Ancestry informative alleles captured with reduced representation library sequencing in *Theobroma cacao*

Jaime A. Osorio-Guarín[1], Corey R. Quackenbush[2,3], Omar E. Cornejo[2]

[1]Centro de Investigación Tibaitatá, Corporación Colombiana de Investigación Agropecuaria – Agrosavia, Mosquera, Cundinamarca, Colombia.

[2]School of Biological Sciences, Washington State University, Pullman, WA, United States.

[3]Division of Molecular and Translational Sciences, U. S. Army Medical Research Institute of Infectious Diseases (USAMRIID), Fort Detrick, Md, United States.

* Corresponding author: Omar E. Cornejo

E-mail: omar.cornejo@wsu.edu (OC)

Abstract

As the source of chocolate, cacao has become one of the most important crops in the world. The identification of molecular markers to understand the demographic history, genetic diversity and population structure plays a pivotal role in cacao breeding programs. Here, we report the use of a modified genotyping-by-sequencing (GBS) approach for large-scale single nucleotide polymorphism (SNP) discovery and allele ancestry mapping. We identified 12,357 bi-allelic SNPs after filtering, of which, 7,009 variants were ancestry informative. The GBS approach proved to be rapid, cost-effective, and highly informative for ancestry assignment in this species.

Introduction

Theobroma cacao L, the tree from which chocolate is confectioned, originated in the Amazonian basin [1] and has become a globally cultivated crop in recent times [2]. There are more than 20 species within the *Theobroma* genus, which belongs to the Malvaceae family [3]. *T. cacao* is the only species that is cultivated extensively because it can be manufactured into chocolate, liquors, confections, cosmetics, and animal feed [4,5]. Chemical and archaeological evidence shows that human settlements have been consuming cacao since 1,400 BC [6,7]. Since 900 BC *T. cacao* has been widely cultivated in Mesoamerica, which is considered the center of domestication. Recent genetic work has supported these findings suggesting that domestication of cacao occurred 3.6 Kya [8]. The domestication of cacao has resulted in the development of an important crop that sustains the livelihood of communities throughout Central and South America, Asia, and Africa [1,9,10], and is the principal income

43 for about six million smallholder farmers globally that produce up to 90% of the world's
44 cocoa [11].
45
46 Cacao is a diploid species with a chromosome complement of 2n=20 [12]. Recently, genomic
47 resources have been developed to facilitate the study of the species evolutionary history, the
48 inference of the genetic basis of phenotypes of interest, and the development of marker
49 assisted breeding programs. There are two reference genomes that have been sequences from
50 different cultivars, the B97-61/B2 genome (Criollo), with an estimated size of 409 million
51 base pairs (Mbp) [13] and the Matina 1-6 genome (Amelonado), with an estimated genome
52 size of 445 Mbp [14]. In addition, Cornejo et al. 2017 [8], performed whole genome
53 sequencing on 200 cultivars in order to understand the domestication process in this species.
54 Recently, approaches to infer genetic diversity in *T. cacao* have been completed using a small
55 panel of single nucleotide polymorphisms (SNPs, ranging from 87 to 6,000) developed with
56 a Fluidigm array system [15–17]. Also, using the information generated from a subset of the
57 200 cultivars project, an Ilumina Infinity II genotyping array of 13,530 markers was recently
58 developed [18]. However, it is well known that genotyping strategies based on arrays are
59 biased towards the capture of known specific loci in reference populations which introduces
60 ascertainment bias in the genotyping of samples from different origins [19]. There is a
61 growing interest in developing strategies to rapidly genotype accessions at a reduced cost in
62 order to perform extensive association studies and better identify accessions with specific
63 markers that will accelerate assisted breeding programs.
64
65 The generation of genomic resources in cacao has benefited from advances in sequencing
66 chemistry, cost reductions, and increased computational capacity to analyze data [20]. Yet,

67 despite the general reduction of costs, the sequencing of complete cacao genomes remains a
68 costly approach for breeders and researchers. Due to the high cost of whole genome
69 sequencing, genotyping arrays are typically used. However, genotyping arrays suffer from
70 the aforementioned issues of ascertainment bias, which hinders the identification of novel
71 untapped genetic variation in the species. It has been shown that genotyping-by-sequencing
72 (GBS) is a cost effective and efficient methodology for the generation of a large number of
73 SNPs, while avoiding ascertainment biases [21,22]. The genetic information generated by
74 GBS has been shown to facilitate robust studies of genetic diversity, population structure,
75 phylogeny, and association in a large number of plant and animal species [23–29] but have
76 never been applied to the study of *T. cacao*. The original protocol proposed for the digestion
77 and generation of reduced representation genomic libraries used the *ApeKI* enzyme [21].
78 However, Cooke et al. (2016) [30], demonstrated that the use of enzymes that cut far from
79 the recognition site, such as *BsaXI*, increases the diversity of the library and therefore the
80 number and quality of reads sequenced. In the original protocol the restriction enzymes
81 would produce libraries with nearly identical initial sequence causing saturation of the image
82 capture during sequencing. To avoid this issue, the original protocol incorporated the barcode
83 in one of the adapter sequences upstream of the cut-site to introduce variation and minimize
84 the problems associated with saturation [31]. In the proposed protocol (taken from Cooke et
85 al. 2016 [30]), the chosen enzyme generates libraries in which the initial sequence is random
86 due to the fact that the enzyme cuts 14 base pairs (bp) away from its recognition site.
87
88 In this work, we adapt and modify an existing GBS method [30] to generate a large number
89 of ancestry informative SNPs in cacao. We further show that the data generated produced
90 genetic diversity estimates consistent with what has been shown in previous work. We also

91 provide an estimate of the proportion of novel versus common SNPs (based on the relative
92 frequency of SNPs) that can be generated using the proposed methodology. Due to the lower
93 cost of generating this type of data we propose that this strategy could be widely used to
94 perform genetic diversity analysis on thousands of samples, facilitating association mapping
95 studies aimed at identifying the genetic basis of important agronomic traits in cacao such as
96 disease resistance and bean quality.
97

98 # Methods

99 ## Plant materials

100 A total of 30 accessions from the germplasm bank from Cocoa Research Centre at the
101 University of West Indies in Trinidad **(S1 Table)** were included in the analysis. Young leaves
102 from one individual per accession were sampled in Trinidad and sent to the Cornejo Lab at
103 Washington State University. Small strips (100 mg) were cut with sterile scissors (washed
104 with 1% hydrochloric acid and ethanol prior to and in between sample processing), placed in
105 1.5 ml Eppendorf tubes and stored at -80 ºC for DNA extraction. Five samples (GU 175/P,
106 ICS 1, IMC 67, M 8[SUR], SCA 6) from the 200 genomes project were genotyped in this
107 study in order to compare the concordance between SNPs identified using GBS to those
108 identified using full genome sequencing [8].
109

110 ## DNA extraction and GBS library preparation

111 Total DNA was isolated from 100 mg of young leaves collected from each accession using
112 a DNeasy Plant Mini Kit (QIAGEN, Germany) according to manufacturer's instructions. The

