High density linkage to physical mapping in a unique Tall x Dwarf Coconut (*Cocos nucifera* L.) F_2 population uncovers a major QTL for flowering time colocalized with the *FLOWERING LOCUS T (FT)*

Wellington Bruno dos Santos Alves[1,2], Cleso Antônio Patto Pacheco[3] and Dario Grattapaglia[1,2]

[1]EMBRAPA Recursos Genéticos e Biotecnologia, Brasília, DF, Brazil
[2]Programa de Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília, DF, Brazil
[3]EMBRAPA Tabuleiros Costeiros, Aracaju, SE

ORCID
Dario Grattapaglia: 0000-0002-0050-970X

*Corresponding author: dario.grattapaglia@embrapa.br

Keywords: Coconut, *FLOWERING LOCUS T*, hybrid breeding, SNPs, flowering

Running title: A flowering time QTL colocalized with FLOWERING LOCUS T in coconut

Abstract

The coconut tree crop (*Cocos nucifera* L.) provides vital resources for millions of people worldwide. Coconut germplasm is largely classified into 'Tall' (Typica) and 'Dwarf' (Nana) types. While Tall coconuts are outcrossing, more productive, stress tolerant and late flowering, Dwarf coconut are inbred, flower early with a high rate of bunch emission. Precocity determines earlier production of a plantation and facilitates management and harvest. We used a unique F_2 population generated by open pollination seeds collected in a commercial plantation of F_1 hybrids between Brazilian Green Dwarf from Jiqui (BGDJ) and West African Tall (WAT) to build a linkage map with ~3,000 SNPs, anchored to the existing chromosome-level genome assemblies. The map established the chromosomes numbering correspondence between the two reference genome versions and the relationship between recombination rate, physical distance and gene density. Leveraging the strong segregation for precocity inherited from the Dwarf cultivar in the F_2, a major effect QTL with incomplete dominance was mapped for flowering time. *FLOWERING LOCUS T* (*FT*) gene homologs of coconut previously described as putatively involved in flowering time by alternative splice variant analysis, were colocalized within a ~200kb window of the major effect QTL (LOD=11.86). Our work provides strong phenotype-based evidence for the role of the FT locus as the putative underlying functional variant for the flowering time difference between Dwarf and Tall coconuts. Major effect QTLs were also detected for developmental traits of the palm, plausibly suggesting pleiotropism of the *FT* locus for other precocity traits. Haplotypes of the two SNPs flanking the flowering time QTL inherited from the Dwarf parent BGDJ caused a reduction in the time to flower of around 400 days. These SNPs

could be used for high throughput marker assisted selection for early flowering and higher productivity recombinant lines, providing novel genetic material to the coconut industry.

Introduction

The coconut palm (*Cocos nucifera* L.) is a worldwide important tropical tree crop providing vital resources for millions of farmers and industries with a diverse use of its oil and fruit products. The only species belonging to the genus *Cocos*, the coconut palm is a diploid (2n = 32) monoecious, perennial monocotyledon of the family *Arecaceae* (Palmae), with a relatively large genome size (1C = 2.74 Gb). With its most likely center of origin in Southeastern Asia, it is currently cultivated in more than 92 tropical countries [1, 2] covering more than 12 million hectares worldwide [3].

Coconut germplasm is largely classified into 'Tall' (Typica) and 'Dwarf' (Nana) types. Tall, a.k.a. Giant coconuts, are preferentially outcrossing, genetically diverse and display high productivity, broad adaptability and tolerance to specific pests and diseases. Dwarf coconuts are mostly self-pollinating with limited genetic diversity within each variety but are chosen for their precocity and high rate of bunch emission. Early flowering of the dwarf varieties is a feature of great agronomic relevance that results not only in precocity of nut harvest, but also increased production, while facilitating coconut tree management and harvest operations. While the tall (late) variety takes an average of 8 years to flower, the early flowering dwarfs typically flower on average 2 to 4 years after planting [4]. The origin of dwarf coconuts has been a matter of debate, although current consensus based on molecular studies points toward a process of domestication from the ancestral tall coconut possibly following the appearance of

69 mutations in flowering time genes resulting in autogamy [4], as well gibberellin
70 biosynthesis affecting height growth [5, 6].

71 Despite its global socioeconomic importance, advances in coconut tree
72 improvement have been modest, hindered by its perennialism, long generational time,
73 late flowering and large tree size, which demand extensive experimental areas and
74 continuous long-term investments. Genetic improvement programs have adopted two
75 main breeding strategies: (1) intravarietal breeding coupled to mass selection and (2)
76 intervarietal hybridization, the most widely used strategy, which consolidates, into an
77 intermediate F_1 hybrid, the key features of each variety, such as precocity and short
78 stature from the dwarfs and the rusticity and production of larger fruits from the tall
79 varieties [2, 4]. Rare have been studies in F_2 populations derived either from selfing or
80 cross pollination among F_1 Tall x Dwarf hybrids. The wide segregation for production
81 traits reported in F_2 generations has discouraged such endeavors. Time to flowering,
82 pollination behavior, height and the presence of bole - an expansion at the base of the
83 stem - have been considered to segregate independently. While the former two traits
84 showed the involvement of several genes, height and bole character segregation
85 suggested the presence of at least one major gene each, with incomplete dominance [4,
86 7, 8], although no genetic follow-up testing of such hypothesis was ever reported.

87 Despite the socioeconomic importance of the coconut tree, advances in mapping
88 quantitative trait loci (QTLs) of economically important traits have been limited,
89 restricted to a few studies using Tall x Dwarf F_1 hybrid populations that were typically
90 readily available [9-13]. In such F_1 experimental populations, segregating variation can
91 only be captured from the outbred tall coconut, because the dwarf parent, being

92 essentially an inbred line, will not contribute allelic segregation. Flowering time as well
93 as other traits related to the dwarf behavior typically display an intermediate phenotype
94 in the F_1 progeny with no segregation that would allow genetic mapping of major effects
95 contributed by the dwarf genotype. Studies of F_2 or backcross populations are needed
96 to advance the molecular understating of key traits in coconut. Furthermore, genetic
97 mapping in the cited studies has been carried out either with low portability AFLP
98 markers or a limited set of microsatellites, which do not provide adequate sequence-
99 level information for deeper investigation of putative genes underlying mapped QTLs.
100 Only recently has this scenario begun to change with studies employing single
101 nucleotide polymorphism (SNP) markers for diversity studies [14], linkage mapping to
102 support genome assembly [15], and whole genome sequencing for genome-wide
103 studies (GWAS) [6]. Understanding the genetic basis of morphological traits that
104 determine precocity with a particular interest in early flowering, would represent an
105 important step towards breeding advancement of coconut.

106 In the present study, we used a unique segregating F_2 population generated by
107 open pollination of F_1 hybrids between Brazilian Green Dwarf from Jiqui (BGDJ) and
108 West African Tall (WAT) to dissect the quantitative variation underlying the genetic
109 divergence between tall and dwarf coconut types. We built a high-density linkage map
110 using SNPs derived from genotyping by sequencing with DArTseq, aligned the map to
111 the current genome assemblies and located QTLs for traits related to precocity, including
112 a major effect locus for flowering time colocalized with the *FLOWERING LOCUS T* (*FT*).

113 # Materials and methods

114 ## Plant material and DNA extraction

115 The study was carried out with 182 F_2 plants, generated by open pollination
116 among PB-141 F_1 coconut hybrid plants in a ten-year-old commercial plantation
117 managed by Embrapa Tabuleiros Costeiros, Aracaju – SE (10.9513° South, 37.0528°
118 West). PB-141 is an F_1 hybrid derived from the cross between the Brazilian Green Dwarf
119 from Jequi (BGDJ) and West African Tall (WAT). Coconuts were collected from F_1 palms
120 in the central area of the commercial plantation in order to maximize the probability of
121 being effectively derived from intercrossing of F_1 plants. F_2 plants were therefore
122 derived either from selfing or cross pollination between different F_1 plants. Samples of
123 young leaves were collected and used to extract total genomic DNA using an optimized
124 Sorbitol-CTAB-based protocol [16]. DNA concentrations and purity were estimated
125 using a Nanodrop 2000 spectrophotometer (Thermo Scientific).

126 ## SNP genotyping

127 Genomic DNA samples were shipped to the Service of Genetic Analysis for
128 Agriculture (SAGA) laboratory in Mexico for high-throughput genotyping using the
129 DArTSeq method developed by Diversity Arrays Technology Pty Ltd. [17]. DNA
130 complexity reduction was performed using the SbfI/MseI enzyme combination. Samples
131 were processed in digestion/ligation reactions and the scrambled fragments were
132 amplified using two primers with sequence complementary to the ligated adapters and
133 oligonucleotides from the sequencing platform. Subsequently, the clusters were
134 sequenced on an Illumina HiSeq2500 sequencing platform. SNP markers were detected
135 following mapping of sequence reads sampled at adequate depth on the Coconut

136 reference genome [15] according to standard parameters defined in DArTsoft14, an
137 automated genomics data analysis program and DArTdb, a laboratory management
138 system developed and patented by DArT for generating SNP marker data. Codominant
139 SNP markers were scored as “0” for homozygous reference allele, “1” for homozygous
140 alternative allele, and “2” for presence of both the reference and the alternative allele.

141 Morphological and flowering trait evaluations

142 Due to restrictions in field availability for planting, only 125 of the 182 F_2 plants
143 were ultimately planted in a field trial in a site known for its soil homogeneity at
144 Embrapa Tabuleiros Costeiros research center in June, 2018. Since to replicate plants
145 were possible for the 125 F_2 plants (n =125 entries), these were deployed in an
146 augmented randomized complete block design (ARCBD) [18] with c = 3 checks (random
147 plants of the parents BGDJ, WAT and the F_1 hybrid BGDJxWAT), r= 10 blocks extending
148 in rows. Because the number of entries was not a multiple of the number of blocks, nine
149 blocks had 13 entries and the 10th block had 8 entries. Nine traits were ultimately
150 evaluated in 121 live coconut trees: Number of Leaves (NL), Height of Second Leaf Hem
151 (cm) (HSLH), Petiole Length (cm) (PL), Rachis Length (cm) (RL), Petiole Width (cm) (PW),
152 Petiole Thickness (cm) (PT), emergence of the base of the stipe (BOLE), Number of
153 Bunches with fruits (NBF) and Days to Launch of the First Spathe (DLFS). BOLE was
154 recorded in a binary fashion with presence of bole when it measured at least 70 cm from
155 the ground, and absence when at a lower that 70 cm height from the ground. DLFS was
156 used as a direct quantitative measure of time to flowering. These traits, but BOLE, are
157 typically used to differentiate dwarf from tall plants in the early stages of life.
158 Measurements were taken according to the standardized methods described by Santos,

159 Batugal (19). Trait evaluations were carried out at one and two years after planting, on
160 June 30, 2019 and July 30, 2020 with evaluation of morphological leaf traits. Flowering
161 initiation was measured on each individual plant according to the emergence of the first
162 spathe starting on October 11, 2020 and then by a monthly monitoring to precisely catch
163 the launch of the first spathe. The number of days for the emergence and opening of
164 the subsequent spathes until the tenth spathe was recorded until the least evaluation
165 on August 11, 2022. The final evaluations of the morphological traits were carried out
166 on May 10, 2022, with the exception of NBF, which was evaluated on August 11, 2022.
167 Pearson correlations were estimated to assess the correlation between traits using the
168 Corrplot package available in R [20]. Phenotypic data for all traits but BOLE was analyzed
169 with the following mixed model: $Y_{ij} = \mu + B_i + E_j + C(_i)_j + e_{ij}$ where Y is the observed
170 phenotypic value of the genotype entry (j) in block (i), μ is the overall mean, B is the
171 fixed block effect, E is the random genotype entry effect, C is the random effect of the
172 common checks in blocks and e is the random residual effect. Best linear unbiased
173 predictors (BLUPs) were estimated using the software SELEGEN [21] under model 76.

174 Linkage map construction

175 The raw SNP calls were filtered for quality for the linkage mapping analyses. The
176 first filtering step was based on a call rate >0.95 followed by selection of SNP markers
177 that fit the expected segregation proportions in an F_2, by a Chi-square test under the
178 null hypothesis of a 1:2:1 segregation ratio with a more relaxed threshold (α>0.001) to
179 allow including SNPs even with some level of segregation distortion given the
180 particularity of the mapping population. Linkage analysis and construction of the genetic
181 map were performed according to the standard F_2 model using JoinMap v3.0 [22]. The

182 genotypes of the F_2 population were coded, according to the format <hkxhk> for an
183 unknown linkage phase configuration and three possible genotypic classes. Markers
184 were assigned to linkage groups (LG) via the clustering module with log-of-probability
185 (LOD) scores greater than 15, maximum recombination fraction of 0.4, under a Kosambi
186 mapping function. Linkage maps were drawn and aligned to the assembled genomes
187 using MAPCHART [23]. For the alignments, recombination distances were expressed in
188 centiMorgans and physical distances in Mbp.

