Development of core-collections for Guizhou tea genetic resources and GWAS of leaf size using SNP developed by genotyping-by-sequencing

Suzhen Niu [1,2], **Hisashi Koiwa** [3], **Qinfei Song** [2], **Dahe Qiao** [1], **Juan Chen** [1], **Degang Zhao** [Corresp., 1], **Zhengwu Chen** [Corresp., 1], **Ying Wang** [4], **Tianyuan Zhang** [4]

[1] Guiyang Station for DUS Testing Center of New Plant Varteties (MOA) / Institute of Tea, Guizhou Academy of Agricultural Sciences, Guiyang, China

[2] The Key Laboratory of Plant Resources Conservation and Germplasm Innovationin Mountainous Region (Ministry of Education), Institute of Agro-Bioengineering / College of Tea Science, Guizhou University, Guiyang, China

[3] Vegetable and Fruit Improvement Center, Department of Horticultural Sciences, Molecular and Environmental Plant Sciences Program, Texas A&M University, College Station, Texas, USA

[4] Wuhan Benagen Tech Solutions Company Limited, Wuhan, China

Corresponding Authors: Degang Zhao, Zhengwu Chen
Email address: dgzhao@gzu.edu.cn, zwchentea@163.com

An accurate depiction of the genetic relationship, the development of core collection, and genome-wide association analysis (GWAS) are key for the effective exploitation and utilization of genetic resources. Here, genotyping-by-sequencing (GBS) was used to characterize 415 tea accessions mostly collected from the Guizhou region in China. A total of 30,282 high-quality SNPs was used to estimate the genetic relationships, develop core collections, and perform GWAS. We suggest 198 and 148 accessions to represent the core set and mini-core set, which consist of 47% and 37% of the whole collection, respectively, and contain 93-95% of the total SNPs. Furthermore, the frequencies of all alleles and genotypes in the whole set were very well retained in the core set and mini-core set. The 415 accessions were clustered into 14 groups and the core and the mini-core collections contain accessions from each group, species, cultivation status and growth habit. By analyzing the significant SNP markers associated with multiple traits, nine SNPs were found to be significantly associated with four leaf size traits, namely MLL, MLW, MLA and MLSI ($P<1.655E-06$).This study characterized the genetic distance and relationship of tea collections, suggested the core collections, and established an efficient GWAS analysis of GBS result.

1 **Development of core-collections for Guizhou tea genetic resources and**

2 **GWAS of leaf size using SNP developed by genotyping-by-sequencing**

3 **Suzhen Niu[1,2,3], Hisashi Koiwa[3], Qinfei Song[2], Dahe Qiao[1], Juan Chen[1], Degang Zhao[1,2], Zhengwu Chen[1],**

4 **Ying Wang[4], Tianyuan Zhang[4]**

5 [1] Guiyang Station for DUS Testing Center of New Plant Varteties (MOA) / Institute of Tea, Guizhou Academy of

6 Agricultural Sciences, Guiyang 550006, Guizhou Province, RP China

7 [2] The Key Laboratory of Plant Resources Conservation and Germplasm Innovationin Mountainous Region (Ministry

8 of Education), Institute of Agro-Bioengineering / College of Tea Science, Guizhou University, Guiyang 550025,

9 Guizhou Province, RP China

10 [3] Vegetable and Fruit Improvement Center, Department of Horticultural Sciences, Molecular and Environmental

11 Plant Sciences Program, MS2133 Texas A&M University, College Station, 77843-2133, TX, USA

12 [4] Wuhan Benagen Tech Solutions Company Limited, Wuhan 430070, Hubei Province, RP China

13 Corresponding Author: Degang Zhao,Email:dgzhao@gzu.edu.cn; Zhengwu Chen, Email: zwchentea@163.com

14   **Abstract**

15   An accurate depiction of the genetic relationship, the development of core collection, and

16   genome-wide association analysis (GWAS) are key for the effective exploitation and utilization

17   of genetic resources. Here, genotyping-by-sequencing (GBS) was used to characterize 415 tea

18   accessions mostly collected from the Guizhou region in China. A total of 30,282 high-quality

19   SNPs was used to estimate the genetic relationships, develop core collections, and perform

20   GWAS. We suggest 198 and 148 accessions to represent the core set and mini-core set, which

21   consist of 47% and 37% of the whole collection, respectively, and contain 93-95% of the total

22   SNPs. Furthermore, the frequencies of all alleles and genotypes in the whole set were very well

23   retained in the core set and mini-core set. The 415 accessions were clustered into 14 groups and

24   the core and the mini-core collections contain accessions from each group, species, cultivation

25   status and growth habit. By analyzing the significant SNP markers associated with multiple traits,

26   nine SNPs were found to be significantly associated with four leaf size traits, namely MLL,

27   MLW, MLA and MLSI ($P<1.655E-06$).This study characterized the genetic distance and

28   relationship of tea collections, suggested the core collections, and established an efficient GWAS

29   analysis of GBS result.

30   **Keywords:** Core Collections, GWAS, Leaf Size, Tea Plant, GBS

31 **Introduction**

32 Tea, coffee and cocoa are considered as the world's three best-known beverages and are also the

33 most produced or consumed beverages worldwide (Wambulwa et al., 2016; Liu et al., 2017). To

34 produce tea, people need to infuse processed tender shoots of the tea plant (*Camellia sinensis*)

35 with connected species, all of which belong to *Thea* of genus *Camellia* in the family Theaceae

36 (Ma et al., 2018). Tea is a diploid ($2n = 2x = 30$) with a genome size of 3.02 Gb and is an

37 endemic species in southwest China (Wei et al., 2018). Tea is highly nutritious with medicinal

38 properties, offering a wealth of health benefits. Daily tea consumption is beneficial for reducing

39 the risks of various cancers, diseases related to obesity, and neurological and cardiovascular

40 dysfunctions (Naghma et al., 2013; Hayat et al., 2015). The rich flavors and multiple health-

41 promoting functions of tea are conferred by 700 bioactive compounds such as catechins, caffeine,

42 theanine, and volatiles (Xia et al., 2017).

43 Tea has a long utilization history in the Yunnan-Guizhou Plateau and is now cultivated

44 worldwide (Wei et al., 2018; Chen et al., 2012). In this region, the diversity of tea germplasm is

45 well preserved with abundant wild tea plants, ancient landraces and modern landraces with

46 different morphological characteristics due to the region's unique geology, diverse climates,

47 plentiful rainfall and the cross-pollination nature of tea (Niu, 2014). In addition, due to the slow

48 socio-economic development and land use in Guizhou Plateau, elimination of various tea species

49 at a large spatial scale has not occurred.

50 The ever-growing popularity of tea products calls for new tea varieties that meet the needs

51 of market diversification (Liang et al., 2015). Similar to other perennial woody species, tea plants

52 have a long growth period and the genome is highly heterozygous, making breeding difficult and

53 costly. Tea producers also face new challenges such as the sustainability of high-quality tea

54 production, environmental change, pest invasion, and diseases. Hence, it is vital to explore the

55 genetic basis of complex traits and identify favorable alleles for breeding new tea cultivars that

56 can overcome these threats by marker-assisted breeding (MAB) (Tan et al., 2016).

57 Compared with other ways, association mapping is faster and more effective in dissecting

58   the genetic basis of complex traits and identifying favorable genetic resource. The results yielded

59   from association mapping have greatly facilitated MAB programs (Iso-Touru et al., 2016).

60   Association mapping is also promising in evading limitations of linkage mapping. The difference

61   between traditional linkage mapping which is often based on bi-parental populations and

62   association mapping is that the latter takes advantage of the ancestral recombination in natural

63   populations to identify loci that significantly associate with traits of interest based on the linkage

64   disequilibrium (LD) (Buckler et al., 2002; Pace et al., 2015). Analysis of a large number of

65   alleles in various populations is made possible by association mapping (Pace et al., 2015;

66   Suwarno et al., 2015; Motilal et al., 2016). Other advantages of GWAS compared with

67   traditional linkage mapping (Tan et al., 2016; Bali et al., 2015) include high-resolution and

68   providing a less time-consuming approach for developing the mapping population (Heena et al.,

69   2018).

70   The population used for association mapping is of great importance, it must have a wide

71   range of diversity that represents most historical recombination events (Heena et al., 2018).

72   Although species diversity with abundant tea genetic collections is well preserved in the center

73   of origin, recent tea breeding programs have not exploited the allelic diversity of many traits. As

74   a result, characterizing tea plant collections should be the first step towards stimulating the use of

75   genetic resources. Germplasm preservation can be costly and time-consuming, therefore,

76   developing 'core collections' with a minimal number of tea varieties that retain the genetic

77   diversity is a cost-effective strategy (Frankel.,1984 ; Brown.,1989; Wang et al.,2011; Taniquchi

78   et al., 2014; Campoy et al.,2016; Hintum et al.,2000; Ndjiondjop et al., 2017; Lassois et al.,2016).

79   In addition, 'core-collections' are useful for genetic association analysis and the identification of

80   genomic variation (Hugginsa et al., 2018; Zhang et al., 2018; Muleta et al., 2016; Cunff et al.,

81   2008; Heena et al., 2018; Ma et al., 2018). Criteria based on the genetic distance between

82   accessions have been proven sueful in evaluating and creating 'core collections' (Odong et al.,

83   2013; Campoy et al., 2016; Ndjiondjop et al., 2017).

84   Previous studies on association mapping and core-collection development mainly focused on

85   maize (Pace et al., 2015; Suwarno et al., 2015; Coan et al., 2018), sunflower (Heena et al., 2018),

86   wheat (Muleta et al., 2017), pine (Bartholome et al., 2017), sorghum (Bouchet et al., 2017),

87   finger millet (Babu et al., 2018), melon (Hou et al., 2018), pea (Desgroux et al., 2018) and cotton

88   (Ma et al., 2018). By contrast, little has been done on tea germplasm with high density SNPs

89   distributed on the whole genome. GBS is a large-scale approach for identifying high density SNP

90   markers that are suitable for association mapping. In this study, GBS was used to analyze the

91   genetic diversity and population structure of 415 tea accessions including wild types, ancient and

92   modern landraces from the Guizhou Plateau, and breeding varieties from Zhejiang, Fujian,

93   Hunan and Guizhou. We aim to (1) analyze the genetic purity, distance and relationships, (2)

94   establish and evaluate the core collections, and (3) perform GWAS on the tea population. Our

95   findings provide a valuable resource for developing new molecular markers that can be used for

96   MAB of tea varieties.

97

98   **Materials and Methods**

99   **Plant materials**

100   A total of 415 accessions including 159 wild type and 256 cultivation type varieties (174 ancient

101   landraces, 77 modern landraces and 5 breeding varieties) were used for this study (Fig. S1; Table.

