TITLE:

Genomic insight into the developmental history of southern highbush blueberry populations

AUTHOR(S):

Nishiyama, Soichiro; Fujikawa, Mao; Yamane, Hisayo; Shirasawa, Kenta; Babiker, Ebrahiem; Tao, Ryutaro

1 **Genomic insight into the developmental history of southern highbush blueberry populations**

2

3 Soichiro Nishiyama[1], Mao Fujikawa[1], Hisayo Yamane[1], Kenta Shirasawa[2], Ebrahiem Babiker[3],

4 Ryutaro Tao[1]

5 [1] Graduate School of Agriculture, Kyoto University, Sakyo-Ku, Kyoto 606-8502, Japan

6 [2] Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan

7 [3] U.S. Department of Agriculture, Agricultural Research Service, Thad Cochran Southern

8 Horticultural Laboratory, 810 Hwy 26W, Poplarville, MS 39470, USA

9

10 Corresponding author: Soichiro Nishiyama (nishiyama.soichiro.8e@kyoto-u.ac.jp)

11

12 **Abstract**

13 Interspecific hybridization is a common breeding approach for introducing novel traits and

14 genetic diversity to breeding populations. Southern highbush blueberry (SHB) is a blueberry cultivar

15 group that has been intensively bred over the last 60 years. Specifically, it was developed by multiple

16 interspecific crosses between northern highbush blueberry [NHB; *Vaccinium corymbosum* L. ($2n = 4x$

17 $= 48$)] and low-chill *Vaccinium* species to expand the geographic limits of highbush blueberry

18 production. In this study, we genotyped polyploid blueberries including 105 SHB, 17 NHB, and 10

19 rabbiteye blueberry (RE) (*V. virgatum* Aiton), from the accessions planted at Poplarville, Mississippi,

20 and accessions distributed in Japan, based on the double-digest restriction site-associated DNA

21 sequencing (ddRAD-seq). The genome-wide SNP data clearly indicated that RE cultivars were

22 genetically distinct from SHB and NHB cultivars, whereas NHB and SHB were genetically

23 indistinguishable. The population structure results appeared to reflect the differences in the allele

24 selection strategies breeders used for developing germplasm adapted to local climates. The genotype

25 data implied there are no or very few genomic segments that were commonly introgressed from low-

26 chill *Vaccinium* species to the SHB genome. PCA-based outlier detection analysis found a few loci

27 associated with a variable that could partially differentiate NHB and SHB. These SNP loci were

28 detected in Mb-scale haplotype blocks and may be close to the functional genes related to SHB

29 development. Collectively, the data generated in this study suggest a polygenic adaptation of SHB to

30 the southern climate and may be relevant for future population-scale genome-wide analyses of

31 blueberry.

32

33

34

35

36 **(5,601 words)**

Introduction

Interspecific hybridization is commonly used to increase the genetic diversity of crop species. Breeders have applied interspecific hybridization to improve crop tolerance to abiotic and biotic stresses, and enhance economically important traits (Tanksley and McCouch 1997; Nicotra et al. 2010; Ceccarelli et al. 2010). Successful outcomes of interspecific hybridization can be seen in blueberry breeding. Cultivated blueberries (*Vaccinium* spp.) have variation in ploidy level and include the tetraploid lowbush (*V. angustifolium* Aiton) and highbush (*V. corymbosum* L.) blueberries ($2n = 4x = 48$) and the hexaploid rabbiteye blueberry [RE; *V. virgatum* Aiton ($2n = 6x = 72$)] (Lyrene and Ballington 1986; Chavez and Lyrene 2009). Highbush cultivars are further separated into northern and southern types depending on their chilling requirement and winter hardiness. Multiple interspecific hybridizations were involved in establishing highbush blueberry cultivars, and currently cultivated highbush blueberry is one of the most successful outcomes of interspecific hybridization breeding.

Blueberry breeding has been extensive for only the last 100 years, and can be described to have a very short history, considering its long generation time as a shrub/tree crop. However, the fruits of wild edible *Vaccinium* species have been harvested and consumed by humans for thousands of years in North America (Moerman, 1998; Song and Hancock, 2011). Highbush blueberry breeding began in the early 20th century in the USA, and the interspecific hybridization of *Vaccinium* species has played a major role in the development of highbush blueberry cultivars. Southern highbush blueberry (SHB) is a cultivar group that is better adapted to warm climates than the original northern highbush blueberry (NHB). Additionally, SHB was derived from crosses between tetraploid NHB and low-chill *Vaccinium* species native to Florida, USA (including *V. darrowii* Camp.), and has helped expand the geographic limits of highbush blueberry production (Sharpe and Darrow, 1959). Because interspecific hybridizations have been commonly used for breeding purposes, all SHB cultivars are assumed to contain genomic segments from one or more of the other *Vaccinium* species, resulting in phenotypically diverse germplasm regarding specific traits (Brevis et al. 2008). Although the definition of SHB varies among researchers, we in this study referred SHB as a tetraploid highbush with at least one *Vaccinium* species native to the southeastern USA in its pedigree (Brevis et al, 2008).

Conventional blueberry breeding typically requires approximately 15 years for a new cultivar to be released and approximately 8 years for good germplasm to be developed (Hancock et al. 2008; Ferrão et al. 2018). The application of molecular breeding technologies may accelerate the improvement of polyploid blueberry cultivars. Polyploidy is common in plant species. Polyploidization events often resulted in phenotypic diversification and the appearance of elite phenotypes, including increased size, possibly because of gene duplications and redundancies (Comai 2005). Several recent investigations effectively correlated phenotypic variations with genome-wide molecular markers in polyploid blueberry populations (Ferrão et al. 2018; Cappai et al. 2018; Campa and Ferreira 2018; de Bem Oliveira et al. 2019; Benevenuto et al. 2019). Despite years of research on

73   developing low-chill SHB cultivars and elucidating the genotype–phenotype relationships, little is
74   known about the population structure and genomic evolution of cultivated blueberry. Clarifying the
75   genetic diversity and population structure of blueberry may generate fundamental information relevant
76   for association studies and useful for efficiently selecting suitable parents for hybridizations.

77   The objectives of this study were to (1) characterize the population structure of blueberries;
78   (2) characterize the linkage disequilibrium (LD) in the tetraploid SHB population; and (3) correlate
79   various genotype patterns (i.e. allele frequency of the different subpopulations) among blueberry
80   cultivar groups with physical genomic positions using the double-digest restriction site-associated
81   DNA sequencing (ddRAD-seq) approach and a recently developed chromosome-scale tetraploid
82   blueberry reference genome (Colle et al. 2019). We herein discuss the study results in terms of the
83   SHB developmental history and highlight the general genomic features of blueberry, especially of the
84   SHB cultivar group.

85

86   **Materials and Methods**

87   **Plant materials and genotyping**

88   Leaves were collected from 105 SHB accessions, 17 NHB accessions, 10 RE accessions, 3
89   half highbush (HH) accessions, and 2 complex hybrid (CH) accessions from the USDA-ARS Southern
90   Horticultural Laboratory (Poplarville, MS, USA), the experimental orchard of the Kyoto University
91   (Kyoto, Japan), the Miyagi Prefectural Institute of Agriculture and Horticulture (Miyagi, Japan), and
92   the Shizuoka Institute of Agriculture and Forestry (Shizuoka, Japan) in April 2018 (Table 1). We
93   included as many available accessions as possible to ensure that almost all of the currently cultivated
94   accessions in their pedigree were represented. The pentaploid 'Robeson' and hexaploid 'Pink
95   Lemonade', which are not pure *V. virgatum* but are from crosses with *V. corymbosum* according to the
96   pedigree record, were grouped as CH in this study. The analyzed plant materials are listed in
97   Supplementary Table S1. Total genomic DNA was isolated from young leaf tissue using a modified
98   hexadecyltrimethylammonium bromide (CTAB) protocol (Doyle and Doyle 1990). The ddRAD-seq
99   libraries were constructed as previously described (Shirasawa et al. 2016). Equal amounts of each
100 library were combined and sequenced with a lane of the Illumina HiSeq 4000 system (Illumina, San
101 Diego, CA, USA) to generate 100-bp paired-end reads.

102   All sequences were pre-processed with a custom Python script
103 (http://comailab.genomecenter.ucdavis.edu/index.php/Barcoded_data_preparation_tools). Sequences
104 with a base quality Phred score lower than 20 and with N bases were trimmed and reads shorter than
105 35 bp were discarded. Clean reads were mapped to the *V. corymbosum* 'Draper' reference genome
106 (Colle et al. 2019) using BWA-MEM (version 0.7.17) (Li and Durbin 2009). Despite the published
107 'Draper' tetraploid genome sequence consisted of four phased sets of the genome (Colle et al. 2019),
108 the diversity across the homoeologous chromosomes remains to be clarified in a population level and

109 subgenome partitions are undistinguishable in the polyploid nature. Therefore, we selected the longest

110 scaffolds set representing each of twelve 'Draper' homoeologous groups (Scaffolds 1, 2, 4, 6, 7, 11,

111 12, 13, 17, 20, 21, and 22, representing chromosomes 1–12) as representing 'Draper' genomic

112 sequences to minimize the complexity. All sequences were confirmed to satisfy the following criteria:

113 > 1,000,000 mapped read counts and > 0.5 mapping rate for each accession. The SAMtools program

114 (version 1.9) (Li et al. 2009) with the mpileup -q 20 option and VarScan (version 2.3.9) (Koboldt et al,

115 2012) with the mpileup2snp mode were used to create the initial VCF file. We applied two types of

116 SNP calling strategies, namely the diploid model and the continuous model. The genotype data based

117 on the diploid model did not include an allelic dosage for each variant. Regarding the continuous

118 model, SNP genotypes were assigned a value between 0 and 1 based on ALT/(ALT + REF), where

119 ALT and REF are the counts for the alternative allele-supporting reads and the reference allele-

120 supporting reads, respectively (de Bem Oliveira et al. 2019).

121 Before producing the genotype matrix with the diploid model, we visualized the distribution

122 of the alternative allele frequency for all RAD sites according to ploidy levels (Supplementary Fig.

123 S1). Although a prominent simplex peak was detected, applying a single threshold for calling

124 heterozygous variants was considered inappropriate because the homozygosity peak at 0% overlapped

125 with the simplex peak. Thousands of ambiguous loci were detected even at the falling point of

126 inflection between peaks. Therefore, to create high-confidence SNP genotype sets, we masked the

127 ambiguous loci. In the diploid model, SNPs were called as heterozygous (5% < ALT% < 95%) or

128 homozygous (0% ≤ ALT% ≤ 0.01% and 99.99% ≤ ALT% ≤ 100%), and the rest were masked

129 (i.e., missing). The SNP loci were further filtered with VCFtools (Danecek et al, 2011) according to

130 the following criteria: (i) minimum depth of coverage for each individual, 20; (ii) biallelic locus only;

131 (iii) maximum missing data, 0.7; and (iv) minor allele frequency, 0.1. Loci that were heterozygous for

132 all individuals were further filtered with a custom Python script. These filtering steps were performed

133 independently for a subset comprising all cultivars (SHB, NHB, RE, CH, and HH), a subset with SHB,

134 NHB, and RE, a subset with only highbush cultivars, and a subset with only SHB to modulate the

135 effect of the minor allele frequency and missing cutoffs. The SNP loci selected based on the diploid

136 model were used in the continuous model. In addition to the SNP selection based on the diploid model,

137 there was no further filtering specific to the continuous model to ensure a fair comparison between the

138 models.

