

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts

Bing Cheng, Agnelo Furtado, Robert J. Henry¹

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

29 **Abstract**

30 **Background:** Polyploidization contributes to the complexity of gene expression resulting in
31 numerous related but different transcripts. This study explored the transcriptome diversity
32 and complexity of tetraploid Arabica coffee (*Coffea arabica*) bean. Long-read sequencing
33 (LRS) by Pacbio Isoform sequencing (Iso-seq) was used to obtain full-length transcripts
34 without the difficulty and uncertainty of assembly required for reads from short read
35 technologies. The tetraploid transcriptome was annotated and compared with data from the
36 sub-genome progenitors. Caffeine and sucrose genes were targeted for case analysis.

37 **Findings:** An isoform-level tetraploid coffee bean reference transcriptome with 95,995
38 distinct transcripts (average 3,236 bp) was obtained. A total of 88,715 sequences (92.42%)
39 were annotated with BLASTx against NCBI non-redundant plant proteins, including 34,719
40 high quality annotations. Further BLASTn to NCBI non-redundant nucleotide sequences, *C.*
41 *canephora* coding sequences with UTR, *C.arabica* ESTs and Rfam resulted in 1,213
42 sequences without hits, were potential novel genes in coffee. Longer UTRs were captured,
43 especially in the 5'UTRs, facilitating the identification of upstream ORFs (uORFs). The LRS
44 also revealed more and longer transcript variants in key caffeine and sucrose metabolism
45 genes from this polyploid genome. Long sequences (>10kb) were poorly annotated.

46 **Conclusions:** LRS technology shows the limitation of previous studies. It provides an
47 important tool to produce a reference transcriptome including more of the diversity of full-
48 length transcripts to help understand the biology and support the genetic improvement of
49 polyploid species such as coffee.

50 **Keywords:** coffee, transcriptome, full-length cDNA, long sequences, isoform, polyploid,
51 UTR

52 **Background**

53 Polyploidy creates a complicated transcriptome with diverse transcript isoforms. As an
54 important evolutionary process in plants, polyploidization generates new species and
55 increases biodiversity [1]. A balance of genetic and biochemical features is required for the
56 polyploid to survive while carrying multiple genomes in the same nucleus [2]. Genetic
57 changes associated with the formation of polyploids include gene function, which may
58 remain unchanged, or diversification among the multiple homeologs, leading to
59 neofunctionalization, subfunctionalization, or pseudogenization [3]. Alternative splicing and
60 polyadenylation also contribute further to the diversity of transcripts [4, 5]. Additionally,
61 different 5'UTRs account for transcript variation, however, limited information is available
62 on this for most genes. This diversity may include different functional motifs, like upstream
63 open reading frames, or introns harboured in this area, influencing post-transcription
64 expression [6, 7]. All these influences contribute to the diversity and complexity of a
65 polyploid transcriptome.

66 The transcriptome represents all the genes expressed in the cell or tissue. RNA sequencing
67 (RNA-Seq) makes it possible to capture the identity of these genes. Generating a reference
68 transcriptome is essential for studying variation in expression of genes and the influence of
69 genotype or environment on their expression [8, 9]. Most studies generate a reference
70 transcriptome by short-read sequencing and reconstruct the transcriptome by the assembly
71 and/or mapping of reads to other available reference genomes [10-12]. However, this is
72 difficult for long transcripts, repetitive sequences and transposable elements. It is particularly
73 challenging for complex polyploid genomes [13]. The LRS technology (e.g. PacBio) has
74 recently become available and this technology overcomes these difficulties by generating
75 sequence information for the full length as a single sequence read, including very long
76 transcripts (*e.g.* those exceeding 10kb) without the need for further assembly. This technique

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

77 has been applied in a few plant studies and provides further information on transcript
78 diversity, including alternative splicing and alternative polyadenylation [4, 5].
79 Arabica coffee is a recent allotetraploid ($2n=4x=44$; ~50,000 years old) derived from *C.*
80 *canephora* and *C. eugenioides*. A high-quality reference genome and annotation are not yet
81 available for Arabica coffee. However, a draft genome is available for one of the sub-
82 genomes, *C. canephora* [14]. Arabica coffee is highly regarded by coffee consumers, is of
83 great economic value and accounts for almost 70% of world coffee traded [15] However, it is
84 produced in limited high altitude tropical environments and is threatened by climate change.
85 Understanding the genetic and environmental control of coffee quality will be facilitated by
86 the availability of detailed knowledge of the transcriptome of the coffee bean. This study
87 used LRS by Pacbio Iso-seq to characterise the Arabica coffee bean transcriptome including
88 beans from immature, intermediate and mature stages in order to explore the complex
89 polyploid system and establish a reference transcriptome for future studies of gene
90 expression.

91 **Data Description**

92 *RNA sample preparation*

93 Fruits at different development stages (immature, intermediate and mature fruits) of *Coffea*
94 *arabica* var. K7 (see Supporting Information 1 Fig. S1) were harvested from Green Cauldron
95 Coffee, Federal, Australia. Ten coffee trees were selected randomly and 5 coffee fruits were
96 collected separately for each tree and each stage of development. Samples were collected in
97 triplicate. In total, 450 coffee fruits (900 beans) from 45 trees were collected. Once each fruit
98 was harvested, the pericarp was removed immediately with a scalpel in 20 s or less. The
99 coffee beans were immediately frozen in liquid nitrogen, transported on dry ice and stored at
100 -80 °C until further use. Total RNA was extracted from coffee fruits as described by Furtado

101 [16]. Isolated RNA was assessed for integrity using an Agilent RNA 6000 nano kit and chips
102 on a Bioanalyzer 2100 (Agilent Technologies, California, USA) and processed further for
103 cDNA preparation.

104 *cDNA preparation*

105 The Pacbio Iso-seq protocol was used for cDNA preparation. cDNA was synthesised using a
106 Clontech SMARTer PCR cDNA Synthesis kit (ClonTech, Takara Bio Inc., Shiga, Japan) and
107 amplified using a KAPA HIFI PCR kit (Kapa Biosystems, Boston, USA). The double-
108 stranded cDNA was split into two sub-samples. One was used directly for sequencing. The
109 other set was normalised to equalise transcript abundance and obtain rare sequences.

110 The cDNA was purified for normalisation using a QIAquick PCR Purification Kit (Qiagen).

111 The purified cDNA was precipitated and normalised with a Trimmer-2 cDNA normalisation
112 kit (Evrogen, Moscow, Russia). The resulting cDNA was evaluated and quantified using an
113 Agilent DNA 12000 Kit and Chips on a Bioanalyzer 2100 (Agilent Technologies, California,
114 USA). The same amount of non-normalized and normalised cDNA was used as input for
115 Pacbio Iso-seq.

116 Samples were subjected to a Pacbio Iso-Seq protocol through purification, size selection
117 (Blue Pippin system), re-amplification, SMRTbell template preparation and Iso-seq on a
118 Pacbio RS II platform. A size selection protocol was applied as smaller cDNAs are more
119 abundant and would otherwise be preferentially sequenced. Four Bluepippin bins were
120 selected for non-normalized cDNA sequencing, with size ranges of 0.5-2.5kb, 2-3.5kb, 3-
121 6.5kb and 5-10kb, respectively since Pacbio sequencing preferentially sequences short DNA
122 fragments. Two bins were selected for normalised cDNA sequencing, 2-3.5kb and 3-6.5kb, as
123 the normalisation biases against longer sequences.

124 *Raw read processing and error correction*

1
2 125 Sequence data was processed through the RS IsoSeq (version 2.3) pipeline [17]. The first step
3
4
5 126 was to remove adapters and artefacts to generate reads of insert (ROIs) consensus sequences.
6
7 127 Short sequences less than 300 bp were removed as the Bluepippin cDNA size selection starts
8
9 128 from 500 bp, where some sequences less than 500 bp have a chance to be sequenced. Non-
10
11 129 Chimeric ROIs sequences were filtered into two groups of sequences comprised of full-length
12
13 130 ROIs sequences and non-full length ROIs sequences. Full-length (FL) ROIs sequences were
14
15 131 identified based on the presence of the 5'-adaptor sequence, the 3' adapter sequences (both
16
17 132 used in the library preparation) and poly (A) tail. Further, FL ROIs sequences were passed
18
19 133 through the isoform-level clustering (ICE). ROI sequences were used to correct errors
20
21 134 (polish) the isoform sequences using the Quiver software module. The polishing process of
22
23 135 Quiver generated two isoform sequence files, one with high quality (HQ) isoform sequences
24
25 136 and the other with low quality (LQ) isoform sequences corresponding to an expected
26
27 137 accuracy of $\geq 99\%$ or below respectively. LQ output (or non-FL coverage sequences) is
28
29 138 useful in some cases, as it may result from rare transcripts or lower coverage sequences. And
30
31 139 these low coverage sequences can be further used to correct errors in HQ output. The Primer
32
33 140 IIA sequence motifs (used in the library preparation) which escaped removal at the ROIs
34
35 141 stage corresponded to 11 sequences were trimmed using CLC genomic workbench 9.0 (CLC,
36
37 142 QIAGEN, CLC Bio, Denmark). After combining the HQ and LQ transcripts, further
38
39 143 clustering was processed with CD-HIT-EST (c=0.99) [18].
40
41
42
43
44
45
46
47
48

49 144 In the following step, the contaminant sequences were removed by CLC stepwise as follows.
50
51

52 145 1) Chloroplast transcript sequences were identified by BLASTn to the *C.arabica* complete
53
54 146 chloroplast genome (GenBank: EF044213.1). 2) Mitochondrial transcripts were characterised
55
56 147 by BLASTn to *N. tabacum* and *V. vinifera* complete mitochondrial genomes (BA000042.1
57
58 148 and FM179380.1). 3) Ribosomal sequences were detected by BLASTn to the reported
59
60
61
62
63
64
65

149 *C.arabica*, *C.canephora* and *C. eugenoides* ribosomal genes (AJ224846, EU650386,
150 DQ153609, AF416459, EU650384, EU650385, AF542981, AF542990, JX459583,
151 JX459584, JX459585, JX459586, JX459587, DQ153593, AF542982, DQ423064,
152 DQ153588, DQ153621, AF542986). 4) Virus, viroid and prokaryote contaminants were
153 identified with BLASTn to their reference genomes from the NCBI database (April 4th,
154 2017). Prokaryotic contaminants were screened with available reference genomes from NCBI
155 (Feb 9th, 2017). 5) Fungal sequences were investigated by BLASTn to fungal proteins (April
156 4th, 2017). All the above analyses were processed one after another with a maximum E-value
157 threshold of 1e-10.

