-03	0.040
36	TF4	5136945	4	28.1	6.26E-07	0.099	1.59E-04	0.060
37	CL	5132432	7	72.7	1.31E-03	0.047	7.30E-04	0.049
38	ECG	5059563	12	49.0	8.91E-09	0.111	2.29E-05	0.068
39	ECG	5136077	12	42.8	1.91E-07	0.094	6.34E-03	0.027
40	ECG	5136790	12	50.4	7.87E-11	0.137	1.58E-04	0.051
41	EGC	5135536	12	50.4	2.25E-04	0.053	2.82E-06	0.086
42	EGCG	5087581	12	48.3	1.51E-04	0.065	1.75E-03	0.040
43	EGCG	5088456	12	47.9	7.70E-04	0.053	2.46E-04	0.060
44	EGCG	5133629	12	47.7	7.17E-04	0.054	1.95E-03	0.041
45	EC	5085963	15	25.4	3.52E-04	0.049	9.09E-03	0.029
46	EC	5132852	15	24.8	1.09E-04	0.056	2.50E-03	0.036
47	EC	5099958_E-25	15	25.1	1.00E-04	0.056	6.37E-03	0.031
48	TF3	5083854	15	15.2	3.52E-03	0.043	2.77E-03	0.035
49	TF3	5136247	15	16.3	1.57E-03	0.049	5.40E-04	0.046

^c Putative QTL identified in General Linear Model and Mixed Linear Model methods using TASSEL software v5.2.43

LG- Linkage group

Marker R²- percentage phenotypic variation explained by markers in both GLM (Q) and MLM (Q+K) methods

Q- Population structure matrix

K- Kinship matrix

CAFF- Caffeine; CAT- Catechin; EC- Epicatechin; ECG- Epicatechin gallate; EGC- Epigallocatechin gallate; EGCG- Epigallocatechin gallate; TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate; CL- Colour; BRT- Brightness; AST- Astringency; BRK- Briskness; AR- Aroma; RWC- percent relative water content

5.5.4 A comparison of linkage mapping and association mapping

The putative QTLs detected using IM, and MQM mapping in MapQTL 6.0 were compared with marker-trait association results obtained using GLM (Q) and MLM (Q+K) models in AM. The significant marker-trait association results obtained using AM were similar to the results obtained using MapQTL 6.0 (Table 5.4 and Table 5.5). The IM analysis produced 14 significant QTLs ($LOD > 3.0$) that were also found to be associated with the phenotypic traits in AM analysis (Table 5.4). From the 14 QTLs, a total of 11 markers were associated with four individual catechins (qCAT, qEC, qECG, and qEGC) and one marker each was associated with qTF2, qCL, and qRWC, respectively. On the other hand, the MQM mapping analysis produced eight significant QTLs ($LOD > 3.0$) which were also found to be significantly associated with the phenotypic traits in AM analysis (Table 5.5). Of these, five markers were associated with four individual catechins (qCAT, qEC, qECG, and qEGC) and one marker each was associated with qTF2, qCL, and qRWC, respectively. Furthermore, the four mapping models were compared with each other to find markers that could associate with phenotypic traits in all four mapping models (Table 5.6). A total of 13 markers were found to be associated with four individual catechins (qCAT, qEC, qECG, and qEGC) and one marker each was associated with qTF2, qCL, and qRWC, respectively. Most of the markers that were associated with different phenotypic traits were only located in LG01, LG02, LG04, LG12 and LG15 with a majority of markers showing both positive and negative male additive effects.

Table 5.4: Association mapping and linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St.504 and St.524 using GLM (Q), MLM (Q + K) methods ($p < 0.01$) in TASSEL software v5.2.43 and interval mapping in MapQTL 6.0

Nr	^{ac} Trait	Marker	LG	Position (cM)	p-Value	Marker R ² GLM (Q)	p-Value	Marker R ² MLM (Q+K)
1	ECG	5128890	1	96.4	1.18E-03	0.042	5.31E-03	0.031
2	EGC	5133866	1	87.1	8.94E-03	0.030	6.62E-04	0.046
3	RWC	5136794	2	60.9	8.58E-03	0.032	3.02E-03	0.033
4	CAT	5063001	4	30.3	2.34E-09	0.098	6.67E-08	0.109
5	EC	5123257	4	26.7	3.89E-08	0.101	7.16E-07	0.089
6	EC	5134490	4	26.4	5.15E-19	0.228	6.21E-15	0.246
7	ECG	5087113	4	17.7	3.74E-03	0.035	2.92E-04	0.048
8	EGC	5123257	4	26.7	9.59E-13	0.162	6.95E-10	0.167
9	TF2	5063001	4	30.3	9.76E-04	0.050	4.22E-03	0.035
10	CL	5132432	7	72.7	1.31E-03	0.047	7.30E-04	0.049
11	ECG	5136790	12	50.4	7.87E-11	0.137	1.58E-04	0.051
12	EGC	5135536	12	50.4	2.25E-04	0.053	2.82E-06	0.086
13	EC	5085963	15	25.4	3.52E-04	0.049	9.09E-03	0.029
14	EC	5099958_E-25	15	25.1	1.00E-04	0.056	6.37E-03	0.031

^{ac} Putative QTL identified in both Interval Mapping (MapQTL 6.0), General Linear Model and Mixed Linear Model methods (TASSEL software v5.2.43)

Marker R²- percentage phenotypic variation explained by markers in GLM and MLM

CAT- Catechin; EC- Epicatechin; ECG- Epicatechin gallate; EGC- Epigallocatechin gallate; TF2- Theaflavin-3-gallate; CL- Colour; RWC- Percent relative water content

Table 5.5: Association mapping and linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St.504 and St.524 using GLM (Q), MLM (Q + K) methods ($p < 0.01$) in TASSEL software v5.2.43 and multiple QTL model mapping in MapQTL 6.0

Nr	^{bc} Trait	Marker	LG	Position (cM)	p-Value	Marker R ² GLM (Q)	p-Value	Marker R ² MLM (Q+K)
1	ECG	5128890	1	96.4	1.18E-03	0.042	5.31E-03	0.031
2	EGC	5133866	1	87.1	8.94E-03	0.030	6.62E-04	0.046
3	RWC	5136794	2	60.9	8.58E-03	0.032	3.02E-03	0.033
4	CAT	5063001	4	30.3	2.34E-09	0.098	6.67E-08	0.109
5	TF2	5063001	4	30.3	9.76E-04	0.050	4.22E-03	0.035
6	CL	5132432	7	72.7	1.31E-03	0.047	7.30E-04	0.049
7	ECG	5136790	12	50.4	7.87E-11	0.137	1.58E-04	0.051
8	EC	5085963	15	25.4	3.52E-04	0.049	9.09E-03	0.029

^{bc} Putative QTL identified in both multiple QTL model mapping (MapQTL 6.0), General Linear Model and Mixed Linear Model methods (TASSEL software v5.2.43)

Marker R²- percentage phenotypic variation explained by markers in GLM and MLM

CAT- Catechin; EC- Epicatechin; ECG- Epicatechin gallate; EGC- Epigallocatechin gallate; TF2- Theaflavin-3-gallate; CL- Colour; RWC- Percent relative water content

Table 5.6: Association mapping and linkage mapping results for putative QTLs in black tea quality and drought tolerance traits in St.504 and St.524 using GLM (Q), MLM (Q + K) methods ($p < 0.01$) in TASSEL software v5.2.43 and interval mapping and multiple QTL model mapping in MapQTL 6.0

Nr	^{abc} Trait	Marker	LG	Position (cM)	p-Value	Marker R ² GLM (Q)	p-Value	Marker R ² MLM (Q+K)	LOD	PEV	Af	Am	D	Allelic effects
1	ECG	5128890	1	96.4	1.18E-03	0.042	5.31E-03	0.031	6.8	11.7	0.326	0.046	0.011	Af
2	EGC	5133866	1	87.1	8.94E-03	0.030	6.62E-04	0.046	3.3	5.8	0.109	0.218	0.054	Am
3	RWC	5136794	2	60.9	8.58E-03	0.032	3.02E-03	0.033	4.3	7.3	0.482	1.13	0.283	Am
4	CAT	5063001	4	30.3	2.34E-09	0.098	6.67E-08	0.109	5.3	4.1	-17.6	-53.45	-13.37	Am
5	EC	5123257	4	26.7	3.89E-08	0.101	7.16E-07	0.089	13.4	21.1	1.703	43.81	10.95	Am
6	ECG	5087113	4	17.7	3.74E-03	0.035	2.92E-04	0.048	3.8	5.8	-2891.86	3715.23	928.81	Af; Am
7	EGC	5123257	4	26.7	9.59E-13	0.162	6.95E-10	0.167	13.4	21.1	1.703	43.81	10.95	Am
8	TF2	5063001	4	30.3	9.76E-04	0.050	4.22E-03	0.035	5.3	4.1	-17.604	-53.47	-13.37	Am
9	CL	5132432	7	72.7	1.31E-03	0.047	7.30E-04	0.049	3.2	6.5	-0.138	-0.118	-0.029	Af; Am
10	ECG	5136790	12	50.4	7.87E-11	0.137	1.58E-04	0.051	15	23.3	-22.37	60.82	15.21	Am
11	EGC	5135536	12	50.4	2.25E-04	0.053	2.82E-06	0.086	15	23.2	-22.32	60.81	15.20	Am
12	EC	5085963	15	25.4	3.52E-04	0.049	9.09E-03	0.029	4.4	7.5	11.67	-24.47	-6.12	Am
13	EC	5099958_E-25	15	25.1	1.00E-04	0.056	6.37E-03	0.031	4.4	7.4	11.69	-24.1	-6.03	Am

^{abc} Putative QTL identified using both Interval Mapping and Multiple QTL Model Mapping (MapQTL 6.0) and General Linear Model and Mixed Linear Model method (TASSEL software v5.2.43)

Marker R²- percentage phenotypic variation explained by markers in GLM and MLM

LG- Linkage group

LOD- Logarithm of odds ratio for putative QTL

PEV- percentage phenotypic variation explained by markers in IM and MQM

Af- Female additivity

Am- Male additivity

D- Dominance

CAT- Catechin; EC- Epicatechin; ECG- Epicatechin gallate; EGC- Epigallocatechin gallate; TF2- Theaflavin-3-gallate; CL- Colour; RWC- Percent relative water content

5.5.5 Functional annotation of putative QTLs in linkage and association mapping

In chapter 4, putative QTLs that were identified in both IM, and MQM mapping were identified by the BLAST algorithm and searched on the reference tea genome (NCBI) and Blast2GO database for functional annotation. A total of 53 QTLs were detected in both interval and MQM mapping methods, of which 45 proteins were assigned functions. An additional six markers in LG02, LG04, and LG12 were found to be associated with caffeine, EC, ECG and EGC traits in AM (Table 5.7). The six additional QTLs detected in AM were also functionally annotated using the above databases. The putative candidate genes identified were involved in purine or thiamine biosynthesis, phenylalanine biosynthesis, tyrosine and tryptophan biosynthesis, carbon fixation (photosynthesis) and abiotic stress. The putative QTL, qCaffeine in LG02 was putatively annotated N-(5'phosphoribosyl) anthranilate (PRA) isomerase enzyme. The two other putative QTLs identified for qEC and qEGC located in LG04 were annotated as phosphoribulokinase/uridine kinase family proteins. Two putative QTLs, for qEC in LG04, were annotated as autophagy-related protein 11 and protein tyrosine kinase, respectively.

Table 5.7: Association mapping, linkage mapping and functional annotation protein of putative QTLs in black tea quality traits in St.504 and St.524 based on IM and MLM (Q + K) method ($p < 0.01$) in MapQTL 6.0 and TASSEL software v5.2.43, respectively.

Nr	Trait	Marker	LG	Pos (cM)	p-Value	A	p-Value	B	p-Value	Annotated protein	Function
1	^{abc} ECG	5128890	1	96.4	1.2E-03	0.042	5.3E-03	0.031	2.0E-25	['Actin']	Cell signalling (response to cold stress, dehydration)
2	^c CAFF	5137282	2	50.7	2.3E-03	0.042	9.7E-03	0.026	8.0E-09	['N-(5'phosphoribosyl) anthranilate (PRA) isomerase']	Involved in the first and intermediate pathway in the purine/thiAmine biosynthesis; It also participates in the phenylalanine, tyrosine and tryptophan biosynthesis pathway
3	^{abc} ECG	5087113	4	17.7	3.7E-03	0.035	2.9E-04	0.048	6.0E-22	['impB/mucB/sAmB fAmily']	Involved in UV protection through DNA repair
4	^{ac} TF1	5106352	4	26	2.0E-04	0.061	8.5E-03	0.032	2.0E-06	['Thiolase, C-terminal domain']	Benzoic acid biosynthesis (volatile compounds)
5	^c EC	5112438	4	27.4	4.9E-07	0.087	1.5E-05	0.067	9.0E-21	['Phosphoribulokinase / Uridine kinase fAmily']	Carbon fixation (photosynthesis) which are products for shikimate pathway (phenolics)
6	^c EGC	5112438	4	27.4	1.5E-11	0.148	2.0E-08	0.134	9.0E-21	['Phosphoribulokinase / Uridine kinase fAmily']	Carbon fixation (photosynthesis) which are productsfor shikimate pathway (phenolics)
7	^{ac} EC	5134490	4	26.4	5.2E-19	0.228	6.2E-15	0.246	3.0E-08	['Aminotransferase class I and II']	Phenylalanine, tyrosine and tryptophan biosynthesis
8	^{ac} EGCG	5134853	4	37.6	7.0E-14	0.205	2.0E-09	0.17	2.0E-06	['Diacylglycerol kinase catalytic domain']	Water, salt, ROS, cold and freezing stress response
9	^c EC	5136058	4	27.6	1.0E-06	0.083	5.3E-06	0.073	8.0E-09	['Autophagy-related protein 11']	Abiotic stress response (stomatal closure)
10	^{ac} EGC	5136058	4	27.6	3.4E-11	0.143	3.6E-09	0.146	8.0E-09	['Autophagy-related protein 11']	Abiotic stress response (stomatal closure)
11	^c EC	5136410	4	29.9	2.4E-05	0.065	1.8E-05	0.065	7.0E-06	['Protein tyrosine kinase']	ABA signalling
12	^{ac} EGCG	5088456	12	47.9	7.7E-04	0.053	2.5E-04	0.06	4.0E-23	['Protein kinase domain']	Abiotic stress response
13	^c ECG	5136077	12	42.8	1.9E-07	0.094	6.3E-03	0.027	5.0E-13	CSA026168	-

^{abc} Putative QTL identified using both Interval Mapping and Multiple QTL Model Mapping (MapQTL 6.0), General Linear Model and Mixed Linear Model (TASSEL software v5.2.43), respectively.

^c Putative QTL identified in General Linear Model and Mixed Linear Model (MLM) using TASSEL software v5.2.43;

A- Marker R^2 GLM (Q); B- Marker R^2 MLM (Q+K)

5.6 Discussion

In this current study, a combination of linkage mapping and AM was used to provide a mutual and precise authentication of QTL, which will enable more reliable results to be obtained. Linkage mapping method has a relatively low genome resolution while AM method is affected by population structure and individual relationships. The combination of linkage mapping and AM has successfully been applied to many plant studies (Fan and Xiong, 2003; Jung *et al.*, 2005; Lu *et al.*, 2010; Lou *et al.*, 2015; Cao *et al.*, 2017).

The population structure analysis in discovery mapping panel used in the current study identified two groups or clusters. The dendrogram demonstrated two subspecies of the two parental clones, TRFK 303/577 and GW Ejulu, which are of *C. sinensis* var. *assamica* and *C. sinensis* var. *sinensis* variety, respectively. The neighbour-joining tree for the 261 tea accessions grouped them into two major clusters on the base of Nei's genetic distance. The phylogenetic relationships for the parental cultivars and their progenies clustering into their respective groups in separate clades based on their parental genotypes were consistent with their genetic backgrounds (Kamunya *et al.*, 2010). The parental clone TRFK 303/577, which is an open-pollinated progeny of clone TRFK 6/8, is high-yielding, drought-tolerant, medium in black tea quality, caffeine, and individual catechins. On the other hand, parental clone GW Ejulu is a low-yielding, high black tea quality and moderate levels of caffeine, but high in total catechins and individual catechin contents (TRFK, 2012). The bulk of tea cultivars used for green and black tea production have been derived through individual selection, hybridisation and molecular breeding of *C. sinensis* var. *assamica* and *C. sinensis* var. *sinensis* varieties to produce tea with desirable characteristics (Yao *et al.*, 2008). Therefore, developing a few markers that have the capability of discriminating the two subspecies and their cultivars is of paramount interest. The results on the linkage mapping for all the phenotypic traits and the QTL positions in different linkage groups in this study were similar as reported in (Koech *et al.*, 2018).

In this study, AM was conducted with three different models, Q, K, and Q+K. The observed $-\log_{10}(P)$ values for QTL deviated from the expected $-\log_{10}(P)$ values in the Q method (GLM), indicated that there might be a few false-positive. However, the addition of genetic relatedness K (kinship) used in a mixed linear model has proven to be more powerful and reduces the number of false-positive associations (Finno *et al.*, 2014). This is because a lot more QTLs were

identified in this current study using Q-model as compared with K and Q + K-model. Therefore, there could be higher chances of false-positive or negative errors in AM due to complex population structure (Müller *et al.*, 2011; George, 2013). In this study, a mixed linear model approach using the K-matrix or a combination (Q+K) performed better than a GLM (Q). However, the K or the Q + K methods did show to be too strict and resulted in the missing of some possibly useful QTLs. Nearly all the significant QTLs identified in AM using the Q, and Q+K models were in line with those identified in MapQTL 6.0 (Koech *et al.*, 2018). The percentages of phenotypic variation explained (R^2) by various markers analysed using AM were significant as those analysed using MapQTL 6.0. This is in agreement with previous studies reported in plants on efficiency and robustness of combining linkage mapping and AM for precise identification of QTL (Mammadov *et al.*, 2015; Li *et al.*, 2016).

The six additional QTLs identified in AM were found to be associated with purine or thiamine biosynthesis, phenylalanine, tyrosine and tryptophan biosynthesis, carbon fixation (photosynthesis) and abiotic stress. The putative QTL, qCaffeine located in LG02 was putatively annotated N-(5'phosphoribosyl) anthranilate (PRA) isomerase enzyme. The enzyme is involved in the first and intermediate pathway in the purine or thiamine biosynthesis and tryptophan biosynthesis pathway (Li *et al.*, 1995). Caffeine (1,3,7-trimethylxanthine) and other methylxanthines such as theobromine (3,7-dimethylxanthine) and methyluric acids present in tea are classified as purine alkaloids (Ashihara *et al.*, 2008). Therefore, the putative QTL, qCaffeine could be associated with the biosynthesis of caffeine, theobromine and methyluric acids in tea. Also, phenylalanine, tyrosine, and tryptophan are products of shikimate pathway which is involved in the biosynthesis of plant flavonoids, including catechins (Ghasemzadeh and Ghasemzadeh, 2011).

The putative QTLs, qEC, and qEGC in LG04 were annotated as phosphoribulokinase/uridine kinase family proteins which are involved in carbon fixation or photosynthesis in plants (Miziorko, 2000). Photosynthesis is an important process in plants for provision of carbon skeletons to the shikimate pathway (Janacek *et al.*, 2009). Catechins are synthesised in the leaves of the tea plant through the acetic-malonic acid and shikimic-cinnamic acid metabolic pathways (Vuong *et al.*, 2011). The chalcone and gallic acid are produced from the shikimic acid pathway, which then produce the different catechins (Chu and Juneja, 1997). The two putative QTLs, qEC

annotated putatively as autophagy-related protein 11 and protein tyrosine kinase are involved in abiotic stress response through a process of stomatal closure and ABA signalling (Sah *et al.*, 2016; Batoko *et al.*, 2017). Stress conditions can interfere with photosynthetic energy production in plants, which leads to stomatal closure, which inhibits CO₂ intake and thereby reduces photosynthetic activity in leaves (Chaves *et al.*, 2009). Absciscic acid has been reported to confer abiotic stress-tolerance in crop plants (Sah *et al.*, 2016). In stress conditions like drought, extreme temperature, and high salinity, content in plants increases considerably, inspiring stress-tolerance effects that help plants, adapt, and survive under these stressful situations (Ng *et al.*, 2014). Also, ABA is also required for plant growth and development under non-stress conditions (Nakashima and Yamaguchi-Shinozaki, 2013).

5.7 Conclusion

The tea plant is a woody perennial crop with long generation intervals, which usually hinders its genetic improvement by conventional breeding methods. The approach of combining the high QTL detection power of genetic linkage mapping with the high-resolution power of AM allowed identification and precise authentication of putative QTLs controlling black tea quality and drought tolerance traits. Based on the two mapping approaches, the putative QTLs associated with caffeine content, catechins fractions, theaflavins fractions, tea tasters' scores, and %RWC detected in linkage mapping and AM were not significantly different except that an additional six more QTLs were detected using AM method. Therefore, the combination of linkage and association strategies can be used to identify closely linked molecular markers for robust MAS and trait introgression. This will accelerate tea breeding progress in terms of reduced breeding time, reduced cost, increased efficiency, and precision of selection.

References

- Ashihara, H., Sano, H. & Crozier, A. (2008). Caffeine and related purine alkaloids: biosynthesis, catabolism, function and genetic engineering. *Phytochemistry* 69(4): 841-856.
- Bartholomé, J., Bink, M. C., van Heerwaarden, J., Chancerel, E., Boury, C., Lesur, I., Isik, F., Bouffier, L. & Plomion, C. (2016). Linkage and association mapping for two major traits used in the maritime pine breeding program: height growth and stem straightness. *PloS one* 11(11): e0165323.
- Batoko, H., Dagdas, Y., Baluska, F. & Sirko, A. (2017). Understanding and exploiting autophagy signaling in plants. *Essays in Biochemistry* 61(6): 675-685.
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y. & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19): 2633-2635.
- Cao, Y., Li, S., Wang, Z., Chang, F., Kong, J., Gai, J. & Zhao, T. (2017). Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and genome-wide association mapping. *Frontiers in Plant Science* 8: 1222.
- Chaves, M. M., Flexas, J. & Pinheiro, C. (2009). Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. *Annals of Botany* 103(4): 551-560.
- Chu, D. & Juneja, L. (1997). General chemical composition of green tea and its infusion. *Chemistry and Applications of Green Tea*: 13-22.
- Du, Q., Gong, C., Wang, Q., Zhou, D., Yang, H., Pan, W., Li, B. & Zhang, D. (2016). Genetic architecture of growth traits in *Populus* revealed by integrated quantitative trait locus (QTL) analysis and association studies. *New Phytologist* 209(3): 1067-1082.
- Earl, D. A. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4(2): 359-361.
- Fan, R. & Xiong, M. (2003). Combined high-resolution linkage and association mapping of quantitative trait loci. *European Journal of Human Genetics* 11(2): 125.
- Finno, C. J., Aleman, M., Higgins, R. J., Madigan, J. E. & Bannasch, D. L. (2014). Risk of false positive genetic associations in complex traits with underlying population structure: a case study. *The Veterinary Journal* 202(3): 543-549.