102   S1). Based on the classification system established by Chen et al. (2000) and Min (1992), 251

103   *C.sinensis* (L.) O. Ktze, 100 *C.tachangensis* F.C. Zhang, 59 *C. remotiserrata* Zhang, and five

104   near *C.taliensis* W.W. Smith were identified (Table S1). Hereafter, wild varieties and their

105   natural offspring will be referred to as "wild type", the cultivation tea varieties aged more than

106   one hundred years old will be called "ancient landraces", and the tea garden landraces will be

107   named "modern landraces" (Table. S1). The ancient landraces, modern landraces and "breeding

108   varieties" that have undergone artificial selection are collectively referred to as the "cultivation

109   type".

110   One hundred and sixty-eight samples were collected from areas suitable for tea growth in

111   north Guizhou (Ia) (Fig. S1), 51 samples were collected from areas suitable for tea growth in east

112   Guizhou (Ib), 57 samples were collected from areas suitable for tea growth in south Guizhou (Ic),

113   83 samples were collected from areas suitable for tea growth in central Guizhou (II), 41 samples

114   were collected from areas with a minor suitable climate for tea growth in west Guizhou (III), ten

115   samples were collected from areas with an unsuitable climate for tea growth in west Guizhou,

116   one cultivar was collected from Guizhou, two cultivars were collected from Fujian, one cultivar

117   was collected from Zhejiang, and one cultivar was collected from Hu'nan (Table. S1). The

118   samples were planted in the city of Guiyang, China. The healthy tender shoots were harvested,

119   snap-frozen in liquid N_2, and kept at -80 °C until use (Jin et al., 2018).

120   **DNA extraction**

121   Genomic DNA was extracted from the plant samples with a kit for rapid extraction of genomic

122   DNA. DNA integrity was tested by 1% agarose gel electrophoresis; a Qubit Fluorometer was

123   used to check the purity and measure the concentration of DNA samples (Niu et al., 2019).

124   **Sequencing of the GBS library**

125   One-hundred nanogram genomic DNA was double digested with 5 U of *Sac*I and *Mse*I (NEB) in

126   a 25 μl reaction containing 1 × restriction buffer. The resulting samples were ligated with the

127   restriction fragment using the SacAD and MseAD adaptors, which contained different barcode

128   combinations to distinguish the samples. Equal volumes of the ligated products of 12 individuals

129   were pooled and processed using the QIA quick PCR Purification Kit (Qiagen). PCR was

130   performed using PCR Master Mix and the PCR Primer Cocktail, which could enrich DNA

131   fragments with the adapters. The PCR products of each mixture were pooled and separated on a

132   2% agarose gel by electrophoresis. Fragments of 500 to 550 bp (including the 120 bp adaptor)

133   were recovered with the QIA quick Gel Extraction Kit (Qiagen). The average length of DNA

134   fragment was selected for final library construction with the Agilent DNA 12,000 kit using the

135   2100 Bioanalyzer system. To quantify the final library, quantitative real-time PCR with a Taq

136   Man probe was employed. Then the libraries were sequenced on the IlluminaHiSeq X Ten

137   platform based on the paired-end 150 (PE150) strategy. Each library contains 48 samples, and

138   clean data were then parsed into different units that exactly match the barcodes and the

139   restriction sites at both ends (Elshire et al., 2011).

140   **SNP genotyping basing on the sequence data**

141   Original IlluminaHiSeq X Ten reads were de-multiplexed based on the barcodes, and a custom

142   Perl script was used to shear the barcoded sequences. Reads in which >50% of the bases have

143   quality values ≤5 were discarded; the clean reads were mapped to the tea reference genome

144   (http://www.plantkingdomgdb.com/tea_tree/) (Xia et al., 2017) using BWA-MEM (v.0.7.10)

145   with parameters '-*T 20 -k 30*' (Li, 2013). SNPs and InDels were called using GATK (v.3.7.0).

146   The SNPs were filtered according to the methods described by Hussain et al (2017), Chen et

147   al (2017) and Eltaher et al. (2018) based on the following criteria: (1) variants must be bi-allelic

148   SNPs, (2) "QUAL < 50.0 || QD < 2.0 || FS > 60.0 || MQ < 40.0 || Mapping Quality Rank Sum < -

149   12.5 || Read Pos Rank Sum < -8.0" was used in Variant Filtration in GATK (V 3.7.0) to filter the

150   SNPs, (3) SNPs with MAF < 0.05 or >10% missing data were filtered out by VCFtools (V

151   0.1.15); (4) a window of 50 SNPs, a step size of ten SNPs, and an r^2 threshold of 0.2 were used

152   to prune the SNPs using Plink (v1.9). As a result, a set of 415 accessions and 30,282 high-quality

153   SNPs were retained and used for subsequent analyses (Table S2). We connected all the scaffolds

154   into 20 pseudo-groups (Table. S7).

155   **Phenotypic data collection**

156   The 415 accessions were measured for leaf size for association mapping. We measured the ten

157   representative mature leaves length (MLL) and mature leaves width (MLW) of each individual

158   and calculated their means to represent the traits of MLL and MLW, respectively. Leaf size

159   measurements were performed in Spring (monthly average temperature is 14.68℃) and Fall

160   (monthly average temperature is 21.29℃), 2018, seperately. Leaf shape index (MLSI),

161   calculated as MLL/MLW, was the third trait we measured. The fourth trait, mature leaf area

162   (MLA), was calculated as 0.75 × MLW × MLL. We modified our data collection method

163   according to guidelines of UPOV (International Union for the Protection of New Varieties of

164   Plants 2008) for tea. Statistical analysis of the phenotypic data was conducted using Microsoft

165   Excel 2010 and SPSS 15.0 software. Pearson's correlation coefficients (r) were calculated for the

166   four traits. Means, standard deviations, and ranges were calculated for trait distribution. To

167   evaluate whether the data followed a normal distribution, Skewness and Kurtosis were calculated

168   using the Descriptive Statistics model-based Frequencies distribution of analysis implemented in

169   SPSS15.0 software.

170   Genetic purity, distance, and relationships of accessions

171   The following equation was used to calculate the polymorphism information content (PIC)

172   values for the SNP data (Bostein et al., 1980). *Pi* and *Pj* were the frequencies of the number of *i*

173   and *j* alleles respectively, and n was the number of alleles (Botstein et al.,2008). We can evaluate

174   each marker locus for its PIC by summing the mating frequencies multiplied by the probability

175   that an offspring will be informative. Under our assumptions, the expected value of PIC can be

176   calculated as

177
$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} P_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2P_i^2 P_j^2$$

178   The alleles, genetic distance and the observed heterozygosity (Ho) were calculated on every

179   groups with TASSEL v.5.2.37 (Bradbury et al., 2007). The neighbor-joining cluster analysis

180   were performed using DARwin v.6.0.17.

181   Development of Core Collections

182   As subsets of larger genetic collections, core collections contain the smallest number of

183   accessions that represent the maximum diversity of the raw collection. DARwin (v.6.0. 17) was

184   used to construct the diversity trees (Hamon et al., 2003). We used 10,000 bootstraps to

185   determine dissimilarities and transformed them into Euclidean distances. The phylogenetic tree

186   was constructed based on the 30,282 SNPs using the unweighted Neighbor Joining (NJ) method.

187   Then, the 'maximum length subtree function' was used to generate the core collection as

188   described previously for cowpea (Egbadzor et al., 2014), prunus (Campoy et al., 2016), sorghum

189   (Claire et al., 2013) and rice (Ndjiondjop et al., 2017). The maximum length subtree

190   implemented a stepwise procedure that consecutively prunes redundant units. In this procedure,

191 sample size determination, which maintains the largest diversity, is allowed, and the procedure is

192 visualized in the phylogenetic tree of the original population of all 415 accessions. If the distance

193 between two accessions, as judged by the edge length, is small, they are considered redundant.

194 Because of more uncommon characters, accessions with the longest edge are considered the most

195 diverse. A 'removed edge value' offered by the NJ tree was used to identify presumptive clusters

196 of synonym accessions, and a threshold value of 0.0008 (Euclidean distance) was required to

197 confirm the synonym mentioned. We determined the final core set that represents the maximum

198 genetic diversity based on the pruned edge length of the initial tree length and the sphericity

199 index (Hamon et al., 2003).

200 **Association analysis**

201 Association analysis was carried out on the four measurable traits, namely MLL, MLW, MLSI,

202 and MLA, in two seasons, independently. We used the mixed linear model (MLM) implemented

203 in TASSEL (v.5.2.37) because the GLM model has a high false positive rate (Bradbury et al.,

204 2007) following the user manual. To overcome this limitation, a comparison was drawn between

205 the two association analysis models using TASSEL 5.2.43, and the most optimized model for

206 each analyzed quality trait was identified and used for subsequence analyses (Bradbury et al.,

207 2007; Heena et al.,2018). To control the potential false-positives result from the confounding of

208 population structure, PCA-matrix or Q-matrix was used as the fixed effect in MLMs. Kinship

209 matrix (K) was considered as a covariate factor in MLMs to cut down the rate of false positives

210 among genotypes (Yu et al., 2006). If the significance threshold of a SNP has the lowest P-value

211 in the peak area ($P \leqslant 0.05/30,282=1.65\times10^{-6}$), it is considered significantly associated with the

212 targeted trait. Correlation coefficient (R^2) explained the phenotypic variation by every marker-

213 trait association analyzed.

214 **Results**

215 **Genetic purity, distance, and relationships**

216 GBS analysis of the 415 tea accessions was conducted using IlluminaHiSeq X Ten and the

217 relevant information was described in our previous study (Niu et al., 2019). Observed

218   heterozygosity in each sample ranged from 0.041 to 0.386 (Fig. S2), with an average

219   heterozygosity of 0.225. Among the 415 accessions, 73.5% had an observed heterozygosity >

220   0.201 (Table. S1). The pairwise genetic distance between two accessions ranged from 0.088 to

221   0.298 (Table. S2), with an average genetic distance of 0.249 among all accessions (Fig. 1; Table.

222   S2). The genetic distances of 36.6% accession pairs were between 0.200 and 0.250, and those of

223   3.73% and 59.68% accession pairs were < 0.200 and > 0.250, respectively (Fig.1; Table S2).