158 From the BLASTn results, significant matches were filtered with a bit score (A) \geq 300 as
159 well as identity \geq 80%. In each step, the filtered significant sequences were processed further
160 with cloud BLASTn to the NCBI non-redundant database (bit score: B) to further confirm the
161 matches. This validation step was confirmed by comparison of the bit score (comparison of
162 value A and B). If the higher bit score was associated with a contaminant sequence in the
163 BLASTn (A>B), then the sequence was discarded. In total 526 sequences corresponding to
164 chloroplastic (200), mitochondrial (264), ribosomal (37), viral (0), viroid (0), prokaryotic (0)
165 and fungal (25) contaminant sequences, respectively, were removed in this process. Sequence
166 quality was then accessed with the Fasta Statistics through Galaxy/GVL 4.0 [19]. This set of
167 Iso-seq processed isoforms was used for further analysis and hereafter named the “Coffee
168 long read sequencing (coffee-LRS) isoforms”. The term ‘isoforms’, or ‘isoform sequence’ or
169 ‘transcript’ used in this study represent individual sequences from the coffee-LRS isoforms,
170 while “transcript variants” indicate different transcript of a gene, including alternative spliced
171 variants, homeologs, etc.

172 *Transcriptome annotation*

173 A number of databases were used for annotation of the coffee-LRS isoforms described as
174 follows. 1) The plant Geninfo identifier (GI) list was downloaded from NCBI Protein Entrez
175 (May 2nd, 2017, 8,431,379 items). The plant proteins were retrieved from the NR database
176 using this GI list, yielding 5,099,147 sequences (NR-plant). Then, the full set of the coffee-
177 LRS isoforms was submitted to stand-alone BLASTx against the NR database below 1e-10.
178 2) Sequences without hits from step 1 were submitted further to NCBI non-redundant
179 nucleotide sequences (NT, May 5th, 2017) BLASTn at 1e-10. 3) Sequences without a hit from
180 step 2 were processed further with BLASTn (1e-20) to *C. canephora* coding sequences
181 (CDS) with UTR and *C. arabica* EST database [20, 21]. 4) The output of BLASTx was
182 filtered with query coverage (qcovs), cumulative identity (ID) and sequence length into three
183 categories, high, medium and low quality annotation. Query coverage indicates the input
184 coffee-LRS isoforms covered by the matched sequences. Cumulative identity represents the
185 identity length to the aligned length (AL). ID can be expressed as the ratio of the sum of
186 identity length to the sum of the aligned length of all the Hsps (High-scoring Segment Pairs)
187 of a subject. The four databases above, NR plant, NT, *C. canephora* CDS with UTR and *C.*
188 *arabica* EST database, are named as FOUR databases in this manuscript. Finally, all the
189 BLASTx and BLASTn results were processed by function annotation with BLAST2GO.
190 The Blast2GO Pro 4.0 (North America, US: USA2 Version: b2g_Sep 16) pipeline was based
191 on default settings [22]. InterProScan (IPS) was used to search sequence protein domains
192 from EBI databases to improve annotation (North America, US: USA2, Version: b2g_Sep
193 16). In the follow-up phase, Blast2GO Mapping, Annotation and Annex functions were
194 applied to retrieve GO (gene ontology) terms, select reliable annotations and increase the
195 number of annotated isoforms respectively. The GO-slim tool was used against the plant

196 database to provide plant generic GOs. Finally, GO enzyme mapping and KEGG (Kyoto
197 encyclopaedia of genes and genomes) pathway maps were loaded.

198 *Case studies with the caffeine and sucrose genes*

199 Two case studies were performed with genes encoding caffeine and sucrose biosynthesis
200 pathway (caffeine and sucrose genes) to investigate specifically the quality, advantage and
201 additional potential of the coffee-LRS isoforms. Reported coffee caffeine and sucrose
202 candidate genes were downloaded from the European Nucleotide Archive (EMBL-EBI)
203 (Table 3 and Table 4).

204 For potential caffeine candidate genes, coffee-LRS isoforms were processed with BLASTn
205 (1e-20) against the reported caffeine genes. Sequences with hits to the reported caffeine genes
206 were submitted to BLASTx (1e-20) with the NR database to confirm whether they were
207 caffeine genes (higher bit score). Confirmed transcripts (potential caffeine isoforms) and
208 sucrose isoforms annotated by Blast2GO (potential sucrose transcripts) were further
209 evaluated with Geneious 10.0.4 by aligning back to the reported candidate genes in allele
210 level [24]. Motif analysis was conducted with default parameters except for “ten motifs”
211 selected with MEME 4.11.2 [25]. UTRscan was used for UTR functional motifs annotation
212 [26].

213 *Comparison to other available coffee databases*

214 To compare with available coffee sequences, the full coffee-LRS isoforms were processed
215 with BLASTn (1e-20) to *C.canephora* CDS with UTR and *C. arabica EST database*,
216 respectively and the other way around [20, 21]. The *C. eugenioides* transcriptome (young
217 leaves and mature fruits) from Illumina was also used in the comparison [23].

218 *Novel genes*

219 Coffee-LRS isoforms without hits to the FOUR databases were submitted to the Rfam
220 database by Blast2GO Pro package for non-coding RNA analysis [27]. Sequences without
221 hits to Rfam were probably novel genes in coffee.

222 *Analysis of long sequences*

223 In order to explore the advantage of using the LRS PacBio platform to obtain long sequences,
224 the BLASTx and BLAST2GO functional annotation result for the coffee-LRS isoforms
225 longer than 10kb were extracted from the total dataset.

226 **Analyses**

227 *Overview of full-Length RNA molecules from long-read sequencing*

228 A total of 2,618,905 raw reads were generated from LRS platform, which yielded 443,877
229 reads of insert. After 8,842 short sequences (less than 300 bp) were removed, 233,464 full-
230 length (FL) and 201,571 non-full-length (NFL) reads were generated. The individual
231 isoforms were sequenced in average five times. In total, 95,995 coffee-LRS isoforms were
232 recovered after sequences representing chloroplast, mitochondrial and ribosomal transcripts
233 were removed (see Table 1). The length of the sequences in this dataset ranged from 301 bp
234 to 23,335 bp, with an average length of 3,236 bp. The GC content was 41.4% and the N50
235 was 4,865 bp.

236 The BLASTx output (against NR plant) was divided into three groups, high, medium and low
237 quality based on Qcov, ID and sequence length (see Data description and Table 2). There
238 were 34,719 (high), 13,655 (medium) and 40,314 (low) sequences were grouped into each
239 quality groups, respectively (see supporting information 1 Table S1). Thereafter, 7,280
240 sequences without hits were processed with BLASTn to NT database and resulting in 1,981
241 sequences with hits. A total of 5,299 sequences without a hit were further accessed with C.

242 *canephora* CDS and UTR and *C.arabica* contigs. Finally, there were 1,217 sequences with no
243 hits to any of the above databases (FOUR databases).

244 *Functional Annotation*

245 Functional annotation of the coffee-LRS isoforms was investigated using different databases.

246 The data in Table 2 shows that 88,715 sequences (92.42%) had hits to NR plant proteins. A

247 total of 70,774 sequences (73.73%) matched to IPS protein domains with 33,605 IPS GOs

248 (35.01%). A number of 78,571 sequences (81.85%) had identified GOs. After the GOs were

249 merged, GOs of 58,050 sequences (60.47%) matched with GO-slim (plant).

250 Of all the hits to the NR plant proteins from BLASTx, the coffee-LRS isoforms (maximum

251 50 hits to each sequences) had the highest number of hits to the *Nicotiana tabacum* (tobacco,

252 174,6308 hits), followed by *C. canephora* (142,656 hits), *Vitis Vinifera* (grape, 134,025 hits)

253 and *Theobroma cacao* (cacao, 132,336 hits) proteins (see supporting information 1 Figure

254 S1). Most hits found in tobacco were probably because the tobacco database is more

255 extensive and well annotated than those of other related species, like *C. canephora*. For top-

256 hit species, there is no doubt the majority of the sequences has top-hit with the progenitor,

257 *C.canephora* (73,587 sequences), followed by *Sesamum idicum* (1,321 sequences), *Nicotiana*

258 *tabacum* (767 sequences), etc. (see supporting information Figure S2). The NR-plant

259 database consists of few proteins sequences from *Coffea arabica* as reflected by just 485

260 protein sequence hits and is ranked seventh in the top-species hit list. This indicates the limit

261 information on *Coffea arabica*. Of the 33,512 sequences (34.91%) with IPS GOs, cytochrome

262 P450 (IPR001128, 353 matches) had the most sequence matches among the IPS families (see

263 supporting information 1 Fig. S3).

264 Biological process (BP, 56,230 sequences) was more abundant than cellular component (CC,

265 44,528 sequences) and molecular function (MF, 45,604 sequences) (see supporting

266 information 1 Fig. S4). Within these functional groups, the highest number of sequences were
267 annotated with the biosynthetic process (11,627 sequences, 20.68%), membrane component
268 (21,175 sequences, 47.55%) and transferase activity (11,921 sequences, 26.14%). A total of
269 156 pathways with 921 enzymes were annotated by KEGG, associated with 11.97% of the
270 whole dataset (11,489 sequences). Among these, starch and sucrose metabolism ranked as the
271 fifth most abundant pathways, with 36 encoding enzymes and 766 isoforms annotated (see
272 supporting information 1 Fig. S5). The average number of coffee LRS isoforms encoding the
273 921 enzymes was 18 while the highest number was found in phosphatase (EC: 3.6.1.15,
274 2,969 sequences), encoding the purine metabolism and thiamine metabolism pathway. In
275 comparison, only 802 sequences were associated with 142 pathways and 374 enzymes in *C.*
276 *eugenioides* transcriptome and starch and sucrose pathway relating to 450 contigs was the
277 most encoded pathway [23].

