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

Enrique J. Schwarzkopf[1], Juan C. Motamayor[2], and Omar E. Cornejo[1,*]

[1]School of Biological Sciences, Washington State University, 100 Dairy Road, Pullman, WA 99164, USA

[2]Universal Genetic Solutions, LLC

[*]Corresponding author: ocornejo@gmail.com

April 20, 2019

Abstract

Our study investigates the possible drivers of recombination hotspots in *Theobroma cacao* using ten genetically differentiated populations. This constitutes the first time that recombination rates from more than two populations of the same species have been compared, providing a novel view of recombination at the population-divergence time-scale. For each population, a fine-scale recombination map was generated using under the coalescent with a standard method based on linkage disequilibrium (LD). They revealed higher recombination rates in a domesticated population and a population that has undergone a recent bottleneck. We address whether the pattern of recombination rate variation along the chromosome is sensitive to the uncertainty in the per-site estimates. We find that uncertainty, as assessed from the Markov chain Monte Carlo iterations is orders of magnitude smaller than the scale of variation of the recombination rates genome-wide. We inferred hotspots of recombination for each population and find that the genomic locations of these hotspots correlate with genetic differentiation between populations (F_{ST}). We developed novel randomization approaches to generate appropriate null models for understanding the association between hotspots of recombination and both DNA sequence motifs and genomic features. Hotspot regions contained fewer known retroelement sequences than expected, and were overrepresented near transcription start and termination sites. Our findings indicate that recombination hotspots are evolving in a way that is consistent with genetic differentiation, but are also preferentially driven to regions of the genome that are up or downstream from coding regions.

Introduction

Genetic variation is fundamental for evolutionary forces like selection and genetic drift to act. Selection and drift also contribute to a loss of variation, which means that they must act in conjunction with forces that maintain variation along the genome in order for populations to continue evolving over prolonged periods of time. Recombination's rearranging of genetic material onto different backgrounds generates a larger set of haplotype combinations on which selection can act, reducing the magnitude of Hill-Robertson interference (Felsenstein, 1974). Different regimes of recombination can strongly influence how efficient selection is at purging deleterious mutations and increasing the frequency of beneficial mutations in the population (Felsenstein, 1974).

One way to elucidate the distribution of recombination events along the genome is by using fine-scale recombination maps (Myers et al., 2005; Auton et al., 2012; Brunschwig et al., 2012; Paape et al., 2012; Choi et al., 2013; Hellsten et al., 2013; Singhal et al., 2015; Stevison et al., 2016). These maps are constructed with methods that leverage current patterns of linkage disequilibrium (LD) using the coalescent, in order to estimate historical rates of recombination between sites along the genome (Auton and McVean, 2007). Studies in a wide range of species have shown that recombination rates are not uniform along the genome and general patterns of variation have been described (Begun and Aquadro, 1992; Akhunov et al., 2003; Wu et al., 2003; Anderson et al., 2004; McVean et al., 2004; Mézard, 2006; Kim et al., 2007; Gore et al., 2009; Schnable et al., 2009; Branca et al., 2011; Paape et al., 2012). One of these patterns is the reduced recombination rate in centromeric regions of the chromosomes and the progressive increase of recombination rates as the physical distance to telomeres decreases (Begun and Aquadro, 1992; Akhunov et al., 2003; Wu et al., 2003; Anderson et al., 2004; Gore et al., 2009; Schnable et al., 2009). This pattern has also been shown to arise in simulation studies (e.g. Mackiewicz et al., 2010). Another interesting pattern that has been

observed is that of regions with unusually high rates of recombination spread throughout chromosomes: recombination hotspots (McVean et al., 2004; Brunschwig et al., 2012; Paape et al., 2012; Hellsten et al., 2013; Stevison et al., 2016; Shanfelter et al., 2018). In this study, we define hotspots locally, requiring that their recombination rate be unusually high when compared to neighboring regions. The importance of recombination hotspots lies in their ability to shuffle genetic variation at higher rates than the rest of the genome, profoundly impacting the dynamics of selection for or against specific mutations (Felsenstein, 1974).

A variety of genomic features have been identified as being associated with regions of high recombination. Recombination hotspots have been linked to transcriptional start sites (TSSs) and transcriptional termination sites (TTSs) in *Arabidopsis thaliana*, *Taeniopygia guttata*, *Poephila acuticauda*, and humans (Myers et al., 2005; Choi et al., 2013; Singhal et al., 2015). In *Mimulus guttatus* hotspots were found to be associated with CpG islands (short segments of cytosine and guanine rich DNA, associated with promoter regions) (Hellsten et al., 2013). CpG islands were also associated with increased recombination rates in humans and chimpazees (Auton et al., 2012). These patterns point to recombination occurring frequently near, but not within coding regions. The formation of chiasmata is important for the proper disjunction of chromosomes during meiosis (Martinez-perez et al., 2008), but repeated double-strand breaks can lead to an increased mutation rate (Rodgers and McVey, 2015). In coding regions in particular, this excess mutation rate can have a high evolutionary cost, due to the likelihood of novel deleterious mutations being higher than that of beneficial ones (Haldane, 1937; Crow and Kimura, 1970; Wloch et al., 2001; Sanjuán et al., 2004; Eyre-Walker and Keightley, 2007). Recombination hotspots have also been found to be correlated with particular DNA sequence motifs. In some mammals, including *Mus musculus* (Brunschwig et al., 2012) and apes (Auton et al., 2012; Stevison et al., 2016) binding sites for PRDM9, a histone trimethylase with a DNA zinc-finger binding domain, have been found to correlate with recombination hotspots. In *Arabidopsis thaliana*, proteins that limit overall

recombination rate have been identified, leading to a genome-wide increase in recombination rate in knockout mutants (Fernandes et al., 2018). However, these *Arabidopsis* proteins have not been shown to direct recombination to particular regions, and are therefore not expected to affect the location of recombination hotspots.

Comparisons of recombination hotspots between pairs of populations have yielded varying results. Hinch et al. (2011) found that, at finer scales, the genetic maps of European and African human populations were significantly different. They also found that, when looking at hotspots in the major histocompatibility complex, the African populations showed a hotspot that was not present in Europeans, but all European hotspots were found in African populations (Hinch et al., 2011). Recent work on recombination in apes (Stevison et al., 2016) found little correlation of recombination rates in orthologous hotspot regions when looking between species, but a strong correlation when comparing between two populations of the same species. Other studies have also found very little sharing of hotspots between humans and chimpanzees Ptak et al. (2005); Winckler et al. (2005). Additionally, the dynamic of changing hotspot locations observed in humans and other apes has been observed in simulations Mackiewicz et al. (2013). This suggests that recombination hotspots are potentially changing in ways that match demographic patterns, differentiating at a similar rate as genomic sequences.

The identification of ten genetically differentiated populations of the cocoa tree, *Theobroma cacao* (Motamayor et al., 2008; Cornejo et al., 2018) can be leveraged to study population-level drivers of recombination hotspots. These ten populations originate from different regions of South and Central America, and include one fully domesticated population (Criollo), used in the production of fine chocolate, and nine wilder, more resilient populations which generate higher cocoa yield than the Criollo variety (Fig. 1) (Motamayor et al., 2008; Henderson et al., 2007; Cornejo et al., 2018). These ten populations have been shown to have strong signatures of differentiation between them (F_{ST} values ranging from

0.16 to 0.65) and they separate into clear clusters of ancestry (Cornejo et al., 2018). Comparing the locations of hotspots between these ten populations of *T. cacao* can contribute to the understanding of hotstpot turnover at the population-divergence time-scale. These comparisons also contribute to our understanding of how demographics impact the turnover of recombination hotspot locations.

Fine-scale, LD-based recombination maps have been constructed for a number of plant models (Paape et al., 2012; Choi et al., 2013; Hellsten et al., 2013), identifying a variety of features correlated to recombination rate. Unlike these model plants with short generation times, *T. cacao* is a perennial woody plant with a five-year generation time (Henderson et al., 2007). The size and long generation time of *T. cacao* makes direct measurements of recombination impractical. However, historical recombination can be estimated for *T. cacao* using coalescent based methods (Auton and McVean, 2007). Theoretical studies have shown that population structure can generate artificially inflated measures of LD (Ohta, 1982; Li and Nei, 1974), which would be detrimental to our estimates of recombination. For this reason recombination maps were constructed independently for each population. In contrast to previous studies, which have focused primarily on recombination rates, this study attempts to describe the relationship between recombination hotspots and a variety of factors.

We used an LD-based method to estimate recombination rates, which we then analyzed with a maximum likelihood statistical framework to infer the location of recombination hotspots. The location of hotspots were compared across populations and a novel resampling scheme tailored to the genomic architecture of *T. cacao* was used to generate null assumptions for the distribution of hotspots along the genome. These null distributions were used to identify differential representation of known DNA sequence motifs in ubiquitous recombination hotspots, and of overlap between recombination hotspots and genomic traits for each population. The re-sampling schemes used to identify these associations are

novel in the context of this work and were designed to take into account the size and distribution of elements in the genome. In this work we aimed to answer the following questions: (i) How are recombination rates distributed within 10 highly differentiated populations of *T. cacao*, and how do they compare to each other? (ii) How are hotspots distributed along the genome of each of the ten populations of *T. cacao*, and can these distributions be explained by patterns of population genetic differentiation? (iii) Are there identifiable DNA sequence motifs that are associated with the location of recombination hotspots along the *T. cacao* genome? (iv) Are there genomic features (e.g. TSSs, TTSs, exons, introns) consistently associated with recombination hotspot locations across *T. cacao* populations? Our findings suggest that recombination hotspot locations generally follow patterns of diversification between populations, while also having a strong tendency to occur close to TSSs and TTSs. Moreover, we find a strong negative association between the occurrence of recombination hotspots and the presence of retroelements.