189 Marey maps and recombination rates in the genome of *Cocos*
190 *nucifera*

191 To estimate local recombination rates across the 16 coconut chromosomes, the
192 targeted SNP marker sequences were aligned to the two existing chromosome-level
193 coconut genome assemblies of Hainan Tall coconut: (1) a draft assembly built using
194 exclusively short Illumina paired-end sequencing [15] hereafter called the YANG genome
195 and (2) a reference-grade assembled using long Nanopore single-molecule sequencing
196 and Hi-C technology [6] hereafter called the WANG genome. The FASTA sequences of
197 the chromosome assemblies were downloaded from their respective repositories. The
198 YANG genome from GenBank BIOPROJECT PRJNA374600, and the WANG genome from
199 the Genome Warehouse in the National Genomics Data Center (Beijing Institute of
200 Genomics) accession number GWHBEBT00000000, BIOPROJECT PRJCA005463. The
201 DArTseq sequences with 69 bases were mapped to the sequences of the 16 coconut
202 pseudo-chromosomes assembled using BLASTn for highly similar sequences (megablast)
203 with a cutoff E-value <1E-05. DArTSeq sequences that mapped to a unique location in
204 both genome versions were used in further analyses. The relationships between the

genetic and physical positions of the mapped SNPs were represented by the local slope of the curve adjusted to the datapoints in MAREY maps [24]. Curves were fitted to the data using locally weighted scatterplot smoothing (LOESS) polynomial regression with a smoothing parameter of 0.3 using MareyMap Online [25]. Sporadic abnormal points from the map that did not fit the expected monotonously increasing relationship between physical and genetic positions were removed using the available functionality in the software. The physically mapped markers were used to obtain an estimate of the effective physical coverage achieved by the genetic map and the recombination rates in cM/Mb along the chromosomes. Pearson's correlations along windows of 1 Mb were calculated between the average recombination rate in the 1 Mb interval and the number of genes as annotated in the coconut chromosomes in the YANG genome [15].

QTL mapping

QTLs detection was performed using the BLUPs for eight traits and the binary 0/1 value for BOLE with R/qtl version 3.5.3, using different methods: interval mapping (IM), Haley Knott regression (HK) and non-parametric interval mapping for those traits that violated the assumption of normality [26]. The probability of genotype error (error.prob = 0.01) was calculated and the threshold logarithm of odds (LOD) score for QTL detection was determined by genome-wide LOD significance thresholds, setting LOD thresholds at different alpha levels (α = 0.01; 0.05 and 0.1) using 1900 permutations to allow for the declaration of QTLs with variable stringency levels in order to control also for Type II errors.

Co-localization of the FLOWERING LOCUS T (FT) gene with the flowering time QTL

A set of 198 expressed gene sequences in coconut were recently identified as homologues of flowering-time genes in *Arabidopsis*. Five of them, annotated as FT homologs to the *Arabidopsis thaliana* FT gene (AT1G65480) were found to be differentially expressed between seedling and reproductive leaf samples in both tall and dwarf coconut. Alternative splicing analysis showed that the FT gene produces different transcripts in tall compared to dwarf coconut [27]. We took all the 87 *Arabidopsis thaliana* genes present in the annotation of the 198 coconut expressed genes and aligned them to the QTL region on the corresponding chromosomes assembled in the two genome assemblies using BLASTn specifically in the genome stretch from 65 Mbp to 75 Mbp where the major effect QTL was mapped (see below).

Results

SNP markers and linkage mapping

Genotyping by sequencing using DArTseq generated 86,719 raw SNP calls based on the standard parameters used in DArTsoft. After applying a call rate threshold of 95% and a test for adherence to a null hypothesis of 1:2:1 segregation, 3,714 SNPs (4.28% of the raw SNP calls) were retained for further linkage analyses (S1 File). Dominant presence/absence in silico DArTseq variants were not used due to their low information content and considering that an adequate number of co-dominant informative SNPs was already available for linkage mapping purposes. A linkage map was ultimately built with

247 182 F_2 individuals that correspond to 364 meiotic events sampled in the gametes from
248 the recombined F_1 parents. All 3,714 SNPs were mapped with LOD ≥15 and maximum
249 recombination fraction of 0.4, in 16 linkage groups matching the expected 16 coconut
250 chromosomes. However, for the subsequent assessment of the recombination rates and
251 QTL mapping, only the SNPs that were physically positioned on the genome assemblies
252 (see below) were ultimately considered on the linkage map (S2 File). Linkage group
253 numbering used for our map was based on the physical positioning of SNP markers on
254 the YANG genome chromosomes which was published first in the literature. The
255 chromosome numbering used in the WANG genome was different but the
256 correspondence was established and reported throughout (Table 1). An average of
257 184.5 SNPs were mapped to each chromosome, varying from a maximum of 329 SNPs
258 on chromosome 1 to a minimum of 101 SNPs on chromosome 15. The genetic map
259 spanned a total of 2,124.3 cM in recombination size, with an average linkage map length
260 of 132.8 cM, and an average intermarker recombination distance between adjacent
261 marker pairs of 0.742 cM. Chromosome 15 displayed a large recombination gap of 31.2
262 cM between two denser linkage clusters although still linked by the minimum LOD 15
263 score threshold adopted (Fig 1). Except for this large gap observed on chromosome 15,
264 maximum intermarker distance was 6.7 cM on chromosome 7.

265 **Fig 1.** Alignment of the linkage map (green) to the two chromosome-level genome
266 assemblies, the YANG genome assembly (red) [15] and the WANG genome assembly
267 (blue) [6].

268 Alignment of the linkage map to the genome assembly

269 The 3,714 DArTseq sequences containing the linkage mapped SNPs were
270 submitted to BLASTn against the chromosome assemblies. On the YANG genome out of
271 the 3,714 linkage mapped DArTSeq SNPs, 2,952 DArTSeq sequences were mapped to
272 unique locations with E-values <7.3e-05. The sequences for the remaining SNPs either
273 aligned to multiple locations or did not align to the assembled coconut chromosomes
274 and were removed from further analyses. On the WANG genome 3,027 DArTSeq
275 sequences were mapped to unique locations with E-values <3.6e-11 (S3 File). On the
276 YANG genome, the linkage map covered 1,035 Mb corresponding to a coverage of
277 approximately 43.14% of the 2,400 Mb *Cocos nucifera* Tall cultivar genome [6]. On the
278 WANG genome, the linkage mapped SNPs covered a total of 2,382 Mb, a 99.27%
279 coverage, more than twice the coverage obtained on the YANG genome with a
280 corresponding lower density of mapped SNPs and long stretches with no segregating
281 SNPs mapped to them. The average estimated physical inter SNP distance were 349kb
282 and 779kb, SNP density of 2.85 and 1.27 SNPs/Mb and maximum physical distances of
283 9.7 and 63.4 Mb respectively on the YANG and WANG genomes (Table 1).

Table 1. Statistics of the linkage map and its alignment to the two available chromosome-level genome assemblies of coconut (*Cocos nucifera* L.).

LINKAGE MAP				YANG GENOME						WANG GENOME			
LG #	# of link. map SNPs	Total Size (cM)	Mean interSNP dist (cM)	Chr. #	Phys. dist. covered (Mb	Mean interSNP phys. dist. (Mb)	Number of annotated genes#	Average recomb. rate (cM/Mbp)	r^2 (Number of genes and recomb. rate)	Chr. #	Phys. dist. covered (Mb)	Mean interSNP phys. dist. (Mb)	Average recomb. rate (cM/Mbp)
1	329	196.6	0.599 ± 0.470	1	119.197	0.362 ± 0.652	2277	2.27 ± 1.17	0.68	5	177.52	0.570 ± 1.533	2.12 ± 0.94
2	311	199.1	0.642 ± 0.492	2	85.353	0.275 ± 0.411	1997	2.44 ± 0.66	0.72	1	213.40	0.656 ± 2.751	1.97 ± 0.649
3	209	149.2	0.717 ± 0.625	3	82.325	0.396 ± 0.576	1705	1.98 ± 0.65	0.73	4	179.12	0.835 ± 2.527	1.67 ± 1.26
4	222	161.2	0.729 ± 0.536	4	66.748	0.302 ± 0.454	1610	3.05 ± 1.52	0.87	6	170.88	0.762 ± 2.005	2.56 ± 1.46
5	150	147.5	0.989 ± 0.654	5	57.717	0.385 ± 0.624	1104	3.77 ± 2.06	0.89	9	156.70	1.073 ± 5.397	2.98 ± 1.87
6	262	193.4	0.729 ± 0.764	6	81.468	0.311 ± 0.463	2095	2.42 ± 0.74	0.46	3	187.69	0.693 ± 2.894	1.84 ± 0.50
7	215	148.8	0.695 ± 0.686	7	77.453	0.361 ± 0.758	1583	2.13 ± 0.60	0.21	2	188.13	0.776 ± 2.790	1.60 ± 0.76
8	176	95.7	0.547 ± 0.505	8	58.631	0.326 ± 0.668	1214	2.09 ± 1.16	0.77	13	95.96	0.508 ± 1.273	1.56 ± 0.75
9	106	89.9	0.856 ± 0.725	9	59.333	0.554 ± 0.792	1026	2.37 ± 1.33	0.89	11	137.92	1.378 ± 3.871	2.12 ± 1.38
10	144	113.0	0.790 ± 0.493	10	57.105	0.399 ± 0.663	1066	2.90 ± 1.45	0.87	10	144.69	0.969 ± 4.013	2.39 ± 1.53
11	165	72.4	0.785 ± 0.750	11	61.809	0.375 ± 0.617	1379	3.01 ± 1.93	0.61	7	173.86	1.073 ± 3.547	2.64 ± 1.77
12	133	94.9	0.728 ± 0.489	12	42.367	0.302 ± 0.425	868	3.21 ± 1.86	0.86	14	89.87	0.681 ± 2.625	2.88 ± 1.76
13	127	102.6	0.814 ± 0.617	13	60.488	0.479 ± 1.032	847	3.77 ± 2.55	0.87	15	82.46	0.641 ± 1.677	3.63 ± 2.05
14	189	132.8	0.706 ± 0.544	14	56.198	0.296 ± 0.351	1309	3.15 ± 1.35	0.73	8	155.26	0.754 ± 3.799	2.66 ± 1.30
15	101	135.7	1.357 ± 3.139	15	31.114	0.311 ± 0.800	646	5.18 ± 3.33	0.04	12	145.36	1.075 ± 3.394	3.04 ± 1.43
16	113	91.5	0.819 ± 0.664	16	38.139	0.313 ± 0.623	867	3.97 ± 1.48	0.91	16	83.33	0.686 ± 2.239	3.34 ± 1.58
Total	2,952	2,124.3	-		1035.445	-	21,593	2.79 ± 1.65			2,382.141	-	2.32 ± 1.40
Mean ± s.d.	184.5 ± 69.9	132.8 ± 40.8	0.742 ± 0.471		43.14%	0.349 ± 0.652					99.27%	0.779 ± 2.947	

Number of annotated genes from (Yang et al. 2021; BioProject: PRJNA374600)

287 ## Marey maps and recombination rates

288 The linkage maps alignment to the 16 pseudochromosomes was largely colinear
289 with the physical assembly, with only sporadic inconsistencies of the map versus
290 genome positions, 2 in the YANG genome and 20 in the WANG genome (Fig 1; S4 File).
291 As expected, on all linkage groups the alignment to the physically larger WANG genome
292 showed long sections of genome sequence with few or no SNP markers mapped. The
293 Marey maps showed plateaus of restricted recombination, where the recombination
294 distance does not follow the increase in physical distance along the chromosome (Fig 2).
295 These plateaus spanned considerably longer stretches for the Marey maps on the WANG
296 genome. From the Marey maps data, estimates of local recombination rate were
297 obtained for both genomes using LOESS (Locally Weighted Scatterplot Smoothing) (S4
298 File) and plotted for the YANG genome (Fig 2). Following LOESS treatment of the data,
299 the average local recombination rates were similar: 2.78 cM/Mb for the YANG genome
300 and 2.32 cM/Mb for the WANG genome. The smoother recombination rate curves
301 closely followed expectations. Chromosomal segments of restricted recombination in
302 the Marey maps coincided with local valleys of low recombination rate notably spotted
303 on YANG chromosomes 1, 4, 5, 11 and 14, for example and more accentuated on the
304 corresponding WANG chromosomes.

305

306 **Fig 2.** Marey maps showing the correspondence between the physical position (x-axis,
307 Mb) and the recombination-based position (y-axis, cM) on the 16 Coconut chromosomes
308 of the YANG genome assembly [15] and the WANG genome assembly [6]. Also shown in

309 the intermediate panels the combined pattern of gene density distribution and the
310 LOESS curve of recombination rate for the YANG genome.

311 Average recombination rates per chromosomes varied across chromosomes up
312 to more than twofold (e.g., 1.98 cM/Mb on chromosome 3 and 5.18 cm/Mb on
313 chromosome 15 on the YANG genome). However, the correlation between the rates
314 estimated on the two genome versions was high and significant (r = 0.85; p-value = 3E-
315 5). The highest rate on chromosome 15 is likely the result of the large linkage gap
316 observed on this chromosome. The average recombination rate across the entire
317 genome were similar, at 2.79 and 2.32 cM/Mb for the two genomes (Table 1).
318 Furthermore, taking advantage of the gene annotation provided for the YANG genome
319 sequence, we underlaid the frequency distribution of the number of genes in 1 Mb
320 intervals, equivalent to gene density, to the smoothed recombination rate curve in the
321 same 1 Mb segment (Fig 2). Visually an almost perfect direct relationship was seen
322 between gene density and recombination rate. This visual relationship was further
323 validated by observing high and significant Pearson correlation between average
324 recombination rate and number of genes in the 1 Mb intervals estimated for the YANG
325 genome. For all chromosomes but chromosomes 7 and 15 these correlations were high
326 and significant (p-value <0.001) (Table 1) suggesting, once again, some different
327 behavior as far as recombination for these two chromosomes.