139

140 **Population structure analysis**

141 The SNP genotypes called with the diploid and continuous models underwent a probabilistic

142 principal component analysis (PCA) with the R package pcaMethods (Stacklies et al. 2007). The

143 probabilistic PCA is a probabilistic formulation of PCA model with maximum likelihood estimation

144 and could deal with dataset with missing value. The results based on the diploid and continuous models

145   were compared by calculating the Pearson correlation coefficient for each PC.

146   The SNPs called with the diploid model for all accessions were used to construct a

147   phylogenetic tree according to the neighbor-joining method of MEGA X (Kumar et al. 2018), with

148   1,000 bootstrap replications and a pairwise deletion option for missing data. Each SNP locus was

149   represented by two bases in the input sequence, AA or BB for the homozygous genotype and AB for

150   the heterozygous genotype. To evaluate the population structure of the blueberry collection, we

151   performed a structure analysis with the STRUCTURE software (version 2.3.4) (Pritchard et al, 2000),

152   which is reportedly more robust than other commonly used clustering programs for analyzing mixed

153   ploidy populations (Stift et al. 2019). Regarding the input data, we coded the genotypes based on the

154   diploid model as co-dominant markers with an unknown dosage as described by Meirmans et al.

155   (2018) and Stift et al. (2019). For example, the genotype AB of a tetraploid individual based on the

156   diploid model, which is a genotype derived from AAAB or AABB or ABBB, was coded as marker

157   phenotype AB. To decrease the computation requirements, the loci were thinned so two or more sites

158   were not within 10 kb, and the resulting 6,495 SNPs were used as the input data for the STRUCTURE

159   software. Additionally, SHB, NHB, and HH were coded as tetraploid, whereas RE and 'Pink

160   Lemonade' were coded as hexaploid and 'Robeson' was coded as pentaploid. The K values ranging

161   from 1 to 10 were evaluated using 100,000 MCMC iterations after 10,000 burn-in iterations to infer

162   the population ancestry of genotypes in K predefined clusters. At least five runs for each K were

163   conducted as replicates and the replicates were summarized with CLUMPP (Jakobsson and Rosenberg

164   2007). The delta K method (Evanno et al. 2005) of STRUCTURE HARVESTER (Earl and VonHoldt

165   2012) was used to infer the optimal K value.

166   To analyze the genomic differentiation among cultivar groups, we performed PCA-based

167   outlier detection analysis implemented with the R package pcadapt (version 4.0.2) (Luu et al. 2017),

168   using SNP genotypes called with the diploid model. The assumption of pcadapt is that markers

169   excessively related to the population structure are responsible for local adaptations. Notably, pcadapt

170   can deal with the continuous separation of groups, which is expected in blueberry populations. To

171   explore the loci driving genomic differentiation, the component-wise genome scans in pcadapt were

172   applied for the PCs with distinct separation patterns among cultivar groups. The q-value was used to

173   control the false positive discovery errors and was calculated with the R package qvalue (Storey et al,

174   2019). Loci with a q-value lower than 0.1 were considered as candidate adaptive loci. To examine the

175   distribution of outlier loci across the HB/RE and NHB/SHB genomes, Manhattan plots depicting the

176   genomic positions of outlier SNPs and their respective significant association values $[-\log_{10}(P)]$ were

177   prepared with the R package qqman (Turner 2018). Pairwise Weir and Cockerham's F_{st} estimate was

178   calculated using VCFtools (Danecek et al, 2011).

179

180   **Linkage disequilibrium**

181   Squared correlation coefficients (r^2) of the SNP genotypes in each pair of SNPs on

182 chromosomes were calculated based on the diploid and continuous models with the PLINK software

183 (version 1.9) (Chang et al. 2015). The r^2 value was regressed with the physical distance via loess

184 smoothing implemented in the R package ggplot2 (Wickham 2009), with span = 0.1.

185   To further evaluate potential associations between distant pairs, a haplotype block

186 estimation based on the quantile regression was applied to the genotype matrix of the SHB group

187 created with the continuous model. First, r^2 values for the correlation between each SNP and all other

188 SNPs on a chromosome were calculated. The r^2 values were regressed against physical distance for

189 each SNP. The regression was conducted based on quantile regression and smoothed with a cubic

190 spline using the qsreg function implemented in the fields R package (version 9.8.6) (Nychka et al.

191 2017), with lambda = 1e10. An evaluation of several quantile values for the regression revealed the

192 95th percentile regression was the best fit for the observed maximum distance of associations

193 (Supplementary Fig. S2). On the basis of the regression, the point where the 95th percentile regression

194 curve first reached r^2 = 0.2 was recorded for each SNP.

195

196 **Results**

197 **Genotyping and population structure**

198   With our SNP selection criteria, 47,254 and 46,511 SNPs were detected in all populations

199 and in the highbush populations, respectively. The overall average read depth across all the individuals

200 in the SNPs loci was 72.8. The PCA results based on the diploid and continuous models were highly

201 correlated at least up to 10 PCs (Supplementary Fig. S3), suggesting even very minor population

202 structure could be detected by the diploid model in this population. Considering the relatively low read

203 depth, the ease in handling and compatibility with diverse software, we applied the diploid model

204 genotype calling for most of the following experiments.

205   The phylogenetic tree revealed a distinct genetic cluster of RE cultivars (Fig. 1). NHB

206 cultivars also clustered together with some exceptions. Some NHB accessions were far from the NHB

207 cluster and some SHB accessions were found in the NHB cluster. Specifically, the NHB cultivars

208 'Bluecrop' and 'Bounty' were far from the NHB cluster. The HH cultivars, which exhibit cold

209 hardiness, clustered with the NHB cultivars, except for 'TopHat' which was far from the NHB cultivars.

210 CH cultivars, which were bred with SHB accessions based on the pedigree, were found together with

211 SHB.

212   The population structure analysis with the STRUCTURE software (Fig. 2) suggested that

213 RE and NHB are relatively homogenous but SHB contains a considerably more admixed genetic

214 background than RE and NHB. Considering the pedigree record of blueberry, the deep blue part of

215 Fig. 2A corresponds to the *V. virgatum* genome, whereas the orange corresponds to the *V. corymbosum*

216 genome. The origin of the other ancestral genomes was unclear, but the gray is most likely the *V.*

217   *angustifolium* genome because it represents half of the HH genomes. Moreover, the yellow, green, and

218   light blue in K = 5 probably correspond to the wild *Vaccinium* genomes including *V. darrowii* and *V.*

219   *elliottii* because most of the SHB individuals possess these genomes. We also analyzed the genomic

220   ancestry according to the different selection sites in the USA for the SHB group. The cultivars bred in

221   North Carolina tend to have more of the *V. corymbosum* genome, whereas the cultivars bred in Florida

222   and Georgia tend to be more admixed. The cultivars 'O'Neal' and 'Reveille', which are widely

223   distributed as SHB, largely consisted of the presumed *V. corymbosum* ancestral genome (Fig. 2B). A

224   single high delta K value was obtained at K = 9, and the delta K values were stably low for the other

225   tested K values (Fig. 2C), providing a possibility that the nine ancestral genomes underly the blueberry

226   gene pool.

228   **Characterization of the genomic differentiation among cultivar groups**

229   We first analyzed the genomes to identify diagnostic loci that could distinguish between

230   SHB and NHB based on the SNP data. Despite the interspecific origin of SHB, there was no allele

231   present in all SHB, but lacking in NHB. This was confirmed with allowing 50% missing data in each

232   group based on the diploid model matrix. Therefore, we assumed that the genetic differentiation might

233   not be significant among blueberry cultivar groups, at least between SHB and NHB. Pairwise F_{st}

234   estimate indicated lower genetic differentiation between SHB and NHB than between SHB and RE,

235   or between NHB and RE (Supplementary Table S2). We next applied a PCA-based whole-genome

236   scan to uncover the genomic differentiation between cultivar groups. In a PCA plot for HB and RE

237   populations, HB and RE were clearly distinguishable along the first PC (PC1) (Fig. 3A). A Manhattan

238   plot depicting the significant association values $[-\log_{10}(P)]$ of the outlier loci revealed many peaks

239   spanning all chromosomes based on the pcadapt component-wise mode for PC1 (Fig. 3B). In contrast,

240   NHB and SHB were not divided into independent clusters, but were continuously distributed along

241   the PC1 score in a PCA plot for the NHB and SHB populations (Fig. 4A). There were 11 SNPs

242   fulfilling the q-value threshold on chromosomes 1, 2, and 8 (Fig. 4B). Although the detected four SNPs

243   on chromosome 1 and the six SNPs on chromosome 8 spanned across 4.9 Mb and 8.7Mb, respectively,

244   the genotypes in the population were highly correlated (Supplementary Tables S3 and S4). On the

245   basis of the genotype correlations, we considered the four SNPs on chromosome 1 and the six SNPs

246   on chromosome 8 to be on the haplotype blocks. The genotype scores of the three loci were modestly

247   correlated with the PC1 value (Fig. 4C). At the outlier locus (13:23005240) with the lowest *p*-value,

248   most of the SHB with the same genotype as NHB (homozygous for the alternative allele) were bred

249   with NC 1528, NC 1524, 'Bluechip', or 'Sharpblue' according to their pedigree records

250   (Supplementary Table S5). Among NHB cultivars, the heterozygous genotype in the outlier loci was

251   observed only in 'Bluecrop' at 1:25303016 and 2:26391780, and 'Bounty' at 13:23005240

252   (Supplementary Table S5).

253

Characterization of the chromosome-wide allelic association

254

255   The LD in the NHB and SHB populations decayed to $r^2 = 0.2$ in less than 10 kb

256 (Supplementary Fig. S4). Although the LD decay occurred slightly faster in SHB than in NHB, the

257 LD decay patterns were similar between these two populations (Supplementary Fig. S4). Despite the

258 observed rapid LD decay, there were many substantially associated SNP pairs with Mb-scale distances

259 (Supplementary Fig. S2). Figure 5 presents a plot of the maximum distances of the substantial allelic

260 associations for each SNP. In the SHB population, long potential associations in SNP pairs separated

261 by more than 5 Mb were found on all chromosomes (Fig. 5a). These associations tended to be located

262 at the center of chromosomes, except for chromosomes 6 and 11. Additionally, apparent secondary

263 peaks were also detected for several chromosomes, including chromosomes 5, 6, and 11. The genome-

264 wide median maximum association distance calculated based on the 95th percentile was 474 kb,

265 ranging from 231 kb on chromosome 12 to 871 kb on chromosome 4 (Fig. 5b). The outlier loci

266 associated with the separation between SHB and NHB were located on the Mb-scale haplotype blocks

267 (Fig. 6), with a considerably greater distance than the genome-wide median.

268

269 **Discussion**

270   Highbush blueberry originated and was domesticated in northern USA, but it is now

271 cultivated worldwide. The available highbush blueberry cultivars adapted to warm climates are the

272 result of extensive breeding, including interspecific hybridizations, which explains the mixture of

273 genomes in these cultivars. To increase the efficiency of crossing and selection strategies, blueberry

274 breeders, especially those with limited genetic resources, may benefit from the genetic characterization

275 of the extremely diverse *Vaccinium* population. In this study, we examined blueberry population

276 genetics using genome-wide SNP data of cultivars/accessions representing most of the currently

277 cultivated lines in their pedigree. We also analyzed the possible genomic differentiation among

278 blueberry cultivar groups.

