THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Genomic studies of biochemical compounds determining arabica coffee (*Coffea arabica* L.) quality

Thi Minh Hue Tran

Bachelor of Plant Science

Master of Science in agricultural biotechnology

A Thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2017

Queensland Alliance for Agriculture and Food Innovation (QAAFI)

Centre for Nutrition and Food Sciences

Abstract

Coffee is an important crop globally. Improving coffee quality is a primary breeding target that would benefit from understanding of trait inheritance and ancestral genetic information. However, comparatively little information exists on the genetics of quality and the relationships between wild and cultivated coffee genotypes, especially arabica coffee (*Coffea arabica* L.) which accounts for almost 60% of coffee production globally. The complex polyploidy genome and limited genomic information for *C. arabica* have impeded the progress of genetic studies. Arabica coffee is among the major organisms that do not have a reference genome. This project aimed to develop and utilise genomic resources in genetic studies of arabica coffee quality. The relationship between coffee species and the position of arabica within the *Coffeeae* tribe were investigated using complete chloroplast genome sequences to determine the value of these genetic resources in breeding. A draft genome sequence of arabica coffee was developed. Marker-trait associations were studied for caffeine and trigonelline, the two principal compounds known to be related to coffee quality, paving the way for developing new improved varieties with preferred levels of these compounds in the coffee beans.

Chloroplast genomes of 16 coffee species were sequenced and their phylogeny constructed. Results support distinct *Psilanthus* and *Coffea* clades. It is likely that *C. canephora* is a hybrid that has *a Psilanthus* maternal genome but received much of its nuclear genome from *Coffea*. The maternal genomes of *C. arabica* and *C. canephora* are divergent. This result is in agreement with the fact that the chloroplast genome of arabica should be that of the maternal parent i.e. *C. eugenioides*. There were two species (*C. humblotiana* and *C. tetragona*) close to *C. arabica* and one species (*P. ebracteolatus)* close to *C. canephora* containing almost no caffeine. They could serve as important materials in arabica quality breeding and research.

For the association study, 232 diverse arabica coffee accessions originating from 27 countries were harvested from the germplasm collection at CATIE (Tropical Agricultural Research and Higher Education Centre), Costa Rica. Substantial variation between genotypes was observed for bean morphology attributes. Non-volatiles including caffeine and trigonelline showed larger variation in range than was previously reported. Results of targeted analysis of 18 volatiles from 35 accessions also showed significant variation. No strong correlation was found between bean morphology and the levels of non-volatile or volatile compounds, implying that it is difficult to select for low or high non-volatile and volatile compounds based on bean physical characteristics. However, it also indicates that breeding for desirable combinations of traits (i.e. large bean size, low caffeine, high trigonelline, and favourable volatiles) is possible.

The genome of the most popular arabica variety (K7) in Australia was sequenced. Genome assembly was performed using both Illumina short reads and PacBio long reads. Assembly was performed using a range of assembly tools resulting in 76,409 scaffolds with a scaffold N50 of 54,544 bp and a total scaffold length of 1,448 Mb. Validation of the genome assembly showed high completeness of the genome in which BWA analysis demonstrated that > 98% of the short reads mapped to the genome and > 93% were marked as properly paired. GMAP analysis indicated that > 99% of the CDS and transcriptome sequences mapped to the *C. arabica* draft genome and 89% of BUSCOs were present. The assembled genome was annotated using AUGUSTUS and yielded 99,829 gene models. The assembly outcomes were used as reference for association analysis.

Extreme-phenotype genome-wide association study (XP-GWAS) was performed to identify loci affecting the caffeine and trigonelline content of *C. arabica* beans. DNA extracted from individuals with extreme phenotypes (high vs. low caffeine, and high vs. low trigonelline) was bulked based on biochemical analysis of the germplasm collection. Sequencing and mapping using the combined reference genomes of *C. canephora* (CC) and *C. eugenioides* (CE) identified 1,351 non-synonymous SNPs that distinguished the low- and high-caffeine bulks. Gene annotation analysis with Blast2GO revealed that these SNPs corresponding to 908 genes with 56 unique KEGG pathways and 49 unique enzymes. Based on KEGG pathway-based analysis, 40 caffeine-associated SNPs were discovered, among which nine SNPs were tightly associated with genes encoding enzymes involved in the conversion of substrates (i.e., SAM, xanthine and IMP) which participate in the caffeine biosynthesis pathways. Likewise, 1,060 non-synonymous SNPs were found to distinguish the low- and high-trigonelline bulks. They were associated with 719 genes involved in 61 unique KEGG pathways and 51 unique enzymes. The KEGG pathway-based analysis revealed 24 trigonelline-associated SNPs tightly linked to genes encoding enzymes involved in the conversion of substrates (i.e. SAM, L-tryptophan) which participate in the trigonelline biosynthesis pathways. Association analysis using the K7 arabica reference genome identified several additional SNPs linked to genes encoding enzymes involved in caffeine and trigonelline synthesis pathways. These SNPs could be useful targets for further functional validation and subsequent application in arabica quality breeding.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including samples collection, sequencing raw data, genome assembly, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the Copyright Act 1968.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Peer-reviewed papers

1. Tran, H. T., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. J. (2016). Advances in genomics for the improvement of quality in coffee. *Journal of the Science of Food and Agriculture* **96**, 3300-12.

2. Tran, H. T.M., Vargas, C.A.C., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. J. (2017). Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (Coffea arabica L.). *Tree Genetics & Genomics*. **13:54**, 1-14.

3. Tran, H. T. M., Ramaraj, T., Furtado, A., Lee, L. S., and Henry, R. J. (2018). Use of a draft genome of coffee (Coffea arabica) to identify SNPs associated with caffeine content. *Plant Biotechnology Journal.* https://doi.org/10.1111/pbi.12912.

4. Tran, H. T.M., Furtado, A., Vargas, C.A.C., Smyth, H., Lee, L. S., and Henry, R. J. (2018). SNP in the *Coffea arabica* genome associated with coffee quality. *Tree Genetics & Genomics* (accepted subject to minor revision).

Conference abstracts

1. Hue T.M. Tran, Tal Cooper, Agnelo Furtado, Darren Crayn, Robert J. Henry and Perla Hamon (2015). Whole chloroplast genome sequence of coffee species reveals their relationships. Plant and Animal Genome Asia Conference, Singapore. Available at

   https://pag.confex.com/pag/asia2015/webprogram/Paper17984.html

2. Hue T.M. Tran, Carlos Alberto Cordero Vargas, Slade Lee, Agnelo Furtado, Heather Smyth, Robert Henry (2015). Variation of physical characters of green bean in coffee (*Coffea arabica)* germplasm. Tropical Agricultural conference (TropAg 2015). Brisbane, Australia. Available at

   https://tropagconference.org/d/TropAg2015-Abstract-Book.pdf

3. Hue T.M. Tran, Carlos Alberto Cordero Vargas, L. Slade Lee, Agnelo Furtado, Heather Smyth, Robert Henry (2016). Genetic diversity of arabica coffee *(C. arabica)* based on morphological and biochemical analysis. Plant and Animal Genome (PAG XXV) Conference. San Diego, USA. Available at https://pag.confex.com/pag/xxiv/webprogram/Paper19979.html

4. Tran, Hue T.M., Vargas, Carlos Alberto Cordero, Lee, L. Slade, Furtado, Agnelo, Smyth, Heather, Henry, Robert (2016). Diversity and genetic analysis of bean morphology and quality compounds in arabica coffee (*Coffea arabica* L.). The 26th International Conference on Coffee Science (*ASIC 2016).* Yunnan, China. Available at

http://www.asic2016china.org/upload/file/20160902/14727996717695860.pdf

5. Tran, Hue T.M., Vargas, Carlos Alberto Cordero, Lee, L. Slade, Furtado, Agnelo, Smyth, Heather, Henry, Robert (2017). Association mapping and SNP discovery using extreme phenotypes for non-volatile compounds selected from a *C. arabica* diversity population. Tropical Agricultural conference (TropAg 2017). Brisbane, Australia.

Publications included in this thesis

1. Tran, H. T., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. J. (2016). Advances in genomics for the improvement of quality in coffee. *Journal of the Science of Food and Agriculture* **96**, 3300-12. A review article - incorporated as Chapter 2.

Contributor	Statement of contribution
Tran, H. T.	Conceptualised and outlined the manuscript (80%) Wrote the paper (100%)
Henry, R. J.	Conceptualised and outlined the manuscript (20%) Edited paper and contributed to the completion of the manuscript (70%)
Lee, L. S.	Contributed to the completion of the manuscript (10%)
Furtado, A.	Contributed to the completion of the manuscript (10%)
Smyth, H	Contributed to the completion of the manuscript (10%)

2. Tran, H. T., Vargas, C.A.C., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. J. (2017). Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (*Coffea arabica* L.). *Tree Genetics & Genomics*. **13:54**, 1-14 - incorporated as Chapter 4.

Contributor	Statement of contribution
Tran, H. T.	Conceptualised and outlined the manuscript (70%) Collected coffee samples (70%) Carried out bean measurement and sample preparation for HPLC and GC-MS analysis (100%) Conducted the volatile compounds identification and quatification (70%) Performed data analyses (90%) Wrote the paper (100%)
Henry, R. J.	Conceptualised and outlined the manuscript (30%)

	Edited paper and contributed to the completion of the manuscript (60%)
Vargas, C.A.C.	Collected coffee samples (30%)
Smyth, H	Conducted the volatile compounds identification and quatification (30%) Performed data analyses (10%) Contributed to the completion of the manuscript (20%)
Lee, L. S.	Contributed to the completion of the manuscript (10%)
Furtado, A.	Contributed to the completion of the manuscript (10%)
All authors	Reviewed and approved the final manuscript.

3. Tran, H. T. M., Ramaraj, T., Furtado, A., Lee, L. S., and Henry, R. J. (2018). Use of a draft genome of coffee (Coffea arabica) to identify SNPs associated with caffeine content. *Plant Biotechnology Journal*. https://doi.org/10.1111/pbi.12912.

Contributor	Statement of contribution
Tran, H. T.	Conceptualised and outlined the manuscript (70%) Performed data analyses on genome assembly (10%) Performed data analyses on SNP identification (100%) Wrote the paper (100%)
Henry, R. J.	Conceptualised and outlined the manuscript (30%) Edited paper and contributed to the completion of the manuscript (50%)
Ramaraj, T.	Performed data analyses on genome assembly (90%) Performed data analyses on genome annotation (100%) Edited paper and contributed to the completion of the manuscript (30%)
Lee, L. S.	Contributed to the completion of the manuscript (10%)
Furtado, A.	Contributed to the completion of the manuscript (10%)
All authors	Reviewed and approved the final manuscript.

4. Tran, H. T.M., Furtado, A., Vargas, C.A.C., Smyth, H., Lee, L. S., and Henry, R. J. (2018). SNP in the *Coffea arabica* genome associated with coffee quality. *Tree Genetics & Genomics* (accepted subject to minor revision).

Contributor	Statement of contribution
Tran, H. T.	Conceptualised and outlined the manuscript (70%) Performed data analyses (80%)

	Wrote the paper (100%)
Henry, R. J.	Conceptualised and outlined the manuscript (30%) Edited paper and contributed to the completion of the manuscript (60%)
Vargas, C.A.C.	Contributed to the completion of the manuscript (10%)
Smyth, H	Contributed to the completion of the manuscript (10%)
Lee, L. S.	Contributed to the completion of the manuscript (10%)
Furtado, A.	Contributed to the completion of the manuscript (10%) Performed data analyses (20%)
All authors	Reviewed and approved the final manuscript.

Contributions by others to the thesis

Contributor	Statement of contribution
Coffee Consortium	Provided raw reads of 13 coffee species Provided the reference genome
Dr Thiruvarangan Ramaraj	Run genome assembly using different software packages. Validated the assembled genome Annotated the assembled genome

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

In this section, I would like to acknowledge my gratitude for the general academic support that I received during the past four years, while studying at the University of Queensland (UQ). So many people have helped me in so many ways that it will be difficult to name all.

I shall start by acknowledging the support of the original advisory team: Prof. Robert J. Henry, Dr Agnelo Furtado, A/Prof. Slade Lee and Dr Heather Smyth. I thank A/Prof. Slade Lee for the very first time he suggested I should contact and work with Prof. Robert J. Henry, and continuously supporting me and giving me advice, not only in my study, but also in my personal life or family aspects. I thank Prof. Robert J. Henry – my principal supervisor - for the supervision and guidance from the very first days till the end of the study; and for sustained enthusiasm and patience throughout all phases of the project. I extend my very warmest and most sincere thanks to him for bringing me not only the knowledge but also the confidence and encouragement. I would like to thank Dr Agnelo Furtado for helping me out with all technical issues, as well as calming me down when I got stressed. I would like to thank Dr Heather Smyth for her support and guidance in chemistry analysis.

Furthermore, I thank the CATIE team who helped me with sample collection and treatment, especially thanks to Carlos Cordero and Manrique Gonzalez for their great help. I would like to acknowledge Peter Wolff – Wolff Coffee - for bean roasting and explaining to me all about roasting with his broad knowledge. I would like to express my gratitude to Wenny Sunarharum, Steve Fuller and Dr Kent Fanning for their great advice and laboratory analysis in relation to chemistry – the part in which I do not have much background and found so stressful.

I would like to thank Dr Thiru Ramaraj and Dr Anitha Sundararajan for their welcoming nature, caring and for making me feel as part of their team as soon as I set foot in New Mexico, thanks to Dr Andrew Farmer, Dr Joann Mudge and Dr Diego Fajardo for their advice to my research direction, and Jennifer Jacobi for guiding me on Illumina and PacBio lab tour.

I would like to acknowledge Coffee Consortium for providing me coffee sequence and allowing me to access to coffee reference.

I would like to acknowledge the financial support received from UQ through 1) Australia Awards Scholarship (AAS) for my allowance during my PhD, conference in Singapore and field trip in Vietnam, and partly funding for ASIC conference (International Conference on Coffee Science) in China; 2) Australian Research Council (ARC) and Green Cauldron for my research and sample collection trip to Costa Rica, as well as partly funding for ASIC conference. I am also grateful for the financial support provided by the Graduate School via the Graduate School International Travel

Award (GSITA) and Queensland Alliance for Agriculture & Food Innovation via QAAFI travel award to attend the conference in San Diego, California and lab visit in Santa Fe, New Mexico.

I acknowledge QAAFI staff for administration support, especially thanks to Annie Morley and Emma Linnell, Robert Henry's research team for numerous scientific discussions and technical guidance throughout the PhD, namely: Hoang Van Nam, Peterson Wambugu, Adam Healey, Marta Brozynska, Bing Cheng, Ravi Nirmal; Tiparat Tikapunya; Tom Kukhang ….

I thank my home institution Western Highlands Agro-Forestry Science and Technical Institute (WASI) for permission, and Dr Le Ngoc Bau, Dr Truong Hong, and Dr Trinh Duc Minh for his advice, encouragement and moral support in my study.

I am grateful to my parents and my parents-in-law for considerate understanding and support, the rest of my family, my siblings and in-laws for moral support.

Finally but most importantly and decisively, my beloved husband Tran Anh Hung for his endless encouragement, patience and sacrifice when leaving his job in Vietnam and staying here in Australia to share with my work and life and take care of our children. I would like to thank my cherished son Tran Anh Minh and beloved daughter Tran Ngoc Minh Dan for their understanding not having enough time with them, and tolerance when their mum was getting stress.

I would like to thanks my friends Tran My Hanh, Tran Chi Trung and Nguyen Thu Van for their support to my children when I was away on field trips or conferences and when my husband was in Vietnam.

I am indebted to all of you.

Financial support

This research was supported as a Linkage Project by the Australian Research Council and Australia Awards Scholarship (AAS)

Keywords

Coffee, genomics, genetics, quality, GWAS, SNPs, caffeine, trigonelline, volatile compounds

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060408 Genomics - 70%

ANZSRC code: 060101 Analytical Biochemistry - 25%

ANZSRC code: 070602 Horticultural Crop Improvement (Selection and Breeding) - 5%

Fields of Research (FoR) Classification

FoR code 0604: Genetics - 70%

FoR code 0601: Biochemistry and cell biology - 25%

FoR code 0706: Horticultural production - 5%

Table of Contents

Tables

Figures

List of abbreviations

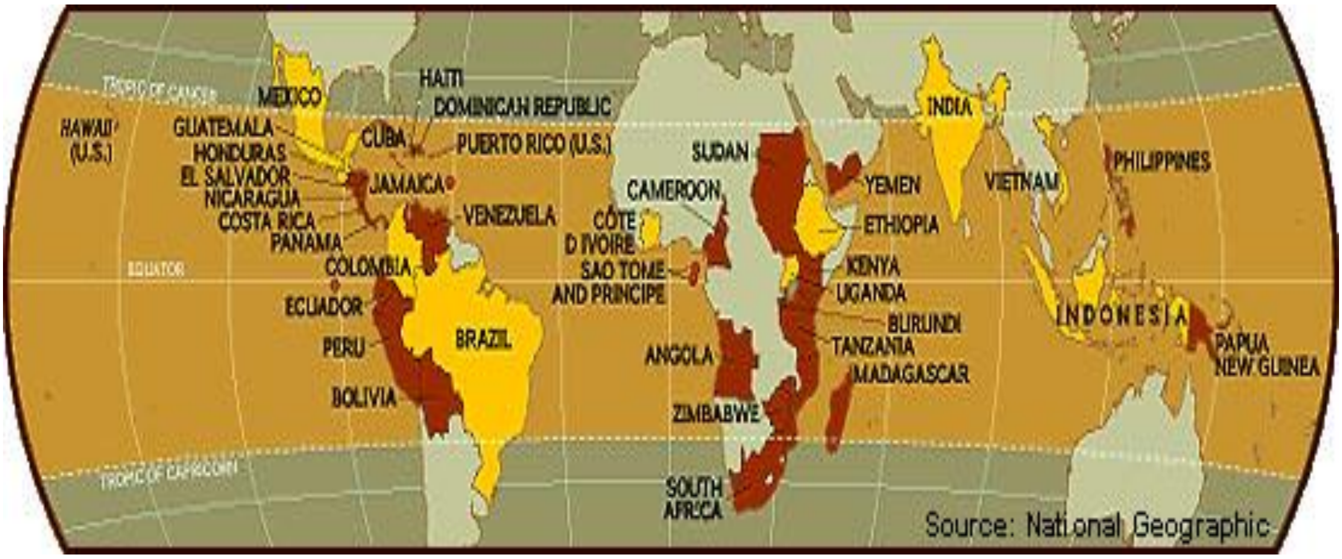
ADP	Adenosine diphosphate
AFLP	Amplified fragment length polymorphism
ANOVA	Analysis of variance
ATP	Adenosine triphosphate
BAC	Bacterial Artificial Chromosome
BSA	Bulk Segregant Analysis
BSH	Broad-sense heritability
BUSCO	Benchmarking Universal Single-Copy Orthologs
BWA	Burrows-Wheeler Aligner
CATIE	Centro Agronómico Tropical de Investigación y Enseñanza – Centre for

	Tropical Agricultural Research and Higher Education
CC	*C. canephora* genome
cDNA	Complementary DNA
CDS	Coding DNA Sequence
CE	*C. eugenioides* genome
CEGMA	Core Eukaryotic Genes Mapping Approach
CGAs	Chlorogenic acids
CP	Chloroplast
CTAB	Cetyltrimethylammonium bromide
CV	Coefficient of variation
DH	Doubled haploid
DNA	Deoxyribonucleic acid
EaEaCaCa	*C. arabica* genome
EDTA	Ethylenediaminetetraacetic acid disodium salt
EIC	Extracted ion chromatogram
EST	Expressed Sequence Tag
F/R	forward and reverse reads
FID	Flame ionization detectors
GC-MS	Gas Chromatography - Mass spectrometry
GM	Genetically Modified
GMAP	Genomic mapping and alignment program
GO	Gene Ontology
GTR	General Time Reversible Model
GWAS	Genome-wide association studies
GWB	Genomics Workbench
HPLC	High Performance Liquid Chromatography
HS-SPME	Headspace - Solid phase microextraction
ICO	International Coffee Organization
IGS	intergenic spacer
indels	Insertions and deletions
IR	Inverted repeat
IRD	Institute of Research for Development
ISO	International Organization for Standardization
ISSR	Inter-simple sequence repeat
ISSR	inter simple sequence repeat
ITS	Internal transcribed spacer
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage disequilibrium
LDS	Least Significant Difference
LF	Length fraction
LSC	Large single copy
MAS	Marker-assisted selection
MCMC	Monte Carlo Markov Chain
ML	Maximum likelihood
MNV	Multi-nucleotide variant

MP	Maximum parsimony
MSD	Mass spectrometry detector
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
NMTs	N-methyltransferases
NSH	Narrow-sense heritability
PacBio	Pacific Biosciences
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PE	Paired end reads
PID	Photoionisation detector
PVP	Polyvinylpyrrolidone
QAAFI	Queensland Alliance for Agriculture & Food Innovation
QC	Quality Control
QTL	Quantitative Trait Locus
RAPD	Random amplified polymorphic DNA
RLFP	Restriction fragment length polymorphism
Rt	Retention time
SAH	S-adenosyl-L-homocysteine
SAM	S-adenosyl-Lmethionine
SD	Standard Deviation
SF	Similarity fraction
SIDA	Stable isotope dilution analysis
SIM	Selected ion monitoring
SMRT	Single-molecule, real-time
SNP	Single Nucleotide Polymorphism
SPE	Solid phase extraction
SRAP	Sequence-related amplified polymorphism
SSC	Small single copy
SSR	Simple Sequence Repeat
TASs	Trait-associated SNPs
TAVs	Trait-associated variants
TEs	Transposable elements
TIC	Total ion current
TRAP	Target Region Amplification Polymorphism
UPGMA	Unweighted pair group methods using arithmetic averages
UQ	The University of Queensland
UTR	Untranslated region
XP-GWAS	Extreme Phenotype Genome-Wide Association Studies

CHAPTER 1: GENERAL INTRODUCTION

Coffee is an important crop and the second most traded commodity in the world (after petroleum) providing a living to more than 125 million people in the world. Coffee is consumed in the form of roasted coffee (77.4%) and soluble coffee (22.6%) and this consumption has increased rapidly over the last ten years (ICO, 2013). Coffee is grown primarily between the Tropics of Cancer and Capricorn (Figure 1.1). The top coffee producing countries are Brazil (32%), Vietnam (18%), Indonesia (6.5%), Columbia, Ethiopia, India and Honduras. The world leading coffee importing countries are France (7%), Germany (4.8%), Netherlands, USA and Belgium. World production of coffee in 2014/15 was 143.3 million 60-kg bags of which 57.85% was arabica and the remaining robusta coffee (http://www.ico.org).



Source: National Geographic

Figure 1.1 The coffee belt bounded by the Tropics of Cancer and Capricorn

(Brown: the Tropics of Cancer and Capricorn; Yellow: coffee growing areas; Source: National Geographic)

Coffee belongs to the Rubiaceae family and the *Coffeeae* tribe, but its taxonomy and classification is still controversial. Many previous studies divided *Coffeeae* tribe into two genera *Coffea* L. and *Psilanthus* Hook.f (summarised by Anthony et al., 2011) while Davis et al. (2011) grouped the *Coffea* and *Psilanthus* into one genus with 125 species distributed in Africa, Madagascar, the Comoros Islands, the Mascarene Islands, tropical Asia, and Australia (reviewed by Krishnan et al., 2013). Commercial coffee production is dominated by only two species belonging to the *Coffea* genus: *C. arabica* and *C. canephora* (generally referred to as robusta coffee) (Figures 1.1 and 1.2) and to a much lesser extent, *C. liberica* Hiern (Liberian, or Liberica, or excelsa coffee) (Davis et al., 2006). All coffee species are diploid (2n=2x=22) and generally self-incompatible, except for two diploid *C. heterocalyx* and *C. anthonyi* (Stoffelen et al., 2009) and the only tetraploid (2n=4x=44) and self-fertile *C. arabica* which was derived from a spontaneous hybridisation between *C.*

canephora (as paternal progenitor) and *C. eugenioides* (as maternal progenitor) (Charrier and Eskes, 2004).

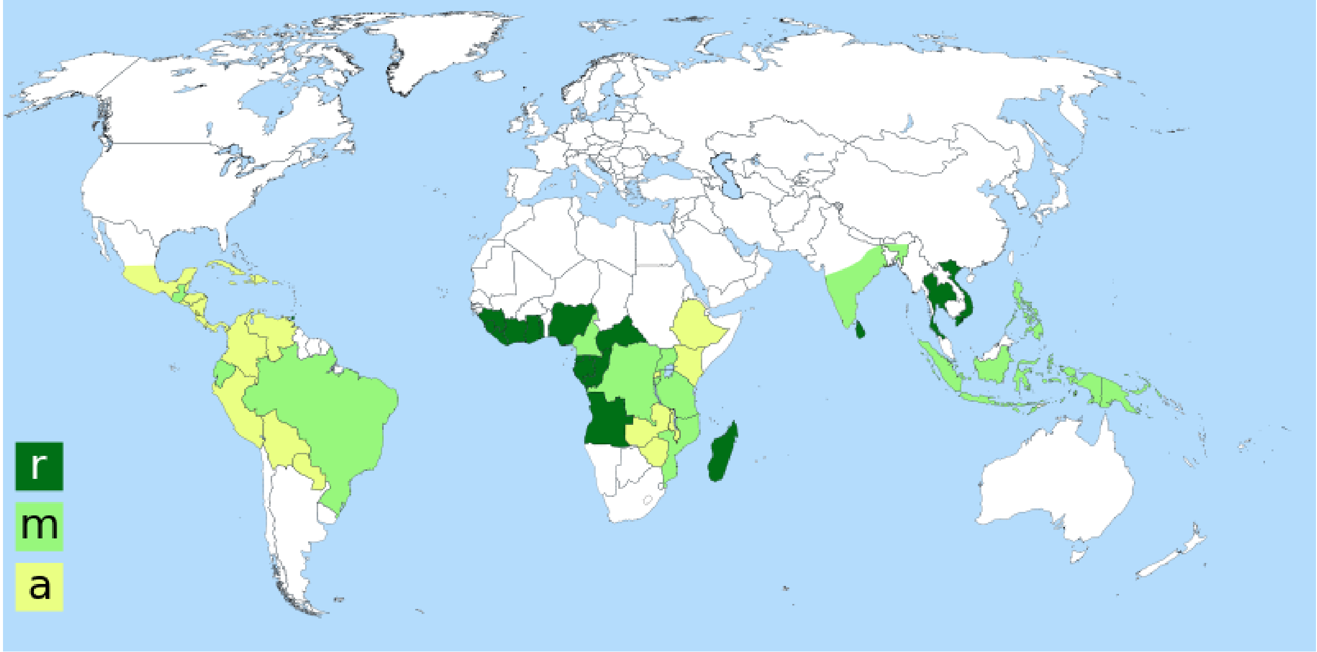


Figure 1.2 Map of world coffee cultivation areas

(r: robusta, m: robusta and arabica, a: arbica; Source: Wikimedia Commons)

Genetic improvement of complex-genome species like arabica can benefit from understanding of ancestral information and trait inheritance at the genomic level. However, such information has been limited for arabica coffee compared to its diploid relative *C. canephora*. It is among the significant organisms that lack a reference genome. This project aims to develop and ulilize new genome resources in genetic studies of arabica coffee. It is hypothezied that applying whole chloroplast genome sequences can resolve the relationship between coffee species and position of arabica within *Coffeeae* tribe, while complete nuclear sequences will enable association genetics to identify markers or genes affecting biochemical compounds linking to coffee quality in arabica. The project will (1) investigate the relationship between coffee species within Coffeeae tribe at the chroloplast genome level, (2) examine the diversity within arabica and wild relatives in terms of bean morphology, physical and chemical characteristics, (3) develop a draft genome sequence of arabica coffee, and (4) use this new resource and other existing sequence information in association analysis to identify loci associated with known quality-determining compounds, with a focus on caffeine and trigonelline. The ultimate goal is to provide potential favorable haplotypes for breeding programs that focus on quality improvement through designing different quality ideotypes within arabica and/or in interspecific hybrids between arabica and wild relatives (*C. humblotiana*, *C. tetragona*) to combine different quality traits. The project will also contribute knowledge and development on coffee genetic and genomic resources.

The overall research plan is outlined in Figure 1.3. First, diversity of arabica vs. other coffee species was examined using chloroplast genome sequences (Chapter 3). Second, a global germplasm set of arabica coffee also including wild relatives was examined for variation in bean morphology and biochemical compounds including volatiles and non-volatiles (Chapter 4). In parallel, leaves of K7 coffee – the most popular arabica genotype in Australia - were collected for DNA extraction then sequenced, followed by assembly, validation and annotation (Chapter 5). Based on results from Chapter 4, leaf samples of individuals that showed extreme phenotypic values in non-volatile compound contents were selected for DNA extraction and bulked for sequencing and mapping to reference sequences (generated from Chapter 5) to identify SNPs associated with the target non-volatile compounds (Chapter 6).

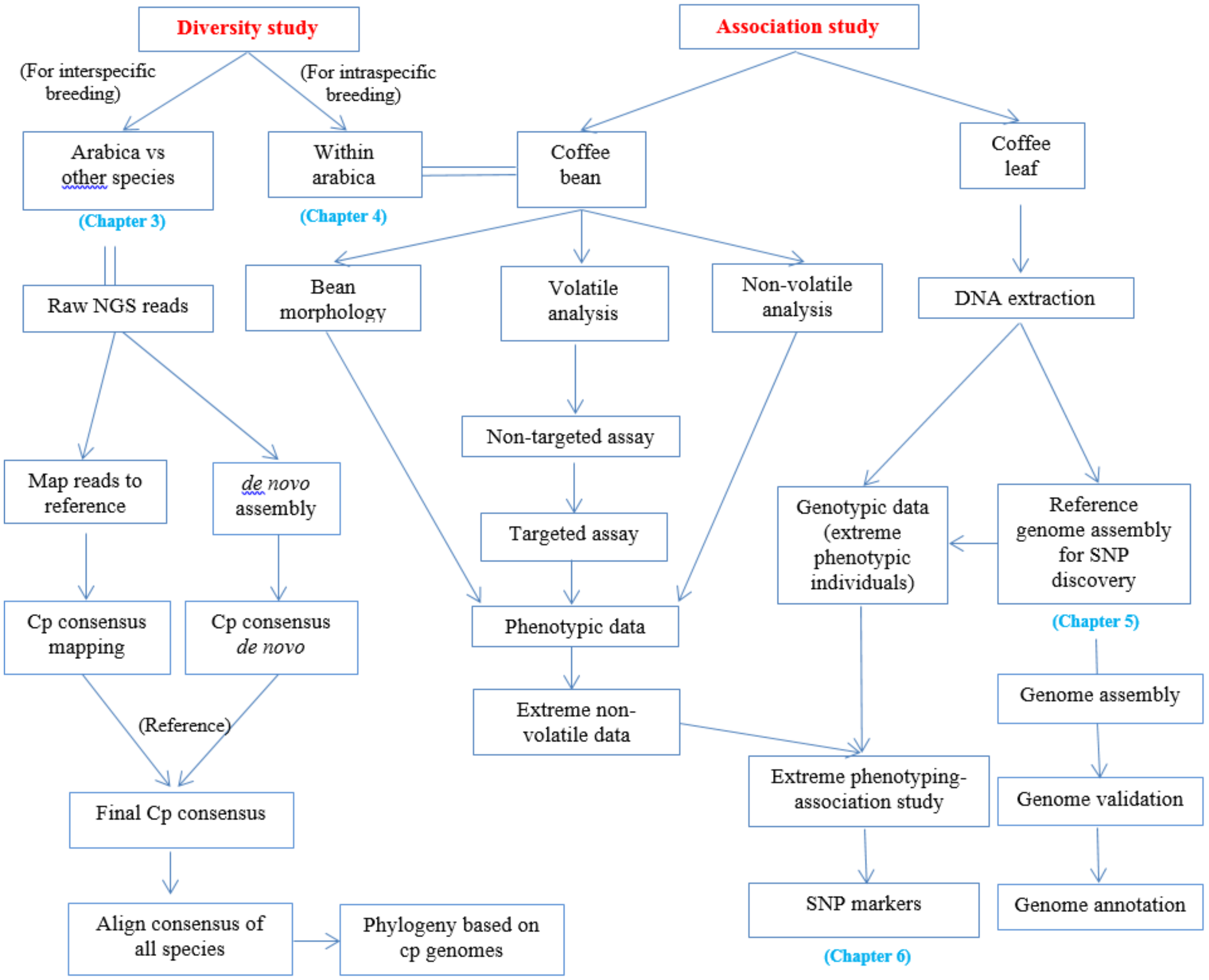


Figure 1.3 Overall research plan

The thesis has seven chapters:

Chapter 1 provides a general introduction on the overall project, including research background, hypothesis, objectives, significance, research plan, and thesis structure.

Chapter 2 provides a comprehensive review on the importance of coffee, the coffee taxonomy, up-to-date information on coffee quality and its determinants, factors affecting them, methods for biochemical compound analysis, and coffee genomic resources.

Chapter 3 provides phylogenetic analysis using available *Coffea* chloroplast genomes. This will help understand the relationship between species within Coffea genus and give the overall pictures of *Coffea* genomes and relationship among them and with the studied species – arabica. The chapter provides information for breeders in relation to quality improvement, especially in interspecific hybridisation. This chapter also provides information on the specific compounds in each species studied, such as those with extreme levels of compounds mentioned in the literature review that could be utilised in arabica breeding.

Chapter 4 presents the variation in bean morphology and biochemical compounds, as well as chemistry analysis of a global germplasm collection for arabica coffee. It reflects the variation of quality determinants including volatiles and non-volatiles among individuals within the arabica germplasm. The result enables selection of groups of individuals with extreme phenotypic data for use in Chapter 6.

Chapter 5 involved arabica genome assembly which will be used as reference for mapping to find trait-associated SNPs in Chapter 6. This chapter described how genome assembly was performed, and how to validate and annotate the draft genome.

Chapter 6 presents the application of XP-GWAS in discovery of trait-associated SNPs for caffeine and trigonelline. This state-of-the art technique was developed for other crops but used in coffee genomic study for the first time in this thesis, so its suitability for arabica will also be discussed.

Chapter 7 provides general discussion and suggestions for future studies.

CHAPTER 2: LITERATURE REVIEW [1]

2.1 Introduction

Although *C. arabica* is considered to have better quality and flavour than *C. canephora*, improving quality of both commercial species remains a target for most coffee improvement programs. Arabica is the dominant coffee species commercially and thus has a higher priority for genetic improvement. With the advances in genomic and sequencing technology, it is feasible to understand the coffee genome and molecular inheritance underlying coffee quality, thereby helping improve the efficiency of breeding programs.

This review will cover aspects of coffee quality including characters used in assessment of green and roasted coffee bean quality, main compounds determining quality, with a focus on caffeine and trigonelline, how they are measured, their inheritance and non-genetic factors affecting their presence in the bean. Available genomic resources as well as gene technology for manipulating biochemical compounds determining coffee quality will also be discussed. Finally, association mapping approaches using next generation sequencing and its potential application for the study of genetics of coffee quality will also be reviewed.

2.2 Coffee quality

Coffee quality is comprised of two components: physical and organoleptic. Physical quality reflects moisture content, defects, bean size and bean colour; while organoleptic quality reflects aroma, taste, flavour, body, acidity and preference of tasters (Leroy et al., 2006).

2.2.1 Physical quality and its link to organoleptic quality.

- Moisture: Coffee is ideally kept at a moisture content of 12% for arabica and 13% for robusta (Rojas, 2004). If the beans have low moisture content (below 8%), they will change colour, as well as cup taste and consistency (Leroy et al., 2006; Rojas, 2004).

- Defects: A standard named ISO 10470 has been established by ISO (2004) that describes defect including foreign materials of non-coffee origin (sticks, stones), non-bean origin (pieces of parchment or husks), abnormal beans for shape regularity (damaged bean, broken bean, pea bean, elephant bean), for visual appearance (black beans) and the cup taste after proper roasting and brewing.

- Bean size (or so-called ‘grade’ in coffee commerce): Bean size is an important factor in commerce since it relates to price; lower price is paid for small beans of the same variety. Beans of consistent

[1] This chapter contains information that is also published in: Tran, H. T., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. J. (2016). Advances in genomics for the improvement of quality in coffee. *Journal of the Science of Food and Agriculture* **96**, 3300-12.

size are ideal for roasting since with unevenly size of beans, the result is burned for the smallest beans and under-roasted for the largest beans, which affects the visual appearance of the beans and, more seriously, affects the cup quality (Barel and Jacquet 1994, originally in French cited by Leroy et al. (2006). Genotypes play a role in defining bean size, where arabica beans are larger (18-22 g/ 100 beans) and denser than robusta beans (12-15 g/100 beans) (Wintgens, 2012)

- Bean colour: Bean colour reflects the freshness and homogeneity of coffee. Washed arabica often has green-bluish colour as the indication of high quality (Wintgens, 2004b) while robusta has browner beans. There are minor differences between arabica varieties in bean colour (Wintgens, 2004b).

2.2.2 Organoleptic quality

Organoleptic quality is difficult to assess. It is often assessed via the aroma, the evaluation of body and perception of taste and flavours. The measurement of organoleptic quality relies on sensory evaluation (Leroy et al., 2006). This is commonly referred to as 'cupping quality' or simply 'cup'.

According to Wintgens (2004b), the characteristics of coffee cup could be described via:

- Aroma: The fragrance or odour perceived by the nose such as the scents of burnt or smoky, chemical or medicinal, and chocolate-like presence.

- Taste: the bitterness, sweetness and/or saltiness perceived by the tongue.

- Flavour: the combination of aroma and taste, described as good flavour (spicy, winey and fragrant) and off flavour (onion, musty, grassy, earthy, etc.). Four basic flavours commonly applied for sensory evaluation are acidity, sweetness, saltiness and bitterness. Sweetness and the interaction between sweetness and acidity provide a broad spectrum of flavours for arabica coffee. Bitterness and saltiness are more common with robusta and low-quality arabica or arabica processed by dry method.

- Body: a feeling of the heaviness or richness on the tongue.

- Acidity: a sharp and pleasing taste, ranging from sweet to fruity/citrus and is considered as a favourable attribute.

- Preference: the liking of the taster and has a good linear correlation with acidity and aroma (Wintgens, 2004b).

Recently, SCAA and WCR developed a new Coffee Taster's Flavor Wheel to standardize the description of coffee flavors (Spencer et al., 2016). In addition, WCR also completed the Sensory Lexicon - a universal language of coffee's sensory qualities, and a tool for measuring them (WCR, 2017). These help to describe precisely the characteristics of coffee cup.

Among several influencing factors, biochemical compounds in coffee beans are most important to the quality of cup.

2.3 The biochemistry of coffee quality and methods of analysis

2.3.1 Chemical composition in coffee bean and its association with coffee quality

2.3.1.1 Non-volatiles

Green or unroasted coffee beans have chemical components in the non-volatile fraction such as carbohydrates and fibre (sucrose, reducing sugars, polysaccharides, lignin and pectins), nitrogenous compounds (protein, free amino acids, caffeine, trigonelline), lipids (coffee oil, diterpene esters), minerals (K and P), acids and esters (total chlorogenic acids, aliphatic acids and quinic acid) (Farah, 2009). These compounds play several decisive roles in the roasting chemistry (Flament, 2002). Proteins and amino acids, for example, are imperative in converting reducing sugars into aroma precursors. Chlorogenic acids and caffeine, are responsible for bitterness (Joët et al., 2009). Among these components, sucrose, caffeine, trigonelline, lipid and chlorogenic acids are major compounds that contribute importantly to the flavour of the beverage after the roasting of the beans (Farah, 2009).

Caffeine and trigonelline are two of key compounds in coffee. They play an important role in the bioactive effects of coffee including anti-proliferative, antioxidant, and antimicrobial effects (Nuhu, 2014).

2.3.1.1.1. Caffeine

Caffeine (1,3,7-trimetylxanthine, $C_8H_{10}N_4O_2$) is one of the main alkaloids contributing to the strength, body and bitterness of brewed coffee (Trugo, 1984). Caffeine can be found in seeds and leaves of coffee and its biosynthesis occurs in the cytoplasm of cells in buds and young leaves (Ogawa et al., 2001) and immature fruits, then gradually accumulates during the maturation of these organs (Ashihara et al., 2008). The variation of caffeine during fruit development was reported by Keller et al. (1972) that the relative caffeine content of the pericarp falls during fruit development (from 1.68% to 0.24% on a dry weight basis) but remains more or less constant in the seed (of about 1.25%). Caffeine content of the pericarp at the beginning of fruit development was twice as much as much at maturity and that of the seed was twenty times (Keller et al., 1972). Biochemical and molecular genetic studies using leaves of tea and coffee show that the major route of caffeine biosynthesis is from xanthosine → 7-methylxanthosine → 7-methylxanthine → theobromine (3,7N-dimethylxanthine) → caffeine with three methylation and one nucleosidase reactions (Ashihara et al., 2008) (See Figure 2.1). Koshiro et al. (2006) found that the pattern of caffeine synthesis during fruit development is not much different between *C. arabica* and *C. canephora*. Active caffeine biosynthesis occurs in several parts of the fruit (from pericarp expansion to endosperm formation, young seeds through to maturation) (Koshiro et al., 2006). It is reported that the caffeine content has a negative correlation with cup-quality attributes, while no correlation was observed between caffeine content and the physical characteristics of green bean (Dessalegn et al., 2008). The caffeine content of green coffee varies widely among species (Campa et al., 2005a) and thus could be the target for breeding programs to modify the level of caffeine in coffee.

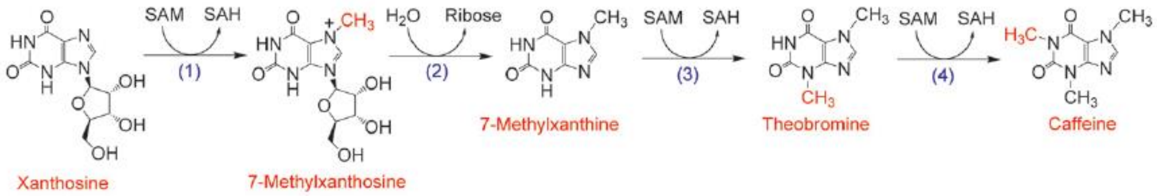


Figure 2.1 The major biosynthetic pathway of caffeine from xanthosine (adapted from Ashihara et al., 2008).

Several studies show that during postharvest processing and roasting, caffeine content remains almost unchanged or even has a slight relative increase in percentage due to the loss of other compounds (Farah et al., 2006b; Oestreich-Janzen, 2010). Under severe roasting, Trugo and Macrae (1989) found that caffeine had 5.4% loss and Franca et al. (2005) found a 30% loss of caffeine with arabica samples of different qualities.

2.3.1.1.2. Trigonelline

Trigonelline is a methyl-betaine biologically derived from enzymatic methylation of nicotinic acid (Farah, 2009). Trigonelline synthesis occurs in all parts of coffee seedlings but higher levels in young tissues (leaves, flower buds, pericarp and beans) were found (reviewed by Ashihara, 2006; De Castro and Marraccini, 2006). In addition, the net biosynthesis of trigonelline takes place in the pericarp and is then possibly transported further from this tissue to the endosperm (Zheng et al., 2004). The pattern of trigonelline biosynthesis during fruit development is very similar to that of caffeine (Koshiro et al., 2006). The biosynthesis pathway of trigonelline involves eight steps, starting from nicotinic acid mononucleotide and trigonelline is synthesised mainly from nicotinic acid which is produced by the degradation of *Nicotinamide adenine dinucleotide (*Ashihara et al., 2011*)* (see Figure 2.2). The levels of trigonelline in green and roasted coffee beans are strongly correlated with high quality (Farah et al., 2006b). It contributes indirectly to the formation of different desirable aromas during the roasting process (Ky et al., 2001b), and thus increasing trigonelline is also a breeding objective.

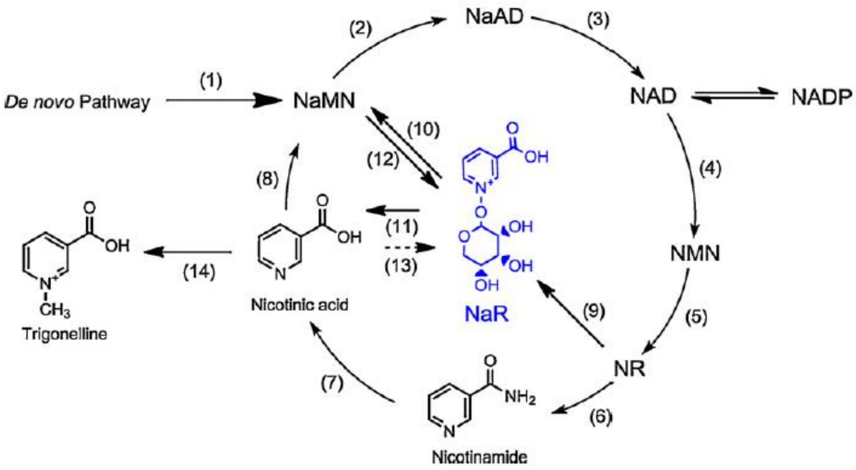


Figure 2.2 Possible pyridine nucleotide cycle and the trigonelline synthesis pathway in coffee fruits (adapted from Ashihara et al., 2011).

Dart and Nursten (1985) found that trigonelline degraded readily during roasting. The loss of trigonelline, which associated with nicotinic acid formation, was strongly dependent on the degree of roasting and was higher in the robusta coffee (Trugo and Macrae, 1989). Trigonelline and caffeine contents were used as indicators to distinguish *arabica* and *robusta* roasted coffees from several geographical origins (Casal et al., 2000). Trigonelline content may also potentially be used as a marker to discriminate growing environments, but not for genotype differentiation (Figueiredo et al., 2013).

2.3.1.2 Volatiles

There are about 230 volatiles detected so far in green coffee bean (Holscher and Steinhart, 1995) of which 2-methoxy-3-isopropylpyrazine and corresponding isobutyl-derivatives are associated with the aroma of coffee (Vitzthum and Werkhoff, 1976).

One of the crucial steps towards a good cup of coffee is the roasting process, where various physical and chemical changes lead to the formation of the desired coffee aroma molecules. During the initial endothermic phase of roasting, the green beans dry, reducing the water content to a few percent. In this phase, there is a change in smell of the beans from green to peasy to bread-like, and the colour turns yellowish. Further heating of the beans initiates the exothermic pyrolysis reactions. The chemical composition of the beans is drastically modified, with release of large amounts of carbon dioxide and the formation of the many hundreds of substances associated with coffee aroma and taste. The beans change colour to dark brown. This phase can be perceived as a popping sound, called the first crack (at about 175–185^{0}C). At the end of the roasting process the beans are cooled quickly by flushing with air after removing the beans from the roaster (Buffo and Cardelli-Freire, 2004; Gloess et al., 2014). During roasting the coffee beans expand in size. The unit volumetric increase in roasted bean compared to that of the green bean ranged from 58.16% to 92.54% (under different roasting temperature between 340°C and 260°C) for coffee bean of mocha variety (Campos et al., 2014). Roasting temperature, time, method of roasting and cooling, and type of roaster used all affect the volatile composition. A number of studies on the effect of roasting time on the concentration of volatiles showed that some volatiles reach the peak of their concentration during a commercial roast, while others do not. Some volatiles increase significantly with time of roasting, whereas others decrease. Some volatiles even decrease and then increase again as roasting proceeds due to the volatiles being derived from two or more precursors, which degrade at different rates during roasting, or two different pathways may be involved, one requiring more energy than the other (summarised by Dart and Nursten, 1985).

Coffee flavour and aroma as well as the formation of these coffee quality determinants are extremely complex; however, numerous studies on the contribution of volatile and non-volatile compounds to coffee quality and their relationship to coffee sensory properties has been thoroughly reviewed (Flament, 2002; Sunarharum et al., 2014). The reactions involved in the formation of coffee aroma including Maillard reaction (non-enzymatic browning), Strecker degradation, breakdown of sulphur amino acids, hydroxy-amino acids and proline and hydroxyproline, degradation of trigonelline, quinic acid moiety and pigments, and its mechanisms have been reviewed by Buffo and Cardelli-Freire (2004). Also in this paper volatile compounds that contribute significantly to coffee aroma as well as classes of volatile compounds identified in roasted coffee were reviewed. However, the change in bean composition, especially in aroma profile, from green to roasted bean is complicated and not all formation pathways are fully understood under coffee roasting conditions (Mestdagh et al., 2014). The use of green or roasted bean in the study of coffee quality genetics should be considered carefully since aroma in roasted bean and brew is of customers' subjective interest, while the green bean attributes are more directly affected by genetics. This is a challenge in the study of genetics of bean composition, which is ascribing links to coffee quality.

Among 1,000 volatile compounds known in roasted bean, 20 were identified as key aroma compounds (Oestreich-Janzen, 2010) and belong to several chemical classes including furans, pyrazines, ketones, and aldehydes etc. (Flament, 2002).

Pyrazine itself has a bitter, sweet and corn-like aroma, with alkyl substitution leading to aromas such as nutty, roasted, burnt, pungent and grassy (Dart and Nursten, 1985). Which compounds are the most abundant pyrazines and how the aroma imparted by pyrazines was described by Dart and Nursten (1985). Furans are numerous in coffee, including aldehydes, ketones, esters, alcohols, ethers, acids, thiols and sulphides. Owing to the wide range of structures, the aroma qualities they impart are also numerous and varied. This class has the aroma of grassy, hay-like (furfural), caramel and burnt sugar (furfuryl ketones) or fragrant caramel, burnt sweet taste, burnt pineapple-like, nutty sweet aroma of almonds, like strawberry preserve and like beef broth (4-hydroxy-2,5- dimethyl-3(2H)-furanone). The same author as above explained which specific compounds of this group are important and what aroma they possess. Ketones vary considerably in aroma. They give aroma of sweet, pungent and fruity (propanone) or buttery (2,3-butanedione) (Dart and Nursten, 1985).

2.3.2 Methods of non-volatile and volatile analysis in coffee

2.3.2.1 Non-volatiles analysis

There are many analytical methods available for the determination of caffeine and trigonelline in coffee. However, high performance liquid chromatography (HPLC) has been a common technique due to its accuracy, precision and high-throughput (De Maria et al., 1995). The analysis of caffeine and trigonelline in coffee by HPLC was reported by a number of authors (Casal et al., 1998; Casal et al., 2000; De Maria et al., 1995; Trugo and Macrae, 1989).

The extraction methods, analytical and detection systems have been thoroughly reviewed by Belay (2011), Jeszka-Skowron et al. (2015); Nuhu (2014).

2.3.2.2 Volatiles analysis

Gas Chromatography (GC) is the most common analytical method used for the analysis of volatiles due to its sensitivity and selectivity to resolve trace compounds (Hinshaw, 2003). Prior to chromatography, appropriate sample preparation and extraction are important. Among different methods of extraction such as static or dynamic headspace extraction, liquid-liquid or solvent extraction, solid phase extraction (SPE) and a solid phase microextraction (SPME), the latter has become the method of choice due to its reliability, selectivity, sensitivity, rapidness, and solvent-free properties (Vas and Vekey, 2004). Then the target analytes can be adsorbed on the fibre by immersing it in the sample or by exposing it to the sample headspace (HS). HS-SPME has also been reported as a promising technique for a variety of fruits (Riu-Aumatell et al., 2004) and other complex horticulture products including wine, tea and chocolate (Ducki et al., 2008; Lv et al., 2014; Siebert et al., 2005). It has been also applied in coffee volatile identification and quantification (Akiyama et al., 2007; Akiyama et al., 2003; Bicchi et al., 2011; Bicchi et al., 1997; Ribeiro et al., 2010).

Following separation on a GC column, volatile compounds can be analysed by different detectors such as flame ionization detectors (FID), photoionisation detector (PID) or mass spectrometry detector (MSD). The MSD combined with GC is the most popular technique used for routine volatile analysis of plant material (Tholl et al., 2006) due to its high separation power, reproducible retention times and sensitive selective mass detection (Koek et al., 2011; Tikunov et al., 2005). Mass spectral identification is based on ionisation of the molecules which are converted into ions in the gas phase before separation based on mass-to-charge ratio (*m/z*) and subsequent detection (Tholl et al., 2006).

The MSD has the added advantage of being operated in either scan mode or with selected ion monitoring (SIM). In full scan, ions are usually collected in the mas scan range of *m/z* 30 – 350 and allows for complete mass spectral data to be collected for each resolved component and comparisons to a mass spectral library can be made for the purpose of peak identification. The MSD operating in full scan can be applied for both targeted and untargeted analysis of components and provides qualitative as well as quantitative data (if standard addition calibrations are developed). SIM is typically applied where target components of interest are present in small amounts and require increased MSD sensitivity. With fewer selected ions being scanned by the MSD, the detector spends more dwell time on fewer ions, thus significantly increasing overall sensitivity. SIM will not allow a full peak profile of the sample but will be intentionally targeted on certain components of interest (Niessen, 2001).

Full scan is one of the most common modes of acquiring LC/MS data which results in the typical total ion current plot (TIC). The full scan is quite useful when identifying unknown compounds in a sample. A full mass spectrum provides more information about the sample composition compared to the single ion monitoring mode. This approach helps highlight certain regions of the chromatogram which is important when interpreted simultaneously with the entire peak profile. TIC has been applied successfully for predicting the composition of biodiesel blends (Pierce et al., 2011), classifying edible oil type (Bagur-González et al., 2015), discriminating counterfeit medicines (Custers et al., 2014) or classifying and differentiating agarwoods (Hung et al., 2014). These studies confirmed this method as a high-throughput fingerprint analysis technique (Pierce et al., 2011), as valuable tools to discriminate between genuine and counterfeit medicines (Bagur-González et al., 2015) or as a simple, convenient and robust method to extract volatile compounds successfully for evaluating agarwoods and sandalwoods (Hung et al., 2014).

This full scan “fingerprinting” method should be applied with caution due to its drawbacks including base line drifts, peak shift between samples, low signal-to-noise ratio and overlapping/co-elution (Amigo et al., 2010; Parastara et al., 2012). Besides, this method is based on the information given in the applied databases which could limit the detection between natural and anthropogenic compounds (Schlabach, 2013). Another limitation of the method is that compounds of low concentration might be masked by compounds of very high concentration eluting with the same retention time and nearby masses or the chromatographic peak deconvolution (Schlabach, 2013).

To quantify volatiles, stable isotope dilution analysis (SIDA) combined with HS-SPME/GC-MS has recently been applied to coffee quality analysis with advantages of high repeatability, sensitivity, automation, speed of analysis and avoiding the drawbacks related to the matrix effect (Bicchi et al., 2011; Pickard et al., 2013).

2.4 Factors affecting chemical compounds in green bean coffee

Coffee quality is influenced by non-genetic factors (e.g. growing conditions, agriculture practices, harvest and post-harvest, storage condition and preparation) and genetic factors (Wintgens, 2012). Understanding the influence of each factor and their interaction would help defining realistic breeding strategies.

2.4.1 Non-genetic factors

According to Clarke and Macarae (1985), the effects of environmental and agricultural factors are less important than genetic variation in controlling the caffeine contents of green beans. Elevation had a positive effect on bean biochemical composition. More caffeine in all stages of fruit development of *C. canephora* in lower elevation and bean caffeine content was reduced 32% when grown at higher elevations (Sridevi and Giridhar, 2014). Fertilising also has an effect on the biochemical compounds in the bean for example an excess of nitrogen increased the caffeine content which results in a more bitter taste in the coffee cup (Mendoza, 1995, originally in Spanish, cited by Wintgens, 2012). However, potassium, phosphate, magnesium and calcium do not have a significant effect on caffeine.

Trigonelline is also affected by environments. Shade delays berry flesh ripening by up to one month which has positive effects on bean size and composition, and beverage quality. However, it was found that higher trigonelline concentrations occurred in beans grown with no shade than those grown with shade. This suggests the incompletion of bean maturation and therefore accounts for the increased bitterness and astringency of the coffee beverage (Muschler, 2001). Altitude has an influence to trigonelline: more trigonelline content of robusta beans from high altitude was observed (Avelino et al., 2005; Sridevi and Giridhar, 2013).

2.4.2 Genetic factors

Extrinsic aspects such as soil composition, climate, agricultural practices and storage conditions affect the bean’s physiology and consequently its chemical composition. However, the basic chemical composition of green coffee affecting cup quality is predominantly controlled by genetic factors (Farah and Donangelo, 2006) with several findings indicating environmental factors playing a minor role. Studies on G x E for coffee cup quality concluded that the best variety in one environment remains the best in others even if the overall cup quality varies (Bertrand et al., 2006; Moschetto et al., 1996)

2.4.2.1 Intra-and inter-species variation of biochemical compounds.

Significant genetic variability in bean chemical composition and organoleptic characteristics exists between species and to a lesser extend within species (Leroy et al., 2006).

2.4.2.1.1. Caffeine

Clarke and Macarae (1985) found that genetic variation was more important than the effects of environmental and agricultural factors in controlling the caffeine contents of green beans. Caffeine contents of green coffee among coffee species varies from 0% (*Coffea* sp. Bakossi, *C. pseudozanguebariae*, *C. humblotiana* and *C. salvatrix*) to 2.64% dry matter basis (dmb) (*C. canephora)* (Campa et al., 2005a) and most of the variation (about 94 %) is under genetic control (Barre et al., 1998; Montagnon et al., 1998). The caffeine content in *C. canephora* is 2.5% dmb while in *C. arabica* is 1.2% dmb (Ky et al., 2001b). In other less important species such as *C. liberica* and *C. arabusta* the caffeine content is reported to be 1.35% dmb and 1.72 % dmb on average, respectively (Clarke and Macarae, 1985). The related *Paracoffea* genus in Africa and Asia has very low caffeine content of about 0.2 % dmb (Clarke and Macarae, 1985). Variation in caffeine among arabica and robusta is summarised in Table 2.1.

Table 2.1 Caffeine variation in *C. arabica* and *C. canephora*

Coffee species	Caffeine content (dmb)		Number of genotypes	Reference
	Lowest	Highest		
C. arabica	0.91	1.32	42	Dessalegn et al. (2008)
	0.96	1.62	38	Ky et al. (2001b)
	0.90	1.40	28	Martín et al. (1998)
	0.62	1.21	9	Mazzafera and Carvalho (1992)
C. canephora	1.51	3.30	38	Ky et al. (2001b)
	1.60	3.20	13	Martín et al. (1998)
	0.81	3.27	Unspecified	summarised by Priolli et al. (2008)

Recently, three naturally low-caffeine arabica coffee trees (0.07%) were identified from 300 accessions collected in Ethiopia (Silvarolla et al., 2004). Unexpectedly large variation in caffeine content among individual green seeds of five arabica cultivars were also observed (Mazzafera and Silvarolla, 2010).

2.4.2.1.2. Trigonelline

Trigonelline content varied from 0.39% (*C. liberica Koto*) to 1.77% dmb (*C. kapakata*) among 14 species of *Coffea* (Campa et al., 2004). However, trigonelline is generally lower in *C. canephora* and is about two-thirds of that found in *C. arabica* (Farah, 2009). Table 2.2 summarizes trigonelline variation reported in various studies.

Table 2.2 Trigonelline variation in *C. arabica* and *C. canephora*

Coffee species	Trigonelline content (dmb)		No of genotypes	Reference
	Lowest	Highest		
C. arabica	0.95	1.29	4	Campa et al. (2004)
	0.88	1.77	38	Ky et al. (2001b)
	1.00	1.92	28	Martín et al. (1998)
	1.52	2.90	16	Mazzafera (1991)
C. canephora	0.52	1.06	4	Campa et al. (2004)
	0.75	1.24	38	Ky et al. (2001b)
	0.91	1.94	13	Martín et al. (1998)

However, Mazzafera (1991) reported a much higher trigonelline content in *C. canephora* (3.08% dm). Trigonelline content is also reported in *C. pseudozanguebariae* (1.02 % of dmb) and *C. liberica* var. *dewevrei* (0.57 % of dmb). Ky et al. (2001) compared beans harvested in different years and environments in interspecific hybrids and found that there was no change in trigonelline content with bean maturity differences and drying conditions.

2.4.2.1.3. Flavours

Significant differences in flavour between different coffee types were also reported. Robusta beans have a bitter, full bodied taste, but low acidity while arabica is more aromatic, with more perceptible acidity, but less body. Within robusta, Moschetto et al. (1996) found important differences in cup quality between the two main genetic groups for preference, aroma, acidity, body and bitterness. Coffee genotypes from Guinea exhibit better cup quality (preference and aroma) than the genotypes from Congo which also show greater variability for this trait. Within arabica, different varieties can be associated with specific flavour profiles, such as SL28 producing a milder brew than Kent, Blue Mountain and Bourbon being finer coffees than Catimor while Liberica being bitter and without finesse (Wintgens, 2012).

2.4.2.2 Genetics of biochemical compounds

Improving both yield and quality is the target for coffee breeding programs. In a study by Montagnon et al. (1998), genetic correlations between coffee yield and quality determinants (e.g., fat content, sucrose, trigonelline, caffeine) and cup tasting components measured in two groups of *C. canephora* were not significant, indicating that the coffee quality is not negatively associated with yield. Understanding genetic inheritance of quality traits can benefit breeding programs but there is little research on genetic control of biochemical compounds determining coffee quality.

2.4.2.2.1. Caffeine

Caffeine content has often been described as an additive trait in intraspecific studies with *C. arabica*, *C. canephora* and interspecific hybrids from several *Coffea* species irrespective of the ploidy level (summarised by Priolli et al., 2008). Le Pierres (1988, cited by Priolli et al., 2008) reported high broad-sense heritability (BSH = 0.76) and intermediate narrow-sense heritability (NSH = 0.33) in *C. canephora* varieties. Montagnon et al. (1998) also found higher NSH heritability (h^2 = 0.80) for caffeine content in a factorial crossing scheme of *C. canephora* and attributed this value to the analytical method used and the genetic origin of the parents. However, Mazzafera and Carvalho (1992) found that the caffeine content in F1 hybrids was additive in some of the crosses. Barre et al. (1998) studied caffeine inheritance in interspecific hybrids from a species with caffeine (*Coffea liberica*) and one without caffeine (*C. pseudozanguebariae*) in the seeds and all F1 hybrids had caffeine content (0.2% dmb). They concluded that one major gene with two alleles may be involved in the control of caffeine biosynthesis and the absence of caffeine was apparently controlled by one recessive gene (Barre et al., 1998). Using other backcross hybrids derived from an interspecific cross between *Coffea pseudozanguebariae* and *Coffea canephora* to analyse the inheritance of caffeine, Akaffou et al. (2012) also found that the absence or presence of caffeine was controlled by one major gene. *Caf2* is the allele responsible for the presence of caffeine and it dominated over the absent one (*caf1*). Priolli et al. (2008) investigated caffeine inheritance in F2 and backcross F1 generations between *Coffea arabica* var. Bourbon Vermelho (BV) and *Coffea canephora* var. Robusta 4x (R4x). Conversely, their genetic analysis showed that caffeine content in seeds was controlled by multi-genes with additive effects. The broad-sense heritability of caffeine content in seeds estimated for both generations was high. The results indicate that there are distinct mechanisms controlling the caffeine content in seeds and leaves, the gene exchange between the *C. arabica* BV and *C. canephora* R4x genomes and favourable conditions for the improvement of caffeine content in this coffee population (Priolli et al., 2008). The authors concluded that both major and minor genes with additive gene effects appear to be involved in the variation of caffeine content in the seeds of the segregating populations studied (Priolli et al., 2008). Within *C. canephora*, Leroy et al. (2011) also found female additive effects were quite significant for caffeine.

2.4.2.2.2. Trigonelline

Nuclear intraspecific effects on trigonelline content were noted in *C. canephora* crosses, with narrow sense heritability estimated as 0.38 (Montagnon et al., 1998). However, the trait heritability measured in F1 between *C. liberica* spp. dewevrei (anciently cultivated cultivar) and *C. pseudozanguebariae* (wild species) and in backcross hybrids were high (h^2 = 0.71) (Ky et al., 2001). This study also found that trigonelline content was attributable to maternal inheritance in which a QTL is located on the G linkage group (Ky et al., 2001). According to this author, the intermediate heritability in the previous study was due to an intraspecific nucleocytoplasmic interaction and epistatic effects. In one study QTLs identified for trigonelline content explained the male additive and dominant effects on this compound in *C. canephora* (Leroy et al., 2011), but in other study, female additive was found (Mérot-L'Anthoëne et al., 2014).

2.5 Coffee genomic resources

2.5.1 Coffee genetic diversity in *C. arabica*

Arabica cultivars have been known for low genetic diversity and the reasons are probably due to the narrow geographical origin, the selection bottleneck and self-pollination (summarised by Tran, 2005). A number of works on the assessment of arabica diversity have been done with differing results. In general, among three main types of material (cultivar/varieties; accessions/introgression/hybrids; and spontaneous/sub-spontaneous) almost all studies show a very low genetic variation amongst arabica cultivars using different marker systems (RAPD, AFLP, SSR, ISSR, SRAP, TRAP) and coffee collected from different regions (Brazil, Yemen, Indian, Australia and Vietnam, Tanzania, Hawaii) (Al-Murish et al., 2013; Maluf et al., 2005; Masumbuko et al., 2003; Mishra et al., 2012b; Motta et al., 2014; Steiger et al., 2002; Tran, 2005; Vieira et al., 2010). The second group (accessions/introgression/hybrids) shows low to significant variation, more diverse in the introgressed lines (Dessalegn et al., 2009; Dessalegn et al., 2008b; Geleta et al., 2012; Lashermes et al., 2000a; Teressa et al., 2010; Tessema et al., 2011). The last group, in which one may expect to see the most diversity, has diversity ranging from low (Aerts et al., 2012; Anthony et al., 2001; Dessalegn et al., 2008b) to moderate (Aga et al., 2003) to high (Silvestrini et al., 2007). The correlation of genetics with geography/origin clearly shows in some studies (Aga et al., 2003; López-Gartner et al., 2009; Silvestrini et al., 2007) but not in the others (Anthony et al., 2001; Dessalegn et al., 2008b; Geleta et al., 2012). In the most recent work presented in the World Coffee Research annual report (WCR, 2014), genetic diversity assessment of 800 arabica coffee accessions from the arabica collection at CATIE, Costa Rica shows the least genetic diversity of this species compared to other major crops. This study also found that cultivated coffee varieties contain approximately 45% of the genetic diversity found in the 800 aforementioned accessions (while only 2% of tomato diversity found in cultivated tomato) indicating the limitation of variability for breeding programs. This also is problematic for association mapping work in arabica for breeding improvement, and thus the selection of appropriate germplasm for association mapping is very critical. Therefore it is essential to understand the population structure and its genetic diversity. In addition, the understanding of relationship among coffee species, especially those that are close to arabica, could also help guide efficient hybridisation schemes in coffee breeding programs at genus level.

2.5.2 Coffee whole genome and synteny

The genome size of coffee was estimated to be 1.3 Gb for arabica (Kochko et al., 2010) and 710Mb for robusta (Denoeud et al., 2014). Using flow cytometry, the DNA content of arabica was estimated as 2.47 pg (equivalent to 1.2 Gb) and canephora as 1.43 pg (equivalent to 700 Mb), while other diploid coffee species varied from 0.96 pg (469 Mb) (*C. mauritiana*) to 1.84 pg (900 Mb) (*C. humilis*) (Noirot et al., 2003; Razafinarivo et al., 2012). Several studies also indicated that species native to dry areas (mostly in East Africa) have a smaller genome size (<1.3 pg) than those native to evergreen forest (up to 1.76 pg) (summarised by Kochko et al., 2010). In breeding, coffee species with genomes differing more than 0.25 pg are unlikely to cross and in fact marked sterility of hybrids was reported (Kochko et al., 2010). Despite its economic importance, the first high-quality draft genome of robusta coffee has just been completed and reported in 2014 (Denoeud et al., 2014). To generate a whole genome sequence of *C. canephora*, Roche 454 single and mate-pair reads and Sanger BAC-end reads were used, representing 29.5X coverage of the genome size (710 Mb). In addition, Illumina sequence data, corresponding to 60X of coverage of the coffee genome (710 Mb), was also used to correct sequencing errors and fill gaps. The resulting genome assembly consists of 25,216 contigs and 13,345 scaffolds with a total length of 568.6 Mb (Table 2.3). Using a high-density genetic map, a total of 349 scaffolds covering approximately 364 Mb were anchored to the 11 *C. canephora* chromosomes, among which 139 were both anchored and oriented. Coffee displays the most conservative chromosomal gene order among asterid angiosperms. Transposable elements (TEs) in the *C. canephora* genome were identified and classified accounting for almost half (49.2%) of the reference genome length. Organelle to nuclear genome transfers were also analysed with a typical amount of chloroplast-derived fragments (2,014 insertions, accounting for 0.16% of the draft genome) and an unusually large 750kb mitochondrion-derived fragment more than 100 kb longer than those known from other plant genomes. In total, 25,574 protein-coding gene models were obtained in which several gene functions involved in secondary compound biosynthesis and coffee flavour and aroma were characterised. Species-specific gene family expansions were examined such as N-methyltransferases (NMTs), defence-related genes, and alkaloid and flavonoid enzymes involved in secondary compound synthesis. This is the most valuable genomics resource for further genomics study in coffee, especially as all of the genomics data, transciptomic data, SNP polymorphisms and genotyping data, gene families and metabolic pathways are publicity available via the Coffee Genome Hub (Dereeper et al., 2014).

The whole chloroplast genome sequence of *Coffea arabica* L. was first reported by Samson et al. (2007) (Table 2.3) and this also provides an important reference for studying other species and helps build up the phylogeny among *Coffeeae* tribe which is still questionable. Several research groups are working on an *C. arabica* genome sequence using different biological materials and sequencing platforms (Deynze, 2017; Kochko et al., 2015; Morgante et al., 2015; Mueller et al., 2014; Yepes et al., 2016). This genomic resource is extremely important for the study of genetics and genomics of the high quality coffee species, *C. arabica*.

Table 2.3 Genome sequence data available for the two main domesticated coffee species

<table>
<tr><td colspan="7"><b>Whole genome sequence of <i>C. canephora</i></b></td></tr>
<tr><td>Genome size (Mb)</td><td>Coverage</td><td>No of contigs</td><td>No of scaffolds</td><td>No of genes</td><td>MicroRNA precursors</td><td>Genome transfers[a]</td></tr>
<tr><td>Est. 710</td><td>30 x</td><td>25,216</td><td>13,345</td><td>25,574</td><td>95</td><td>2,573</td></tr>
<tr><td colspan="7"><b>Chloroplast genome sequence of <i>C. arabica</i></b></td></tr>
<tr><td rowspan="2">Genome size (bp)</td><td rowspan="2">IRs</td><td rowspan="2">No of genes</td><td colspan="4">Coding region</td></tr>
<tr><td>protein genes</td><td>tRNA genes</td><td>rRNA genes</td><td>genes containing introns</td></tr>
<tr><td>155,189</td><td>25,943</td><td>130</td><td>79</td><td>29</td><td>4</td><td>18</td></tr>
</table>

(a) Organellar-to-nuclear genome transfers

Several studies (Guyot et al., 2012; Lefebvre-Pautigny et al., 2010; Mahe et al., 2007) have compared genomes between coffee and other crops. The most recent study reported that coffee chromosomal regions showed unique one-to-one correspondences with grape chromosomes, proving the lack of any events changing ploidy intervening in their separate histories. Coffee and grape genomes show one-to-three correspondence with the tomato genome (Denoeud et al., 2014).

Among coffee species, comparison of genomes showed a high gene synteny between *C. arabica* (EaEaCaCa) and *C. canephora* (CC) and among *C. arabica* (EaEaCaCa), *C. eugenioides* (EE) and *C. liberica* (LL) (Cenci et al., 2012; Herrera et al., 2012; Yu et al., 2011), but numerous chromosomal rearrangements were detected (Yu et al., 2011). The two sub-genomes of *C. arabica* (Ca and Ea) indicated sufficient sequence differences, and thus obtaining a whole-genome sequence would be an important resource for the allotetraploid genome of *C. arabica*. Results also revealed that *C. eugenioides* was more closely related to *C. arabica* than to *C. liberica*, which was in agreement with the ancestral history of the allotetraploid *C. arabica* (Herrera et al., 2012).

2.5.3 Other genomic resources

2.5.3.1 Molecular markers

As for other crops, determination of molecular predictors for coffee quality traits would help reduce the length of breeding selection cycles and thus phenotypic evaluation cost. However, the use of DNA technology in coffee quality improvement is still in its infancy and therefore limited. Pot *et al.* (2007) used polymorphisms generated from SNPs, INDELs (insertions and deletions) and SSRs (simple sequence repeats) to identify the nucleotide diversity of four sucrose metabolism enzymes in *C. canephora* genotypes using direct sequencing. The variation of these genes was also analysed between different *Coffea* species to allow the identification of more polymorphic sites using parallel *in silico* analysis of expressed sequence tag (EST) resources (Pot et al., 2007). AFLP (amplified fragment length polymorphism) and SSR markers were used to construct a genetic map of an F2 population between *C. arabica* and *C. canephora* (artificial tetraploid). The number of markers associated with quality traits identified was nineteen for sugar content, eight for caffeine, eight for chlorogenic acids (CGAs) and one for caffeine and CGAs (Priolli et al., 2009). These markers need to be validated in other genotypes for consistency before they can be used in marker assisted selection (MAS) in coffee breeding. Recently, a total of 33,239 SNPs specific to *C. arabica* and 87,271 SNPs specific to *C. canephora* were developed using targeted genome capture strategies and next-generation sequencing, and were evaluated on 72 samples from *C. canephora* and 72 from *C. arabica*. These genomic resources will provide support for genome assemblies, accelerate breeding of interested traits as well as manage genetic diversity in coffee species (Resende et al., 2016).

2.5.3.2 Bacterial artificial chromosome (BAC) library

Three BAC libraries were developed for each species of *C. canephora* and *C. arabica* in relation to sucrose metabolism (Leroy et al., 2005), genome composition and evolution (Dereeper et al., 2013), nematode and leaf rust resistant (Noir et al., 2004), bean size and cup quality (Jones et al., 2006) and mannose-6-phosphate reductase gene (CaM6PR) (Cacao et al., 2013) (Table 2.4). These available BAC libraries of *Coffea* are an important resource to support genomic studies such as comparative genomics and the identification of genes and regulatory elements controlling important traits in coffee.

Table 2.4 BAC libraries of two main coffee species

Genotypes	No of clones	Ave. size of clones (kb)	Coverage	Authors
C. canephora				
IF 126	55,296	135	10.6 x	Leroy et al., 2005
IF 200 DH	36,864	166	8.6 x	Dereeper et al., 2013
IF 200 DH	36,864	121	6.3 x	Dereeper et al., 2013
C. arabica				
IAPAR 59	80,813	130	8 x	Noir et al., 2004
Hybrid (Mokka x Catimor)	52,416	94	4 x	Jones et al., 2006
Timor Hybrid CIFC 832/2	56,832	118	5-6 x	Cacao et al., 2013

2.5.3.3 Expressed sequence tags (ESTs)

The development of large EST sequencing projects has supported the identification of potential genes contributing to quality traits and their interplay with other genes through essential biochemical pathways. ESTs facilitate the development of whole transcriptome analysis and the identification of the biosynthetic pathways associated with the expression of quality and the genes within these pathways (Leroy et al., 2006). The EST resources available in coffee have been reviewed by Kochko *et al.* (2010) and are summarised in Table 2.5.

Table 2.5 Expressed sequence tags (ESTs) from the three main coffee species

Project/ organisation	**Tissues**	**ESTs of species**			Unigenes/ genes
		C. arabica	*C. canephora*	*C.racemosa*	
Fernandez et al. (2004)	leaves	527	-	-	
Nestle and Cornell Uni Lin et al. (2005)	seeds	-	47,000	-	13,175
Brazilian Genome Vieira et al. (2006)	leaves and fruits	130,792	12,381	10,566	33,000
CENICAFE Montoya et al. (2006)	leaf, flower and fruit	32,961	-	-	10,799
De Nardi et al. (2006)	leaves and roots	1,587	-	-	-
French IRD Poncet et al. (2006)	leaves and fruits	-	10,420		5,534
Salmona et al. (2008)	seeds	266	-	-	-
Total (246,500)		166,133	69,801	10,566	62,508

A total of 246,500 ESTs have been reported with 69,801 for *C. canephora*, 166,133 for *C. arabica* and 10,566 for *C. racemosa* using different organs of the plant (leaves, flower, fruits and roots) at different stages of development and maturation (De Nardi et al., 2006; Fernandez et al., 2004; Lin et al., 2005; Montoya et al., 2006; Poncet et al., 2006; Salmona et al., 2008; Vieira et al., 2006). However, not all these resources are published or available, which limits their use. A number of studies have used the available resource of ESTs for further research such as microsatellite marker discovery (Aggarwal et al., 2007), analysis of transcriptome divergence and analysis of genes involved in the biosynthesis pathways of lipids and main storage proteins (Privat et al., 2011), gene structure prediction and functional annotation with the detection of a total of 345 pathways in *C. arabica* and 300 pathways in *C. canephora* (Mondego et al., 2011). This genomic resource will be very important for further research on coffee quality improvement.

2.5.3.4 Genetic maps and quantitative trait loci

The construction of genetic maps is often the first step for molecular dissection of complex traits (yield components, quality and disease). Genetic maps enable the mapping of quantitative trait loci (QTLs) controlling these traits and thus provide a framework for isolation of candidate genes followed by manipulation of genes for yield increase and quality improvement (Vega et al., 2008). In many cases, markers linked to QTLs are directly used in MAS without the need of understanding gene functionality underlying the QTLs.

Linkage mapping in coffee in general requires more effort and cost than in annual crops due to its long life cycle, low polymorphism (in the case of *C. arabica* – with the exception of Ethiopian cultivars and germplasm), and the absence of a large collection of DNA markers as well as genomic resources (Vega et al., 2008). The first genetic maps were constructed in the diploid *C. canephora* by Paillard et al. (1996) followed by several others constructed for different coffee populations as described in Table 2.6. For *C. arabica*, linkage analysis is hindered by low polymorphism caused by the extremely narrow genetic base and its reproductive nature of self-pollination. The first *C. arabica* linkage map has been constructed by Pearl et al. (2004). In this study, a pseudo-F2 population of two significant differences in beverage quality of two parental cultivars (Tall Mokka and Catimor) was used for mapping QTLs controlling quality traits in coffee (Pearl et al., 2004). Nagai et al. (2007) used the F2 population of the same parental cultivars and identified more linkage groups. Later, an interspecific F2 population of *C. arabica* and *C. canephora* was used for genetic map construction. Forty-six markers associated with quality (sugar content, caffeine, CGA) and yield were found. Genetic maps from other species were also constructed (Table 2.6). The genetic map from backcross of F1 (*C. pseudozanguebariae* x *C. liberica* var. *dewevrei* (DEW)) and DEW allowed interspecific polymorphisms to detect a gene encoding caffeoyl-coenzyme A 3-O-methyltransferase (*CcCoAOMT1*) relating to CGAs pathway in *C. canephora* (Campa et al., 2003) and a gene encoding phenylalanine ammonia-lyase (*CcPAL1*) relating to phenylpropanoid pathway in *C. canephora* (Mahesh et al., 2006) to be located.