224   Based on the NJ cluster analysis on the genetic distance matrix, the 415 accessions could be

225   clustered into fourteen groups (Fig. 2A; Table. S1) based on cultivation status (ancient landraces,

226   wild tree, modern landraces and breeding cultivars), growth habits (cultivation type and wild

227   type), and species, and the here-derived core and mini-core collections of origin. Group one

228   consisted of 71 accessions, most of which were cultivation type (of the group one, 85.92% are

229   cultivation type), ancient landraces (of the group one, 74.65% are ancient landraces), and *C.*

230   *sinensis* (of the group one, 85.92% are *C. sinensis*). The second group contained 88 accessions,

231   most of which belong to *C. tachangensis* (of second group, 93.18% are *C. tachangensis*) and

232   wild type (of the second group, 96.59% are wild type). The third group consisted of 54

233   accessions, most of which were wild type (of the third group, 98.15% are wild type) and *C.*

234   *remotiserrata* varieties (of the third group, 91.48% are *C. remotiserrata*). Group four had a total

235   of 76 accessions and most were cultivation type (of the group four, 97.37% are cultivation type),

236   modern landraces (of the group four, 61.84% are cultivation type) and *C. sinensis* (of the group

237   four, 98.68% are *C. sinensis*). Group five had 54 accessions, all of which were from *C. sinensis*

238   and cultivation type, and the number of modern landraces were almost the same as that of the

239   ancient landraces. Groups six, seven, eight, 12, 13 and 14 consisted of 15, 15, two, seven, one

240   and one accessions, respectively, all of which were ancient landraces, cultivation type, and *C.*

241   *sinensis* varieties. Groups nine and ten consisted of three and six accessions, respectively, which

242   were wild tree and *C. remotiserrata* varieties. Group 11 had a total of 20 accessions and most

243   were cultivation type (of the Group 11, 95.45% are cultivation type), ancient landraces (of the

244   Group 11, 95.45% are ancient landraces) and *C. sinensis* varieties (of the group 11, 90.91% are *C.*

245   *sinensis* varieties) (Fig. 2A; Table. S1).

Creating core and mini-core collections

247   We proposed genetic core and mini-core sets to represent the genetic diversity of a large tea

248   population. These core and mini-core sets can be used for association studies, breeding, and

249   other purposes (Brown, 1989). The maximum length subtree method implicated in DARwin

250   v.6.0.17. was repeatedly applied to remove the most redundant accessions until the sphericity

251   index percentage and pruned edge were linear with a relatively low slope, corresponding to 299

252   accessions (Fig. S4). The 299 core accessions could represent the 415 accessions (referred to as

253   'core-1' from hereafter) (Table S1; Fig. 2B). Then, the sphericity index increases stably and

254   slowly from 301 to 198 on the x-axis, indicating no significant difference in the information of

255   the 104 accessions, suggesting that the elimination of these accessions had no significant impact

256   on sphericity index (Fig. S4). The 198 core accessions (referred to as 'core-2' hereafter, which

257   contains less accessions than 'core-1') were selected to represent the 415 accessions (Fig. 2C;

258   Table. S1). The percentage of sphericity index and pruned edge flattened until reaching a sample

259   size of 146 accessions on the x-axis (Fig. S4). Thus, the 146 remaining accessions constitute the

260   mini-core set. We suggest that these 148 accessions, consisting of the 146 retained accessions

261   and two cultivars ('Fudingdabaicha' and 'Tieguanyincha'), represent the mini-core set of the

262   Guizhou Plateau tea germplasm (Fig. 2D; Table. S1). The selected cores and mini-core sets

263   originated from 14 different clusters yielded from NJ tree, respectively, although some

264   individuals were assigned to other clusters (Fig. 2), which suggested that edges constituting the

265   phylogenetic tree backbone have been retained.

Diversity assessment of the cores and mini-core sets

267   The alleles and genotypes, PIC, genetic distances and minor allele frequency of the 415

268   accessions were compared to those of the two core sets and the mini-core set. With regard to the

269   average polymorphic sites in the 415 tea accessions, the reduction in sample size from 415 to

270   299 and 198 accessions in the two core sets, and 148 in the mini-core set reduced SNPs numbers

271   by only 103 (0.36%), 1,403 (4.93%), and 1,722 (6.04%), respectively (Table. 1). Ho, PIC and

272   MAF in Core-1 and Core-2 sets were almost the same as for whole tea accessions, suggesting

273   that both core-1 and core-2 can represent the 415 accessions. Ho, PIC and MAF in mini-Core set

274   were 89.77%, 98.89% and 98.25% of those of the whole population, respectively (Table. 1).

275   Allele and genotype frequencies in the two core sets and one mini-core sets almost made no odds

276   from the entire set (Fig. S3). Consistent with our hypothesis, the genetic distances between

277   accession pairs increased slightly in the core and mini-core sets (Fig. 1; Table S2~S5). The

278   proportion of accession pairs with a genetic distance greater than 0.15 was 98.59% for the entire

279   population, compared to 99.95%, 99.97% and 99.95% in core-1, core-2 and the mini-core set

280   (Fig. 1; Table. 1), respectively. The proportion of samples number removed per group based on

281   cluster analysis of NJ tree was not consistent in the two core and one mini-core sets (Fig. 2;

282   Table. 2; Table. S1). The proportion of accessions removed from group two and group four were

283   significantly higher than that from groups 1, 3, 5, 6, 7, 10, 11 and 12. The numbers of accessions

284   excluded from groups 8, 9, 13 and 14 in the two core sets and the mini-core set were same as that

285   from the entire tea collection (Fig. 3). The mini-core collection consisted of the accessions from

286   all 14 groups, as revealed by the cluster analysis of NJ tree, and represented 14 groups, three

287   cultivation status, two growth habits, and four different species (Fig. 3). These results indicated

288   that the core and mini-core collections can well represent the 415 accessions.

289   **Genome-wide association studies**

290   Skewness and Kurtosis for MLL, MLW and MLSI did not differed from those of a normal

291   distribution (Table. 3). Despite the existence of left obliqueness, MLA still exhibited normal

292   distribution. (Table. 3). The four traits were significant correlated ($P < 0.0001$), as revealed by

293   the Pearson correlation analysis (Table. 4). MLSI was highly correlated with MLL and MLW but

294   not with MLA. The same trait was significant correlated ($P < 0.0001$) between two environments

295   (Table. 4).

296   The population structure of the 415 accessions was studied with STRUCTURE 2.3.4 on the

297   basis of the LD-pruned 1,135 high-quality SNPs. Three subpopulations were acquired, which

298   was coincident with the preceding reports (Niu et al., 2019). As a result, the Q-matrix was

299 generated from k=3 using STRUCTURE 2.3.4 and used for GWAS as the fixed effect. P-matrix,

300 as the first three PCs value given by the PCA analysis, was used as covariates in the GWAS

301 model. We also calculated kinship matrix using TASSEL. Therefore, the MLMQ+K and

302 MLMP+K models were compared based on the Q-Q plot (Fig. S5) and the MLMP+K model best

303 fitted the four traits.

304   We identified nine SNPs that significantly associated with MLL, MLW, MLA and MLSI (P

305 < 1.655E-06) (Fig. S6; Table. 5; Table. S6). Among these, three were significantly associated

306 with multiple traits. In the case of MLL, significant associations were detected for SNPs

307 PG_5:79474508 (on xfSc0000122) and PG_1:32346865 (on Sc0000000), which explained 3.77%

308 and 3.23% of the phenotypic variation, respectively, and had a deduced effect of major alleles

309 from -0.72 to -2.94 (Table. 5; Table. S6).

310   For MLW, only one significant SNP was detected on the same locus as MLL, which

311 explained 8.24% of the phenotypic variation; this SNP has a negative major allele effect of -1.31

312 and a positive minor allele effect of 0.44 for MLW. Six loci, located at Sc0000000, Sc0000365,

313 Sc0002172, Sc0002452, Sc0004093 and xfSc0000122, significantly correlated with MLA,

314 explaining 3.49% to 4.50% of the total phenotypic variation, with the main influence varying

315 from -18.23% to 4.462% and minor effects varying from -18.76% to 42.47% (Table. 5; Table.

316 S6). Three SNPs located at Sc0000037, Sc0000011 and Sc0000046 were found to significantly

317 associate with MLSI, explaining 4.012%, 4.33%, and 3.83% of the MLSI variation at 82,813,415,

318 30,871,283, and 98,891,957 bp, respectively; while the deduced effect of the major alleles

319 ranged from -0.08 to 0.06 (Table. 5; Table. S6). Among these SNPs, two were co-associated with

320 multiple traits. PG_5:79474508, which had a negative effect of major alleles, was co-associated

321 with MLL and MLA. PG_1:32346865 with an increased effect of major alleles was co-

322 associated with MLL, MLW and MLA (Table. 5; Table. S6). Three, three and two significantly

323 associated SNPs identified in the whole population were retained in core-1, core-2 and mini-core

324 set, separately (Table. S7).

325 **Discussions**

326 **Genetic purity, genetic distance and relationships**

327 GBS is a cost-efficient and effective genotyping method (Yang et al., 2017; Bhattarai et al., 2018;

328 Hackett et al., 2018). In this study, 390.30 Gb clean reads were generated. Among them, 0.208-

329 3.32 Gb clean reads were generated after the quality filtering step in 415 accessions, and 30,282

330 high-quality SNPs were retained with strict filter conditions. More SNPs were identified in this

331 study compared with previous studies, which meets the requirements of our objectives (Chen et

332 al., 2017; Eltaher et al., 2018)

333 *Camellia sinensis* and related species are cross-pollinated species with a low self-pollination

334 rate ranging from 2% to 6% (Chen et al., 2000; Ma et al., 2018). Consistently, the heterozygosity

335 rate of 73.5% were > 20%. Most alterations in allele frequencies occurred during natural

336 regeneration without human intervention, which is in agreement with previous studies (Niu,

337 2014), indicating that the tea population from the center of origin maintained a high level of

338 genetic diversity. Genetic distance measures the genetic divergence between a given pair of

339 accessions or populations; the pairs usually share many alleles with a small genetic distance

340 (Ndjiondjop et al., 2017). The results in this study suggested remarkably different genetic

341 distances between pairs of 415 accessions, with 3.73% of the pairs in similarity, 36.6% of the

342 pairs moderately distant, and 59.68% significantly distant. We detected very low redundancy in

343 our collection, which consists with the low observed genetic divergence and high genetic

344 variation retained from the original tea population. Using clustering, we identified 14 groups or

345 clusters to establish the core accessions. In general, the groups clustered according to the

346 cultivation status, growth habits, and species.