328 Trait variation and correlations

329 The F_2 individuals displayed extensive segregation for all traits measured, as
330 expected in a recombinant F_2 population of divergent cultivars. Raw phenotypic values
331 and BLUPs are provided in S5 File. The continuous traits showed an approximately

332 normal distribution, although with some skewness, most notably for NBF and DLFS (Fig
333 3). BOLE expression fit a 3:1 segregation ratio with 98 plants showing no bole and 23
334 with bole (Chi-square 2.31; p-value = 0.128). Days to Launch of the First Spathe (DLFS)
335 varied from a minimum of 812 days for the precocious flowering plants to a maximum
336 of 1608 days for the late blooming ones with an average of 1287 ± 214.5 days. One plant
337 therefore displayed 2.2 standard deviations below the mean and 13 flowered with less
338 than 1000 days, corresponding to more than one standard deviation below the mean.
339 Eight of these very early flowering plants displayed the highest numbers (≥13) fruit per
340 bunch (NBF). Significant negative correlations were seen between DLFS and NL, HSLH,
341 BOLE and NBF, and a significant positive correlation with PL (Fig 4). Positive significant
342 correlations were seen between five of the six leaf morphology traits (NL, HSHL, RL, PW
343 and PT), except PL that showed no correlation with all other five traits. Positive and
344 significant correlations were also observed between NBF and NL, HSLH, RL and BOLE.
345 Coconut trees with a recorded bole flowered on average after 1098 days and produced
346 an average of 9.5 fruit per bunch while plants without bole flowered on average after
347 1347 days, and produced 5.96 fruit per bunch.

348

349 **Fig 3.** Histogram of the coconut phenotypic trait distributions in the F_2 mapping
350 population.

351

352 **Fig 4.** Matrix of genotypic correlations among the measured traits in the F_2 mapping
353 population

354

355 **Table 2.** Summary of QTL mapping results in the coconut F2 population.

	Days to Launch of First Spathe (DLFS)	Bole emergence (BOLE)	Number of bunches with fruit (NBF)	Petiole Length (PL)
Chromosome	3	3	1	3
LOD score threshold (percentile)	5.14 (99%)	4.47 (95%)	5.14 (99%)	5.14 (99%)
# SNPs in QTL region	52	1	1	22
QTL region position (cM)	118.928 - 149.166	132.62	57.16	123.503 – 136.125
Right SNP	100287206_T>A-55	-	-	100287206_T>A-55
Left SNP	100273632_T>G-15	-	-	100273632_T>G-15
Peak LOD	11.86	4.64	5.68	7.79
Peak LOD SNP	100226499_C>T-34	100228320_G>A-40	100237259_A>G-48	100226499_C>T-34
Target SNP allele (Origin; effect)	T (Dwarf; reduction)	A (Dwarf; increase)	A (Dwarf; increase)	T (Dwarf; reduction)
Genome position (bp)	73,147,443	75,613,473	19,206,916	73,147,443
Variance explained (%)	64.0	3.12	0.43	22.3

356

357 ## Mapping and haplotype analysis of a major flowering time QTL

358 QTL detection was carried out with three methods, IM, CIM and HK to assess
359 consistency of results and minimize type I errors. Significant QTL peaks were declared
360 based on the LOD limit analysis for the different alpha levels to allow the detection of
361 suggestive QTLs in order to minimize Type II errors. The significance thresholds indicated
362 LOD limits of 4.15, 4.49 and 5.22 respectively for the different alpha levels (α = 0.01;
363 0.05 and 0.1) (Table 2). Results with CIM showed a considerably larger number of 16
364 putatively significant QTLs in 11 chromosomes when compared to the results with IM
365 and HK. These CIM results are not discussed further but are provided (S6 File). Genome-
366 wide QTL profile results across the 16 chromosomes obtained with the three QTL
367 detection methods, for the four traits for which significant QTLs were detected are
368 provided for a full visualization of results. The LOD limit thresholds are specified by the
369 differentially dotted lines (Fig 5).

370

371 **Fig 5.** Genome-wide QTL mapping profiles for the mapped traits (Y axis) obtained with
372 the three different QTL mapping methods implemented (X axis)

373

374 Overall, consistent QTLs across the methods were detected on chromosome 3
375 for DFLS, BOLE and PL and on chromosome 1 for NBF. A genomic segment at the tip of
376 chromosome 3 spanning 52 significantly linked SNPs to DFLS stands out as a major effect
377 QTL for flowering time, explaining an estimated 64% of the variation in DLFS. The
378 alignment of the linkage map to the chromosome 3 of the YANG genome and

379 chromosome 4 of the WANG genome allowed a detailed haplotype analysis of this
380 segment. In the YANG genome starts at SNP 100285601_C>G-14 with LOD 4.99 located
381 at 117.99 cM corresponding to 68.84 Mbp, and ends at SNP 100270789_A>G-59 with
382 LOD 6.53 located at 148.17 cM corresponding to position 82.32 Mbp (S3 File). In the
383 WANG genome it spans 55 SNPs including the 52 mapped to the YANG genome form
384 position 162.7 Mb to the end of the chromosome at 178.5 Mb. Within this segment, the
385 highest QTL peak (LOD 11.86) is located at SNP 100226499_C>T-34, positioned at 129.11
386 cM and 73.15 Mbp, flanked by SNPs 100287206_T>A-55 and 100273632_T>G-15. At SNP
387 100226499_C>T the T allele inherited from the Dwarf parent AVeBrJ causes a major
388 reduction in the time to flower, suggesting incomplete dominance of the locus.
389 Heterozygous genotype G/T confers a reduction from 1460 to 1390 days to the launch
390 of the first spathe, while the homozygous T/T genotype confers a further reduction of
391 almost 300 days, as the coconut trees bearing this genotype flowered on average after
392 1110 days (Fig 6). In addition to acting on early flowering, the same genomic segment is
393 involved in the control of BOLE and PL. The LOD peak for both QTLs for BOLE and PL was
394 located at the same SNP 100228320_G>A-40, at 132.62 cM and 75.61 Mbp. The co-
395 localization of QTLs for three traits suggests a pleiotropic action of this major effect QTL
396 segment on chromosome 3. Some additional suggestive QTLs were detected for PL on
397 chromosomes 1, 2, 4, 6 that were significant at $\alpha = 0.1$ based on the IM analysis but that
398 did not reach significance with the two other QTL detection methods. Finally, a single
399 QTL was also found for NBF on chromosome 1 whose Dwarf allele confers an increase
400 in the number of fruits per bunch.

401 **Fig 6.** Detailed view of the major effect QTL region for flowering time, petiole length and
402 bole presence in the stem on chromosome 3 of the YANG genome. The *A. thaliana*
403 FLOWERING LOCUS T was colocalized with the leading SNPs for the QTL.

404 Co-localization of the flowering time QTL with the FLOWERING 405 LOCUS T (FT)

406 Out of the 87 *Arabidopsis thaliana* genes involved in flowering-time pathways
407 that were found to be homologous to 198 coconut genes [27], only the complete coding
408 sequence of the *Arabidopsis thaliana* putative flowering signals mediating protein FT
409 (At1g65480) produced strongly significant alignment to the flowering time QTL mapped
410 on linkage group 3. The FT gene (At1g65480) complete cds mRNA (accession
411 AY065378.1) with 840 bp produced two hits in the YANG genome: (i) at 70,554,038 to
412 70,554,249 bp with bit score 108, 71% identity, e-value = 3e-23 and 0/212 gaps (0%) and
413 (ii) at 73,265,231 to 73,265,450 bp with bit score 88.7, 70% identity, e-value = 3e-17 and
414 3/222 (1%) gaps (Fig 6; S7 File). These hits are located respectively at 137 kb from SNP
415 100241292_A>G-33 LOD 9.55 and 117 kb from SNP 100226499_C>T-34 this last one had
416 the highest LOD score (LOD 11.86) in the QTL region. When the FT gene (At1g65480)
417 was blasted against the corresponding chromosome 4 in the WANG genome it also
418 produced two significant hits at 164.688 Mb (e-value = 7E-21) and 168.065 Mb (e-value
419 = 7E-15) that correspond to the genomic positions at the tip of the chromosome (Fig 1;
420 S8 File), further corroborating the finding. This co-localization result between the
421 flowering time QTL and the FLOWERING LOCUS T (FT) gene when taken together with
422 the splice variant analysis reported earlier [27], strongly suggests that the underlying

423 candidate gene to the flowering time QTL mapped in this study could be an FT homolog
424 in *Cocos nucifera.*

425 Discussion

426 We have generated a dense linkage map for *Cocos nucifera* based on a novel
427 approach for the crop, by using an F_2 population derived from open pollination
428 intercrossing of F_1 hybrids between Brazilian Green Dwarf from Jequi (BGDJ) and West
429 African Tall (WAT). Almost 3,000 genotyping by sequencing DArTSeq sequence derived
430 SNPs were positioned on the existing chromosome-level coconut genome assemblies.
431 By aligning the linkage map to the genome sequences, we provided an independent
432 assessment of the contiguity of the current coconut reference genomes, and a detailed
433 analysis of the relationship between recombination rate and physical distance for the
434 corresponding coconut chromosomes. Furthermore, based on the segregation observed
435 for traits related to precocious behavior inherited from the Dwarf cultivar, we mapped
436 a major effect QTL for flowering time on chromosome 3. Flowering Locus T (FT) gene
437 homologs of coconut, previously described as involved in flowering time by alternative
438 splice variant analysis, were co-localized with the molecular phenotype mapped QTL,
439 providing evidence for its role as the putative underlying gene.

440 Exploiting the unique genetics of the Tall x Dwarf coconut hybrid
441 for linkage mapping in an outbred F_2.

442 The linkage map reported in our study is the first one in coconut built from an F_2
443 segregating population from a Tall x Dwarf hybrid. Previous maps either used hybrid F_1
444 or backcross (BC1) populations [9, 10, 13, 15] or a cross between Tall cultivars[11].

445 Differently from a conventional F_2 populations generated by selfing an individual F_1
446 plant, the particular genetics of coconut allowed using an open pollinated progeny of an
447 F_1 hybrid field. The parental F_1 plants composed a commercial production plantation
448 with seeds resulting from controlled pollination between a pollen mix of several Tall
449 coconut parents on female flowers of several Dwarf plants. In the particular case of
450 coconut, we took advantage of the fact that the Dwarf variety is essentially a fixed inbred
451 line with very little if any residual heterozygosity, while the Tall parents are allogamous,
452 heterozygous and genetically diverse [6, 14, 28]. The F_1 plants SNP genotypes would
453 potentially differ only in what was inherited from the Tall parents, but SNP markers fixed
454 in the Tall variety and polymorphic in relation to the Dwarf would segregate in a 1:2:1
455 ratio in the F_2, irrespective of what Tall parent generated each different F_1 plant
456 intercrossed to produce the F_2 population. This premise proved valid, and 3,714 SNPs
457 called with a call rate >95%, fit the expected segregation and were linkage mapped.
458 Approximately 20% of the sequences of linkage mapped DArTSeq mapped to multiple
459 locations in the two genomes and were left out of subsequent analysis, as no objective
460 criteria allowed choosing one location over another. Therefore, for QTL mapping
461 analyses we used only around 3,000 linkage mapped SNPs that could be anchored to
462 unique positions on the 16 assembled chromosomes of the two genome versions.

463 The linkage map assembled in the 16 expected linkage groups spanned a total of
464 2,124.3 cM, slightly smaller than the linkage map reported earlier by Yang et al. [15] with
465 2,365 cM built from a backcross (BC_1) population with 216 individuals and 8,402 SNPs
466 retained with >80% call rate. Linkage group recombination sizes of our map were also
467 similar to those reported by Yang et al. in the BC_1 map (Spearman rank correlation 0.72),
468 including the only large recombination gap in chromosome 15, starting at around the 40

469 cM position (Fig 1). This recombination gap was hypothesized to be due to reduced
470 heterozygosity resulting from a selection sweep in favor of the allele transmitted from
471 the Dwarf parent observed in the YANG genome assembly [15].

472 Although our map was built from a smaller number of individuals than the BC_1
473 map used to anchor the YANG genome [15], those were derived from two recombined
474 gametes from the F_1 parents, therefore sampling twice as many meiotic events as in a
475 BC1 population. The number of recombination events per meiosis captured in the
476 mapping population is one of the main drivers of mapping accuracy, with an F_2 being the
477 most efficient configuration [29, 30]. Additionally, for improved linkage map quality, we
478 decided to use only SNPs with call rate >95% aware of the potential reproducibility
479 challenges of marker data derived from genotyping by sequencing methods in
480 heterozygous genomes [31]. From the raw data, 10,576 DArTSeq markers were retained
481 at >95% call rate, and the main detractor of the final SNPs number was the test for the
482 1:2:1 segregation. This substantiates the variability in the F_1 parents contributed by the
483 Tall grandparental pollen pool. This observation indicates that a truly self-pollinated
484 progeny from a single F_1 plant would possibly allow mapping a considerably larger
485 number of DArTSeq SNP markers.

486 The F_2 linkage map provides an independent assessment of the 487 properties of the coconut genome assemblies

488 The DArTseq method using the rare cutter SbfI combined to the frequent MseI
489 enzyme proved effective for complexity reduction and enrichment of the sequenced
490 pool for low copy genomic regions across most of the large coconut genome, generating

491 a large number of segregating markers [17]. Our mapping and data quality control
492 approach proved effective to order SNPs markers along the linkage groups into robust
493 maps, highly colinear with the genome sequences (Table 1). The physical distance
494 covered by our F_2 linkage map was 1035.4 Mb on the YANG genome, slightly higher than
495 the 1,020 Mb of the Yang et al. BC_1 map, while on the WANG genome, the mapped SNPs
496 covered 2,382 Mb, reaching a 99.27% coverage. As expected, very few segregating
497 DArTSeq sequences and the corresponding SNPs were mapped to long physical
498 stretches covering several hundred Mb of DNA of the reference-grade WANG genome
499 assembly produced by nanopore long read sequencing (Fig 1). These regions were not
500 sampled in the YANG genome, assembled exclusively with short read sequencing. The
501 extensive genome gaps with no mapped SNPs and low gene density (Fig 2), likely
502 correspond to heterochromatin, typically found at centromeres and telomeres, known
503 to be relatively gene poor, mostly consisting of repetitive DNA sequences [32], where
504 the DArTSeq method by principle tends not to sample restriction fragments. Performing
505 linkage mapping in those heterochromatic regions will prove challenging, requiring
506 approaches that can target the occasional low-copy regions scattered in the repetitive
507 DNA. Using a selfed F_1 progeny might prove valuable in this respect to sample much
508 larger number of segregating DArTSeq SNPs, possibly targeting low-copy interspersed
509 segments in these heterochromatic regions of the coconut genome.

