

1 **Genetic differentiation and intrinsic genomic features explain variation in recombination**
2 **hotspots among cocoa tree populations**

3
4 Enrique J. Schwarzkopf¹, Juan C. Motamayor², Omar E. Cornejo^{1,a}

5 **Affiliations:**

6 ¹ School of Biological Sciences, Washington State University, Pullman, WA, USA

7 ² Universal Genetic Solutions, LLC

8 ^a Corresponding author: ocornejo@gmail.com

9

10 **Correspondence:**

11

12 **Running title:** Recombination hotspots in ten *Theobroma cacao* populations

13

14 **Keywords:** recombination; recombination hotspots; domestication

15

16 **Abstract:**

17

18 Our study investigates the possible drivers of recombination hotspots in *Theobroma cacao*
19 using ten genetically differentiated populations. By comparing recombination patterns between
20 multiple populations, we obtain a novel view of recombination at the population-divergence
21 timescale. For each population, a fine-scale recombination map was generated using the
22 coalescent with a standard method based on linkage disequilibrium (LD). These maps revealed
23 higher recombination rates in a domesticated population and a population that has undergone a
24 recent bottleneck. We inferred hotspots of recombination for each population and find that the
25 genomic locations of hotspots correlate with genetic differentiation between populations (F_{ST}).
26 We used randomization approaches to generate appropriate null models to understand the
27 association between hotspots of recombination and both DNA sequence motifs and genomic
28 features. We found that hotspot regions contained fewer known retroelement sequences than
29 expected and were overrepresented near transcription start and termination sites. Our findings
30 indicate that recombination hotspots are evolving in a way that is consistent with genetic
31 differentiation but are also preferentially driven to near coding regions. We illustrate that,
32 consistent with predictions in plant domestication, the recombination rate of the domesticated
33 population is orders of magnitude higher than that of other populations. More importantly, we
34 find two fixed mutations in the domesticated population's FIGL1 protein. FIGL1 has been shown
35 to increase recombination rates in *Arabidopsis* by several orders of magnitude, suggesting a
36 possible mechanism for the observed increased recombination rate in the domesticated
37 population.

38

39 **Introduction:**

40

41 Genetic recombination is an important source of genome-wide genetic variation fundamental for
42 evolutionary forces like selection and genetic drift to act. Selection and drift contribute to a loss
43 of variation, which means that in the absence of forces that maintain variation along the
44 genome, populations would be incapable of evolving over prolonged periods of time.
45 Recombination rearranges genetic material onto different backgrounds generating a larger set
46 of haplotype combinations on which selection can act, reducing the magnitude of Hill-Robertson
47 interference (Felsenstein 1974). Different regimes of recombination can strongly influence how
48 efficient selection is at purging deleterious mutations and increasing the frequency of beneficial
49 mutations in the population (Felsenstein 1974).

50

51 Studies in a wide range of species have shown that recombination rates are not uniform along
52 the genome and general patterns of variation have been described (Begun and Aquadro 1992,
53 Akhunov et al. 2003, Wu et al. 2003, Anderson et al. 2004, McVean et al. 2004, Mézard 2006,
54 Kim et al. 2007, Gore et al. 2009, Schnable et al. 2009, Branca et al. 2011, Paape et al. 2012).
55 One pattern that has been observed in multiple species is the reduced recombination rate in
56 centromeric regions of the chromosomes and the progressive increase of recombination rates
57 as the physical distance from the telomeres decreases (Begun and Aquadro 1992, Akhunov et
58 al. 2003, Wu et al. 2003, Anderson et al. 2004, Gore et al. 2009, Schnable et al. 2009). This
59 pattern has also been shown to arise in simulation studies (e.g. Mackiewicz et al. 2010).
60 Another interesting pattern that has been observed is that of regions with unusually high rates of
61 recombination spread throughout chromosomes: recombination hotspots (McVean et al. 2004,
62 Brunschwig et al. 2012, Paape et al. 2012, Hellsten et al. 2013, Stevison et al. 2016, Shanfelter
63 et al. 2018). The importance of recombination hotspots lies in their ability to shuffle genetic
64 variation at higher rates than the rest of the genome, profoundly impacting the dynamics of
65 selection for or against specific mutations (Felsenstein 1974). In this study, we focus on locally
66 defined recombination hotspots, requiring that their recombination rate be unusually high when
67 compared to neighboring regions.

68
69 A variety of genomic features have been identified as being associated with regions of high
70 recombination. Recombination hotspots have been linked to transcriptional start sites (TSSs)
71 and transcriptional termination sites (TTSs) in *Arabidopsis thaliana*, *Taeniopygia guttata*,
72 *Poephila acuticauda*, and humans (Myers et al. 2005, Choi et al. 2013, Singhal et al. 2015). In
73 *Mimulus guttatus* hotspots were found to be associated with cpG islands (short segments of
74 cytosine and guanine rich DNA, associated with promoter regions) (Hellsten et al. 2013). CpG
75 islands were also associated with increased recombination rates in humans and chimpanzees
76 (Auton et al. 2012). These patterns point to recombination occurring frequently near, but not
77 within, coding regions. The formation of chiasmata is important for the proper disjunction of
78 chromosomes during meiosis (Martinez-Perez et al. 2008), but repeated double-strand breaks
79 can lead to an increased mutation rate (Rodgers and McVey 2015). In coding regions in
80 particular, this excess mutation rate can have a high evolutionary cost, due to the likelihood of
81 novel deleterious mutations being higher than that of beneficial ones (Haldane 1937, Crow and
82 Kimura 1970, Wloch et al. 2001, Sanjuán et al. 2004, Eyre-Walker and Keightley 2007).
83 Recombination hotspots have also been found to be correlated with particular DNA sequence
84 motifs. In some mammals, including *Mus musculus* (Brunschwig et al. 2012) and apes (Auton et
85 al. 2012, Stevison et al. 2016) binding sites for PRDM9, a histone trimethylase with a DNA zinc-
86 finger binding domain, have been found to correlate with recombination hotspots. In *A. Thaliana*,
87 proteins that limit overall recombination rate have been identified, leading to a genome-wide
88 increase in recombination rate in knockout mutants (Fernandes et al. 2018b). However, these
89 *Arabidopsis* proteins have not been shown to direct recombination to particular regions and are
90 therefore not expected to affect the location of recombination hotspots.

91
92 The dynamics of recombination hotspots shared between related species or populations of the
93 same species have been investigated in apes, yielding varying results. Hinch et al. (2011) found
94 that, at finer scales, the genetic maps of European and African human populations were
95 significantly different. They also found that, when looking at hotspots in the major
96 histocompatibility complex, the African populations showed a hotspot that was not present in
97 Europeans, but all European hotspots were found in African populations (Hinch et al. 2011).
98 Recent work on recombination in apes found little correlation of recombination rates in
99 orthologous hotspot regions when looking between species, but a strong correlation when
100 comparing between two populations of the same species (Stevison et al. 2016). Other studies
101 have also found very little sharing of hotspots between humans and chimpanzees (Ptak et al.

102 2005, Winckler et al. 2005). Additionally, the dynamic of changing hotspot locations observed in
103 humans and other apes has been observed in simulations (Mackiewicz et al. 2013). The
104 disparity of empirical results regarding hotspots shared between related populations suggest
105 that further work is required to disentangle the relationship between demographics and shared
106 hotspots.

107
108 The identification of ten genetically differentiated populations of the cocoa tree, *Theobroma*
109 *cacao*, (Motamayor et al. 2008, Cornejo et al. 2017) can be leveraged to study population-level
110 dynamics of recombination patterns. The ten *T. Cacao* populations originate from different
111 regions of South and Central America, and include one fully domesticated population (Criollo),
112 used in the production of fine chocolate, and nine wilder, more resilient populations which
113 generate higher cocoa yield than the Criollo variety (Fig. S1) (Motamayor et al. 2008,
114 Henderson et al. 2007, Cornejo et al 2017). These ten populations have been shown to have
115 strong signatures of differentiation between them (F_{ST} values ranging from 0.16 to 0.65) and
116 they separate into clear clusters of ancestry (Cornejo et al. 2017). During domestication,
117 recombination plays an important role in the segregation of traits, and for this reason it has been
118 hypothesized that recombination rates will increase during the process of domestication
119 (Moyers et al. 2018). Domestication can be a rapid process and there is theoretical evidence for
120 the increase of recombination rates during periods of rapid evolutionary change (Otto and
121 Barton 1997). Empirical evidence for this prediction has been shown in a limited number of
122 herbaceous plant species with short generation times (Ross-Ibarra 2004). It is not clear if plant
123 species with longer generation times are also expected to experience increased recombination
124 rates, and it is also unclear what mechanisms could explain these differences. One possible
125 explanation for differences in recombination rates between wild and domesticated populations is
126 polymorphism in genes like those previously demonstrated to suppress recombination in
127 *Arabidopsis thaliana* (Girard et al. 2015, Fernandes et al. 2018a). The differences in
128 recombination rates between wild and domesticated populations is just one of the possible
129 questions that can be touched on with this system.

130
131 The ten populations of *T. cacao* also allow us to compare the locations of hotspots between
132 them, potentially contributing to the understanding of hotspot turnover at the population-
133 divergence timescale. These comparisons can also contribute to our understanding of how
134 demographics impact the turnover of recombination hotspot locations. *T. cacao* is unique in this
135 case for being a long-lived organism with no known driver of recombination hotspots (e.g.
136 PRDM9). What contributes to the location of recombination hotspots in such a species is, of
137 course, contingent on our being able to detect recombination hotspots in the different
138 populations of *T. cacao*.

139
140 In order to locate recombination hotspots for *T. Cacao* populations, we must first obtain fine-
141 scale recombination maps for each population, which we did using an LD-based method. Fine-
142 scale, LD-based recombination maps have been constructed for a number of plant models
143 (Paape et al. 2012, Choi et al. 2013, Hellsten et al. 2013), identifying a variety of features
144 correlated to recombination rate. Unlike these model plants with short generation times, *T.*
145 *Cacao* is a perennial woody plant with a five-year generation time (Henderson et al. 2007). The
146 size and long generation time of *T. Cacao* makes direct measurements of recombination
147 impractical. However, historical recombination can be estimated for *T. Cacao* using coalescent
148 based methods (Auton and McVean 2007). Theoretical studies have shown that population
149 structure can generate artificially inflated measures of LD (Li and Nei 1974, Ohta 1982), which
150 would be detrimental to our estimates of recombination. For this reason, recombination maps
151 were constructed independently for each population. For each population we aim to describe

152 the relationship between recombination hotspots and a variety of evolutionary and genomic
153 factors.

154
155 We used an LD-based method to estimate recombination rates for ten populations of *T. Cacao*,
156 which we then analyzed with a maximum likelihood statistical framework to infer the location of
157 recombination hotspots. The locations of hotspots were compared across populations and a
158 novel resampling scheme tailored to the genomic architecture of *T. Cacao* was used to generate
159 null assumptions for the distribution of hotspots along the genome. These null distributions were
160 used to identify differential representation of known DNA sequence motifs in ubiquitous
161 recombination hotspots, and of overlap between recombination hotspots and genomic traits for
162 each population. The re-sampling schemes used to identify these associations are novel in the
163 context of this work and were designed to take into account the size and distribution of elements
164 in the genome. In this work we aimed to answer the following questions: (i) How are
165 recombination rates distributed within 10 highly differentiated populations of *T. Cacao*, and how
166 do they compare to each other? (ii) How are hotspots distributed along the genome of each of
167 the ten populations of *T. Cacao*, and can these distributions be explained by patterns of
168 population genetic differentiation? (iii) Are there identifiable DNA sequence motifs that are
169 associated with the location of recombination hotspots along the *T. Cacao* genome? (iv) Are
170 there genomic features (e.g. TSSs, TTSSs, exons, introns) consistently associated with
171 recombination hotspot locations across *T. Cacao* populations? Our findings suggest that
172 recombination hotspot locations generally follow patterns of diversification between populations,
173 while also having a strong tendency to occur close to TSSs and TTSSs. Moreover, we find a
174 strong negative association between the occurrence of recombination hotspots and the
175 presence of retroelements.