Table 2.6 Genetic linkage maps of coffee

Markers	Population used	Length (cM)	No of linkage groups	Traits/ purpose	Authors
C. arabica					
AFLP	pseudo-F2 population (Mokka hybrid x Catimor)	1,802.8	31	source-sink traits	Pearl et al. (2004)
AFLP	F2 of Tall Mokka and Catimor	1,042.4	40	cupping quality and morphology	Nagai et al. (2007)
AFLP and SSR	F2 of *C. arabica* x *C. canephora*	1,011	37	quality and productivity	Priolli et al. (2009)
RAPD	F2 of (Mundo Nov x Hybrido de Timor) x Hybrido de Timor	540.6	8	Partial linkage map	Teixeira-Cabral et al. (2004)
SSR	F2 and F3 of Caturra x CCC1046	3800	22	yield, plant height and fruit size	Moncada et al. (2014)
C. canephora					
RFLP and RAPD	doubled haploid (IF200)	1,402	15	QTL analysis	Paillard et al. (1996)
RFLP and SSR	doubled haploid and TC (IF200 x DH)	1,041	11	segregation distortion	Lashermes et al. (2001)
RFLP and SSR	(*C. heterocalyx* x *C. canephora*) x *C. canephora*	1,360	15	QTL analysis	Coulibaly et al. (2003)
RFLP and SSR	BP409 x Q121	1,258	11	yield and vigor	Crouzillat et al. (2005)
AFLP and ISSR	*(C.canephora* x *C. liberica)* x *C. liberica*	1,502	16	morphological traits	Amidou et al. (2007)
Conserved Ortholog Set (COS)	Intraspecific hybrid of *C. canephora*	1,331	11	synteny maps between coffee and tomato	(Lefebvre-Pautigny et al., 2010)
SSR	pseudo-backcross of *C. canephora*	1,290	11	yield and quality	Leroy et al. (2011)
SSR	2 populations of intraspecific hybrid of *C. canephora*	1,201	11	yield and quality	(Mérot-L'Anthoëne et al., 2014)
Other species					
AFLP and RFLP	(*C.pseudozanguebariae* x *C. liberica* var. *dewevrei* (DEW)) x DEW	1,144	14	biochemical traits (caffeine, CGAs, sucrose)	Ky et al. (2000)
SSR	*C. liberica* x *C. eugenioides*	798.68	11	QTL analysis	Gartner et al. (2013)

To date, QTL analyses relating to quality compounds and cup have been largely performed on *C. canephora* and other species, but none has been reported for *C. arabica* as indicated in Table 2.7. One QTL contributing to trigonelline accumulation in green beans was mapped on the linkage group G of the cross between *Coffea pseudozanguebariae* and *C. liberica* var. dewevrei (Ky et al., 2001). Campa et al. (2003) identified a QTL for CGA content on the linkage group A of the same cross. The study of caffeine variation in caffeinated seeds was further done by Ky et al. (2013)

using the same cross. Two QTLs *RCQ1* and *CQA1* which are responsible for the caffeine variation in caffeinated seeds and allowed the explanation up to 97 % of the caffeine content variance were mapped on two different linkage groups (A and G, respectively), indicating their genetic independence. *RCQ1* explained the variation in the caffeine/CGA ratio while *CQA1* explained the part of the caffeine content depending on the CGA content. The findings also confirmed that only a part of the CGAs were trapped by caffeine, as in wild species. For each hybrid, the caffeine detected in their seeds was the result of both *RCQ1* and *CQA1* (Ky et al., 2013). Another 27 QTL regions for several biochemical traits including trigonelline, 3-CQA and 4-CQA, 5-CQA, 5-FQA, 3,4di-CQA, 3,5di-CQA, caffeine, and sucrose in *C. canephora* were also reported (Leroy et al., 2011). Most recently, Mérot-L'Anthoëne et al. (2014) identified seven QTLs for sucrose on LGs of A, E, I, and J; eleven QTLs for caffeine in which one is the same as previous study on LGs of A and E; three QTLs for trigonelline on LGs of F, G and K; seven QTLs for lipids on LGs of B, C, E, G and I; eight QTLs for CGAs in which four are the same as previous study on LGs of A and E. Regarding coffee cup quality, another six QTLs were identified for organoleptic traits in the same study. One QTL was identified for the global note in LG H. For bitterness, QTLs were identified in LG D and I while a QTL for acidity was also located in LG I. This result is in agreement with the negative correlations of phenotypes observed between these traits (Leroy et al., 2011). Another eight QTLs for cup quality on A, B, E, G, H and I were also recently identified in *C. canephora* (Mérot-L'Anthoëne et al., 2014).

Table 2.7 Quantitative Trait Loci and Linkage Groups in coffee

Traits	**No of QTLs and LGs**	
	C. canephora	**Other species** (*)
Sucrose	9 QTLs/A, E, I, J	
Caffeine	14 QTLs/A, C, E, I, K	2 QTLs/A, G
Trigonelline	5 QTLs/F, G, I, K	1 QTL/G
Lipids	7 QTLs/B, C, E, G, I	
CGAs	23 QTLs/A, B, D, E, F, I, J, K	1 QTL/A
Cup quality	14 QTLs/A, B, D, E, G, H, I	

(*) hybrid between *Coffea pseudozanguebariae* and *C. liberica* var. dewevrei.

Co-location of some QTLs and genes involved in different metabolic pathways related to coffee quality were noted (Leroy et al., 2011). Using such candidate genes in association mapping may determine the allelic variation of coffee quality (Hendre and Aggarwal, 2007). The co-localisation between QTLs relating to organoleptic traits and genes associating with caffeine or CGA biosynthesis may explain the fact that both CGA and caffeine are involved in conferring bitterness in the coffee cup (Leroy et al., 2011).

2.5.3.5 Gene identification.

The identification of genes relating to quality is one of the main objectives of several coffee research groups around the world. Thanks to their concerted efforts on coffee genomics, a number of coffee candidate genes have been identified and some of them have been cloned and characterised. These results are useful to the coffee genetics community, especially those on genes encoding the enzymes of key metabolic processes. These are candidate genes which may control the variability of coffee quality (Leroy et al., 2006). Genes regulating the main chemical components that are thought to be involved in the flavour and sensory quality of coffee are listed in Table 2.8 (Campa et al., 2003; Denoeud et al., 2014; Geromel et al., 2006; Joët et al., 2009; Kato and Mizuno, 2004; Koshiro et al., 2006; Lepelley et al., 2012; Leroy et al., 2011; Leroy et al., 2005; Mahesh et al., 2006; Mizuno et al., 2014; Mizuno et al., 2003b; Ogawa et al., 2001; Ogita et al., 2004; Perrois et al., 2014; Privat et al., 2011; Privat et al., 2008; Simkin et al., 2006; Uefuji et al., 2003). Recently, 36,935 unigenes from leaves and fruits of *C. eugenioides,* one of *C. arabica*'s ancestors, were identified by Yuyama et al. (2014). A sub-set of these genes related to sugar metabolism and fruit ripening were examined for their expression. This will be a valuable resource for molecular and genetics studies of commercial coffee species. Currently there are works on transcriptomics to understand the transcriptional regulations during seed development (Joët et al., 2014b; Joët et al., 2009) or more specifically for caffeine accumulation (Privat et al., 2014) and chlorogenic acids (Joët et al., 2010b) or galactomannan (Joët et al., 2014a). These could facilitate the development of robust genomics/metabolomics fingerprints of coffee bean quality and ultimately be useful in breeding programs for coffee quality.

Table 2.8 Genes controlling biochemical compounds determining quality traits in coffee

Species	Sucrose	Caffeine	Trigonelline	Fatty acids	CGAs
C. arabica	*CaSUS1* & *CaSUS2*[(1)], *CaInv3*[(2)]	*CmXRS1, CTS1-2*[(1)], *CCS1*[(2)], *CaXMT1-2*[(3)], *CaMXMT1-2*[(4)], and *CaDXMT1-2*[(3)]	*CTgS1* & *CTgS2*[(1)]	*OLE2, DGAT*[(1)]	*4CL8, HQT, F5H1* & *POD*[(1)]
C. canephora	CcSUS1, *CcSUS2*[(3)] & 8 new genes[(2)]	*CcXMT1, CcMXMT1, CcDXMT1*[(5)] & 23 genes relating to alkaloid catalytic[(6)]	N/A	*CcOLE1-5, CcSTO1*[(2)] & 15 other genes[(3)]	Many genes involved in phenylpropa-noid pathway[(2)]

(***Sucrose***: (1) Geromel et al. 2006; (2) Privat et al. 2008; (3) Leroy et al., 2005; ***Caffeine***: (1) Koshiro et al., 2006; (2) Mizuno et al., 2003; (3) Ogita et al., 2004; Uefuji et al., 2003 (4) Ogawa et al., 2001; Ogita et al., 2004; Uefuji et al., 2003; (5) Kato and Mizuno, 2004; Perrois et al. 2014; (6) Denoeud et al. 2014; ***Trigonelline***: (1) Mizuno et al., 2014;

Fatty acid: (1) Privat et al., 2011; (2) Simkin et al., 2006; (3) Denoeud et al. 2014; ***CGAs***: (1) Joët et al. 2012; (2) Mahesh et al. 2006; Lepelley et al. 2012; Denoeud et al. 2014; Campa et al., 2003; Lepelley et al., 2007).

Although several genes encoding the biosynthesis of biochemical compounds in coffee have been identified in *C. arabica* and *C. canephora*, there are no genes identified for trigonelline synthesis and no studies on allelic variation (e.g., SNPs) associated with low and high levels of the key biochemical compounds which can be utilised in MAS. However, sequences from the known genes can serve as useful references in re-sequencing to detect polymorphisms for genetic mapping of candidate genes contributing to genetic variation in biochemical compounds in *C. arabica*.

2.6 Genetically Modified (GM) coffee relating to quality

Several studies have been implemented with β-glucuronidase gene transformation using different approaches, which resulted in transient expression in *C. arabica* (summarised by Zamarripa and Petiard, 2012) or stable expression in transgenic plants of *C. arabica*, *C. canephora* and *C. arabusta* (Spiral and Petiard 1993, originally in French, cited by Zamarripa and Petiard, 2012). Ogita et al. (2003) and Ogita et al. (2004) obtained transgenic coffee plants in which caffeine synthesis was suppressed using RNA interference (RNAi) technology to inhibit a theobromine synthase gene (*CaMXMT1*). Low-caffeine plantlets of *C. canephora* were produced showing 70% reduction of both theobromine and caffeine in the leaves compared to the control plants. Transcript down-regulation was observed for the *CaMXMT1* and *CaMXT1* genes involved in the methylation steps of caffeine biosynthesis. Theobromine and caffeine reduction in transformed *C. arabica* embryogenic tissue was also analysed. The transgenic tissue showed a significant reduction in theobromine (65-85%) and caffeine (65 to 100%) compared to the nulls. Sequences and functions of the caffeine genes have thus been elucidated. However, since coffee is a beverage that is consumed by humans and GM is still under public debate, it is wise to search and utilise natural variation of those genes in coffee breeding. The GM gene sequences can serve as a valuable reference to detect gene polymorphisms in natural populations, from which favourable alleles conferring high caffeine content can be introgressed in modern cultivars via crossing.

2.7 Whole genome sequencing and association studies

A number of previous studies involved genome-wide association studies (GWAS) using next generation sequencing (NGS) to detect candidate genes/QTLs of complex traits such as physical wood properties, resistant genes, drought tolerance, cold hardiness and timing of bud set (summarised by Sexton et al., 2012). Such approaches have been successfully applied in various crops including annuals (rice, maize, soybean, wheat), perennial crops and forestry trees (pine, cottonwood) (summarised by Hall et al., 2010).

Application of NGS in plant genomics studies has been extensively reviewed by a number of authors (Henry, 2011; Henry et al., 2012; Kumar et al., 2012; Varshney et al., 2014). NGS lowers the cost of sequencing and generates a large amount of data in a single sequencing run, making it feasible to study genetics at the whole genome level. The steps involved in the NGS strategies (shearing of DNA, ligation, library preparation) as well as different sequencing platforms (Illumina, Life Technologies, PacBio) have been described and their advantages and disadvantages have been reviewed by several authors (Edwards, 2013; Gao et al., 2012). NGS can be used via an approach of whole genome sequencing or targeted sequencing.

Two methods of whole genome sequencing can be used including (1) *de novo* sequencing - applied for species that do not have pre-existing sequence data, followed by the use of bioinformatics tools to assemble the sequences and obtain the genomic map for that species; and (2) re-sequencing performed on individuals of a species that has already known genome sequence (Gao et al., 2012).

Targeted sequencing includes two approaches, amplicon sequencing and target enrichment. Amplicon sequencing will sequence amplified regions (PCR amplicons) of small and selected regions of the genome with lengths of hundreds of base pairs and with a high depth of coverage to identify common and rare sequence variations. Sequencing PCR amplicons of sets of candidate genes from DNA bulks will help to identify the available variation in these genes exploited in a population (Henry et al., 2012). Target enrichment is quite similar to amplicon sequencing in terms of using only selected regions or genes enriched in the library from genomic DNA. However, target enrichment allows for larger DNA insert sizes and enables a greater amount of total DNA to be sequenced per sample. For this method, regions of interest in the genome can be targeted, making it an ideal approach for examining specific gene pathways, or as a follow-up analysis to GWAS (Illumina, 2013).

When making decisions on the approaches of candidate gene or of whole-genome, one of the most critical factors is the extent of linkage disequilibrium (LD) in the organism of interest. In genome-wide studies, the extent of LD determines the mapping resolution that is achieved and the numbers of markers covered (Whitt and Buckler IV, 2003). Genetic, biochemical, physiological and phenotypic information must be used to support the selection of candidate genes (Neale and Savolainen, 2004). Where a developmental or biochemical pathway is well-understood, candidate gene selection is straightforward; but the researcher is typically limited to the field of known genes, which presents the risk of overlooking causal gene mutations not previously identified (Hall et al., 2010)

Association studies differ from quantitative trait locus mapping by utilising samples from genetically diverse populations that contain short stretches of linkage disequilibrium due to generations of recombination (Sexton et al., 2012). Such studies thus enable relationships between phenotypic traits and underlying DNA sequence variation to be understood. Association studies may be suitable for perennial species such as coffee because they can be carried out on pre-existing populations, in collections or in selection trials. This approach does not require specific populations created by controlled crossing (like conventional genetic mapping approaches) which is very difficult for perennial crops (Neale and Savolainen, 2004). Association or linkage disequilibrium mapping has become a very popular method for dissecting the genetic basis of complex traits in plants (Hall et al., 2010). For the study on volatile and non-volatile compounds in coffee and that are likely to be playing a role in coffee flavour, once the data on biochemical compounds and SNPs variation are available, association genomics will be applied to identify key genes associating coffee quality. The integration of NGS technologies and association study to analyse complex traits will be a powerful strategy complementing the traditional method of parent-hybrid map construction (Gao et al., 2012).

XP-GWAS is a modification of GWAS in which only individuals with extreme phenotypes from natural populations are bulked for sequencing (Yang et al., 2015). This approach helps reduce the cost of genotyping for every single individual in the population required in GWAS while ensuring the detection of trait-associated variants (TAVs) with high mapping resolution (Yang et al., 2015).

2.8 Conclusions and implications for the study of genetics of coffee quality.

Research on genetics and genomics for coffee, especially in relation to quality is relatively limited compared to other crops and thus belies its potential and economic contribution. The reasons could possibly be due to the lack of funding – most coffee growing regions being developing countries – the complexity of the quality traits and the limitations of the technology used. More studies have been conducted for robusta coffee than arabica coffee due to its lower ploidy level and greater genetic diversity. Previous studies showed considerable genetic variation in compounds defining coffee quality among coffee species, especially in diploids. In addition, a number of genes relating to the synthesis of sucrose, caffeine, several sub-groups of CGAs and lipids have also been identified. These studies provided gene sequences via ESTs that will facilitate further studies on bean composition relating to coffee quality. However, there has been no research focusing on allelic variation of these genes in natural populations except bi-parental mapping from intra or interspecific hybrids (Nagai et al., 2007; Pearl et al., 2004; Priolli et al., 2009).

NGS technology has been used to assemble the whole genome sequence of *C. canephora* (Denoeud et al., 2014) and to examine genetic change in allopolyploidisation of arabica (Lashermes et al., 2014). Based on knowledge of biochemical compounds (and their reference gene sequences) that are thought to be involved in the flavour and sensory quality, it should be possible to apply NGS and

GWAS combined with bi-parental QTL mapping to detect valuable haplotypes from natural populations for use in breeding coffee varieties with better quality.

To obtain the genomic sequence of *C. arabica,* as only the draft genome of *C. canephora* is available, a *de novo* approach can be applied. The other approach for obtaining the genomic sequence of *C. arabica* is whole genome re-sequencing as the available draft genomic sequence of *C. canephora* can be used as reference since it is one of progenitors of *C. arabica*. The whole genome sequence will be a key resource of data to identify gene sequences and SNP markers for most genes including novel and published genes. EST and BAC sequences can be used as reference sequences to assemble sequence reads from NGS data to identify genes/alleles and their positions on the chromosome. In addition, whole-genome re-sequencing and SNP genotyping data can be used to generate a dense SNP genetic map of the population for QTL mapping.

Amplicon sequencing in a larger set of samples could be an appropriate approach to genetic association study of quality traits, since the biosynthesis pathways, QTLs and candidate genes of sucrose, caffeine and several subgroups of CGAs are known (summarised in Section 2.5.3.4). This will help to confirm or validate the candidate genes from the literature and identify the alleles that are useful for breeding. For those compounds for which candidate genes have never been identified or very few identified, for example trigonelline and some compounds belonging to the lipid group, the approach could be to use whole genome sequencing via NGS to detect SNPs, then using an association genomics approach to detect the link between SNPs and the traits of interest.

While whole genome technology will provide a powerful resource, a good start has been made with AFLP and SSR markers associated with quality traits identified for sugar content, caffeine and CGAs in *C. arabica*. These are already helpful in studies aimed at genetic improvement of coffee quality, but importantly, they will provide useful reference points as the genomic approach is developed. Existing coffee BAC libraries (summarised in Section 2.5.3.2) are beneficial in comparative genomics, the identification of genes and regulatory elements controlling important traits in coffee, and will play an important role in validating new *C. arabica* genomic data as it emerges. With the current knowledge of biochemical compounds that govern coffee bean composition that relates to beverage quality and available genetic resources, it is possible to apply an association genomics approach using both whole genome sequencing and targeted sequencing to detect the link between SNPs and the traits of quality determinants from *C. arabica* natural populations for use in quality improvement via intraspecific breeding programs. However, caution should be taken when using this approach for *C. arabica* since this species has a narrow genetic base, lacks a reference genome, is a polyploid (4n) and lacks populations from controlled crosses for analysis of specific traits and the validation of markers or genes once detected. The aforementioned difficulties can be overcome by the use of wild arabica accessions from Ethiopia, development of a reference genome sequence for *C. arabica*, and the complementary use of the two approaches of targeted re-sequencing and whole genome re-sequencing in variant discovery respectively. To overcome the narrow genetic base in *C. arabica*, another approach for interspecific breeding could be an alternative by understanding the relationship among coffee species that possess good quality traits and are close to arabica. There is no doubt that the availability of whole genome sequences will provide the greatest stimulus yet to the understanding of the genetic basis of coffee quality.

CHAPTER 3: PHYLOGENETIC POSITION OF ARABICA COFFEE BASED ON ANALYSIS OF CHLOROPLAST GENOME SEQUENCES [2]

Abstract

Coffee belongs to the Rubiaceae family and the *Coffeeae* tribe. The taxonomy and classification of coffee has been controversial. The understanding of relationships among coffee species has implications for the improvement of coffee. Many previous studies divided *Coffeeae* tribe into two genera *Coffea* L. and *Psilanthus* Hook.f. Whole chloroplast genome sequences of 16 samples representing 15 species have been assembled using Illumina platform in order to investigate the relationships among coffee species in the C*offeeae* tribe. Coffee species were collected from eleven countries representing different regions of origin. Two approaches of reference-guided mapping assembly and *de novo* assembly were performed on CLC Genomics Workbench 7.0.4 from thousands millions reads to get consensus. Different methods of phylogeny construction (PAUP, Mr. Bayes and UPGMA) were applied. Results showed the discrimination of coffee species into two clades with high bootstrap value and probability. Four *Psilanthus* sources from Oceania formed into one clade and the other two *Psilanthus* sources from Africa were included in a clade that grouped all *Coffea*. This suggested that the origin of *Psilanthus* is in Asia. Species grouped according to their biogeographic origin and almost all sub-groups are in agreement with previous studies. Surprisingly, *C. canephora* groups with other two supposed *Psilanthus* from West/Central Africa, indicating that the ancestor of *C. canephora* might capture the chloroplast genome from a maternal *Psilanthus* donor (through hybridization). The maternal genomes of *C. arabica* (Arabica) and *C. canephora* (Robusta) were divergent. This result is in agreement with the fact that the chloroplast genome of Arabica should be that of the maternal parent i.e. *C. eugenioides.* This is the first phylogeny of *Coffea* constructed from whole chloroplast genome sequence for coffee trees.

[2] This chapter contains information that is in manuscript submitted to American Journal of Botany (under review): Lan, T., Aprea, G., Giuliano, G., **Tran, H T.M.**, et al. (2017). Homoploid hybrid speciation at the origin of Robusta coffee.

3.1 Introduction

Coffee belongs to the Rubiaceae family and the *Coffeeae* tribe. Research on the taxonomy and classification of coffee has been controversial. Many previous studies divided *Coffeeae* tribe into two genera *Coffea* L. and *Psilanthus* Hook.f, each consisting of two sub-genera: (1) *Coffea* with subgenus *Coffea* (103 species) and *Baracoffea* (J.-F. Leroy) J.-F. Leroy (9 species), and (2) *Psilanthus* with subgenus *Psilanthus* (2 species) and *Afrocoffea* (Moens) (20 species) (summarised by Anthony et al., 2011) (Figure 3.1). However, Davis et al. (2011) grouped the *Coffea* and *Psilanthus* into one genus which includes more than 124 species. While *Coffea* has restricted geographic distribution to tropical humid regions of Africa and islands in the West Indian Ocean, *Psilanthus* is distributed widely in tropical humid regions of Africa, India, South-East Asia and Pacific (Charrier et al., 2012).

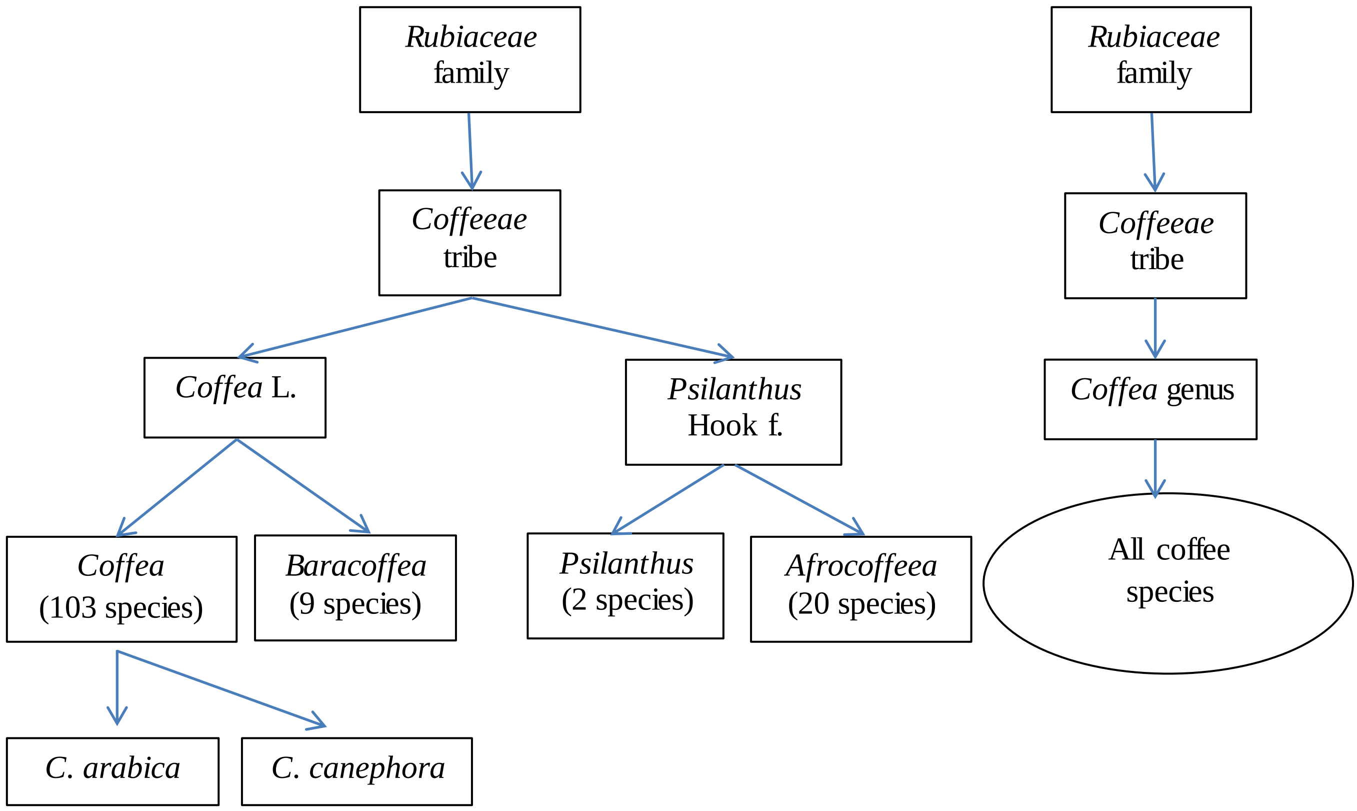


Figure 3.1 Taxonomy of coffee with separate subgenus of *Coffea* L and *Psilanthus* Hook f. (A) (summarised by Anthony et al., 2011) and all in one genus of *Coffea* (B) (Davis et al., 2011).

The relationship between species was determined using several methods such as morphology, chemotaxonomy, hybridisation and cytogenetic studies, and molecular markers. Bridson (1988a, b) (cited by Davis et al., 2011) separated *Psilanthus* and *Coffea* based on five morphological characters, mainly floral structure, such as filaments, the length of anthers, the emergence of anthers, styles and mean number of pollen apertures. Davis et al. (2005) added the calyculi as a key character. This author also described morphological characters present in both genera such as sympodial growth pattern and terminal florescence position, corolla tube as well as the anther. Maurin et al. (2007) also used two characters of style and anthers to separate *Psilanthus* and *Coffea*. However, there is an overlap between species within the two genera for these morphological characters (reviewed by Davis et al., 2011).

Chemotaxonomy is another method to determine the relationship between coffee species. The first two compounds used for this purpose were CGA and purine alkaloid analysed in 30 *Coffea* species. The results enabled three groups to be distinguished, but it did not strictly match the five phylogenetic clades of the *Coffea* genus (Anthony et al., 1993; Clifford et al., 1989). Later, seed diterpenoids were used to examine the relationship among nine African coffee species (de Roos et al., 1997). No clear taxonomic structure was revealed from this study. More recently, two main seed lipid classes, fatty acids (FA) and sterols, were used to investigate the relationships of 17 distinct *Coffea* species except those from *Psilanthus* (Dussert et al., 2008). In this study, all species were classified in seven groups with both fatty acids and sterols; however, only groups of similar seed fatty acid composition exhibit noticeable ecological and geographical coherence and congruence with the clades inferred from nuclear and plastid DNA sequences of previous works (Cros et al., 1998; Lashermes et al., 1997; Maurin et al., 2007).

Hybridisation has also been used to assess the relationship between different species of the two genera, for example, Couturon et al. (1998) hybridised *C. arabica* and tetraploid genotypes of *P. ebracteolatus* (2*n* = 22). The survival rate and the fertility of these hybrids were comparable with those reported for intrageneric crosses between *Coffea* spp. indicating that *Coffea* and *Psilanthus* have no separation at the genetic level.

Lombello and Pinto-Maglio (2003) conducted cytogenetic studies on *Coffea* (*C. brevipes* Hiern, *C. racemosa* Lour.) and *Psilanthus* (*P. ebracteolatus*, *P. benghalensis* and *P. travancorensis*) using chromomycin A3/4′,6-diamidino-2-phenylindole (CMA/DAPI) and fluorescence *in situ* hybridisation (FISH) as cytogenetic markers. Results showed no remarkable cytological variation between the species of the two genera.

Several molecular sequences have been used to clarify the relationship among species belonging to the two genera. The first study used the internal transcribed spacer (ITS) region (ITS2) for 37 *Coffea* and three *Psilanthus* accessions (Lashermes et al., 1997). The results showed little sequence variation between the two genera and some *Psilanthus* species were placed as sisters to *Coffea* species. Plastid sequences from the *trnL–trnF* intergenic spacer (IGS) were used by Cros et al. (1998) for 23 *Coffea* species and two *Psilanthus* species, again revealing low levels of sequence variation between two genera. More recently, Maurin et al. (2007) used a much larger number of samples with 84 species of *Coffea* and 7 species of *Psilanthus* and also a larger number of markers from four plastid regions (*trnL–F* intron, *trnL–F* IGS, *rpl16* intron and *accD–psa1* IGS) and the ITS region (ITS1/5.8S/ITS2) to resolve the relationship among species of the two genera. However, results indicated low levels of sequence divergence and as a result the relationship between two genera remained unresolved. Davis et al. (2007) combined molecular data and molecular–morphological data from three *Coffea* species and 4 *Psilanthus* species and concluded that *Coffea* and *Psilanthus* formed a separated clade. Results from Tosh et al. (2009) using Plastid sequences (trnL-F, rpl16, petD, and accDpsa1) separated *Coffea* and *Psilanthus* in a monophyletic clade. However, this study used only three species of *Coffea* and three species of *Psilanthus.* Anthony et al. (2010) used plastid sequences from *trnL–F*, *trnT–L* and *atpB–rbcL* IGS for 24 *Coffea* taxa and two *Psilanthus* spp. Once again because of the low levels of sequence variation, there were no new insights into the relationship between two genera. The most recent study used the same markers as those reported by Maurin et al. (2007) for 45 *Coffea and* 10 *Psilanthus* species leading to the conclusion that *Psilanthus* should be subsumed into *Coffea* (Davis et al., 2011). Apparently, most authors in previous studies (Cros et al., 1998; Davis et al., 2007; Lashermes et al., 1997; Maurin et al., 2007) concluded that *Coffea* and *Psilanthus* should be in a single genus. Applying data of the whole chloroplast genome sequence would help fully resolve the relationship between *Coffea* and *Psilanthus*.

The use of chloroplast genome in phylogenetic studies in plant species is well known due to (1) its small and relative constant size compared with mitochondrial and nuclear; (2) low frequency of structural changes and conservative rate of sequence evolution; (3) maternal inheritance, but still possible to reveal introgression (reviewed by Cros et al., 1998; Palmer, 1985); (4) haploid nature (reviewed by Rogalski et al., 2015).

Molecular markers have been widely used in phylogenetic studies. Recently, with the advances of next-generation sequencing techniques due to its high-throughput, time-savings, and low-cost (reviewed by Goodwin et al., 2016), phylogenetic studies at the genome-wide level have become feasible and significantly increasing (reviewed by Rogalski et al., 2015). In this study, we present the complete chloroplast sequence of 14 species Illumina sequencing technology of total DNA. The objectives of this research is to find out:

(1) What is the level of *Coffea* divergence at the genome level;

(2) whether *Psilanthus* is part of *Coffea* or a very closely-related genus;

(3) what is the phylogenetic position of arabica and canephora (the two commercial coffee species) in *Coffea* and the implication for breeding using introgression schemes.

3.2 Materials and Methods

3.2.1 Genetic materials and sequencing

Thirteen species (one individual per species) were collected from different sites and sequenced on an Illumina platform (Sequencing mode: 2 x 100. Instrument: HiSeq. Software: HiSeq Control Software 2.0.12.0) by various collaborators in the "genome13" project, an international coffee consortium of researchers working on the analysis of relationships in the genus. The project involved the Coffee Board in India and 13 institutions from 8 countries: France (2 institutions), the US (3), Australia (1), Brazil (3), Cote D'Ivoire (1), Madagascar (1), Ethiopia (1), and Italy (1). Sequencing data were provided from different sources as follows:

(1) 13 species (one duplicate for validation making 14 sequences) were provided by the "genome13" project.

(2) The species *C. brassii* was collected in Australia by Prof. Darren Crayn (Director of the Australian Tropical Herbarium in Cairns - North Queensland); DNA extraction was performed by Tal Cooper, an honours student from QAAFI/UQ, and sequencing was performed on an Illumina platform.

(3) Sequences of *C. canephora* were derived from the *C. canephora* genome project (Denoeud et al., 2014) by Prof. Robert Henry.

(4) The second sequence database (Bud 15) of *C. canephora* (of distinct origin) was provided by Alexandre de Kochko (Institute of Research for Development, France).

(5) The chloroplast whole genome sequence of *C. arabica* reported by Samson et al. in 2007 was used as a reference (available at http://chloroplast.ocean.washington.edu/tools/cpbase/run). In total, 18 samples representing for the 16 species (included one reference) were used in the analysis as presented in Table 3.1. These species possess distinct features in relation to flower morphology, biochemical compounds variation (Table 3.2), habit (small shrubs to 20 m high) habitat or growing altitude (sea level up to 2000 m) (Table 3.1), variation in fruit duration (2 to 14 months), colour (yellow to black) and shape, and pest resistance ("genome13" project description). For example *Coffea* with exerted anthers and stigma while that of *Psilanthus* is inserted, resulting in the change in mating system: *Psilanthus* is autogamous, while allogamy is possible for *Coffea* species. *C. pseudozanguebariae, C. humblotiana* and *C. tetragona* have no caffeine; *C. stenophylla* can grow at low altitude or dry condition. In addition, these species are listed as near threatened or endangered implying the need for conservation.

Table 3.1 The sources of species' sequences used in the study

No	Species Name	Abbrevia-tion, when used	Plant, voucher/herbarium code	Country of origin [African sub-region]	Germplasm collection source
Coffea, formerly Psilanthus					
1	*C. benghalensis* var bababudanii (Sivar., Biju & P.Mathew) A.P.Davis	BABA	PBT1 (CCRI)	IND	CBI
2	*C. benghalensis* (Heyne ex J.A. Schult.) Leroy)	BENG	PBT5 (CCRI)	IND	CBI
3	*C. brassii* (J.-F.Leroy) A.P.Davis	BRA	D. Crayn 1196 (CNS)	AU	CNS
4	*C. horsfieldiana* (Miq.) J.-F. Leroy	HORS	HOR (K)	ID	ICCRI
5	*C. mannii* Hook.f.	MA45A	2003 1365-45 (BR)	CAM [W/WCA]	BR
		MA45G			
6	*C. ebracteolata* Hiern	PSI	PSI11 (K,P)	Ivory Coast [W/WCA]	BRC
Coffea					
7	*C. canephora* Pierre ex A.Froehner	CAN1	DH200-94	DRC [W/WCA]	BRC
	C. canephora Pierre ex A.Froehner	CAN2	BUD15 (K)	UG [CA]	BRC
8	*C. stenophylla* G.Don. (1)	FB	FB55 (K)	Ivory Coast [W/WCA]	BRC
9	*C. humilis* A.Chev. (2)	GH	G57 (K)	Ivory Coast [W/WCA]	BRC
10	*C. pseudozanguebariae* Bridson (3)	HP	H53 (K)	Kenya [EA]	BRC
11	*C. racemosa* Lour (4)	IB	IB62 (K)	MOZ [EA]	BRC
12	*C. dolichophylla* J.-F.Leroy	DOL	A.206 (P)	MAD	KCRS
13	*C. humblotiana* Baill. (5)	HUMB	BM19/20 (K, MO, TAN)	Comoros	BRC
14	*C. macrocarpa* A.Rich. (5)	MAC	PET (P, K)	MAU	BRC
15	*C. tetragona* Jum. & H.Perrier (6)	TET	A.252 (K, MO, TAN)	MAD [NW]	KCRS
16	*C. arabica*	ARA	Samson et al. 2007	SWEt/NEA	

Abbreviations Countries: Cameroon (CAM); Democratic Republic of Congo (DRC); UG (Uganda); Madagascar (MAD); Mauritius (MAU), India (IND) and Mozambique (MOZ). African regions: Central Africa (CA), West and West-Central Africa (W/WCA), East Africa (EA), SWEt: South-West Ethiopia, NEA: North East Africa. Origin of germplasm material: unknown germplasm collection from Brazil (BRA); Indonesian Coffee and Cocoa Research Institute (ICCRI); Centre de Ressources Biologiques *Coffea*, Saint Pierre, Reunion (BRC); Kianjavato Coffee Research Station, Madagascar (KCRS); Coffee Board of India (CBI); National Botanic Garden of Belgium (BR); Australian Tropical Herbarium, (CNS); Remarks: (1) *C. affinis*, grown in dry condition and low altitude (100m), (2) near threatened, (3) grown at 500m, vulnerable, (4) grown at 400m, near threatened, (5) vulnerable, (6) endangered. Species in endangered, threatened or vulnerable was classified by Davis et al. (2006).

Table 3.2 Quality characteristics of the sequenced species

No	Species	Sucrose	Trigonelline	Caffeine	Lipids	CGAs
1	*P. bababudani*			0.00 [3]		
2	*P. bengalensis*			0.00 [3]		
3	*P. brassii*			0.00 [3]		
4	*P. horsfieldiana*			0.00 [3]		
5	*P. mannii*			0.00 [3]		
6	*P. ebracteolatus*			0.03 [3]		0.27 [3]
7	*C. canephora*	6.10 [1]	0.82 [1]	2.64 [2]	9.83 [6]	11.34 [2]
8	*C. stenophylla*	7.5 [1]	1.09 [1]	2.27 [2]	8.0 - 12.8 [5]	8.23 [2]
9	*C. humilis*	6.89 [1]	0.52 [1]	1.93 [2]		8.65 [2]
10	*C. pseudozangue-bariae*	7.95 [1]	1.02 [1]	0.00 [2]		1.47 [2]
11	*C. racemosa*	6.44 [1]	1.02 [1]	1.06 [2]	10.39 [6]	5.33 [2]
12	*C. dolichophylla*					
13	*C. humblotiana*	5.73 [1]	0.81 [1]	0.00 [2]		1.00 [2]
14	*C. macrocarpa*					
15	*C. tetragona*			0.03 [4]		1.40 [4]
16	*C. arabica*	9.32 [1]	1.13 [1]	1.2 [2]	17.3 [5]	4-8.4 [2]

Sources: [1] Campa 2004; [2] Campa 2005; [3] Clifford et al. 1989; [4] Anthony et al. 1993; [5] Crisafulli 2013; [6] Mazzafera et al. 1998

3.2.2 Data analysis and verification

Chloroplast sequence assembly pipeline

All steps, unless indicated, were carried out using the CLC Genomics Workbench (CLC-GWB) software (CLC Genomics Workbench 7.0.4, http://www.clcbio.com). Whole genomic DNA of 16 *Coffea* samples was subjected to whole shotgun next generation sequencing (NGS) using the Illumina HiSeq2500 to obtain 100 bp paired-end reads. All raw NGS reads for the seventeen samples (fifteen species) were imported into CLC-GWB. In addition, sequence of *C. arabica* chloroplast genome from NCBI was also imported to CLC as a standard import and used as a reference chloroplast sequence. Raw reads were subjected to Quality Control (QC) analysis which was used as a guide for trimming the reads. Low-quality paired-end sequence reads were trimmed using default parameters. The quality score limit was set to 0.05 (corresponding to Phred quality value >15) and minimum number of nucleotides in reads of 15 bp. These trimmed reads were used in the assembly of the chloroplast genomic sequence using two approaches: reference-guided mapping assembly and *de novo* assembly.

+ For reference-guided mapping assembly, the trimmed reads were subjected to read mapping using *C. arabica* as the reference sequence. In addition, the consensus sequence derived from mapping step was also used as the reference sequence. Indel structural variants analysis was performed based on the mapping files with P-Value threshold of 0.0001. The result of indel structural variants was used as guidance-variant track for local re-alignment to create the stand-alone mapping and the consensus sequence was then used as the mapping-derived chloroplast assembled sequence. Trimmed reads were then mapped to the mapping-derived chloroplast sequence to obtain the mapping-Cp-mapping file.

+ For *de novo* assembly, different combinations of settings for word sizes and bubble sizes as well as minimum contig lengths were used to get the best contigs. The contigs were then subjected to BLAST-analysis against the *C. arabica* chloroplast as the reference for the selection of long contigs. The selected contigs were then aligned to the *C. arabica* chloroplast reference sequence using Clone Manager (SciEd, USA) to determine the correct orientation and their complete coverage of the reference chloroplast sequence. The selected and correctly oriented contigs were updated using the Update Contig tool in the CLC genomics Workbench. The Updated Contigs were aligned back to the reference sequence, the overlaps determined and stitching of the contigs at the overlaps was undertaken to obtain a *de novo*-derived chloroplast sequence. The trimmed reads were mapped to the *de novo*-derived chloroplast sequence to obtain corresponding *de novo*-Cp-mapping file. The *de novo*-derived chloroplast sequence and the mapping-derived chloroplast sequence were aligned to determine any discrepancies between these two assembled chloroplast sequences. Any discrepancies observed were manually curated by observing the reads mapped at the corresponding nucleotide position in the mapping-Cp-mapping file and the *de novo*-Cp-mapping. Clone manager 9 was also used to align contigs, process the molecule (invert the sequence). All analysis steps are presented in Figure 3.2.

Dealing with gaps in de novo assembly

Several samples obtained enough contigs to generate the whole chloroplast sequence after a few *de novo* assemblies. However, other samples had gaps up to thousands of bases due to missing contigs. In this case, more *de novo* assemblies were run with different settings of word sizes and bubble sizes, and using contigs from several *de novo* assembly results to fill the gaps.

Dealing with conflicts between mapping consensus and de novo consensus

If after running the first local realignment, the first two consensuses from reference-guided mapping assembly (*C. arabica* and consensus derived from mapping as reference) and consensus from *de novo* assembly had numerous discrepancies, additional mapping of trimmed reads using the consensus sequence from the preceding mapping steps as a reference sequence followed by local realignments was used which led to further improvement to the consensus.

3.2.3 Phylogeny construction

To create a phylogeny of coffee species, several out-group species belonging to closely related families were selected. Literature showed grape (*Vitis vinifera)* and tomato (*Solanum lycopersicum*) were close to coffee (Guyot et al., 2012) and they were used as out-groups in the phylogenetic construction. Tobacco (*Nicotiana tabacum)* sharing the same family of Solanaceae as tomato and a further plant Ardisia *polyticta* - a basal asterid genus - were also selected as out-groups in the analysis. Final consensus of all species were removed the IRb region and then subjected to Geneious 7.0.5 software platform (www.geneious.com/) for alignment under fasta file format using the plug-in named MAFFT (Katoh et al., 2002). The alignment result was checked for credibility then subjected to several plug-ins and methods to construct the phylogeny trees including PAUP 4.0 (Maximum Likelihood (ML), Maximum Parsimony (MP)), MrBayes and UPGMA. For PAUP - Maximum Likelihood and MrBayes, the model tests were run with software jModeltest 2 (Darriba et al., 2012) and Akaike information criterion to find the best-fit model of nucleotide substitution prior to running the phylogenetic tree. Two models of GTR + G and GTR + I + G were chosen for maximum likelihood calculation. ML, MP analyses were completed in PAUP* 4.0 Beta software package (Swofford, 2002) choosing the heuristic search type to find the optimal tree, and using the stepwise addition to obtain a starting tree. Branch-swapping algorithm was set for tree-bisection-reconnection (TBR). Bootstrapping was performed with random seed type and number of replications of 100. Gaps were treated as missing data. For ML, no assumed proportion of invariable sites and equal distribution of rates at variable sites were set.

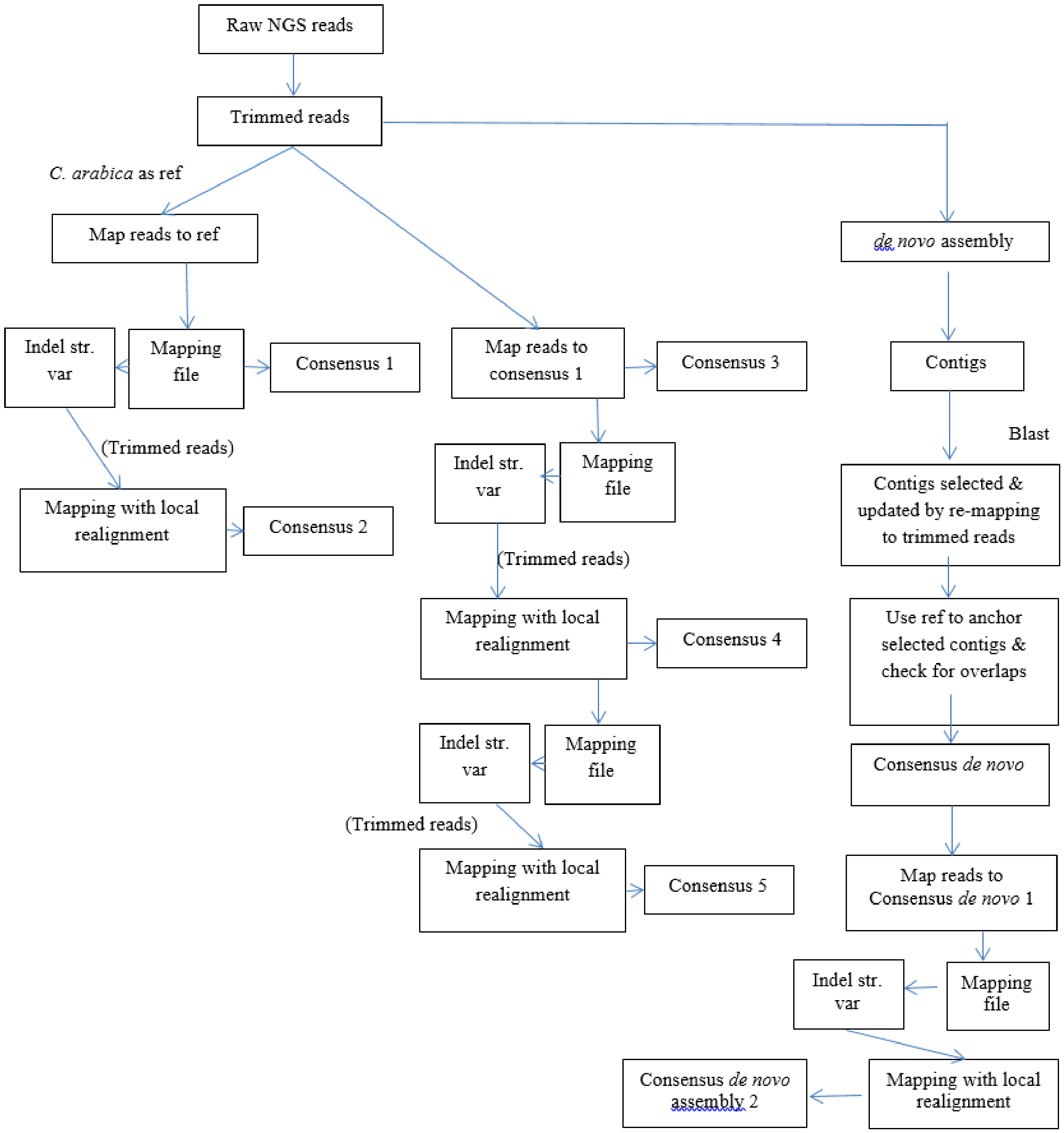


Figure 3.2 CLC analysis workflow.

Bayesian analysis was performed using MrBayes (Huelsenbeck and Ronquist, 2001) through the Geneious 7.0.5 software platform (www.geneious.com/). The evolutionary models used were GTR + G (the General Time Reversible Model with G gamma-shaped for rate variation) and GTR + I + G. The branch length prior was set to exponential with parameter 10. Settings for MCMC (Monte Carlo Markov Chains) include chain length of 1,100,000; heated chains of 4; heated chain of 0.2; subsampling frequency of 200; burn-in length of 100,000 and random seed of 30.403 (for GTR + I + G) and 12,035 (for GTR + G). Consensus nodal support was assessed by posterior probability distribution. All trees were rooted using the outgroup method. The UPGMA method was applied for tree construction incorporating genetic distance models and bootstraps for resampling. Random seed was set at 313,227 with 100 replicates to create the consensus tree. The percentage of support threshold is 50.

3.3 Results and discussion

3.3.1 Reference-guided mapping assembly and de novo assembly

The analysis was performed with a big data set of more than 100 million reads (except *C. canephora 1* using a different sequencing platform). After trimming, almost all samples had reads with a length of more than 99 bases indicating a good quality of the sequence (Table 3.3). The number of *de novo* assemblies performed for each sample varied from 3 to 12. A consensus was built from contigs of one *de novo* assembly (7 samples) to 8 *de novo* assemblies (in *P. mannii* G). In some samples, only a few contigs allowed generation of the whole chloroplast consensus (4 contigs in *P. bengalensis*, *P. brassii*, *C. racemosa* and *C. humblotiana*) (Figure 3.3a) while others needed up to 17 contigs from different *de novo* assemblies to complete the whole consensus (*P. mannii* G) (Figure 3.3b). The discrepancies among different consensuses generated from reference-guided mapping assembly and *de novo* assembly varied from only five (*C. macrocarpa)* to 39 (in *P. bengalensis)* (Table 3.3). When there were discrepancies between the consensuses of reference-guided mapping assembly and *de novo* assembly, the latter seemed to be more reliable after checking carefully with the mapping file (Figure 3.4a). However, in some cases, the other gave much better and clearer results (Figure 3.4b). Every discrepancy of every sample was checked to get the final consensus for each sample.

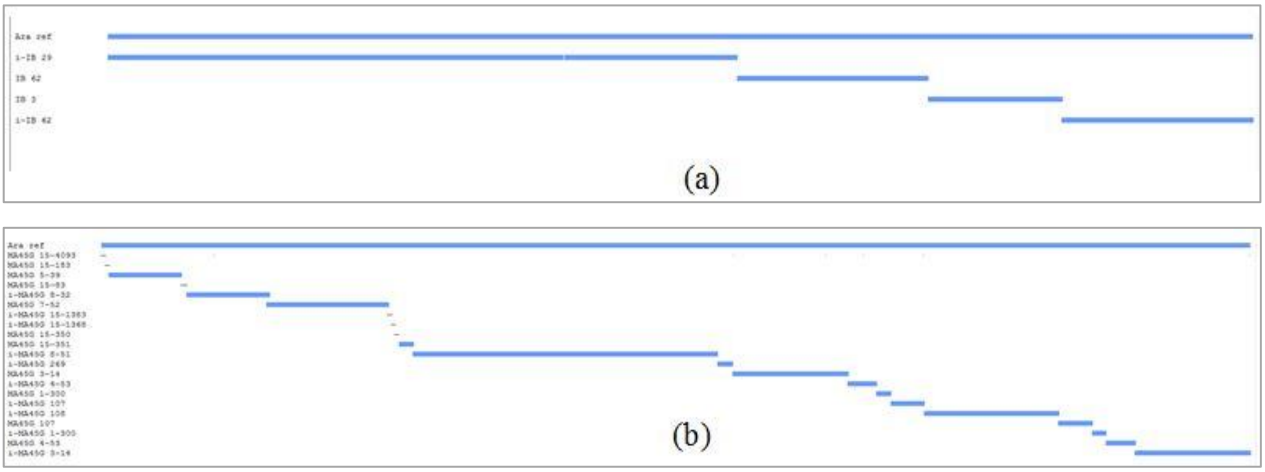


Figure 3.3 Number of contigs used to generate whole chloroplast consensus in (a) *C. racemosa* (4 contigs) and (b) *P. mannii* (17 contigs).

Table 3.3 Quality parameters of reads and *de novo* assembly outcomes

No	Species	No of reads	Ave. length after trimming	No of *de novo* run	No of *de novo* used	No of contigs used	No of conflicts
1	*P. bababudani*	165,748,509	99.8	9	5	13	16
2	*P. bengalensis*	163,620,144	99.8	3	1	3	39
3	*P. brassii*	144,329,882	97.6	6	1	3	19
4	*P. horsfieldiana*	163,133,183	99.7	6	3	12	24
5	*P. mannii* A	121,618,582	99.7	8	5	15	17
	P. mannii G	111,543,674	99.7	12	8	17	20
6	*P. ebracteolatus*	169,186,834	98.3	5	2	7	17
7	*C. canephora 1*	35,032,562	75.8	6	4	8	19
	C. canephora 2	331,774,597	99.8	4	3	16	11
8	*C. stenophylla*	151,177,652	99.7	3	1	5	31
9	*C. humilis*	143,478,290	99.7	6	2	5	12
10	*C. pseudozangue-bariae*	151,973,371	99.7	11	2	6	16
11	*C. racemosa*	145,503,020	99.7	3	1	3	11
12	*C. dolichophylla*	150,138,476	99.7	6	1	5	9
13	*C. humblotiana*	135,572,346	99.7	6	1	3	12
14	*C. macrocarpa*	190,904,048	99.7	3	1	4	5
15	*C. tetragona*	129,525,188	99.7	4	2	7	15

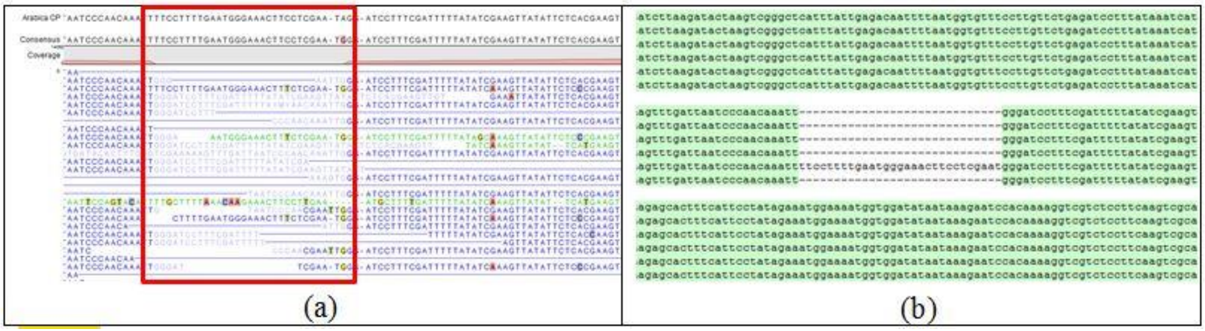


Figure 3.4 (a) Mismatch in reference-guided mapping assembly (b) *de novo* assembly shows discrepancies to other consensus.

Chloroplast genome is structured into a single circular chromosome that contains a large single copy region (LSC), a small single copy region (SSC) and two copies of an inverted repeat (IRs) (Palmer 1991 in Liu 2013). Data on chloroplast genome structure of 16 species (18 genomes) were presented in Table 3.4.

Table 3.4 Information of the chloroplast sequence of all studied species

No	Species	Genome size* (pg)	Total (bp)	LSC (bp)	IRs (bp)	SSC (bp)
1	*P. bababudani*	NA	154,713	84,817	25,877	18,142
2	*P. bengalensis*	NA	154,821	84,955	25,871	18,124
3	*P. brassii*	NA	154,545	84,800	25,770	18,205
4	*P. horsfieldiana*	NA	154,978	85,118	25,862	18,136
5	*P. mannii 1*	NA	155,016	85,113	25,874	18,155
	P. mannii 2	NA	155,016	85,113	25,874	18,155
6	*P. ebracteolatus*	NA	155,084	85,134	25,905	18,140
7	*C. canephora 1*	1.43-1.55	154,726	84,840	25,888	18,110
	C. canephora 2		155,097	85,028	25,946	18,177
8	*C. stenophylla*	1.28-1.35	155,087	85,228	25,854	18,151
9	*C. humilis*	1.76-1.84	155,129	85,277	25,858	18,136
10	*C. pseudozangue-bariae*	1.13-1.23	155,031	85,124	25,901	18,105
11	*C. racemosa*	0.95-1.05	154,951	85,136	25,849	18,117
12	*C. dolichophylla*	NA	155,168	85,288	25,880	18,120
13	*C. humblotiana*	0.97	155,109	85,227	25,884	18,114
14	*C. macrocarpa*	1.17	155,186	85,271	25,882	18151
15	*C. tetragona*	1.07	155,134	85,273	25,880	18,101
16	*C. arabica*	2.47-2.61	155,189	85,166	25,943	18,138

* Genome size was estimated by 2 C nuclear DNA content (pg) in Cros et al. 1995, Noirot et al. 2003, Razafinarivo et al. 2012. Total: Total length of chloroplast; LSC: Large Single Copy; SSC: Small Single Copy; IRs: Inverted Repeat Regions.

P. brassii had the smallest chloroplast genome (154,545 bp) while *C. macrocarpa* was highest (155,186 bp) for the diploids studied (Table 3.4). The proportion of each region was not necessarily even among the species. Compared to other closely related plants such as tomato (155,461 bp), tobacco (155,943 bp) and grape (160,928 bp), coffee had a smaller chloroplast genome. It seems that the size of the chloroplast genome (total length in bp) did not correspond to the size of the whole genome (measured in pg). Nuclear DNA content estimated by flow cytometry showed *C. racemosa* has smallest value (0.95 - 1.05 pg) and *C. humilis* has highest value (1.76 -1.84 pg) of diploid species studied; other diploid species such as *C. pseudozanguebariae* (1.13-1.23 pg), *C. stenophylla* (1.28-1.35 pg) and *C. canephora* (1.43-1.55 pg) have middle values. *C. arabica* is a tetraploid species and has highest value of 2.47-2.61 pg (Cros et al., 1995; Noirot et al., 2003; Razafinarivo et al., 2012). However, in this study, the chloroplast genome of *C. racemosa* (154,951 bp) was bigger than that of *C. canephora 1* (154,726 bp) and *C. humilis* size was smaller than several diploid species (*C. tetragona, C. dolichophylla and C. macrocarpa*). There was also variation in the size between the two *C. canephora* samples (154,726 bp and 155,097 bp). Razafinarivo et al. (2012) concluded that genome size and genetic divergence or genome size and phylogeny have no relationship. In this study, it seems that the size of the chloroplast also did not explain the relationship among coffee species.

3.3.2 Phylogenic relationships between species

The consensuses of all species were subjected to Geneious 7.0.5 software platform for alignment and the alignment file was used to construct the phylogeny trees. Alignment statistics include: length of 137,380 bases, identical sites of 106,538 bases (77.6%), pairwise identity of 95.6%, and mean of ungapped sequence length of 129,231.6 bases with standard deviation of 383.2.

Although the topology may be different among four methods, the species within each clade or sub-clade and relationship between small groups remain unchanged. All methods of Maximum Parsimony (Figure 3.5a), UPGMA (Figure 3.5b), Maximum Likelihood (Figure 3.5c), and Mr. Bayes (Figure 3.5d) gave consistent results which separate species into two clades with very high bootstrap value and probability and clear distinction between coffee and the out-groups. The first clade includes 4 *Psilanthus* species; the second clade consisted of two sub-clades: (1) two distinct samples of *C. canephora* grouped with other two supposed *Psilanthus* from West/Central Africa, and (2) other *Coffea* species. Species grouped according to their biogeographic origin such as West/Central Africa, or East/Central Africa, or Madagascar, or Oceania (Figure 3.6). This is the first phylogeny of *Coffea* constructed from whole chloroplast genome sequence which gives much better resolution compared to other methods. *P. brassii* groups with three other *Psilanthus* accessions from India and Indonesia to form one clade, which is similar to results reported by Davis et al. (2011) in which Asian and Australian *Psilanthus* were grouped in one clade. Surprisingly, *C. arabica* and *C. canephora* were far apart from each other, even though *C. canephora* is supposed to be an ancestral parent of *C. arabica.* However, chloroplast inheritance is generally maternal so a close relationship between *C. arabica* and *C. canephora* chloroplast genomes should not be expected. This finding is in agreement with several studies using molecular markers which placed these two species in different clades (Anthony et al., 2010; Cros et al., 1998). That *C. canephora* had a cpDNA closer to *Psilanthus* than *Coffea* could seem to indicate that the ancestor of *C. canephora* captured the chloroplast genome from a maternal *Psilanthus* donor (through hybridisation). The hybridisation between *P. ebracteolatus* and *C. arabica* confirmed the close relationship between *P. ebracteolatus* and *Coffea* species (Couturon et al., 1998). Similarly, in this study, *P. ebracteolatus* was placed in the same sub-clade with *C. canephora* indicating that *P.*

ebracteolatus was close to *Coffea* species and the previous classification may need to be justified. The relationship between *P. ebracteolatus* and *P. mannii* was very inconsistent between studies. While several studies showed distance between *P. ebracteolatus* and *P. mannii* (Anthony et al., 2010; Cros et al., 1998), other studies which are in agreement with the current study show the close relationship between them (Davis et al., 2011; Maurin et al., 2007). Sub-groups for *C. racemosa* and *C. pseudozanguebariae*, *C. humilis* and *C. stenophylla* were always in agreement with other studies (Anthony et al., 2010; Cros et al., 1998; Davis et al., 2011; Lashermes et al., 1997; Maurin et al., 2007). The placement of *C. arabica*, *C. humblotiana*, *C. dolichophylla*, *C. macrocarpa* and *C. tetragona* were slightly different among trees, especially the relationship among *C. humblotiana*, *C. dolichophylla* and *C. tetragona*. Within *Coffea*, *C. macrocarpa* seems to be the most ancestral chloroplast genome.

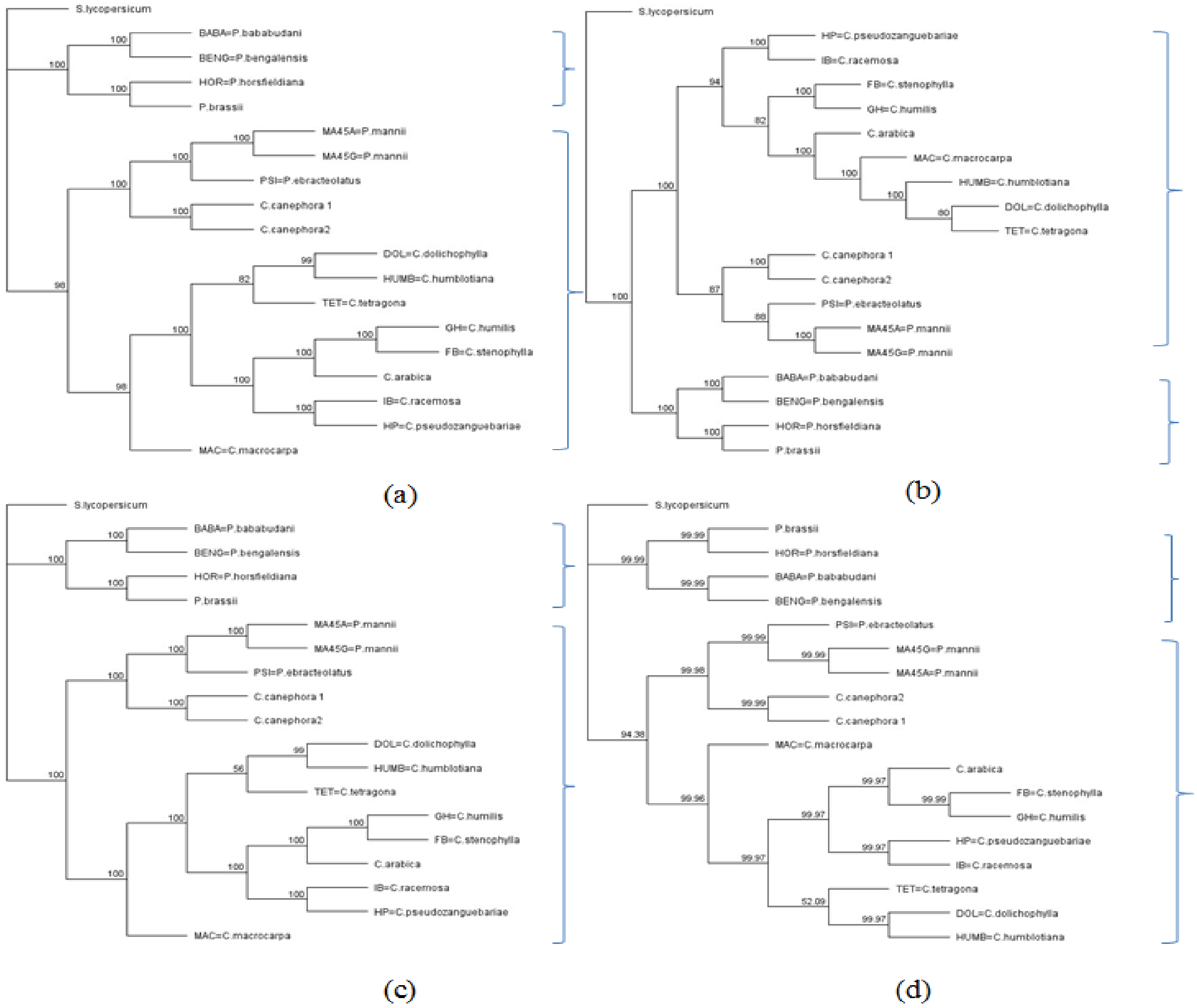


Figure 3.5 Phylogenetic tree constructed by different methods: Maximum Parsimony (a), UPGMA (b), Maximum Likelihood (c), and Mr. Bayes (d) using 4 out-groups of forest tree (*Ardisia polyticta)*, grape (*Vitis vinifera)*, tomato (*Solanum lycopersicum*) and tobacco (*Nicotiana tabacum).*

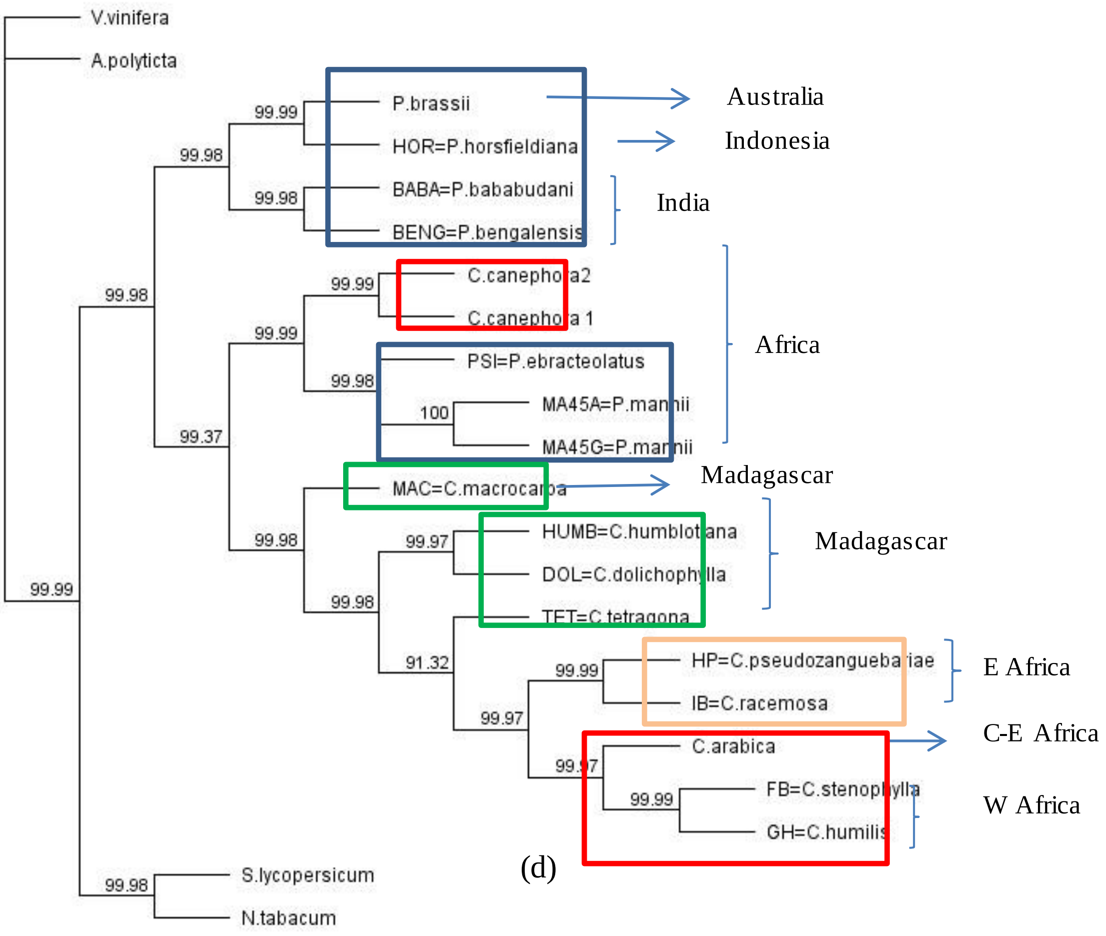


Figure 3.6 Relationship between coffee species using chloroplast genome constructed by Mr. Bayes method

3.4 Conclusions

To the student's knowledge, this is the first phylogeny of *Coffea* constructed from whole chloroplast genome sequence for coffee trees. While almost all previous studies did not show distinct clades between *Psilanthus* and *Coffea*, and tend to subsume *Psilanthus* to *Coffea*, this study discriminated coffee species to two clades with high bootstrap value and probability. Four *Psilanthus* sources from Asia formed into one separate clade, while the other two *Psilanthus* sources from Africa were included in a clade that grouped all *Coffea*. This suggested that the origin of *Psilanthus* is probably in Asia. If *Psilanthus* and *Coffea* belong to the same genus, the origin of coffee trees probably from Asia. Species grouped according to their biogeographic origin and almost all sub-groups are in agreement with previous studies. The position of *C. canephora* is interesting and should be considered as more complex than previously thought. As chloroplast genome is maternal inheritance, it is likely that canephora has *Psilanthus* genome but received nuclear genome from *Coffea*. Chloroplast genome showed that a hybridisation event between an African *Psilanthus* and a canephora ancestor. The maternal genomes of *C. arabica* (arabica) and *C.*

canephora (robusta) are divergent. This result is in agreement with the fact that the chloroplast genome of Arabica should be that of the maternal parent i.e. *C. eugenioides.*

This study was based on sequence data available for a certain number of species. Due to a deceasing trend in the cost of sequencing, more data from additional species is expected to be available from various research groups in the future. This perspective would bring in a more concrete conclusion on the genomic synteny between species, especially those close to the two domesticated species *C. arabica* and *C. canephora* that can be utilised in coffee breeding programs. Especially, species containing almost no caffeine including *C. humblotiana*, *C. tetragona* (close *C. arabica)* and *P. ebracteolatus* (close to *C. canephora*) have been an important breeding focus. Promising results have been reported for low-caffeine hybrids derived from *Stenophyla* hybridised with arabica varieties (Mundo Novo and Catuai Red 1 and 2). Likewise, the caffeine level in *C. canephora* was also increased in its hybrids derived from *C. eugenioide* (Clifford and Kazi, 1987). According to Hamon et al. (2015), interspecies crosses are possible for *Coffea* species, but there may be high variations in the success rate of hybridization. This first phylogeny of *Coffea* constructed from whole chloroplast genome sequences contributes to the understanding of relationships among coffee species that could help guide efficient hybridisation schemes in coffee breeding programs for quality improvement.

CHAPTER 4: GENETIC DIVERSITY OF ARABICA BASED ON BEAN MORPHOLOGY AND BIOCHEMICAL COMPOUNDS [3]

Abstract

Selection of improved quality coffee genotypes can be supported by the use of association genetics linking phenotypic data (morphological and biochemical components in the coffee bean determining coffee quality) with genotypic data. The choice of germplasm representing diversity is critical for the success of an association study. The narrow genetic base resulting from a genetic bottle neck during domestication and self-pollination of commercial arabica has been well documented. However, the sub/spontaneous accessions have higher genetic diversity as assessed by molecular markers in other research. Beans of 232 diverse arabica coffee accessions originating from 27 countries were harvested from the germplasm collection at CATIE, Costa Rica. Substantial variation was observed for bean morphology including 100-bean weight, bean length, width, thickness, and bulk density. Non-volatiles including caffeine and trigonelline were analysed and showed larger variation in range than has previously been reported. Results of targeted analysis of 18 volatiles from 35 accessions also showed significant variation, with coefficients of variation from 140% for 4-vinylguaiacol to 62% for geraniol. There were strong correlations between some volatile compounds, suggesting that representative volatiles used in selection would save analytical costs. However, no strong correlation was found between bean morphology and the levels of non-volatile or volatile compounds, implying that it is difficult to select for low or high composition of these compounds based on bean physical characteristics. Utilizing the large variation observed for bean morphology and biochemical traits, it should be possible to select for desirable combinations of traits in arabica coffee breeding.

[3] This chapter contains information that is also published in: Tran, H. T. M., Vargas, C. A. C., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. (2017). Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (*Coffea arabica* L.). *Tree Genetics & Genomes* **13**, 1-14

4.1 Introduction

C. arabica is known to have a very narrow genetic base (Charrier and Eskes, 2004) which is probably due to the narrow geographical origin, the selection bottleneck and self-pollination (summarised by Tran, 2005). A number of studies have focussed on assessment of arabica diversity using diverse groups of materials such as cultivar, mutants, spontaneous/sub-spontaneous accessions (terms defined in Charrier and Eskes, 2004) and hybrids as well as different marker systems (Aerts et al., 2012; Aga et al., 2003; Dessalegn et al., 2008b). These studies indicated a limited genetic variability available for arabica breeding programs, even with the wild accessions (Aga et al., 2003; Anthony et al., 2001; WCR, 2014). Further assessment in a wider germplasm collection is required to identify genetically diverse material to be then used as a useful sources for improving arabica coffee quality.

Coffee quality is assessed based on both physical and organoleptic quality. The physical quality characteristics include length, width, thickness or weight, shape or colour of coffee beans while the organoleptic quality relates to the fragrance or aroma, flavour, sweetness, acidity or overall taste (Leroy et al., 2006). Both the physical and chemical attributes of coffee beans are important criteria which determine the value of arabica coffee beans in world markets (Belete, 2014).

For the physical bean quality, weight of 100 beans is used as an indicator of bean size. Alternatively, the beans can be measured directly on the sieve in industry. Coffee with larger beans usually get good grade and fetch higher price than smaller ones even though the former do not necessarily produce desirable roast or liquor (Belete, 2014). Length, width and thickness or ratio of length/width is the reflection of the bean shape. Data on bean size and shape of coffee as well as its genetic research was not well documented. There is also limited information on the correlation between physical bean and the chemical compounds or cup quality in coffee. The amount of moisture in coffee parchment and beans is important because coffee too high or too low in moisture will not maintain high cupping quality. Green coffee that is high in moisture (greater than 12% wet basis) can deteriorate due to bacteria, mould, or yeast, especially if the seed is killed. If the seed remains alive, enzymatic activity will cause the cupping quality to change. In any case, the parchment coffee moisture level should be lowered to below 12% soon after harvest (Gautz et al., 2008). International Coffee Organization (ICO) recommended the standard moisture for green bean is 9-12% (Subedi, 2011).

Caffeine and trigonelline are key non-volatile compounds in coffee. Caffeine is one of the main alkaloids contributing to the strength, body and bitterness of brewed coffee (Trugo, 1984). Trigonelline is a pyridine alkaloid and considered to be important for both taste and nutrition (Ky et al., 2001). High quality coffee correlated with high level of trigonelline in green and roasted coffee beans (Farah et al., 2006b). Trigonelline contributes indirectly to the formation of different desirable aromas during the roasting process (Ky et al., 2001b).

Coffee flavour and its formation is extremely complex; however, a number of studies on coffee flavour and its constituents have been reported. The contribution of volatile and non-volatile compounds to coffee quality and their relationship to coffee sensory properties has been thoroughly reviewed by Flament (2002) and more recently by Sunarharum et al. (2014).

The analytical methods for the determination of non-volatile (i.e. caffeine and trigonelline) and volatile compounds (both non-targeted and targeted analysis) have been mentioned in Chapter 2 (2.3.2 Methods of non-volatile and volatile analysis in coffee).

To analyse volatiles in roasted coffee bean, green coffee bean must first be roasted, as roasting is a crucial step towards a good cup of coffee. Factors affecting the volatile composition as well as its variation during the roasting process been reviewed in Chapter 2 (2.3.1.2 Volatiles).

In this study, we report the analysis of genetic diversity of a *C. arabica* population collected from one of the world's largest coffee collections based on measurement of bean physical properties and composition using state-of-the-art analytical technologies. The outcome would help guide breeding programs in selection for favourable bean size, caffeine, trigonelline and volatile compounds.

As mentioned above, arabica has a limited genetic variability while association studies require a diverse germplasm resource. The evaluation of genetic diversity of arabica germplasm collection is the first step in association mapping. The aims of this study are:

1/ To collect a wide germplasm set of arabica coffee and evaluate their genetic diversity with regards to bean morphology, non-volatiles (caffeine and trigonelline) and volatiles (analysed by non-targeted and targeted approaches).

2/ To determine if there is any correlation between bean physical properties and chemical composition (non-volatiles and volatiles), so that it can be used for indirect selection of desired quality genotypes.

3/ To select desirable genotypes for use in breeding and association analysis to identify genes or markers linked to coffee quality.

4.2 Materials and Methods

4.2.1 Materials

4.2.1.1 Genetic material conserved in CATIE/collection site and its diversity

CATIE (Centro Agronómico Tropical de Investigación y Enseñanza – Centre for Tropical

Agricultural Research and Higher Education, Costa Rica) is holding one of the largest *C. arabica* coffee genebank in the world, containing 1730 accessions (updated Nov 2014). It comprises wild coffee trees collected in the centre of origin, varieties and mutants selected in various research centres, as well as intra- and interspecific hybrids from 27 countries in total. This is the only gene bank of coffee available in Latin American and Caribbean countries (Anthony et al., 2007). The collection is located in an area of approximately of 8.5 ha at latitude 9^0 53' 57" N, longitude -83^0 39' 43" W, and at an elevation of 616 m above sea level. It has an average annual rainfall of 2,779 mm and an average annual temperature of 22.2^0C. Its soil has a pH of 4.76 and contains 3.62% organic carbon, 0.32% total nitrogen, and 11.3 mg/l phosphorus (provided by CATIE, obtained in 2012). Most of the accessions were grown from seed each of which constitutes a genotype. Each accession has four to eight trees. The spacing between rows and between trees within rows is generally 2.5 m x 2 m. *Erythrina popeppigiana* trees were grown as shade trees at the spacing of 6 m x 6 m (Figure 4.1). The collection's maintenance was described in detail by Anthony et al. (2007). Harvesting is conducted 4 times in a year from July to November. Each individual accession was marked by a concrete board (Figure 4.1). Accessions were classified into three groups based on their genetic origins: (1) Accessions from the diversity centre of *C. arabica*; (2) Accessions from Typica and Bourbon-derived varieties and mutants; and (3) Accessions from introgressed lines selected from interspecific hybrid progeny (*C. arabica* x *Coffea* spp.) (Anthony et al., 2007b). Genetic markers revealed a higher genetic diversity in group 1 than in group 2 while group 3 is significantly different from the other two groups, with especially high variation in bean chemical contents and cup quality (Anthony et al., 2007b).