Results

Comparing recombination rates between populations

Populations show a mean recombination rate r/kb between 2.1×10^{-5} and 5.25×10^{-3} (Table 1), with a variety of distributions (Fig. 2). We observe a higher mean than median r/kb for all populations, indicating that extreme high values are present for all populations. The extreme recombination rate values affect the mean, driving it to values consistently higher than the median. The pattern of recombination rates along the genome varied between populations, as can be seen in the comparison of the Nanay and Purus third chromosome (Fig. 3). Purus appears to have a higher average recombination rate than Nanay for chromosome

three. More specifically, particular regions of the chromosome present peaks in one population that are absent in the other. A similar patter can also be observed for the density of recombination hotspots, e.g. Purus presenting a high density of hotspots in certain regions that is not observed in Nanay. The median 95% probability interval for recombination rate across the genome for each population was found to be several orders of magnitude larger than the uncertainty per site, estimated as the median 95% Credibility Interval of the trace for each position in the genome for that population (Table 2).

Overall, the mean recombination rate for most of the populations was higher than estimated mutation rates for multicellular eukaryotes of 10^{-6} changes per kb per generation (Lynch, 2010; Exposito-Alonso et al., 2018) (Table 1). Two populations, Guianna and Criollo, were notable exceptions, having higher average recombination rates than the other populations by one and two orders of magnitude respectively. Guianna and Criollo also have been estimated to have a lower effective population size (N_e) (Cornejo et al., 2018) by one and two orders of magnitude respectively. However, there was no significant linear trend between mean N_e and r/kb ($p = 0.1119$), indicating that, for a high enough N_e, the ability to detect recombination events is not dictated by the effective population size. When Criollo and Guianna were excluded, the relationship was also not present ($p = 0.3886$). When all populations were included, the inbreeding coefficient (F, from Cornejo et al., 2018) showed no significant linear association with mean r/kb ($p = 0.3361$). We also found no linear trend between sample size and mean r/kb ($p = 0.2333$). The average recombination rate per population was transformed from r/kb to cM/Mb (Table 1) using the Kosambi mapping function Kosambi (1943). The average cM/Mb was 4.6×10^{-04}.

In order to compare the average recombination rates (r/kb) of the different populations, a Kruskal-Wallis test was performed for every pair of populations. The only pair of populations that did not show a significant difference in mean recombination rate was the pair of Nacional and Nanay ($p = 0.3$). All other pairwise comparisons were highly significant

183 $(p < 2 \times 10^{-16})$.

184

185 *Comparing recombination hotspot locations between populations*

186

187 The majority (55.5%) of hotspots identified were not shared between populations. The 25
188 most numerous sets of hotspots are represented in Fig. 5. The nine largest of these are sets
189 of hotspots unique to single populations. The hotspots unique to the remaining population
190 (Criollo) formed the eleventh largest set. Effective population size (N_e) is not a good linear
191 predictor of the amount of detected hotspots ($p = 0.1489$), nor is sample size ($p = 0.351$).
192 The recombination rate in hotspot regions for nine of the populations was on average be-
193 tween 22 and 237% higher than the average recombination rate of the genome. The exception
194 was Guianna, which only showed an approximately 1% increase in average recombination
195 rate in hotspots regions when compared to that of the non-hotspot regions. A 1% higher
196 average recombination rate in hotspots may be due to an increased ability to detect hotspots
197 in regions of low recombination for this population. Additionally, Guianna presents unusu-
198 ally large hotspots (average 8.9 kb, Table 6), which points to an especially low resolution in
199 hotspot detection for this population.
200 Despite the majority of hotspots not being shared between populations, we conducted
201 pairwise Fisher's exact tests to verify whether there was significantly more hotspot overlap
202 than expected (if hotspots were randomly distributed along the genome) between popu-
203 lations. For most pairs of populations we found significantly more hotspot overlap than
204 expected (Table 3). There were three comparisons that did not show significantly more
205 overlap than expected: Amelonado-Nacional, Amelonado-Purus, and Criollo-Nacional. A
206 Mantel test comparing distances between populations based on shared hotspots and F_{ST}
207 values between populations resulted in a significant correlation between them ($r = 0.66$,
208 $p = 0.002$). The correlation between eigenvectors from a correlation matrix and those of the

genetic covariance matrix were also explored. When all populations were included, we found that the first eigenvector from the genetic covariance matrix was not significantly correlated with the first eigenvector from the hotspot correlation matrix ($p = 0.7055$), but the second genetic eigenvector was ($p = 0.009007$, $r = 0.7711638$). However, the first eigenvector of the genetic covariance matrix captured the difference between the Criollo population (the only domesticated variety) and the rest of the populations. The second eigenvector explains most of the natural differentiation across populations (Cornejo et al., 2018). For that reason, we decided to exclude Criollo and repeat the analysis. We found that the first eigenvector from the correlation matrix constructed from shared hotspot information was not significantly correlated with either of the first two eigenvectors of the genetic covariance matrix when Criollo was excluded (eigenvector 1: $p = 0.1314$, eigenvector2: $p = 0.3376$).

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

Identifying DNA sequence motifs associated with the locations of recombination hotspots

RepeatMasker was used to analyze the set of recombination hotspots that were present in at least eight *T. cacao* populations (17 total hotspots), as well as the consensus set of recombination hotstpots, and the reference genome. In order to determine whether a particular set of DNA sequence repeats was overrepresented in the regions of ubiquitous

recombination hotspots, the percentage of DNA sequence that was identified as potentially being from retroelements or DNA transposon was compared to an empirical distribution. The percentage of observations from the distribution which were greater than the observed are reported in Table 4. While retroelements were found to be underrepresented in the ubiquitous hotspots, DNA transposons were marginally overrepresented.

Identifying genomic features associated with the location of recombination hotspots

An overrepresentation of recombination hotspots was found in all ten of the populations at transcriptional start sites (TSSs) and transcriptional termination sites (TTSs)(Table 5). The level of overrepresentation of hotstpots in particular regions was compared to a null expectation based on simulations of hotspots of the same size as the ones detected, distributed randomly along the chromosomes. For all populations, all 1000 simulations showed a lower proportion of overlap with TSSs and TTSs than the observed. In the case of exons and introns, seven populations (Contamana, Criollo, Iquitos, Maranon, Nacional, Nanay, Purus) had an observed value that was lower than all, or almost all (Purus for exons), simulations. Three of the remaining four populations (Amelonado, Curaray, and Nanay) had no clear trend in either direction (Table 5). The final population (Guianna) showed an overrepresentation of hotspots in both exons and introns.

Discussion

Understanding how recombination rates vary between genetically differentiated populations of the same species is an important step toward disentangling the role of recombination in genetic differentiation. This set of *T. cacao* populations presents a unique opportunity to

infer recombination in wild, long- and recently established populations, as well as a domesticated population (Criollo) (Cornejo et al., 2018; Bartley, 2005). This system has allowed us to explore differences in recombination hotspot locations between populations of the same species. Our results point to a conservation of hotspots between populations that generally mirrors the patterns of genetic differentiation between populations. Also, we find that TSSs and TTSs are strongly associated with recombination hotspots in all populations, which is consistent with previous findings in plants (Paape et al., 2012; Choi et al., 2013; Hellsten et al., 2013). This factor seems to play an important role in determining the location of novel hotspots. Finally, hotspots that are shared by at least eight populations appear to be associated with DNA transposons, pointing to a potential mechanism for the maintenance of recombination hotspots at the population-divergence time-scale.

Comparing recombination rates between populations

We found that the eight long-established, wild *T. cacao* populations show an average recombination rate (r/kb) greater than multicellular eukaryotic mutation rates (Table 1), while the other two populations (Criollo and Guianna) show unusually high average recombination rates in comparison. Despite a small sample for some populations, we found no linear trend between sample size and recombination rate. Additionally, the rates calculated for the two wild, small-sample populations (Curaray and Nacional) were consistent with those of other wild populations. This makes us confident in our estimates, particularly for the domesticated Criollo population. For all populations, the mean recombination rate was found to be greater than the median. This is consistent with high rate outlier values; an expected result in the presence of recombination hotspots. Using the effective population size for *Medicago truncatula* from Siol et al., 2007 and the estimate of rho from Paape et al., 2012, we calculated r/kb ($= 4 \times 10^{-3}$) and found that it was comparable with the rate found

for the Criollo population (Table 1). We also calculated the median recombination rate in cM/Mb for each chromosome using the Kosambi mapping function (Kosambi, 1943) over non-overlapping, 100 SNP windows. The average cM/Mb for all populations was 4.6×10^{-04}, which is lower than has been measured for any Malvale (Kundu et al., 2015), but not as low as the lowest measured for conifers (Chen et al., 2010; Stapley et al., 2017). This places *T. cacao* on the high end of known recombination rates for its order but comfortably in the range of other long-lived, woody plants. Average recombination rates in cM/Mb varied between populations from Amelonado (4.04×10^{-06}) to Criollo ($3.91 \times^{-03}$). Previous work has shown that Criollo is the only population showing a strong signature of domestication, as revealed by much higher drift parameter than that observed for other populations (Cornejo et al., 2018). Domestication has been observed to increase recombination rates, particularly in plants (Ross-Ibarra, 2004), and is a possible explanation for the higher recombination rate observed for the Criollo population. The high recombination rate observed in Guianna can be explained in a similar way; while Guianna does not show a strong signature of domestication, it is the most recently established population (Bartley, 2005), and it has also undergone a recent bottleneck (Cornejo et al., 2018). We hypothesize from this result that the Guianna population is undergoing the initial stages of domestication and its increased recombination is an early indicator of this. It is possible that the high recombination rates estimated for Criollo and Guianna can be explained by biases in estimation caused by errors associated to small samples or low genetic variation; yet, the recombination rates for Amelonado (another population with low variation) or Purus (a population with small sample size) did not present this problem. Analyses exploring mutations of putative recombination suppression genes (Fernandes et al., 2018) could help disentangle the nature of this extreme variation in recombination rate in the Criollo and Guianna populations.

Despite recombination rates for eight of the ten populations being of the same order of magnitude, pairwise comparisons of average rates indicated that most populations have

a significantly different rate of recombination from the others. The only exception were Nacional and Nanay whose average rates were not significantly different from each other. These two populations, however, are not more closely related to each other than they are to other populations, based on genetic differentiation (Cornejo et al., 2018). We interpret this result as suggestive that their similarity is not due to genetic similarity, but some other factors, e.g. epigenetics.