510 The alignment of the linkage maps to the genome assembly indicated consistent
511 ordering between the Joinmap estimated positions of the SNPs, and their relative
512 physical order along the assembled chromosomes, with only sporadic inconsistencies
513 seen most notably at the tip of some chromosomes (Fig 1). These localized
514 inconsistencies could be due either to erroneous linkage map ordering or misassembled

515 genome scaffolds. Historically, the majority of the inconsistencies between the physical
516 and genetic map order pointed to errors in the physical map order [33], making linkage
517 maps effective tools to correct genome assemblies [34]. However, sequencing
518 technologies and assembly algorithms have improved tremendously in recent years. The
519 alignment of the linkage map to two independent genome assemblies, one of them built
520 with long reads, indicates that those few SNPs that show inconsistent positions in
521 relation to both genome assemblies, for example at the tips of chromosomes 8, 9 and
522 12, most likely were inaccurately mapped.

523 The largely consistent genetic to physical position of the mapped SNPs led us to
524 carry out further analysis of this relationship for each coconut chromosome by
525 estimating local recombination rates in cM/Mb, given by the slope of the curve built
526 with graphical Marey maps. Additionally, the gene density histogram distribution
527 annotated in the YANG genome was merged to the Marey maps for a combined view
528 (Fig 2). In all chromosomes, the Marey maps showed genomic stretches of constrained
529 recombination, where the recombination distance does not follow the linear increase in
530 physical distance and gene density drops. This observation was evidently more distinct
531 in the WANG genome Marey maps that included the heterochromatic regions of the
532 genome. These regions are typically considered to be the putative location of
533 centromeres. Considering the WANG chromosome numbering for now, these were
534 located more toward the extremes of the chromosome (chromosomes 3, 7, 8, 9, 10, 12,
535 13 and 16) that would correspond to acrocentric or even telocentric chromosomes. In
536 the other chromosomes these regions are found more toward the middle (Fig 2).
537 Centromeres are essential for faithful segregation of chromosomes both in mitosis and
538 meiosis and are typically rich in repetitive sequences in many eukaryotes [35].

539 Equivalent observations to ours on the putative centromere positions was reported in
540 the YANG genome [15], establishing their correlation to the presence of repetitive rich
541 DNA regions that were challenging to assemble. A few localized negative recombination
542 rate estimates were also observed when the LOESS curve under passed the red dotted
543 line, for example on chromosomes 7, 8, 9 and 12 (Fig 2). These localized negative
544 estimates coincide with the inconsistent genetic to physical map alignments as
545 described above and are usually located at the tip of chromosomes. All our results taken
546 together, on the collinearity between our linkage map and both physical genomes
547 provide an independent experimental validation of the genome assemblies currently
548 available for coconut and establish the correspondence of chromosome numbering used
549 in the two reports (Fig 2).

550 Colocation of a major effect QTL with the FLOWERING LOCUS T 551 (FT) provides evidence for its role in early flowering

552 Little is known about the underlying mechanisms of early flowering in dwarf
553 coconut. Knowledge on the complex physiology and genetics of flowering initiation has
554 come mainly from annual plants including the model species *Arabidopsis* as well as crop
555 plants, which typically flower and set seeds within a single year [36, 37]. Less is known
556 about flowering time in perennial plants which typically have distinguishable transition
557 from a juvenile to an adult phase, the developmental stage that determines the
558 emergence of flower meristems. It is now well established, however, that the formation
559 of flower structures is largely conserved across flowering plants, annual and perennial,
560 with relatively minor variations in the role of key developmental genes [38]. Amongst
561 the several genes identified as involved in regulation of flowering time, the expression

562 of FLOWERING LOCUS T (FT) and its homologues has been shown to accelerate the onset
563 of flowering in a number of plant species [39]. During the course of our study, we came
564 across a transcriptomics report in coconut showing that the *FLOWERING LOCUS T* (FT)
565 gene generates different transcripts in tall compared to dwarf coconut, with the shorter
566 splice variant of FT present exclusively in dwarf varieties but absent in most tall coconuts
567 [27]. The authors suggested that this FT splice variant could be the putative causal
568 mechanism for flowering time differentiation between the dwarf and tall coconut types,
569 although they recommended further investigation and validation to confirm this effect.
570 By collocating a phenotype-derived major QTL, detected agnostically to any prior gene-
571 centered hypothesis, our study provides independent validation of the FT gene as the
572 underlying major functional variants controlling flowering time between the two
573 coconut varieties.

574 The *FT* locus likely controls other precocity traits in Dwarf 575 coconut

576 A number of traits were targeted in previous QTL mapping studies in coconut,
577 such as yield, fruit components, early germination and cuticular wax [9-13]. Flowering
578 time, however could not be approached because it did not segregate in intermediate F_1
579 hybrid families and less so in progeny of crosses between Tall cultivars. Two studies to
580 date have reported on the patterns of genetic variation in F_2 populations of coconut
581 hybrids, although without attempting QTL mapping. Ample segregation was reported
582 for a number of vegetative and morphological flowering traits, including height growth
583 and the presence of a bole, for which the presence of at least one major gene each, with
584 incomplete dominance was suggested [4, 7]. Curiously, despite its importance as a trait

585 directly linked to production, time to flower was not assessed. Our study is therefore
586 the first one to tackle flowering time in coconut by a QTL mapping approach facilitated
587 by the strong segregation seen in the F_2. In our experiment we could not, however,
588 measure height growth with confidence and include this trait in the mapping work.
589 Recently a GWAS was carried out in a collection of Tall and Dwarf cultivars identifying a
590 major effect QTL for height growth [6]. The leading SNP for the GWAS hit was mapped
591 to WANG chromosome 12, co-located in the promoter region of the gibberellin (GA)
592 biosynthetic enzymes GA 20-oxidases (GA20ox) further backed by expression analysis,
593 validating previous reports [5]. These results strongly support the role of GA20ox in
594 controlling height growth difference between Tall and Dwarf coconuts. Taken together
595 with the detection of a major flowering time QTL on chromosome 4 of WANG genome
596 (chromosome 3 of YANG) (Figs 5 and 6) and its colocation with the FT gene, these results
597 also suggest that height growth and flowering time in coconut involve different major
598 effect loci in line with the early proposal of two independent major loci [4] , although it
599 cannot be ruled out that the FT gene is also involved in height growth (see below).

600 Petiole length (PL) and the presence of a bole (BOLE) were also assessed in our
601 study. Both traits are typically used as morphological markers to identify dwarf plants
602 [2]. In fact, PL and BOLE showed significant direct (0.466) and inverse (-0.502) correlation
603 with flowering time (DLFS) (Fig 4). Major effect QTLs were also detected for these traits
604 in the same exact region on chromosome 3 suggesting that the *FT* gene might also be
605 affecting these traits (Fig 5). The *FT/TFL1* gene family has been extensively studied in
606 flowering plants. It constitutes a major target of evolution in nature, as a single essential
607 gene that has played a central role in plant diversification and adaptation [40]. FT
608 homologs have been identified in many species, demonstrating general conservation of

609 functions across gymnosperms and angiosperms. It encodes a major florigen, a key
610 flowering hormone in controlling flowering time. Studies in a number of species have
611 revealed diverse roles of the FT/TFL1 gene family in plant developmental processes
612 other than flowering regulation, with a distinct effect on stem and leaf development,
613 mainly in the inhibition of stem and leaf growth [41, 42]. It is therefore plausible that *FT*
614 is directly affecting PL and the formation of a bole structure in coconut as well.

615 Conclusions: implications for coconut advanced breeding lines

616 From the conventional coconut breeding perspective, offspring of F_1 hybrids
617 would in principle not be desirable for commercial production due to the wide variation
618 in yield traits among F_2 coconut trees. Nevertheless, the two previous studies of F_2
619 populations have reported individuals displaying high husked nut weights, indicating
620 little or no inbreeding depression, likely resulting from the inbred nature of the ancestral
621 Dwarf grandparent. These results indicated good potential of extracting recombined
622 lines from F_2 populations with desirable characters [7, 8]. Still, no developments have
623 been made to implement this breeding strategy in coconut possibly due to the long time
624 necessary to develop and eventually propagate such recombined lines.

625 The results of our study open a fundamentally new perspective in terms of
626 marker assisted selection (MAS) in coconut breeding. Knowledge of the position and
627 the genotypes at the SNP markers linked to the flowering time QTL would easily allow
628 selecting hybrid F_2 plants homozygous for the early flowering haplotype. Homozygosity
629 for the favorable alleles at SNPs 100226499_C>T-34, 100287206_T>A-55 and
630 100273632_T>G-15, inherited from the Dwarf parent BGDJ caused a reduction in the
631 time to flower around 400 days in our experiment. Our data also showed a significant

632 negative correlation between the necessary days to flowering (DLFS) and NBF (Fig 4).
633 Early flowering plants displayed some of the highest numbers of fruit per bunch and a
634 separate QTL was mapped in chromosome 1 which could be selected by a set of linked
635 SNPs flanking the lead SNP 100237259_A>G-48 (Table 2). This small set of SNPs could be
636 used for high throughput inexpensive MAS screening of F_2 plants at ultra-early stages of
637 development or even at the embryo stage. Juvenile tissue of these selected individuals
638 could be induced to somatic embryogenesis [43] or, more successfully, to shoot culture
639 [44] for clonal propagation. This would capture the full genetic value of these plants
640 allowing regenerating thousands of plants indefinitely, providing the coconut industry
641 with a solution for their current need of homogeneous quality planting material.

642 Acknowledgments

643 We would like thank the entire staff of field technicians EMBRAPA Tabuleiros Costeiros
644 who were involved in establishing and managing the several field trials, and collecting
645 and organizing the data used in this study. We thank Dr. Orzenil B. Silva-Junior and Dr.
646 Roberto Togawa for assistance with the BLAST analysis against the YANG and Wang et
647 al. genomes respectively.

648 References

649 1. Gunn BF, Baudouin L, Olsen KM. Independent Origins of Cultivated Coconut (Cocos
650 nucifera L.) in the Old World Tropics. Plos One. 2011;6(6). doi: 10.1371/journal.pone.0021143.
651 PubMed PMID: WOS:000292033700039.
652 2. Batugal P, Bourdeix R, Baudouin L. Coconut Breeding. In: Jain SM, Priyadarshan PM,
653 editors. Breeding Plantation Tree Crops: Tropical Species. New York, NY: Springer New York;
654 2009. p. 327-75.
655 3. FAOSTAT statistical database [Internet]. Food and Agricultural Organization, Rome (Date
656 accessed: March 10, 2022). 2021 [cited March 10, 2022]. Available from:
657 https://search.library.wisc.edu/catalog/999890171702121.
658 4. Perera L, Baudouin L, Mackay I. SSR markers indicate a common origin of self-pollinating
659 dwarf coconut in South-East Asia under domestication. Scientia Horticulturae. 2016;211:255-62.
660 doi: 10.1016/j.scienta.2016.08.028. PubMed PMID: WOS:000386409800031.

5. Boonkaew T, Mongkolsiriwatana C, Vongvanrungruang A, Srikulnath K, Peyachoknagul S. Characterization of GA20ox genes in tall and dwarf types coconut (Cocos nucifera L.). Genes & Genomics. 2018;40(7):735-45. doi: 10.1007/s13258-018-0682-4.

6. Wang S, Xiao Y, Zhou Z-W, Yuan J, Guo H, Yang Z, et al. High-quality reference genome sequences of two coconut cultivars provide insights into evolution of monocot chromosomes and differentiation of fiber content and plant height. Genome Biology. 2021;22(1):304. doi: 10.1186/s13059-021-02522-9.

7. Namboothiri CGN, Niral V, Parthasarathy VA, Muralidharan K. Segregation, phenotypic and genotypic variance and heritability in the F2 population of coconut (*Cocos nucifera*). Indian Journal of Agricultural Sciences. 2011;81(10):954-6.

8. Fernando WMU, Perera P. Evaluation of genotypes arising from F2 generation segregations in dwarf × tall crosses of cocos nucifera. Cord. 1997;13(2):26-35.

9. Herrán A, Estioko L, Becker D, Rodriguez MJB, Rohde W, Ritter E. Linkage mapping and QTL analysis in coconut (Cocos nucifera L.). Theoretical and Applied Genetics. 2000;101(1):292-300. doi: 10.1007/s001220051482.

10. Lebrun P, Baudouin L, Bourdeix R, Konan JL, Barker JHA, Aldam C, et al. Construction of a linkage map of the Rennell Island Tall coconut type (Cocos nucifera L.) and QTL analysis for yield characters. Genome. 2001;44(6):962-70. doi: 10.1139/gen-44-6-962. PubMed PMID: WOS:000172264900006.

11. Riedel M, Riederer M, Becker D, Herran A, Kullaya A, Arana-Lopez G, et al. Cuticular wax composition in Cocos nucifera L.: physicochemical analysis of wax components and mapping of their QTLs onto the coconut molecular linkage map. Tree Genetics & Genomes. 2009;5(1):53-69. doi: 10.1007/s11295-008-0168-7. PubMed PMID: WOS:000261265100005.

12. Baudouin L, Lebrun P, Konan JL, Ritter E, Berger A, Billotte N. QTL analysis of fruit components in the progeny of a Rennell Island Tall coconut (Cocos nucifera L.) individual. Theoretical and Applied Genetics. 2006;112(2):258-68. doi: 10.1007/s00122-005-0123-z. PubMed PMID: WOS:000234026000007.