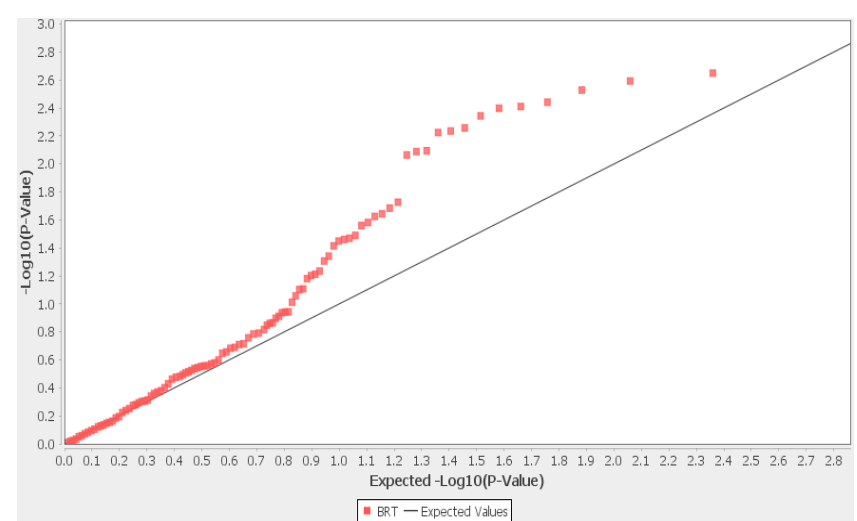
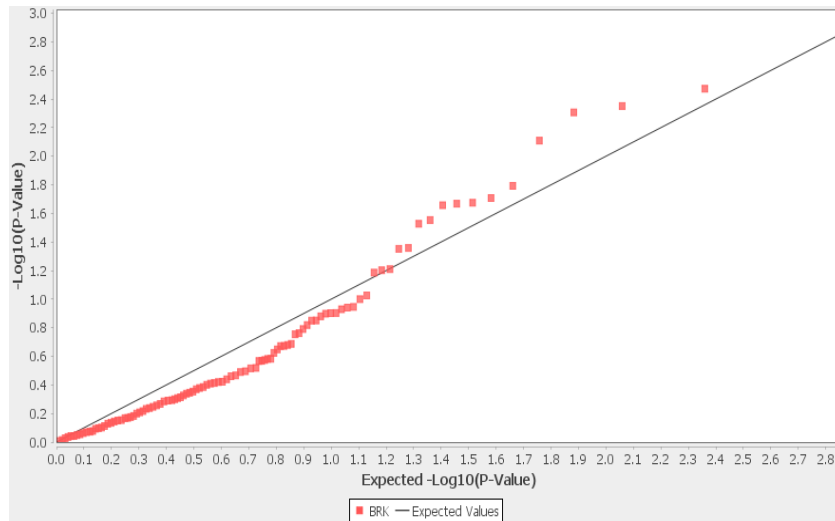
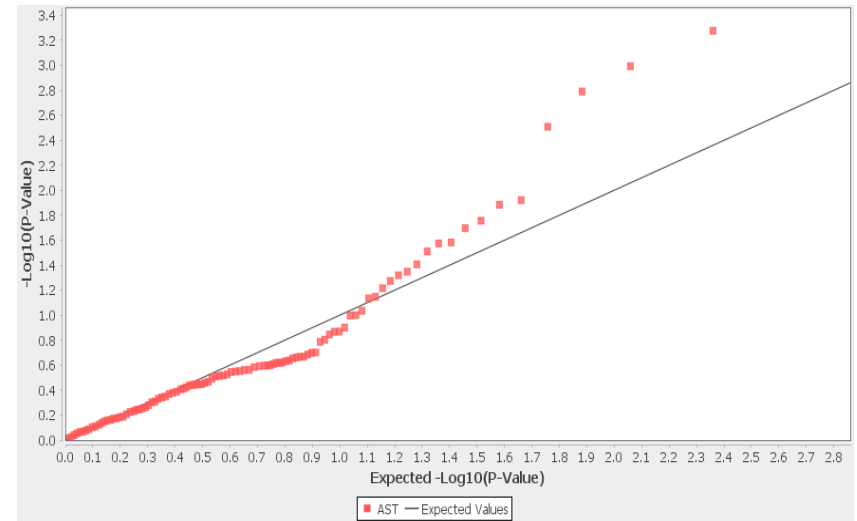
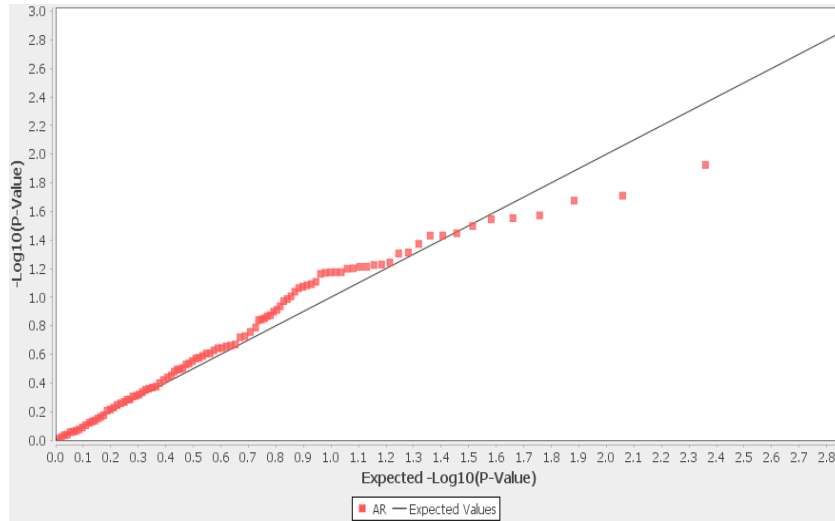
- Fournier-Level, A., Lacombe, T., Le Cunff, L., Boursiquot, J. & This, P. (2010). Evolution of the VvMybA gene family, the major determinant of berry color in cultivated grapevine (*Vitis vinifera* L.). *Heredity* 104(4): 351.
- George, A. (2013). Controlling type 1 error rates in genome-wide association studies in plants. *Heredity* 111(1): 86.
- Ghasemzadeh, A. & Ghasemzadeh, N. (2011). Flavonoids and phenolic acids: Role and biochemical activity in plants and human. *Journal of Medicinal Plants Research* 5(31): 6697-6703.
- Jakobsson, M. & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23(14): 1801-1806.
- Janacek, S. H., Trenkamp, S., Palmer, B., Brown, N. J., Parsley, K., Stanley, S., Astley, H. M., Rolfe, S. A., Paul Quick, W. & Fernie, A. R. (2009). Photosynthesis in cells around veins of the C3 plant *Arabidopsis thaliana* is important for both the shikimate pathway and leaf senescence as well as contributing to plant fitness. *The Plant Journal* 59(2): 329-343.
- Jin, J. Q., Yao, M. Z., Ma, C. L., Ma, J. Q. & Chen, L. (2016). Association mapping of caffeine content with TCS1 in tea plant and its related species. *Plant Physiology and Biochemistry* 105: 251-259.
- Jung, J., Fan, R. & Jin, L. (2005). Combined linkage and association mapping of quantitative trait loci by multiple markers. *Genetics* 170(2): 881-898.
- 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 and Genomes* 6(6): 915-929.
- 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 and Genomes* 14(1): 9.
- Kraakman, A. T., Niks, R. E., Van den Berg, P. M., Stam, P. & Van Eeuwijk, F. A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 168(1): 435-446.

- Li, J., Zhao, J., Rose, A. B., Schmidt, R. & Last, R. L. (1995). *Arabidopsis* phosphoribosyl anthranilate isomerase: molecular genetic analysis of triplicate tryptophan pathway genes. *The Plant Cell* 7(4): 447-461.
- Li, X., Zhou, Z., Ding, J., Wu, Y., Zhou, B., Wang, R., Ma, J., Wang, S., Zhang, X. & Xia, Z. (2016). Combined linkage and association mapping reveals QTL and candidate genes for plant and ear height in maize. *Frontiers in Plant Science* 7: 833.
- Lou, Q., Chen, L., Mei, H., Wei, H., Feng, F., Wang, P., Xia, H., Li, T. & Luo, L. (2015). Quantitative trait locus mapping of deep rooting by linkage and association analysis in rice. *Journal of Experimental Botany* 66(15): 4749-4757.
- Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., Farkhari, M., Ribaut, J. M., Cao, M. & Rong, T. (2010). Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences* 107(45): 19585-19590.
- Mammadov, J., Sun, X., Gao, Y., Ochsenfeld, C., Bakker, E., Ren, R., Flora, J., Wang, X., Kumpatla, S. & Meyer, D. (2015). Combining powers of linkage and association mapping for precise dissection of QTL controlling resistance to gray leaf spot disease in maize (*Zea mays* L.). *BMC Genomics* 16(1): 916.
- Mezmouk, S., Dubreuil, P., Bosio, M., Décousset, L., Charcosset, A., Praud, S. & Mangin, B. (2011). Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. *TAG Theoretical and Applied Genetics* 122(6): 1149-1160.
- Miziorko, H. M. (2000). Phosphoribulokinase: current perspectives on the structure/function basis for regulation and catalysis. *Advances in Enzymology and Related Areas of Molecular Biology, Volume 74*: 95-127.
- Müller, B., Stich, B. & Piepho, H. (2011). A general method for controlling the genome-wide type I error rate in linkage and association mapping experiments in plants. *Heredity* 106(5): 825.
- Nakashima, K. & Yamaguchi-Shinozaki, K. (2013). ABA signaling in stress-response and seed development. *Plant Cell Reports* 32(7): 959-970.
- Ng, L. M., Melcher, K., Teh, B. T. & Xu, H. E. (2014). Absciscic acid perception and signaling: structural mechanisms and applications. *Acta Pharmacologica Sinica* 35(5): 567.

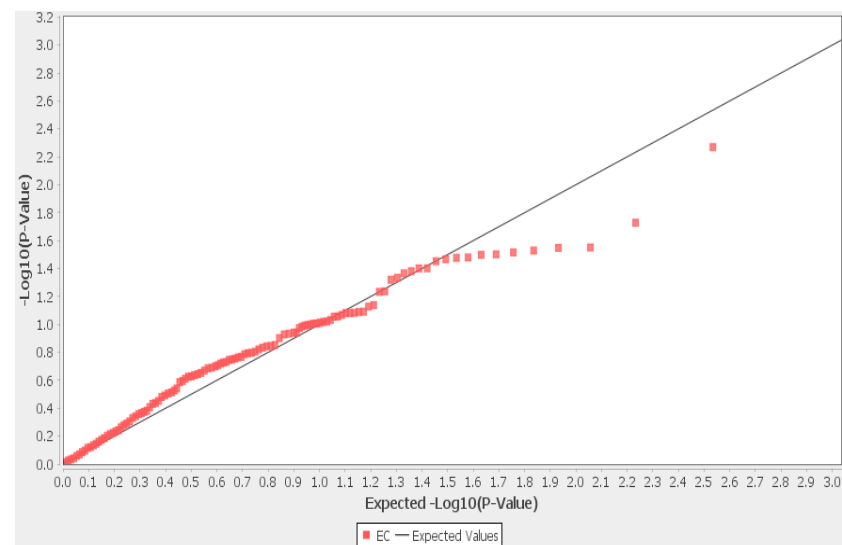
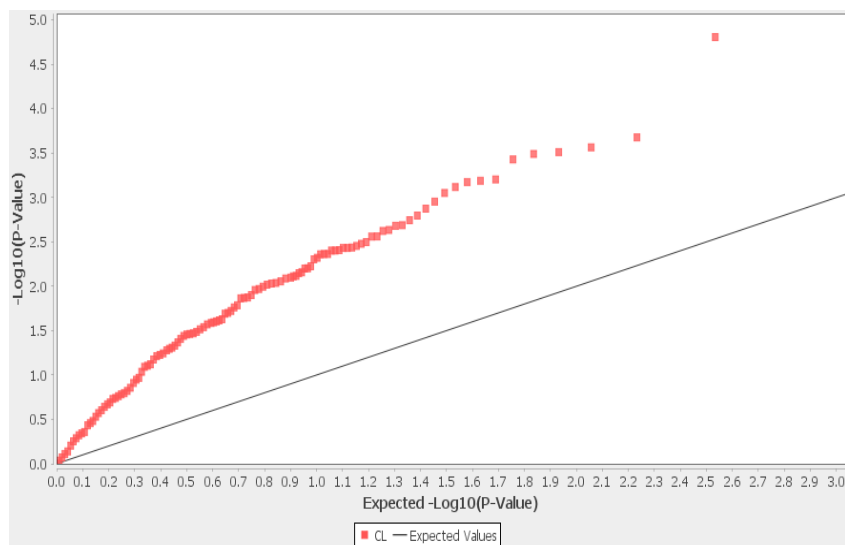
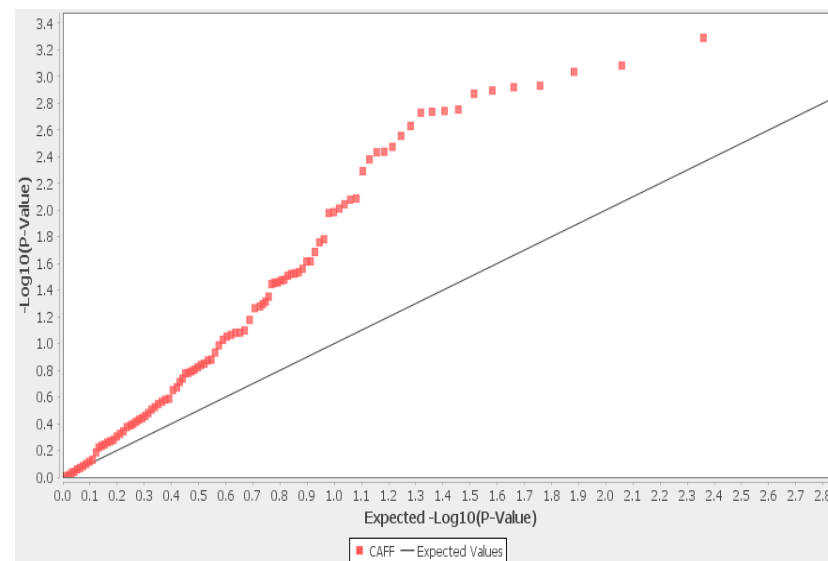
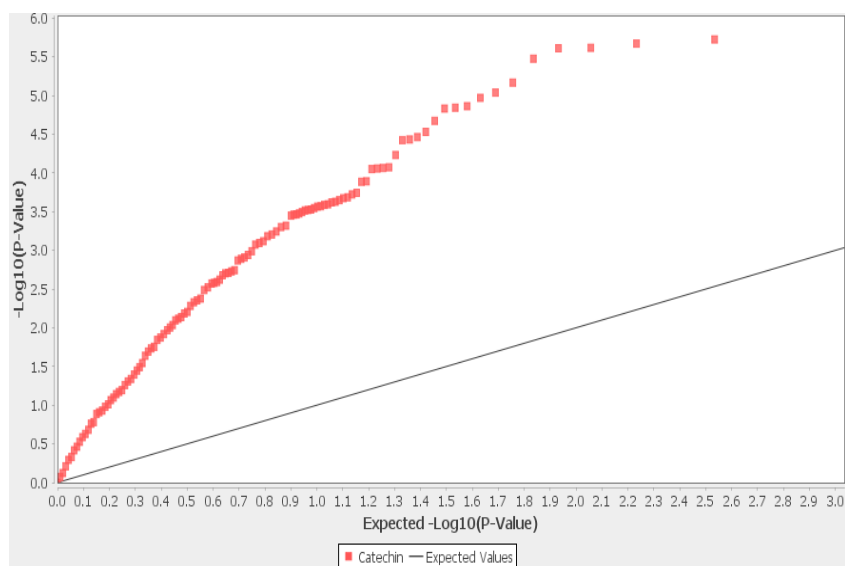
- Pritchard, J., Wen, X. & Falush, D. (2010). Documentation for STRUCTURE software.[Documentation file]. Available with the program at <http://pritch.bsd.uchicago.edu/structure.html>.
- Ramasamy, R. K., Ramasamy, S., Bindroo, B. B. & Naik, V. G. (2014). STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user-friendly interface. *SpringerPlus* 3(1): 431.
- Sadok, I. B., Celton, J.-M., Essalouh, L., El Aabidine, A. Z., Garcia, G., Martinez, S., Grati-Kamoun, N., Rebai, A., Costes, E. & Khadari, B. (2013). QTL mapping of flowering and fruiting traits in olive. *PloS one* 8(5): e62831.
- Sah, S. K., Reddy, K. R. & Li, J. (2016). Absciscic acid and abiotic stress tolerance in crop plants. *Frontiers in Plant Science* 7: 571.
- Shi, W., Hao, C., Zhang, Y., Cheng, J., Zhang, Z., Liu, J., Yi, X., Cheng, X., Sun, D. & Xu, Y. (2017). A Combined Association Mapping and Linkage Analysis of Kernel-Number Per Spike in Common Wheat (*Triticum aestivum* L.). *Frontiers in Plant Science* 8: 1412.
- Soto-Cerda, B. J. & Cloutier, S. (2012). Association mapping in plant genomes. In *Genetic diversity in plants*: InTech.
- Sterken, R., Kiekens, R., Boruc, J., Zhang, F., Vercauteren, A., Vercauteren, I., De Smet, L., Dhondt, S., Inzé, D. & De Veylder, L. (2012). Combined linkage and association mapping reveals CYCD5; 1 as a quantitative trait gene for endoreduplication in *Arabidopsis*. *Proceedings of the National Academy of Sciences* 109(12): 4678-4683.
- TRFK (2012). TRFK clonal catalog. (Ed TRFK).
- Van Ooijen, J. (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. *Kyazma BV, Wageningen* 33: 10.1371.
- Van Ooijen, J. W., Boer, M., Jansen, R. & Maliepaard, C. (2000). MapQTL 4.0: software for the calculation of QTL positions on genetic maps (user manual). Plant Research International.
- Vuong, Q. V., Stathopoulos, C. E., Nguyen, M. H., Golding, J. B. & Roach, P. D. (2011). Isolation of green tea catechins and their utilization in the food industry. *Food Reviews International* 27(3): 227-247.

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.

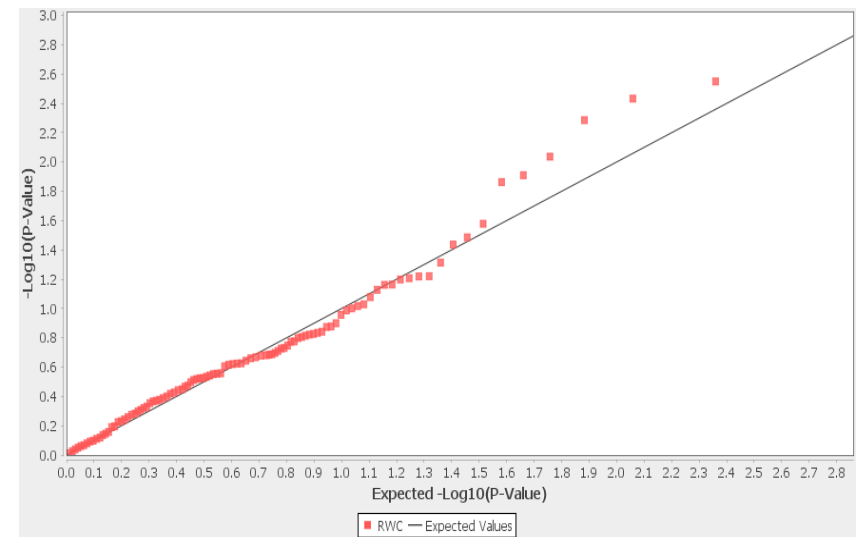
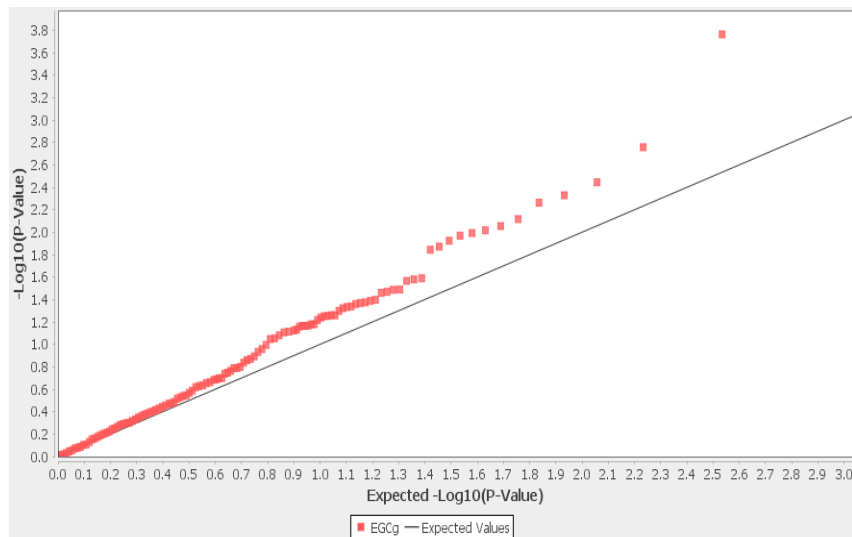
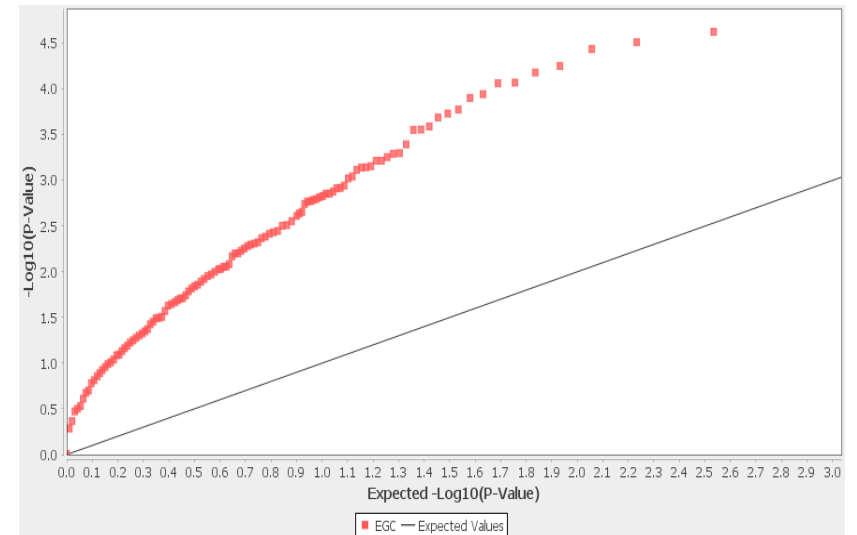
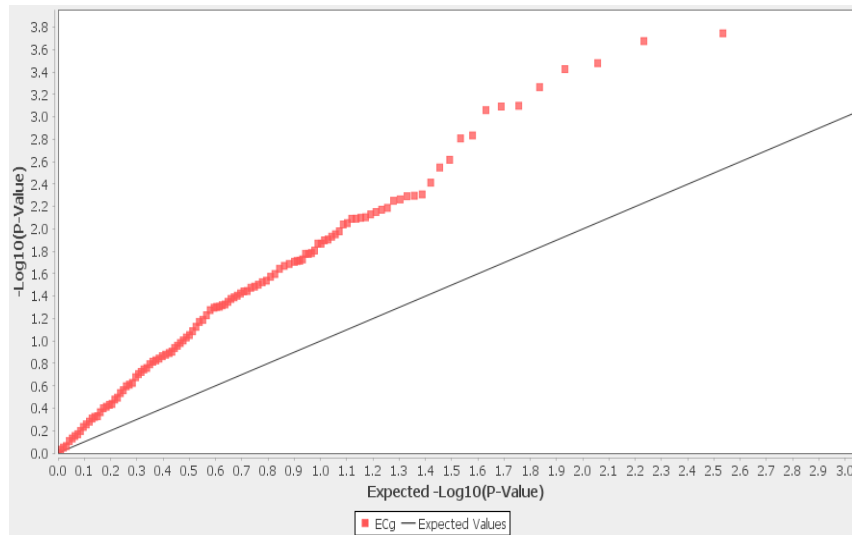
Appendices