The likelihood of detecting hotspots of recombination in the genome will likely be affected by the amount uncertainty in the estimates of recombination across sites or regions. Yet, we have been unable to identify any study where the magnitude of the uncertainty in the estimates of recombination are assessed to address this issue. We have performed careful comparisons and assessed the magnitude of the uncertainty in the estimation of recombination rates to show that this uncertainty is several orders of magnitude smaller than the variation in recombination rates across the genome (Table 2).

Comparing recombination hotspot locations between populations

Similarly to recombination rates, the location of recombination hotspots can be very informative to questions of divergence between populations. Understanding the pattern and rate of change of recombination hotspots at the population level can elucidate their role in shaping genome architecture, impacting how effectively selection operates (Felsenstein, 1974). We found that a large proportion (55.5%) of hotspots detected are unique to a single population. While we do not detect all the hotspots in these populations and not all the hotspots detected are necessarily true positives, this proportion of unique hotspots can be seen as an indicator that the turnover rate for hotspots is faster than the time it took the 10 populations to differentiate. The detection rate for LDhot is approximately 55% under constant population conditions, and greater when a recent bottleneck has occurred (Auton

et al., 2014; Dapper and Payseur, 2017). Only two of the populations in this study (Criollo and Guianna) have a known recent bottleneck (Cornejo et al., 2018). However, Criollo was the only one of these two with an unusually low hotspot count (Table 6). Criollo's low number of detected hotspots can be a product of its increased genome-wide recombination rate, making the signal of hotspots less pronounced. The observed variability of hotspot location between populations points to demographic history not being the main driver of recombination hotspot location. However, the hotspots tend to appear in similar regions, as demonstrated by the Fisher's exact tests (Table 3). This dichotomy can be explained by considering that the proportion of the genome occupied by recombination hotspots is very low, so even a small proportion of hotspots from two different populations being in the same region is enough for the Fisher's exact test to recognize them as significantly similar. This small but significant similarity can occur by recombination being limited in its possible positioning along the genome, but not to the point of forcing hotspots to occur consistently in the same locations, and thus maintaining some level of stochasticity. It is important to note that our hotspots are unusually large (Table 6). This is likely a product of our low sample size leading to low resolution when resolving hotspot regions.

Given the significant proportion of overlapping hotspots between populations, it was still important to explore whether the similarities can be explained by shared genetic history. If demographic history explains the evolution of hotspot location, we would expect that more closely related populations would have a higher percent of overlapped hotspots. A significant relationship was found between population differentiation (F_{ST}) and the distance between populations based on shared hotspots (Mantel test, $r = 0.66$, $p = 0.002$). The comparison between the hotspot correlation matrix and the genetic covariance matrix supports what was found when comparing the hotspot correlation matrix to the F_{ST} matrix. One caveat is that the first genetic eigenvector, which separated Criollo from the other populations, was not correlated with the first hotspot correlation eigenvector, indicating that Criollo's

domestication generated a genetic pattern that deviates from the pattern of shared hotspots. This indicates that, to some extent, the genetic differentiation and the location of hotspots are mirroring each other, which could be due to recombination hotspots being a product of the shared history between the populations. However, since recombination rates were estimated using a coalescent-based method, we expect historical relationships to be represented in our findings. We transformed the information of hotspot overlap to model hotspots as quantitative traits changing along a population tree (Cornejo et al., 2018). Our results, show that a Brownian motion model (AIC=79.4) better fits the data than a model with stabilizing selection Ornstein-Uhlenbeck model (AIC=81.4) and suggest that, in principle, drift alone could explain the evolution of the location of recombination hotspots. However, the absolute number of hotspots that are shared among populations indicates that demographic history alone is insufficient to explain the evolution of recombination hotspots in this species.

One conclusion that follows from these results is that, while shared recombination hotspots can to some extent be explained by patterns of genetic differentiation, some of the sharing can simply be due to a tendency for hotspots to arise near TSSs and TTSs. It has been observed in other organisms that hotspots of recombination are frequently associated to specific genomic features (including TSSs and TTSs) (Auton et al., 2013; Choi et al., 2013; Hellsten et al., 2013; Myers et al., 2005; Singhal et al., 2015) or DNA sequence motifs (Auton et al., 2012; Brunschwig et al., 2012; Stevison et al., 2016). These factors can affect the landscape of recombination, contributing to the patterns of shared hotspot locations between populations that we are observing in *T. cacao*. Previous studies looking at apes and finches have explored recombination hotspots in multiple species and as many as two populations of the same species (Singhal et al., 2015; Stevison et al., 2016; Shanfelter et al., 2018), but this study is the first to compare hotspots in more than two populations of the same species at once. The increased number of populations allows us to analyze the relationship between population genetic processes and recombination. Our results suggest that the pattern of

gains and losses of recombination hotspots is very dynamic and the landscape of recombination changes rapidly during the process of diversification within a species. This dynamism can have a tremendous impact on the adaptive dynamics of a species, and it should be taken into account, considering that theoretical studies tend to assume that recombination rates are constant during the evolution of populations (Hudson and Kaplan, 1988; Donnelly and Kurtz, 1999).

Identifying DNA sequence motifs associated with the locations of recombination hotspots

The analysis of 17 hotspots shared between at least eight populations of *T. cacao* found an underrepresentation of retroelements and a marginal overrepresentation of DNA transposons when compared to the entire genome (Table 4). These results are not entirely surprising as it has already been suggested that transposable elements (TEs) tend to be enriched in areas of low recombination in *Drosophila* as a consequence of selection against TEs (Rizzon et al., 2002). However, the marginal over-representation of DNA transposons in the most conserved recombination hostspot is unexpected, given that all previous observations have shown a reduced representation of mobile elements in areas with high recombination rate (Rizzon et al., 2002). It is possible that DNA transposons are at least partly responsible for the maintenance of recombination hotspots as populations diverge, from which we expect that site-directed recombination is more frequent in these locations of the genome. However, the low percentage of these sequences observed in the set of all hotspots (Table 4) indicates that these sequences only have a small effect on the maintenance of hotspots. It has been observed in humans that short DNA motifs enriched for repeat sequences determine the location of 40 per cent of hotspots enriched for recurrent non-allelic homologous recombination (McVean, 2010). One potential explanation for why natural selection does not eliminate hotspots in these regions is the possibility that these regions do not produce a large enough

mutational load for natural selection to remove them from the population (McVean, 2010).

Identifying genomic features associated with the location of recombination hotspots

For all ten populations, an overrepresentation of hotspots was found in the areas immediately preceding and following transcribed regions of the chromosome. This matches the findings of previous studies in *Arabidopsis thaliana* (Choi et al., 2013), *Taenipygia guttata* and *Poephila acuticauda* (Singhal et al., 2015), and humans (Myers et al., 2005). The most likely explanation is that recombination events within genes are selected against. The rationale being that a recombinant chromosome that undergoes a double-strand break in the middle of a coding region will have a higher risk of being inviable, and therefore not represented in the current set of chromosomes for its population. Recombination occurring in transcription start and stop sites, on the other hand, does a much better job at breaking up haplotypes or shuffling alleles in different genomic backgrounds, while preserving the functionality of coding regions. This rational is supported by previous findings of increased recombination rates in these regions (Choi et al., 2013). It is also supported by results from PRDM9 knock-out *Mus musculus*, which has shown a reversion to hotspots located near TSSs (Brick Kevin et al., 2012). The enrichment of *T. cacao* hotspots in TSSs and TTSs is thus a reasonable result given that zinc-finger binding motifs and potential modifiers like PRDM9 have not been identified in this species.

Implications for the evolutionary history of T. cacao

Overall, our results show a large consistent pattern where recombination rates in the ten populations of *T. cacao* are of a similar magnitude as mutation rates, but show a high diversity in location and number of hotspots of recombination that cannot be explained solely

by the process of diversification of the populations. In fact, the results are indicative of the turnover rate of hotspots being faster than the process of divergence among populations. A potential hypothesis that could explain the rapid turnover of hotspots of recombination and the relative differences in recombination among populations is that epigenetic changes are involved in controlling the turnover of recombination in plants. This hypothesis is not unreasonable given the recent observation of epigenetic control of recombination in plants (Yelina et al., 2015). Further theoretical and simulation work should be done in order to better understand the implications of the rapidly changing recombination hotspots in adaptive dynamics. We also show that there is an overall underrepresentation of hotspots in exons and introns for most populations, which is consistent with purifying selection acting against changes that could result in disruptions of gene function. On the other hand, we observed an overrepresentation of hotspots in TTSs and TSSs for all ten populations. This could impact the maintenance and spread of beneficial traits in the population by shuffling allelic variants of genes without causing disruption of their function. We hypothesize that the enrichment of hotspots of recombination in TTSs and TSSs can have an important impact in the spread of beneficial mutations across different genomic backgrounds; increasing the rate of adaptation to selective pressures (e.g. selection for improved pathogen response).

Materials and Methods

Comparing recombination rates between populations

Sequence data were downloaded from the Cacao Genome Database and NCBI (Accession PRJNA486011), including the reference sequence for each chromosome and the full genome annotation (*Theobroma cacao* cv. Matina 1-6 v1.1)(Motamayor et al., 2013). Processing

was done using the pipeline from (Cornejo et al., 2018) available at the github repository *oeco*28/*Cacao_Genomics*. Full genome data was used from a total of 73 individuals across 10 populations (Cornejo et al., 2018): Criollo (N = 4, #SNPs = 309,818), Curaray (N = 5, #SNPs = 1,106,871), Contamana (N = 9, #SNPs = 2,097,618), Amelonado (N = 11, #SNPs = 373,789), Maranon (N = 14, #SNPs = 1,783,226), Guianna (N = 9, #SNPs = 770,729), Iquitos (N = 7, #SNPs = 1,575,711), Purus (N = 6, #SNPs = 1,184,181), Nanay (N = 10, #SNPs = 830,885), and Nacional (N = 4, #SNPs = 718,099). We filtered the single nucleotide polymorphism data and excluded rare variants (minor allele frequency <= 0.05) per population. Separate variant files per population per chromosome were then phased using default conditions with SHAPEIT2 (Delaneau et al., 2011) under default parameters. Haplotype files were converted back to phased variant calling format (vcf) for its downstream analysis. We have also phased the data with Beagle (Browning and Browning, 2007), using a burnin of 10000 iterations, and estimations done over 10000 iterations. No appreciable differences were observed between the two methods and Beagle phasing was maintained for the analyses. The reason for performing the phasing separately for each population is that linkage disequilibrium patterns are expected to be affected by population structure. The ten populations have been shown to be unique clusters with very little admixture between them Cornejo et al. (2018), and the individuals used in this study were those whose ancestry was clearly from a single population. VCFTools (Danecek Petr et al., 2011) was used to remove all singletons and doubletons. Only bi-allelic single nucleotide polymorphisms (SNPs) were retained and were exported in LDhat format.