279

280 **Genetic differentiation between RE and highbush populations**

281   The clear separation of highbush cultivars from the RE group based on the phylogenetic

282 relationships and PCA results is consistent with the findings of previous studies (Bian et al. 2014;

283 Campa and Ferreira 2018; Bassil et al. 2020). Despite the widespread contribution of RE in the SHB

284 pedigree (Brevis et al. 2008), the outlier SNPs associated with the separation between RE and highbush

285 blueberries were not localized to specific genomic regions, but were distributed throughout the

286 genome (Fig. 3B). These results are consistent with the notion that the initial NHB and RE cultivars

287 developed independently and RE was subsequently used to generate SHB in the NHB genomic

288 background. In the STRUCTURE analysis, the presumed *V. virgatum* genome was separated at K = 2

289 (Fig. 2). This is in accordance with the PCA result, and suggests the existence of a distinct feature in

290 the RE genome. Notably, RE cultivars appeared to comprise mostly the presumed *V. virgatum* genome,

291 with no contribution from the *V. corymbosum* genome, although the inverse was previously reported

292 (Brevis et al. 2008) and revealed in the clustering data (Fig. 2).

293

294 **Considerable admixture of the SHB population and its relationship to the allele selection**

295 **preferences**

296   In contrast to the clear separation between RE and the highbush blueberries, the relationship

297 between SHB and NHB is complex, with neither the PCA nor the phylogenetic analysis uncovering a

298 clear separation. The continuous relationship between SHB and NHB may be explained by the

299 complex interspecific crosses and recurrent backcrosses related to SHB development. The detection

300 of only three significant outlier loci associated with the continuous relationship further suggests a

301 weak population differentiation. The results also imply that the genotype information includes the

302 record of the directed selection of SHB, and the outlier loci may have been functional during SHB

303 development. The genotype patterns of the outlier loci appear to be associated with specific functions.

304 For example, at the most significant locus (13:23005240), all SHB accessions with the same allele as

305 NHB accessions (homozygous for the alternative allele) have NC 1528, NC 1524, 'Bluechip', or

306 'Sharpblue' in their pedigree (Supplementary Table S5). This may indicate that breeders favored

307 alleles from *V. angustifolium* and *V. corymbosum* over those from low-chill *Vaccinium* species. This is

308 also consistent with the known SHB breeding history, in which the initial low-chill SHB cultivars

309 developed in Florida were further crossed to adapt to colder regions (Ehlenfeldt et al. 1995). The

310 genomic admixture in SHB revealed by the STRUCTURE analysis (Fig. 2) likely reflects the genomic

311 segments introgressed from other *Vaccinium* species. The STRUCTURE data further revealed that the

312 ancestral genomic composition varied depending on the original selection locations (Fig. 2). The

313 cultivars bred in North Carolina, which is closer to the NHB production area than the other examined

314 regions, possessed more of the presumed *V. corymbosum* genome than the cultivars bred in other

315 regions. This is probably because of the targeted selection of the expected cold hardiness of *V.*

316 *corymbosum*. In contrast, the cultivars bred in Florida, which is the southernmost region examined in

317 this study, had a more mixed ancestry (i.e., admixed population). Florida is where SHB breeding was

318 initiated because breeders needed to develop cultivars adapted to the climate in this state, which is far

319 from where highbush blueberries originated. Therefore, the observed admixture can be attributed to

320 the local adaptation efforts. Thus, SHB is difficult to define at the genome level; however, we identified

321 different breeding directions during SHB development, likely because of the diversity in local breeding

322 centers. Additionally, we determined that the ancestry can be traced based on genomics, even in

323 polyploid blueberry.

324   We also confirmed the absence of a genomic region satisfying a strict threshold for

325   distinguishing SHB from NHB (i.e., a homozygous site in all NHB accessions that was heterozygous

326   or homozygous for the alternative allele in all SHB accessions). This suggests a lack of or only a few

327   introgressed genomic segments that are shared by all SHB accessions. The same result was obtained

328   when we excluded a relatively high-chill SHB cultivar ('Summit') from the analysis. Considering the

329   low-chill SHB accessions used in this study could not be distinguished from the other accessions in

330   the highbush population, we hypothesized that the adaptation of SHB to the southern region was

331   achieved through factors under polygenic control. Local adaptations with polygenic factors are

332   common in many plant species (Flood and Hancock 2017; Wisser et al. 2019). In this situation, a shift

333   in the allele frequency at many loci drives the adaptation (Stephan 2016), which is consistent with the

334   observed genotype patterns and population structure results. The outlier loci detected in the genome

335   scan may include loci controlling the traits mediating the adaptation of SHB to the southern region.

336   The observed long-range genotype associations of the outlier loci (Fig. 6) support the allele selection

337   preferences of the outlier loci. Our preliminary examination of the chilling requirement phenotype

338   indicated a lack of a significant association between the chilling requirement and the genotype of the

339   loci (data not shown). Ongoing association studies will hopefully elucidate the adaptation process.

340   Some of the results that were inconsistent with the general population features in the

341   clustering, phylogenetic analysis, and genome scan (Fig. 1, 2, and 4) can be explained by the

342   hybridization history. Pedigree of HH cultivar 'TopHat', which was far from the NHB cluster (Fig. 1),

343   is Mich. 19-H x 'Berkeley'. 'Berkeley' was developed with three of the four parents ('Stanley',

344   'Jersey', and 'Pioneer') of 'Bluecrop', which was extensively used for the development of SHB

345   (Brevis et al. 2008). The distinction of 'TopHat' from the NHB cultivars is also consistent with the

346   previous study (Bian et al. 2014). The mixture of SHB and NHB in the phylogenetic analysis (Fig. 1)

347   is probably related to the repeated hybridizations or shared polymorphisms in their ancestors. The

348   detection of NHB 'Bluecrop' and 'Bounty' in the SHB cluster (Fig. 1) is likely due to the contribution

349   of 'Bluecrop' and Crabbe-4 genomes to the SHB population, as previously suggested (Brevis et al.

350   2008). Crabbe-4, a wild *V. corymbosum* clone that is not present in the pedigree of most of NHB

351   cultivars, was used to develop NHB 'Murphy', a parent of 'Bounty'. This notion is also consistent

352   with the detection of the heterozygous genotype of 'Bluecrop' and 'Bounty' at the outlier loci (Fig. 4,

353   Supplementary Table S5). Moreover, SHB cultivars 'O'Neal' and 'Reveille', which appeared to largely

354   consist of the ancestral *V. corymbosum* genome based on the clustering analysis (Fig. 2), were likely

355   to have lower than expected (according to the pedigree records) genomic contribution from the other

356   *Vaccinium* species (Brevis et al. 2008). This can be explained by the elimination of alleles derived

357   from interspecific hybridizations during the development, considering that the interspecific

358   hybridizations were made several generations prior to the development of 'O'Neal' and 'Reveille'

359   (Ballington et al. 1990; Cummins 1991).

360

Mb-scale linkage disequilibrium in the SHB population

361

362   The pattern and extent of LD are important factors for explaining the past events in a

363   population and for designing association mapping studies. Additionally, LD is a sensitive indicator of

364   the population genetic forces influencing genomic structures (Slatkin 2008), and it is affected by

365   multiple factors, including the ploidy level and introgression. Regarding blueberry, although several

366   association studies have been attempted, only a few investigations have focused on the extent of LD.

367   Ferrão et al. (2018) reported that the estimated genome-wide LD decay in a tetraploid blueberry

368   breeding population was 73–80 kb, which was based on genotypic correlations, with genotypes called

369   with the diploid and tetraploid models. In the current study, we estimated a less extensive LD for the

370   SHB and NHB groups with the diploid model (Supplementary Fig. S4). However, this may have

371   severely underestimated the population LD extent because repulsion-phase marker pairs, which are

372   less informative in polyploids, were averaged together with more informative pairs. In fact, we

373   identified SNP pairs with allelic associations with distances of several Mb in the SHB population.

374   Therefore, we applied quantile regression with empirically determined parameters to characterize the

375   genome-wide pattern of allelic genotype correlations. A similar methodology using quantile regression

376   was previously applied in the LD survey of sugar beet and tetraploid potato (Adetunji et al. 2014;

377   Sharma et al. 2018). By using this method, we proved the existence of long-lasting association pairs

378   with distances of up to several Mb in all chromosomes (Fig. 5). These long-lasting associations should

379   be consistent with the SHB breeding history, considering the recent origin and the widespread genetic

380   contribution of wild *Vaccinium* clones (Brevis et al. 2008). The pattern of the distribution of the LD

381   estimates across the genome may be related to different recombination frequencies and large structural

382   variations. The predominant localization of the long-lasting association pairs at the center of

383   chromosomes may be due to the suppression of recombination in the centromeric region. In contrast,

384   the distinct distribution pattern observed for chromosome 6 may be related to the rearrangement or

385   mis-assembly in the 'Draper' reference genome (Colle et al. 2019). Moreover, apparent secondary

386   peaks in addition to those at the centromeric regions were detected for several chromosomes. The data

387   also suggest the existence of haplotype blocks that are longer than expected (Supplementary Figs S2

388   and S3), which may decrease the genotyping costs of a future genome-wide association study (GWAS).

389   Considered together, the allelic associations detected by the quantile regression method in this study

390   appear to be useful for characterizing the genomic features of tetraploid blueberry. To increase the

391   resolution and the accuracy of LD estimates, it is essential that future studies elucidate the inheritance

392   mode and produce genetic maps on a genome-wide scale, as has been done for potato (Vos et al. 2017).

393   Our data revealed a less extensive LD in SHB than in NHB in our highbush population

394   (Supplementary Fig. S4). There are two potential explanations for this finding. First, compared with

395   the SHB accessions, there were fewer and less diverse NHB accessions. Second, the SHB accessions

396   had more founder haplotypes than the NHB accessions because of interspecific hybridizations. There

397   are reportedly two different genotypes for Florida 4B, which contributed considerably to the SHB

398   genome (Bassil et al. 2018). Specifically, CVAC 1790, which is one of the Florida 4B genotypes that

399   has been widely used during SHB development, is the result of an interspecific hybridization between

400   the wild diploid species in Florida (Bassil et al. 2018).

401

402   **Population structure inference of cultivated polyploid blueberries**

403   It is known that allele dosage of polyploid species significantly affects calculation of allele

404   frequency, which is fundamental to many population genetic based inferences (Cockerham 1973;

405   Dufresne et al. 2014). However, in many cases, there are still difficulties regarding the dosage

406   genotyping especially in genotyping accuracy, costs, and software/parameter compatibility (Gerard et

407   al. 2018; Meirmans et al. 2018). Herein, as the result of PCA highly matched between the diploid and

408   continuous models (Supplementary Fig S3), we considered that the genotype matrix based on the

409   diploid model represented most of the population structural information present in the population. The

410   observed high consistency between the two can relate to diversity in the presence/absence of alleles,

411   which is assumed in the situation of less generation cycles from the domestication and the potential

412   allopolyploid origin of blueberry (Colle et al. 2019).