113 final elution volume was adjusted to 60 µl with 1X TE solution buffer. Integrity of the total
114 DNA was checked on a 0.7% agarose gel. DNA concentration was measured using a Qubit
115 Fluorometer (Life Technologies, Thermo Fisher Scientific Inc., Waltham, MA, USA) with a
116 dsDNA HS Assay Kit (Life Technologies). All samples were normalized to the same volume
117 (50 µl) containing 500 ng of DNA. For each sample, genomic DNA was digested with 2.4
118 units of *BsaXI* (New England Biolabs, NEB, R0609S) at 37 ºC overnight in a 100 µl reaction
119 buffered with Cut Smart NEBuffer (NEB, B7204S). After digestion, we performed a 0.7X
120 size selection using AMPure XP beads to remove any large undigested fragments. The DNA
121 was then prepared for sequencing using an NEBNext Ultra DNA Library Prep Kit for
122 Illumina (NEB E7370S; version 1.2) with the following modifications. Adapter ligation was
123 followed with a bead clean up without size selection. Library enrichment was accomplished
124 using 10 cycles of PCR amplification. Finally, we evaluated the distribution of fragment
125 lengths for each library using an Agilent 2100 BioAnalyzer. Samples were sequenced on an
126 Illumina HiSeq 2500 (Illumina Inc. San Diego, CA), with paired-end 100 bp reads.
127

128 SNP discovery and data processing

129 Raw sequence reads were demultiplexed and converted to Fastq files by the Spokane
130 Sequencing Core at Washington State University using Casava v1.8.2 (Illumina Inc.). SNP
131 quality control and trimming were conducted using FastQC [32] and Trim Galore v0.5.0 [33],
132 respectively. Reads were trimmed to remove 3' bases with a quality score less than 25. Reads
133 in which complete or partial adapter sequences were identified were trimmed accordingly.
134 Reads less than 60 bp in length after trimming were removed. The trimmed sequences were
135 aligned to the Matina reference genome using BWA v0.17.0 [34] with the default parameters

136 (except for mismatch penalty -B 6). We performed quality assessment of the BAM files,
137 validation read groups, removal of PCR duplicates, and correction of mapping quality
138 assignment with Picard v2.18.9 [35]. The genome analysis toolkit (GATK) v3.8.1 [36] was
139 used to perform local realignment and base quality recalibration. We performed SNP calling
140 using the unified genotyper module in GATK to facilitate comparison with previous work.
141 Hard filters were applied to the data (quality by depth higher than 2, strand bias FS 60, Root
142 mean square of mapping quality 40.0) and only SNPs that passed the filters were retained.
143
144 For population structure analysis we used VCFtools v0.1.13 [37] to performed additional
145 variant filtering in which only bi-allelic variants with a minimum minor allele frequency
146 (MAF) greater than 0.05 were retained. Coverage statistics were obtained by analyzing these
147 loci with Plink-Seq v0.10 [38]. To analyze the types of mutations we used the program
148 SnpEff with the genome built in database [39]. The R scripts are provided in the GitHub of
149 the project.
150

151 *In silico* digest

152 *In silico* digestion of the cacao chromosomes were performed with the *T. cacao* cv. Matina
153 reference genome [14]. To identify all restriction cut site positions for *BsaXI*, the program
154 RESTRICT from the Emboss package v6.5.7.0 was used [40]. Digested fragments were
155 ordered per chromosome and then were summed using two size ranges (200-700 bp and 200-
156 1000 bp). The desired range for the libraries to be sequenced on an Illumina HiSeq is between
157 200-700 bp, however fragments of up to 1000 bp can be sequenced [41]. We report the

158 distribution of the fragments as a histogram and the location of the *in silico* digested
159 fragments along the genome.
160
161 To estimate the number of fragments produced from the *BsaXI* GBS experiment and to
162 compare them with the *in silico* predictions, the alignment files were analyzed using
163 INTERSECT included in BEDtools v2.27.0 [42]. Using the output files from this command
164 we estimated the number of *BsaXI* fragments for all chromosomes and compared the results
165 with those from the *in silico* digestion.
166

167 Genetic diversity

168 Observed heterozygosity (Ho), expected heterozygosity (He), and polymorphism
169 information content (PIC) were calculated with Power Marker v1.25 [43]. Tajima's D was
170 computed with VCFtools v0.1.13. We also estimated allele frequencies using VCFtools for
171 each of the sites that met the following criteria: i) bi-allelic and ii) no missing data. For each
172 site we used folded site frequency spectrum (SFS) for the cacao GBS dataset, which describes
173 the genetic variation as a distribution of allele frequencies across the genome. The site
174 frequency spectrum was generated in R.
175

176 Admixture and population genetic structure analysis

177 The population analysis was carried out on the filtered set of SNPs. We analyzed the ancestry
178 and admixture patterns for the 30 unrelated accessions using cacao samples recently
179 sequenced [8]. The intersected dataset of MAF filtered SNPs between our cacao GBS dataset
180 (test set) and the 79 individuals from the whole sequence dataset (reference set) consisted of

181 5,225 SNPs. We ran ADMIXTURE v1.3 [44] in supervised mode assuming ten populations,
182 following Motamayor et al. 2008 [45] and Cornejo et al [8]. Principal component analysis
183 (PCA) was performed with the algorithm implemented in EIGENSOFT [46]. All figures
184 were generated using R version v3.5.1 [47]. The precision of the ancestry analysis was
185 evaluated by obtaining bias and standard error for each ancestry out of a bootstrap with 2000
186 pseudo-replicates in ADMIXTURE v1.3 [44].
187

188 # Results

189 ## *In silico* restriction enzyme digestion

190 To determine i) what the expected number of fragments to be sequenced was; and ii) to infer
191 how many of the sequenced fragments matched locations in the genome that were expected
192 from the digestion, we evaluated the distribution of fragments throughout the genome (**S1**
193 **Fig**). Those sequenced regions that did not match the locations identified from the *in silico*
194 digest were considered off target. The *in silico* digestion was done considering two sets: 1)
195 lengths ranging from 200 to 700 bp generated a total of 5,479,477 bp **(Fig 1A)** and 2) lengths
196 ranging from 200 to 1000 bp generated a total of 10,217,990 bp **(Fig 1B)**. These represent 1
197 and 3% of the genome, which suggests that *BsaXI* is an appropriate cutter. We used the larger
198 range to evaluate overlap of sequenced target and predicted target.
199

200 **Fig 1. *In silico* analysis of restriction enzyme sites in the *T. cacao* genome.** (**A)** Number
201 of fragments computed in the size range between 200 and 700 bp. **(B)** Number of fragments
202 computed in the size range between 200 and 1000 bp.
203

204 The comparison between the predicted fragment-size distribution with the fragments
205 produced through our assay **(Fig 2)** suggests high variability in the accuracy of sequenced
206 targets across samples. Despite the high variability, 73% of the samples had an overlap of
207 over 57% with the *in silico* digest, and 53% of samples had an overlap over 76% with the
208 expectation
209
210 **Fig 2. Assessment of sequencing on target per sample.** The orange bar corresponds to the
211 total number of fragments that are predicted to be digested *in silico* within the length
212 distribution of 200 to 700 bp. Blue bars correspond to the number of sequenced fragments
213 that meet our callability criteria overlapping with those predicted *in silico*.
214

215 SNP Discovery

216 We obtained a total of 366 million paired-end 100 bp reads. The number of raw sequence
217 reads per individual ranged from 0.4 million to 52.2 million. After removal of low-quality
218 sequences and adapter trimming, 341 million reads remained. Of these reads, 322 million
219 aligned to the Matina reference genome, and were used for SNP calling. The mean read depth
220 ranged from 8.9 to 160.1. The observed differences in read depth across samples is most
221 likely due to differences in quality and/or variations in barcode efficiencies. Individual data
222 information is shown in **Table 1**.
223
224 **Table 1.** Summary statistics of sequenced data per individual.