347 **Selection of core and mini-core**

348 Establishing the core and the mini-core sets with the lowest level of redundancy that represent

349 the maximum potential genetic diversity from the total collection facilitates the identification of

350 suitable variations for GWAS and MAB (Brown, 1989; Ndjiondjop et al., 2017). Core

351 collections could be evaluated based on genetic markers or phenotypic traits, including pairwise

352 distances (Hamon et al., 2003; Franco et al., 2005; Leroy et al., 2014) and allelic

353   richness/diversity (Beukelaer et al., 2012; Beukelaer et al 2018). Selection of the most suitable

354   evaluation method depends upon the purpose of core collections (Odong et al., 2013). Preserving

355   most alleles is an ideal way to conserve germplasm, while the approaches based on distance

356   majorly hammer the retention of most combinations of alleles in specific genotypes, which are

357   suitable for GWAS and MAB (Campoy et al., 2016; Ndjiondjop et al., 2017). Thus, we used the

358   distance-based methods applied on GBS data to propose the first core collection and the mini-

359   core collection from the tea origin center, the Guizhou plateau, and accounting for wild trees,

360   ancient landraces, modern landraces and cultivars in this study. Core-1 and Core-2 collections

361   and one mini-core collection were created based on Sphericity index and the length of pruned

362   values, which contain 72%, 48% and 36% of the total number of accessions. Ho, PIC, MAF,

363   allele and genotype frequencies in Core-1 and Core-2 sets were almost the same as for the whole

364   population, which suggested that both core-1 and core-2 can well represent the 415 accessions

365   for further study. The genetic diversity parameters, genetic distances , allele information and

366   retained SNPs of core-1 and core-2 revealed that the core-2 of 198 accessions was equally

367   appropriate for representing the whole population as core-1 of 299 accessions   Therefore, the

368   core-2 of 198 accessions was chosen as the appropriate set, considering the costs of future

369   research. The mini-core set selected in this study is easy to manage for phenotypic and

370   physiological evaluation in the field and under controlled conditions when selecting parent lines

371   for improving the traits of interest and identifying genes associated with these traits using GWAS.

372   **Genome-wide association studies**

373   Based on morphology, tea plants can be divided into two subgroups. Small-leaf shrubs that are

374   cold tolerant constitute group one, whereas group two consists of large-leaf arbor trees that are

375   less resistant to cold (Yao et al., 2011). Leaf size is associated to the fitness and temperature-

376   adaptation history of tea varieties (Tan et al., 2016). Leaf size may also relate to tea production

377   and the shape of dry tea. The mapping population used in this study included two sub groups

378   based on leaf size. Therefore, we observed a high level of segregation in all four leaves size-

379   related traits. Quantitative leaf traits are influenced by the environment (Tan et al., 2016; Baker

380   et al., 2016). In this study, the same trait was significantly correlated between two different

381   environments, suggesting that there are little false positive in the significantly associated SNPs.

382   MLM was employed for association mapping and two models (MLMQ+K and MLMP+K)

383   to control false positive caused by population structure. Leaf size is associated with local

384   adaptation, and it can actually decrease the false rate by removing accessions with extreme

385   genetic diversity and phenotypic expression (Pace et al., 2015). We did not test the GLM model

386   but used the combination of MLMQ+K and MLMP+K models, which reduce the rate of false

387   positives for detecting SNPs, and are proven to be more successful than using each strategy alone

388   (Song et al., 2019). Q denotes the Q-matrix produced from the population structure (k = 3) and P

389   represents P-matrix from the top three PCs, that are used as covariates in GWAS. K refers to

390   kinship matrix and is used to determine the correlations between individuals. Both the Q+K and

391   P+K matrices fit the MLM to control spurious associations resulted from relatedness and

392   population structure, separately (Zhao et al., 2019).

393   In previous studies, tea germplasm showed significant variations for leaf size (Yao et al.,

394   2011). However, to our knowledge, few genetic loci influencing this trait have been identified.

395   Tan et al. (2016) reported one major QTL (*qLSI13*) for leaf size in tea plant. In this study, nine

396   SNPs significantly associated with leaf traits were identified using the P+K MLM. Six, three,

397   two and one SNPs were significantly associated with MAL, MLSI, MLL and MLW, respectively.

398   SNP PG_1:32346865 was significantly associated with three traits, MLL, MLW and MLA, all

399   were significantly positively correlated ($P < 0.0001$),   Which was consistent with the results of

400   significant positive correlation among the three characters (Table. 4). Three SNPs significantly

401   associated with MLSI were carrying with only 1 minor allele, which need to keep the work in

402   exploring the truth in next step study.

403   Our study provides evidence that leaf size is affected by major effect genes, which was

404   consistent with previous reports (Tan et al., 2016). These results lay a foundation for future study

405   and will inform leaf size-related gene mining and MAB in tea breeding.

406   **Conclusions**

407   This study illustrates that GBS is efficient for analyzing the genetic purity, distance and

408   relationship of tea germplasm, and for creating the core sets. The core and mini-core collections

409   account for approximately 47% and 37% of the entire collection, respectively, and contains 93-

410   95% of all SNPs and almost all genotype frequencies and alleles that were observed in the entire

411   tea collection. These core sets are highly valuable for identifying favorable alleles and selecting

412   parent lines to improve agronomically important traits in tea varieties. Our study identified SNP

413   markers associated with leaf size. These SNPs can be employed in MAB for tea improvement.

414   **Figure and Table legends**

415   **Fig. 1** Frequency distribution categories of pairwise genetic distance of 415 tea accessions (red

416   fill); a core set of 299 accessions (orange fill); a core set of 198 accessions (green fill), and a

417   mini-core set of 148 accessions (purple fill), based on 30,282 polymorphic SNPs.

418   **Fig. 2** Neighbor joining tree for (A) 415 accessions (G1 = Group 1, pink; G2 = Group 2, red;G3=

419   Group 3, light green; G4 = Group 4, blue; G5 = Group 5, brown ; G6 = Group 6 , orange red; G7

420   = Group 7 , orange yellow; G8 = Group 8, light blue; G9 = Group 9 ,yellow; G10 = Group 10,

421   light purple; G11= Group 11, dark green; G12 = Group12, dark blue; G13 = Group 13, black;

422   G14 = Group 14, purple), (B) 299 accessions selected for a core set, (C) 198 accessions selected

423   for a core set and (D) 148 accessions selected for a mini-core set based on 30,282 polymorphic

424   SNPs.

425   **Fig. 3** Summary of accessions elected to conform core-1, core-2 and mini-core collection

426   compared with the whole tea accessions by (A) groups predicted based on cluster analysis (G1 =

427   Group 1,G2 = Group 2, G3= Group 3, G4 = Group 4, G5 = Group 5, G6 = Group 6 , G7 = Group

428   7 , G8 = Group 8, G9 = Group 9 , G10 = Group 10, G11= Group 11, G12 = Group 12, G13 =

429   Group 13, G14 = Group 14) ; (B) summary of accessions selected to account for core-1, core-2

430   and mini-core collection compared with the whole tea accessions in each group by cultivation

431   status (AL, Ancient landraces; ML, Modern landraces; WT, Wild tree) , by Growth Habits (CT,

432   cultivation type; WT, wild type)   and Species of origin (S1, *C.sinensis*; S2, *C.tachangensis*; S3,

433   *C.remotiserrata*; S4, *C.taliensis*).

434   **Table 1.** Genetic differentiation of inferred populations of tea accessions from Guizhou

435   Province.

436   **Table 2.** The number of whole accessions and core collections distributed in 14 groups.

437   **Table 3.** Trait statistics collected for mature leaf length (MLL), mature leaf width (MLW),

438   mature leaf shape index (MLSI) and mature leaf area (MLA).

439   **Table 4**. The correlation coefficient between two environment and among four traits.

440   **Table 5.** SNPs significantly associated with mature leaf length (MLL), mature leaf width
441   (MLW), mature leaf shape index (MLSI) and mature leaf area (MLA) detected by GWAS of 415
442   accessions

443   **Supplemental Figure and Table legends**

444   **Fig. S1** Geographic distribution of tea accessions analyzed in this study according to the

445   collection.

446   **Fig. S2** Frequency distribution of observed heterozygosity for 415 tea accessions based on

447   30,282 polymorphic SNPs.

448   **Fig. S3** Summary of (A) allele frequencies and (B) genotype frequencies of the entire tea

449   accessions (N = 415) based on 30,282 polymorphic SNPs as compared with a core set of 299

450   accessions, a core set of 198 accessions, and a mini-core set of 148 accessions. Nucleotide codes

451   are as follows: A, adenine; C, cytosine; G, guanine; T, thymine; Y, C or T; R, A or G; K, G or T;

452   M, A or C; W, A or T; S, G or C (http://www.bioinformatics.org/sms/iupac.html; accessed in

453   Sept. 2017).

454   **Fig. S4** Sphericity index and the length of pruned values for the selected core collection

455   individuals and mini-core collection for 415 tea accessions using 30,282 polymorphic SNPs.

456   **Fig. S5** Q-Q plot of P-values analyzed using MLMQ+K and MLMP+K models.   MLL: mature

457   leaf Length, MLW: mature leaf width, MLA: mature leaf area, MLSI: mature leaf shape index.

458   **Fig. S6** Manhattan and quantile–quantile plots generated from GWAS results for leaf traits.

459   **Table. S1** Information of 415 tea accessions used in this study, including the accession /clone

460   /collection, the accession name, the zone, the cultivation status, growth habits, the species, the

461   observed heterozygosity, the core collection (N=299), the core collection (N=198) and the mini-

462   core collection (N=148).

463   **Table. S2** Identity-by-state (IBS)-based genetic distance matrix for 415 accessions using 30,282

464   polymorphic SNPs.

465 **Table. S3** Identity-by-state (IBS)-based genetic distance matrix for 299 accessions using 30,282

466 polymorphic SNPs.

467 **Table. S4** Identity-by-state (IBS)-based genetic distance matrix for 198 accessions using 30,282

468 polymorphic SNPs.

469 **Table. S5** Identity-by-state (IBS)-based genetic distance matrix for 148 accessions using 30,282

470 polymorphic SNPs.

471 **Table. S6** The information of scaffold and SNP position in the genome Corresponding to the

472 name of pseudo-group (PG).

473 **Table. S7** SNPs significantly associated with mature leaf length (MLL), mature leaf width

474 (MLW), mature leaf shape index (MLSI) and mature leaf area (MLA) detected by genome-wide

475 association study (GWAS) of core-1, core-2 and mini-core.

476 **Availability of data and material**

477 The 415 tea plant materials used in this study are planted in our germplasm resource nursery, and

478 the requests for materials should be addressed to Suzhen Niu (email: niusuzhen@163.com).