413   Up to this time, seven *Vaccinium* species, *V. darrowii*, *V. elliottii*, *V. tenellum*, *V.*

414   *angustifolium*, *V. corymbosum*, *V. constablaei*, and *V. virgatum*, are recognized as a genomic backbone

415   of cultivated polyploid blueberries (Brevis et al. 2008; Ballington 2009). In addition, *V. myrtilloides*

416   and *V. pallidum* have partially but substantially contributed to the blueberry gene pool (Ballington

417   2009). Thus, it is possible to interpret that the optimal K value nine in the clustering (Fig. 2) is fairly

418   matched with the number of species underlying the development of cultivated polyploid blueberries.

419   However, species delimitation within the *Vaccinium* genus is still controversial, and hybridization

420   among species in section *Cyanococcus* is common in nature. Thus, this point is unable to be

421   experimentally assessed unless the diversity of the ancestral species is clarified. Future works with

422   combining the ancestral species and full dosage information of cultivated blueberries may facilitate

423   deeper understanding of the genomic origin of cultivated blueberries.

424

425   **Conclusion**

426   In this study, an analysis of the population genetics of diverse blueberry populations clarified

427   the genomic ancestry of blueberry. The general trends revealed by the results presented herein include

428   a homogenous genomic background in RE and NHB, in contrast to the admixed background of SHB,

429   which is consistent with the recorded history of blueberry breeding. The structural characterization

430   and scanning of the genomes indicate that SHB development likely involved directed selection. This

431   is probably related to the independence of the breeding projects conducted by various breeding centers,

432   which were influenced by the local climate and breeder strategies. Despite the extensive breeding and

433   admixed nature of the SHB population, there appears to be no introgressed genomic segment common

434   to all SHB cultivars. Collectively, we hypothesize that polygenic factors affected the adaptation of

435   SHB to the climate in the southern USA. The detected outlier loci were associated with the continuous

436   separation between NHB and SHB, and may be considered as part of the alleles mediating the

437   adaptation of SHB. To the best of our knowledge, none of the loci presented in this study match loci

438   detected in previous GWAS/mapping studies. Future population-scale genomic investigations of

439   diverse NHB accessions as well as accurate association analyses regarding the adaptive traits may help

440   to further clarify the process underlying the adaptation of SHB.

441

442   **Data archiving**

443   The raw ddRAD-seq data analyzed in this study have been submitted to the DDBJ Sequence Read

444   Archive (accession number DRA009951).

445

446   **Acknowledgments**

453

454   **Conflict of interest**

455   The authors declare that they have no conflict of interest.

456

457   **References**

458   Adetunji I, Willems G, Tschoep H, Bürkholz A, Barnes S, Boer M, et al. (2014) Genetic diversity and

459   linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. Theor

460   Appl Genet. 127:559–571. https://doi.org/10.1007/s00122-013-2239-x

461   Ballington JR, Mainland CM, Duke SD, Draper AD, Galletta GJ (1990) 'O'Neal' southern highbush

462   blueberry. HortSci. 25:711–712

463   Ballington JR (2009) The role of interspecific hybridization in blueberry improvement. Acta Hortic.

464   810:49–60

465   Bassil N, Bidani A, Hummer K, Rowland LJ, Olmstead J, Lyrene P, et al. (2018) Assessing genetic

466   diversity of wild southeastern North American *Vaccinium* species using microsatellite markers.

467   Genet Resour Crop Evol 65:939–950

468 Bassil N, Bidani A, Nyberg A, Rowland LJ, Olmstead J, Lyrene P, et al. (2020) Microsatellite markers
469   confirm identity of blueberry (*Vaccinium* spp.) plants in the USDA-ARS National Clonal
470   Germplasm Repository collection. Genet Resour Crop Evol 67:393–409

471 Benevenuto J, Ferrão LF, Amadeu RR, Munoz P (2019) How can a high-quality genome assembly
472   help plant breeders? GigaScience 8:1–4. https://doi.org/10.1093/gigascience/giz068

473 Bian Y, Ballington J, Raja A, Brouwer C, Reid R, Burke M, et al. (2014) Patterns of simple sequence
474   repeats in cultivated blueberries (*Vaccinium* section *Cyanococcus* spp.) and their use in revealing
475   genetic diversity and population structure. Mol Breed. 34:675–689.
476   https://doi.org/10.1007/s11032-014-0066-7

477 Brevis PA, Bassil N, Ballington JR, Hancock JF (2008) Impact of wide hybridization on highbush
478   blueberry breeding. J Amer Soc Hort Sci. 133:427–437.
479   https://doi.org/10.21273/JASHS.133.3.427

480 Campa A, Ferreira JJ (2018) Genetic diversity assessed by genotyping by sequencing (GBS) and for
481   phenological traits in blueberry cultivars. PLOS ONE 13:e0206361.
482   https://doi.org/10.1371/journal.pone.0206361

483 Cappai F, Benevenuto J, Ferrão L, Munoz P (2018) Molecular and Genetic Bases of Fruit Firmness
484   Variation in Blueberry—A Review. Agronomy 8:174.
485   https://doi.org/10.3390/agronomy8090174

486 Ceccarelli S, Grando S, Maatougui M, Michael M, Slash M, Haghparast R, et al. (2010) Plant breeding
487   and climate changes. J Agric Sci. 148:627–637. https://doi.org/10.1017/S0021859610000651

488 Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation
489   PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7.
490   https://doi.org/10.1186/s13742-015-0047-8

491 Chavez DJ, Lyrene PM (2009) Interspecific crosses and backcrosses between diploid *Vaccinium*
492   *darrowii* and tetraploid southern highbush blueberry. J Amer Soc Hort Sci. 134:273–280.
493   https://doi.org/10.21273/JASHS.134.2.273

494 Cockerham CC (1973) Analyses of gene frequencies. Genetics 74:679–700.

495 Colle M, Leisner CP, Wai CM, Ou S, Bird KA, Wang J, et al. (2019) Haplotype-phased genome and
496   evolution of phytonutrient pathways of tetraploid blueberry. GigaScience 8:1–15.
497   https://doi.org/10.1093/gigascience/giz012

498 Comai L (2005) The advantages and disadvantages of being polyploid. Nat Rev Genet 6:836–846.
499   https://doi.org/10.1038/nrg1711

500 Cummins JN (1991) Register of new fruit and nut varieties. HortSci. 26:951-986

501 de Bem Oliveira I, Resende MFR, Ferrão LF v., Amadeu RR, Endelman JB, Kirst M, et al. (2019)
502   Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and
503   continuous genotyping calls in phenotype prediction. G3 9:g3.400059.2019.

504   https://doi.org/10.1534/g3.119.400059

505   Doyle JJ, Doyle LH (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15.

506   Dufresne F, Stift M, Vergilino R, Mable BK (2014) Recent progress and challenges in population
507   genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical
508   tools. Mol Ecol 23:40–69.

509   Ehlenfeldt MK, Draper AD, Clark JR (1995) Performance of southern highbush blueberry cultivars
510   released by the U.S. Department of Agriculture and cooperating state agricultural experiment
511   stations. HortTechnol. 5:127–130. https://doi.org/10.21273/HORTTECH.5.2.127

512   Ferrão LF, Benevenuto J, Oliveira I de B, Cellon C, Olmstead J, Kirst M, et al. (2018) Insights into
513   the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS
514   context. Front Ecol Evol 6:107. https://doi.org/10.3389/fevo.2018.00107

515   Flood PJ, Hancock AM (2017) The genomic basis of adaptation in plants. Curr Opin Plant Biol. 36:88–
516   94. https://doi.org/10.1016/j.pbi.2017.02.003

517   Gerard D, Ferrão LFV, Garcia AAF, and Stephens M (2018) Genotyping polyploids from messy
518   sequencing data. Genetics 210: 789–807

519   Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for
520   dealing with label switching and multimodality in analysis of population structure.
521   Bioinformatics 23:1801–1806.

522   Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular eolutionary genetics
523   analysis across computing platforms. Mol Biol Evol. 35:1547–1549.
524   https://doi.org/10.1093/molbev/msy096

525   Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.
526   Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

527   Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. (2009) The Sequence
528   Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
529   https://doi.org/10.1093/bioinformatics/btp352

530   Luu K, Bazin E, Blum MGB (2017) pcadapt : an R package to perform genome scans for selection
531   based on principal component analysis. Mol Ecol Resour. 17:67–77.
532   https://doi.org/10.1111/1755-0998.12592

533   Lyrene PM, Ballington JR (1986) Wide hybridization in *Vaccinium*. HortSci. 21:52–57

534   Meirmans PG, Liu S, van Tienderen PH (2018) The Analysis of polyploid genetic data. J Hered.
535   109:283-296. https://doi.org/10.1093/jhered/esy006

536   Nicotra AB, Atkin OK, Bonser SP, Davidson AM, Finnegan EJ, Mathesius U, et al. (2010) Plant
537   phenotypic plasticity in a changing climate. Trend Plant Sci. 15:684–692.
538   https://doi.org/10.1016/j.tplants.2010.09.008

539   Sharma SK, MacKenzie K, McLean K, Dale F, Daniels S, Bryan GJ (2018) Linkage disequilibrium

540   and evaluation of genome-wide association mapping models in tetraploid potato. G3 8:3185–
541   3202. https://doi.org/10.1534/g3.118.200377
542 Shirasawa K, Hirakawa H, Isobe S (2016) Analytical workflow of double-digest restriction site-
543   associated DNA sequencing based on empirical and in silico optimization in tomato. DNA Res.
544   23:145–153. https://doi.org/10.1093/dnares/dsw004
545 Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the
546   medical future. Nat Rev Genet. 9:477–485. https://doi.org/10.1038/nrg2361
547 Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007) pcaMethods: a bioconductor package
548   providing PCA methods for incomplete data. Bioinformatics 23:1164–1167.
549   https://doi.org/10.1093/bioinformatics/btm069
550 Stephan W (2016) Signatures of positive selection: from selective sweeps at individual loci to subtle
551   allele frequency changes in polygenic adaptation. Mol Ecol. 25:79–88.
552   https://doi.org/10.1111/mec.13288
553 Stift M, Kolář F, Meirmans PG (2019) Structure is more robust than other clustering methods in
554   simulated mixed-ploidy populations. Heredity 123:429–441. https://doi.org/10.1038/s41437-
555   019-0247-6
556 Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: Unlocking genetic potential from
557   the wild. Science 277:1063–1066. https://doi.org/10.1126/science.277.5329.1063
558 Turner SD (2018) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots.
559   J Open Source Software. 3:731. https://doi.org/10.21105/joss.00731
560 Vos PG, Paulo MJ, Voorrips RE, Visser RGF, van Eck HJ, van Eeuwijk FA (2017) Evaluation of LD
561   decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato.
562   Theor Appl Genet. 130:123–135. https://doi.org/10.1007/s00122-016-2798-8
563 Wisser RJ, Fang Z, Holland JB, Teixeira JEC, Dougherty J, Weldekidan T, et al. (2019) The genomic
564   basis for short-term evolution of environmental adaptation in maize. Genetics 213:1479–1494.
565   https://doi.org/10.1534/genetics.119.302780

566

567 **Figure legends**

568

569 Fig. 1. Consensus neighbor-joining phylogenetic tree of the blueberry population. The tree was
570 constructed based on the genotype data for 47,254 genome-wide SNPs in 137 accessions. Black, blue,
571 and red represent rabbiteye blueberry (RE), southern highbush blueberry (SHB), and northern
572 highbush blueberry (NHB) cultivars, respectively. Green circles and squares represent half highbush
573 (HH) and complex hybrid (CH) cultivars, respectively. Branches reproduced in less than 50% of the
574 bootstrap replicates are collapsed.