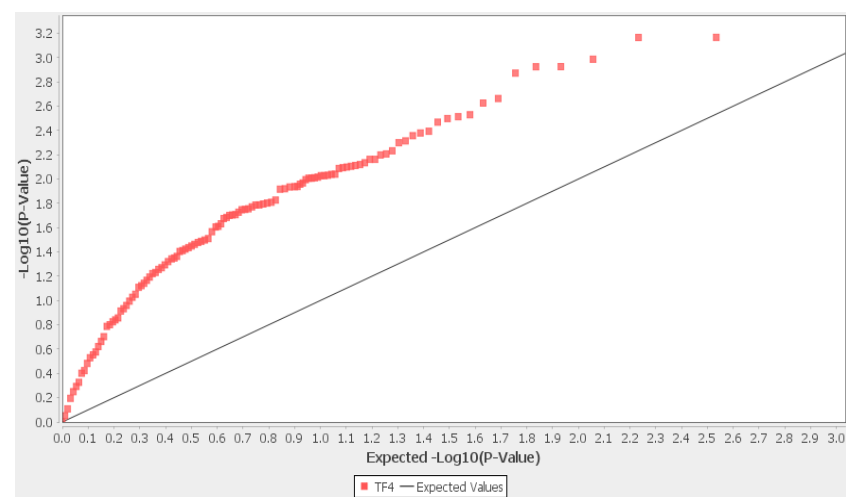
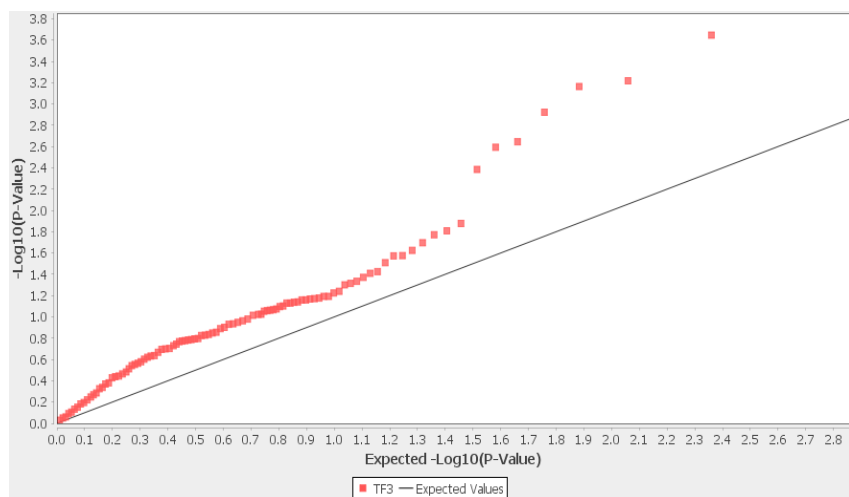
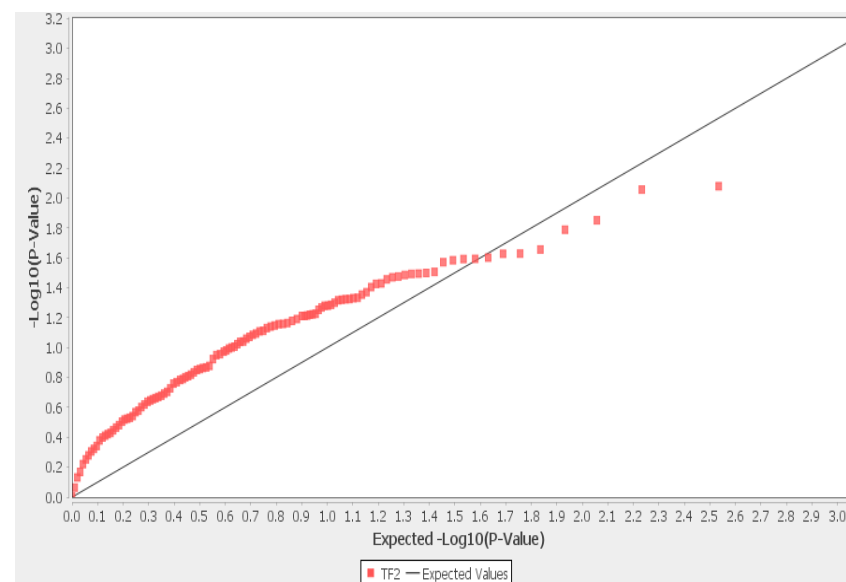
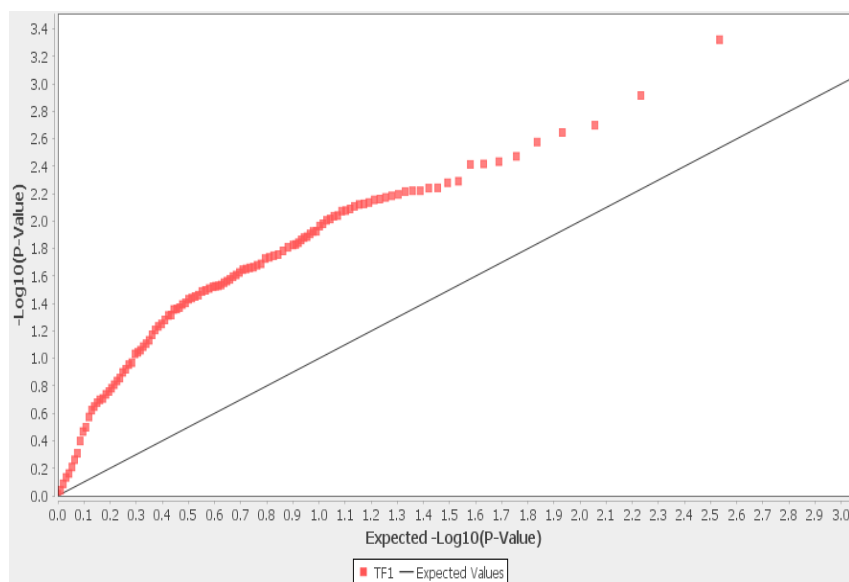
Appendix 5.1: Trait-wise association Q-Q plots for GLM. AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness



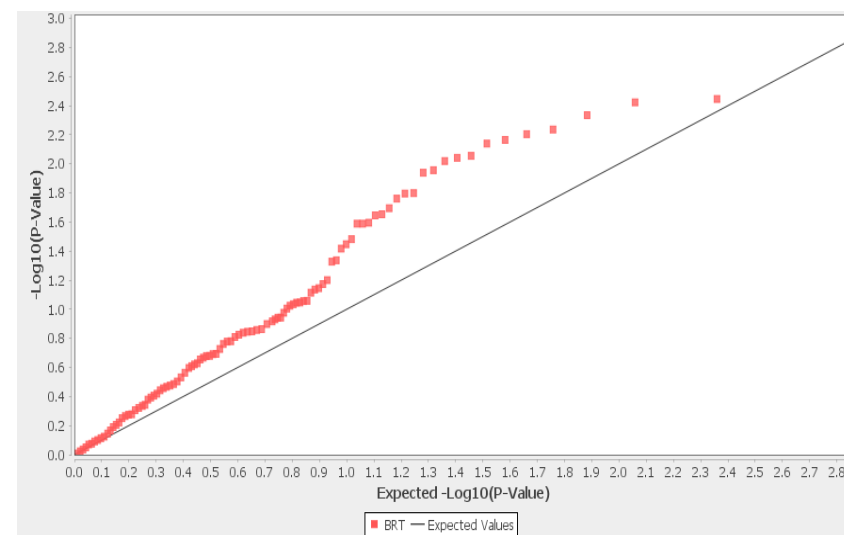
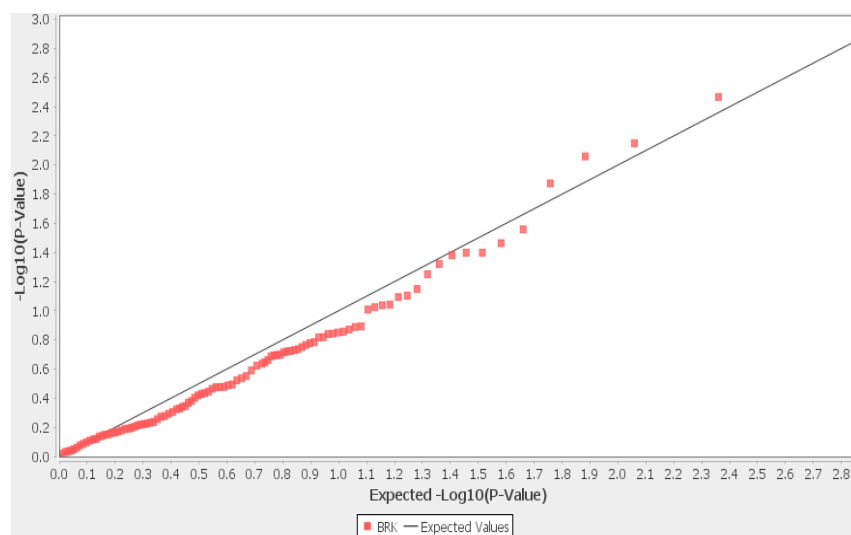
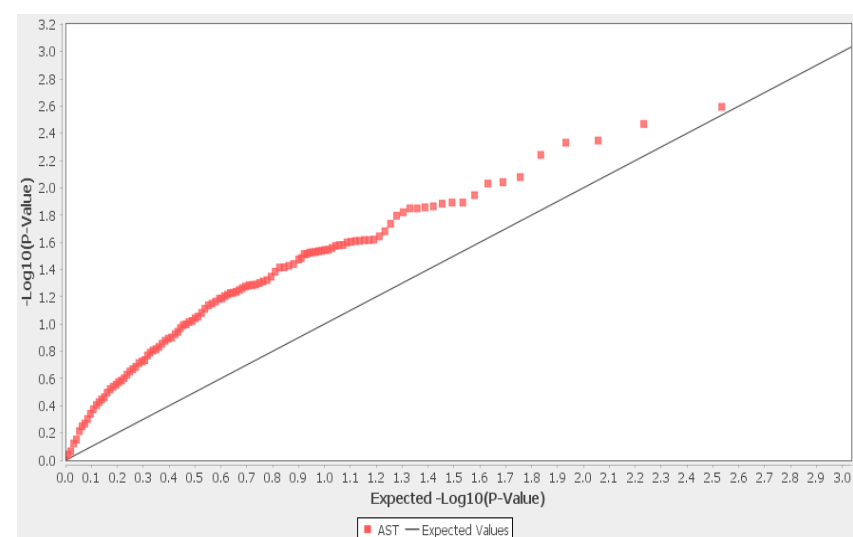
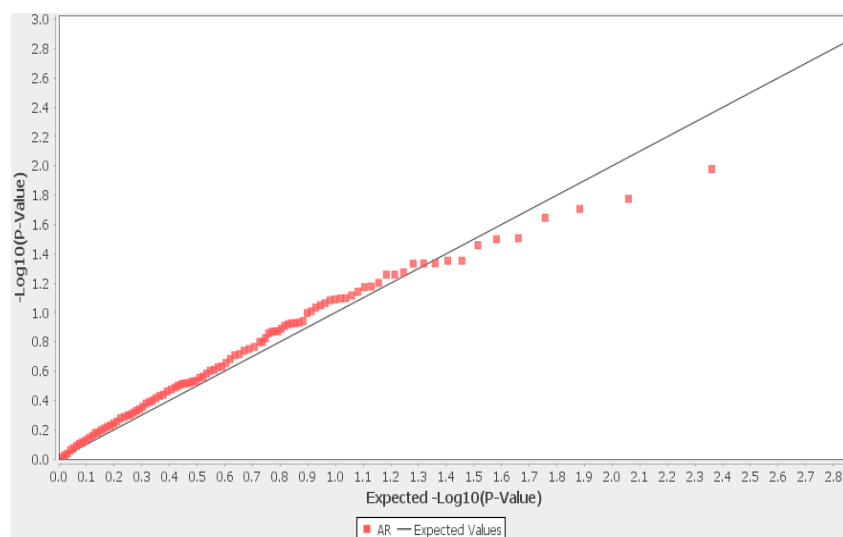
Continued Appendix 5.1: Trait-wise Q-Q association plots for GLM. C- Catechin; CAFF- Caffeine; CL- Colour; EC- Epicatechin



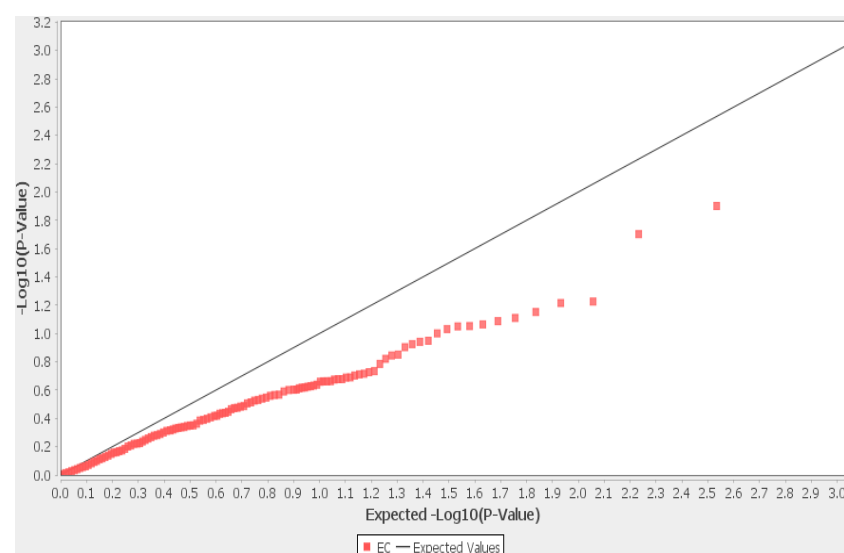
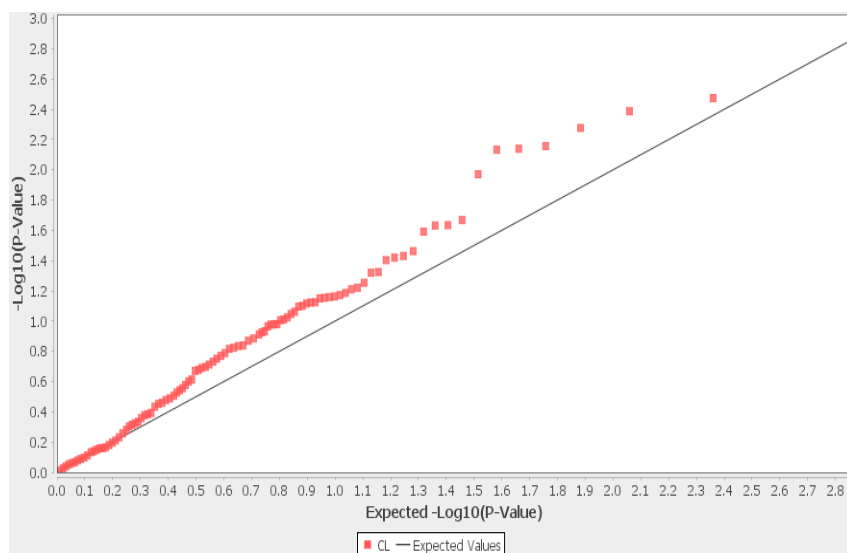
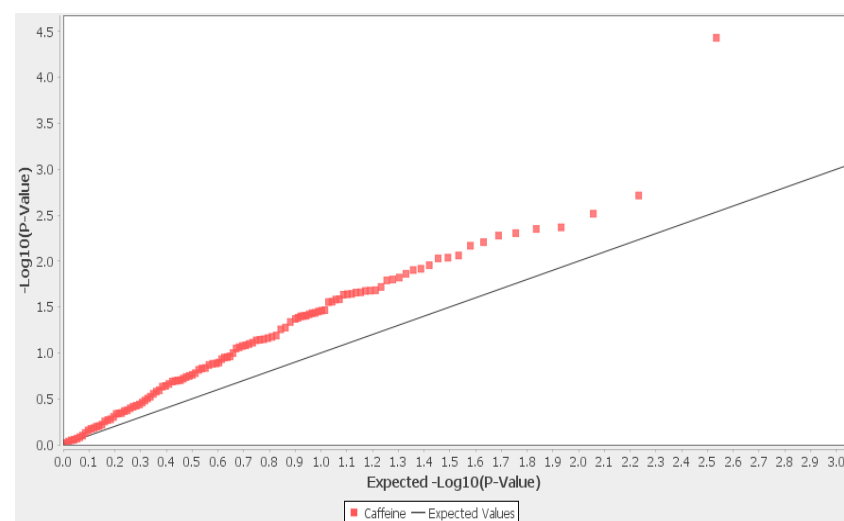
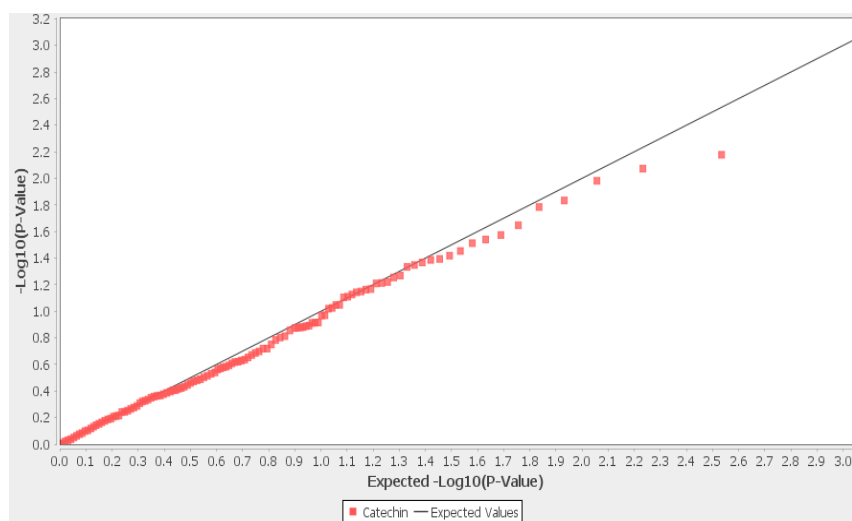
Continued Appendix 5.1: Trait-wise association Q-Q plots for GLM. ECG- Epicatechin gallate; EGC- Epicatechin gallate; EGCg- Epigallocatechin gallate; RWC- Percent relative water content



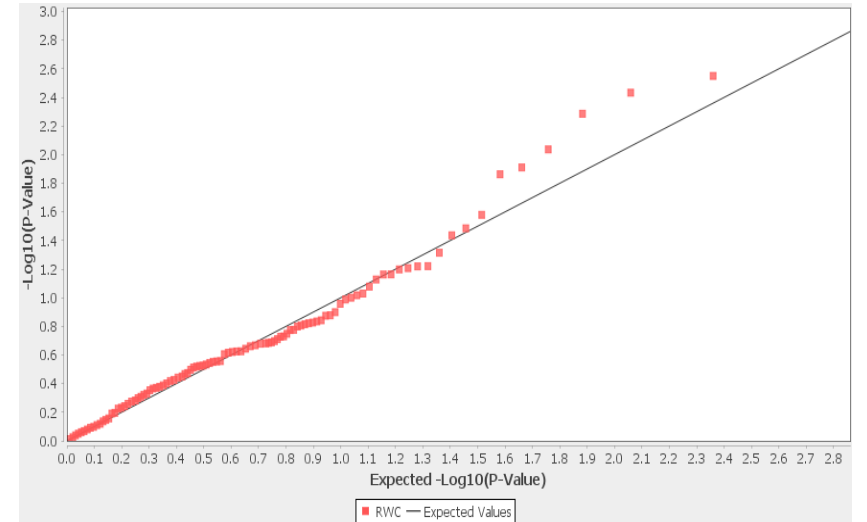
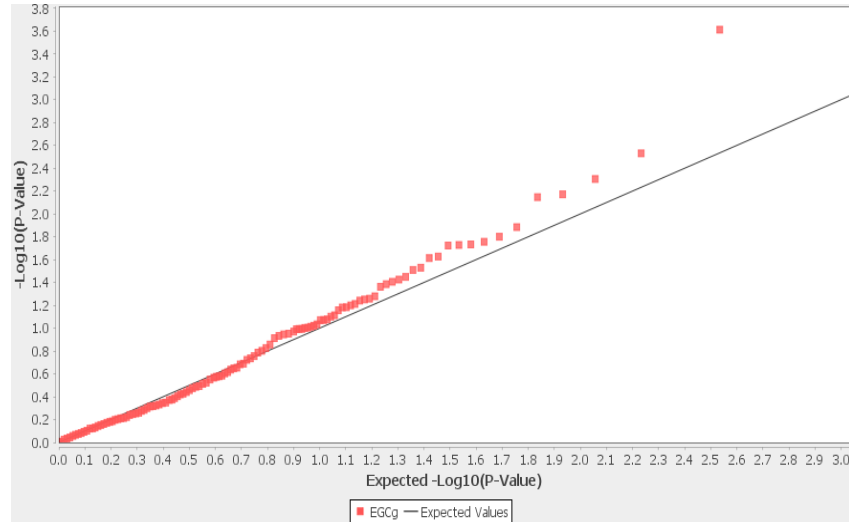
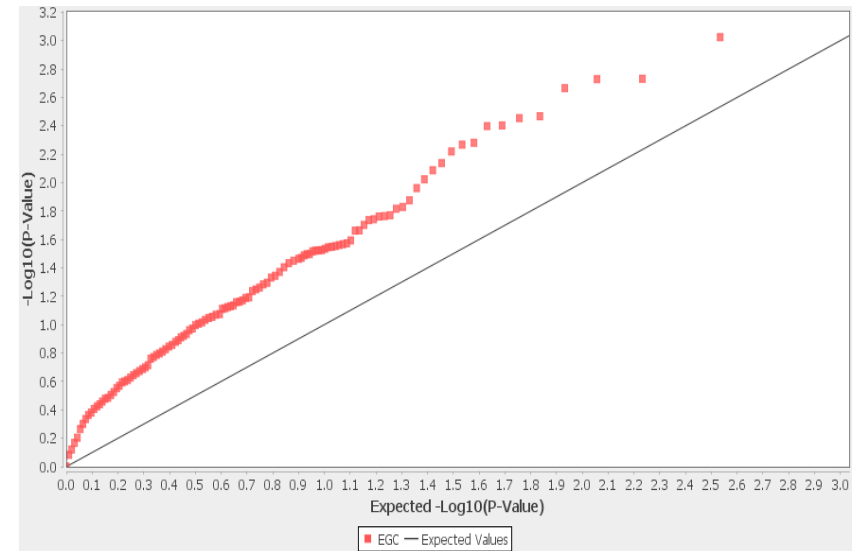
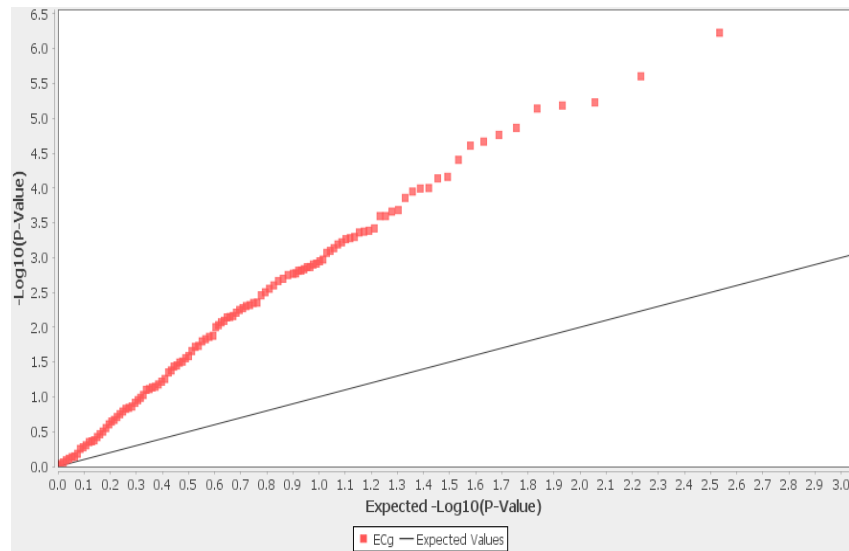
Continued Appendix 5.1: Trait-wise association Q-Q plots for GLM. TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate



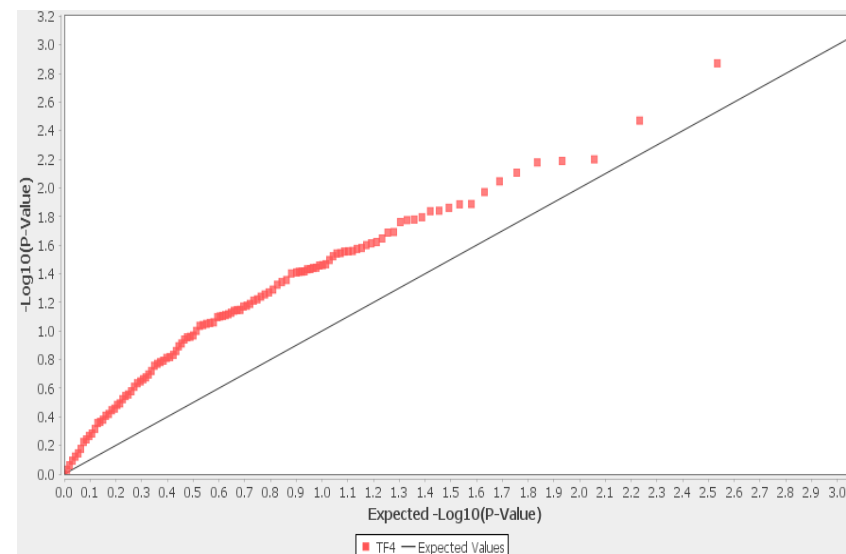
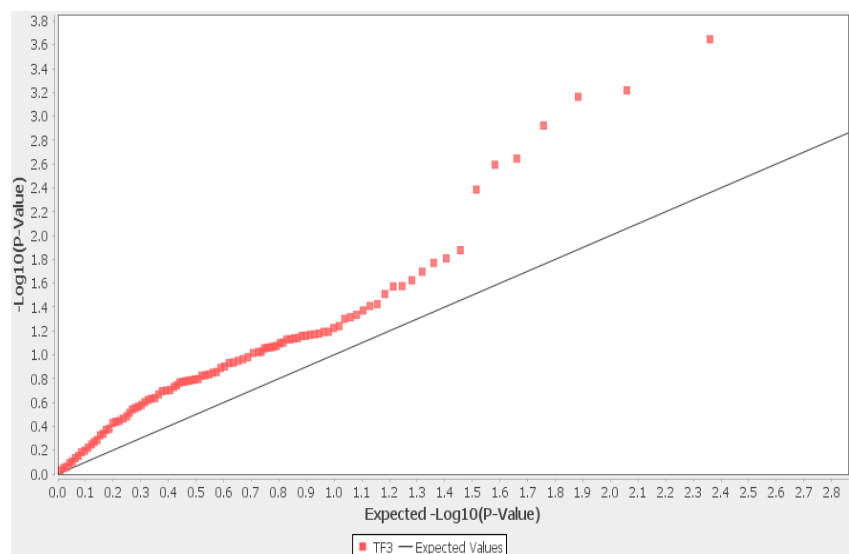
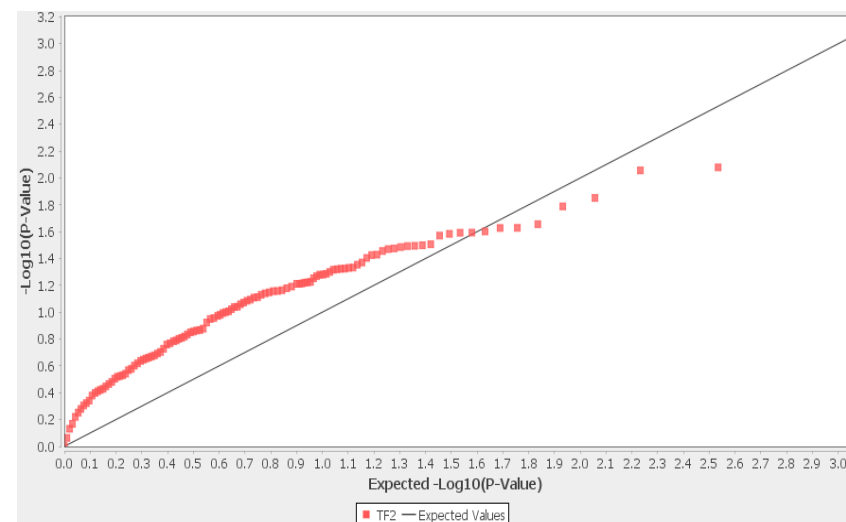
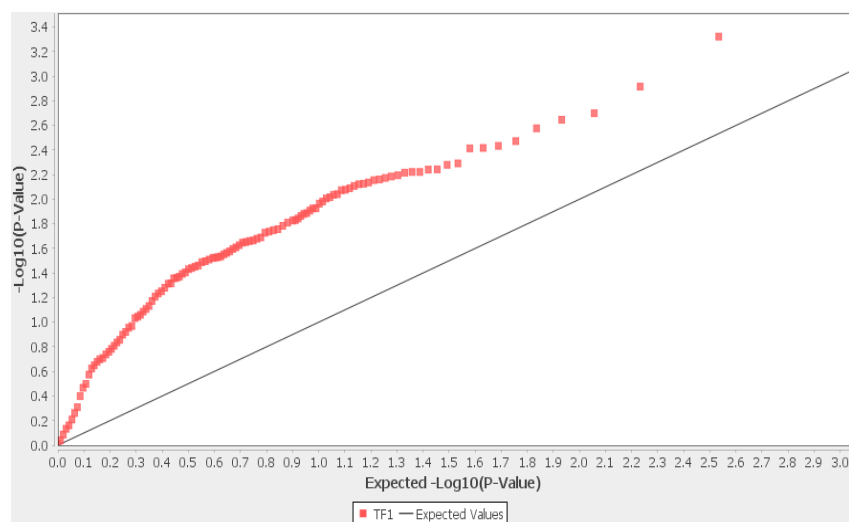
Appendix 5.2: Trait-wise association Q-Q plots for GLM (Q). AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness



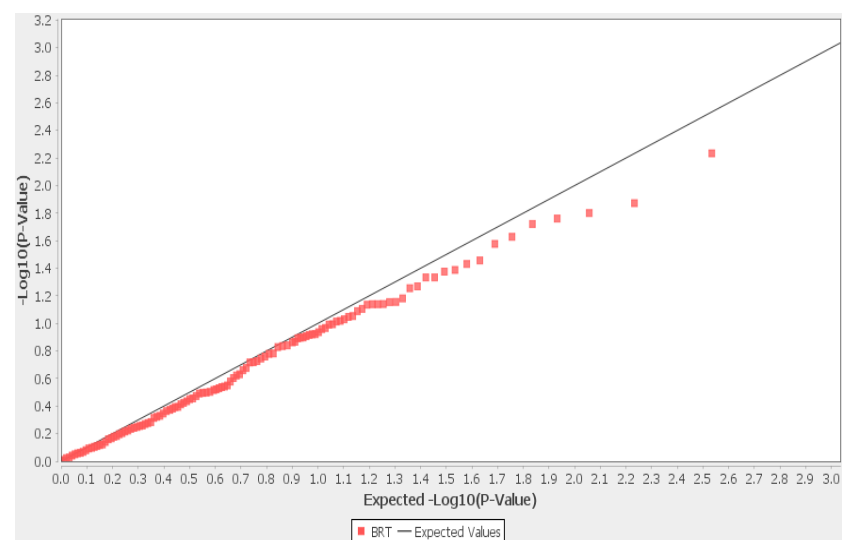
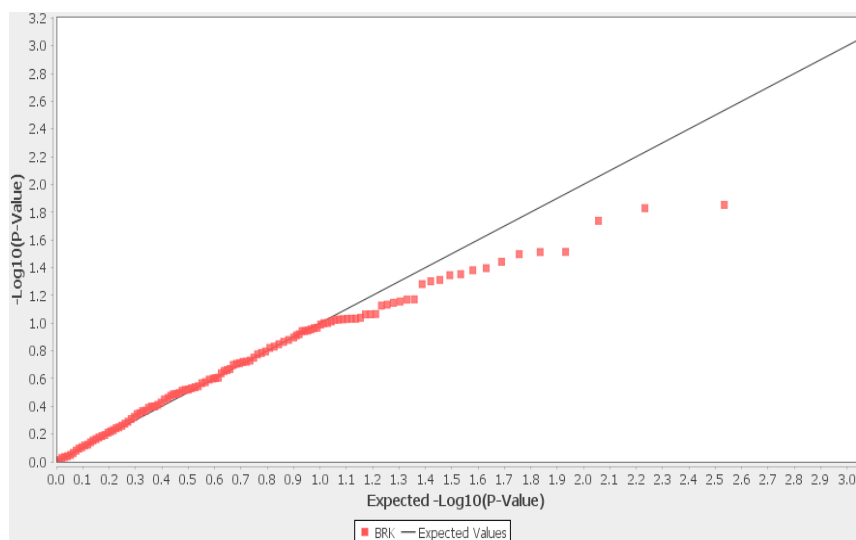
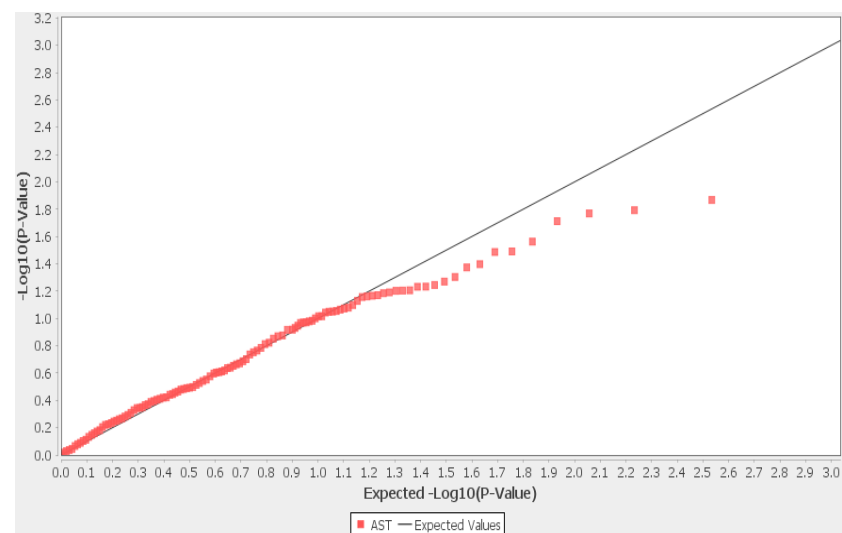
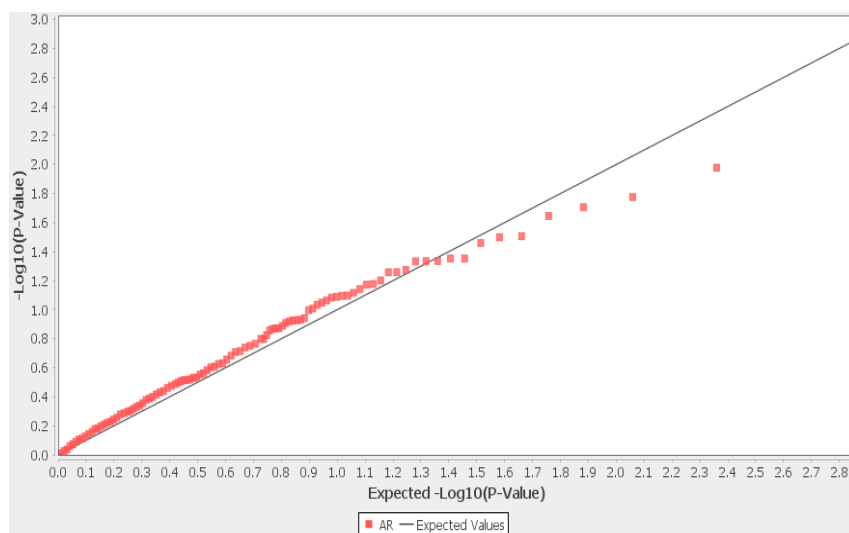
Continued Appendix 5.2: Trait-wise Q-Q association plots for GLM (Q). C- Catechin; CAFF- Caffeine; CL- Colour; EC- Epicatechin



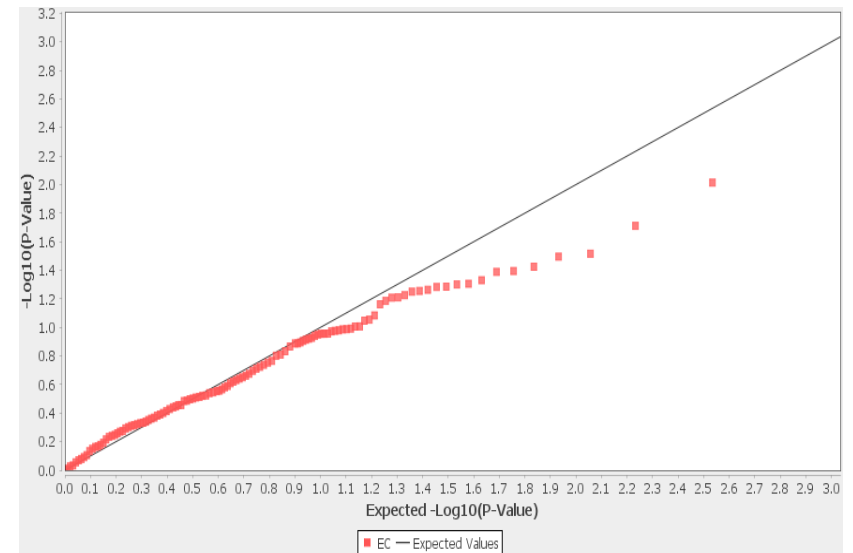
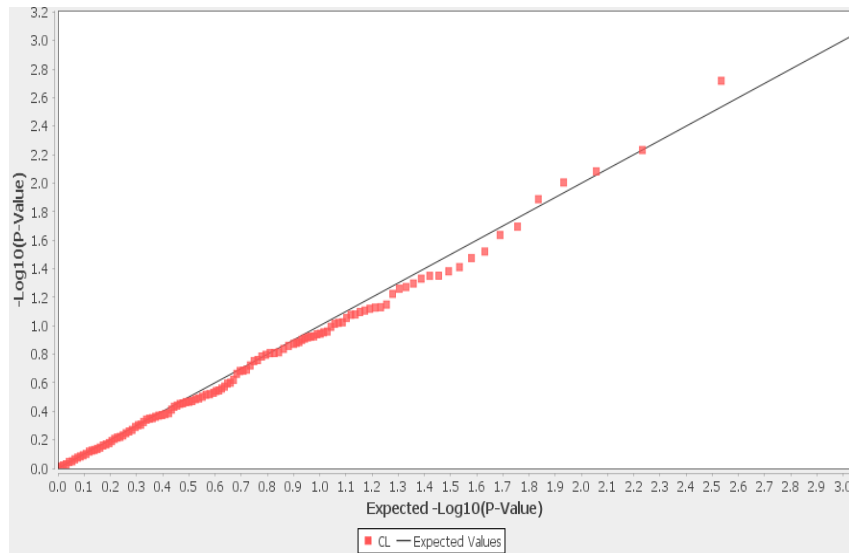
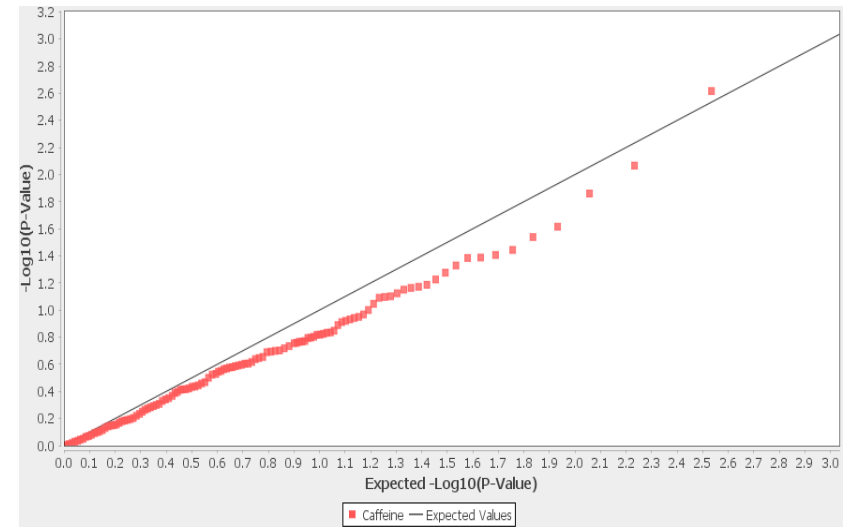
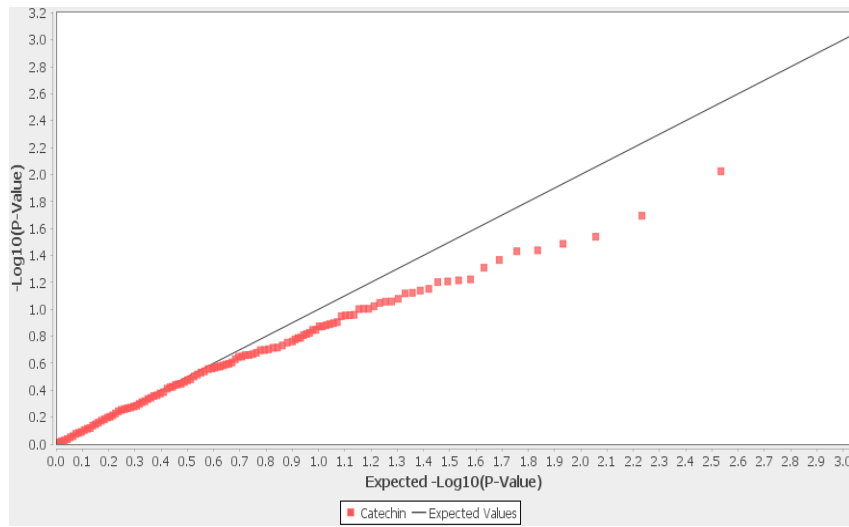
Continued Appendix 5.2: Trait-wise Q-Q association plots for GLM (Q). ECG- Epicatechin gallate; EGC- Epicatechin gallate; EGCG- Epigallocatechin gallate; RWC- percent relative water content



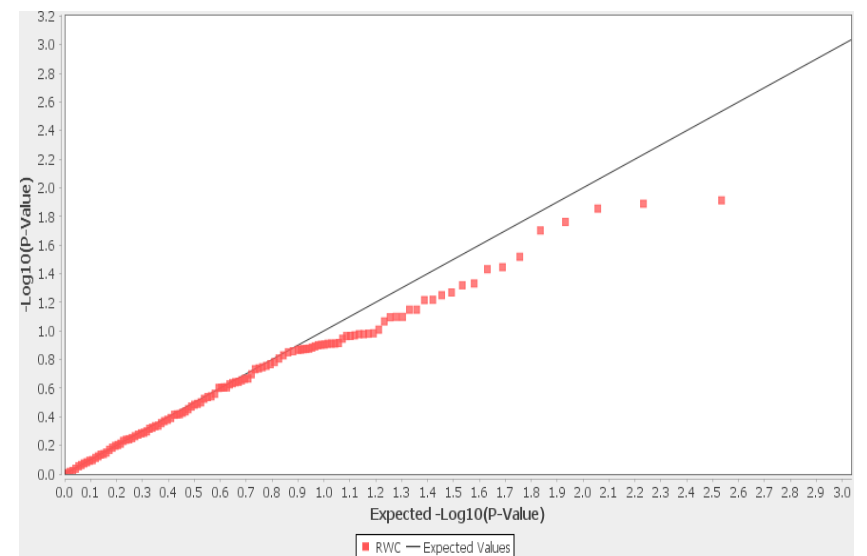
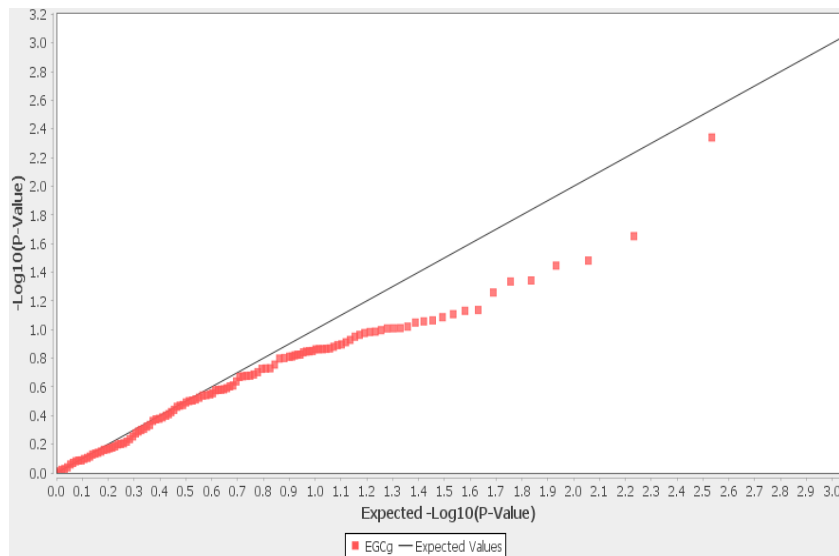
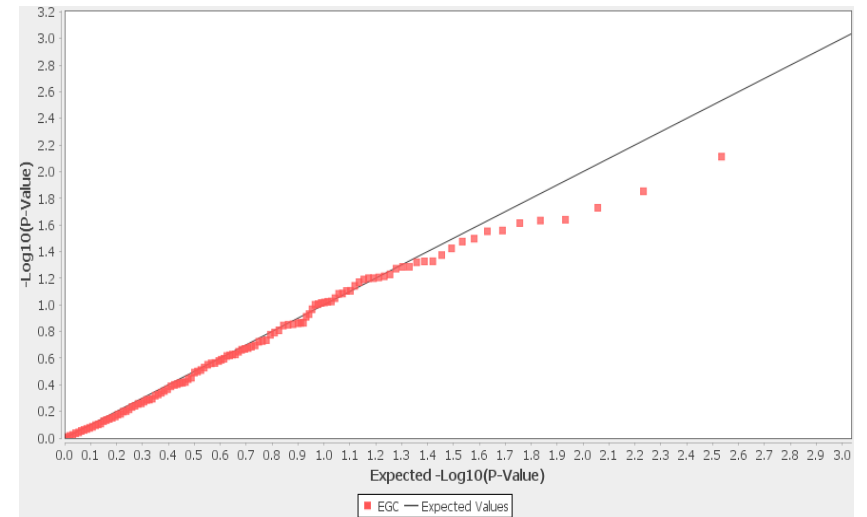
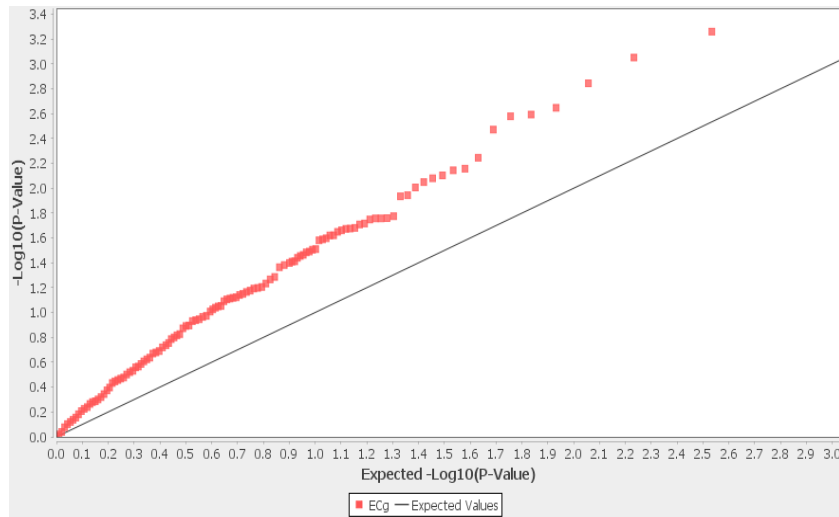
Continued Appendix 5.2: Trait-wise Q-Q association plots for GLM (Q). TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate



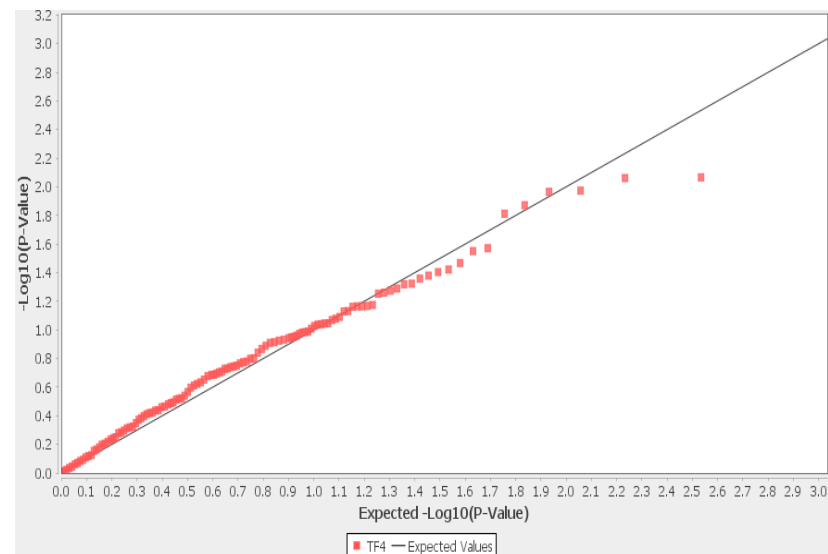
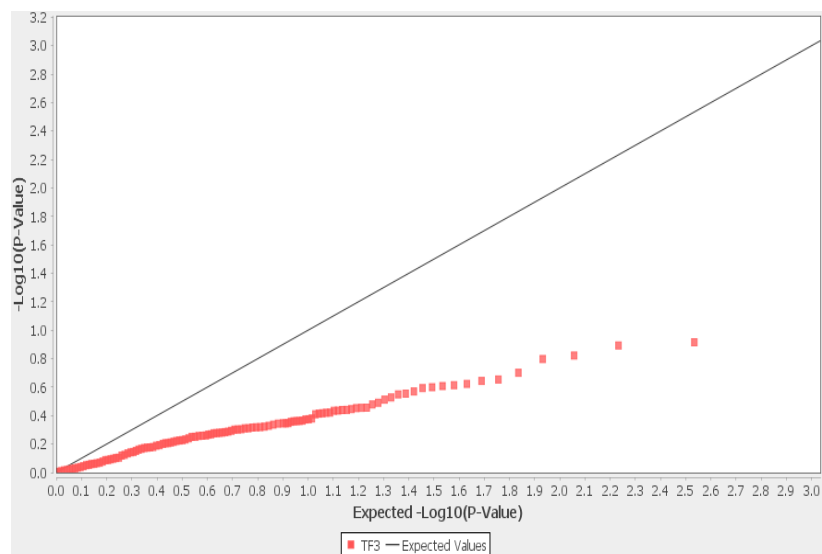
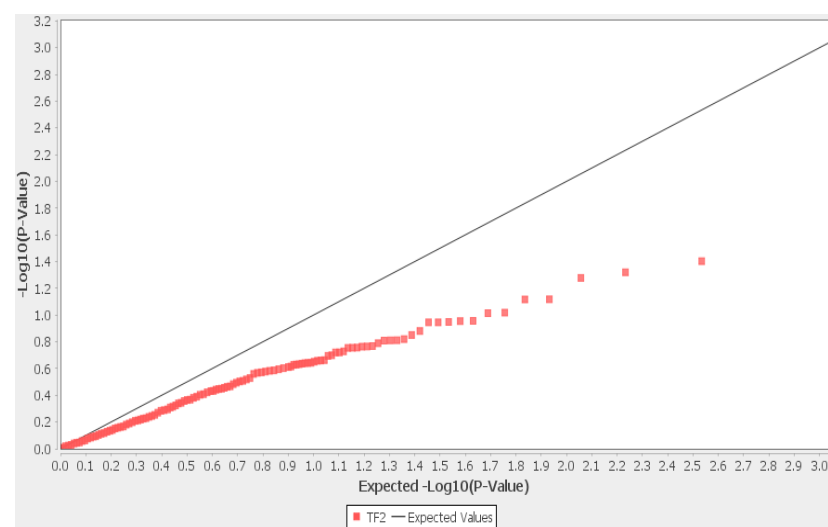
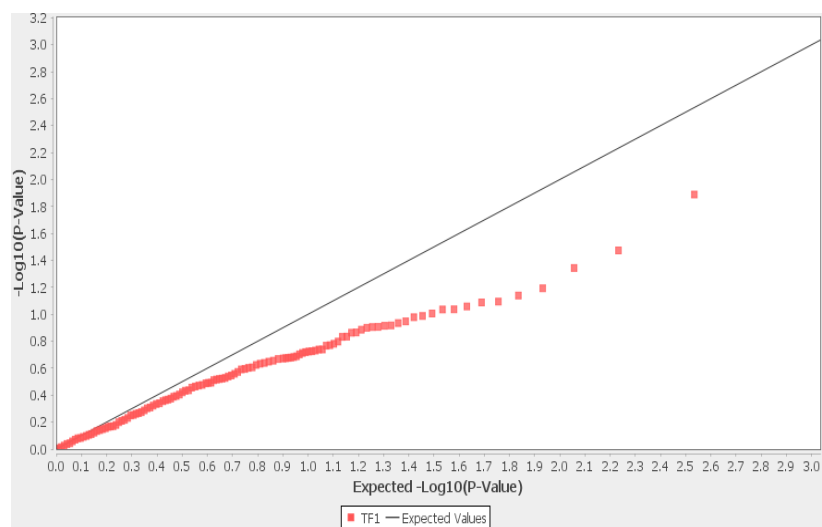
Appendix 5.3: Trait-wise Q-Q association plots for mixed linear model MLM (K). AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness



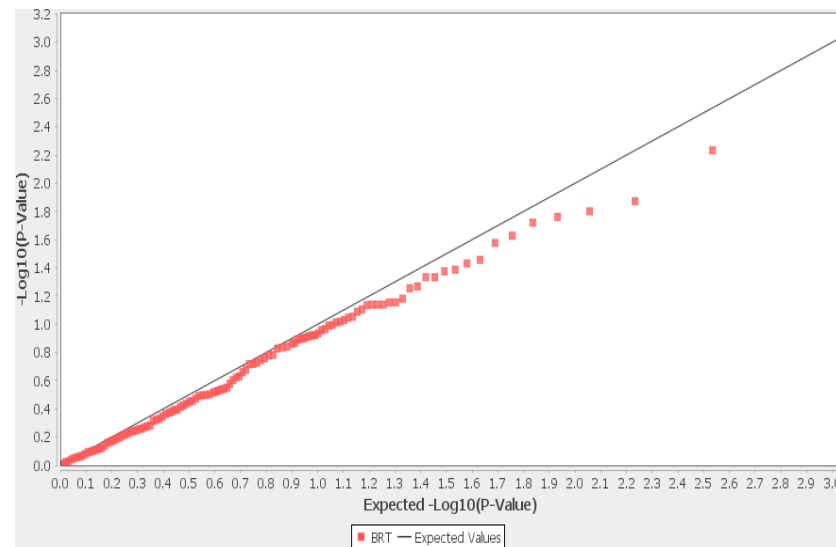
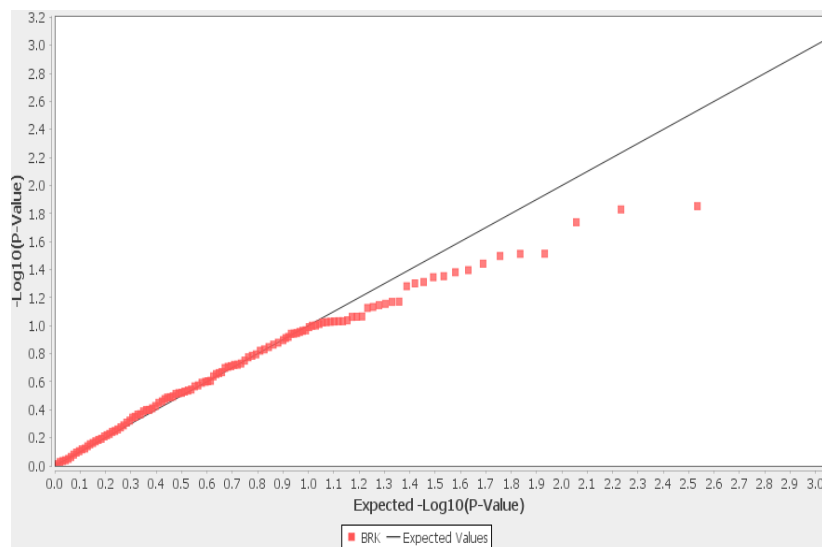
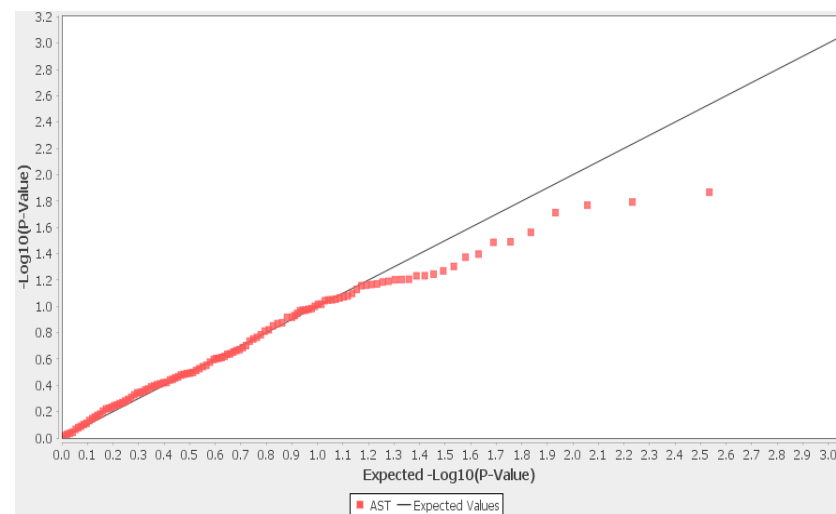
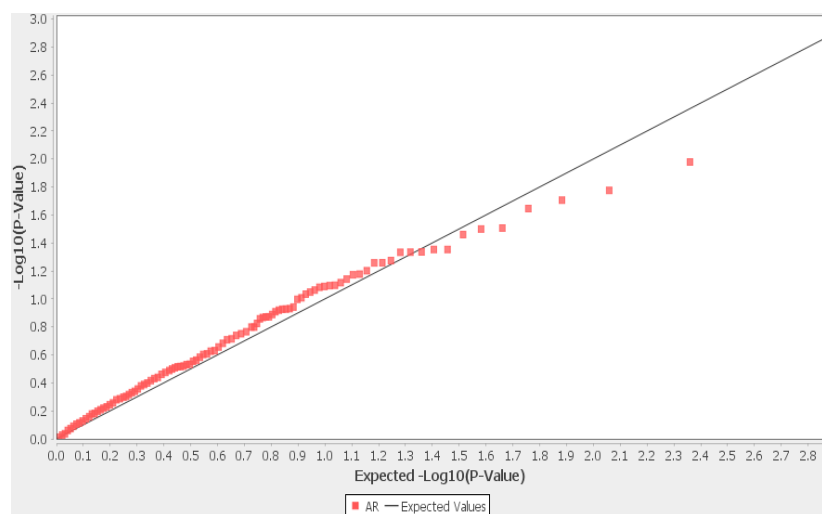
Continued Appendix 5.3: Trait-wise Q-Q association plots for mixed linear model MLM (K). C- Catechin; CAFF- Caffeine; CL- Colour; EC- Epicatechin



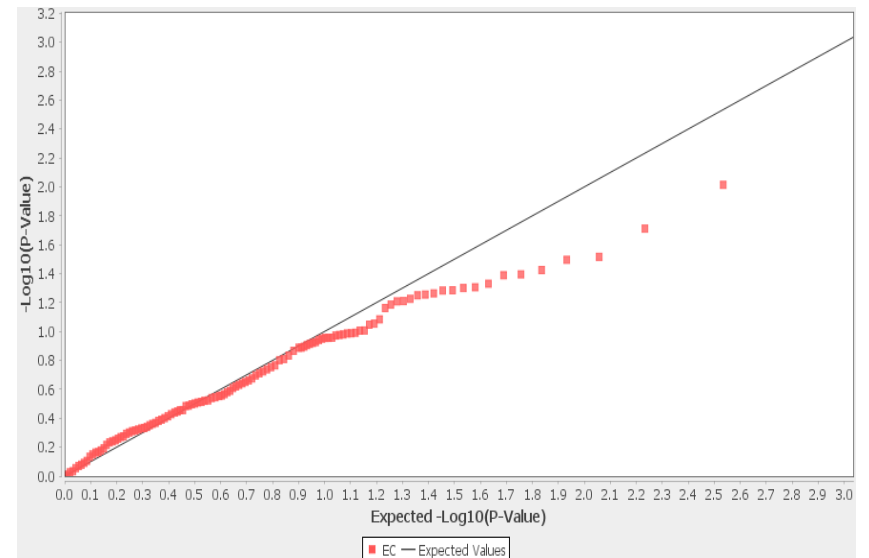
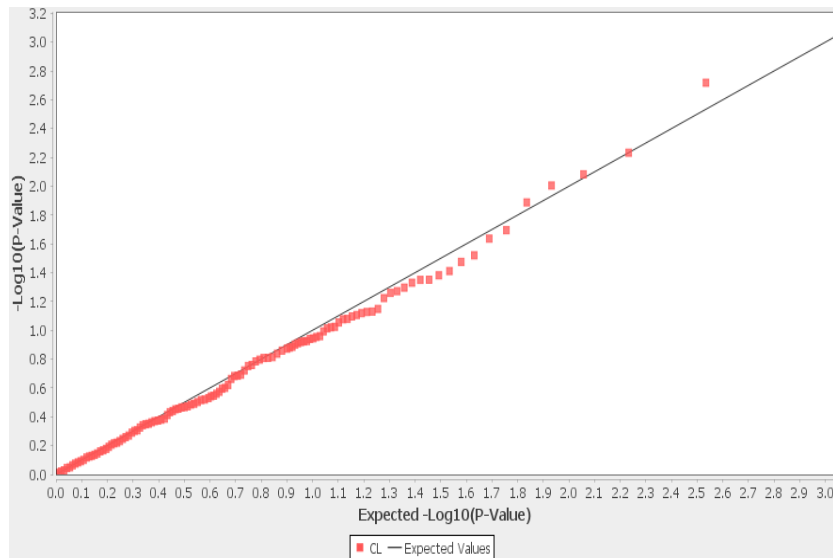
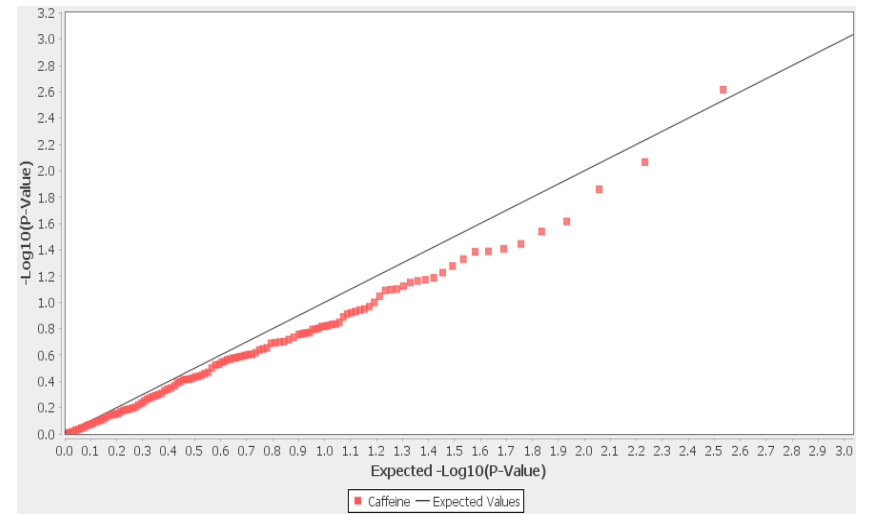
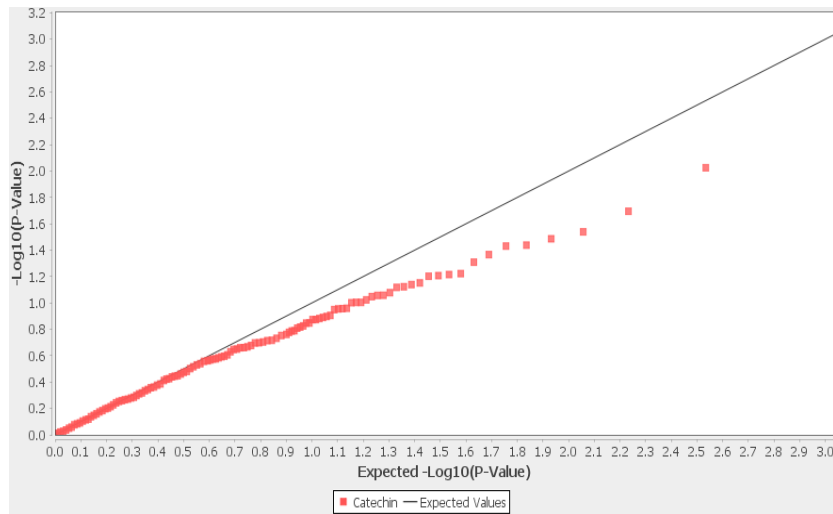
Continued Appendix 5.3: Trait-wise Q-Q association plots for mixed linear model MLM (K). ECG- Epicatechin gallate; EGC- Epicatechin gallate; EGCg- Epigallocatechin gallate; RWC- Percent relative water content



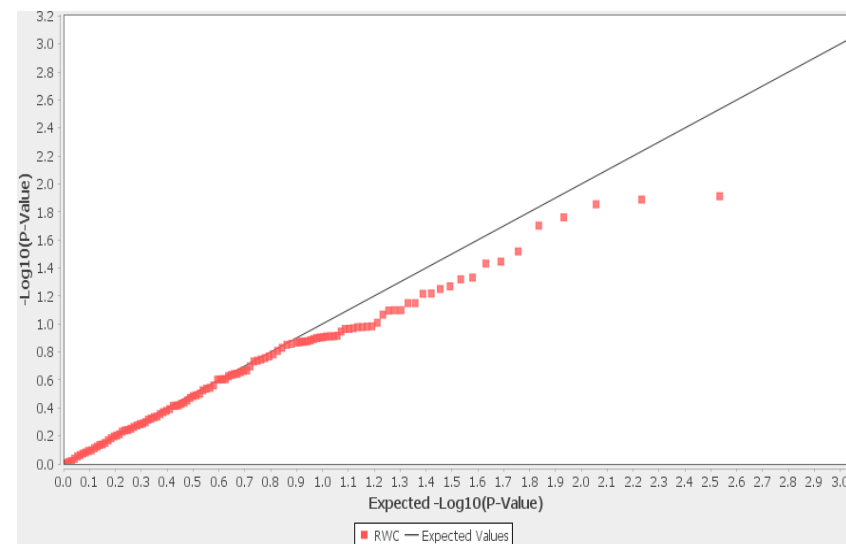
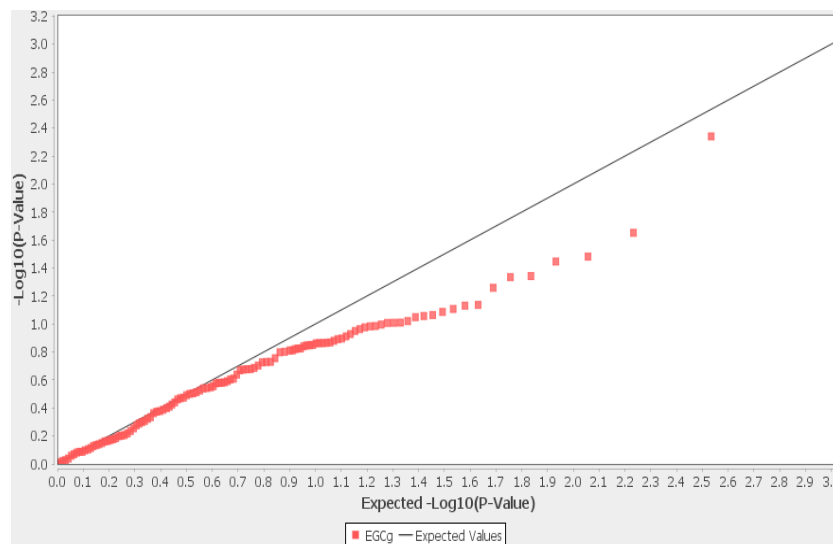
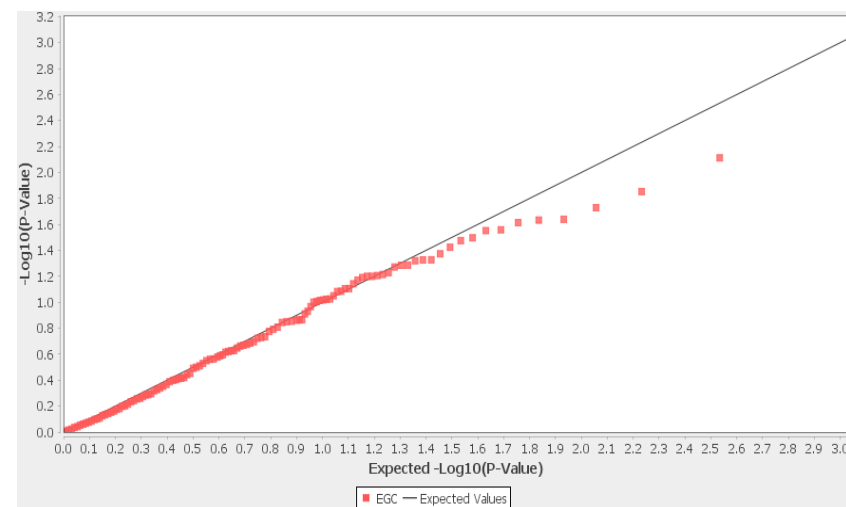
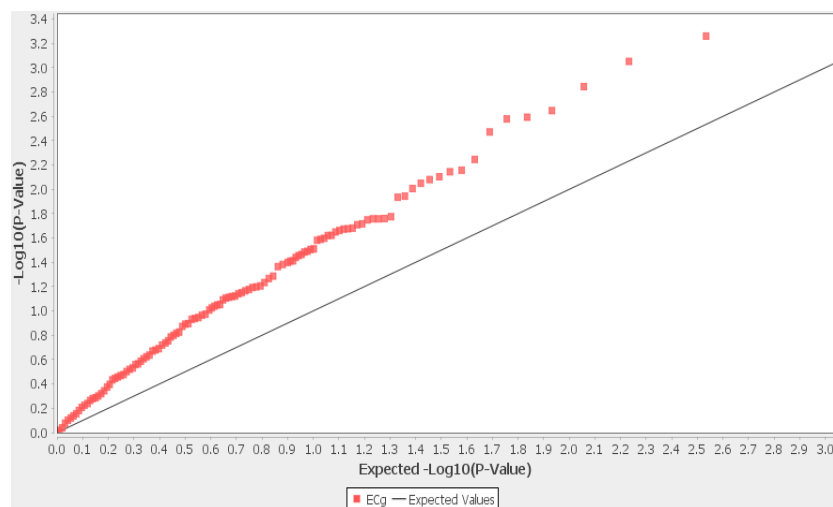
Continued Appendix 5.3: Trait-wise Q-Q association plots for mixed linear model MLM (K). TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate



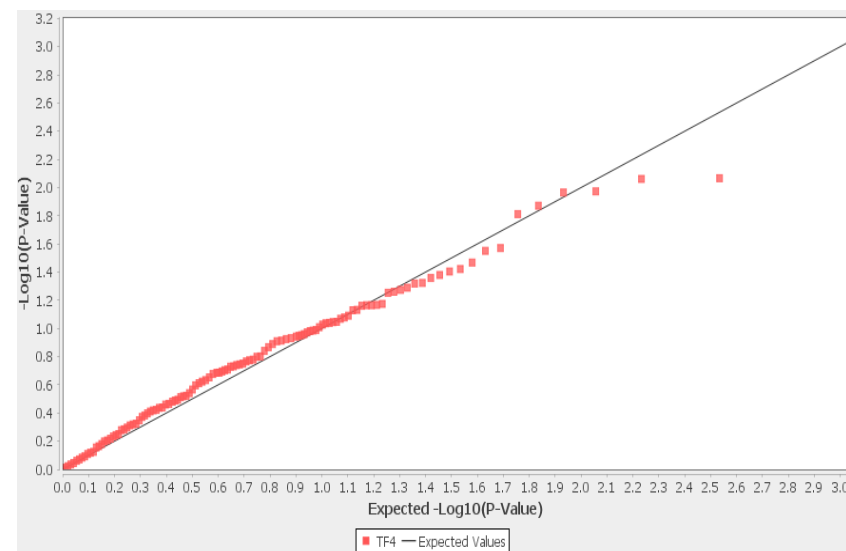
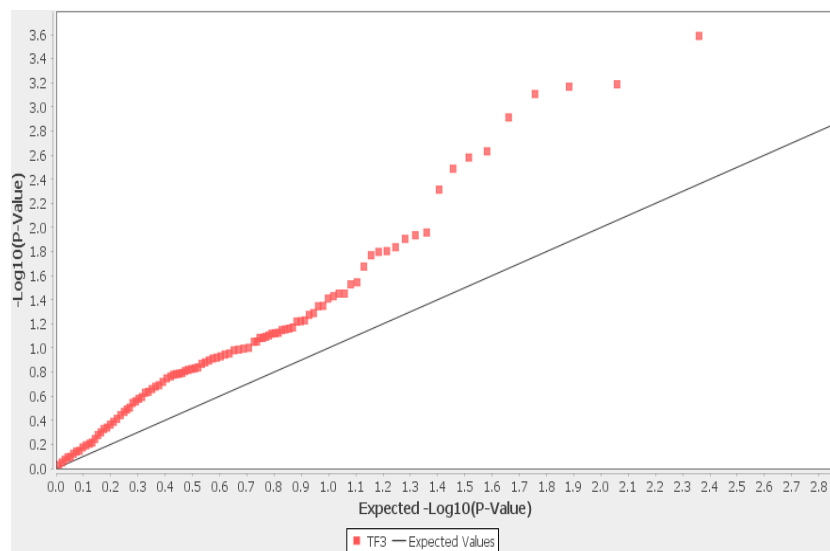
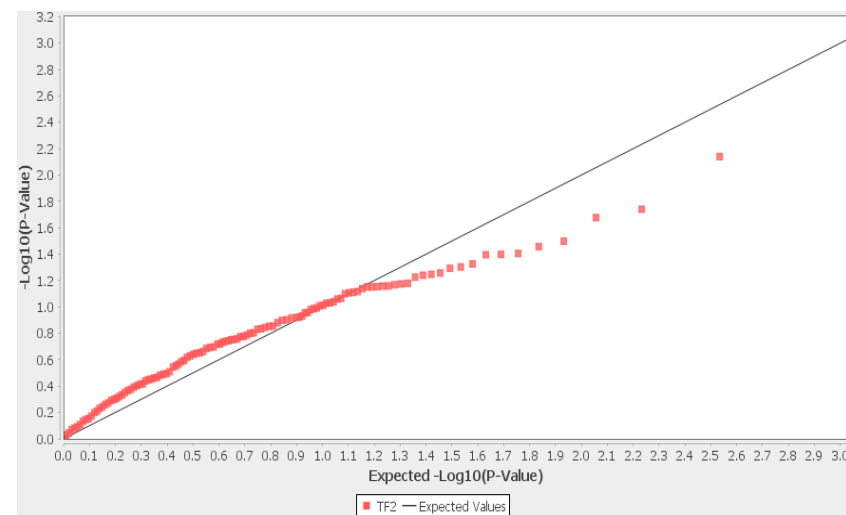
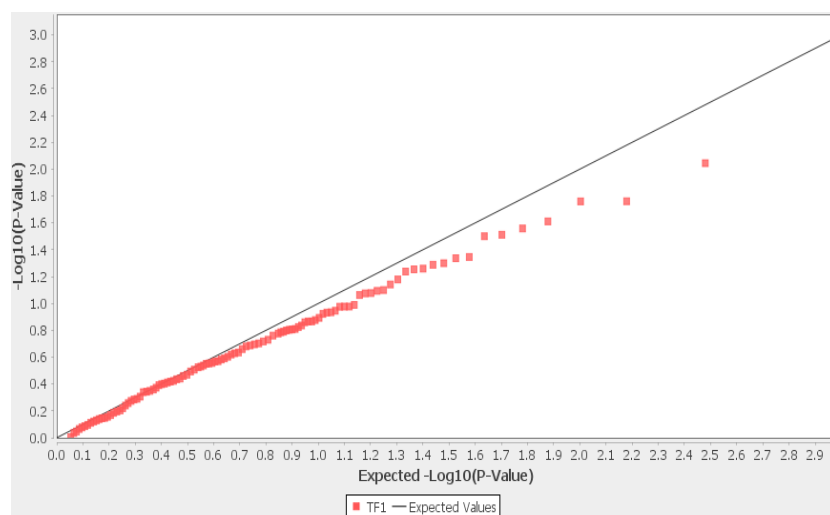
Appendix 5.4: Trait-wise Q-Q association plots for mixed linear model MLM (Q+K). AR- Aroma; AST- Astringency; BRK- Briskness; BRT- Brightness



Continued Appendix 5.4: Trait-wise Q-Q association plots for mixed linear model MLM (Q+K). C- Catechin; CAFF- Caffeine; CL- Colour; EC- Epicatechi



Continued Appendix 5.4: Trait-wise Q-Q association plots for mixed linear model MLM (Q+K). ECG- Epicatechin gallate; EGC- Epicatechin gallate; ECGG- Epigallocatechin gallate; RWC- Percent relative water content



Continued Appendix 5.4: Trait-wise Q-Q association plots for mixed linear model MLM (Q+K). TF1- Theaflavin; TF2- Theaflavin-3-gallate; TF3- Theaflavin-3'-gallate; TF4- Theaflavin-3,3'-digallate

CHAPTER 6

APPLICATION OF MACHINE LEARNING METHODOLOGIES IN GENOMIC SELECTION OF BLACK TEA QUALITY AND DROUGHT TOLERANCE TRAITS

Abstract

Genomic selection can be used to choose parents with desirable traits for further testing as potential commercial cultivars at a very early-stage. Thus, genomic selection has the potential to accelerate breeding efficiency significantly because of increased selection intensity. This study evaluated the different machine learning methodologies for genomic selection of black tea quality traits and drought tolerance trait. All 255 F₁ progenies in the discovery population from two segregating tea populations (TRFK St. 504 and TRFK St. 524) and 56 tea cultivars in the validation population were genotyped using 1 421 DArTseq markers. Two-fold cross-validation was used for training the models in the discovery population, and the best models were consequently, fitted to the validation population. A total of 98 models were fitted to the entire dataset and were evaluated on their predictive ability using cross-validation on the eight different phenotypic traits. Of all the four based prediction approaches, putative QTLs + annotated proteins + KEGG pathway-based prediction approach, showed more robustness and usefulness in prediction of phenotypes than individual prediction based models. The predictive ability of all models varied across the phenotypic traits. ELM model had better prediction ability for catechin, astringency, brightness, briskness, and colour based on putative QTLs + annotated proteins + KEGG pathway approach. The percent variables of importance of putatively annotated proteins and KEGG pathways were associated with the phenotypic traits. The findings from this study have for the first time opened up a new avenue for future application of genomic selection in tea breeding by early selection of parents with desirable traits and thus reducing the time from initial crosses to planting in commercial tea fields.

6.1 Introduction

In plants, QTL mapping has successfully been used in genetic analysis of several non-complex traits where much of the genetic variance is conditioned by one or a few loci. However, exploiting QTL for more complex traits to predict candidates for selection in practical breeding programs has had limited impact because of the small genetic variance accounted for by QTLs owing to the model being used, and many detected QTLs are specific to particular genetic background. Due to the limitations posed by QTLs mapping technique, machine learning methods provide an alternative approach in plant breeding. Machine learning methods have been used in many areas such as genomics, transcriptomics, proteomics, and systems biology (Larranaga *et al.*, 2006; Tarca *et al.*, 2007; Kelchtermans *et al.*, 2014). In animals, machine learning methods have been applied widely in genome research studies which include, genome assembly, genome annotation and gene regulatory network studies, and in prediction of gene function studies (Larranaga *et al.*, 2006; Tarca *et al.*, 2007). However, only a few machine-learning-based applications have been used in plant science studies.

Plants can also provide an excellent framework in machine learning methodologies because of their ability to provide a large number of offspring to be tested or studied. Thus, the currently available plant research genome data can be used to integrate genotype and phenotype data and provide a valuable resource to plant breeders to uncover novel trait-associated candidate genes (Hu *et al.*, 2018). Recently, machine learning approaches were used in the identification of healthy and non-healthy citrus leaves (ur Rahman *et al.*, 2017), disease resistance in wheat (Ornella *et al.*, 2017), breeding of drought-tolerant maize (Ornella *et al.*, 2012), grain yield in maize and wheat (Shekoofa *et al.*, 2014; Pantazi *et al.*, 2016) and in stress severity studies in soybean (Naik *et al.*, 2017). Machine learning has also been used in fruit trees (apples) which have a long breeding cycle to select for fruit quality traits (Kumar *et al.*, 2012). In tea plants, tremendous progress has also been made regarding the availability of good genetic resources and linking molecular markers to QTL of complex traits (Hazra *et al.*, 2018). However, tea breeding programs have had challenges in breeding tea cultivars with desirable traits, because tea plants are self-incompatible, highly heterozygous, and has a long juvenile period of 4-5 years and about 22-25 years to breed a new cultivar (Chen *et al.*, 2007). This, therefore, has led to slow progress in breeding tea cultivars for production of high-quality black tea and drought-tolerant.

The price of black tea in the world market depends primarily on the physical appearance of the made tea and the infused leaf which include the aroma and liquor characters (colour, brightness, strength, and briskness) (Bhatia, 1963). Therefore, this implies that only producers of high-quality black teas are likely to survive in the global tea market. On the other hand, climate change and population growth have put more pressure on traditional tea growing regions in Kenya with less land available to conduct field trials on newly developed tea cultivars (FAO, 2015b). This has contributed to tea farming and production being moved to marginal areas which are unsuitable for tea farming and production (FAO, 2015b). Therefore, the objective of tea breeding programs is to develop tea cultivars for production of high-quality black tea and tea cultivars that can adapt to changing environmental conditions such as drought stress. This can be accelerated by applying new and novel techniques such as machine learning methodologies which are more robust, accurate, efficient, and capable of evaluating a much wider set of variables within a short period of time (Crossa *et al.*, 2017). Machine learning methodologies can also help in the identification of genomic regions that are of agronomic value by facilitating functional annotation of genomes and enable real-time high-throughput phenotyping of agronomic traits. This is a study carried out in tea for the first time and is expected to increase the genetic gain per unit time and reduce cost in tea breeding especially when applied on traits with low heritability due to its long selection cycle.

6.2 Research objective

The main objective of this study was to assess the predictive ability of different machine learning methodologies based on four approaches. The first approach was based on both the putative QTLs identified, the putatively annotated proteins and the KEGG pathways for the different traits as described in chapter 3 and chapter 4. The second approach was based on the identified putative QTLs alone for the different traits as described in chapter 3. The third approach was based on the putatively annotated proteins alone for the different traits as described in chapter 4. The fourth and last approach was based on the KEGG pathways alone for the different traits as described in chapter 4.

6.3 Hypotheses

Alternative hypothesis (H_1): The predictive ability of different machine learning methodologies applied are significantly different based on putative QTLs + annotated proteins + KEGG pathways, putative QTLs, annotated proteins and KEGG pathways approaches.

Null hypothesis (H_0): The predictive ability of different machine learning methodologies applied are not significantly different based on putative QTLs + annotated proteins + KEGG pathways, putative QTLs, annotated proteins and KEGG pathways approaches.

6.4 Materials and methods

6.4.1 Genome selection dataset

The genomic dataset used in this study comprises of 255 F_1 individuals (discovery population) from two segregating tea populations (TRFK St. 504 and TRFK St. 524) and 56 tea cultivars (validation population) with each individual phenotyped for black tea quality traits and drought tolerance trait. All the individuals in both the discovery and validation populations were genotyped using 1 421 DArTseq markers as described in chapter 2 (Diversity Array Technology Pty Ltd, Canberra, Australia).

6.4.2 Transforming and ranking phenotypic data

The phenotypic data (y) in 311 tea datasets were centred at zero and standardised to unit variance. The y in 310 datasets were then grouped into three classes (upper, middle and lower), based on 15-85 % percentiles of each phenotypic trait analysed. For example, for 15-85 % percentiles, 0.15 and 0.85 quantiles were used to split y into three classes: upper class, if $y > 0.85$; middle class, if $0.15 < y \leq 0.85$; and lower class; if $y \leq 0.15$.

6.4.3 Partition designs and cross-validation of the dataset

In this study, all analyses were carried out using R v 3.5.0 (2018). The caret package (Williams *et al.*, 2018) was the main package used for training and fitting the various models. Two-fold cross-validation was used for training the models on the discovery population, while the “best” models were consequently fitted to the validation population. The discovery population consisted of all the observations labelled “Discovery”, except for two parents (GW Ejulu and TRFK 303/577) that were added to the validation set. The markers were averaged across the putative QTLs, annotated proteins and the KEGG pathways. For example, if, “pathway A” “QTL A” or “protein A” is expressed in ten out of the 13 000 DArTseq markers consisting of 0 and 1 score

values, i.e. absence or presence of polymorphism in the genomic representation of the sample, then the ten, 0 and 1 score values are summed and divided by ten, which is the average of ten DArTseq markers. If only nine DArTseq markers are present in the data, then the 0 and 1 score values are summed and divided by nine, which is the average of nine DArTseq markers (Appendix 6.1). The missing values in the dataset were imputed with 0 (i.e. not-expressed). The trained prediction models obtained were then applied to predict phenotypic trait values of individuals from the testing set labelled “Validation” using genomic, phenotypic, putative QTLs, annotated proteins and KEGG pathways data.

6.4.4 Prediction models

The prediction models were built for the various phenotypes as dependent variables which resulted in four models namely: putative QTLs + annotated protein + KEGG pathway, putative QTLs, annotated protein and KEGG pathway. In all cases, the aggregated markers were used as independent variables (inputs to the models).

6.4.5 Comparison of performance of different prediction models

The prediction models were ranked according to the root-mean-squared-error (RMSE) values, from smallest to largest. The variable of importance was calculated as a percentage between 0 and 100, depending on the model used from the varImp function (Williams *et al.*, 2018). The average variable of importance for each variable across all the prediction models fitted were calculated and weighted by the RMSE value for that particular model. The reason was to minimise overfitting the different prediction models. The percentage concordance was calculated by creating a binary version of each phenotype: high ≥ 0 , and low = otherwise.

6.5 Results

Out of the 16 measured phenotypic traits as described in chapter 2, only eight phenotypic traits were used to build predictive QTLs + annotated proteins + KEGG pathways based models, only putative QTLs based models only annotated proteins based models and only KEGG pathways based models. The eight phenotypic traits used were, %RWC, caffeine, catechin and tea tasters' scores (aroma, astringency, brightness, briskness, and colour) which are associated with black tea quality and drought tolerance traits. A total of 98 models were fitted to the entire dataset and were evaluated on their predictive ability using cross-validation on the eight different phenotypic

traits. Also, the prediction models were sorted according to their percentage concordance, from the highest to lowest based on validation population.

6.5.1 Percent RWC

The best prediction models based on QTLs + annotated proteins + KEGG pathway for %RWC trait in both discovery and validation datasets based on ranked RMSE and percent concordance were the Principal Component Analysis, Supervised Principal Component Analysis and Independent Component Regression (Table 6.1). The percent concordance for all the three prediction models for the discovery population was 56.9%, and the RMSE ranged between 0.995 and 0.988. However, the two best prediction models for the validation population were Principal Component Analysis and Supervised Principal Component Analysis with RMSE of 0.986 and 0.990, respectively and 60.3% concordance. Principal Component Analysis and Supervised Principal Component Analysis models performed better in the validation population as compared to discovery population.

Table 6.1: The best prediction models based on putative QTLs for %RWC trait + annotated protein + KEGG pathway in both discovery and validation datasets

Model	RMSE	% concordance	RMSE*	% concordance*
Principal Component Analysis	0.996	56.9	0.986	60.3
Supervised Principal Component Analysis	0.995	56.9	0.990	60.3
Independent Component Regression	0.988	56.9	0.985	58.6

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The prediction models for %RWC trait based on putative QTLs only were cross-validated in both the discovery and validation populations. The 13 prediction models that performed better based on putative QTLs are shown in Table 6.2. The two best predictive models with high percent concordance in the discovery population were Polynomial Kernel Regularized Least Squares (69.4%) and Conditional Inference Random Forest (66.9%) with RMSE of 0.993 and 0.989, respectively. However, in the validation population, both models had a percent concordance of 56.9% and RMSE of 0.992 and 0.984, which was slightly lower than in the best prediction models in discovery population. The two prediction models that performed better in the validation population with a 60.3% concordance were Independent Component Regression

and Support Vector Machines with Polynomial Kernel. The RMSE for Independent Component Regression and Support Vector Machines with Polynomial Kernel models was 0.981 and 0.982, respectively. However, the predictive ability of the two models was slightly lower in the discovery population with 54.4% and 57.3% concordance, respectively.

Table 6.2: The best prediction models based on putative QTLs identified for %RWC trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Independent Component Regression	0.989	54.4	0.981	60.3
Support Vector Machines with Polynomial Kernel	0.992	57.3	0.982	60.3
Relevance Vector Machines with Radial Basis Function Kernel	0.933	56.5	0.970	58.6
Conditional Inference Random Forest	0.989	66.9	0.984	56.9
Gaussian Process with Polynomial Kernel	0.990	58.1	0.978	56.9
Extreme Learning Machine	0.991	58.5	0.971	56.9
Radial Basis Function Kernel Regularized Least Squares	0.992	61.3	0.990	56.9
Bayesian Ridge Regression (Model Averaged)	0.992	57.7	0.987	56.9
Polynomial Kernel Regularized Least Squares	0.993	69.4	0.992	56.9
Bayesian Ridge Regression	0.994	56.5	0.991	56.9
Boosted Generalized Linear Model	0.994	56.9	0.985	56.9
glmnet	0.999	55.2	0.985	56.9
Penalized Linear Regression	0.990	55.2	0.984	55.2

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The prediction models that performed better based only on functional annotation of putative proteins identified for %RWC trait were cross-validated in both datasets (Table 6.3). Conditional Inference Random Forest (68.5%) and Radial Basis Function Kernel Regularized Least Squares (67.7%) models had better predictive ability in the discovery population while Bayesian Ridge Regression, Support Vector Machines with Polynomial Kernel and Conditional Inference Random Forest models had better predictive ability of 60.3%, 58.6%, and 58.6%, respectively, in the validation population. The RMSE for prediction models that performed better in both dataset

ranged between 0.928 and 0.991. Conditional Inference Random Forest was the only model that performed better in both the discovery and validation populations.

Table 6.3: The best prediction models based on annotated protein for putative QTLs identified for %RWC trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Bayesian Ridge Regression	0.929	61.7	0.990	60.3
Support Vector Machines with Polynomial Kernel	0.992	63.3	0.987	58.6
Conditional Inference Random Forest	0.991	68.5	0.989	58.6
Gaussian Process with Polynomial Kernel	0.995	60.1	0.987	56.9
Bayesian Ridge Regression (Model Averaged)	0.928	61.7	0.988	56.9
Elastic Net	0.999	61.7	0.991	56.9
Radial Basis Function Kernel Regularized Least Squares	0.928	67.7	0.990	55.2

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The models that performed better in prediction of putative QTLs based only on KEGG pathway in both datasets were Least Angle Regression, Penalized Linear Regression and Polynomial Kernel Regularized Least Squares (Table 6.4). The percent concordance for both datasets ranged between 54.8% and 58.6% with RMSE of between 0.991 and 0.999.

Table 6.4: The best prediction models based on KEGG pathway for putative QTLs identified for %RWC trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Least Angle Regression	0.997	54.8	0.992	58.6
Penalized Linear Regression	0.992	58.1	0.999	56.9
Polynomial Kernel Regularized Least Squares	0.994	57.7	0.991	56.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The percent variable of importance of annotated proteins associated with %RWC trait in both discovery and validation datasets are listed in Table 6.5. The putatively annotated proteins with the high percentage of importance were those associated with abiotic stress response.

Table 6.5: Percent variable of importance of annotated proteins related to %RWC trait in both discovery and validation dataset

Annotated protein	% Importance
Actin	84.2
Armadillo beta-catenin-like repeat	78.8
CSA002263	71.2
Isocitrate isopropyl malate dehydrogenase	66.3
Peptidase C65 Otubain	56.1
BT1 family	54.8
Catalase	48.9
Autophagy-related protein 11	47.4

For the KEGG pathway, the metabolic pathways with the high percentage of importance are listed in Table 6.6. The KEGG pathways were those associated with arginine and proline metabolism (78.3%), glycerolipid metabolism (67.4%), and fructose and mannose metabolism (53.2%).

Table 6.6: Percent variable of importance of KEGG pathways related to %RWC trait in both discovery and validation dataset

KEGG pathway	% variable importance
Arginine and proline metabolism	78.3
Glycerolipid metabolism	67.4
Alanine aspartate and glutamate metabolism	55.8
Tryptophan metabolism	53.4
Fructose and mannose metabolism	53.2

6.5.2 Caffeine

The four prediction models based on QTLs + annotated proteins + KEGG pathway for caffeine trait that performed better in the discovery dataset were the Tree-Based Ensembles (61.5%), Linear Regression with Forward Selection (59.1%), Tree Models from Genetic Algorithms (57.5%) and Least Angle Regression (57.5%). The four models also performed better in the validation dataset with a 65% concordance for each model (Table 6.7).

Table 6.7: The best prediction models based on putative QTLs identified for caffeine trait + annotated protein + KEGG pathway in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Tree Models from Genetic Algorithms	0.995	57.5	0.950	65.5
Linear Regression with Forward Selection	0.983	59.1	0.970	65.5
Tree-Based Ensembles	0.983	61.5	0.976	65.5
Least Angle Regression	0.990	57.5	0.984	65.5

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The best prediction model based on only putative QTLs for caffeine trait in both the discovery and validation datasets were Polynomial Kernel Regularized Least Squares, Self-Organising Maps, and Stochastic Gradient Boosting. Polynomial Kernel Regularized Least Squares had a high percent concordance of 68.7% in the discovery dataset and 56.9% in the validation dataset. Stochastic Gradient Boosting performed better in the validation dataset with percent concordance of 58.6% than in the discovery dataset which had percent concordance of 58.3%. The RMSE for both datasets ranged between 0.991 and 0.995 (Table 6.8).

Table 6.8: The best prediction models based on putative QTLs identified for caffeine trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Stochastic Gradient Boosting	0.995	58.3	0.970	58.6
Polynomial Kernel Regularized Least Squares	0.992	68.7	0.990	56.9
Self-Organising Maps	0.991	58.7	0.999	56.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The prediction models for caffeine trait based only on annotated proteins were cross-validated in both the discovery and validation datasets. A total of five models which performed better in both discovery and validation datasets are listed in Table 6.9. The Polynomial Kernel Regularized Least Squares model had a high percent concordance of 82.1% and 62.1% in discovery and validation dataset, respectively. However, the best prediction models in validation dataset were

Classification and Regression Tree (CART) (65.5%), and Least Angle Regression (65.5%) which both had a percent concordance of 57.5% in the discovery dataset.

Table 6.9: The best prediction models based on annotated protein for putative QTLs identified for caffeine trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
CART	0.985	57.5	0.950	65.5
Least Angle Regression	0.924	57.5	0.984	65.5
Linear Regression with Forward Selection	0.982	58.3	0.983	63.8
Linear Regression with Stepwise Selection	0.992	58.3	0.983	63.8
Polynomial Kernel Regularized Least Squares	0.993	82.1	0.991	62.1

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

A total of seven KEGG pathway-based prediction models for caffeine trait performed better in both discovery and validation datasets. For the discovery dataset, the percent concordance ranged between 56% and 61.9% for Self-Organising Maps and Penalized Linear Regression, respectively. However, for the validation dataset, Penalized Linear Regression had percent concordance of 53.4% and 67.2% for Self-Organising Maps (Table 6.10). The RMSE for both discovery and validation dataset ranged between 0.958 and 0.999.