In order to estimate recombination rates we used the *interval* routine of LDhat (Auton and McVean, 2007), a program that implements coalescent resampling methods to estimate historical recombination rates from SNP data. To reduce computation time, each chromosome was split into windows, each containing 2000 SNPs. To counteract the overestimation of recombination rate produced at the ends of the windows, an overlap of 500 SNPs was left

between consecutive windows. The final window for each chromosome did not always match the general scheme, so the final 2000 SNPs were taken (making the overlap with the second to last window variable, but never less than 500 SNPs) (Fig. 6). Once these windows were generated, LDhat was run over each window using 100 million iterations, sampling every 10000 iterations (10000 total points sampled), with a block penalty of 5. Lookup tables with a grid of 100 points, a population mutation rate parameter (θ) of 0.1 and a number of sequences (n) of 50 were used for all populations. We used the same θ for all populations since estimates from Cornejo et al. (2018) ranged from $\pi = 0.27\%$ to $\pi = 0.37\%$, all comfortably within an order of magnitude of each other. The first 50 million iterations were discarded as burn-in. Once recombination rates were calculated, 250 positions were cut off from both windows involved in each overlap, so that the estimates for the first half of the overlap was taken from the end of the preceding window and the estimates for the second half of the overlap were taken from the beginning of the following window. The final overlap in each chromosome was split in order to remove 250 SNPs from the second to last window, regardless of the remaining size of the last window. The remaining rate estimates were then merged in order to obtain recombination rates for the entire chromosome. This was done for each chromosome of each population.

The estimation of recombination rates with LDhat is approximated using a sampling scheme with a Markov Chain Monte Carlo (MCMC) algorithm as implemented in the *interval* routine. The inference of recombination rates is the result of the integration of estimated parameter values across iterations with the routine *stats*. In the majority of recent studies where LDhat or LDhelmet are used (Myers et al., 2005; Auton et al., 2012; Brunschwig et al., 2012; Paape et al., 2012; Auton et al., 2013; Choi et al., 2013; Singhal et al., 2015; Stevison et al., 2016), whether there is convergence of the Markov chains has not been explicitly investigated. One study that we are aware of has used simulations to asses whether their small sample size affected their ability to obtain reliable estimates of recombination using

LDhelmet (Booker et al., 2017), but did not assess the uncertainty of the estimates from the MCMC process itself. We argue that evaluation of convergence is important to assess the confidence in the estimated reported values, especially if there is interest in analyzing the differences in recombination rate along the genome. Visual inspection of pilot runs of the analysis demonstrated that convergence was not achieved after running 40M iterations, which is why the length of the chains was increased to 100M iterations. Additionally we explored the uncertainty in the estimates of recombination site-wise by integrating over the trace of the estimates for recombination rate to infer the 95% Credibility Interval. We then estimated the 95% interval of recombination estimates range across all sites in the genome to have an overall measure of uncertainty that we compared to the median 95% Credibility Interval for the trace of each position.

In order to compare recombination rates, the effective population size (N_e) calculated for each population (Cornejo et al., 2018) was used to convert rates in $N_e r/kb$ to r/kb. Differences in the mean genome-wide recombination rate between populations were then tested using the Kruskal-Wallis test (`kruskal.test` function from the `stats` package in `R`) (R Core Team, 2018). There were 45 comparisons, making the Bonferroni correction cutoff value: $\alpha = 0.0011$. To transform per population recombination rates from r/kb to cM/Mb, we divided each chromosome into windows of 100 SNPs and used the Kosambi mapping function (Kosambi, 1943). The median for the windows of a chromosome was then calculated, and the average of each population's chromosomes was taken as that population's average recombination rate in cM/Mb.

Comparing recombination hotspot locations between populations

Recombination hotspots were estimated with LDhot (Auton and McVean, 2007), a likelihood-based program that tests whether a single distribution model or a two distribution model

better explains the observed recombination rates in 1 kb sliding windows (default), for each chromosome. Each chromosome was run in its entirety, with the number of simulations (nsims) set to 1000. The resulting potential hotspots were refined by an alpha of 0.001, and overlapping hotspots were merged. This method therefore detects hotspots by comparing rates in 1 kb windows to the rates in the surrounding regions.

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

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

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

The Boolean matrix for shared hotspots was also used to explore the relationship between hotspot similarities and genetic covariances from a previous study (Cornejo et al., 2018). Singletons were removed from the hotspot matrix, which was converted to a corre-

lation matrix using the `mixed.cor` function from the `psych` package in `R` (Revelle, 2018). The `mixed.cor` function was used due to its ability to calculate Pearson correlations from dichotomous data. We then used the `eigen` function in `R` (R Core Team, 2018) to generate eigenvectors for the hotspot correlation matrix and the genetic covariance matrix. Pearson correlations between the first and second eigenvector of the genetic covariance matrix and the hotspot correlation matrix were then calculated (`cor.test` function, `stats` package, `R`)(R Core Team, 2018). This analysis was done once with all populations included, and once with the Criollo population excluded before correlations were calculated.

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

Identifying DNA sequence motifs associated with the locations of recombination hotspots

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

595 from the results.

596

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

598

599 Testing whether recombination hotspots were overrepresented near particular genomic
600 features was done by using a resampling scheme to establish null expectations and then
601 comparing the observed value to the empirical distribution. For each feature, locations were
602 retrieved and the number of observed hotspots that overlap with this feature were counted.
603 To determine whether this amount of overlapping hotspots was unusually high or low, a set of
604 hotspots that matched the number of hotspots and the size of each hotspot was simulated.
605 These simulated hotspots were placed randomly along the chromosome, using a uniform
606 distribution. The simulation was run 1000 times and the number of simulated hotspots that
607 overlap with the true genomic features was measured for each simulation. The simulations
608 generate an expected distribution of overlap with the genomic feature, and the true value
609 was then compared to the distribution. When simulated hotspots overlapped, the location
610 of one of them was sampled again. Features tested were: Transcriptional start sites (TSSs),
611 transcriptional termination sites (TTSs), exons, and introns. TSSs and TTSs are considered
612 to be the 500bp upstream and downstream of coding regions respectively.

613 The reason for the proposed novel resampling scheme is that, if the size and distribution
614 of genomic features and hotspots were not taken into account, it would set unrealistic expec-
615 tations for the overlap between features under a null model of no association. In this sense,
616 the null model would be inappropriate and potentially inflate the false positive rate.

617

618 *Data and code availability*

619

620 Rate and summary files from LDhat runs as well as hotspots for each population will

be placed in a Dryad repository. Scripts for LDhat and LDhot runs as well as the re-sampling schemes used and additional analysis is available in the following github repository $ejschwarzkopf/recombination-map$.

1 Acknowledgments

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

Tables

Population	Mean 4Ner/kb	Mean Ne	Mean r/kb	Median r/kb	Lower Bound (Mean r/kb)	Upper Bound (Mean r/kb)	Mean cM/Mb
Amelonado	1.58	15744	2.51e-05	2.40e-09	2.48e-05	2.54e-05	4.04e-06
Contamana	8.53	61102	3.49e-05	4.92e-06	3.48e-05	3.50e-05	7.74e-05
Criollo	14.60	695	5.25e-03	4.27e-03	5.23e-03	5.27e-03	3.91e-03
Curaray	10.36	58213	4.45e-05	1.78e-05	4.44e-05	4.46e-05	1.18e-04
Guianna	8.66	4651	4.65e-04	7.74e-06	4.63e-04	4.67e-04	2.74e-04
Iquitos	4.23	49984	2.11e-05	5.88e-09	2.10e-05	2.12e-05	1.84e-05
Maranon	4.09	34037	3.01e-05	1.64e-08	2.99e-05	3.02e-05	1.68e-05
Nacional	4.66	26060	4.47e-05	9.76e-08	4.44e-05	4.49e-05	4.10e-05
Nanay	6.82	42429	4.02e-05	1.51e-07	4.00e-05	4.04e-05	1.33e-05
Purus	5.95	17357	8.57e-05	7.74e-06	8.54e-05	8.60e-05	1.23e-04

Table 1: Recombination rates in $4N_e r/kb$, r/kb, and cM/Mb for all ten *T. cacao* populations. The N_e that was used for the transformation is also reported for each population, as are the lower and upper bounds of a 95% confidence interval for r/kb.

Population	Position L95	Position U95	Genome L95	Genome U95	Position Range Quotient	Genome Range Quotient
Amelonado	6.35e-10	7.67e-08	2.33e-10	3.13e-04	120.75	1.34e+06
Contamana	1.63e-06	1.34e-05	1.40e-09	2.64e-04	8.22	1.88e+05
Criollo	8.75e-04	4.31e-03	5.35e-07	1.66e-02	4.92	3.11e+04
Curaray	5.15e-06	2.63e-05	2.72e-09	2.02e-04	5.11	7.40e+04
Guianna	9.76e-06	1.26e-04	1.81e-09	2.96e-03	12.90	1.63e+06
Iquitos	3.50e-10	6.52e-08	1.58e-10	2.45e-04	186.29	1.55e+06
Maranon	1.98e-09	7.40e-07	2.31e-10	3.52e-04	373.35	1.52e+06
Nacional	4.80e-10	6.50e-08	2.76e-10	3.66e-04	135.60	1.33e+06
Nanay	1.65e-09	3.06e-07	2.06e-10	3.52e-04	185.32	1.71e+06
Purus	3.98e-08	5.24e-06	2.00e-09	6.35e-04	131.87	3.18e+05

Table 2: The median of the upper and lower bounds of the 95% Credibility Interval for the trace of estimates of r from all positions in the genome are presented for each population (i.e. Position L95 and Position U95). The upper and lower bounds of the 95% probability interval for the median estimate of r for each population is also presented (i.e. Genome L95 and Genome U95). The quotients of the upper and lower bounds for each of the two intervals point to a much larger genome-wide variation in r than per-position variation in the trace for the estimate of r.