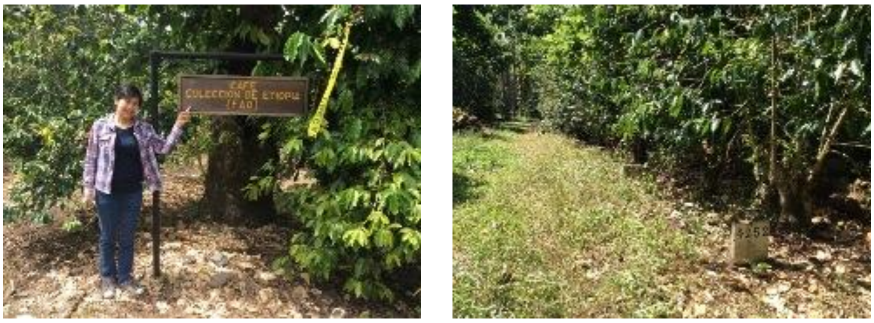
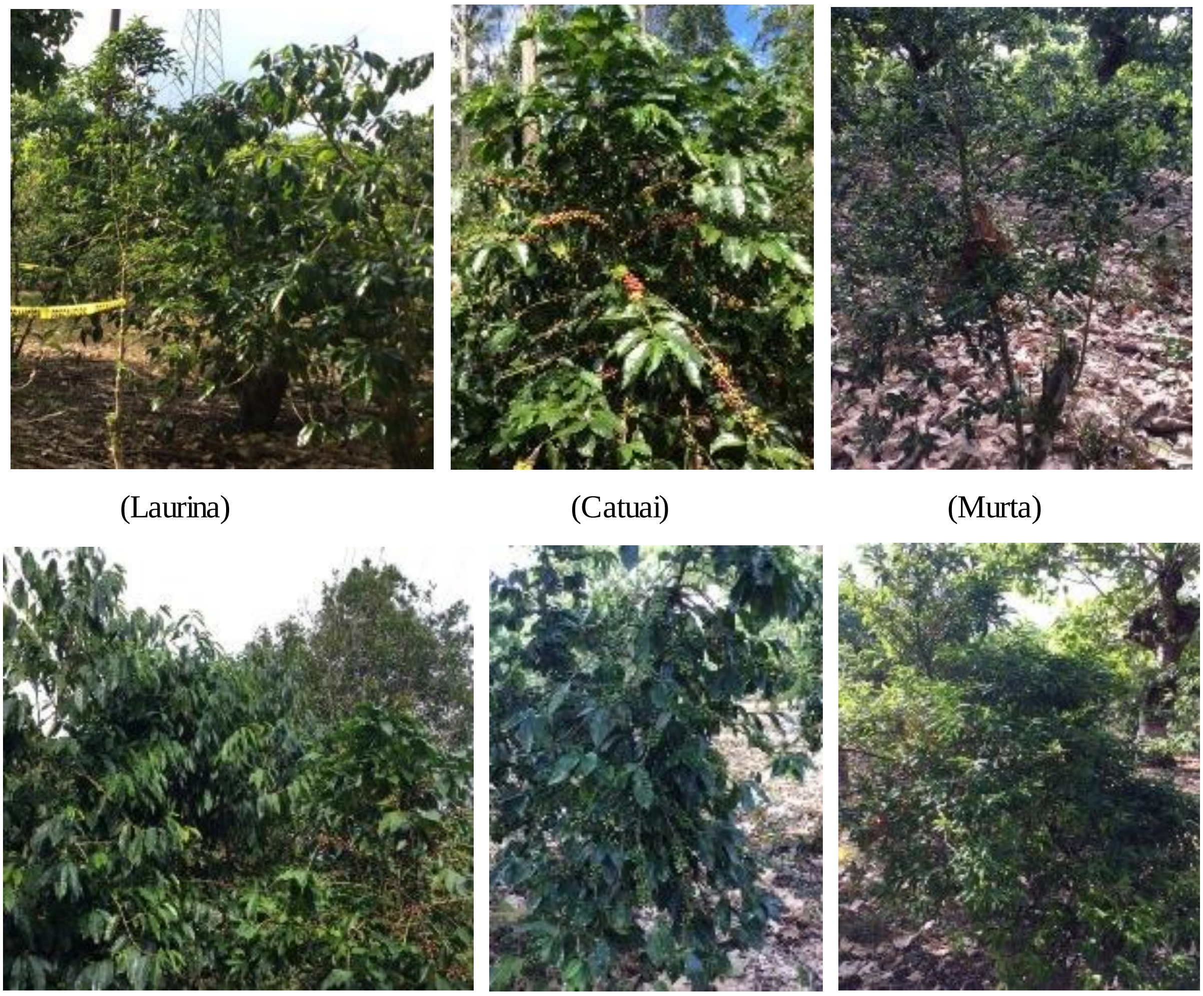


Figure 4.1 Coffee collected from Ethiopia by FAO and the concrete board to mark genotypes

4.2.1.2 The selection of germplasm

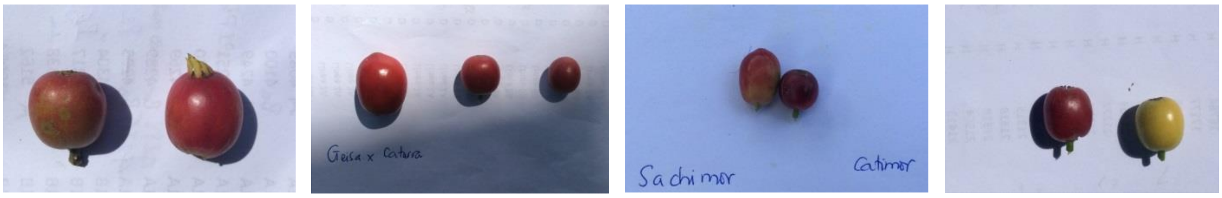
Materials for the study were selected based on these criteria: (1) they are listed in core collection, (2) they are known for biochemical compounds or quality, (3) they came from different coffee growing countries, (4) priority is given to wild type Ethiopia and hybrids, and (5) their fruits are available at the collection time. A large diversity in tree morphology such as leaf, canopy, branch and fruit appearance was observed at the collection field site (Figure 4.2 and 4.3).



(Laurina) (Catuai) (Murta)

(Wild type Ethiopia) (San Ramon) (Mokka Hybrid)

Figure 4.2 Example of tree morphological diversity in the coffee germplasm collection at CATIE, Costa Rica



(1) (2) (3) (4)

Figure 4.3 Example of diversity in fruit morphology: (1) common Arabica and Goiaba (2) Geisa and Caturra (3) Sachimor and Catimor (4) Red and yellow Catuai

The selection of leaf and fruit followed these consecutive steps: (1) located the accessions, (2) selected the tree within accession that was bearing fruits, (3) marked the selected tree with colour band and a sign with name and code, and (4) collected leaf and fruit samples from the selected tree.

In total, 235 accessions originated from 27 countries were selected for this study (Table 4.1 and Appendix Table S4.1). Among them, 232 accessions are *C. arabica*, two *C. canephora* and one *C. brevipes*. The arabica collection was divided into three groups as showed in Table 4.2.

Table 4.1 Coffee accessions used in the study and their countries of origin

No	Countries	Total accessions available at CATIE	Accessions collected for this study
1	Brazil	293	19
2	Cambodia	1	1
3	Cameroon	47	7
4	Sri Lanka	3	2
5	Colombia	36	7
6	Congo	8	2
7	Belgian Congo	43	3
8	Costa Rica	291	16
9	Cuba	2	1
10	Ecuador	4	2
11	El Salvador	24	5
12	Ethiopia	517	115
13	Indonesia	73	10
14	Guatemala	28	12
15	India	24	2
16	Kenya	20	4
17	Madagascar	1	1
18	Malawi	24	2
19	México	26	2
20	Panama	1	1
21	Portugal	163	5
22	Puerto Rico	60	8
23	Sudan	1	1
24	Taiwan	2	2
25	Tanzania	24	3
26	The US	12	1
27	Venezuela	2	1
	Total	**1730**	**235**

Table 4.2 Constituents of the coffee germplasm collection

Species	Group	Types	Quantity	%
C. arabica	1	Wild type from Ethiopia	112	47.66
	2	Variety	42	34.89
		Cultivar	3	
		Selection	29	
		Natural mutant	8	
	3	Hybrid	38	16.17
C. canephora			2	0.85
C. brevipes			1	0.43
Total			**235**	**100.00**

4.2.2 Methods

4.2.2.1 Bean collection, processing and storage

Fruits were collected from one single tree per accession. Ripe cherry (as the mature fruit is called) was harvested by hand, and then taken to a processing house for pulping. After removing the pulp, fresh coffee bean was stored in nylon (PE) bags with water (due to the small amount of coffee bean) for the removal of mucilage. After 12 hours, the fresh coffee bean was taken out of the plastic bag and washed with clean water. Then the fresh cleaned bean was dried in the shade and open air for 1-2 weeks. The parchment was then de-husked. The bean was measured for the moisture content using a MD1229 SEEDBURO 1200S at 12^{o}C, then placed in a hermetic box and shipped to the University of Queensland (UQ), Australia. Bean was stored in the refrigerator at 4^{o}C prior to measurement, roasting and chemical analysis.

4.2.2.2 Bean physical measurement

Bean physical characteristics were measured by applying the method developed by IPGRI (1996) with modification for the weight of 100 beans as follows:

- Weight of 100 beans at 12% moisture content = [(1-x%)/(1-12%)] * weight of 100 beans calculated at x% moisture, measured in triplicate.

- Bean length and width (ratio of length/width), and bean thickness: Measured by the calliper in cm at the longest, widest and thickest parts, respectively, and in quintuplicate.

- Bulk density of green bean (in duplicate): measured by weighing the dried beans in a 100-ml cylinder, and then converted to kg/m^3.

- Moisture of green and roasted bean (in duplicate): measured by Coffee moisture tester named Firmware 2.10 FX (EEPROM 8150 v 1.1 - Coffeelabequipement.com) and also by oven set at 100^{o}C for 22 hours with 2 g of coffee bean.

4.2.2.3 Non-volatile analysis

4.3.2.3.1 Sample preparation

In total 235 samples of green bean and 38 samples of roasted bean were extracted in duplicate. Samples were extracted using a method based on that of Casal et al. (1998) with minor modification. Approximately 4.5 g of roasted or green coffee bean samples, which were stored in the pouches at -20°C (roasted beans) or the fridge at 4°C (green beans), were subjected to milling using Retsch Mixer Mill 4000, frequency 30/s for 25 s (roasted beans) and 1.2 min (green beans) (Figure 4.7). Two grams of green (or roasted) coffee powder of each sample was placed in a 100 ml Erlenmeyer flask with a magnetic stirrer and boiled in 20 ml of MilliQ water for 2.5 min, and then allowed to cool for about 2.5 min before transferring the extract into a 100 ml volumetric flask. The boiling cycle was repeated two more times and extracts mixed together in the volumetric flask and diluted to a final volume of 100 ml with MilliQ water. The extracts was transferred into a 50 ml tube and then cooled down by placing in a -20°C freezer for 12 min before filtered into HPLC vials using 0.13 µm syringe filter. The cooled extract was first centrifuged at 1,500 rpm (252 *xg*) for 5 min at 4°C before the supernatant was filtered. The filtered extracts were stored in HPLC vials at -20°C for further analysis. Analyses were conducted in duplicate with separated bean samples.

4.2.2.3.2 Instrumental Conditions

Samples were analysed on a HPLC/diode-array detector system. The HPLC conditions followed those of Casal et al. (2000) using a Spherisorb S5 ODS2 (0.46 x 25.0 cm) column, with a µBondapak C18 (10 µm) guard column (Waters), and a Shimadzu chromatograph (modelLC-10 VP). The temperature of the column and the autosampler was 23°C and 4°C, respectively. The injection volume was 20 µl with a gradient as follows: (A) phosphate buffer (KH_2PO_4) 0.1 M (pH 4.0), (B) methanol at 0 min (7% B), 4 min (9% B), 6 min (25% B), 13 min (29% B), 21 min (50% B), 30 min (7% B) at a flow rate of 1ml/min. Detection in both cases was achieved using a diode-array detector at 265 nm for trigonelline and 273 nm for caffeine (Figure 4.5). The compounds under study were identified by chromatographic comparisons with authentic standards (caffeine and trigonelline hydrochloride standards were from Sigma-Aldrich, Sydney, Australia). Analytical grade methanol and KH_2PO_4 were also obtained from Sigma-Aldrich (Sydney, Australia).

Quantification was based on an external standard curve. Four calibration points were used to create a linear calibration curve in order to quantify the compounds (caffeine: 0.05-500 µg/ml; trigonelline: 0.15-450 µg/ml) (Casal et al., 1998). The correlation coefficient between the external standard concentrations and absorbance values for each standard curve invariably exceeded 0.999 (Figure 4.4).

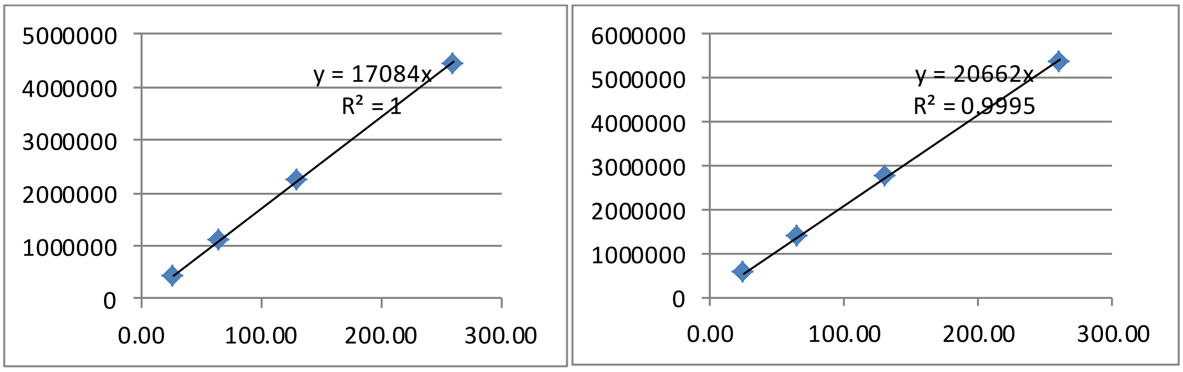


Figure 4.4 Calibration curves for trigonelline (left) and caffeine (right)

(The X and Y axes represent external standard concentrations and absorbance values, respectively).

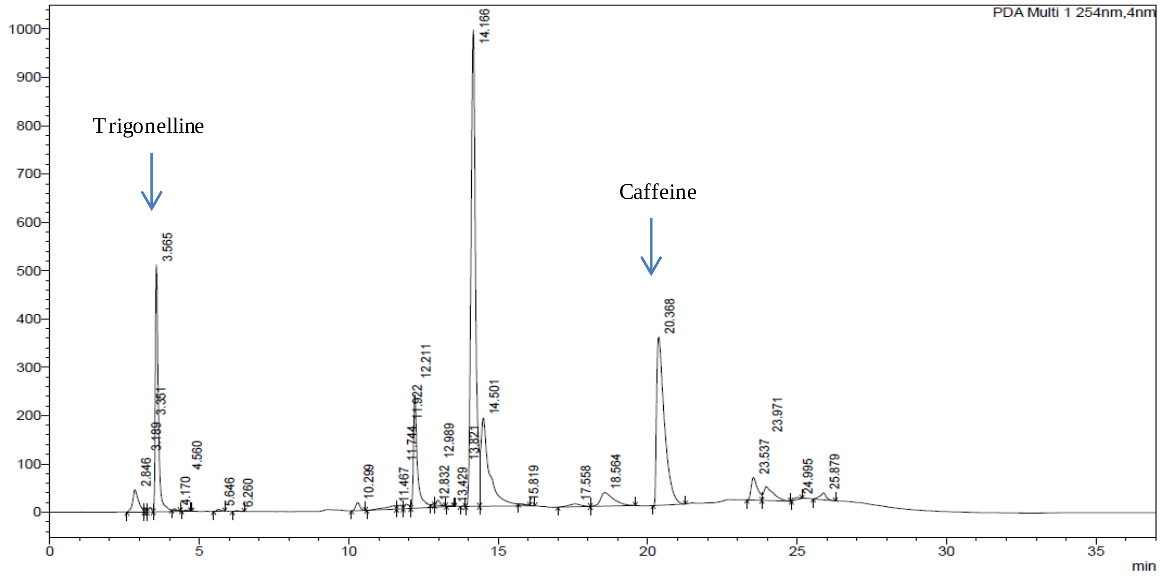


Figure 4.5 Trigonelline and caffeine detection at the absorbance of 265 nm and 273 nm, respectively

4.2.2.4 Bean roasting

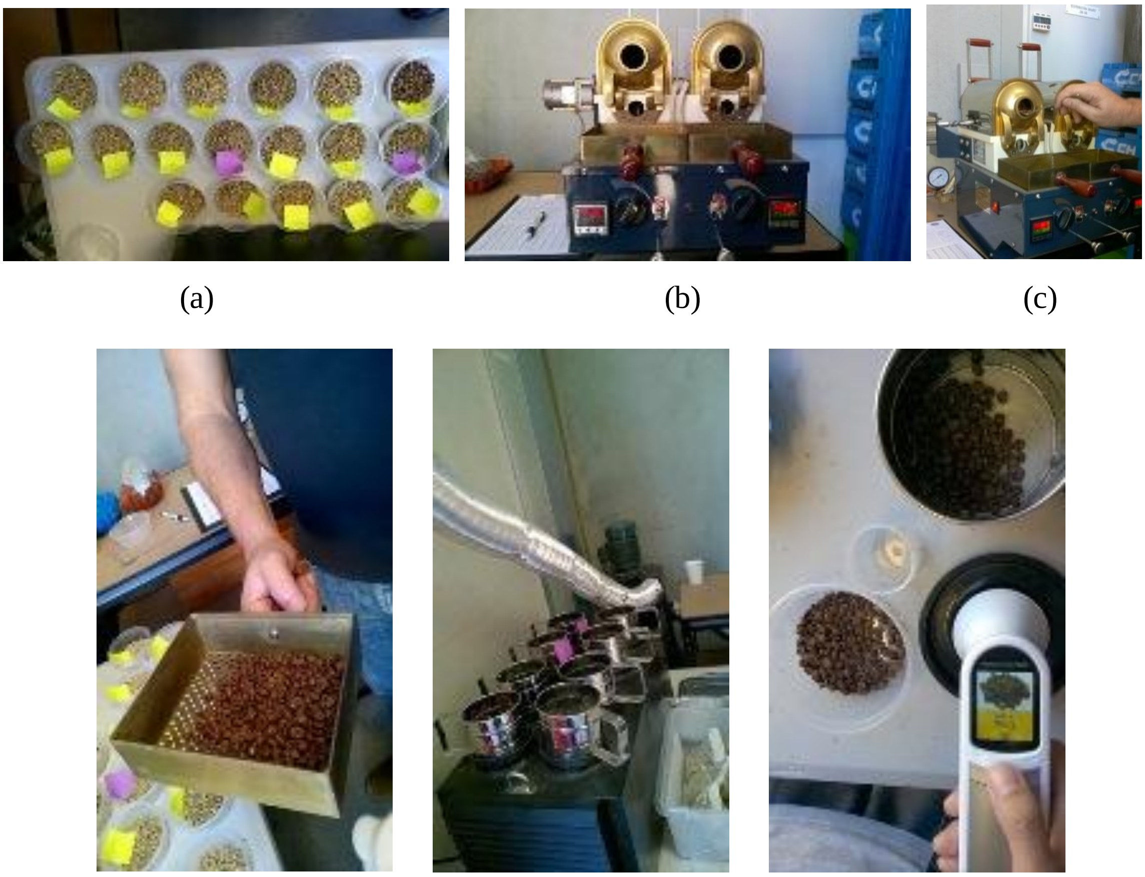
To analyse volatiles, green coffee bean was roasted at Ashtan Place, Banyo, QLD 4014 by Peter Wolff (Wolff Coffee Roasters) who has been at the forefront of the coffee industry in Australia for more than thirty years.

The density and moisture of each green bean sample was measured in order to optimise the temperature and time of roasting to accommodate for the variation in bean size across the samples. In general, each sample was roasted at a starting temperature of 180°C then increased to 188°C, adjusted to 185°C for 6-7 mins total time, and finally cooled for 4 min. The roasting machine used was designed for research scale use and was a Coffee PRO, sample Pro 100 Gas, capacity 2 x 100 g, date of production of Dec 2010 (Nexu International Ltd, coffeetrays.com) (Figure 4.6). Roasted bean was measured for the percentage of weight loss, bean density and especially bean colour.

+ Percentage of weight loss calculated by the subtraction the weight before and after roast.

+ Bean density was measured as for green bean.

+ The colour of the roasted bean was measured using a Roast Analyser – RoAmi, Roaster's Friend (TRUE systems, true-systems.com). The colour range was based on SCAA/Agtron (Agtron, 2004).



(a) (b) (c)

(d) (e) (f)

Figure 4.6 Bean roasting procedure: (a) samples preparation, (b) roaster, (c) checking roasting temperature, (d) bean roasted, (e) cooling system, and (f) bean colour measurement.

Colour, density and % of weight loss of the roasted bean were used to control the consistency among the samples. Coffee bean was subjected to volatile analysis after 24 hours of roasting. The remainder was stored in the aluminium pouches and vacuumed, and then placed in the -20^{o}C for non-volatile analysis (Cheong et al., 2013) at the Health and Food Sciences Precinct, Coopers Plains, Queensland.

4.2.2.5 Non-targeted volatile profile analysis.

4.2.2.5.1 Sample preparation and extraction

The 221 roasted coffee samples were prepared for volatile analysis within 48 hours of roasting. Approximately 4-5 g of fresh roasted coffee beans of each sample were subjected to milling using Retsch Mixer Mill 4000, frequency 30/s for 30 s (Figure 4.7). Two grams of ground coffee was weighted and directly put into a screw capped glass headspace sample vial (20 ml) ready for analysis. The headspace of prepared samples were analysed by solid-phase microextraction (SPME) using a Combi-PAL autosampler (CTC Analytics, Zwingen, Switzerland) controlled by Cycle Composer software (CTC Analytics, version 1.5.2). The SPME fibre was a 50/30 µm divinylbenzene / carboxen/ polydimethylsiloxane (DVB/ PDMS/CAR), StableFlex, Supelco, Bellefonte, PA. Prior to headspace sampling, the vials containing the coffee samples were equilibrated at 60 °C for 20 min. During extraction, the SPME fibre was exposed to the sample headspace for 30 min at 60 °C, and then inserted into the GC inlet heated at 250°C and set to a 1:25 split ratio.

Figure 4.7 Coffee bean miller, samples extracted and GC-MS analysis

4.2.2.5.2 Instrumental conditions

SPME extracts were analysed using a Shimadzu GC-2010 gas chromatograph coupled with a Shimadzu GC-MS-QP2010S mass selective detector (MSD). The system was controlled by Shimadzu GC-MS Solutions software (version 2.53).

The GC column oven was fitted with a DB-Wax capillary column (60 m× 0.25 mm i.d., 0.25 μm phase) Agilent J&W, USA. The carrier gas was helium set to a flow rate of 1.6 mL/min, linear velocity 32 cm/s. The initial oven temperature was 40°C for 1 min, then ramped at 10 °C/min to

250°C and held for 15 minutes. The interface temperature was set to 280°C and the MSD set to scan a *m/z* range of 20 to 300. The ion source was set at 70 eV and electron multiplier at 1350 V.

4.2.2.5.3 Data pre-treatment and analysis

For data analysis, the software GC-MS Solutions (version 4.2, Shimadzu Corporation) and the database NIST21 and NIST107 were used. Two types of data were collected:

Total ion chromatographs (TICs) with full scan (35-350 *m/z*) electron ionisation mass spectral data which was exported as time versus Total Ion Current response from Chemstation to Microsoft Excel.

Extracted Ion Chromatograms were also obtained with mass spectral ions (*m/z*) of specific or particular interest, such as *m/z* 42, 67 and 135 common for many pyrazines, *m/z* for ketones (43) or *m/z* for furans (95) were singularly extracted. In total about 10 extracted ions corresponding to 10 key compounds of coffee were exported to Microsoft Excel. Chemical identification was performed via the respective mass spectrum using the NIST MS library and retention time (Rt).

Using Microsoft Excel, all chromatographic fingerprints were manually trimmed at the beginning (before 3.75 min retention time) and at the end (after 21.4 min retention time) since there was no useful information observable in these regions (Figure 4.14). Due to slight chromatographic shift from run to run, peaks were manually aligned with reference to certain peaks which were identified using the NIST library to ensure the alignments were correct (Figure 4.8). After a pre-treatment of the GC chromatogram data including baseline correction and smoothing signal, data was subjected to principal component analysis (PCA).

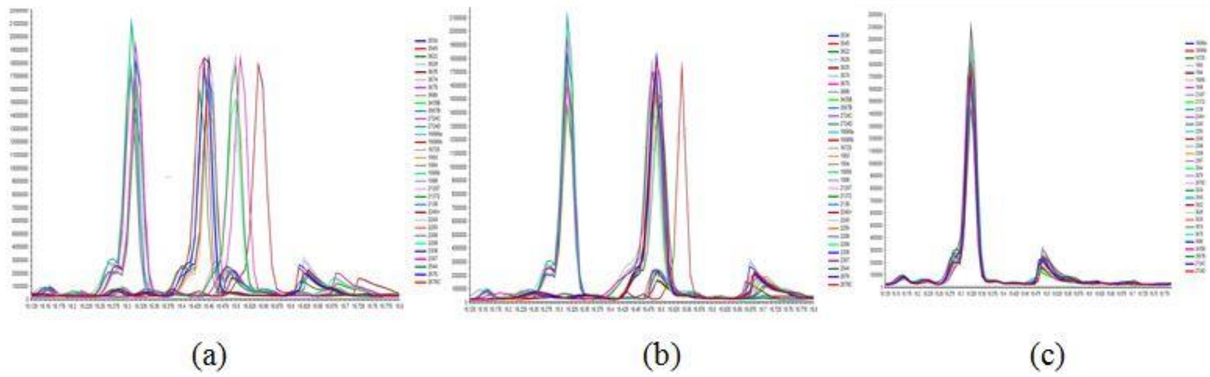


Figure 4.8 Data smoothing: (a) original data (b) data after using Correlation Optimized Warping

(COW) (c) data after manual alignment/smoothing.

4.2.2.6 Targeted volatile profile analysis

35 samples showed contrasting characters for principal component 1 and 2 with some outlying wild type genotypes based on chromatographic fingerprints (total ion chromatographs -TICs) were selected for targeted analysis (SIDA combined with HS-SPME/GC-MS).

The targeted method involved SIDA together with HS SPME/GC-MS for 18 important volatile flavour compounds. The details of the extraction method, instrumental conditions, and calibration parameters of the method applied are described in detail by Sunarharum (2016).

4.2.2.6.1. Sample and standard preparation and extraction

The 35 roasted bean samples were stored at -20°C until prepared for analysis. Samples were ground using a Retsch Mixer Mill 4000 with a frequency of 30/s for 30 s and weighed (2 g) in triplicate into a screw cap glass sample vial together with 2 ml of saturated brine (NaCl), 2 ml milliQ water and a magnetic stirrer flea. Similarly, spiked standard addition calibration solutions were prepared and weighed (2 g) into vials together with brine, water and a stir flea. A spike of internal standard solution was added which contained isotope analogues of the target volatiles. Sample extraction was conducted using a Gerstel MPS2 Autosampler (Gerstel, Germany). SPME fibre type, extraction time and temperature was as described by Sunarharum (2016).

4.2.2.6.2. Development of calibrations and instrumental conditions

Instrumental analysis was performed using an Agilent 6890N gas chromatograph (GC) and an Agilent 5975 series Mass Selective Detector (MSD) unit (Agilent Technologies Inc., California, USA) with a MPS-2 multipurpose sampler (Gerstel, Germany) installed according to the instrumental parameters and conditions described by Sunarharum (2016). All analyses were performed in triplicate. A four-point calibration function was developed for each target compound.

Quantitation was achieved using selected ion monitoring and qualifying ions were used to identify target compounds, together with matching retention times, with authentic reference standards. Co-elution with other compounds was common as coffee has a great number of volatile compounds; however, there was no co-elution observed for target ions used for quantification of each compound. The data were collected using the Enhanced ChemStation software MSD ChemStation G1701EA revision E.02.02 and compound concentration data were exported into Excel prior to analysis.

The analysis of volatile data was summarised in the following diagram (Figure 4.9)

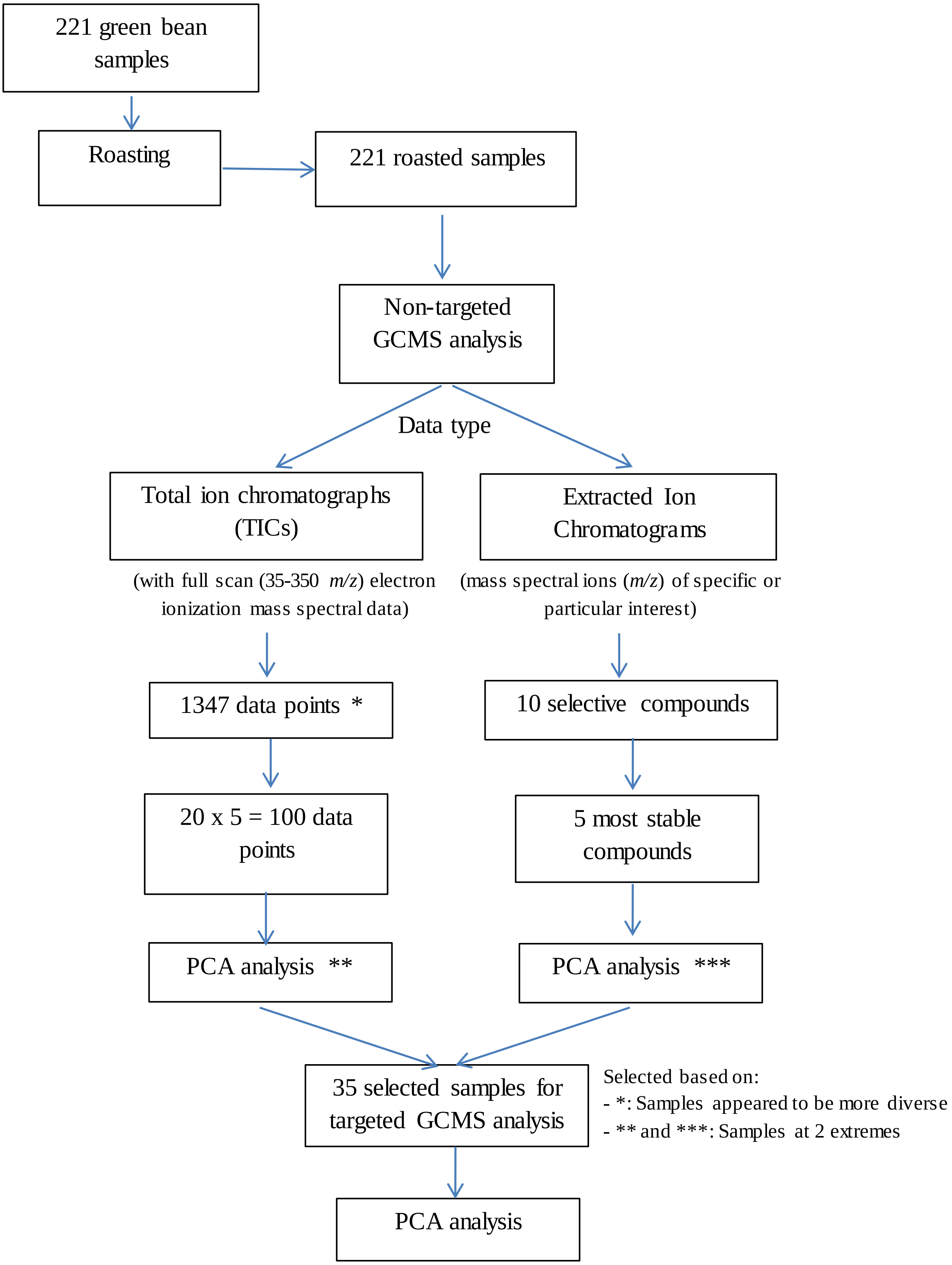

221 green bean samples
Roasting
221 roasted samples
Non-targeted GCMS analysis
Data type
Total ion chromatographs (TICs)
(with full scan (35-350 m/z) electron ionization mass spectral data)
Extracted Ion Chromatograms
(mass spectral ions (m/z) of specific or particular interest)
1347 data points *
10 selective compounds
20 x 5 = 100 data points
5 most stable compounds
PCA analysis **
PCA analysis ***
35 selected samples for targeted GCMS analysis
Selected based on:
- *: Samples appeared to be more diverse
- ** and ***: Samples at 2 extremes
PCA analysis

Figure 4.9 Analysis of volatiles in roasted coffee bean using non-targeted and targeted analytical approaches.

(*) Samples that appeared to be more diverse were selected (see Figure 4.15 left, samples in circle); (**) Samples at 2 extremes were selected (see Figure 4.16, samples in circle); (**) Samples at 2 extremes were selected (see Figure 4.17, samples in circle).

4.2.2.7 Data analysis

One-way analysis of variance (ANOVA) was performed using the software GenStat version 11 (Payne et al., 2008) using category of accessions (cultivars, hybrids and wild) as a factor. Mean values between two groups were compared using linear contrasts. Spearman rank correlations was used to examine the relationship between variables after testing for normality using Shapiro-Wilk test.

For each variable, the value of measurement for each individual was standardised by subtracting it from the population mean and dividing by the standard deviation in order to reduce the influence of the scale differences. The standardised data were then used in PCA performed with the software XLSTAT (Addinsoft, 2007) for targeted analysis in which 18 variables (18 volatile compounds) and 35 samples were included. For non-targeted analysis (supplementary data), due to the large data set (100 selected chromatographic variables and 221 samples), principal component analysis (PCA) multivariate exploration was conducted using The Unscrambler X, version 10.3 (Martens et al., 1987).

4.3 Results and discussion

4.3.1 Variation in physical characters of green bean and its changes after roasting

4.3.1.1. Variation in green bean physical characteristics

Physical characteristics of green coffee bean have been reported to affect beverage quality to some extent. In this study green bean physical characteristics and coffee quality of 232 arabica coffee.

Measurement of 232 arabica accessions for green bean physical characteristics showed that 100 bean weight (W100), length, width, ratio between length and width, thickness and bulk density were on average 15.53 g, 9.62 mm, 6.83 mm, 1.41, 3.95 mm and 600 kg/m^3, respectively. W100 and bean length had the largest variation, with the coefficient of variation (CV) of 12.20% and 7.73%, respectively (Table 4.3). There was a significant difference ($P < 0.001$) between accessions for all variables. Although the three arabica groups had similar mean weighs of 100 beans (15.35, 15.80 and 15.57 g), the hybrid group comprising both intraspecific and interspecific hybrids had the largest variation (CV = 15.54%), followed by the cultivar group (CV = 12.50%) and the least variable was the wild type group (CV = 10.69%) even though this group accounts for almost half of the population. This group was also the least variable for length (6.99% vs 8.33 and 8.34%) and width (5.01% vs 5.27 and 6.13%). For thickness and bulk density, hybrids had the least variability. However, the difference between groups was not significant for W100, length, and bulk density. For the ratio between length and width, width and thickness, the wild type group was significantly different to the cultivar and hybrid groups ($P < 0.05$), while there was no significant difference between the cultivar and hybrid groups. The wild group had slimmer beans and a greater thickness than the cultivars and hybrids (Table 4.3).

Table 4.3 Variation in green bean physical quality measured in 232 arabica coffee accessions.

Variable	Min	Max	Mean	CV (%)	LSD $_{0.05}$	Literature
W100 (g)	10.16	23.13	15.53	12.20	0.54	9.77 - 21.82 (1)
+ cultivar (n=83)	10.16	23.13	15.35 a	12.50		
+ hybrids (n = 36)	12.17	22.71	15.80 a	15.54		
+ wild type (n = 113)	11.87	19.88	15.57 a	10.69		
Length (L) (mm)	6.64	13.20	9.62	7.73	1.12	8.19 - 11.04 (2)
+ cultivar (n=83)	6.64	13.20	9.50 a	8.33		
+ hybrids (n = 36)	8.32	12.52	9.75 a	8.34		
+ wild type (n = 113)	8.01	12.05	9.66 a	6.99		
Width (W) (mm)	5.97	8.08	6.83	5.43	0.59	6.11 - 8.27 (3)
+ cultivar (n=83)	5.97	8.07	6.91 a	5.27		
+ hybrids (n = 36)	6.22	7.92	6.94 a	6.13		
+ wild type (n = 113)	6.12	8.08	6.74 b	5.01		
Ratio L/W	1.11	1.75	1.41	7.68	0.15	1.33 – 1.35 (4)
+ cultivar (n=83)	1.11	1.67	1.38 b	7.13		
+ hybrids (n = 36)	1.18	1.65	1.41 b	7.58		
+ wild type (n = 113)	1.22	1.75	1.44 a	7.45		
Thickness (mm)	3.27	5.18	3.95	7.04	0.49	4.60 - 5.13 (5)
+ cultivar (n=83)	3.27	4.84	3.88 b	7.54		
+ hybrids (n = 36)	3.41	4.49	3.87 b	5.80		
+ wild type (n = 113)	3.42	5.18	4.01 a	6.66		
Bulk density (kg/m^3)	454	679	600	4.57	20.86	635 - 707 (6)
+ cultivar (n=83)	458	675	603 a	5.24		
+ hybrids (n = 36)	566	664	605 a	3.76		
+ wild type (n = 113)	505	679	597 a	4.54		

Values with the same alphabet along the same column are not significantly different (P>0.05). CV (%): coefficient of variation. LSD $_{0.05}$: Least Significant Difference at P value less than 0.05. Ranges of phenotypic values reported from literature: (1) Belete, 2014; Bertrand et al., 2005; Montagnon and Bouharmont, 1996; (2) (3) (4) (5) Olukunle and Akinnuli, 2012; Ghosh and Gacanja, 1970; (6) Franca et al., 2005; Rodrigues et al., 2003.

In this study green bean physical characteristics and coffee quality of 232 arabica coffee genotypes was assessed and the relationship between these attributes was determined. For the physical bean quality, the weight of 100 beans is used as an indicator of bean size. Alternatively, the bean size can be measured directly on the sieve in industry. Coffee with larger beans usually gets a good grade and fetches a higher price than coffee with smaller beans even though the former do not necessarily produce a more desirable roast or liquor (Belete, 2014).

The large variation in bean size in the population studied based on visual assessment and weight of 100 beans (W100) (Figure 4.10 and Table 4.3) indicated the potential for selection in breeding. In general, the population used in this study had a smaller average size compared to those in previous studies, but the range of W100 and bean length was larger than that reported in the literature. Previous studies showed that the weight of 100 beans of arabica hybrids and maternal lines ranged from 17.00 - 20.01g (Bertrand et al., 2005). A more diverse population of 148 accessions from Ethiopia showed a mean 100 bean weight of 2 groups of 16.17 - 16.63 g with a minimum of 12.30 g and a maximum of 20.8 g (moisture content of 10%) (Montagnon and Bouharmont, 1996). A larger variation in 100 bean weight was observed for 30 accessions at 4 locations in Ethiopia with average of 13.12 to 19.13 with minimum of 9.77 g and maximum of 21.82 g (at 11% moisture content) (Belete, 2014). With a lower limit for W100 of 10.16 g and upper of 23.13 exceeding other studies at both ends of the spectrum, this study has captured a greater range of diversity in bean weight than previous, smaller investigations. Accessions with exceptionally large bean size identified in the current study could be a useful source for genetic improvement of arabica physical quality.

Bean length, width, length/width ratio, and thickness reflect the bean shape and also contribute to the 100-bean weight. In a previous study involving 52 samples collected from Kenya, the average bean length, width and thickness were reported as 11.04 mm, 8.27 mm and 5.13 mm, respectively (Ghosh and Gacanja, 1970). Similarly, Olukunle and Akinnuli (2012) also reported the measurement for beans collected from Nigeria with average length, width and thickness of 8.19 mm, 6.11 mm and 4.60 mm, respectively. The current study involving a much larger population exhibited smaller mean size in each of these dimensions.

Density is a parameter that relates to bean weight as well as the roasting properties. The population studied had relatively low density compared to previous studies (Franca et al., 2005; Rodrigues et al., 2009). This may be the result of the higher moisture content of studied samples than that of the standard which relates to bean weight and volume.

Among the three groups of arabica, the hybrids had the largest average bean size and the most variable size (Table 4.3), which could be a result of cross breeding from both parents having large bean size, or selection of hybrid population from a subset of the progenies of the cross which was selected based on large bean size. The groups of cultivars and wild relatives had almost the same value for bean size of 15.35 g and 15.57 g while the wild groups included more accessions and would be expected to include more variation. However, among three groups, the differences are not significant for weight of 100 beans, length and bulk density. For width and thickness which relates to bean size and ratio of L/W which relates to bean shape, the wild type group was significantly different from cultivar and hybrid groups suggesting its potential use in selection for bean size and shape in quality breeding. Furthermore, the mean L/W ratio in our study was larger than reported in the literature (Table 4.3), indicating a more elongated bean shape was observed when a large diverse sample population was examined.

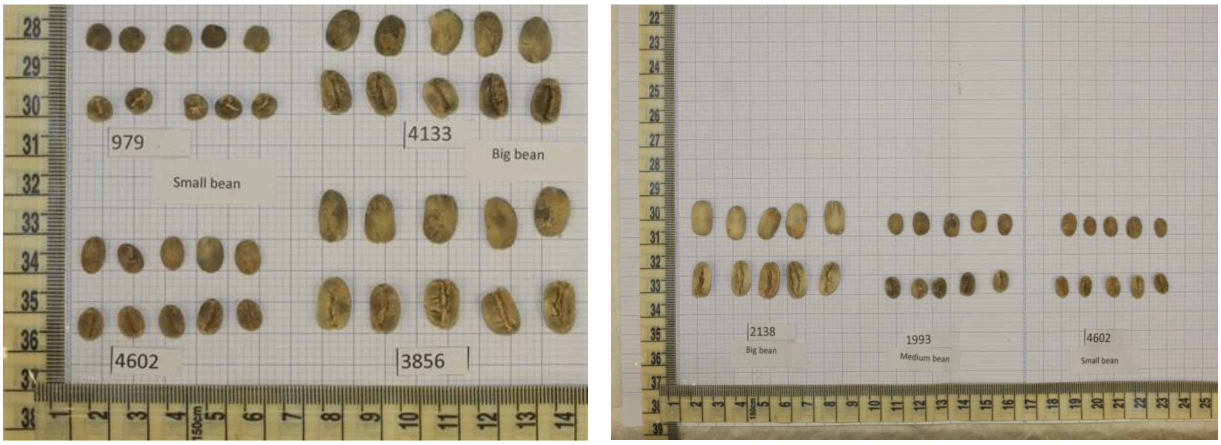


Figure 4.10 Variation in bean size: small bean vs large bean (left); different sizes of bean from small-medium to large.

(numbers are the code of the samples: 979 - W100 = 10.16 g; 4602 - W100 = 12.81 g; 4133 - W100 = 19.52 g; 3856 - W100 = 22.71 g; 2138 - W100 = 23.13 g; 1993 - W100 = 16.37 g)

The green bean size could be consider an important criterion for coffee plant selection aiming to improve the green bean quality since the evaluations and comparisons have occurred in the same environmental conditions and in the same post-harvest processing procedures.

4.3.1.2 Bean roasting and the change of bean physical properties after roast

The population studied has much higher bean moisture compared to that required by the coffee industry with average of 13.97% (Table 4.4). The majority of the samples have moisture of 14 – 15%. The reason may be that after getting bean from Costa Rica, it was stored in the fridge at 4°C waiting for further analysis. While this may affect the concentration as well as the profile of compounds in coffee, it is expected that the quality of bean may not be affected at this temperature.

After roasting, the 42 samples' bean moisture content reduced to 1.7% on average with range of 1.11-2.37% (Table 4.4) which was very close to that reported by Campos et al. (2014), ranging between 1.5 and 2.2%.

Bulk density of green bean was also measured to modify the roasting profile, the higher density the longer time of roast as high density beans are more resistant to heat (https://bootcampcoffee.com/coffee-density-roasting-strategy/). Variation in bulk density in the population of 235 genotypes was quite high, ranging between 458 – 679 kg/m^3 (Table 4.4). The accessions named Arabigo Puerto Rico had the lowest bean density (458 kg/m^3) while a wild accessions from Ethiopia called E-190 was the highest accession in bean density of 679 kg/m^3. Only 221 genotypes were roasted, with the mean bulk density after roast of 482 kg/m^3, reducing 21% compared to bulk density of green bean. While the bulk density of the green bean was lower than reported in the literature, the bulk density of roasted bean was much higher (Table 4.4). This may be the result of high moisture content in the green bean. According to Franca et al. (2005), bean density was higher for the soft (high-quality) sample and lower for the rio (low-quality) sample. The lowest quality sample (rio) presented the highest values of density after roasting and the lowest increase in volume. This parameter should be taken into account to see if there is any link between this type of bean property with bean biochemical compounds.

Table 4.4 Changes of bean properties after roast

Bean moisture								
Green bean (235 samples)						Roasted (42 samples)		
Mean (%)	SD	Range (%)	Frequency			Mean (%)	SD	Range (%)
			<13%	13-15%	>15%			
13.97	0.66	9.53-15.65	9	209	14	1.7	0.30	1.11-2.37

Bulk density									
Green bean* (232 samples)				Roasted bean (221 samples)					
Mean kg/m^3	SD	Range kg/m^3	Literature [1]	Mean kg/m^3	SD	Range kg/m^3	Literature [2]	% reduce
600	28.19	458 - 679	635 - 707	482	29.83	400 - 590	250 - 369	21

Weight loss (221 samples)					
Mean (%)	14%	16%	18%	20%	22%
17	8	79	98	35	1

Bean size				
	No of samples	Length (mm)	Width (mm)	Thickness (mm)
Green bean	232	9.62	6.83	3.95
Roasted bean	42	11.11	8.26	4.98
% of increase		13.48	15.53	20.60

Bean colour (221 samples)									
Degree of roast**	Undeve-loped	Extra light	Very light	Light	Med light	Med	Med dark	Dark	Very dark
Colour range	100	95	80-90	70-80	60-70	50-60	40-50	30-40	25
Frequency	0	3	31	122	63	2	0	0	0

* bulked density for green bean was standardised at 12% moisture; ** see appendix Table S4.2 for more details

[1] & [2]: Franca et al. (2005) & Rodrigues et al. (2003). (SD: standard deviation); Green bean is the bean before roasting; Roasted is the bean after roasting; Number in brasket is the number of samples measured.

After roast, percentage of weight loss of roasted bean, density of roasted bean and bean colour were measured to make sure the samples were roasted consistently and in the range of the industry requirement. For the population studied, the average weight loss of roasted bean was 17%, in which 80% of samples had weight loss of 16-18%, 8 samples have 14% and one sample has 22% of weight loss. Even though the moisture of the green bean was higher than standard, the weight loss of roasted bean was in the range of that of industry which is from 14-23% (Table 4.4) and other studies (Campos et al., 2014; Gloess et al., 2014). These studies also found that weight loss of roasted bean was higher for darker roast degrees or with the increase of roasting temperature.

Measurement of 42 samples show that the increase in bean size was from 13.38% for length, 15.53% for width and 20.60% for thickness. In the study of Gloess et al. (2014), the gain in volume of the beans ranged from 43% to 82%, being higher for darker roast degrees and lower for longer roasting times.

One of the measurements reflecting the consistency of roast is colour of roasted bean. Bean should be roasted from light to medium to develop proper flavour. In this study, almost all samples (97.74%) were roasted from very light to medium light corresponding to colour range of 60 to 90 or Agtron tiles of 65 to 85 indicating a comparable roasting of the samples had been achieved (Figure 4.11). Only two samples were with medium roast and three samples with extreme light roast (Table 4.4)

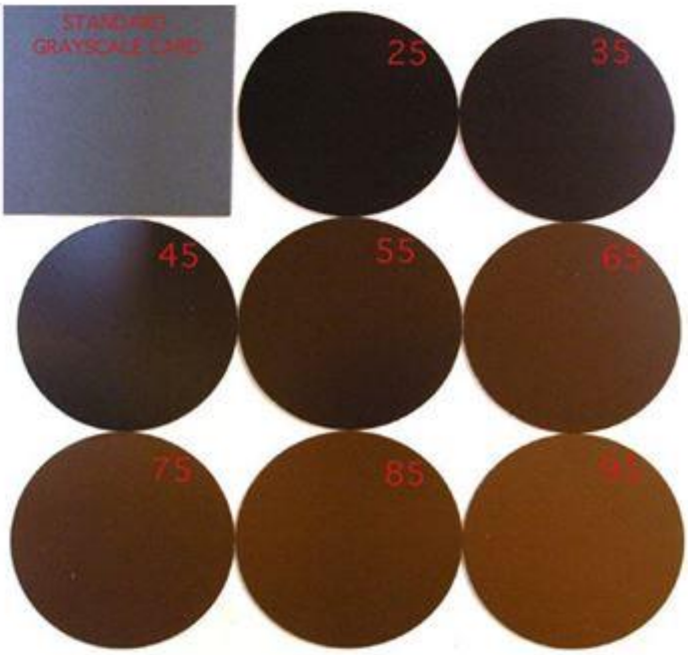


Figure 4.11 Degree of roast and corresponding colour

Source: https://legacy.sweetmarias.com/library/content/using-sight-determine-degree-roast (based on Agtron Roast Color Tiles)

4.3.2 Variation of non-volatile compounds in 232 arabica accessions

4.3.2.1 Variation of caffeine and trigonelline in green bean.

The green bean caffeine content of the 232 arabica accessions was on average 1.25% on a dry matter basis (d.m.b), ranging from 0.82% to 1.76%, while the trigonelline content was 1.11% on average, ranging from 0.80% to 1.38% . The coefficient of variation was 9.64% for caffeine and 9.55% for trigonelline (Table 4.5). Significant genotypic variation (P < 0.001) in caffeine contents was observed among accessions (Table 4.5); they appeared to follow a normal distribution that was slightly skewed towards the lower value (Figure 4.12). The cultivar group had a significantly (P < 0.05) lower caffeine content than the hybrid and wild groups (Table 4.5). With regards to within-group variation, the hybrid group was the most diverse in caffeine (1.01 – 1.76%, CV = 11.29%) despite having the smallest number of accessions (38), followed by the cultivar group (0.82 – 1.52%, CV = 9.18%) and the wild type group from Ethiopia (1.18 – 1.51%, CV = 9.15%) (Table 4.5). For trigonelline, however, there was not a significant difference (*P* < 0.05) between the three groups.

Table 4.5 Variation in caffeine and trigonelline content among different arabica coffee groups

Caffeine (% d.m.b.)								
Species	**Group**	**Types**	**Size**	**Min**	**Max**	**Mean**	**CV%**	**LSD $_{0.05}$**
C. arabica	1	Variety/ selection	83	0.82	1.52	1.23 b	9.18	0.13
	2	Hybrid	36	1.01	1.76	1.28 a	11.29	0.07
	3	Wild type	113	1.18	1.51	1.26 a	9.15	0.11
		Total/Average	**232**	**0.82**	**1.76**	**1.25**	**9.64**	**0.12**
C. canephora			2	1.19	2.48	1.84		
C. brevipes			1			2.37		
Trigonelline (% d.m.b.)								
Species	**Group**	**Types**	**Size**	**Min**	**Max**	**Mean**	**CV%**	**LSD $_{0.05}$**
C. arabica	1	Variety/ selection	83	0.89	1.38	1.12 a	9.66	0.11
	2	Hybrid	36	0.89	1.31	1.13 a	9.69	0.10
	3	Wild type	113	0.94	1.31	1.10 a	9.46	0.10
		Total/Average	**232**	**0.80**	**1.38**	**1.11**	**9.55**	**0.10**
C. canephora			2	0.91	1.24	1.08		
C. brevipes			1			0.76		

Values with the same alphabet along the same column are not significantly different (P>0.05). CV (%): coefficient of variation. LSD $_{0.05}$: Least Significant Difference at P value less than 0.05.

Although caffeine and trigonelline content in coffee have been analysed in a number of reports, this is the first study using a relatively large arabica population of 232 genotypes with many wild accessions. Most other studies involved fewer accessions with different sample sizes and species and focused more on caffeine than on trigonelline. Previous studies showed caffeine content ranging from 0.62 – 1.82% (Anthony et al., 1993; Avelino et al., 2005; Ky et al., 2001b; Martín et al., 1998; Mazzafera and Carvalho, 1992; Mehari et al., 2016; Taveira et al., 2014) which is very similar to the current study (0.82 – 1.76%). The study by Silvarolla et al. (2000) showed a substantial variation from 0.42 – 2.90% but included a very large population of inter and intra hybrid progenies with 499 plants from 68 progenies (Kaffa region of Ethiopia) and 166 plants from 22 progenies (Illubabor region of Ethiopia). For trigonelline, the current study showed smaller variation (0.80 – 1.38%) compared to previous studies, especially that reported in the study by Mazzafera (1991) (1.52 – 2.90%). However, the current study gave results that were very close to those of the recent study by Mehari et al. (2016) using 99 arabica coffee samples from eight varieties from Ethiopia. The differences could be genetic or due to different methods of extraction and/or analytical instrument used in different studies. For caffeine, the significantly lower concentration in the cultivars relative to the hybrid and wild type groups indicated the result of selection for lower caffeine in the cultivar group.

The demand for decaffeinated coffee is increasing and now accounts for 10% of the total coffee consumed in the world. Selection and breeding for low caffeine content is thus a new target of the coffee industry. Accessions that had low caffeine content in this study may serve as a potential source of desirable genes for development of varieties with low caffeine content.

For breeding, varieties with low caffeine and high trigonelline should be a target. Among the 232 accessions studied, the 10 lowest in caffeine content accessions were mostly from the variety group (6 accessions), only two were from wild type group and two from the hybrid group (Table 4.6). Laurina, a natural dwarf mutant of *C. arabica* cv. 'Bourbon' and controlled by lrlr alleles, which is very famous for low caffeine (Baumann et al., 1998; Joët et al., 2010; Mazzafera and Carvalho, 1992), was found in the lowest caffeine group together with two other accessions from wild types. Maragogype and its hybrid were also in this group with the lowest caffeine. These accessions could be useful in breeding programs aiming to reduce caffeine in coffee. Although very low caffeine content accessions originated from another Ethiopia germplasm collection were found (Benatti et al., 2012), no breeding program would want to be restricted to a single parent line for a desired trait; it makes it difficult to introgress desired traits that may be absent in that parent line, and a diversity of parental lines is always a good strategy. Among the 10 highest in caffeine content, two accessions were hybrids followed by six accessions from wild types and two from the cultivar group.

Table 4.6 Top 10 genotypes of lowest and highest caffeine content of studied population

Lowest content	No	Code	Name	Country	Type	Value
	1	2299	Laurina	Costa Rica	Natural mutant	0.82
	2	4755	E-232	Ethiopia	Wild	0.93
	3	4713	E-190	Ethiopia	Wild	0.98
	4	4276	Mibirizi 49-1848	Congo Belga	Selection	1.00
	5	2676	Laurina	Cameroon	Natural mutant	1.00
	6	3722	Geisha Castañer, Puerto Rico x 54375	Puerto Rico	Hybrid	1.01
	7	3432	Maragogipe rojo Brazil x 47127	Brasil	Hybrid	1.01
	8	2733	S.L 34	Kenya	Selection	1.02
	9	2919	Jackson 2	Congo	Selection	1.03
	10	978	Maragogipe	Guatemala	Variety	1.04
			Average			**0.98**
Highest content	**No**	**Code**	**Name**	**Country**	**Type**	**Value**
	1	3622	Mibirizi No.49, Belgian Congo x 51149	Belgium Congo	Hybrid	1.76
	2	3545	No. 31 Etiopia x 48989	Ethiopia	Hybrid	1.55
	3	3956	Babbaka	Ethiopia	Variety	1.52
	4	4875	E-508	Ethiopia	Wild	1.51
	5	4837	E-437	Ethiopia	Wild	1.51
	6	4569	E-146	Ethiopia	Wild	1.49
	7	21315	ET-59	Indonesia	Wild	1.47
	8	4133	Seleccion 353 4/5 CRRC 35/9	Portugal	Selection	1.47
	9	4555	E-482	Ethiopia	Wild	1.47
	10	4494	E-46	Ethiopia	Wild	1.46
			Average			**1.52**

Seven out of the 10 accessions with the lowest trigonelline come from Ethiopia wild types (Table 4.7). The two common varieties, Maragogype (famous for large bean and special quality) and Geisha (excellent reputation for quality) (Montagnon et al., 2012), should be considered as valuable materials for breeding purposes as they possess low caffeine and high trigonelline content. It is interesting to observe that there were fewer wild type accessions found in the groups which were low caffeine and high trigonelline as expected. Accessions that had a low caffeine content (such as Laurina, E-232, or the Maragogype hybrids or Geisha hybrids) or a high trigonelline content (such as Maragogype amarillo, Geisha VC-496, Purpurascens) may serve as a potential source of desirable genes for development of variety types with relatively low caffeine and high trigonelline content. A biological replicate experiment designed for these accessions for further assessment would give a more concrete conclusion.

From the distribution of population, if 10 genotypes of two extremes are selected, the difference between two extremes of lowest and highest in caffeine content (0.98 and 1.52% for caffeine and 0.89 and 1.33% for trigonelline) will be more than 30% (Table 4.5, 4.7 and Figure 4.4). This variation may be sufficient for association genetics.

Table 4.7 Top 10 genotypes of lowest and highest trigonelline content of studied population

Lowest content	**No**	**Code**	**Name**	**Country**	**Type**	**Value**
	1	16729	ET47	Ethiopia	Wild	0.80
	2	4900	E-531	Ethiopia	Wild	0.82
	3	4755	E-232	Ethiopia	Wild	0.85
	4	4685	Limnu E-171	Ethiopia	Wild	0.85
	5	12841	Caturra rojo x Hibrido de Timor (INMC 32-3) Catimor Cenife	Mexico	Hybrid	0.89
	6	4276	Mibirizi 49-1848	Congo Belga	Selection	0.89
	7	4483	E-305	Ethiopia	Wild	0.93
	8	17232	L-334 ET45 C7	Cameroon	Wild	0.94
	9	16690	ET02	Ethiopia	Wild	0.94
	10	16782	Catuaí rojo UFV 2197/194 CH 2077/2 5-72	Brasil	Selection	0.94
			Average			**0.89**
Highest content	**No**	**Code**	**Name**	**Country**	**Type**	**Value**
	1	4080	Maragogipe amarillo	Colombia	Variety	1.38
	2	2722	Geisha VC-496	Tanzania	Variety	1.38
	3	3411	Columnaris	Panama	Mutant	1.38
	4	4133	Seleccion 353 4/5 CRRC 35/9	Portugal	Selection	1.35
	5	986	Purpurascens	Guatemala	Mutant	1.32
	6	4712	E-189	Ethiopia	Wild	1.31
	7	4569	E-146	Ethiopia	Wild	1.31
	8	4497	E-67	Ethiopia	Wild	1.30
	9	4874	E-507	Ethiopia	Wild	1.30
	10	3628	Red tipped, Etiopia x 51356	Ethiopia	Hybrid	1.30
			Average			**1.33**

Environmental effects on caffeine and trigonelline content of coffee beans and genotype x environment interactions have been reported (Avelino et al., 2005; Casal et al., 2000; Figueiredo et al., 2013; Sridevi and Giridhar, 2013; Sridevi and Giridhar, 2014; Taveira et al., 2014). Such environmental effects may explain the difference in the range of concentrations of these non-volatiles observed between the current study and others.

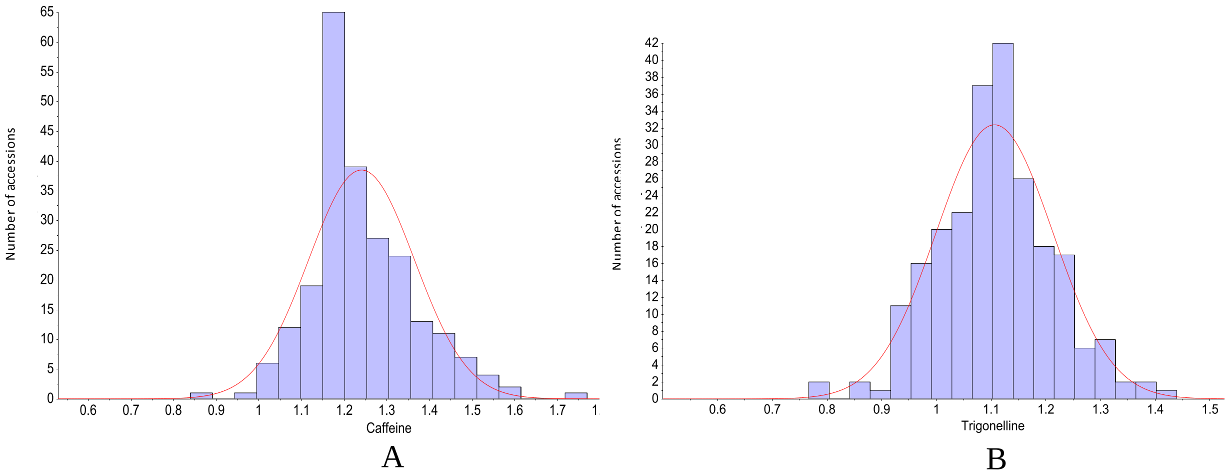


Figure 4.12 The distribution of caffeine (A) and trigonelline (B) in the arabica coffee population (n = 232).

4.3.2.2 The change of caffeine and trigonelline content after roast.

Roasted coffee bean was also subjected to HPLC to examine the change of its caffeine and trigonelline content after roasting. Results are presented in Table 4.8

Table 4.8 Caffeine and trigonelline content before and after roast observed in 36 samples

Caffeine (% d.m.b.)									
Green	**Roasted**	**Reduced (%)**				**Increased (%)**			
		No of samples	Mean	Min	Max	No of samples	Mean	Min	Max
1.26	1.17	27	11.01	1.08	27.14	9	4.09	0.15	10.86
Trigonelline (% d.m.b.)									
Green	**Roasted**	**Reduced (%)**				**Increased (%)**			
		No of samples	Mean	Min	Max	No of samples	Mean	Min	Max
1.05	0.85	36	18.74	1.93	44.03	0	0	0	0

After roasting, 76% samples reduced caffeine content from 1.08 – 27.14% after roast and 23.68% samples had it increased in the range of 0.15 – 10.86%. Previous studies show that caffeine content remains almost unchanged or even slightly increased in percentage due to the loss of other compounds (Farah et al., 2006b; Oestreich-Janzen, 2010) or can be lost 30% (Franca et al., 2005). Trigonelline content was reduced between 1.93 and 44.03% (Table 4.8) which was similarly to previous studies (reviewed by Farah et al., 2006b). However, the reduction is smaller than previous studies (reviewed by Farah et al., 2006b) which was up to 90%. These differences may be attributed to distinct roasting conditions, which include differences in colorimetric standards, since trigonelline degradation was reported to be strongly dependent upon the degree of roast (reviewed by Farah et al., 2006b). The result also indicates the selection of green bean or roasted bean for genetic analysis should be considered carefully. Using green bean can reflect the original concentration of the accessions and control the variables due to the roasting process, especially for trigonelline. However, roasted bean could be also be used because the reason of selection could be targeted at the accessions that less affected by roasting process for trigonelline (i.e. less reduced trigonelline during roasting) and more affected for caffeine (i.e. reduced caffeine after roasting).

4.3.3 Variation in coffee bean volatile profiles using a non-targeted analytical approach.

4.3.3.1 Protocol improvement

4.3.3.1.1 Determination of sample size for roast

Since the cherry samples were harvested from single trees at one harvest time, the quantity available was very limited, sample size ranging from 24 to 188 grams. A preliminary experiment was done to determine the lowest sample size for roasting using Red Catuai variety from Brazil. The experiment involved different roasting sample sizes at 25, 50, 75 and 100 g and assessed for aroma and volatile compounds using Gas Chromatography - Mass Spectrometry (GC-MS). As a result, the sample roasted with 25 g appeared to be different from the others, so the 50 g was taken as a standard roasting size for all samples. The number of samples thus reduces to 221 (out of the 235 harvested). The roasting of coffee bean was modified from sample to sample based on the moisture, density and bean size. Higher moisture content and density required longer time of roast.

The roasting capacity of the roaster used in this study was 100 g. An initial roasting experiment was conducted for 25, 50, 75 and 100 g. Two data types including total ion current (TIC) and extracted ion chromatogram (EIC), and fast evaluation of the aroma and taste of ground and brew coffee were used to determine the coffee sample size for roast.

Results of total ion current (TIC) using 239 data points showed that the batch size of 25 g was grouped distinctly from the other 3 batch sizes based on the principal component analysis (Appendix Figure S4.1). Data of extracted ion (EI) for 10 selected compounds using peak area (which correlates to relative concentration) also showed similar results (Appendix Figure S4.2). The powder of roasted coffee was used to evaluate the aroma and taste of ground and brewed coffee with the participation of five trained tasters. Similarly, general discussion and notes for each sample was recorded and 25 g was noted as “different” “slightly burned” “smoky” and “more bitter”. Analysis from different data types as well as aroma evaluation showed that the treatment of 25 g is the most different from the other treatments and was not the acceptable size for roasting. Given the varied small amount of beans in the samples collected, this information confirmed the validity of the minimum batch size of 50 g used for roasting most of the collection leading to the reduction of total samples to 221.

4.3.3.1.2 Determination of bean freshness

Due to the large population size, for convenience, all samples were proposed to be stored in -20^{o}C after roasting awaiting measurement. This is a typical approach used for the compositional analysis of large numbers of agricultural samples. While volatiles are particularly known to be effected by frozen storage, the relative changes are usually consistent, and thus comparisons within a set of samples stored the same way are possible. However, commercial advice from one roaster was that coffee taste and odour analysis should be performed within 24 - 48 hours after roasting. Therefore, an experiment was implemented to compare the profiles of bean compounds derived from the two methods involving frozen or fresh sample.

Similar to the determination of sample size for roast, two data types including TIC and EIC were used to determine the bean volatile differences. TIC data analysed by PCA showed that freshly-roasted and frozen roasted beans were grouped differently and that more fluctuation was observed in the latter (Appendix Figure S4.3), indicating that using freshly roasted samples might be more reliable for volatile analysis. Similarly, data from 10 selected compounds showed peak area was different between fresh and stored samples (Appendix Figure S4.4).

Results from both TIC and EIC indicated that even stored at -20^{o}C, the volatile profile was changing, and thus the fresh samples were used for the analysis even though it was a logistical challenge to conduct roasting and analyse within 24-48 hours. It should be noted, however, that more careful consideration should be taken with this approach, which does not allow for internal standards or calibration, as many other external factors may also affect the consitency of GC-MS results. This may include availability of instrumentation over such an extended period as this approach required. The need for prepared samples to sit in the autosampler tray awating analysis for varying durations was also less than ideal as volatile composition in the headspace of samples themselves may change over time. Further, instrumental shift is well documented for mass spectral instruments which require regular tuning. Mass spectrometers may indeed give variations in results over time, and an approach requiring large time-gaps between samples sequences where the instrument is sitting idle, is far from ideal. Indeed, these issues were encoutered during the execution of this experiment, indicating the gap between industry advice and the scientific approach. Thus, it is recommend that freezing samples is a more feasible and controllable approach for volatile analysis.

4.3.3.1.3 Verification of the ‘fingerprinting’ approach and its reproducibility.

In the current study, a novel chromatography ‘fingerprinting’ approach was applied for the rapid detection of the entire volatile profile ‘fingerprint’ which allowed for variation in almost any of the components across the coffee population to be observed. This approach has been used by others for secondary metabolites in citrus fruits peels (Parastara et al., 2012), medicines (Custers et al., 2014) and edible oils (Bagur-González et al., 2015). The advantages of applying this approach in the current study were:

(1) The aroma of coffee is not formed from a few target compounds, but rather a combination of numerous (>1000) compounds (Sunarharum et al., 2014) so a total fingerprint of the volatile profile would give a more ‘holistic’ picture of the diversity of components in each sample of the population, (2) the population studied had almost no prior information or data on the volatile compounds present with no known specific target compounds, so a fingerprint approach could be suitable for identification of groups of samples with contrasting volatile profiles for further investigation, and (3) with the large population of more than 220 samples, a targeted approach was unsuitable as it takes extensive sample preparation and data processing time to develop the method for even a small number of target compounds. Rapid whole chromatographic fingerprinting therefore could provide a potential fast approach to exlore and select representative samples for a more targeted volatile analysis.

Due to the fact that the fingerprinting approach is not a typical application for gas chromatography, data from target components of likely importance to coffee flavour, were extracted from the chromatograms and also used to check the reproducibility of the fingerprinting approach. TIC data of 1,347 points plotted on PCA showed that the majority of samples (21 out of 26 samples, or 85%) showed consistent data between two replicates, indicating the method was fairly reproducible (Figure 4.13). However, this result was for duplicates in consecutive runs; that of non-consecutive runs was lower (Table 4.9).

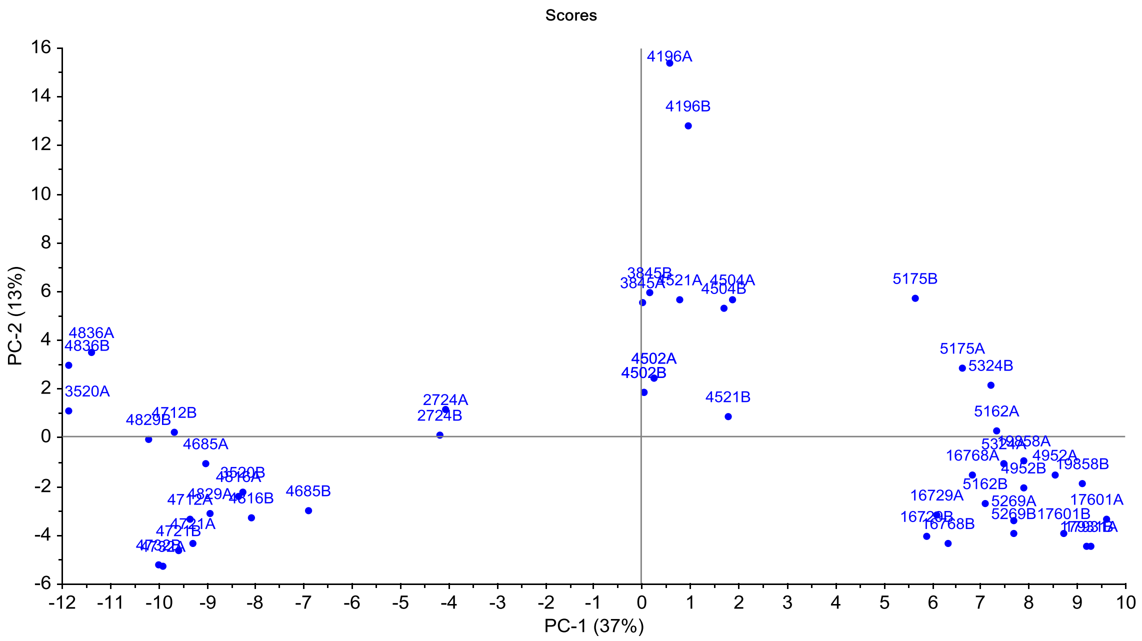


Figure 4.13 Method reproducibility test based on PCA of the duplicates (suffixes A and B) measured by fingerprinting approach.

The method reproducibility was also examined by using EIC data for specific target compounds. When running GC-MS, vials of duplicates were placed close to each other (for consecutive runs) or far apart from each other (for non-consecutive runs). Not surprisingly, results of non-consecutive duplicates showed more variation than those of consecutive duplicates (Table 4.9). Among 10 selected compounds, five compounds (2,5-dimethylpyrazine, 2,3-dimethylpyrazine, 2-ethyl-3,6-dimethyl-pyrazine, 2- furaldehyde and 3-methylbutyric acid) always gave low CV% indicating the reliability of the method, and these compounds could be selected for further analysis. The reason why larger variation was observed for 2,3-butanedione, and 2,3-pentanedione is likely due to these compounds being highly volatile and early eluting components (Akiyama et al., 2003).

According to Dart and Nursten (1985), the volatile composition is very complex and influenced by a large number of factors, particularly species, degree of roast, previous and subsequent processing and storage. The first three factors are more readily controlled than the fourth, and the short shelf-life of freshly roasted coffee is due primarily to the rapid changes which occur to the volatiles after roasting. The volatiles differ enormously in aroma quality, potency, concentration, and the influence each has on the other volatiles present. This explained why the replicates varied and the duplicates of non-consecutive samples (replicates far apart from each other) had more variation than the duplicates of consecutive samples (replicates next to each other) on the GC-MS tray when running. We suggest that the duplicates should be placed close to each other, and the fewer samples placed in the GC-MS tray waiting for being analysed the better, as there are 32 vials/tray which take more than a day to be analysed. The variation between samples may not be entirely from the genetic variation, but to some degree from the change and interaction of volatiles during this extended waiting period.

Table 4.9 Duplicate variation based on peak area of selective compounds extracted from TIC data

Rt	Compound name (class)	Consecutive run (39 samples)			Non consecutive run (18 samples)		
		Mean	SD	CV%	Mean	SD	CV%
3.925	Acetaldehyde (aldehyde)	1,478,056	40,708	2.65	1,579,966	137,521	10.25
6.025	2,3-Butanedione (Ketones)	1,018,368	134,123	13.59	1,015,094	164,454	16.01
7.458	2,3-pentanedione (Ketones)	1,551,263	166,617	10.72	1,526,029	181,284	11.84
11.312	2,5-dimethylpyrazine (Pyrazines)	2,050,122	35,227	1.73	2,008,868	73,271	3.61
11.667	2,3-dimethylpyrazine (Pyrazines)	260,815	4,532	1.73	265,215	10,276	3.90
13.050	2-ethyl-3,6-dimethylpyrazine (Pyrazines)	207,440	7,125	3.21	236,543	17,844	7.13
13.333	2- furaldehyde (Furans)	4,388,589	56,166	1.25	4,204,808	151,304	3.74
15.842	3-methylbutyric acid (Acids)	1,246,434	43,276	3.27	1,276,722	95,687	7.79
19.950	Furaneol (Furanones)	117,672	7,738	6.67	121,259	10,822	9.08
21.478	4-vinylguaiacol (Phenolic compounds)	292,046	15,436	5.23	270,143	29,679	12.07

For coffee, GC-MS with HS-SPME method has been used in several studies to discriminate arabica/robusta blends coming from different geographical origins based on their volatile profiles (Freitas et al., 2001; Zambonin et al., 2005). The method has also been used in identification of compounds in coffee brew based on coffee varieties and preparation methods (Rocha et al., 2003) or to differentiate coffee of different origins or in mixtures of different compositions (Bicchi et al., 1997) or defective beans from healthy roasted coffee beans (Agresti et al., 2008).

Rocha et al. (2003) mentioned about the combined technique of headspace SPME/GC/PCA as a lower-cost, fast, and reliable technique for the screening and distinction of coffee brews. Nebesny et al. (2007) used HS-SPME-GCMS for robusta coffee roasted by three different methods to find the effects of roasting methods on the formation and retention of volatile aroma compounds. The study found there were contradictory observations between results obtained by SPME-GC analysis and by evaluation of sensory attributes. Bertrand et al. (2012) applied the same technique and analysis instrument to determine the influence of different climatic conditions on coffee beverage and volatile compounds in green coffee. Results showed that high temperature has negative effect on coffee beverage and some volatile compounds could be used as indicators for growing conditions. However, those studies used EIC with authentic standards, none of them using a total volatile fingerprinting for a large population. This study used the chromatographic fingerprinting approach with inclusion of specific target compound data for selection of the subset of samples for further analysis – quantification approach, and result showed consistency. However, as mentioned above, one should take into account the problems of availability and certainty of the instrument, and the analysis time gap among samples to get the consistency between analyses. In addition, as mentioned in the method (see 4.2.2.5.3.) and next section (see 4.3.3.1.3), chromatographic fingerprinting faced the issue of shifting retention time (Rt) (see Amigo et al., 2010 for the drawbacks of fingerprinting approach). Due to the very novel nature of this approach, there has been no automatic procedure available that can help overcome this shifting problem even though COW (an option in the Unscrambler to align the chromatogram data) was applied (Figure 4.8b), the shifting in Rt was not necessarily proportional across the chromatographic period. Manual aligment/smoothing therefore was required (Figure 4.8c) which was impossible for such a large data set (32 minutes with 3,840 data points in total) with a single shift. This problem leads to the narrowing down of the number of data used in analysis as only the certain peaks (which were clear and shiftable) and data points nearby were used in the analysis and thus did not reflect the full fingerprinting. It seems that the method of using chromatographic fingerprints is in its infancy, and in need of further development before it can be used reliably for large sample population comparisons. Certainly, it has proved in this study to only be suitable for the first step in rough selection of important volatiles or the most diverse genotypes in the coffee population for the targeted SIM method.

4.3.3.2 Variation revealed by TIC and Selective compounds (EIC)

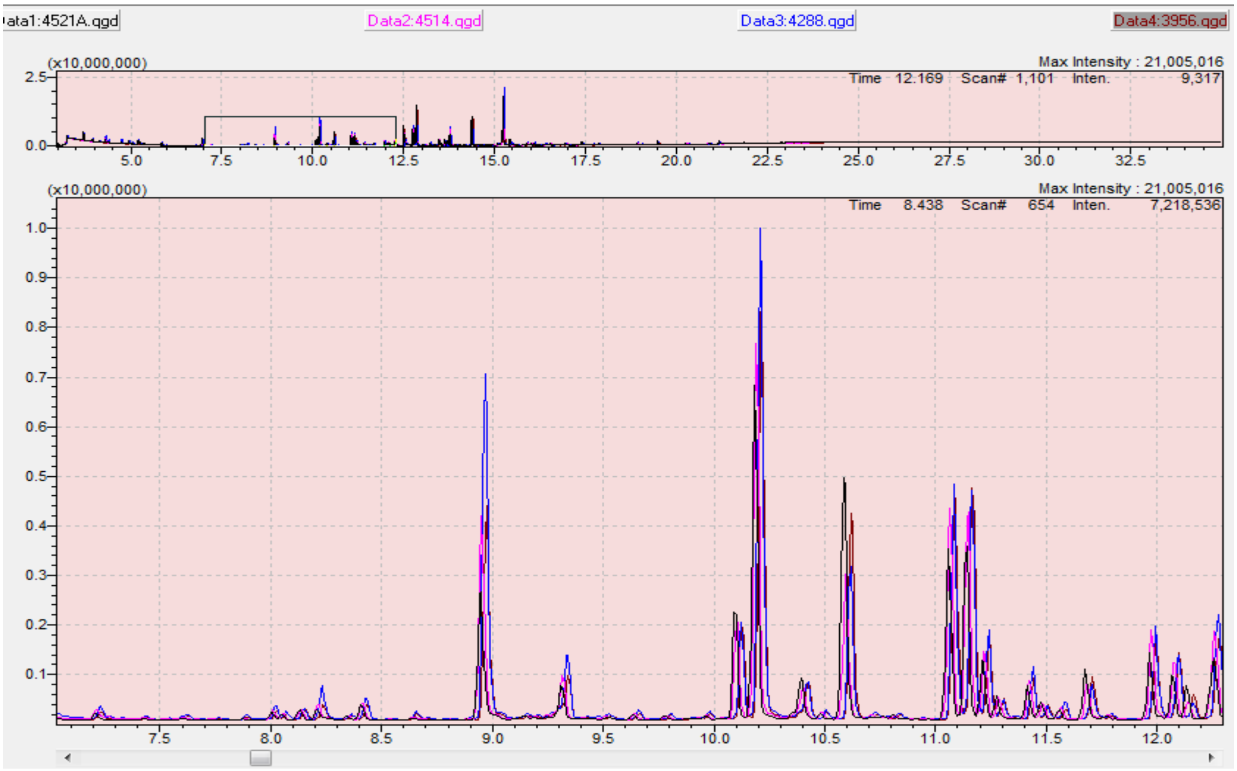


Figure 4.14 Example of volatile chromatographic fingerprints of 4 arabica coffee accessions.

Unexpectedly, the comparison among chromatographic fingerprints of volatiles of the arabica population shows that they seemed identical between individuals (Figure 4.14) even though genotypes used were selected from a diverse germplasm. The difference among samples is the peak area which was not measured correctly by this method. At first, full chromatogram (scanning 35-350 m/z) of 1,347 data points was used to examine the distribution of the population based on volatile chromatographic fingerprints (Figure 4.15). The population was divided into three groups and the most diverse groups come from wild type Ethiopia genotypes expressed by the outlying genotypes from the group (Figure 4.15, left). These genotypes were treated as outliers and removed from analysis and the new results were also with other wild type Ethiopia genotypes indicating the diversity or difference of these genotypes from the population. It would be interesting to quantify volatiles of these genotypes by SIM to confirm their variation. Because too many data points were used, the resolution was not good enough to identify which Rt corresponding to which compounds contributing most or influencing most to the population (Figure 4.15, right). The data was then narrowed down to the compounds that were identified as key compounds in coffee (Table 4.9). Among 140 peaks identified in full scan, 20 were clearly seen in the chromatography, had a high match to the NIST library and were in the list of key compounds (Table 4.10) and five data points around each of the compounds formed a set of 100 data points for use in PCA of the population (Figure 4.16). Identification of compounds was done by similarity searches in the NIST mass database and verified with retention indices.

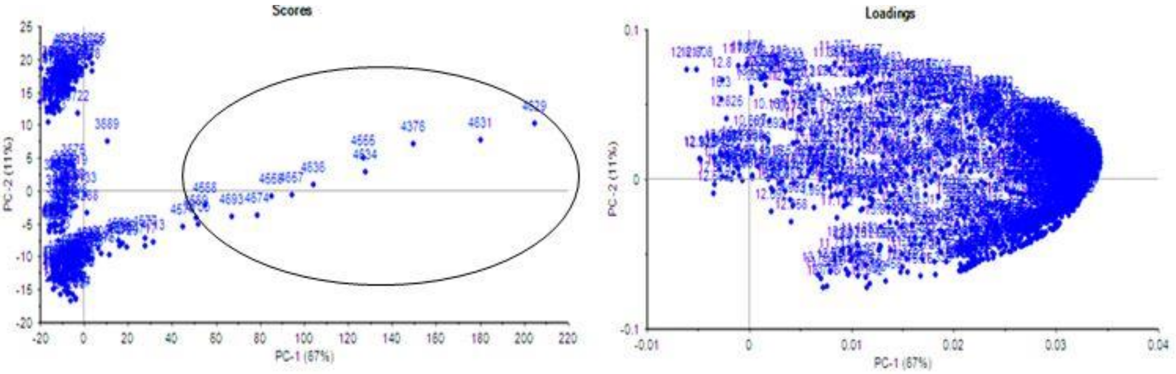


Figure 4.15 PCA of the population based on full chromatographic data (left) and the variables (Rt) (right).