13. Ritter E, Rodriguez MJB, Herran A, Estioko L, Becker D, Rohde W. Analysis of quantitative trait locis (QTL) based on linkage maps in Coconut (Cocos nucifera L.). In: Arencibia AD, editor. Plant Genetic Engineering: Towards the Third Millennium. Developments in Plant Genetics and Breeding. 52000. p. 42-8.

14. Muñoz-Pérez JM, Cañas GP, López L, Arias T. Genome-wide diversity analysis to infer population structure and linkage disequilibrium among Colombian coconut germplasm. Scientific Reports. 2022;12(1):2958. doi: 10.1038/s41598-022-07013-w.

15. Yang Y, Bocs S, Fan H, Armero A, Baudouin L, Xu P, et al. Coconut genome assembly enables evolutionary analysis of palms and highlights signaling pathways involved in salt tolerance. Communications Biology. 2021;4(1):105. doi: 10.1038/s42003-020-01593-x.

16. Inglis PW, Pappas MdCR, Resende LV, Grattapaglia D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLosOne. 2018;13(10):e0206085. doi: 10.1371/journal.pone.0206085.

17. Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, et al. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. BMC Proceedings. 2011;5(Suppl 7):P54. PubMed PMID: doi:10.1186/1753-6561-5-S7-P54.

18. Crossa J, Federer † W. I.4 Screening Experimental Designs for Quantitative Trait Loci, Association Mapping, Genotype-by Environment Interaction, and Other Investigations. Frontiers in Physiology. 2012;3.

19. Santos GA, Batugal P, Othman A, Baudouin L, Labouisse J-P. Manual on standardized research techniques in coconut breeding. 1996.

20. Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J, et al. R package 'corrplot': Visualization of a Correlation Matrix (Version 0.92). 2021.

21. Resende MDV. Software Selegen-REML/BLUP: a useful tool for plant breeding. Crop Breed Appl Biotechnol. 2016;16(4):330-9. doi: 10.1590/1984-70332016v16n4a49. PubMed PMID: WOS:000391702900009.
22. Van Ooijen JW, Voorrips RE. JoinMap 3.0, Software for the calculation of genetic linkage maps. P R International. 2001; Wageningen, The Netherlands.
23. Voorrips RE. MapChart: Software for the graphical presentation of linkage maps and QTLs. Journal of Heredity. 2002;93(1):77-8. doi: DOI 10.1093/jhered/93.1.77. PubMed PMID: ISI:000175627300018.
24. Chakravarti A. A Graphical Representation of Genetic and Physical Maps - the Marey Map. Genomics. 1991;11(1):219-22. doi: Doi 10.1016/0888-7543(91)90123-V. PubMed PMID: ISI:A1991GA48800029.
25. Siberchicot A, Bessy A, Guéguen L, Marais GA. MareyMap Online: A User-Friendly Web Application and Database Service for Estimating Recombination Rates Using Physical and Genetic Maps. Genome Biology and Evolution. 2017;9(10):2506-9. doi: 10.1093/gbe/evx178.
26. Broman KW, Wu H, Sen Ś, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19(7):889-90. doi: 10.1093/bioinformatics/btg112 %J Bioinformatics.
27. Xia W, Liu R, Zhang J, Mason AS, Li Z, Gong S, et al. Alternative splicing of flowering time gene FT is associated with halving of time to flowering in coconut. Scientific Reports. 2020;10(1):11640. doi: 10.1038/s41598-020-68431-2.
28. Santos P, Venancio TM, dos Santos PHD, Ramos HCC, Aredes FAS, Azevedo AON, et al. Genotyping-by-sequencing technology reveals directions for coconut (Cocos nucifera L.) breeding strategies for water production. Euphytica. 2020;216(3). doi: 10.1007/s10681-020-02582-1. PubMed PMID: WOS:000519369500001.
29. Ferreira A, da Silva MF, Silva L, Cruz CD. Estimating the effects of population size and type on the accuracy of genetic maps. Genetics and Molecular Biology. 2006;29(1):187-92. doi: 10.1590/s1415-47572006000100033. PubMed PMID: WOS:000236280300033.
30. Liu HB. Statistical Genomics, Linkage, Mapping and QTL Analysis. Boca Raton, Florida: CRC; 1998. 611 p.
31. Myles S, Chia JM, Hurwitz B, Simon C, Zhong GY, Buckler E, et al. Rapid Genomic Characterization of the Genus Vitis. Plos One. 2010;5(1):e8219. doi: 10.1371/Journal.Pone.0008219. PubMed PMID: ISI:000273554600001.
32. Allshire RC, Madhani HD. Ten principles of heterochromatin formation and function. Nature Reviews Molecular Cell Biology. 2018;19(4):229-44. doi: 10.1038/nrm.2017.119.
33. DeWan AT, Parrado AR, Matise TC, Leal SM. The Map Problem: A Comparison of Genetic and Sequence-Based Physical Maps. The American Journal of Human Genetics. 2002;70(1):101-7. doi: https://doi.org/10.1086/324774.
34. Fierst JL. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Frontiers in Genetics. 2015;6. doi: 10.3389/fgene.2015.00220.
35. Achrem M, Szucko I, Kalinka A. The epigenetic regulation of centromeres and telomeres in plants and animals. Comparative Cytogenetics. 2020;14(2):265-311. doi: 10.3897/CompCytogen.v14i2.51895. PubMed PMID: WOS:000546682500001.
36. Komeda Y. GENETIC REGULATION OF TIME TO FLOWER IN ARABIDOPSIS THALIANA. Annual Review of Plant Biology. 2004;55(1):521-35. doi: 10.1146/annurev.arplant.55.031903.141644.
37. Roux F, Touzet P, Cuguen J, Le Corre V. How to be early flowering: an evolutionary perspective. Trends in Plant Science. 2006;11(8):375-81. doi: https://doi.org/10.1016/j.tplants.2006.06.006.
38. Khan MRG, Ai X-Y, Zhang J-Z. Genetic regulation of flowering time in annual and perennial plants. WIREs RNA. 2014;5(3):347-59. doi: https://doi.org/10.1002/wrna.1215.

39. Putterill J, Varkonyi-Gasic E. FT and florigen long-distance flowering control in plants. Current Opinion in Plant Biology. 2016;33:77-82. doi: https://doi.org/10.1016/j.pbi.2016.06.008.

40. Pin PA, Nilsson O. The multifaceted roles of FLOWERING LOCUS T in plant development. Plant, Cell & Environment. 2012;35(10):1742-55. doi: https://doi.org/10.1111/j.1365-3040.2012.02558.x.

41. Wickland DP, Hanzawa Y. The FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family: Functional Evolution and Molecular Mechanisms. Molecular Plant. 2015;8(7):983-97. doi: 10.1016/j.molp.2015.01.007. PubMed PMID: WOS:000357615600003.

42. Freytes SN, Canelo M, Cerdan PD. Regulation of Flowering Time: When and Where? Current Opinion in Plant Biology. 2021;63. doi: 10.1016/j.pbi.2021.102049. PubMed PMID: WOS:000711158100002.

43. Pérez-Núñez MT, Chan JL, Sáenz L, González T, Verdeil JL, Oropeza C. Improved somatic embryogenesis from Cocos nucifera (L.) plumule explants. In Vitro Cellular & Developmental Biology - Plant. 2006;42(1):37-43. doi: 10.1079/IVP2005722.

44. Wilms H, De Bièvre D, Longin K, Swennen R, Rhee J, Panis B. Development of the first axillary in vitro shoot multiplication protocol for coconut palms. Scientific Reports. 2021;11(1):18367. doi: 10.1038/s41598-021-97718-1.

Supporting information

S1 File. DArTSeq SNP genotypic data for the full 3714 linkage mapped SNPs in the F_2 population

S2 File. Genotypic data for the 2952 DArTSeq SNP markers linkage mapped and physically mapped to unique positions on the coconut genome assemblies

S3 File. BLAST results of the DArTSeq SNP corresponding sequences, target SNP type and position, against the YANG and WANG genomes.

S4 File. Local recombination rates estimated following the Marey maps construction and LOESS smoothing between the linkage map and the two assembled genomes

S5 File. Raw phenotypic data and estimated BLUPs following the mixed model analysis for the evaluated traits in the F_2 population

S6 File. QTL mapping results obtained with the composite interval mapping method

S7 File. BLAST alignment result of the *Arabidopsis thaliana* putative flowering signals mediating protein FT (At1g65480) mRNA, complete cds against the YANG chromosome 3

S8 File. BLAST alignment result of the *Arabidopsis thaliana* putative flowering signals mediating protein FT (At1g65480) mRNA, complete cds against the WANG chromosome 4

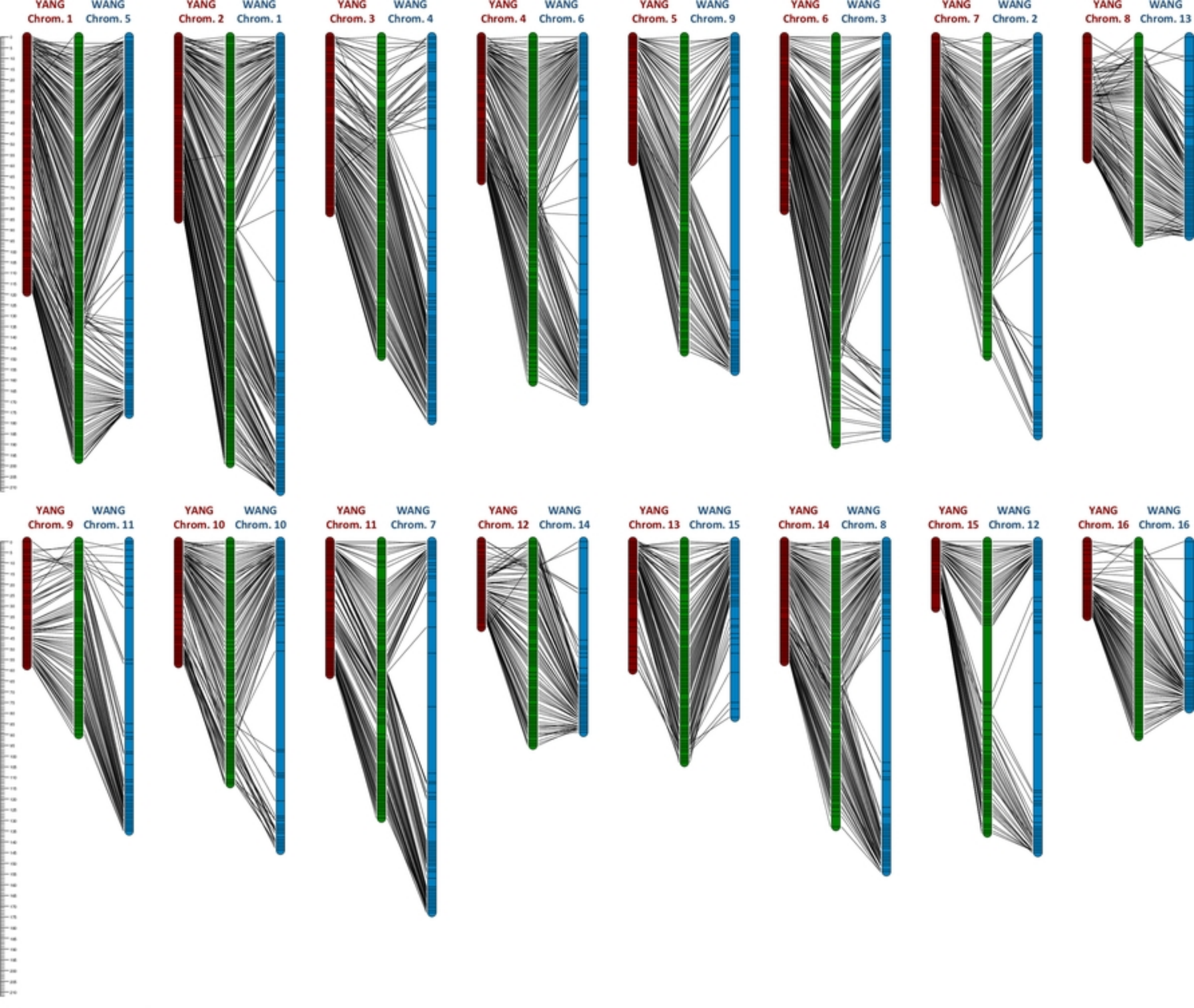

YANG Chrom. 1
WANG Chrom. 5
YANG Chrom. 2
WANG Chrom. 1
YANG Chrom. 3
WANG Chrom. 4
YANG Chrom. 4
WANG Chrom. 6
YANG Chrom. 5
WANG Chrom. 9
YANG Chrom. 6
WANG Chrom. 3
YANG Chrom. 7
WANG Chrom. 2
YANG Chrom. 8
WANG Chrom. 13
YANG Chrom. 9
WANG Chrom. 11
YANG Chrom. 10
WANG Chrom. 10
YANG Chrom. 11
WANG Chrom. 7
YANG Chrom. 12
WANG Chrom. 14
YANG Chrom. 13
WANG Chrom. 15
YANG Chrom. 14
WANG Chrom. 8
YANG Chrom. 15
WANG Chrom. 12
YANG Chrom. 16
WANG Chrom. 16

