Heredity (2021) 126:194–205
https://doi.org/10.1038/s41437-020-00362-0

the geneticssociety

ARTICLE

Genomic insight into the developmental history of southern highbush blueberry populations

Soichiro Nishiyama[1] · Mao Fujikawa[1] · Hisayo Yamane[1] · Kenta Shirasawa[2] · Ebrahiem Babiker[3] · Ryutaro Tao[1]

Received: 4 May 2020 / Revised: 16 August 2020 / Accepted: 18 August 2020 / Published online: 1 September 2020

Abstract

Interspecific hybridization is a common breeding approach for introducing novel traits and genetic diversity to breeding populations. Southern highbush blueberry (SHB) is a blueberry cultivar group that has been intensively bred over the last 60 years. Specifically, it was developed by multiple interspecific crosses between northern highbush blueberry [NHB, *Vaccinium corymbosum* L. ($2n = 4x = 48$)] and low-chill *Vaccinium* species to expand the geographic limits of highbush blueberry production. In this study, we genotyped polyploid blueberries, including 105 SHB, 17 NHB, and 10 rabbiteye blueberry (RE) (*Vaccinium virgatum* Aiton), from the accessions planted at Poplarville, Mississippi, and accessions distributed in Japan, based on the double-digest restriction site-associated DNA sequencing. The genome-wide SNP data clearly indicated that RE cultivars were genetically distinct from SHB and NHB cultivars, whereas NHB and SHB were genetically indistinguishable. The population structure results appeared to reflect the differences in the allele selection strategies that breeders used for developing germplasm adapted to local climates. The genotype data implied that there are no or very few genomic segments that were commonly introgressed from low-chill *Vaccinium* species to the SHB genome. Principal component analysis-based outlier detection analysis found a few loci associated with a variable that could partially differentiate NHB and SHB. These SNP loci were detected in Mb-scale haplotype blocks and may be close to the functional genes related to SHB development. Collectively, the data generated in this study suggest a polygenic adaptation of SHB to the southern climate, and may be relevant for future population-scale genome-wide analyses of blueberry.

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 (*Vaccinium angustifolium* Aiton) and highbush (*Vaccinium corymbosum* L.) blueberries ($2n = 4x = 48$) and the hexaploid rabbiteye blueberry [RE, *Vaccinium 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,

Associate editor: Pår Ingvarsson

Supplementary information The online version of this article (https://doi.org/10.1038/s41437-020-00362-0) contains supplementary material, which is available to authorized users.

✉ Soichiro Nishiyama
nishiyama.soichiro.8e@kyoto-u.ac.jp

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

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

[3] US Department of Agriculture, Agricultural Research Service, Thad Cochran Southern Horticultural Laboratory, 810 Hwy 26W, Poplarville, MS 39470, USA

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 twentieth century in the United States, 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). In addition, SHB was derived from crosses between tetraploid NHB and low-chill *Vaccinium* species native to Florida, USA (including *Vaccinium 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 United States in its pedigree (Brevis et al 2008).

Conventional blueberry breeding typically requires ~15 years for a new cultivar to be released and ~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 developing low-chill SHB cultivars and elucidating the genotype–phenotype relationships, little is known about the population structure and genomic evolution of cultivated blueberry. Clarifying the genetic diversity and population structure of blueberry may generate fundamental information relevant for association studies and useful for efficiently selecting suitable parents for hybridizations.

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

Materials and methods

Plant materials and genotyping

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

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 southern highbush blueberry, *NHB* northern highbush blueberry, *RE* rabbiteye blueberry, *HH* half-highbush blueberry, *CH* complex hybrid.

All sequences were preprocessed with a custom Python script (http://comailab.genomecenter.ucdavis.edu/index.php/Barcoded_data_preparation_tools). Sequences with a base-quality Phred score lower than 20 and with N bases were trimmed, and reads shorter than 35 bp were discarded. Clean reads were mapped to the *V. corymbosum* 'Draper' reference genome (Colle et al. 2019) using BWA-MEM (version 0.7.17) (Li and Durbin 2009). Despite the fact that the published 'Draper' tetraploid genome sequence consisted of four phased sets of the genome (Colle et al. 2019), the diversity across the homoeologous chromosomes remains to be clarified in a population level, and subgenome partitions are undistinguishable in the polyploid nature. Therefore, we selected the longest scaffold set representing each of 12 'Draper' homoeologous groups (Scaffolds 1, 2, 4, 6, 7, 11, 12, 13, 17, 20, 21, and 22, representing chromosomes 1–12) as representing 'Draper' genomic sequences to minimize the complexity. All sequences were confirmed to satisfy the following criteria: >1,000,000 mapped read counts and >0.5 mapping rate for each accession. The SAMtools program (version 1.9) (Li et al. 2009) with the mpileup -q 20 option and VarScan (version 2.3.9) (Koboldt et al. 2012) with the mpileup2snp mode were used to create the initial VCF file. We applied two types of SNP calling strategies, namely, the diploid model and the continuous model. The genotype data based on the diploid model did not include an allelic dosage for each variant. Regarding the continuous model, SNP genotypes were assigned a value between 0 and 1 based on ALT/(ALT + REF), where ALT and REF are the counts for the alternative allele-supporting reads and the reference allele-supporting reads, respectively (de Bem Oliveira et al. 2019).