Table 4.10 List of key coffee compounds identified from full scan.

No	Rt	Cas#	Compound name	Class	MW	% match	Aroma description(*)
1	3.717	75-07-0	Acetaldehyde	Aldehydes	44.05	88	fruity, pungent
2	5.900	431-03-8	2,3-Butanedione	Ketones	86.09	95	buttery, buttery-oily
3	7.008	600-14-6	2,3-Pentanedione	Ketones	100.12	92	buttery, buttery-oily

4	9.008	110-86-1	Pyridine	Pyridines	79.10	95	burnt rubber, rancid, fishy amine
5	9.333	138-86-3	D-Limonene	Terpenes	136.24	84	citrus, orange, fresh sweet
6	10.167	3188-00-9	3(2H)-Furanone, dihydro-2-methyl	Furanones	100.12	95	sweet caramel
7	10.258	109-08-0	Pyrazine, methyl-	Pyrazines	94.11	95	earthy, roasty, nutty-roast
8	11.125	123-32-0	Pyrazine, 2,5-dimethyl	Pyrazines	108.14	94	nutty
9	11.208	108-50-9	Pyrazine, 2,6-dimethyl	Pyrazines	108.14	94	
10	11.292	13925-00-3	Pyrazine, ethyl	Pyrazines	108.14	86	sweaty, peanut butter, musty, nutty, woody, roasted cocoa
11	11.483	5910-89-4	Pyrazine, 2,3-dimethyl-	Pyrazines	108.14	90	chocolate
12	12.042	15707-23-0	Pyrazine, 2-ethyl-3-methyl	Pyrazines	122.17	84	earthy
13	12.825	431-03-8	2,3-Butanedione	Ketones	86.09	92	buttery, buttery-oily
14	12.933	98-01-1	Furfural	Furans	96.09	94	bread, caramel, sweet woody, almond, fragrant, baked bread
15	13.775	600-14-6	2,3-Pentanedione	Ketones	100.12	89	buttery, buttery-oily
16	13.842	623-17-6	2-Furanmethanol, acetate	Furans	140.14	93	sweet fruity, banana, horseradish
17	14.458	620-02-0	2-Furancarboxal-dehyde, 5-methyl	Furans	110.11	93	spicy, candy, slight caramel, spice, caramel, maple
18	15.325	98-00-0	2-Furanmethanol	Furans	98.1	92	sweet fruity, banana, horseradish
19	15.492	503-74-2	Butanoic acid, 3-methyl	Acids	102.13	85	acidic, sweaty
20	21.250	7786-61-0	4-vinylguaiacol	Phenolic compound	150.18	91	Spicy, phenolic, clove-like

(*) Summarised by Sunarharum et al. (2014)

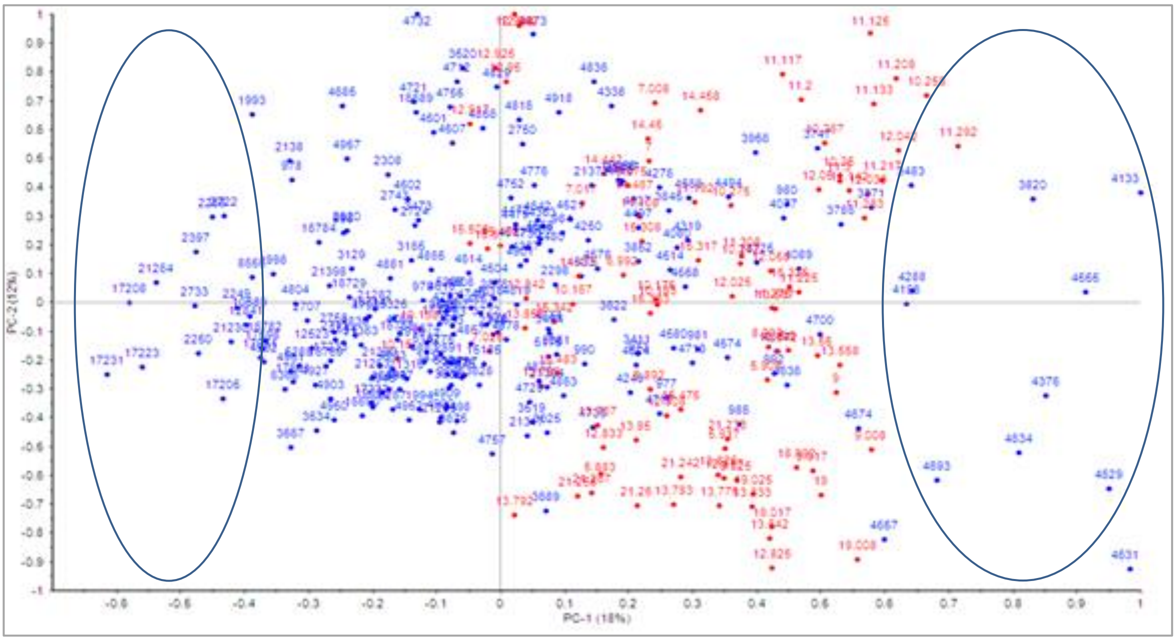


Figure 4.16 PCA of the coffee germplasm collection based on 100 selected chromatographic data

(Blue: Accession code; Red: Retention time; Circle: accessions at two extremes).

The Rt value of 11.292 corresponding to pyrazine showed the most influence on the population distribution followed by Rt values of 10.258 and 9.008 corresponding to pyrazine and pyridine, respectively (Figure 4.16). This result guides the selection of a subset of samples at two extremes for measurement of these variables or compounds using a quantification method (SIM). It is not really surprising that the extreme phenotypes are from wild types, because domestication would have primarily selected for preferred morphology (observable traits). The genotypes with lowest pyrazine and pyridine (on the left of the Y axis) were 17231, 17223, 17208 and 21264, all being wild types from Cameroon. The genotypes with the highest above mentioned compounds (on the right of the Y axis) were 4133, 4631, 4555, 4629 and 4376, three of which were wild type from Ethiopia. It would be worth quantifying the concentration of these samples to confirm their values for quality breeding. Even when the morphology and non-volatile (caffeine and trigonelline) data were added in the analysis, the distribution of the population kept almost unchanged as the substantial influence of the aforementioned Rt values to the population (Appendix, Figure S4.5).

For the EIC data, among 10 selective compounds, five of them appeared most stable. They were used for PCA to check if the subset of samples at two extremes were consistent between two data types (TIC vs EIC). Results showed that a number of samples at two extremes (in circles) were in common between two data types (Figure 4.16 and 4.17).

The volatile fingerprinting results showed that arabica is fairly similar in the presence of peaks or number of peaks presenting in each sample. Fast evaluation of the coffee powder aroma shows clearly that each genotype has different aroma. However, using this method, no difference regarding presence and absence of the peaks or number of peak was detected among the population. The difference may come from peak height or peak areas which link to the concentration of each compound. Other reasons could be due to the combination of the compounds, or the presence of important compounds (dominant aroma) below their detection threshold. As mentioned in Dart and Nursten (1985), the aroma of roasted coffee is very complex, being composed of a large number of volatiles with different odour qualities, some pleasant, some unpleasant, and many probably below their detection threshold. The concentration, proportion and influence of one volatile on another all affect the final aroma quality. For example, 2,4,5-trimethyloxazole has an earthy, potato-like aroma but in the presence of 2,3-butanedione the aroma is like sweet pyridine. This illustrates the fact that the aroma is not due simply to an additive effect of the volatiles, but involves synergism and antagonism. The effects that one volatile has on another are not well understood, and nearly impossible to elucidate in a mixture as complex as the volatile fraction of coffee, where several hundred compounds are present at widely different concentrations (Dart and Nursten, 1985).

If the reasons were from the difference of concentration of compounds, not the number of peaks presenting in each sample, this qualitative and semi-quantitative method would not have given the exact answer. Therefore, two groups of samples from two extremes on the axis (Figure 4.16 and 4.17) were selected for the quantification method (SIM). Outliers in the PCA, which are mainly wild type from Ethiopia, were also selected for quantification (Figure 4.15, left and Appendix, Figure S4.5).

The variables of the PCA showed pyrazines, furans and ketones mostly influenced the PCA grouping in which accessions with contrasting bean volatile fingerprinting were mostly clustered in different groups (Figure 4.16). Data from EIC also showed compounds belong to these classes are quite stable and consistent (Table 4.9) and were thus selected for SIM approach (Appendix, Table S4.4). In total, 35 samples were selected for quantification including 15 from low compounds group (four from group of cultivation/selection, four from hybrid group, seven from wild type), 15 from high compounds group (four from group of cultivation/selection, two from hybrid group, eight from wild type, and one robusta) and five from outlying wild type accessions (Appendix, Table S4.4).

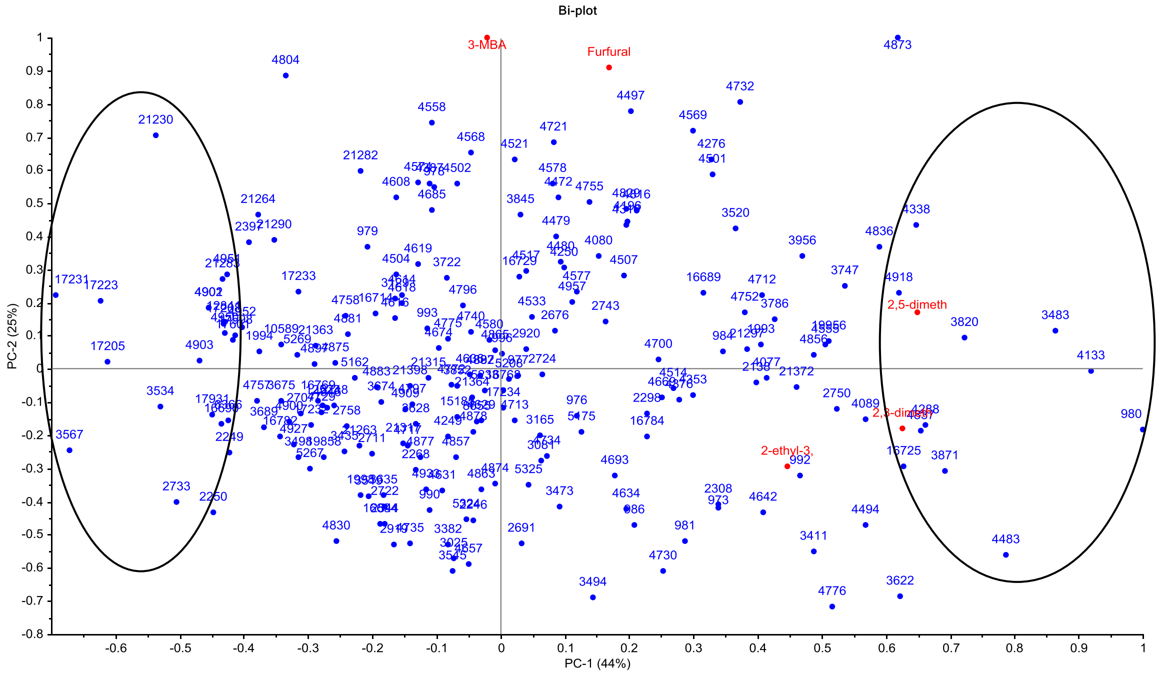


Figure 4.17 Population distribution based on PCA involving five selective compounds (EIC)

(Accession code: blue; Selective compounds: red)

4.3.4 Variation in volatiles using targeted analysis

As mentioned above, the principal components analysis of the chromatographic fingerprints derived from the non-targeted scanning of 221 samples allowed identification of 35 accessions with contrasting distribution. Targeted analysis of these accessions showed significant variation ($P < 0.001$), with CV% varying from 14% (for 4-vinylguaiacol) to 62% (for geraniol). The volatile concentration ranged from 9 ppb for beta damascenone to 53,300 ppb for 2,5-dimethylpyrazine, and all volatiles quantified were far above their reported sensory threshold concentration which verifies their importance as target aroma volatiles in arabica (Table 4.11).

Table 4.11 Variation in 18 key volatile compounds measured in a representation of 35 arabica coffee accessions selected from non-targeted analysis.

Compounds	Min (ppb)	Max (ppb)	Mean (ppb)	% CV	LSD $_{0.05}$	Literature (ppb)	Sensory threshold (ppb)	Aroma description
Ketones 2,3-butanedione	1,426	9,158	4,255	48	3,158	48,400 - 50,800[a]	0.3[a]	buttery[a] oily, fruity, caramel-like[bd]
2,3-pentanedione	2,256	27,932	12,258	48	4,330	3,540 - 39,600[a]	20[a]	buttery oily[a], caramel-like[bd]
Aldehydes 2-methylpropanal	1,050	4,773	2,344	36	859	320 – 430[c]	3[c]	buttery oily[a]
3-methylbutanal	896	3,723	2,014	33	541	210 - 18,600[a, c]	0.35[a]	malty[ac], chocolate[c], caramel-like[d]
2-methylbutanal	257	1,052	523	36	169	200 - 20,700[ac]	1.3[a]	chocolate[c], caramel-like[d]
Phenolic compounds guaiacol	381	3,384	1,479	43	504	2,000 - 17,970[a]	2.5[a]	phenolic, burnt[ab], smoky, phenolic[d]
4-ethylguaiacol	64	371	186	44	63	800 - 24,800[a]	25[a]	spicy[a], sweet[b], smoky, phenolic[d]
4-vinylguaiacol	22,723	38,773	30,253	14	4818	8,000 - 64,800[a]	0.75[a]	spicy[a], clove-like[b], smoky, phenolic[d]
Norisoprenoids beta damascenone	7	14	9	19	2	195 – 255[a]	0.00075[a]	honey-line, fruity[ad]
Pyrazines 2,5-dimethylpyrazine	19,201	123,612	53,312	38	19,484	4,550 – 14,070[ac]	80[a]	-
2-ethyl-3,6-dimethylpyrazine	2,685	12,320	6,390	36	20	2,570 - 5,980[a]	8.6[a]	-
2,3-diethyl-5-methylpyrazine	41	134	75	31	25	73 – 95[a]	0.09[a]	nutty-roast[a], earthy[d]
2,3-dimethylpyrazine	3,131	33,662	11,963	50	10,192	2,580 – 8,890[ac]	800[a]	-
2-ethyl-3,5-dimethylpyrazine	479	2,136	996	34	326	55 – 840[ac]	0.04[a]	nutty-roast[a], earthy[d]
3-isobutyl-2-methoxypyrazine	7	38	14	56	6	59 – 97[a]	0.002[a]	peasy[a], earthy[d]
Terpenes linalool	15	143	57	52	20	780-1310[c]	0.17[a]	flowery[a], floral, fruity (citrus)[c]
geraniol	5	75	36	62	16	-	1.1[a]	-
limonene	29	169	80	44	43	1080-1320[c]	4[a]	-

CV (%): coefficient of variation. LSD $_{0.05}$: Least Significant Difference at P value less than 0.05.
[a]: summarised by Sunarharum et al., 2014; [b]: Amanpour et al., 2015; [c]: Piccino et al., 2014; [d]: Belits et al., 2009

Five of the 18 compounds (2,3-pentanedione, 3-methylbutanal, 2-methylbutanal, 4-vinylguaiacol and 2,3-diethyl-5-methylpyrazine) had concentrations in the range previously reported. Eight compounds (2,3-butanedione, 3-methylbutanal, guaiacol, 4-ethylguaiacol and beta damascenone, 3-isobutyl-2-methoxypyrazine, linalool, limonene) had concentrations lower than those reported in the literature. Four compounds belonging to the pyrazines class (2,5-dimethylpyrazine, 2-ethyl-3,6-

dimethylpyrazine, 2,3-dimethylpyrazine and 2-ethyl-3,5-dimethylpyrazine) and aldehydes (2-methylpropanal) had concentrations far exceeding those reported in literature (Belitz et al., 2009; Cheong et al., 2013; Czerny et al., 1999; Piccino et al., 2014).

Significant variation in volatile concentrations among the 35 diverse accessions indicates their potential in quality improvement. Among 18 compounds quantified, five had concentrations in the range previously reported, and eight were lower than that reported in the literature. However, these studies were limited in the number of accessions examined (Cheong et al., 2013; Czerny et al., 1999; Czerny and Grosch, 2000; Piccino et al., 2014; Semmelroch and Grosch, 1996; Semmelroch et al., 1995). It is interesting that most compounds belonging to the pyrazines class (2,5-dimethylpyrazine, 2-ethyl-3,6-dimethylpyrazine, 2,3-dimethylpyrazine and 2-ethyl-3,5-dimethylpyrazine) were much higher than in the literature (summarised by Sunarharum et al., 2014) even with the same roast degree (light to medium), for example 2,5-dimethylpyrazine of 53,312 ppb in this study vs 662 ppb or 2,3-dimethylpyrazine of 11,963 ppb vs 119 ppb in another study (Toci and Farah, 2014). A study investigating the influence of environment on coffee volatiles found that aldehydes and ketones appeared to be positively linked to elevated temperatures and high solar radiation (Bertrand et al., 2012). The coffee used in the current study was collected from an elevation of 616 m (above sea level) and average annual temperature of 22.2^{o}C. This may partly explain the reason for the lower concentration of these compounds compared to those in the literature and the high pyrazines. Low variation among replicates (or %CV) suggests the reliability of the volatile analysis and that these volatiles such as 4-vinylguaiacol, beta damascenone, a few volatile pyrazines (2-ethyl-3,6-dimethylpyrazine and 2,3-diethyl-5-methylpyrazine) and aldehydes (3-methylbutanal) can be used as criteria for breeding purposes.

Although the concentrations were lower or in the range previously reported for some volatiles, they are all above the sensory threshold (see definition in Lawless and Heymann, 2010, p127) and significantly varied among arabica accessions. Other volatiles in the pyrazines class (2,5-dimethylpyrazine, 2-ethyl-3,6-dimethylpyrazine, 2,3-dimethylpyrazine and 2-ethyl-3,5-dimethylpyrazine) and the aldehydes class (2-methylpropanal) had concentration ranges exceeding those reported in the literature (Table 4.10), which could be good sources for quality improvement in breeding.

Principal components analysis of 18 volatiles measured in 35 accessions showed that the first two components PC1 and PC2 explained 61% of the total variation (Figure 4.18). The population was distributed along the PC1 with contrasting volatile profile, either low or high in almost all volatiles, but mostly affected by pyrazines (2,5-dimethylpyrazine, 2-ethyl-3,5-dimethylpyrazine and 2-ethyl-3,6-dimethylpyrazine) and aldehydes (2-methylpropanal , 3-methylbutanal and 2-methylbutanal)

(Figure 4.18). PC2 showed that coffee accessions that were either higher in phenolic compounds (guaiacol, 4-ethylguaiacol) and lower in linalool (terpenes) and 3-isobutyl-2-methoxypyrazine (pyrazines) or vice versus. Almost all volatile compounds of the same chemical class were clustered into groups such as aldehydes, phenolic compounds, pyrazines, and ketones. Some accessions showed high concentrations in most compounds (e.g., 4857, 4909 and 4918) while others (e.g., 4288, 3483, 3747 and 4387) were low in most compounds. Accessions 1993 and 2268 were highest in linalool (terpenes) and 3-isobutyl-2-methoxypyrazine (pyrazines) and lowest in phenolic compounds (guaiacol, 4-ethylguaiacol), while accessions 4631, 4634 and 4693 were the opposite.

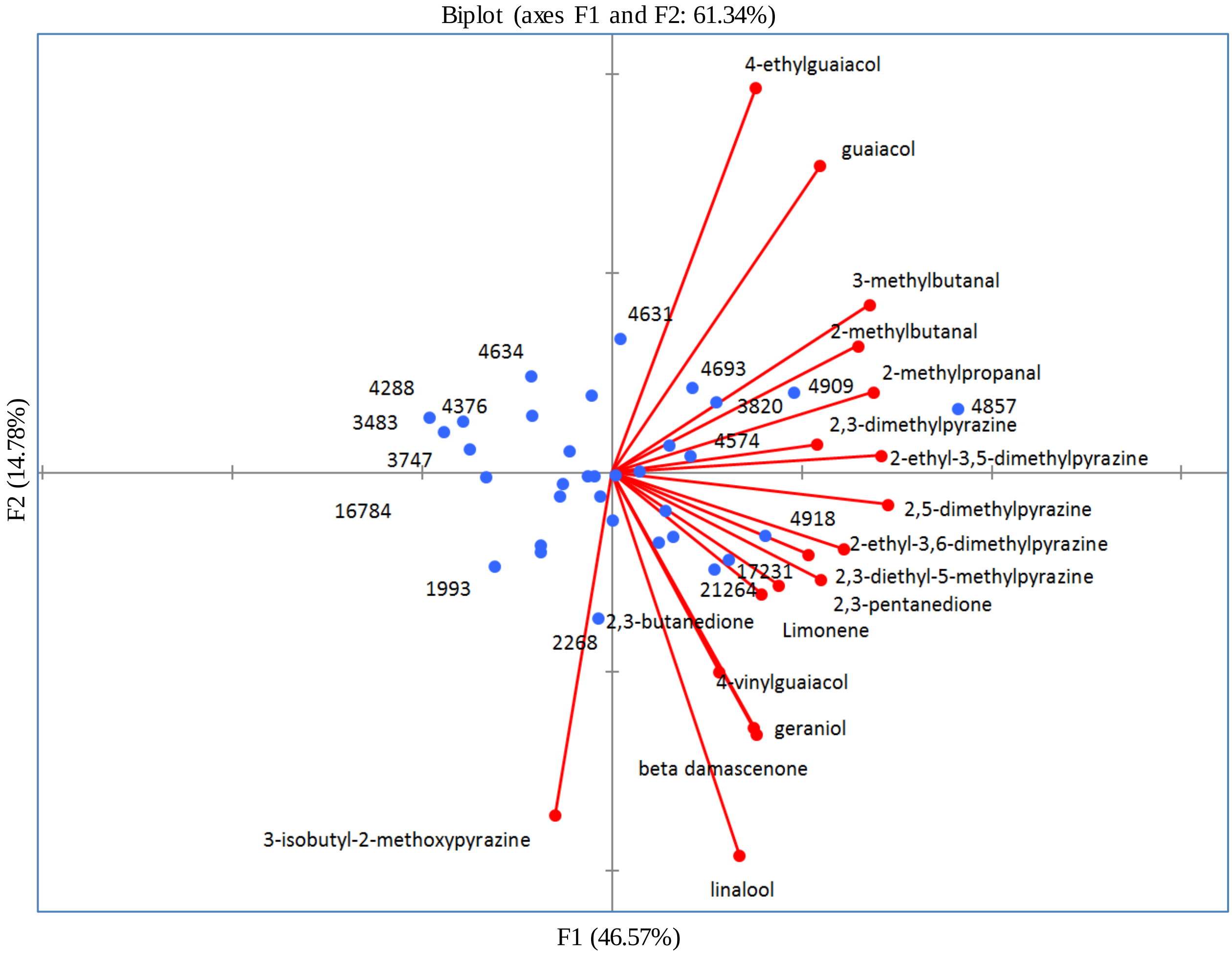


Figure 4.18 Principle component analysis of 18 volatiles measured in roasted coffee bean for 35 arabica accessions using targeted analysis

(Red lines and dots are volatile compounds, blue dots are accessions codes)

The multivariate analysis (PCA) of the 18 volatiles showed that almost all volatile compounds of the same chemical class were clustered into groups as expected indicating the reliability of the method used. The population was distributed with contrasting volatile profile (along the PC1), either low or high in almost all volatiles suggested that the selection of samples for low or high volatiles for breeding is possible. It is interesting to observe that samples with high volatiles contributing most to the PC1 were from the wild type group such as 4857 (E-457, Ethiopia, Wild), 4909 (E-540, Ethiopia, Wild), 4918 (E-549, Ethiopia, Wild) and 17231 (L334 et 45 c7, Cameroon, Wild) while samples with low volatile were mainly from the variety/cultivar group and hybrid group, for example 4288 (Irgalen Kella Sidano, Belgian Congo, cultivar), 3747 (San Rafael, Costa Rica, Variety), 4387 (Hibrido de Timor CRRC 1343/80, Portugal, Hybrid) and 16784 (Sarchimor F3 IAC 1669/31-1 CIFC H-361/971-10 Villasarchi x 832/2, Brazil, Hybrid). This indicates that the previous selection of varieties for cultivation was mainly based on bean morphology and that wild type accessions could be a valuable source for breeding. Similarly, accessions contributing most to PC2 which are higher in phenolic compounds (guaiacol – phenolic, burnt and 4-ethylguaiacol - spicy) and lower in linalool (terpenes) - flowery aroma and 3-isobutyl-2-methoxypyrazine (pyrazines) - peasy such as 4631 (E-358, Ethiopia, Wild), 4634 (E-361, Ethiopia, Wild), 4693 (Limnu E-188, Ethiopia, Wild) and 1993 (Goiaba colección 11 (552), Brazil, selection), or accessions which were opposite in the concentration of these compounds such as 2268 (San Martin, Guatemala, Cultivar), 21264 (ET-08, Indonesia, Wild), could be useful materials for genetic improvement of the corresponding compounds.

4.3.5 Trait correlations and implication for quality breeding

Spearman rank correlation analyses of the green bean morphology and non-volatiles among 232 accessions showed that there were strong correlations between bean physical parameters (Table 4.12). However, there was a small but significant positive correlation between caffeine and trigonelline (r = 0.141, P < 0.05) and between caffeine and 100 bean weight (r = 0.199, P < 0.001), while trigonelline had a slight negative correlation with 100 bean weight (r = -0.157, P < 0.05) (Table 4.12).

Table 4.12 Spearman rank correlation between caffeine and trigonelline content (in green bean) and green bean physical characteristics (n = 232).

Variables	Caffeine	Trigonelline	W100	Density	Length	Width
Trigonelline	0.141*					
W100	0.199**	-0.157*				
Density	-0.106	0.000	-0.232***			
Length	0.077	0.041	0.567***	-0.195**		
Width	0.118	-0.106	0.564***	-0.217***	0.267***	
Thickness	0.160*	-0.056	0.598***	-0.222***	0.445***	0.382***

*,**,*** Significantly different from zero at the 0.05, 0.01 and 0.001 significance level, respectively.

The lack of strong correlation between physical bean properties and chemical compounds in this study agrees with other reports. Trugo et al. (1983) found no significant correlation between caffeine and trigonelline contents in 13 instant coffees. Similarly, Casal et al. (2000) analysed 20 roasted robusta coffees and 9 roasted arabica coffees and found that while there was a strong negative correlation between caffeine and trigonelline when all data from both arabica and robusta samples was analysed simultaneously, their correlation was not significant when coffee varieties were analysed separately. Kathurima et al. (2009) studied the relationship between beverage quality and green bean physical characteristics of 42 arabica coffee genotypes from Kenya and found a relatively better beverage quality in the beans with low 100-bean weight compared to that of larger beans, but the correlation was not significant. Similarly, Dessalegn et al. (2008) indicated beans with a low 100-bean weight had relatively more caffeine than heavy beans in 42 *C. arabica* genotypes from Ethiopia. This study also showed that correlations between caffeine content and green bean physical characteristics were not significant while caffeine content had a significant negative correlation with all cup quality attributes of coffee such as acidity, body, flavour, and overall standard of the liquor. However, these studies were implemented with only 42 genotypes.

The large variation observed for bean size of the germplasm collection (Table 4.3) provides a useful opportunity for improving genetic gain, as it is also a highly heritable trait and may have an effect on quality. Giomo et al. (2012) used 24 *Coffea arabica* genotypes from Brazil to evaluate physical characteristics of the green beans and to describe the sensory profile. The results indicated that there were significant genotype effects both on coffee bean size and overall sensory quality, and the genotype × environmental interaction was not significant. Thus, the green bean size could be consider an important criterion for coffee plant selection aiming to improve the green bean quality since the evaluations and comparisons have occurred in the same environmental conditions and in the same post-harvest processing procedures. In our study, however, the correlation between 100 bean weight and caffeine and trigonelline, and between caffeine and trigonelline was not highly significant when investigating in a larger number of samples (n = 232). This implied that it is difficult to select for low or high non-volatile compounds such as caffeine and trigonelline based on bean physical characteristics. However, it also indicates that the selection for this trait will not be affected by the selection of the others, or that it is possible to select at the same time a variety that combines large bean size, low caffeine and high trigonelline. Development of molecular markers for each of these traits would enable them to be selected simultaneously in breeding. The investigation of a smaller subset of samples (n = 35) based on two volatile extremes show a strong negative correlation between weight of 100 beans and trigonelline ($r = -0.57$, $p<0.001$) indicating that trigonelline as aroma precursors may link to volatiles presented in the samples such as furans, pyrazines, pyrroles, pyridines (Franca et al., 2005). Based on the phenotypic data generated for this diverse arabica germplasm collection, further genotyping of the population will enable marker-trait association analysis for application in marker-assisted selection.

Among the 35 accessions selected for volatile-compound analysis, there was no strong correlation between volatile profiles and either bean morphology or non-volatile levels, except for a strong correlation between bean roasting degree (or roasting colour) and 4-ethylguaiacol (r = -0.79, P < 0.001) or guaiacol (r = -0.75, P < 0.001) (i.e., higher 4-ethylguaiacol or guaiacol in darker roasted bean) or less strong correlation with 3-methylbutanal (r = -0.42, P < 0.001). Strong correlations existed between compounds belong to the same class (aldehydes, phenolic compounds, pyrazines) (up to r = 0.93, P < 0.001) or between compounds of different classes, such as aldehydes and ketones (r = 0.73, P < 0.001), aldehydes and phenolic compounds (r = 0.82, P < 0.001), and aldehydes and pyrazines (r = 0.72, P < 0.001) (Table 4.13).

To the student's knowledge, this is the first report where correlations between bean volatile compounds have been examined. A number of previous studies merely focused on the relationship between bean non-volatile and cupping quality which was controlled by the volatiles. Franca et al. (2005) and Farah et al. (2006b) found caffeine had a positive correlation with high quality beverages in arabica. Similarly, Figueiredo et al. (2013) reported that higher trigonelline was found in the best sensory score genotypes, and Barbosa et al. (2012) confirmed that both trigonelline and caffeine correlate with better sensory scores. However, Figueiredo et al. (2013) showed no correlation between caffeine content and beverage quality. Similarly, Avelino et al. (2005) studied arabica grown in two regions of Costa Rica and found both caffeine and trigonelline were not well correlated with the sensory characteristics. The effect of environment to beverage quality score was also reported in several studies (Avelino et al., 2005; Barbosa et al., 2012; Rodrigues et al., 2009). The current study showed that both caffeine and trigonelline have weak positive and negative correlation with almost all volatiles even though trigonelline was considered as an important precursor of volatile compounds that link to coffee aroma and taste (Malta & Chagas 2009 in Barbosa et al., 2012). This would be hard to link to the sensory attributes as the complexity and the interaction of the volatiles in the aroma matrix exists.

He et al. (2015) found that darker roast correlated with aromatic compounds such as heptyl ether, 4-ethyl-2-methoxyphenol, 5-methylfuran-2-carbaldehyde and 1-(furan-2-ylmethyl)-1H-pyrrole, while Belitz et al. (2009) reported that 2-furfurylthiol and guaiacol increase with increasing degree of roasting. Mayer et al. (1999) found that pyrazines (2,3-diethyl-5-methylpyrazine, 2-ethenyl-3-ethyl-5-methylpyrazine, 3-isobutyl-2-methoxypyrazine), 4-vinylguaiacol, vanillin, 2-methyl-3-furanthiol and dimethyl trisulfide had almost no significant change with degree of roasting while propanal, 2(5)-ethyl-4-hydroxy-5(2)-methyl-3(2H)-furanone, guaiacol, 4-ethylguaiacol, 2-furfurylthiol, 3-methyl-2-buten-1-thiol and methanethiol were affected by roasting. However Toci's study (2014) showed almost all pyrazines concentration was decreased with roast degree. Holscher and Steinhart

(1992) indicated that ketones (2,3-butanedione and 2,3-pentanedione) and aldehydes (2-methylpropanal, 3-methylbutal, 2-methylbutanal) were decreased with degree of roasting. In the current study, we found that colour of roasted bean had a strong correlation with 4-ethylguaiacol and guaiacol, and intermediate correlation with 3-isobutyl-2-methoxypyrazine and 3-methylbutanal (Table 4.13) which is in agreement with Belitz et al. (2009) and Mayer et al. (1999) for guaiacol and most of pyrazines. However, it was different from the study of Holscher and Steinhart (1992) for ketones and aldehydes. The results again implied that the roasted bean colour could serve as a simple indicator for selecting certain favourable volatiles.

From the breeding perspective, the 35 individuals selected based on their distinct variation in complete volatile fingerprints can serve as useful founder parents for genetic improvement of volatile compounds. The strong and significant correlations (Table 4.13) between targeted volatile compounds that belong to the same class (aldehydes, phenolic compounds, pyrazines) or between compounds of different classes (aldehydes and ketones, aldehydes and phenolic compounds, and aldehydes and pyrazines) suggests the selection of certain compounds for analysis should be the indication for many more other volatiles present in this germplasm subset. Such indirect selection could help reduce the labour, effort and cost for volatile analysis by focusing only on representing compounds that are easier and more stable to measure. For aldehydes, 3-methylbutanal can be selected for analysis as it is highly correlated with 2-methylbutanal (r = 0.91, P < 0.001) and 2-methylpropanal (r = 0.91, P < 0.001), and has lower variation among replicates. For phenolic compounds, 4-vinylguaiacol had a weak correlation with both guaiacol and 4-ethylguaiacol, while 4-vinylguaiacol had a highly reproducible result, it can be selected for analysis along with either guaiacol or 4-ethylguaiacol, which were strongly correlated. For pyrazines, among the six volatiles that were strongly correlated, 2-ethyl-3,6-dimethylpyrazine and 2,3-diethyl-5-methylpyrazine had the least variation among replicates (i.e., most reproducible) and thus could be used for analysis of this chemical class.

Table 4.13 Spearman rank correlation between bean morphology, non-volatiles and volatiles of 35 samples.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(2)	-0.12																						
(3)	0.13	-0.22			Bean morphology																		
(4)	0.23	-0.41	0.25																				
(5)	0.07	0.44	-0.08	-0.08																			
(6)	-0.47	0.37	-0.26	-0.06	0.20		Non-volatiles																
(7)	-0.11	-0.29	0.15	-0.03	-0.21	-0.31																	
(8)	0.10	-0.21	0.07	-0.04	-0.16	-0.37	**0.68**	Ketones															
(9)	0.12	-0.12	0.01	-0.22	-0.09	-0.28	**0.53**	**0.82**															
(10)	-0.04	0.04	-0.06	-0.42	-0.07	-0.24	**0.50**	**0.75**	**0.93**			Aldehydes											
(11)	-0.09	-0.03	0.01	-0.27	-0.14	-0.23	**0.54**	**0.73**	**0.92**	**0.92**													
(12)	-0.28	0.30	-0.28	**-0.75**	0.02	0.04	0.20	0.35	**0.55**	**0.69**	**0.57**												
(13)	-0.20	0.44	-0.39	**-0.79**	0.16	0.13	-0.09	0.05	0.31	**0.49**	0.35	**0.86**			Phenolic compounds								
(14)	0.18	-0.29	0.36	0.01	-0.09	-0.18	0.32	0.37	0.27	0.25	0.22	0.19	0.13										
(15)	0.17	-0.30	0.41	0.08	-0.11	-0.24	0.40	**0.56**	**0.46**	0.38	0.36	0.11	-0.06	**0.51**	Norisoprenoids								
(16)	-0.03	-0.09	0.14	-0.11	0.13	-0.14	**0.48**	**0.67**	**0.72**	**0.68**	**0.58**	**0.51**	0.25	0.38	**0.54**								
(17)	-0.01	-0.08	0.06	0.09	0.30	-0.04	0.41	**0.52**	**0.55**	**0.48**	0.43	0.34	0.07	0.30	0.38	**0.81**							
(18)	0.00	-0.08	0.02	0.16	0.35	0.00	0.26	0.34	0.38	0.31	0.25	0.26	0.03	0.18	0.21	**0.69**	**0.93**					Pyrazines	
(19)	0.11	-0.11	0.23	-0.21	0.03	-0.06	0.13	0.27	**0.48**	0.38	0.37	0.42	0.28	0.34	0.45	**0.61**	0.41	0.38					
(20)	-0.08	0.00	-0.01	-0.15	0.25	-0.08	0.44	**0.56**	**0.66**	**0.64**	**0.56**	**0.57**	0.33	0.33	0.37	**0.87**	**0.94**	**0.86**	**0.52**				
(21)	0.27	**-0.50**	**0.50**	0.37	0.09	-0.24	0.10	0.05	-0.07	-0.18	-0.18	-0.30	-0.39	0.47	0.37	0.13	0.17	0.20	0.12	0.03			
(22)	-0.03	-0.35	0.44	0.12	-0.17	-0.13	0.44	**0.50**	0.32	0.27	0.23	0.06	-0.26	0.32	**0.62**	**0.50**	0.42	0.33	0.32	0.38	0.43	(Terpenes)	
(23)	-0.12	-0.24	0.25	-0.04	-0.26	-0.03	0.37	0.27	0.28	0.21	0.24	0.19	-0.03	0.07	0.46	0.38	0.21	0.15	0.37	0.26	0.29	**0.72**	
(24)	-0.12	-0.12	0.14	0.05	0.03	-0.22	0.33	0.46	0.40	0.38	0.44	0.29	0.04	0.16	0.22	**0.47**	0.36	0.33	0.21	0.42	0.10	0.43	0.32

(1) Weight of 100 bean (2) Bean moisture loss during roasting (3) Roasted bean density (4) Bean roasting colour (5) caffeine (6) trigonelline (7) 2,3-butanedione (8) 2,3-pentanedione (9) 2-methylpropanal (10) 3-methylbutanal (11) 2-methylbutanal (12) guaiacol (13) 4-ethylguaiacol (14) 4-vinylguaiacol (15) beta damascenone (16) 2,5-dimethylpyrazine (17) 2-ethyl-3,6-dimethylpyrazine (18) 2,3-diethyl-5-methylpyrazine (19) 2,3-dimethylpyrazine (20) 2-ethyl-3,5-dimethylpyrazine (21) 3-isobutyl-2-methoxypyrazine (22) linalool (23) geraniol (24) limonene (values in bold are different from 0 at the 0.001 significance level).

4.4 Conclusions

The observed high diversity in bean morphology, non-volatiles and volatiles in the worldwide arabica collection demonstrated its potential application as a valuable genetic resource to quality improvement in coffee breeding. For bean morphology, substantial variation was observed for weight of 100 beans, bean length, width and thickness, and bulk density. Non-volatiles including caffeine and trigonelline also showed large variation in the range previously reported. There was small variation among different groups of arabica in both caffeine and trigonelline content. The genotypes with both low caffeine and high trigonelline were not found in any wild type coffee accessions even though they presented in a large number and were more genetically diverse than the cultivated collection as reported in the literature. The analysis allowed identification of accessions with contrasting levels of caffeine and trigonelline. These groups will be bulked and subject to whole genome sequence for identification of trait-associated DNA markers to assist arabica coffee breeding.

Non-targeted analysis revealed a low diversity in the entire arabica population based on chromatographic fingerprints but more variation was observed in the wild type as expected. These results were also confirmed by targeted analysis of 18 volatiles from 35 selected accessions, in which most volatiles were lower or in the range previously reported, except for several compounds in the pyrazine and aldehyde classes that could help identify useful sources for genetic improvement of coffee quality.

There was insignificant correlation between bean morphology and non-volatile or volatile compounds, or between caffeine and trigonelline content. This implied that biochemical analysis would still be needed to quantify both volatile and non-volatile compounds because bean morphology cannot be used as their predictor based on our correlation analysis. The lack of strong correlations also indicates that it should be possible to breed for desirable combinations of traits independently (i.e. large bean size, low caffeine, high trigonelline, and favourable volatiles). The strong correlations existing within several volatile groups provide useful direction for targeted analyses focusing on reproducible and representing compounds so as to improve analytical accuracy and efficiency in coffee bean quality research and industry application.

CHAPTER 5: A DRAFT GENOME SEQUENCE OF *C. ARABICA* FOR DOWNSTREAM GENOMIC ANALYSIS [4]

Abstract

Despite the fact that *C. arabica* accounts for higher proportion of the world's coffee production than *C. canephora*, only recently has the first high-quality draft genome of *C. canephora* coffee been completed while *C. arabica* still lags behind. Obtaining a whole genome sequence for Arabica coffee will assist in identifying new options for the genetic improvement of coffee. A whole genome assembly of K7 – the most popular arabica coffee variety in Australia was achieved by sequencing using both Illumina short reads (Pair-End and Mate-Pair) and PacBio long reads. Assembly was performed using a range of assembly tools resulting in 76,409 scaffolds with a scaffold N50 of 54,544 bp and a total scaffold length of 1,448 Mb. Validation of the genome assembly using different tools showed high completeness of the genome. BWA analysis demonstrated that > 98% of the short reads mapped to the genome and > 93% were marked as properly paired; GMAP analysis indicated that > 99% of the CDS and transcriptome sequences mapped to the *C. arabica* draft genome and 89 % of BUSCOs were present. The assembled genome was annotated using AUGUSTUS yielded 99,829 gene models. The outcome of the study could be beneficial for downstream genomic analysis in *C. arabica*.

[4] This chapter contains information that is in Tran, H. T. M., Ramaraj, T., Furtado, A., Lee, L. S., and Henry, R. J. (2018). Use of a draft genome of coffee (Coffea arabica) to identify SNPs associated with caffeine content. *Plant Biotechnology Journal*. https://doi.org/10.1111/pbi.12912.

5.1 Introduction

Coffee is an important crop and world coffee production relies on only two species: *Coffea canephora* (robusta) and *C. arabica* (arabica). Recently, the first high-quality draft genome of robusta coffee was completed (Denoeud et al., 2014). This genome sequence provides a reference for analysis of the genomes of other *Coffea* species and genotypes. However, *C. canephora* is a diploid species while *C. arabica* is tetraploid. Obtaining a whole genome sequence for Arabica coffee will assist in identifying new options for the genetic improvement of coffee.

In recent years, sequencing technologies and assembly methods have been rapidly developed. Short read technology (Illumina) has been the most popular due to the low cost, simplicity, accuracy and high throughput (Mavromatis et al., 2012). However, short read sequencing technologies have been problematic when dealing with genomes that have large repetitive regions even within relatively small genomes (Nagarajan and Pop, 2013). This leads the incomplete assembly of genomes due to high fragmentation in regions of repeat sequences (Chain et al., 2009). The benefit of long-range mate-pairs was demonstrated in the assembly of the sacred lotus (*Nelumbo nucifera*) genome (Abbott and Butcher, 2012). Long sequencing reads generated by single-molecule sequencing technology offer the possibility of dramatically improving the contiguity of genome assemblies (Zimin et al., 2016). Combination of different sequencing platforms has become popular and recent studies have combined data of different read length and from several different sequencing platforms (Koren et al., 2012). Long sequencing reads alone can be used to generate complete and accurate *de novo* assemblies of the genomes of bacteria (Brown et al., 2014; Chin et al., 2013; English et al., 2012; Koren et al., 2013), chloroplast genome (Ferrarini et al., 2013; Li et al., 2014; Stadermann et al., 2015), and plant and human genomes (Berlin et al., 2015; Pendleton et al., 2015).

The completeness of genome assembly depends on several factors including genome size, repeat content, paralogy, and heterozygosity (Michael and VanBuren, 2015; Sims et al., 2014). Compared to animal genomes, plant genome are often more complex due to their larger genome size, higher ploidy (occurs in 80% of all plant species, but only a few of them have been sequenced), and higher rates of heterozygosity and repeats and complex gene content. This is further complicated by the presence of a large number of chloroplasts and mitochondria organelles and the challenge of producing from plants the high quality DNA required for sequencing (Schatz et al., 2012). Plant species have different levels of ploidy. For example, *C. arabica* - a tetraploid species - has multiple genome copies which will become a major challenge for discovering true polymorphisms between genomes and for discriminating them from artificial polymorphisms between genome copies within the same genotype (Trick et al., 2009).

Assembly paradigms such as greedy approaches, graph-based approaches, OLC graphs, De Bruijn graphs and String graphs, and corresponding assembly software as well as validation of genome assembly have been well developed (Nagarajan and Pop, 2013; Paszkiewicz and Studholme, 2010; Simpson and Pop, 2015). Genome assembly typically has two-stages: assembly of reads to contigs, and joining contigs into scaffolds (Hunt et al., 2014). Additional steps are gap closing to fill gaps generated by the scaffolding, and anchoring onto a genetic map to build the final pseudo-molecules (Madoui et al., 2016). Several assemblers (mappers and scaffolders) and their weaknesses as well as strengths have been evaluated (Hunt et al., 2014; Salzberg et al., 2012; Zhang et al., 2011). There was a large variation in the quality of results found depending on the tool and dataset used. The quality of assembly depends largely on sequence depth, sequencing methodology and genome complexity (Hunt et al., 2014). Algorithms and software for genome assembly have been developed to addresses these issues. Examples include, software for genome assembly using PacBio long reads alone (Koren and Phillippy, 2015) or to overcome the problem of high heterozygosity (Kajitani et al., 2014; Pryszcz and Gabaldon, 2016), or to deal with the issue of inappropriate collapsed contigs in polyploid plants by estimating the ploidy of contigs based on the allele proportion (ConPADE) (Margarido and Heckerman, 2015).

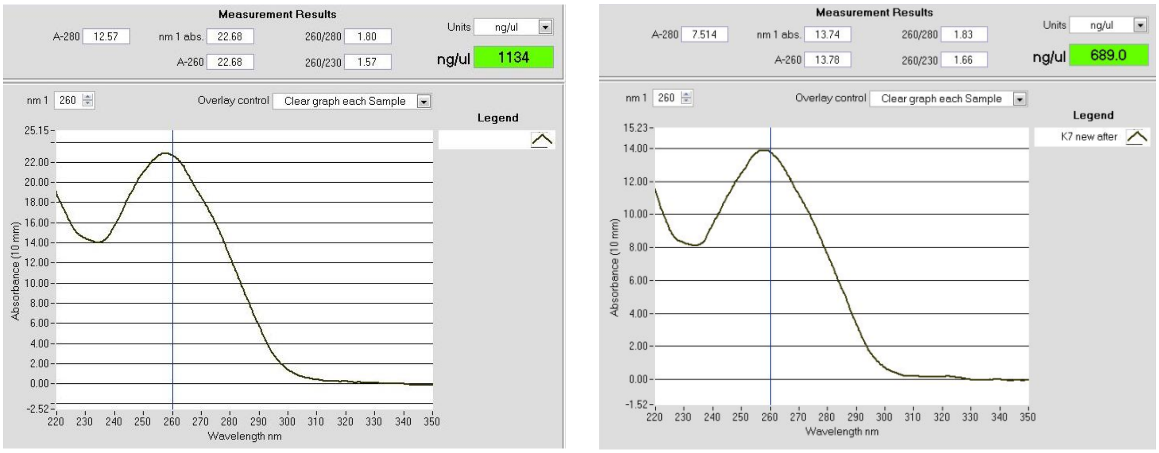
Evaluating the outputs of a *de novo* assembly can be difficult in the absence of a reference genome and additional genetic resources. Commonly used statistics such as N50 only provide sizing information, with no indication of correctness (Abbott and Butcher, 2012; Ekblom and Wolf, 2014). Information from remapped paired-end or mate-pair data can be used to detect errors in the assembly (Ekblom and Wolf, 2014). Using available genomic resources such as CDSs, cDNA, ESTs or protein data of the same species or closely related species is another approach to validate sequence accuracy and for correcting the scaffolding where genes span across contigs (Ekblom and Wolf, 2014). CEGMA (Core Eukaryotic Genes Mapping Approach) (Parra et al., 2007; Parra et al., 2009) and BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simao et al., 2015) are comparative genomic approaches and complement the commonly used N50 length to assess completeness of genome assembly for genomes that may have little or no annotation. CEGMA is based on a set of 248 core proteins that are highly conserved in a wide range of eukaryotes (Parra et al., 2009), while BUSCO uses 3,023 genes for vertebrates, 2,675 for arthropods, 843 for metazoans, 1,438 for fungi and 429 for eukaryotes (Simao et al., 2015). Genome annotation includes two distinct phases: the ‘computation” phase where available genomics resources such as ESTs, cDNA, proteins are aligned to the genome and *ab initio* and/or evidence-driven gene predictions are generated; and the “annotation” phase where these data are synthesised into gene annotations (Yandell and Ence, 2012).

This chapter describes steps involved in arabica coffee genome assembly, validation and annotation. Outputs of different genome assemblers were also compared. The results would enable downstream genomic analysis toward genetic dissections of arabica coffee quality, and also allow this genotype to be linked to international coffee genomics research outcomes.

5.2 Materials and Methods

5.2.1 DNA extraction and sequencing

A leaf sample of Arabica variety – K7 – the most widely grown coffee variety in NSW and a Kenyan commercial variety (Omondi et al., 2016) was collected from Green Cauldron 330 Federal Rd, Federal NSW 2480 in Oct 2014. DNA extraction was performed following the method described by Healey et al. (2014) with some modification in storage time and the use of 2-mercaptoethanol (0.3%) and PVPP (2%). The DNA from several extractions was mixed and precipitated to obtain an adequate concentrated of DNA required for Illumina and especially Pac Bio sequencing. DNA quality and quantity was assessed using a Nanodrop and agarose gel electrophoresis (Figure 5.1 and 5.2).



Measurement Results

A-280 12.57 nm 1 abs. 22.68 260/280 1.80 Units ng/ul

A-260 22.68 260/230 1.57 ng/ul 1134

nm 1 260 Overlay control Clear graph each Sample

Measurement Results

A-280 7.514 nm 1 abs. 13.74 260/280 1.83 Units ng/ul

A-260 13.78 260/230 1.66 ng/ul 689.0

nm 1 260 Overlay control Clear graph each Sample

Legend

K7 new after

Figure 5.1 DNA quantity and quality measured on Nanodrop

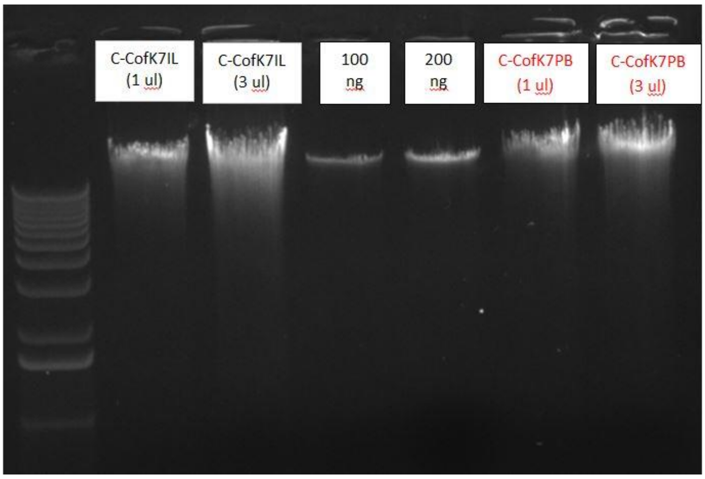


Figure 5.2 DNA quality and quantity measured on agarose gel

5.2.1.1. Illumina sequencing: At least 25 µg of total DNA was dissolved in 10 mM Tris and used to prepare sequencing libraries. Sequencing was performed using an Illumina HiSeq 2000 with three libraries including:

(1) C-CofK7IL for Truseq PCR free Library HiSeq 2X 100bp paired-end sequencing using adapter 04 with an insert size of 350 bp, concentration of 355 ng/µl with volume of 60 µl to make a total of 21,300 ng of DNA;

(2) C-CofK7IL3 for Nextera 3kb mate pair (indexed) library using adapter 13 and then HiSeq 2X 100bp paired-end sequencing, concentration of 447 ng/µl with volume of 90 µl to make a total of 40,230 ng DNA

(3) C-CofK7IL8 for Nextera 8kb mate pair (indexed) library using adapter 18 and then HiSeq 2X 100bp paired-end sequencing, concentration of 422 ng/µl with volume of 90 µl to make a total 37,980 ng of DNA.

For more precise measurement of DNA quality and quantity, DNA was examined on Qubit QC to confirm sufficient quantity for sequencing (Table 5.1):

Table 5.1 DNA passed the Qubit QC.

Sample Name	Stock Concentration (ng/ul)	Sample Volume Available (ul)	Total DNA Available (ng)
C-CofK7IL	355.0	60.0	21300
C-CofK7IL3	447.0	90.0	40230
C-CofK7IL8	422.0	90.0	37980

Gel portions for the 3kb and 8kb bands cut from the gel during library preparation were estimated to have an average size of 5.3kb and 10.4kb, respectively using a Bioanalyzer. The two Nextera Mate Pair libraries were run together in one lane and de-multiplexed evenly with 49% of reads identified for each library. The single TruSeq PCR-Free library was run in a lane of its own.

5.2.1.2. Pac Bio sequencing: A sample at a concentration of 303 ng/µl (Qubit QC) and total DNA of 31.8µg was sequenced in 15 SMRT (Single Molecule Real-Time) cells using P6-C4 chemistry (20 KB protocol) resulting in 927,726 reads with an N50 read length of 13,127 bp and a mean read length of 8,312 bp (Table 5.2).

Table 5.2 Information of PacBio sequencing

QC	20 KB protocol	P6-C4 Chemistry
<1%	Short inserts (11-100 bp)	0.01%
Approx. 50K	Number of reads	927,726
	N50 read length	13,127
Ave. 10 KB	Mean read length	8,312

5.2.2 Genome Assembly

5.2.2.1. Assembly of Pair End (PE) and Mate Pair (MP) Illumina data: Initially, assemblies of Illumina PE and MP reads were generated using CLC Genomics Workbench Version 9.5 (CLC Bio, www.clcbio.com). The same dataset was assembled with alternative assembly algorithms such as ABySS (Simpson et al., 2009), PLATANUS (Kajitani et al., 2014) and SOAP*denovo*2 (Luo et al., 2012). SOAP*denovo*2 resulted in improved assembly statistics, and was subjected to GAPCloser (a module in SOAP), then SSPACE Standard (Boetzer et al., 2011) for scaffolding, then GAPCloser again using Illumina PE reads.

5.2.2.2. Assembly of PacBio data: The final Illumina assembly was further scaffolded with PacBio long reads using SSPACE-LR (Long Reads) (Boetzer and Pirovano, 2014) for scaffolding, followed by GAPCloser using Illumina reads, and finally gap filled with PBJelly2 (English et al., 2012) using PacBio long reads (Figure 5.3)

5.2.2.3. Settings of each assembler:

CLC settings: Raw reads of Illumina data were imported to CLC GWB for *de novo* assembly using different settings as follows: Mode: Fast (create simple contig sequence) and slow (map reads back to contigs). Word size, bubble size and minimum contig length: 45-98-1000 and 64-120-1000 applied scaffolding and no scaffolding.

ABySS, PLATANUS, SOAPdenovo2: To attempt assembly of Illumina reads, three short read assemblers, ABySS (Simpson et al., 2009), PLATANUS (Kajitani et al., 2014), and SOAP*denovo*2 (Luo et al., 2012) were selected based on their heavy use in the genome assembly community. All three were primarily developed to assemble short reads (50 – 200 base pairs) from next generation sequence data, mainly Illumina sequencing technology. They all use de bruijn graph theory to effectively handle and assemble millions of short reads. For graph- based methods, the choice of k-mer size (base string length) can affect the contiguity and/or completeness of an assembly (Chikhi and Medvedev, 2014)

Firstly, ABySS assembler version 1.3.7 was used to assemble Illumina reads with different k-mer values (from 50 bp up to 95 bp with interval of 5 bp). The rest of the parameters were set to default.

Secondly we used PLATANUS to assemble Illumina reads with number of CPUs set to 16 and memory limit for k-mer distribution was set to 100 GB. The rest of the parameters were set to default. Lastly we used SOAP*denovo*2 to assemble the Illumine data set using default settings. Assemblies from these three assemblers were compared in terms of basic assembly metrics, such as (1) Number of sequences, (2) Genome size, (3) N50, and (4) Max and Min sequence length. GAEMR (Genome Assembly Evaluation and Metrics Reporting) was used to generate the basic assembly metrics. Comparison clearly indicated that assembly from SOAP*denovo*2 was superior to assembly from ABySS and PLATANUS based on the basic assembly metrics. Hence SOAP*denovo*2 assembly was used for all further downstream analysis.

GAPCloser, SSPACE and PBJelly2: During the process of assembling short reads it is very common for assemblers to introduce N's to scaffold two or more contigs (referred to as the scaffolding process). Final assembly from any one of the short reads assemblers (in this case, final assembly from SOAP*denovo*2) will contain several stretched of N's. One common practice in genome assembly process is to try to replace these N's with actual base pairs (A, C, G, T). There are several tools that can accomplish this. We used GAPCloser, which takes as input, the final assembly and a configuration file which has details on the read length, read type, insert size and the path to the short reads from which the assembly was constructed. GAPCloser was designed to work only with short reads data. Attempting to use long reads from technologies like, Sanger, 454, PacBio, MinION results in a failed run. GAPCloser maps short reads back to assembly in attempting to replace N's with A, C, G, or T. In this process most of the gaps were filled in but larger gaps still remain since Illumina short reads cannot span those really long regions.

SSPACE (Boetzer et al., 2011; Boetzer and Pirovano, 2014) has two modes, SSPACE-STANDARD & SSPACE-LR. SSPACE-STANDARD is used for scaffolding using short reads mainly data from Illumina sequence technology. SSPACE-LR, LR stands for Long Reads which is used for scaffolding NGS assemblies using data from long read technologies such as PacBio and MinION. Here, the gap filled SOAP*denovo* assembly is further scaffolded using SSPACE-STANDARD program. It takes in as input the gap filled SOAP*denovo* assembly and a configuration file which has details on the read length, read type, insert size and the path to the short reads. The final output will be an assembly FASTA file which will also have stretches of N's that got introduced during the SSPACE-STANDARD scaffolding process. After this another GAPCloser was run for the second time in an attempt to fill in gaps now introduced during the SSPACE-STANDARD scaffolding step. Following this, 6X PacBio data was incorporated into the assembly process by using it to scaffold the assembly from the previous step using SSPACE-LR.

Finally to finish the assembly, the assembly generated from SPPACE-LR was gap filled with GAPCloser (Illumina Data) and PBJelly2 (PacBio Long Reads). PBJelly2 software is used to fill gaps in the assembly using PacBio long reads. The program is designed to handle PacBio data taking its error model into consideration. It uses a PacBio read data specific aligner called BLASR (Chaisson and Tesler, 2012) to map PacBio reads to the assembly and attempt to replace N's with A, C, G, or T. The final output from PBJelly2 was the first draft version of the genome.

5.2.3 Validation of genome assembly

Validation of assembly was assessed using three different approaches:

(1) PE Illumina reads were remapped to detect errors in the assembly using BWA (Burrows-Wheeler Aligner) (Li and Durbin, 2009).

(2) Available coffee genomic resources such as *C. canephora* CDS sequences (http://coffee-genome.org/coffeacanephora) and *C. arabica* (K7) PacBio transcriptome data (not published) were used to map back to the draft genome using GMAP (Genomic mapping and alignment program) (Wu and Watanabe, 2005).

(3) The BUSCO (Benchmarking Universal Single-Copy Ortholog) (Simao et al., 2015) strategy was used to test the completeness of the genome assembly and gene space using the plant specific profile. This approach makes use of single copy genes expected to be present in plants (956 genes).

5.2.4 Genome annotation

Final genome assembly was repeatmasked using REPEATMODELER (Smit and Hubley, 2008-2015) and REPEATMASKER (Smit et al., 1996–2010). MAKER-P (Campbell et al., 2014) was run on the repeatmasked genome with SNAP (Korf, 2004) and AUGUSTUS (Stanke and Morgenstern, 2005). The gene prediction programs, SNAP and AUGUSTUS used Arabidopsis thaliana HMM (Hidden Markov Model) and tomato respectively. ESTs of *C. arabica* (http://www.ncbi.nlm.nih.gov/nucest), and CDSs of *C. canephora* (http://coffee-genome.org/coffeacanephora) and 96,521 in-house *C. arabica* (K7) PacBio transcriptome sequences were used as evidence to guide the annotation process.

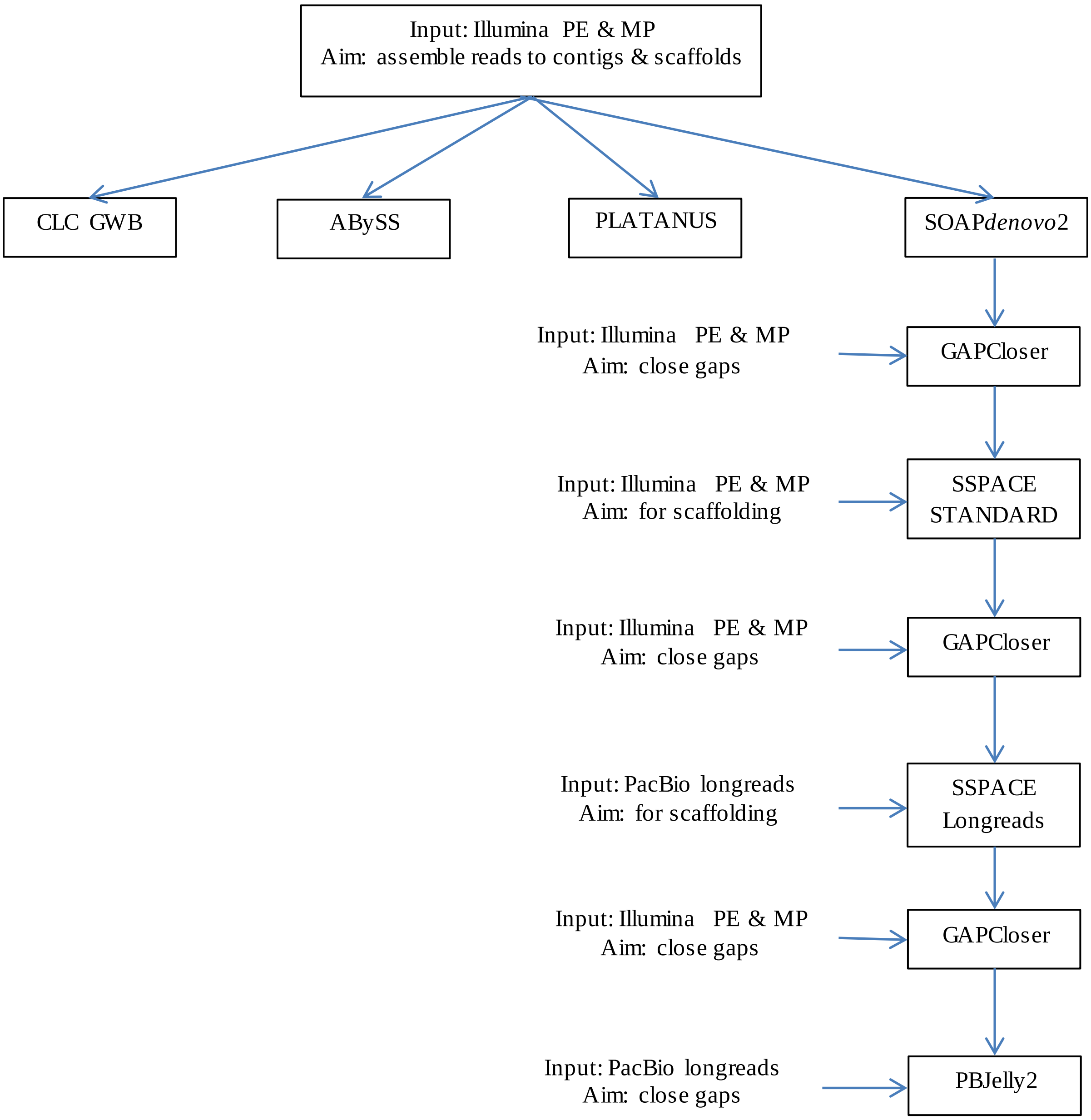


Figure 5.3 Work flow in coffee genome assembly with Illumina and PacBio data

5.3 Results and discussion

5.3.1 Sequencing details

A total of 385.10^6 reads of 100 bp Illumina PE estimated at 64x coverage (insert size of 350 bp) and 220.10^6 reads (equivalent to 36x) and 223.10^6 reads (equivalent to 37x coverage) of 100 bp Illumina MP of insert size of 3 kb and 8 kb, respectively were generated on an Illumina HiSeq 2000. Over 1.37 million PacBio reads, equivalent to 6x genome coverage were also generated using the SMRT P6-C4 chemistry (Table 5.3 and Figure 5.4). These data sets were the input sequence used for hybrid assembly of *C. arabica*.

Table 5.3 Parameters of Illumina data and PacBio data

Platform	Library	Insert size (bp)	Read length (bp)	No of reads	Coverage (*)	GC content	Ave. Phred score
Illumina	Paired end	350	100	385.10^{6}	64 x	37%	39
	Mate pair	3 Kb	100	220.10^{6}	36 x	40%	39
		8 Kb	100	223.10^{6}	37 x	39%	39
PacBio	SMRT (15)	- No of bases: 7,688,079,960 - N50 read length: 8,180 - Mean read length: 5,579		1,377,852	6 x	37%	

(*) Based on a genome size estimate of 1.3 Gbp

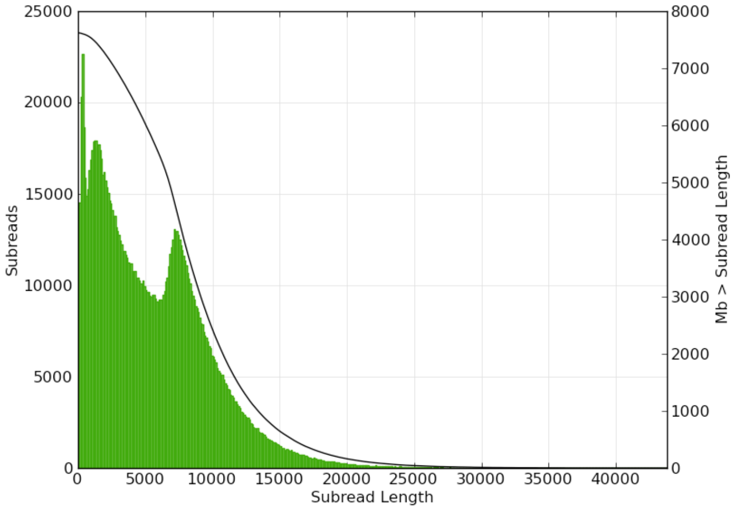


Figure 5.4 Length distribution of extracted subreads.

(The length distribution of all the subreads extracted from the complete dataset by using SMRT Analysis. Subread length is given in bp).

Phred score is an indicator of the read quality, Phred score of 30 indicates a probability of 0.1% of a wrong base call (Dohm et al., 2008). In this study, the Phred score of 39 for both paired end and mate pair indicated good quality of reads. GC content in K7 arabica variety is relatively low (37-40%). The GC content difference is a primary factor for non-random sequencing-depth distribution (Li et al., 2010a). The GC content was found related to read coverage and to the effect of GC bias which has an effect on the completeness of assembly (Chen et al., 2013). Both very low (<20%) and very high GC regions (>80%) had a relatively lower sequencing depth. However, these regions only account for a minor fraction (e.g. 0.004% for the panda genome, 0.079% for dog genome, 0.095% for human genome and 0.015% for mouse genome) (Li et al., 2010a). The sequencing irregularities due to sequence-dependent coverage biases and non-uniform error rates will cause unexpectedly low coverage regions (e.g., Illumina sequencers have lower coverage in low-GC regions) and consequently more gaps in an assembly (Schatz et al., 2010).

The coverage of Illumina data of 137x (both PE and MP) should be sufficient for assembly (Table 5.3) based on literature. Ekblom and Wolf (2014) suggested that the total read coverage should be more than

100x for large and complex genomes, while for bacteria and yeast it required only 35x – 50x (Desai et al., 2013). Illumina input data are short reads which may result in fragmented assembly (Schatz et al., 2010), but this is complemented by an additional long reads of PacBio albeit with a low coverage (6.0x). According to Faino and Thomma (2014), a combination of different sequencing platforms including Illumina reads at 30x coverage (PE reads of 100 –150 bp derived from a library with 500 bp inserts and MP reads of 50–100 bp from a 5 kb insert library) and 5–10x coverage using SMRT sequencing of a 20 kb insert would give optimal assembly statistics.

5.3.2 Genome assembly

5.3.2.1 Comparison among different genome assemblers using Illumina data

Genome assembly was done using different software. First, CLC GWB was employed. CLC GWB is commercial software for genome assembly which is usually more user-friendly than freely available programs and thus readily used by researchers with limited bioinformatics skills. However, its drawbacks are licensing cost involved, and it acts even more like ‘black box’ solutions as the algorithms are unknown (Ekblom and Wolf, 2014). Other highly used assembly software that relies on De Bruijn graph algorithms includes ABySS and SOAP*denovo*, and more recently, PLATANUS. The assembly outputs of these four assemblers are provided in Table 5.4.

Table 5.4 Assembly statistics among different assemblers with Illumina sequencing reads

Parameters	**CLC** (1)	**ABySS** (2)	**PLATANUS** (3)	**SOAP*denovo*2** (4)
Contig metrics				
# contig	189,330	628,389	62,453	276,322
Contig N50	5,490	1,188	3,627	6,284
Ave contig size		798.3	2,631.1	2,944.1
Scaffold metrics				
# scaffold	189,330	139,505	101,960	120,578
Scaffold N50	5,490	8,987	17,991	16,694
Min scaffold size	985	1,000	1000	1000
Max scaffold size	105,805	218,433	501,363	176,350
Ave scaffold size incl. gaps	4,230	4,265.7	4,654.4	8,154.9
Ave scaffold size w/o gaps	1,263	3,596.0	4,295.5	6,746.9
Genome incl. gaps *	800,884,967	595,091,719	290,679,196	983,303,576
Genome w/o gaps **	604,221,306	501,664,397	268,267,094	813,528,413

* Total genome length including gaps; ** Total genome length without gaps; (1) *De novo* settings of slow mode, word size, bubble size and minimum contig length of 64-120-1000, and including scaffolding; (2) k-mer of 80 bp, minimum scaffold size of 1000 bp, scaffolded with MP; (3) (4) minimum scaffold sizes of 1000 bp.

The CLC GWB applied to both PE and MP reads assembly, yielding a total length of 800,884,967 bp (61.61% of the genome size) and a fairly good N50 value (5,490 bp) (Table 5.4). According to

Ribeiro et al. (2012), MP libraries are capable of resolving repetitive regions and structural variants while increasing the accuracy and size of assembled contigs. In this study, the MP helps improve both contig N50 and genome length. The CLC assembler has previously been used to assemble bacteria genomes (Brown et al., 2012; Hwang et al., 2014) and more complex plant genomes, such as barley (International-Barley-Genome-Sequencing-Consortium et al., 2013) and rubber (Rahman et al., 2013). However, in the present study, the output of this assembler seems to be not satisfactory. Besides, this assembler is not suitable for hybrid assembly where Illumina reads and PacBio long reads can be combined to improve assembly quality (i.e. resolved ambiguities). Other assemblers were therefore performed for improvement.

ABySS is a *de novo,* parallel, paired-end sequence assembler that is designed for short reads and large genomes, and low RAM occupancy as it transfers the sequence reads into binary format. The ABySS algorithm proceeds in two stages. First, all possible substrings of length k (termed k-mers) are generated from the sequence reads. The k-mer data set is then processed to remove read errors and initial contigs are built. In the second stage, mate-pair information is used to extend contigs by resolving ambiguities in contig overlaps (Simpson et al., 2009). Although a number of previous studies found that ABySS assembler generated some of the best assembly statistics when only PE Illumina reads were used for different bacteria (Boetzer et al., 2011; Boisvert et al., 2010; Utturkar et al., 2014) or 20 Gb white spruce (Birol et al., 2013), the present study showed that ABySS assembler seems to fail giving sufficient results. After the second step of scaffolding, the scaffold N50 was reasonably high (8,987 bp). However, the total genome length without gaps only covered 38.5% of the estimated genome of arabica, which is too low.

PLATANUS, another assembler for short reads recently developed by Kajitani et al. (2014) to assemble heterogeneous diploid genomes, was used to seek improvement. The number of scaffolds and scaffold N50 in PLATANUS were slightly better than others, however, the total genome length was small, only covering 20.6% of the expected genome size (Table 5.4). Patel et al. (2015) compared PLATANUS with ALLPATHS-LG in genome assembly of highly heterozygous grape species using sequences from three libraries (180, 600, 3000 bp). Results showed that the number of N/L50 and length the scaffolds and contigs were greater for PLATANUS than found for the ALLPATHS-LG assemblies. However, Chin et al. (2016) found genome assembly results using PLATANUS were poor for Arabidopsis, grape and fungus, especially only less than 1% of the expected genome size assembled for grape. PLATANUS was also tested on yeasts, fungus and Arabidopsis using PE and MP reads (Pryszcz and Gabaldon, 2016). Results showed that PLATANUS deals well over the full spectrum of loss of heterozygosity, but it fails at divergences above 10%. Moreover, the assemblies returned by PLATANUS are more fragmented than those produced by another new assembler called Redundans (Pryszcz and Gabaldon, 2016). When using additional Illumina mate-pair reads and long reads from PacBio, the neem tree genome quality was improved with PLATANUS, and PLATANUS performed better than SOAP*denovo*2 in regards to assembly statistics for this heterozygous genome (Krishnan et al., 2016). In this study, as the results of assembly using PE and MP were not promising, PacBio reads were not used and therefore another assembler was performed.

SOAP*denovo*2 outperformed other assemblers with a total genome length (without gaps) of 813,528,413 bp covering 62.6% of the estimated genome size of 1.3 Gb even though the scaffold N50 (16,694 bp) and number of scaffold (120,578) were slightly lower than PLATANUS (Table 5.4). SOAP*denovo*2 performed well on short reads (Zhang et al., 2011) and being a popular software (Ekblom and Wolf, 2014). SOAP was first developed by Li et al. (2010b) and it was compared with other assemblers (ABySS, Velvet, EULER-SR, SSAKE and Edena) on human genome data. Results showed SOAP obtained better N50 contig, higher genome coverage and shorter running time, higher assembly accuracy; however, it requires a much higher peak memory usage (Li et al., 2010b). SOAP*denovo*2 was then designed to (1) reduce memory consumption in graph construction, (2) resolve more repeat regions in contig assembly, (3) increase coverage and length in scaffold construction, (4) improve gap closing, and (5) optimise for large genomes (Luo et al., 2012). According to Simpson and Durbin (2012), SOAP*denovo*2 gave N50 smaller than ABySS and lower assembly completeness, but higher assembly accuracy on nematode genome assembly. In this study, SOAP*denovo*2 outperformed other software and was applied to next steps for assembly improvement.

5.3.2.2 A draft genome assembled using both Illumina and PacBio reads

GAPCloser (a module in SOAP) was deployed to address some of the gaps (N's) emerging from scafolding steps in SOAP*denovo*2. SSPACE Standard (Boetzer et al., 2011) was then used for scaffolding, followed by GAPCloser again to fill N regions in the scaffolds using Illumina PE and MP reads. Basic assembly metrics (i.e. Contig N50, scaffold N50, total genome length, the total gap length) were improved after each step of gap filling and scaffolding (Table S5.1). The addition of PacBio data, even with low coverage, showed improvement in terms of assembly metrics and resulted in the draft genome as presented in Table 5.5.

The total size of the assembled contigs was 1,167 Mb which was 90% of the estimated genome size (i.e. 1,300 Mb) while that of scaffolds was 1,448 Mb - 11% larger than the estimated genome (Table 5.5). The draft genome included 76,409 scaffolds with N50 scaffolds of 54,544 bp and longest scaffold of 769,411 bp. The draft genome had a total size of gaps of 281 Mb, accounting for 21% of the estimated genome.

Table 5.5 Characteristics of the K7 arabica draft genome assembly

Estimated genome size (Mb)	1,300
Chromosome number (2n = 4x)	44
Total size of assembled contigs (Mb)	1,167
Number of contigs	265,687
Largest contig (bp)	186,701
N50 length (contigs) (bp)	12,184
Number of scaffolds	76,409
Total size of assembled scaffolds (Mb)	1,448
N50 length (scaffolds) (bp)	54,544
Longest scaffold (bp)	769,411
Number of gaps	189,278
Mean gaps length (bp)	1,485
Total size of gaps (Mb)	281
GC content (%)	37

GapCloser was a module of SOAP*denovo*2 which assembled sequences iteratively in the gaps to fill large gaps (Luo et al., 2012). SSPACE and SOAP*denovo*2 are fast to run on all datasets and perform well with respect to correct versus incorrect joins. SSPACE comfortably has the most citations of any of the scaffolding tools and is also the easiest of the tools to install and run. SOAP*denovo*2 and SSPACE were found to generally outperformed other scaffolders (Hunt et al., 2014). SOAP*denovo* was also used to assemble the giant panda (~2.4 Gb) with short reads resulting in an assembled N50 contig size reaching 40 kb, and an N50 scaffold size of 1.3 Mb and with 54.2 Mb remained unclosed (Li et al., 2010a).

PacBio long-read data has emerged as a way of filling N regions in scaffolds (English et al., 2012). According to Ekblom and Wolf (2014), the coverage and sequencing platform selection are specific to each project which requires basic knowledge on genome size, sequencing error rates, repeat content and the degree of genome duplications to make decision. If the genome has a high repeat content or a high degree of duplications, a larger amount of long-insert data is needed for correct assembly (Ekblom and Wolf, 2014). In this study, PacBio data has low coverage and was used for scaffolding using SSPACE Longreads. The output of this step was subjected to GapCloser again using Illumina (PE and MP) reads. The PBJelly method (English et al., 2012), a gap-filling approach that takes scaffolds generated by SSPACE Longreads and fills scaffold gaps using long reads. PBJelly2 can use long reads to correct erroneously scaffolded contigs and to close gaps, provided that a quality score is associated to each read (English et al., 2012). The addition of PacBio data, even with low coverage, showed a significant improvement in terms of assembly metrics (Table S5.2). With the addition of only 10x coverage of PacBio longreads (20-kb library) to the Illumina assembly, genome assembly statistics for fungus showed the contig N50 length increased up to 25 times while the gaps reduced approximately three times (Faino and Thomma, 2014). In the current study, scaffold N50 was only double with the addition of 6x PacBio longreads. The total scaffold length was 11% higher than estimated genome size of arabica of 1.3 Gb. This expansion phenomenon was also observed in several crops such as grapevine genome assembly (4.8%) (Jaillon et al., 2007), walnut genome assembly (10-24%) (Martınez-Garcıa et al., 2016) or in pineapple (32-48%) after the first and second draft assembly (Redwan et al., 2016). One reason could be that the high ploidy level or the high heterozygosity rate in arabica made the genome assemblers assume that the genome is diploid (Redwan et al., 2016). Another reason could be that SOAP*denovo*2 over-estimated gap sizes during scaffolding. The gap (N content) seems to be slightly high, but since several rounds of scaffolding were performed, this issue could be happened. Besides, it is most likely that these gaps are complex regions that the assembler could not resolve, or there was not enough sequence coverage, which could be overcome by adding more sequence data.

In summary, the K7 arabica genome assembled was higher than estimated arabica genome. After being optimised with many steps using different assemblers, scaffolders, and gap closers and with different sequencing reads, the gap length was still large and accounts for 21% of estimated genome. This is however an encouraging outcome given the challenges in arabica genome assembly such as high ploidy level, heterozygosity, low coverage and fairly large genome. Fragmented genome assembly was also reported for *C. canephora* (13,345 scaffolds) (Denoeud et al., 2014) and other species. In date palm with half the genome size of arabica, the draft genome was still in 57,277 scaffolds and N50 scaffold of 30,480 bp (Al-Dous et al., 2011). Similarly, walnut with half the genome size of arabica, the draft genome was with 186,636 scaffolds (Martınez-Garcıa et al., 2016). For larger genome size like rubber (2.15 Gb haploid genome), the genome assembly was with 608,017 scaffolds and N50 scaffold of 2,972 bp (Rahman et al., 2013). Even the Arabidopsis genome, which is arguably the best-assembled plant genome, is still in 102 contigs with a total gap length of at least 185,644 bp (Michael and VanBuren, 2015). In the study of panda genome, Li et al. (2010a) estimated that around 3.6Mb (0.15%) of tandem repeat sequences might be missing in the current panda genome assembly.