575

576 Fig. 2. Proportion of the ancestry of blueberry. (A) Proportion of the ancestry of the individuals
577 inferred with the STRUCTURE software. K values ranging from 2 to 10 was plotted. Each individual
578 is presented as a vertical bar. RE, SHB, NHB, HH, CH represent rabbiteye, southern highbush,
579 northern highbush, half highbush, and complex hybrid blueberries, respectively. FL, GA, MS, and NC
580 represent Florida, Georgia, Mississippi, and North Carolina, respectively, and indicate the USA states
581 producing the SHB cultivars. (B) Plot of the proportion of the ancestry of SHB bred at North Carolina
582 inferred with K = 9. (C) Evanno's delta K plotted against K.

584 Fig. 3. Population differentiation between RE and highbush blueberry. (A) Principal component
585 analysis based on the SNP data of RE, SHB, and NHB individuals generated with the diploid model.
586 (B) Manhattan plot of the outlier loci associated with the first principal component in panel A, inferred
587 with the component-wise genome scan implemented in the pcadapt software. Green dots represent
588 significantly associated SNPs.

590 Fig. 4. Population differentiation among highbush blueberries. (A) Principal component analysis based
591 on the SNP data generated with the diploid model for a subpopulation of SHB and NHB. (B)
592 Manhattan plot of the outlier loci associated with the first principal component in panel A, inferred
593 with the component-wise genome scan implemented in the pcadapt software. Green dots represent
594 significantly associated SNPs. (C) Plot of the PC1 scores according to the genotypes of the three outlier
595 loci.

597 Fig. 5. Chromosome-wide allelic associations in the SHB population. (A) Genome-wide distribution
598 of allelic associations. The maximum distance with a substantial association ($r^2 = 0.2$) estimated with
599 the 95th percentile regression was plotted for each SNP. (B) Violin plot indicating the maximum
600 association distance for each chromosome. Red dots represent the median value.

602 Fig. 6. Pairwise allelic genotype correlations of the outlier loci associated with the separation of SHB
603 and NHB. Red and green lines represent a cubic spline fitted for the 95th and 50th percentiles,
604 respectively. Red dots represent the points where the fitted curve of the 95th percentile first decayed
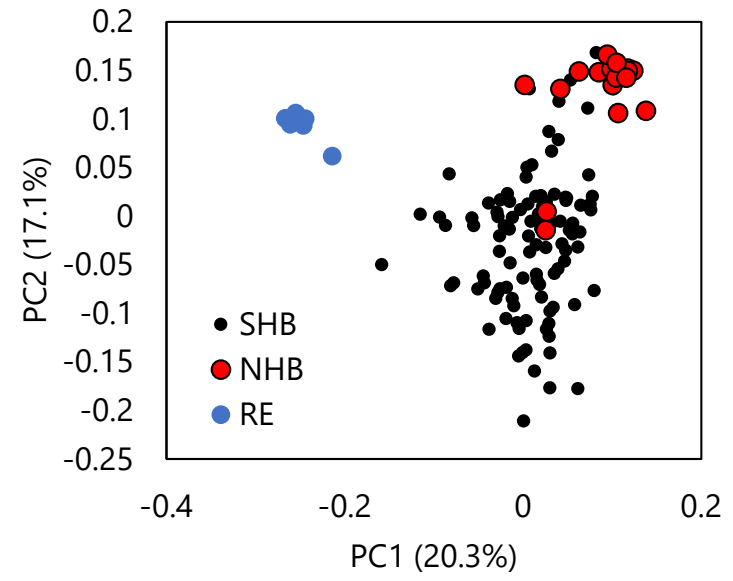605 to $r^2 = 0.2$.

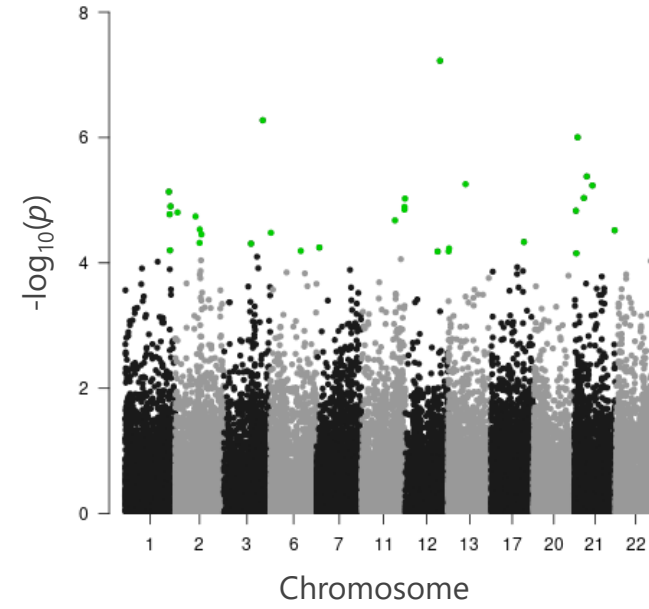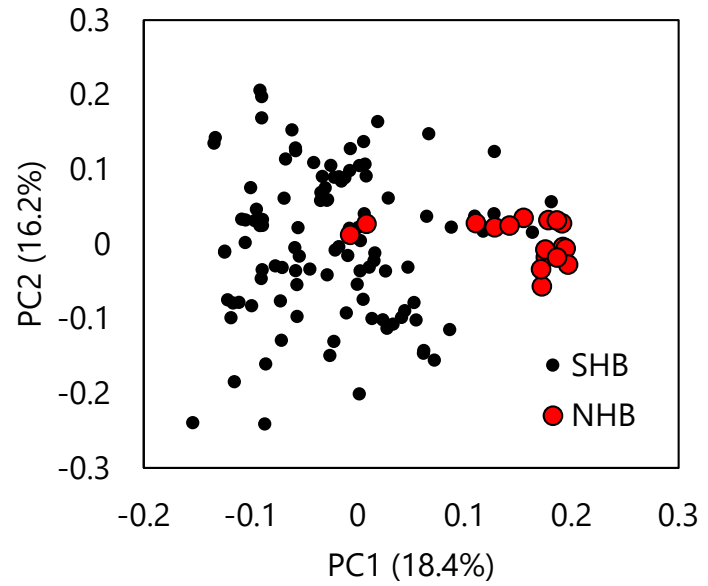607 Table 1. The number of accessions by the sampled locations