Before producing the genotype matrix with the diploid model, we visualized the distribution of the alternative allele frequency for all RAD sites according to ploidy levels (Supplementary Fig. S1). Although a prominent simplex peak was detected, applying a single threshold for calling heterozygous variants was considered inappropriate because the homozygosity peak at 0% overlapped with the simplex peak. Thousands of ambiguous loci were detected even at the falling point of inflection between peaks. Therefore, to create high-confidence SNP genotype sets, we masked the ambiguous loci. In the diploid model, SNPs were called as heterozygous (5% < ALT% < 95%) or homozygous (0% ≤ ALT% ≤ 0.01% and 99.99% ≤ ALT% ≤ 100%), and the rest were masked (i.e., missing). The SNP loci were further filtered with VCFtools (Danecek et al. 2011) according to the following criteria: (1) minimum depth of coverage for each individual, 20, (2) biallelic locus only, (3) maximum missing data, 0.7, and (4) minor allele frequency, 0.1. Loci that were heterozygous for all individuals were further filtered with a custom Python script. These filtering steps were performed independently for a subset comprising all cultivars (SHB, NHB, RE, CH, and HH), a subset with SHB, NHB, and RE, a subset with only highbush cultivars, and a subset with only SHB to modulate the effect of the minor allele frequency and missing cutoffs. The SNP loci selected based on the diploid model were used in the continuous model. In addition to the SNP selection based on the diploid model, there was no further filtering specific to the continuous model to ensure a fair comparison between the models.

Population structure analysis

The SNP genotypes called with the diploid and continuous models underwent a probabilistic principal component analysis (PCA) with the R package pcaMethods (Stacklies et al. 2007). The probabilistic PCA is a probabilistic formulation of the PCA model with maximum likelihood estimation, and could deal with a dataset with the missing value. The results based on the diploid and continuous models were compared by calculating the Pearson correlation coefficient for each PC.

The SNPs called with the diploid model for all accessions were used to construct a phylogenetic tree according to the neighbor-joining method of MEGA X (Kumar et al. 2018), with 1000 bootstrap replications and a pairwise deletion option for missing data. Each SNP locus was represented by two bases in the input sequence, AA or BB for the homozygous genotype and AB for the heterozygous genotype. To evaluate the population structure of the blueberry collection, we performed a structure analysis with the STRUCTURE software (version 2.3.4) (Pritchard et al. 2000), which is reportedly more robust than other commonly used clustering programs for analyzing mixed ploidy populations (Stift et al. 2019). Regarding the input data, we coded the genotypes based on the diploid model as codominant markers with an unknown dosage as described by Meirmans et al. (2018) and Stift et al. (2019). For example, the genotype AB of a tetraploid individual based on the diploid model, which is a genotype derived from AAAB or AABB or ABBB, was coded as marker phenotype AB. To decrease the computation requirements, the loci were thinned, so two or more sites were not within 10 kb, and the resulting 6495 SNPs were used as the input data for the STRUCTURE software. In addition, SHB, NHB, and HH were coded as tetraploid, whereas RE and 'Pink Lemonade' were coded as hexaploid and 'Robeson' was coded as pentaploid. The *K* values ranging from 1 to 10 were evaluated using 100,000 MCMC iterations after 10,000 burn-in iterations to infer the population ancestry of genotypes in *K* predefined clusters. At least five runs for each *K* were conducted as replicates, and the replicates were summarized with CLUMPP (Jakobsson and

Rosenberg 2007). The delta K method (Evanno et al. 2005) of STRUCTURE HARVESTER (Earl and von Holdt 2012) was used to infer the optimal K value.

To analyze the genomic differentiation among cultivar groups, we performed PCA-based outlier detection analysis implemented with the R package pcadapt (version 4.0.2) (Luu et al. 2017), using SNP genotypes called with the diploid model. The assumption of pcadapt is that markers excessively related to the population structure are responsible for local adaptations. Notably, pcadapt can deal with the continuous separation of groups, which is expected in blueberry populations. To explore the loci driving genomic differentiation, the componentwise genome scans in pcadapt were applied for the PCs with distinct separation patterns among cultivar groups. The q value was used to control the false-positive discovery errors and was calculated with the R package q value (Storey et al. 2020). Loci with a q value lower than 0.1 were considered as candidate adaptive loci. To examine the distribution of outlier loci across the HB/RE and NHB/SHB genomes, Manhattan plots depicting the genomic positions of outlier SNPs and their respective significant association values $[-\log_{10}(P)]$ were prepared with the R package qqman (Turner 2018). Pairwise Weir and Cockerham's F_{st} estimate was calculated using VCFtools (Danecek et al. 2011).

Linkage disequilibrium

Squared correlation coefficients (r^2) of the SNP genotypes in each pair of SNPs on chromosomes were calculated based on the diploid and continuous models with the PLINK software (version 1.9) (Chang et al. 2015). The r^2 value was regressed with the physical distance via loess smoothing implemented in the R package ggplot2 (Wickham 2016), with span = 0.1.

To further evaluate potential associations between distant pairs, a haplotype block estimation based on the quantile regression was applied to the genotype matrix of the SHB group created with the continuous model. First, r^2 values for the correlation between each SNP and all other SNPs on a chromosome were calculated. The r^2 values were regressed against physical distance for each SNP. The regression was conducted based on quantile regression and smoothed with a cubic spline using the qsreg function implemented in the field R package (version 9.8.6) (Nychka et al. 2017), with lambda = 1e10. An evaluation of several quantile values for the regression revealed that the 95th percentile regression was the best fit for the observed maximum distance of associations (Supplementary Fig. S2). On the basis of the regression, the point where the 95th percentile regression curve first reached $r^2 = 0.2$ was recorded for each SNP.