176
177 **Results:**

178
179 *Comparing recombination rates between populations*

180
181 Populations show a mean recombination rate (r) between 2.1 and 525 cm/Mb (Table 1), with a
182 variety of distributions (Fig. S2). We observe a higher mean than median r indicating that
183 extreme high values are present for all populations. The extreme recombination rate values
184 affect the mean, driving it to values consistently higher than the median. The pattern of
185 recombination rates along the genome varied between populations, as can be seen in the
186 comparison of the Nanay and Purus third chromosome (Fig. 1). Purus appears to have a higher
187 average recombination rate than Nanay for chromosome three. More specifically, particular
188 regions of the chromosome present peaks in one population that are absent in the other. A
189 similar pattern can also be observed for the density of recombination hotspots, e.g. Purus
190 presenting a high density of hotspots in certain regions that is not observed in Nanay. The
191 median 95% probability interval for recombination rate across the genome for each population
192 was found to be several orders of magnitude larger than the uncertainty per site, estimated as
193 the median 95% Credibility Interval of the trace for each position in the genome for that
194 population (Table S1).

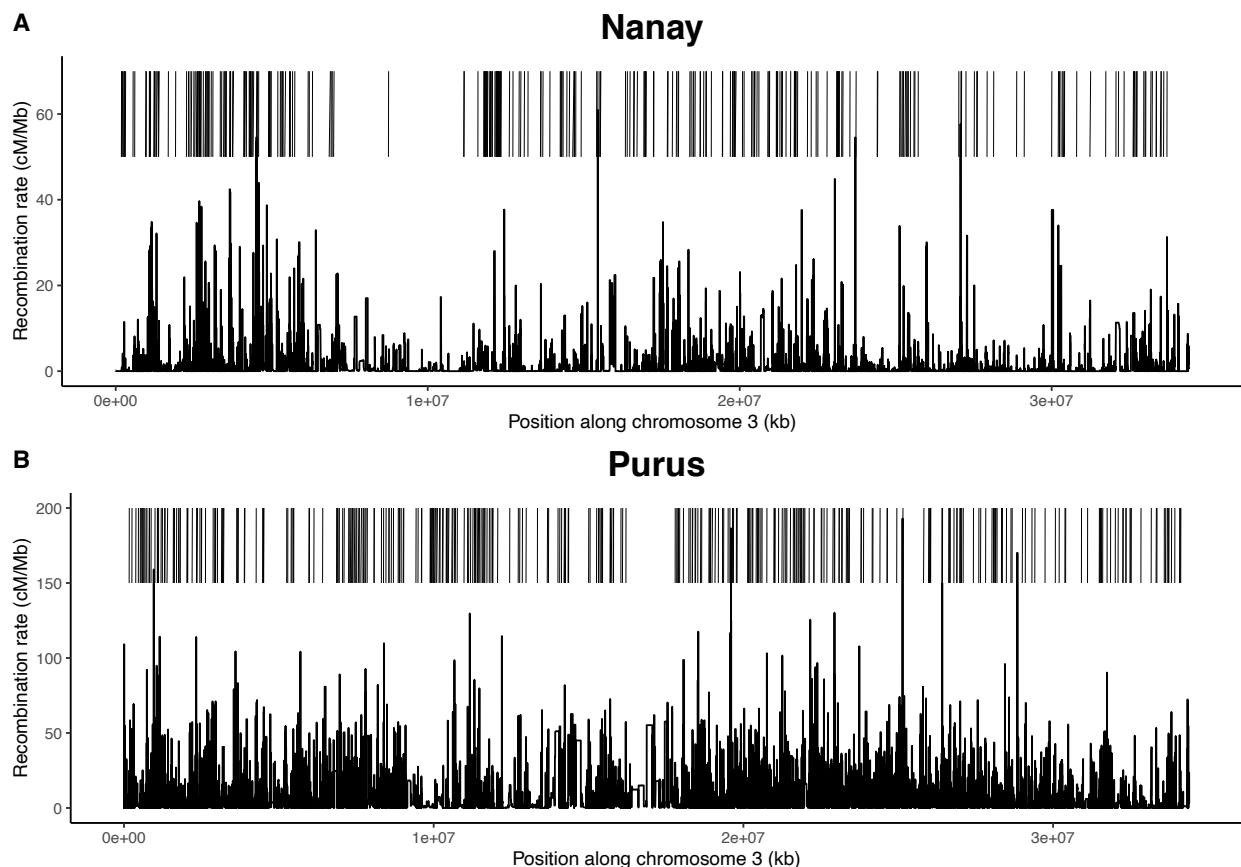
195
196 Overall, the mean recombination rate for most of the populations is similar to that found for
197 *Arabidopsis thaliana* using LDhat, when using $\theta=0.1$ (Choi et al. 2013) (Table 1). The LDhat
198 estimates using $\theta=0.001$ were slightly higher than the estimates using $\theta=0.1$ for each
199 population. We chose to proceed with analyses using the results from the $\theta=0.1$ since more of
200 the mean population recombination rates fell within the range of values identified in plants
201 (Stapley et al. 2016) (Table S2).

202

203 **Table 1.** Recombination rates in $\rho = 4N_e r$ (in Morgans per base) and r (in cM/Mb) for all ten *T. cacao*
 204 populations. The N_e (from Cornejo et al., 2018) used to transform ρ to r for each population is also
 205 reported, as are the lower and upper bounds of a 95% confidence interval for mean r .
 206

Population	Mean $4N_e r$	Mean N_e	Mean r (cM/Mb)	Median r (cM/Mb)	L bound (Mean r)	U bound (Mean r)
Amelonado	1.58e-03	15,744	2.51	2.40e-04	2.48	2.54
Contamana	8.53e-03	61,102	3.49	4.92e-01	3.48	3.50
Criollo	1.46e-04	695	525	427	523	527
Curaray	1.04e-04	58,213	4.45	1.78	4.44	4.46
Guianna	8.66e-03	4,651	46.5	7.74e-01	46.3	46.7
Iquitos	4.23e-03	49,984	2.11	5.88e-04	2.10	2.12
Maranon	4.09e-03	34,037	3.01	1.64e-03	2.99	3.02
Nacional	4.66e-03	26,060	4.47	9.76e-03	4.44	4.49
Nanay	6.82e-03	42,429	4.02	1.51e-02	4.00	4.04
Purus	5.95e-03	17,357	8.57	7.74e-01	8.54	8.60

207
 208
 209
 210



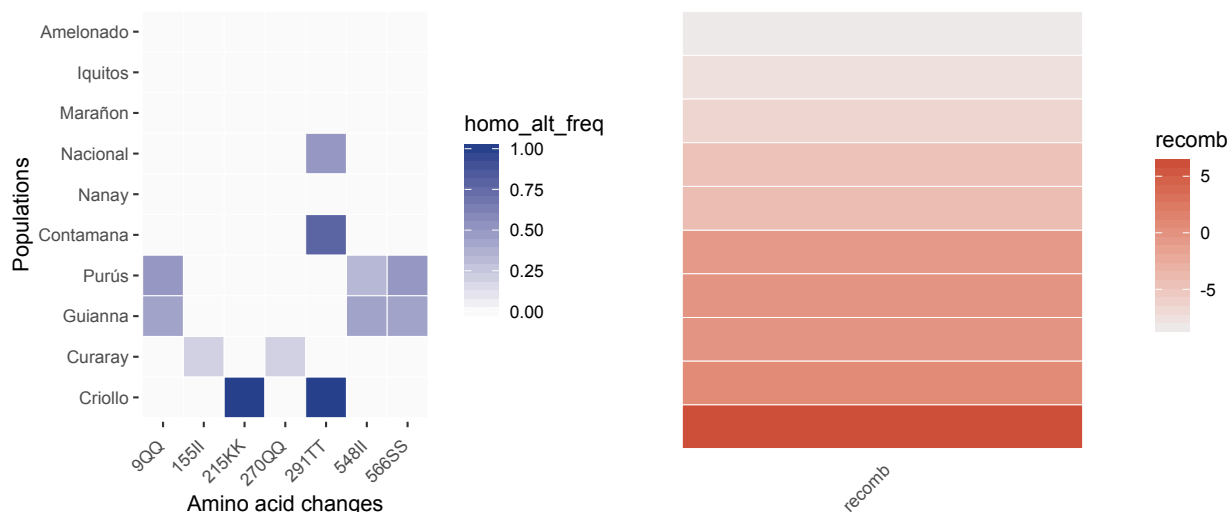
211 **Figure 1.** The third chromosomes of the Nanay (A) and Purus (B) populations were selected to exemplify
 212 the differences between populations in recombination rates (r) and recombination hotspot locations (bars
 213 above rates).
 214

215 In order to compare the average recombination rates of the different populations, a Wilcoxon
 216 signed-rank test was performed for every pair of populations. The only pair that did not show a
 217

218 significant difference in mean recombination rate was that of Nacional and Nanay ($p=0.3$). All
 219 other pairwise comparisons were highly significant ($p<2e-16$).
 220

221 Two populations, Guianna and Criollo, have a higher average recombination rate than the other
 222 populations by one and two orders of magnitude respectively (Table 1). Guianna and Criollo
 223 also have been estimated to have a lower effective population size (N_e) (Cornejo et al. 2018) by
 224 one and two orders of magnitude respectively. However, there was no significant association
 225 between mean N_e and r ($p=0.1119$), indicating that, for a high enough N_e , the ability to detect
 226 recombination events is not dictated by the effective population size. When Criollo and Guianna
 227 were excluded, the relationship was also not present ($p=0.3886$). When all populations were
 228 included, the inbreeding coefficient (F , from Cornejo et al. 2018) showed no significant linear
 229 association with mean r ($p=0.3361$). We also found no linear trend between sample size and
 230 mean r ($p=0.2333$).
 231