Figure 1

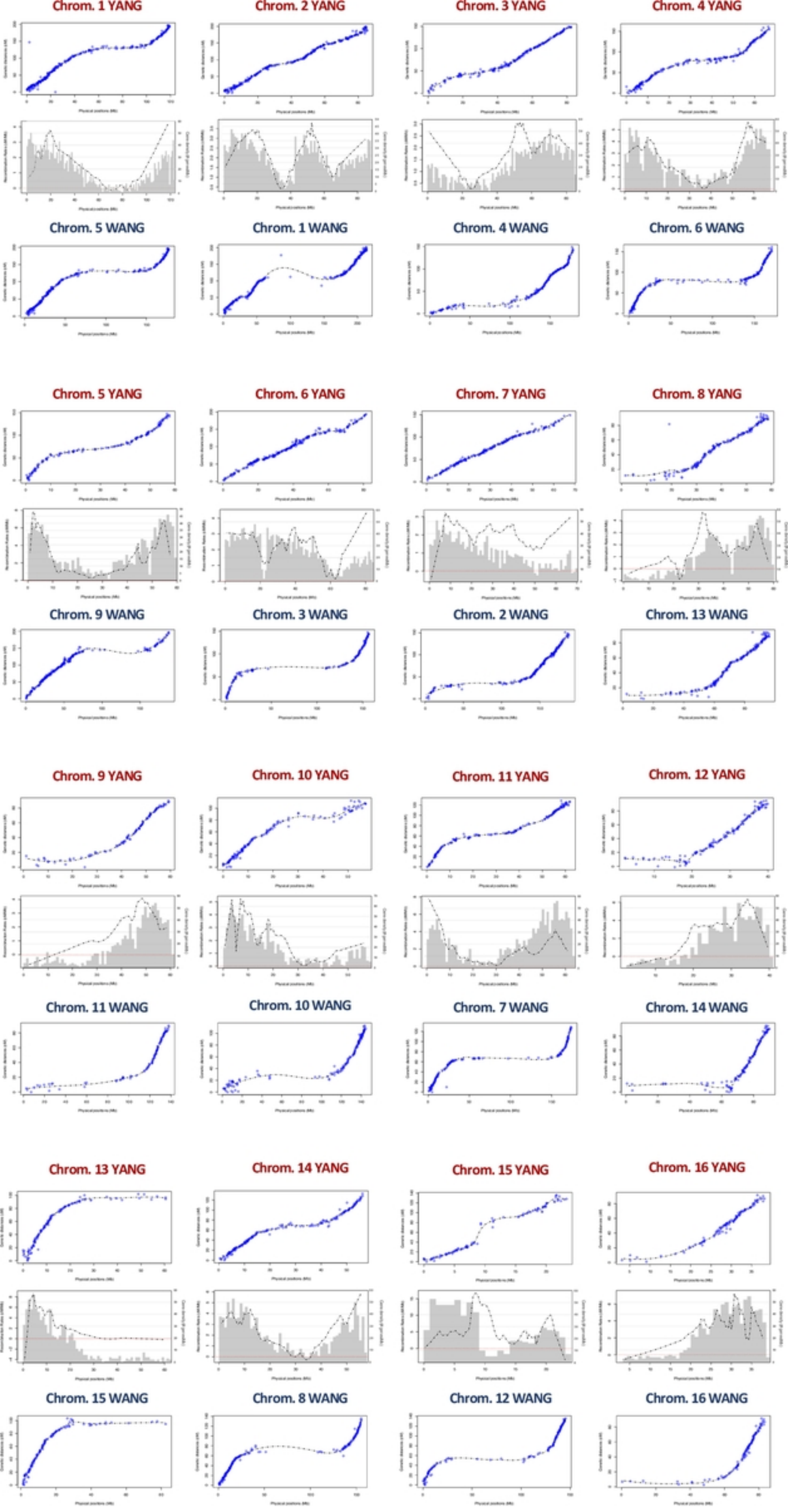
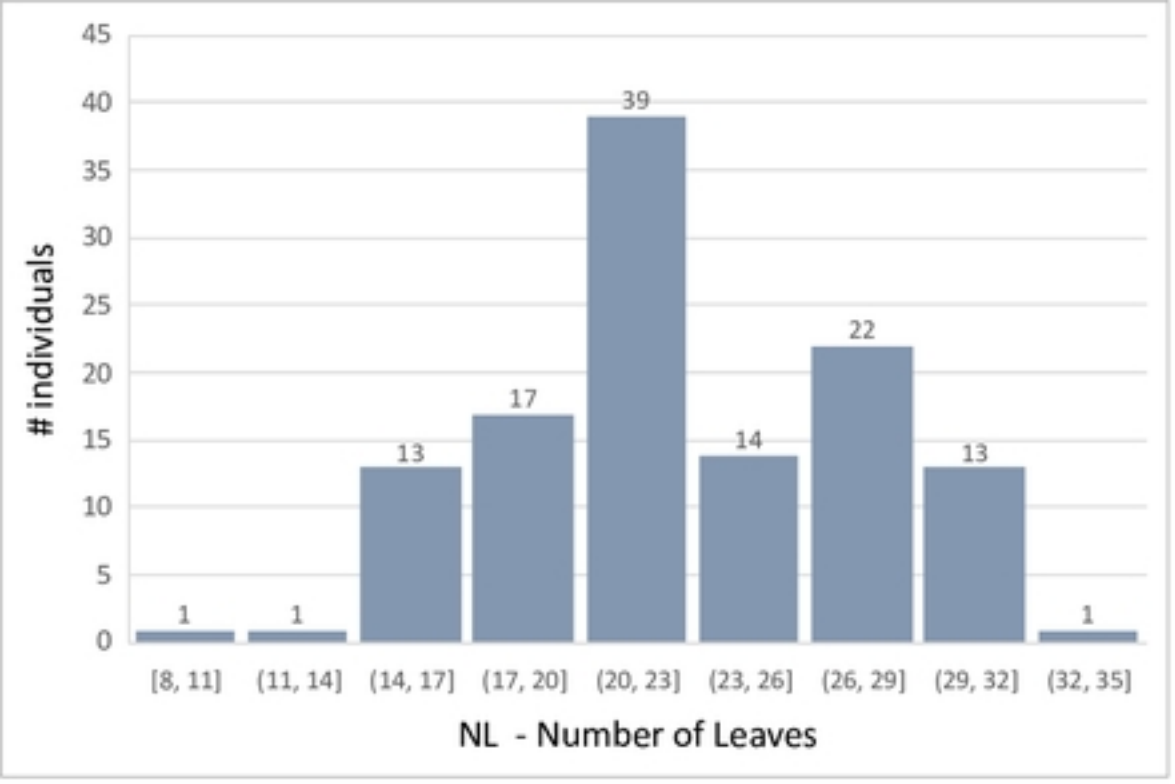
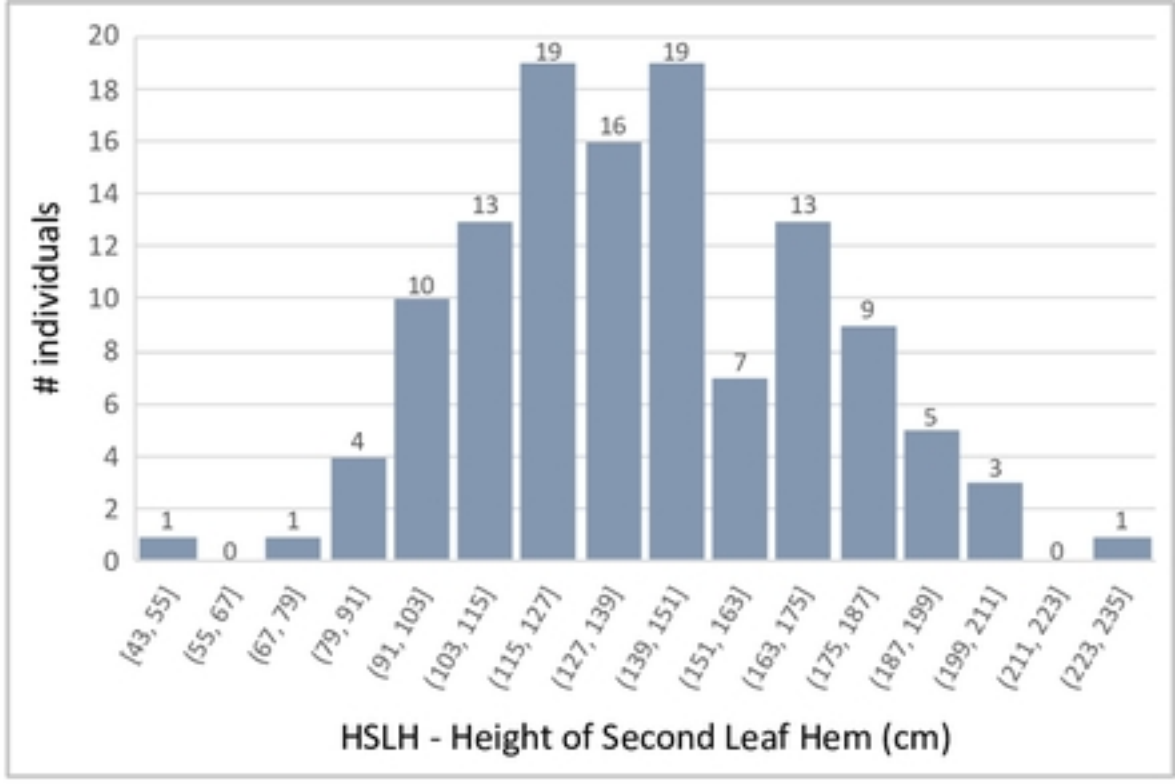
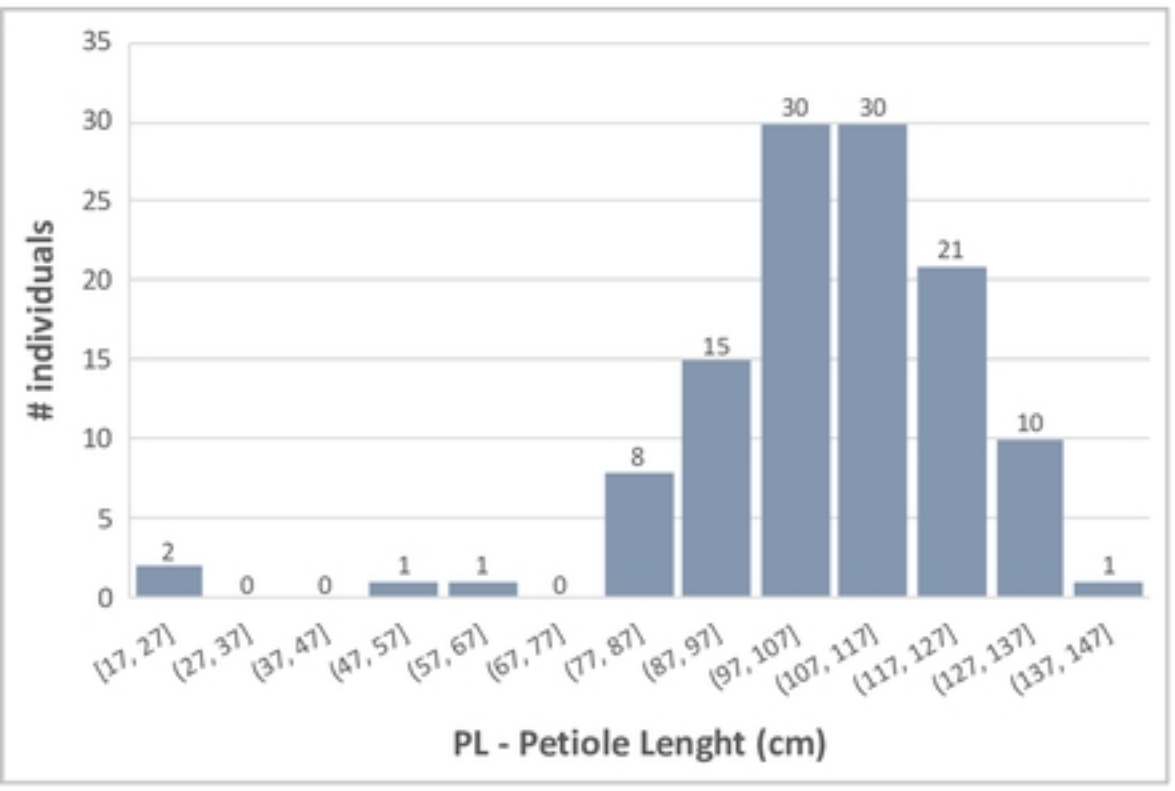
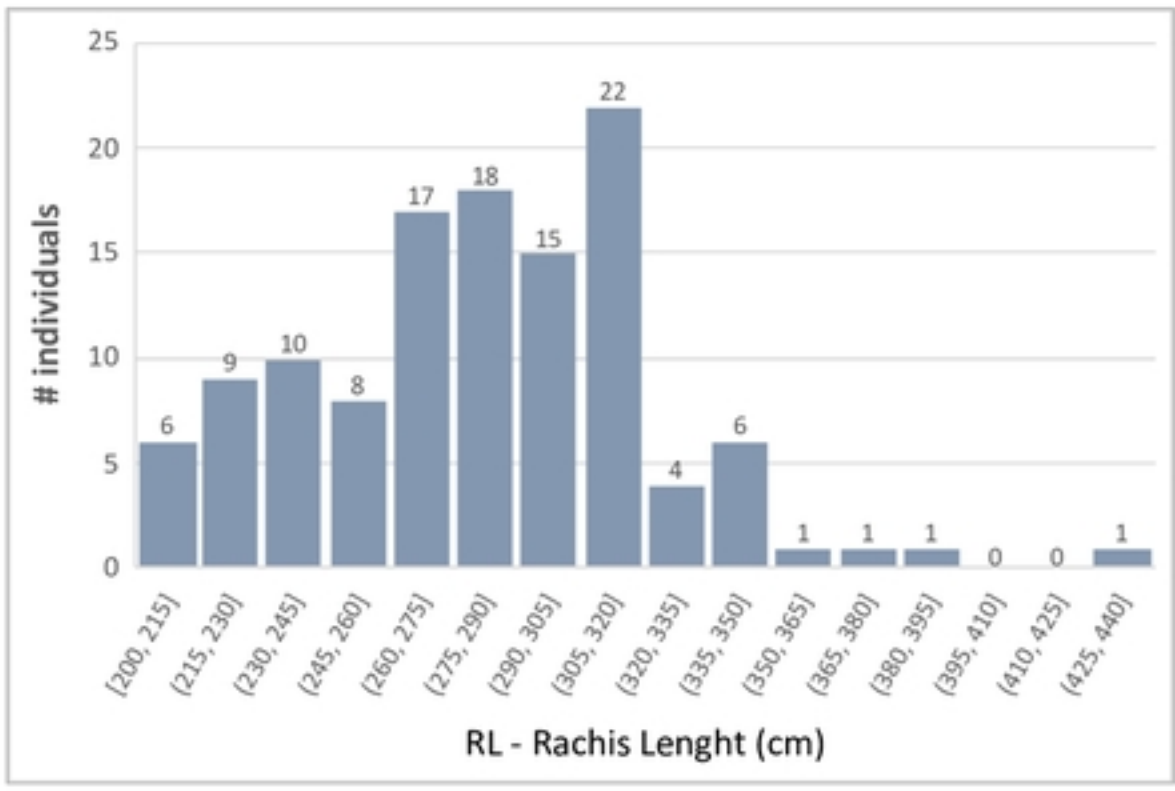

