SNP in the *Coffea arabica* genome associated with coffee quality

Hue T. M. Tran^{1,2} · Agnelo Furtado¹ · Carlos Alberto Cordero Vargas³ · Heather Smyth¹ · L. Slade Lee¹ · Robert Henry¹

Received: 8 April 2018 / Revised: 3 August 2018 / Accepted: 8 August 2018 / Published online: 5 September 2018
© The Author(s) 2018

Abstract

Association analysis was performed at the whole genome level to identify loci affecting the caffeine and trigonelline content of *Coffea 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. Sequencing and mapping using the combined reference genomes of *C. canephora* and *C. eugenoides* (CC and CE) identified 1351 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 biosynthetic pathways. Likewise, 1060 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. These SNPs could be useful targets for further functional validation and subsequent application in arabica quality breeding.

Keywords Caffeine · Trigonelline · Association mapping · Extreme phenotypes · TAVs

Key message Forty caffeine-associated SNPs and 24 trigonelline-associated SNPs tightly linked to genes encoding enzymes involved in the conversion of substrates which participate in the caffeine and trigonelline biosynthetic pathways were

Communicated by V. Decroocq

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11295-018-1282-9>) contains supplementary material, which is available to authorized users.

✉ Hue T. M. Tran
t.tran22@uq.edu.au

✉ Robert Henry
robert.henry@uq.edu.au

¹ Queensland Alliance for Agriculture and Food Innovation (QAAFI), The University of Queensland, QLD, St Lucia 4072, Australia

² Western Highlands Agriculture & Forestry Science Institute (WASI), Buon Me Thuot, Vietnam

³ Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Turrialba 30501, Costa Rica

Introduction

Caffeine and trigonelline are two of the five most important non-volatiles in coffee beans. Their biosynthetic pathways have been well documented (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. *Coffea canephora* or *C. eugenoides*) (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 in an *C. arabica* plant (Silvarolla et al. 2004) has been reported. For trigonelline, enzymes catalysing the conversion of nicotinate to trigonelline, coffee trigonelline synthases (termed CTgS1 and CTgS2), have been characterised (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 detectable by association mapping.

Association (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 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 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, despite reducing costs. Recently, a method called “extreme-phenotype genome-wide association study” (XP-GWAS) was developed (Yang et al. 2015) that allows pooling or bulking the individuals from a diverse population that exhibits 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 (quantitative trait locus) 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 study applied XP-GWAS to identify SNPs (single nucleotide polymorphisms) associated with bean caffeine and trigonelline content. The wide phenotypic variation observed for these traits in the *C. arabica* germplasm collection (Tran et al. 2017) enables selection of extreme phenotypic groups for use in this method. Tran et al. (2018) identified SNPs associated with caffeine content using the same

populations and a relative low-quality draft reference genome of K7 arabica. We have now extended this study to include analysis of the SNP associated with trigonelline content and employed a new reference genome from the Arabica Coffee Genome Consortium comprising the *C. canephora* and *C. eugenioides* subgenomes as a framework for sequence comparison between DNA bulks and subsequent identification of TAVs and genes that may be involved in the biosynthesis pathways of caffeine and trigonelline.

Materials and methods

Biological materials

A total of 72 individuals, each representing one accession, were selected for bulk sequencing from a diverse population of 232 arabica accessions (Tran et al. 2017). They were divided into four bulks, each containing 18 individuals from each extreme phenotypic group (lowest/highest caffeine, and lowest/highest trigonelline) (Supplement Tables S1 and 2).

Methods

DNA extraction, pooling and quality

Leaf samples were collected from the germplasm plantation at CATIE (Centro Agronómico Tropical de Investigación y Enseñanza – Centre for Tropical Agricultural Research and Higher Education), Costa Rica. DNA extraction, pooling and quality were described in Tran et al. (2018) except the 18 samples in each extreme-phenotypic group were finally mixed to form a DNA bulk, resulting in four DNA bulks of low caffeine (B1), high caffeine (B2), low trigonelline (B3) and high trigonelline (B4).

Library preparation and sequencing

DNA samples were sent to Queensland Brain Institute (QBI), University of Queensland for sequencing. Four DNA bulks were prepared with four indexed PCR (polymerase chain reaction)-free libraries, then mixed and sequenced over two lanes of a HiSeq 2000 (v4) flowcell representing 90% of the sequencing.

Mapping reads to reference

Paired-end reads with insert sizes of 150 bp from four DNA bulks were imported using CLC Genomics Workbench Version 10.0 (CLC Bio, www.clcbio.com). The whole genome sequences of double haploid (DH) *C. canephora* (published) and *C. eugenioides* (provided by Arabica Coffee Genome Consortium) were imported to CLC as a standard

import and used as a reference for mapping. Raw reads were subjected to quality control analysis which was used as a guide for trimming the reads. Low-quality paired-end sequence reads were trimmed using default parameters. These trimmed reads were mapped to the reference which contained two genomes *C. canephora* (CC) and *C. eugenoides* (CE). 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 for call variant, and also to create track to examine the variant calling.