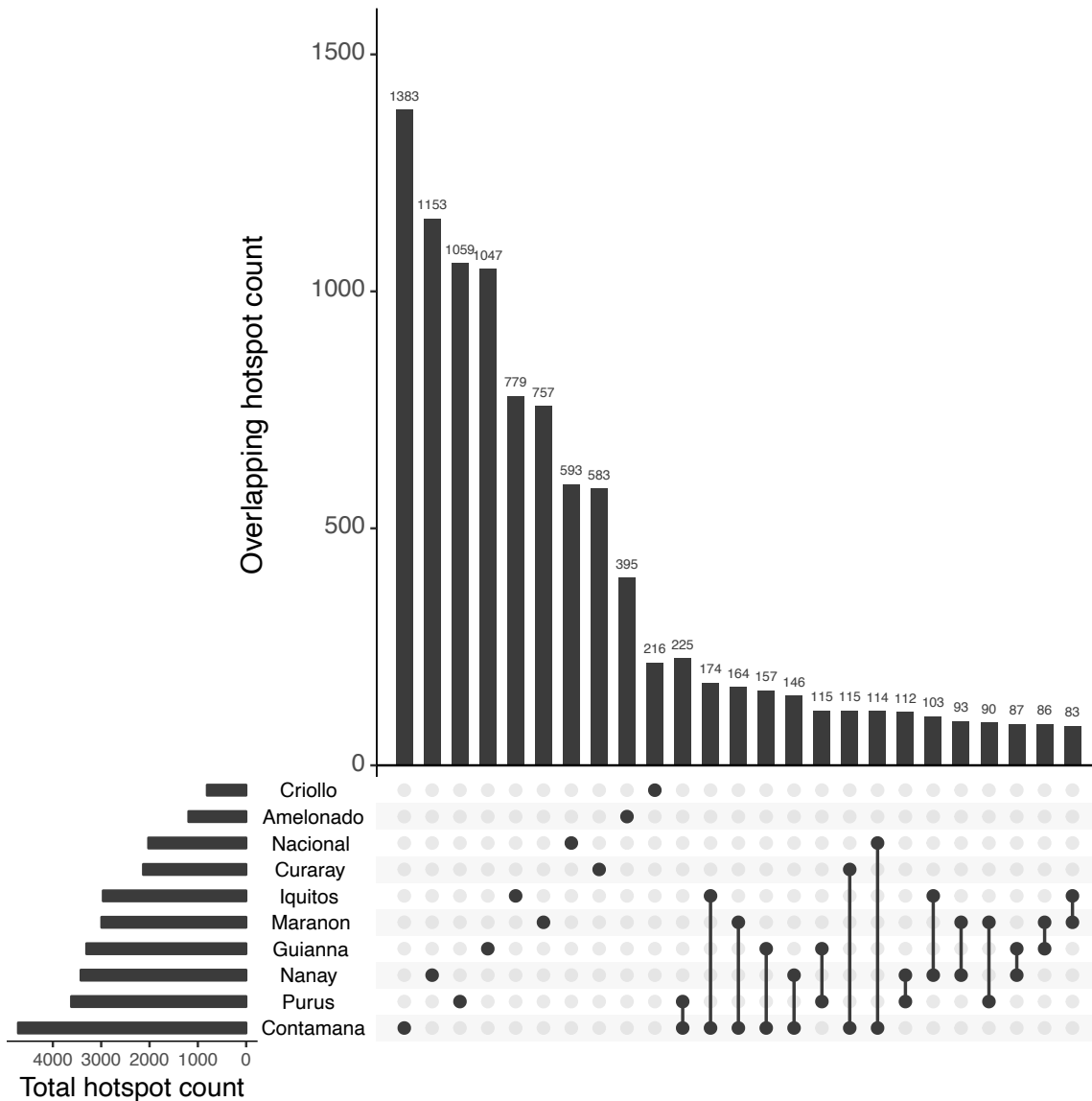
232 The FIGL1 and FLIP proteins characterized by Fernandes et al. (2018a) were found to be
 233 responsible for recombination suppression in *Arabidopsis*. Plants with a FIGL1 knockout were
 234 found to increase recombination rates significantly and FLIP knockouts show increases of
 235 recombination at a much lesser extent (Fernandes et al. 2018). Therefore, we explored the
 236 possibility that missense FIGL1 and FLIP orthologs in *T. cacao* explain the between-population
 237 differences in recombination rate. We used a reciprocal BLAST search to identify the orthologs
 238 for both genes and used annotation data from Cornejo et al. (2018) to identify 15 missense
 239 mutations in FIGL1 and 18 missense mutations in FLIP (Fig. 2, Fig. S3, Table S3). We then
 240 used a generalized linear model framework to infer the impact of the 8 uncorrelated missense
 241 mutations found in the *T. cacao* FIGL1 ortholog under the assumption of a full recessive model.
 242 We find that mutations 215KK (Coeff.=426.54, $p=2.41e-13$), 155II (Coeff.=8.97, $p=0.000231$),
 243 and 291TT (Coeff.=0.47, $p=0.047$) significantly explain changes in the recombination rate, but
 244 all other mutations made no significant impact. The same model was run for FLIP but returned
 245 no significant coefficients (after eliminating perfectly correlated mutations with those found in
 246 FIGL1).
 247



248 **Figure 2.** The left panel shows the frequency of individuals that are homozygous for the alternative allele
 249 of amino acid mutations in a *T. cacao* FIGL1 ortholog. Alternative allele is defined in terms of the
 250 Amelonado reference genome. The right panel shows the \log_e transformed recombination rates (r). The
 251 populations are in the same order in both panels.
 252
 253

254 *Comparing recombination hotspot locations between populations*

255
 256 The majority (55.5%) of hotspots identified were not shared between populations. The 25 most
 257 numerous sets of hotspots are represented in Fig. 3. The nine largest of these are sets of
 258 hotspots unique to single populations. The hotspots unique to the remaining population (Criollo)
 259 formed the eleventh largest set. Effective population size (N_e) is not a good linear predictor of
 260 the number of detected hotspots ($\rho=0.1489$), nor is sample size ($\rho=0.351$).
 261



262 **Figure 3.** Upset plot showing number of hotspots in different subsets. Horizontal bars
 263 represent total hotspots detected in a population, each dot on the matrix indicate that
 264 the vertical bar above it is the count of hotspots unique to that population, connected dots
 265 indicate that the vertical bar above them represents hotspots shared between the populations
 266 represented by the connected dots. The 25 largest subsets are shown.
 267
 268

269 The recombination rate in hotspot regions for nine of the populations was on average between
 270 22 and 237% higher than the average recombination rate of the genome. The exception was
 271 Guianna, which only showed an approximately 1% increase in average recombination rate in

272 hotspots regions when compared to that of the non-hotspot regions. For Guiana, we also
 273 compared the recombination rate inside hotspots to their surrounding regions (+/- 5kb). We
 274 found that hotspot regions had a rate ~42% higher rate than their neighboring regions. This
 275 result leads us to believe that the 1% higher average recombination rate in the Guiana
 276 hotspots when compared to the entire genome may be due to an increased ability to detect
 277 hotspots in regions of low recombination for this population. Additionally, Guiana presents
 278 unusually large hotspots (average 8.9 kb, Table S4), which points to an especially low resolution
 279 in hotspot detection for this population.

280
 281 Despite the majority of hotspots not being shared between populations, we conducted pairwise
 282 Fisher's exact tests to verify whether there was significantly more hotspot overlap than expected
 283 (if hotspots were randomly distributed along the genome) between populations. For most pairs
 284 of populations, we found significantly more hotspot overlap than expected (Table 2). There were
 285 three comparisons that did not show significantly more overlap than expected: Amelonado-
 286 Nacional, Amelonado-Purus, and Criollo-Nacional. A Mantel test comparing distances between
 287 populations based on shared hotspots and F_{ST} values between populations resulted in a
 288 significant correlation between them ($r=0.66$, $p=0.002$).

289
 290 **Table 2.** Fisher's exact test p-values for pairwise comparisons of recombination hotspot locations
 291 between populations of *T. cacao*. We conducted 45 comparisons, corresponding to a Bonferroni
 292 correction cutoff value of $\alpha=0.0011$.

Population	Ame	Con	Cri	Cur	Gui	Iqu	Mar	Nac	Nan
Amelonado	-	-	-	-	-	-	-	-	-
Contamana	<2e-07	-	-	-	-	-	-	-	-
Criollo	<9e-05	<5e-13	-	-	-	-	-	-	-
Curaray	<3e-05	<3e-37	<5e-08	-	-	-	-	-	-
Guiana	<3e-06	<1e-37	<7e-07	<4e-20	-	-	-	-	-
Iquitos	<4e-08	<6e-87	<2e-11	<3e-16	<2e-29	-	-	-	-
Maranon	<6e-13	<7e-77	<2e-11	<2e-20	<5e-33	<4e-64	-	-	-
Nacional	0.0015	<2e-43	0.0212	<7e-14	<3e-06	<6e-14	<3e-13	-	-
Nanay	0.0004	<2e-44	<9e-11	<4e-16	<2e-21	<3e-39	<2e-38	<9e-06	-
Purus	0.1782	<4e-117	<2e-05	<2e-29	<1e-33	<2e-39	<8e-43	<6e-27	<2e-21

293
 294
 295 To study the effects of demographic history more closely, shared hotspots were converted to
 296 dimensions of a multiple correspondence analysis and modeled along a previously constructed
 297 drift tree (Cornejo et al. 2018). Modeling the dimension as a Brownian motion was a better fit
 298 (AIC=79.4) than modeling it as an Ornstein-Uhlenbeck (OU) process (AIC=81.4), which is
 299 consistent with the small number of hotspots shared between populations. The model assuming
 300 Brownian motion is consistent with pure drift driving differentiation of a trait along a genealogy,
 301 while an OU process is consistent with a higher trait maintenance (stabilizing selection).

302 303 *Identifying DNA sequence motifs associated with the locations of recombination hotspots*

304
 305 We used RepeatMasker to analyze the set of recombination hotspots that were present in at
 306 least eight *T. Cacao* populations (17 total hotspots; referred to as ubiquitous hotspots), as well
 307 as the consensus set of recombination hotspots and the reference genome. In order to
 308 determine whether a particular set of DNA sequence repeats was overrepresented in ubiquitous
 309 hotspots, the percentage of DNA sequence that was identified as potentially being from
 310 retroelements or DNA transposon was compared to an empirical distribution. The percentage of
 311 observations from the distribution which were greater than the observed are reported in Table 3.

312 While retroelements were found to be underrepresented in the ubiquitous hotspots, DNA
 313 transposons were marginally overrepresented.

314
 315 **Table 3.** Percentage of DNA sequences identified as either retroelements or DNA transposons, and total
 316 interspersed repeats. Observed values for the entire *T. cacao* genome, for all recombination hotspots
 317 (HS), and ubiquitous hotspots (hotspots in the same location in at least eight different populations). Also
 318 presented are mean percentage of these sequences for 1000 simulations of hotspots equivalent in size
 319 and count as the ubiquitous set and the percentile at which the observed value for the ubiquitous set is
 320 found in the distribution of the simulated set (Sim).
 321

Measures	Observed % ubiquitous HS	Observed % all HS	Observed % whole genome	Mean % Sim	% Sim > ubiquitous
Retroelements	2.34	9.45	11.12	11.11	99.9
DNA transposons	1.94	1.64	1.10	1.10	5.4
Total	4.28	11.09	12.21	12.22	99.7

322
 323
 324
 325 *Identifying genomic features associated with the location of recombination hotspots*
 326
 327 We found an overrepresentation of recombination hotspots at transcriptional start sites (TSSs)
 328 and transcriptional termination sites (TTSs) in all ten of the *T. Cacao* populations (Table 4). The
 329 level of overrepresentation of hotspots in particular regions was compared to a null expectation
 330 based on simulations of hotspots of the same size as the ones detected, distributed randomly
 331 along the chromosomes. For all populations, all 1000 simulations showed a lower proportion of
 332 overlap with TSSs and TTSs than the observed. In the case of exons and introns, seven
 333 populations (Contamana, Criollo, Iquitos, Maranon, Nacional, Nanay, Purus) had an observed
 334 value that was lower than all, or almost all (Purus for exons), simulations. Three of the
 335 remaining four populations (Amelonado, Curaray, and Nanay) had no clear trend in either
 336 direction (Table 4). The final population (Guianna) showed an overrepresentation of hotspots in
 337 both exons and introns.

338
 339 **Table 4.** Proportion of simulated chromosomes that presented a lower number of hotspots intersecting
 340 with TSSs, TTSs, exons, and introns than the observed chromosomes. TSSs and TTSs are considered to
 341 be the 500bp upstream and downstream of transcribed regions, respectively.
 342

Population	TSSs (500bp)	TTSs (500bp)	Exons	Introns
Amelonado	1	1	0.602	0.527
Contamana	1	1	0	0
Criollo	1	1	0	0
Curaray	1	1	0.346	0.058
Guianna	1	1	1	1
Iquitos	1	1	0	0
Maranon	1	1	0	0
Nacional	1	1	0	0
Nanay	1	1	0.027	0.237
Purus	1	1	0.004	0

343
 344
 345
 346

347 **Discussion:**

348

349 The set of ten *T. cacao* populations, which includes wild, long-established and recently
350 established populations as well as a domesticated population, has provided us a unique
351 opportunity to study differences in recombination in populations of the same species with
352 varying evolutionary histories (Cornejo et al. 2018, Bartley et al. 2005). We also explored
353 differences in the location of recombination hotspots between the populations and found that the
354 conservation of hotspots between them generally mirrors their patterns of pairwise genetic
355 differentiation. Additionally, we found that TSSs and TTSs are strongly associated with
356 recombination hotspots in all populations, which is consistent with previous findings in plants
357 (Paape et al. 2012, Choi et al. 2013, Hellsten et al. 2013). This factor seems to play an
358 important role in determining the location of novel hotspots. Finally, hotspots that are shared by
359 at least eight populations appear to be associated with DNA transposons, pointing to a potential
360 mechanism for the maintenance of recombination hotspots at the population-divergence
361 timescale. Understanding how recombination rates vary between genetically differentiated
362 populations of the same species is an important step toward disentangling the role of
363 recombination in genetic differentiation.