479 The raw sequence data reported in this study have been deposited in the Genome Sequence

480 Archive (Wang et al., 2017) in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese

481 Academy of Sciences, under accession number CRA001438 that is publicly accessible at

482 http://bigd.big.ac.cn/gsa. Genotyping of 30,282 SNPs based on GBS in 415 tea accessions in this

483 study were deposited into the figshare website: https://doi.org/10.6084/m9.figshare.8019737.

484 **Competing interests**

485 The authors declare that they have no competing interests and consent for publication.

486 **Funding**

487 This work was supported by the National Science Foundation of China (31560222), the Science

488 and Technology Plan Project of Guizhou province (20172558, 20172557, 20175788, 20191404),

489 the USDA-NIFA SCRI grant (2017-51181-26834), the Guizhou Top Level Innovation Talents

490 Cultivation Project (20164003), the Talent Base for Germplasm Resources Utilization and

491 Innovation of Characteristic Plant in Guizhou (RCJD2018-14) and the Genetically Breeding

492 Major Project of the Ministry of Agriculture of China (2016ZX08010-003).

Authors' contributions

S.Z.N, D.G.Z and Z.W.C conceived and supervised the study. H.K led writing and reviewed the manuscript. Q.F.S analyzed and interpreted GWAS. D.H.Q and J.C performed the DNA extraction and filtered the genotyping data. Y.W and T.Y.Z analyzed and interpreted Core Collection. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank tea office of Guiding, Huishui, Liping, Renhuai, Sandu, Wuchuan, hishui, Daozhen, Dejiang,Duyun, Guian, Jinsha, Liuzhi, Nayong, Pu an, Puding ,Qinglong, Qixingguan, Sandu, Shiqian, Shuicheng, Tongzi, Wuchuan, Xingren, Xingyi, Xishui, Yanhe, Yinjiang, Yuqing, Zhenfeng, zheng'an for their help in teas collection. We thank College of tea science of Guizhou University and Department of Horticultural Sciences of Texas A&M University for providing research facilities and computing facilities.

References

Arora, S., Cheema, J., Poland, J., Uauy, C., & Chhuneja, P. (2019). Genome-wide association mapping of grain micronutrients concentration in *aegilops tauschii*. *Frontiers in Plant Science*, 10.

Babu, B.K., Sood, S., Chandrashekara, C., Pattanayak, A., and Kant, L. (2018). Mapping quantitative trait loci for important agronomic traits in finger millet (*Eleusine coracana*) mini core collection with genomic and genic SSR markers. *Journal of Plant Biochemistry & Biotechnology* (1), 1-14.

Bac-Molenaar, J A., Fradin, E.F., Becker, F.M., Rienstra, J.A., Schoot, J., Vreugdenhil, D., Keurentjes J.J.B. (2015). Genome-wide association mapping of fertility reduction upon heat stress reveals Ddevelopmental stage-specific QTLs in *Arabidopsis thaliana*. *The Plant Cell, 27(7)*: 1857-1874.

Brown, AHD. (1989). Core collections-a practical approach togenetic-resources management. *Genome, 31*(2):818-824.

Baker, R.L., Fung, L.W., Brock, M.T., Markelz, R.J.C., Covington, M.F., Devisetty,

520   U.K.,Edwards,   C.E., Maloof, J., Welch, S., Weinig, C. (2015). Modeling development and

521   quantitative trait mapping reveal independent genetic modules for leaf size and shape. *New*

522   *Phytologist, 208* (1)**,** 257-268.

523   Bali, S., Mamgain, A., Raina, S.N., Yadava, S.K., Bhat, V., Das, S., Pradhan, A.K., Goel, S.

524   (2015). Construction of a genetic linkage map and mapping of drought tolerance trait in

525   Indian beveragial tea. *Molecular Breeding, 35* (5)**,** 1-20.

526   Bartholomé, J., Bink, M.C., Van, H.J., Chancerel, E., Boury, C., Lesur, I., Isik, F., Bouffier, L.,

527   Plomion, C. (2016). Correction: Linkage and association mapping for two major traits used

528   in the maritime pine breeding program: Height growth and stem straightness. *Plos One ,11*

529   (11).

530   Beukelaer, H. D., Smýkal, P., Davenport, G. F., and Fack, V. (2012). Core Hunter II: fast core

531   subset selection based on multiple genetic diversity measures using Mixed Replica search.

532   *BMC Bioinformatics*,13:312.

533   Beukelaer, H. D., Davenport, G. F., Veetle, F. (2018). Core Hunter 3: flexible core subset

534   selection. *BMC Bioinformatics*, *19*:203. Bhattarai, U., and Subudhi, P.K. (2018).

535   Identification of drought responsive QTLs during vegetative growth stage of rice using a

536   saturated GBS-based SNP linkage map. *Euphytica, 214* (2), 38.

537   Botstein, D., White, R, L., Skolnick, M., Davis, R.W. (2008). Construction of a genetic linkage

538   map in man using restriction fragment length polymorphisms. *American journal of human*

539   *genetics, 32*(3), 314-331.

540   Bouchet, S., Olatoye, M.O., Marla, S.R., Perumal, R., Tesso, T., Yu, J., Tuinstra, M., Morris,

541   G.P.   (2017). Increased power to dissect adaptive traits in global sorghum diversity using a

542   nested association mapping population. *Genetics ,206* (2)**,** 573-585.

543   Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S.

544   (2007). TASSEL: software for association mapping of complex traits in diverse samples.

545   *Bioinformatics*, *23* (19)**,** 2633-2635.

546   Brown, A. H. D. (1989). Core collections: a practical approach to genetic resources management.

547   *Genome*, 31, 818–824. Brown, A.H.D., Hintum, V., Th., J.L., Morales, and E., A.V. (2000).

548   Core collections of plant genetic resources. *IPGRI Technical Bulletin No.3. International*

549   *Plant Genetic Resources Institute, Rome, Italy, 48 pp* (96).

550 Buckler, E.S., and Thornsberry, J.M. (2002). Plant molecular diversity and applications to

551   genomics. *Current Opinion in Plant Biology 5* (2)**,** 107-111.

552 Campoy, J.A., Lerigoleurbalsemin, E., Christmann, H., Beauvieux, R., Girollet, N., Querogarcía,

553   J., Dirlewanger, E., Barreneche, T. (2016). Genetic diversity, linkage disequilibrium,

554   population structure and construction of a core collection of *Prunus avium* L. landraces and

555   bred cultivars. *BMC Plant Biology ,16* (1)**,** 49.

556 Chen, L., Yu, F.L., and Tong, Q.Q. (2000). Discussions on phylogenetic classification and

557   evolution of *Sect.Thea. Journal of Tea Science, 20* (2): 89-94.

558 Chen, L., Apostolides, Z., and Chen, Z.M. (2012). Global tea breeding: Achievements,

559   challenges and perspectives. Springer-Zhejiang University Press

560 Chen, W., Hou, L., Zhang, Z., Pang, X., and Li, Y. (2017). Genetic diversity, population

561   structure, and linkage disequilibrium of a core collection of *Ziziphus jujuba* assessed with

562   genome-wide SNPs developed by genotyping-by-sequencing and SSR markers. *Front Plant*

563   *Sci ,8*, 575.

564 Claire, B., Punna, R., Sophie, B., Jacques, C., Monique, D., Laetitia, G., Jean-Louis, N., Jean-

565   Fran?Ois, R., Ronan, R., Yu, L. (2013). Massive sorghum collection genotyped with SSR

566   markers to enhance use of global genetic resources. *Plos One ,8* (4).

567 Coan, M.M.D., Senhorinho, H.J.C., Pinto, R.J.B., Scapim, C.A., and Warburton, M.L. (2018).

568   Genome-wide association study of resistance to ear rot by in a tropical field maize and

569   popcorn core collection. *Crop Science, 58* (2), 564-.

570 Cunff, L.L., Fournier-Level, A., Laucou, V., Vezzulli, S., Lacombe, T., Adam-Blondon, A.F.,

571   Boursiquot, J.M., This, P. (2008). Construction of nested genetic core collections to

572   optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. sativa. *BMC Plant*

573   *Biology, 8* (1) , 31.

574   Desgroux, A., Baudais, V.N., Aubert, V., Le, R.G., De, L.H., Miteul, H., Aubert, G., Boutet, G.,
575   Duc, G., Baranger, A. (2017). Comparative genome-wide-association mapping identifies
576   common loci controlling root system architecture and resistance to aphanomyces euteiches
577   in pea. *Frontiers in Plant Science, 8*, 2195.

578   Egbadzor, K.F., Ofori, K., Yeboah, M., Aboagye, L.M., Opoku-Agyeman, M.O., Danquah, E.Y.,
579   Offei, S.K. (2014). Diversity in 113 cowpea [*Vigna unguiculata* (L) Walp] accessions
580   assessed with 458 SNP markers. *Springer plus ,3* (1), 541.

581   Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E.
582   (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity
583   species. *PloS One ,6* (5).

584   Eltaher, S., Sallam, A., Belamkar, V., Emara, H., Nower, A., Salem, K., Poland, J., and
585   Baenziger, P.S. (2018). Genetic diversity and population structure of F Nebraska winter
586   wheat genotypes using genotyping-by-sequencing. *Front Genet, 9*, 76.

587   Franco, J., Crossa, J., Taba, S., and Shands, H. (2005). A sampling strategy for conserving
588   genetic diversity when forming core subsets. *Crop Science ,45* (3)**,** 1035-1044.

589   Frankel, OH. (1984). Genetic perspectives of germplasm conservation. In: Arber, W.K.,
590   Llimensee, K., Peacock, W.J., Starlinger, P.   (eds) Genetic manipulation: impact on manand
591   society. Cambridge University Press, Cambridge, England, 161-170.

592   Gao, X. Y., Joshua, S., and Martin, E.R. (2010). A multiple testing correction method for genetic
593   association studies using correlated single nucleotide polymorphisms. *Genetic*
594   *Epidemiology ,32* (4)**,** 361-369.

595   Hackett, C.A., Milne, L., Smith, K., Hedley, P., Morris, J., Simpson, C.G., Preedy, K., Graham, J.
596   (2018). Enhancement of Glen Moy x Latham raspberry linkage map using GBS to further
597   understand control of developmental processes leading to fruit ripening. *BMC Genetics ,19*.

598   Hamon, P., Seguin, M., Perrier, X., and Glaszmann, J.C. (2003). Genetic diversity of cultivated
599   tropical plants. *Diversite Genetique Des Plantes Tropicales Cultivees*.