SNP identification and filtering

Variants were called using “Basic variant detection” tool with ploidy level of 4. To ensure that the SNPs identified are of high quality, SNPs selected must meet the following requirements: (i) the site must have a minimum coverage of at least 20 and maximum coverage of 1000, (ii) ignore broken pairs and remove non-specific matches, (iii) minimum of reads to be called as a variant of 4 to 6 or 15–30%, (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, 20 and 35%. Various settings for SNPs call were performed in order to determine the optima in which 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 for minimum coverage-minimum count-% of minimum frequency-% of read direction frequency, respectively.

Identification of trait-associated SNPs

The tool “Identify known mutations from sample mappings” was performed with the CLC program 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 $\geq 97\%$ were considered as “hom”. All categories were analysed with the tool “Amino acid changes” in CLC which required a number of input annotation tracks 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 spreadsheet. 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 (Fig. 1).

Functional annotation of SNPs

The final non-synonymous SNP lists were extracted from annotation track to get CDS (coding DNA sequence) sequences and imported into Blast2GO (version 4.0.7) (Conesa et al. 2005) to obtain more information of the functional annotation and the biological meaning of these SNPs. Steps involved in this analysis were run using default settings. Output data of all KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathways were examined thoroughly in order to narrow down the number of KEGG pathways and enzymes involved directly or indirectly in caffeine and trigonelline biosynthesis.

Results

Genotype selection, bulking and sequencing

For caffeine and trigonelline, the lowest and highest groups selected for XP-GWAS were about 34–35% difference in the average content (Table 1). Details of selected accessions per group are provided in Supplement Tables S1 and S2. The low caffeine group (Bulk 1, hereinafter B1) comprised mainly individuals from the cultivar/selection/natural mutant collection, while the high caffeine group (Bulk 2, hereinafter B2) included mainly wild accessions. In contrast, high trigonelline (Bulk 4, hereinafter B4) was more common in the cultivar/selection/natural mutant accessions (Table 1). Sequencing of each DNA bulk resulted in 193 million high-quality reads in B4 or up to 324 million high-quality reads in B2, with the coverage (depth) ranging from 24× for B4 to 40× for B2 (Table 1). The GC content, average PHRED score and average length after trim were comparable among bulks.

Fig. 1 Highly confident TAVs for caffeine that distinguish between two bulks 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 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 indicate location of SNPs

Read mapping and identification of SNP associated with caffeine

Mapping of reads to reference genomes

Results showed the length fraction (LF) of 1.0 and similarity fraction (SF) of 0.8 were optimal giving rise to more than 80% mapped reads and 57× average coverage (Supplement Table S3). The default setting (LF 0.5 and SF 0.8) gave the highest percentage of mapped reads (> 95%). More stringent settings in mapping (LF 1.0 and SF 0.9 or 0.95) gained less percentage of mapped reads (79 and 63%, respectively).

The identification of optimum setting for the discovery of SNP associated with caffeine

Different variant-calling settings were applied in order to determine the optimum. The minimum coverage was set

at 20 for all settings. 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) and the balance between forward and reverse reads (F/R balance) was set from 0%, 20 to 35% (Table S4). Setting 1 was the most stringent and thus resulted in the least number of TAVs (397,331 SNPs and 6101 non-synonymous SNPs), whereas setting 5 was the least stringent resulting in more TAVs (2,080,035 SNPs and 29,890 non-synonymous SNPs) (Table S4). The setting 1 required at least 20 reads at the position of SNPs called with minimum count of 6 (30%) and it also required F/R balance of 35%. The setting 5 required at least 20 reads at the position of SNPs called with minimum count of 4 (15%) and no F/R balance applied. When manually checked the SNP to examine the accuracy of the SNPs called, setting 3 gave highest score of correct call of 90%.

Table 1 Composition and sequencing statistics of four DNA bulks for low/high caffeine and trigonelline

	Low	High
Caffeine		
Total individuals (groups 1, 2, 3) ^a	18 (11; 2; 5)	18 (2; 4; 12)
Content (% dmb)	1.03	1.48
Total 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		
Total individuals (groups 1, 2, 3)	18 (2; 8; 8)	18 (10; 4; 4)
Content (% dmb)	0.93	1.31
Total 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

^a Group 1: cultivar/selection/natural mutant; group 2: introgression/hybrids; group 3: wild

KEGG pathway-based analysis for detection of trait-associated SNPs for caffeine

There were a large number of SNPs on the same CDS resulting in only 908 CDS from 1351 SNPs and 61 were hom-het or het-hom type (Supplement Fig. S1). Blast2GO outputs showed only 176 genes containing KEGG pathways in which 56 pathways were unique. One hundred four enzymes were involved in 176 KEGG pathways with 49 unique enzymes (Supplement Fig. S1 and Tables S5 and S6). 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 (Supplement Table S7). The 56 KEGG pathways were thoroughly examined to record the processes and enzymes linking to caffeine biosynthesis. There were 10 pathways with 11 enzymes presented in 63 sequences or SNPs linking to caffeine biosynthesis in which seven substrates or precursors entering the caffeine pathway were formed (Table S7 and Fig. 2). Out of 63 sequences where SNPs residing, 39 were unique sequences (CDS) with 40 unique SNPs with the majority of SNPs (37 out of 40) in the het-het form. The average coverage at the location where SNPs were called was 30× for B1 and 45× for B2. The F/R balance in reads was 0.38 in B1 and 0.39 in B2 and the average base quality score was 36 in B1 and 35 in B2 (Supplement Table S7).