Accession Number	Reads raw-data (Millions)	Reads after trimming (Millions)	Mapping reads (and percent)	Depth
CL 10/11	0.53	0.50	0.49 (97.6%)	15.2

CL 10/3	4.8	4.7	4.5 (96.8%)	67.1
CLM 100	14.9	14.6	14.2 (97.4%)	131.2
GS 6	14.5	14.1	13.8 (97.8%)	122.9
GS 61	10.2	9.9	9.5 (96.4%)	104.2
GU 175/P_1	5.3	5.1	5.0 (96.7%)	73.6
GU 175/P_2	1.1	1.04	1.01 (96.6%)	9.0
ICS 1_1	3.2	3.1	3.0 (96.9%)	52.6
ICS 1_2	4.9	4.7	4.5 (95.6%)	86.7
ICS 28	1.9	1.88	1.82 (96.8%)	26.6
ICS 87	3.1	1.67	1.62 (96.9%)	21.6
IMC 6	7.9	5.3	5.1 (95.7%)	71.6
IMC 67_1	1.91	1.90	1.83 (96.4%)	26.6
IMC 67_2	0.4	0.36	0.35 (97.6%)	8.9
LCT EEN23	10.4	8.3	8.2 (98.2%)	126.6
LCT EEN246	8.2	7.8	7.6 (97.1%)	100.4
LCT EEN31	7.8	6.2	6.0 (96.3%)	74.8
LP 3/29[POU]	7.4	5.2	5.0 (96 %)	62.3
LZ 33	8,0	6.2	6.0 (97.4%)	66.3
M8 [SUR]_1	4.6	3.4	3.3 (96.6%)	52.1
M8 [SUR]_2	15.2	14.8	14.4 (97.4%)	125.9
NA 189	7.2	5.2	5.1 (96.9%)	72.4
NA 26	12,0	11.7	11.2 (95.9%)	128.5
NA 68	6.6	6.4	6.2 (96.7%)	86.6
RIM 113[MEX]	10.8	10.5	10.1 (96.1%)	39.9
SCA 6_1	2.6	2.5	2.4 (96.4%)	39.9
SCA 6_2	8.4	8.1	7.9 (96.6%)	101.2
SLC 18	13.9	13.1	12.7 (97.1%)	123.9
SLC 8	20.6	19.5	18.8 (96.6%)	141.9
UF 38	52.2	50.9	49.05 (96.2%)	160.1

A total of 12,357 SNPs were identified in the collection. The SNPs were classified into transitions (Ti) and transversions (Tv) based on nucleotide substitution. The number of A/G transitions was very similar to the number of C/T transitions. The numbers of A/C and A/T transversions were relatively higher than the C/G and similar to G/T transversions. Transitions are the most common type of nucleotide substitutions [48]. In

231 this dataset 62.7% of the base changes were transitions and 37.3% were transversions
232 **(Table 2)**.
233
234 **Table 2.** Summary of polymorphisms identified in cacao dataset.

Total number of polymorphisms		
SNPs	12,357	100%
Transition		
A/G	3,856	31.2%
C/T	3,894	31.5%
Transversion		
A/C	1,243	10.1%
A/T	1,403	11.3%
C/G	758	6.2%
G/T	1,203	9.7%
Ti/Tv ratio	1.79	

235

236 # Genetic diversity analysis

237 The dataset was reduced to 7,009 SNPs after excluding markers showing: 1) more than 20%
238 missing data, 2) a MAF ≤ 0.05, 3) multi-allelic sites, 4) insertions and deletions. Expected
239 heterozygosity (He) (mean 0.272), Ho (mean 0.200) and PIC (mean 0.226) values estimated
240 per position are listed in **S2 Table**. He, Ho and PIC ranged from 0.117 to 0.500, from 0 to
241 0.437, and from 0.374 to 0.375, respectively. Tajima's D was 0.458, indicating an excess of
242 intermediate frequency alleles that could result from demographic processes such as
243 population bottlenecks, population subdivision or migration. The SFS showed a significantly
244 high proportion of rare alleles (45.3% of identified alleles). These results quantitatively show
245 the proportion of rare variants that would be missed if genotyping arrays were employed for

246 detecting and characterizing genetic variation **(Fig 3)**. The SFS shows a variable pattern of
247 frequencies of alleles at intermediate minor allele counts, which is consistent with
248 unaccounted for population structure.
249
250 **Fig 3. The site frequency spectrum of *T. cacao* samples used in the study.** The orange
251 bars correspond to alleles found in one, two or three chromosomes in the sample, considered
252 to be low frequency in this study. The green bars correspond to alleles with a minor allele
253 count of 4 or more and are considered common variants for the purpose of this experiment.
254 Conservative estimates suggest that nearly 50% of the SNPs are rare and would be missed if
255 an array were used to genotype these accessions.
256

257 Admixture and population genetic structure analysis

258 The filtered set of SNPs was used for admixture and population structure analysis. The 30
259 accessions were compared with ten reference populations that are related to the cacao groups
260 previously defined [45]. Fifty percent of analyzed individuals were assigned to unique
261 ancestry in the Curaray, Nanay, Contamana, Amelonado, Nacional, Guianna and Iquitos
262 populations. The other half of the individuals presented a variety of mixed ancestries **(Fig 4)**.
263 The analysis of bias and error supports a high level of precision in the assignment of ancestry
264 for the majority of the samples (**S2 Fig A**). For the most part, bias in the assignment was
265 lower than 5%. However, LCTEEN23 presented a relatively higher bias (10%) in the
266 assignment to Nacional ancestry (**S2 Fig B**). Standard error (SE) is relatively low overall
267 with the majority being below 0.05. Sample LCTEEN23 presented a relatively higher SE in
268 the assignment of Nacional ancestry (**S2 Fig C**).

269 **Fig 4. Supervised ADMIXTURE using ten reference populations.** Each bar corresponds
270 to an individual and the height to the proportion of ancestry explained by one of the 10
271 previously described populations. The first uniformly colored 79 columns correspond to
272 individuals in the reference set. The remaining 30 columns show the proportions of ancestry
273 assignment for each one of the newly sequenced cacao samples with our GBS approach.
274
275 In addition, pairwise allele (SNP) sharing distances were computed between all samples and
276 reference populations and PCA was used to illustrate the resulting pairwise distances **(Fig 5)**.
277 Plotting the two first axes showed that they capture the previously described genetic groups
278 with distinct clusters reflecting the different origin in cacao today. The first principal
279 component (13.74% of the variation) showed a clear relationship between the samples and
280 the Criollo and Amelonado ancestral populations whereas the remaining populations
281 (Curaray, Nanay, Contamana, Maranon, Nacional, Gianna, Iquitos, and Purus) separated
282 along the second principal component (11.77% of the variation).
283
284 **Fig 5. Principal component analysis of reference and cacao GBS samples.** PCA plots of
285 the same cacao collection according to subgroups, as identified by ADMIXTURE software.
286 Although reference samples are closer together, the newly sequenced individuals are placed
287 close to other individuals showing similar ancestry. Admixed individuals in the new cacao
288 GBS set are more variable and overdispersed when compared to the reference set.
289