278 The candidate genes for the major caffeine candidate genes were not identified by KEGG
279 pathway. To evaluate the annotated isoforms and their diversity, further analysis was
280 performed with caffeine pathway. The sucrose pathway was also analysed as a case study as
281 sucrose candidate genes were relatively long and highly diverse. Both of these pathways are
282 important for the understanding of coffee quality [28].

283 *Case study I: Isoform diversity in the caffeine biosynthesis pathway*

284 The caffeine pathway has been widely studied previously (Fig. 2a). Candidate genes and
285 complete coding sequences of both transcripts and genomic DNA are available in public
286 databases and can be used as well-established references for caffeine candidate gene analysis
287 (Table 3). From the BLASTn output, 25 long-read transcripts were annotated and related to
288 candidate caffeine genes. Further alignment suggests ten high quality isoforms were likely to
289 be putative caffeine genes, including three transcript variants of *XMT1*, one of *MXMT1*, one
290 of *MXMT2* together with two of *DXMT1* and three of *DXMT2*. All genes encoding caffeine

291 the primary pathway except the *XMT2* gene were present in this bean transcriptome (Fig. 2
292 and Table 3). The length distribution of these isoforms ranged between 977 and 1,517bp.
293 Importantly, all ten isoforms were extended at the 5' UTR region compared to the
294 corresponding sequences reported in Arabica and Robusta coffee (see Table 3), while eight
295 isoforms were longer at the 3' end (Fig. 2b, 2c, 2e and supporting information 2 Fig. S6). The
296 most extended isoform (c695597/f1p2/1421) was 136 bp longer than the previously reported
297 candidate genes (*CaXMT1*, Fig. 2b). Nine isoforms were found to be longer than the reported
298 genomic DNA sequences. The other isoform was likely to have resulted from an alternative
299 polyadenylation event (c25904/f2p0/977, Fig. 2c) as two potential polyadenylation signals
300 (AAUAAA) were identified in the 3' UTR (Fig 2d). Alternative splicing was also presented
301 in caffeine isoforms, for example, intron retention was detected in one of the putative *DXMT2*
302 isoforms (Fig. 2e).

303 Coffee LRS isoforms encoding *XMT1* (Fig 2b), *MXMT1* (supporting information Fig. S6)
304 and *DXMT2* (Fig 2c, 2e) were better aligned to the corresponding *C.canephora* isoforms,
305 individually (higher identity, see Fig. 2 and supporting information Fig. S6). This indicates
306 these transcript variants were potentially *C.canephora* sub-genome copies. In contrast,
307 isoforms encoding *XMT2* (Fig 2c), *MXMT2* (supporting information Fig. S6) and *DXMT1*
308 (Fig 2c, 2e) were poorly aligned with *C. canephora* isoforms (more variants) and were
309 probably *C.eugenioides* sub-genome copies.

310 *Case study II: Long sucrose isoforms provide insight into the complexity of the polyploid* 311 *system*

312 Sucrose genes were used to investigate the transcriptome sequence diversity of the polyploidy
313 system. For the sucrose synthase 1 gene (*SSI*), one of the important genes in the sucrose
314 metabolism, nine transcript variants were identified (Fig. 3, Table 4 and Fig. 4a). Compared
315 to c86432/f7p9/4842, the other eight transcript variants varied in motif replacement (motif 7

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

316 replaced motif 9 in c106591/f2p0/4381), deletion (for example c92344/f1p26/4662) and
317 relocation (intron retention, c92296/f1p5/4676 and c91298/f1p1/3137), etc. (Fig. 4b). The
318 length of these nine putative *SSI* transcript variants ranged from 2,961 to 4,842 bp.
319 Importantly, all the sucrose transcript variants studied in this research were extended in the
320 5'UTR region relative to previous reports, except for *SPSI*, c51110/f2p0/3136 (Table 4).
321 Some transcript variants, such as the longest putative *SSI* sequence identified,
322 c86432/f2p7/4842 (4,842 bp), extended 2,131 bp upstream of the *C.canephora* coding
323 sequence (*G-CcSSI*) and 1,994 bp upstream of the Arabica sucrose synthase 1 mRNA coding
324 sequence (*CaSSI*). The length of the 5' leading region of the *SSI* transcript variants ranged
325 between 218 and 2,131 bp (Table 5). To understand the diversity in this region, the 5' leading
326 sequences of the nine putative *SSI* transcript variants were scanned using the UTRdb online
327 server. A maximum of 12 upstream open reading frames (uORFs) were identified and the
328 number was positively correlated with the length of the sequences. No uORFs were identified
329 in the two transcript variants with short 5'UTR, c62911/f29p21/2965 (218 bp leader
330 sequence) and c72639/f25p28/2961 (232 bp leader sequence).
331 The nine *SSI* transcript variants revealed transcript diversity that resulted largely from
332 different copies from the progenitors. When aligned to *G-CcSSI* (*C. canephora SSI* genomic
333 sequence), the top four putative *SSI* transcript variants showed high identity and consistent
334 nucleotide variants (like the guanine highlighted at 3,726 bp in the consensus sequence, Fig.
335 4c), suggesting that these were copies from the *C. canephora* sub-genome. For example,
336 compared to the consensus sequence, the same indels were present in 3,707bp and 3,733bp, a
337 cytosine at 3,713bp and guanine at 3,715bp, etc. Consistently, the sequence of intron
338 retention in one of the top four sequences, c91298/f1/p1/3137 (Fig. 4d) shows high homology
339 to the intron sequence of *C. canephora*. However, the bottom five transcript variants had a
340 higher number of variations compared to *G-CcSSI* that are likely to be *C. eugenioides* sub-

341 genome derived copies. The lower five transcripts had lots of variations compared to
342 *C.canephora* intron 10, further indicating this group was from a different copy, probably *C.*
343 *eugenioides* (Fig. 4e). Additionally, some alleles of *G-CcSSI* were common in nine putative
344 Arabica *SSI* transcript variants and Arabica sucrose synthase 1(*CaSSI*), such as the variant at
345 3,666 bp (Fig. 4e). This type of allele probably results from different genotypes. Polyploid
346 expression patterns were also observed in *SPI* transcript variants, the top two alignments
347 were similar to *C.canephora* and the other two were slightly different but related. All of the
348 four transcript variants were longer in the upstream sequences while three extended further
349 downstream than had previously been reported.

350 Another essential potential of LRS is to explore sequences not yet complete or published. For
351 instance, four transcript variants were identified from this research while *SPS2* has only been
352 identified in *C. canephora* rather than *C. arabica* (Fig. 4f).

353 *Comparison to other available coffee databases*

354 To understand the advantage and the diversity of this polyploid coffee transcriptome, a
355 comparison was made with the available coffee database. More than twice the number of
356 isoforms were identified in the tetraploid Arabica LRS transcriptome (immature, intermediate
357 and mature fruits) compared with the *C. eugenioides* contigs (36,935 de novo assembled
358 contigs, average length: 701 bp, from immature leaves and mature fruits), *C.canephora* CDS
359 with UTR (25,570 sequences, from a variety of tissues, including fruits) and *C. arabica* EST
360 database (35,153 contigs, including fruits) (see Table 2) [14, 20, 23]. The coffee-LRS
361 isoforms show greater transcript length, diversity and a lower GC content. The N50 of the
362 Pacbio dataset (4,865 bp) was more than three times longer and the average length was more
363 than twice that of the other databases. The sequence distribution of *C.arabica* contigs peaks
364 at 655 bp while *C.canephora* CDS with UTR reaches the largest number of sequences at

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

365 1,490 bp (Fig 5). Most of the sequences from the *C. canephora* CDS with UTR and the *C.*
366 *arabica* EST database were less than 3,770 bp. By comparison, 39,917 coffee LRS isoforms
367 (41.6%) were longer than 3,770 bp

368 Results of the BLASTn analysis indicated that of the 95,995 coffee-LRS isoforms, 9,308
369 (9.7%) had no matches to the *C.canephora* CDS with UTR while 3,682 (3.8%) isoforms had
370 no hits to the *C.arabica* contigs. This indicates coffee-LRS isoforms are very diverse
371 compared to these two databases. Conversely, 9,167 (26.1%) of *C.canephora* CDS with UTR
372 and 4,830 (18.9%) of *C.arabica* contigs had no hits to the coffee-LRS isoforms. These two
373 sets of sequences without hits are probably sequences from leaf or other tissues not expressed
374 in the tissues investigated in this study.

375 *Novel genes*

376 The 1,217 sequences without hits to the FOUR databases (NR plant proteins, NT database, *C.*
377 *canephora* CDS with UTR and *C. arabica* EST database) were submitted to the Rfam server
378 to predict non-coding RNAs (ncRNA). The four isoforms that matched were in three
379 biotypes, two transcripts were identified as CD-box snoRNA, one as HACA-box snoRNA
380 and the other one as a miRNA (see supporting information Table S2). Other than these, the
381 other 1,213 sequences had no hit to the FOUR databases and Rfam are likely to be novel
382 genes that have not been discovered in coffee or contaminants from other organisms with no
383 sequence information to date. Length distribution of this new dataset ranged from 325 to
384 19,189 bp.