The 40 TAVs were plotted onto each chromosome of the two subgenomes (Fig. 3). 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 conversion of ATP to ADP are rather generic and distributed across almost all chromosomes (grey squares) and are not subject to further analysis. TAVs linking to enzymes involved in IMP synthesis pathway (black and blue squares) were distributed in chromosomes 0, 2 and 9. The distribution of TAVs across the genome indicates the presence of multi-genes controlling the caffeine trait.

Steps in identification of trigonelline associated SNPs

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% (Supplement Table S8). In total, 1060 non-synonymous SNPs were detected. Similar to caffeine, only 0.17% of tri-allele SNPs were detected and most of SNPs (99%) were het-het (bi-allele) confirming the loss of heterozygosity within the two ancestral genomes (Supplement Table S8).

There were 719 CDS extracted from 1060 SNPs in which 14 TAVs were different in the type of alleles (hom-het and het-hom) at the SNP loci (Supplement Fig. S3). These CDS sequences were imported to Blast2GO to achieve 125 genes containing KEGG pathways with 106 enzymes in which 61 pathways and 51 enzymes were unique (Supplement Fig. S3 and Table S9). 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 (Supplement Table S10). The 61 KEGG pathways were thoroughly examined to determine the pathways and enzymes linking to trigonelline biosynthesis (Fig. 4) which were proposed in several studies (Ashihara et al. 2015; Lee et al. 2013; Zheng et al. 2004). There were seven pathways with nine enzymes associating with 36 sequences where TAVs are located. Four substrates entering the trigonelline pathway were found (Table S11 and Fig. 4) with 24 unique SNPs (Supplement Table S11) in which 23 TAVs were het-het between two bulks. The average coverage (49× for B3 and 30× for B4), 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) were high for both bulks (Supplement Table S11).

The 24 unique SNPs were plotted on the chromosomes of two subgenomes (Fig. 5). SNPs associated with enzymes involved in the reaction to produce SAM (highlighted in red) were most noticeable, followed by SNPs linked with enzymes involved in the reaction to produce L-tryptophan (black

Fig. 2 Enzymes (encoded by genes carrying the TAVs) and substrates involved in caffeine biosynthesis pathways and its related pathways. **a** The biosynthetic pathways of caffeine from xanthosine adapted from Ashihara et al. (2011b) with modification. **b** Simplified de novo biosynthetic pathway of IMP and xanthosine adapted from Ashihara and Suzuki

(2004) with modification. Red squares: enzymes encoded by genes carrying the TAVs identified from this study and catalysed for the reactions. Solid blue circles: substrates involved in caffeine synthesis pathway. Dotted blue squares: other substrates involved in the reactions

squares). There were five SNPs involved in the metabolism of glucose which eventually enters trigonelline synthesis (blue squares) and 17 were associated with the conversion of ATP to ADP (grey squares), which is very generic in chemical reactions.

Common SNPs associated with both caffeine and trigonelline contents

There were five individuals shared between low caffeine and low trigonelline bulk, and four individuals shared between

Fig. 3 Location 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))

Fig. 4 Enzymes (encoded by genes carrying the TAVs) and substrates involved in trigonelline biosynthesis pathways and its related pathways. **a** Possible metabolic pathways of biosynthesis of trigonelline and pyridine nucleotides in *Coffea arabica* (adapted from Zheng et al. 2004). **b** Simplified pathway for conversion of excess L-serine to L-aspartate (adapted from Lee et al. 2013). **c** Simplified aspartate pathway of the

pyridine nucleotides biosynthesis de novo in plants (adapted from Ashihara et al. 2015); Red squares: enzymes encoded by genes carrying the TAVs identified from this study and catalysed for the reactions. Solid blue circles: substrates involved in trigonelline synthesis pathway. Dotted blue squares: other substrates involved in the reactions

high caffeine bulk and high trigonelline bulk. Two individuals were low in this trait and high in the other. 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 (Fig. 6).