364

365 *Comparing recombination rates between populations*

366

367 We found that the eight long-established, wild *T. cacao* populations show an average
368 recombination rate (r) within the range of recombination rates measured for other suporrosids
369 (Table 1; Stapley et al. 2016). The average recombination rate for these *T. cacao* populations,
370 was 4.07 cM/Mb, which is very similar to that of *Juglans regia* (4.05 cM/Mb; Zhu et al. 2015),
371 and comparable to other woody rosids (e.g. *Populus deltoides*, 8.13 cM/Mb (Mousavi et al.
372 2016); *Mangifera indica*, 7.15 cM/Mb (Luo et al. 2016); *Citrus clementina*, 3.6 cM/Mb (Ollitrault
373 et al 2012)) (Stapley et al. 2016). This places *T. cacao* on the high end of known recombination
374 rates for its order but comfortably in the range of other long-lived, woody plants. For all ten
375 populations, the mean recombination rate was found to be greater than the median. This is
376 consistent with high rate outlier values; an expected result in the presence of recombination
377 hotspots.

378

379 The other two *T. cacao* populations (Criollo, the domesticated variety and Guianna, the recently
380 established wild population) show unusually high average recombination rates when compared
381 to the eight long-established wild populations. Despite a small sample for some populations,
382 including Criollo, we found no linear trend between sample size and recombination rate (Table
383 S5). Additionally, the rates calculated for the two wild, small-sample populations (Curaray and
384 Nacional) were consistent with those of other wild populations. This makes us confident in our
385 estimates for the Criollo and Guianna populations. We also used the effective population size
386 for *Medicago truncatula* from Siol et al. 2007 and the estimate of ρ from Paape et al. 2012, to
387 calculate r for *M. truncatula* (=433 cM/Mb) and found that it was comparable with the mean rate
388 found for the Criollo population (Table 1). One possible explanation for the higher recombination
389 rate observed for the Criollo population is domestication; which has been observed to increase
390 recombination rates, particularly in plants (Ross-Ibarra 2004). Previous work has shown that
391 Criollo is the only population showing a strong signature of domestication, as revealed by its
392 much higher drift parameter compared to other populations (Cornejo et al. 2018). The high
393 recombination rate observed in Guianna (46.5 cM/Mb) can be explained in a similar way; while
394 Guianna does not show a strong signature of domestication, it is the most recently established
395 population (Bartley et al. 2005) and has undergone a recent bottleneck (Cornejo et al. 2018).
396 We hypothesize from this result that the Guianna population is undergoing the initial stages of
397 domestication, and its increased recombination is an early indicator of this. Another possibility is

398 that the high recombination rates estimated for Criollo and Guianna can be explained by biases
399 in estimation caused by errors associated to small samples or low genetic variation. However,
400 the recombination rates for Amelonado (another population with low variation) and Purus (a
401 population with small sample size) did not present this problem.

402
403 Analyses exploring mutations of putative recombination suppression genes (Fernandes et al.
404 2018b) could help disentangle the nature of this extreme variation in recombination rate in the
405 Criollo and Guianna populations. It has been widely discussed that although actual
406 recombination rates (chiasmata per bivalent) might increase in domestic plant populations,
407 effective recombination is likely to be reduced in domestic populations as a result of increased
408 LD and long runs of homozygosity. The interpretation for this observation is that it is likely that
409 chromosomes might physically recombine at higher rates, but if homologous chromosomes
410 contain the same sequence then no appreciable exchange of variants among different
411 chromatid backgrounds can occur (Moyers et al. 2018). We find significantly higher estimates of
412 median recombination rates for domesticated Criollo populations across the genome. Given
413 previous research suggesting the key function of FIGL-1 on suppressing recombination rates in
414 *Arabidopsis*, we investigated if mutations in this protein could explain the differences in
415 recombination rate. Our results suggest that missense mutations in FIGL-1 could impair the
416 activity of this protein in domesticated populations, reducing its efficacy as a recombination
417 suppressor. The mutation showing the largest effect (KK_215) is close to the domain that has
418 been previously shown to be involved in the interaction of FIGL-1 with RAD51 and DMC1
419 (Fernandes et al. 2018a) and thus lead us to propose that FIGL-1 might be responsible for the
420 increase recombination rate in domesticated cacao populations. Further work using yeast two-
421 hybrid systems similar to the work performed by Fernandes et al. (2018a) will be necessary in
422 the future to show if the mutations do indeed interfere with the normal function of FIGL-1. These
423 mutations were not found in increased frequency in Guianna, suggesting that the underlying
424 mechanisms leading to increased recombination rates in this population are different to those in
425 the Criollo population.

426
427
428 Despite recombination rates for eight of the ten populations being of the same order of
429 magnitude, pairwise comparisons of average rates indicated that most populations have a
430 significantly different rate of recombination from the others. The only exception were Nacional
431 and Nanay whose average rates were not significantly different from each other. These two
432 populations, however, are not more closely related to each other than they are to other
433 populations, based on genetic differentiation (Cornejo et al. 2017). We interpret this result as
434 suggestive that their similarity is not due to genetic similarity, but some other factors, e.g.
435 epigenetics.

436
437 The likelihood of detecting hotspots of recombination in the genome will likely be affected by the
438 amount of uncertainty in the estimates of recombination across sites or regions. Yet, we have
439 been unable to identify any study where the magnitude of the uncertainty in the estimates of
440 recombination are assessed to address this issue. We have performed careful comparisons and
441 assessed the magnitude of the uncertainty in the estimation of recombination rates to show that
442 this uncertainty is several orders of magnitude smaller than the variation in recombination rates
443 across the genome (Table S1).

444
445 *Comparing recombination hotspot locations between populations*

446
447 Understanding the pattern and rate of change of recombination hotspots at the population level
448 can elucidate their role in shaping genome architecture, impacting how effectively selection

449 operates (Felsenstein 1974). We found that a large proportion (55.5%) of hotspots detected are
450 unique to a single population. The observed variability of hotspot location between populations
451 points to demographic history not being the main driver of recombination hotspot location.
452 However, the hotspots tend to appear in similar regions, as demonstrated by the Fisher's exact
453 tests (Table 2). This dichotomy can be explained by considering that the proportion of the
454 genome occupied by recombination hotspots is very low, so even a small proportion of hotspots
455 from two different populations being in the same region is enough for the Fisher's exact test to
456 recognize them as significantly similar. This small but significant similarity can occur by
457 recombination being limited in its possible positioning along the genome, but not to the point of
458 forcing hotspots to occur consistently in the same locations, and thus maintaining some level of
459 stochasticity.

460
461 Given the significant proportion of overlapping hotspots between populations, it was still
462 important to explore whether the similarities can be explained by shared genetic history. If
463 demographic history explains the evolution of hotspot location, we would expect that more
464 closely related populations would have a higher percent of overlapped hotspots. A significant
465 relationship was found between population differentiation (F_{ST}) and the distance between
466 populations based on shared hotspots (Mantel test, $r=0.66$, $p=0.002$). This result indicates that,
467 to some extent, the genetic differentiation and the location of hotspots are mirroring each other,
468 which could be due to recombination hotspots being a product of the shared history between the
469 populations. However, since recombination rates were estimated using a coalescent-based
470 method, we expect historical relationships to be represented in our findings to some extent. We
471 transformed the information of hotspot overlap to model hotspots as quantitative traits changing
472 along a population tree (Cornejo et al. 2018). Our results show that a Brownian motion model
473 (AIC=79.4) better fits the shared hotspot data than a model with stabilizing selection Ornstein-
474 Uhlenbeck model (AIC=81.4), suggesting that, in principle, drift alone could explain the
475 evolution of the location of recombination hotspots. However, the absolute number of hotspots
476 that are shared among populations indicates that demographic history is insufficient to explain
477 the evolution of recombination hotspots in this species.

478
479 While we do not detect all the hotspots in these populations and not all the hotspots detected
480 are necessarily true positives, this proportion of unique hotspots can be seen as an indicator
481 that the turnover rate for hotspots is faster than the time it took the 10 populations to
482 differentiate. The detection rate for LDhot is approximately 55% under constant population
483 conditions, and greater when a recent bottleneck has occurred (Auton et al. 2014, Dapper and
484 Payseur 2017). Only two of the populations in this study (Criollo and Guianna) have a known
485 recent bottleneck (Cornejo et al. 2018). However, Criollo is the only one of these two with an
486 unusually low hotspot count (Table S4). Criollo's low number of detected hotspots can be a
487 product of its increased genome-wide recombination rate, making the signal of hotspots less
488 pronounced. In a similar fashion, the increased overall recombination rate of Guianna may be
489 affecting the detectability of hotspot regions, limiting our ability to resolve the limits of hotspots.
490 It is important to note that our hotspots are unusually large for all populations (Table S4), which
491 is likely a product of our low sample size leading to low resolution when resolving hotspot
492 regions. Further work, increasing the sample size per population will contribute to increasing the
493 resolution of these estimates.

494
495 While shared recombination hotspots can, to some extent, be explained by patterns of genetic
496 differentiation, some of the sharing can simply be due to a tendency for hotspots to arise near
497 TSSs and TTSSs. It has been observed in other organisms that hotspots of recombination are
498 frequently associated to specific genomic features (including TSSs and TTSSs) (Auton et al.
499 2013, Choi et al. 2013, Hellsten et al. 2013, Myers et al. 2005, Singhal et al. 2015) or DNA

500 sequence motifs (Auton et al. 2012, Brunshwig et al. 2012, Stevison et al. 2016). These factors
501 can affect the landscape of recombination, contributing to the patterns of shared hotspot
502 locations between populations that we observe in *T. cacao*.

503
504 Our ability to measure recombination in ten distinct populations allows us to analyze the
505 relationship between population genetic processes and recombination. Our results suggest that
506 the pattern of gains and losses of recombination hotspots is very dynamic and the landscape of
507 recombination changes rapidly during the process of diversification within a species. This
508 dynamism can have a tremendous impact on the adaptive dynamics of a species, and it should
509 be taken into account, considering that theoretical studies tend to assume that recombination
510 rates are constant during the evolution of populations (Hudson and Kaplan 1988, Donnelly and
511 Kurtz 1999).

512
513 *Identifying DNA sequence motifs associated with the locations of recombination hotspots*

514
515 The analysis of 17 hotspots shared between at least eight populations of *T. cacao* found an
516 underrepresentation of retroelements and a marginal overrepresentation of DNA transposons
517 when compared to the entire genome (Table 3). These results are not entirely surprising as it
518 has already been suggested that transposable elements (TEs) tend to be enriched in areas of
519 low recombination in *Drosophila* as a consequence of selection against TE activity that could
520 lead to chromosome instability (Rizzon et al. 2002). However, the marginal over-representation
521 of DNA transposons in the most conserved recombination hotspot is unexpected, given that all
522 previous observations have shown a reduced representation of mobile elements in areas with
523 high recombination rate (Rizzon et al. 2002). It is possible that DNA transposons are at least
524 partly responsible for the maintenance of recombination hotspots as populations diverge, from
525 which we expect that site-directed recombination is more frequent in these locations of the
526 genome. However, the low percentage of these sequences observed in the set of all hotspots
527 (Table 3) indicates that these sequences only have a small effect on the maintenance of
528 hotspots. It has been observed in humans that short DNA motifs enriched for repeat sequences
529 determine the location of 40 per cent of hotspots enriched for recurrent non-allelic homologous
530 recombination (McVean 2010). One potential explanation for why natural selection does not
531 eliminate hotspots in these regions is the possibility that these regions do not produce a large
532 enough mutational load for natural selection to remove them from the population (McVean2010
533 et al.).