Fig5
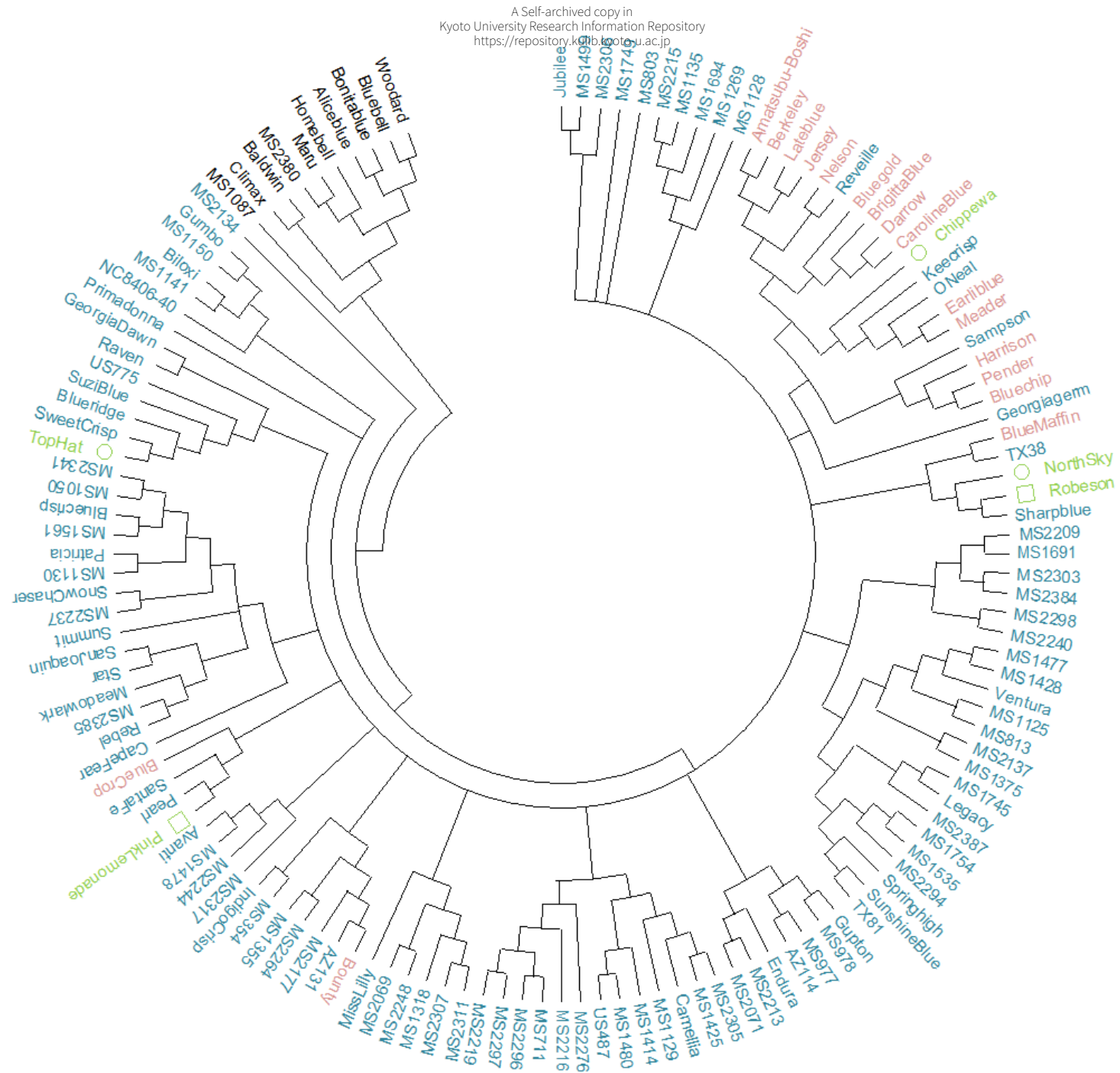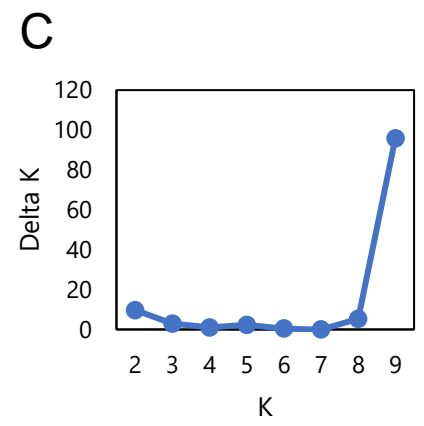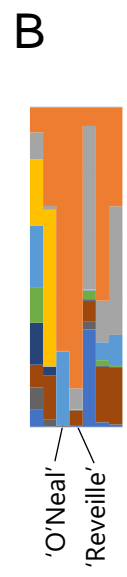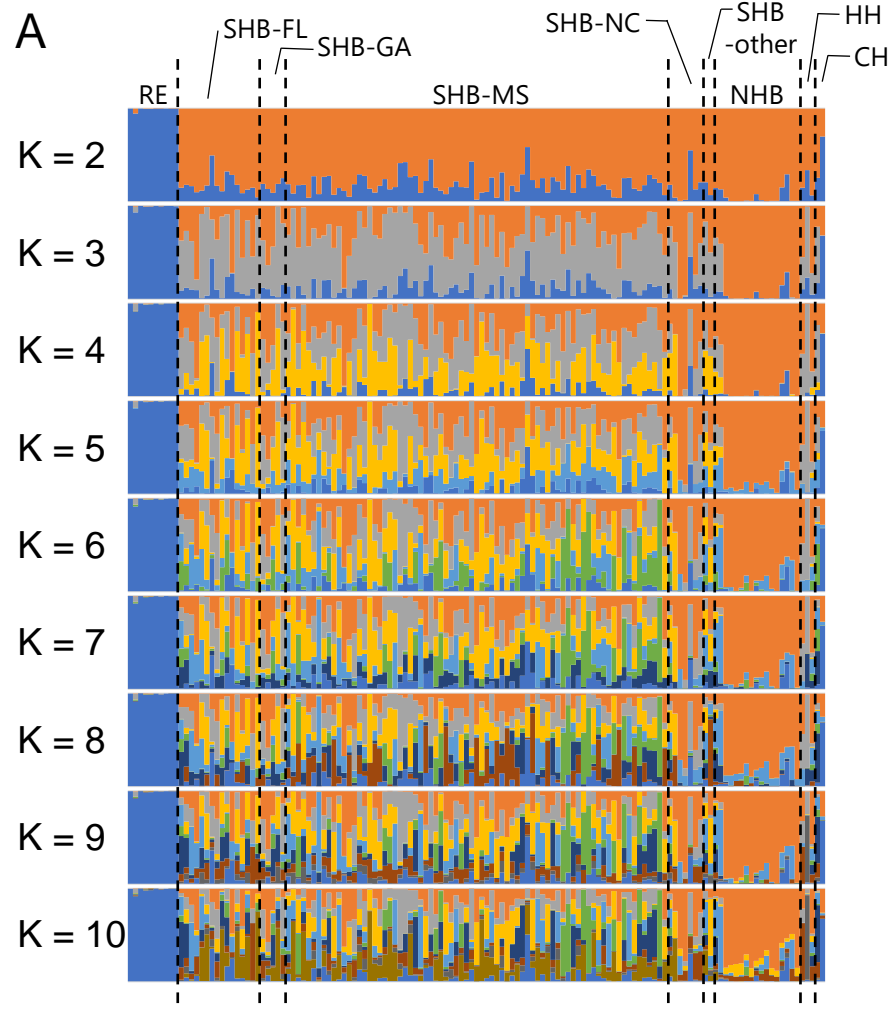
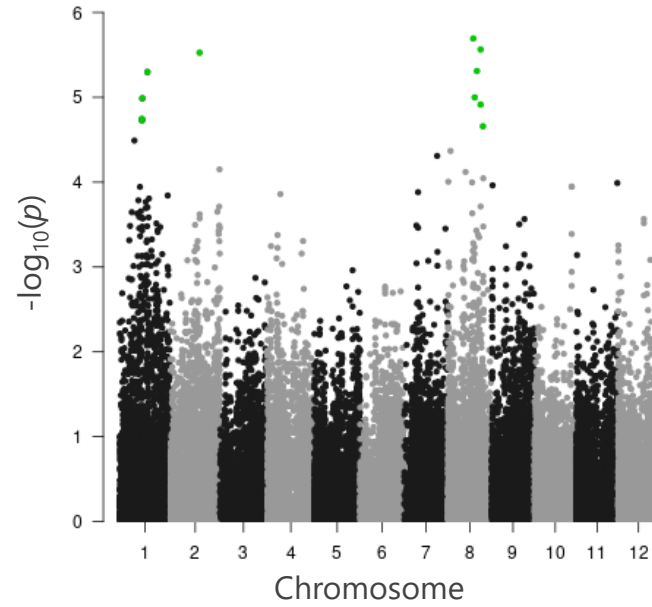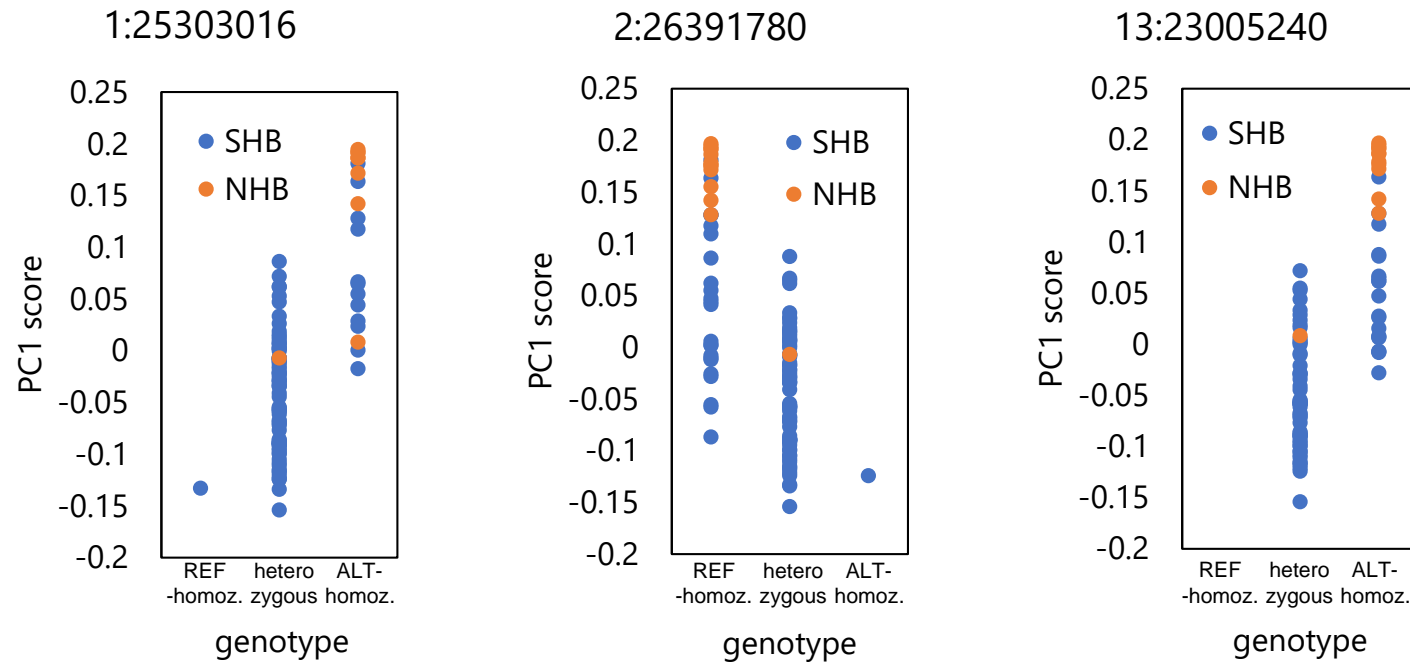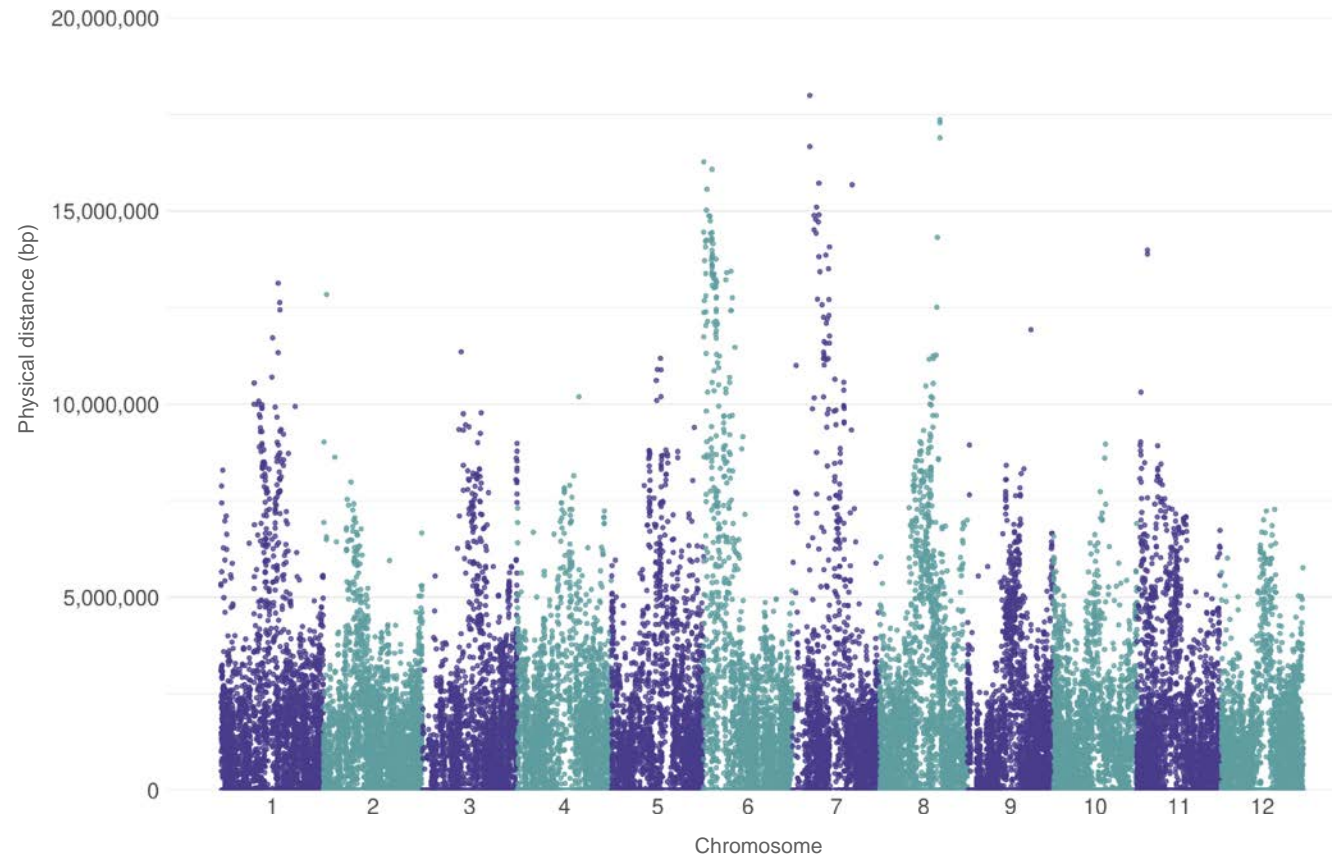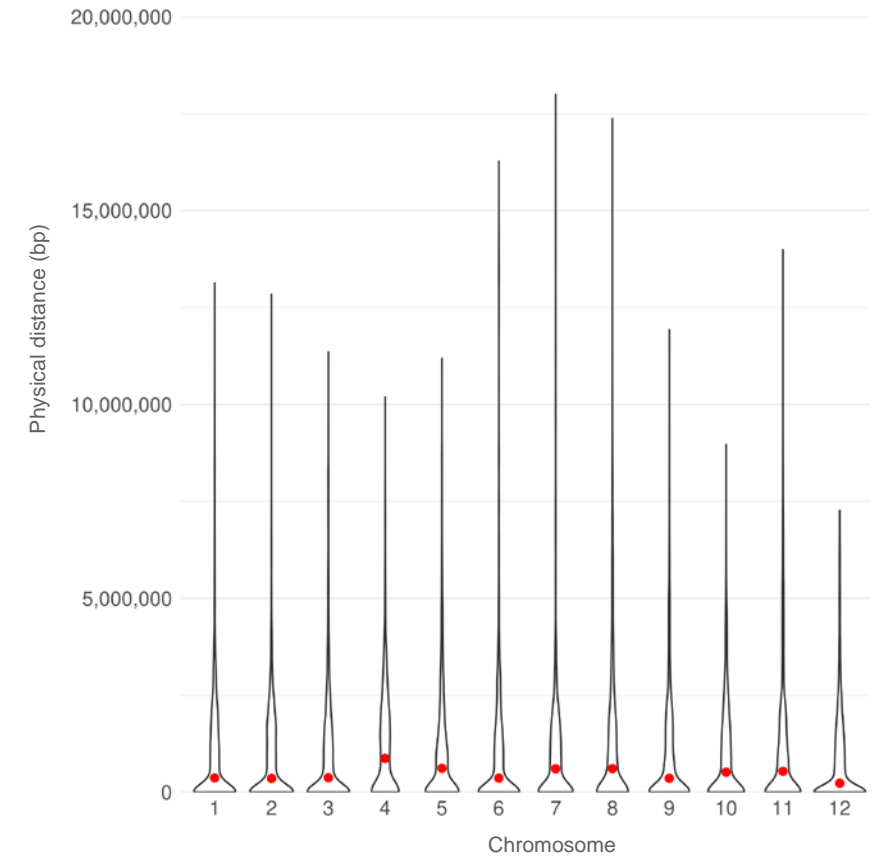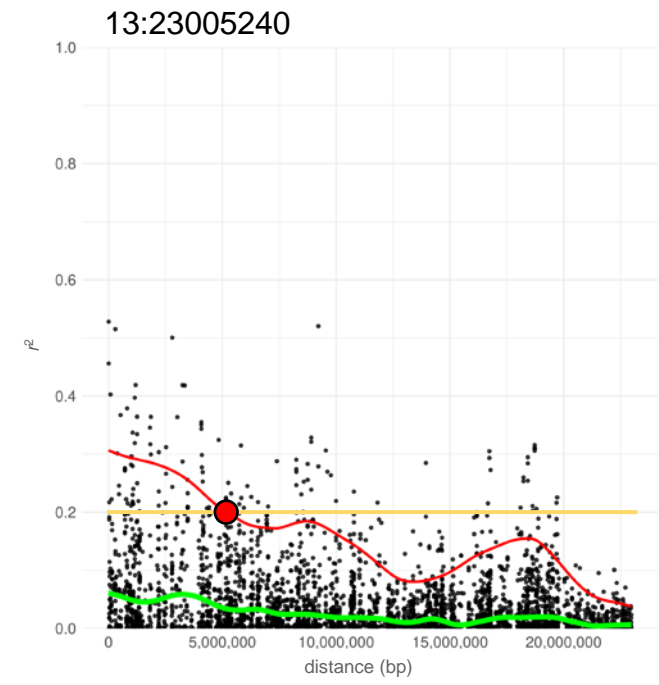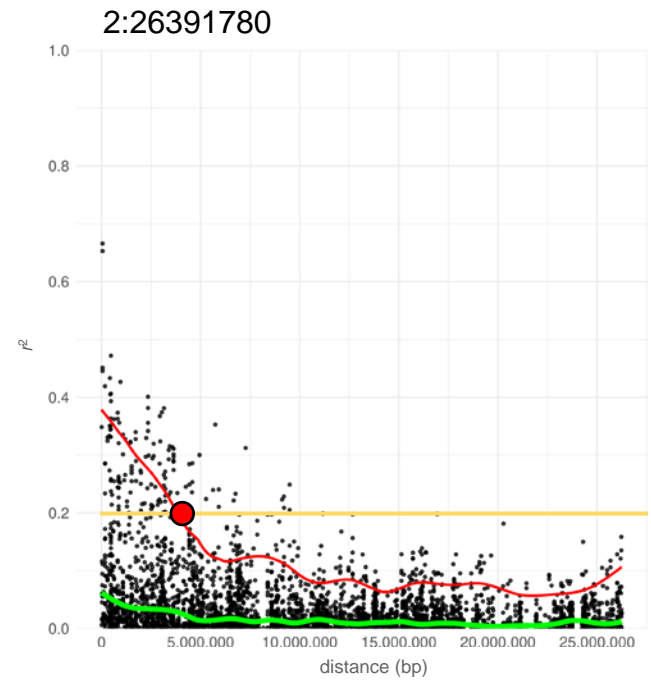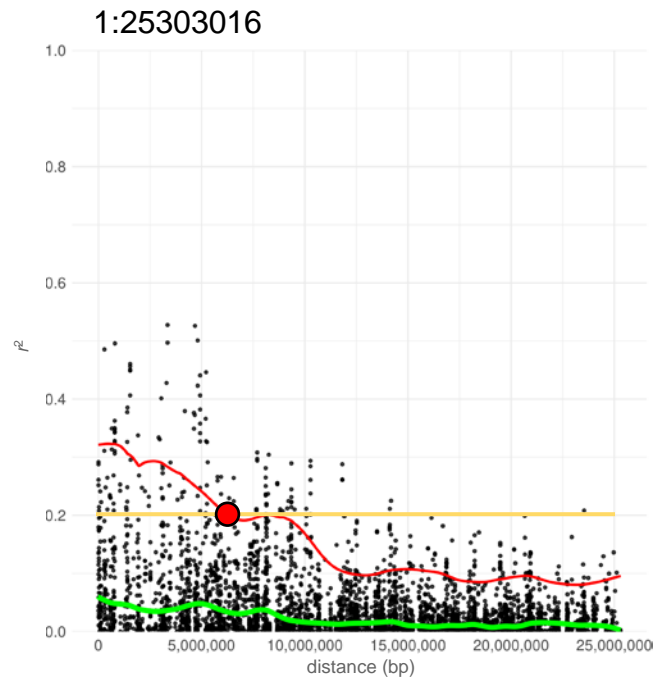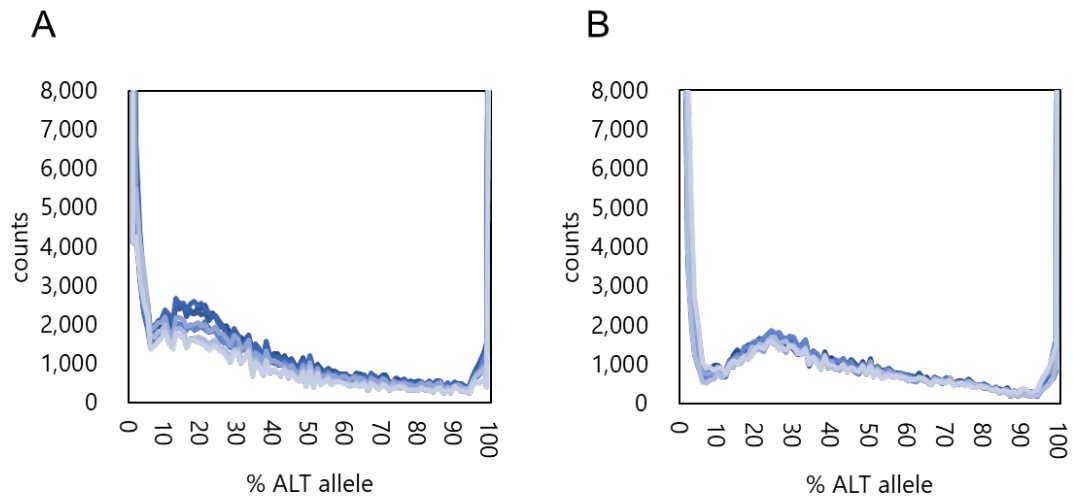Supplementary Fig. S1. Distribution of the alternative allele frequencies of (A) rabbiteye blueberry and (B) highbush blueberry. Sites were counted based on their alternative allele frequency in 1% bins. Lines represent different cultivars. Eight representative cultivars are presented in each figure.
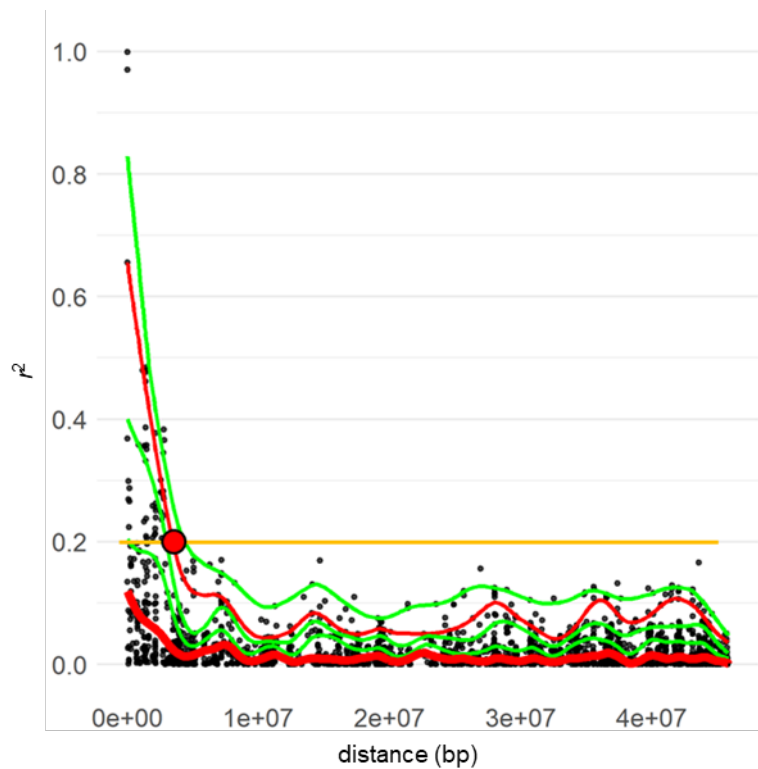
Supplementary Fig. S2. Representative patterns of the detection of the maximum association distance. The r^2 value was plotted for all SNPs on the same