290 # Discussion

291 Marker assisted breeding in cacao currently faces a great challenge in the development of
292 alternative fast, cheap, and accurate genotyping strategies for a large number of accessions.
293 The need for inexpensive and accurate genotyping strategies is especially true when
294 considering that most cacao producers are located in developing countries with limited
295 resources. Some of the technologies developed for array genotyping, which have been
296 advanced by us and others, work well for the identification of known variants but are limited
297 in the ability to discover new and potentially informative variants. Genotyping-by-
298 sequencing is a high-throughput, low-cost technology, which is useful for variant
299 identification in diverse species and populations and allows for the discovery of previously
300 unidentified variants [21,49]. In this work, we have adapted and modified a GBS genotyping
301 assay for cacao. The most important changes to the library preparation are: 1) the use of the
302 restriction enzyme *BsaXI* that cut 14 bp away from the recognition motif [30], 2) the use of
303 the Illumina Y-adapters allowing for PCR incorporation of dual-indexed barcodes, which
304 facilitates large-scale, inexpensive multiplexing, and 3) the efficient cleaning of non-ligated
305 adapter excess and PCR duplicates using cleaning beads, which is a widely used technique
306 already employed in previous work [8].
307
308 There was variation across samples in the number of mapped reads that overlapped with the
309 fragments predicted from the *in silico* digestion. The performance of our genotyping in on-
310 target regions is similar to other studies [21], while we produced a large number of reads per
311 sample (0.5 M – 14.3 M reads), only a fraction overlapped with predicted regions for a
312 number of samples. The lack of overlap for some samples suggests that a large number of
313 fragments sequenced were off target. For example, GU 175.2 had a low percent of overlap
314 (5%), whereas UF 38 had 95% overlap with the *in silico* digest (**Fig 2**). The number of

315 sequenced fragments overlapping with *in silico* predicted regions is likely explained by
316 variations in DNA quality that result from fragmentation during DNA extraction, prior to the
317 digestion.
318
319 We identified a large number of variants, 12,357 SNPs, that were retained after filtering.
320 Estimates of genetic variability are comparable with those estimated in previous studies of
321 cacao [18]. Enhanced SNP discovery, SNP quality, production steps, and optimization of
322 parameters improved the SNP detection when compared to those obtained from 96 SNPs
323 from small Fluidigm arrays or the 6K SNP array previously reported [15–17]; or even from
324 the 15K SNP arrays considering the fact that a large proportion of SNPs cannot be identified
325 with array systems. According to Romay et al. 2013 [50], one important advantage of GBS
326 over the SNP array systems, is the MAF distribution. We are interested in the identification
327 of novel genetic variation that might prove to be useful in the identification of traits of interest
328 and much of this genetic variation may not be captured with genetic arrays. The allele
329 frequency distribution of variants identified using GBS followed an expected pattern with an
330 excess of rare variants **(Fig 3)**. More than 45% of bi-allelic variants were rare and the
331 distribution of intermediate variants revealed signatures of population structure.
332
333 The mean expected heterozygosity value was higher than the observed heterozygosity,
334 indicating a deficit of heterozygotes. Compared with recent published genetic diversity
335 studies in cacao [16,17,51], our results present slightly lower values of He, Ho and PIC,
336 probably due to the high presence of homozygous genotypes of the Criollo and Amelonado
337 populations and the impact of a well-known effect of allelic dropout in GBS datasets [30,52].

338 The pattern of population structure was further validated when ancestry and structure were
339 analyzed together with the reference panel set. To study the ancestry of the newly sequenced
340 samples 5,225 bi-allelic SNPs were successfully intersected with a reference panel, which
341 consisted of 79 fully sequenced genomes. The samples used for the reference set were
342 recently sequenced [8] and represent the 10 different genetic groups from the Upper Amazon,
343 Lower Amazon, Orinoco, and Guyana [45]. Our ancestry characterization of the newly
344 sequenced 30 samples was consistent with the known passport data for the accessions **(S1**
345 **Table)**. More specifically, accessions NA 26 and NA 189 were successfully assigned to the
346 Nanay group and both SCA 6 accessions were correctly assigned to Contamana, showing
347 that the procedure allows us to correctly identify their population of provenance. In the case
348 of admixed individuals, we could correctly assign ancestry to individuals with known
349 admixture patterns. For example, for ICS or UF individuals which are known to be a mix
350 between Criollo and Amelonado. These results make us confident that the generated data and
351 procedure could be used to disentangle more complex patterns of genetic admixture like those
352 obtained for previously uncharacterized individuals such as LCT EEN 246. The accession
353 LCT EEN 246 is comprised of Criollo, Marañon, Iquitos, and Contamana ancestry, which
354 could open new possibilities for studying traits of interest.
355
356 Overall, the observed patterns of ancestry are consistent with a known history of the
357 hybridization process in cacao breeding programs and natural hybridization [8]. Studies with
358 a limited set of markers have already hinted at the possibility of further substructure within
359 the ten main populations described for cacao [45,53]. Our results show the usefulness of GBS
360 to address these questions.

Conclusions

We have shown the increased potential for the identification of ancestry informative and novel variants, which are useful for addressing questions about population demographic history and genomic variation in *Theobroma cacao*. We were able to identify 12,357 SNPs markers from only 30 samples, out of which 7,009 could be used for ancestry determination. There is a need to increase our knowledge about cacao accessions throughout the world to produce high yielding and disease resistance genotypes. Our results have practical implications for the development of strategies for genotyping large numbers of accessions with the aim of performing association mapping studies. In our future work, the GBS technique will be refined to include a two digestion enzymes protocol to genotype larger cacao collections and increase the number of common targets/SNPs identified with the method.

Acknowledgments

We thank Path Umaharan, Lambert Motilal, and Annelle Holder-John for providing the germplasm material and associated information used in this study. The authors thank Joanna Kelley for assistance in revising the final version of the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Jaime Osorio was supported through funds assigned from Agrosavia-Colombia to the Cornejo Lab to work on the library preparation of the samples.

Data availability statement

383 Raw sequence reads (data to SRA) is available at NCBI Bioproject ID No. PRJNA486150.
384 The scripts are available at GitHub repository (https://github.com/jaog224/cacao_gbs).
385

386 Supporting information

387 **S1 Table. Summary of the accessions of the *Theobroma cacao* collection used in this**
388 **study.**
389
390 **S2 Table. Summary statistics of genetic diversity of the 7,009 SNPs.**
391
392 **S1 Fig. Distribution of *BsaXI* enzyme cut sites across the *T. cacao* cv. Matina genome.**
393
394 **S2 Fig. Precision of the ancestry analysis using 200 bootstrap pseudoreplicates. (A)**
395 ADMIXTURE bar plot, **(B)** bias bar plot and **(C)** standard error for each ancestry.
396

397 References

398 1. Motamayor J, Risterucci A, Lopez P, Ortiz C, Moreno A, Lanaud C. Cocoa
399 domestication I: the origin of the cocoa cultivated by the Mayas. Heredity (Edinb).
400 2002;89: 380–386.
401 2. Bartley BGD. The genetic diversity of cacao and its utilization [Internet]. Bartley
402 BGD, editor. Genetic Diversity of Cacao and Its Utilization. Wallingford: CABI;
403 2005. doi:10.1079/9780851996196.0000

3. Wood GAR, Lass R. Cocoa. 4th ed. Longman, London; 1985.

4. Donald PF. Biodiversity Impacts of Some Agricultural Commodity Production Systems. Conserv Biol. 2004;18: 17–38. doi:10.1111/j.1523-1739.2004.01803.x

5. Othman A, Ismail A, Abdul Ghani N, Adenan I. Antioxidant capacity and phenolic content of cocoa beans. Food Chem. 2007;100: 1523–1530. doi:https://doi.org/10.1016/j.foodchem.2005.12.021

6. Henderson JS, Joyce RA, Hall GR, Hurst WJ, McGovern PE. Chemical and archaeological evidence for the earliest cacao beverages. Proc Natl Acad Sci. National Academy of Sciences; 2007;104: 18937–18940. doi:10.1073/pnas.0708815104

7. Ogata N, Gómez-pompa A, Taube KA. The domestication and distribution of *Theobroma cacao* L. in the Neotropics. In: McNeil C, editor. Chocolate in Mesoamerica. Gainesville, USA: University Press of Florida; 2009. pp. 69–89. doi:10.5744/florida/9780813029535.003.0003

8. Cornejo OE, Yee M-C, Dominguez V, Andrews M, Sockell A, Strandberg E, et al. Genomic insights into the domestication of the chocolate tree, *Theobroma cacao* L. bioRxiv. Cold Spring Harbor Laboratory; 2017; doi:10.1101/223438

9. Cuatrecasas J. Cacao and its allies: a taxonomic revision of the genus *Theobroma*. Contrib US Herb. 1964;35: 379–614.