385 *Long transcripts*

386 In order to assess the value of LRS in discovering long sequences, 577 transcripts longer than
387 10 kb were further analysed. Functional annotation of this extremely long dataset shows the
388 majority of the sequences (564 sequences, 97.8%) matched to the FOUR databases. The

389 HSP/Hit coverage distribution was relatively evenly distributed from 0 to 100% compared to
390 the HSP/Seq coverage. In parallel, the majority of sequences distributed less than 50%
391 HSP/Seq coverage and peaked at 6%, representing limited information of long sequences in
392 the NR database. IPS matches were found for 352 sequences (61.0%) while 61 of them had
393 IPS GOs. A total of 446 sequences (77.3%) were retrieved with GO terms, while 201
394 isoforms (34.8%) from these were also annotated with GO-Slim.
395 In total, 144 sequences were classified into the biological process, with 92 sequences into
396 cellular component and 79 into molecular function (see supporting information Table Fig.
397 S7). Among them, biosynthetic process (31 sequences), member (43 sequences) and
398 hydrolase activity (25 sequences) were the top groups, separately, from the three functional
399 process. Among the annotated isoforms, 18 sequences encoding 12 enzymes from 13
400 pathways were annotated with a KEGG pathway. The starch and sucrose metabolism ranking
401 the third most encoded pathway with two isoforms encoding two enzymes.

402 **Discussion**

403 Full-length transcripts generated by LRS in this study provided an isoform level polyploid
404 coffee bean reference transcriptome. Compared to its sub-genome progenitors, the Arabica
405 coffee bean transcriptome was more diverse and complicated with more isoforms, enzymes
406 and pathways. Case studies in caffeine and sucrose identified that this diversity and
407 complexity were a result of alternative splicing, polyadenylation, 5'UTR extension and sub-
408 genome copies. Discovery of novel genes and long transcripts was also an advantage of using
409 the LRS technology.

410 *Polyploid expression*

411 Different transcript variants may vary in function within the cell and be differentially
412 expressed in tissues or environmental conditions. The abundance of variants in the Arabica

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

413 transcriptome and case studies of caffeine and sucrose genes compared to the sub-genome
414 progenitors clearly shows the complexity of the polyploid expression.

415 Generally, polyploidy results in three main expression patterns of non-additive expression,
416 dominant expression in which total gene expression in the hybrid is similar to one of the
417 parents, transgressive expression compared to the progenitors or unequal homeolog
418 expression [1]. Previously, it was proposed in coffee that the lower caffeine in Arabica coffee
419 was due to the *C. eugenioides* sub-genome attributes. Based on phylogenetic analysis,
420 CaXMT1, CaMXMT1 and CaDXMT2 were believed to be from the *C. canephora* sub-
421 genome while CaXMT2, CaMXMT2 and CaDXMT1 were from the *C. eugenioides* sub-
422 genome[29]. *C. eugenioides* has a very low caffeine biosynthesis together with a rapid
423 catabolism [30]. The expression of sub-genome copies from *C. eugenioides* suggested lower
424 caffeine in Arabica coffee compared to Robusta coffee. This study supports this hypothesis of
425 transcript variants from sub-genome copies controlling the trait.

426 Using the LRS isoforms, further studies are now possible at the isoform level (this study was
427 at the transcript variant level) to understand sub-genome gene expression in the polyploid *C*
428 *arabica*. First, it will be possible to determine directly whether the expression of Arabica
429 caffeine genes follows a non-additive expression pattern. Secondly, it would be interesting to
430 determine the reason for more transcript variants identified is similar to those from the *C.*
431 *canephora* sub-genome copies of the XMT and DXMT genes and their differential
432 expression in tissues and at development stages. Thirdly, it will be possible to determine
433 whether this expression pattern is influenced by environment, influencing coffee quality.
434 Fourth, whether these different gene expression patterns result in different phenotypes.
435 Similar analysis could be applied to many other genes or pathways of interest. Isoforms and
436 transcript variants found in LRS tetraploid Arabica coffee bean transcriptome in this study
437 were assigned to a number of functional groups, pathways and to specific enzyme functions.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

438 Arabica is believed to be more adaptive to temperature change than its diploid parents [31].
439 This study may also help elucidate the genetic basis of the higher sucrose in Arabica coffee.
440 More generally, the complete polyploid transcriptome from this study will improve our
441 understanding of the evolutionary adaptation and plasticity of polyploid species. However,
442 further improvement is still needed in LRS technologies to improve the sequencing depth.
443 Candidate genes in the caffeine pathway are reported to be expressed at low levels in fruits
444 compared to leaves, especially XMT2 (detected by quantitative RT-PCR) [29]. Transcripts
445 were not detected in this study probably due to low expression of XMT2 and the PacBio iso-
446 seq technology not being sensitive enough to capture these transcripts. This is likely to
447 happen in the case of other isoforms expressed at low levels that may not be captured by the
448 Iso-Seq technology even after application of the cDNA library normalisation step, as was
449 applied in this study.

450 *5'UTR extension*

451 Full-length transcripts captured in this study show the advantage of LRS. All the caffeine and
452 sucrose isoforms annotated in this study, except for *SPS1*, were extended in the 5'UTR
453 compared to those available from public databases. Previously, it was difficult to sequence
454 the 5' end as cDNA library preparation starts from the 3' end and normally fails to reach the
455 5' end. Further, it was not easy to assemble the non-coding parts of transcripts as limited
456 cDNA sequence was available to guide the assembly and confirm the contigs obtained.
457 Therefore, less information is available on the 5'UTRs, especially for plants. Generally, the
458 length of the 5'UTR ranges from 100 up to a few thousand bp [32]. This length difference is
459 proposed because of the complex gene regulation maintained in eukaryotes [33]. Few post-
460 transcriptional mechanisms have been studied in 5'UTRs, including the regulation by the pre-
461 initiation complex and uORF re-initiation.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

462 uORFs are common in 5'UTRs that have critical regulation. They contain their own set of
463 start and stop codons that can be scanned by ribosomes and translated. This regulation can
464 inhibit translation of the main ORF transcript and reduce the amount of protein translated.
465 Regulation of re-initiation of uORF translation was found to be associated with the length of
466 sequence between the uORF and the main ORF, suggesting interactions with translation
467 factors are required before initiation of translation [34]. This was also shown to be influenced
468 by stress conditions [34]. However, not all uORF may have a role in translation control. In
469 the leucine zipper transcription factor (*bZIP*) 11 gene, for example, harbouring four uORFs,
470 only uORF2 was required for this regulation and this uORF is relatively conserved [35].
471 Other types of 5'UTR regulation may also be found such as that due to introns in the 5'UTRs.
472 This happens to approximately 35% of human genes [6].
473 Understanding the mechanism of 5'UTR regulation will be greatly facilitated by the use of
474 the full-length transcripts. In this study, multiple uORFs were characterised in the *SSI* 5'
475 UTR and these may contribute to diverse functions and regulation that may be influenced by
476 stress conditions. Climate change is a threat to Arabica coffee, which grows at high altitude.
477 It may be possible to influence 5'UTR regulation in Arabica coffee and have the potential to
478 influence coffee quality. To confirm this, further phenotype, proteome and metabolome
479 studies are required.

480 *Long transcripts*

481 LRS also has potential in discovering long transcripts, such as the sucrose synthase genes
482 annotated here. Even though numerous studies have defined the sucrose pathways, not all the
483 candidate genes have been identified. Many sucrose metabolism genes are too long to be
484 captured by short read sequencing without significant *de novo* assembly. For example, the *C.*
485 *arabica* *SS2* coding sequence is 2,889 bp and the genomic DNA sequence (exon 1 to 15) is
486 5,672 bp (Table 4). Sucrose synthase genes (6-7 different isoforms) were previously

1
2
3 488 [36-38]. For genes that were only previously available for *C. canephora*, (e.g. *SPI*), this
4
5 489 study also identified isoforms in Arabica. For genes that previously only had partial
6
7 490 sequences available, (e.g. *SPS2*), the transcripts identified in this study will guide further
8
9 491 studies and improve current databases. Furthermore, the low coverage annotation of long
10
11 492 sequences (>10kb) by BLASTx and BLASTn against the FOUR databases indicated the
12
13
14 493 limited information on long sequences requiring further study.
15
16

17 494 *Transcriptome analysis of polyploids using long-read sequencing*

18
19
20 495 LRS technologies show advantages in understanding complex transcriptomes, especially
21
22 496 from polyploid species [4, 39, 40]. First, this eliminates transcriptome reconstruction and that
23
24
25 497 reduces the computation time. This is an essential goal for bioinformatics data analysis and
26
27 498 software development [41]. To avoid obsolescence, transcriptome analysis calls for rapid
28
29
30 499 genomics and bioinformatics to reduce the time from experiment to publication. Secondly, as
31
32 500 there is no assembly of reads with LRS, there are no erroneous results due to misassemblies
33
34
35 501 caused by complex polyploid transcriptomes with a large number of repeats or homeolog
36
37 502 genes. For example, almost 80% of the wheat genome is repetitive [40]. Last but not least, it
38
39
40 503 shows the potential to capture rare or long sequences to provide an overview of the
41
42 504 transcriptome and fully characterise RNA diversity, like 5'UTR extension in this study,
43
44 505 alternative splicing, polyadenylation, etc. [4, 42].
45
46

47
48 506 However, LRS technologies have been normally biased with high error rates, for example,
49
50 507 previously released PacBio single molecule real-time sequencing (SMRT) reads had a very
51
52 508 high error rate, 11-14%, therefore, numerous methods have been proposed to correct the
53
54
55 509 sequences [43]. One common approach was to map back to a reference genome and (or) use
56
57 510 hybrid sequencing, for example, using short reads with high throughput to correct LRS
58
59
60 511 isoform sequences [5, 44]. However, caution is necessary when using this strategy. The
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

512 reference genome is often far from 100% accurate: 1) most draft genomes have numerous
513 fragmented contigs or scaffolds with huge imbedded gaps. Even genomes previously
514 considered well assembled have had many gaps[45]. 2) Problems also exist in poorly
515 assembled gene loci. Few recently released genomes have been re-visited to generate
516 improved assemblies [13]. 3) LRS isoform sequences normally come from different sources
517 (e.g. genotype) to the reference genomes that they can be compared with. Hybrid sequencing
518 correction may have system bias and result in loss of isoforms/transcript variants or generate
519 a “compromised” consensus. Previously, it has been estimated that there was no approach
520 that has achieved more than 60 % accuracy for transcript reconstruction, even for the most
521 studied human genome [46]. For instance, short read platforms deliver data that is less
522 representative of rare or long isoforms and there is a high chance of losing these reads from
523 the long-read dataset when correcting.