600   Hayat, K., Iqbal, H., Malik, U., Bilal, U., and Mushtaq, S. (2015). Tea and its consumption:

601   benefits and risks. *Crit Rev Food Sci Nutr, 55* (7), 939-954.

602 Heena, A., Shivendra, K., Amar, K., Manu, A., Arun, J., & Shailendra, G. (2018). Association
603   mapping for important agronomic traits in safflower (*carthamus tinctorius* L.) core
604   collection using microsatellite markers. *Frontiers in Plant Science, 9*, 402-.

605 Hou, J., Zhou, Y.-F., Gao, L.-Y., Wang, Y.-L., Yang, L.-M., Zhu, H.-Y., Wang, J.-M., Zhao,
606   S.-J., Ma, C.-S., Sun, S.-R., Hu, J.-B. (2018). Dissecting the genetic architecture of
607   melon chilling tolerance at the seedling stage by association mapping and identification of
608   the elite alleles. *Frontiers in Plant Science ,9* (1577).

609 Huggins, T.D., Chen, M.-H., Fjellstrom, R.G., Jackson, A.K., McClung, A.M., and Edwards, J.D.
610   (2018). Association analysis of three diverse rice (Oryza sativa L.) germplasm collections
611   for loci regulating grain quality traits. *The Plant Genome*.
612   doi:10.3835/plantgenome2017.09.0085

613 Hussain, W., Baenziger, P.S., Belamkar, V., Guttieri, M.J., Venegas, J.P., Easterly, A., Sallam,
614   A., Poland, J. (2017). Genotyping-by-sequencing derived high-density linkage map and its
615   application to QTL mapping of flag leaf traits in bread wheat. *Sci Rep, 7* (1).

616 Iso-Touru, T., Sahana, G., Guldbrandtsen, B., Lund, M.S., and Vilkki, J. (2016). Genome-wide
617   association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome
618   sequence variants. *BMC Genetics, 17* (1), 55.

619 Jin J, Chai Y, Liu Y, Zhang J, Yao M, Chen L. (2018). Hongyacha, a naturally caffeine-free tea
620   plant from Fujian, China. *Journal of agricultural and food chemistry*, *100*:10.

621 Lassois, L., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Hibrandsaintoyant, L., Poncet, C.,
622   Lasserrezuber, P., Feugey, L., Durel, C.E. (2016). Genetic diversity, population structure,
623   parentage analysis, and construction of core collections in the French apple germplasm
624   based on SSR markers. *Plant Molecular Biology Reporter ,34* (4)**,** 827-844.

625 Li, D., Zhao, X., Han, Y., Li, W., and Xie, F. (2018). Genome-wide association mapping for
626   seed protein and oil contents using a large panel of soybean accessions. *Genomics***.** doi:
627   10.1016/j.ygeno.2018.01.004

628   Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with *BWA-MEM*.
629   arXiv:1303.3997.

630   Liang, Y.R., and Shi, M. (2015). Advances in tea plant genetics and breeding. *Journal Tea*
631   *Science ,35*:103-109.

632   Liu, S., Liu, H., Wu, A., Hou, Y., An, Y., and Wei, C. (2017). Construction of fingerprinting for
633   tea plant (Camellia sinensis) accessions using new genomic SSR markers. *Molecular*
634   *Breeding, 37* (8).

635   Thierry, L., Fabien, D.B., Hyacinthe, L., Pascal, M., Adrien, K., Rey Gastón, L.S., Philippe, C.
636   (2014). Developing core collections to optimize the management and the exploitation of
637   diversity of the coffee Coffea canephora. *Genetica, 142*, 185-199.

638   Luo, X., Ma, C., Yue, Y., Hu, k., Li, Y., Duan, Z., Wu, M.,Tu, J., Shen, J., Yi, B., Fu, T.(2015).
639   Unravelling the complex trait of harvest index in rapeseed (*Brassica napus* L.) with
640   association mapping. *BMC Genomics, 16*:379.

641   Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., Wu, L., Li, Z., Liu, Z., and Sun, G.
642   (2018). Resequencing a core collection of upland cotton identifies genomic variation and
643   loci influencing fiber quality and yield. *Nature Genetics, 50* (6).

644   Ma, J.Q., Jin, J.J., Yao, M.Z., Ma,C.L., Xu,Y.X., Hao, W.G., Chen, L. (2018). Quantitative trait
645   loci mapping for theobromine and caffeine contents in tea plant (*Camellia sinensis*). *Journal*
646   *of Agricultural and Food Chemistry, 66* (50), 13321–13327.

647   Ma, Q., Chen, C., Zeng, Z., Zou, Z., Li, H., Zhou, Q., Chen, X., Sun, K., Li, X. (2018).
648   Transcriptomic analysis between self- and cross-pollinated pistils of tea plants (*Camellia*
649   *sinensis*). *BMC Genomics, 19*:289.

650   Mckenna, A., Hanna, M.E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K.,
651   Altshuler, D., Gabriel, S., Daly, M., Depristo, M.A.   (2010). The genome analysis toolkit:
652   a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Res,*
653   *20*, 1297-1303.

654   Min, T. L. (1992). A revision of Camellia sect.thea. *Acta Botanica Yunnanica ,14* (2): 115-132.

655 Motilal, L.A., Zhang, D., Mischke, S., Meinhardt, L.W., Boccara, M., Fouet, O., Lanaud, C.,
656   Umaharan, P. (2016). Association mapping of seed and disease resistance traits in
657   *Theobroma cacao* L. *Planta*, *244* (6)**,** 1265-1276.

658 Muleta, K.T., Bulli, P., Zhang, Z., Chen, X., and Pumphrey, M. (2017). Unlocking diversity in
659   germplasm collections via genomic selection: a case study based on quantitative adult plant
660   resistance to stripe rust in spring wheat. *Plant Genome, 10* (3).

661 Naghma, K., and Hasan, M. (2013). Tea and health: studies in humans. *Current Pharmaceutical*
662   *Design, 19* (34).

663 Ndjiondjop, M., Semagn, K., Gouda, A., Kpeki, S., Dro Tia, D., Sow, M., Goungoulou, A., Sie,
664   M., Perrier, X., Ghesquiere, A. (2017). Genetic variation and population structure of and
665   development of a mini-core collection using DArTseq. *Front Plant Science, 8***,** 1748.

666 Niu, S.Z. (2014). Studies on genetic diversity and resistance of wild tea germplasm (Camellia
667   spp.) in Guizhou province. [doctoral thesis]: Guizhou university.

668 Niu, S., Song, Q,Koiwa,H., Qiao, D., Zhao, D., Chen, Z., Liu,X., Wen,X. (2019) . Genetic
669   diversity, linkage disequilibrium, and population structure analysis of the tea plant
670   (Camellia sinensis) from an origin center, Guizhou plateau, using genome wide SNPs
671   developed by genotyping-by sequencing. *BMC Plant Biology ,19*:328.

672 Odong, T.L., Jansen, J., Eeuwijk, F.A., Van, and Hintum, T.J.L., Van (2013). Quality of core
673   collections for effective utilisation of genetic resources review, discussion and interpretation.
674   *Theoretical & Applied Genetics, 126* (2)*,* 289-305.

675 Pace, J., Gardner, C., Romay, C., Ganapathysubramanian, B., and Lübberstedt, T. (2015).
676   Genome-wide association analysis of seedling root development in maize (Zea mays L.).
677   *BMC Genomics*, *16* (1) *,* 47.

678 Song, J. M., Arif, M., Zhang, M., Sze, S.H., Zhang, H.B. (2019). Phenotypic and molecular
679   dissection of grain quality using the USDA rice mini-core collection. *Food Chemistry, 284*
680   (30):312-322.

681 Suwarno, W.B., Pixley, K.V., Palacios-Rojas, N., Kaeppler, S.M., and Babu, R. (2015).

682   Genome-wide association analysis reveals new targets for carotenoid biofortification in

683   maize. *Theoretical & Applied Genetics, 128* (5)**,** 851-864.

684   Taniguchi, F., Kimura, K., Saba, T. (2014) Worldwide core collections of tea (*Camellia sinensis*)

685   based on SSR markers. *Tree Genetics & Genomes*, *10*, 1555-1565.

686   Tan, L.Q., Wang, L.Y., Xu, L.Y., Wu, L.Y., Peng, M., Zhang, C.C., Wei, K., Bai, P.-X., Li, H.-

687   L., Cheng, H., Qi, G. N. (2016). SSR-based genetic mapping and QTL analysis for timing

688   of spring bud flush, young shoot color, and mature leaf size in tea plant (*Camellia sinensis*).

689   *Tree Genetics & Genomes, 12* (3)**,** 52.

690   Taniguchi, I., McCloskey, A., Ohno, M. (2014). Chapter 18 - Analysis of RNA transport in

691   xenopus oocytes and mammalian cells. *Methods in Cell Biology, 122*: 395-413.

692   Wambulwa, M.C., Meegahakumbura, M.K., Kamunya, S., Muchugi, A., Moller, M., Liu, J., Xu,

693   J.C., Ranjitkar, S., Li, D.Z., Gao, L.M. (2016). Insights into the genetic relationships and

694   breeding patterns of the African tea germplasm based on nSSR markers and cpDNA

695   sequences. *Front Plant Science, 7,* 1244.

696   Wang, X.C., Liu, Z., Yao, M.Z., Ma, C.L., Chen, L., Yang, Y.J. (2009). Sampling strategy to

697   establish a primary core collection of Chinese tea germplasms. *Journal of Tea Science*, *29*

698   (2),159-167.

699   Wang, XC., Chen, L., Yang, YJ (2011). Establishment of core collection for Chinese tea

700   germplasm based on cultivated region grouping and phenotypic data. *Front. Agric. China,*

701   *5*(3): 344-350.

702   Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., Tang, B., Dong, L., Ding, N.,

703   Zhang, Q., Bai, Z., Dong, X., Chen, H., Sun, M., Zhai, S., Sun, Y., Yu, L.,

704   Lan, L., Xiao, J., Fang, X., Lei, H., Zhang, Z., Zhao, W. (2017). GSA: Genome

705   sequence archive*. *Genomics, Proteomics & Bioinformatics, 15* (1), 14-18.