Results

Genotyping and population structure

With our SNP selection criteria, 47,254 and 46,511 SNPs were detected in all populations and in the highbush populations, respectively. The overall average read depth across all the individuals in the SNP loci was 72.8. The PCA results based on the diploid and continuous models were highly correlated at least up to ten PCs (Supplementary Fig. S3), suggesting that even a very minor population structure could be detected by the diploid model in this population. Considering the relatively low read depth, the ease in handling, and compatibility with diverse software, we applied the diploid model genotype calling for most of the following experiments.

The phylogenetic tree revealed a distinct genetic cluster of RE cultivars (Fig. 1). NHB cultivars also clustered together with some exceptions. Some NHB accessions were far from the NHB cluster, and some SHB accessions were found in the NHB cluster. Specifically, the NHB cultivars 'Bluecrop' and 'Bounty' were far from the NHB cluster. The HH cultivars, which exhibit cold hardiness, clustered with the NHB cultivars, except for 'TopHat' that was far from the NHB cultivars. CH cultivars, which were bred with SHB accessions based on the pedigree, were found together with SHB.

The population structure analysis with the STRUCTURE software (Fig. 2) suggested that RE and NHB are relatively homogeneous, but SHB contains a considerably more admixed genetic background than RE and NHB. Considering the pedigree record of blueberry, the deep-blue part of Fig. 2a corresponds to the *V. virgatum* genome, whereas orange corresponds to the *V. corymbosum* genome. The origin of the other ancestral genomes was unclear, but gray is most likely the *V. angustifolium* genome because it represents half of the HH genomes. Moreover, the yellow, green, and light blue in $K = 5$ probably correspond to the wild *Vaccinium* genomes, including *V. darrowii* and *Vaccinium elliottii* because most of the SHB individuals possess these genomes. We also analyzed the genomic ancestry according to the different selection sites in the United States for the SHB group. The cultivars bred in North Carolina tend to have more of the *V. corymbosum* genome, whereas the cultivars bred in Florida and Georgia tend to be more admixed. The cultivars 'O'Neal' and 'Reveille', which are widely distributed as SHB, largely consisted of the presumed *V. corymbosum* ancestral genome (Fig. 2b). A single high delta K value was obtained at $K = 9$, and the delta K values were stably low for the other tested K values (Fig. 2c), providing a possibility that the nine ancestral genomes underlie the blueberry gene pool.