534
535 *Identifying genomic features associated with the location of recombination hotspots*

536
537 For all ten populations, an overrepresentation of hotspots was found in the areas immediately
538 preceding and following transcribed regions of the chromosome. This matches the findings of
539 previous studies in *Arabidopsis thaliana* (Choi et al. 2013), *Taenipygia guttata* and *Poephila*
540 *acuticauda* (Singhal et al. 2015), and humans (Myers et al. 2005). The most likely explanation is
541 that recombination events within genes are selected against. The rationale being that a
542 recombinant chromosome that undergoes a double-strand break in the middle of a coding
543 region will have a higher risk of being inviable, and therefore not represented in the current set
544 of chromosomes for its population. Recombination occurring in transcription start and stop sites,
545 on the other hand, does a much better job at breaking up haplotypes or shuffling alleles in
546 different genomic backgrounds, while preserving the functionality of coding regions. This
547 rationale is supported by previous findings of increased recombination rates in these regions
548 (Choi et al. 2013). It is also supported by results from PRDM9 knock-out *Mus musculus*, which
549 has shown a reversion to hotspots located near TSSs (Brick et al. 2011). The enrichment of *T.*

550 *cacao* hotspots in TSSs and TTSs is thus a reasonable result given that zinc-finger binding
551 motifs and potential modifiers like PRDM9 have not been identified in this species.

552

553 *Implications for the evolutionary history of T. cacao*

554

555 Overall, our results show a large, consistent pattern where recombination rates in the ten
556 populations of *T. cacao* are of a similar magnitude as mutation rates but show a high diversity in
557 location and number of hotspots of recombination that cannot be explained solely by the
558 process of diversification of the populations. A potential hypothesis that could explain the rapid
559 turnover of hotspots of recombination and the relative differences in recombination among
560 populations is that epigenetic changes are involved in controlling the turnover of recombination
561 in plants. This hypothesis is not unreasonable given the recent observation of epigenetic control
562 of recombination in plants (Yelina et al. 2015). Further theoretical and simulation work should be
563 done in order to better understand the implications of the rapidly changing recombination
564 hotspots in adaptive dynamics. We also show that there is an overall underrepresentation of
565 hotspots in exons and introns for most populations, which is consistent with purifying selection
566 acting against changes that could result in disruptions of gene function. On the other hand, we
567 observed an overrepresentation of hotspots in TTSs and TSSs for all ten populations. This
568 could impact the maintenance and spread of beneficial traits in the population by shuffling allelic
569 variants of genes without causing disruption of their function. We hypothesize that the
570 enrichment of hotspots of recombination in TTSs and TSSs can have an important impact in the
571 spread of beneficial mutations across different genomic backgrounds; increasing the rate of
572 adaptation to selective pressures (e.g. selection for improved pathogen response).

573

574 **Materials and Methods:**

575

576 *Comparing recombination rates between populations*

577

578 Sequence data were downloaded from the Cacao Genome Database and NCBI (Accession
579 PRJNA486011), including the reference sequence for each chromosome and the full genome
580 annotation (*Theobroma cacao* cv. Matina 1-6 v1.1) (Motamayor et al. 2013). Processing was
581 done using the pipeline from (Cornejo et al 2018) available at the github repository
582 *oeco28/Cacao_Genomics*. Full genome data was used from a total of 73 individuals (146
583 chromosomes) across 10 populations (Cornejo et al. 2018). We filtered single nucleotide
584 polymorphism data and excluded rare variants (minor allele frequency ≤ 0.05) per population.
585 Separate variant files per population per chromosome were then phased using default
586 conditions with SHAPEIT2 (Delaneau et al. 2011) under default parameters. Haplotype files
587 were converted back to phased variant calling format (vcf) for its downstream analysis. We have
588 also phased the data with Beagle (Browning and Browning. 2007), using a burn-in of 10000
589 iterations, and estimations done over 10000 iterations. No appreciable differences were
590 observed between the two methods and Beagle phasing was maintained for the analyses. The
591 reason for performing the phasing separately for each population is that linkage disequilibrium
592 patterns are expected to be affected by population structure. The ten populations have been
593 shown to be unique clusters with very little admixture between them (Cornejo et al. 2018), and
594 the individuals used in this study were those whose ancestry was clearly from a single
595 population. VCFTools (Danecek et al. 2011) was used to remove all singletons and doubletons
596 (i.e. SNPs where the minor allele count is less than three). Only bi-allelic single nucleotide
597 polymorphisms (SNPs) were retained and were exported in LDhat format. The sample size and
598 post-filtering SNP count for each population can be found in Table S5.

599

600 In order to estimate recombination rates, we used the *interval* routine from LDhat (Auton and
601 McVean 2007), a program that implements coalescent resampling methods to estimate
602 historical recombination rates from SNP data. To reduce computation time, each chromosome
603 was split into windows, each containing 2000 SNPs. To counteract the overestimation of
604 recombination rate produced at the ends of the windows, an overlap of 500 SNPs was left
605 between consecutive windows. The final window for each chromosome did not always match
606 the general scheme, so the final 2000 SNPs were taken (making the overlap with the second to
607 last window variable, but never less than 500 SNPs) (Fig. S4). Once these windows were
608 generated, LDhat was run over each window using 100 million iterations, sampling every 10000
609 iterations (10000 total points sampled), with a block penalty of 5. Lookup tables with a grid of
610 100 points, population mutation rate parameters (θ) of 0.1 and 0.001, and a number of
611 sequences (n) of 50 were used for all populations. Downstream analyses were conducted using
612 the results from the $\theta=0.1$ runs of LDhat. We used the same θ for all populations since
613 estimates from (Cornejo et al. 2018) ranged from $\pi=0.27\%$ to $\pi=0.37\%$, all comfortably within an
614 order of magnitude of each other. The first 50 million iterations were discarded as burn-in. Once
615 recombination rates were calculated, 250 positions were cut off from both windows involved in
616 each overlap, so that the estimates for the first half of the overlap was taken from the end of the
617 preceding window and the estimates for the second half of the overlap were taken from the
618 beginning of the following window. The final overlap in each chromosome was split in order to
619 remove 250 SNPs from the second to last window, regardless of the remaining size of the last
620 window. The remaining rate estimates were then merged in order to obtain recombination rates
621 for the entire chromosome. This was done for each chromosome of each population.

622
623 The estimation of recombination rates with LDhat is approximated using a sampling scheme
624 with a Markov Chain Monte Carlo (MCMC) algorithm as implemented in the *interval* routine. The
625 inference of recombination rates is the result of the integration of estimated parameter values
626 across iterations with the routine *stats*. In the majority of recent studies where LDhat or
627 LDhelmet are used (Myers et al. 2005, Auton et al. 2012, Brunshwig et al. 2012, Paape et al.
628 2012, Auton et al. 2013, Choi et al. 2013, Singhal et al. 2015, Stevison et al. 2016), whether
629 there is convergence of the Markov chains has not been explicitly investigated. One study that
630 we are aware of has used simulations to assess whether their small sample size affected their
631 ability to obtain reliable estimates of recombination using LDhelmet (Booker et al. 2017) but did
632 not assess the uncertainty of the estimates from the MCMC process itself. We argue that
633 evaluation of convergence is important to assess the confidence in the estimated reported
634 values, especially if there is interest in analyzing the differences in recombination rate along the
635 genome. Visual inspection of pilot runs of the analysis demonstrated that convergence was not
636 achieved after running 40M iterations, which is why the length of the chains was increased to
637 100M iterations. Additionally, we explored the uncertainty in the estimates of recombination site-
638 wise by integrating over the trace of the estimates for recombination rate to infer the 95%
639 Credibility Interval. We then estimated the 95% interval of recombination estimates range
640 across all sites in the genome to have an overall measure of uncertainty that we compared to
641 the median 95% Credibility Interval for the trace of each position.

642
643 In order to compare recombination rates, the effective population size (N_e) calculated for each
644 population (Cornejo et al. 2018) was used to convert rates in $N_e r$ to r . We tested the difference
645 between the whole-genome mean rate of recombination (r) between populations using the
646 Wilcoxon signed-rank test `wilcox.test` function from the *stats* package in R) (R core team
647 2018). There were 45 comparisons, making the Bonferroni correction cutoff value: $\alpha=0.0011$.

648
649 Suppressors of recombination identified in Arabidopsis and other systems, FIGL1 and FLIP,
650 have not yet been identified in *T. cacao* (Motamayor et al. 2013). We performed a reciprocal

651 BLAST search using a newly generated databased containing sequences of FIGL1 and FLIP
652 obtained from ncbi (Altschul et al. 1990, Altschul et al. 1994b, Shiryev et al. 2007). After
653 identification of orthologous copies of FIGL1 and FLIP, we extracted the annotated variants
654 responsible for missense mutations in the proteins from data generated in Cornejo et al. 2018.
655 We then used these mutations to estimate the frequency of homozygous genotypes with
656 alternative alleles (different to the reference). Because the reference genome belongs to the
657 Amelonado populations, the population that presents the lowest estimated median
658 recombination rate in our work, we used the frequency of homozygous alternatives to infer the
659 impact of missense mutations under the assumption of a recessive model on the recombination
660 rate. For this, we estimated the homozygosity for missense mutations and eliminated those
661 found in complete correlation ($r^2 = 1$) and fitted a generalized linear model of the form:

$$y = \beta_0 + \beta_j + \varepsilon_j.$$

662
663
664
665 Where y is a vector of the median recombination rate accross populations, b_0 is the intersect (in
666 this case the Amelonado median recombination rate) and b_j is the effect of the homozygosity in
667 position j . We assumed a Gaussian link function for the model.

668 669 670 *Comparing recombination hotspot locations between populations*

671
672 Recombination hotspots were estimated with LDhot (Auton and McVean 2007), a likelihood-
673 based program that tests whether a single distribution model or a two-distribution model better
674 explains the observed recombination rates in 1 kb sliding windows (default), for each
675 chromosome. Each chromosome was run in its entirety, with the number of simulations (nsims)
676 set to 1000. The resulting potential hotspots were refined by an α of 0.001, and overlapping
677 hotspots were merged. This method therefore detects hotspots by comparing rates in 1 kb
678 windows to the rates in the surrounding regions.

679
680 To determine the set of consensus hotspots, the hotspots from all populations were merged.
681 Two hotspots from different populations were considered to be shared if they both overlapped
682 with the same hotspot in the consensus set. To summarize all shared hotspots, a Boolean
683 matrix was constructed, in which a population having a hotspot that overlaps with a hotspot in
684 the consensus list leads to an indication of presence of the consensus hotspot in that
685 population. This matrix was used to determine hotspots shared by two or more populations.

686
687 A Fisher's exact test was run for each pair of populations in order to determine whether hotspots
688 for the pair of populations overlap significantly more than expected. The BED files containing
689 the location of the recombination hotspots for each pair of populations were compared using
690 Bedtools:fisher (Quinlan and Hall. 2010). The number of comparisons was 45, making the
691 Bonferroni correction cutoff value: $\alpha=0.0011$.

692
693 In order to compare the relationships between populations based on shared hotspots we
694 calculated Jaccard distances (`distance` function, `phylentropy` package, R) (Drost 2018) and
695 compared them to a published F_{ST} matrix (Cornejo et al. 2018) using a Mantel test
696 (`mantel.rtest` function, `ade4` package, R) (Chessel et al. 2004, Dray et al. 2007a, Dray et al.
697 2007b, Bougeard and Dray 2018). The F_{ST} estimates from (Cornejo et al. 2018) were generated
698 using Weir and Cockerham's estimator (Weir and Cockerham 1984).
699

700 In order to model the presence or absence of hotspots along a drift tree, a multiple
701 correspondence analysis was used on the Boolean matrix of shared hotspots using the `MCA`
702 function from the `FactoMineR` package in R (Le et al. 2008). Nine dimensions were retained
703 and used as traits along a previously generated drift tree (Cornejo et al. 2018). Using the
704 `Rphylopars` package in R (Goolsby et al. 2016), the dimensions were modeled as Brownian
705 motion and as an Ornstein-Uhlenbeck process. The fit of the two models were compared using
706 the AIC values for the best fitting models of each type.