524 Improved accuracy may be generated from the platform itself, for example, Pacbio Iso-seq
525 generates improved accuracy from CCS reads. This allows multiple passes of each transcript.
526 Each pass can be used to correct the others with their random errors (mainly indels). The
527 isoform clustering and polishing in this protocol is expected to deliver 99% accuracy. Prior to
528 size selection, normalisation was further applied in parallel to the dataset in this study to
529 decrease the frequency of abundant reads and produce a more even representation of the
530 transcriptome and to capture rare sequences. A highly diverse transcriptome has resulted. The
531 abundance of genes that had not been previously sequenced (1,213), transcript variants and
532 longer isoforms indicate the limits of previous studies and potential of LRS technologies.
533 However, the limitation shows in detecting short sequences less than 300bp (raw data cut-
534 off). The chances of large errors due to indels from Pacbio sequencing may produce reads
535 shorter than the actual reads. Additionally, the Blue pippin size selection system starts from

536 500bp in the cDNA library preparation, with few sequences from the boundary (400-500bp).

537 Therefore, improvement is needed to capture a broader transcriptome.

538 In conclusion, this study will improve the understanding of the biology and genetic
539 improvement of polyploid species such as coffee. It provides a useful technique to generate a
540 full-length reference transcriptome and improve understanding of UTR regions.

541 **Additional information**

542 New sequence data used in this manuscript has been submitted to European Nucleotide
543 Archive at EMBL database with accession number: PRJEB19262. Additionally information
544 on specific selected sequence IDs, such as high quality annotated sequences, novel genes in
545 coffee, etc, are shown in supporting information 2.

546 **Completing interests**

547 All authors have no conflicts of interest to this manuscript.

548 **Funding**

549 This study was funded by Australian Research Council (PROJECT ID: LP130100376) and
550 Chinese Scholarship Council (2014-2018).

551 **Author's contributions**

552 B.C., A.F. and R.H. designed the research and discussed the results. B.C performed the
553 experiment and analysis. B.C drafted the manuscript, R.H and AG refined it.

554 **Acknowledgements**

555 We thank Green Cauldron Coffee (Australia) for providing coffee materials, Prathima
556 Perumal Thirugnanasambandam for assistance in sucrose synthase analysis, Kevin Smith and
557 Erli Wang for help in the informatics pipeline and the Research Computing Center of the

558 University of Queensland, Australia for access to high-performance computers. We also
1
2 559 appreciated the help from Poss Reading, Marta Brozynska, Adam Healey, Tiparat Tikapunya,
3
4
5 560 Ravi Nirmal, Nam Hoang and Hayba Badro in coffee sampling.
6
7

8 561 Reference 9

- 10 562 1. Yoo, M.-J., et al., *Nonadditive gene expression in polyploids*. Annual review of genetics, 2014.
11 563 **48**: p. 485-517.
- 12 564 2. Adams, K.L., et al., *Genes duplicated by polyploidy show unequal contributions to the*
13 565 *transcriptome and organ-specific reciprocal silencing*. Proceedings of the National Academy
14 566 of sciences, 2003. **100**(8): p. 4649-4654.
- 15 567 3. Levasseur, A. and P. Pontarotti, *The role of duplications in the evolution of genomes*
16 568 *highlights the need for evolutionary-based approaches in comparative genomics*. Biology
17 569 direct, 2011. **6**(1): p. 11.
- 18 570 4. Wang, B., et al., *Unveiling the complexity of the maize transcriptome by single-molecule*
19 571 *long-read sequencing*. Nature Communications, 2016. **7**.
- 20 572 5. Abdel-Ghany, S.E., et al., *A survey of the sorghum transcriptome using single-molecule long*
21 573 *reads*. Nature Communications, 2016. **7**.
- 22 574 6. Bicknell, A.A., et al., *Introns in UTRs: why we should stop ignoring them*. Bioessays, 2012.
23 575 **34**(12): p. 1025-1034.
- 24 576 7. Mignone, F., et al., *Untranslated regions of mRNAs*. Genome biology, 2002. **3**(3): p.
25 577 reviews0004. 1.
- 26 578 8. Van Veen, H., et al., *Transcriptomes of eight Arabidopsis thaliana accessions reveal core*
27 579 *conserved, genotype-and organ-specific responses to flooding stress*. Plant physiology, 2016:
28 580 p. pp. 00472.2016.
- 29 581 9. Garg, R., et al., *Transcriptome analyses reveal genotype-and developmental stage-specific*
30 582 *molecular responses to drought and salinity stresses in chickpea*. Scientific reports, 2016. **6**.
- 31 583 10. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a*
32 584 *reference genome*. Nature biotechnology, 2011. **29**(7): p. 644-652.
- 33 585 11. Wang, X.-W., et al., *De novo characterization of a whitefly transcriptome and analysis of its*
34 586 *gene expression during development*. BMC genomics, 2010. **11**(1): p. 400.
- 35 587 12. Li, P., et al., *The developmental dynamics of the maize leaf transcriptome*. Nature genetics,
36 588 2010. **42**(12): p. 1060-1067.
- 37 589 13. Michael, T.P. and R. VanBuren, *Progress, challenges and the future of crop genomes*. Current
38 590 opinion in plant biology, 2015. **24**: p. 71-81.
- 39 591 14. Denoeud, F., et al., *The coffee genome provides insight into the convergent evolution of*
40 592 *caffeine biosynthesis*. science, 2014. **345**(6201): p. 1181-1184.
- 41 593 15. Fridell, G., *Coffee*. 2014: John Wiley & Sons.
- 42 594 16. Furtado, A., *RNA Extraction from Developing or Mature Wheat Seeds*. Cereal Genomics:
43 595 Methods and Protocols, 2014: p. 23-28.
- 44 596 17. PacificBiosciences. *RS_IsoSeq (v2.3) Tutorial 2*. 2. Isoform level clustering (ICE and Quiver)
45 597 2015; Available from:
46 598 [https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-%28v2.3%29-](https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-%28v2.3%29-Tutorial-%232.-Isoform-level-clustering-%28ICE-and-Quiver%29)
47 599 [Tutorial-%232.-Isoform-level-clustering-%28ICE-and-Quiver%29](https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-%28v2.3%29-Tutorial-%232.-Isoform-level-clustering-%28ICE-and-Quiver%29).
- 48 600 18. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data*.
49 601 Bioinformatics, 2012. **28**(23): p. 3150-3152.
- 50 602 19. Afgan, E., et al., *Genomics Virtual Laboratory: a practical bioinformatics workbench for the*
51 603 *cloud*. PloS one, 2015. **10**(10): p. e0140829.

- 604 20. Mondego, J.M., et al., *An EST-based analysis identifies new genes and reveals distinctive*
1 605 *gene expression features of Coffea arabica and Coffea canephora*. BMC plant biology, 2011.
2 606 **11**(1): p. 1.
- 3 607 21. Dereeper, A., et al., *The coffee genome hub: a resource for coffee genomes*. Nucleic acids
4 608 research, 2015. **43**(D1): p. D1028-D1035.
- 5 609 22. Götz, S., et al., *High-throughput functional annotation and data mining with the Blast2GO*
6 610 *suite*. Nucleic acids research, 2008. **36**(10): p. 3420-3435.
- 7 611 23. Yuyama, P.M., et al., *Transcriptome analysis in Coffea eugenioides, an Arabica coffee*
8 612 *ancestor, reveals differentially expressed genes in leaves and fruits*. Molecular Genetics and
9 613 Genomics, 2016. **291**(1): p. 323-336.
- 10 614 24. Kearse, M., et al., *Geneious Basic: an integrated and extendable desktop software platform*
11 615 *for the organization and analysis of sequence data*. Bioinformatics, 2012. **28**(12): p. 1647-
12 616 1649.
- 13 617 25. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching*. Nucleic acids
14 618 research, 2009: p. gkp335.
- 15 619 26. Grillo, G., et al., *UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory*
16 620 *motifs of the untranslated regions of eukaryotic mRNAs*. Nucleic acids research, 2010.
17 621 **38**(suppl 1): p. D75-D80.
- 18 622 27. Nawrocki, E.P., et al., *Rfam 12.0: updates to the RNA families database*. Nucleic acids
19 623 research, 2014: p. gku1063.
- 20 624 28. Cheng, B., et al., *Influence of genotype and environment on coffee quality*. Trends in Food
21 625 Science & Technology, 2016.
- 22 626 29. Perrois, C., et al., *Differential regulation of caffeine metabolism in Coffea arabica (Arabica)*
23 627 *and Coffea canephora (Robusta)*. Planta, 2015. **241**(1): p. 179-191.
- 24 628 30. Ashihara, H. and A. Crozier, *Biosynthesis and catabolism of caffeine in low-caffeine-*
25 629 *containing species of Coffea*. Journal of agricultural and food chemistry, 1999. **47**(8): p. 3425-
26 630 3431.
- 27 631 31. Combes, M.C., et al., *Contribution of subgenomes to the transcriptome and their intertwined*
28 632 *regulation in the allopolyploid Coffea arabica grown at contrasted temperatures*. New
29 633 phytologist, 2013. **200**(1): p. 251-260.
- 30 634 32. Lodish, H., *Molecular cell biology*. 2008: Macmillan.
- 31 635 33. Rhind, N., et al., *Comparative functional genomics of the fission yeasts*. Science, 2011.
32 636 **332**(6032): p. 930-936.
- 33 637 34. Somers, J., T. Pöyry, and A.E. Willis, *A perspective on mammalian upstream open reading*
34 638 *frame function*. The international journal of biochemistry & cell biology, 2013. **45**(8): p.
35 639 1690-1700.
- 36 640 35. Hummel, M., et al., *Sucrose-mediated translational control*. Annals of botany, 2009: p.
37 641 mcp086.
- 38 642 36. Chen, A., et al., *Analyses of the sucrose synthase gene family in cotton: structure, phylogeny*
39 643 *and expression patterns*. BMC plant biology, 2012. **12**(1): p. 85.
- 40 644 37. Hirose, T., G.N. Scofield, and T. Terao, *An expression analysis profile for the entire sucrose*
41 645 *synthase gene family in rice*. Plant Science, 2008. **174**(5): p. 534-543.
- 42 646 38. Bieniawska, Z., et al., *Analysis of the sucrose synthase gene family in Arabidopsis*. The Plant
43 647 Journal, 2007. **49**(5): p. 810-828.
- 44 648 39. Minoche, A.E., et al., *Exploiting single-molecule transcript sequencing for eukaryotic gene*
45 649 *prediction*. Genome biology, 2015. **16**(1): p. 1.
- 46 650 40. Dong, L., et al., *Single-molecule real-time transcript sequencing facilitates common wheat*
47 651 *genome annotation and grain transcriptome research*. BMC genomics, 2015. **16**(1): p. 1039.
- 48 652 41. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity*
49 653 *platform for reference generation and analysis*. Nature protocols, 2013. **8**(8): p. 1494-1512.
- 50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