Fig. 5 Location 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-SA

Fig. 6 Commonality between caffeine and trigonelline in (input) genotypes, SNPs, genes and enzymes

Discussion

High-quality and comparable sequencing among bulks

The difference in the average content of two extreme-phenotypic groups of caffeine and trigonelline is similar to that reported for the kernel row number between bulks in maize XP-GWAS (Yang et al. 2015) indicating the sufficient differences of the materials to work with. That more individuals from the cultivar/selection/natural mutant collection in low caffeine group while more wild accessions in the high caffeine group indicates that high caffeine was more common in the wild. In contrast, high trigonelline was more common in the cultivar/selection/natural mutant accessions indicating that there may be an involvement of selection for low caffeine and high trigonelline. The presence of all types of arabica (i.e. cultivar/selection/natural mutant, hybrid and wild) in each bulk reduced the risks of population stratification causing spurious allelic association (Cardon and Palmer 2003; Price et al. 2006). The coverage (24 to 40×) 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 30× 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 polyploids like arabica coffee. Generally, for tetraploids, a sequencing depth greater than 20×, as obtained from bulk sequencing in this study, would have minimal impact on random polymerase errors on variant calls (summarised 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 and library preparation. 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 147 bp in length 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 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. 2010).

Read mapping and identification of SNP associated with caffeine

Results of mapping of reads to reference genomes with different settings

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 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).

To obtain the optimal setting for mapping, reads were mapped against the *C. canephora* genome using different length fraction and similarity fraction settings. The default setting gave the highest percentage of mapped reads. 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 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, used as guidance for local realignment to improve mapping results. This advantage was confirmed by Liu et al. (2012) in which

the false-positive rate was efficiently reduced for high coverage regions.

The identification of SNP associated with caffeine

SNPs that distinguish the low- and high-caffeine bulks were identified using different variant-calling settings. Various settings were assessed to maximise the accurate detection of SNPs which may otherwise be masked due to different subgenome homologues. Since the average coverage of the two bulks of low/high caffeine was 28 and 40 (Table 1), minimum coverage therefore was set at 20 which is higher than in several studies of allotetraploids (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 apparent alleles due to the two sub-genomes (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 unbalanced representation of individuals in the pool. 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%) to less stringent (of 20% or 15%) 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 to allow high quality reads (paired reads) in the mapping (Table S4).

Setting 1 was the most stringent in almost all parameters. Therefore, 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 and 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 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 (5), 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 (Denoëud 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 indicates that the two subgenomes of *C. arabica* might both be homozygous. Heterozygosity within the two ancestral genomes appears to have been lost, since only one allele from each genome remains in *arabica* (Cubry et al. 2008). 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. 2000; 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 (1282 from the 18,408 in setting 3) (Table S4). Chi-square test is applied to ensure the difference between two bulks are significant while If command (allele one or two in bulk 1 must be $\geq 50\%$ while the corresponding allele in the other bulk must be $\leq 50\%$) is applied to ensure the non-synonymous SNPs (caused by changing amino acid) was called between the two bulks, not between each bulk with reference. The final SNPs were checked manually using the mapping and variant calling files in the form of tracks to gain confidence (Fig. 1). Among five settings, setting 3 gave highest score for correct call (90%) with 1351 TAVs (Table S4) and was subjected to Blast2GO for functional analysis.

Results of KEGG pathway-based analysis for detection of trait-associated SNPs for caffeine

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 differences in frequency of alleles of these biosynthesis pathway loci was common between bulks of low and high caffeine and suggesting that many of these genes may influence caffeine content. 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 two bulks 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 and the average base quality score 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). 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 1000-bp error rate) (Quinlan et al. 2007).

Among the ten candidate pathways linked to caffeine, purine metabolism was the most common that generated four substrates (Supplement Table S7) 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). The 11 enzymes (where TAVs were located) were involved in three biosynthesis pathways: (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; and (3) caffeine (Fig. 2) (Ashihara et al. 2011b; Ashihara and Suzuki 2004). These enzymes catalyse conversion of SAM, a methyl donor for methylation reactions in the caffeine biosynthesis pathway, to S-adenosyl-L-homocysteine (SAH) or conversion of SAM to spermidine and spermine, the conversion of hypoxanthine to xanthine, a cascade of reactions involved in the conversion of IMP to XMP, SAICAR to AICAR, ATP to ADP, the conversion of glutamine to glutamate, and the formation of PRPP (Table S7). 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.

Among 40 TAVs potentially linked to the caffeine biosynthesis pathway, the most noticeable one is the SNP associating with enzyme that participated in the conversion of SAM to SAH in SAM cycle (Fig. S2 - A1). The SNP was located in chromosome 7 at the 13,417,576 bp position. The low-caffeine bulk has higher A-allele frequencies than the high caffeine bulk which has higher G-allele frequencies. 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. More A in B1 while more G in B2 would lead to a change in amino acid from alanine to threonine (Supplement Table S7). This enzyme would have an influence on the formation of the methyl group and thus might affect the synthesis of caffeine (Fig. 2). 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 caffeine synthesis. The second SNP which was also involved in the metabolism (breakdown) of SAM to spermidine and spermine (Fig. S2 - 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 higher C-alleles frequencies and lower A-alleles frequencies in B1 than in B2 results in the change of amino acid from mainly alanine to mainly serine (Supplement Table S7).