Genome assembly for ploidy crop like arabica is faced with certain challenges including genome size, repeat content, paralogy, and heterozygosity (Michael and VanBuren, 2015). In addition, the low coverage of input data in the present study made the arabica genome assembly even more challenging. Arabica has the genome size estimated at 1.3 Gb (Kochko et al., 2010) which is fairly large compared to other common genome species such as *Arabidopsis*, rice, grape and sorghum (Michael and VanBuren, 2015). A large and complex genome has large repetitive elements and covers a large fraction of the genome resulting in ambiguities in the scaffolding step (Madoui et al., 2016). A tetraploid species like arabica has different “copies” which tend to be less similar while the algorithms and software developed for assembly were mainly developed for haploid or diploid genomes may lead to the risk of information loss when using it for genome assembly in polyploid crops (Margarido and Heckerman, 2015). Since arabica is an allotetraploid crop, it is expected to show heterozygosity conditioned by two different alleles derived from two different progenitors (Mishra et al., 2011). Paralogous regions and heterozygous sites create 'bubbles' during genome assembly where two or more regions that are highly similar assemble together, and the adjacent dissimilar regions assemble separately but eventually merge again (Michael and VanBuren, 2015) make it hard for assembly. It is suggested that coverage for finished assemblies was 50x for long reads (Koren and Phillippy, 2015) while there was only 6x coverage for PacBio data in the present study.

5.3.3 Validation of genome assembly

Illumina short reads (PE) were aligned back to the assembled genome using BWA to evaluate the genome completeness and to detect errors in the assembly. More than 98% of the short reads mapped to the genome and more than 93% were marked as properly paired (Table 5.6).

The *C. canephora* CDS sequences and *C. arabica* (K7) PacBio transcriptome data were used to map back to the draft genome using GMAP. More than 99% of the CDS and transcriptome sequences were mapped to the *C. arabica* draft genome in which 85.1% of the CDS sequences and 88% of transcriptome sequences with >= 90% identity and query coverage were mapped (Table 5.6).

Table 5.6 Validation of draft genome using BWA, GMAP and BUSCO

Results of read remapping using BWA		
Read alignment metrics		
Total number of reads mapping back	98.4%	
Reads properly paired	93.0%	
Gene capture analysis using GMAP (*C. canephora* CDS sequences)		
Parameters	CDS sequences in *C. canephora*	Transcriptome of *C. arabica* (K7)
Total number of sequences	25,574	96,521
Total number of sequences mapping to *C. arabica* (draft genome)	> 99.0%	> 99.0%
Total number of sequences mapping to *C. arabica* (draft genome) (>= 90% identity and query coverage)	> 85.0%	> 88.0%
BUSCO analysis		
Parameters	Number of BUSCOs mapped	Percentage
Complete single-copy BUSCOs (C)	858	89%
Complete duplicated BUSCOs [D]	553	57%
Fragmented BUSCOs (F)	23	2.4%
Missing BUSCOs (M)	75	7.8%
Total BUSCO groups searched (n)	956	

BUSCO analysis against a plant-specific database of 956 genes was also used to assess the completeness of the draft genome and identified 858 (89 %) complete BUSCOs, of which 553 (57 %) were duplicated. A further 23 fragmented BUSCOs were identified (Table 5.6)

Although the assembly metrics such as N50 and contig numbers are widely used for the assembly evaluation, they may not accurately reflect the quality of an assembly (Baker, 2012; Nagarajan and Pop, 2013). They merely indicate contiguity and contain no information on assembly accuracy (Ekblom and Wolf, 2014). To assess the genome completeness or to detect errors in the assembly, Illumina short reads (PE) were aligned back to the assembly outcomes using BWA (Li and Durbin, 2009). The genome quality assessment was evaluated by remapping short reads to the final draft using BWA. Results showed high percentage of reads mapped back to the draft suggesting that most of the reads were incorporated into the genome and thus most of the genome were assembled.

Validation of the assembly or gene space completeness can be based on the genomic resources available for coffee such as *C. canephora* genes or ESTs data base of *C. arabica* (187,739 ESTs from NCBI) aligned to *C. arabica* draft genome using GMAP. In this study, *C. canephora* CDS sequences and *C. arabica* (K7) PacBio transcriptome data were used to map back to the draft genome. High percentage of the CDS and transcriptome sequences were mapped to *C. arabica* draft genome (99%) with high identity (>= 90%) and query coverage (90%) indicating the completeness of the draft genome (Table 5.6).

Another approach to assess the completeness of genome assembly was running CEGMA (Core Eukaryotic Genes Mapping Approach) (Parra et al., 2007; Parra et al., 2009) or BUSCO (Benchmarking Universal Single-Copy Orthologsplant conserved genes) (Simao et al., 2015) in order to identify putative core eukaryotic genes (CEGs) and universal single copy orthologs (USCOs) in the assembly. CEGMA has been replaced with BUSCO which is newly developed and more comprehensive than CEGMA (Simao et al., 2015) and was used in this study. BUSCO analysis against a plant-specific database of 956 genes identified 858 (89 %) complete BUSCOs (Table 5.6). The high percentage of BUSCOs mapped to the draft genome indicated the high completeness of the assembly. According to Simao et al. (2015), the high amount of duplicated complete BUSCOs indicated the erroneous assembly of haplotypes. However, Lee et al. (2016) proved that this links to genome duplication or recent hybridisation between seagrass species which is also the case of arabica (Lashermes et al., 1999; Tesfaye et al., 2007). Sayadi et al. (2016) also stated that this may represent allelic variation (heterozygosity) in the sample used to construct the assembly, gene duplication and/or mechanisms such as alternative splicing. Using different approaches of genome completion assessment indicted the completion of the genome assembly. The draft genome was then subjected to annotation to facilitate the downstream analysis.

5.3.4 Genome annotation

The gene prediction programs, SNAP and AUGUSTUS were used with tomato genome as reference. ESTs of *C. arabica* and CDSs of *C. canephora* and 96,521 in-house *C. arabica* (K7) PacBio transcriptome sequences were used as evidence to guide the annotation process. Altogether, 24,478 gene models were predicted consistently with different parameters when using MAKER. When performed with SNAP and AUGUSTUS, the number of gene models reached to 99,829 using tomato as reference. Mean length of gene was 2,612 bp with min length as low as 49 bp to max length of up to 51,554 bp (Table 5).

Table 5: Statistics of genome annotation using Blast2GO

Type	Total number
CDS (Median : Mean; Min-Max)	384,252 (138 : 228; 2 - 9,191)
Genes (Median : Mean; Min-Max)	99,829 (1,523 : 2,612; 49 - 51,554)
Intron (Median : Mean; Min-Max)	292,819 (258 : 590; 5 - 26,764)
Start codon	92,584
Stop codon	93,413
Transcript	99,829

The number of genes in the first annotation is lower than expected as it is close to the number of one of its ancestor *C. canephora* (25,574 protein-coding genes) (Denoeud et al., 2014). The low number of genes is probably because the genomic database used was limited. In the final annotation, when using other public databases as reference, especially with tomato – the closest plant species to coffee, the number of genes was almost four times higher than its double haploid progenitor. This number may be explained by the "true" tetraploid nature of K7 which is four times larger in genome size compared to the double haploid canephora. Compared to other close related plant species such as grape (30,425 genes), tomato (34,771 genes) and potato (35,004 genes) (Tomato-Genome-Consortium, 2012), K7 arabica is almost three times higher.

Functional annotation probably reveals more insight the K7 genome, especially when comparing with one of its progenitors – *C. canephora*.

5.4 Conclusion

The *C. arabica* whole genome has successfully been sequenced using short read (Illumina) and long read (PacBio) technology. Using the hybrid approach in assembly with several assemblers, gap fillers and scaffolders resulted in 76,409 scaffolds with scaffold N50 of 54,544 bp. The total scaffold length was 1,448 Mb which is 11% higher than the estimated arabica genome (1.3 Gb). This expansion could be attributable to the effect of ploidy and heterozygosity levels in arabica, which could not be resolved using the existing genome assemblers. Development and deployment of software that is suitable for highly heterozygous genomes or polyploidy combined with longer-read sequencing technology (e.g., Nanopore) may help to reduce the expansion and fragmentation of the sequenced genome. Validation of the genome assembly using different tools showed high completeness of the genome. BWA analysis demonstrated that > 98% of the short reads mapped to the genome and > 93% were marked as properly paired; GMAP analysis indicated that > 99% of the CDS and transcriptome sequences mapped to the *C. arabica* draft genome and 89% of BUSCOs were present. Altogether, 99,829 gene models have been annotated when using public database as reference which is four times higher than that of the double haploid canephora. Currently, there are several groups working on arabica genome assembly using different genotypes, sequencing platforms, assemblers and scaffolders (Deynze, 2017; Gaitan et al., 2015; Morgante et al., 2015; Strickler, 2015; Yepes et al., 2016). It seems that the progress was slow as it is a challenging genome to work with, and none of these genomes have been published in any peer-review journals and none are open source for public access. Future access to those pending resources would help refine the genome assembly reported in this study by a mapping approach. Those resources could also help to detect the variation among arabica germplasm. A reference genome sequence for *C. arabica* will have an important impact on coffee genetics and breeding.

CHAPTER 6: ASSOCIATION MAPPING AND SNP DISCOVERY IN EXTREME PHENOTYPES SELECTED FROM *C. ARABICA* POPULATION [5]

Abstract

Association analysis was performed at the whole genome level to identify loci affecting the caffeine and trigonelline content of *C. arabica* beans. DNA extracted from extreme phenotypes was bulked (high and low caffeine, and high and low trigonelline) based on biochemical analysis of the germplasm collection (Chapter 4). Sequencing and mapping using the combined reference genomes of *C. canephora* and *C. eugenioides* (CC and CE) identified 1,351 non-synonymous SNPs that distinguished the low- and high-caffeine bulks. Gene annotation analysis with Blast2GO revealed that these SNPs corresponding to 908 genes with 56 unique KEGG pathways and 49 unique enzymes. Based on KEGG pathway-based analysis, 40 caffeine-associated SNPs were discovered, among which nine SNPs were tightly associated with genes encoding enzymes involved in the conversion of substrates (i.e. SAM, xanthine and IMP) which participate in the caffeine biosynthesis pathways. Likewise, 1,060 non-synonymous SNPs were found to distinguish the low- and high-trigonelline bulks. They were associated with 719 genes involved in 61 unique KEGG pathways and 51 unique enzymes. The KEGG pathway-based analysis revealed 24 trigonelline-associated SNPs tightly linked to genes encoding enzymes involved in the conversion of substrates (i.e. SAM, L-tryptophan) which participate in the trigonelline biosynthesis pathways. Analysis of the K7 arabica reference genome (Chapter 5) identified several additional SNPs linked to genes encoding enzymes involved in caffeine and trigonelline synthesis pathways. These SNPs could be useful targets for further functional validation and subsequent application in arabica quality breeding.

[5] This chapter contains information that is in manuscript submitted to Tree Genetics and Genomes (accepted subject to minor revision): Tran, H. T. M., Furtado, A., Lee, L. S., Smyth, H., Vargas, C. A. C. and Henry, R. (2018). Association mapping and SNP discovery using extreme phenotypes for non-volatile compounds selected from a *C. arabica* diversity population

6.1 Introduction

Caffeine and trigonelline are two of the five most important non-volatiles in the coffee bean. Trigonelline is an important precursor of volatile compounds that link to coffee aroma and taste (Malta & Chagas 2009 in Barbosa et al., 2012) and strongly correlated with high quality (Farah et al., 2006b). Caffeine is also a compound of the consumer's concern as it contributes to the strength, body and bitterness of brewed coffee (Trugo, 1984). Decaf coffee can be obtained by the decaffeination process. However, the flavour tends to be influenced by this process. Thus, development of low-caffeine cultivars can help preserve other natural flavor compounds for various consumer preferences. The identification of single nucleotide polymorphism (SNPs) associated with bean caffeine content in diverse populations will support the manipulation of this compound in arabica coffee utilising molecular markers. The biosynthetic pathways of these two compunds have been well documented (reviewed in Chapter 2 sections 1.3.1.1) (Ashihara et al., 2011; Ashihara et al., 2011b; Ashihara et al., 2008). Genes involved in the metabolism of caffeine have been widely studied and described (Ashihara, 2006; Ogawa et al., 2001; Ogita et al., 2004; Ogita et al., 2003; Salmona et al., 2008; Uefuji et al., 2003), some of these have been mapped to subgenomes (e.g. *C. canephora* or *C. eugenioides*) (Perrois et al., 2015). In particular, a nucleotide mutation in the CADXMT1 gene (Maluf et al., 2009) coding for an N-methyltransferase (NMT) enzyme which caused a natural caffeine-free mutation (called "caffeine-free") in an *C. arabica* plant (Silvarolla et al., 2004) has been reported. For trigonelline, enzymes catalysed the conversion of nicotinate to trigonelline, coffee trigonelline synthases (termed CTgS1 and CTgS2), have been isolated and characterized (Mizuno et al., 2014). However, these caffeine/trigonelline biosynthesis genes were identified based on sequences derived from a limited number of specific arabica cultivars (e.g., Caturra and Laurina). Sequencing of a broader germplasm would help discover gene polymorphisms underlining quantitative variation that may be detected by association mapping.

Association mapping, or population mapping, involves searching for genotype-phenotype correlations in unrelated individuals. The main advantage of population mapping is that it exploits all of the recombination events that have occurred in the evolutionary history of a sample, which almost invariably results in a much higher mapping resolution compared to family mapping. The use of unrelated populations is particularly significant for research on organisms that are difficult to cross or clone, or have long generation intervals (Nordborg and Weigel, 2008) as is found in tree crops like coffee. Association mapping has advantages of increased mapping resolution, reduced research time, and greater allele number (Yu and Buckler, 2006). One of the limitations of association mapping approaches is that they require genotyping of large numbers of individuals, which may be expensive for large populations, even with the cost reduction in sequencing technology. Recently, a method called "extreme-phenotype genome-wide association study" (XP-GWAS) has been developed (Yang et al., 2015) that allows pooling or bulking the individuals from a diverse population that exhibit extreme phenotypes to reduce the cost of genotyping every member of the population while ensuring the high resolution of mapping of trait-associated variants (TAVs) (Yang et al., 2015). The method was built on previous suggestion by Sun et al. (2010) that pooled DNA analysis could be used for two contrasting groups of individuals from any population, not just for those from bi-parental segregating populations used in bulked segregant analysis (BSA) (Michelmore et al., 1991). Individuals with extreme phenotypes from natural populations have been bulked for sequencing and genome wide association study (GWAS) (Bastide et al., 2013; Turner et al., 2010; Yang et al., 2015). Schlötterer et al. (2014) also used the terms "Pool-seq" or "Pool-GWAS" for this approach and promoted its best practice in terms of the number of individuals included in a pool, depth of coverage, sequencing technology and downstream analysis. With XP-GWAS, allele frequencies in the extreme pools are measured, thus enabling discovery of associations between genetic variants and traits of interest. Empirical study showed that XP-GWAS outweighed conventional QTL mapping by reducing the number of samples while ensuring detection of small-effect loci if a sufficient number of samples per pool are used, enriching for rare alleles and increasing allele effects (Yang et al., 2015).

This chapter aims to apply XP-GWAS to identify SNPs associated with bean caffeine and trigonelline content. The wide phenotypic variation observed for these traits in the *C. arabica* germplasm collection (Chapter 4) enables selection of extreme phenotypic groups for use in this method. The reference genomes of *C. canephora* (published) (Denoeud et al., 2014) and *C. eugenioides* (not yet published and provided by Coffee Consortium) combined with the *C. arabica* draft genome developed as part of this thesis (Chapter 5) will provide a framework for sequence comparison between DNA bulks for subsequent identification of TAVs and genes that may be involved in the biosynthetic pathways of caffeine and trigonelline.

6.2 Materials and Methods

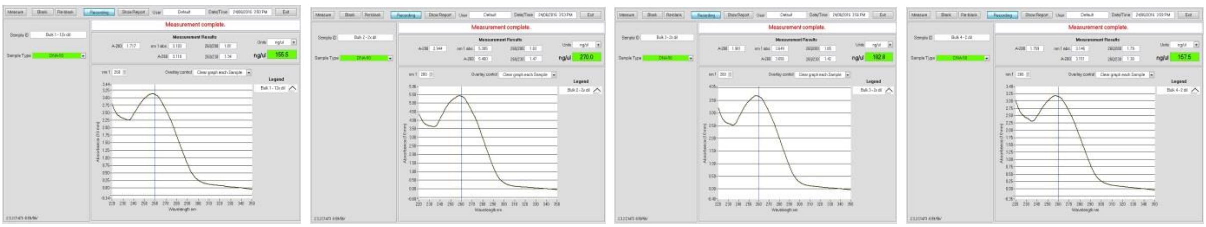
6.2.1 Biological materials

A total of 72 individuals, each representing one accession, were selected for bulk sequencing from a diverse population of 232 arabica accessions as described in Chapter 4,. They were divided into four bulks containing 18 individuals from each extreme phenotypic group (lowest/highest caffeine, and lowest/highest trigonelline) (Appendix table S6.1 and S6.2)

6.2.2 DNA extraction, pooling and quality

Leaf samples were collected from the germplasm plantation at CATIE. They were stored in two forms: preserved and fresh leaves. For the preserved form, 1.5 mg of coffee leaf from each tree was finely ground with liquid nitrogen in a pre-chilled mortar and pestle. The ground tissue was transferred into a 15 mL falcon tube, then 13 mL of extraction buffer was added to the ground tissue and mixed by inversion. The extraction buffer contained 100 mM Tris-HCl pH 8.0 ($C_4H_{11}NO_3$*HCl), 25 mM EDTA pH 8.0 ($C_{10}H_{16}N_2O_8$) and 1.5 M NaCl and 2% CTAB ($C_{19}H_{42}BrN$). Each tube was sealed with tape around the cap to avoid leakage of the extraction buffer. For the fresh-leaf form, coffee leaves of each genotype were wrapped in moistened tissue and placed in plastic bags. Each preserved/fresh sample was labelled with unique sample names, botanical names, collection date, and contact details. They were wrapped in bubble plastic sheets, placed in a foam box and shipped to the University of Queensland subject to international quarantine processes.

Before performing the next steps as described by Healey et al. (2014), 3% mercaptoethanol and 2% PVP (polyvinylpyrrolidone) ((C_6H_9NO)n) were added to the preserved samples. DNA quality and quantity were measured on Nanodrop, agarose gel and Qbit meter for every sample. The concentration was then standardized and diluted to include an equal amount of DNA for all individual DNA samples. The 18 samples in each extreme-phenotypic group were finally mixed to form a DNA bulk, resulting in four DNA bulks in total. The DNA quality and quantity of each bulk was then checked using Nanodrop, agarose gel and Qbit meter (Figs. 6.1 and 6.2, Table 6.1) before sequencing.



Measurement complete.
ng/ul 155.5

Measurement complete.
ng/ul 270.0

Measurement complete.
ng/ul 182.8

Measurement complete.
ng/ul 157.5

Figure 6.1 DNA check on Nanodrop

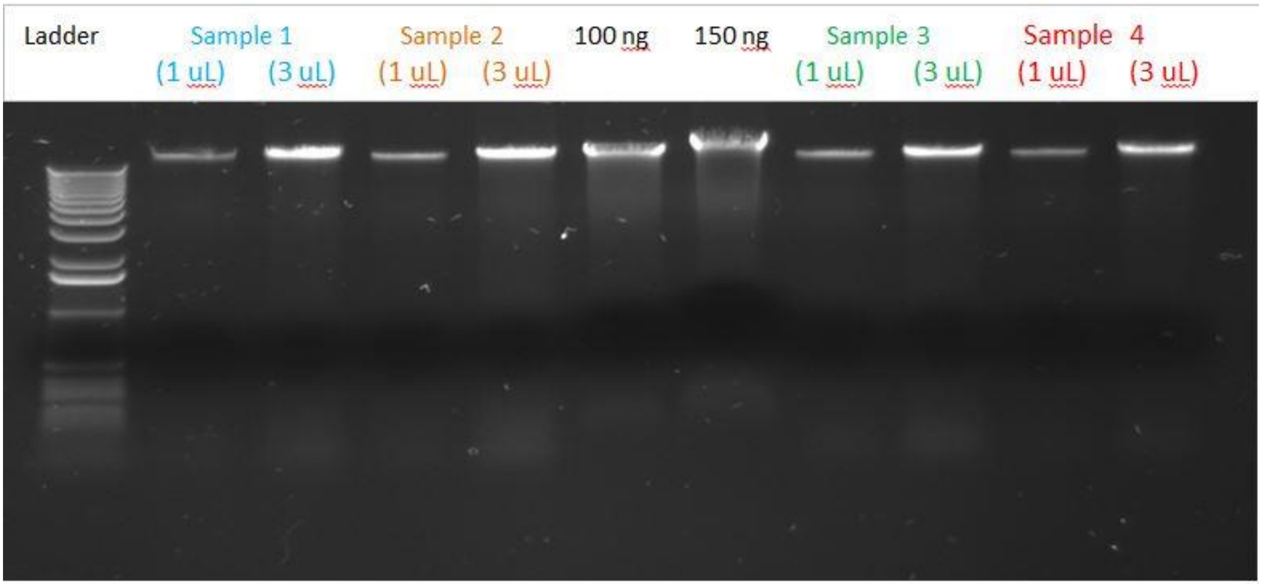


Figure 6.2 DNA check on agarose gel

Table 6.1 DNA quality and quantity

Samples	260/280	260/230	Conc. on Qubit (ng/ul)	Final volume in TE (ul)	Total amount (ng)
Low caffeine	1.81	1.34	46.9	95	4712
High caffeine	1.83	1.47	53.0	95	5035
Low trigonelline	1.85	1.42	55.0	95	5225
High trigonelline	1.79	1.30	57.0	95	5415

6.2.3 Library preparation and sequencing

DNA samples were sequenced as 4 indexed PCR-free libraries, using a HiSeq 2000 (v4) flowcell of an Illumina platform (Queensland Brain Institute (QBI), University of Queensland).

6.2.4. Trait-associated SNPs discovery

Paired-end reads with insert sizes of 150 bp from four DNA bulks were imported to CLC Genomics Workbench Version 10.0 (CLC Bio, www.clcbio.com). The whole genome sequences of Double Haploid (DH) *C. canephora* (published), *C. eugenioides* (provided by Coffee Consortium) and *C. arabica* (assembled as part of my research and explained in Chapter 5) were also imported to CLC as a standard import and used as a reference for mapping. Raw reads were subjected to Quality Control (QC) analysis which was used as a guide for trimming the reads. Low-quality paired-end sequence reads were trimmed using default parameters. The quality score limit was set to 0.05 (corresponding to Phred quality value >15) and minimum number of nucleotides in reads of 15 bp. These trimmed reads were mapped independently to the two references (1) Combined *C. canephora* (CC) and *C. eugenioides* (CE) genome; (2) *C. arabica* draft genome. The mapping was performed using default settings (match score: 1; mismatch cost: 2; insertion cost: 3; deletion cost: 3) except the length fraction (LF) and similarity fraction (SF) being set at 1.0 and 0.8, respectively. Indel structural variants analysis was performed based on the mapping files with P-Value threshold of 0.0001. The output of indel structural variants was used as guidance-variant track for local re-alignment to improve on the alignments of the reads in an existing read mapping with setting of the Multi-pass local realignment of 2. The output of the local re-alignment was the stand-alone mapping which was used to call variant, and also to create track to examine the variant calling.

6.2.5. SNPs identification and filtering

The stand-alone mapping file was used for variant calling. Variants were called using the "Basic variant detection" tool with ploidy level of 4. Various settings for SNPs call were implemented in order to determine the optimal settings. To ensure that the SNPs identified were of high quality, two types of filters were applied including general filters and noise filters. There are two types of general filters: reads filters (ignore broken pairs and ignore non-specific matches) and coverage and count filters which were set at different settings (Table 6.2). There are two types of noise filters: base quality filters and read direction and position filters which were applied as default settings except for read direction frequency (Table 6.2). In summary, the SNPs were selected as follows: (i) minimum coverage of at least 20 and maximum coverage of 1,000, (ii) broken pairs and non-specific matches were removed, (iii) minimum of reads to be called as a variant of 4 to 6 or 15-30% (Table 6.2), (iv) base quality filter was applied with minimum central quality of 20, neighbourhood radius of 5, minimum neighbourhood quality of 15, and (v) read direction and read position filters were applied with read direction frequency of 0.0%, 20% and 35%.

Table 6.2 Different settings of coverage and count filters

Parameters	Minimum coverage	Minimum count	Minimum frequency (%)	Read direction frequency (%)
Setting 1	20	6	30	35
Setting 2	20	6	30	0
Setting 3	20	4	20	20
Setting 4	20	4	15	35
Setting 5	20	4	15	0

6.2.6. Identification of trait-associated SNPs

The tool "Identify known mutations from sample mappings" was used to identify variants between two bulks of the same trait (i.e. caffeine or trigonelline) in which variant files from bulks of targeted trait (i.e. low caffeine and high trigonelline) were used as reference variant tracks. The minimum coverage and detection frequency were set consistently with those in variant calling step. Broken pairs and non-specific matches were removed. The outputs of this mutation test were filtered with the zygosity (i.e. homozygous or heterozygous) to create a list of SNPs under four categories: 1) hom B1- hom B2; 2) hom B1- het B2; 3) het B1- hom B2; 4) het B1 – het B2 (where hom:

homozygous; het: heterozygous; B1: Bulk 1; B2: Bulk 2). Similar approaches were applied for B3 and B4. Only alleles that had frequency of ≥ 97% were considered as "hom". All categories were analysed with the tool "Amino acid changes" which required a number of input annotation tracks (i.e. CDS track, genome sequence track, and mRNA track of corresponding genome) (Figure 6.3a) to detect the coding region change, amino acid change and non-synonymous substitution of the SNPs. In the case of the SNPs identified in the category "het B1 – het B2", SNP identified and their frequency details were exported to an Excel spread sheet. The chi-square test was applied to the "het-het" allele frequencies from two bulks. The "If command" tool in Excel was used to select the non-synonymous SNPs that distinguish two bulks. Each SNP was manually checked using "track tools" which integrated the variant call track (for each bulk) and the mapping track to examine the accuracy of the SNPs called (Figure 6.3a). The entire process for SNP identification and filtering is described in Figure 6.3b

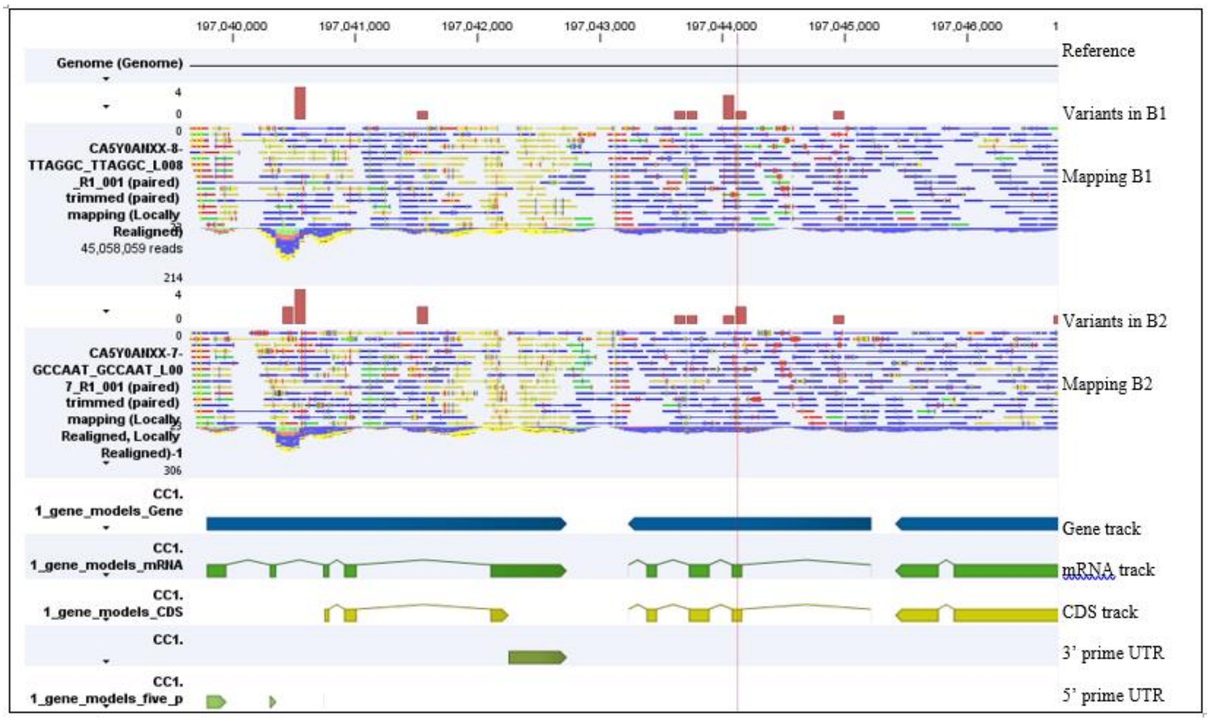


Figure 6.3a Illustration of how a non-synonymous SNP was manually checked to ensure the accuracy of the SNPs called by using "track tools"

(Red line indicates the location of SNP; pair reads were in blue in mapping; forward reads were in green in mapping; reverse reads were in red in mapping; non-specific reads were in yellow in mapping).

6.2.7. Functional annotation of SNPs.

The final set of non-synonymous SNP were extracted from the annotation track to obtain CDS sequences and imported into Blast2GO (version 4.0.7) (Conesa et al., 2005) so as to obtain more information on the functional annotation and biological role of these SNPs. Steps involved in this analysis (including blast, interproscan, mapping, and annotation) were run using default settings.

The Blast2GO output included Gene Ontology names, Enzyme Codes, Kyoto Encyclopedia of Genes and Genomes (KEGG) maps and statistics of biological process, cellular component and molecular function of SNPs. Output data of all KEGG pathways were examined thoroughly in order to identify those involved in caffeine and trigonelline biosynthesis, as well as those involved in the metabolism of substrates that eventually enter into the caffeine and trigonelline biosynthetic pathways (Figure 6.3b).

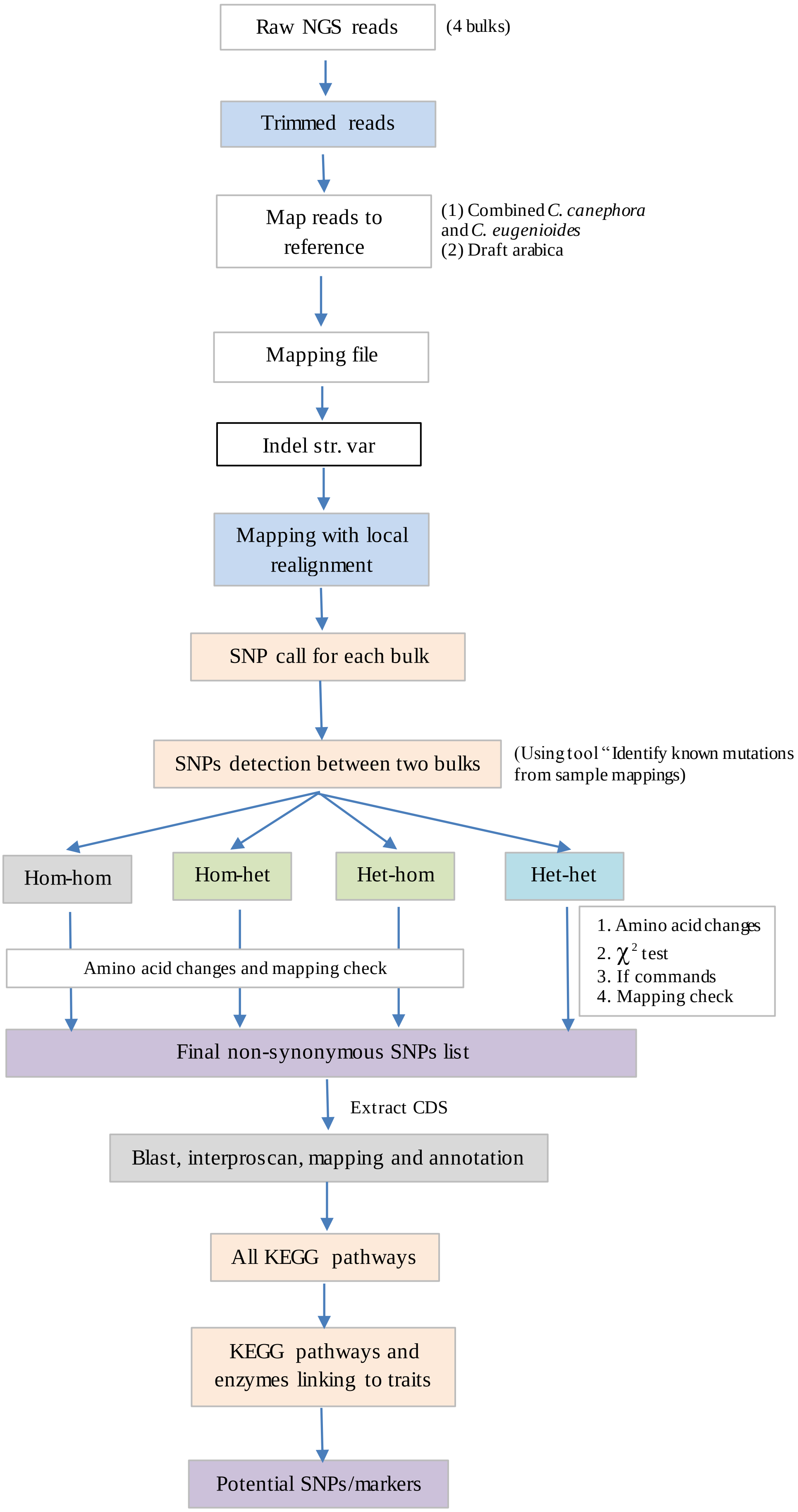

Raw NGS reads
(4 bulks)
Trimmed reads
Map reads to reference
(1) Combined *C. canephora* and *C. eugenioides*
(2) Draft arabica
Mapping file
Indel str. var
Mapping with local realignment
SNP call for each bulk
SNPs detection between two bulks
(Using tool "Identify known mutations from sample mappings)
Hom-hom
Hom-het
Het-hom
Het-het
Amino acid changes and mapping check
1. Amino acid changes
2. χ^2 test
3. If commands
4. Mapping check
Final non-synonymous SNPs list
Extract CDS
Blast, interproscan, mapping and annotation
All KEGG pathways
KEGG pathways and enzymes linking to traits
Potential SNPs/markers

Figure 6.3b The workflow for SNP detection and filtering using CLC Work Bench and functional annotation of SNPs using Blast2GO.

6.3 Results and Discussions

6.3.1 Genotype selection, bulking and sequencing

For each trait (caffeine and trigonelline), the lowest and highest groups selected for XP-GWAS were about 34-35% different in the average content (Table 6.3), which is similar to that reported for the kernel row number between bulks used in maize XP-GWAS (Yang et al., 2015). Details of selected accessions per group were provided in Appendix (Tables S6.1 and S6.2). The low caffeine group comprised mainly individuals from the cultivar/selection/natural mutant collection, while the high caffeine group included mainly wild accessions (Table 6.3), indicating that high caffeine was more common in the wild. In contrast, high trigonelline was more common in the cultivar/selection/natural mutant accessions (Table 6.3). Anyhow, the presence of all types of arabica in each bulk reduced the risks of population stratification causing spurious allelic association (Cardon and Palmer, 2003; Price et al., 2006).

Table 6.3 The population constituent and sequencing statistics of four DNA bulks for low/high caffeine and trigonelline

Caffeine	**Low**	**High**
Total individuals (Groups 1, 2, 3)*	18 (11; 2; 5)	18 (2; 4; 12)
Content (% dmb)	1.03	1.48
Total of reads after trimmed (#)	230,140,744	324,144,616
Average coverage (x)	28	40
GC content (%)	36	36
Average Phred score	37	37
Average length after trim (bp)	147	147
Trigonelline	**Low**	**High**
Total individuals (Groups 1, 2, 3)	18 (2; 8; 8)	18 (10; 4; 4)
Content (% dmb)	0.93	1.31
Total of reads after trimmed (#)	306,657,574	193,093,468
Average coverage (x)	38	24
GC content (%)	36	36
Average Phred score	37	37
Average length after trim (bp)	148	147

*Group 1: Cultivar/selection/natural mutant; Group 2: Introgression/Hybrids; Group 3: Wild

Sequencing of each DNA bulk resulted in more than 190 million high-quality reads, with the coverage (depth) ranging from 24 to 40x (Table 6.3). This met the recommended depth for pooled sequencing, which should be at least equal to or higher than the number of individuals in a pool (Magwene et al., 2011). According to Ries et al. (2016), sequencing coverage of 30x for each pool would allow the identification of causative SNPs without requiring prior knowledge or additional sequencing of single offspring genotypes or parental lines. High sequencing depth is a key consideration in genomic analysis for polyploidy like arabica coffee. Generally, for tetraploids, a sequencing depth greater than 20x, as obtained from bulk sequencing in this study, would have minimal impact of random polymerase errors on variant calls (sumarised by Olson et al., 2015).

According to Clevenger et al. (2015), to accurately identify SNPs in a polyploid, one should also take into account the sequencing technology, read length, library preparation. Among several sequencing technologies, Illumina paired-end sequencing was used in this study due to its affordability and ability to produce the high read depth required for unambiguously detecting alleles in a polyploidy, and for the improvement of read mapping and assembly accuracy (Clevenger et al., 2015). The use of paired-end library preparation methodology can also reduce variant calling error as it reduces duplicate mapping errors (Olson et al., 2015; Schlötterer et al., 2014). These authors suggested longer reads (paired-end 150 or greater) were suitable for species with highly similar subgenomes (e.g. allopolyploids with lowly diverged progenitors like *C. arabica*) (Lashermes et al., 2014; Pearl et al., 2004) in that their homeologs can be distinguished. Sequencing reads used in the current study were 151 bp in length before trimming and 4 bp shorter after trimming, which is very close to the above recommended read length.

The use of PCR-free protocols to generate libraries for sequencing in this study reduced the erroneous polymorphisms due to PCR error, which can be easily mistaken as true allelic or homeologous SNPs within a polyploid species that possesses little genetic variation within and among subgenomes (Clevenger et al., 2015). The pre-processing of reads by trimming was also performed so as to reduce the error rate towards the 3´ end of Illumina reads that could impair downstream analyses such as variant calling (Schlötterer et al., 2014). As a result, Phred score (an indicator of the read quality) was 37, which is higher than the recommended Phred score of 30 giving a 0.1% probability of a wrong base call (Dohm et al., 2008). The GC bias (GC-poor or GC-rich), which is related to uneven read coverage across genome (Chen et al., 2013), should be considered. The GC content of the current study (37%) was not in the range of GC-poor (< 20%) or GC-rich (> 80%) (Li et al., 2010a).

6.3.2. Mapping of reads to reference genomes

The correct alignment of reads to the genome reference provides a framework for SNP detection. Since the reference genome of *C. arabica* has not been completed (currently in the form of contigs), its ancestral species (*C. canephora* and *C. eugenioides*) were first used as references for read alignment. In fact, for the inbreeding allopolyploids with known ancestral relationships like coffee, derivation of reference genomic sequences for each subgenome and their diploid counterparts is both desirable and feasible (Kaur and Francki, 2012).

Read alignment with the *C. canephora* reference using different length fraction and similarity fraction settings showed the length fraction (LF) of 1.0 and similarity fraction (SF) of 0.8 were the best settings which gave rise to more than 80% mapped reads and 57x average coverage (Appendix Table S6.3). The default setting (LF 0.5 and SF 0.8) gave the highest percentage of mapped reads (> 95%). However, it seems that this setting was not stringent enough and might result in low quality mapping, and consequently result in the large number and possible false SNPs identified. More stringent settings in mapping (LF 1.0 and SF 0.9 or 0.95) would help eliminate false SNPs. However, because the genome reference used in the current study are partly related to the arabica samples, applying these stringent settings resulted in the reduced number of mapped reads and SNPs identified arising from the other subgenome. So, the less stringent setting of LF 1.0 and SF 0.8 was then used for mapping.

According to Olson et al. (2015), the two most common sources of true read mapping errors are genomic duplication and structural variation. A careful evaluation of mapping parameters therefore has resulted in guidelines that will considerably reduce errors due to alignment problems (Schlötterer et al., 2014). Therefore, in this study the mapping file was subjected to Indels and structural variants analysis which was used as guidance for local realignment to improve mapping results. This advantage was confirmed by Liu et al. (2012) in which efficiently reduce the false-positive rate was efficiently reduced for high coverage regions.

When *C. canephora* and *C. eugenioides* were used separately as references in the mapping, the percentage of mapped reads were comparable between two genomes (88.34% and 86.85% respectively) (Appendix table S6.4), indicating that they and the arabica genome share a considerable syntenic regions. Indeed, Cenci et al. (2012) analysed nucleotide substitutions among the three orthologous regions and found the similarity between sequences resulting from the two *C. arabica* subgenomes was 95% (based on 780 kb). Yu et al. (2011) also found a high level of sequence similarity between BACs from *C. arabica* and *C. canephora*.

When the two genomes *C. canephora* and *C. eugenioides* were combined, the percentage of mapped reads (93.75%) was higher than those when analysed separately) (Appendix table S6.4). This percentage of mapped reads was also equivalent to that derived from the alignment using *C. arabica* as reference (93.06%, Appendix table S6.4). Therefore, the mapping files derived from alignment to *C. arabica* and the combined *C. canephora* and *C. eugenioides* genomes were used for variant calling.

6.3.3. Identification of caffeine-associated SNPs

6.3.3.1. SNP detection using CC and CE genomes as reference

SNPs that distinguish the low- and high-caffeine bulks (TAVs) were identified using different variant-calling settings (Table 6.4). A combination of different settings was assessed to maximise the accurate detection of SNPs which may otherwise be masked due to different subgenome homologs. Since the average coverage of two bulks of low/high caffeine is 28 and 40 (Table 6.3), setting for minimum coverage therefore was at 20 which is higher than in several studies of allotetraploids (only 4 to 8) (Byers et al., 2012; Nagy et al., 2013; Peace et al., 2012; Schmutzer et al., 2015). For tetraploid species like *C. arabica*, identification of high confidence SNPs is challenging as one locus could potentially have up to 4 alleles (i.e., allele frequencies could be 25, 50, 75 or 100%) (Castle et al., 2014). However, in pool-sequencing, the allele frequency will not follow the theoretical scenarios due to possible experimental noise during sampling, chemical analysis and DNA mixture which might result in unbalance representation of individuals in the pool. Schlötterer et al. (2014) found that the impact of differential representation of individuals on the accuracy of allele frequency estimates is not large if appropriate sample sizes are used. The identification of the threshold of SNP frequency that is significantly different between bulks therefore becomes challenging and needs standardisation involving assessing the use of various combinations of settings for variant calling. In the present study, the minimum count or minimum of reads to be called as a variant was set at more stringent setting of 30% (6 per 20 reads) to less stringent of 20% or 15% (or 4 per 20 reads) to examine for the presence of tri-alleles or tetra-alleles in a locus. The balance between forward and reverse reads (F/R balance) was set from 0.0%, 20% to 35% to allow high quality reads (paired reads) in the mapping (Table 6.4).

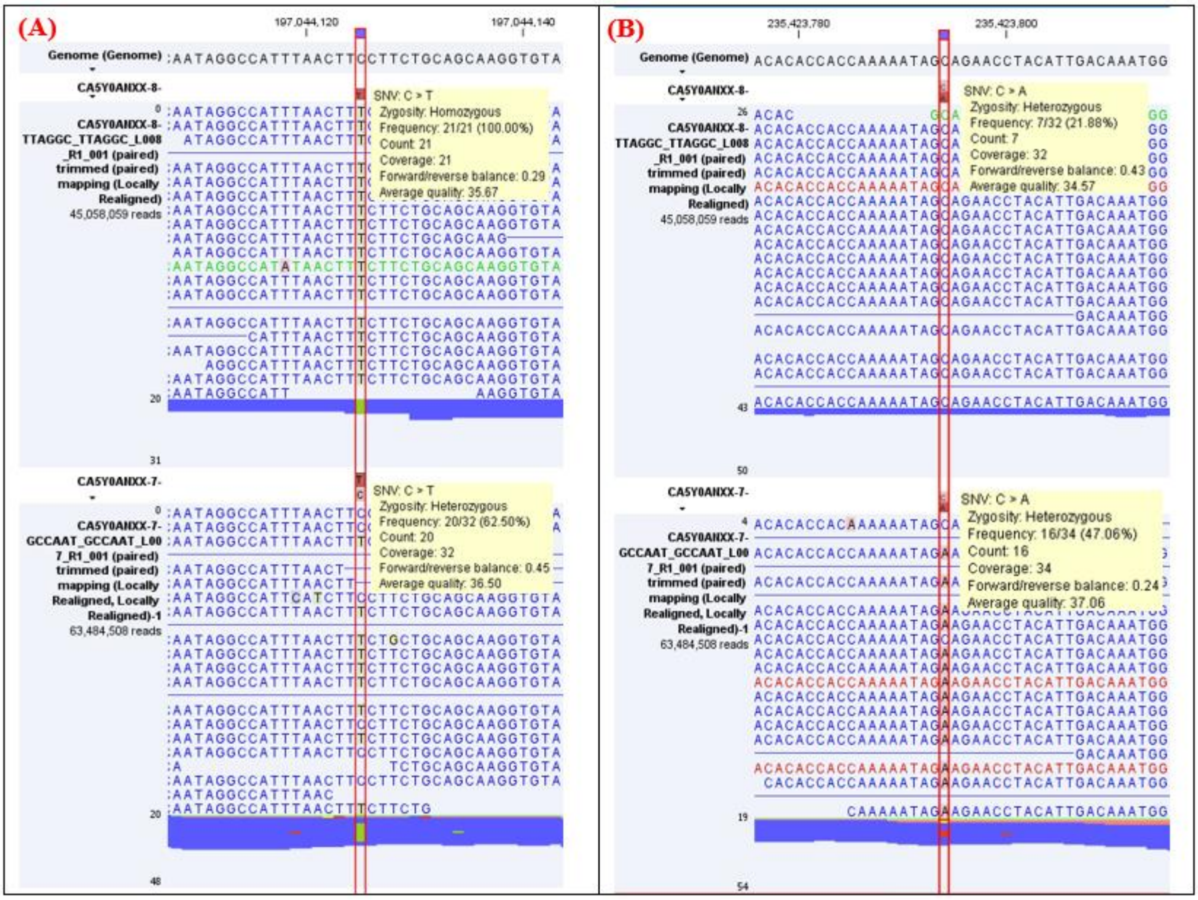
Table 6.4 Number of SNPs distinguishing the low- and high-caffeine bulks based on different settings of variant calling.

Settings	Type (*)	No of variant	Non-synonymous	Di-allele	Tri-allele	% tri-allele	Filter (**)	Check mapping	Correct call (%)
Setting 1 20-6-30-0.35	HomB1-HetB2	564	8					8	
	HetB1-HomB2	3,711	57					4	
	HetB1-HetB2	393,056	6,036	6,036	0	0	339	339	
	Total	**397,331**	**6,101**					**351**	**87**
Setting 2 20-6-30-0.0	HomB1-HetB2	765	11					11	
	HetB1-HomB2	9,443	149					17	
	HetB1-HetB2	921,335	13,954	13,954	0	0	778	778	
	Total	**931,543**	**14,114**					**806**	**86**
Setting 3	HomB1-HetB2	1,909	40					35	

20-4-20-0.20	HetB1-HomB2	15,891	182					34	
	HetB1-HetB2	1,201,432	18,408	18,380	28	0.15	1,282	1,282	
	Total	**1,219,232**	**18,630**					**1,351**	**90**
Setting 4 20-4-15-0.35	HomB1-HetB2	2,603	68					32	
	HetB1-HomB2	11,289	91					8	
	HetB1-HetB2	710,321	10,577	10,567	10	0.09	624	624	
	Total	**724,213**	**10,736**					**664**	**85**
Setting 5 20-4-15-0.0	HomB1-HetB2	3,458	81					42	
	HetB1-HomB2	39,496	286					38	
	HetB1-HetB2	2,037,081	29,523	29,451	72	0.24	1,708	1,708	
	Total	**2,080,035**	**29,890**					**1,788**	**86**

(*) Hom: Homozygous; Het: Heterozygous; B1: Bulk 1; B2: Bulk 2; (**) Filtered by using Chi-square test and If command to keep the SNPs that are significant difference in frequency between two bulks and being non-synonymous between two bulks.

Setting 1 was the most stringent and thus resulted in the least number of TAVs (6,101 non-synonymous SNPs), whereas setting 5 was the least stringent resulting in more TAVs (29,890 non-synonymous SNPs) (Table 6.4). The setting 1 required at least 20 reads at the position of SNPs was called with minimum count of 6 (30%) and it also required F/R balance of 35%. When checking with mapping, this setting showed reliable variant call as it only called the main alleles of the locus which would be kept if the two bulks were actually different. However, because the minimum was set at high percentage (30%), alleles with low frequency were not called thus unable to compare the frequency of the alleles between two bulks. In addition, SNPs were not called at the loci that the F/R balance was less than 35%. This led to the risk of missing out the important SNPs that presented in low frequency or did not pass the F/R balance. In theory, the marker alleles can be present in different dosages, ranging from 0 (nulliplex) to 4 (quadruplex) in tetraploid species (Voorrips et al., 2014) and each allele will account for 25% on average. However, no tetra-alleles were observed in any setting even when the frequency was set as low as 15% to be called a variant. No tri-alleles were observed in setting 1 and 2 when the frequency was set at 30%, and even with the least stringent setting (setting5), the percentage of tri-alleles was very low (0.24%). Since *C.* arabica was constituted by two subgenomes *C. canephora* and *C. eugenioides* and *C. canephora* was highly heterozygous (Denoeud et al., 2014), it is expected that *C. arabica* gets heterozygosity from *C. canephora* leading to the common status of tri-alleles. However, the low percentage of tri-allelic loci observed in this study indicates that the two subgenomes of *C. arabica* might be both homozygous. This result is consistent with previous findings using cytological or molecular markers (Cenci et al., 2012; Lashermes et al., 2014; Pearl et al., 2004) and that *C.* arabica has a diploid-like meiotic behaviour (Krug and Mendes, 1940; Lashermes et al., 2000b; Teixeira-Cabral et al., 2004) would facilitate the identification and interpretation of the SNPs. This also suggests that it is sufficient to set the minimum count at 20 or 30%. There was no hom-hom between two bulks (e.g. A in Bulk1 and G in Bulk 2) observed. The most common type was het-het which accounted for 95% in most settings while the hom-het or het-hom is more meaningful. For the category of hom-het or het-hom, the variant file was subjected to amino acid change analysis to identify non-synonymous SNPs since the difference in allele types and frequency between two bulks was significant. For the het-het category, applying the chi-square test and "If" command resulted in a considerable reduction in the number of significant TAVs (1,282 from the 18,408 in setting 3) (Table 6.4). Chi-square test is applied to ensure the difference between two bulks are significant while the "If command" tool in Excel (allele one or two in bulk 1 must be ≥ 50% while the corresponding allele in the other bulk must be ≤ 50%) was applied to select the non-synonymous SNPs (caused by changing amino acid) between the two bulks and not between each bulk and the reference. The final SNPs were checked manually using the mapping and variant calling files in the form of tracks to gain confidence (Figure 6.4). Among five settings, setting 3 gave highest score for correct call (90%) with 1,351 TAVs (Table 6.4) and was subjected to Blast2GO for functional analysis.


(A)
197,044,120
197,044,140
Genome (Genome)
AATAGGCCATTTAACTTCCTTCTGCAGCAAGGTGTA
CA5Y0ANXX-8-
CA5Y0ANXX-8-TTAGGC_TTAGGC_L008_R1_001 (paired) trimmed (paired) mapping (Locally Realigned)
45,058,059 reads
0
20
31
CA5Y0ANXX-7-
CA5Y0ANXX-7-GCCAAT_GCCAAT_L007_R1_001 (paired) trimmed (paired) mapping (Locally Realigned, Locally Realigned)-1
63,484,508 reads
0
20
48
(B)
235,423,780
235,423,800
Genome (Genome)
ACACACCACCAAAAAATAGCAGAACCTACATTGACAAATGG
CA5Y0ANXX-8-
CA5Y0ANXX-8-TTAGGC_TTAGGC_L008_R1_001 (paired) trimmed (paired) mapping (Locally Realigned)
45,058,059 reads
26
43
50
CA5Y0ANXX-7-
CA5Y0ANXX-7-GCCAAT_GCCAAT_L007_R1_001 (paired) trimmed (paired) mapping (Locally Realigned, Locally Realigned)-1
63,484,508 reads
4
19
54

(A)

SNV: C > T
Zygosity: Homozygous
Frequency: 21/21 (100.00%)
Count: 21
Coverage: 21
Forward/reverse balance: 0.29
Average quality: 35.67

SNV: C > T
Zygosity: Heterozygous
Frequency: 20/32 (62.50%)
Count: 20
Coverage: 32
Forward/reverse balance: 0.45
Average quality: 36.50

(B)

SNV: C > A
Zygosity: Heterozygous
Frequency: 7/32 (21.88%)
Count: 7
Coverage: 32
Forward/reverse balance: 0.43
Average quality: 34.57

SNV: C > A
Zygosity: Heterozygous
Frequency: 16/34 (47.06%)
Count: 16
Coverage: 34
Forward/reverse balance: 0.24
Average quality: 37.06

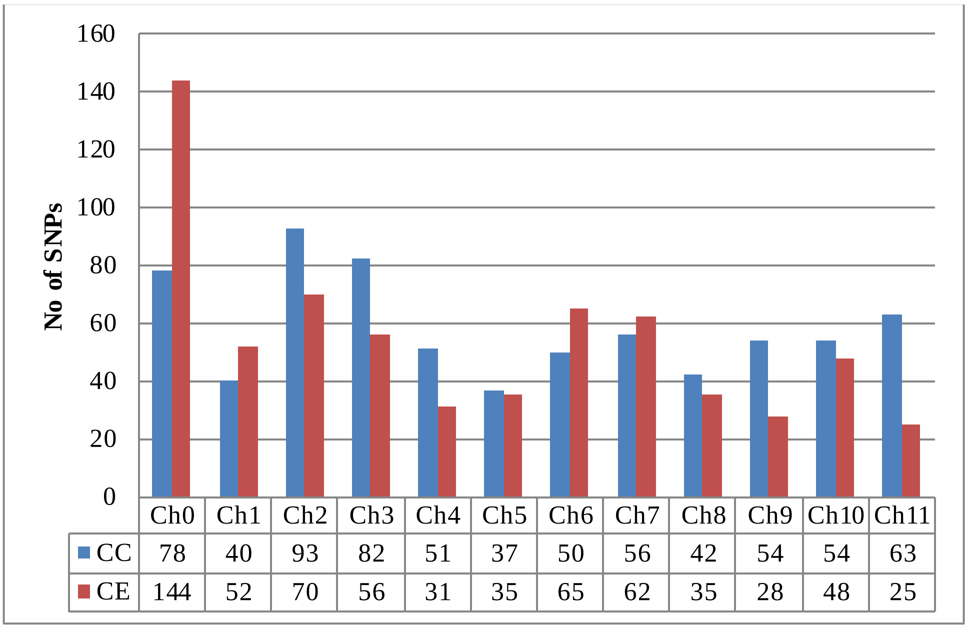
Figure 6.4 Examples of highly confident TAVs for caffeine that distinguish between two bulks (B1 and B2) in the types of hom-het (A) and het-het (B)

(A) hom-het type: B1 was homozygous with 100% of T with coverage of 21, F/R balance of 0.29 and average quality of 35.67; B2 was heterozygous with 63% of T and 37% of C with coverage of 32, F/R balance of 0.45 and average quality of 36.50 (B) het-het type (both bulks are heterozygous) B1 with 22% of A and 78% of C with coverage of 32, F/R balance of 0.43 and average quality of 34.57 while B2 with 47% of A and 53% of C with coverage of 34, F/R balance of 0.24 and average quality of 37.06. Red lines are the location of SNPs

6.3.3.2. Distribution of the caffeine-associated SNPs across subgenomes of arabica.

The analysis detected 1,351 TAVs with high average sequencing depth of 31x for Bulk 1 and 46x for Bulk 2. This indicates the confidence of variant call as the minor alleles (at the setting of 20%) would have 6 to 9 reads. The high value of F/R balance (0.38) showed the high read quality at the SNP location. The base quality assigned to each base is the probability that the base in question is not an over-call and no SNP call should be made from a single allele with a base quality lower than 30 (1 in 1,000 bp error rate) (Quinlan et al., 2007). However, it is very common that this parameter was set at threshold of ≥ 20 in several studies for tetraploids (Nagy et al., 2013; Peace et al., 2012; Schmutzer et al., 2015). In the present study, the average base quality score (36 and 34 for Bulk 1 and 2, respectively) was higher than the recommended threshold.

The detected TAVs were plotted to reference genomes of *C. canephora* and *C. eugenioides* to examine their distribution. There were slightly more SNPs on the subgenome of *C. canephora* (700 SNPs) than *C. eugenioides* (651 SNPs) (Fig. 6.5), probably due to the closeness of *C. arabica* to *C. eugenioides*. There were more SNPs on chromosome 0, 2, 3 and 11 for CC, and chromosome 0, 2, 6 and 7 for CE. Chromosome 0 and 2 had highest number of SNPs for CE and CC, respectively (144 and 93) (Figure 6.5). The unequal distribution of TAVs between two subgenomes of arabica provides additional evidence for the phylogenetic contribution of each sub-genome revealed by other genomic research. Arabica coffee is a recent allotetraploid (C^aE^a genome). It is estimated that the two genomes diverged approximately < 50,000 years ago (Cenci et al., 2012; Lashermes et al., 1999). Examination of C^a - E^a homeologous genome regions at the DNA sequence level reveals the two parental species are closely related, and the two subgenomes have low sequence divergence (Cenci et al., 2012; Yu et al., 2011). Lashermes et al. (2014) using *C. arabica* mRNA-seq in comparative analysis of the number of SNPs per unigenes and found the number of SNPs between the two diploid progenitors was rather similar. However, other studies (Cotta et al., 2014; Vidal et al., 2010) showed the two subgenomes contained in the *C. arabica* allotetraploid genome (subgenome *C. canephora* - CaCc and subgenome *C. eugenioides* - CaCe) do not contribute equally to the transcriptome.



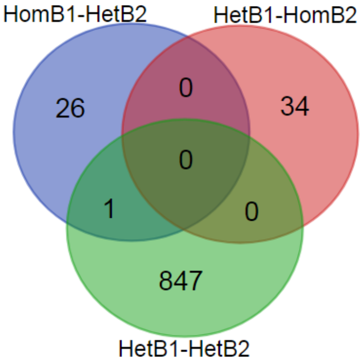
	Ch0	Ch1	Ch2	Ch3	Ch4	Ch5	Ch6	Ch7	Ch8	Ch9	Ch10	Ch11
🟦 CC	78	40	93	82	51	37	50	56	42	54	54	63
🟥 CE	144	52	70	56	31	35	65	62	35	28	48	25

Figure 6.5 Genome-wide distribution of TAVs for caffeine in the two subgenomes *C. canephora* (CC) and *C. eugenioides* (CE) of *C. arabica*

The distribution of TAVs across the genome indicates the presence of multi-genes controlling the caffeine trait. Het-het in the form of bi-allele accounted for 95% of the total SNPs indicated the high heterozygous rate in arabica. As *C. arabica* is well known for having a narrow genetic base and being a self-pollinating species, it is expected to be highly homozygous. However, due to its allotetraploid origin, *C. arabica* contains a considerable amount of fixed heterozygosity and shows high levels of heterozygosity at individual loci (Lashermes et al., 1999). Similarly, Mishra et al. (2011) suggested that since arabica is an allotetraploid crop, it is expected to show heterozygosity conditioned by two different alleles derived from two different progenitors. Heterozygosity within the two ancestral genomes appears to have been lost, since only one allele from each genome remains in arabica, indicating a possible lack of recombination between the ancestral genomes, while recombination within each genome occurs normally (Cubry et al., 2008). More specifically, Perrois et al. (2015) found that the three genes involved in caffeine biosynthesis (XMT, MXMT, DXMT) are clearly encoded and assigned to each subgenome in arabica. This helps determine their respective importance for caffeine accumulation. In addition, great differences in caffeine metabolism depending on the organ and the species were found (Perrois et al., 2015).

6.3.3.3 KEGG pathway-based analysis using progenitor genomes as reference

There were a large number of SNPs on the same CDS resulting in only 908 CDS from 1,351 SNPs. Blast2GO outputs showed only 176 SNPs/genes encoded enzymes mapped to KEGG pathways accounting for 19%, in which 56 pathways were unique. 104 enzymes were involved in 176 KEGG pathways with 49 unique enzymes (Figure 6.6 and Appendix table S6.5). Among 908 CDS, only one was shared between hom-het and het-het. The majority of SNPs (93%) linking to caffeine was different in frequency of the same alleles in each locus in the bulk, not the type of alleles.



No of SNP	No of CDS	Genes with KEGG	Unique KEGG pathways	No of enzyme in pathways	No of unique enzymes
1,351	908	176	56	104	49

Figure 6.6 Metrics on number of SNPs, CDSs, KEGG pathways and enzymes associating with caffeine.

In the present study, all KEGG pathways recorded were examined thoroughly to narrow down to only pathways and enzymes involved directly or indirect to caffeine biosynthesis pathways. This method has originally been developed to identify disease-related pathways. It uses prior biological knowledge on gene functions to facilitate more powerful analysis of GWA study data sets and examine whether a group of related genes in the same functional pathway are jointly associated with a trait of interest (Wang et al., 2010). The key intention of pathway-based analysis is to identify novel pathways associated with traits instead of candidate causal pathways that represent the way in which the candidate causal SNPs affect traits (Zhang et al., 2011).

Among the 56 KEGG pathways recorded, purine metabolism is the most common pathway with 36 sequences and 7 enzymes involved, followed by thiamine metabolism with 30 sequences. The other pathways with large number of SNPs were biosynthesis of antibiotics (9 sequences), Pyrimidine metabolism (6 sequences) and Starch and sucrose metabolism (5 sequences) (Appendix table S6.5). The 56 KEGG pathways were thoroughly examined to record the processes and enzymes linking to caffeine biosynthesis. There were 10 pathways with 11 enzymes present in 63 sequences or SNPs linked to caffeine biosynthesis through seven substrates or precursors that entered the caffeine pathway (Table 6.5 and Figure 6.7).

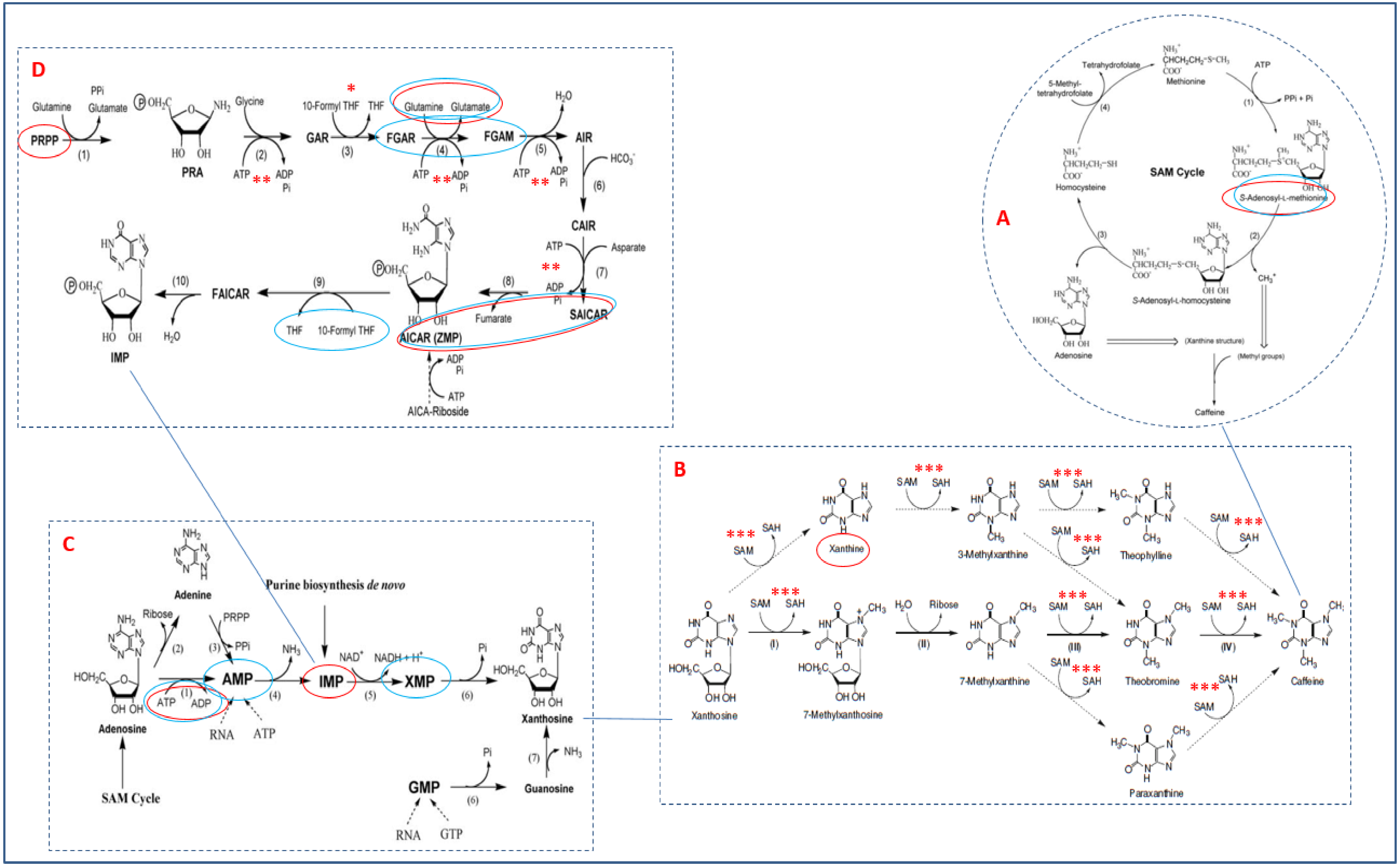

D
PRPP
Glutamine
PPi
Glutamate
(1)
PRA
Glycine
(2)
ATP
ADP
Pi
GAR
10-Formyl THF
THF
(3)
FGAR
Glutamine
Glutamate
(4)
FGAM
(5)
H$_2$O
AIR
HCO$_3^-$
(6)
CAIR
ATP
Asparate
(7)
SAICAR
(8)
Fumarate
AICAR (ZMP)
AICA-Riboside
(9)
THF
10-Formyl THF
FAICAR
(10)
IMP
A
SAM Cycle
Methionine
Tetrahydrofolate
5-Methyl-tetrahydrofolate
(4)
ATP
(1)
PPi + Pi
S-Adenosyl-L-methionine
(2)
CH$_3^+$
S-Adenosyl-L-homocysteine
(3)
Homocysteine
Adenosine
(Xanthine structure)
(Methyl groups)
Caffeine
C
Adenine
Ribose
(2)
PRPP
(3)
PPi
Adenosine
(1)
ATP
ADP
AMP
RNA
ATP
NH$_3$
(4)
Purine biosynthesis de novo
IMP
NAD$^+$
NADH + H$^+$
(5)
XMP
Pi
(6)
Xanthosine
SAM Cycle
GMP
RNA
GTP
Guanosine
(7)
NH$_3$
B
Xanthine
SAM
SAH
3-Methylxanthine
Theophylline
Xanthosine
(I)
7-Methylxanthosine
H$_2$O
Ribose
(II)
7-Methylxanthine
(III)
Theobromine
(IV)
Caffeine
Paraxanthine

Figure 6.7 Substrates involved in caffeine biosynthesis pathways and its other related pathways.

(A) The SAM (S-adenosyl-Lmethionine) cycle (the activated methyl cycle) in plants (Adapted from Ashihara and Suzuki, 2004); (B) The biosynthetic pathways of caffeine from xanthosine (Adapted from Ashihara et al., 2011b); (C) The "provider pathways" for xanthosine synthesis in purine alkaloid forming plants (Adapted from Ashihara and Suzuki, 2004); (D) *De novo* biosynthetic pathway of IMP in plants (Adapted from Ashihara and Suzuki, 2004); Circles: Location of substrates/precursors which were formed in the KEGG pathways catalysed by enzymes encoded by genes carrying the TAVs identified from this study; Red circles: detected using CC and CE as reference; Blue circles: detected using draft arabica as reference; (*) (**) (***) indicated the alternative locations of the same substrate of THF, ATP-ADP and SAM, respectively.

Out of 63 sequences, 39 were unique (CDS) with 40 unique SNPs (Appendix table S6.7). The majority of SNPs (37 out of 40) corresponding to enzymes that are involved in the caffeine biosynthesis pathway were the het-het form. This indicated that the difference in frequency of allele at each locus between two bulks of low and high caffeine is common. According to Schlötterer et al. (2014) if the pools of individuals are large enough and the population under study is randomly crossing, genetic variants that do not contribute to the trait are expected to have the same frequency in both pools while the causal variants or linked polymorphisms will differ in frequency between pools.

The average coverage at the location where SNPs were called was very high for the two bulks (30x for B1 and 45x for B2) compared to a number of studies both pooled-seq (Das et al., 2015; Takagi et al., 2013) and non-pooled-seq (Byers et al., 2012; Clarke et al., 2016; Hamilton et al., 2011; Hulse-Kemp et al., 2015; Zhu et al., 2014) indicating the high quality and confidence of these SNPs. The confidence was supported with the high F/R balance in reads (0.38 in B1 and 0.39 in B2) and the average base quality score (36 in B1 and 35 in B2) as these are a reflection of read quality used in SNP detection. The parameter of average base quality score was much higher than what was usually suggested (i.e. 20) (Clevenger et al., 2015; Das et al., 2015; Nagy et al., 2013). Only three out of 40 SNPs were hom-het which link to an enzyme (EC:6.3.1.2 – synthetase) that form L-glutamate which enter to the IMP biosynthetic pathways (Figure 6.7 – D and Table 6.5), and two enzymes (EC:3.6.1.3 – adenylpyrophosphatase and EC:3.6.1.15 – phosphatase) in purine metabolism pathways that converts ATP to ADP which entered into xanthosine synthesis pathways (Figure 6.7 - C and Table 6.5).

Among the 10 candidate pathways linked to caffeine, the purine metabolism was the most common and generated four substrates (Table 6.5) entering the caffeine pathway. This is not surprising as xanthosine is synthesised via purine. Caffeine is one of the purine alkaloids, and biosynthetic pathways to these purine alkaloids from purine nucleotides in tea and coffee plants have been proposed (Ashihara et al., 1996). Pathways related to amino acid metabolism (cysteine and methionine metabolism; arginine and proline metabolism; alanine, aspartate and glutamate metabolism and arginine biosynthesis) were the most popular, followed by nucleotide metabolism (purine metabolism and pyrimidine metabolism). Other pathways were carbohydrate metabolism (glyoxylate and dicarboxylate metabolism), energy metabolism (nitrogen metabolism) and metabolism of cofactors and vitamins (folate biosynthesis).

Table 6.5 Substrates/precursors, pathways and enzymes involved in caffeine biosynthesis pathway associated with the TAVs identified using CC and CE subgenomes as reference.

Substrates	Pathways	Enzymes	Metabolism[a]	# seq[b]
SAM	- Cysteine and methionine metabolism	- EC:2.5.1.16 - synthase	C	1
		- EC:2.1.1.37 - (cytosine-5-)-methyltransferase	A	1
	- Arginine and proline metabolism	- EC:2.5.1.16 - synthase	C	1
Xanthine	- Purine metabolism	- EC:1.17.1.4 - dehydrogenase	A	1
IMP-XMP	- Purine metabolism	- EC:1.1.1.205 - dehydrogenase	A	1[×]
SAICAR-AICAR	- Purine metabolism	- EC:4.3.2.2 - lyase	A	1
ATP-ADP	- Purine metabolism	- EC:3.6.1.3 – adenylpyrophosphatase	A	19*
		- EC:3.6.1.15 – phosphatase	A	30*
Glutamine-Glutamate	- Pyrimidine metabolism	- EC:6.3.5.5 - synthase (glutamine-hydrolysing)	C	1
	- Alanine, aspartate and glutamate metabolism	- EC:6.3.5.5 - synthase (glutamine-hydrolysing)	C	1
		- EC:6.3.1.2 – synthetase		1
	- Glyoxylate and dicarboxylate metabolism	- EC:6.3.1.2 – synthetase	C	1
	- Arginine biosynthesis	- EC:6.3.1.2 – synthetase	A	1
	- Nitrogen metabolism	- EC:6.3.1.2 – synthetase	A	1
	- Folate biosynthesis	- EC:6.3.2.17 - synthase	A	1
PRPP	- Phenylalanine, tyrosine and tryptophan biosynthesis	- EC:2.4.2.18 - phosphoribosyltransferase	A	1
	10 pathways	**11 enzymes**		**63****

[a] Type of metabolism: C: Catabolism (breakdown) of the substrate; A: Anabolism (synthesis) of the substrate; [b] Sequences where TAVs are located: * Among 49 sequences, 17 was duplicates; ** 63 sequences have 39 unique sequences (CDS) with 40 SNPs; [×] 1 sequence with 2 SNPs; Number in the same colour come from the same sequences (CDS).

The eleven enzymes (where TAVs were located) were involved in three biosynthesis pathways of (1) IMP – an intermediate of the de novo purine biosynthesis pathway and a precursor of xanthosine, (2) xanthosine, the initial purine compound in the caffeine biosynthesis pathway, acting as a substrate for the methyl group donated by SAM (Ashihara and Crozier, 2001) and (3) caffeine (Figure 6.7) including the formation of PRPP (EC:2.4.2.18 – phosphoribosyltransferase in Phenylalanine, tyrosine and tryptophan biosynthesis), a cascades of reactions involved in the conversion of ATP to ADP, SAICAR-AICAR, IMP to XMP in purine metabolism, the conversion of SAM, a methyl donor for methylation reactions in the caffeine biosynthesis pathway, to S-adenosyl-L-homocysteine (SAH) (EC:2.1.1.37 - (cytosine-5-)-methyltransferase in cysteine and methionine metabolism), conversion of SAM to spermidine and spermine (EC:2.5.1.16 - synthase in arginine and proline metabolism) and conversion of hypoxanthine to xanthine (EC:1.17.1.4 – dehydrogenase in purine metabolism). The precursors of caffeine are derived from purine nucleotides and low caffeine accumulation is due mainly to the low biosynthetic activity of purine alkaloids, possibly the extremely weak N-methyltransferase reactions in caffeine biosynthesis (Ashihara et al., 2011b). This is confirmed by the detected TAVs that generally link to the biosynthetic activity of purine alkaloids. The up- and down-regulation of these alleles may have led to the difference in the caffeine content between the two extreme groups.

The 40 TAVs were plotted onto each chromosome of the two subgenomes (Figure 6.8). The most significant TAVs linking to enzymes involved in caffeine synthesis pathway reside in chromosome 7 and 11 (red squares), followed by TAVs linking to enzymes participated in xanthosine synthesis pathway in chromosome 4 (green squares). TAVs associated with enzymes involved in the conversion of ATP to ADP is rather generic and distributed across almost all chromosomes (grey squares), which are not subject to further analysis. TAVs linking to enzymes involved in IMP synthesis pathway (black and blue squares) was distributed in chromosomes 0, 2 and 9. The distribution of 23 caffeine synthase-related NMT genes in *Coffea canephora* was found to be mainly on two chromosomes - Chr1 and Chr9 (Denoeud et al., 2014) while only two and three TAVs were on chr1 and chr9 of the CC subgenome, respectively. This indicates the TAVs controlling the content of caffeine were not in the region of caffeine synthase-related NMT genes.

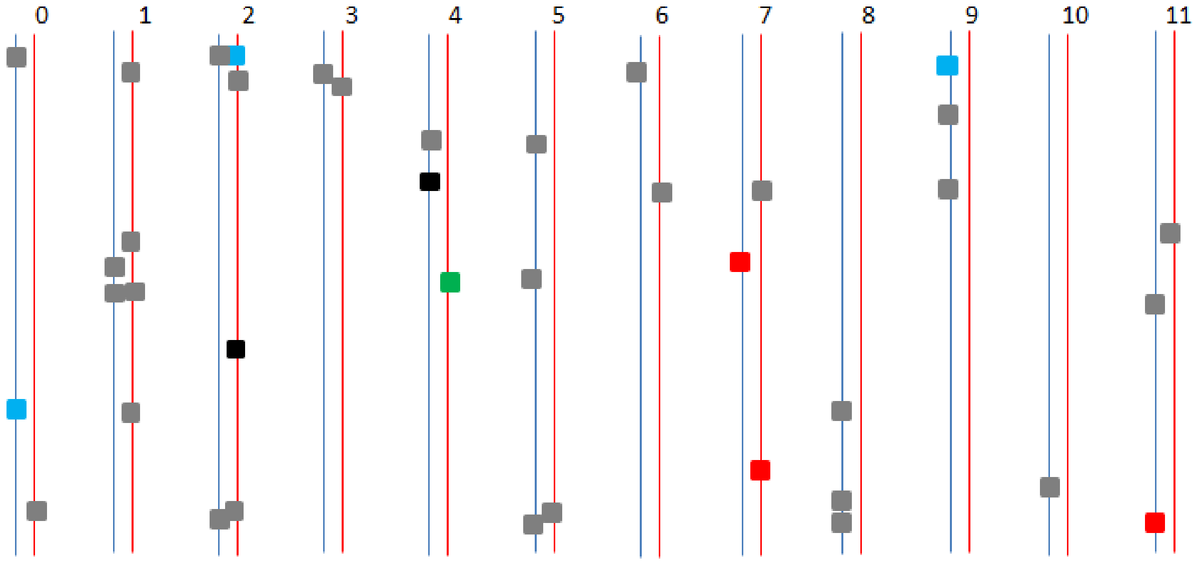


Figure 6.8 Locations of SNPs tightly associated with caffeine synthesis pathway on two subgenomes

(Blue line: CC genome; Red line: CE genome; Black squares: IMP synthesis pathway (PRPP and SAICAR-AICAR; Blue squares: glutamine-glutamate; Grey squares: xanthosine synthesis pathway (ATP-ADP); Green squares: xanthosine synthesis pathway (IMP-XMP); Red squares: caffeine synthesis pathway (xanthine and SAM-SAH).

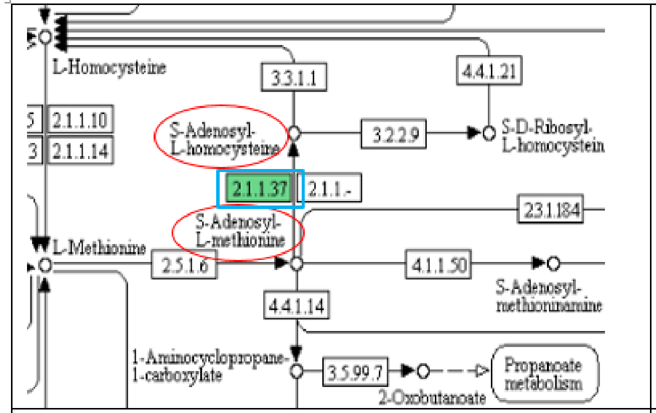
Among 40 SNPs potentially link to the caffeine biosynthesis pathway, the most noticeable one is the SNP associated with enzyme cytosine-5-methyltransferase (EC:2.1.1.37) participating in the conversion of SAM to SAH in the SAM cycle (Figure 6.9 - A1). The SNP was located in chromosome 7 at the 13,417,576-bp position. The low caffeine bulk has more A-allele and less G-allele than the high caffeine bulk. As explained by Guo et al. (1996), the differences in allele dosage may result in differences in the RNA levels of a particular allele and in phenotypic differences. The change of more A in B1 while more G in B2 would lead to a change in amino acid from alanine to threonine (Appendix table S6.7). This enzyme would have an influence on the formation of the methyl group and thus might influence the synthesis of caffeine (Figure 6.7 – A). Ashihara and Crozier (2001) also found that caffeine synthase is inhibited completely by low concentrations of SAH. Therefore, control of the intracellular SAM:SAH ratio is one possible mechanism for regulating the activity of caffeine synthase *in vivo* (Ashihara and Crozier, 2001). This SNP seems to play an important role in the caffeine synthesis. The second SNP which was also involved in the metabolism (breakdown) of SAM to spermidine and spermine (Figure 6.9 – A2) while these two compounds have effects on salinity and drought tolerance recorded in a number of crops (Kasukabe et al., 2006; Li et al., 2016; Roychoudhury et al., 2011). The SNP located in chromosome 11 at 30,965,526 with more C-alleles and less A-alleles in B1 than in B2 (56% vs 30% and 44% vs 70%, respectively) resulting in the change of amino acid from mainly alanine to mainly serine (Appendix table S6.7).