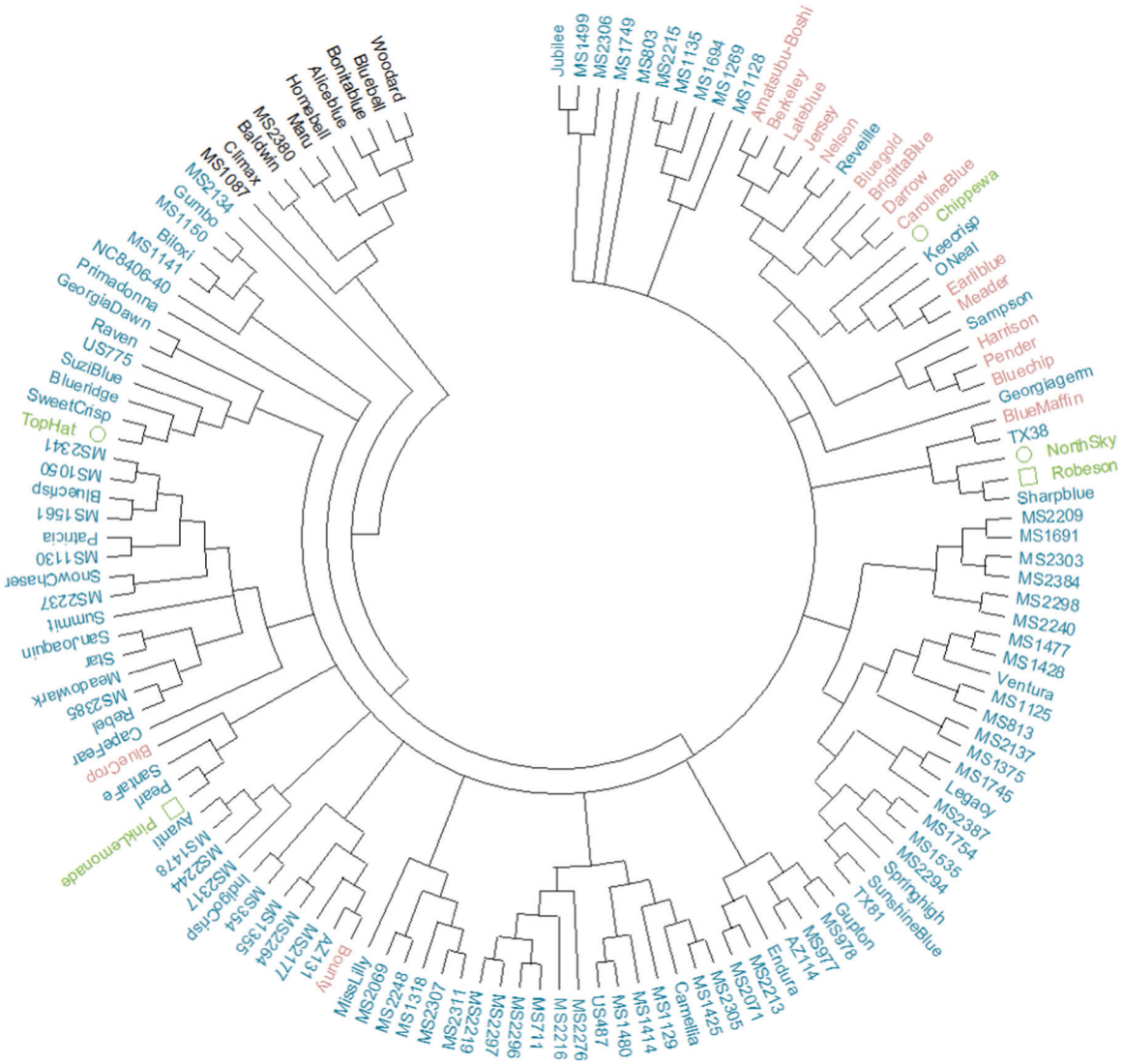

Fig. 1 Consensus neighbor-joining phylogenetic tree of the blueberry population. The tree was constructed based on the genotype data for 47,254 genome-wide SNPs in 137 accessions. Black, blue, and red represent rabbiteye blueberry (RE), southern highbush blueberry (SHB), and northern highbush blueberry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

Characterization of the genomic differentiation among cultivar groups

We first analyzed the genomes to identify diagnostic loci that could distinguish between SHB and NHB based on the SNP data. Despite the interspecific origin of SHB, there was no allele present in all SHBs, but lacking in NHB. This was confirmed with allowing 50% missing data in each group based on the diploid model matrix. Therefore, we assumed that the genetic differentiation might not be significant among blueberry cultivar groups, at least between SHB and NHB. Pairwise F_{st} estimate indicated lower genetic differentiation between SHB and NHB than between SHB and RE, or between NHB and RE (Supplementary Table S2). We next applied a PCA-based whole-genome scan to uncover the genomic differentiation between cultivar groups. In a PCA plot for HB and RE populations, HB and RE were clearly distinguishable along the first PC (PC1) (Fig. 3a). A Manhattan plot depicting the significant association values [$-\log_{10}(P)$] of the outlier loci revealed many peaks spanning all chromosomes based on the pcadapt componentwise mode for PC1 (Fig. 3b). In contrast, NHB and SHB were not divided into independent clusters, but were continuously distributed along the PC1 score in a PCA plot for the NHB and SHB populations (Fig. 4a). There were 11 SNPs fulfilling the q value threshold on chromosomes 1, 2, and 8 (Fig. 4b). Although the detected four SNPs on chromosome 1 and the six SNPs on chromosome 8 spanned across 4.9 and 8.7 Mb, respectively, the genotypes in the

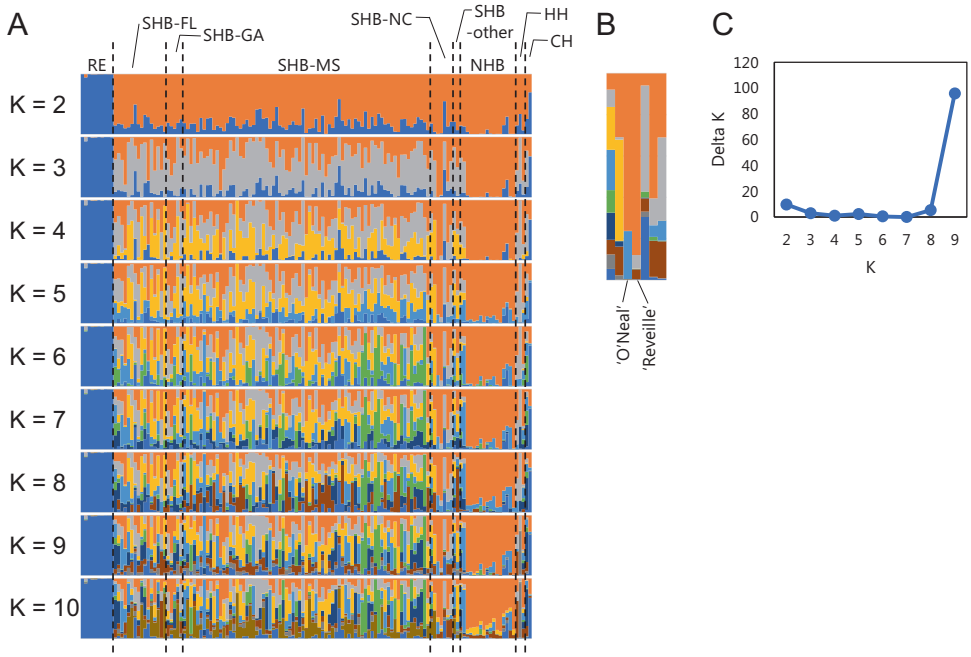

Fig. 2 Proportion of the ancestry of blueberry. **a** Proportion of the ancestry of the individuals inferred with the STRUCTURE software. *K* values ranging from 2 to 10 were plotted. Each individual is presented as a vertical bar. RE, SHB, NHB, HH, and CH represent rabbiteye, southern highbush, northern highbush, half-highbush, and complex hybrid blueberries, respectively. FL, GA, MS, and NC represent Florida, Georgia, Mississippi, and North Carolina, respectively, and indicate the USA states producing the SHB cultivars. **b** Plot of the proportion of the ancestry of SHB bred at North Carolina inferred with $K = 9$. **c** Evanno's delta *K* plotted against *K*.