The enzyme converting hypoxanthine to xanthine (Fig. S2 - B) is another significant SNP. Xanthine is converted to 3-methylxanthine before being converted to theophylline—a possible direct precursor of caffeine (Fig. 2). In young and mature leaves of *C. eugenoides* which contain low levels of caffeine, [8-14C] caffeine is catabolised rapidly primarily by the main caffeine catabolic pathway via theophylline. This suggests that low caffeine accumulation in *C. eugenoides* is a consequence of rapid degradation of caffeine perhaps accompanied by a slow rate of caffeine biosynthesis (Ashihara and Crozier 1999). In tea and maté (*Ilex paraguariensis*), large amounts of theophylline are also converted to theobromine and caffeine (Ito et al. 1997). Xanthine may break down to urate or convert to xanthosine (Fig. S2 - B) which is the precursor in the main biosynthesis pathway of caffeine (Fig. 2). It seems that the frequency of C and G allele in the two bulks affects the enzyme which catalyses the formation of more or less xanthine, and eventually affects the concentration of caffeine. More C at this region 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 different in frequency of alleles resulting in the change in amino acids (Supplement Table S7). These substrates enter the xanthosine biosynthesis pathway (Fig. 2), and thus, the presence of more or less of the specific SNP alleles in each individual may have contributed to caffeine concentration difference between 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 (Fig. 2). One out of three SNPs links to enzymes that catalyse the formation of

glutamine or glutamate (Supplement Fig. S2 – E1, E2, and E3) is hom-het SNP (Supplement Table S7) which could be a positive marker.

PRPP is the first substrate in the IMP biosynthetic pathway which generates xanthosine—the first substrate in the caffeine synthesis pathway (Fig. S2). The SNP linking to the enzyme that metabolises PRPP has higher A-allele frequencies in the low bulk than in the high bulk which has higher T-allele frequencies (i.e. 75% A in B1 vs 43% A in B2 and 25% T in B1 vs 57% T in B2), resulting in the change of more threonine than serine in the high bulk (Supplement Table S7).

AMP is formed from ADP and being a substrate that participates in the synthesis xanthosine. However, ATP and ADP are generic substrates for chemical reactions so their TAVs will not be focussed further. However, two SNPs associated with two enzymes in purine metabolism pathways that convert ATP to ADP which enter the xanthosine synthetic pathways (Fig. 2 and Table S7) are significant as they are in the hom-het form (Supplement Table S7), which required more attention as potential markers. According to Koshiro et al. (2006) active supply of the substrates from purine nucleotides (i.e. ATP, ADP and AMP) are important for the biosynthesis of caffeine in coffee fruits.

Identification of trigonelline-associated SNPs

Similar to TAVs linking to caffeine, the majority (23/24) of TAVs linking to enzymes involving in trigonelline biosynthesis were het-het between two bulks revealing that the allele frequency difference at each locus between two bulks of low and high trigonelline is dominant. The high average coverage, F/R balance in reads and average base quality score indicated the high quality and confidence of these SNPs. Assuming that the two homologous genes encoding caffeine synthases were similar to trigonelline synthases, Mizuno et al. (2014) used the N-methyltransferase assay with S-adenosyl [methyl-14C] methionine to confirm these recombinant enzymes catalyse the conversion of nicotinate to trigonelline. In the present study, there was one SNP associating with an enzyme involved in the metabolism of S-adenosyl-methioninamine—a breakdown substrate of SAM which also participated in the caffeine biosynthesis pathway (Fig. S4 – A1 and A2). Similar to caffeine, the breakdown of SAM to S-adenosyl-methioninamine, and to spermidine and spermine was recorded (Fig. S4 - A2) while these two compounds have effects on salinity and drought tolerance. Tramontano and Jouve (1997) also found that trigonelline has a role as an osmoregulator in salt-stressed legume. 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 high trigonelline bulk (B4) but vice versa for low trigonelline bulk (Bulk 3) (Supplement

Table S11). The presence of more C and less A in B4 probably causes more trigonelline accumulation.

L-tryptophan was converted from serine or from indole by the same enzyme (Fig. S4 - B1 and B2). L-tryptophan is the substrate converting to L-kynurenine followed by quinolinic acid which is a precursor for trigonelline synthesis (Fig. 4). 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 (Supplement Table S11) which may be the explanation for trigonelline content in this bulk.

Glucose is converted from different substrates such as sucrose, β -glycan, D-glucoside and melibiose (Fig. S4 -C1 and C2) by two pathways with five enzymes (Table S11). Glucose enters into the pathways to form pyruvate, followed by the conversion of L-aspartate—a precursor in trigonelline pathway (Fig. 2). Among six SNPs detected from these pathways, a hom-het SNP between B4 and B3 was the most noticeable one with only G (100%) for B4 (arginine) while both G (77%) and A (23%) were present in B3 (arginine and lysine) making this attractive as a candidate marker.

There were 17 SNPs associated with enzymes involved in the conversion of ATP to ADP. 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 in chemical reactions and therefore there was investigated no further for these SNPs.