Another significant SNP is the one that is associated with the enzyme that converts hypoxanthine to xanthine (EC:1.17.1.4 – dehydrogenase) (Figure 6.9 - B). This substrate entered the pathways of caffeine biosynthesis to be converted to 3-methylxanthine before being converted to theophylline - a possible direct precursor of caffeine (Figure 6.7 - B). In young and mature leaves of *C. eugenioides* which contain low levels of caffeine, [8-14C] caffeine is catabolised rapidly primarily by the main caffeine catabolic pathway via theophylline. This suggests that the low caffeine accumulation in *C. eugenioides* is a consequence of rapid degradation of caffeine perhaps accompanied by a slow rate of caffeine biosynthesis (Ashihara and Crozier, 1999). In tea and mate (*Ilex paraguariensis),* large amounts of theophylline are also converted to theobromine and caffeine via a theophylline -> 3-methylxanthine -> theobromine -> caffeine salvage pathway (Ito et al., 1997). Xanthine may break down to urate or convert to xanthosine (Figure 6.9 - B) which is the precursor in the main biosynthesis pathway of caffeine (Figure 6.7 – B). It seems that the frequency of C and G allele in the two bulks is affecting the enzyme which catalyses the formation of more or less xanthine, and eventually affects the concentration of caffeine. More individuals with C at this region in a bulk may favour the formation of xanthine and vice versa.

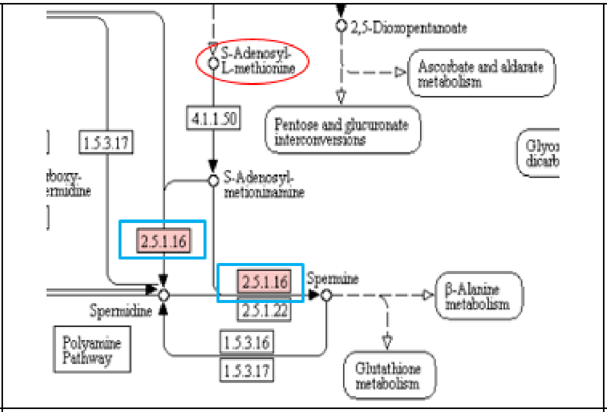
The three SNPs associating with enzymes participating in purine metabolism converting SAICAR to AICAR and IMP to XMP were located in chromosome 2 (at 35,626,636) and chromosome 4 (11,359,176 and 11,364,138) subgenome CE (Appendix table S6.7). The enzyme EC:4.3.2.2 – lyase can convert SAICAR to AICAR or vice versa while enzyme EC:1.1.1.205 – dehydrogenase converts IMP to XMP (Figure 6.9 – C and D) and both enter the xanthosine biosynthesis pathway (Figure 6.7). SNP alleles at these locations were het-het and different in frequency resulting in a change in amino acid. The presence of more or less of the specific SNP alleles in each individual may have contributed to the synthesis and conversion of the aforementioned substrates causing the difference in the concentration of caffeine between the two extreme groups.

Glutamine is converted to glutamate in the IMP biosynthetic pathway which eventually produces xanthosine – the first substrate in the caffeine synthesis pathway (Figure 6.7 –B, C and D). One out of three SNPs links to enzymes that catalyse the formation of glutamine or glutamate (Figure 6.9 – E1, E2 and E3) is hom-het SNP (Appendix table S6.7). At this locus (chromosome 1, at 197,044,125), the low-caffeine bulk has T-allele only (100%) while the high-caffeine bulk has T-allele (61%) and C-allele (39%) resulting in the change of amino acid (lysine only in the low bulk, both glutamic acid and lysine in the high bulk) (Appendix table S6.7).

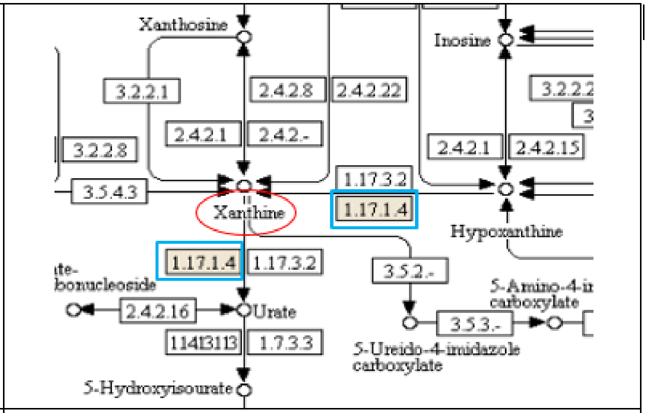
PRPP is the first substrate in the IMP biosynthetic pathway which at last produce xanthosine - the first substrate in the caffeine synthesis pathway (Figure 6.7 –B, C and D). The SNP linking to the enzyme (EC:2.4.2.18 – phosphoribosyltransferase) that metabolises PRPP (Figure 6.9 - F) has more A-alleles and less T-alleles in the low bulk than in the high bulk (75 vs 43% and 25 vs 57%, respectively), resulting in the change of amino acid (more threonine than serine in the high bulk) (Appendix table S6.7).



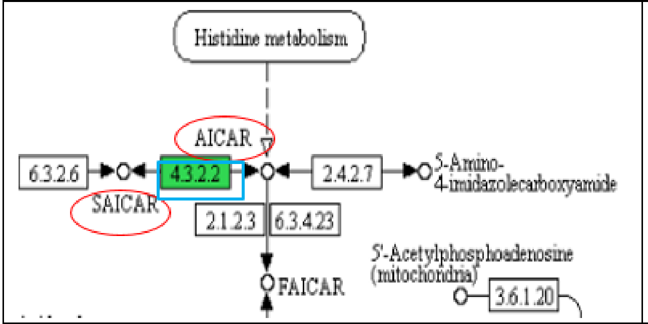
(A1) SAM - Cysteine and methionine metabolism - EC:2.1.1.37



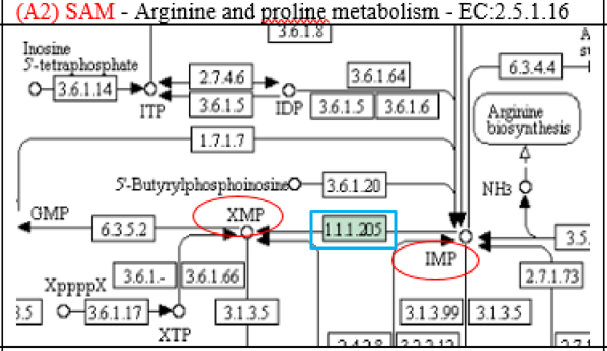
(A2) SAM - Arginine and proline metabolism - EC:2.5.1.16



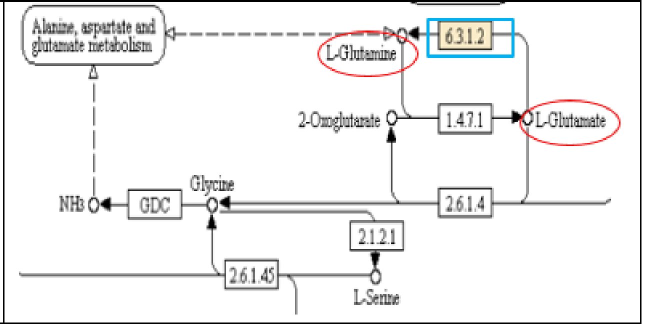
(B) Xanthine - Purine metabolism - EC:1.17.1.4



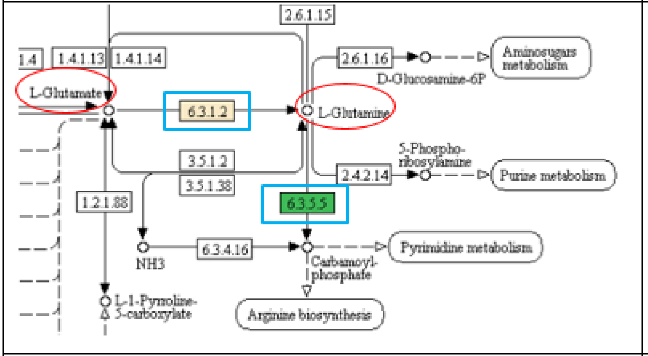
(C) AICAR-SAICAR - Purine metabolism - EC:4.3.2.2



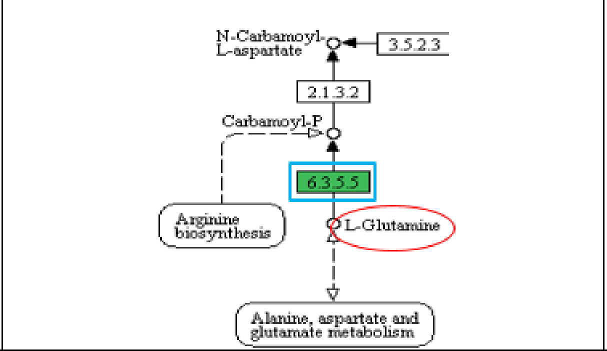
(D) IMP-XMP - Purine metabolism - EC:1.1.1.205



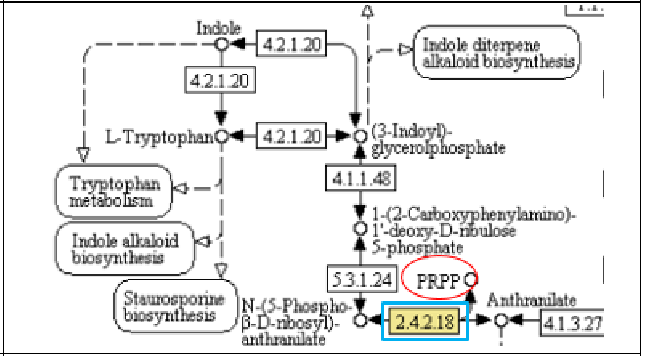
(E1) Glutamine-Glutamate - Glyoxylate and dicarboxylate metabolism - EC:6.3.1.2



(E2) Glutamine-Glutamate - Alanine, aspartate and glutamate metabolism - EC:6.3.1.2



(E3) Glutamine-Glutamate - Pyrimidine metabolism - EC:6.3.5.5


Indole
4.2.1.20
Indole diterpene alkaloid biosynthesis
4.2.1.20
L-Tryptophan
4.2.1.20
(3-Indoyl)-glycerolphosphate
Tryptophan metabolism
4.1.1.48
Indole alkaloid biosynthesis
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose 5-phosphate
5.3.1.24
PRPP
Staurosporine biosynthesis
N-(5-Phospho-β-D-ribosyl)-anthranilate
2.4.2.18
Anthranilate
4.1.3.27

(F) PRPP - Phenylalanine, tyrosine and tryptophan biosynthesis - EC:2.4.2.18

Figure 6.9 Snap shots of the KEGG pathways (obtained from Blast2GO analysis) at the location where SNPs associating with enzymes involved in the metabolism of the substrates that entered to the caffeine biosynthesis pathway.

(Substrates are circled in red; enzymes are highlighted in blue squares).

ATP (adenosine triphosphate) and ADP (adenosine diphosphate) are organic nucleotide molecules. ATP is converted to ADP in the cells of plants and animals when energy is required to power processes in the cell with the energy released. Energy is also released when a phosphate is removed from ADP to form adenosine monophosphate (AMP). Although AMP is a substrate that participate in the synthesis pathway of xanthosine, ATP and ADP are generic substrates for chemical reactions so their TAVs will not be discussed further. Instead, two SNPs associated with two enzymes (EC:3.6.1.3 – adenylpyrophosphatase and EC:3.6.1.15 – phosphatase) in purine metabolism pathways that convert ATP to ADP interacting with the xanthosine synthetic pathways (Figure 6.7 – C and Table 6.5) were considered as they are in the hom-het form (Appendix Table S6.7). This SNP at 11,004,936 (chromosome 11 of *C. eugenioides* subgenome) was homozygous allele in the low caffeine bulk (100% G-allele) while heterozygous alleles (G-and A-allele of 74% and 26%, respectively) were identified in the high caffeine bulk. The presence of the A-allele in the high bulk probably elevated the conversion from ATP to ADP. At the other SNP at 10,465,912 (chromosome 6 of *C. eugenioides* subgenome), only C-allele (100%) was found in the low caffeine bulk while in the high caffeine bulk both C-allele (32%) and A-allele (68%) were recorded. Similarly, the presence of the A-allele in the high bulk probably elevated the conversion from ATP to ADP resulting in more caffeine being synthesised. Accoding to Koshiro et al. (2006) active supply of the substrates from purine nucleotides (i.e. ATP, ADP and AMP) is important for the biosynthesis of caffeine in coffee fruits.

It is obvious that the availability of SNPs within coding sequences is a very powerful tool for molecular geneticists to detect a causative mutation (Varshney, 2009). However, often QTLs are found located in noncoding regulatory sequences such as enhancers or locus control regions, which could be located several megabases away from genes within intergenic spaces (Dean, 2006). Promoter, intron, exon, and 5′/3′-untranslated regions are all reasonable targets for identifying candidate gene SNPs, with non-coding regions expected to have higher levels of nucleotide diversity than coding regions (Zhu et al., 2008). Therefore, SNPs in other regions (5’ UTR, 3’UTR, regulatory and intron) near these shortlisted SNPs were investigated. There were nine SNPs in the 5’ prime UTR and 3’ prime UTR of genes linking to the formation of three substrates targeting PRPP and XMP (Table 6.6). Further functional validation of these SNPs may explain the up or downregulation of these alleles, which may affect to the caffeine content, especially the het-hom SNP in PRPP.

Table 6.6 TAVs in non-coding regions of the genes involved in caffeine pathways

Substrates	Location in gene [(*)]	Chromosome	Region	Reference	Allele	Frequency in B1	Frequency in B2
PRPP	a	CC1.1ch04	12730620	C	T	22	100
					C	78	0
IMP-XMP	a	CE1.1ch04	11359176	C	G	59	42
					C	41	58
	a	CE1.1ch04	11362237	A	C	35	55
					A	65	45
ATP-ADP	a (5); b (5); c(7) - not provide in details						

(*) a - SNPs in 5' or 3'UTR; b – SNPs possible in regulatory region near 3' or 5' prime UTR (100-1000 bp from 3' or 5' prime UTR); c - Intron retention: SNPs in intron region but aligned with in-house transcripts. SNPs link to ATP-ADP were not reported in details as they are so common.

The sequences of published caffeine genes were also imported in the CLC to check the SNPs within these genes between two bulks. Three non-synonymous SNPs from one gene of 3,7-dimethylxanthine methyltransferase were found. However, this gene has no KEGG pathway from the database developed in Blast2GO probably because the KEGG pathways include only well-known enzymes. This also indicates the advantages of applying whole genome sequencing approach over the targeted genome sequencing approach in SNPs detection. If targeted sequencing has been applied based on these published caffeine genes, there would be few or no TAS detected.

6.3.3.4. KEGG pathway-based analysis using arabica draft genome as reference

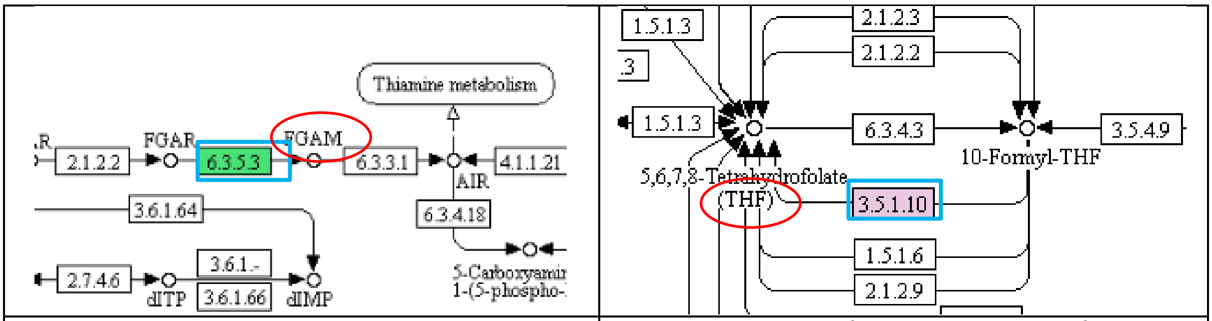
When using K7 arabica draft genome as reference in mapping and variant call, 1,444 non-synonymous SNPs were identified in which 39 was hom-het or het-hom and 1,405 was het-het. 1,086 CDS was extracted and imported to Blast2GO detecting 189 genes containing KEGG pathways in which 70 pathways and 80 enzymes were unique. KEGG pathway-based analysis showed several additional pathways and enzymes linking to caffeine synthesis pathways which were not detected when using CC and CE as reference, and the most important enzymes catalysing SAM to SAH were also confirmed (Table 6.7).

Table 6.7 Additional pathways and enzymes involved in caffeine biosynthesis process where TAVs were identified when using draft arabica genome as reference.

Substrates	Pathways	EC	Metabolism [(*)]	# seq
SAM (confirmed)	Cysteine and methionine metabolism	EC:2.1.1.37-(cytosine-5-)-methyltransferase	A	1
FGAM	Purine metabolism	EC:6.3.5.3 - synthase	A	2
XMP	Purine metabolism	EC:6.3.5.2 - synthase (glutamine-hydrolysing)	C	2
AMP	Purine metabolism	EC:4.3.2.2 - lyase	C	1
10-Formyl-THF	One carbon pool by folate	EC:3.5.1.10 - deformylase	A	1
	3 pathways	**5 enzymes**		**7**

[(*)] A: Anabolism (synthesis) of the substrate

The conversion of SAM-SAH by enzyme EC:2.1.1.37 - (cytosine-5-)-methyltransferase is very important as it is a methyl donor for the formation of caffeine. Enzymes catalysed for this reaction was confirmed when using the arabica reference. However, this SNP is different from the one detected by CC and CE genomes. The other four new TAVs associating with enzymes involved in caffeine biosynthesis were detected using K7 as a reference (Table 6.7). Three enzymes convert FGAR to FGAM (Figure 6.10 – A), AMP to GMP and adenylo-succinate to AMP were EC:6.3.5.3 – synthase, EC:6.3.5.2 - synthase (glutamine-hydrolysing) and EC:4.3.2.2 – lyase, respectively. These substrates are involved in the IMP and xanthosine synthesis pathway (Figure 6.7 - C and D). 10-Formyl-THF acts as a donor of formyl groups in anabolism of THF (Figure 6.10 - B) and participates in the conversion of GAR to FGAR or AICAR to FAICAR in the synthetic pathway of IMP (Figure 6.7 - D).



(A) Purine metabolism - EC:6.3.5.3 - synthase (B) One carbon pool by folate - EC:3.5.1.10 - deformylase

Figure 6.10 Snap shots of the KEGG pathways at the location where TAVs associating with enzymes involved in the metabolism of the substrates that entered to the caffeine biosynthesis when using arabica as reference genome.

(Substrates are circled in red; enzymes are highlighted in blue squares).

6.3.4. Identification of Trigonelline associated SNPs

6.3.4.1. SNP detection and KEGG pathway-based analysis for trigonelline trait using two progenitor genomes as reference.

Since the average coverage of the two DNA bulks for low and high trigonelline (38 and 24) is comparable to those for caffeine (28 and 40), the best setting of variant calling for caffeine trait (setting 3) was also applied to trigonelline. With this setting the percentage of correct call after manually checking the SNPs with variant call track and mapping track was 89%. In total 1,060 non-synonymous SNPs were detected. Similar to caffeine, only 0.17% of tri-allele SNPs were detected and most of SNPs were het-het (bi-allele) (99%) confirming the loss of heterozygosity within the two ancestral genomes and only one allele from each genome remains in arabica. Out of 1,060 SNPs, only 14 SNPs were different in the type of alleles at the SNP loci or hom-het and het-hom type (Table 6.8).

Table 6.8 Number of SNPs distinguishing the low- and high-trigonelline bulks based on setting 3.

Settings	Type (*)	Variant	Non-synonymous	Di-allele	Tri-allele	% tri-allele	Filter (**)	Check mapping	Correct call (%)
Setting 3 20-4-20-0.20	HomB1-HetB2	664	14					9	
	HetB1-HomB2	10,815	130					5	
	HetB1-HetB2	1,088,441	16,017	15,990	27	0.17	1,046	1,046	
	Total	**1,099,920**	**16,161**					**1,060**	**89**

(*) Hom: Homozygous; Het: Heterozygous; B1: Bulk 1; B2: Bulk 2; (**) Filtered by using Chi-square test and If command to identify the SNPs that were significant different in frequency between the two bulks and non-synonymous between the two bulks.

Like caffeine, there were more TAVs for trigonelline on the subgenome of *C. canephora* (562 SNPs) than *C. eugenioides* (498 SNPs). The distribution of TAVs across genome indicates that trigonelline is a quantitative trait controlled by multiple genes (Figure 6.11).

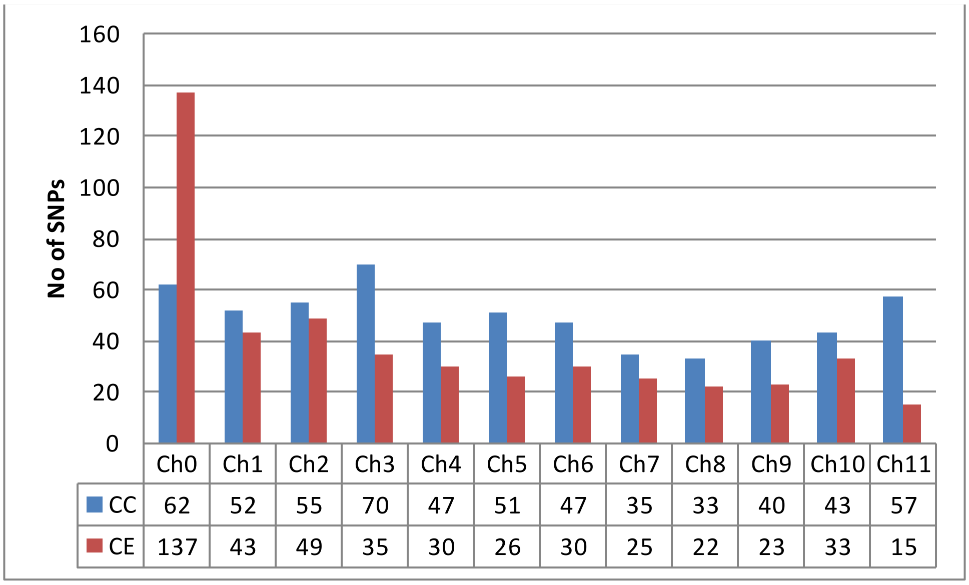


Figure 6.11 The genome-wide distribution of TAVs for trigonelline in two CC and CE subgenomes.

1,060 SNPs were detected with high average sequencing depth (47x for Bulk 3 - B3 and 29x for Bulk 4 - B4), high value of F/R balance (0.39 for B3 and 0.38 for B4) and high average base quality score (35 and 36 for B3 and B4, respectively) (Appendix table S6.8) indicating the high quality and confidence of SNP identified (Figure 6.12).

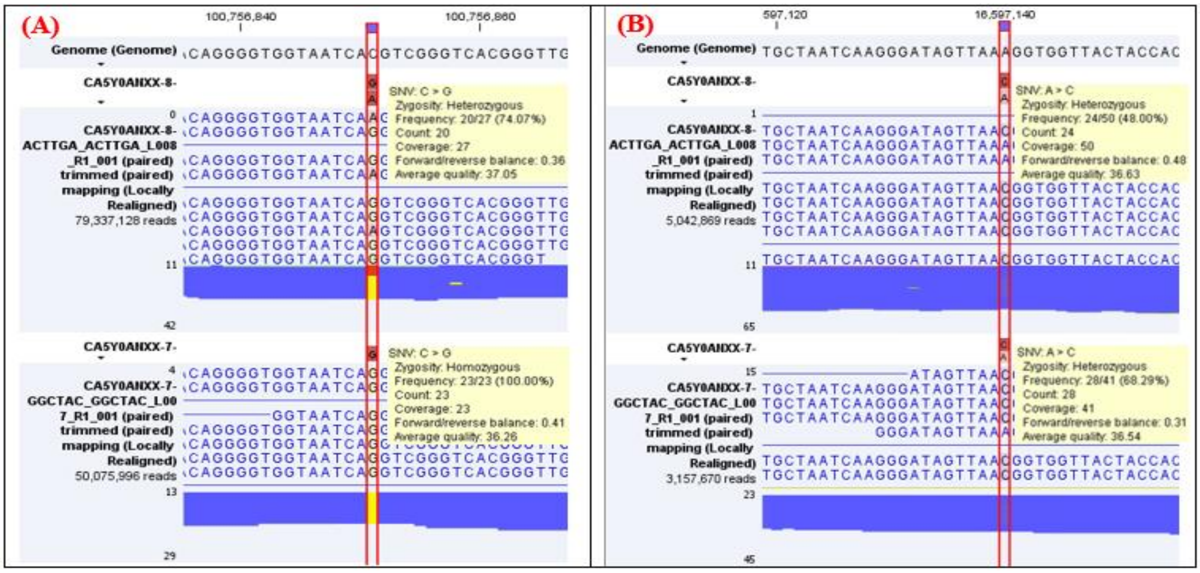
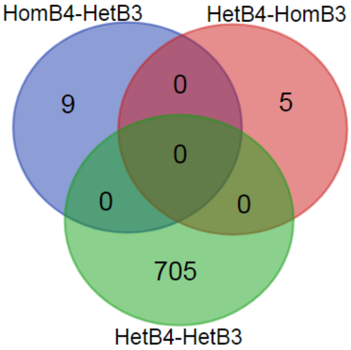

(A)
100,756,840
100,756,860
Genome (Genome)
CAGGGGTGGTAATCACGTCGGGTCACGGGTTG
CA5Y0ANXX-8-ACTTGA_ACTTGA_L008_R1_001 (paired) trimmed (paired) mapping (Locally Realigned)
79,337,128 reads
SNV: C > G
Zygosity: Heterozygous
Frequency: 20/27 (74.07%)
Count: 20
Coverage: 27
Forward/reverse balance: 0.36
Average quality: 37.05
CA5Y0ANXX-7-GGCTAC_GGCTAC_L007_R1_001 (paired) trimmed (paired) mapping (Locally Realigned)
50,075,996 reads
SNV: C > G
Zygosity: Homozygous
Frequency: 23/23 (100.00%)
Count: 23
Coverage: 23
Forward/reverse balance: 0.41
Average quality: 36.26
(B)
597,120
16,597,140
Genome (Genome)
TGCTAATCAAGGGATAGTTAAAGGTGGTTACTACCAC
CA5Y0ANXX-8-ACTTGA_ACTTGA_L008_R1_001 (paired) trimmed (paired) mapping (Locally Realigned)
5,042,869 reads
SNV: A > C
Zygosity: Heterozygous
Frequency: 24/50 (48.00%)
Count: 24
Coverage: 50
Forward/reverse balance: 0.48
Average quality: 36.63
CA5Y0ANXX-7-GGCTAC_GGCTAC_L007_R1_001 (paired) trimmed (paired) mapping (Locally Realigned)
3,157,670 reads
SNV: A > C
Zygosity: Heterozygous
Frequency: 28/41 (68.29%)
Count: 28
Coverage: 41
Forward/reverse balance: 0.31
Average quality: 36.54

Figure 6.12 Examples of highly confident TAVs for trigonelline that distinguish between two bulks (B3 and B4) in the types of hom-het (A) and het-het (B)

(A) hom-het type: B4 was homozygous with 100% of G with coverage of 23, F/R balance of 0.41 and average quality of 36.26; B3 was heterozygous with 74% of G and 26% of A with coverage of 27, F/R balance of 0.36 and average quality of 37.05 (B) het-het type (both bulks are heterozygous) B3 with 48% of C and 54% of A with coverage of 50, F/R balance of 0.48 and average quality of 36.63 while B4 with 68% of C and 34% of A with coverage of 41, F/R balance of 0.31 and average quality of 36.57. Red lines are the location of SNPs.

There were a number of SNPs on the same CDS resulting in only 719 CDS from 1,060 SNPs. Blast2GO outputs showed only 125 genes containing KEGG pathways in which 61 pathways were unique. 106 enzymes were involved in 125 KEGG pathways with 51 unique enzymes (Figure 6.13). Among 719 CDS, none was shared between different categories of SNPs. The majority of SNPs (98%) linking to trigonelline is het-het (i.e. different in frequency of the same alleles in each locus) between two bulks of low and high trigonelline.



No of SNP	No of CDS	Genes with KEGG	Unique KEGG pathways	No of enzyme in pathways	No of unique enzymes
1,060	719	125	61	106	51

Figure 6.13 Metrics on the number of SNPs, CDSs, KEGG pathways and enzymes associated with trigonelline biosynthesis.

Among the 61 KEGG pathways recorded, purine metabolism was the most frequent pathway with 20 sequences and five enzymes involved, followed by thiamine metabolism with 17 sequences. The other popular pathways were biosynthesis of antibiotics (10 sequences), Phenylpropanoid biosynthesis (5 sequences) and starch and sucrose metabolism (5 sequences) (Appendix Table S6.6). The 61 KEGG pathways were thoroughly examined to determine the pathways and enzymes linking to trigonelline biosynthesis pathways. There were seven pathways with nine enzymes located in 36 sequences where TAVs are located. Among these pathways, four were related to amino acid metabolism (cysteine and methionine metabolism; arginine and proline metabolism; glycine, serine and threonine metabolism and phenylalanine, tyrosine and tryptophan biosynthesis) and two were related to carbohydrate metabolism (starch and sucrose metabolism and galactose metabolism). Only one pathway relating to nucleotide metabolism (purine metabolism) was involved. The links to trigonelline biosynthesis pathways in which four substrates or precursors entered the trigonelline pathway were found (Table 6.6 and Figure 6.14).

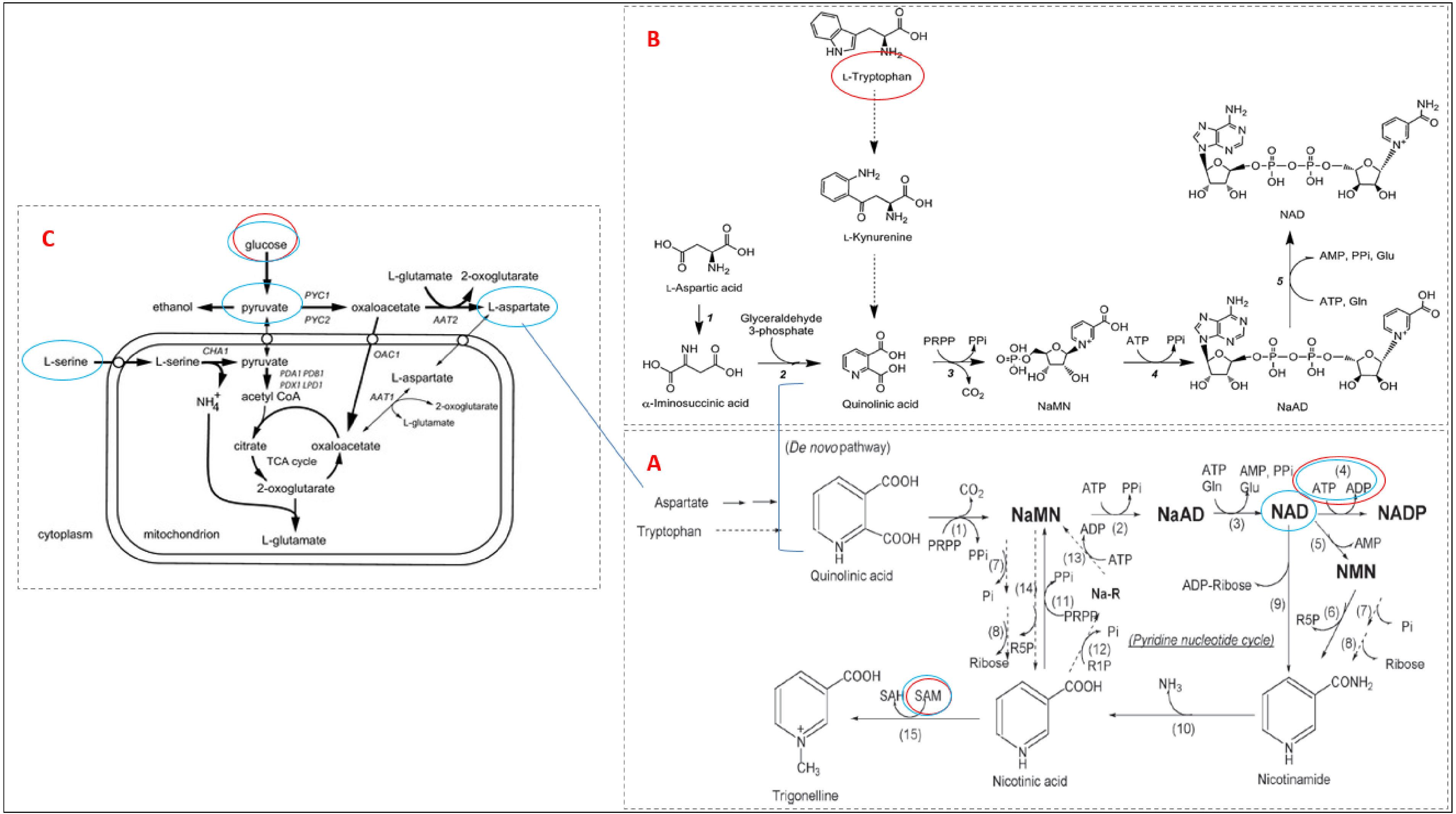

C
glucose
ethanol
pyruvate
PYC1
PYC2
oxaloacetate
L-glutamate
2-oxoglutarate
AAT2
L-aspartate
L-serine
CHA1
pyruvate
OAC1
PDA1 PDB1
PDX1 LPD1
acetyl CoA
NH_4^+
AAT1
2-oxoglutarate
L-glutamate
citrate
oxaloacetate
TCA cycle
2-oxoglutarate
cytoplasm
mitochondrion
L-glutamate
B
L-Tryptophan
L-Kynurenine
L-Aspartic acid
1
α-Iminosuccinic acid
Glyceraldehyde 3-phosphate
2
Quinolinic acid
PRPP
PPi
3
CO_2
NaMN
ATP
PPi
4
NaAD
AMP, PPi, Glu
5
ATP, Gln
NAD
A
(De novo pathway)
Aspartate
Tryptophan
Quinolinic acid
CO_2
(1)
PRPP
PPi
NaMN
ATP
PPi
ADP (2)
NaAD
ATP
Gln
AMP, PPi
Glu
(3)
NAD
(4)
ATP
ADP
NADP
(5)
AMP
NMN
(13)
ATP
(7)
Pi
(14)
PPi
(11)
PRPP
Na-R
(8)
R5P
Ribose
(12)
Pi
R1P
ADP-Ribose
(9)
R5P
(6)
(7)
Pi
(8)
Ribose
(Pyridine nucleotide cycle)
SAH
SAM
(15)
Trigonelline
Nicotinic acid
NH_3
(10)
Nicotinamide

Figure 6.14 Substrates involved in trigonelline biosynthesis and its other related pathways.

(A - Possible metabolic pathways of biosynthesis of trigonelline and pyridine nucleotides in *Coffea arabica* plants (Adapted from Zheng et al., 2004); B- Aspartate pathway of the pyridine nucleotides biosynthesis de novo in plants (Adapted from Ashihara et al., 2015); C - Pathway for conversion of excess L-serine to L-aspartate (Adapted from Lee et al., 2013). Circles: Location of substrates/precursors which were formed in the KEGG pathways catalysed by enzymes that associating with SNPs; Red circles: detected using CC and CE as reference; Blue circles: detected using draft arabica as reference).

Among 36 sequences, 24 had CDS with 24 unique SNPs (Appendix table S6.8). Similar to caffeine, the majority of SNPs (23/24) associated with enzymes that involved in the trigonelline biosynthesis pathway were het-het between two bulks, indicating that the allele frequency difference at each locus between two bulks of low and high trigonelline is dominant. The average coverage was 49x for B3 and 30x for B4, indicating the high quality and confidence of these SNPs. The F/R balance in reads (0.33 in B3 and 0.38 in B4) and the average base quality score (35 in B3 and 36 in B4) supported for the confidence of SNPs quality (Appendix table S6.8). Only one out of 24 SNPs were hom-het which links to enzyme (EC:3.2.1.26 – invertase in starch and sucrose metabolism pathways) that conversed sucrose-6-P to D-glucose-6P while glucose is a substrate that enters into the pathways to form pyruvate, followed by the conversion of L-aspartate – a precursor in the trigonelline pathway (Figure 6.14 - C and Table 6.9).

Table 6.9 Potential SNPs associating with enzymes directly or indirectly involved in trigonelline biosynthesis pathway detected by using CC and CE as reference.

Substrates	**Pathways**	**EC**	**Metabolism**	**# seq**
SAM	- Cysteine and methionine metabolism - Arginine and proline metabolism	- EC:2.5.1.16 – synthase - EC:2.5.1.16 - synthase	C C	1 1
L-tryptophan	- Glycine, serine and threonine metabolism - Phenylalanine, tyrosine and tryptophan biosynthesis	- EC:4.2.1.20 – synthase - EC:4.2.1.20 - synthase	A A	1 1
Glucose	- Starch and sucrose metabolism - Galactose metabolism	- EC:3.2.1.21 – gentiobiase - EC:3.2.1.3 - 1,4-alpha-glucosidase - EC:3.2.1.26 – invertase - EC:3.2.1.39 - endo-1,3-beta-D-glucosidase - EC:3.2.1.22 - melibiase	A A A A A	2 1 1 1 1
ATP-ADP	- Purine metabolism	- EC:3.6.1.15 - phosphatase - EC:3.6.1.3 - adenylpyrophosphatase	A A	17** 9**
	7 Pathways	**9 enzymes**		**36***

C: Catabolism (breakdown) of the substrate; A: Anabolism (synthesis) of the substrate; * Among 26 sequences, 9 was duplicates; ** 36 sequences have 24 unique sequences (CDS) with 24 SNPs; Number in the same colour come from the same sequences (CDS).

There were nine enzymes participated to seven pathways catalysing action on substrates that directly or indirectly entered into the pathways of trigonelline biosynthesis. 24 SNPs were plotted on the chromosomes of two subgenomes (Figure 6.15). SNPs associated with enzymes involved in the reaction to produce SAM (highlighted in red) were most noticeable, followed by SNPs associated with enzyme involved in the reaction to produce L-tryptophan (black squares). There were five SNPs involved in the metabolism of glucose which eventually enters trigonelline synthesis (blue squares) and 17 were associating with the conversion of ATP to ADP (grey squares), which is very generic in involvement in chemical reactions.

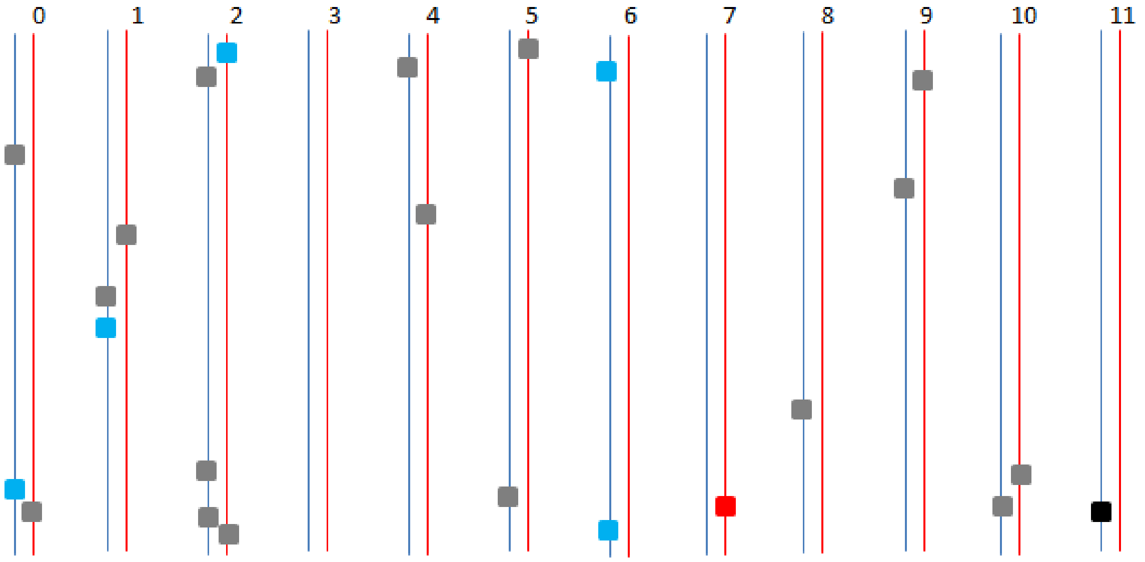
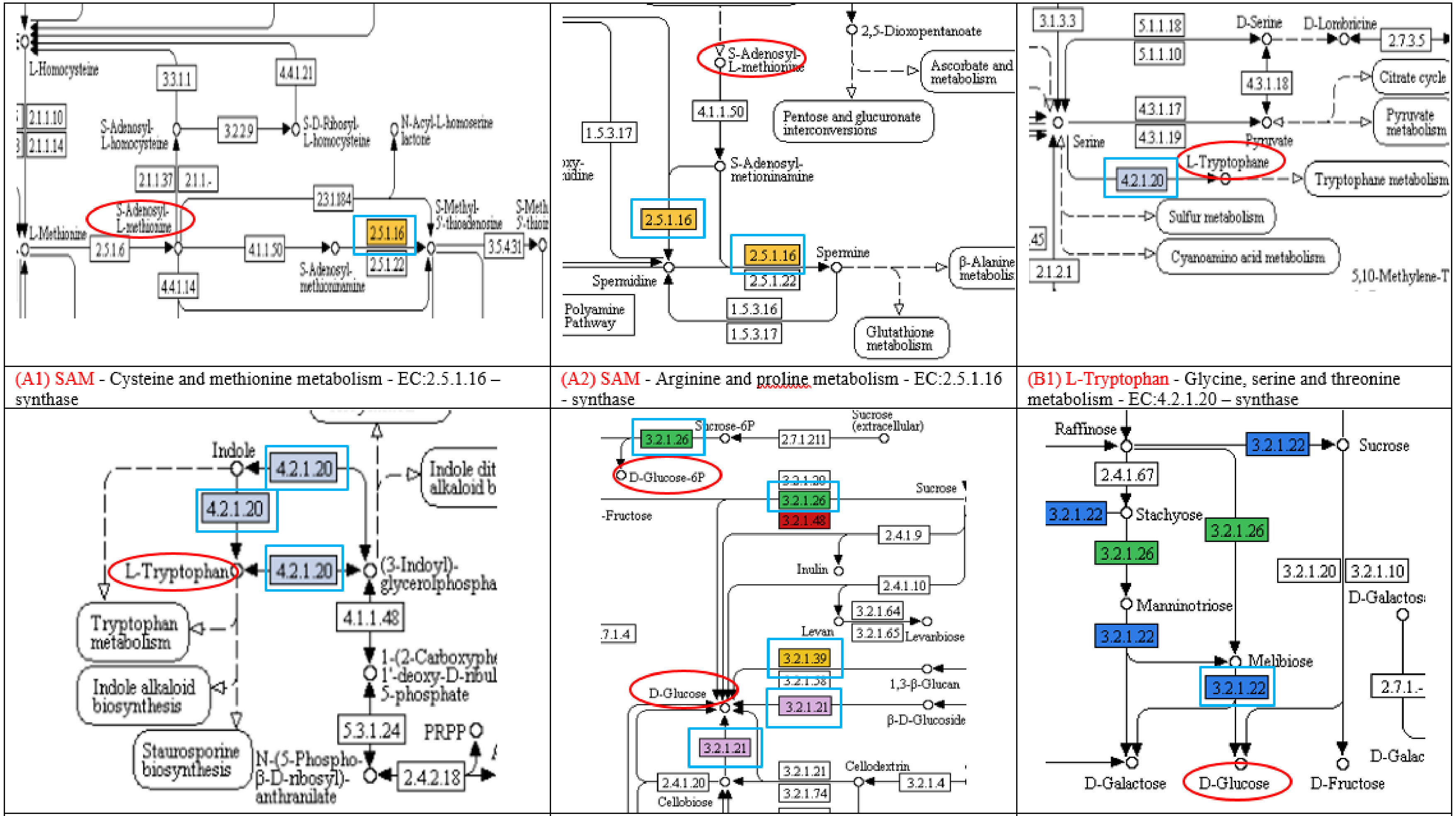


Figure 6.15 Locations on two subgenomes where SNPs are tightly associated with trigonelline synthesis pathway.

(Blue line: CC genome; Red line: CE genome; Blue squares: Glucose; Black squares: L-tryptophan; Grey squares: ATP-ADP; Red squares: SAM-SAH).

Mizuno et al. (2014) found that production of trigonelline from nicotinate is catalyzed by a N-methyltransferase, as is caffeine synthase. In addition, the expression profiles for two genes homologous to caffeine synthases were similar to the accumulation profile of trigonelline. With the assumption that these two homologous genes encoding both caffeine synthases and trigonelline synthases, using the N-methyltransferase assay with S-adenosyl[methyl-14C]methionine, these recombinant enzymes catalyzing the conversion of nicotinate to trigonelline was confirmed. The coffee trigonelline synthases (termed CTgS1 and CTgS2) were highly identical (over 95% identity) to each other. The sequence homology between the coffee trigonelline synthases and coffee caffeine synthase (CCS1) was 82%. In the present study, there was one SNP associating with an enzyme (EC:2.5.1.16 – synthase) involved in the metabolism of S-adenosyl-methioninamine - a breakdown substrate of SAM which also participated in the caffeine biosynthesis pathway detected in two pathways (Cysteine and methionine metabolism and Arginine and proline metabolism) (Figure 6.16 – A1 and A2). Similar to caffeine, the breakdown of SAM to S-adenosyl-methioninamine, and to spermidine and spermine was recorded (Figure 6.16 – A2) while these two compounds have effects on salinity and drought tolerance recorded in a number of crops (Kasukabe et al., 2006; Li et al., 2016; Roychoudhury et al., 2011). Tramontano and Jouve (1997) also found that trigonelline has a role as an osmoregulator in salt-stressed legumes. SNP at this locus was het-het for two bulks with the presence of allele C (75% in B4 and 48% in B3) and A (25% in B4 and 52% in B3) resulting in the change of amino acid with more alanine than serine in B4 (high trigonelline) but more serine than alanine in B3 (low trigonelline) (Appendix table S6.8).

L-tryptophan was converted from serine (in glycine, serine and threonine metabolism) or from indole (in phenylalanine, tyrosine and tryptophan biosynthesis) by the same enzyme (EC:4.2.1.20 – synthase) (Figure 6.16 – B1 and B2). L-tryptophan is the substrate converting to L-kynurenine followed by quinolinic acid which is a precursor for trigonelline synthesis (Figure 6.14). SNP at this locus was het-het for two bulks with the presence of allele C (45% in B4 and 74% in B3) and T (55% in B4 and 26% in B3) resulting in the change of amino acid with more serine than glycine in B4 (high trigonelline) but more glycine than serine in B3 (low trigonelline) (Appendix table S6.8).



(A1) SAM - Cysteine and methionine metabolism - EC:2.5.1.16 – synthase

(A2) SAM - Arginine and proline metabolism - EC:2.5.1.16 - synthase

(B1) L-Tryptophan - Glycine, serine and threonine metabolism - EC:4.2.1.20 – synthase

(B2) L-Tryptophan - Phenylalanine, tyrosine and tryptophan biosynthesis - EC:4.2.1.20 – synthase

(C1) Glucose - Starch and sucrose metabolism - EC:3.2.1.21 – gentiobiase, EC:3.2.1.26 – invertase, EC:3.2.1.39 - endo-1,3-beta-D-glucosidase.

(C2) Glucose - Galactose metabolism - EC:3.2.1.22 – melibiase.

Figure 6.16 Snap shots of the KEGG pathways at the location where SNPs associated with enzymes involved in the metabolism of the substrates that entered the trigonelline biosynthesis pathway. (Substrates are circled in red; enzymes are highlighted in blue square).

Glucose was converted from different substrates such as sucrose, β-glycan, D-glucoside, and melibiose (Figure 6.16 – C1 and C2) by two pathways with five enzymes (starch and sucrose metabolism with four enzymes and galactose metabolism with one enzyme) (Table 6.9). Glucose enters into the pathways to form pyruvate, followed by the conversion of L-aspartate – a precursor in trigonelline pathway (Figure 6.11). Six SNPs were detected from these pathways in which the most noticeable SNP was hom-het between high and low trigonelline bulks. G was the only allele (100%) for B4 while both G (77%) and A (23%) presented in B3 resulting in the change in amino acid with only Arginine in B4 and both arginine and lysine in B3.

There were 17 SNPs associated with enzymes involved in the conversion of ATP to ADP. As explained previously, AMP could be the product of ATP-ADP conversion when a phosphate is removed from ADP and AMP is the substrate in xanthosine synthetic pathway. However, ATP-ADP is too generic for chemical reaction and therefore there was no further investigation for these SNPs.

6.3.4.2. SNP detection and KEGG pathway-based analysis for trigonelline trait using arabica as reference.

When using K7 arabica as a reference in mapping and variant calling, 1,024 non-synonymous SNPs were identified of which only five were hom-het or het-hom and 1,019 was het-het. 796 CDS was extracted and imported to Blast2GO detecting 142 genes contained in KEGG pathways of which 66 pathways and 68 enzymes were unique. KEGG pathway-based analysis showed several additional pathways and enzymes linking to trigonelline synthesis pathways which were not detected when using CC and CE as reference. TAV associating with enzyme (EC:4.1.1.50 – decarboxylase) catalysing SAM to S-adenosyl-methioniamine (Figure 6.17 – A) was the most important TAV, followed by TAVs linking to two enzymes (EC:1.6.99.3 - dehydrogenase and EC:1.6.5.3 - reductase (H+-translocating) converting NADH to NAD (Table 6.10 and Figure 6.17 - B). Another four important TAVs were the ones associating with five enzymes involved in the metabolism of three substrates (L-aspartate, pyruvate and L-serine) which enter to an L-aspartate pathway – a substrate for trigonelline biosynthesis (Figure 6.14 – C, Table 6.10 and Figure 6.17 – C and D).

Table 6.10 Additional pathways and enzymes involved in the trigonelline biosynthesis process with TAVs identified when using the draft arabica genome as reference.

Substrates	Pathways	EC	Metabolism	# seq
SAM	Cysteine and methionine metabolism	EC:4.1.1.50 – decarboxylase (*)	C	1
NAD	Oxidative phosphorylation	EC:1.6.99.3 - dehydrogenase	A	1
		EC:1.6.5.3 - reductace (H+-translocating)	A	1
L-aspartate	Alanine, aspartate and glutamate metabolism	EC:4.3.2.2 - lyase	C	1
Pyruvate	Glycolysis / Gluconeogenesis	EC:4.2.1.11 - hydratase	A	1
	Methane metabolism		A	1
	Glycine, serine and threonine metabolism	EC:4.3.1.17 - ammonia-lyase	A	1
		EC:4.3.1.18 - ammonia-lyase	A	1
L-serine	Glyoxylate and dicarboxylate metabolism	EC:2.1.2.1 – hydroxymethyltransferase	A	1
	Methane metabolism		A	1
	Glycine, serine and threonine metabolism		A	1
	Cyanoamino acid metabolism		A	1
	8 pathways	**7 enzymes**		**6**

C: Catabolism (breakdown) of the substrate; A: Anabolism (synthesis) of the substrate; Number in the same colour come from the same sequences (CDS); (*) different enzymes to the TAV detected by using CC and CE.

These findings indicate the value of using the genome of the species (*C. arabica*) (despite its incomplete sequence) in detecting important SNPs associated with the two traits based on pathway-based analysis including some were not detected when using the genomes of the two progenitors as a reference.

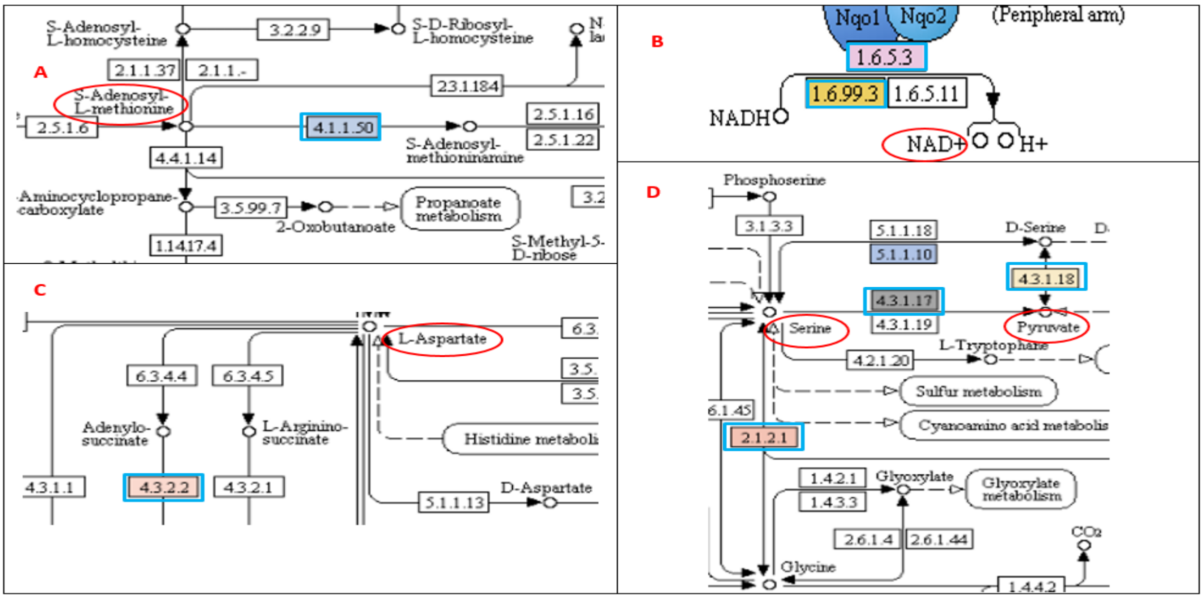


Figure 6.17 Snap shots of the KEGG pathways at the location where TAVs associated with enzymes involved in the metabolism of substrates that entered into the trigonelline biosynthesis and detected when using arabica as a reference genome.

(A) SAM - Cysteine and methionine metabolism - EC:4.1.1.50; (B) NAD - Oxidative phosphorylation EC:1.6.99.3 and EC:1.6.5.3; (C) L-aspartate - Alanine, aspartate and glutamate metabolism - EC:4.3.2.2; (D) Pyruvate and L-serine - Glycine, serine and threonine metabolism - EC:4.3.1.17, EC:4.3.1.18 - ammonia-lyase and EC:2.1.2.1 – hydroxymethyltransferase; substrates are circled in red; enzymes are highlighted in blue square.

6.3.5. Common SNPs associated with both caffeine and trigonelline contents.

There were five individuals shared between low caffeine bulk and low trigonelline bulk, and four individuals shared between high caffeine bulk and high trigonelline bulk. The number of individuals which low in this trait and high in the other was two. For TAVs, there were 75 common TAVs detected in both caffeine and trigonelline bulking analyses, and 162 common genes between two traits. Their biosynthesis also shares a fairly large number of common enzymes (20 out of 80) based on KEGG analysis (Figure 6.18). This is consistent with results from the study of bean variation of arabica in which caffeine and trigonelline contents were somewhat positively correlated (Chapter 3 and Tran et al., 2017). In fact, caffeine and trigonelline are both major nitrogenous alkaloids found in coffee seeds, and the pattern of trigonelline biosynthesis during fruit development is very similar to that of caffeine (Koshiro et al., 2006).

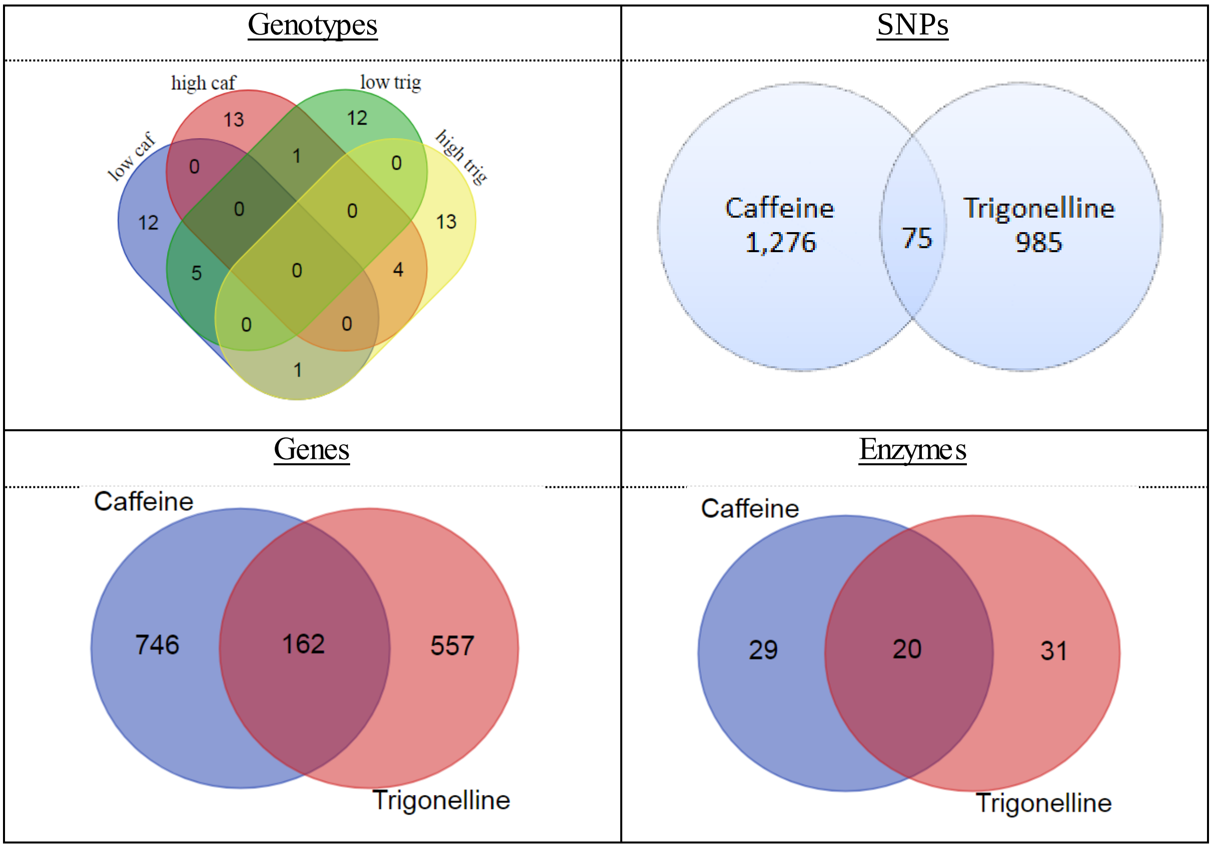


Figure 6.18 Common between caffeine and trigonelline in (input) genotypes, SNPs, genes and enzymes.

Among the common TAVs involved in the biosynthesis pathways of caffeine and trigonelline, the most significant TAV is the one links to the enzyme EC:2.5.1.16 – synthase which catalyses S-adenosyl-methioniamine to S-methyl 5'-thiodenosine (in cysteine and methionine metabolism) or S-adenosyl-methioniamine to spermine or spermidine (in arginine and proline metabolism) while S-adenosyl-methioniamine is a breakdown substrate of SAM which also participates in both caffeine and trigonelline biosynthesis pathways. The finding supports the study by Mizuno et al. (2014) who used the N-methyltransferase assay with S-adenosyl[methyl-14C]methionine to confirm these enzymes catalyzing the conversion of nicotinate to trigonelline and the expression profiles for two genes homologous to caffeine synthases were similar to the accumulation profile of trigonelline. The common TAV associating with the enzymes EC:2.5.1.16 – synthase was on chromosome 11 (at 30,965,526) subgenome CC with more C-allele in low caffeine and trigonelline while more A-allele in high caffeine and trigonelline resulting in a change of ratio of the amino acids serine and alanine in the two bulks of the two traits (Appendix table S6.7 and S6.8). Moreover this TAV linked to enzymes that break down S-adenosyl-methioniamine to spermine or spermidine – the two compounds having effects on salinity and drought tolerance. Targeting common TAVs associated with common pathways and enzymes could help improve both traits.

6.4. Conclusions and suggestions

SNP loci associated with pathways of caffeine and trigonelline biosynthesis were first discovered in *C. arabica* using extreme phenotyping GWAS. Previous studies had mainly used the mining of expressed sequence tags (ESTs) or transcriptome data for SNP discovery (Combes et al., 2013; Kochko et al., 2010; Vidal et al., 2010; Yuyama et al., 2016; Zhou et al., 2016). QTL mapping for quality traits has been conducted for *C. canephora* only (reviewed by Tran et al., 2016). Such studies involved limited genetic variation contributed from two parents. The current study was based on a diverse population of *C. arabica* for identification of trait-associated markers. The use of a wider genetic base is particularly valuable for improvement of cultivated arabica with a very narrow genetic base that has hampered the progress of genetic studies. The current study also makes use of wild species with a rich source of new alleles that might have been lost through domestication. This effort led to the discovery of a number of TAVs for caffeine and trigonelline with XP-GWAS. Targeting these TAVs, especially the common TAVs between the two traits in breeding will thus potentially help manipulate these compounds in domesticated arabica coffee utilising molecular markers.

Findings and certain limitations from this chapter open opportunities for further research. The detected TAVs and their flanking sequences could be used for PCR-based markers or incorporated in high-throughput genotyping platforms. This study used selective pooling sequencing. Genotyping of the entire population would help validate the TAVs and calculate linkage disequilibrium for association analysis at the haplotype level. Alternatively, sequencing of the germplasm collection at targeted regions harbouring putative TAVs will help define favorable haplotypes for caffeine and trigonelline.

CHAPTER 7: GENERAL DISCUSSIONS, CONCLUSSIONS AND FUTURE DICRECTIONS

7.1 General discussions

Despite its economic importance, coffee has received very limited research attention compared to other crops, especially in the field of genetics and genomics. Perception about the taxonomy and classification of coffee has been controversial: some argue that the *Coffeeae* tribe has two genera *Coffea* L. and *Psilanthus* Hook.f., while others believe that it is comprised of only one genus of *Coffea*. The presentation of genomic relationships among coffee species in this study supports the former argument. Results of whole chloroplast genome sequencing of 15 species collected from eleven countries showed a clear separation of coffee species into two clades. Four *Psilanthus* from Oceania grouped together suggesting that the origin of *Psilanthus* is in Asia which has not been proposed before. *C. canephora* was classified as a coffee species belonging to the *Coffea* genus. However, in the present study, *C. canephora* was grouped with two supposed *Psilanthus* species from West/Central Africa, indicating that the ancestor of *C. canephora* might have captured the chloroplast genome from a maternal *Psilanthus* donor through historical hybridization. As the evolutionary origin of *C. canephora* is not yet fully resolved, this study may guide further research on the origin of *C. canephora*. Although *C. canephora* is *C. arabica*'s progenitor, the maternal genomes of *C. arabica* and *C. canephora* are divergent, implying that the chloroplast genome of *C. arabica* was not inherited from *Psilanthus* but from *C. eugenioides*. The presence of a number of species with special features in relation to quality such as low or no caffeine, high trigonelline or low CGAs could be an important breeding focus. However, the success of interspecific hybridisation may vary due to genetic barriers between species, so genetic improvement focusing on using materials within the same species (Arabica) is therefore the priority.

Using next generation sequencing in association analysis to dissect complex traits has become popular, but the success can be influenced by the choice of germplasm. *C. arabica* is well known for its narrow genetic base, especially in the cultivated arabica varieties, making it difficult for genetic analysis and improvement. In this study, a survey on 232 individuals from a world collection demonstrated substantial variation observed for bean morphology, non-volatiles and volatiles that could be used as a valuable genetic resource for quality improvement in arabica coffee breeding. The correlation between bean morphology and non-volatile or volatile compounds was positive but not significant, implying that it may be possible to select for desirable combinations of traits in arabica coffee breeding. Most practically, the strong correlations existing within several volatile groups provide useful direction for targeted analyses applied to in the studied population by focusing on a smaller number of representative compounds so as to improve analytical accuracy and efficiency in coffee bean quality research and industry application. The wide variation in the levels of caffeine and trigonelline also led to identification of accessions with extreme phenotypic values for use in pooled sequencing to identify trait-associated markers to assist breeding.

Arabica coffee is among major commercial species that lack a reference genome sequence. Only recently, the first high-quality draft genome of *C. canephora* was completed and published. Obtaining a whole genome sequence for *C. arabica* has been a challenge but will assist in identifying new options for its genetic improvement of arabica coffee. Genomic analysis of *C. arabica* is complex for a number of reasons. Genome assembly for polyploid plants is problematic, and exacerbated in *C. arabica* by its relatively large genome size, repeat content, paralogy and heterozygosity. Further, the relatively low density of sequence data, particularly PacBio data, presented difficulty in sequence analysis. In this study, the K7 arabica cultivar was sequenced and assembled resulting in 76,409 scaffolds containing 99,829 gene models. Although the draft genome had gaps and is incomplete, it was helpful in detecting a number of TASs for caffeine and trigonelline which may eventually have an important impact on coffee genetics and breeding for these traits.

An association study carried out on pre-existing populations, in collections or in selection trials is a suitable approach for coffee because this does not require specific populations created by controlled crossing, which is very difficult for coffee – a perennial crop. Association mapping has two approaches, candidate-gene association mapping and genome-wide association mapping. In the present study, the latter was employed surveying genetic variation in the whole genome to find signals of association for traits linking to coffee quality (i.e. caffeine and trigonelline) using extreme phenotypes. SNP identification in polyploids is challenging due to the difficulty of distinguishing between homeologous SNPs (co-resident genomes) and allelic SNPs (SNPs segregating in a Mendelian fashion). To optimize the quality of the data used to generate the variant calls, a number of strategies was applied in the present study in order to effectively reducing false-positive SNP calls such as sequencing technology, read length, library preparation. In the present study, Illumina PE reads of 150 bp and PCR-free library was applied which helped reduce the possible false-positive SNP. In addition, quality was improved by retaining only the uniquely mapping reads, high minimum coverage, base quality of 20, maximum read depth filters and other more sophisticated filters (tail distance bias, map quality bias, base quality bias, strand bias). Using the arabica draft genome combined with progenitor genomes as references in mapping and variant calling allowed the identification of 79 caffeine-associated SNPs and 62 trigonelline-associated SNPs. These TAVs were proved to be associated with genes involved in the caffeine and trigonelline biosynthesis pathways based on KEGG pathway-based analysis. Interestingly, there was a common TAV associated with an enzyme that is critically involved in both caffeine and trigonelline biosynthesis. This enzyme catalyses the conversion of SAM to SAH for the formation of trigonelline from nicotinate and also serves as a methyl donor for the formation of caffeine. The enzyme also breaks down SAM to spermine or spermidine, both of which have an effect on salinity and drought tolerance. Targeting this common TAV could help improve both quality and stress tolerance.

7.2 Contributions to scientific knowledge

Previously, the relationship between species was determined based on morphology, chemotaxonomy, hybridisation and cytogenetic studies, and molecular markers. This is the first phylogeny of *Coffea* constructed from whole chloroplast genome sequence for coffee trees. The study of phylogeny of coffee species provides the overall picture of *Coffea* genomes, relationship among them and will help guide efficient interspecific hybridisation schemes in coffee.

The substantial variation in bean morphology, non-volatiles and volatiles of arabica population demonstrates the significant diversity in bean compounds within the *C. arabica* species. The outcomes also allowed the identification of accessions of extremes (in the levels of caffeine and trigonelline) for pooling sequencing to identify trait-associated DNA markers to assist arabica coffee breeding. Further, the phenotypic survey identified groups of accessions with distinct variation in quality-related compounds that can be used as founder parents to generate pre-breeding populations. The correlation among traits relating to bean morphology and compositions implies that the selection for desirable combinations of traits in arabica coffee breeding is possible.

Genome assembly and annotation of a "real" arabica variety contributes valuable genomic resources to genetic analysis and improvement of *C. arabica*. Although it is an initial draft genome, using it in the mapping and variant calling to identify polymorphic SNPs between two DNA pools was proved efficient.

To the author's knowledge, this is the first research where SNP markers affecting chemical compounds have been identified for arabica using extreme-phenotyping association study. Targeting these TAVs, especially the common TAVs between two traits in breeding will thus potentially help manipulate these compounds in domesticated arabica coffee utilising molecular markers.

7.3 Future directions

Based on the genomic relationship, at the genus level, interspecific hybridisation between *C. arabica* and other close species with desirable quality traits such as low caffeine (*C. pseudozanguebariae, C. tetragona*) or high sucrose (*C. pseudozanguebariae)* should be investigated for quality improvement. Likewise, the species *P. ebracteolatus* close to *C. canephora* (i.e. *P.*

ebracteolatus) but low in caffeine and CGAs could also be used in arabica breeding. This study has involved sequencing of many but not all coffee species. A globally coordinated effort to gather the sequence of additional coffee species from various research coffee groups would be needed to bring in a more concrete conclusion on the genomic synteny between coffee species. Additional long reads sequences would help to improve the genome assembly for K7 arabica. In addition, functional annotation would help provide more details of the genome assembled and will be more useful for downstream analysis. The functional annotation can be done using Blast2GO and the outcomes will reveal genes and pathways as well as the biological functions (biological process, molecular function, cellular component) of the genes where significant TAVs were found.

At the species level, breeding for high/low levels of caffeine and trigonelline could be achieved through marker-assisted selection subject to marker validation. Primers flanking each significant TAV have been designed (Table S7.1). They can be used to assay contrasting individuals (high vs. low in caffeine or trigonelline) for PCR amplicon sequencing. The analysis of sequences of the PCR products will help confirm the polymorphism detected by bulked sequencing and may also reveal which sub-genome it comes from. This information combined with additional phenotyping (at different growth stages, environments and genotypes) will enable an estimation of individual SNP additive effects as well as their possible interaction with caffeine and trigonelline contents and with the environment. Although each single SNP may confer only a small effect on the caffeine or trigonelline content, their joint actions are likely to have a significant role. Furthermore, the designed primers could also be assayed in additional genetic populations including other diversity panels and bi-parental mapping populations that vary in caffeine and trigonelline for validation and *de novo* QTL mapping.