706   Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., Xia, E., Lu, Y., Tai, Y., She, G.,

707   Sun, J., Cao, H., Tong, W., Gao, Q., Li, Y., Deng, W., Jiang, X., Wang, W.,

708   Chen, Q., Zhang, S., Li, H., Wu, J., Wang, P., Li, P., Shi, C., Zheng, F., Jian,

709   J., Huang, B., Shan, D., Shi, M., Fang, C., Yue, Y., Li, F., Li, D., Wei, S.,

710   Han, B., Jiang, C., Yin, Y., Xia, T., Zhang, Z., Bennetzen, J.L., Zhao, S., Wan,

711   X. (2018). Draft genome sequence of Camellia sinensis var. sinensis provides insights into

712   the evolution of the tea genome and tea quality. *Proc Natl Acad Sci U S A, 115* (18)**,** E4151-

713   E4158.

714 Xia, E.H., Zhang, H.B., Sheng, J., Li, K., Zhang, Q.J., Kim, C., Zhang, Y., Liu, Y., Zhu, T.,

715   Li, W., Huang, H., Tong, Y., Nan, H., Shi, C., Shi, C., Jiang, J.J., Mao, S.Y.,

716   Jiao, J.Y., Zhang, D., Zhao, Y., Zhao, Y.J., Zhang, L.P., Liu, Y.L., Liu, B.Y.,

717   Yu, Y., Shao, S.F., Ni, D.J., Eichler, E.E., Gao, L.Z. (2017). The tea tree genome

718   provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol*

719   *Plant, 10* (6), 866-877.

720 Yang, Z., Chen, Z., Peng, Z., Yu, Y., Liao, M., and Wei, S. (2017). Development of a high-

721   density linkage map and mapping of the three-pistil gene (Pis1) in wheat using GBS

722   markers. *BMC Genomics, 18* (1), 567.

723 Yao, M.-Z., Ma, C.-L., Qiao, T.-T., Jin, J.-Q., and Chen, L. (2012). Diversity distribution and

724   population structure of tea germplasms in China revealed by EST-SSR markers. *Tree*

725   *Genetics & Genomes, 8* (1), 205-220.

726 Zhang, D.Z., Rabia, B.P., Liu, J.J., Gong, X.W., Liang, J.B., Liu, M.X., Lu, P., Gao, X. L., Feng,

727   B.L. (2018). Morphological diversity and correlation analysis of phenotypes and quality

728   traits of proso millet (*Panicum miliaceum* L.) core collections. *Journal of Integrative*

729   *Agriculture, 17* (0), 60345-60347.

730 Zhao, X., Dong, H., Chang,H., Zhao,J., Teng,W., Qiu, L., Li, W., Han,Y (2019). Genome wide

731   association mapping and candidate gene analysis for hundred seed weight in soybean

732   [*Glycine max* (L.) Merrill]. *BMC Genomics,20*:648.

733

Figure 1

Frequency distribution categories of pairwise genetic distance of 415 tea accessions

Frequency distribution categories of pairwise genetic distance of 415 tea accessions (red fill); a core set of 299 accessions (orange fill); a core set of 198 accessions (green fill), and a mini-core set of 148 accessions (purple fill), based on 30,282 polymorphic SNPs.

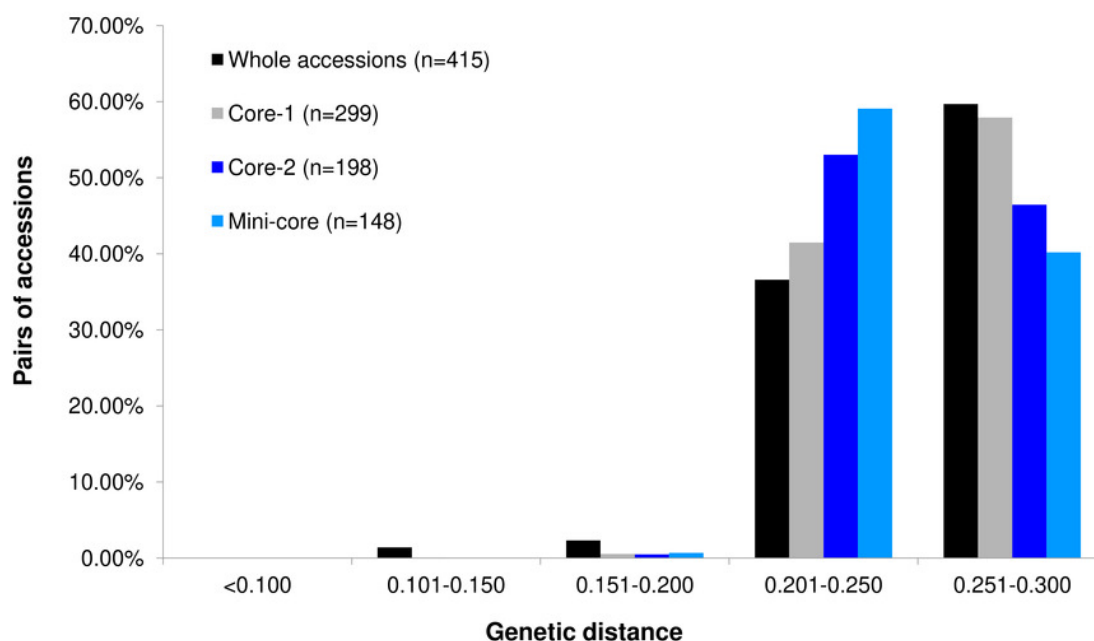



Figure 2

Neighbor joining tree

Neighbor joining tree for (A) 415 accessions (G1 = Group 1, pink; G2 = Group 2, red;G3= Group 3, light green; G4 = Group 4, blue; G5 = Group 5, brown ; G6 = Group 6 , orange red; G7 = Group 7 , orange yellow; G8 = Group 8, light blue; G9 = Group 9 ,yellow; G10 = Group 10, light purple; G11= Group 11, dark green; G12 = Group12, dark blue; G13 = Group 13, black; G14 = Group 14, purple), (B) 299 accessions selected for a core set, (C) 198 accessions selected for a core set and (D) 148 accessions selected for a mini-core set based on 30282 polymorphic SNPs.

A

G2
G3
G10
G14
G 9
G13
G8
G6
G7
G11
G12
G5
G1
G4

B

G3
G2
G10
G14
G9
G14
G8
G13
G1
G7 G12 G6
G11
G5
G4

C

G2
G3
G4
G14 G10
G9
G8
G1
G1
G1
G4
G4
G1
G13
G5
G6 G12
G11 G7 G6
G11

D

G2
G3
G8 G10
G14
G9
G4
G5
G1
G12
G7 G6 G11

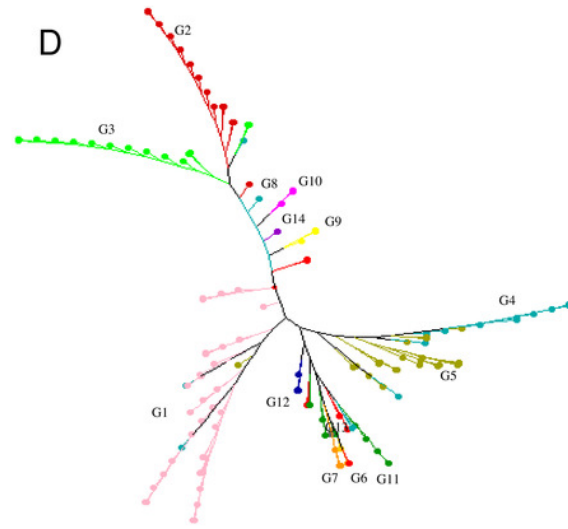
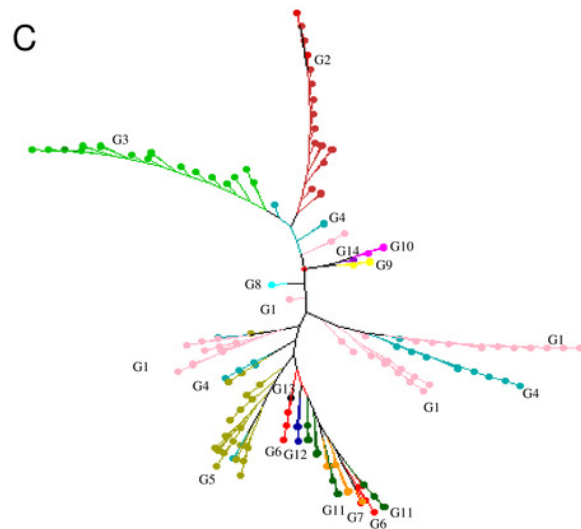
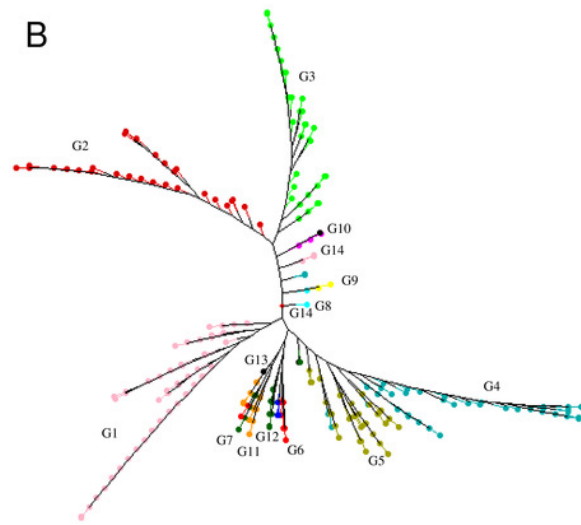
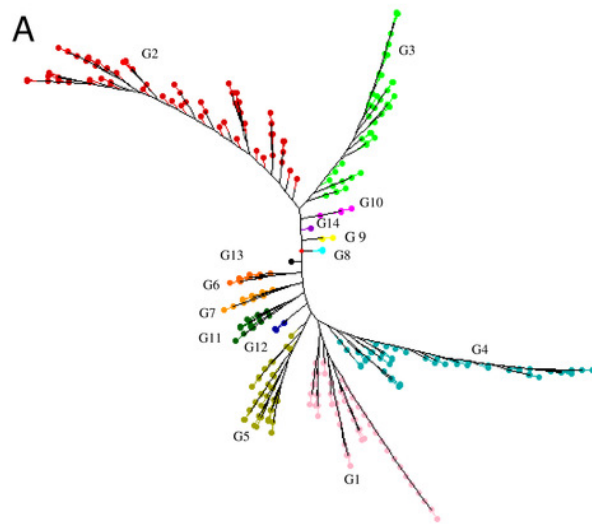
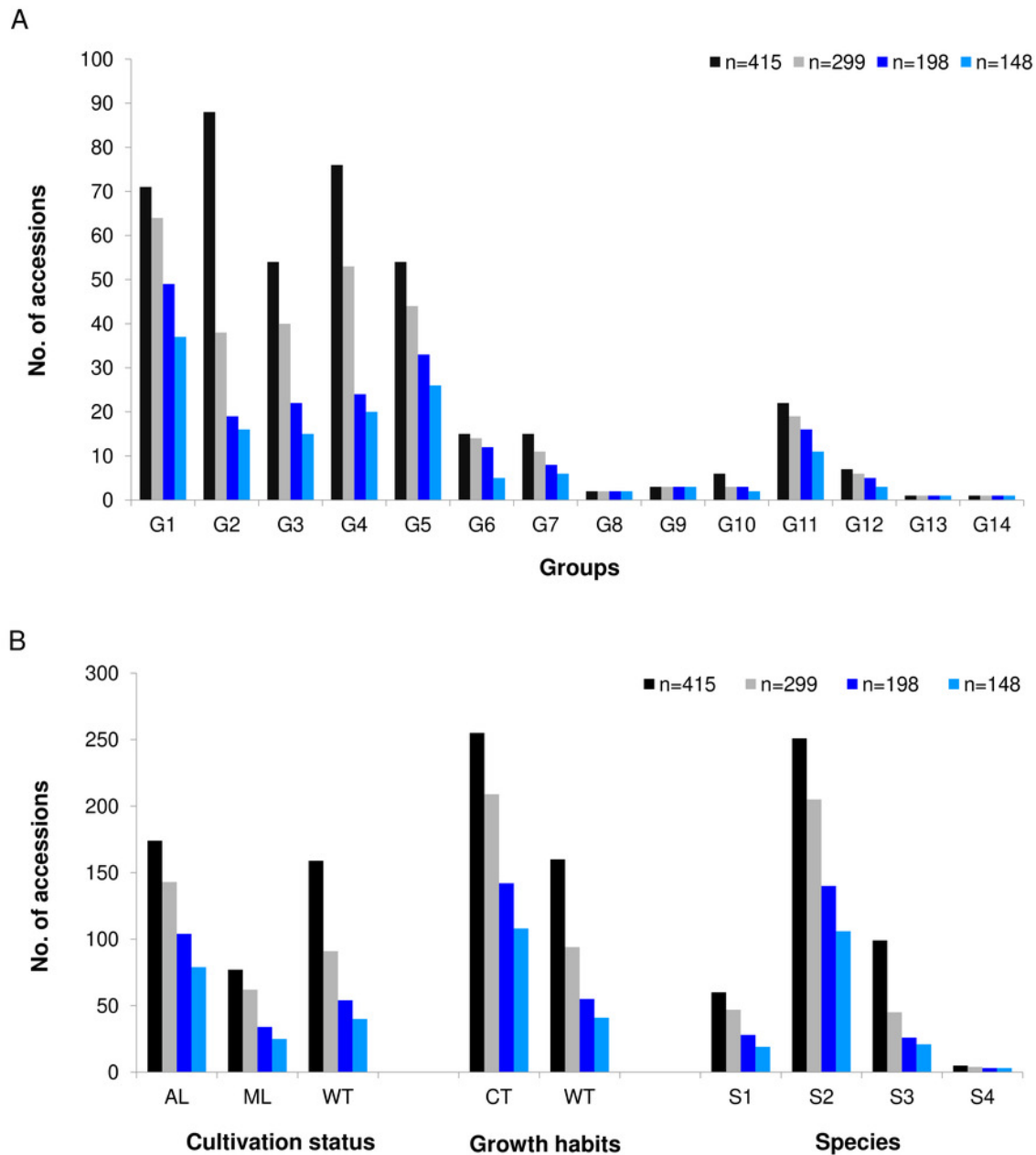