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	-	-	-	-	-	-
Guianna	<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

Table 3: Fisher's exact test p-values for pairwise comparisons of recombination hotspot locations between populations of *T. cacao*

Measures	Observed % ubiquitous HS	Observed % all HS	Observed % whole genome	Mean % Sim	% Sim >ubiquitous HS
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

Table 4: Percentage of DNA sequences identified as either retroelements or DNA transposons, and total interspersed repeats. Observed values for the entire *T. cacao* genome, for all recombination hotspots (HS), and ubiquitous hotspots (hotspots in the same location in at least eight different populations). Also presented are mean percentage of these sequences for 1000 simulations of hotspots equivalent in size and count as the ubiquitous set and the percentile at which the observed value for the ubiquitous set is found in the distribution of the simulated set (Sim).

	TSSs (500bp)	TTSs (500bp)	Exon	Intron
Amelonado	1	1	0.602	0.527
Contamana	1	1	0.000	0.000
Criollo	1	1	0.000	0.000
Curaray	1	1	0.346	0.058
Guianna	1	1	1.000	1.000
Iquitos	1	1	0.000	0.000
Maranon	1	1	0.000	0.000
Nacional	1	1	0.000	0.000
Nanay	1	1	0.027	0.237
Purus	1	1	0.004	0.000

Table 5: Proportion of simulated chromosomes that presented a lower amount of hotspots intersecting with TSSs, TTSs, exons, and introns than the observed chromosomes. TSSs and TTSs are considered to be the 500bp upstream and downstream of transcribed regions, respectively.

Population	Mean Hotspot Size (kb)	Hotspot Count
Amelonado	6.9	1324
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

Table 6: Average hotspot size (in kb) and count for hotspots detected in each population and average for all populations.

Figures

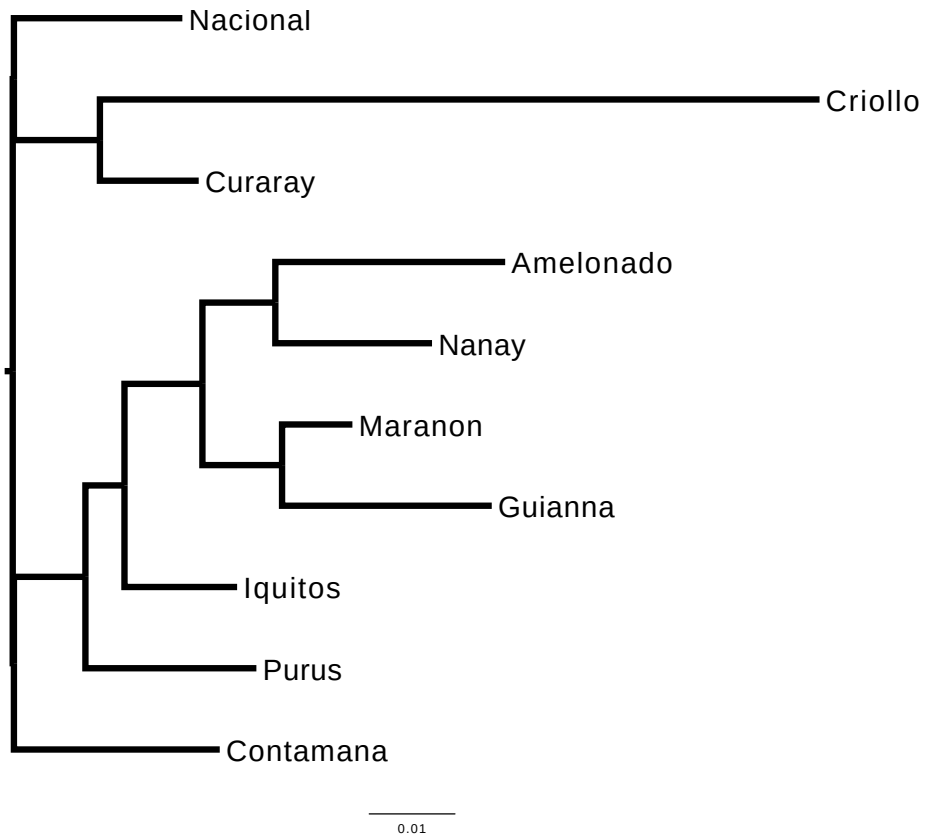

Figure 1: Drift tree constructed using TreeMix (Pickrell and Pritchard, 2012) for the 10 *T. cacao* populations. Distances between populations are based on the drift parameter. Modified from Cornejo et al. (2018)

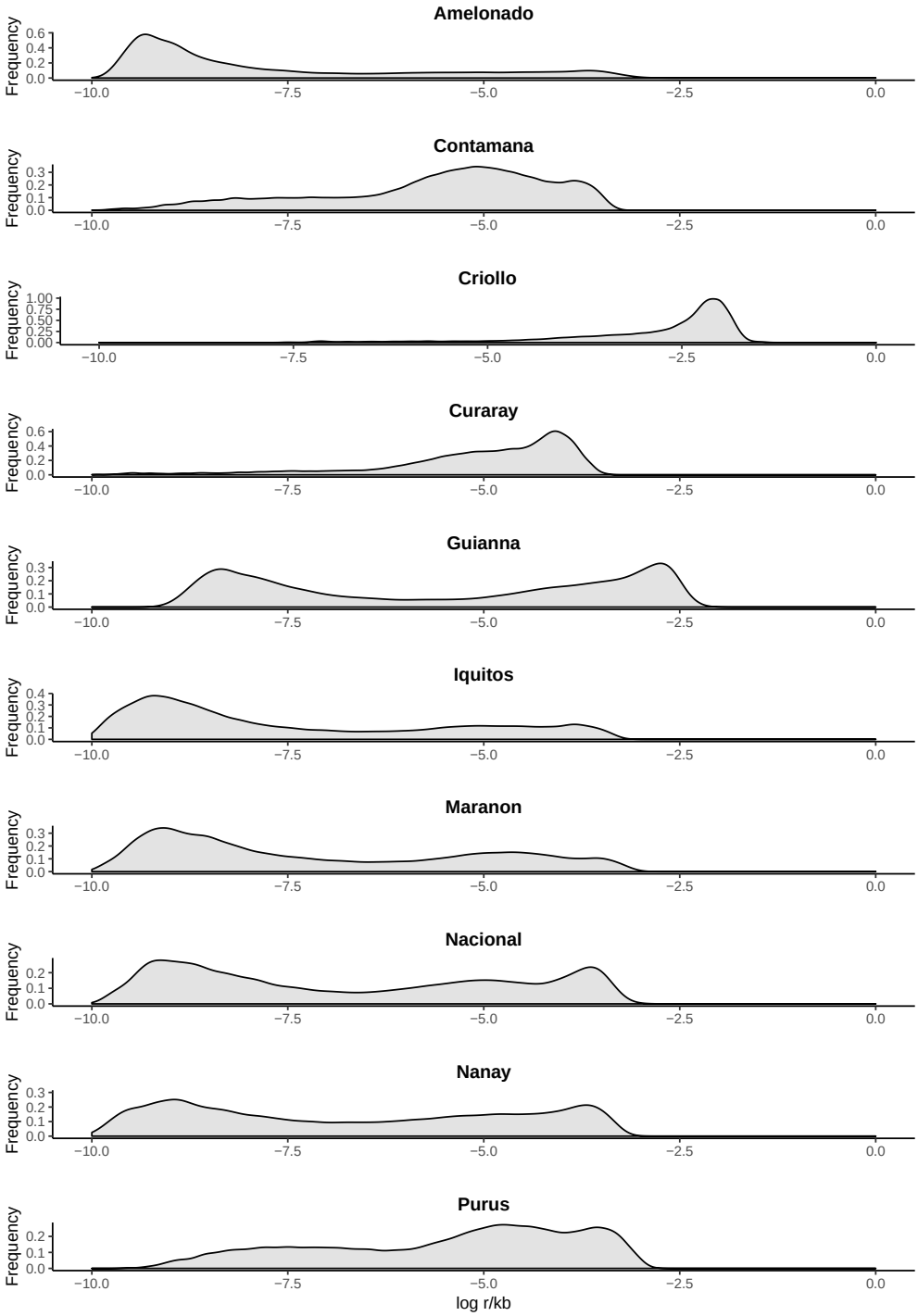

Figure 2: Distribution of log_{10} recombination rates ($\log(r/kb)$) along the genomes of the ten *T. cacao* populations.