The percent variable of importance of annotated proteins associated with caffeine trait in both discovery and validation datasets are listed in Table 6.11. The putatively annotated protein with the high percentage of importance was N-(5' phosphoribosyl) anthranilate (PRA) isomerase (91.9%).

Table 6.10: The best prediction models based on KEGG pathway for putative QTLs identified for caffeine trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Self-Organising Maps	0.996	56.0	0.960	67.2
Tree Models from Genetic Algorithms	0.958	58.3	0.972	65.5
Least Angle Regression	0.992	58.3	0.989	65.5
Boosted Linear Model	0.986	59.5	0.986	62.1
Relevance Vector Machines with Polynomial Kernel	0.928	58.7	0.992	62.1
Extreme Learning Machine	0.989	58.3	0.976	56.9
Penalized Linear Regression	0.999	61.9	0.996	53.4

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Table 6.11: Percent variable of importance of annotated proteins related to caffeine trait in both discovery and validation dataset

Annotated protein	% variable importance
N-(5' phosphoribosyl anthranilate) (PRA) isomerase	91.9
Acyltransferase	61.6
Glutaminyl-tRNA synthetase	60.6
Glycosyl hydrolase family 9	41.8
Peptidase C65 Otubain	32.1

For the KEGG pathway, the metabolic pathways with the high percentage of importance are listed in Table 6.12. The three KEGG pathways with DArTseq markers that had a high percentage of importance were metabolic pathways involved in carbon fixation in photosynthetic organisms (79.3%), purine metabolism (63.7%) and arginine biosynthesis (60.5%).

Table 6.12: Percent variable of importance of KEGG pathways related to caffeine trait in both discovery and validation dataset

KEGG Pathway	% variable importance
Carbon fixation in photosynthetic organisms	79.3
Purine metabolism	63.7
Arginine biosynthesis	60.5
Thiamine metabolism	48.1
Amino acid and nucleotide metabolism	47.8
Glyoxylate and dicarboxylate metabolism	44.1
Pyruvate metabolism	30.9
Cysteine and methionine metabolism	30.1

6.5.3 Catechin

Similarly, the prediction models based on QTLs + annotated proteins + KEGG pathway for catechin trait were built and cross-validated in the discovery and validation datasets. Out of the 98 build models, Extreme Learning Machine was the only prediction model that had low RMSE (0.942 and 0.978) and performed better in both discovery and validation datasets for prediction of catechin trait. The percent concordance in the discovery and validation datasets was 55.6% and 69%, respectively (Table 6.13).

Table 6.13: The best prediction model based on putative QTLs identified for catechin trait + annotated protein + KEGG pathway in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Extreme Learning Machine	0.942	55.6	0.978	69.0

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Also, Principal Component Analysis was the only best prediction model based on putative QTLs identified for catechin trait in both discovery and validation dataset (Table 6.14). The prediction ability of Principal Component Analysis model was 68% with RMSE of 0.924 in the discovery dataset and 65.5% with RMSE of 0.991 in the validation dataset.

Table 6.14: The best prediction model based on putative QTLs identified for catechin trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Principal Component Analysis	0.924	68.7	0.991	65.5

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The prediction models for catechin trait based only on annotated QTLs were cross-validated in both the discovery and validation datasets. Neural Network was the only prediction model that performed better in both discovery and validation datasets (Table 6.15) with 55.6% and 65.6% concordance, respectively. However, the prediction model in both discovery and validation dataset had a high RMSE value of 1.22.

Table 6.15: The best prediction model based on annotated protein for putative QTLs identified for catechin trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Neural Network	1.22	55.6	1.12	65.5

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The two models that performed better in prediction of catechin trait based only on KEGG pathway in both discovery and validation datasets were Self-Organising Maps and Independent Component Regression models (Table 6.16). In the discovery dataset, Self-Organising Maps model had 54% concordance with RMSE of 0.972 while Independent Component Regression model had 54.4% concordance with RMSE of 0.974. However, in the validation dataset, Self-Organising Maps and Independent Component Regression prediction models had high percent concordance of 75.9% and 69% but with high RMSE of 0.969 and 0.994, respectively.

Table 6.16: The best prediction models based on KEGG pathway for putative QTLs identified for catechin trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Self-Organising Maps	0.972	54.0	0.969	75.9
Independent Component Regression	0.974	54.4	0.994	69.0

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The percent variables of importance of annotated proteins associated with catechin trait in both discovery and validation dataset are listed in Table 6.17. The putatively annotated proteins with the high percentage of importance were those associated with aminotransferase I and II (94.8%), phosphoribulokinase or uridine kinase family (45.1%) and autophagy-related protein 11 (32.9%).

Table 6.17: Percent variable of importance of annotated proteins related to catechin trait in both discovery and validation dataset

Annotated protein	% variable importance
Diacylglycerol kinase catalytic domain	105.6
Aminotransferase class I and II	94.8
Phosphoribulokinase Uridine kinase family	45.1
CSA016461	34.4
Autophagy-related protein 11	32.9
CSA024230	31.5
Thiolase C-terminal domain	30.6

The metabolic KEGG pathways with the high percentage of importance and are associated with catechin trait are listed in Table 6.18. The KEGG pathways are linked to arginine biosynthesis (88.4%) and Alanine aspartate and glutamate metabolism (38.3%).

Table 6.18: Percent variable of importance of KEGG pathways related to catechin trait in both discovery and validation dataset

KEGG Pathway	% variable importance
Arginine biosynthesis	88.4
Alanine aspartate and glutamate metabolism	38.3

6.5.4 Tea tasters' scores

A total of 98 prediction models for tea tasters' evaluation parameters which included black tea liquor aroma, black tea liquor astringency, black tea liquor brightness, black tea liquor briskness and black tea liquor colour were cross-validated in both the discovery and validation datasets.

6.5.4.1 Aroma

The best prediction models for aroma based on QTLs + annotated proteins + KEGG pathway in the discovery dataset were Bayesian Ridge Regression (63.8%), The Bayesian LASSO (61.9%) and Least Angle Regression (59%). The three prediction models also performed better in the validation dataset but with 62.7% concordance for Least Angle Regression model (Table 6.19).

Table 6.19: The best prediction model based on putative QTLs identified for tea liquor aroma trait + annotated protein + KEGG pathway in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Least Angle Regression	0.996	59.0	0.988	62.7
The Bayesian LASSO	0.964	61.9	0.986	56.9
Bayesian Ridge Regression (Model Averaged)	0.969	63.8	0.985	54.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Twenty-two best prediction models based on putative QTLs identified for aroma trait were cross-validated in both discovery and validation datasets (Table 6.20). The percent concordance in the discovery dataset ranged between 56.2% and 100% with Random Forest being the best prediction model for aroma trait. The Bayesian LASSO was the best prediction model for aroma trait based on putative QTLs in the validation dataset with 66.7% concordance.

Table 6.20: The best prediction model based on putative QTLs identified for tea liquor aroma trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
The Bayesian LASSO	0.994	56.2	0.986	66.7
Bayesian Ridge Regression (Model Averaged)	0.994	61.0	0.986	66.7
Relevance Vector Machines with Polynomial Kernel	0.994	62.4	0.957	66.7
Radial Basis Function Kernel Regularized Least Squares	0.994	73.8	0.987	64.7
Penalized Linear Regression	0.990	59.5	0.966	64.7
Polynomial Kernel Regularized Least Squares	0.994	79.5	0.989	62.7
Elasticnet	0.989	59.0	0.979	62.7
Least Angle Regression	0.989	59.0	0.979	62.7
Boosted Linear Model	0.991	60.5	0.975	62.7
Gaussian Process with Polynomial Kernel	0.994	61.9	0.972	62.7
Independent Component Regression	1.000	50.5	0.971	62.7
The LASSO	0.990	59.0	0.971	62.7
Least Angle Regression	0.990	59.0	0.971	62.7
Boosted Generalized Linear Model	1.000	60.0	0.966	62.7
Conditional Inference Tree	0.994	59.0	0.965	62.7
Multi-Step Adaptive MCP-Net	0.948	59.0	0.960	62.7
Ridge Regression with Variable Selection	0.999	59.0	0.981	60.8
Relaxed LASSO	0.994	60.5	0.966	60.8
Random Forest	0.992	100.0	0.993	58.8
Linear Regression with Backwards Selection	0.992	59.0	0.981	58.8
Linear Regression with Forward Selection	0.992	59.0	0.981	58.8
Linear Regression with Stepwise Selection	0.992	59.0	0.981	58.8
Tree-Based Ensembles	0.978	74.8	0.986	54.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The Parallel Random Forest and Regularized Random Forest models were the best prediction models for black tea aroma trait based on annotated proteins in both discovery and validation datasets (Table 6.21). The percent concordance for the two prediction models in the discovery dataset was 99.5% and 98.6%, respectively and 60.8% and 56.9% in the validation dataset,

respectively. Although the discovery dataset had high percent concordance, the RMSE was slightly high than in the validation dataset.

Table 6.21: The best prediction model based on annotated protein for putative QTLs identified for tea liquor aroma trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Parallel Random Forest	1.02	99.5	0.987	60.8
Regularized Random Forest	1.01	98.6	0.992	56.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The models that performed better in prediction of aroma trait based only on the KEGG pathway in both discovery and validation datasets were the Principal Component Analysis and Tree-Based Ensembles models (Table 6.22). In the discovery dataset, Principal Component Analysis and Tree-Based Ensembles models had 54.3% and 59.5% concordance, respectively. The prediction ability of Principal Component Analysis model and Tree-Based Ensembles model in the validation dataset was 62.7% and 58.8%, respectively.

Table 6.22: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor aroma trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Principal Component Analysis	0.994	54.3	0.974	62.7
Tree-Based Ensembles	1.001	59.5	0.982	58.8

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The percent variables of importance of annotated proteins related to black tea aroma trait in both discovery and validation dataset are listed in Table 6.23). One of the putatively annotated protein was 2OG Fe II oxygenase superfamily, which had 42% concordance with markers associated with black tea aroma trait is linked to tea catechins biosynthesis.

Table 6.23: Percent variable of importance of annotated proteins related to tea liquor aroma trait in both discovery and validation dataset

Annotated protein	% variable importance
Actin	78.1
CSA003424	49.0
WD domain	44.2
2OG Fe II oxygenase superfamily	42.0
14-3-3 protein	41.9
NB-ARC domain	40.3
Pectinesterase	39.2
CSA033214	38.6
N-(5' phosphoribosyl) anthranilate (PRA) isomerase	38.5
ATPase family associated with various cellular activities AAA	36.4

The metabolic KEGG pathways with a high percentage of importance and were associated with black tea aroma trait are listed in Table 6.24). Flavone and flavonol biosynthesis pathways which are pathways involved in catechins biosynthesis had 52.8% concordance with markers associated with black tea aroma trait.

Table 6.24: Percent variable of importance of KEGG pathways related to tea liquor aroma trait in both discovery and validation dataset

KEGG Pathway	% variable Importance
Tyrosine metabolism	73.5
Alanine aspartate and glutamate metabolism	62.7
Flavone and flavonol biosynthesis	52.8
Fructose and mannose metabolism	48.6
Pentose and glucuronate interconversions	39.2
Arginine biosynthesis	38.5

6.5.4.2 Astringency

Extreme Learning Machine model was the only best prediction model for black tea astringency trait based on QTLs + annotated proteins + KEGG pathway in both the discovery and validation dataset with 56.2% and 60.8% concordance, respectively (Table 6.25).

Table 6.25: The best prediction model based on putative QTLs identified for tea liquor astringency trait + annotated protein + KEGG pathway in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Extreme Learning Machine	0.984	56.2	0.962	60.8

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The best prediction models based on putative QTLs identified for black tea astringency trait had high percent concordance in the discovery dataset (Table 6.26). In the discovery dataset, Random Forest, Tree-Based Ensembles, glmnet, and Least Angle Regression prediction models had 93.3%, 70%, 60% and 59.5% concordance, respectively. However, in the validation dataset, percent concordance ranged between 54.9% and 56.9% with RMSE of between 1.004 and 1.054.

Table 6.26: The best prediction model based on putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
glmnet	0.999	60.0	1.054	56.9
Least Angle Regression	0.986	59.5	1.051	54.9
Random Forest	0.986	93.3	1.029	54.9
Tree-Based Ensembles	0.936	70.0	1.004	54.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The prediction models that performed better in both datasets based on annotated proteins for putative QTLs identified for black tea astringency were Monotone Multi-Layer Perceptron Neural Network and CART (Table 6.27). Monotone Multi-Layer Perceptron Neural Network and CART had 76.2% and 64.3% concordance, respectively in the discovery dataset. In the validation dataset, the two prediction models had 56.9% and 54.9% concordance. However, the two models had a slightly high RMSE in both the discovery and validation datasets.

Table 6.27: The best prediction model based on annotated protein for putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Monotone Multi-Layer Perceptron Neural Network	1.186	76.2	1.115	56.9
CART	1.149	64.3	1.053	54.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

On KEGG pathway-based prediction models for putative QTLs identified for black tea astringency traits; the best four prediction models in both datasets are listed in Table 6.28. Polynomial Kernel Regularized Least Squares model had high percent concordance of 69% and 64.7% with the markers in the discovery and validation dataset, respectively.

Table 6.28: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Polynomial Kernel Regularized Least Squares	0.994	69.0	0.990	64.7
Self-Organising Maps	0.927	55.2	0.979	58.8
Least Angle Regression	0.992	51.9	0.996	56.9
Radial Basis Function Kernel Regularized Least Squares	0.993	62.4	0.992	54.9

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Aminotransferase class I and II enzymes (81.6%) were the annotated protein with high percent variable of importance related to black tea astringency trait (Table 6.29). The variables of importance based on KEGG pathways related to black tea astringency as listed in Table 6.30 were arginine biosynthesis (78.3%) and arginine and proline metabolism (71.1%).

Table 6.29: Percent variable of importance of annotated proteins related to tea liquor astringency trait in both discovery and validation dataset

Annotated protein	% variable importance
Aminotransferase class I and II	81.6
Autophagy-related protein 11	42.5

Table 6.30: Percent variable of importance of KEGG pathways related to tea liquor astringency trait in both discovery and validation dataset

KEGG Pathway	% variable importance
Arginine biosynthesis	78.3
Arginine and proline metabolism	71.1
Glycin, serine and threonine metabolism	51.0

6.5.4.3 Brightness

All the four prediction models for black tea brightness trait based on QTLs + annotated proteins + KEGG pathway in the discovery dataset had low percent concordance of 51% with genomic DArT markers (Table 6.31) However, in the validation dataset, all the four prediction models had high percent concordance with genomic DArTseq markers. The prediction ability of Extreme Learning Machine model was 71.2% while Non-Informative Model, partDSA and Stacked AutoEncoder Deep Neural Network model was each 69.5% (Table 6.31).

Table 6.31: The best prediction model based on putative QTLs identified for tea liquor brightness trait + annotated protein + KEGG pathway in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Extreme Learning Machine	0.974	51.9	0.966	71.2
Non-Informative Model	0.998	51.0	0.990	69.2
partDSA	0.998	51.0	0.990	69.2
Stacked AutoEncoder Deep Neural Network	0.994	51.0	0.999	69.2

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

The prediction models based on putative QTLs identified for black tea brightness trait in the discovery dataset also had low percent concordance with genomic DArTseq markers (Table

6.32). Polynomial Kernel Regularized Least Squares was the best prediction model in both datasets with 79% and 53.8% concordance in the discovery and validation dataset, respectively.

Table 6.32: The best prediction model based on putative QTLs identified for tea liquor brightness trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Non-Informative Model	0.998	51.0	0.990	69.2
CART	0.985	51.0	0.990	69.2

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Table 6.33: The best prediction model based on annotated protein for putative QTLs identified for tea liquor brightness trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Polynomial Kernel Regularized Least Squares	0.997	79.0	0.990	53.8

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Supervised Principal Component Analysis model was the only best prediction model based on the KEGG pathway for putative QTLs identified for black tea brightness trait (Table 6.34). The model prediction ability was 56.2% and 59.6% in the discovery and validation dataset, respectively.

Table 6.34: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor brightness trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Supervised Principal Component Analysis	0.991	56.2	0.993	59.6

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Aminotransferase class I and II (83.8%) were the important variables for annotated proteins that are associated with black tea liquor brightness trait in both discovery and validation datasets

(Table 6.35.). Similarly, arginine biosynthesis (83.1%) was the important variable associated with black tea liquor brightness trait (Table 6.36).

Table 6.35: Percent variable of importance of annotated proteins related to tea liquor brightness trait in both discovery and validation dataset

Annotated protein	% variable importance
Aminotransferase class I and II	83.8
Autophagy-related protein 11	66.0
Armadillo beta-catenin-like repeat	65.6
CSA024230	61.8
CSA002263	61.1
Phosphoribulokinase Uridine kinase family	47.5
Diacylglycerol kinase catalytic domain	46.7
ATPase family associated with various cellular activities AAA	40.9

Table 6.36: Percent variable of importance of KEGG pathways related to tea liquor brightness trait in both discovery and validation dataset

KEGG Pathway	% variable importance
Arginine biosynthesis	83.1
Arginine and proline metabolism	27.1
Glycine, serine and threonine metabolism	23.7
Tyrosine metabolism	21.1

6.5.4.4 Briskness

A total of eight prediction models had high prediction ability for black tea liquor briskness trait based on QTLs + annotated proteins + KEGG pathway in both discovery and validation datasets (Table 6.37). Stacked AutoEncoder Deep Neural Network, Random Forest Rule-Based Model, Quantile Regression with LASSO penalty, Non-Convex Penalized Quantile Regression, and Cubist had 65.2% and 74% concordance in both discovery and validation dataset, respectively.

Table 6.37: The best prediction model based on putative QTLs identified for tea liquor briskness trait + annotated protein + KEGG pathway in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Stacked AutoEncoder Deep Neural Network	0.994	65.2	0.995	74.0
Random Forest Rule-Based Model	0.997	65.2	0.992	74.0
Quantile Regression with LASSO Penalty	0.997	65.2	0.992	74.0
Non-Convex Penalized Quantile Regression	0.997	65.2	0.992	74.0
Cubist	0.997	65.2	0.992	74.0
Linear Regression with Forward Selection	0.999	57.6	0.979	66.0
Linear Regression with Stepwise Selection	0.999	57.6	0.979	66.0
Extreme Learning Machine	0.980	56.2	0.990	56.0

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Of all the prediction models, Quantile Regression with LASSO penalty, Non-Convex Penalized Quantile Regression, and Cubist were the best prediction models based only for putative QTLs with 65.2% and 74% concordance in both datasets (Table 6.38). Similarly, Quantile Regression with LASSO penalty and Non-Convex Penalized Quantile Regression models were the best prediction models for black tea briskness based on annotated proteins (Table 6.39). There were seven best prediction models based on the KEGG pathway for putative QTLs identified for black tea liquor briskness trait (Table 6.40).

Table 6.38: The best prediction model based on putative QTLs identified for tea liquor briskness trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Quantile Regression with LASSO penalty	0.997	65.2	0.992	74.0
Non-Convex Penalized Quantile Regression	0.997	65.2	0.992	74.0
Cubist	0.997	65.2	0.992	74.0

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Table 6.39: The best prediction model based on annotated protein for putative QTLs identified for tea liquor astringency trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Quantile Regression with LASSO penalty	0.997	65.2	0.992	74.0
Non-Convex Penalized Quantile Regression	0.997	65.2	0.992	74.0
Polynomial Kernel Regularized Least Squares	0.998	80.0	0.990	54.0

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Table 6.40: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor briskness trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Quantile Regression with LASSO penalty	0.997	65.2	0.992	74.0
Non-Convex Penalized Quantile Regression	0.997	65.2	0.992	74.0
Cubist	0.997	65.2	0.992	74.0
Multi-layer Perceptron Network by Stochastic Gradient Descent	0.994	65.2	0.998	74.0
Multi-Layer Perceptron	0.994	65.2	0.990	74.0
Multi-Layer Perceptron, multiple layers	0.994	65.2	0.990	74.0
Extreme Learning Machine	0.991	65.2	0.990	74.0

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

All the seven prediction models had high percent concordance of between 65.2% and 74%. Aminotransferase class I and II (86.3%) were the important variables for annotated proteins that are associated with black tea liquor briskness trait in both discovery and validation datasets (Table 6.41). Similarly, arginine biosynthesis (83.2%) was the important variable associated with black tea liquor briskness trait (Table 6.42).

Table 6.41: Percent variable of importance of annotated proteins related to tea liquor briskness trait in both discovery and validation dataset

Annotated protein	% variable importance
Aminotransferase class I and II	86.3
Armadillo beta-catenin-like repeat	70.7
CSA002263	66.0
Autophagy-related protein 11	58.7
CSA024230	57.0

Table 6.42: Percent variable of importance of KEGG pathways related to tea liquor briskness trait in both discovery and validation dataset

KEGG Pathway	% variable importance
Arginine biosynthesis	83.2
Arginine and proline metabolism	33.7
Terpenoid backbone biosynthesis	29.0
Phosphatidylinositol signalling system	24.3
Alanine aspartate and glutamate metabolism	22.9

6.5.4.5 Colour

The prediction models for black tea liquor colour trait based on QTLs + annotated proteins + KEGG pathway were cross-validated in both the discovery and validation populations (Table 6.43). Polynomial Kernel Regularized Least Squares and Extreme Learning Machine were the only two models in both datasets that had high percent concordance with genomic DArTseq markers. In the discovery dataset, percent concordance of Polynomial Kernel Regularized Least Squares and Extreme Learning Machine models were 55.7% and 54.3% concordance, respectively. In the validation dataset, the two prediction models had 58.6% and 53.4% concordance, respectively.

Table 6.43: The best prediction model based on putative QTLs identified for tea liquor colour trait + annotated protein + KEGG pathway in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Polynomial Kernel Regularized Least Squares	0.994	55.7	0.991	58.6
Extreme Learning Machine	0.980	54.3	1.000	53.4

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

For prediction models based only on putative QTLs identified for tea liquor colour trait, Multi-Layer Perceptron, with multiple layers and Multi-layer Perceptron Network by Stochastic Gradient Descent were the best prediction in both datasets with 54.3% and 60.3% concordance, respectively (Table 6.44). The Multi-layer Perceptron Network by Stochastic Gradient Descent and Extreme Learning Machine models were the best prediction models based on annotated proteins for black tea colour trait in both datasets (Table 6.45). In the discovery dataset, the Multi-layer Perceptron Network by Stochastic Gradient Descent model and the Extreme Learning Machine model were 54.3% and 56.2% concordance, respectively. Also, the two models had a better prediction ability in the validation dataset with 60.3% and 55.2% concordance, respectively.

Table 6.44: The best prediction model based on putative QTLs identified for tea liquor colour trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Multi-Layer Perceptron, with multiple layers	0.988	54.3	1.028	60.3
Multi-layer Perceptron Network by Stochastic Gradient Descent	0.991	54.3	1.000	60.3
Extreme Learning Machine	0.981	58.6	1.005	53.4

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Table 6.45: The best prediction model based on annotated protein for putative QTLs identified for tea liquor colour trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Multi-layer Perceptron Network by Stochastic Gradient Descent	0.99	54.3	0.99	60.3
Extreme Learning Machine	0.99	56.2	0.99	55.2

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Polynomial Kernel Regularized Least Squares was the best prediction model based on KEGG pathway for putative QTLs identified for black tea liquor colour trait in both datasets (Table 6.46). The percentage concordance of Polynomial Kernel Regularized Least Squares model was 54.3% and 62.1% in the discovery and validation datasets, respectively

Table 6.46: The best prediction models based on KEGG pathway for putative QTLs identified for tea liquor colour trait in both discovery and validation dataset

Model	RMSE	% concordance	RMSE*	% concordance*
Polynomial Kernel Regularized Least Squares	0.994	54.3	0.991	62.1
Extreme Learning Machine	0.991	51.0	0.988	58.6
Bayesian Regularized Neural Networks	0.979	51.4	0.992	58.6

RMSE - Root Mean Square Error for discovery dataset

% concordance - percent concordance for discovery dataset

RMSE* - Root Mean Square Error for validation dataset

% concordance* - percent concordance for validation dataset

Aminotransferase class I and II enzymes (79.4%) was the annotated protein with high percent variable of importance related to black tea colour trait in both datasets (Table 6.47). The variable of importance based on KEGG pathways related to black tea colour trait was arginine biosynthesis (79.6%) as shown in Table 6.48.