- 654 42. Gonzalez-Garay, M.L., *Introduction to isoform sequencing using pacific biosciences*
 1 655 *technology (Iso-Seq)*, in *Transcriptomics and Gene Regulation*. 2016, Springer. p. 141-160.
 2 656 43. Roberts, R.J., M.O. Carneiro, and M.C. Schatz, *The advantages of SMRT sequencing*. *Genome*
 3 657 *biology*, 2013. **14**(7): p. 1.
 4 658 44. Xu, Z., et al., *Full - length transcriptome sequences and splice variants obtained by a*
 5 659 *combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza*
 6 660 *and tanshinone biosynthesis*. *The Plant Journal*, 2015. **82**(6): p. 951-961.
 7 661 45. Lamesch, P., et al., *The Arabidopsis Information Resource (TAIR): improved gene annotation*
 8 662 *and new tools*. *Nucleic acids research*, 2012. **40**(D1): p. D1202-D1210.
 9 663 46. Korf, I., *Genomics: the state of the art in RNA-seq analysis*. *Nature methods*, 2013. **10**(12): p.
 10 664 1165-1166.

13
 14 **665 Tables and figure legends**

15 **666** Table 1 Arabica long-read sequencing transcriptome annotation with different databases

Databases	Number of sequences annotated	% of sequences annotated
Long-read sequencing transcriptome	95,995	-
BLAST	94,709	98.66
Mapped	78,571	81.85
InterProScan	70,774	73.73
InterProScan GOs	33,605	35.01
GO slim	58,050	60.47
KEGG	11,489	11.97

31
 32 **667**

33
 34 **668** Table 2 Arabica long-read sequencing isoforms compared to *Coffea canephora* coding sequences and *Coffea*
 35 **669** *arabica* EST sequences

Different datasets	GC content %	N50 (bp)	average length (bp)	min length (bp)	max_length (bp)	Number of sequences
<i>Coffea arabica</i> EST database ¹ [20]	44.7	734	662	32	3,584	35,153
<i>Coffea canephora</i> coding sequences with UTR ²	42.6	2,046	1,616	45	17,206	25,570
<i>Coffea arabica</i> long-read sequencing isoforms	41.4	4,865	3,236	301	23,335	95,995

48 **670**

49 **671** Note: ¹ <http://bioinfo03.ibi.unicamp.br/coffea/data/CA.fasta>; ² [http://coffee-genome.org/sites/coffee-](http://coffee-genome.org/sites/coffee-genome.org/files/download/coffea_cds.fna.gz)
 50 **672** [genome.org/files/download/coffea_cds.fna.gz](http://coffee-genome.org/files/download/coffea_cds.fna.gz).

51
 52 **673**

53
 54 **674**

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 3 Details of caffeine candidate genes, putative transcript variants annotated and 5'UTR extension information

Candidate genes	Accession number	Species	Source	Abbreviation	length (bp)	completeness	Putative transcript variants from LRS isoform sequences	5'UTR extension
xanthosine methyltransferase 1	AB048793	<i>C. arabica</i>	mRNA	CaXMT1	1,316	YES	c69597/f1p2/1421 c154338/f1p2/1360 c71416/f3p3/1376	YES
	JX978514	<i>C. arabica</i>	Genomic DNA	G-CaXMT1	1,987	YES		
	DQ422954	<i>C. canephora</i>	mRNA	CcXMT1	1,316	YES		
	JX978509	<i>C. canephora</i>	Genomic DNA	G-CcXMT1	1,994	YES		
xanthosine methyltransferase2	JX978515	<i>C. arabica</i>	Genomic DNA	G-CaXMT2	2,038	YES	Not identified	-
7-methylxanthine N-methyltransferase 1	AB048794	<i>C. arabica</i>	mRNA	CaMXMT1	1,298	YES	c20397/f5p1/1361	YES
	JX978511	<i>C. arabica</i>	Genomic DNA	G-CaMXMT1	1,838	YES		
	HQ616707	<i>C. canephora</i>	mRNA	CcMXMT1	1,222	YES		
	JX978507	<i>C. canephora</i>	Genomic DNA	G-CcMXMT1	1,829	YES		
7-methylxanthine N-methyltransferase 2	AB084126	<i>C. arabica</i>	mRNA	CaMXMT2	1,155	YES	c10402/f2p3/1277	YES
	JX978512	<i>C. arabica</i>	Genomic DNA	G-CaMXMT2	2,010	YES		
3,7-dimethylxanthine N-methyltransferase 1	AB084125	<i>C. arabica</i>	mRNA	CaDXMT1	1,155	YES	c25904/f2p0/977 c71881/f6p2/1386	YES
	JX978510	<i>C. arabica</i>	Genomic DNA	G-CaDXMT1	2,063	YES		
3,7-dimethylxanthine N-methyltransferase 2	KJ577793	<i>C. arabica</i>	mRNA	CaDXMT2	1,155	YES	c63815/f1p2/1273 c48759/f1p1/1517 c26870/f6p5/1402	YES
	KJ577792	<i>C. arabica</i>	Genomic DNA	G-CaDXMT2	2,006	YES		
	DQ422955	<i>C. canephora</i>	mRNA	CcDXMT1	1,364	YES		

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 4 Details of sucrose candidate genes, putative transcript variants annotated and 5'UTR extension information

Candidate genes	Accession number	Species	Source	Abbreviation	length (bp)	completeness	Putative transcript variants from LRS isoform sequences	5'UTR extension
Sucrose synthase 1	AM087674.1	<i>C. arabica</i>	mRNA	CaSS1	2,979	YES	c86432/f7p9/4842 c91298/f1p1/3137 c84406/f3p18/2975	YES
	DQ834312.1	<i>C. canephora</i>	mRNA	CcSS2	2,989	YES	c62911/f29p21/2965 c92344/f1p26/4662 c92296/f1p5/4676 c89510/f1p6/4592	
	AJ880768.2	<i>C. canephora</i>	Genomic DNA	G-CcSS1	3,957	exon 1-13	c106591/f2p0/4381 c72639/f25p28/2961	
Sucrose synthase 2	AM087675.1	<i>C. arabica</i>	mRNA	CaSS2	2,889	YES	c73322/f3p2/3080	YES
	AM087676.1	<i>C. canephora</i>	Genomic DNA	G-CcSS2	5,672	exon 1-15	c75363/f3p2/2906	
Sucrose phosphate synthase 1	DQ834321.1	<i>C. canephora</i>	mRNA	CcSPS1	3,150	YES	c51110/f2p0/3136	YES
	DQ842233.1	<i>C. canephora</i>	Genomic DNA	G-CcSPS1	8,215	YES		
Sucrose phosphate synthase 2	DQ842234.1	<i>C. canephora</i>	Genomic DNA	G-CcSPS2	1,550	NO	c103631/f1p2/4695 c88660/f2p0/4282 c106342/f1p4/4274 c104672/f1p1/4440 (reverse)	YES

Table 5 Results of 5' UTRs from long-read sequencing scanned with UTRdb. uORF, Upstream Open Reading Frame.

No.	Sequence name	5' UTR length (bp)	uORF
1	c86432/f2p7/4842	2,131	12
2	c91298/f1p1/3137	347	2
3	c84406/f3p18/2975	242	2
4	c62911/f29p21/2965	218	0
5	c92344/f1p26/4662	1,981	10
6	c92296/f1p5/4676	1,884	12
7	c89510/f1p6/4592	1,871	11
8	c106591/f2p0/4381	1,683	11
9	c72639/f25p28/2961	224	0

Figure 1 Coffee fruits of immature, intermediate and mature stages

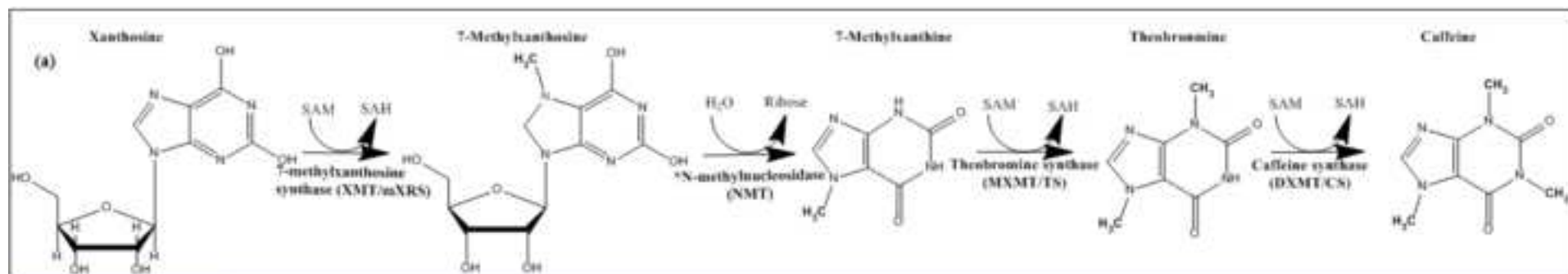
Figure 2 Sequence distribution, comparison of the number of sequences and their length. Coffee long-read sequencing isoforms, *C.canephora* coding sequences with UTR and *C. arabica* EST database were included.