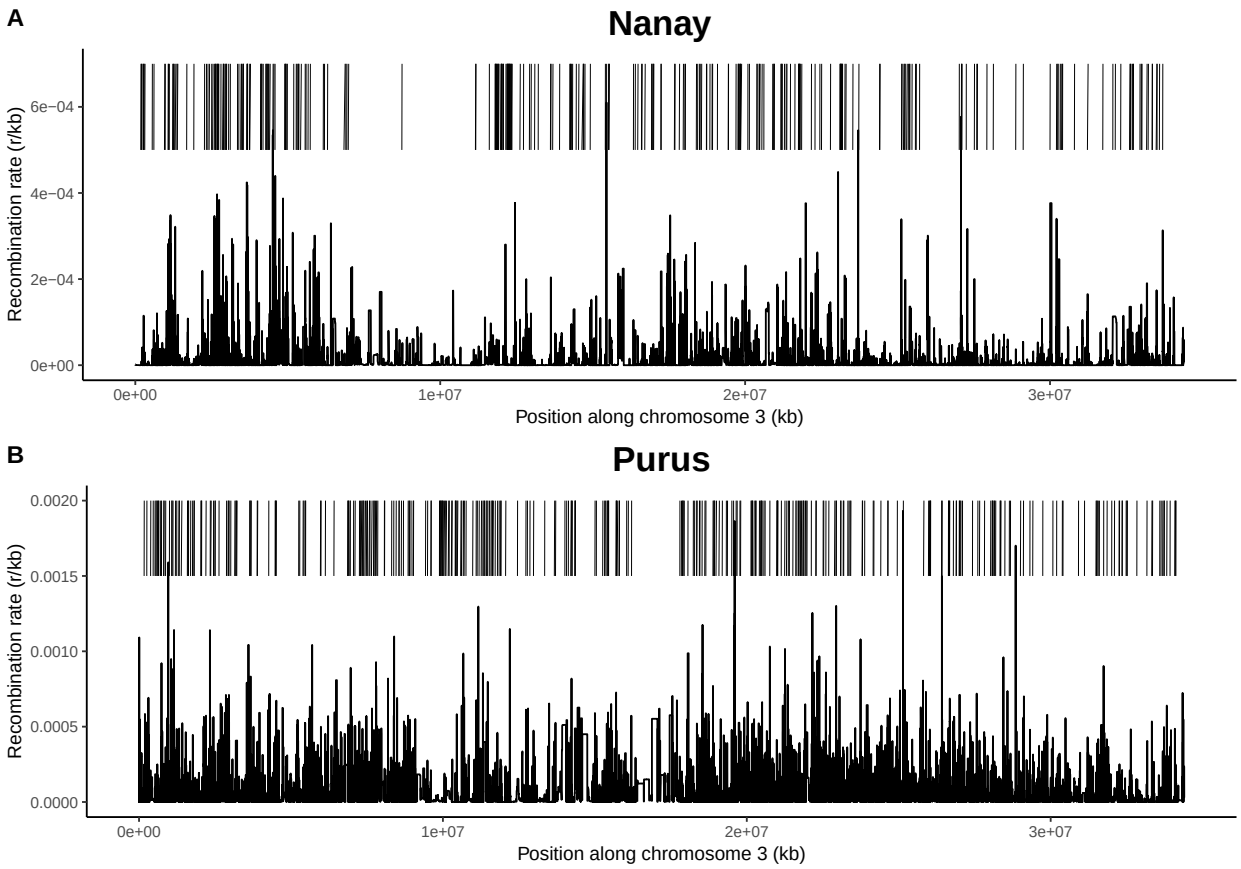

Figure 3: The third chromosomes of the Nanay (A) and Purus (B) populations were selected to exemplify the differences between populations in recombination rates (r/kb) and recombination hotspot locations (vertical bars above rates).

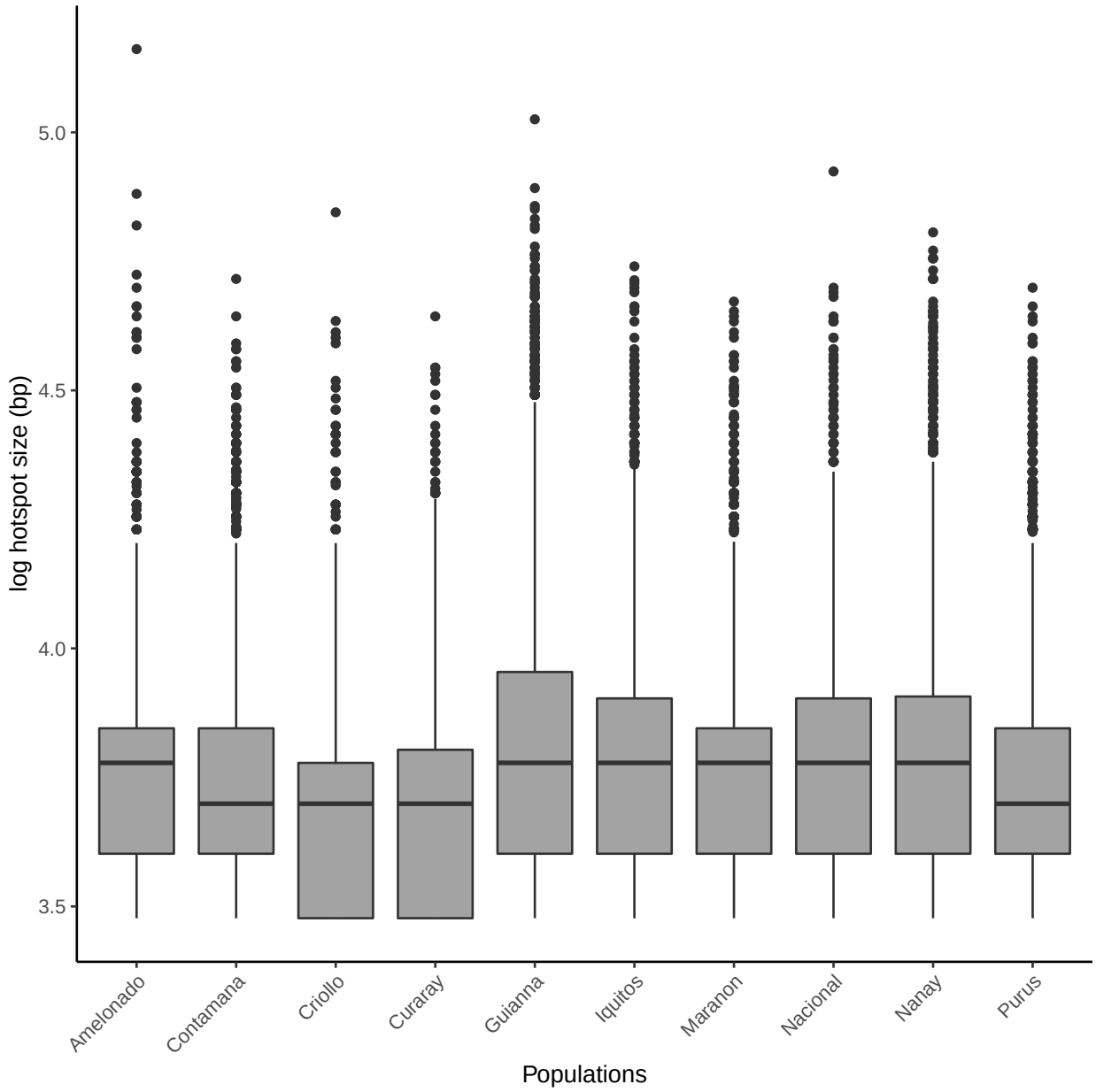

Figure 4: Boxplots of recombination hotspot sizes ($log_{10}(bp)$) by population. The horizontal line in the box represents the median value, while the points represent potential outliers.

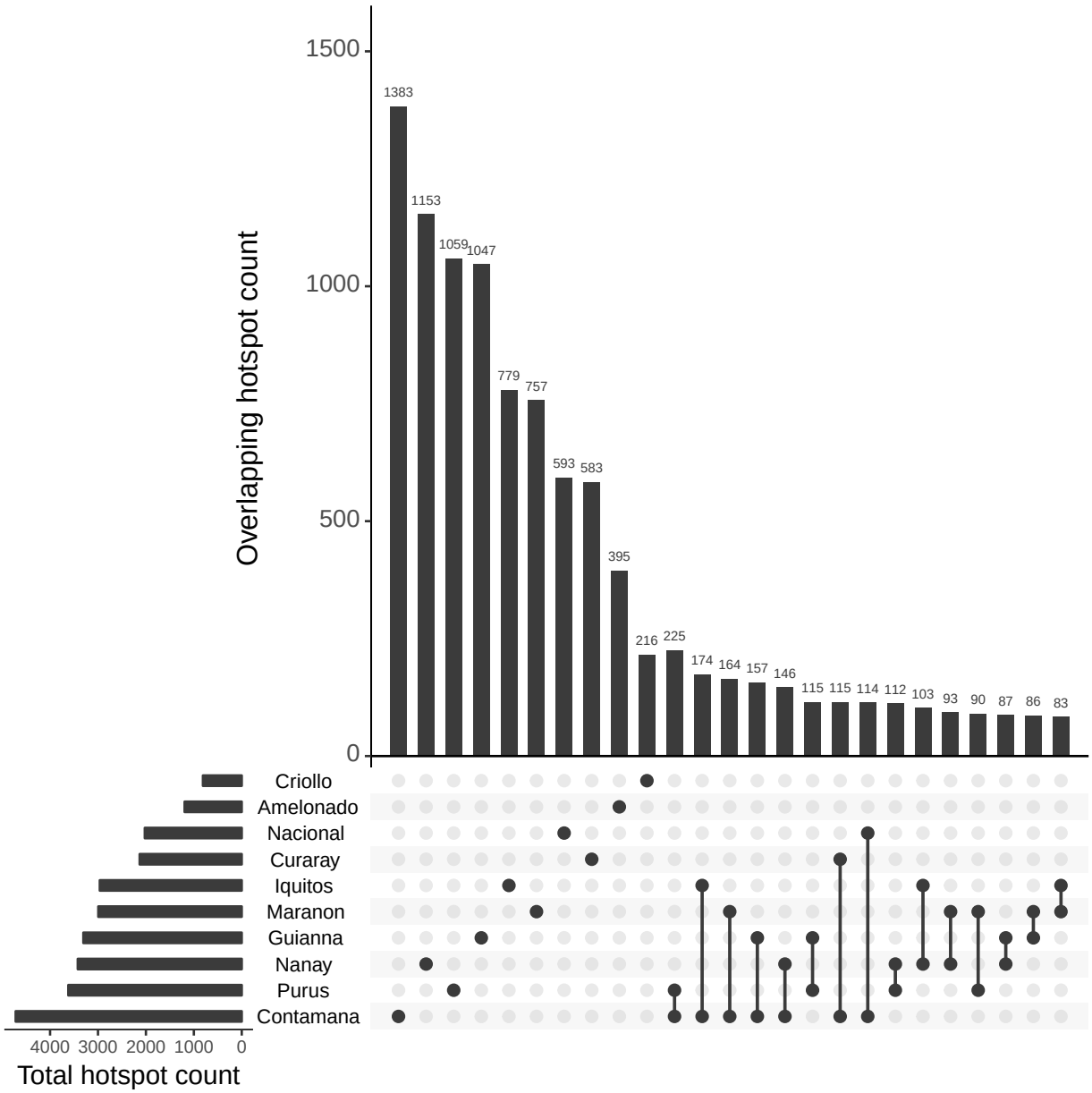

Figure 5: Upset plot showing number of hotspots in different subsets. Horizontal bars represent total hotspots detected in a population, each dot on the matrix indicate that the vertical bar above it is the count of hotspots unique to that population, connected dots indicate that the vertical bar above them represents hotspots shared between the populations represented by the connected dots. The 25 largest subsets are shown.