707 *Identifying DNA sequence motifs associated with the locations of recombination hotspots*

708 Motifs associated with hotspots were found using RepeatMasker (Smith et al. 2016). The entire
709 genome, the set of consensus hotspots, and a set of ubiquitous hotspots (hotspots shared by at
710 least eight of the populations) were examined with RepeatMasker, using normal speed and
711 "theobroma cacao" in the species option. In order to determine whether ubiquitous hotspots
712 were enriched for particular DNA sequences, a set of the same number and size of sequences
713 was randomly selected from the genome using `Bedtools:shuffle` (Quinlan and Hall. 2010) and
714 examined with RepeatMasker. This simulation was repeated one thousand times and a null
715 distribution against which observed values were compared was constructed from the results.

716 *Identifying genomic features associated with the location of recombination hotspots*

717 Testing whether recombination hotspots were overrepresented near particular genomic features
718 was done by using a resampling scheme to establish null expectations and then comparing the
719 observed value to the empirical distribution. For each feature, locations were retrieved and the
720 number of observed hotspots that overlap with this feature were counted. To determine whether
721 this number of overlapping hotspots was unusually high or low, a set of hotspots that matched
722 the number of hotspots and the size of each hotspot was simulated. These simulated hotspots
723 were placed randomly along the chromosome, using a uniform distribution. The simulation was
724 run 1000 times and the number of simulated hotspots that overlap with the true genomic
725 features was measured for each simulation. The simulations generate an expected distribution
726 of overlap with the genomic feature, and the true value was then compared to the distribution.
727 When simulated hotspots overlapped, the location of one of them was sampled again. Features
728 tested were: Transcriptional start sites (TSSs), transcriptional termination sites (TTSs), exons,
729 and introns. TSSs and TTSs are considered to be the 500bp upstream and downstream of
730 coding regions respectively.

731 The reason for the proposed novel resampling scheme is that, if the size and distribution of
732 genomic features and hotspots were not taken into account, it would set unrealistic expectations
733 for the overlap between features under a null model of no association. In this sense, the null
734 model would be inappropriate and potentially inflate the false positive rate.

735 **Data Access:**

736 Rate and summary files from LDhat runs as well as hotspots for each population will be placed
737 in a Dryad repository. Scripts for LDhat and LDhot runs as well as the resampling schemes
738 used and additional analysis is available in the following github repository:
739 [ejschwarzkopf/recombination-map](https://github.com/ejschwarzkopf/recombination-map).

740
741
742
743
744
745
746
747
748
749
750

751 **Acknowledgements:**

752
753 The authors would like to thank the Noe Higinbotham endowment and the WSU College of Arts
754 and Science for travel funds to EJS to present earlier versions of this work. We would like to
755 thank the Kamiak High Performance Computing Cluster at WSU for the infrastructure support to
756 run the analyses, and the Cornejo, Kelley, and Busch labs at WSU for feedback and edits on
757 the manuscript.

758
759 **References:**

- 760
761 Akhunov E, Goodyear A, Geng S, Qi L, Echaliier B, Gill B, Gustafson J, Lazo G, Chao S,
762 Anderson O, et al. 2003. The Organization and Rate of Evolution of Wheat Genomes
763 Are Correlated With Recombination Rates Along Chromosome Arms. *Genome Res.*
764 **13**:753–763. Doi:10.1101/gr.808603.
- 765 Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol*
766 *Biol.* **215**:403-410.
- 767 Altschul S, Boguski M, Gish W, Wootton J. 1994. Issues in searching molecular sequence
768 databases. *Nature Genet.* **6**:119-129
- 769 Anderson L, Salameh N, Bass H, Harper L, Cande W, Weber G, Stack S. 2004. Integrating
770 Genetic Linkage Maps With Pachytene Chromosome Structure in Maize. *Genetics.*
771 **166**:1923–1933.
- 772 Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. *Genome*
773 *res.* **17**:1219–1227.
- 774 Auton A, Myers S, McVean G. 2014. Identifying recombination hotspots using population
775 genetic data. bioRxiv doi: 1403.4264v1.
- 776 Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, Leffler EM, Bowden R, Aneas I,
777 Broxholme J, et al. 2012. A fine-scale chimpanzee genetic map from population
778 sequencing. *Science.* **336**:193–198. doi:10.1126/science.1216872.
- 779 Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Greally JM,
780 Wang J, et al. 2013. Genetic Recombination Is Targeted towards Gene Promoter
781 Regions in Dogs. *PLoS Genet.* **9**. doi:10.1371/journal.pgen.1003984.
- 782 Bartley BGD. 2005. *The genetic diversity of cacao and its utilization*. CABI. Wallingford, United
783 Kingdom.
- 784 Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with
785 recombination rates in *D. melanogaster*. *Nature.* **356**. doi:10.1038/356519a0.
- 786 Booker TR, Ness RW, Keightley PD. 2017. The Recombination Landscape in Wild House Mice
787 Inferred Using Population Genomic Data. *Genetics.* **207**:297–309.
788 doi:10.1534/genetics.117.300063.
- 789 Bougeard S, Dray S. 2018. Supervised Multiblock Analysis in R with the ade4 Package. *J Stat*
790 *Softw.* **86**:1–17. doi:10.18637/jss.v086.i01.
- 791 Branca A, Paape T, Zhou P, Briskine R, Farmer A, Mudge J, Bharti A, Woodward J, May G,
792 Gentzbittel L, et al. 2011. Whole-genome nucleotide diversity, recombination, and
793 linkage disequilibrium in the model legume *Medicago truncatula*. *P Natl Acad Sci USA.*
794 **108**:E864–E870.
- 795 Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. 2012. Genetic recombination
796 is directed away from functional genomic elements in mice. *Nature.* **485**:642–645.
797 doi:10.1038/nature11089.
- 798 Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data
799 inference for whole-genome association studies by use of localized haplotype clustering.

- 800 *Am J Hum Genet.* **81**.
- 801 Brunschwig H, Levi L, Ben-David E, Williams R, Yakir B, Shifman S. 2012. Fine-Scale Maps of
802 Recombination Rates and Hotspots in the Mouse Genome. *Genetics.* **191**:757–64.
- 803 Chessel D, Dufour A, Thioulouse J. 2004. The ade4 Package – I: One-Table Methods. *R News*,
804 **4**, 5–10.
- 805 Choi K, Zhao X, Kelly K, Venn O, Higgins J, Yelina N, Hardcastle T, Ziolkowski P, Copenhaver
806 G, Franklin F, et al. 2013. *Arabidopsis* meiotic crossover hot spots overlap with H2A.Z
807 nucleosomes at gene promoters. *Nature Genet.* **45**:1327–36.
- 808 Cornejo OE, Yee M-C, Dominguez V, Andrews M, Sockell A, Strandberg E, Livingstone D,
809 Stack C, Romero A, Umaharan P, et al. 2018. Population genomic analyses of the
810 chocolate tree, *Theobroma cacao* L., provide insights into its domestication process.
811 *Commun Biol.* **1**:167–167. doi:10.1038/s42003-018-0168-6.
- 812 Crow JF, Kimura, Motoo. 1970. An introduction to population genetics theory. Harper & Row,
813 New York.
- 814 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G,
815 Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics.*
816 **27**:2156–2158. doi:10.1093/bioinformatics/btr330.
- 817 Dapper AL, Payseur BA. 2018. Effects of Demographic History on the Detection of
818 Recombination Hotspots from Linkage Disequilibrium. *Mol Biol Evol.* **35**:335–353.
819 doi:10.1093/molbev/msx272.
- 820 Delaneau O, Marchini J, Zagury J. 2012. A linear complexity phasing method for thousands of
821 genomes. *Nat Methods.* **9**:179–81. doi:10.1038/nmeth.1785.
- 822 Donnelly P, Kurtz TG. 1999. Genealogical Processes for Fleming-Viot Models with Selection
823 and Recombination. *Ann Appl Probab.* **9**:1091–1148.
- 824 Dray S, Dufour A. 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. *J*
825 *Stat Softw*, **22**:1–20. doi: 10.18637/jss.v022.i04.
- 826 Dray S, Dufour A, Chessel D. 2007. The ade4 Package – II: Two-Table and K-Table Methods. *R*
827 *News*, **7**:47–52.
- 828 Drost HG. 2018. Philentropy: Information Theory and Distance Quantification with R. *J Open*
829 *Source Softw.* **3**:765.
- 830 Eyre-Walker A, Keightley P. 2007. The distribution of fitness effects of new mutations. *Nat Rev*
831 *Genet.* **8**:610–618. doi:10.1038/nrg2146.
- 832 Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics.* **78**:737–56.
- 833 Fernandes J, Duhamel M, Seguéla-Arnaud M, Froger N, Girard C, Choinard S, et al. 2018a.
834 FIGL1 and its novel partner FLIP form a conserved complex that regulates homologous
835 recombination. *PLoS Genet.* **14**:e1007317. doi:10.1371/journal.pgen.1007317
- 836 Fernandes JB, Séguéla-Arnaud M, Larchevêque C, Lloyd AH, Mercier R. 2018b. Unleashing
837 meiotic crossovers in hybrid plants. *P Natl Acad Sci USA.* **115**:2431–2436.
838 doi:10.1073/pnas.1713078114.
- 839 Girard C, Chelysheva L, Choinard S, Froger N, Macaisne N, Lehmemdi A, et al. 2015. AAA-
840 ATPase FIDGETIN-LIKE 1 and Helicase FANCM Antagonize Meiotic Crossovers by
841 Distinct Mechanisms. *PLoS Genet.* **11**: e1005369. doi:10.1371/journal.pgen.1005369
- 842 Goolsby EW, Bruggeman J, Ané C. 2017. Rphylopars: fast multivariate phylogenetic
843 comparative methods for missing data and within-species variation. *Methods in Ecol*
844 *Evol.* **8**:22–27. doi:10.1111/2041-210X.12612.
- 845 Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills

- 846 GS, Ross-Ibarra J, et al. 2009. A First-Generation Haplotype Map of Maize. *Science*.
847 **326**:1115–1117. doi:10.1126/science.1177837.
- 848 Haldane J. 1937. The Effect of Variation on Fitness. *Am Nat*. **71**.
- 849 Hellsten U, Wright K, Jenkins J, Shu S, Yuan Y, Wessler S, Schmutz J, Willis J, Rokhsar D.
850 2013. Fine-scale variation in meiotic recombination in *Mimulus* inferred from population
851 shotgun sequencing. *P Natl Acad Sci USA*. **110**.
- 852 Henderson J, Joyce R, Hall G, Hurst W, McGovern P. 2007. Chemical and archaeological
853 evidence for the earliest cacao beverages. *P Natl Acad Sci USA*. **104**.
- 854 Hinch A, Tandon A, Patterson N, Song Y, Rohland N, Palmer C, Chen G, Wang K, Buxbaum S,
855 Akylbekova E, et al. 2011. The landscape of recombination in African Americans.
856 *Nature*. **476**:170–5. doi:10.1038/nature10336.
- 857 Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and
858 recombination. *Genetics*. **120**:831–840.
- 859 Kim S, Plagnol V, Hu T, Toomajian C, Clark R, Ossowski S, Ecker J, Weigel D, Nordborg M.
860 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet*.
861 **39**:1151–1155. doi:10.1038/ng2115.
- 862 Lê S, Josse J, Husson F. 2008. FactoMineR : An R Package for Multivariate Analysis. *J Stat*
863 *Softw*. **25**. doi:10.18637/jss.v025.i01.
- 864 Li W-H, Nei M. 1974. Stable linkage disequilibrium without epistasis in subdivided populations.
865 *Theor Popul Biol*. **6**:173–183. doi:10.1016/0040-5809(74)90022-7.
- 866 Luo C, Shu B, Yao Q, Wu H, Xu W, Wang S. 2016. Construction of a High-Density Genetic Map
867 Based on Large-Scale Marker Development in Mango Using Specific-Locus Amplified
868 Fragment Sequencing (SLAF-seq). *Front Plant Sci*. **7**. doi:10.3389/fpls.2016.01310.
- 869 Mackiewicz D, de Oliveira PMC, Moss de Oliveira S, Cebrat S, Lustig AJ. 2013. Distribution of
870 Recombination Hotspots in the Human Genome – A Comparison of Computer
871 Simulations with Real Data. *PLoS ONE*. **8**. doi:10.1371/journal.pone.0065272.
- 872 Mackiewicz D, Zawierta M, Waga W, Cebrat S. 2010. Genome analyses and modelling the
873 relationships between coding density, recombination rate and chromosome length. *J*
874 *Theor Biol*. **267**:186–192. doi:10.1016/j.jtbi.2010.08.022.
- 875 Martinez-Perez E, Schvarzstein M, Barroso C, Lightfoot J, Dernburg AF, Villeneuve AM. 2008.
876 Crossovers trigger a remodeling of meiotic chromosome axis composition that is linked
877 to two-step loss of sister chromatid cohesion. *Gene Dev*. **22**:2886–901.
878 doi:10.1101/gad.1694108.
- 879 McVean G. 2010. What drives recombination hotspots to repeat DNA in humans? *Philos T Roy*
880 *Soc B*. **365**:1213–1218. doi:10.1098/rstb.2009.0299.
- 881 Mcvean G, Myers S, Hunt S, Deloukas P, Bentley D, Donnelly P. 2004. The fine-scale structure
882 of recombination rate variation in the human genome. *Science*. **304**:581–4.
- 883 Mézard C. 2006. Meiotic recombination hotspots in plants. *Biochem Soc T*. **34**:531–4.
- 884 Motamayor J, Lachenaud P, Loor R, Kuhn D, Brown J, Schnell R. 2008. Geographic and
885 Genetic Population Differentiation of the Amazonian Chocolate Tree (*Theobroma cacao*
886 L). *PLoS One*. **3**. doi:10.1371/journal.pone.0003311.
- 887 Motamayor J, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, Findley S, Zheng
888 P, Utró F, Royaert S, et al. 2013. The genome sequence of the most widely cultivated
889 cacao type and its use to identify candidate genes regulating pod color. *Genome Biol*.
890 **14**:r53–r53. doi:10.1186/gb-2013-14-6-r53.
- 891 Mousavi M, Tong C, Liu F, Tao S, Wu J, Li H, Shi J. 2016. De novo SNP discovery and genetic
892 linkage mapping in poplar using restriction site associated DNA and whole-genome
893 sequencing technologies. *BMC Genomics*. **17**. doi:10.1186/s12864-016-3003-9.
- 894 Moyers B, Morrell P, McKay J. 2018. Genetic Costs of Domestication and Improvement. *J*
895 *Hered*. **109**: 103–116.
- 896 Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of

- 897 recombination rates and hotspots across the human genome. *Science*. **310**:321–324.
898 Ohta T. 1982. Linkage disequilibrium due to random genetic drift in finite subdivided
899 populations. *P Natl Acad Sci USA*. **79**:1940–1944. doi:10.1073/pnas.79.6.1940.
900 Ollitrault P, Terol J, Chen C, Federici C, Lotfy S, Hippolyte I, Ollitrault F, Bérard A, Chauveau A,
901 Cuenca J, et al. 2012. A reference genetic map of *C. clementina* hort. ex Tan.; citrus
902 evolution inferences from comparative mapping. *BMC Genomics*. **13**. doi:10.1186/1471-
903 2164-13-593.
904 Otto S, Barton N. 1997. The evolution of recombination: removing the limits to natural selection.
905 *Genetics*. **147**:879–906.
906 Paape T, Zhou P, Branca A, Briskine R, Young N, Tiffin P. 2012. Fine-scale population
907 recombination rates, hotspots, and correlates of recombination in the *Medicago*
908 *truncatula* genome. *Genome Biol Evol*. **4**:726–737. doi:10.1093/gbe/evs046.
909 Pickrell J, Pritchard J. 2012. Inference of Population Splits and Mixtures from Genome-Wide
910 Allele Frequency Data. *PLoS Genetics*. **8**. doi: 10.1371/journal.pgen.1002967.
911 Ptak S, Hinds D, Koehler K, Nickel B, Patil N, Ballinger D, Przeworski M, Frazer K, Pääbo S.
912 2005. Fine-scale recombination patterns differ between chimpanzees and humans.
913 *Nature Genet*. **37**:429–434.
914 Quinlan A, Hall I. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
915 *Bioinformatics*. **26**:841–842. doi:10.1093/bioinformatics/btq033.
916 R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for
917 Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
918 Revelle W. 2019. psych: Procedures for Psychological, Psychometric, and Personality
919 Research. Northwestern University, Evanston, Illinois. R package version 1.9.12,
920 <https://CRAN.R-project.org/package=psych>.
921 Rizzon C, Marais G, Gouy M, Biéumont C. 2002. Recombination rate and the distribution of
922 transposable elements in the *Drosophila melanogaster* genome. *Genome Res*. **12**:400–
923 407.
924 Rodgers K, McVey M. 2016. Error-Prone Repair of DNA Double-Strand Breaks. *J Cell Physiol*.
925 **231**:15–24. doi:10.1002/jcp.25053.
926 Ross-Ibarra J. 2004. The Evolution of Recombination under Domestication: A Test of Two
927 Hypotheses. *Am Nat*. **163**:105–112. doi:10.1086/380606.
928 Sanjuán R, Moya A, Elena SF, Ohta T. 2004. The Distribution of Fitness Effects Caused by
929 Single-Nucleotide Substitutions in an RNA Virus. *P Natl Acad Sci USA*. **101**:8396–8401.
930 Schnable P, Ware D, Fulton R, Stein J, Wei F, Pasternak S, Liang C, Zhang J, Graves L, Minx
931 T, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science*.
932 **326**:1112–1115. doi:10.1126/science.1178534.
933 Shanfelter A, Archambeault S, White M, Hurst L. 2019. Divergent Fine-Scale Recombination
934 Landscapes between a Freshwater and Marine Population of Threespine Stickleback
935 Fish. *Genome Biol Evol*. **11**:1573–1585. doi:10.1093/gbe/evz090.
936 Shiryev S, Papadopoulos J, Schäffer A, Agarwala R. 2007. Improved BLAST searches using
937 longer words for protein seeding. *Bioinformatics*. **23**:2949–51.
938 Singhal S, Leffler E, Sannareddy K, Turner I, Venn O, Hooper D, Strand A, Li Q, Raney B,
939 Balakrishnan C, et al. 2015. Stable recombination hotspots in birds. *Science*. **350**:928–
940 932. doi:10.1126/science.aad0843.
941 Siol M, Bonnin I, Olivieri I, Prosperi J, Ronfort J. 2007. Effective population size associated with
942 self-fertilization: lessons from temporal changes in allele frequencies in the selfing
943 annual *Medicago truncatula*. *J Evolution Biol*. **20**:2349–2360. doi:10.1111/j.1420-
944 9101.2007.01409.x.
945 Smith A, Hubley R, Green P. 2016. *RepeatMasker Open-4.0* (2013-2015).
946 <<http://www.repeatmasker.org>>.
947 Stapley J, Feulner P, Johnston S, Santure A, Smadja C. 2017. Variation in recombination

948 frequency and distribution across eukaryotes: patterns and processes. *Philos T Roy Soc*
949 *B.* **372**. doi:10.1098/rstb.2016.0455.
950 Stevison L, Woerner A, Kidd J, Kelley J, Veeramah K, McManus K, Bustamante C, Hammer M,
951 Wall J. 2016. The Time Scale of Recombination Rate Evolution in Great Apes. *Mol Biol*
952 *Evol.* **33**:928–945. doi:10.1093/molbev/msv331.
953 Weir B, Cockerham C. 1984. Estimating f-statistics for the analysis of population structure.
954 *Evolution.* **38**:1358–1370. doi:10.1111/j.1558-5646.1984.tb05657.x.
955 Winckler W, Myers S, Richter D, Onofrio R. 2005. Comparison of Fine-Scale Recombination
956 Rates in Humans and Chimpanzees. *Science.* **308**:107–11.
957 Wloch D, Szafraniec K, Borts R, Korona R. 2001. Direct estimate of the mutation rate and the
958 distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. *Genetics.* **159**:441–
959 452.
960 Wu J, Mizuno H, Hayashi-Tsugane M, Ito Y, Chiden Y, Fujisawa M, Katagiri S, Saji S, Yoshiki S,
961 Karasawa W, et al. 2003. Physical maps and recombination frequency of six rice
962 chromosomes. *Plant J.* **36**:720–730. doi:10.1046/j.1365-313X.2003.01903.x.
963 Yelina N, Diaz P, Lambing C, Henderson IR. 2015. Epigenetic control of meiotic recombination
964 in plants. *Sci China Life Sci.* **58**:223–231. doi:10.1007/s11427-015-4811-x.
965 Zhu Y, Yin Y, Yang K, Li J, Sang Y, Huang L, Fan S. 2015. Construction of a high-density
966 genetic map using specific length amplified fragment markers and identification of a
967 quantitative trait locus for anthracnose resistance in walnut (*Juglans regia* L.). *BMC*
968 *Genomics.* **16**:614.
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995

996 **Supplementary Materials:**

997
998 Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots
999 among cocoa tree populations

1000
1001 Enrique J. Schwarzkopf, Juan C. Motamayor, Omar E. Cornejo (ocornejo@gmail.com)

1002
1003 **Tables:**

1004
1005 **Table S1.** The median of the upper and lower bounds of the 95% Credibility Interval for the trace of
1006 estimates of r (in cM/Mb) from all positions in the genome are presented for each population (i.e. Position
1007 L95 and Position U95). The upper and lower bounds of the 95% probability interval for the median
1008 estimate of r for each population is also presented (i.e. Genome L95 and Genome U95). The quotients of
1009 the upper and lower bounds for each of the two intervals point to a much larger genome-wide variation in
1010 r than per-position variation in the trace for the estimate of r .

1011

Population	Position L95	Position U95	Genome L95	Genome U95	Position Range Quotient	Genome Range Quotient
Amelonado	6.35e-05	7.67e-03	2.33e-05	31.3	120.75	1.34e+06
Contamana	1.63e-01	1.34	1.40e-04	26.4	8.22	1.88e+05
Criollo	87.5	43.1	5.35e-02	1,660	4.92	3.11e+04
Curaray	5.15e-01	2.63	2.72e-04	20.2	5.11	7.40e+04
Guianna	9.76e-01	12.6	1.81e-04	296	12.90	1.63e+06
Iquitos	5.50e-05	6.52e-03	1.58e-05	24.5	186.29	1.55e+06
Maranon	1.98e-04	7.40e-02	2.31e-05	35.2	373.35	1.52e+06
Nacional	4.80e-05	6.50e-03	2.76e-05	36.6	135.60	1.33e+06
Nanay	1.65e-04	3.06e-02	2.06e-05	35.2	185.32	1.71e+06
Purus	3.98e-03	5.24e-01	2.00e-04	63.5	131.87	3.18e+05

1012
1013 **Table S2.** Mean and median genome-wide recombination rates (r) in cM/Mb for all ten *T. cacao*
1014 populations obtained using LDhat with $\theta = 0.001$.