10. Hunter JR. The status of cacao (*Theobroma cacao*, sterculiaceae) in the western hemisphere. Econ Bot. 1990;44: 425–439. doi:10.1007/BF02859775

11. World Cocoa Foundation. Cocoa market update [Internet]. 2014 [cited 9 Jul 2018]. Available: http://www.worldcocoafoundation.org/wp-content/uploads/Cocoa-Market-Update-as-of-4-1-2014.pdf.

12. Dantas LG, Guerra M. Chromatin differentiation between *Theobroma cacao* L. and *T.*

428 *grandiflorum* Schum. Genet Mol Biol. scielo; 2010;33: 94–98. Available:
429 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-
430 47572010000100017&nrm=iso
431 13. Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, et al. The genome of
432 *Theobroma cacao*. Nat Genet. The Author(s); 2010;43: 101. doi:10.1038/ng.736
433 14. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, et
434 al. The genome sequence of the most widely cultivated cacao type and its use to
435 identify candidate genes regulating pod color. Genome Biol. BioMed Central Ltd;
436 2013;14: r53. doi:10.1186/gb-2013-14-6-r53
437 15. Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, et al. Making a
438 chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*.
439 DNA Res An Int J Rapid Publ Reports Genes Genomes. Oxford University Press;
440 2015;22: 279–291. doi:10.1093/dnares/dsv009
441 16. Cosme S., Cuevas HE, Zhang D, Oleksyk TK, Irish BM. Genetic diversity of
442 naturalized cacao (*Theobroma cacao* L.) in Puerto Rico. Tree Genet Genomes. Tree
443 Genetics & Genomes; 2016;12: 88. doi:10.1007/s11295-016-1045-4
444 17. Osorio-Guarín JA, Berdugo-Cely J, Coronado RA, Zapata YP, Quintero C, Gallego-
445 Sánchez G, et al. Colombia a source of cacao genetic diversity as revealed by the
446 population structure analysis of germplasm bank of *Theobroma cacao* L. Front Plant
447 Sci. 2017;8: 1994. doi:10.3389/fpls.2017.01994
448 18. Livingstone D, Stack C, Mustiga GM, Rodezno DC, Suarez C, Amores F, et al. A
449 larger chocolate chip—development of a 15k *Theobroma cacao* L. SNP array to create
450 high-density linkage maps. Front Plant Sci. 2017;8: 2008.
451 doi:10.3389/fpls.2017.02008

452 19. Ganal MW, Polley A, Graner E-M, Plieske J, Wieseke R, Luerssen H, et al. Large
453 SNP arrays for genotyping in crop plants. J Biosci. 2012;37: 821–828.
454 doi:10.1007/s12038-012-9225-3
455 20. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al.
456 Performance comparison of benchtop high-throughput sequencing platforms. Nat
457 Biotechnol. Nature Publishing Group, a division of Macmillan Publishers Limited. All
458 Rights Reserved.; 2012;30: 434. Available: http://dx.doi.org/10.1038/nbt.2198
459 21. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust,
460 simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS
461 One. Public Library of Science; 2011;6: 1–10. doi:10.1371/journal.pone.0019379
462 22. Poland JA, Rife TW. Genotyping-by-Sequencing for plant breeding and genetics.
463 Plant Genome J. 2012;5: 92–102. doi:10.3835/plantgenome2012.05.0005
464 23. Johnson JL, Wittgenstein H, Mitchell SE, Hyma KE, Temnykh S V, Kharlamova A
465 V, et al. Genotyping-By-Sequencing (GBS) detects genetic structure and confirms
466 behavioral qtl in tame and aggressive foxes (*Vulpes vulpes*). PLoS One. Public Library
467 of Science; 2015;10: 1–22. doi:10.1371/journal.pone.0127013
468 24. Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, et al. High-
469 resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun.
470 The Author(s); 2015;6: 6914. Available: http://dx.doi.org/10.1038/ncomms7914
471 25. Osorio-Guarín JA, Enciso-Rodríguez FE, González C, Fernández-Pozo N, Mueller
472 LA, Barrero LS. Association analysis for disease resistance to *Fusarium oxysporum*
473 in cape gooseberry (*Physalis peruviana* L). BMC Genomics. BMC Genomics;
474 2016;17: 248. doi:10.1186/s12864-016-2568-7
475 26. Torkamaneh D, Laroche J, Belzile F. Genome-wide snp calling from genotyping by

476 sequencing (GBS) data: a comparison of seven pipelines and two sequencing
477 technologies. PLoS One. Public Library of Science; 2016;11: 1–14.
478 doi:10.1371/journal.pone.0161333
479 27. Alipour H, Bihamta MR, Mohammadi V, Peyghambari SA, Bai G, Zhang G.
480 Genotyping-by-Sequencing (GBS) revealed molecular genetic diversity of iranian
481 wheat landraces and cultivars. Front Plant Sci. 2017;8: 1293.
482 doi:10.3389/fpls.2017.01293
483 28. Bastien M, Boudhrioua C, Fortin G, Belzile F. Exploring the potential and limitations
484 of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato.
485 Genome. 2018;61: 449–456. doi:10.1139/gen-2017-0236
486 29. Gurgul A, Miksza-Cybulska A, Szmatoła T, Jasielczuk I, Piestrzyńska-Kajtoch A,
487 Fornal A, et al. Genotyping-by-sequencing performance in selected livestock species.
488 Genomics. 2018; doi:https://doi.org/10.1016/j.ygeno.2018.02.002
489 30. Cooke TF, Yee M-C, Muzzio M, Sockell A, Bell R, Cornejo OE, et al. GBStools: a
490 statistical method for estimating allelic dropout in reduced representation sequencing
491 data. PLOS Genet. Public Library of Science; 2016;12: 1–18.
492 doi:10.1371/journal.pgen.1005631
493 31. He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z. Genotyping-by-sequencing (GBS), an
494 ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant
495 Sci. 2014;5: 484. doi:10.3389/fpls.2014.00484
496 32. Andrews S. FastQC: a quality control tool for high throughput sequence data.
497 [Internet]. 2018 [cited 9 Jul 2018]. Available:
498 http://www.bioinformatics.babraham.ac.uk/projects/fastqc
499 33. Krueger F. Trim Galore [Internet]. 2018 [cited 9 Jul 2018]. Available:

500 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
501 34. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler
502 transform. Bioinformatics. Oxford University Press; 2009;25: 1754–1760.
503 doi:10.1093/bioinformatics/btp324
504 35. Picard toolkit [Internet]. Broad Institute; 2018. Available:
505 http://broadinstitute.github.io/picard/
506 36. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
507 Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation
508 DNA sequencing data. Genome Res. Cold Spring Harbor Laboratory Press; 2010;20:
509 1297–1303. doi:10.1101/gr.107524.110
510 37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The
511 variant call format and VCFtools. Bioinformatics. Oxford University Press; 2011;27:
512 2156–2158. doi:10.1093/bioinformatics/btr330
513 38. PLINK/SEQ: A library for the analysis of genetic variation data [Internet]. 2018.
514 Available: https://atgu.mgh.harvard.edu/plinkseq/
515 39. Cingolani P, Platts A, Wang LLL, Coon M, Nguyen T, Wang LLL, et al. A program
516 for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:
517 SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly
518 (Austin). Landes Bioscience; 2012;6: 80–92. doi:10.4161/fly.19695
519 40. Rice P, Longden I, Bleasby A. EMBOSS: the european molecular biology open
520 software suite. Trends Genet. 2000;16: 276–277. doi:https://doi.org/10.1016/S0168-
521 9525(00)02024-2
522 41. Bronner IF, Quail MA, Turner DJ, Swerdlow H. Improved protocols for illumina
523 sequencing. Curr Protoc Hum Genet. 2009;0 18: 10.1002/0471142905.hg1802s62.

524 doi:10.1002/0471142905.hg1802s62
525 42. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
526 features. Bioinformatics. Oxford University Press; 2010;26: 841–842.
527 doi:10.1093/bioinformatics/btq033
528 43. Liu K, Muse S V. PowerMarker: an integrated analysis environment for genetic
529 marker analysis. Bioinformatics. 2005;21: 2128–9. doi:10.1093/bioinformatics/bti282
530 44. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
531 unrelated individuals. Genome Res. Cold Spring Harbor Laboratory Press; 2009;19:
532 1655–1664. doi:10.1101/gr.094052.109
533 45. Motamayor JC, Lachenaud P, Wallace J, Loor R, Kuhn DN, Brown S, et al.
534 Geographic and genetic population differentiation of the amazonian chocolate tree
535 (*Theobroma cacao* L.). PLoS One. Public Library of Science; 2008;3: 1–8.
536 doi:10.1371/journal.pone.0003311
537 46. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLOS Genet.
538 Public Library of Science; 2006;2: e190. Available:
539 https://doi.org/10.1371/journal.pgen.0020190
540 47. R development core team. R: a language and environment for statistical computing.
541 [Internet]. Vienna: R Foundation for Statistical Computing; 2008. Available:
542 http://www.r-project.org
543 48. Duchêne S, Ho SYW, Holmes EC. Declining transition/transversion ratios through
544 time reveal limitations to the accuracy of nucleotide substitution models. BMC Evol
545 Biol. London: BioMed Central; 2015;15: 36. doi:10.1186/s12862-015-0312-6
546 49. Chung YS, Choi SC, Jun TH, Kim C. Genotyping-by-sequencing: a promising tool
547 for plant genetics research and breeding. Hortic Environ Biotechnol. 2017;58: 425–