chromosome. Curves represent cubic splines fitted for the 99th, 95th, 90th, 75th, and 50th percentiles (top to bottom). The point where the fitted curve of the 95th percentile first decayed to $r^2 = 0.2$ was recorded for all SNPs.
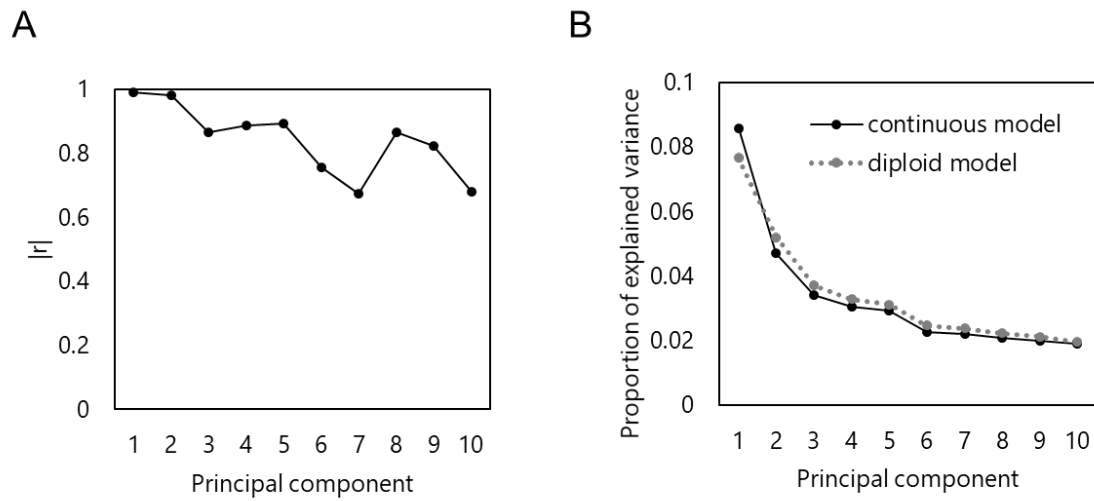
Supplementary Fig. S3. Comparison of the genotyping models used in this study. (A) Plot of the absolute Pearson correlation coefficients for the principal component scores based on the continuous and diploid models. The principal component analysis (PCA) involved the genotype matrices of 47,424 SNPs for the population with RE, NHB, and SHB. The principal component scores were calculated based on the probabilistic PCA. (B) Scree plot of the explained variance in the probabilistic PCA.
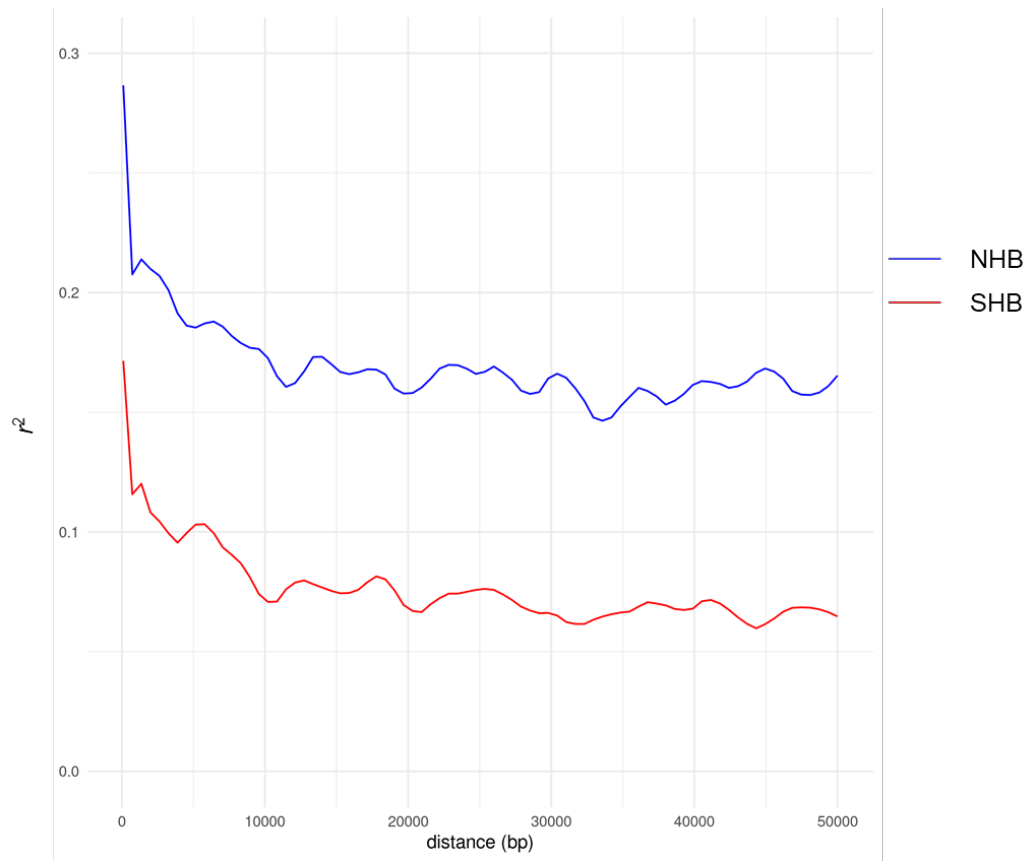
Supplementary Fig. S4. Patterns of LD decay in the highbush population. The r^2 value was plotted against the physical distance and regressed via loess smoothing, with span = 0.1.