population were highly correlated (Supplementary Tables S3 and S4). On the basis of the genotype correlations, we considered the four SNPs on chromosome 1 and the six SNPs on chromosome 8 to be on the haplotype blocks. The genotype scores of the three loci were modestly correlated with the PC1 value (Fig. 4c). At the outlier locus (13:23005240) with the lowest *P* value, most of the SHB with the same genotype as NHB (homozygous for the alternative allele) was bred with NC 1528, NC 1524, 'Bluechip', or 'Sharpblue' according to their pedigree records (Supplementary Table S5). Among NHB cultivars, the heterozygous genotype in the outlier loci was observed only in 'Bluecrop' at 1:25303016 and 2:26391780, and 'Bounty' at 13:23005240 (Supplementary Table S5).

Characterization of the chromosome-wide allelic association

The LD in the NHB and SHB populations decayed to $r^2 = 0.2$ in <10 kb (Supplementary Fig. S4). Although the LD decay occurred slightly faster in SHB than in NHB, the LD decay patterns were similar between these two populations (Supplementary Fig. S4). Despite the observed rapid LD decay, there were many substantially associated SNP pairs with Mb-scale distances (Supplementary Fig. S2). Figure 5 presents a plot of the maximum distances of the substantial allelic associations for each SNP. In the SHB population, long potential associations in SNP pairs separated by more than 5 Mb were found on all chromosomes (Fig. 5a). These associations tended to be located at the center of chromosomes, except for chromosomes 6 and 11. In addition, apparent secondary peaks were also detected for several chromosomes, including chromosomes 5, 6, and 11. The genome-wide median maximum association distance calculated based on the 95th percentile was 474 kb, ranging from 231 kb on chromosome 12 to 871 kb on chromosome 4 (Fig. 5b). The outlier loci associated with the separation between SHB and NHB were located on the Mb-scale haplotype blocks (Fig. 6), with a considerably greater distance than the genome-wide median.

Discussion

Highbush blueberry originated and was domesticated in northern United States, but it is now cultivated worldwide. The available highbush blueberry cultivars adapted to warm climates are the result of extensive breeding, including interspecific hybridizations, which explains the mixture of genomes in these cultivars. To increase the efficiency of crossing and selection strategies, blueberry breeders, especially those with limited genetic resources, may benefit from the genetic characterization of the extremely diverse *Vaccinium* population. In this study, we examined blueberry population genetics using genome-wide SNP data of cultivars/accessions representing most of the currently cultivated lines in their pedigree. We also analyzed the possible genomic differentiation among blueberry cultivar groups.

Genetic differentiation between RE and highbush populations

The clear separation of highbush cultivars from the RE group based on the phylogenetic relationships and PCA

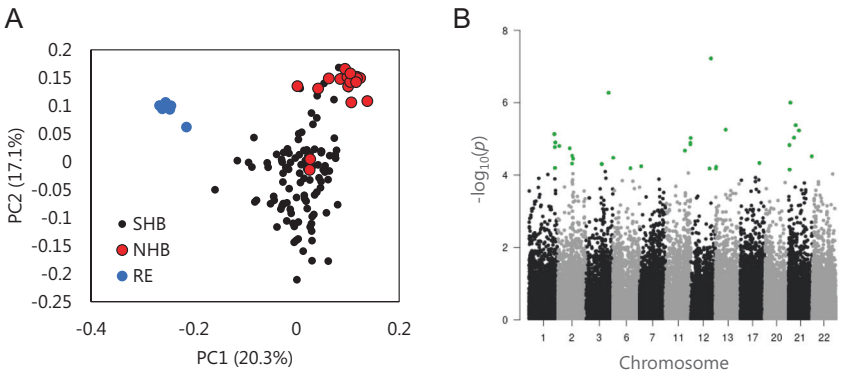

Fig. 3 Population differentiation between RE and highbush blueberry. a Principal component analysis based on the SNP data of RE, SHB, and NHB individuals generated with the diploid model. **b** Manhattan plot of the outlier loci associated with the first principal component in panel (**a**), inferred with the componentwise genome scan implemented in the pcadapt software. Green dots represent significantly associated SNPs.