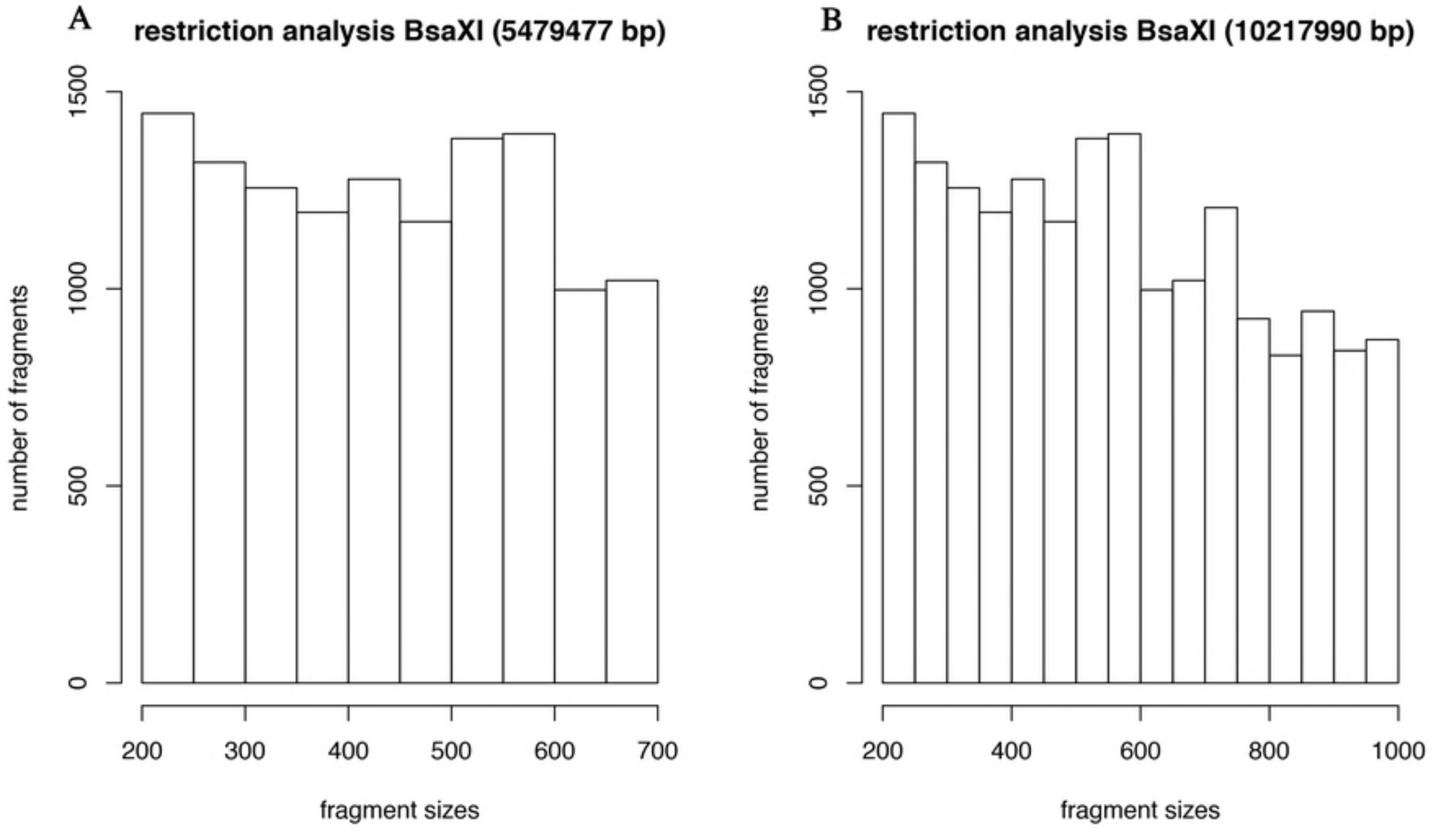
548 431. doi:10.1007/s13580-017-0297-8

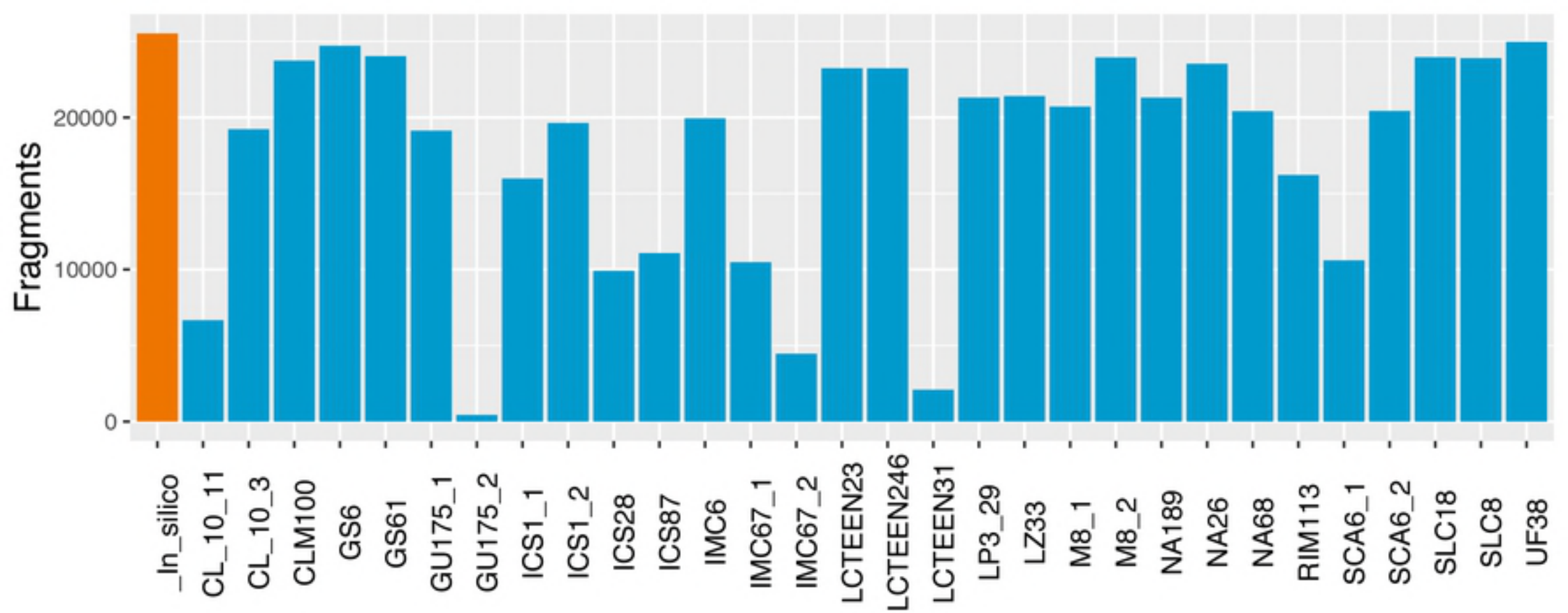
549 50. Romay MC, Millard MJ, Glaubitz JC, Peiffer J a, Swarts KL, Casstevens TM, et al.
550 Comprehensive genotyping of the USA national maize inbred seed bank. Genome
551 Biol. BioMed Central Ltd; 2013;14: R55. doi:10.1186/gb-2013-14-6-r55

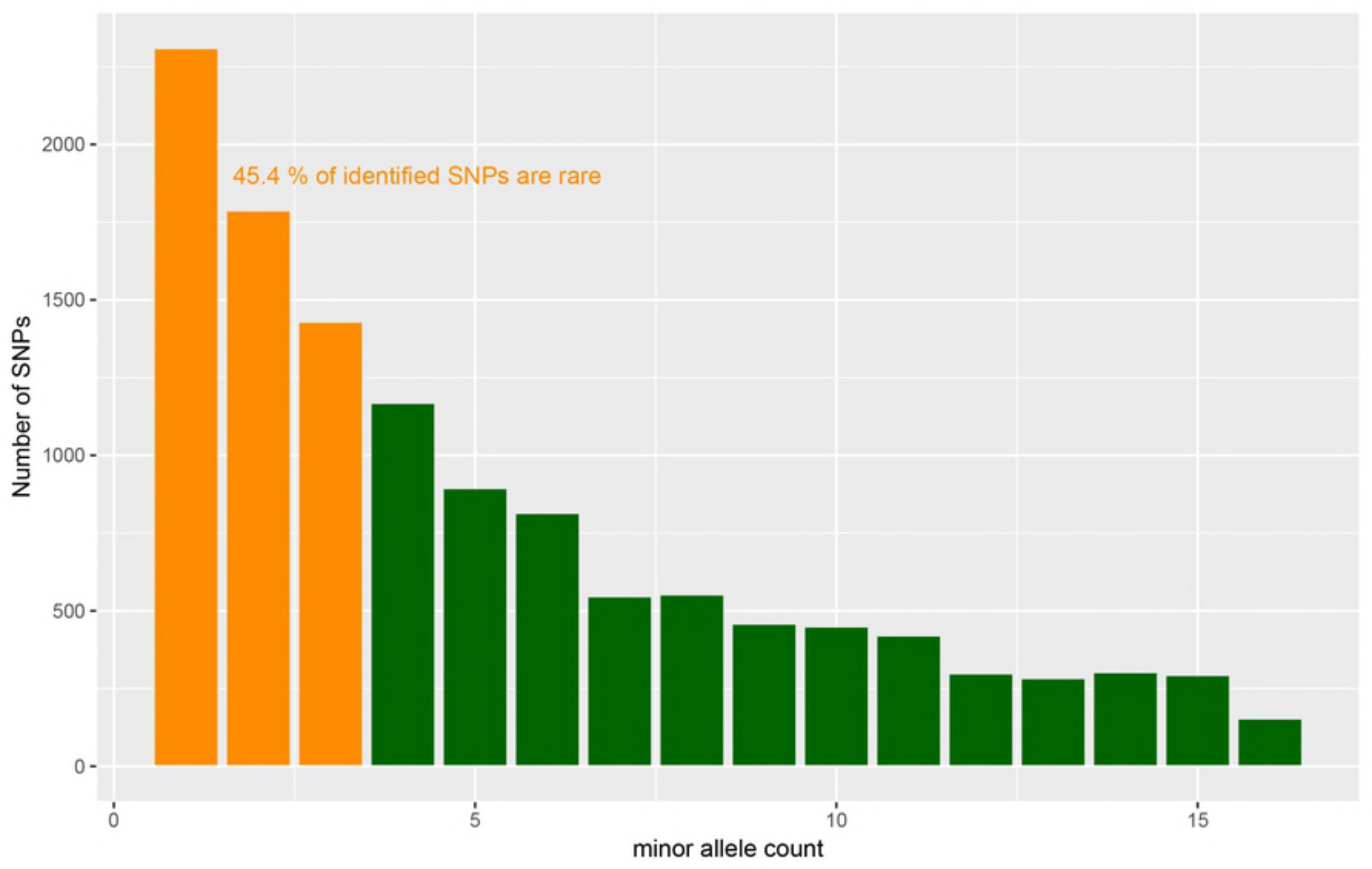
552 51. Ji K, Zhang D, Motilal LA, Boccara M, Lachenaud P, Meinhardt LW. Genetic
553 diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from
554 Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP)
555 markers. Genet Resour Crop Evol. 2013;60: 441–453. doi:10.1007/s10722-012-9847-
556 1