Figure 3

Summary of accessions elected to conform core-1, core-2 and mini-core

Summary of accessions elected to conform core-1, core-2 and mini-core collection compared with the whole tea accessions by (A) groups predicted based on cluster analysis (G1 = Group 1,G2 = Group 2, G3= Group 3, G4 = Group 4, G5 = Group 5, G6 = Group 6 , G7 = Group 7 , G8 = Group 8, G9 = Group 9 , G10 = Group 10, G11= Group 11, G12 = Group 12, G13 = Group 13, G14 = Group 14) ; (B) summary of accessions selected to account for core-1, core-2 and mini-core collection compared with the whole tea accessions in each group by cultivation status (AL, Ancient landraces; ML, Modern landraces; WT, Wild tree) , by Growth Habits (CT, cultivation type; WT, wild type) and Species of origin (S1, *C.sinensis*; S2, *C.tachangensis*; S3, *C.remotiserrata*; S4, *C.taliensis*).

Genetic differentiation of inferred populations of tea accessions from Guizhou Province.

Genetic differentiation of inferred populations of tea accessions from Guizhou Province.

1   **Table 1** Genetic differentiation of core and mini-core sets of tea plant in Guizhou Province.

Group	S	ASC	Ho	PIC	MAF	GDR	AGD
Whole Set	415	26,810	0.215±0.006a	0.359±0.006a	0.171±0.002a	0.088−0.298	0.249
Core-1	299	26,592	0.223±0.005a	0.360±0.008a	0.172±0.001a	0.110−0.298	0.252
Core-2	198	24,917	0.206±0.006a	0.363±0.006a	0.170±0.001ab	0.115−0.298	0.250
Mini-Core	148	23,731	0.193±0.006b	0.355±0.005a	0.168±0.001b	0.115−0.292	0.248

2   Note: S, Sample size; ASC, Average Site Count; Ho, observed heterozygosity; PIC, polymorphism
3   information content; MAF, Minor Allele Frequency; GDR, Genetic distance range; AGD, average genetic
4   distance. The different letters indicate a significant difference in a column at *p = 0.05* levels by T-test.

The number of whole accessions and core collections distributed in 14 groups.

The number of whole accessions and core collections distributed in 14 groups.

1   **Table 2** The number of whole accessions and core collections distributed in 14 groups

Group	Number of accessions of whole set	Number of accessions retained in the 299 core set	Number of accessions retained in the 198 core set	Number of accessions retained in the 148 mini-core set
Group1	71	64	49	37
Group2	88	38	19	16
Group3	55	40	22	15
Group4	75	53	24	20
Group5	54	44	33	26
Group6	15	14	12	5
Group7	15	11	8	6
Group8	2	2	2	2
Group9	3	3	3	3
Group10	6	3	3	2
Group11	22	19	16	11
Group12	7	6	5	3
Group13	1	1	1	1
Group14	1	1	1	1
Total	415	299	198	148

Trait statistics collected for mature leaf length (MLL), mature leaf width (MLW), mature leaf shape index (MLSI) and mature leaf area (MLA).

Trait statistics collected for mature leaf length (MLL), mature leaf width (MLW), mature leaf shape index (MLSI) and mature leaf area (MLA).

1 Table 3 Trait statistics collected for mature leaf length (MLL), mature leaf width (MLW), mature leaf shape
2 index (MLSI) and mature leaf area (MLA)

Trait	Mean	Std.dev	Minimum	Maximum	Skewness(Std.Err)	Kurtosis(Std.Err)
MLL	8.655	2.251	3.886	15.200	0.497(0.122)	0.002(0.243)
MLW	3.748	0.933	1.986	6.830	0.687(0.122)	0.415(0.244)
MLA	25.700	12.929	6.570	75.566	1.223(0.122)	1.582(0.244)
MLI	2.326	0.316	0.830	3.457	0.043(0.122)	1.949(0.244)

4
5
6
7
8
9
10
11
12
13
14

The correlation coefficient between two environment and among four traits.

The correlation coefficient between two environment and among four traits.

2   Table 4 The correlation coefficient between two environment and among four traits

Trait/ Environment	MLL	MLW	MLA	M1SI
MLL	0.896**			
MLW	0.857**	0.960**		
MLA	0.953**	0.951**	0.905**	
MLSI	0.338**	−0.174**	0.081	0.806**

3   Note: The diagonal line is the correlation coefficient of two environments; * indicate a significant difference in a column at $p =$

4   0.05 levels, ** indicate a significant difference in a column at $p = 0.01$ levels

SNPs significantly associated with mature leaf length (MLL), mature leaf width (MLW), mature leaf shape index (MLSI) and mature leaf area (MLA) detected by GWAS of 415 accessions

SNPs significantly associated with mature leaf length (MLL), mature leaf width (MLW), mature leaf shape index (MLSI) and mature leaf area (MLA) detected by GWAS of 415 accessions

1 Table 5 SNPs significantly associated with mature leaf length (MLL), mature leaf width (MLW), mature leaf shape index (MLSI) and mature leaf area

2 (MLA) detected by GWAS of 415 accessions

3 Note: Pre-Group, Pseudo Group; MAF minor allele frequency.

Trait	SNP Marker	Pse-Group	Position(bp)	Scaffold	Major / minor allele	MAF	P-value	Major/minor allele effect	R-square	No. of accessions with homozygous major/minor allele	No. of accessions with heterozygous allele
MLL	PG_13:79474508	13	79474508	xfSc0000122	G/T	0.07	5.41E-07	-0.72/-3.36	3.77%	307/14	22
	PG_1:32346865	1	32346865	Sc0000000	T/A	0.05	1.03E-06	-2.94/1.09	3.23%	345/8	23
MLW	PG_1:32346865	1	32346865	Sc0000000	T/A	0.05	3.59E-07	-1.31/0.44	8.24%	344/8	23
MLA	PG_1:32346865	1	32346865	Sc0000000	T/A	0.05	1.95E-08	-18.23/8.25	4.50%	344/8	23
	PG_4:40545730	4	40545730	Sc0000365	G/A	0.07	8.80E-07	-1.47/14.72	3.79%	316/13	20
	PG_12:21050315	12	21050315	Sc0002172	C/T	0.06	9.47E-07	-4.12/11.61	3.58%	327/19	8
	PG_12:133979444	12	133979444	Sc0002452	G/A	0.06	1.07E-06	1.06/42.47	3.49%	322/2	36
	PG_16:74477057	16	74477057	Sc0004093	G/A	0.08	1.40E-06	4.46/31.33	3.81%	283/4	42
	PG_13:79474508	13	79474508	xfSc0000122	G/T	0.07	1.45E-06	-5.57/-18.76	3.68%	306/14	22
MLI	PG_13:82813415	1	82813415	Sc0000037	T/C	0.13	3.24E-07	-0.08/-1.25	4.01%	266/1	96
	PG_17:30871283	17	30871283	Sc0000011	A/G	0.05	5.84E-07	-1.49/0.09	4.33%	280/1	32
	PG_10:98891957	10	98891957	Sc0000046	G/A	0.06	1.40E-06	0.06/-1.50	3.83%	306/1	36