Table 6.47: Percent variable of importance of annotated proteins related to tea liquor colour trait in both discovery and validation dataset

Annotated protein	% variable importance
Aminotransferase class I and II	79.4
Phosphoribulokinase Uridine kinase family	63.0
Autophagy-related protein 11	54.6
CSA024230	52.8
Diacylglycerol kinase catalytic domain	38.5
Adaptor complexes medium subunit family	31.0
CSA016461	30.7
ATPase family associated with various cellular activities AAA	30.6
Isocitrate isopropylmalate dehydrogenase	26.9

Table 6.48: Percent variable of importance of KEGG pathways related to tea liquor colour trait in both discovery and validation dataset

KEGG Pathway	% variable importance
Arginine biosynthesis	79.6
Porphyrin and chlorophyll metabolism	46.9
Arginine and proline metabolism	41.7
Pentose and glucuronate interconversions	33.1
Fructose and mannose metabolism	28.6
Glycine, serine and threonine metabolism	27.9

6.6 Discussion

This study reports the first initiative of applying machine learning techniques to develop putative QTLs + annotated proteins + KEGG pathway, putative QTLs, annotated proteins and KEGG pathway-based prediction models for tea breeding. While previous research, particularly on tea breeding, has largely been focused on the conventional linkage and AM methods in selection of tea cultivars with desirable traits, no work has focused on prediction models to improve tea breeding. This study compared the performance of 98 different machine learning models which comprised of both supervised and unsupervised techniques. A few models were chosen as the best prediction models based on low RMSE and high percent concordance in both datasets. The low values of RMSE indicate high model quality (Alexander *et al.*, 2015; Lopatin *et al.*, 2016).

Moreover, despite the rigorous and stringent validation procedures followed in this study for all the four based prediction approaches, putative QTLs + annotated proteins + KEGG pathway-based prediction approach, showed more robustness and usefulness in prediction than use of individual prediction based models. This important contribution also demonstrates the obvious underlying principle of the basic quantitative relationship between identified QTLs, functions of annotated proteins and the role of identified QTLs in the KEGG pathways.

In this study, different genomic prediction models were compared in terms of their predictive ability, or accuracy for different traits. However, the predictive ability of all models varied across traits. The difference in the prediction values of different prediction models for the same trait confirms that only traits that are easier to phenotype should be considered for genomic predictions (Nyine *et al.*, 2018). Similarly, the difference in model performance between traits suggests that variation in trait architecture, number of QTLs controlling the trait and LD between markers and QTL influence the performance of the models (Clark *et al.*, 2011). The results in this study are in agreement with previous studies in that prediction model that perform variable selection (possibly irrelevant variables) such as Extreme Learning Machine give better predictive ability values (Huang *et al.*, 2006b). For example, Extreme Learning Machine model which is robust and sensitive to variable selection had better prediction ability for catechin, astringency, brightness, briskness, and colour based on putative QTLs + annotated proteins + KEGG pathway approach. The argument on the superiority of the model could also be highly dependent on the presence of large QTL effects. It is likely that the phenotypic traits that are controlled by large effect QTL in tea were selected by Extreme Learning Machine model in all cross-validations. The high level of concordance in the four prediction approaches also suggests that some molecular markers could be associated with the phenotypic traits in question. The same molecular markers are linked to proteins and metabolic pathways of different phenotypic traits. In addition, variable of importance for each trait based on either annotated proteins or KEGG pathway confirmed the predictive ability of each model.

The percent variable of importance of annotated proteins associated with %RWC trait were actin, armadillo beta-catenin-like repeat and isocitrate isopropyl malate dehydrogenase which are associated with abiotic stress (Drøbak *et al.*, 2004; Kamal *et al.*, 2012; Sharma *et al.*, 2014). The KEGG pathways were arginine and proline metabolism, glycerolipid metabolism and fructose

and mannose metabolism, which are primary metabolites reported to be associated with abiotic stress (Upchurch, 2008; Keunen *et al.*, 2013; Majumdar *et al.*, 2016). The high percentage of importance of putatively annotated proteins for caffeine trait was N-(5' phosphoribosyl) anthranilate (PRA) isomerase, which is involved in the first and intermediate pathway in the purine biosynthesis (D'Mello, 2017). Caffeine and other methylxanthines such as theobromine (3, 7-dimethylxanthine), methyluric acids and theophylline are present in tea and are classified as purine alkaloids (Ashihara *et al.*, 2008). Therefore, the putative protein, N-(5' phosphoribosyl) anthranilate (PRA) isomerase may be associated with the biosynthesis of caffeine, theobromine, theophylline and methyluric acids in tea. Also, phenylalanine, tyrosine, and tryptophan are products of shikimate pathway which is involved in the biosynthesis of plant flavonoids, including catechins (Ghasemzadeh and Ghasemzadeh, 2011). For KEGG pathways, DArT markers with a high percentage of importance were those associated with carbon fixation in photosynthetic organisms, arginine biosynthesis and purine metabolism, which is associated with caffeine biosynthesis (Ashihara *et al.*, 2013). The putatively annotated proteins for catechin trait were those associated with phenylalanine, tyrosine and tryptophan biosynthesis and carbon fixation, which are products for shikimate pathway (Maeda and Dudareva, 2012) and abiotic stress response (stomatal closure). Hence, it was in agreement with this study since previous reports have shown that catechins play a role as indicators of desiccation tolerance in tea plants (Cheruiyot *et al.*, 2008b). Aminotransferases have been shown to play a major role in a variety of metabolic pathways, including, amino acid biosynthesis and photorespiration. Aminotransferase I and II, putatively annotated for catechin trait, are involved in the biosynthesis of phenylalanine and tyrosine, which are precursors of flavonoid biosynthesis (Maeda *et al.*, 2011). Also, aminotransferase I and II had high percent variable of importance of annotated proteins associated with astringency, brightness, briskness and colour traits. In relation to this argument, theaflavins, and thearubigins present in black tea are formed from the oxidation of catechins during the manufacture of black tea. Besides caffeine and catechins, amino acids such as arginine contribute to astringency in tea infusion (Zhen, 2003). The findings corroborate with the current study on the variable of importance in the KEGG pathways which is involved in arginine biosynthesis.

6.7 Conclusion

The comparison of the performance of different prediction models was based on ranked RMSE values. Therefore, predictive ability of different machine learning methodologies applied were significantly different based on putative QTLs + annotated proteins + KEGG pathways, putative QTLs, annotated proteins and KEGG pathways approaches which was in agreement with the alternative hypothesis of this study. However, the discovery and validation tea populations were phenotyped in different locations and in different years. Thus, the putative genotype by environment interactions due to differences in years and locations were not considered, and this might have affected the prediction accuracy. However, since the %concordance in the validation population were quite high, this indicates that the models are robust across locations and years. The parents with their F₁ progenies (discovery population) and the validation population were planted in the same plots, so there was an insufficient number of reference genotypes available to estimate genotype by environment interaction. Additional phenotyping of the discovery population in different locations is in progress, and using the average over multiple years may yield more stable estimates of the phenotypic performance, which could further increase the accuracy of the prediction models. Moreover, the results in this study suggest that for an outbreeding plant species such as tea, increasing the training population size could improve the predictive ability of the models as compared to using only two parental clones. In other tree species such as eucalyptus, pine and apples more than two parents have been used to improve on prediction ability and accuracy of the models. However, this research study has opened up a new avenue for future application of the machine learning techniques in tea breeding. Also, genomic selection can be applied earlier than phenotypic selection and thus reducing the time from initial crosses to planting in commercial tea fields. Implementation of genomic selection in an early-stage in tea breeding program can increase the selection intensity which is relevant concerning related traits which are often costly to phenotype.

References

- Alexander, D., Tropsha, A. & Winkler, D. A. (2015). Beware of R²: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling* 55(7): 1316-1322.
- Ashihara, H., Sano, H. & Crozier, A. (2008). Caffeine and related purine alkaloids: biosynthesis, catabolism, function and genetic engineering. *Phytochemistry* 69(4): 841-856.
- Ashihara, H., Yokota, T. & Crozier, A. (2013). Biosynthesis and catabolism of purine alkaloids. In *Advances in botanical research*, Vol. 68, 111-138: Elsevier.
- Bhatia, I. (1963). Composition of leaf in relation to liquor characters of made tea. *Two and a Bud* 8(3): 11-14.
- 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. & Wanyoko, J. K. (2008). Shoot epicatechin and epigallocatechin contents respond to water stress in tea [*Camellia sinensis* (L.) O. Kuntze]. *Bioscience, Biotechnology, and Biochemistry* 72(5): 1219-1226.
- Clark, S. A., Hickey, J. M. & Van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution* 43(1): 18.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., Camacho-González, J. M., Pérez-Elizalde, S. & Beyene, Y. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*.
- D'Mello, J. F. (2017). *The Handbook of Microbial Metabolism of Amino Acids*. CABI.
- Drøbak, B., Franklin-Tong, V. & Staiger, C. (2004). The role of the actin cytoskeleton in plant cell signaling. *New Phytologist* 163(1): 13-30.
- FAO (2015b). Kenya's tea sector under climate change: An impact assessment and formulation of a climate smart strategy. (Eds A. Elbehri, B. Cheserek, A. Azapagic, D. Raes, M. Mwale, J. Nyengena, P. Kiprono and C. Ambasa). Rome, Italy.

- Ghasemzadeh, A. & Ghasemzadeh, N. (2011). Flavonoids and phenolic acids: Role and biochemical activity in plants and human. *Journal of Medicinal Plants Research* 5(31): 6697-6703.
- 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*.
- Hu, H., Scheben, A. & Edwards, D. (2018). Advances in Integrating Genomics and Bioinformatics in the Plant Breeding Pipeline. *Agriculture* 8(6): 75.
- Huang, G. B., Zhu, Q.-Y. & Siew, C. K. (2006). Extreme learning machine: theory and applications. *Neurocomputing* 70(1-3): 489-501.
- Kamal, A. H. M., Cho, K., Kim, D. E., Uozumi, N., Chung, K. Y., Lee, S. Y., Choi, J. S., Cho, S. W., Shin, C. S. & Woo, S. H. (2012). Changes in physiology and protein abundance in salt-stressed wheat chloroplasts. *Molecular Biology Reports* 39(9): 9059-9074.
- Kelchtermans, P., Bittremieux, W., Grave, K., Degroeve, S., Ramon, J., Laukens, K., Valkenborg, D., Barsnes, H. & Martens, L. (2014). Machine learning applications in proteomics research: How the past can boost the future. *Proteomics* 14(4-5): 353-366.
- Keunen, E., Peshev, D., Vangronsveld, J., Van Den Ende, W. & Cuypers, A. (2013). Plant sugars are crucial players in the oxidative challenge during abiotic stress: extending the traditional concept. *Plant, cell & environment* 36(7): 1242-1255.
- Kumar, S., Chagné, D., Bink, M. C., Volz, R. K., Whitworth, C. & Carlisle, C. (2012). Genomic selection for fruit quality traits in apple (*Malus× domestica* Borkh.). *PloS one* 7(5): e36674.
- Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armananzas, R., Santafé, G. & Pérez, A. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*: 86-112.
- Lopatin, J., Dolos, K., Hernández, H., Galleguillos, M. & Fassnacht, F. (2016). Comparing generalized linear models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. *Remote sensing of environment* 173: 200-210.
- Maeda, H. & Dudareva, N. (2012). The shikimate pathway and aromatic amino acid biosynthesis in plants. *Annual Review of Plant Biology* 63: 73-105.

- Maeda, H., Yoo, H. & Dudareva, N. (2011). Prephenate aminotransferase directs plant phenylalanine biosynthesis via aroenate. *Nature Chemical Biology* 7(1): 19.
- Majumdar, R., Barchi, B., Turlapati, S. A., Gagne, M., Minocha, R., Long, S. & Minocha, S. C. (2016). Glutamate, ornithine, arginine, proline, and polyamine metabolic interactions: the pathway is regulated at the post-transcriptional level. *Frontiers in Plant Science* 7: 78.
- Naik, H. S., Zhang, J., Lofquist, A., Assefa, T., Sarkar, S., Ackerman, D., Singh, A., Singh, A. K. & Ganapathysubramanian, B. (2017). A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant Methods* 13(1): 23.
- Nyine, M., Uwimana, B., Blavet, N., Hřibová, E., Vanrespaille, H., Batte, M., Akech, V., Brown, A., Lorenzen, J. & Swennen, R. (2018). Genomic prediction in a multiploid crop: genotype by environment interaction and allele dosage effects on predictive ability in banana. *The Plant Genome*.
- Ornella, L., Cervigni, G. & Tapia, E. (2012). Applications of Machine Learning for Maize Breeding.
- Ornella, L., González-Camacho, J. M., Dreisigacker, S. & Crossa, J. (2017). Applications of Genomic Selection in Breeding Wheat for Rust Resistance. In *Wheat Rust Diseases*, 173-182: Springer.
- Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture* 121: 57-65.
- Sharma, M., Singh, A., Shankar, A., Pandey, A., Baranwal, V., Kapoor, S., Tyagi, A. K. & Pandey, G. K. (2014). Comprehensive expression analysis of rice Armadillo gene family during abiotic stress and development. *DNA Research* 21(3): 267-283.
- Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M. & Ebrahimie, E. (2014). Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. *PloS one* 9(5): e97288.
- Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R. & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology* 3(6): e116.
- Upchurch, R. G. (2008). Fatty acid unsaturation, mobilization, and regulation in the response of plants to stress. *Biotechnology Letters* 30(6): 967-977.

- ur Rahman, H., Ch, N. J., Manzoor, S., Najeeb, F., Siddique, M. Y. & Khan, R. A. (2017). A comparative analysis of machine learning approaches for plant disease identification. *Advancements in Life Sciences* 4(4): 120-126.
- Williams, C. K., Engelhardt, A., Cooper, T., Mayer, Z., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T. & Kuhn, M. M. (2018). Package ‘caret’.
- Zhen, Y. S. (2003). *Tea: bioactivity and therapeutic potential*. CRC Press

CHAPTER 7

CONCLUDING DISCUSSION AND RECOMMENDATION

7.1 Concluding discussion

The present study constructed a linkage map using 1 421 DArTseq markers on two segregating tea populations TRFK St 504 and TRFK St 524. The two populations were derived from a reciprocal cross of two parental clones TRFK 303/577 and GW Ejulu which comprised of a total of 261 progenies. The map consisted of 15 linkage groups which correspond to chromosome haploid number of tea plant ($2n = 2x = 30$) and spanned 1 260.1 cM with a mean interval of 1.1 cM between markers indicating saturation of the genome. However, there was a gap greater than 20 cM between the adjacent markers on LG06 and LG15. The presence of a large gap may be as a result of genome regions corresponding to the gap regions of the genetic map or homozygous could be a result of gaps in both mapping parents which leads no recombination events.

A total of 16 phenotypic data was collected for the two segregating populations and used to identify QTLs influencing both black and green tea biochemical traits as well as drought stress trait. The analysis of tea biochemical molecules in black and green tea using UPLC revealed a wide range of variation in individual catechins, theaflavins and caffeine contents in the two parents and F_1 progeny. Green tea had high amounts of individual catechins and black tea high amounts of individual theaflavins. However, there was no significant difference in caffeine levels in both green and black tea. Also, besides the five major peaks identified, other several peaks were detected but unidentified. This is an indication that there are several other biochemical compounds present in the tea extracts. There were variations in tea taster scores in the parental clones and F_1 progenies. For the parents, phenotypic trait for tea liquor colour scored highly in parental clone GW Ejulu as compared to parental clone TRFK 303/577. There were also variations the %RWC in F_1 progenies and parental clones in the two tea populations. The %RWC ranged from 49 to 99% which indicate that there were tea plants that were either tolerant or susceptible to drought. Parental clone GW Ejulu which is drought susceptible tea cultivar had low %RWC of 55.1% while parental clone TRFK 303/577 which is drought-tolerant cultivar had high %RWC of 74.1%. In overall, a large number of phenotypic traits for the progeny fell outside to those of the parents indicating transgressive segregation, hence polygenic inheritance

of the traits in question. Further, the normal distribution observed for the phenotypic traits shows the involvement of several genes in the genetic control of phenotypic traits.

In general, significant correlations were observed between the individual catechins, theaflavins and tea tasters' scores in the two tea population. A significant correlation was observed between EC and EGC, TF1 and TF2, TF1 and TF4, TF2 and TF4, EGC and EGCG, EC and EGC, TF1 and EGC, TF2 and EC, TF1 and EGCG, TF2 and EGC, TF4 and EGC. For tea taster's score, a significant correlation was observed between individual AST and BRK, AR and CL. However, there was no correlation between %RWC, all biochemical compounds, and tea tasters' scores. The strong correlation of the digallate fraction observed in this current study and the changes in catechin content to theaflavins in relation to sensory characteristics can be utilised to enhance the quality of black tea.

Both interval, multiple QTL and association mapping revealed a total of 59 putative QTLs in the 15 LGs associated with black and green tea biochemical traits and drought tolerance trait at a significant genome-wide threshold of 5%. A total of 70 putative QTLs were detected for 16 phenotypic traits measured using the interval mapping. Of all the putative QTLs detected, most QTLs were located in all the other linkage groups except in LG03, LG05, and LG11. For black and green tea biochemical traits, 67 putative QTL were detected and were located on all linkage groups, except on LG03, LG05, and LG11. The remaining three putative QTLs which are associated with drought tolerance were identified in LG02, LG06, and LG09. The highest number of putative QTLs were associated with catechin (13), EGC (12), tea tasters' score (12), caffeine (10) and EGC (9), respectively. The phenotypic variance for all the identified putative QTLs ranged from 2.8% for EGC to 23.3% for ECG, respectively. The phenotypic variance for %RWC ranged from 5.7 to 7.3%, while for the tea tasters' score it ranged from 5.8 to 9.1%. The putative QTLs identified in interval mapping were used as cofactors to measure multiple QTL model mapping. A total of 40 putative QTLs were identified in 16 phenotypic traits measured using MQM mapping. Similarly, no putative QTLs in LG03, LG05, and LG11 were associated with any of the traits. Of these 40 putative QTLs identified, 37 putative QTLs were associated with black and green biochemical traits while three were associated with %RWC. The putative QTLs detected were similar to those detected in interval mapping. The phenotypic variance for all putative QTLs ranged from 4.1% (CAT) to 23% (ECG). The phenotypic variance for %RWC

ranged from 5.7 to 7.3%, while the tea tasters' score ranged from 6.0 to 9.1%. The association mapping analysis using the Q-model and K-model detected a total of 49 DArT markers associated with 15 different black tea quality traits and one drought-tolerant trait. Several pleiotropic loci associated with C, EC, EGC, TF1, TF2 and TF4 were also identified which included, 5136945, 5112438, 5136058, 5123257, 5106022 and 5064764 DArTseq markers.

All 59 putative QTLs detected in 15 linkage groups derived from the DArTseq linkage map were identified and annotated function using the tea genome. Of 1 421 DArTseq markers derived from the linkage map, 53 markers were identified to be linked to black tea quality and %RWC. A total of 5,592 unigenes were assigned GO terms, of which, 56% comprised of BP, 29% were CCs, and 15% were associated MFs. The BP category had the highest representation of unigenes with 35% (BP), 16% (cellular process), 12% (metabolic process), 11% (response to stimulus) 9% (developmental process) and 6% (localization and signalling). Among the top 20 functional subgroups in BP, multicellular organism development, response to stress and response to endogenous stimulus represented the largest group. For the MF category, binding and catalytic activities were the most abundant groups. The transcripts that corresponded to the cell, cell parts, and cell organelles were the most abundant groups within the cellular component category.

A total of 53 unigenes linked to putative QTLs detected were assigned two different KEGG database pathways; thiamine and purine biosynthesis pathways. Additional 84 unigenes in all 15 linkage groups were assigned to 25 different KEGG database pathways based on categories of secondary metabolite biosynthesis. The most abundant unigenes were for carbohydrate and amino acids biosynthesis, which are mostly involved in plant hormone signal transduction pathways in response to abiotic stress and biosynthesis of other secondary metabolites such as phenylpropanoid biosynthesis, flavonoid biosynthesis, and alkaloid biosynthesis. Also, major enzymes that are involved in different metabolic pathways were categorised into the six different categories; oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. Functional annotations of detected putative QTL led to the putative identification of functions of 37 putative candidate genes that could be involved in the expression of the targeted traits. Most of the candidate genes that were putatively identified were involved in phenylalanine biosynthesis, tyrosine biosynthesis, tryptophan biosynthesis, flavonoid biosynthesis, purine or thiamine

biosynthesis, phenylpropanoid and shikimate pathway, and proteins related to cell signalling and abiotic stress.

The different genomic prediction models were compared in terms of their predictive ability for different traits. Of all the four based prediction approaches, putative QTLs + annotated proteins + KEGG pathway-based prediction approach, showed more robustness and usefulness in prediction of phenotypes than individual prediction based models. In addition, the predictive ability of all models varied across all the phenotypic traits. However, Extreme Learning Machine model had a better prediction ability for catechin, astringency, brightness, briskness, and colour based on putative QTLs + annotated proteins + KEGG pathways approach. The percent variables of importance of putatively annotated proteins and KEGG pathways were associated with the phenotypic traits. The percent variable of importance of annotated proteins associated with %RWC trait were actin, armadillo beta-catenin-like repeat and isocitrate isopropyl malate dehydrogenase which are associated with abiotic stress. The KEGG pathways were arginine and proline metabolism, glycerolipid metabolism and fructose and mannose metabolism, which are primary metabolites also associated with abiotic stress. The high percentage of importance of putatively annotated proteins for caffeine trait was N-(5' phosphoribosyl) anthranilate (PRA) isomerase, which is involved in the first and intermediate pathway in the purine biosynthesis. For KEGG pathways, the high percentage of importance were those associated with carbon fixation in photosynthetic organisms, arginine biosynthesis, and purine metabolism, which are associated with caffeine biosynthesis. The putatively annotated proteins for catechin trait were those associated with phenylalanine, tyrosine and tryptophan biosynthesis and carbon fixation, which are products for shikimate pathway. Aminotransferase I and II enzymes which are involved in the biosynthesis of phenylalanine and tyrosine, the precursors of flavonoid biosynthesis, were putatively annotated for catechin trait. Furthermore, aminotransferase I and II had high percent variable of importance of annotated proteins associated with tea tasters' scores which include, astringency, brightness, briskness, and colour.

7.2 Recommendations

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

- The detection of several other peaks besides the five major peaks identified using UPLC is an indication that there are several other biochemical compounds present in the tea extracts. The identification of the unknown biochemical compounds or peaks may provide more information on their role in determining black tea quality and drought tolerance.
- This current study was conducted at one site in Timbilil Estate, Kericho and therefore the generated information and knowledge are only applicable to that region. Therefore, there is a need to replicate the same experiment carried out in this current study in other sites to precisely strengthen the reliability, and the stability of putative QTLs detected and to precisely estimate phenotypic, genotype and environment variations which are major determinants of heritability of a trait.
- To validate the effects of detected QTLs using independent and large populations constructed either from the same parental genotypes or closely-related genotypes which are established in different environments.
- Future application of the machine learning techniques in tea breeding with additional phenotyping of discovery and validation populations in different locations over multiple years to yield more stable estimates of the phenotypic performance, which could further increase the accuracy of the prediction models.
- To increase the training population size so as to improve the predictive ability of the models as opposed to using only a few parental clones for outbreeding plant species such as tea.
- To plant the parents with their F₁ progenies of the discovery population and validation population in different geographical locations, so as to increase the number of reference genotypes available and also to estimate genotype by environment interaction.