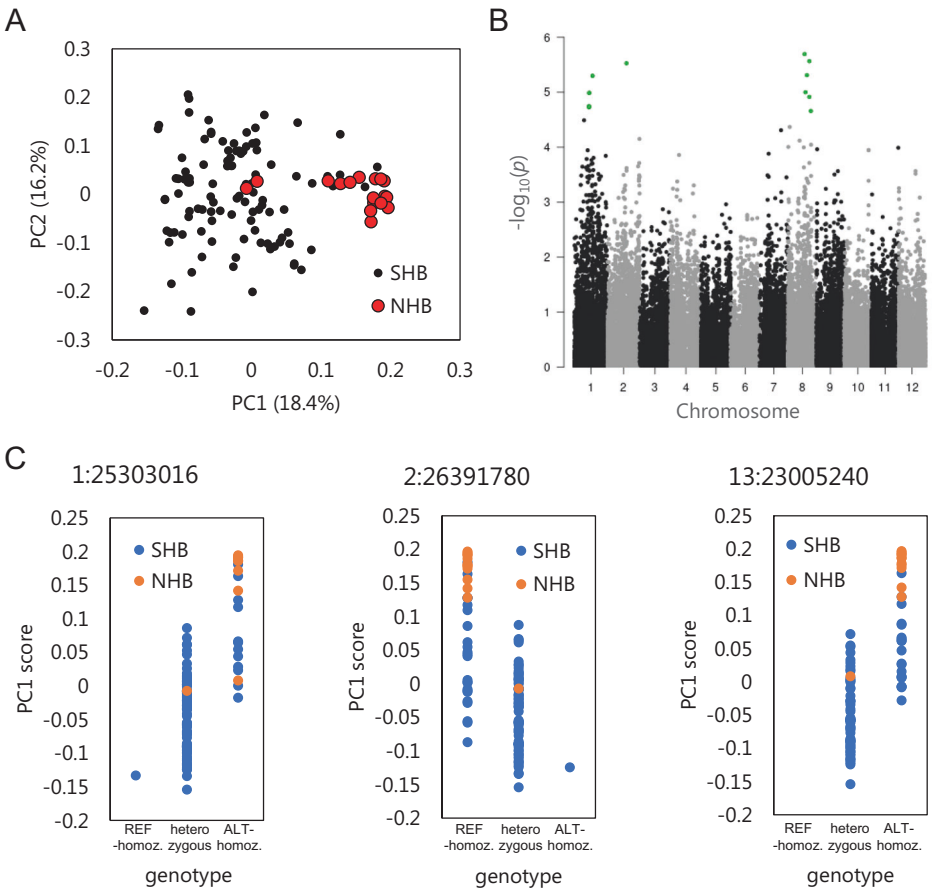

Fig. 4 Population differentiation among highbush blueberries. a Principal component analysis based on the SNP data generated with the diploid model for a subpopulation of SHB and NHB. **b** Manhattan plot of the outlier loci associated with the first principal component in panel (**a**), inferred with the componentwise genome scan implemented in the pcadapt software. Green dots represent significantly associated SNPs. **c** Plot of the PC1 scores according to the genotypes of the three outlier loci.

results is consistent with the findings of previous studies (Bian et al. 2014; Campa and Ferreira 2018; Bassil et al. 2020). Despite the widespread contribution of RE in the SHB pedigree (Brevis et al. 2008), the outlier SNPs associated with the separation between RE and highbush blueberries were not localized to specific genomic regions, but were distributed throughout the genome (Fig. 3b). These results are consistent with the notion that the initial NHB and RE cultivars developed independently, and RE was subsequently used to generate SHB in the NHB genomic background. In the STRUCTURE analysis, the presumed *V. virgatum* genome was separated at $K = 2$ (Fig. 2). This is in accordance with the PCA result, and suggests the existence of a distinct feature in the RE genome. Notably, RE

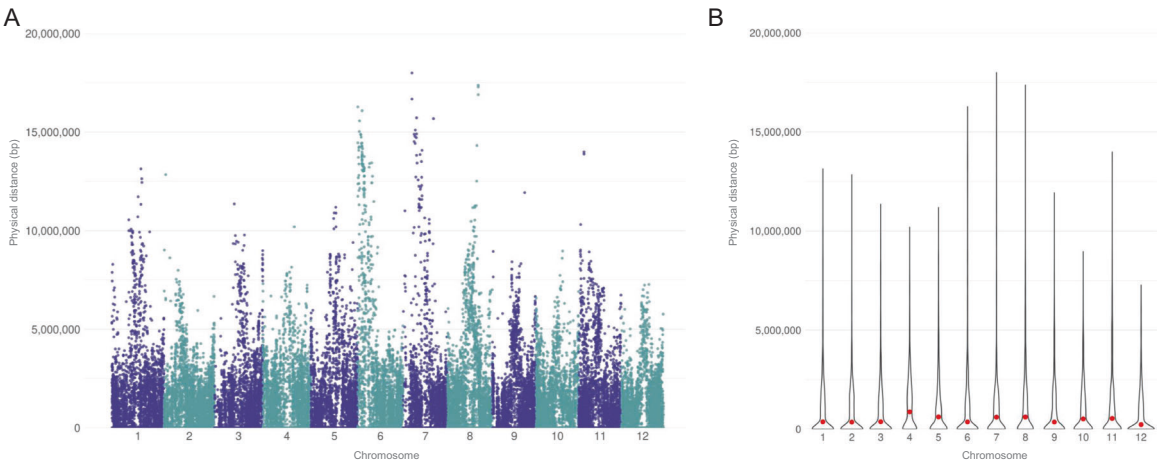

Fig. 5 Chromosome-wide allelic associations in the SHB population. a Genome-wide distribution of allelic associations. The maximum distance with a substantial association ($r^2 = 0.2$) estimated with the 95th percentile regression was plotted for each SNP. (**b**) Violin plot indicating the maximum association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