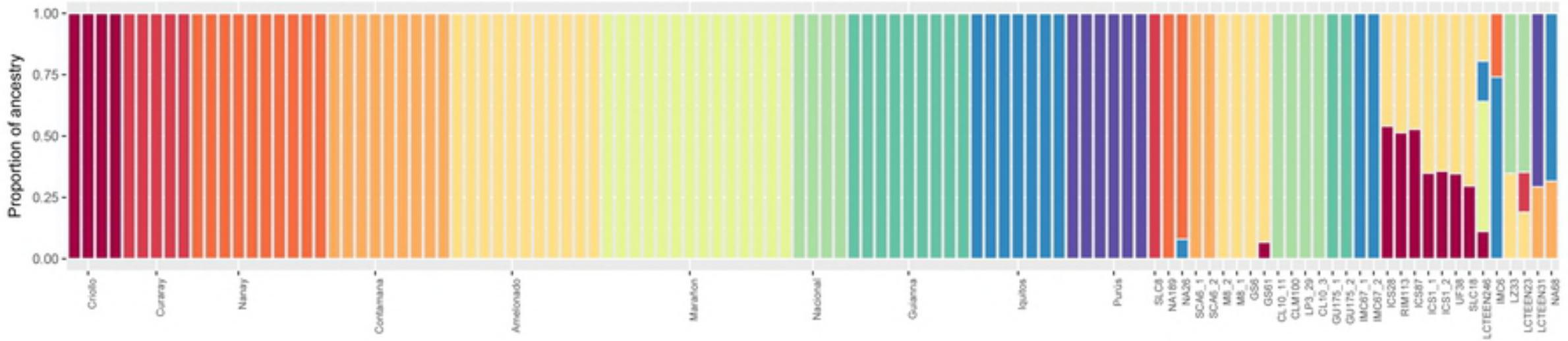
557 52. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, et al. The effect of
558 RAD allele dropout on the estimation of genetic variation within and between
559 populations. Mol Ecol. Wiley/Blackwell (10.1111); 2012;22: 3165–3178.
560 doi:10.1111/mec.12089

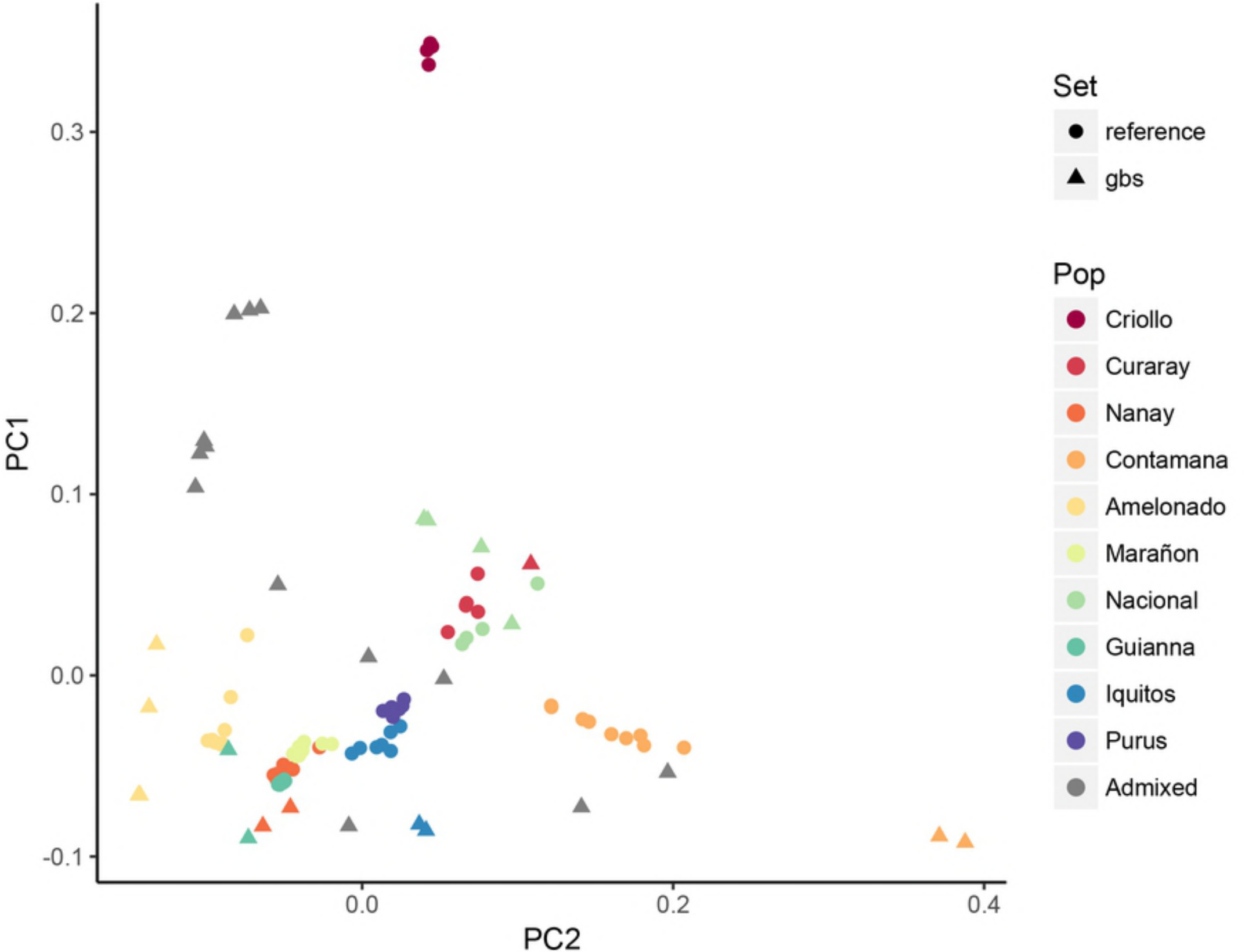
561 53. Zhang D, Martínez WJ, Johnson ES, Somarriba E, Phillips-Mora W, Astorga C, et al.
562 Genetic diversity and spatial structure in a new distinct *Theobroma cacao* L.
563 population in Bolivia. Genet Resour Crop Evol. 2012;59: 239–252.
564 doi:10.1007/s10722-011-9680-y


A
restriction analysis BsaXI (5479477 bp)
number of fragments
0
500
1000
1500
200
300
400
500
600
700
fragment sizes
B
restriction analysis BsaXI (10217990 bp)
number of fragments
0
500
1000
1500
200
400
600
800
1000
fragment sizes


Fragments
20000
10000
0
_In_silico
CL_10_11
CL_10_3
CLM100
GS6
GS61
GU175_1
GU175_2
ICS1_1
ICS1_2
ICS28
ICS87
IMC6
IMC67_1
IMC67_2
LCTEEN23
LCTEEN246
LCTEEN31
LP3_29
LZ33
M8_1
M8_2
NA189
NA26
NA68
RIM113
SCA6_1
SCA6_2
SLC18
SLC8
UF38


45.4 % of identified SNPs are rare
Number of SNPs
2000
1500
1000
500
0
0
5
10
15
minor allele count


Proportion of ancestry
1.00
0.75
0.50
0.25
0.00
Criollo
Curaray
Naway
Contamana
Amelionado
Marañon
Nacional
Guianna
Iquitos
Purús
SLC8
NA189
NA36
SCA6_1
SCA6_2
M8_2
M8_1
GS6
GS61
CL10_11
CLM100
LP3_29
CL10_3
GU175_1
GU175_2
IMD67_1
IMD67_2
ICS28
RIM113
ICS87
ICS1_1
ICS1_2
UF38
SLC18
LCTEEN246
IMC6
LZ33
LCTEEN23
LCTEEN31
NA68


Set
reference
gbs
Pop
Criollo
Curaray
Nanay
Contamana
Amelonado
Marañon
Nacional
Guianna
Iquitos
Purus
Admixed
PC1
0.3
0.2
0.1
0.0
-0.1
PC2
0.0
0.2
0.4