Chrom. 1 YANG
Chrom. 2 YANG
Chrom. 3 YANG
Chrom. 4 YANG
Chrom. 5 WANG
Chrom. 1 WANG
Chrom. 4 WANG
Chrom. 6 WANG
Chrom. 5 YANG
Chrom. 6 YANG
Chrom. 7 YANG
Chrom. 8 YANG
Chrom. 9 WANG
Chrom. 3 WANG
Chrom. 2 WANG
Chrom. 13 WANG
Chrom. 9 YANG
Chrom. 10 YANG
Chrom. 11 YANG
Chrom. 12 YANG
Chrom. 11 WANG
Chrom. 10 WANG
Chrom. 7 WANG
Chrom. 14 WANG
Chrom. 13 YANG
Chrom. 14 YANG
Chrom. 15 YANG
Chrom. 16 YANG
Chrom. 15 WANG
Chrom. 8 WANG
Chrom. 12 WANG
Chrom. 16 WANG

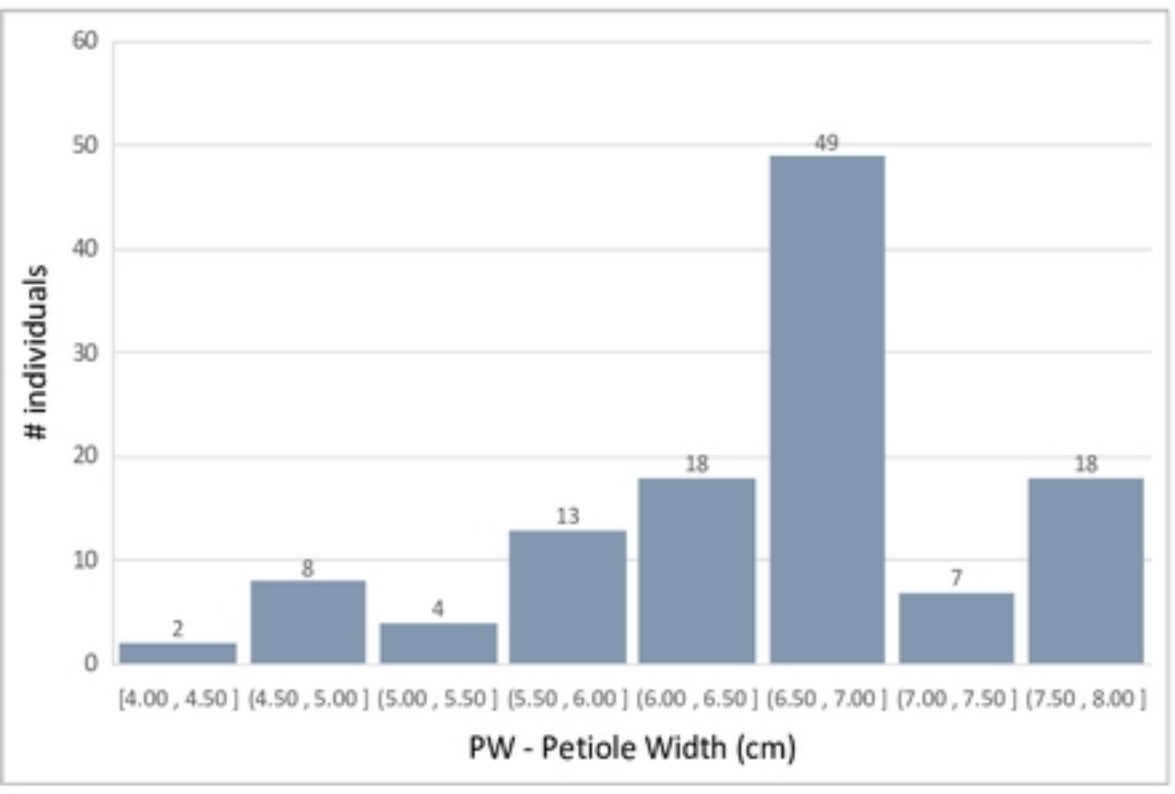
Figure 2

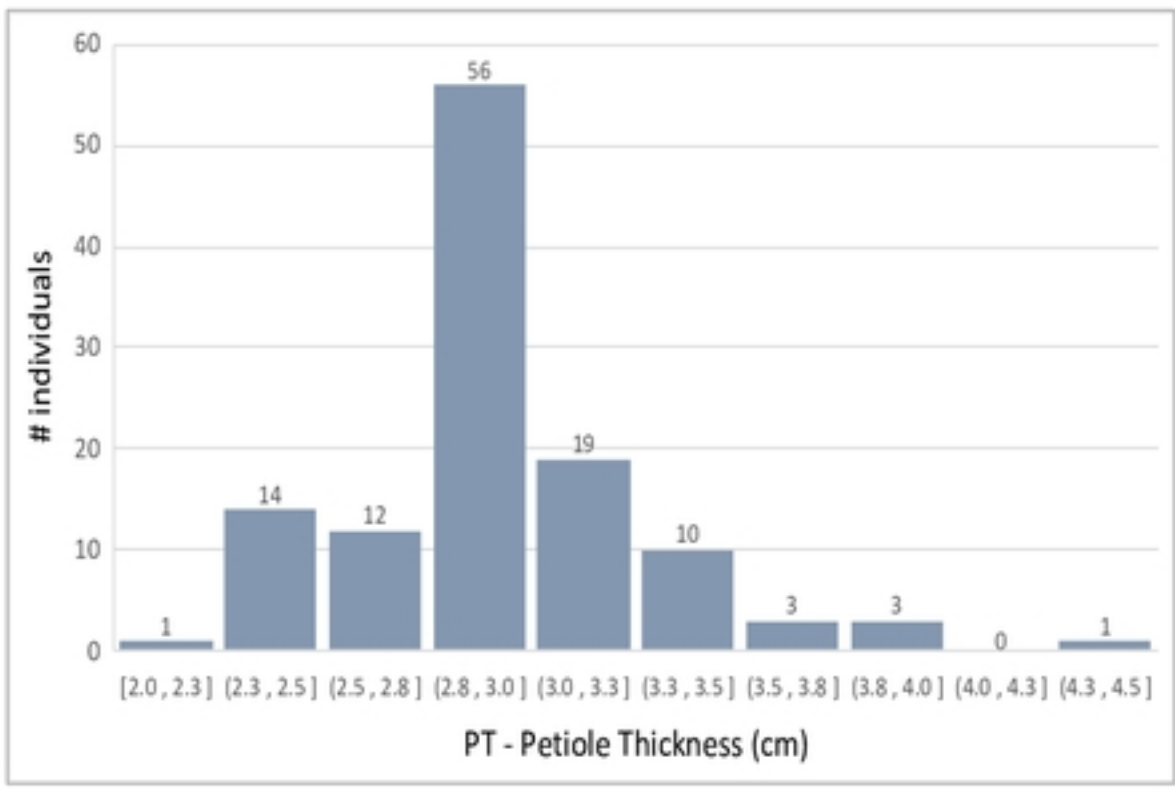

individuals
45
40
35
30
25
20
15
10
5
0
1
1
13
17
39
14
22
13
1
[8, 11]
(11, 14]
(14, 17]
(17, 20]
(20, 23]
(23, 26]
(26, 29]
(29, 32]
(32, 35]
NL - Number of Leaves

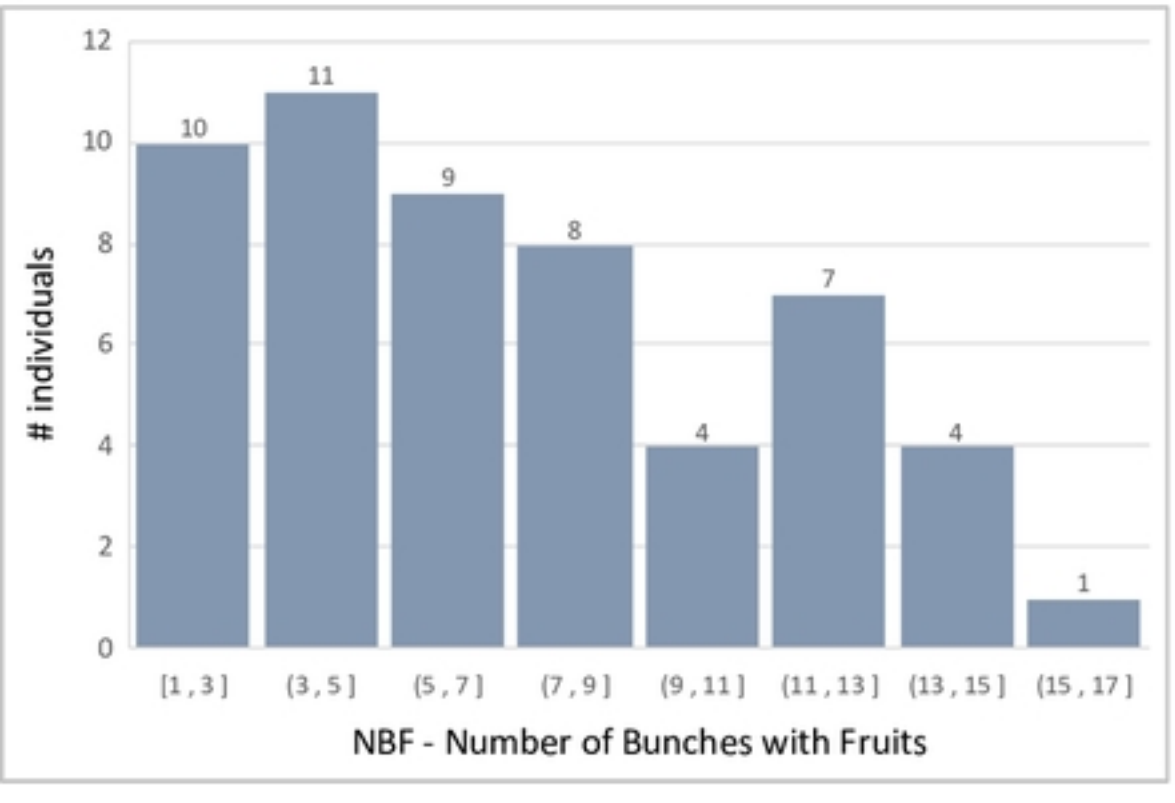

individuals
20
18
16
14
12
10
8
6
4
2
0
1
0
1
4
10
13
19
16
19
7
13
9
5
3
0
1
[43, 55]
(55, 67]
(67, 79]
(79, 91]
(91, 103]
(103, 115]
(115, 127]
(127, 139]
(139, 151]
(151, 163]
(163, 175]
(175, 187]
(187, 199]
(199, 211]
(211, 223]
(223, 235]
HSLH - Height of Second Leaf Hem (cm)


individuals
35
30
25
20
15
10
5
0
2
0
0
1
1
0
8
15
30
30
21
10
1
[17, 27]
(27, 37]
(37, 47]
(47, 57]
(57, 67]
(67, 77]
(77, 87]
(87, 97]
(97, 107]
(107, 117]
(117, 127]
(127, 137]
(137, 147]
PL - Petiole Lenght (cm)


individuals
25
20
15
10
5
0
6
9
10
8
17
18
15
22
4
6
1
1
1
0
0
1
[200, 215]
(215, 230]
(230, 245]
(245, 260]
(260, 275]
(275, 290]
(290, 305]
(305, 320]
(320, 335]
(335, 350]
(350, 365]
(365, 380]
(380, 395]
(395, 410]
(410, 425]
(425, 440]
RL - Rachis Lenght (cm)


individuals
60
50
40
30
20
10
0
2
8
4
13
18
49
7
18
[4.00 , 4.50]
(4.50 , 5.00]
(5.00 , 5.50]
(5.50 , 6.00]
(6.00 , 6.50]
(6.50 , 7.00]
(7.00 , 7.50]
(7.50 , 8.00]
PW - Petiole Width (cm)


individuals
60
50
40
30
20
10
0
1
14
12
56
19
10
3
3
0
1
[2.0 , 2.3]
(2.3 , 2.5]
(2.5 , 2.8]
(2.8 , 3.0]
(3.0 , 3.3]
(3.3 , 3.5]
(3.5 , 3.8]
(3.8 , 4.0]
(4.0 , 4.3]
(4.3 , 4.5]
PT - Petiole Thickness (cm)


individuals
12
10
8
6
4
2
0
10
11
9
8
4
7
4
1
[1 , 3]
(3 , 5]
(5 , 7]
(7 , 9]
(9 , 11]
(11 , 13]
(13 , 15]
(15 , 17]
NBF - Number of Bunches with Fruits