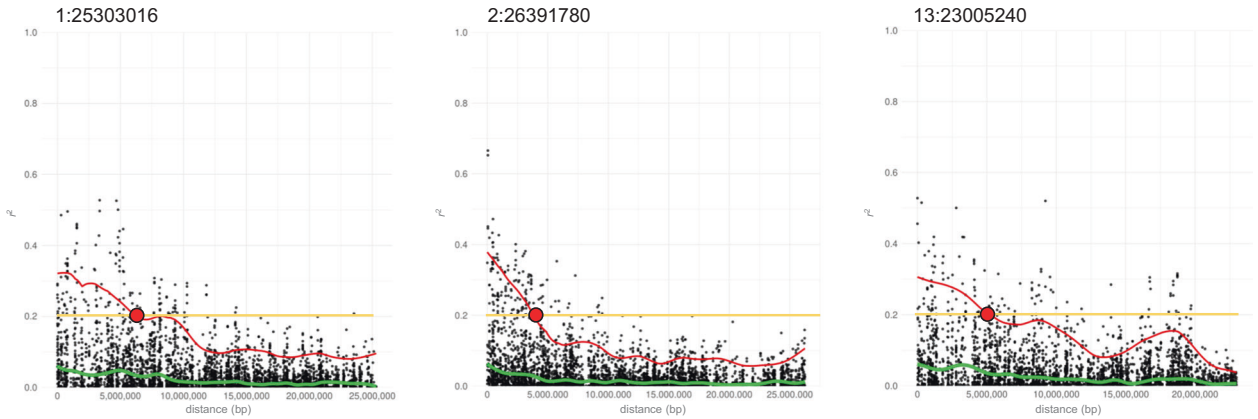

Fig. 6 Pairwise allelic genotype correlations of the outlier loci associated with the separation of SHB and NHB. Red and green lines represent a cubic spline fitted for the 95th and 50th percentiles, respectively. Red dots represent the points where the fitted curve of the 95th percentile first decayed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

cultivars appeared to comprise mostly the presumed *V. virgatum* genome, with no contribution from the *V. corymbosum* genome, although the inverse was previously reported (Brevis et al. 2008) and revealed in the clustering data (Fig. 2).

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

In contrast to the clear separation between RE and the highbush blueberries, the relationship between SHB and NHB is complex, with neither the PCA nor the phylogenetic analysis uncovering a clear separation. The continuous relationship between SHB and NHB may be explained by the complex interspecific crosses and recurrent backcrosses related to SHB development. The detection of only three significant outlier loci associated with the continuous relationship further suggests a weak population differentiation. The results also imply that the genotype information includes the record of the directed selection of SHB, and the outlier loci may have been functional during SHB development. The genotype patterns of the outlier loci appear to be associated with specific functions. For example, at the most significant locus (13:23005240), all SHB accessions with the same allele as NHB accessions (homozygous for the alternative allele) have NC 1528, NC 1524, 'Bluechip', or 'Sharpblue' in their pedigree (Supplementary Table S5). This may indicate that breeders favored alleles from *V. angustifolium* and *V. corymbosum* over those from low-chill *Vaccinium* species. This is also consistent with the known SHB breeding history, in which the initial low-chill SHB cultivars developed in Florida were further crossed to adapt to colder regions (Ehlenfeldt et al. 1995). The genomic admixture in SHB revealed by the STRUCTURE analysis (Fig. 2) likely reflects the genomic segments introgressed from other *Vaccinium* species. The STRUCTURE data

further revealed that the ancestral genomic composition varied depending on the original selection locations (Fig. 2). The cultivars bred in North Carolina, which is closer to the NHB production area than the other examined regions, possessed more of the presumed *V. corymbosum* genome than the cultivars bred in other regions. This is probably because of the targeted selection of the expected cold hardiness of *V. corymbosum*. In contrast, the cultivars bred in Florida, which is the southernmost region examined in this study, had a more mixed ancestry (i.e., admixed population). Florida is where SHB breeding was initiated because breeders needed to develop cultivars adapted to the climate in this state, which is far from where highbush blueberries originated. Therefore, the observed admixture can be attributed to the local adaptation efforts. Thus, SHB is difficult to define at the genome level; however, we identified different breeding directions during SHB development, likely because of the diversity in local breeding centers. In addition, we determined that the ancestry can be traced based on genomics, even in polyploid blueberry.

We also confirmed the absence of a genomic region satisfying a strict threshold for distinguishing SHB from NHB (i.e., a homozygous site in all NHB accessions that was heterozygous or homozygous for the alternative allele in all SHB accessions). This suggests a lack of or only a few introgressed genomic segments that are shared by all SHB accessions. The same result was obtained when we excluded a relatively high-chill SHB cultivar (Summit) from the analysis. Considering that the low-chill SHB accessions used in this study could not be distinguished from the other accessions in the highbush population, we hypothesized that the adaptation of SHB to the southern region was achieved through factors under polygenic control. Local adaptations with polygenic factors are common in many plant species (Flood and Hancock 2017; Wisser et al. 2019). In this situation, a shift in the allele frequency at many loci drives the adaptation (Stephan 2016), which is consistent with the observed genotype patterns and population structure results. The outlier loci detected in the genome scan may include loci controlling the traits mediating the adaptation of SHB to the southern region. The observed long-range genotype associations of the outlier loci (Fig. 6) support the allele selection preferences of the outlier loci. Our preliminary examination of the chilling requirement phenotype indicated a lack of a significant association between the chilling requirement and the genotype of the loci (data not shown). Ongoing association studies will hopefully elucidate the adaptation process.