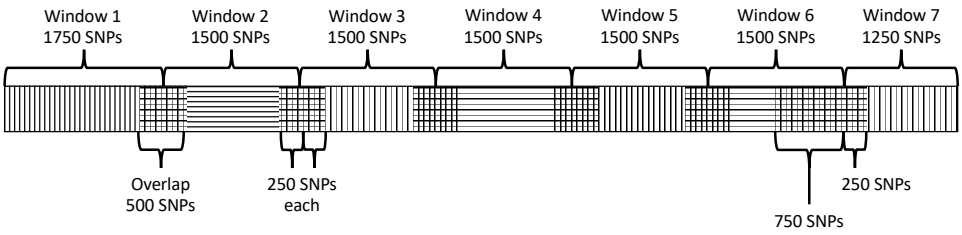

Figure 6: Example of the window layout for a 10,750 SNP chromosome. The 2,000 SNP long windows are represented by alternating horizontal and vertical lines and the overlaps between them are represented by square crosshatches. Braces above the chromosome indicate the regions from which recombination rates are extracted to generate the chromosome-wide recombination rates.

References

Akhunov, E. D., and Coauthors, 2003: The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome research*, **13 (5)**.

Anderson, L. K., N. Salameh, H. W. Bass, L. C. Harper, W. Z. Cande, G. Weber, and S. M. Stack, 2004: Integrating genetic linkage maps with pachytene chromosome structure in maize. *Genetics*, **166 (4)**, 1923–1933, doi:10.1534/genetics.166.4.1923, URL http://www.genetics.org/content/166/4/1923, http://www.genetics.org/content/166/4/1923.full.pdf.

Auton, A., and G. McVean, 2007: Recombination rate estimation in the presence of hotspots. *Genome Research*, **17 (8)**, 1219–1227, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933511/.

Auton, A., S. Myers, and G. McVean, 2014: Identifying recombination hotspots using population genetic data. *arXiv preprint arXiv:1403.4264*.

Auton, A., and Coauthors, 2012: A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*, **336 (6078)**, 193–198, doi:10.1126/science.1216872, URL http://science.sciencemag.org/content/336/6078/193, http://science.sciencemag.org/content/336/6078/193.full.pdf.

Auton, A., and Coauthors, 2013: Genetic recombination is targeted towards gene promoter regions in dogs. *PLOS Genetics*, **9 (12)**, 1–9, doi:10.1371/journal.pgen.1003984, URL https://doi.org/10.1371/journal.pgen.1003984.

Bartley, B., 2005: *The Genetic Diversity of Cacao and Its Utilization*. CAB books, CABI, URL https://books.google.com/books?id=_I40iGVJD64C.

661 Begun, D. J., and C. F. Aquadro, 1992: Levels of naturally occurring dna polymorphism
662 correlate with recombination rates in d. melanogaster. *Nature*, **356 (6369)**.

663 Booker, T. R., R. W. Ness, and P. D. Keightley, 2017: The recombination landscape in
664 wild house mice inferred using population genomic data. *Genetics*, **207 (1)**, 297–309,
665 doi:10.1534/genetics.117.300063, URL http://www.genetics.org/content/207/1/297, http:
666 //www.genetics.org/content/207/1/297.full.pdf.

667 Bougeard, S., and S. Dray, 2018: Supervised multiblock analysis in R with the ade4 package.
668 *Journal of Statistical Software*, **86 (1)**, 1–17, doi:10.18637/jss.v086.i01.

669 Branca, A., and Coauthors, 2011: Whole-genome nucleotide diversity, recombination, and
670 linkage disequilibrium in the model legume medicago truncatula. *Proceedings of the Na-*
671 *tional Academy of Sciences*, **108 (42)**.

672 Brick Kevin, Smagulova Fatima, Khil Pavel, Camerini-Otero R. Daniel, and Petukhova
673 Galina V., 2012: Genetic recombination is directed away from functional ge-
674 nomic elements in mice. *Nature*, **485 (7400)**, 642–645, doi:http://dx.doi.org/
675 10.1038/nature11089, URL http://www.nature.com/nature/journal/v485/n7400/abs/
676 nature11089.html#supplementary-information, 10.1038/nature11089.

677 Browning, S. R., and B. L. Browning, 2007: Rapid and accurate haplotype phasing and
678 missing-data inference for whole-genome association studies by use of localized haplo-
679 type clustering. *The American Journal of Human Genetics*, **81 (5)**, 1084 – 1097, doi:
680 https://doi.org/10.1086/521987, URL http://www.sciencedirect.com/science/article/pii/
681 S0002929707638828.

682 Brunschwig, H., L. Levi, E. Ben-David, R. W. Williams, B. Yakir, and S. Shifman, 2012:
683 Fine-scale Map of Recombination Rates and Hotspots in the Mouse Genome. *Genet-*
684 *ics*, doi:10.1534/genetics.112.141036, URL http://www.genetics.org/content/early/2012/

685 05/04/genetics.112.141036, http://www.genetics.org/content/early/2012/05/04/genetics.
686 112.141036.full.pdf.

687 Chen, M.-M., F. Feng, X. Sui, M.-H. Li, D. Zhao, and S. Han, 2010: Construction of
688 a framework map for pinus koraiensis sieb. et zucc. using srap, ssr and issr markers.
689 *Trees*, **24 (4)**, 685–693, doi:10.1007/s00468-010-0438-5, URL https://doi.org/10.1007/
690 s00468-010-0438-5.

691 Chessel, D., A.-B. Dufour, and J. Thioulouse, 2004: The ade4 package – I: One-table meth-
692 ods. *R News*, **4 (1)**, 5–10, URL https://cran.r-project.org/doc/Rnews/.

693 Choi, K., and Coauthors, 2013: *Arabidopsis* meiotic crossover hot spots overlap with
694 h2a.z nucleosomes at gene promoters. *Nature Genetics*, **45 (11)**, 1327–1336, doi:
695 10.1038/ng.2766.

696 Cornejo, O. E., and Coauthors, 2018: Population genomic analyses of the chocolate tree,
697 Theobroma cacao L., provide insights into its domestication process. *Communications*
698 *Biology*, doi:10.1038/s42003-018-0168-6.

699 Crow, J. F., and M. Kimura, 1970: (Harper international editions.). An introduction to
700 population genetics theory. Harper & Row, New York.

701 Danecek Petr, and Coauthors, 2011: The variant call format and VCFtools. *Bioinformatics*,
702 **27 (15)**, 2156–2158, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/.

703 Dapper, A. L., and B. A. Payseur, 2017: Effects of Demographic History on the Detection of
704 Recombination Hotspots from Linkage Disequilibrium. *Molecular Biology and Evolution*,
705 **35 (2)**, 335–353, doi:10.1093/molbev/msx272, URL https://doi.org/10.1093/molbev/
706 msx272, http://oup.prod.sis.lan/mbe/article-pdf/35/2/335/24367711/msx272.pdf.

Delaneau, O., J. Marchini, and J.-F. Zagury, 2011: A linear complexity phasing method for thousands of genomes. *Nature Methods*, **9 (2)**.

Donnelly, P., and T. G. Kurtz, 1999: Genealogical processes for fleming-viot models with selection and recombination. *Ann. Appl. Probab.*, **9 (4)**, 1091–1148, doi:10.1214/aoap/1029962866, URL https://doi.org/10.1214/aoap/1029962866.

Dray, S., and A.-B. Dufour, 2007: The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22 (4)**, 1–20, doi:10.18637/jss.v022.i04.

Dray, S., A.-B. Dufour, and D. Chessel, 2007: The ade4 package – II: Two-table and K-table methods. *R News*, **7 (2)**, 47–52, URL https://cran.r-project.org/doc/Rnews/.

Drost, H.-G., 2018: *philentropy: Similarity and Distance Quantification Between Probability Functions*. URL https://CRAN.R-project.org/package=philentropy, r package version 0.2.0.

Exposito-Alonso, M., and Coauthors, 2018: The rate and potential relevance of new mutations in a colonizing plant lineage. *PLOS Genetics*, **14 (2)**, 1–21, doi:10.1371/journal.pgen.1007155, URL https://doi.org/10.1371/journal.pgen.1007155.

Eyre-Walker, A., and P. D. Keightley, 2007: The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, **8 (8)**.

Felsenstein, J., 1974: The evolutionary advantage of recombination. *Genetics*, **78 (2)**, 737–756, URL http://www.genetics.org/content/78/2/737, http://www.genetics.org/content/78/2/737.full.pdf.

Fernandes, J. B., M. Séguéla-Arnaud, C. Larchevêque, A. H. Lloyd, and R. Mercier, 2018: Unleashing meiotic crossovers in hybrid plants. *Proceedings of the National Academy of Sci-*

ences, **115 (10)**, 2431–2436, doi:10.1073/pnas.1713078114, URL http://www.pnas.org/content/115/10/2431, http://www.pnas.org/content/115/10/2431.full.pdf.

Goolsby, E. W., J. Bruggeman, and C. Ane, 2016: *Rphylopars: Phylogenetic Comparative Tools for Missing Data and Within-Species Variation*. URL https://CRAN.R-project.org/package=Rphylopars, r package version 0.2.9.

Gore, M. A., and Coauthors, 2009: A first-generation haplotype map of maize. *Science*, **326 (5956)**, 1115–1117, doi:10.1126/science.1177837, URL http://science.sciencemag.org/content/326/5956/1115, http://science.sciencemag.org/content/326/5956/1115.full.pdf.

Haldane, J. B. S., 1937: The effect of variation on fitness. *The American Naturalist*, **71 (735)**, 337–349, URL http://www.jstor.org/stable/2457289.

Hellsten, U., and Coauthors, 2013: Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *PNAS*, **110 (48)**, 19 478–19 482, doi:10.1073/pnas.1319032110.