Interesting common SNPs associated with both traits

The share in common TAVs, genes and enzymes between two traits is consistent with results from the study of bean variation of arabica in which caffeine and trigonelline contents were somewhat positively correlated (Tran et al. 2017). 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).

Among the common TAVs, that linking to the enzyme (EC:2.5.1.16 – synthase) catalysing S-adenosyl-methioniamine to S-methyl 5'-thiodenosine or S-adenosyl-methioniamine to spermine or spermidine is the most significant one, while S-adenosyl-methioniamine is a breakdown substrate of SAM which participated in both caffeine and trigonelline biosynthesis pathways. The finding supports work by Mizuno et al. (2014) who used the N-methyltransferase assay with S-adenosyl [methyl-14C] methionine to confirm these enzymes catalyse 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. This enzyme was on chromosome 11 (at 30,965,526) subgenome CC with higher C-allele frequencies in low caffeine and trigonelline while

higher A-allele frequencies in high caffeine and trigonelline resulting to the change of ratio of amino acids of serine and alanine in two bulks of the two traits (Supplement Tables S7 and S10). Moreover, this TAV linking to enzymes that breaks 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.

Conclusions and suggestions

SNP loci associated with pathways of caffeine and trigonelline biosynthesis were first discovered in *C. arabica* using extreme phenotyping GWAS. Other studies have 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 two traits in breeding will thus potentially help manipulate these compounds in domesticated arabica coffee utilising molecular markers.

Findings from this study 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 favourable haplotypes for caffeine and trigonelline.

Acknowledgements The authors acknowledge Coffee Consortium for providing the genomic resources. This research was supported under Australian Research Council's *Linkage Projects* (project number LP130100376). We acknowledge Australia Awards Scholarship (AAS) and QAAFI/UQ for their ongoing support of this research.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Data archiving statement All NGS sequence data as raw data was submitted to NCBI at the Sequence Read Archive (SRA) and is available under BioProject# PRJNA477601 and BioSample #SUB4196385 and #SUB4196264.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Ashihara H (2006) Metabolism of alkaloids in coffee plants. *Braz J Plant Physiol* 18:1–8
- Ashihara H, Crozier A (1999) Biosynthesis and catabolism of caffeine in low-caffeine-containing species of *Coffea*. *J Agric Food Chem* 47:3425–3431
- Ashihara H, Crozier A (2001) Caffeine: a well known but little mentioned compound in plant science. *Trends Plant Sci* 6:407–413
- Ashihara H, Suzuki T (2004) Distribution and biosynthesis of caffeine in plants. *Front Biosci* 9:1864–1876
- Ashihara H, Monteiro AM, Gillies FM, Crozier A (1996) Biosynthesis of caffeine in leaves of coffee. *Plant Physiol* 111:747–753
- Ashihara H, Sano H, Crozier A (2008) Caffeine and related purine alkaloids: biosynthesis, catabolism, function and genetic engineering. *Phytochemistry* 69:841–856
- Ashihara H, Deng WW, Nagai C (2011) Trigonelline biosynthesis and the pyridine nucleotide cycle in *Coffea arabica* fruits: metabolic fate of [carboxyl-¹⁴C] nicotinic acid riboside. *Phytochem Lett* 4:235–239
- Ashihara H, Kato M, Crozier A (2011b) Distribution, biosynthesis and catabolism of methylxanthines in plants. *Handb Exp Pharmacol* 200:11–31
- Ashihara H, Ludwig IA, Katahira R, Yokota T, Fujimura T, Crozier A (2015) Trigonelline and related nicotinic acid metabolites: occurrence, biosynthesis, taxonomic considerations, and their roles in planta and in human health. *Phytochem Rev* 14:765–798
- Bastide H, Betancourt A, Nolte V, Tobler R, Stobe P, Futschik A, Schlotterer C (2013) A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS Genet* 9:e1003534
- Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA (2012) Development and mapping of SNP assays in allotetraploid cotton. *Theor Appl Genet* 124:1201–1214
- Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. *Lancet* 361:598–604
- Castle JC, Loewer M, Boegel S, Tadmor AD, Boisguerin V, Graaf Jd, Paret C, Diken M, Kreiter S, Tureci O, Sahin U (2014) Mutated tumor alleles are expressed according to their DNA frequency. *Scientific reports* 4
- Cenci A, Combes M-C, Lashermes P (2012) Genome evolution in diploid and tetraploid *Coffea* species as revealed by comparative analysis of orthologous genome segments. *Plant Mol Biol* 78:135–145
- Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C (2013) Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* 2013:4
- Clarke YE, Higgins IE, Plieske J, Wieseke R, Sidebottom C, Khedikar Y, Batley J, Edwards D, Meng J, Li R, Lawley CT, Pauquet J, Laga B, Cheung W, Iniguez-Luy F, Dyrszka E, Rae S, Stich B, Snowden RJ, Sharpe AG, Ganai MW, IAP 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 SA, Ozias-Akins P, Jackson SA (2015) Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. *Mol Plant* 8:831–846
- Combes M-C, Dereeper A, Severac D, Bertrand B, Lashermes P (2013) Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. *New Phytol* 200:251–260
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676
- Cubry P, Musoli P, Legnate H, Pot D, de Bellis F, Poncet V, Anthony F, Dufour M, Leroy T (2008) Diversity in coffee assessed with SSR markers: structure of the genus *Coffea* and perspectives for breeding. *Genome* 51:50–63
- Das S, Upadhyaya HD, Bajaj D, Kujur A, Badoni S, Laxmi KV, Tripathi S, Gowda CLL, Sharma S, Singh S, Tyagi AK, Parida SK (2015) Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. *DNA Res* 22:193–203
- Denoeud F, Carretero-Paulet L, Dereeper A, Droc G (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345:1181–1184
- Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:1–10
- Guo M, Davis D, Birchler JA (1996) Dosage effects on gene expression in a maize ploidy series. *Genetics* 142:1349–1355
- Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, Deynze AV, Jong WSD, Douches DS, Buell CR (2011) Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics* 12:1–11
- Hulse-Kemp AM, Ashrafi H, Stoffel K, Zheng X, Saski CA, Scheffler BE, Fang DD, Chen ZJ, Deynze AV, Stelly DM (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
- Ito E, Crozier A, Ashihara H (1997) Theophylline metabolism in higher plants. *Biochim Biophys Acta* 1336:323–330
- Kasukabe Y, He L, Watakabe Y, Otani M, Shimada T, Tachibana S (2006) Improvement of environmental stress tolerance of sweet potato by introduction of genes for spermidine synthase. *Plant Biotechnol J* 23:75–83
- Kaur S, Francki MGJWF (2012) Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant Biotechnol J* 10:125–138
- Kochko AD, Akaffou S, Andrade AC, Campa C, Crouzillat D, Guyot R, Hamon P, Ming R, Mueller LA, Poncet V, Tranchant-Dubreuil C, Hamon S (2010) Advances in *Coffea* genomics. *Adv Bot Res* 53:23–63
- Koshiro Y, Zheng X-Q, Wang M-L, Nagai C, 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 Sci* 171:242–250
- Krug CA, Mendes AJT (1940) Cytological observations in *Coffea*. *J Genet* 39:189–203