APPENDICES

Table S4.1 List of germplasm used in the study

No	Acc No	Name	Country	Coffee type
1	972	Philippine	Guatemala	Variety
2	973	Chocolá	Guatemala	Variety
3	976	Mocha	Guatemala	Variety
4	977	Blue Mountain	Guatemala	Variety
5	978	Maragogipe	Guatemala	Variety
6	979	Kona	Guatemala	Variety
7	980	Sumatra	Guatemala	Variety
8	981	Preanger	Guatemala	Variety
9	984	San Ramón	Guatemala	Variety
10	986	Purpurascens	Guatemala	Natural mutant
11	990	Surinam	El Salvador	Variety
12	992	Padang	El Salvador	Variety
13	993	Nacional Salvadoreña	El Salvador	Variety
14	996	Typica	El Salvador	Variety
15	1993	Goiaba colección 11 (552)	Brazil	selection
16	1994	Erecta colección 11 (H-1048-9)	Brazil	Natural mutant
17	1998	Semperflorens	El Salvador	selection
18	2138	Arabigo Puerto Rico	Guatemala	Variety
19	2246	Jimna-1	Ethiopia	selection
20	2249	Dessie	Ethiopia	Variety
21	2250	Batie-1	Ethiopia	selection
22	2268	San Martin	Guatemala	Variety
23	2298	Coorg	Kenya	Variety
24	2299	Laurina	Costa Rica	Natural mutant
25	2308	Caturra Rojo, coleccion 818 PQ x 28603	Brazil	Hybrid
26	2397	Colunmaris	Puerto Rico	Natural mutant
27	2540	Bourbon Amarillo	Brazil	Variety
28	2544	Mundo novo	Brazil	Variety
29	2676	Laurina	Cameroon	Natural mutant
30	2691	BA-21 Arabica x Liberica	India	Hybrid
31	2707	F-840	Tanzania	Variety
32	2711	Seleccion 2 Ennarea	Ethiopia	selection
33	2722	Geisha VC-496	Tanzania	Variety
34	2724	Rune Sudan, Resistente CBD	Tanzania	Variety
35	2733	S.L 34	Kenya	selection
36	2739	Seleccion L.28	Kenya	selection
37	2743	Series L	Kenya	selection
38	2748	S-8 Tafari Kela	Ethiopia	selection
39	2750	S-1 Erythean Moca	Ethiopia	selection
119	4618	Coleccion E-116	Ethiopia	Wild
120	4619	Coleccion E-118	Ethiopia	Wild
121	4629	E-356	Ethiopia	Wild
122	4631	E-358	Ethiopia	Wild
123	4634	E-361	Ethiopia	Wild
124	4636	E-363	Ethiopia	Wild
125	4642	E-369	Ethiopia	Wild
126	4649	E-130	Ethiopia	Wild
127	4657	E-405	Ethiopia	Wild
128	4669	E-162	Ethiopia	Wild
129	4674	E-471	Ethiopia	Wild
130	4685	Limnu E-171	Ethiopia	Wild
131	4693	Limnu E-188	Ethiopia	Wild
132	4700	E-312	Ethiopia	Wild
133	4712	E-189	Ethiopia	Wild
134	4713	E-190	Ethiopia	Wild
135	4717	E-194	Ethiopia	Wild
136	4721	E-198	Ethiopia	Wild
137	4729	E-206	Ethiopia	Wild
138	4730	E-207	Ethiopia	Wild
139	4732	E-209	Ethiopia	Wild
140	4734	E-211	Ethiopia	Wild
141	4735	E-212	Ethiopia	Wild
142	4740	E-217	Ethiopia	Wild
143	4752	E-229	Ethiopia	Wild
144	4755	E-232	Ethiopia	Wild
145	4757	E-223-A	Ethiopia	Wild
146	4758	E-237	Ethiopia	Wild
147	4772	E-268	Ethiopia	Wild
148	4775	E-272	Ethiopia	Wild
149	4776	E-273	Ethiopia	Wild
150	4796	E-381	Ethiopia	Wild
151	4797	E-382	Ethiopia	Wild
152	4800	E-393	Ethiopia	Wild
153	4804	E-397	Ethiopia	Wild
154	4816	E-416	Ethiopia	Wild
155	4829	E-429	Ethiopia	Wild
156	4830	E-430	Ethiopia	Wild
157	4836	E-436	Ethiopia	Wild

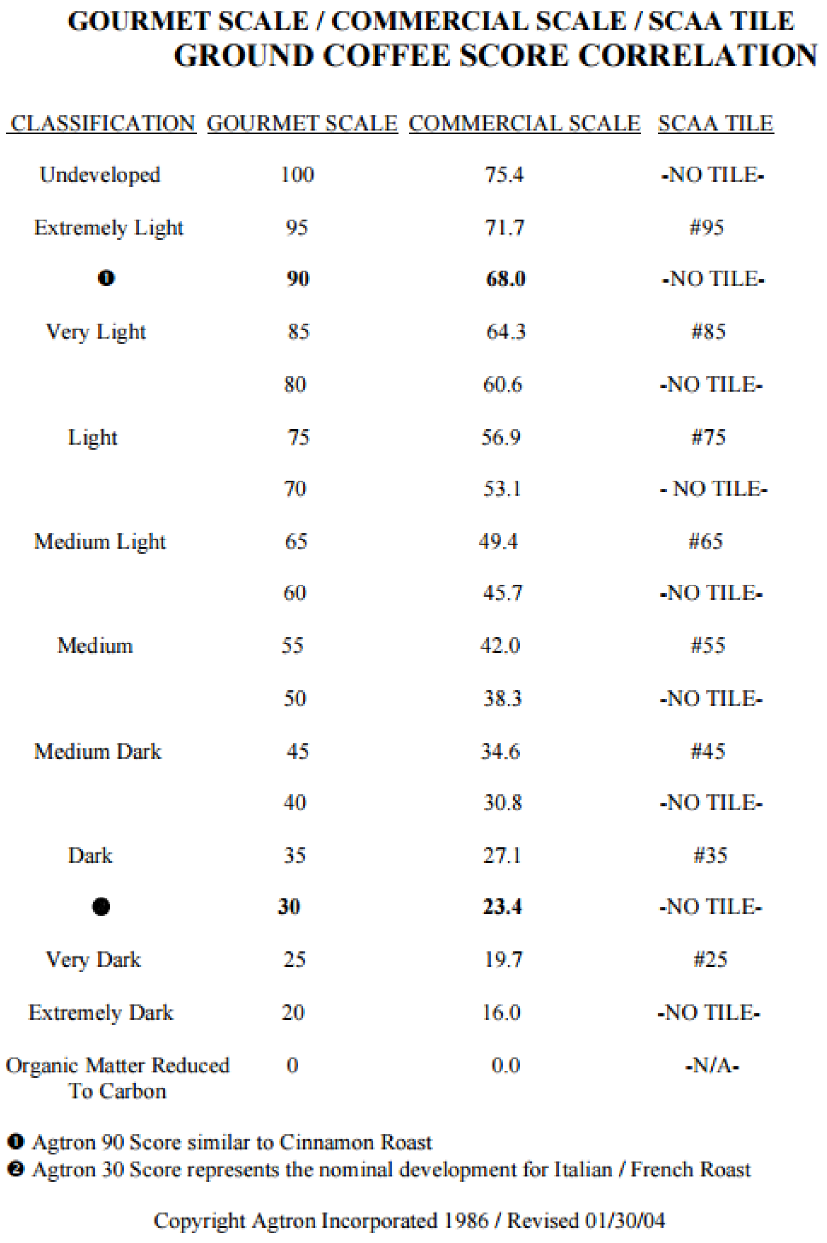
40	2758	Barbuk Sudan 2	Sudan	selection
41	2919	Jackson 2	Congo	selection
42	2920	Local Bronza 8	Congo	selection
43	3025	Villasarchi	Costa Rica	Variety
44	3081	Carrizal	Costa Rica	Variety
45	3129	Dos tiempos amarillo	Cuba	Variety
46	3165	Cuebaya	Ecuador	Variety
47	3382	LCP-376 Mundo Novo	Brazil	Variety
48	3411	Columnaris	Panama	Natural mutant
49	3432	Maragogipe rojo Brazil x 47127	Brazil	Hybrid
50	3433	Purpurascens Brazil x 47127	Brazil	Hybrid
51	3435	Xanthocarpa Brazil x 47127	Brazil	Hybrid
52	3473	Kents Ceylon x 48705	Ceylan	Hybrid
53	3483	Robusta	Sri Lanka	Robusta
54	3494	Lejeune 15 line 0144 Zona Strada Bandera ovest x 48849	Ethiopia	Hybrid
55	3498	Lejeune 26 line Zona Starda Bandera Ovest x 48849	Ethiopia	Hybrid
56	3519	No.2 Etiopia x 48989	Ethiopia	Hybrid
57	3520	No. 4 Etiopia x 48989	Ethiopia	Hybrid
58	3534	No. 19 Etiopia x 48989	Ethiopia	Hybrid
59	3545	No. 31 Etiopia x 48989	Ethiopia	Hybrid
60	3567	Kents India x 49406	India	Hybrid
61	3622	Mibirizi No.49, Belgian Congo x 51149	Belgian Congo	Hybrid
62	3628	Red tipped, Etiopia x 51356	Ethiopia	Hybrid
63	3635	Cumbaya	Ecuador	Variety
64	3674	Ta-Ku(Java), Sample a Taiwan x 53083	Taiwan	Hybrid
65	3675	Var-X, Sample B taiwan x 53083	Taiwan	Hybrid
66	3682	Seleccion Geisha (4-4)(205928),Castañer Puerto Rico x 53632	Puerto Rico	Hybrid
67	3689	Seleccion Castañar 22-1196, Puerto Rico x 53	Puerto Rico	Hybrid
68	3690	Cambodia x 53688	Cambodia	Robusta
69	3722	Geisha Castañer, Puerto Rico x 54375	Puerto Rico	Hybrid
70	3747	San Rafael	Costa Rica	Variety
71	3786	Bealanana Madasgascar x 59831	Madagascar	Hybrid
72	3820	Bourbon Amarillo, Puerto Rico x 60396	Puerto Rico	Hybrid
73	3845	H-66 self tree 875	Puerto Rico	selection
74	3852	KP-532 self tree 38	Puerto Rico	selection
75	3856	Caturra X Geisha R4-T4	Puerto Rico	Hybrid
76	3871	Zeghie Etiopia x 62275	Ethiopia	Hybrid
158	4837	E-437	Ethiopia	Wild
159	4856	E-456	Ethiopia	Wild
160	4857	E-457	Ethiopia	Wild
161	4863	E-463	Ethiopia	Wild
162	4865	E-465	Ethiopia	Wild
163	4873	E-506	Ethiopia	Wild
164	4874	E-507	Ethiopia	Wild
165	4875	E-508	Ethiopia	Wild
166	4877	E-510	Ethiopia	Wild
167	4878	E-511	Ethiopia	Wild
168	4881	E-514	Ethiopia	Wild
169	4883	E-516	Ethiopia	Wild
170	4887	E-520	Ethiopia	Wild
171	4892	E-524	Ethiopia	Wild
172	4897	E-529	Ethiopia	Wild
173	4900	E-531	Ethiopia	Wild
174	4901	E-532	Ethiopia	Wild
175	4902	E-533	Ethiopia	Wild
176	4903	E-534	Ethiopia	Wild
177	4909	E-540	Ethiopia	Wild
178	4918	E-549	Ethiopia	Wild
179	4927	E-559	Ethiopia	Wild
180	4933	E-565	Ethiopia	Wild
181	4946	E-1	Ethiopia	Wild
182	4950	E-12	Ethiopia	Wild
183	4951	E-16	Ethiopia	Wild
184	4952	E-20	Ethiopia	Wild
185	4957	E-324	Ethiopia	Wild
186	5136	Moka	Ethiopia	Natural mutant
187	5162	Icatu, Mundo Novo x (Robusta x Bourbon rojo) x Caturra	Brazil	Hybrid
188	5175	CIFC HW-26/13 19/1 Caturra x 832/1	Portugal	Hybrid
189	5208	Agaro S-4 SB-473	Colombia	selection
190	5233	CIFC H-151/1 S-288-23 33/1 x S-4 Agaro 110/5-1	Portugal	Hybrid
191	5267	Catuai	Costa Rica	Variety
192	5269	Catimor(B-60803 Hibrido HW-26/13)hw-26 19/1	Portugal	Hybrid
193	5324	Catuaí rojo	Costa Rica	Variety
194	5325	Catuaí amarillo	Costa Rica	Variety

77	3956	Babbaka	Ethiopia	Variety
78	4077	Variegata	Colombia	Natural mutant
79	4080	Maragogipe amarillo	Colombia	Variety
80	4089	Arabica	Colombia	selection
81	4133	Seleccion 353 4/5 CRRC 35/9	Portugal	selection
82	4196	Pluma Hidalgo	Mexico	Variety
83	4249	Anomala no.1	Colombia	selection
84	4250	Goiaba	Colombia	Variety
85	4276	Mibirizi 49-1848	Belgian Congo	selection
86	4288	Irgalen Kella Sidano	Belgian Congo	cultivar
87	4305	Geisha	Malawi	Variety
88	4319	KP-228	Malawi	selection
89	4338	M-7817	Ethiopia	selection
90	4353	Cubujuqui o San Joaquin	Costa Rica	cultivar
91	4376	Mosquey	Venezuela	cultivar
92	4387	Hibrido de timor CRRC 1343/80	Portugal	Hybrid
93	4472	E-7	Ethiopia	Wild
94	4475	E-19	Ethiopia	Wild
95	4479	E-301	Ethiopia	Wild
96	4480	E-302	Ethiopia	Wild
97	4483	E-305	Ethiopia	Wild
98	4494	E-46	Ethiopia	Wild
99	4497	E-67	Ethiopia	Wild
100	4501	E-37	Ethiopia	Wild
101	4502	E-38	Ethiopia	Wild
102	4504	E-53	Ethiopia	Wild
103	4507	E-56	Ethiopia	Wild
104	4514	E-72	Ethiopia	Wild
105	4517	E-68	Ethiopia	Wild
106	4521	E-318	Ethiopia	Wild
107	4533	E-495	Ethiopia	Wild
108	4555	E-482	Ethiopia	Wild
109	4558	E-486	Ethiopia	Wild
195	6366	C-68-1.7-33 Selección 31	Costa Rica	selection
196	8655	Catimor UFV 2326 19/1 caturra x 832/1 Hibrido de B-61813 Timor	Brazil	Hybrid
197	10589	Villalobos	Costa Rica	Variety
198	12523	Clon 7357 no.1	The US	selection
199	12841	Caturra rojo x Hibrido de Timor (INMC 32-3) Catimor Cenife	Mexico	Hybrid
200	15185	Cavimor F3 UFV 3552 F2 1091-732 T1 UFV F1 357-28 CIFC H-528	Brazil	Hybrid
201	16689	ET01	Ethiopia	wild
202	16690	ET02	Ethiopia	wild
203	16694	IRCC 206 ET 8	Indonesia	wild
204	16714	ET27	Ethiopia	wild
205	16725	ET41	Ethiopia	wild
206	16729	ET47	Ethiopia	wild
207	16768	Catuaí amarillo UFV 1116/354 F.E. PN CH 2077	Brazil	selection
208	16769	Catuaí amarillo ufv 1117/244 f.e pn ch 2077	Brazil	selection
209	16782	Catuaí rojo UFV 2197/194 CH 2077/2 5-72	Brazil	selection
210	16784	Sarchimor F3 IAC 1669/31-1 CIFC H-361/971-10 Villa sarchi X 832/2	Brazil	Hybrid
211	17205	L 268 et 32 bc5	Cameroon	Wild
212	17208	L 279 et 35 bc2	Cameroon	Wild
213	17223	L-320 ET 38 C10	Cameroon	Wild
214	17231	L 334 et 45 c7	Cameroon	Wild
215	17232	L-334 ET45 C7	Cameroon	Wild
216	17233	L-336 ET46 C3	Cameroon	Wild
217	17234	Q 24787 Caturra amarillo UFV 2199 LCH 2077-2	Brazil	Hybrid
218	17598	Cati F4 UFV 4586 Caturra amarillo X H.Timor CENECAFE	Brazil	Hybrid
219	17601	CH 7313-12 SH2	Brazil	Selection
220	17931	MR-422 Línea de Catimor grano rojo	Colombia	Selection
221	19858	T.H.-40, Hibrido F2, Jimma 4 (T.2253) X Bourbon rojo (T.995)	Costa Rica	Hybrid
222	19956	T.H.-309 (Laurina T.4059 x Maragogipe T.2312) x E-550 (T.4919)	Costa Rica	Hybrid
223	21230	Java	Indonesia	Variety
224	21263	ET-07	Indonesia	Wild
225	21264	ET-08	Indonesia	Wild
226	21282	ET-23	Indonesia	Wild
227	21283	ET-23	Indonesia	Wild

110	4568	E-134	Ethiopia	Wild
111	4569	E-146	Ethiopia	Wild
112	4574	Ennarea E-151	Ethiopia	Wild
113	4577	Ennarea E-154	Ethiopia	Wild
114	4578	Ennarea E-155	Ethiopia	Wild
115	4580	Ennarea E-157	Ethiopia	Wild
116	4608	Ennarea E-350	Ethiopia	Wild
117	4614	Coleccion E-498	Ethiopia	Wild
118	4616	Coleccion E-500	Ethiopia	Wild
228	21290	ET-29	Indonesia	Wild
229	21297	ET-34B	Indonesia	Wild
230	21315	ET-59	Indonesia	Wild
231	21317	Catuaí	Costa Rica	Variety
232	21363	Sarchimor-T.5296	Costa Rica	Variety
233	21364	Catimor-T.05175	Costa Rica	Variety
234	21372	C. brevipes	Indonesia	C. brevipes
235	21398	Hibrido 1, 3, 5 Poblacion segregante	Costa Rica	Hybrid

(Highlighted in red: Different species to *C. arabica*; 221 white background: samples used in non-targeted volatile analysis; 14 grey background: additional samples used in non-volatile analysis)

Table S4.2 Degree of roast



GOURMET SCALE / COMMERCIAL SCALE / SCAA TILE
GROUND COFFEE SCORE CORRELATION

CLASSIFICATION	GOURMET SCALE	COMMERCIAL SCALE	SCAA TILE
Undeveloped	100	75.4	-NO TILE-
Extremely Light	95	71.7	#95
❶	**90**	**68.0**	-NO TILE-
Very Light	85	64.3	#85
	80	60.6	-NO TILE-
Light	75	56.9	#75
	70	53.1	- NO TILE-
Medium Light	65	49.4	#65
	60	45.7	-NO TILE-
Medium	55	42.0	#55
	50	38.3	-NO TILE-
Medium Dark	45	34.6	#45
	40	30.8	-NO TILE-
Dark	35	27.1	#35
●	**30**	**23.4**	-NO TILE-
Very Dark	25	19.7	#25
Extremely Dark	20	16.0	-NO TILE-
Organic Matter Reduced To Carbon	0	0.0	-N/A-

❶ Agtron 90 Score similar to Cinnamon Roast
❷ Agtron 30 Score represents the nominal development for Italian / French Roast

Copyright Agtron Incorporated 1986 / Revised 01/30/04

Table S4.3 List of isotope labelled internal standards and volatiles identified

No	Isotope labelled internal standards	Class	Compounds quantified	Spike (µl)	Conc in 2g coffee (µg)
1	d6-2,3-butanedione	Ketones	2,3-butanedione	100	500
2	d2-3-methylbutanal-2,2		3-methylbutanal	100	500
3	d5-2,3-pentanedione-1,1,1,4,4	Aldehydes	methylpropanal	100	500
			3-methylbutanal		
			2-methylbutanal		
			(E)-2-nonenal		
			2,3-pentanedione		
4	d9-ethyl-2-methylbutyrate	Esters	ethyl-2-methylbutyrate	100	50
5	d6-limonene	Terpenes	D-limonene	100	53
6	d5-4-ethyl-2methoxyphenol	Phenolic compounds	guaiacol	100	50
			4-ethylguaiacol		
			4-vinylguaiacol		
7	d3-linalool (d3-vinyl)	Terpenes	linalool	100	50
			geraniol		
		Norisoprenoids	beta damascenone		
8	d3-2-isobutyl-3-methoxypyrazine	Pyrazines	2,5-dimethylpyrazine	100	50
			2,3-dimethylpyrazine		
			2-ethyl-3,6-dimethylpyrazine		
			2-ethyl-3,5-dimethylpyrazine		
			2,3-diethyl-5-methylpyrazine		
			3-isobutyl-2-methoxypyrazine		

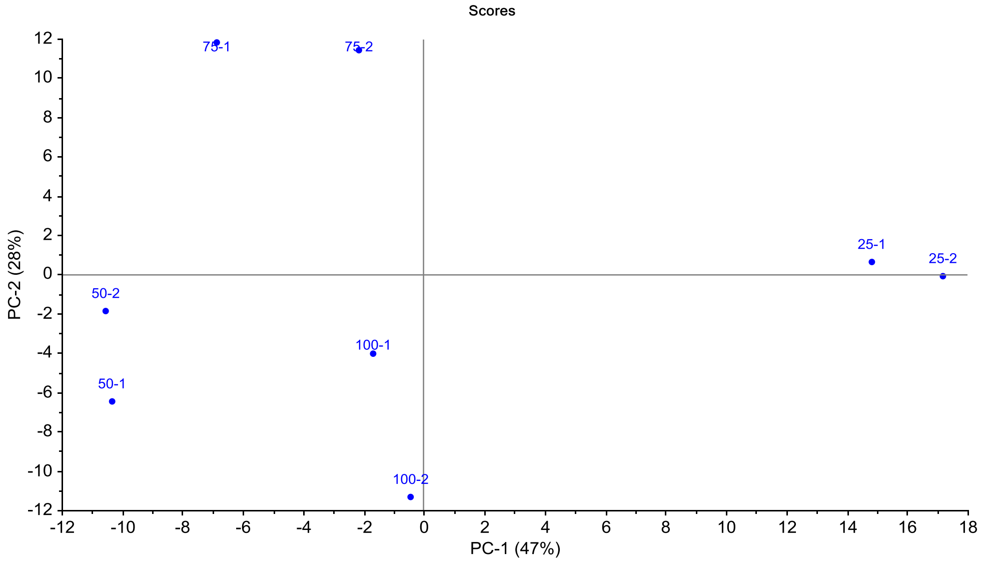


Figure S4.1 PCA of 4 batch sizes (25, 50, 75 and 100 g) in duplicates used in roast based on TIC data.

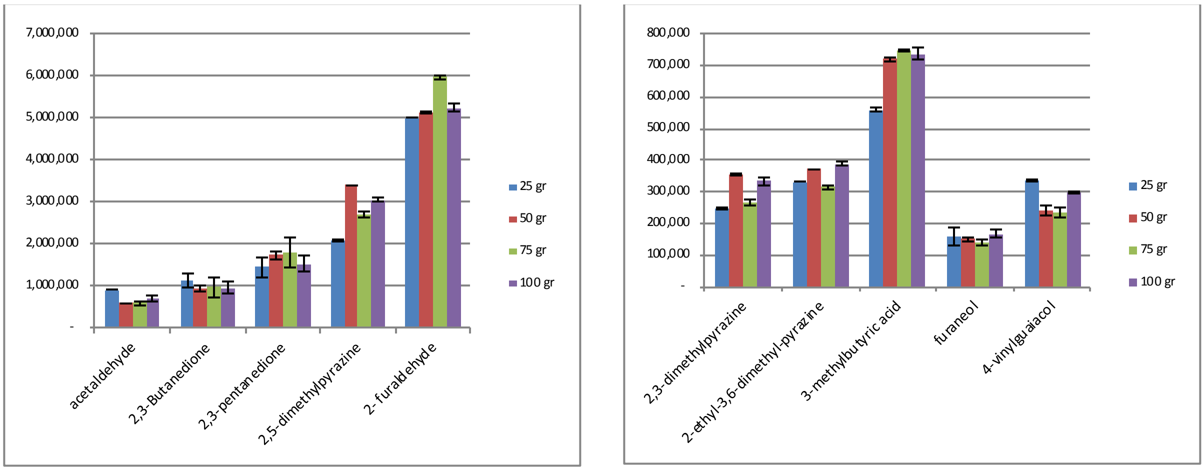


Figure S4.2 Peak area of 10 selective compounds of 4 batch sizes used in roast

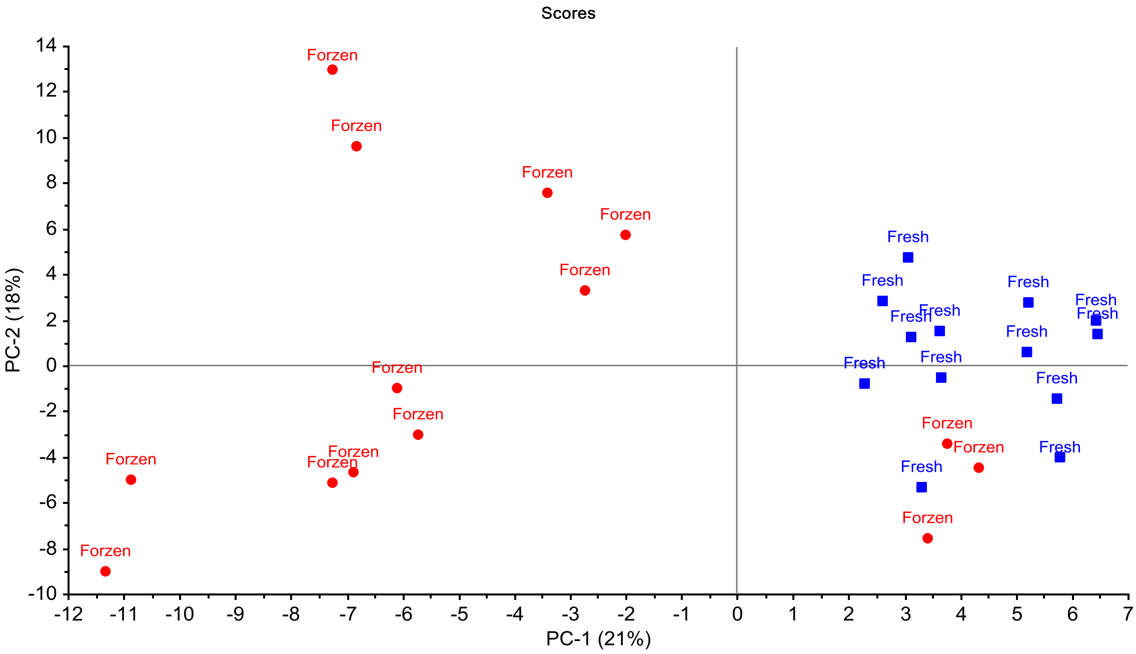


Figure S4.3 PCA of fresh (in blue) and stored (in red) beans based on TIC data.

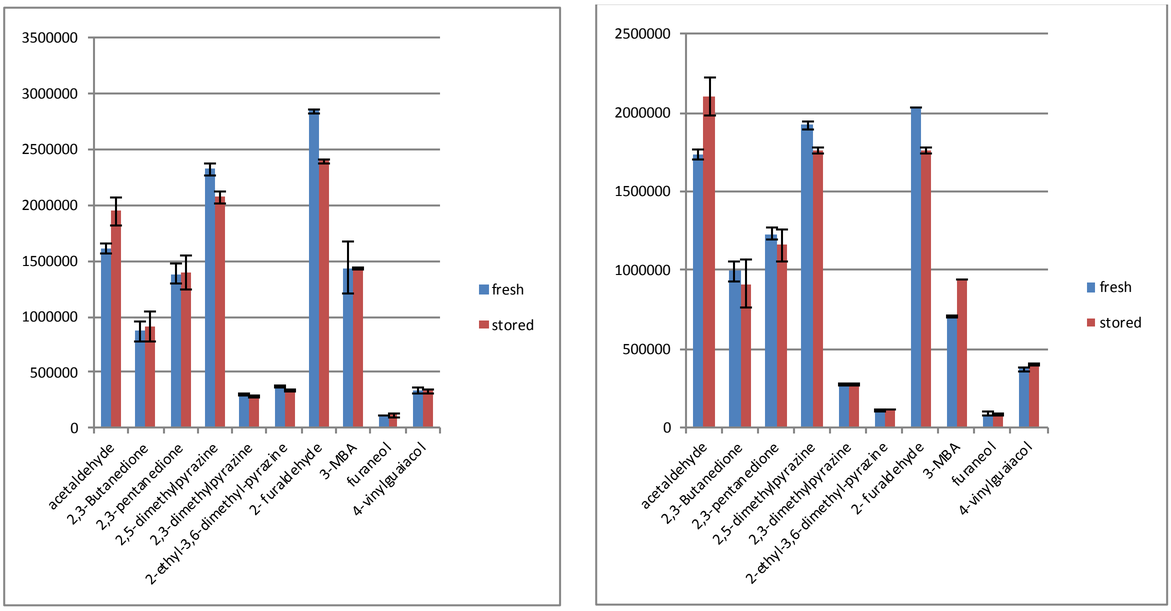


Figure S4.4 Changes of the volatiles between fresh and stored samples (left – accession no. 4700; right – accession no. 4631).

Table S4.4 Samples selected for quantification

No	Code	Name	Country	Type
Low in concentration of selective compounds				
1	1993	Goiaba colección 11 (552)	Brazil	selection
2	2250	Batie-1	Ethiopia	selection
3	2268	San Martin	Guatemala	Variety
4	2733	S.L 34	Kenia	selection
5	3519	No.2 Etiopia x 48989	Ethiopia	Hybrid
6	3567	Kents India x 49406	India	Hybrid
7	8655	Catimor UFV 2326 19/1 Caturra x 832/1 Hibrido de B-61813 Timor	Brazil	Hybrid
8	16784	Sarchimor F3 IAC 1669/31-1 CIFC H-361/971-10 Villasarchi x 832/2	Brazil	Hybrid
9	17205	L 268 et 32 bc5	Cameroon	Wild
10	17208	L 279 et 35 bc2	Cameroon	Wild
11	17223	L-320 ET 38 C10	Cameroon	Wild
12	17231	L 334 et 45 c7	Cameroon	Wild
13	21264	ET-08	Indonesia	Wild
14	4857	E-457	Ethiopia	Wild
15	4909	E-540	Ethiopia	Wild
High in concentration of selective compounds				
1	3483	Robusta	Sri Lanka	Robusta
2	3747	San Rafael	Costa Rica	Variety
3	4196	Pluma Hidalgo	Mexico	Variety
4	4288	Irgalen Kella Sidano	Belgian Congo	cultivar
5	4376	Mosquey	Venezuela	cultivar
6	3820	Bourbon Amarillo, Puerto Rico x 60396	Puerto Rico	Hybrid
7	4387	Hibrido de timor CRRC 1343/80	Portugal	Hybrid

8	4494	E-46	Ethiopia	Wild
9	4555	E-482	Ethiopia	Wild
10	4816	Coleccion E-500	Ethiopia	Wild
11	4634	E-361	Ethiopia	Wild
12	4693	Limnu E-188	Ethiopia	Wild
13	4918	E-549	Ethiopia	Wild
14	4629	E-356	Ethiopia	Wild
15	4631	E-358	Ethiopia	Wild
Outlying wild type genotypes				
1	4636	E-363	Ethiopia	Wild
2	4558	E-486	Ethiopia	Wild
3	4568	E-134	Ethiopia	Wild
4	4574	Ennarea E-151	Ethiopia	Wild
5	4685	Limnu E-171	Ethiopia	Wild
Total: 35 accessions				

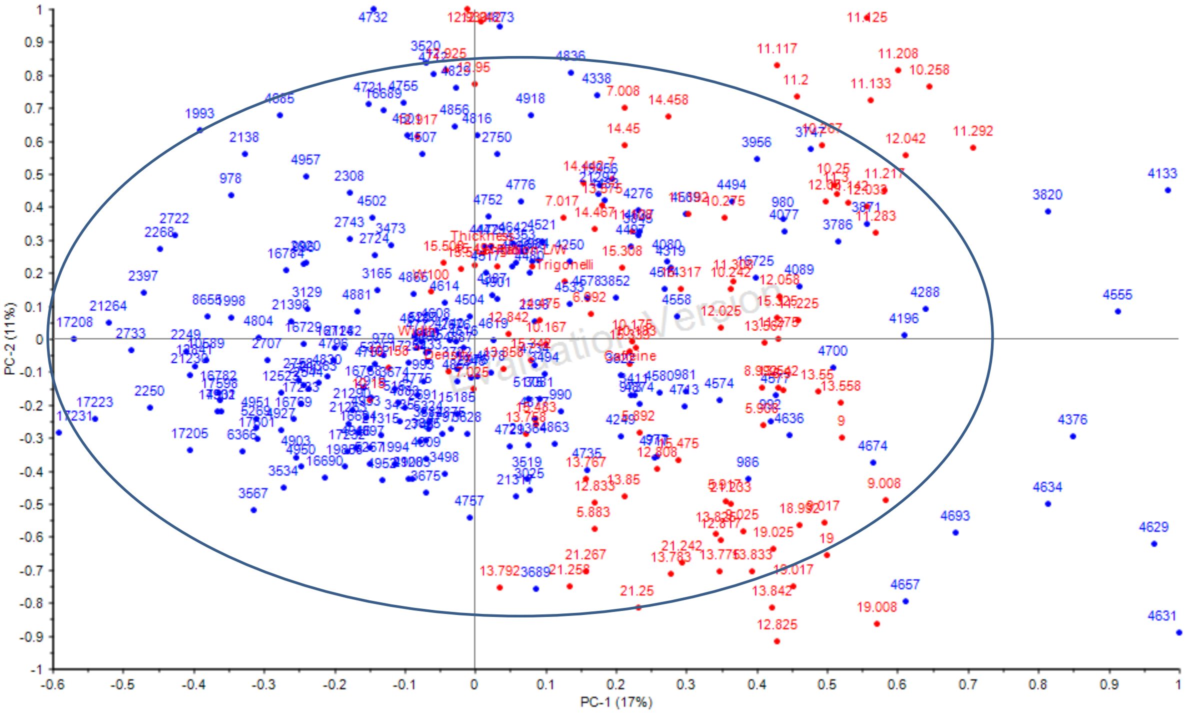


Figure S4.5 PCA of the coffee germplasm collection based on morphology, non-volatile and 100 selected chromatographic data

(Blue: accessions, Red: the variables (Rt), circle: outliers – out of the circle).

Table S5.1 Assembly improvement using GapCloser and Scaffolders with Illumina sequencing reads

Software	SOAP*denovo*	GAPCloser (1st)	SSPACE (standard)	GAPCloser (2nd)
Contig metrics				
# contigs	602,834	276,322	411,313	320,474
Ave. contig size	1,111	2,944	2,236	3,135

Contig N50	2,063	6,284	6,571	8,522
Scaffold metrics				
# scaffolds	120,114	120,578	110,179	110,443
Ave. scaffold size incl. gaps	8,118	8,155	12,281	12,277
Ave. scaffold size w/o gaps	5,575	6,747	8,346	9,098
Scaffold N50	16,569	16,694	27,147	27,310
Max scaffold size	175,948	176,350	389,435	389,501
Min scaffold size	1,000	1,000	1,000	1,000
Total genome length incl. gaps	975,074,759	983,303,576	1,353,166,521	1,355,930,976
Total genome length w/o gaps	669,693,502	813,528,413	919,608,986	1,004,783,410
Gap metrics				
Captured gaps	482,683	155,718	359,134	210,009
Max gap	27,925	27,658	27,658	27,658
Mean gap	633	1,090	1,207	1,672
Gap N50	3,731	4,751	4,957	5,270
Total gap length	305,378,544	169,774,888	433,614,750	351,147,504

Table S5.2 Assembly improvement using GapClosers and Scaffolders with PacBio longreads

Assembler	**GAPCloser (Illumina reads)**	**SSPACE-Longsreads (PacBio reads)**	**GAPCloser (Illumina reads)**	**PBJelly2 (PacBio reads)**
Contig metrics				
# contigs	320,474	320,178	283,660	265,687
Max contig	183,259	183,259	186,627	186,701
Mean Contig	3,135	3,148	3,778	4,393
Contig N50	8,522	8,592	10,133	12,184
Contig N90	1,392	1,400	1,732	2,110
Total contig length	1,004,783,410	1,007,838,579	1,071,793,316	1,167,243,096
Assembly GC (%)	37.00	36.63	36.71	36.76
Scaffold metrics				
# scaffolds	110,443	76,673	76,673	76,409
Max scaffold	389,501	755,558	757,040	769,411
Mean scaffold	12,277	18,292	18,314	18,954
Scaffold N50	27,310	52,292	52,431	54,544
Scaffold N90	5,476	8,105	8,105	8,145
Total scaffold length	1,355,930,976	1,402,516,201	1,404,209,962	1,448,282,977
Gap metrics				
Captured gaps	210,031	243,505	206,987	189,278
Max gap	27,658	27,658	27,141	27,141
Mean gap	1,672	1,621	1,606	1,485
Gap N50	5,270	4,844	5,050	5,282
Total gap length	351,147,566	394,677,622	332,416,646	281,039,881

Table S6.1 Accessions of low- and high-caffeine used in DNA bulking for XP-GWAS

No	Code	Accession name	Origin	Type	Content (%)
18 accessions lowest in caffeine					
1	2299	Laurina 1	Costa Rica	Natural mutant	0.82
2	4755	E-232	Ethiopia	Wild	0.93
3	4713	E-190	Ethiopia	Wild	0.98
4	4276	Mibirizi 49-1848	Congo Belga	Natural mutant	1.00
5	2676	Laurina 2	Cameroon	Natural mutant	1.00
6	3722	Geisha Castañer, Puerto Rico x 54375	Puerto Rico	Hybrid	1.01
7	2733	S.L 34	Kenia	Selection	1.02
8	2919	Jackson 2	Congo	Selection	1.03
9	978	Maragogipe	Guatemala	Variety	1.04
10	2743	Series L	Kenia	Selection	1.05
11	3871	Zeghie Etiopia x 62275	Ethiopia	Hybrid	1.05
12	3845	H-66 self tree 875	Puerto Rico	selection	1.06
13	2722	Geisha VC-496	Tanzania	Variety	1.08
14	2249	Dessie	Ethiopia	Variety	1.09
15	4892	E-524	Ethiopia	Wild	1.09
16	4909	E-540	Ethiopia	Wild	1.10
17	4918	E-549	Ethiopia	Wild	1.10
18	6366	C-68-1.7-33 Selección 31	Costa Rica	Selection	1.10
		Average			**1.03**
18 accessions highest in caffeine					
1	4629	E-356	Ethiopia	Wild	1.38
2	4631	E-358	Ethiopia	Wild	1.41
3	4712	E-189	Ethiopia	Wild	1.42
4	3494	Lejeune 15 line 0144 Zona Strada Bandera ovest x 48849	Ethiopia	Hybrid	1.42
5	5162	Icatu, Mundo Novo x (Robusta x Bourbon rojo) x Caturra	Brazil	Hybrid	1.42
6	4608	Ennarea E-350	Ethiopia	Wild	1.43
7	4533	E-495	Ethiopia	Wild	1.44
8	4877	E-510	Ethiopia	Wild	1.44
9	4494	E-46	Ethiopia	Wild	1.46
10	4555	E-482	Ethiopia	Wild	1.47
11	4133	Seleccion 353 4/5 CRRC 35/9	Portugal	Selection	1.47
12	21315	ET-59	France	Wild	1.47
13	4569	E-146	Ethiopia	Wild	1.49
14	4837	E-437	Ethiopia	Wild	1.51
15	4875	E-508	Ethiopia	Wild	1.51
16	3956	Babbaka	Ethiopia	Variety	1.52
17	3545	No. 31 Etiopia x 48989	Ethiopia	Hybrid	1.55
18	3622	Mibirizi No.49, Belgian Congo x 51149	Congo Belga	Hybrid	1.76
		Average			**1.48**

Table S6.2 Accessions of low- and high- trigonelline used in DNA bulking for XP-GWAS

No	Code	Accession name	Origin	Type	Content (%)
		18 accessions lowest in trigonelline			
1	16729	ET47	Ethiopia	Wild	0.80
2	4755	E-232	Ethiopia	Wild	0.85
3	4685	Limnu E-171	Ethiopia	Wild	0.85
4	4276	Mibirizi 49-1848	Congo Belga	Natural mutant	0.89
5	4483	E-305	Ethiopia	Wild	0.93
6	17232	L-334 ET45 C7	Camerun	Wild	0.94
7	16690	ET02	Ethiopia	Wild	0.94
8	16782	Catuaí rojo UFV 2197/194 CH 2077/2 5-72	Brasil	Selection	0.94
9	16769	Catuaí amarillo ufv 1117/244 fepnch 2077	Brasil	Selection	0.95
10	8655	Catimor UFV 2326 19/1 caturra x 832/1 Hibrido de B-61813 Timor	Brasil	Hybrid	0.96
11	3622	Mibirizi No.49, Belgian Congo x 51149	Congo Belga	Hybrid	0.96
12	5324	Catuaí rojo	Costa Rica	Selection	0.96
13	4897	E-529	Ethiopia	Wild	0.96
14	2743	Series L	Kenia	Selection	0.96
15	6366	C-68-1.7-33 Selección 31	Costa Rica	Selection	0.97
16	4319	KP-228	Malawi	Selection	0.97
17	4479	E-301	Ethiopia	Wild	0.97
18	3845	H-66 self tree 875	Puerto Rico	Selection	0.97
		Average			**0.93**
		18 accessions highest in trigonelline			
1	3682	Seleccion Geisha (4-4)(205928),Castañer Puerto Rico x 53632	Puerto Rico	Hybrid	1.27
2	3433	Purpurascens Brazil x 47127	Brasil	Hybrid	1.27
3	3520	No. 4 Etiopia x 48989	Ethiopia	Hybrid	1.28
4	2722	Geisha VC-496	Tanzania	Variety	1.28
5	2920	Local Bronza 8	Congo	Selection	1.28
6	4338	M-7817	Ethiopia	Selection	1.29
7	990	Surinam	El Salvador	Variety	1.33
8	4494	E-46	Ethiopia	Wild	1.29
9	3473	Kents Ceylon x 48705	Ceylan	Hybrid	1.29
10	4878	E-511	Ethiopia	Wild	1.29
11	4353	Cubujuqui o San Joaquin	Costa Rica	Cultivar	1.30
12	4497	E-67	Ethiopia	Wild	1.30
13	4569	E-146	Ethiopia	Wild	1.31
14	4712	E-189	Ethiopia	Wild	1.31
15	986	Purpurascens	Guatemala	Variety	1.32
16	4133	Seleccion 353 4/5 CRRC 35/9	Portugal	Selection	1.35
17	3411	Columnaris	Panama	Natural mutant	1.38
18	4080	Maragogipe amarillo	Colombia	Variety	1.38
		Average			**1.31**

Table S6.3 Number of mapped reads and coverage when aligned to the *C. canephora* genome reference using different settings

Bulk	**LF 0.5 and SF 0.8 (Default)**		**LS 1.0 and SF 0.8**		**LF 1.0 and SF 0.9**		**LF 1.0 and SF 0.95**	
	Mapped reads (%)	Ave. coverage	Mapped reads (%)	Ave. coverage	Mapped reads (%)	Ave. coverage	Mapped reads (%)	Ave. coverage
1	95.53	50	88.57	46	78.89	44	62.69	35
2	95.17	71	88.27	64	78.68	60	62.68	49
3	95.07	67	88.21	61	78.72	57	62.78	46
4	95.24	42	88.30	58	78.66	36	62.54	29
Ave.	**95.25**	**58**	**88.34**	**57**	**78.74**	**49**	**62.67**	**40**

Table S6.4 Number of mapped reads and coverage when aligned to different reference genomes using the same mapping setting (LF 1.0 and SF 0.8)

Bulk	***C. canephora* (1)**		***C. eugenioides* (2)**		***Combined (1) and (2)***		***C. arabica***	
	Mapped reads (%)	Ave. coverage	Mapped reads (%)	Ave. coverage	Mapped reads (%)	Ave. coverage	Mapped reads (%)	Ave. coverage
1	88.57	46	87.02	43	94.01	23	93.31	29
2	88.27	64	86.76	61	93.65	32	92.92	40
3	88.21	61	86.77	60	93.61	31	92.98	38
4	88.30	58	86.84	36	93.71	19	93.04	24
Ave.	**88.34**	**57**	**86.85**	**50**	**93.75**	**26**	**93.06**	**33**

Table S6.5 Enzymes and pathways recorded from Blast2GO analysis of 1,351 TAVs for caffeine using CC and CE genomes as reference

No	**Enzymes**	**Pathways**
1	EC:1.1.1.205 - dehydrogenase	Drug metabolism - other enzymes, Purine metabolism
2	EC:6.3.5.5 - synthase (glutamine-hydrolysing)	Pyrimidine metabolism, Alanine, aspartate and glutamate metabolism
3	EC:1.11.1.7 - lactoperoxidase	Phenylpropanoid biosynthesis
4	EC:3.6.1.3 - adenylpyrophosphatase	Purine metabolism
5	EC:2.1.1.53 - N-methyltransferase	Tropane, piperidine and pyridine alkaloid biosynthesis
6	EC:2.7.7.9 – uridylyltransferase	Amino sugar and nucleotide sugar metabolism, Pentose and glucuronate interconversions, Galactose metabolism, Starch and sucrose metabolism, Biosynthesis of antibiotics
7	EC:1.17.1.4 - dehydrogenase	Purine metabolism

8	EC:2.4.2.18 - phosphoribosyltransferase	Phenylalanine, tyrosine and tryptophan biosynthesis, Biosynthesis of antibiotics
9	EC:6.3.5.4 - synthase (glutamine-hydrolysing)	Alanine, aspartate and glutamate metabolism
10	EC:3.6.3.6 - ATPase	Oxidative phosphorylation
11	EC:3.1.3.16 - phosphatase	T cell receptor signaling pathway, Th1 and Th2 cell differentiation
12	EC:2.7.1.67 - 4-kinase	Inositol phosphate metabolism, Phosphatidylinositol signaling system
13	EC:3.2.1.26 - invertase	Galactose metabolism, Starch and sucrose metabolism
14	EC:1.10.3.1 - oxidase	Isoquinoline alkaloid biosynthesis, Tyrosine metabolism
15	EC:1.8.1.4 – dehydrogenase	Propanoate metabolism, Valine, leucine and isoleucine degradation, Glycine, serine and threonine metabolism, Glycolysis / Gluconeogenesis, Pyruvate metabolism, Citrate cycle (TCA cycle), Biosynthesis of antibiotics
16	EC:2.7.7.64 – uridylyltransferase	Amino sugar and nucleotide sugar metabolism, Ascorbate and aldarate metabolism, Pentose and glucuronate interconversions, Galactose metabolism, Biosynthesis of antibiotics
17	EC:1.14.11.15 - 3beta-dioxygenase	Diterpenoid biosynthesis
18	EC:4.3.3.2 - synthase	Indole alkaloid biosynthesis
19	EC:1.3.99.6 - 4-dehydrogenase	Steroid hormone biosynthesis
20	EC:1.14.13.8 - monooxygenase	Drug metabolism - cytochrome P450
21	EC:2.4.1.12 - synthase (UDP-forming)	Starch and sucrose metabolism
22	EC:1.11.1.6 - equilase	Tryptophan metabolism, Glyoxylate and dicarboxylate metabolism, Biosynthesis of antibiotics
23	EC:2.1.3.3 - carbamoyltransferase	Biosynthesis of antibiotics, Arginine biosynthesis
24	EC:4.1.1.28 – decarboxylase	Isoquinoline alkaloid biosynthesis, Tyrosine metabolism, Tryptophan metabolism, Histidine metabolism, Phenylalanine metabolism, Betalain biosynthesis, Indole alkaloid biosynthesis
25	EC:5.4.2.8 - mannose phosphomutase	Amino sugar and nucleotide sugar metabolism, Biosynthesis of antibiotics, Fructose and mannose metabolism
26	EC:6.3.1.2 - synthetase	Nitrogen metabolism, Glyoxylate and dicarboxylate metabolism, Arginine biosynthesis, Alanine, aspartate and glutamate metabolism
27	EC:1.3.3.6 - oxidase	Biosynthesis of unsaturated fatty acids, Fatty acid degradation, alpha-Linolenic acid metabolism
28	EC:3.2.1.21 - gentiobiase	Cyanoamino acid metabolism, Starch and sucrose metabolism, Phenylpropanoid biosynthesis
29	EC:2.5.1.1 - geranyl-diphosphate synthase	Biosynthesis of antibiotics, Terpenoid backbone biosynthesis
30	EC:4.3.2.2 - lyase	Biosynthesis of antibiotics, Alanine, aspartate and glutamate metabolism, Purine metabolism
31	EC:2.7.7.6 - RNA polymerase	Pyrimidine metabolism, Purine metabolism
32	EC:2.7.7.7 - DNA polymerase	Pyrimidine metabolism, Purine metabolism
33	EC:2.1.1.37 - (cytosine-5-)-methyltransferase	Cysteine and methionine metabolism
34	EC:2.8.1.8 - synthase	Lipoic acid metabolism
35	EC:2.6.1.83 - aminotransferase	Lysine biosynthesis, Biosynthesis of antibiotics
36	EC:3.6.1.15 - phosphatase	Thiamine metabolism, Purine metabolism
37	EC:3.1.3.41 - nitrophenyl phosphatase	Aminobenzoate degradation
38	EC:2.7.10.2 - protein-tyrosine kinase	T cell receptor signaling pathway
39	EC:3.2.1.15 - pectin depolymerase	Pentose and glucuronate interconversions
40	EC:3.2.1.48 - alpha-glucosidase	Starch and sucrose metabolism
41	EC:2.7.1.158 - 2-kinase	Inositol phosphate metabolism, Phosphatidylinositol signaling system
42	EC:2.7.4.14 - kinase	Pyrimidine metabolism
43	EC:3.5.4.5 - deaminase	Drug metabolism - other enzymes, Pyrimidine metabolism
44	EC:6.3.2.17 - synthase	Folate biosynthesis
45	EC:1.14.13.78 - oxidase	Diterpenoid biosynthesis
46	EC:3.2.1.23 - lactase (ambiguous)	Other glycan degradation, Galactose metabolism, Sphingolipid metabolism, Glycosaminoglycan degradation, Glycosphingolipid biosynthesis - ganglio series
47	EC:3.2.1.14 - ChiC	Amino sugar and nucleotide sugar metabolism
48	EC:2.4.1.11 - synthase	Starch and sucrose metabolism
49	EC:2.5.1.16 - synthase	Cysteine and methionine metabolism, Arginine and proline metabolism, beta-Alanine metabolism, Glutathione metabolism

Table S6.6 Enzymes and pathways recorded from Blast2GO analysis of 1,060 TAVs for trigonelline using CC and CE genomes as reference

No	Enzymes	Pathways
1	EC:6.1.1.17 - ligase	Porphyrin and chlorophyll metabolism, Aminoacyl-tRNA biosynthesis
2	EC:1.14.13.73 - 16-hydroxylase	Indole alkaloid biosynthesis
3	EC:1.1.1.31 - dehydrogenase	Valine, leucine and isoleucine degradation
4	EC:4.99.1.1 - ferrochelatase	Porphyrin and chlorophyll metabolism
5	EC:3.2.1.39 - endo-1,3-beta-D-glucosidase	Starch and sucrose metabolism
6	EC:1.11.1.7 - lactoperoxidase	Phenylpropanoid biosynthesis
7	EC:3.6.1.3 - adenylpyrophosphatase	Purine metabolism
8	EC:3.2.1.37 - 1,4-beta-xylosidase	Amino sugar and nucleotide sugar metabolism
9	EC:2.1.1.53 - N-methyltransferase	Tropane, piperidine and pyridine alkaloid biosynthesis
10	EC:1.1.1.145 - dehydrogenase	Steroid degradation, Steroid hormone biosynthesis
11	EC:1.14.18.1 - monophenol monooxygenase	Betalain biosynthesis, Isoquinoline alkaloid biosynthesis, Tyrosine metabolism
12	EC:1.2.1.38 - reductase	Arginine biosynthesis, Biosynthesis of antibiotics
13	EC:2.7.1.138 - kinase	Sphingolipid metabolism
14	EC:3.2.1.26 - invertase	Galactose metabolism, Starch and sucrose metabolism
15	EC:1.10.3.1 - oxidase	Isoquinoline alkaloid biosynthesis, Tyrosine metabolism
16	EC:2.6.1.16 - transaminase (isomerizing)	Amino sugar and nucleotide sugar metabolism, Biosynthesis of antibiotics, Alanine, aspartate and glutamate metabolism
17	EC:1.14.11.15 - 3beta-dioxygenase	Diterpenoid biosynthesis
18	EC:2.1.1.45 - synthase	One carbon pool by folate, Pyrimidine metabolism
19	EC:2.7.1.33 - kinase	Pantothenate and CoA biosynthesis
20	EC:1.14.13.8 - monooxygenase	Drug metabolism - cytochrome P450
21	EC:6.1.1.18 - ligase	Aminoacyl-tRNA biosynthesis
22	EC:2.1.1.43 - N-methyltransferase	Lysine degradation
23	EC:2.1.3.3 - carbamoyltransferase	Arginine biosynthesis, Biosynthesis of antibiotics
24	EC:2.5.1.18 - transferase	Drug metabolism - cytochrome P450, Metabolism of xenobiotics by cytochrome P450, Glutathione metabolism
25	EC:2.1.2.2 - formyltransferase	One carbon pool by folate, Purine metabolism, Biosynthesis of antibiotics
26	EC:3.2.1.21 - gentiobiase	Starch and sucrose metabolism, Phenylpropanoid biosynthesis, Cyanoamino acid metabolism
27	EC:2.5.1.1 - geranyl-diphosphate synthase	Terpenoid backbone biosynthesis, Biosynthesis of antibiotics
28	EC:2.7.7.6 - RNA polymerase	Purine metabolism, Pyrimidine metabolism
29	EC:4.2.1.20 - synthase	Phenylalanine, tyrosine and tryptophan biosynthesis, Biosynthesis of antibiotics, Glycine, serine and threonine metabolism
30	EC:2.7.7.7 - DNA polymerase	Purine metabolism, Pyrimidine metabolism
31	EC:6.4.1.2 – carboxylase	Aflatoxin biosynthesis, Carbon fixation pathways in prokaryotes, Propanoate metabolism, Pyruvate metabolism, Biosynthesis of antibiotics, Fatty acid biosynthesis
32	EC:1.5.1.3 - reductase	One carbon pool by folate, Folate biosynthesis
33	EC:3.2.1.22 - melibiase	Galactose metabolism, Sphingolipid metabolism, Glycerolipid metabolism, Glycosphingolipid biosynthesis - globo and isoglobo series
34	EC:1.14.11.13 - 2beta-dioxygenase	Diterpenoid biosynthesis
35	EC:2.6.1.83 - aminotransferase	Lysine biosynthesis, Biosynthesis of antibiotics
36	EC:3.6.1.15 - phosphatase	Thiamine metabolism, Purine metabolism
37	EC:3.2.1.55 - end alpha-L-arabinofuranosidase	Amino sugar and nucleotide sugar metabolism
38	EC:3.2.1.15 - pectin depolymerase	Pentose and glucuronate interconversions
39	EC:1.13.11.27 - dioxygenase	Ubiquinone and other terpenoid-quinone biosynthesis, Tyrosine metabolism, Phenylalanine metabolism
40	EC:1.10.3.3 - oxidase	Ascorbate and aldarate metabolism
41	EC:3.2.1.48 - alpha-glucosidase	Starch and sucrose metabolism
42	EC:1.1.1.44 - dehydrogenase (NADP+-dependent, decarboxylating)	Biosynthesis of antibiotics, Pentose phosphate pathway, Glutathione metabolism
43	EC:2.4.1.69 - 1 galactoside alpha-(1,2)-fucosyltransferase	Glycosphingolipid biosynthesis - globo and isoglobo series, Glycosphingolipid biosynthesis - lacto and neolacto series
44	EC:1.1.1.288 - dehydrogenase	Carotenoid biosynthesis
45	EC:3.2.1.3 - 1,4-alpha-glucosidase	Starch and sucrose metabolism
46	EC:3.1.1.1 - ali-esterase	Drug metabolism - other enzymes
47	EC:3.2.1.23 - lactase (ambiguous)	Galactose metabolism, Sphingolipid metabolism, Glycosaminoglycan degradation, Glycosphingolipid biosynthesis - ganglio series, Other glycan degradation
48	EC:3.2.1.14 - ChiC	Amino sugar and nucleotide sugar metabolism

49	EC:2.4.1.11 - synthase	Starch and sucrose metabolism
50	EC:1.1.1.1 – dehydrogenase	Glycolysis / Gluconeogenesis, Chloroalkane and chloroalkene degradation, Naphthalene degradation, Drug metabolism - cytochrome P450, Metabolism of xenobiotics by cytochrome P450, Fatty acid degradation, alpha-Linolenic acid metabolism, Biosynthesis of antibiotics, Retinol metabolism, Tyrosine metabolism, Glycine, serine and threonine metabolism
51	EC:2.5.1.16 - synthase	Arginine and proline metabolism, Cysteine and methionine metabolism, beta-Alanine metabolism, Glutathione metabolism

Table S6.7 Detailed information of 64 TAVs for caffeine detected by using CC and CE genomes as reference followed by KEGG pathway analysis

No	Substrate	Pathways	EC	#seq	Genes	ID
1	PRPP	- Phenylalanine, tyrosine and tryptophan biosynthesis	- EC:2.4.2.18 - phosphoribosyltransferase	1	anthranilate chloroplastic-like	1
2	Glutamine-Glutamate	- Pyrimidine metabolism	- EC:6.3.5.5 - synthase (glutamine-hydrolysing)	1	LOB domain-containing 1-like	2
		- Alanine, aspartate and glutamate metabolism	- EC:6.3.5.5 - synthase (glutamine-hydrolysing)	1	LOB domain-containing 1-like	3
			- EC:6.3.1.2 – synthetase	1	glutamine synthetase	4
		- Glyoxylate and dicarboxylate metabolism	- EC:6.3.1.2 – synthetase	1	glutamine synthetase	5
		- Arginine biosynthesis	- EC:6.3.1.2 – synthetase	1	glutamine synthetase	6
		- Nitrogen metabolism	- EC:6.3.1.2 – synthetase	1	glutamine synthetase	7
		- Folate biosynthesis	- EC:6.3.2.17 - synthase	1	dihydrofolate synthetase	8
3	SAICAR-AICAR	- Purine metabolism	- EC:4.3.2.2 - lyase	1	adenylosuccinate lyase-like	9
4	IMP-XMP	- Purine metabolism	- EC:1.1.1.205 - dehydrogenase	1	inosine-5 -monophosphate dehydrogenase 2-like	10
						11
5	Xanthine	- Purine metabolism	- EC:1.17.1.4 - dehydrogenase	1	xanthine dehydrogenase 1-like	12
6	SAM	- Cysteine and methionine metabolism	- EC:2.5.1.16 - synthase	1	Spermidine synthase 1	13
			- EC:2.1.1.37 - (cytosine-5-)-methyltransferase	1	DNA (cytosine-5)-methyltransferase 1B-like	14
		- Arginine and proline metabolism	- EC:2.5.1.16 - synthase	1	Spermidine synthase 1	15
7	ATP-ADP	- Purine metabolism	- EC:3.6.1.3 – adenylpyrophosphatase	1	P-glyco 9 isoform 1	16
				2	DNA mismatch repair MSH3	17
				3	ABC transporter C family member 10-like	18
				4	ATPase plasma membrane-type	19
				5	calcium-transporting ATPase plasma membrane-type	20
				6	P-glyco 21	21
				7	ATP-dependent RNA helicase DEAH13	22
				8	DEAD-box ATP-dependent RNA helicase	23

					mitochondrial	
				9	probable phospholipid-transporting ATPase 5	24
				10	ABC transporter B family member 11-like	25
				11	DEAD-box ATP-dependent RNA helicase 37	26
				12	RNA polymerase II transcription factor B subunit 2	27
				13	ruvB 1	28
				14	abc transporter i family member chloroplastic	29
				15	ATP-binding cassette transporter	30
				16	calcium-transporting ATPase plasma membrane-type	31
				17	topoisomerase I	32
				18	probable manganese-transporting ATPase PDR2	33
				19	DEAD-box ATP-dependent RNA helicase 17 isoform X1	34
			- EC:3.6.1.15 – phosphatase	1	DNA mismatch repair MSH3	35
				2	ABC transporter C family member 10-like	36
				3	unnamed protein product	37
				4	ATPase plasma membrane-type	38
				5	calcium-transporting ATPase plasma membrane-type	39
				6	P-glyco 21	40
				7	unnamed protein product	41
				8	D -box ATP-dependent RNA helicase D 12-like	42
				9	dynamin-related 5A	43
				10	cell division homolog 2- chloroplastic-like	44
				11	unnamed protein product	45
				12	ATP-dependent RNA helicase DEAH13	46
				13	DEAD-box ATP-dependent RNA helicase mitochondrial	47
				14	probable phospholipid-transporting ATPase 5	48

15	twinkle homolog chloroplastic mitochondrial	49
16	unnamed protein product	50
17	ABC transporter B family member 11-like	51
18	E3 ubiquitin- ligase CIP8	52
19	DEAD-box ATP-dependent RNA helicase 37	53
20	RNA polymerase II transcription factor B subunit 2	54
21	elongation factor-like GTPase 1	55
22	ruvB 1	56
23	abc transporter i family member chloroplastic	57
24	ATP-binding cassette transporter	58
25	calcium-transporting ATPase plasma membrane-type	59
26	topoisomerase I	60
27	DEAD-box ATP-dependent RNA helicase 17 isoform X1	61
28	kinesin KIN-14R	62
29	probable manganese-transporting ATPase PDR2	63
30	P-glyco 9 isoform 1	64

Table S6.7 (continued)

ID	Chrom.	Region			Bulk 1					Bulk 2					
			Ref.	Alle	Zyg	Cov.	Freq.	F/R bal.	Ave. qual.	Zyg	Cov.	Freq.	F/R bal.	Ave. qual.	AAC
1	CC1.1ch04	12730180	A	T	het	24	25	0.33	36.67	het	40	57	0.35	35.17	Ser-Thr
2	CC1.1ch09	2619830	G	C	het	21	52	0.38	35.36	het	48	23	0.33	35.91	Arg-Pro
3	CC1.1ch09	2619830	G	C	het	21	52	0.38	35.36	het	48	23	0.33	35.91	Arg-Pro
4	CC1.1ch00	197044125	C	T	hom	21	100	0.29	35.67	het	36	61	0.43	34.09	Glu-Lys
5	CC1.1ch00	197044125	C	T	hom	21	100	0.29	35.67	het	36	61	0.43	34.09	Glu-Lys
6	CC1.1ch00	197044125	C	T	hom	21	100	0.29	35.67	het	36	61	0.43	34.09	Glu-Lys
7	CC1.1ch00	197044125	C	T	hom	21	100	0.29	35.67	het	36	61	0.43	34.09	Glu-Lys

8	CE1.1ch02	1468986	T	C	het	23	78	0.39	35.83	het	34	24	0.20	34.38	His-Arg
9	CE1.1ch02	35626636	A	T	het	50	38	0.43	35.42	het	60	57	0.37	34.97	Tyr-Phe
10	CE1.1ch04	11359176	C	G	het	22	59	0.23	35.31	het	29	38	0.43	33.91	His-Gln
11	CE1.1ch04	11364138	G	A	het	20	65	0.50	34.46	het	31	42	0.36	36.08	Gly-Ser
12	CE1.1ch07	22201200	G	C	het	23	22	0.40	37.00	het	28	75	0.43	36.24	Gln-Glu
13	CC1.1ch11	30965526	A	C	het	27	56	0.47	36.67	het	57	30	0.39	35.29	Ser-Ala
14	CC1.1ch07	13417576	G	A	het	23	52	0.36	37.17	het	66	44	0.48	34.76	Ala-Thr
15	CC1.1ch11	30965526	A	C	het	27	56	0.47	36.67	het	57	30	0.39	35.29	Ser-Ala
16	CC1.1ch00	39754182	C	A	het	48	69	0.43	36.82	het	70	47	0.38	35.12	Glu-Asp
17	CC1.1ch01	22513197	G	A	het	67	63	0.27	36.05	het	71	37	0.46	34.50	Asp-Asn
18	CC1.1ch02	1650005	A	G	het	30	60	0.33	36.78	het	57	37	0.45	36.10	Thr-Ala
19	CC1.1ch02	55043880	T	A	het	21	76	0.38	35.94	het	42	31	0.43	35.00	Leu-Ile
20	CC1.1ch03	3492096	G	T	het	34	56	0.42	36.89	het	54	35	0.35	32.84	His-Asn, Sys-Phe
21	CC1.1ch04	8795183	A	C	het	21	24	0.20	38.00	het	37	54	0.25	35.65	Leu-Val
22	CC1.1ch08	28614027	A	G	het	23	30	0.29	36.71	het	51	55	0.23	35.21	Ile-Val
23	CC1.1ch08	29580266	A	G	het	32	56	0.28	35.67	het	48	29	0.33	36.43	Gln-Arg
24	CC1.1ch09	8758342	G	T	het	33	61	0.48	37.00	het	47	36	0.41	34.76	Asp-Tyr
25	CC1.1ch11	20642004	T	C	het	21	24	0.40	35.00	het	35	66	0.33	36.39	Val-Ala
26	CE1.1ch01	5099260	A	C	het	37	57	0.42	36.29	het	70	36	0.48	35.40	Ser-Ala
27	CE1.1ch01	14237249	T	A	het	24	46	0.45	37.18	het	43	74	0.41	34.97	Glu-Val
28	CE1.1ch01	23751605	T	C	het	31	65	0.43	36.50	het	44	25	0.42	36.55	Thr-Ala
29	CE1.1ch02	4765425	T	C	het	21	38	0.50	33.88	het	21	71	0.33	34.73	Leu-Pro
30	CE1.1ch02	49050069	A	C	het	26	58	0.47	36.27	het	27	26	0.14	32.86	LeuPhe
31	CE1.1ch03	5014461	T	G	het	33	58	0.30	36.42	het	45	31	0.50	33.50	Trp-Gly
32	CE1.1ch05	25902339	C	G	het	25	80	0.45	34.55	het	29	24	0.50	36.30	ProArg
33	CE1.1ch11	11004936	A	G	hom	28	100	0.43	37.18	het	27	74	0.42	34.65	Val-Ala
34	CE1.1ch06	10465912	A	C	hom	21	100	0.45	36.38	het	42	33	0.38	33.62	Asn-Lys

35	CC1.1ch01	22513197	G	A	het	67	63	0.27	36.05	het	71	37	0.46	34.50	Asp-Asn
36	CC1.1ch02	1650005	A	G	het	30	60	0.33	36.78	het	57	37	0.45	36.10	Thr-Ala
37	CC1.1ch01	22513197	G	A	het	67	63	0.27	36.05	het	71	37	0.46	34.50	AspAsn
38	CC1.1ch02	55043880	T	A	het	21	76	0.38	35.94	het	42	31	0.43	35.00	Leu-Ile
39	CC1.1ch03	3492096	G	T	het	34	56	0.42	36.89	het	54	35	0.35	32.84	His-Asn; Sys-Phe
40	CC1.1ch04	8795183	A	C	het	21	24	0.20	38.00	het	37	54	0.25	35.65	Leu-Val
41	CC1.1ch05	11738903	C	G	het	24	54	0.38	35.92	het	23	30	0.43	36.00	Glu-Gln
42	CC1.1ch05	18227786	G	A	het	38	32	0.38	36.67	het	52	52	0.46	33.26	Ala-Thr
43	CC1.1ch05	29350876	A	C	het	31	32	0.30	36.30	het	55	56	0.44	32.77	Ser-Ala
44	CC1.1ch06	4651265	T	A	het	31	58	0.47	35.94	het	43	35	0.40	36.33	Cys-Ser
45	CC1.1ch08	23379103	A	T	het	35	60	0.35	35.86	het	66	38	0.50	34.44	Lys-Met
46	CC1.1ch08	28614027	A	G	het	23	30	0.29	36.71	het	51	55	0.23	35.21	Ile-Val
47	CC1.1ch08	29580266	A	G	het	32	56	0.28	35.67	het	48	29	0.33	36.43	Gln-Arg
48	CC1.1ch09	8758342	G	T	het	33	61	0.48	37.00	het	47	36	0.41	34.76	Asp-Tyr
49	CC1.1ch09	9842808	G	T	het	24	33	0.38	34.62	het	36	72	0.46	29.23	Gly-Val
50	CC1.1ch10	23529857	C	A	het	22	73	0.31	37.50	het	47	30	0.29	36.79	Val-Phe
51	CC1.1ch11	20642004	T	C	het	21	24	0.40	35.00	het	35	66	0.33	36.39	Val-Ala
52	CE1.1ch00	317137304	A	G	het	27	67	0.37	35.17	het	26	46	0.38	31.42	Phe-Leu
53	CE1.1ch01	5099260	A	C	het	37	57	0.42	36.29	het	70	36	0.48	35.40	Ser-Ala
54	CE1.1ch01	14237249	T	A	het	24	46	0.45	37.18	het	43	74	0.41	34.97	Glu-Val
55	CE1.1ch01	19695940	C	G	het	24	25	0.29	33.33	het	29	52	0.40	34.60	Ala-Pro
56	CE1.1ch01	23751605	T	C	het	31	65	0.43	36.50	het	44	25	0.42	36.55	Thr-Ala
57	CE1.1ch02	4765425	T	C	het	21	38	0.50	33.88	het	21	71	0.33	34.73	Leu-Pro
58	CE1.1ch02	49050069	A	C	het	26	58	0.47	36.27	het	27	26	0.14	32.86	Leu-Phe
59	CE1.1ch03	5014461	T	G	het	33	58	0.30	36.42	het	45	31	0.50	33.50	Trp-Gly
60	CE1.1ch05	25902339	C	G	het	25	80	0.45	34.55	het	29	24	0.50	36.30	Pro-Arg
61	CE1.1ch06	10465912	A	C	hom	21	100	0.45	36.38	het	42	33	0.38	33.62	Asn-Lys
62	CE1.1ch07	9654627	G	T	het	35	23	0.50	37.12	het	29	55	0.35	36.12	Val-Phe

63	CE1.1ch11	11004936	A	G	hom	28	100	0.43	37.18	het	27	74	0.42	34.65	Val-Ala
64	CC1.1ch00	39754182	C	A	het	48	69	0.43	36.82	het	70	47	0.38	35.12	Glu-Asp
						29		**0.38**	**36**		**44**		**0.39**	**35**	

Table S6.8 Detailed information of 36 unique TAVs for trigonelline detected by using CC and CE genomes as reference followed by KEGG pathway analysis

No	Substrates	Pathways	EC	# seq	Genes	ID
1	Glucose	Starch and sucrose metabolism	EC:3.2.1.21 - gentiobiase	1	glucan endo-1,3-beta-glucosidase 6	1
				1	beta-glucosidase 3B-like	2
			EC:3.2.1.3 - 1,4-alpha-glucosidase	1	unnamed protein product	3
			EC:3.2.1.26 – invertase	1	beta- insoluble isoenzyme 1	4
			EC:3.2.1.39 - endo-1,3-beta-D-glucosidase	1	glucan endo-1,3-beta-glucosidase 6	5
		Galactose metabolism	EC:3.2.1.22 - melibiase	1	alpha-galactosidase 3	6
2	L-tryptophan	Glycine, serine and threonine metabolism	EC:4.2.1.20 - synthase	1	tryptophan synthase beta chain 2	7
		Phenylalanine, tyrosine and tryptophan biosynthesis	EC:4.2.1.20 - synthase	1	tryptophan synthase beta chain 2	8
3	SAM	Cysteine and methionine metabolism	EC:2.5.1.16 – synthase	1	Spermidine synthase 1	9
		Arginine and proline metabolism	EC:2.5.1.16 - synthase	1	Spermidine synthase 1	10
4	ATP-ADP	Purine metabolism	EC:3.6.1.15 - phosphatase	1	116 kDa U5 small nuclear ribonucleo component-like	11
				2	D -box ATP-dependent RNA helicase D 3	12
				3	dynamin-related 4C-like	13
				4	unnamed protein product	14
				5	E3 ubiquitin- ligase CIP8	15
				6	DNA mismatch repair MSH3	16
				7	pleiotropic drug resistance 3	17
				8	twinkle homolog chloroplastic mitochondrial	18

				9	translation initiation factor IF- chloroplastic	19
				10	D -box ATP-dependent RNA helicase D 6-like isoform X1	20
				11	dynamin-related 5A	21
				12	pleiotropic drug resistance 3-like	22
				13	ABC transporter B family member 11-like	23
				14	ABC transporter G family member 6-like	24
				15	replication factor C subunit 2	25
				16	unnamed protein product	26
				17	calcium-transporting ATPase plasma membrane-type	27
			EC:3.6.1.3 - adenylpyrophosphatase	1	pleiotropic drug resistance 3-like	28
				2	D -box ATP-dependent RNA helicase D 3	29
				3	ABC transporter B family member 11-like	30
				4	ABC transporter G family member 6-like	31
				5	replication factor C subunit 2	32
				6	DNA mismatch repair MSH3	33
				7	calcium-transporting ATPase plasma membrane-type	34
				8	pleiotropic drug resistance 3	35
				9	D -box ATP-dependent RNA helicase D 6-like isoform X1	36

Table S6.8 (continued)

ID					Bulk 4					Bulk 3					
	Chromosome	Region	Ref.	Alle	Zyg	Cov.	Freq.	F/R bal.	Ave. qual.	Zyg	Cov.	Freq.	F/R bal.	Ave. qual.	AAC
1	CC1.1ch06	39614846	A	G	het	26	65	0.41	37.06	het	43	44	0.5	34.05	Ile-Val
2	CC1.1ch00	235340977	C	G	het	35	31	0.45	36.82	het	64	56	0.37	34.06	Ser-Thr
3	CC1.1ch01	25200184	A	G	het	30	33	0.45	35.1	het	38	53	0.13	34.05	Glu-Gly
4	CC1.1ch06	4874743	A	G	hom	21	100	0.5	36.33	het	30	77	0.38	34.96	Lys-Arg

5	CC1.1ch06	39614846	A	G	het	26	65	0.41	37.06	het	43	44	0.5	34.05	Ile-Val
6	CE1.1ch02	4440080	A	C	het	43	56	0.41	36.42	het	60	33	0.48	34.1	Gln-Pro
7	CE1.1ch07	24812383	T	C	het	20	45	0.44	36.89	het	35	74	0.33	32.62	Ser-Gly
8	CE1.1ch07	24812383	T	C	het	20	45	0.44	36.89	het	35	74	0.33	32.62	Ser-Gly
9	CC1.1ch11	30965526	A	C	het	36	75	0.26	36.48	het	50	48	0.46	34.75	Ser-Ala
10	CC1.1ch11	30965526	A	C	het	36	75	0.26	36.48	het	50	48	0.46	34.75	Ser-Ala
11	CC1.1ch00	69742620	A	G	het	28	61	0.47	36.41	het	60	42	0.36	35.08	Asp-Gly
12	CC1.1ch01	22513197	G	A	het	43	63	0.38	36.7	het	64	44	0.32	32.82	Asp-Asn
13	CC1.1ch02	7154515	G	T	het	29	34	0.27	36.4	het	46	54	0.35	37	His-Asn
14	CC1.1ch02	49107279	T	C	het	28	32	0.3	34.78	het	53	55	0.38	33.52	Gln-Arg
15	CC1.1ch02	53547886	T	C	het	35	60	0.33	35.95	het	36	42	0.07	33.93	Ile-Met
16	CC1.1ch04	5181922	G	A	het	27	59	0.35	36.94	het	51	37	0.29	35.42	His-Tyr
17	CC1.1ch05	29350435	T	A	het	34	32	0.33	36.64	het	42	55	0.46	35.57	Thr-Ser
18	CC1.1ch08	23379633	C	G	het	36	44	0.44	36.44	het	73	64	0.39	35.79	Leu-Val
19	CC1.1ch09	9823406	C	G	het	25	20	0.29	37.2	het	32	56	0.42	36.72	Gln-Glu
20	CC1.1ch10	27110860	A	C	het	30	40	0.38	36.92	het	38	63	0.44	33.67	Ile-Arg
21	CE1.1ch00	317137277	C	A	het	22	55	0.46	35.58	het	36	36	0.38	32.15	Gly-Trp
22	CE1.1ch01	13526143	C	T	het	22	32	0.43	35.29	het	26	54	0.33	36.86	Ser-Asn
23	CE1.1ch02	51146945	T	G	het	34	26	0.4	35.44	het	45	51	0.36	35.13	Ser-Ala
24	CE1.1ch04	8678658	G	T	het	22	55	0.21	35.17	het	50	32	0.21	34.88	Pro-His
25	CE1.1ch05	1646719	A	G	het	37	59	0.48	37.14	het	77	35	0.43	34.78	Lys-Glu
26	CE1.1ch09	2947791	A	T	het	21	52	0.27	36.27	het	24	21	0.2	32.6	Val-Glu
27	CE1.1ch10	19589484	C	T	het	34	32	0.46	36.27	het	74	62	0.49	35.43	Ser-Asn
28	CC1.1ch01	22513197	G	A	het	43	63	0.38	36.7	het	64	44	0.32	32.82	Asp-Asn
29	CC1.1ch02	7154515	G	T	het	29	34	0.27	36.4	het	46	54	0.35	37	His-Asn
30	CC1.1ch02	53547886	T	C	het	35	60	0.33	35.95	het	36	42	0.07	33.93	Ile-Met

31	CC1.1ch04	5181922	G	A	het	27	59	0.35	36.94	het	51	37	0.29	35.42	His-Tyr
32	CC1.1ch10	27110860	A	C	het	30	40	0.38	36.92	het	38	63	0.44	33.67	Ile-Arg
33	CE1.1ch02	51146945	T	G	het	34	26	0.4	35.44	het	45	51	0.36	35.13	Ser-Ala
34	CE1.1ch04	8678658	G	T	het	22	55	0.21	35.17	het	50	32	0.21	34.88	Pro-His
35	CE1.1ch05	1646719	A	G	het	37	59	0.48	37.14	het	77	35	0.43	34.78	Lys-Glu
36	CE1.1ch10	19589484	C	T	het	34	32	0.46	36.27	het	74	62	0.49	35.43	Ser-Asn
						30		**0.38**	**36**		**49**		**0.36**	**35**	

Table S7.1 Primers flanking significant TAVs for caffeine and trigonelline identified in the present study.

No	Name	Sequence (5' to 3')	No of bases	GC%	Amplicon size (bp)
Primers for caffeine markers					
1	SAICAR For	TGGAACTGGAGGCTACTTTTAGGGTATCTG	30	46	201
2	SAICAR Rev	CTACAGCTTCAGTACTGCTCTCT	23	47	
3	SAM1 For	GGAGAGGCACACTCGCTAAAGGTAGA	26	53	305
4	SAM1 Rev	TTGGGCTTGGAATAGAGCAGAGAGGAAGA	29	48	
5	SAM2 For	TCTTCCGGAGGAAATCCTGACTTGAG	26	50	417
6	SAM2 Rev	TGGCTTCATTGAGCGCCAGAGTTC	24	54	
7	XAN For	CCGTGAATGGCTATTACACTTACCT	24	44	604
8	XAN Rev	CGCCTGCAATTCCCAACACCCTCCACT	27	59	
9	XMP1L1 For	GCAAGTCTGGAGCAGGAATATGGCACA	27	51	297
10	XMP1L1 Rev	AGCTGGAGCAACCACAGCACAGA	23	56	
11	XMP1L2 For	TGTCCACGGCCAACAGCGCAGA	22	63	390
12	XMP1L2 Rev	CGAGGTGGTGAACCTGGCCACTT	23	60	
Primers for trigonelline markers					
13	SAMtrig For	GAGGCACACTCGCTAAAGGTAGA	23	52	300
14	SAMtrig Rev	GGGCTTGGAATAGAGCAGAGAGGA	24	54	
15	Tryp For	AGCGATCTCTGAGGCTGTGGAAGTTG	26	53	373
16	Tryp Rev	TGGGTCGGGTATAAAGTCATGC	22	50	

REFERENCE

Abbott, J. C., and Butcher, S. A. (2012). Strategies towards sequencing complex crop genomes. *Genome Biology* **13**, 1-3.

Addinsoft (2007). XLSTAT, Analyse de données et statistique avec MS Excel. *Addinsoft, NY, USA.*

Aerts, R., Berecha, G., Gijbels, P., Hundera, K., Glabeke, S. V., Vandepitte, K., Muys, B., Roldan-Ruiz, I., and Honnay, O. (2012). Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in South-Western Ethiopian montane rainforests. *Evolutionary Applications*, 243-252.

Aga, E., Bryngelsson, T., Bekele, E., and Salomon, B. (2003). Genetic diversity of forest arabica coffee (*Coffea arabica* L.) in Ethiopia as revealed by Random Amplified Polymorphic DNA (RAPD) analysis. *Hereditas* **138**, 36–46.

Aggarwal, R., Hendre, P., Varshney, R., Bhat, P., Krishnakumar, V., and Singh, L. (2007). Identification, characterization and utilization of EST-derived genetic microsatellite markers for genome analyses of coffee and related species. *Theoretical and Applied Genetics* **114**, 359–372.

Agresti, P. D. C. M., Franca, A. S., Oliveira, L. S., and Augusti, R. (2008). Discrimination between defective and non-defective Brazilian coffee beans by their volatile profile. *Food Chemistry* **106**, 787–796.

Agtron (2004). Agtron M-basic/II Coffee Roast Analyser. *Agtron Inc.*, page 10.

Akaffou, D. S., Hamon, P., Doulbeau, S., Keli, J., Legnate, H., Campa, C., Hamon, S., Kochko, A., and Zoro, B. I. A. (2012). Inheritance and relationship between key agronomic and quality traits in an interspecific cross between *Coffea pseudozanguebariae* Bridson and *C. canephora* Pierre. *Tree Genetics & Genomes* **8**, 1149-1162.

Akiyama, M., Murakami, K., Ikeda, M., Iwatsuki, K., Wada, A., Tokuno, K., Onishi, M., and Iwabuchi, H. (2007). Analysis of the headspace volatiles of freshly brewed arabica coffee using Solid-Phase Microextraction. *Journal of Food Science* **72**, C388-C396.

Akiyama, M., Murakami, K., Ohtani, N., Iwatsuki, K., Sotoyama, K., Wada, A., Tokuno, K., Iwabuchi, H., and Tanaka, K. (2003). Analysis of volatile compounds released during the grinding of roasted coffee beans using Solid-Phase Microextraction. *Journal of Agricultural and Food Chemistry* **51**, 1961-1969.

Al-Dous, E. K., George, B., Al-Mahmoud, M. E., Al-Jaber, M. Y., Wang, H., Salameh, Y. M., Al-Azwani, E. K., Chaluvadi, S., Pontaroli, A. C., DeBarry, J., Arondel, V., Ohlrogge, J., Saie, I. J., Suliman-Elmeer, K. M., Bennetzen, J. L., Kruegger, R. R., and Malek, J. A. (2011). De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). *Nat Biotech* **29**, 521-527.

Al-Murish, T. M., Elshafei, A. A., Al-Doss, A. A., and Barakat, M. N. (2013). Genetic diversity of coffee (*Coffea arabica* L.) in Yemen via SRAP, TRAP and SSR markers. *Journal of Food, Agriculture & Environment* **11**, 411-416.

Amidou, N. D., Michel, N., Serge, H., and Valérie, P. (2007). Genetic basis of species differentiation between Coffea liberica Hiern and C. canephora Pierre: Analysis of an interspecific cross. *Genetic Resources and Crop Evolution* **54**, 1011-1021.

Amigo, J. M., Popielarz, M. J., Callejón, R. M., Morales, M. L., Troncoso, A. M., Petersen, M. A., and Toldam-Andersen, T. B. (2010). Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *Journal of Chromatography A* **1217**, 4422–4429.

Anthony, F., Astorga, C., Avendaño, J., and Dulloo4, E. (2007). Conservation of coffee genetic resources in the CATIE field genebank. *In* "Conserving coffee genetic resources" (F. Engelmann, M. E. Dulloo, C. Astorga, S. Dussert and F. Anthony, eds.). Biodiversity International, Maccarese (Fiumicino), Rome, Italy.

Anthony, F., Bertrand, B., Astorga, C., and Lashermes, P. (2007b). Characterization and assessment of *Coffea arabica* L. genetic resources conserved in the CATIE field genebank. *In* "Conserving coffee genetic resources" (F. Engelmann, M. E. Dulloo, C. Astorga, S. Dussert and F. Anthony, eds.). Biodiversity International, Maccarese (Fiumicino), Rome, Italy.

Anthony, F., Bertrand, B., Etienne, H., and Lashermes, P. (2011). *Coffea* and *Psilanthus*. *In* "Wild Crop Relatives: Genomic and Breeding Resources", pp. 41-61. Springer.

Anthony, F., Bertrand, B., Quiros, O., Wilches, A., Lashermes, P., Berthaud, J., and Charrier, A. (2001). Genetic diversity of wild coffee (Coffea arabica L.) using molecular markers. *Euphytica* **118**, 53–65.

Anthony, F., Clifford, M. N., and Noirot, M. (1993). Biochemical diversity in the genus *Coffea* L.: chlorogenic acids, caffeine and mozambioside contents. *Genetic Resources and Crop Evolution* **40**, 61-70.

Anthony, F., Diniz, L. E. C., Combes, M.-C., and Lashermes, P. (2010). Adaptive radiation in *Coffea* subgenus *Coffea* L. (*Rubiaceae*) in Africa and Madagascar. *Plant Systematics and Evolution* **285**, 51–64.

Ashihara, H. (2006). Metabolism of alkaloids in coffee plants. *Brazilian Journal of Plant Physiology* **18**, 1-8.

Ashihara, H., and Crozier, A. (1999). Biosynthesis and Catabolism of Caffeine in Low-Caffeine-Containing Species of *Coffea*. *J. Agric. Food Chem* **47**, 3425-3431.

Ashihara , H., and Crozier, A. (2001). Caffeine: a well known but little mentioned compound in plant science. *TRENDS in Plant Science* **6**, 407-413.

Ashihara, H., Deng, W. W., and Nagai, C. (2011). Trigonelline biosynthesis and the pyridine nucleotide cycle in *Coffea* arabica fruits: Metabolic fate of [carboxyl-^{14}] nicotinic acid riboside. *Phytochemistry Letters* **4**, 235-239.

Ashihara, H., Kato, M., and Crozier, A. (2011b). Distribution, Biosynthesis and Catabolism of Methylxanthines in Plants. *Handbook of Experimental Pharmacology* **200**, 11-31.

Ashihara, H., Ludwig, I. A., Katahira, R., Yokota, T., Fujimura, T., and Crozier, A. (2015). Trigonelline and related nicotinic acid metabolites: occurrence, biosynthesis, taxonomic considerations, and their roles in planta and in human health. *Phytochem Reviews* **14**, 765–798.

Ashihara, H., Monteiro, A. M., Gillies, F. M., and Crozier, A. (1996). Biosynthesis of Caffeine in Leaves of Coffee. *Plant Physiol.* **111**, 747-753.

Ashihara, H., Sano, H., and Crozier, A. (2008). Caffeine and related purine alkaloids: biosynthesis, catabolism, function and genetic engineering. *Phytochemistry* **69**, 841-856.

Ashihara , H., and Suzuki, T. (2004). Distribution and biosynthesis of caffeine in plants. *Frontiers in Bioscience* **9**, 1864-1876.

Avelino, J., Barboza, B., Araya, J. C., Fonseca, C., Davrieux, F., Guyot, B., and Cilas, C. (2005). Effects of slope exposure, altitude and yield on coffee quality in two altitude terroirs of Costa Rica, Orosi and Santa Mar´ıa de Dota. *Journal of the Science of Food and Agriculture* **85**, 1869–1876.

Bagur-González, M. G., Pérez-Castano, E., Sánchez-Vinas, M., and Gázquez-Evangelista, D. (2015). Using the liquid-chromatographic-fingerprint of sterols fraction to discriminate virgin olive from other edible oils. *Journal of Chromatography A* **1380**, 64-70.

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nat Methods* **9**.

Barbosa, J. N., Borém, F. M., Cirillo, M. Â., Malta, M. R., Alvarenga, A. A., and Alves, H. M. R. (2012). Coffee quality and its interactions with environmental factors in Minas Gerais, Brazil. *Journal of Agricultural Science* **4**, 181-190.

Barre, P., Akaffou, S., Louarn, J., Charrier, A., Hamon, S., and Noirot, M. (1998). Inheritance of caffeine and heteroside contents in an interspecific cross between a cultivated coffee species *Coffea liberica var dewevrei* and a wild species caffeine-free *C. pseudozanguebariae*. *Theoretical and Applied Genetics* **96**, 306-311.

Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stobe, P., Futschik, A., and Schlotterer, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *Plos genetics* **9**, e1003534.

Baumann, T. W., Sondahl, M. R., Waldhauser, S. S. M., and Kretschmar, J. A. (1998). Non-destructive analysis of natural variability in bean caffeine content of Laurina coffee. *Phytochemistry* **49**, 1569-1573.

Belay, A. (2011). Some biochemical compounds in coffee beans and methods developed for their analysis. *International Journal of the Physical Sciences* **6**, 6373-6378.

Belete, Y. (2014). Performance evaluations of hundred beans weights of indigenous Arabica coffee genotypes across different environments. *Sky Journal of Agricultural Research* **3** 120 - 127.

Belitz, H.-D., Grosch, W., and Schieberle, P. (2009). Food chemistry (4th ed.). *In* "Chapter 21: Coffee, tea, cocoa" (H.-D. Belitz, W. Grosch and P. Schieberle, eds.), pp. 938-970. Springer-Verlag Berlin Heidelberg, Germany.

Benatti, L. B., Silvarolla, M. B., and Mazzafera, P. (2012). Characterisation of AC1: a naturally decaffeinated coffee. *Bragantia* **71**, 143-154.

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology* **33**, 623-634.