Henderson, J. S., R. A. Joyce, G. R. Hall, W. J. Hurst, and P. E. McGovern, 2007: Chemical and archaeological evidence for the earliest cacao beverages. *Proceedings of the National Academy of Sciences*, **104 (48)**, 18 937–18 940, doi:10.1073/pnas.0708815104, URL http://www.pnas.org/content/104/48/18937.abstract, http://www.pnas.org/content/104/48/18937.full.pdf.

Hinch, A. G., and Coauthors, 2011: The landscape of recombination in african americans. *Nature*, **476 (7359)**.

Hudson, R. R., and N. L. Kaplan, 1988: The coalescent process in models with selection and recombination. *Genetics*, **120 (3)**, 831–840, URL http://www.genetics.org/content/120/3/831, http://www.genetics.org/content/120/3/831.full.pdf.

Kim, S., and Coauthors, 2007: Recombination and linkage disequilibrium in arabidopsis thaliana. *Nature Genetics*, **39 (9)**.

Kosambi, D. D., 1943: The estimation of map distances from recombination values. *Annals of Eugenics*, **12 (1)**, 172–175, doi:10.1111/j.1469-1809.1943.tb02321.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1943.tb02321.x, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1943.tb02321.x.

Kundu, A., A. Chakraborty, N. A. Mandal, D. Das, P. G. Karmakar, N. K. Singh, and D. Sarkar, 2015: A restriction-site-associated dna (rad) linkage map, comparative genomics and identification of qtl for histological fibre content coincident with those for retted bast fibre yield and its major components in jute (corchorus olitorius l., malvaceae s. l.). *Molecular Breeding*, **35 (1)**, 19, doi:10.1007/s11032-015-0249-x, URL https://doi.org/10.1007/s11032-015-0249-x.

Lê, S., J. Josse, and F. Husson, 2008: FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, **25 (1)**, 1–18, doi:10.18637/jss.v025.i01.

Li, W. H., and M. Nei, 1974: Stable linkage disequilibrium without epistasis in subdivided populations. *Theoretical population biology.*, **6 (2)**, 173–183.

Lynch, M., 2010: Evolution of the mutation rate. *Trends in genetics : TIG*, **26 (8)**, 345–352, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2910838/.

Mackiewicz, D., A. J. Lustig, P. M. C. de Oliveira, S. Moss de Oliveira, and S. Cebrat, 2013: Distribution of recombination hotspots in the human genome – a comparison of computer simulations with real data. *PLoS ONE*, **8 (6)**.

Mackiewicz, D., M. Zawierta, W. Waga, and S. Cebrat, 2010: Genome analyses and modelling the relationships between coding density, recombination rate and chromosome length. *Journal of Theoretical Biology*, **267 (2)**, 186 – 192, doi:https://doi.

org/10.1016/j.jtbi.2010.08.022, URL http://www.sciencedirect.com/science/article/pii/S0022519310004315.

Martinez-perez, E., M. Schvarzstein, C. Barroso, J. M. Lightfoot, A. F. Dernburg, and A. M. Villeneuve, 2008: Crossovers trigger a remodeling of meiotic chromosome axis composition that is linked to two-step loss of sister chromatid cohesion. *Genes & development*, **22 20**, 2886–901.

McVean, G., 2010: What drives recombination hotspots to repeat dna in humans? *Philosophical Transactions of the Royal Society B*, **365 (1544)**, 1213–1218.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, 2004: The fine-scale structure of recombination rate variation in the human genome. *Science*, **304 (5670)**, 581–584, doi:10.1126/science.1092500, URL http://science.sciencemag.org/content/304/5670/581, http://science.sciencemag.org/content/304/5670/581.full.pdf.

Mézard, C., 2006: Meiotic recombination hotspots in plants. *Biochemical Society Transactions*, **34 (4)**, 531–534, doi:10.1042/BST0340531, URL http://www.biochemsoctrans.org/content/34/4/531, http://www.biochemsoctrans.org/content/34/4/531.full.pdf.

Motamayor, J. C., P. Lachenaud, da Silva e Mota Jay Wallace, R. Loor, D. N. Kuhn, S. J. Brown, and R. J. Schnell, 2008: Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma cacao L). *PLOS ONE*, **3 (10)**, e3311, doi:https://doi.org/10.1371/journal.pone.0003311.

Motamayor, J. C., and Coauthors, 2013: The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology*, **14 (6)**, r53, doi:10.1186/gb-2013-14-6-r53, URL https://doi.org/10.1186/gb-2013-14-6-r53.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005: A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science*, **310 (5746)**, 321–324, doi:10.1126/science.1117196, URL http://science.sciencemag.org/content/310/5746/321, http://science.sciencemag.org/content/310/5746/321.full.pdf.

Ohta, T., 1982: Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences*, **79 (6)**, 1940–1944, doi:10.1073/pnas.79.6.1940, URL http://www.pnas.org/content/79/6/1940, http://www.pnas.org/content/79/6/1940.full.pdf.

Paape, T., P. Zhou, A. Branca, R. Briskine, N. Young, and P. Tiffin, 2012: Fine-Scale Population Recombination Rates, Hotspots, and Correlates of Recombination in the Medicago truncatula Genome. *Genome Biology and Evolution*, **4 (5)**, 726–737, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3381680/.

Pickrell, J. K., and J. K. Pritchard, 2012: Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genetics*, **8 (11)**, 1–17, doi:10.1371/journal.pgen.1002967, URL https://doi.org/10.1371/journal.pgen.1002967.

Ptak, S. E., and Coauthors, 2005: Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics*, **37 (4)**.

Quinlan, A. R., and I. M. Hall, 2010: Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26 (6)**, 841–842, doi:10.1093/bioinformatics/btq033, URL +http://dx.doi.org/10.1093/bioinformatics/btq033, /oup/backfile/content_public/journal/bioinformatics/26/6/10.1093_bioinformatics_btq033/3/btq033.pdf.

R Core Team, 2018: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL https://www.R-project.org/.

Revelle, W., 2018: *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois, Northwestern University, URL https://CRAN.R-project.org/package=psych, r package version 1.8.10.

Rizzon, C., G. Marais, M. Gouy, and C. Biémont, 2002: Recombination rate and the distribution of transposable elements in the drosophila melanogaster genome. *Genome Research*, **12 (3)**, 400–407, doi:10.1101/gr.210802, URL http://genome.cshlp.org/content/12/3/400.abstract, http://genome.cshlp.org/content/12/3/400.full.pdf+html.

Rodgers, K., and M. McVey, 2015: Error-prone repair of dna double-strand breaks. *Journal of Cellular Physiology*, **231**.

Ross-Ibarra, J., 2004: The Evolution of Recombination under Domestication: A Test of Two Hypotheses. *The American Naturalist*, **163 (1)**, 105–112, doi:10.1086/380606, URL https://doi.org/10.1086/380606, pMID: 14767840, https://doi.org/10.1086/380606.

Sanjuán, R., A. Moya, and S. F. Elena, 2004: The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus. *Proceedings of the National Academy of Sciences*, **101 (22)**, 8396–8401, doi:10.1073/pnas.0400146101, URL http://www.pnas.org/content/101/22/8396, http://www.pnas.org/content/101/22/8396.full.pdf.

Schnable, P. S., and Coauthors, 2009: The b73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, **326 (5956)**.

Shanfelter, A., S. L. Archambeault, and M. A. White, 2018: Fine-scale recombination landscapes between a freshwater and marine population of threespine stickleback fish. *bioRxiv*, doi:10.1101/430249, URL https://www.biorxiv.org/content/early/2018/09/29/430249, https://www.biorxiv.org/content/early/2018/09/29/430249.full.pdf.

Singhal, S., and Coauthors, 2015: Stable recombination hotspots in birds. *Science*,

350 (6263), 928–932, doi:10.1126/science.aad0843, URL http://science.sciencemag.org/content/350/6263/928, http://science.sciencemag.org/content/350/6263/928.full.pdf.

Siol, M., I. Bonnin, I. Oliveri, J. M. Prosperi, and J. Ronfort, 2007: Effective population size associated with self-fertilization: lessons from temporal changes in allele frequencies in the selfing annual medicago truncatula. *Journal of Evolutionary Biology*, **20 (6)**, 2349–2360, doi:10.1111/j.1420-9101.2007.01409.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1420-9101.2007.01409.x, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1420-9101.2007.01409.x.

Smith, A., R. Hubley, and P. Green, 2016: Repeatmasker open-4.0.(2013-2015).

Stapley, J., P. G. D. Feulner, S. E. Johnston, A. W. Santure, and C. M. Smadja, 2017: Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372 (1736)**, doi:10.1098/rstb.2016.0455, URL http://rstb.royalsocietypublishing.org/content/372/1736/20160455, http://rstb.royalsocietypublishing.org/content/372/1736/20160455.full.pdf.

Stevison, L. S., and Coauthors, 2016: The time scale of recombination rate evolution in great apes. *Molecular Biology and Evolution*, **33 (4)**, 928–945, doi:10.1093/molbev/msv331.

Weir, B. S., and C. C. Cockerham, 1984: Estimating f-statistics for the analysis of population structure. *Evolution*, **38 (6)**, 1358–1370, URL http://www.jstor.org/stable/2408641.

Winckler, W., and Coauthors, 2005: Comparison of fine-scale recombination rates in humans and chimpanzees. *Science (New York, N.Y.)*, **308 (5718)**, 107–111, URL http://search.proquest.com/docview/67570665/.

Wloch, D. M., K. Szafraniec, R. H. Borts, and R. Korona, 2001: Direct estimate of the mutation rate and the distribution of fitness effects in the yeast saccharomyces cerevisiae.

Genetics, **159 (2)**, 441–452, URL http://www.genetics.org/content/159/2/441, http://www.genetics.org/content/159/2/441.full.pdf.

Wu, J., and Coauthors, 2003: Physical maps and recombination frequency of six rice chromosomes. *Plant Journal*, **36 (5)**, 720–730.

Yelina, N., P. Diaz, C. Lambing, and I. R. Henderson, 2015: Epigenetic control of meiotic recombination in plants. *Science China Life Sciences*, **58 (3)**, 223–231, doi: 10.1007/s11427-015-4811-x, URL https://doi.org/10.1007/s11427-015-4811-x.