- Lashermes P, Trouslot P, Combes MC, Couturon E, Charrier A (2000) Brief communication. Single-locus inheritance in the allotetraploid *Coffea arabica* L. and interspecific hybrid *C. arabica* x *C. canephora*. *J Hered* 91:81–85
- Lashermes P, Combes M-C, Hueber Y, Severac D, Dereeper A (2014) Genome rearrangements derived from homoeologous recombination following allopolyploidy speciation in coffee. *Plant J* 78:674–685
- Lee JC-Y, Tsoi A, Kornfeld GD, Dawes IW (2013) Cellular responses to L-serine in *Saccharomyces cerevisiae*: roles of general amino acid control, compartmentalization, and aspartate synthesis. *FEMS Yeast Res* 13:618–634
- 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 OA, Leung FC-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 CC, Lam TT-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 MW, 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, Wang Y, Lam T-W, Yiu S-M, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK-S, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463:311–317
- Li S, Jin H, Zhang Q (2016) The effect of exogenous spermidine concentration on polyamine metabolism and salt tolerance in *Zoysia japonica* Steud) subjected to short-term salinity stress. *Front Plant Sci* 7:article 1221
- Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y (2012) Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* 13:1–8
- Magwene PM, Willis JH, Kelly JK (2011) The statistics of bulk Segregant analysis using next generation sequencing. *PLoS Comput Biol* 7:e1002255
- Maluf MP, da Silva CC, de Oliveira MDA, Tavares AG, Silvarolla MB, Guerreiro O (2009) Altered expression of the caffeine synthase gene in a naturally caffeine-free mutant of *Coffea arabica*. *Genet Mol Biol* 32:802–810
- Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci* 66: 9828–9832
- Mizuno K, Matsuzaki M, Kanazawa S, Tokiwano T, Yoshizawa Y, Kato M (2014) Conversion of nicotinic acid to trigonelline is catalyzed by N-methyltransferase belonged to motif B' methyltransferase family in *Coffea arabica*. *Biochem Biophys Res Commun* 452:1060–1066
- Nagy I, Barth S, Mehenni-Ciz J, Abberton MT, 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
- Nordborg M, Weigel D (2008) Next-generation genetics in plants. *Nature* 456:doi:<https://doi.org/10.1038/nature07629>
- Ogawa M, Herai Y, Koizumi N, Kusano T, Sano H (2001) 7-Methylxanthine methyltransferase of coffee plants: gene isolation and enzymatic properties. *J Biol Chem* 276:8213–8218
- Ogita S, Uefuji H, Yamaguchi Y, Koizumi N, Sano H (2003) Producing decaffeinated coffee plants. *Nature*:423
- Ogita S, Uefuji H, Morimoto M, 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 Mol Biol* 54:931–941
- Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in genetics* 6:doi: <https://doi.org/10.3389/fgene.2015.00235>
- Peace C, Bassil N, Main D, Ficklin S, Rosyara UR, Stegmeier T, Sebolt A, Gilmore B, Lawley C, Mockler TC, Bryant DW, Wilhelm L, 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, Ming R (2004) Construction of a genetic map for arabica coffee. *Theor Appl Genet* 108:829–835
- Perrois C, Strickler SR, Mathieu G, Lepelley M, Bedon L, Sp M, Husson J, Mueller L, Privat I (2015) Differential regulation of caffeine metabolism in *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta). *Planta* 241:179–191
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
- Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2007) PyroBayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* 5:179–181
- Ries D, Holtgräwe D, Viehöver P, 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
- Roychoudhury A, Basu S, Sengupta DN (2011) Amelioration of salinity stress by exogenously applied spermidine or spermine in three varieties of indica rice differing in their level of salt tolerance. *J Plant Physiol* 168:317–328
- Salmona J, Dussert S, Descroix F, de Kochko A, Bertrand B, Joet T (2008) Deciphering transcriptional networks that govern *Coffea arabica* seed development using combined cDNA array and real-time RT-PCR approaches. *Plant Mol Biol* 66:105–124
- Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat Rev* 15:749–763
- Schmutzer T, Samans B, Dyrszka E, Ulpinnis C, Weise S, Stengel D, Colmsee C, Lepinasse D, Micic Z, Abel S, Duchscherer P, Breuer F, Abbadi A, Leckband G, Snowdon R, Scholz U (2015) Species-wide genome sequence and nucleotide polymorphisms from the model allopolyploid plant *Brassica napus*. *Scientific Data* 2:150072
- Silvarolla MB, Mazzafera P, Fazuoli LC (2004) A naturally decaffeinated arabica coffee. *Nature* 429:826
- Sun Y, Wang J, Crouch JH, Xu Y (2010) Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. *Mol Breed* 26:493–511
- Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, Innan H, Cano LM, Kamoun S, Terauchi R (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J* 74:174–183
- Teixeira-Cabral TA, Sakiyama NS, Zambolim L, Pereira AA, Schuster I (2004) Single-locus inheritance and partial linkage map of *Coffea arabica* L. *Crop Breed Appl Biotechnol* 4:416–421
- Tramontano WA, 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 HT, Lee LS, Furtado A, Smyth H, Henry RJ (2016) Advances in genomics for the improvement of quality in coffee. *J Sci Food Agric* 96:3300–3312
- Tran HT, Vargas CAC, Lee LS, Furtado A, Smyth H, Henry R (2017) Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (*Coffea arabica* L.). *Tree Genet Genomes* 13:1–14