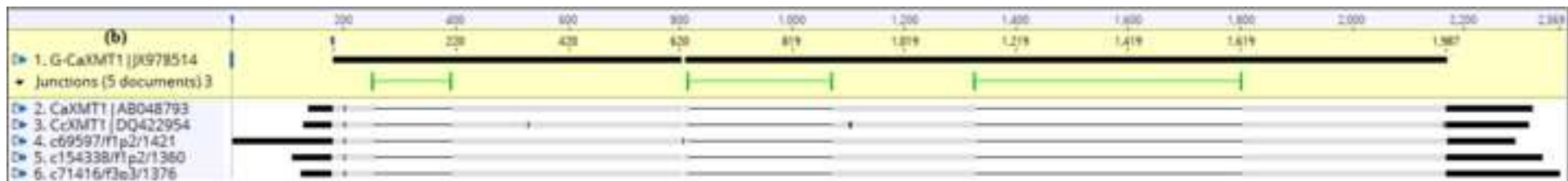
Figure 3 Putative transcript variants from long-read sequencing aligned to reference caffeine genes. a. Main caffeine biosynthesis pathway in coffee, adaptive from Cheng, Furtado [28]. b. Alignment of three Arabica putative XMT1 variants from long-read sequencing (c69597/f1p2/1412, c154338/f1p2/1360 and c71416/f3p3/1376), *Coffea arabica* and *Coffea canephora* XMT1 (CaXMT1 and CcXMT1) to Arabica XMT1 genomic DNA sequence (G-CaXMT1). c. Possible alternative polyadenylation of putative XMT1 Iso-seq variant (c25904/f2p0/977) from long-read sequencing; G-CaDXMT1, Arabica DXMT1 genomic DNA sequence; CaDXMT1, DXMT1 coding sequence; d. Two polyadenylation signals were identified in 3' ends of c25904/f2p0/977; e. Possible alternative splicing (intron retention) in one of the putative DXMT2 variants (c48759/f1p1/1517) from long-read sequencing transcripts; G-CaDXMT2, Arabica DXMT2 genomic DNA sequence; CaDXMT2, Arabica DXMT2 coding sequence. (Note: black colour in the alignment means different nucleotides to reference sequence, Arabica genomic XMT1, while grey colour means the same nucleotides as the reference.).

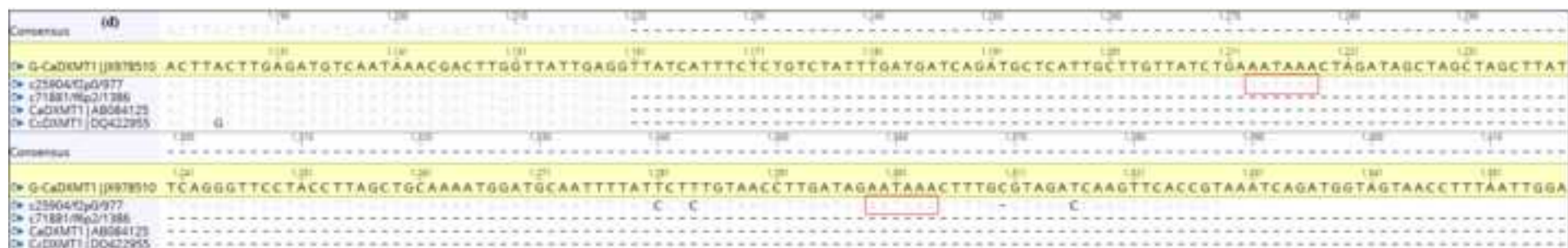
Figure 4 Motif search results of putative sucrose synthase gene 1 from long read sequencing. a. Ten motifs were annotated in 9 putative sucrose synthase 1 variants from long-read sequencing, analysed by MEME 4.11.2. b. Motif location of 9 putative sucrose synthase 1 variants. Different motifs were highlighted with red arrows and intron retention was shown with dashed boxes.

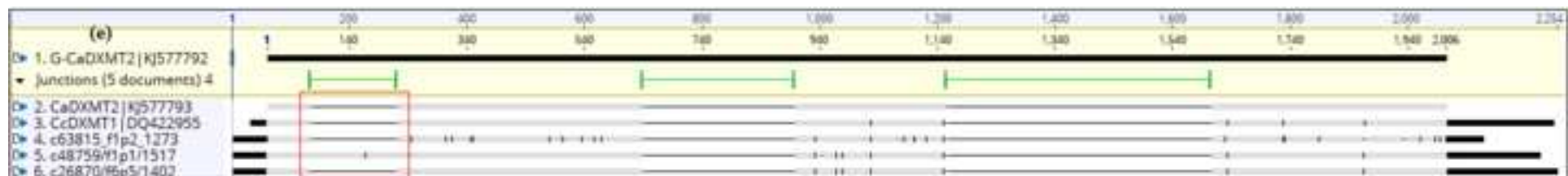
Figure 5 Putative variants from long-read sequencing aligned to the reference sucrose genes. a. Possible sucrose metabolism in coffee; SS, sucrose synthase; SPS, sucrose phosphate synthase; SP, sucrose phosphatase; INV, invertase; CINV, cell wall invertase (modified from Cheng B. et al. (2016)); b. Alignment of 9 Putative Sucrose synthase variants from long-read sequencing and *C.arabica* sucrose synthase gene 1 (CaSS1) to *Coffea canephora* genomic sucrose synthase 1 (exons 1-13) (G-CcSS1 (1-13)); Green box highlights variants result from different sub-genome copies, while intron retention events were marked with the blue box highlight; c. polyploid expression when zooming green area in 100%; d. possible alternative splicing (intron retention) from a *C.canephora* sub-genome copy when zooming blue box in 100%; e. possible intron retention from a *C.eugenioides* sub-genome copy when zooming blue area in 100%.red line classifies two groups of variants as different sub-genome copies. Different nucleotides compared to the consensus were highlighted in black in the alignment; f. Putative variants from long read sequencing aligned with *C.canephora* genomic sucrose phosphate synthase 2 sequence (G-CcSPS2); FWD, forward sequence; REV, reverse sequence. Different nucleotides compared to the consensus were highlighted in black in the alignment.