1015

Population	Mean r (cM/Mb)	Median r (cM/Mb)
Amelonado	4.04	4.78e-03
Contamana	4.55	9.75e-01
Criollo	715.35	668.90
Curaray	5.85	3.12
Guianna	63.24	4.34
Iquitos	2.68	7.95e-03
Maranon	4.16	3.38e-02
Nacional	6.00	1.32e-01
Nanay	5.43	1.61e-01
Purus	11.27	1.96

1016
1017
1018
1019
1020
1021

1022 **Table S3.** Name of *T. cacao* gene coding for FIGL1 and FLIP and amino acid mutations for FIGL1 and
 1023 FLIP orthologs.
 1024

Protein	FIGL1	FLIP
Gene	<i>Thecc1EG017182t2</i>	<i>Thecc1EG020410t1</i>
Mutations	K9Q, F66L, D114Y, T155I, D189V, K195M, E215K, S235N, R270Q, A285S, S291T, V548I, G566S, A567V, M580T	L102F, S153T, K174T, Y140N, V125A, H203R, Q237K, V406L, S430Y, R446Q, M497I, L521M, D696V, L729I, M739I, A746G, K875Q, P906L

1025 **Table S4.** Average hotspot size (in kb) and count for hotspots detected in each population
 1026 and average for all populations.
 1027
 1028

Population	Mean hotspot size (kb)	Hotspot count
Amelonado	6.9	1325
Contamana	6.1	5184
Criollo	6.1	887
Curaray	5.8	2303
Guianna	8.6	3655
Iquitos	7.0	3258
Maranon	6.8	3296
Nacional	6.9	2202
Nanay	7.6	3818
Purus	6.3	3972
Average	6.9	2989.9

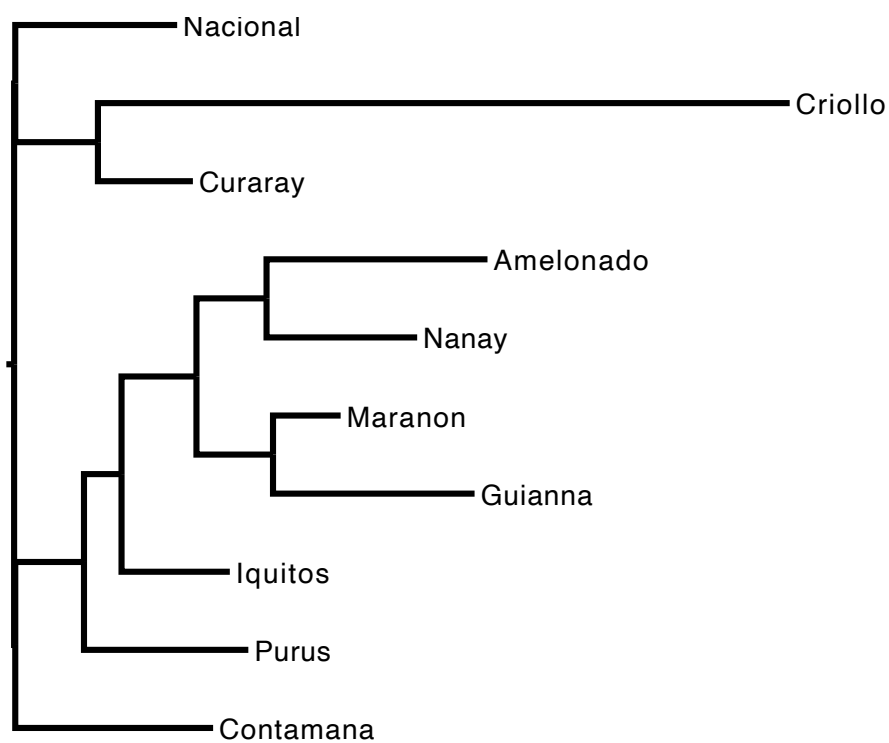
1029 **Table S5.** Sample size and post-filtering SNP count for all ten populations of *Theobroma cacao* for which
 1030 recombination maps were generated. The proportion of the genome that is callable is also reported.
 1031

Population	Sample size (chromosome)	SNP count	Proportion callable
Amelonado	11 (22)	373,789	0.68
Contamana	9 (18)	2,097,618	0.87
Criollo	4 (8)	309,818	0.69
Curaray	5 (10)	1,106,871	0.70
Guianna	9 (18)	770,729	0.915
Iquitos	7 (14)	1,575,711	0.922
Maranon	14 (28)	1,783,226	0.955
Nacional	4 (8)	718,099	0.89
Nanay	10 (20)	830,885	0.94
Purus	6 (12)	1,184,181	0.9

1046
 1047
 1048
 1049
 1050
 1051

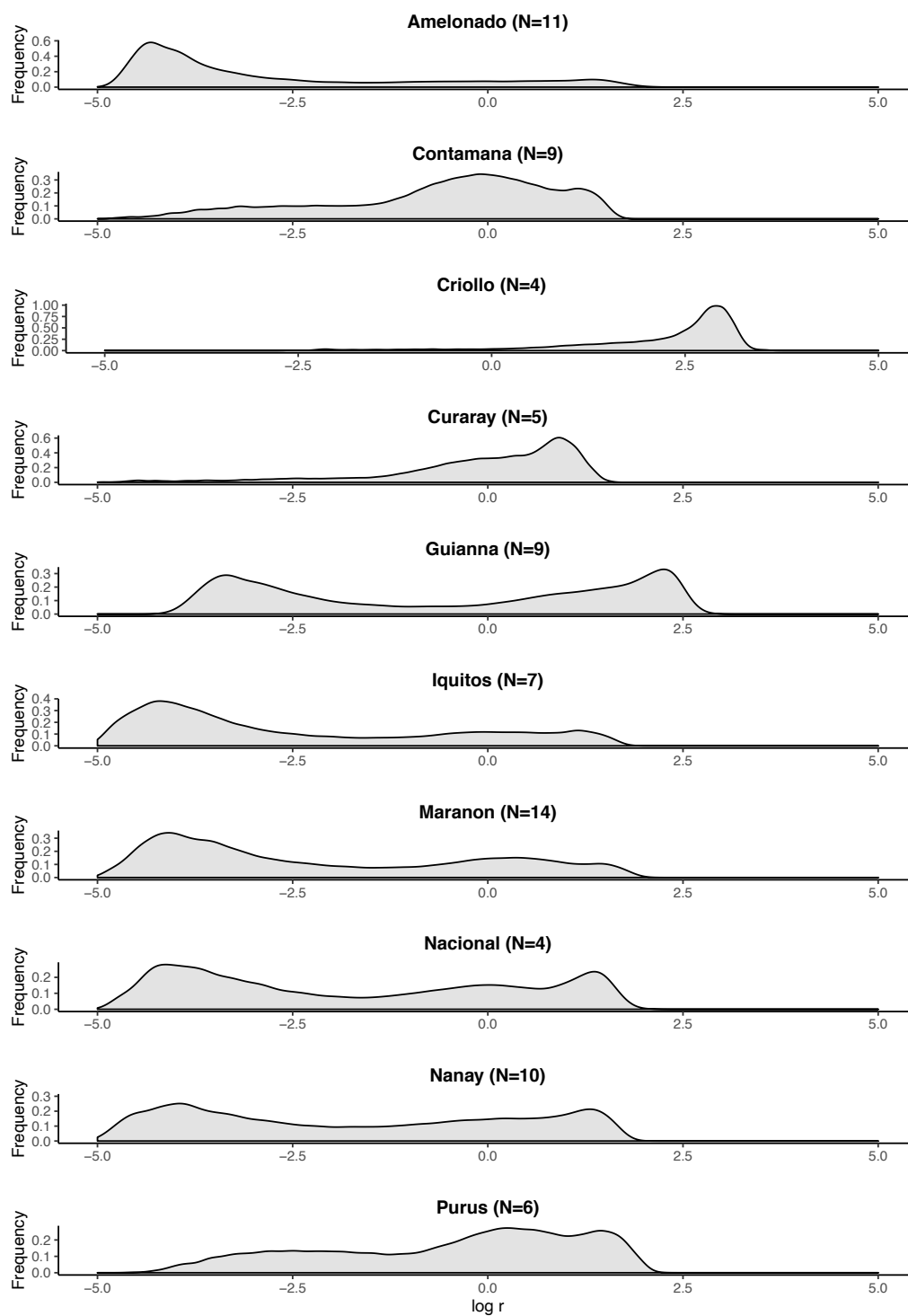
1052
1053
1054
1055
1056

Figures:



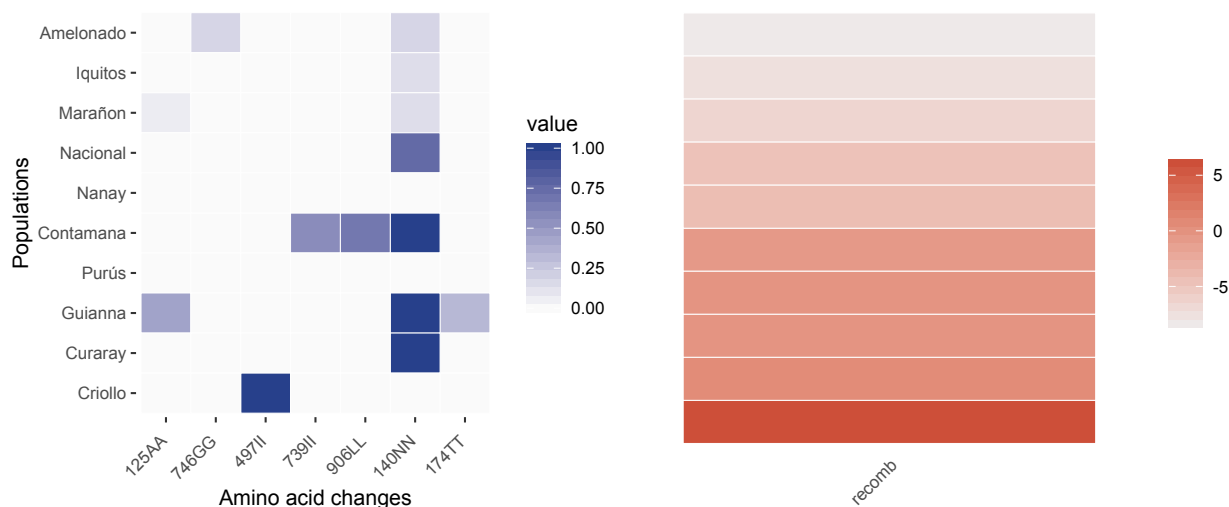
1057
1058
1059
1060
1061

Figure S1. Drift tree constructed using treemix (Pickrell and Pritchard, 2012) for the 10 *T. Cacao* populations. Distances between populations are based on the drift parameter. Modified from Cornejo et al. (2018)

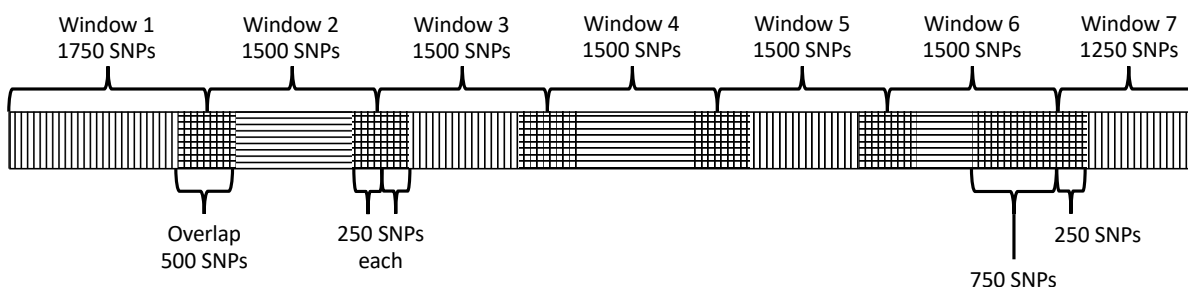


1062
1063
1064
1065

Figure 2. Distribution of \log_{10} recombination rates ($\log_{10}(r)$) along the genomes of the ten *T. Cacao* populations. The sample size (N) is reported for each population.



1066
1067
1068 **Figure S3.** The left panel shows the frequency of individuals that are homozygous for the alternative
1069 allele of amino acid mutations in a *T. cacao* FLIP ortholog. Alternative allele is defined in terms of the
1070 Amelonado reference genome. The right panel shows the \log_e transformed recombination rates (r). The
1071 populations are in the same order in both panels.
1072



1073
1074 **Figure S4.** Example of the window layout for a 10,750 SNP chromosome. The 2,000 SNP
1075 long windows are represented by alternating horizontal and vertical lines and the overlaps
1076 between them are represented by square crosshatches. Braces above the chromosome
1077 indicate the regions from which recombination rates are extracted to generate the chromosome-wide
1078 recombination rates.
1079