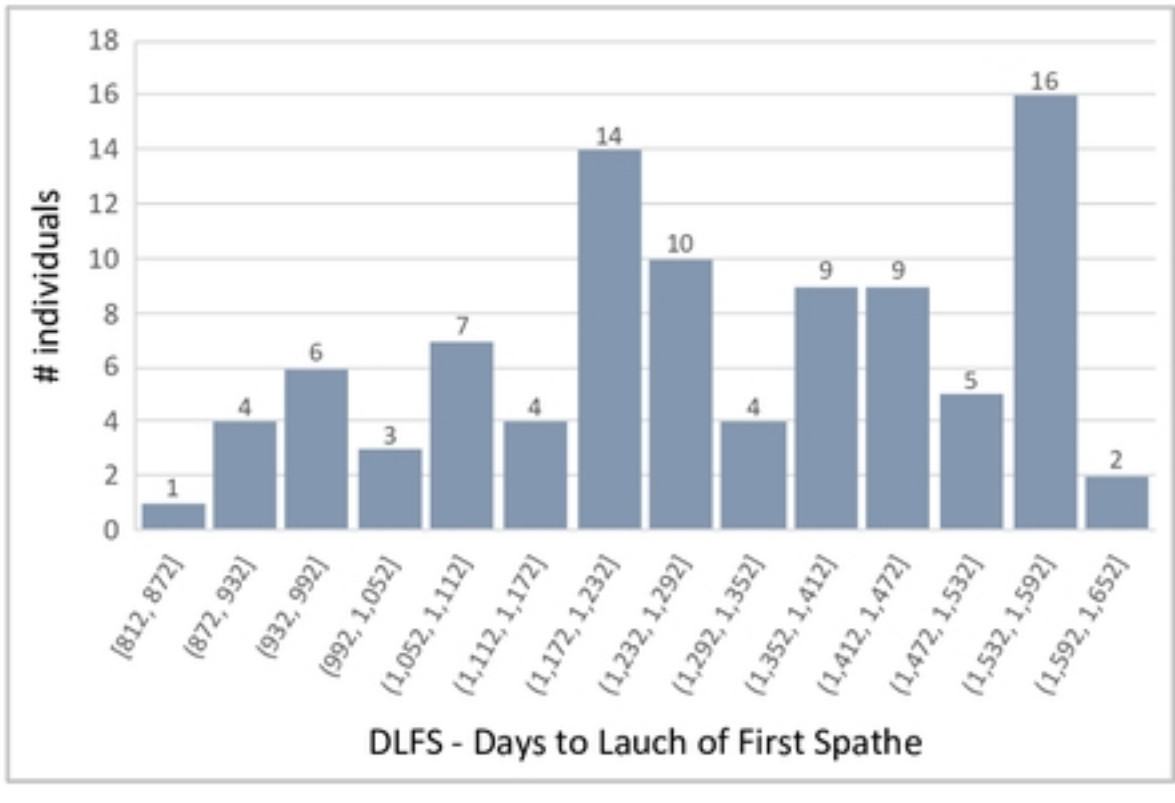

individuals
18
16
14
12
10
8
6
4
2
0
1
4
6
3
7
4
14
10
4
9
9
5
16
2
[812, 872]
(872, 932]
(932, 992]
(992, 1,052]
(1,052, 1,112]
(1,112, 1,172]
(1,172, 1,232]
(1,232, 1,292]
(1,292, 1,352]
(1,352, 1,412]
(1,412, 1,472]
(1,472, 1,532]
(1,532, 1,592]
(1,592, 1,652]
DLFS - Days to Lauch of First Spathe

Figure 3

	HSLH	PL	RL	PW	PT	BOLE	NBF	DLFS
NL	0.772	-0.091	0.519	0.368	0.224	0.368	0.421	-0.484
HLH		0.119	0.706	0.435	0.362	0.430	0.369	-0.349
PL			0.043	0.053	0.146	-0.207	-0.348	0.479
RL				0.514	0.505	0.292	0.255	-0.132
PW					0.403	0.141	0.066	-0.067
PT						0.081	-0.044	0.110
BOLE							0.347	-0.417
NBF								-0.519

Figure 4

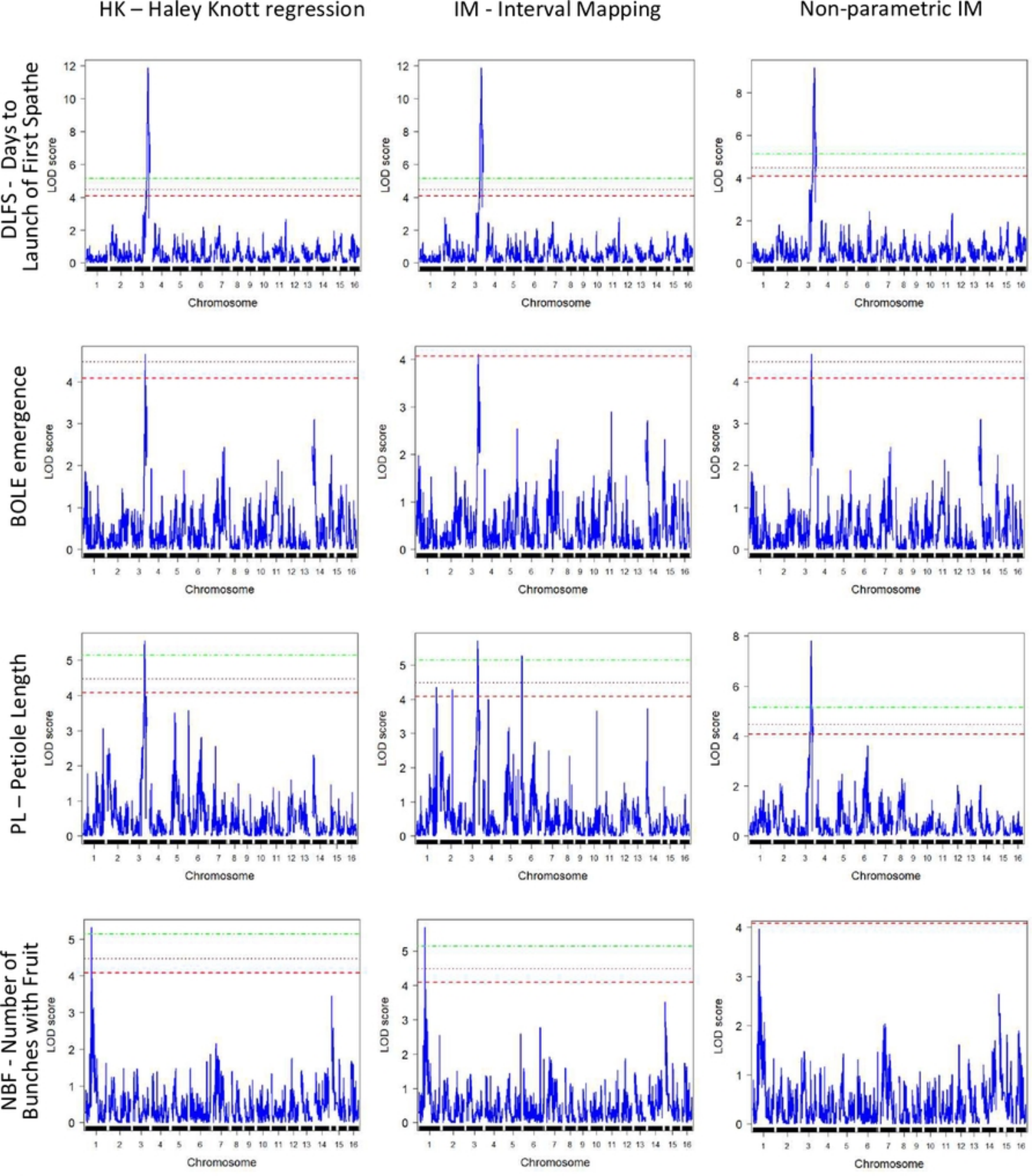

HK – Haley Knott regression
IM - Interval Mapping
Non-parametric IM
DLFS - Days to Launch of First Spathe
BOLE emergence
PL – Petiole Length
NBF - Number of Bunches with Fruit
LOD score
Chromosome

Figure 5

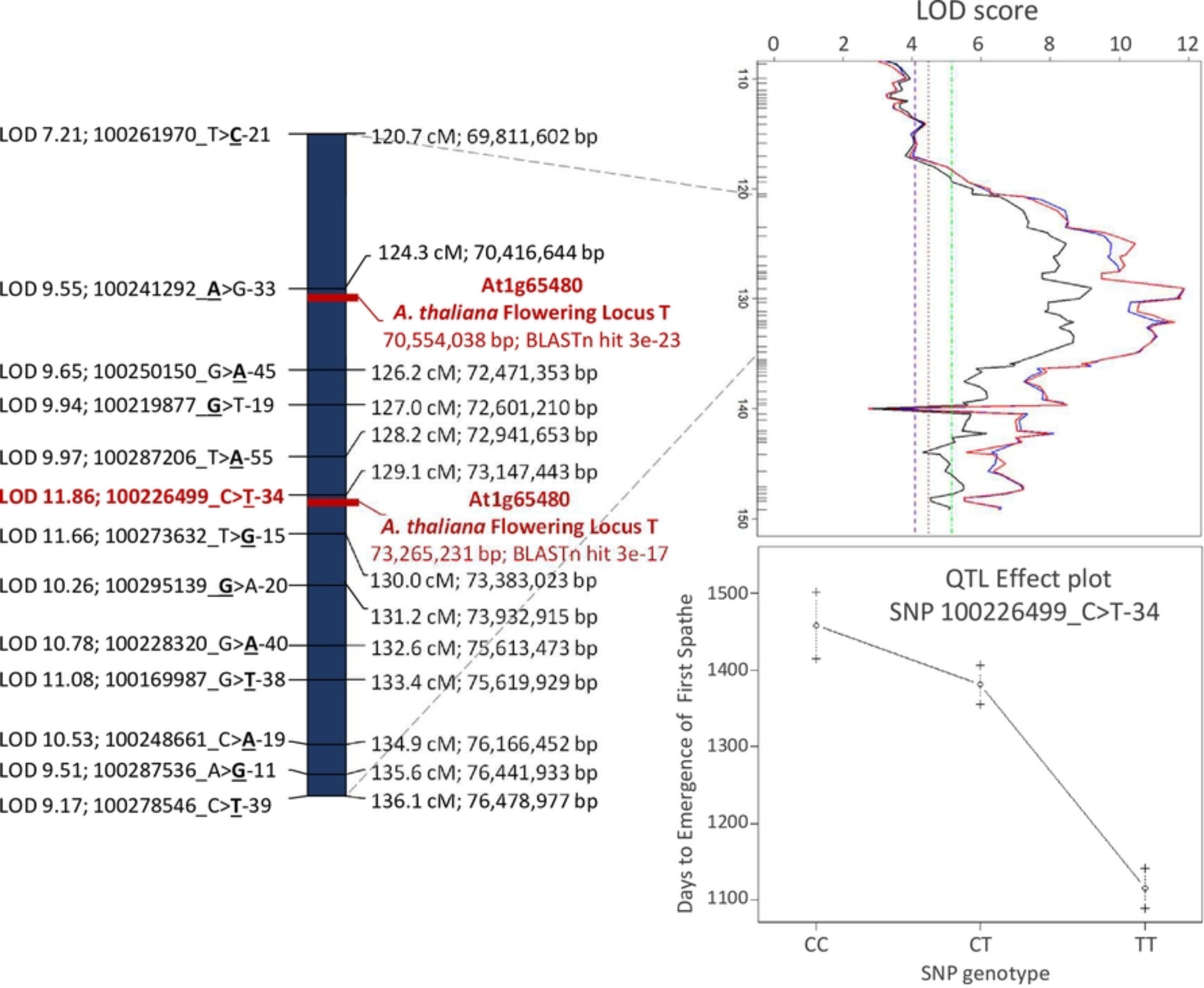

LOD score
0
2
4
6
8
10
12
110
120
130
140
150
LOD 7.21; 100261970_T>C-21
120.7 cM; 69,811,602 bp
124.3 cM; 70,416,644 bp
LOD 9.55; 100241292_A>G-33
At1g65480
A. thaliana Flowering Locus T
70,554,038 bp; BLASTn hit 3e-23
LOD 9.65; 100250150_G>A-45
126.2 cM; 72,471,353 bp
LOD 9.94; 100219877_G>T-19
127.0 cM; 72,601,210 bp
128.2 cM; 72,941,653 bp
LOD 9.97; 100287206_T>A-55
129.1 cM; 73,147,443 bp
LOD 11.86; 100226499_C>T-34
At1g65480
A. thaliana Flowering Locus T
73,265,231 bp; BLASTn hit 3e-17
LOD 11.66; 100273632_T>G-15
LOD 10.26; 100295139_G>A-20
130.0 cM; 73,383,023 bp
131.2 cM; 73,932,915 bp
LOD 10.78; 100228320_G>A-40
132.6 cM; 75,613,473 bp
LOD 11.08; 100169987_G>T-38
133.4 cM; 75,619,929 bp
LOD 10.53; 100248661_C>A-19
134.9 cM; 76,166,452 bp
LOD 9.51; 100287536_A>G-11
135.6 cM; 76,441,933 bp
LOD 9.17; 100278546_C>T-39
136.1 cM; 76,478,977 bp
QTL Effect plot
SNP 100226499_C>T-34
Days to Emergence of First Spathe
1500
1400
1300
1200
1100
CC
CT
TT
SNP genotype

Figure 6