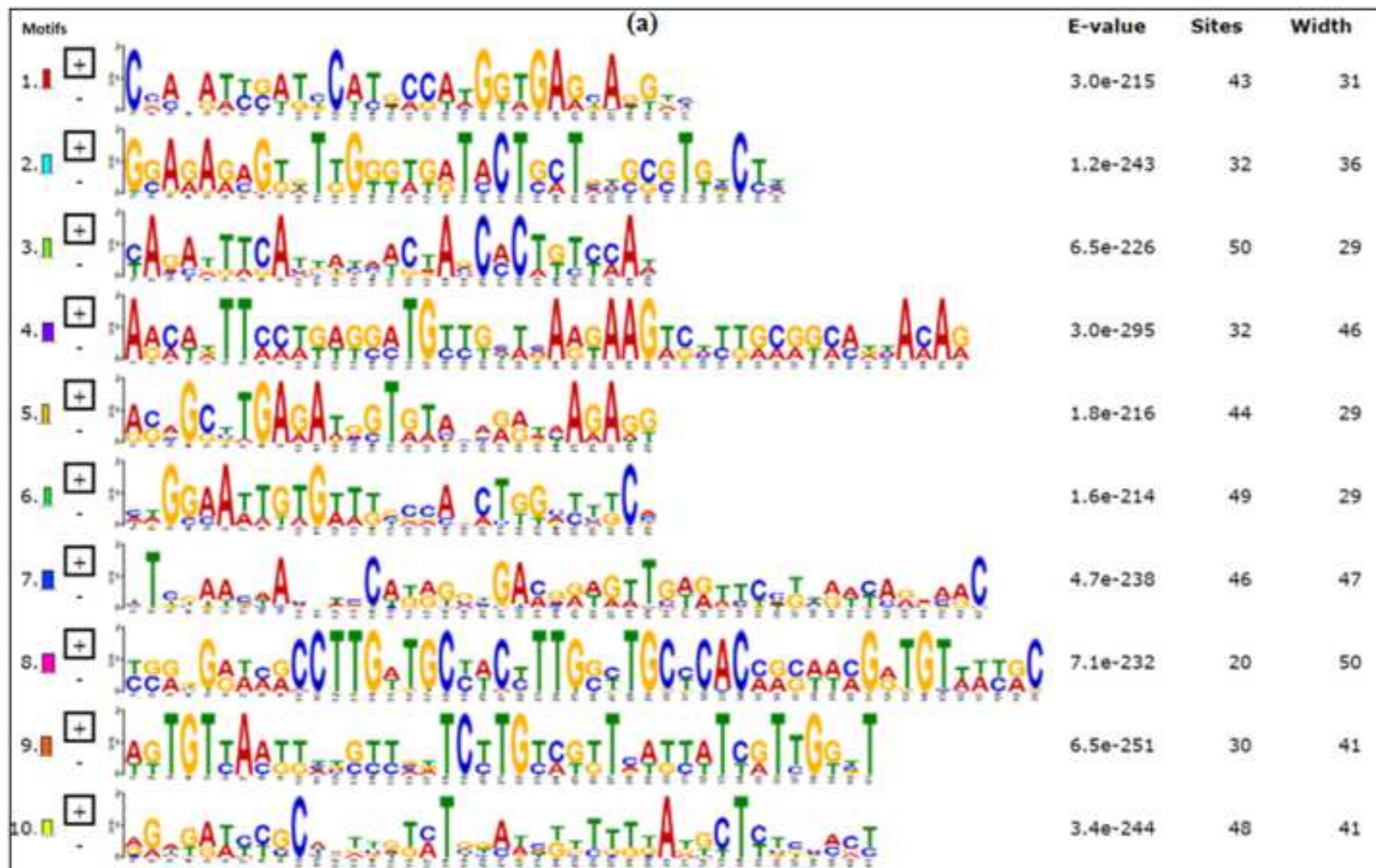


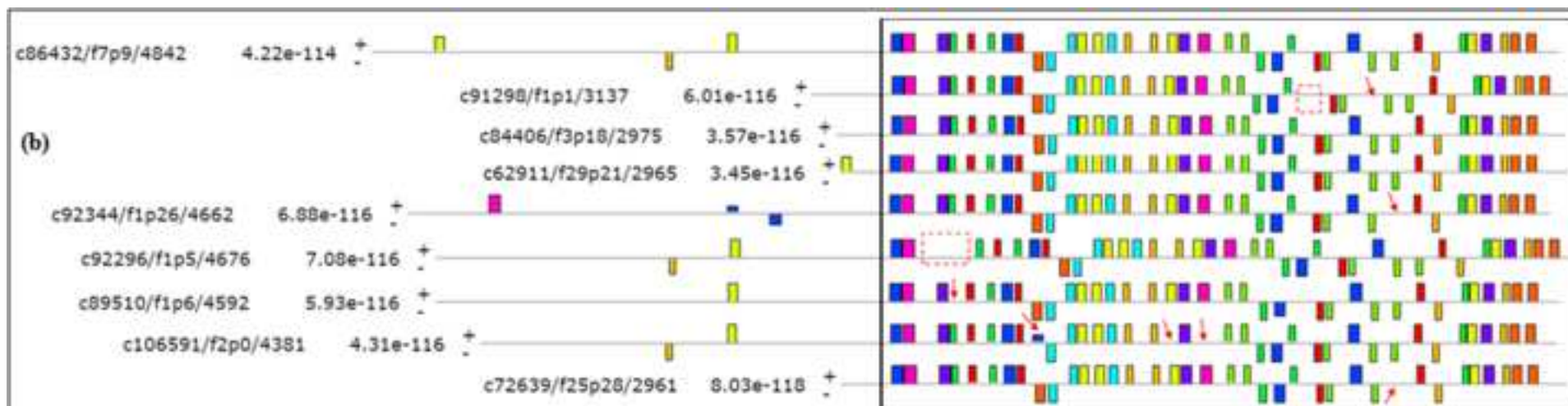


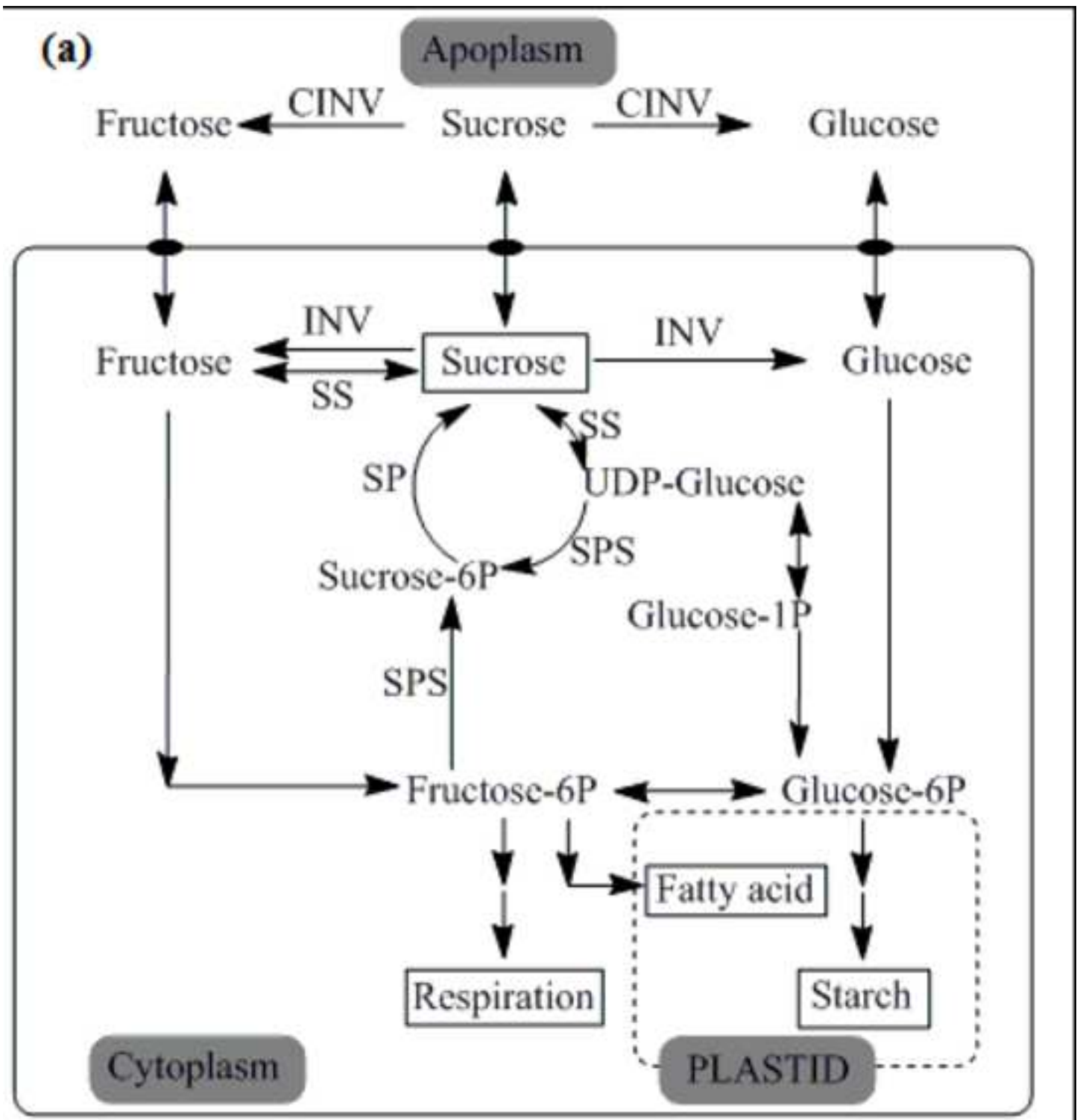






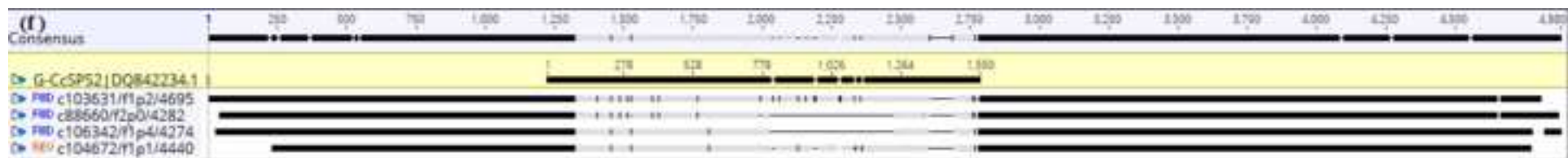




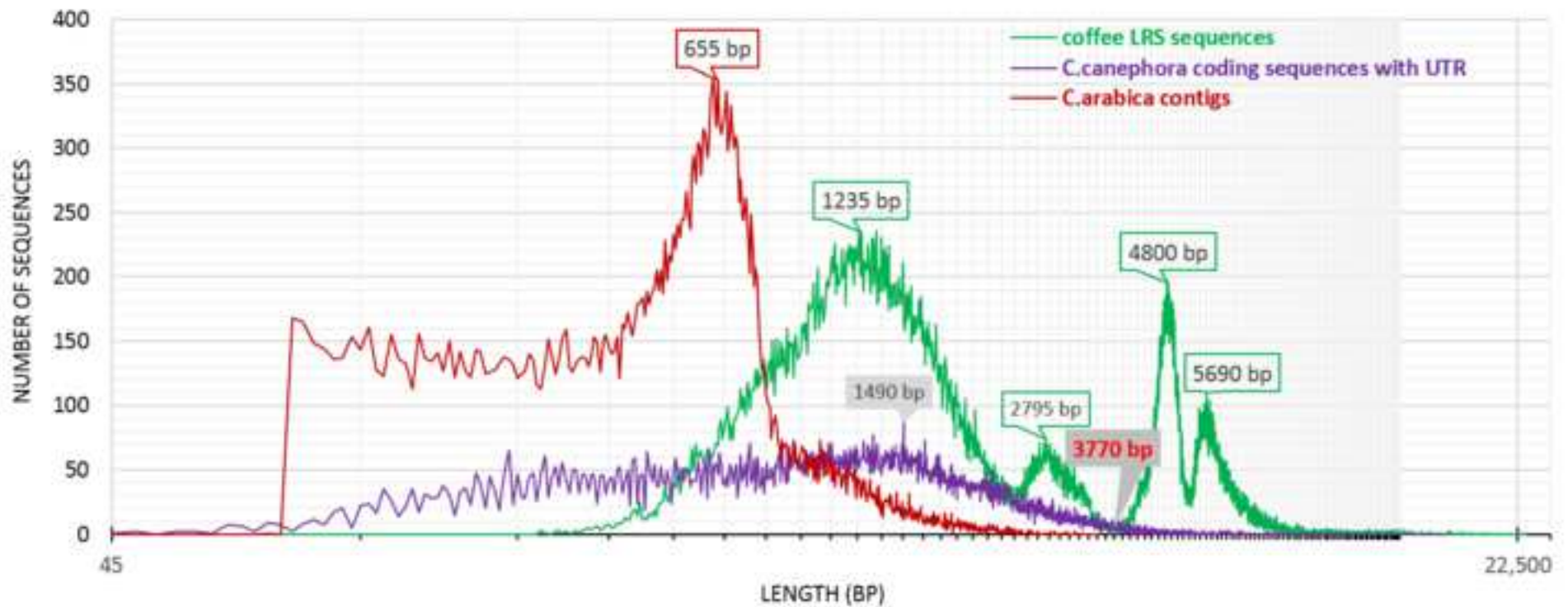








Distribution of number of sequences with length

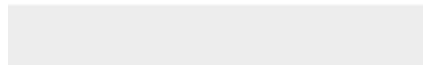




[Click here to access/download](#)

Supplementary Material

3. supporting information 1-for resubmission.pdf





Click here to access/download

Supplementary Material

4. supporting information 2-.xlsx