Table 1. The number of accessions by the sampled locations

Institutions	SHB	NHB	RE	HH	CH
USDA-ARS, Southern Horticultural Laboratory	102	2	2	2	2
Kyoto University	2	3	2	0	0
Miyagi Prefectural Institute of Agriculture and Horticulture	0	10	0	0	0
Shizuoka Institute of Agriculture and Forestry	1	2	6	1	0

SHB: Sourthern highbush blueberry, NHB: Northern highbush blueberry, RE: Rabbiteye blueberry, HH: Half highbush blueberry, CH: complex hybrid

Supplementary Table S5. Genotype of the outlier loci associated with the continuous differentiation of SHB and NHB

Accession name	group	1:25303016	2:26391780	13:23005240	PC1
Amatsubu-boshi	NHB	1	-1	1	0.1863
Berkeley	NHB	NA	NA	NA	0.1721
Bluechip	NHB	1	-1	1	0.1910
BlueCrop	NHB	0	0	NA	-0.0071
Bluegold	NHB	NA	-1	NA	0.1550
BlueMaffin	NHB	NA	NA	NA	0.1105
Bounty	NHB	1	NA	0	0.0082
Brigitta	NHB	1	-1	1	0.1714
CarolineBlue	NHB	NA	-1	1	0.1758
Darrow	NHB	NA	-1	1	0.1966
Earliblue	NHB	1	-1	1	0.1419
Harrison	NHB	NA	-1	1	0.1783
Jersey	NHB	1	-1	1	0.1919
Lateblue	NHB	NA	-1	NA	0.1754
Meader	NHB	1	NA	1	0.1862
Nelson	NHB	1	-1	1	0.1944
Pender	NHB	NA	-1	1	0.1280
Avanti	SHB	0	0	1	-0.0075
AZ114	SHB	0	0	0	-0.0343
AZ131	SHB	0	-1	1	-0.0081
Biloxi	SHB	0	0	NA	-0.1342
Bluecrisp	SHB	0	-1	0	-0.0869
Blueridge	SHB	0	0	0	0.0187
Camellia	SHB	NA	-1	NA	-0.0259
CapeFear	SHB	0	NA	0	-0.0447
Endura	SHB	0	0	0	-0.0287
Georgia Dawn	SHB	0	0	1	0.0070
Georgia Germ	SHB	NA	0	1	0.0875
Gumbo	SHB	0	0	NA	-0.0913
Gupton	SHB	0	0	0	-0.0947
Indigocrisp	SHB	0	0	1	0.0079
Jubilee	SHB	NA	0	0	0.0161
Keecrisp	SHB	NA	-1	NA	0.1275
Legacy	SHB	NA	-1	0	0.0015
Meadowlark	SHB	NA	-1	0	-0.0580
MissLilly	SHB	0	0	1	0.0154
MS1050	SHB	0	0	0	-0.1542
MS1125	SHB	0	-1	0	-0.0106
MS1128	SHB	0	0	0	-0.0092
MS1129	SHB	0	0	0	-0.1084
MS1130	SHB	0	0	0	-0.1183
MS1135	SHB	0	0	0	-0.1050
MS1141	SHB	-1	0	NA	-0.1330
MS1150	SHB	0	0	NA	-0.0897
MS1269	SHB	0	1	NA	-0.1243
MS1318	SHB	0	0	0	-0.0674
MS1355	SHB	NA	0	1	0.0061
MS1375	SHB	0	-1	1	0.0470
MS1414	SHB	0	0	NA	-0.0176
MS1425	SHB	0	0	0	-0.0567
MS1428	SHB	1	-1	0	0.0440
MS1477	SHB	1	0	1	0.0646
MS1478	SHB	0	0	1	-0.0069
MS1480	SHB	0	-1	0	-0.0557
MS1499	SHB	0	NA	0	0.0026
MS1535	SHB	0	NA	0	0.0717
MS1561	SHB	0	0	0	-0.0859
MS1691	SHB	NA	0	1	0.0274
MS1694	SHB	0	0	0	-0.0913
MS1745	SHB	0	-1	1	0.0618
MS1749	SHB	0	0	0	-0.0302
MS1754	SHB	0	0	1	0.0615
MS2069	SHB	0	-1	0	-0.0285
MS2071	SHB	1	0	NA	-0.0175
MS2134	SHB	0	0	0	-0.1242
MS2137	SHB	NA	0	NA	0.0052

MS2177	SHB	0	0	0	-0.0892
MS2209	SHB	0	0	NA	-0.0223
MS2213	SHB	0	0	NA	-0.0152
MS2215	SHB	0	0	NA	-0.0890
MS2216	SHB	1	0	0	0.0286
MS2219	SHB	NA	0	0	-0.0579
MS2237	SHB	0	0	NA	-0.0902
MS2240	SHB	0	NA	1	0.0261
MS2244	SHB	NA	0	1	-0.0281
MS2248	SHB	0	0	0	-0.0004
MS2264	SHB	1	NA	NA	0.0005
MS2294	SHB	0	-1	1	0.0861
MS2296	SHB	0	0	0	-0.0721
MS2296	SHB	0	0	NA	-0.0217
MS2297	SHB	0	0	0	-0.0581
MS2298	SHB	0	-1	NA	-0.0120
MS2303	SHB	0	-1	NA	0.0023
MS2305	SHB	0	NA	0	-0.0705
MS2306	SHB	0	0	0	-0.0587
MS2307	SHB	0	0	0	-0.0999
MS2311	SHB	0	NA	0	-0.0963
MS2317	SHB	0	0	NA	-0.0077
MS2341	SHB	0	0	0	-0.1152
MS2384	SHB	0	0	NA	0.0109
MS2385	SHB	0	0	0	-0.1109
MS2387	SHB	0	0	NA	0.0134
MS354	SHB	0	0	0	-0.0687
MS711	SHB	0	0	NA	-0.0347
MS803	SHB	0	NA	0	-0.0210
MS813	SHB	1	0	0	0.0235
MS977	SHB	0	0	0	-0.0546
MS978	SHB	NA	0	0	-0.1048
NC8406-40	SHB	NA	0	0	-0.0897
O'Neal	SHB	1	-1	1	0.1634
Patricia	SHB	0	0	0	-0.1215
Pearl	SHB	0	0	0	-0.0411
Primadonna	SHB	0	0	NA	-0.0251
Raven	SHB	0	0	NA	0.0016
Rebel	SHB	0	0	0	-0.1165
Reveille	SHB	1	-1	NA	0.1810
Sampson	SHB	1	-1	1	0.1278
SanJoaquin	SHB	NA	0	NA	-0.0712
SantaFe	SHB	0	0	0	-0.0889
Sharpblue	SHB	NA	-1	NA	0.1094
SnowChaser	SHB	0	0	0	-0.0991
Springhigh	SHB	1	-1	0	0.0547
Star	SHB	NA	0	NA	-0.0567
Summit	SHB	0	0	0	-0.0769
SunshineBlue	SHB	0	0	0	0.0332
SuziBlue	SHB	0	0	0	-0.0615
SweetCrisp	SHB	1	0	1	0.0665
TX38	SHB	1	-1	1	0.1173
TX81	SHB	0	NA	0	0.0527
US487	SHB	0	-1	0	0.0056
US775	SHB	0	0	NA	-0.0330
Ventura	SHB	NA	-1	NA	0.0411

-1: homozygous for reference allele, 0: heterozygous, 1: homozygous for alternative allele, NA: missing.