Bertrand, B., Boulanger, R., Dussert, S., Ribeyre, F., Berthiot, L., Descroix, F., and Joet, T. (2012). Climatic factors directly impact the volatile organic compound fingerprint in green Arabica coffee bean as well as coffee beverage quality. *Food Chemistry* **135**, 2575-2583.

Bertrand, B., Etienne, H., Cilas, C., Charrier, A., and Baradat, P. (2005). *Coffea arabica* hybrid performance for yield, fertility and bean weight. *Euphytica* **141**, 255–262.

Bertrand, B., Vaast, P., Alpizar, E., Etienne, H., Davrieux, F., and Charmetant, P. (2006). Comparison of bean biochemical composition and beverage quality of Arabica hybrids involving Sudanese-Ethiopian origins with traditional varieties at various elevations in Central America. *Tree Physiology* **26**, 1239-1248.

Bicchi, C., Ruosi, M. R., Cagliero, C., Cordero, C., Liberto, E., Rubiolo, P., and Sgorbini, B. (2011). Quantitative analysis of volatiles from solid matrices of vegetable origin by high concentration capacity headspace techniques: Determination of furan in roasted coffee. *Journal of Chromatography A* **1218**, 753-762.

Bicchi, C. P., Panero, O. M., Pellegrino, G. M., and Vanni, A. C. (1997). Characterization of Roasted Coffee and Coffee Beverages by Solid Phase Microextraction−Gas Chromatography and Principal Component Analysis. *Journal of Agricultural and Food Chemistry* **45**, 4680-4686.

Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., Yuen, M. M. S., Keeling, C. I., Brand, D., Vandervalk, B. P., Kirk, H., Pandoh, P., Moore, R. A., Zhao, Y., Mungall, A. J., Jaquish, B., Yanchuk, A., Ritland, C., Boyle, B., Bousquet, J., Ritland, K., MacKay, J., Bohlmann, J., and Jones, S. J. M. (2013). Assembling the 20Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Genome analysis*, doi:10.1093/bioinformatics/btt178.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Genome analysis* **27**, 578–579.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 1-9.

Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of computational biology* **17**, 1519-1533.

Brown, S. D., Nagaraju, S., Utturkar, S., Tissera, S. D., Segovia, S., Mitchell, W., Land, M. L., Dassanayake, A., and Köpke, M. (2014). Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia. *Biotechnology for Biofuels* **7**, 1-18.

Brown, S. D., Utturkar, S. M., Klingeman, D. M., Johnson, C. M., Martin, S. L., Land, M. L., Lu, T.-Y. S., Schadt, C. W., Doktycz, M. J., and Pelletiera, D. A. (2012). Twenty-One Genome Sequences from Pseudomonas Species and 19 Genome Sequences from Diverse Bacteria Isolated from the Rhizosphere and Endosphere of Populus deltoides. *Journal of Bacteriology* **194**, 5991–5993.

Buffo, R. A., and Cardelli-Freire, C. (2004). Coffee flavour: an overview. *Flavour And Fragrance Journal* **19**, 99–104.

Byers, R. L., Harker, D. B., Yourstone, S. M., Maughan, P. J., and Udall, J. A. (2012). Development and mapping of SNP assays in allotetraploid cotton. *Theor Appl Genet* **124**, 1201–1214.

Cacao, S. M. B., Silva, N. V., Domingues, D. S., and Vieira, L. G. E. (2013). Construction and characterization of a BAC library from the *Coffea arabica* genotype Timor Hybrid CIFC 832/2. *Genetica* **141**, 217–226.

Campa, C., Ballester, J. F., Doulbeau, S., Dussert, S., Hamon, S., and Noirot, M. (2004). Trigonelline and sucrose diversity in wild Coffea species. *Food Chemistry* **88**, 39-43.

Campa, C., Doulbeau, S., Dussert, S., Hamon, S., and Noirot, M. (2005a). Diversity in bean caffeine content among wild Coffea species: evidence of a discontinuous distribution. *Food Chemistry* **91**, 633–637.

Campa, C., Noirot, M., Bourgeois, M., Pervent, M., Ky, C., Chrestin, H., Hamon, S., and de Kochko, A. (2003). Genetic mapping of a caffeoyl-coenzyme A 3-0-methyltransferase gene in coffee trees. Impact on chlorogenic acid content. *Theoretical and Applied Genetics* **107**, 751–756.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S.-H., Childs, K. L., Sun, Y., Jiang, N., and Yandell, M. (2014). MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. *Plant Physiology* **164**, 513-524.

Campos, R., Corrêa, P. C., Busto, D., Assis, S., and Zaidan, Ú. (2014). Evaluation of capacity expansion of mocha grains and its comparison with the conventional ones. *In* "The 25th International conference on coffee and science ASIC 2014", Colombia.

Cardon, L. R., and Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet* **361**, 598–604.

Casal, S., Oliveira, M. B., and Ferreira, M. A. (1998). Development of an HPLC/Diode-Array Detector method for simultaneous determination of Trigonelline, Nicotinic Acid, and Caffeine in coffee. *Journal of Liquid Chromatography & Related Technologies* **21**, 3187-3195.

Casal, S., Oliveira, M. B. P. P., Alves, M. R., and Ferreira, M. A. (2000). Discriminate analysis of roasted coffee varieties for trigonelline, nicotinic acid, and caffeine content. *Journal of Agricultural and Food Chemistry* **48**, 3420-3424.

Castle, J. C., Loewer, M., Boegel, S., Tadmor, A. D., Boisguerin, V., Graaf, J. d., Paret, C., Diken, M., Kreiter, S., Tureci, O., and Sahin, U. (2014). Mutated tumor alleles are expressed according to their DNA frequency. *Scientific reports* **4**.

Cenci, A., Combes, M.-C., and Lashermes, P. (2012). Genome evolution in diploid and tetraploid *Coffea* species as revealed by comparative analysis of orthologous genome segments. *Plant Molecular Biology* **78**, 135–145.

Chain, P. S. G., Grafham, D. V., Fulton, R. S., FitzGerald, M. G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D. C., Buhay, C., Cole, J. R., Ding, Y., Dugan, S., Field, D., Garrity, G. M., Gibbs, R., Graves, T., Han, C. S., Harrison, S. H., S.Highlander, Hugenholtz, P., Khouri, H. M., Kodira, C. D., Kolker, E., Kyrpides, N. C., Lang, D., Lapidus, A., Malfatti, S. A., Markowitz, V., Metha, T., Nelson, K. E., Parkhill, Pitluck, S., Qin, X., Read, T. D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R. L., Sutton, G., Thomson, N. R., Tiedje, J. M., Weinstock, G., Wollam, A., Jumpstart, C. G. S. C. H. M. P., and Detter, J. C. (2009). Genome project standards in a new era of sequencing. *Science* **326**, doi:10.1126/science.1180614.

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **12**, 1-17.

Charrier, A., and Eskes, A. B. (2004). Botany and Genetics of Coffee. *In* "Coffee: Growing, Processing, Sustainable Production - A Guidebook for Growers, Processors, Traders, and Researchers" (J. N. Wintgens, ed.). WILEY-VCH Verlag GmbH & Co. KCaA, Germany.

Charrier, A., Lashermes, P., and Eskes, A. B. (2012). Botany, genetics and genomics of coffee. *In* "Coffee: Growing, Processing, Sustainable Production - A Guidebook for Growers, Processors, Traders, and Researchers" (J. N. Wintgens, ed.). WILEY-VCH Verlag GmbH & Co. KCaA.

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., and Hwang, C.-C. (2013). Effects of GC bias in Next-Generation-Sequencing data on de novo genome assembly. *PLOS one* **2013**, 4.

Cheong, M. W., Tong, K. H., Ong, J. J. M., Liu, S. Q., Curran, P., and Yu, B. (2013). Volatile composition and antioxidant capacity of Arabica coffee. *Food Research International* **51**, 388-396.

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31–37.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Mature methods* **10**, 563-571.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., and Schatz, M. C. (2016). Phased diploid genome assembly with Single Molecule Real-Time sequencing. *bioRxiv*, http://dx.doi.org/10.1101/056887.

Clarke, R., and Macarae, R. (1985). "Coffee Vol.1: Chemistry," Elsevier, New York.

Clarke, y. E., Higgins, i. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., Batley, J., Edwards, D., Meng, J., Li, R., Lawley, C. T., Pauquet, J., Laga, B., Cheung, W., Iniguez‐Luy, F., Dyrszka, E., Rae, S., Stich, B., Snowdon, R. J., Sharpe, A. G., Ganal, M. W., and Parkin, I. A. P. (2016). A high‐density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single‐locus markers in the allotetraploid genome. *Theor Appl Genet* **129**, 1887–1899.

Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., and Jackson, S. A. (2015). Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. *Molecular Plant* **8**, 831–846.

Clifford, M. N., and Kazi, T. (1987). The influence of coffee bean maturity on the content of chlorogenic acids, caffeine and trigonelline. *Food Chemistry* **26**, 59-69.

Clifford, M. N., Williams, T., and Bridson, D. (1989). Chlorogenic acids and caffeine as possible taxonomic criteria in Coffea and Psilanthus. *Phytochemistry* **28**, 829-838.

Combes, M.-C., Dereeper, A., Severac, D., Bertrand, B., and Lashermes, P. (2013). Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. *New phytologist* **200**, 251–260.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676.

Cotta, M. G., Barros, L. M. G., Almeida, J. D. d., Lamotte, F. d., Barbosa, E. A., Vieir, N. G., Alves, G. S. C., Vinecky, F., Andrade, A. C., and Marraccini, P. (2014). Lipid transfer proteins in coffee: isolation of *Coffea* orthologs, *Coffea arabica* homeologs, expression during coffee fruit development and promoter analysis in transgenic tobacco plants. *Plant Mol Biol* **85**, 11–31.

Coulibaly, I., Revol, B., Noirot, M., Poncet, V., Lorieux, M., Carasco-Lacombe, C., Minier, J., Dufour, M., and Hamon, P. (2003). AFLP and SSR polymorphism in a Coffea interspecific backcross progeny [(*C. heterocalyx* × *C. canephora*) × *C. canephora*]. *Theoretical and Applied Genetics* **107**, 1148-1155.

Couturon, E., Lashermes, P., and Charrier, A. (1998). First intergeneric hybrids (*Psilanthus ebracteolatus* Hiern × *Coffea arabica* L.) in coffee trees. *Canadian Journal of Botany* **76**, 542-546.

Cros, J., Combes, M. C., Chabrillange, N., C. Duperray, Angles, A. M. d., and Hamon, S. (1995). Nuclear DNA content in the subgenus *Coffea* (Rubiaceae): inter- and intra-specific variation in African species. *Canadian Journal of Botany* **73**, 14-20.

Cros, J., Combes, M. C., Trouslot, P., Anthony, F., Hamon, S., Charrier, A., and Lashermes, P. (1998). Phylogenetic analysis of chloroplast DNA variation in *Coffea* L. *Molecular phylogenetics and evolution* **9**, 109–117.

Crouzillat, D., Rigoreau, M., Bellanger, L., Priyono, P., Mawardi, S., Syahrudi, McCarthy, J., Tanksley, S., Zaenudin, I., and Pétiard, V. (2005). A Robusta consensus genetic map using RFLP and microsatellite markers for the detection of QTL. *In* "20th International Conference on Coffee Science (ASIC 2004)", pp. 546-553 Bangalore, India.

Cubry, P., Musoli, P., H, L., D, P., F, d. B., V, P., F, A., M, D., and T., L. (2008). Diversity in coffee assessed with SSR markers: structure of the genus *Coffea* and perspectives for breeding. *genome* **51**, 50-63.

Custers, D., Canfyn, M., P.Courselle, J.O.DeBeer, Apers, S., and Deconinck, E. (2014). Headspace–gas chromatographic fingerprints to discriminate and classify counterfeit medicines. *Talanta* **123**, 78-88.

Czerny, M., F., M., and Grosch, W. (1999). Sensory study on the character impact odorants of roasted arabica coffee. *Journal of Agricultural and Food Chemistry* **47**, 695-699.

Czerny, M., and Grosch, W. (2000). Potent odorants of raw arabica coffee. Their changes during roasting. *Journal of Agricultural and Food Chemistry* **48**, 868-872.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Natural Methods* **9**, 772.

Dart, S. K., and Nursten, H. E. (1985). Chapter 7: Volatile components. *In* "Coffee" (R. J. Clarke and R. Macrae, eds.), pp. 223-265. Elsevier Science Publishers.

Das, S., Upadhyaya, H. D., Bajaj, D., Kujur, A., Badoni, S., Laxmi, Kumar, V., Tripathi, S., Gowda, C. L. L., Sharma, S., Singh, S., Tyagi, A. K., and Parida, S. K. (2015). Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. *DNA Research* **22**, 193–203.

Davis, A., Bridson, D., and Rakotonasolo, F. (2005). A re-examination of *Coffea* subgenus *Baracoffea* and comments on the morphology and classification of *Coffea* and *Psilanthus* (*Rubiaceae–Coffeeae*). *In* "A festschrift for William G. D'Arcy: the legacy of a taxonomist. (Monographs in systematic botany 104)." (K. RC, H. VC and C. T, eds.), pp. 399–420. Missouri Botanical Garden Press, Missouri.

Davis, A. P., Chester, M., Maurin, O., and Fay, M. F. (2007). Searching for the relatives of *Coffea (Rubiaceae, Ixoroideae)*: The circumscription and phylogeny of *Coffeeae* based on plastid sequence data and morphology. *American Journal of Botany* **94**, 313–329.

Davis, A. P., Govaerts, R., Bridson, D. M., and Stoffelen, P. (2006). An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). *Botanical Journal of the Linnean Society* **152**, 465–512.

Davis, A. P., Tosh, J., Ruch, N., and Fay, M. F. (2011). Growing coffee: *Psilanthus* (*Rubiaceae*) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of *Coffea*. *Botanical Journal of the Linnean Society* **167**, 357-377.

De Castro, R. D., and Marraccini, P. (2006). Cytology, biochemistry and molecular changes during coffee fruit development. *Brazilian Journal of Plant Physiology* **18**, 175-199.

De Maria, C. A. B., Trugo, L. C., and Moreira, R. F. A. (1995). Simultaneous determination of total chlorogenic acid, trigonelline and caffeine in green coffee samples by high performance gel filtration chromatography. *Food Chemistry* **52**, 447-449.

De Nardi, B., Dreos, R., Del Terra, L., and Martellossi, C. (2006). Differential responses of *Coffea arabica* L. leaves and roots to chemically induced systemic acquired resistance. *Genome* **49**, 1594–1605.

de Roos, B., Weg, G. v. d., Urgert, R., Bovenkamp, P. v. d., Charrier, A., and Katan, M. B. (1997). Levels of cafestol, kahweol, and related diterpenoids in wild species of the coffee plant *Coffea*. *Journal of Agricultural and Food Chemistry* **45**, 3065-3069.

Dean, A. (2006). On a chromosome far, far away: LCRs and gene expression. *TRENDS in Genetics* **22**, 38-45.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., and Droc, G. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181-1184

Dereeper, A., Bocs, S., Rouard, M., Guignon, V., Ravel, S., Tranchant-Dubreuil, C., Poncet, V., Garsmeur, O., Lashermes, P., and Droc, G. (2014). The coffee genome hub: a resource for coffee genomes. *Nucleic Acids Research*.

Dereeper, A., Guyot, R., Tranchant-Dubreuil, C., and Anthony, F. o. (2013). BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution. *Plant Molecular Biology* **83**, 177-189.

Desai, A., Marwah, V. S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V., and Jere, A. (2013). Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PLOS one* **8**, e60204.

Dessalegn, Y., Herselman, L., and Labuschagne, M. (2009). Comparison of SSR and AFLP analysis for genetic diversity assessment of Ethiopian arabica coffee genotypes. *S. Afr. J. Plant Soil* **26**, 119-125.

Dessalegn, Y., Herselman, L., and Labuschagne, M. T. (2008b). AFLP analysis among Ethiopian arabica coffee genotypes. *African Journal of Biotechnology* **7**, 3193-3199.

Dessalegn, Y., Labuschagne, M. T., Osthoff, G., and Herselman, L. (2008). Genetic diversity and correlation of bean caffeine content with cup quality and green bean physical characteristics in coffee (*Coffea arabica* L.). *Journal of the Science of Food and Agriculture* **88**, 1726-1730.

Deynze, V. A. (2017). Update on the Sequencing of the *Coffea arabica* Variety, Geisha. *In* "Plant and Animal Genome XXV ", pp. http://app.core-apps.com/pag-2017/abstract/1a72fbe5da697beb0993e254b2c68756, San Diego, CA.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**, 1-10.

Ducki, S., Miralles-Garcia, J., Zumbe, A., Tornero, A., and Storey, D. M. (2008). Evaluation of Solid-Phase Micro-Extraction coupled to Gas Chromatography–Mass Spectrometry for the headspace analysis of volatile compounds in cocoa products. *Talanta* **74**, 1166–1174.

Dussert, S., Laffargue, A., Kochko, A. d., and Joët, T. (2008). Effectiveness of the fatty acid and sterol composition of seeds for the chemotaxonomy of *Coffea* subgenus *Coffea*. *Phytochemistry* **69**, 2950–2960.

Edwards, M. (2013). Whole-Genome Sequencing for Marker Discovery. *In* "Molecular Markers in Plants" (R. J. Henry, ed.). Wiley-Blackwell.

Ekblom, R., and Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* **7**, 1026–1042.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Jiaxin Qu, X. Q., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A. (2012). Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLOS one* **7**, e47768.

Faino, L., and Thomma, B. P. H. J. (2014). Get your high-quality low-cost genome sequence. *Trends in Plant Science* **9**, 288-291.

Farah, A. (2009). Coffee as a speciality and functional beverage. *In* "Functional and speciality beverage technology" (P. Paquin, ed.), pp. 370-395. Woodhead Publishing Limited, Cambridge CB21 6AH, UK.

Farah, A., and Donangelo, C. M. (2006). Phenolic compounds in coffee. *Brazilian Journal of Plant Physiology* **18**, 23-26.

Farah, A., Monteiro, M. C., Calado, V., Franca, A. S., and Trugo, L. C. (2006b). Correlation between cup quality and chemical attributes of Brazilian coffee. *Food Chemistry* **98**, 373–380.

Fernandez, D., Santos, P., Agostini, C., Bon, M., and Petito, A. (2004). Coffee (*Coffea arabica* L.) genes early expressed during infection by the rust fungus (*Hemileia vastatrix*). *Molecular Genetics and Genomics* **5**, 527–536.

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A., and Sargent, D. J. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics* **14**, 1-12.

Figueiredo, L. P., Borém, F. M., Cirillo, M. Â., Ribeiro, F. C., Giomo, G. S., and Salva, T. D. J. G. (2013). The potential for high quality bourbon coffees from different environments. *Journal of Agricultural Science* **5**, 87–98.

Flament, I. (2002). "Coffee flavour chemistry," Chichester, UK.

Franca, A. S., Mendonca, J. C. F., and Oliveira, S. D. (2005). Composition of green and roasted coffees of different cup qualities. *Food Science and Technology* **38**, 709–715.

Freitas, A. M. C., Parreira, C., and Vilas-Boas, L. (2001). The use of an electronic aroma-sensing device to assess coffee differentiation-comparison with SPME Gas Chromatography-Mass Spectrometry aroma patterns. *Journal of food composition and analysis* **14**, 513-522.

Gaitan, A., Cristancho, M. A., Gongora, C. E., Moncada, P., Posada, H., Gast, F., Yepes, M., and Aldwinckle, H. (2015). Long-Read Deep Sequencing and Assembly of the Allotetraploid *Coffea arabica* cv. Caturra and its Maternal Ancestral Diploid species *Coffea eugenioides*. *In* "Plant and Animal Genome XXIII", pp. https://pag.confex.com/pag/xxiii/webprogram/Paper17662.html, San Diego, CA.

Gao, Q., Yue, G., Li, W., Wang, J., Xu, J., and Yin, Y. (2012). Recent Progress Using High-throughput Sequencing Technologies in Plant Molecular Breeding. *Journal of integrative plant biology* **54**, 215-227.

Gartner, G. A. L., McCouch, S. R., and Moncada, M. D. P. (2013). A genetic map of an interspecific diploid pseudo testcross population of coffee. *Euphytica* **192**, 305-323.

Gautz, L. D., Smith, V. E., and Bittenbender, H. C. (2008). Measuring coffee bean moisture content. *Engineer's Notebook. College of Tropical Agriculture and Human Resources* **EN-3**.

Geleta, M., Herrera, I., Monzon, A., and Bryngelsson, T. V., Article ID 939820 (2012). Genetic diversity of arabica coffee (*Coffea arabica* L.) in Nicaragua as estimated by Simple Sequence Repeat Markers. *The Scientific World Journal* **2012**.

Geromel, C., Ferreira, L. P., Guerreiro, S. M. C., Cavalari, A. A., Pot, D., Pereira, L. F. P., Leroy, T., Vieira, L. G. E., and Marraccini, P. M. a. P. (2006). Biochemical and genomic analysis of sucrose metabolism during coffee (*Coffea arabica*) fruit development. *Journal of Experimental Botany* **57**, 3243–3258.

Ghosh, B. N., and Gacanja, W. (1970). A study of the shape and size of wet parchment coffee beans. *Journal of Agricultural Engineering Research* **15 (2)**, 91-99.

Giomo, G., Borem, F., Saath, R., Mistro, J., Figueiredo, L., Ribeiro, F., Pereira, S., and Bernardi, M. (2012). Evaluation of green bean physical characteristics and beverage quality of arabica coffee varieties in Brazil. *In* "24th International Conference on Coffee Science. 12th –16th November 2012", San José (CostaRica).

Gloess, A. N., Vietrib, A., Wielanda, F., Smrkea, S., Schönbächlera, B., Lópeza, J. A. S., Petrozzia, S., Bongersc, S., Koziorowskic, T., and Yeretzian, C. (2014). Evidence of different flavour formation dynamics by roasting coffeefrom different origins: On-line analysis with PTR-ToF-MS. *International Journal of Mass Spectrometry* **365–366** 324–337.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of nextgeneration sequencing technologies. *Nature reviews* **17**, 333-351.

Guo, M., Davis, D., and Birchler, J. A. (1996). Dosage Effects on Gene Expression in a Maize Ploidy Series. *Genetics* **142**, 1349-1355.

Guyot, R., Lefebvre-Pautigny, F., Tranchant-Dubreuil, C., Rigoreau, M., Hamon, P., Leroy, T., Hamon, S., Poncet, V., Crouzillat, D., and de Kochko, A. (2012). Ancestral synteny shared between distantly-related plant species from the asterid (*Coffea canephora* and *Solanum* Sp.) and rosid (*Vitis vinifera*) clades. *BMC Genomics* **13**, 103.

Hall, D., Tegstrom, C., and Ingvarsson, P. K. (2010). Using association mapping to dissect the genetic basis of complex traits in plants. *Briefings in functional genomics* **9**, 157-165.

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Deynze, A. V., Jong, W. S. D., Douches, D. S., and Buell, C. R. (2011). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics* **12**, 1-11.

Hamon, P., Hamon, S., Razafinarivo, N. J., Guyot, R., Siljak-Yakovlev, S., Couturon, E., Crouzillat, D., Rigoreau, M., Akaffou, S., Rakotomalala, J.-J., and Kochko, A. d. (2015). Chapter 4: *Coffea* Genome Organization and Evolution. *In* "Coffee in Health and Disease Prevention" (V. R. Preedy, ed.), pp. 29-37. Elsevier.

He, Y., Hu, R., Zhang, H., Wen, N., Cai, T., Peng, J., and Xu, Y. (2015). Characteristic aroma detection of coffee at different roasting degree based on electronic nose. *Transactions of the Chinese Society of Agricultural Engineering* **31**, 247-255.

Healey, A., Furtado, A., Cooper, T., and Henry, R. J. (2014). Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**.

Hendre, P. S., and Aggarwal, R. K. (2007). DNA markers: Development and application for genetic improvement of coffee. *In* "Genomics Assisted Crop Improvement: Vol. 2: Genomics Applications in Crops" (R. K. Varshney and R. Tuberosa, eds.), pp. 399–434. Springer.

Henry, R. J. (2011). Next-generation sequencing for understanding and accelerating crop domestication. *Briefings in functional genomics.* **11**, 51-56.

Henry, R. J., Edwards, M., Waters, D. L., Bundock, P., Sexton, T. R., Masouleh, A. K., Nock, C. J., and Pattemore, J. (2012). Application of large-scale sequencing to marker discovery in plants. *Journal of biosciences* **37**, 829-841.

Herrera, J.-C., Romero, J.-V., Camayo, G.-C., Caetano, C.-M., and Cortina, H.-A. (2012). Evidence of intergenomic relationships in triploid hybrids of coffee (*Coffea* sp.) as revealed by meiotic behavior and genomic in situ hybridization. *Tropical Plant Biology* **5**, 207-217.

Hinshaw, J. V. (2003). Solid-Phase Microextraction. *GC Connections. LC-GC Europe.* .

Holscher, H., and Steinhart, H. (1995). Aroma compounds in green coffee. *In* "Food flavours – Generation, Analysis and Process Influence" (G. Charalambous, ed.), pp. 785-803. Elsevier Science, Amsterdam.

Holscher, W., and Steinhart, H. (1992). Investigation of roasted coffee freshness with an improved headspace technique. *European Food Research and Technology* **195**, 33-38.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5.

Hulse-Kemp, A. M., Ashrafi, H., Stoffel, K., Zheng, X., Saski, C. A., Scheffler, B. E., Fang, D. D., Chen, Z. J., Deynze, A. V., and Stelly, D. M. (2015). BAC-End Sequence-Based SNP Mining in

Allotetraploid Cotton (*Gossypium*) Utilizing Resequencing Data, Phylogenetic Inferences, and Perspectives for Genetic Mapping. *G3 (Bethesda)* **5**, 1095-1105.

Hung, C.-H., Lee, C.-Y., Yang, C.-L., and Lee, M.-R. (2014). Classification and differentiation of agarwoods by using non-targeted HS-SPME-GC/MS and multivariate analysis. *Analytical Methods* **6**, 7449–7456.

Hunt, M., Newbold, C., Berriman, M., and Otto, T. D. (2014). A comprehensive evaluation of assembly scaffolding tools. *Genome Biology* **15**, R42.

Hwang, J. Y., Kim, S. H., Oh, H. R., Cho, Y.-J., Chun, J., Chung, Y. R., and Nam, D. H. (2014). Draft genome sequence of Kitasatospora cheerisanensis KCTC 2395, which produces Plecomacrolide against phytopathogenic fungi. *Genome Announcements* **2**, e00604-14.

ICO (2013). "ICO Annual Review 2011/12."

Illumina (2013). An introduction to Next-generation sequencing technology.

International-Barley-Genome-Sequencing-Consortium, Mayer, K. F. X., and Waugh, R. (2013). A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-716.

IPGRI (1996). Descriptors for Coffee (Coffea spp. and Psilanthus spp.). *International Plant Genetic Resources Institute.* , 28-29.

ISO (2004). International Standard ISO 10470: 2004. Green coffee – Defect reference chart., 15 pp.

Ito, E., Crozier , A., and Ashihara, H. (1997). Theophylline metabolism in higher plants. *Biochimica et Biophysica Acta* **1336**, 323–330.

Jaillon, O., Aury, J.-M., and Noel, B. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-468.

Jeszka-Skowron, M., Zgoła-Grzes´kowiak, A., and Grzes´kowiak, T. (2015). Analytical methods applied for the characterization and the determination of bioactive compounds in coffee. *European Food Research and Technology* **240**, 19–31.

Joët, T., Bertrand, B., and Dussert, S. (2014b). Environmental effects on coffee seed biochemical composition and quality attributes: a genomic perspective. *In* "The 25th International conference on coffee and science ASIC 2014", Colombia.

Joët, T., Laffargue, A., Descroix, F., Doulbeau, S., Bertrand, B., kochko, A. d., and Dussert, S. (2010). Influence of environmental factors, wet processing and their interactions on the biochemical composition of green Arabica coffee beans. *Food Chemistry* **118**, 693-701.

Joët, T., Laffargue, A., Salmona, J., Doulbeau, S., Descroix, F., Bertrand, B., Kochko, A. d., and Dussert, S. (2009). Metabolic pathways in tropical dicotyledonous albuminous seeds: *Coffea arabica* as a case study. *New Phytologist* **182**, 146–162.

Joët, T., Laffargue, A., Salmona, J., Doulbeau, S., Descroix, F., Bertrand, B., Lashermes, P., and Dussert, S. (2014a). Regulation of galactomannan biosynthesis in coffee seeds. *Journal of Experimental Botany* **65**, 323–337.

Joët, T., Salmona, J., Laffargue, A., Descroix, F., and Dussert, S. (2010b). Use of the growing environment as a source of variation to identify the quantitative trait transcripts and modules of co-expressed genes that determine chlorogenic acid accumulation. *Plant, cell & environment* **33**, 1220-1233.

Jones, M. R., Byers, A., Skelton, R. L., Yu, Q., Nagai, C., and Moore, P. H. (2006). Construction of an arabica coffee BAC library for molecular dissection of an allotetraploid genome. *In* "21st International Conference on Coffee Science (ASIC)", pp. 49, Montpellier, France.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., and Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24**, 1384–1395.

Kasukabe, Y., He, L., Watakabe, Y., Otani, M., Shimada, T., and Tachibana, S. (2006). Improvement of environmental stress tolerance of sweet potato by introduction of genes for spermidine synthase. *Plant Biotechnology Journal* **23**, 75–83.

Kathurima, C., Gichimu, B., Kenji, G., Muhoho, S., and Boulanger, R. (2009). Evaluation of beverage quality and green bean physical characteristics of selected Arabica coffee genotypes in Kenya. *African Journal of Food Science* **3**, 365-371.

Kato, M., and Mizuno, K. (2004). Caffeine synthase and related methyltransferases in plants. *Frontiers in Bioscience* **1**, 1833-42.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066.

Kaur, S., and Francki, M. G. J. W. F. (2012). Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant Biotechnology Journal* **10**, 125–138.

Keller, H., Wanner, H., and Baumann, T. W. (1972). Caffeine synthesis in fruits and tissue cultures of *Coffea arabica*. *PlanSa (Berl.)* **108**, 339 - 350.

Kochko, A. D., Akaffou, S., Andrade, A. C., Campa, C., Crouzillat, D., Guyot, R., Hamon, P., Ming, R., Mueller, L. A., Poncet, V., Tranchant-Dubreuil, C., and Hamon, S. (2010). Advances in Coffea Genomics *Advances in Botanical Research* **53**, 23-63.

Kochko, A. d., Crouzillat, D., Rigoreau, M., Lepelley, M., Bellanger, L., l'Anthoene, V. M., Vandecasteele, C., Guyot, R., Poncet, V., Tranchant-Dubreuil, C., Hamon, P., Hamon, S., Couturon, E., Descombes, P., Moine, D., Mueller, L., Strickler, S. R., Andrade, A., Luiz-Filipe, Marraccini, P., Giuliano, G., Fiore, A., Pietrella, M., Aprea, G., Ming, R., Wai, J., Domingues, D. S., Paschoal, A., Kuhn, G., Korlach, J., Chin, J., Sankoff, D., Zheng, C., and Albert, V. A. (2015). Dihaploid *Coffea arabica* genome sequencing and assembly. *In* "Plant and Animal Genomes XXIII", San Diego, CA.

Koek, M. M., Kloet, F. M. v. d., Kleemann, R., Kooistra, T., Verheij, E. R., and Hankemeier, T. (2011). Semi-automated non-target processing in GC 3 GC–MS metabolomics analysis: applicability for biomedical studies. *Metabolomics* **7**, 1-14.

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H., and Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* **14**, R101.

Koren, S., and Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* **23**, 110–120.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* **30**, 693-701.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* **5**.

Koshiro, Y., Zheng, X.-Q., Wang, M.-L., Nagai, C., and Ashihara, H. (2006). Changes in content and biosynthetic activity of caffeine and trigonelline during growth and ripening of *Coffea arabica* and *Coffea canephora* fruits. *Plant Science* **171**, 242-250.

Krishnan, N. M., Jain, P., Gupta, S., Hariharan, A. K., and Panda, B. (2016). An Improved Genome Assembly of *Azadirachta indica* A. Juss. *Genetics Society of America*, doi:10.1534/g3.116.030056.

Krishnan, S., Ranker, T. A., Davis , A. P., and Rakotomalala, J. J. (2013). An assessment of the genetic integrity of ex situ germplasm collections of three endangered species of *Coffea* from Madagascar: implications for the management of field germplasm collections. *Genetics resources and crop evolution journal* **60**, 1021–1036.

Krug, C. A., and Mendes, A. J. T. (1940). Cytological observations in *Coffea*. *Journal of Genetics* **39**, 189-203.

Kumar, S., Banks, T. W., and Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. *International journal of plant genomics* **2012**.

Ky, C.-L., Barre, P., Lorieux, M., Trouslot, P., Akaffou, S., Louarn, J., Charrier, A., Hamon, S., and Noirot, M. (2000). Interspecific genetic linkage map, segregation distortion and genetic conversion in coffee (Coffea sp.). *Theoretical and Applied Genetics* **101**, 669-676.

Ky, C.-L., Barre, P., and Noirot, M. (2013). Genetic investigations on the caffeine and chlorogenic acid relationship in an interspecific cross between *Coffea liberica dewevre*i and *C. pseudozanguebariae*. *Tree Genetics & Genomes* **9**, 1043-1049.

Ky, C.-L., Guyot, B., Louarn, J., Hamon, S., and Noirot, M. (2001). Trigonelline inheritance in the interspecific *Coffea pseudozanguebariae* × *C. liberica var. dewevrei* cross. *Theoretical and Applied Genetics* **102**, 630–634.

Ky, C.-L., Louarn, J., Dussert, S., Guyot, B., Hamon, H., and Noirot, M. (2001b). Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild *Coffea arabica* L. and *C. canephora* P. accessions. *Food Chemistry* **75**, 223–230.

Lashermes, P., Andrzejewski, S., Bertrand, B., Combes, M. C., Dussert, S., Graziosi, G., Trouslot, P., and Anthony, F. (2000a). Molecular analysis of introgressive breeding in coffee (*Coffea arabica* L.). *Theor Appl Genet* **100**, 139–146

Lashermes, P., Combes, M.-C., Hueber, Y., Severac, D., and Dereeper, A. (2014). Genome rearrangements derived from homoeologous recombination following allopolyploidy speciation in coffee. *The Plant Journal* **78**, 674–685.

Lashermes, P., Combes, M.-C., Robert, J., Trouslot, P., D'Hont, A., Anthony, F., and Charrier, A. (1999). Molecular characterisation and origin of the *Coffea arabica* L. genome. *Molecular Genetics and Genomics* **261**, 259-266.

Lashermes, P., Combes, M. C., Prakash, N. S., Trouslot, P., Lorieux, M., and Charrier, A. (2001). Genetic linkage map of *Coffea canephora*: effect of segregation distortion and analysis of recombination rate in male and female meioses. *Genome* **44**, 589-595.

Lashermes, P., Combes, M. C., Trouslot, P., and Charrier, A. (1997). Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. *Theoretical and Applied Genetics* **94**, 947-955.

Lashermes, P., Trouslot, P., Combes, M. C., Couturon, E., and Charrier, A. (2000b). Brief communication. Single-locus inheritance in the allotetraploid *Coffea arabica* L. and interspecific hybrid *C. arabica* x *C. canephora*. *Journal of Heredity* **91**, 81-85.

Lawless, H. T., and Heymann, H. (2010). Chapter 6: Measurement of Sensory Thresholds. *In* "Sensory Evaluation of Food" Principles and Practices" (H. T. Lawless and H. Heymann, eds.), pp. 127. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA.

Lee, H., Golicz, A. A., Philipp E. Bayer, Yuannian Jiao, Haibao Tang, Andrew H. Paterson, Gaurav Sablok, Rahul R. Krishnaraj, Chon-Kit Kenneth Chan, Jacqueline Batley, Gary A. Kendrick, Anthony W.D. Larkum, Ralph, P. J., and Edwards, D. (2016). The genome of a southern hemisphere seagrass species (*Zostera muelleri*). *Plant Physiology Preview*, DOI:10.1104/pp.16.00868.

Lee, J. C.-Y., Tsoi, A., Kornfeld, G. D., and Dawes, I. W. (2013). Cellular responses to L-serine in Saccharomyces cerevisiae: roles of general amino acid control, compartmentalization, and aspartate synthesis. *FEMS Yeast Research* **13**, 618–634.

Lefebvre-Pautigny, F., Wu, F., Philippot, M., Rigoreau, M., Priyono, Zouine, M., Frasse, P., Bouzayen, M., Broun, P., Pétiard, V., Tanksley, S. D., and Crouzillat, D. (2010). High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. *Tree Genetics & Genomes* **6**, 565-577.

Lepelley, M., Mahesh, V., McCarthy, J., Rigoreau, M., Crouzillat, D., Chabrillange, N., de Kochko, A., and Campa, C. (2012). Characterization, high-resolution mapping and differential expression of three homologous PAL genes in *Coffea canephora* Pierre (Rubiaceae). *Planta* **236**, 313-26.

Leroy, T., Bellis, F., Legnate, H., and Kananura, E. (2011). Improving the quality of African robustas: QTLs for yield- and quality-related traits in *Coffea canephora*. *Tree Genetics & Genomes* **7**, 781-798.

Leroy, T., Marraccini, P., Dufour, M., Montagnon, C., Lashermes, P., Sabau, X., Ferreira, L. P., Jourdan, I., Pot, D., Andrade, A. C., Glaszmann, J. C., Vieira, L. G., and Piffanelli, P. (2005). Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes. *Theoretical and Applied Genetics* **111**, 1032-41.

Leroy, T., Ribeyre, F., Bertrand, B., Charmetant, P., Dufour, M., Montagnon, C., Marraccini, P., and Pot, D. (2006). Genetics of coffee quality. *Brazilian Journal of Plant Physiology* **18**, 229-242.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Li, Q., Li, Y., Song, J., Xu, H., Xu, J., Zhu, Y., Li, X., Gao, H., Dong, L., Qian, J., Sun, C., and Chen, S. (2014). High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. *New Phytologist* **204**, 1041–1049.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C. C., Lam, T. T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., et al. (2010a). The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010b). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**, 265–272.

Li, S., Jin, H., and Zhang, Q. (2016). The Effect of Exogenous Spermidine Concentration on Polyamine Metabolism and Salt Tolerance in Zoysia grass(*Zoysia japonica* Steud) Subjected to Short-Term Salinity Stress. *Frontiers in Plant Science* **7**, Article 1221.

Lin, C., Mueller, L. A., Carthy, J. M., Crouzillat, D., Pétiard, V., and Tanksley, S. (2005). Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. *Theoretical and Applied Genetics* **112**, 114-130.

Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., and Shyr, Y. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* **13**, 1-8.

Lombello, R. A., and Pinto-Maglio, C. A. F. (2003). Cytogenetic studies in *Psilanthus ebracteolatus* Hiern., a wild diploid coffee species. *Cytologia* **68**, 425–429.

López-Gartner, G., Cortina, H., McCouch, S. R., and Moncada, M. D. P. (2009). Analysis of genetic structure in a sample of coffee (Coffea arabica L.) using fluorescent SSR markers. *Tree Genetics & Genomes* **5**, 435-446.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., and Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 1-6.

Lv, S., Wu, Y., Zhou, J., Lian, M., Li, C., Xu, Y., Liu, S., Wang, C., and Meng, Q. (2014). The study of fingerprint characteristics of Dayi Pu-Erh tea using a fully automatic HS-SPME/GC–MS and combined chemometrics method. *PLOS one*.

Madoui, M.-A., d'Agata, C. D., Léo , Oeveren, J. v., Vossen, E. v. d., and Aury, J.-M. (2016). MaGuS: a tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling™ Data. *BMC Bioinformatics* **17**, 1-9.

Magwene, P. M., Willis, J. H., and Kelly, J. K. (2011). The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. *PLoS Computational Biology* **7**, e1002255.

Mahe, L., Combes, M. C., and Lashermes, P. (2007). Comparison between a coffee single copy chromosomal region and Arabidopsis duplicated counterparts evidenced high level synteny between the coffee genome and the ancestral Arabidopsis genome. *Plant Molecular Biology* **64**, 699-711.

Mahesh, V., Rakotomalala, J. J., Le Gal, L., Vigne, H., De Kochko, A., Hamon, S., Noirot, M., and Campa, C. (2006). Isolation and genetic mapping of a Coffea canephora phenylalanine ammonia-lyase gene (CcPAL1) and its involvement in the accumulation of caffeoyl quinic acids. *Plant cell reports* **25**, 986-992.

Maluf, M. P., da Silva, C. C., de Oliveira, M. D. A., Tavares, A. G., Silvarolla, M. B., and Guerreiro, O. (2009). Altered expression of the caffeine synthase gene in a naturally caffeine-free mutant of *Coffea arabica*. *Genetics and Molecular Biology* **32**, 802-810.

Maluf, M. P., Silvestrini, M., Ruggiero, L. M. d. C., Filho, O. G., and Colombo, C. A. (2005). Genetic diversity of cultivated coffea arabica inbred lines assessed by RAPD, AFLP and SSR marker systems. *Sci. Agric. (Piracicaba, Braz.)* **62**, 366-373.

Margarido, G. R. A., and Heckerman, D. (2015). ConPADE: Genome assembly ploidy estimation from next-generation sequencing data. *PLOS Computational Biology* **11**, e1004229.

Martens, H., Karstang, T., and Næs, T. (1987). Improved selectivity in spectroscopy by multivariate calibration *Journal of Chemometrics* **1**, 201-219

Martín, M. a. J., Pablos, F., and González, A. G. (1998). Discrimination between arabica and robusta green coffee varieties according to their chemical composition. *Talanta* **46**, 1259-1264.

Martınez-Garcıa, P. J., Crepeau, M. W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J., Stevens, K. A., Paul, R., Butterfield, T. S., Britton, M. T., Reagan, R. L., Chakraborty, S., Walawage, S. L., Vasquez-Gross, H. A., Cardeno, C., Famula, R. A., Pratt, K., Kuruganti, S., Aradhya, M. K., Leslie, C. A., Dandekar, A. M., Salzberg, S. L., Wegrzyn, J. L., Langley, C. H., and Neale, D. B. (2016). The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *The Plant Journal*, doi: 10.1111/tpj.13207.

Masumbuko, L. I., Bryngelsson, T., Mneney, E. E., and Salomon, B. (2003). Genetic diversity in Tanzanian Arabica coffee using Random Amplified Polymorphic DNA (RAPD) markers. *Hereditas* **139**, 56–63.

Maurin, O., Davis, A. P., Chester, M., Mvungi, E. F., Jaufeerally-Fakim, Y., and Fay, M. F. (2007). Towards a phylogeny for coffea (rubiaceae): Identifying well-supported lineages based on nuclear and plastid DNA sequences. *Annals of Botany* **100**, 1565–1583.

Mavromatis, K., Land, M. L., Brettin, T. S., Quest, D. J., Copeland, A., Clum, A., Goodwin, L., Woyke, T., Lapidus, A., Klenk, H. P., Cottingham, R. W., and Kyrpides, N. C. (2012). The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. *PLOS one* **7**, e48837.

Mayer, F., Czerny, M., and Grosch, W. (1999). Influence of provenance and roast degree on the composition of potent odorants in Arabica coffees. *European Food Research and Technology* **209**, 242–250.

Mazzafera, P. (1991). Trigonelline in coffee. *Phytochemistry* **30**, 2309 -2310.

Mazzafera, P., and Carvalho, A. (1992). Breeding for low seed caffeine content of coffee *(Coffea* L.) by interspecific hybridization. *Euphytica* **59**, 55–60.

Mazzafera, P., and Silvarolla, M. B. (2010). Caffeine content variation in single green Arabica coffee seeds. *Seed Science Research* **20**, 163-167.

Mehari, B., Redi-Abshiro, M., Chandravanshi, B. S., Atlabachew, M., Combrinck, S., and McCrindle, R. (2016). Simultaneous determination of alkaloids in green coffee beans from Ethiopia: chemometric evaluation of geographical origin. *Food Analytical Methods* **9**, 1627–1637.

Mérot-L'Anthoëne, V., Mangin, B., Lefebvre-Pautigny, F., Jasson, S., Rigoreau, M., Husson, J., Lambot, C., and Crouzillat, D. (2014). Comparison of three QTL detection models on biochemical, sensory, and yield characters in *Coffea canephora*. *Tree Genetics & Genomes*.

Mestdagh, F., Wocheslander, S., Rodriguez, A., Egli, A., Davidek, T., and Blank, I. (2014). Conversion of green coffee precursors into flavour during roasting of arabica and robusta coffees. *In* "The 25th International conference on coffee and science ASIC 2014", Colombia.

Michael, T. P., and VanBuren, R. (2015). Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology* **21**, 71–81.

Mishra, M. K., Sandhyarani, N., Suresh, N., Kumar, S. S., Soumya, P. R., Yashodha, M. H., Bhat, A., and Jayarama (2012b). Genetic diversity among Indian coffee cultivars determined via molecular markers. *Journal of Crop Improvement* **26**, 727–750.

Mishra, M. K., Tornincasa, P., Nardi, B. D., Asquini, E., Dreos, R., Terra, L. D., Rathinavelu, R., Rovelli, P., Pallavicini, A., and Graziosi, G. (2011). Genome Organization in Coffee as Revealed by EST PCRRFLP, SNPs and SSR Analysis. *Journal of crop science and biotechnology* **14**, 25-37.

Mizuno, K., Matsuzaki, M., Kanazawa, S., Tokiwano, T., Yoshizawa, Y., and Kato, M. (2014). Conversion of nicotinic acid to trigonelline is catalyzed by N-methyltransferase belonged to motif B' methyltransferase family in *Coffea arabica*. *Biochemical and Biophysical Research Communications* **452**, 1060–1066.

Mizuno, K., Okuda, A., Kato, M., Yoneyama, N., Tanaka, H., Ashihara, H., and Fujimura, T. (2003b). Isolation of a new dual-functional caffeine synthase gene encoding an enzyme for the conversion of 7-methylxanthine to caffeine from coffee (*Coffea arabica* L.). *FEBS letters* **534**, 75-81.

Moncada, P., Tovar, E., Montoya, J. C., Gonzalez, A., Spindel, J., and McCouch, S. (2014). A genetic of linkage map of coffee (*Coffea arabica* L) and QTL for yield, plant height and bean size. *In* "The 25th International conference on coffee and science ASIC 2014", Colombia.

Mondego, J. M., Vidal, R. O., Carazzolle, M. F., and Tokuda, E. K. (2011). An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. *BMC Plant Biology* **11**, 30.

Montagnon, C., and Bouharmont, P. (1996). Multivariate analysis of phenotypic diversity of *Coffea arabica*. *Genetic Resources and Crop Evolution* **43**, 221-227.

Montagnon, C., Guyot, B., Cilas, C., and Leroy, T. (1998). Genetic parameters of several biochemical compounds from green coffee, *Coffea canephora*. *Plant Breeding* **117**, 576-578.

Montagnon, C., Marraccini, P., and Bertrand, B. (2012). Breeding for coffee quality. *In* "Specialty Coffee: Managing Quality. International Plant Nutrition Institute", pp. 93-122. Southeast Asia Program

Montoya, G., Vuong, H., Cristancho, M., Moncada, P., and Yepes, M. (2006). Sequence analysis from leaves, flowers and fruits of *Coffea arabica* var. Caturra. *In* "21st International Conference on Coffee Science (ASIC)", Montpellier, France.

Morgante, M., Scalabrin , S., Scaglione , D., Cattonaro, F., Magni, F., Jurman, I., Cerutti, M., Liverani, F. S., Navarini, L., Terra, L. D., Pellegrino, G., Graziosi, G., Vitulo, N., and Valle, G. (2015). Conference proceedings: Progress report on the sequencing and assembly of the allotetraploid *Coffea arabica* var. Bourbon genome. *Conference of Plant and Animal Genome XXIII. January 10 - 14, 2015. San Diego, CA, USA.*

Moschetto, D., Montagnon, C., Guyot, B., Perriot, J. J., Leroy, T., and Eskes, A. (1996). Studies on the effect of genotype on cup quality of *Coffea canephora*. *Tropical Science* **36**, 18- 31.

Motta, L. B., Soares, T. C. B., Ferrão, M. A. G., Caixeta, E. T., Lorenzoni, R. M., and Neto, J. D. d. S. (2014). Molecular characterization of arabica and conilon coffee plants genotypes by SSR and ISSR markers. *Brazilian archives of biology and technology* **57**, 728-735.

Mueller, L., Strickler, S., Somingues, D., Pereira, L., Andrade, A., Marraccini, P., Ming, R., Wai, J., Albert, V., Giuliano, G., Descombes, P., Moine, D., Guyot, R., Poncet, V., Hamon, P., Hamon, S., Tranchant, C., De kochko, A., Lepelley, M., Rigoreau, M., and Crouzillat, D. (2014). Towards a better understanding of the *Coffea arabica* genome structure. *In* "The 25th International conference on coffee and science ASIC 2014", Colombia.

Muschler, R. G. (2001). Shade improves coffee quality in a sub-optimal coffee-zone of Costa Rica. *Agroforestry systems* **51**, 131-139.

Nagai, C., Jones, M. R., Byers, A. E., Adamski, D. J., and Ming, R. (2007). Development and characterization of a true F2 population for genetic and QTL mapping in Arabica. *In* "21st International Conference on Coffee Science ", pp. 771-777 Montpellier, France.

Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. *Nature reviews* **14**, 157-167.

Nagy, I., Barth, S., Mehenni-Ciz, J., Abberton, M. T., and Milbourne, D. (2013). A hybrid next generation transcript sequencing-based approach to identify allelic and homeolog-specific single nucleotide polymorphisms in allotetraploid white clover. *BMC Genomics* **14**, 1-19.

Neale, D. B., and Savolainen, O. (2004). Association genetics of complex traits in conifers. *Trends in Plant Science* **9**, 325-330.

Nebesny, E., Budryn, G., Kula, J., and Majda, T. (2007). The effect of roasting method on headspace composition of robusta coffee bean aroma. *European Food Research and Technology* **225**, 9-19.

Niessen, W. M. A. (2001). Principles and Intrument of Gas Chromatography - Mass Spectrometry. *In* "Current Practice of Gas Chromatography - Mass Spectrometry" (W. M. A. Niessen, ed.). Marcel Dekker, Inc., New York.

Noir, S., Patheyron, S., Combes, M.-C., Lashermes, P., and Chalhoub, B. (2004). Construction and characterisation of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.). *Theoretical and Applied Genetics* **109**, 225-230.

Noirot, M., Poncet, V., Barre, P., Hamon, P., Hamon, S., and Kochko, A. D. (2003). Genome size variations in diploid African *Coffea* species. *Annals of Botany* **92**, 709-714.

Nordborg, M., and Weigel, D. (2008). Next-generation genetics in plants. *Nature* **456**, doi:10.1038/nature07629.

Nuhu, A. A. (2014). Review article: Bioactive micronutrients in coffee: Recent analytical approaches for characterization and quantification. *ISRN Nutrition* **2014**, 1-13.

Oestreich-Janzen, S. (2010). Chemistry of coffee. *In* "Comprehensive Natural Products II: Chemistry & Biology. Volume 3: Development & Modification of Bioactivity" (L. Mander and H.-W. Liu, eds.), pp. 1085–1117. CAFEA GmbH, Hamburg, Germany.

Ogawa, M., Herai, Y., Koizumi, N., Kusano, T., and Sano, H. (2001). 7-Methylxanthine Methyltransferase of coffee plants: gene isolation and enzymatic properties. *Journal of Biological Chemistry* **276**, 8213-8218.

Ogita, S., Uefuji, H., Morimoto, M., and Sano, H. (2004). Application of RNAi to confirm theobromine as the major intermediate for caffeine biosynthesis in coffee plants with potential for construction of decaffeinated varieties. *Plant molecular biology* **54**, 931-941.

Ogita, S., Uefuji, H., Yamaguchi, Y., Koizumi, N., and Sano, H. (2003). Producing decaffeinated coffee plants. *Nature*, 423.

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow, J. B., Salit, M. L., and Zook, J. M. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics* **6**, doi: 10.3389/fgene.2015.00235.

Olukunle, O. J., and Akinnuli, B. O. (2012). Investigating some engineering properties of coffee seeds and beans. *Journal of Emerging Trends in Engineering and Applied Sciences* **3 (5)**, 743-747.

Omondi, C., Gichimu, B., Cheserek, J., and Gimase, J. (2016). Leveraging on germplasm acquisition for Arabica coffee improvement in Kenya *Journal of Agricultural and Crop Research* **4**, 9-16.

Paillard, M., Lashermes, P., and Pétiard, V. (1996). Construction of a molecular linkage map in coffee. *Theoretical and Applied Genetics* **93**, 41-47.

Palmer, J. D. (1985). Review articles: Chloroplast DNA and molecular phylogeny. *BioEssays* **2**, 263-267.

Parastara, H., Jalali-Heravi, M., Sereshtic, H., and Mani-Varnosfaderani, A. (2012). Chromatographic fingerprint analysis of secondary metabolites in citrus fruits peels using gas chromatography–mass spectrometry combined with advanced chemometric methods. *Journal of Chromatography A* **1251**, 176– 187.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Genome analysis* **23**, 1061–1067.

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes *Nucleic Acids Research* **37**, 289–297.

Paszkiewicz, K., and Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in bioinformatics* **II**, 457 - 472.

Patel, S., Swaminathan, P., Fennell, A., and Zeng, E. (2015). De novo genome assembly tool comparison for highly heterozygous species *Vitis vinifera* cv. Sultanina. *IEEE International Conference on Bioinforrnatics and Biomedicine (BIBM). Nov 9 - 12, 2015. Washington D.C., USA*, 1771-1773.

Payne, R., Harding, S. A., Murray, D. A., Soutar, D. M., Baird, D. B., Glaser, A. I., Channing, I. C., Welham, S. J., Gilmour, A. R., Thompson, R., and Webster, R. (2008). GENSTAT release 11 reference manual. Parts 1, 2 and 3. *VSN International: Hemel Hempstead, UK*.

Peace, C., Bassil, N., Main, D., Ficklin, S., Rosyara, U. R., Stegmeir, T., Sebolt, A., Gilmore, B., Lawley, C., Mockler, T. C., Bryant, D. W., Wilhelm, L., and Iezzoni, A. (2012). Development and Evaluation of a Genome-Wide 6K SNP Array for Diploid Sweet Cherry and Tetraploid Sour Cherry. *PLOS one* **7**, e48305.

Pearl, H., Nagai, C., Moore, P., Steiger, D., Osgood, R., and Ming, R. (2004). Construction of a genetic map for arabica coffee. *Theoretical and Applied Genetics* **108**, 829-835.

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., Guo, Y., Paxinos, E. E., Korbel, J. O., Darnell, R. B., McCombie, W. R., Kwok, P.-Y., Mason, C. E., Schadt, E. E., and Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods* **12**, 780–786.

Perrois, C., Strickler, S., G, M., M, L., L, B., S, M., J, H., L, M., and I., P. (2014). Differential regulation of caffeine metabolism in *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta). *Planta*.

Perrois, C., Strickler, S. R., Mathieu, G., Lepelley, M., Bedon, L., Michaux, S. p., Husson, J., Mueller, L., and Privat, I. (2015). Differential regulation of caffeine metabolism in *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta). *Planta* **241**, 179–191.

Piccino, S., Boulanger , R., Descroix, F., and Sing, A. S. C. (2014). Aromatic composition and potent odorants of the "specialty coffee" brew "Bourbon Pointu" correlated to its three trade classifications. *Food Research International* **61**, 264–271.

Pickard, S., Becker, I., Merz, K.-H., and Richling, E. (2013). Determination of the alkylpyrazine composition of coffee using stable isotope dilution-gas chromatography-mass spectrometry (SIDA-GC-MS). *Journal of Agricultural and Food Chemistry* **61**, 6274-81.

Pierce, K. M., Schale, S. P., Le, T. M., and Larson, J. C. (2011). An advanced analytical chemistry experiment using Gas Chromatography Mass Spectrometry, Matlab, and chemometrics to predict biodiesel blend percent composition. *Journal of Chemical Education* **88**, 806–810.

Poncet, V., Rondeau, M., Tranchant, C., Cayrel, A., Hamon, S., de Kochko, A., and Hamon, P. (2006). SSR mining in coffee tree EST databases: potential use of EST–SSRs as markers for the *Coffea* genus. *Molecular Genetics and Genomics* 436–449.

Pot, D., Bouchet, S., Marraccini, P., De bellis, F., Cubry, P., Jourdan, I., Pereira, L. F. P., Vieira, L. G. E., Ferreira, L. P., Musoli, P., Legnate, H., and Leroy, T. (2007). Nucleotide diversity of genes involved in sucrose metabolism. Towards the identification of candidates genes controlling sucrose variability in *Coffea* sp. *In* "International conference on coffee science - ASIC", Montpellier, France.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909.

Priolli, R. H. G., Ramos, L. C. S., Pot, D., and Moller , M. (2009). Construction of a genetic map based on an interspecific F2 population between *Coffea arabica* and *Coffea canephora* and its usefulness for quality related traits. *In* "22nd International Conference on Coffee Science, ASIC 2008", pp. 882-890, Campinas, SP, Brazil.

Priolli, R. H. M., Paulo, Siqueira, W. J., Möller, M., Zucchi, M. I., Ramos, L. C. S., Gallo, P. B., and Colombo, C. A. (2008). Caffeine inheritance in interspecific hybrids of *Coffea arabica* x *Coffea canephora* (Gentianales, Rubiaceae). *Genetics and Molecular Biology* **31**, 498-504.

Privat, I., Bardil, A., Gomez, A. B., and Severac, D. (2011). The 'PUCE CAFE' Project: the first 15K coffee microarray, a new tool for discovering candidate genes correlated to agronomic and quality traits. *BMC Genomics* **12**, 5.

Privat, I., Foucrier, S., Prins, A., Epalle, T., Eychenne, M., Kandalaft, L., Caillet, V., Lin, C., Tanksley, S., and Foyer, C. (2008). Differential regulation of grain sucrose accumulation and metabolism in *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta) revealed through gene expression and enzyme activity analysis. *New Phytologist* **178**, 781-797.

Privat, I., Perrois, C., Mathieu, G., Strickler, S. R., Lepelley, M., Bedon, L., Michaux, S., Husson, J., and Mueller, L. (2014). Differential regulation of caffeine metabolism in *Coffea arabica* and *Coffea canephora*. *In* "The 25th International conference on coffee and science ASIC 2014", Colombia.

Pryszcz, L. P., and Gabaldon, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, doi: 10.1093/nar/gkw294.

Quinlan, A. R., Stewart, D. A., Stromberg, M. P., and Marth, G. T. (2007). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature methods* **5**, 179-181.

Rahman, A. Y. A., Usharraj, A. O., Misra, B. B., Thottathil, G. P., Jayasekaran, K., Feng, Y., Hou, S., Ong, S. Y., Ng, F. L., Lee, L. S., Tan, H. S., Sakaff, M. K. L. M., Teh, B. S., Khoo, B. F., Badai, S. S., Aziz, N. A., Yuryev, A., Knudsen, B., Dionne-Laporte, A., Mchunu, N. P., Yu, Q., Langston, B. J., Freitas, T. A. K., Young, A. G., Chen, R., Wang, L., Najimudin, N., Saito, J. A., and Alam, M. (2013). Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genomics* **14**, 1-15.

Razafinarivo, N. J., Rakotomalala, J.-J., Brown, S. C., Bourge, M., Hamon, S., Kochko, A., Poncet, V., Dubreuil-Tranchant, C., Couturon, E., Guyot, R., and Hamon, P. (2012). Geographical gradients in the genome size variation of wild coffee trees (Coffea) native to Africa and Indian Ocean islands. *Tree Genetics & Genomes* **8**, 1345-1358.

Redwan, R. M., Saidin, A., and Kumar, S. V. (2016). The draft genome of MD-2 pineapple using hybrid error correction of long reads. *DNA Research* **0**, 1-13.

Resende , M., Caixeta, E., Alkimin , E. R., Sousa, T. V., Resende, M. D. V., Chamala, S., and Neves, L. G. (2016). High-throughput targeted genotyping of *Coffea arabica* and *Coffea canephora* using next generation sequencing. *In* "Plant and Animal Genome XXIV", San Diego, CA.

Ribeiro, J. S., Augusto, F., Salva, T. J. G., and Ferreira, M. M. C. (2012). Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics. *Talanta* **101**, 253-260.

Ribeiro, J. S., Teófilo, R. F., Augusto, F., and Ferreira, M. M. C. (2010). Simultaneous optimization of the microextraction of coffee volatiles using response surface methodology and principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **102**, 45-52.

Ries, D., Holtgräwe, D., Viehöver, P., and Weisshaar, B. (2016). Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. *BMC Genomics* **17**, 1-13.

Riu-Aumatell, M., Castellari, M., Lopez-Tamames, E., Galassi, S., and Buxaderas, S. (2004). Characterisation of volatile compounds of fruit juices and nectars by HS/SPME and GC/MS. *Food Chemistry* **87**, 627–637

Rocha, Sılvia, Maeztu, L., Barros, A., Cid, C., and Coimbra, M. A. (2003). Screening and distinction of coffee brews based on Headspace Solid Phase Microextraction/Gas Chromatography/principal component analysis. *Journal of the Science of Food and Agriculture* **84**, 43–51.

Rodrigues, C. I., Maia, R., Miranda, M., Ribeirinho, M., Nogueira, J. M. F., and Maguas, C. (2009). Stable isotope analysis for green coffee bean: a possible method for geographic origin discrimination *Journal of Food Composition and analysis* **22**, 463-471.

Rodrigues, M. A. A., Borges, M. L. A., Franca, A. S., Oliveira, L. S., and Corrêa, P. C. (2003). Evaluation of physical properties of coffee during roasting *Agricultural Engineering International* **V**.

Rogalski, M., Vieira, L. d. N., P.Fraga, H., and P.Guerra, M. (2015). Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. *Frontiers in Plant Science* **6**.

Rojas, J. (2004). Part IV: Storage, Shipment, Quality: Green Coffee Storage. *In* "Coffee: Growing, Processing, Sustainable Production - A Guidebook for Growers, Processors, Traders, and Researchers" (J. N. Wintgens, ed.). WILEY-VCH Verlag GmbH & Co. KCaA.

Roychoudhury, A., Basu, S., and Sengupta, D. N. (2011). Amelioration of salinity stress by exogenously applied spermidine or spermine in three varieties of indica rice differing in their level of salt tolerance. *Journal of Plant Physiology* **168**, 317–328.

Salmona, J., Dussert, S., Descroix, F., de Kochko, A., Bertrand, B., and Joet, T. (2008). Deciphering transcriptional networks that govern *Coffea arabica* seed development using combined cDNA array and real-time RT-PCR approaches. *Plant Molecular Biology* **66**, 105-124.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marcxais, G., Pop, M., and Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* **22**, 557–567.

Samson, N., Bausher, M. G., Lee, S.-B., Jansen, R. K., and Daniell, H. (2007). The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. *Plant Biotechnology Journal* **5**, 339–353.

Sayadi, A., Immonen, E., Helen Bayram, and Arnqvist, G. (2016). The De Novo transcriptome and its functional annotation in the seed beetle *Callosobruchus maculatus*. *PLOS one*, DOI:10.1371/journal.pone.0158565.

Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research* **20**, 1165–1173.

Schatz, M. C., Witkowski, J., and McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. *Genome Biology* **13**, 1-7.

Schlabach, M. (2013). Non-target screening – A powerful tool for selecting environmental pollutants. *Miljo-direktoratet rapport* **M-27**.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature reviews* **15**, 749-763.

Schmutzer, T., Samans, B., Dyrszka, E., Ulpinnis, C., Weise, S., Stengel, D., Colmsee, C., Lespinasse, D., Micic, Z., Abel, S., Duchscherer, P., Breuer, F., Abbadi, A., Leckband, G., Snowdon, R., and Scholz, U. (2015). Species-wide genome sequence and nucleotide polymorphisms from the model allopolyploid plant *Brassica napus*. *Scientific data* **2:150072**.

Semmelroch, P., and Grosch, W. (1996). Studies on character impact odorants of coffee brews. *Journal of Agricultural and Food Chemistry* **44**, 537-543.

Semmelroch, P., Laskawy, G., Blank, I., and Groscht, W. (1995). Determination of potent odourants in roasted coffee by stable isotope dilution assays. *Flavour and Fragrance Journal* **10**, 1-7.

Sexton, T. R., Henry, R. J., Harwood, C. E., Thomas, D. S., McManus, L. J., Raymond, C., Henson, M., and Shepherd, M. (2012). Pectin Methylesterase genes influence solid wood properties of Eucalyptus pilularis. *Plant Physiology* **158**, 531–541.

Siebert, T. E., Smyth, H. E., Capone, D. L., Neuwohner, C., Pardon, K. H., Skouroumounis, G. K., Herderich, M. J., Sefton, M. A., and Pollnitz, A. P. (2005). Stable isotope dilution analysis of wine fermentation products by HS-SPME-GC-MS. *Analytical and Bioanalytical Chemistry* **381**, 937–947.

Silvarolla, M. B., Mazzafera, P., and Fazuoli, L. C. (2004). A naturally decaffeinated arabica coffee. *Nature* **429**, 826.

Silvarolla, M. B., Mazzafera, P., and Lima, M. M. A. d. (2000). Caffeine content of Ethiopian *Coffea* arabica beans. *Genetics and Molecular Biology* **23**, 213-215.

Silvestrini, M., Junqueira, M. G., Favarin, A. C., Guerreiro-Filho, O., Maluf, M. P., Silvarolla, M. B., and Colombo, C. A. (2007). Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. *Genet Resour Crop Evol* **54**, 1367–1379.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Genome analysis*, doi: 10.1093/bioinformatics/btv351.

Simkin, A. J., Qian, T., Caillet, V., Michoux, F., Ben Amor, M., Lin, C., Tanksley, S., and McCarthy, J. (2006). Oleosin gene family of *Coffea canephora*: Quantitative expression analysis of five oleosin genes in developing and germinating coffee grain. *Journal of plant physiology* **163**, 691-708.

Simpson, J. T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**, 549–556.

Simpson, J. T., and Pop, M. (2015). The theory and practice of genome sequence assembly. *Annual Review of Genomics and Human Genetics* **16**, 153–72.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol2, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research* **19**, 1117–1123.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews* **15**, 121-131.

Smit, A., and Hubley, R. (2008-2015). RepeatModeler Open-1.0. *http://www.repeatmasker.org* **Accessed on Jan 16, 2016**.

Smit, A. F., Hubley, R., and Green, P. (1996–2010). RepeatMasker 3.0 *repeatmasker.org* [online]. *http://www.repeatmasker.org/webrepeatmaskerhelp.html* **Accessed on Nov 19, 2015**.

Spencer, M., Sage, E., Velez, M., and Guinard, J.-X. (2016). Using Single Free Sorting and Multivariate Exploratory Methods to Design a New Coffee Taster’s Flavor Wheel. *Journal of Food Science* **81**, S2997-S3005.

Sridevi, V., and Giridhar, P. (2013). Influence of altitude variation on trigonelline content during ontogeny of *Coffea canephora* fruit. *Journal of Food Studies* **2**.

Sridevi, V., and Giridhar, P. (2014). Changes in caffeine content during fruit development in *Coffea canephora* P. ex. Fr. grown at different elevations. *Journal of Biology and Earth Sciences* **4** B168-B175.

Stadermann, K. B., Weisshaar, B., and Holtgräwe, D. (2015). SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome. *BMC Bioinformatics* **16**, 1-10.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**, W465–W467.

Steiger , D. L., Nagai, C., Morden, P. H. M. C. W., Osgood, R. V., and Ming, R. (2002). AFLP analysis of genetic diversity within and among *Coffea arabica* cultivars. *Theor Appl Genet* **105**, 209–215.

Stoffelen, P., Noirot, M., Couturon, E., Bontems, S., De Block, P., and Anthony, F. (2009). *Coffea anthonyi*, a New Self-Compatible Central African Coffee Species, Closely Related to an Ancestor of *Coffea arabica*. *Taxon* **58** 133-140

Strickler, S. R. (2015). Genome assembly strategies of the recent polyploid, *Coffea arabica*. *In* "Plant and Animal Genome XXIII. January 10 - 14, 2015. San Diego, CA, USA". https://pag.confex.com/pag/xxiii/webprogram/Paper17695.html.

Subedi, R. N. (2011). Comparative analysis of dry and wet processing of coffee with respect to quality and cost in kavre district, nepal: A case of Panchkhal village. *International Research Journal of Applied and Basic Sciences. Vol., 2 (5), 181-193, 2011* **2**, 181-193.

Sun, Y., Wang, J., Crouch, J. H., and Xu, Y. (2010). Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. *Molecular Breeding* **26**, 493–511.

Sunarharum, W. (2016). The compositional basis of coffee flavour, PhD thesis at The University of Queensland.

Sunarharum, W. B., Williams, D. J., and Smyth, H. E. (2014). Review: Complexity of coffee flavor: A compositional and sensory perspective. *Food Research International* **62**, 315–325.

Swofford, D. L. (2002). PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0 b10. *Sunderland, MA (USA): Sinauer Associates, Inc*

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S., and Terauchi, R. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal* **74**, 174–183.

Taveira, J. H. d. S., Borém, F. M., Figueiredo, L. P., Reis, N., Franca, A. S., Harding, S. A., and Tsai, C.-J. (2014). Potential markers of coffee genotypes grown in different Brazilian regions: A metabolomics approach. *Food Research International* **61**, 75–82.

Teixeira-Cabral, T. A., Sakiyama, N. S., Zambolim, L., Pereira, A. A., and Schuster, I. (2004). Single-locus inheritance and partial linkage map of *Coffea arabica* L. *Crop Breeding and Applied Biotechnology* **4**, 416-421.

Teressa, A., Crouzillat, D., Petiard, V., and Brouhan, P. (2010). Genetic diversity of Arabica coffee (*Coffea arabica* L.) collections. *EJAST* **1**, 63-79.

Tesfaye, K., Borsch, T., Kim Govers, and Bekele, E. (2007). Characterization of *Coffea* chloroplast microsatellites and evidence for the recent divergence of *C. arabica* and *C. eugenioides* chloroplast genomes. *Genome* **50**, 1112-1129.

Tessema, A., Alamerew, S., Kufa, T., and Garedew, W. (2011). Genetic diversity analysis for quality attributes of some promissing *Coffea arabica* germplasm collections in Southwestern Ethiopia. *Journal of Biological Sciences* **11**, 236-244.

Tholl, D., Boland, W., Hansel, A., Loreto, F., Rose, U. S. R., and Schnitzler, J.-P. (2006). Practical approaches to plant volatile analysis. *The Plant Journal* **45**, 540–560.

Tikunov, Y., Lommen, A., Vos, C. H. R. d., Verhoeven, H. A., Bino, R. J., Hall, R. D., and Bovy, A. G. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology* **139**, 1125–1137.

Toci, A. T., and Farah, A. (2014). Volatile fingerprint of Brazilian defective coffee seeds: corroboration of potential marker compounds and identification of new low quality indicators. *Food Chemistry* **253**, 298-314.

Tomato-Genome-Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641.

Tosh, J., Davis, A. P., Dessein, S., Block, P. D., Huysmans, S., Fay, M. F., Smets, E., and Robbrecht, E. (2009). Phylogeny of *Tricalysia (Rubiaceae)* and its relationships with allied genera based on plastid DNA data: Resurrection of the Genus *Empogona*. *Annals of the Missouri Botanical Garden* **96**, 194-213.

Tramontano , W. A., and Jouve, D. (1997). Trigonelline accumulation in salt-stressed legumes and the role of other osmoregulators as cell cycle control agents. *Phytochemistry* **44**, 1037 - 1040.

Tran, H. T. (2005). Genetic variation in cultivated coffee (Coffea arabica L.) accessions in northern New South Wales, Australia, Southern Cross University, Lismore, NSW.

Tran, H. T., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. J. (2016). Advances in genomics for the improvement of quality in coffee. *Journal of the Science of Food and Agriculture* **96**, 3300-12.

Tran, H. T. M., Vargas, C. A. C., Lee, L. S., Furtado, A., Smyth, H., and Henry, R. (2017). Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (*Coffea arabica* L.). *Tree Genetics & Genomes* **13**, 1-14.

Trick, M., Long, Y., Meng, J., and Bancroft, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant biotechnology journal* **7**, 334-346.

Trugo, L. C. (1984). HPLC in Coffee Analysis, University of Reading, England.

Trugo, L. C., and Macrae, R. (1989). Application of high performance liquid chromatography to the analysis of some non-volatile coffee compounds. *Archivos Latinoamericanos de Nutricion* **39**, 96-107.

Trugo, L. C., Macrae , R., and Dick, J. (1983). Determination of purine alkaloids and trigonelline in instant coffee and other beverages using High Performance Liquid Chromatography. *Journal of the Science of Food and Agriculture* **34**, 300-306.

Turner, T. L., Bourne, E. C., Wettberg, E. J. V., Hu, T. T., and Nuzhdin, S. V. (2010). Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature genetics* **42**, doi:10.1038/ng.515.

Uefuji, H., Ogita, S., Yamaguchi, Y., Koizumi, N., and Sano, H. (2003). Molecular cloning and functional characterization of three distinct N-methyltransferases involved in the caffeine biosynthetic pathway in coffee plants. *Plant physiology* **132**, 372-380.

Utturkar, S. M., Klingeman, D. M., Land, M. L., Schadt, C. W., Doktycz, M. J., Pelletier, D. A., and Brown, S. D. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Genome analysis* **30**, 2709-16.

Varshney, R. K. (2009). Gene-Based Marker Systems in Plants: High Throughput Approaches for Marker Discovery and Genotyping. *In* "Molecular Techniques in Crop Improvement: 2nd Edition" (S. M. Jain and D. S. Brar, eds.), pp. 119-142. Springer Netherlands, Dordrecht.

Varshney, R. K., Terauchi, R., and McCouch, S. R. (2014). Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. *PLOS biology* **12**.

Vas, G., and Vekey, K. (2004). Solid-phase microextraction: a powerful sample preparation tool prior to mass spectrometric analysis. *Journal of Mass Spectrometry*, 233–254

Vega, F. E., Ebert, A. W., and Ming, R. (2008). Coffee germplasm resources, genomics, and breeding. *In* "Plant Breeding Reviews" (J. Janick, ed.), Vol. 30. John Wiley & Sons, Inc.

Vidal, R. O., Mondego, J. M., Pot, D., Ambrosio, A. B., Andrade, A. C., Pereira, L. F., Colombo, C. A., Vieira, L. G., Carazzolle, M. F., and Pereira, G. A. (2010). A high-throughput data mining of single nucleotide polymorphisms in Coffea species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. *Plant Physiol* **154**, 1053-66.

Vieira, E. S. N., Von Pinho, E. V. D., Carvalho, M. G. G., Esselink, D. G., and Vosman, B. (2010). Development of microsatellite markers for identifying Brazilian *Coffea arabica* varieties. *Genetics and Molecular Biology* **33**, 507-U120.

Vieira, L. G. E., Andrade, A. C., and Colombo, C. A. (2006). Brazilian coffee genome project: an EST-based genomic resource. *Brazilian Journal of Plant Physiology* **18**, 95-108.

Vitzthum, O. G., and Werkhoff, P. (1976). Steam volatile aroma constituents of roasted coffee: neutral fraction. *Zeitschrift fur Lebensmittel-Untersuchung und-Forschung* **160**, 277-291.

Voorrips, R. E., Gort, G., and Vosman, B. (2014). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* **12**, 1-11.

Wang, K., Li , M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Natural reviews* **11**, 843-854.

WCR (2014). Assessment of Genetic Diversity in Coffea arabica. *World Coffee Research 2014 Annual Report*.

WCR (2017). "Sensory lexicon: Unabridged Definition and References. Version 2.0-2017." World Coffee Research, 5728 John Kimbrough Blvd., Suite 230 College Station, TX 77843-2477.

Whitt, S. R., and Buckler IV, E. S. (2003). Using natural allelic diversity to evaluate gene function. *In* "Plant Functional Genomics: Methods and Protocols" (E. Grotewold, ed.), pp. 123-139. Humana Press, Totowa, NJ.

Wintgens, J. N. (2004b). Coffee Bean Quality Assessment. *In* "Coffee: Growing, Processing, Sustainable Production - A Guidebook for Growers, Processors, Traders, and Researchers" (J. N. Wintgens, ed.). WILEY-VCH Verlag GmbH & Co. KCaA.

Wintgens, J. N. (2012). Factors influencing the Quality of Green Coffee. *In* "Coffee: Growing, Processing, Sustainable Production - A Guidebook for Growers, Processors, Traders, and Researchers" (J. N. Wintgens, ed.). WILEY-VCH Verlag GmbH & Co. KCaA.

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature reviews* **13**, 329-342.

Yang, J., Jiang, H., Yeh, C.-T., Yu, J., Jeddeloh, J. A., Nettleton, D., and Schnable, P. S. (2015). Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. *The Plant Journal* **84**, 587–596.

Yepes, M., Gaitan, A., Cristancho, M. A., Rivera, L. F., Correa, J. C., Maldonado, C. E., Gongora, C. E., Villegas, A. M., Posada, H., Zimin, A., Yorke, J. A., Mockaitis, K., and Aldwinckle, H. (2016). Building High Quality Reference Genome Assemblies using PACBio long reads for the Allotetraploid *Coffea arabica* and its Diploid Ancestral Maternal Species *Coffea eugenioides*. *In* "Plant and Animal Genome XXIV ", pp. https://pag.confex.com/pag/xxiv/webprogram/Paper22250.html, San Diego, CA.

Yu, J., and Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology* **17**, 155–160.

Yu, Q., Guyot, R., Kochko, A. d., Byers, A., rez, R. N.-P., Langston, B. J., Dubreuil-Tranchant, C., Paterson, A. H., Poncet, V. r., Nagai, C., and Ming, R. (2011). Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*). *The Plant Journal* **67**, 305–317.

Yuyama, P. M., Ivamoto, S. T., Carazzolle, M. F., Reis Júnior, O., Pereira, G. A. G., Sakuray, L. M., Ruiz, M., Leroy, T., Charmetant, P., Domingues, D. S., and Pereira, L. F. P. (2014). Transcriptome analysis of leaves and fruits of *Coffea eugenioides*. *In* "The 25th International conference on coffee and science ASIC 2014", Colombia.

Yuyama, P. M., Júnior, O. R., Ivamoto, S. T., Domingues, D. S., Carazzolle, M. F., Pereira, G. A. G., Charmetant, P., Leroy, T., and Pereira, L. F. P. (2016). Transcriptome analysis in *Coffea eugenioides*, an Arabica coffee ancestor, reveals differentially expressed genes in leaves and fruits. *Mol Genet Genomics* **291**, 323–336.

Zamarripa, C. A., and Petiard, V. (2012). Biotechnologies applied to coffee. *In* "Coffee: Growing, Processing, Sustainable Production - A Guidebook for Growers, Processors, Traders, and Researchers" (J. N. Wintgens, ed.), pp. 141-163. WILEY-VCH Verlag GmbH & Co. KCaA.

Zambonin, C. G., Balest, L., Benedetto, G. E. D., and Palmisano, F. (2005). Solid-Phase Microextraction–Gas Chromatography Mass Spectrometry and multivariate analysis for the characterization of roasted coffees. *Talanta* **66**, 261–265.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLOS one* **6**, e17915.

Zheng, X., Nagai, C., and Ashihara, H. (2004). Pyridine nucleotide cycle and trigonelline (N-methylnicotinic acid) synthesis in developing leaves and fruits of *Coffea arabica*. *Physiologia Plantarum* **122**, 404-411.

Zhou, L., Vega, F. E., Tan, H., Lluch, A. E. R., Meinhardt, L., Fang, W., SueMischke, Irish, B., and Zhang, D. (2016). Developing Single Nucleotide Polymorphism (SNP) Markers for the Identification of Coffee Germplasm. *Tropical Plant Biol.* **9**, 82–95.

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* **1**, 5-20.

Zhu, Q.-H., Spriggs, A., Taylor, J. M., Llewellyn, D., and Wilson, I. (2014). Transcriptome and Complexity-Reduced, DNA-Based Identification of Intraspecies Single-Nucleotide Polymorphisms in the Polyploid *Gossypium hirsutum* L. *G3 (Bethesda)* **4**, 1893-1905.

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Yorke, J. A., Dvorak, J., and Salzberg, S. L. (2016). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the mega-reads algorithm. *bioRxiv*, doi: http://dx.doi.org/10.1101/066100.