- Tran HTM, Ramaraj T, Furtado A, Lee LS, Henry RJ (2018) Use of a draft genome of coffee (*Coffea arabica*) to identify SNPs associated with caffeine content. *Plant Biotechnol J* (**in press**)
- Turner TL, Bourne EC, Wettberg EJV, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature genetics* 42:doi:<https://doi.org/10.1038/ng.1515>
- Uefuji H, Ogita S, Yamaguchi Y, Koizumi N, Sano H (2003) Molecular cloning and functional characterization of three distinct N-methyltransferases involved in the caffeine biosynthetic pathway in coffee plants. *Plant Physiol* 132:372–380
- Vidal RO, Mondego JM, Pot D, Ambrosio AB, Andrade AC, Pereira LF, Colombo CA, Vieira LG, Carazzolle MF, Pereira GA (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* 154:1053–1066
- Voorrips RE, Gort G, Vosman B (2014) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12:1–11
- Yang J, Jiang H, Yeh C-T, Yu J, Jeddalo JA, Nettleton D, Schnable PS (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. *Plant J* 84:587–596
- Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol* 17:155–160
- Yuyama PM, Júnior OR, Ivamoto ST, Domingues DS, Carazzolle MF, Pereira GAG, Charmetant P, Leroy T, Pereira LFP (2016) Transcriptome analysis in *Coffea eugenioides*, an Arabica coffee ancestor, reveals differentially expressed genes in leaves and fruits. *Mol Gen Genomics* 291:323–336
- Zheng X, Nagai C, Ashihara H (2004) Pyridine nucleotide cycle and trigonelline (N-methylnicotinic acid) synthesis in developing leaves and fruits of *Coffea arabica*. *Physiol Plant* 122:404–411
- Zhou L, Vega FE, Tan H, Lluch AER, Meinhardt L, Fang W, SueMischke IB, Zhang D (2016) Developing single nucleotide polymorphism (SNP) markers for the identification of coffee germplasm. *Tropical Plant Biol* 9:82–95
- Zhu Q-H, Spriggs A, Taylor JM, Llewellyn D, Wilson I (2014) Transcriptome and complexity-reduced, DNA-based identification of intraspecific single-nucleotide polymorphisms in the polyploid *Gossypium hirsutum* L. G3 (Bethesda) 4:1893–1905