Some of the results that were inconsistent with the general population features in the clustering, phylogenetic analysis, and genome scan (Figs. 1, 2, and 4) can be explained by the hybridization history. Pedigree of HH cultivar 'TopHat', which was far from the NHB cluster (Fig. 1), is Mich. 19-H x 'Berkeley'. 'Berkeley' was developed with three of the four parents ('Stanley', 'Jersey', and 'Pioneer') of 'Bluecrop', which was extensively used for the development of SHB (Brevis et al. 2008). The distinction of 'TopHat' from the NHB cultivars is also consistent with the previous study (Bian et al. 2014). The mixture of SHB and NHB in the phylogenetic analysis (Fig. 1) is probably related to the repeated hybridizations or shared polymorphisms in their ancestors. The detection of NHB 'Bluecrop' and 'Bounty' in the SHB cluster (Fig. 1) is likely due to the contribution of 'Bluecrop' and Crabbe-4 genomes to the SHB population, as previously suggested (Brevis et al. 2008). Crabbe-4, a wild *V. corymbosum* clone that is not present in the pedigree of most of NHB cultivars, was used to develop NHB 'Murphy', a parent of 'Bounty'. This notion is also consistent with the detection of the heterozygous genotype of 'Bluecrop' and 'Bounty' at the outlier loci (Fig. 4, Supplementary Table S5). Moreover, SHB cultivars 'O'Neal' and 'Reveille', which appeared to largely consist of the ancestral *V. corymbosum* genome based on the clustering analysis (Fig. 2), were likely to have lower-than-expected (according to the pedigree records) genomic contribution from the other *Vaccinium* species (Brevis et al. 2008). This can be explained by the elimination of alleles derived from interspecific hybridizations during the development, considering that the interspecific hybridizations were made several generations prior to the development of 'O'Neal' and 'Reveille' (Ballington et al. 1990; Cummins 1991).

Mb-scale linkage disequilibrium in the SHB population

The pattern and extent of LD are important factors for explaining the past events in a population and for designing association-mapping studies. In addition, LD is a sensitive indicator of the population genetic forces influencing genomic structures (Slatkin 2008), and it is affected by multiple factors, including the ploidy level and introgression. Regarding blueberry, although several association studies have been attempted, only a few investigations have focused on the extent of LD. Ferrão et al. (2018) reported that the estimated genome-wide LD decay in a tetraploid blueberry breeding population was 73–80 kb, which was based on genotypic correlations, with genotypes called with the diploid and tetraploid models. In the current study, we estimated a less extensive LD for the SHB and NHB groups with the diploid model (Supplementary Fig. S4). However, this may have severely underestimated the population LD extent because repulsion-phase marker pairs, which are less informative in polyploids, were averaged together with more informative pairs. In fact, we identified SNP pairs with allelic associations with distances of several Mb in the SHB

population. Therefore, we applied quantile regression with empirically determined parameters to characterize the genome-wide pattern of allelic genotype correlations. A similar methodology using quantile regression was previously applied in the LD survey of sugar beet and tetraploid potato (Adetunji et al. 2014; Sharma et al. 2018). By using this method, we proved the existence of long-lasting association pairs with distances of up to several Mb in all chromosomes (Fig. 5). These long-lasting associations should be consistent with the SHB breeding history, considering the recent origin and the widespread genetic contribution of wild *Vaccinium* clones (Brevis et al. 2008). The pattern of the distribution of the LD estimates across the genome may be related to different recombination frequencies and large structural variations. The predominant localization of the long-lasting association pairs at the center of chromosomes may be due to the suppression of recombination in the centromeric region. In contrast, the distinct distribution pattern observed for chromosome 6 may be related to the rearrangement or misassembly in the 'Draper' reference genome (Colle et al. 2019). Moreover, apparent secondary peaks in addition to those at the centromeric regions were detected for several chromosomes. The data also suggest the existence of haplotype blocks that are longer than expected (Supplementary Figs. S2 and S3), which may decrease the genotyping costs of a future genome-wide association study (GWAS). Considered together, the allelic associations detected by the quantile regression method in this study appear to be useful for characterizing the genomic features of tetraploid blueberry. To increase the resolution and the accuracy of LD estimates, it is essential that future studies elucidate the inheritance mode and produce genetic maps on a genome-wide scale, as has been done for potato (Vos et al. 2017).

Our data revealed a less extensive LD in SHB than in NHB in our highbush population (Supplementary Fig. S4). There are two potential explanations for this finding. First, compared with the SHB accessions, there were fewer and less diverse NHB accessions. Second, the SHB accessions had more founder haplotypes than the NHB accessions because of interspecific hybridizations. There are reportedly two different genotypes for Florida 4B, which contributed considerably to the SHB genome (Bassil et al. 2018). Specifically, CVAC 1790, which is one of the Florida 4B genotypes that has been widely used during SHB development, is the result of an interspecific hybridization between the wild diploid species in Florida (Bassil et al. 2018).

Population structure inference of cultivated polyploid blueberries

It is known that allele dosage of polyploid species significantly affects calculation of allele frequency, which is fundamental to many population genetic-based inferences (Cockerham 1973; Dufresne et al. 2014). However, in many cases, there are still difficulties regarding the dosage genotyping, especially in genotyping accuracy, costs, and software/parameter compatibility (Gerard et al. 2018; Meirmans et al. 2018). Herein, as the result of PCA highly matched between the diploid and continuous models (Supplementary Fig. S3), we considered that the genotype matrix based on the diploid model represented most of the population structural information present in the population. The observed high consistency between the two can relate to diversity in the presence/absence of alleles, which is assumed in the situation of less generation cycles from the domestication and the potential allopolyploid origin of blueberry (Colle et al. 2019).

Up to this time, seven *Vaccinium* species, *V. darrowii*, *V. elliottii*, *Vaccinium tenellum*, *V. angustifolium*, *V. corymbosum*, *Vaccinium constablaei*, and *V. virgatum*, are recognized as a genomic backbone of cultivated polyploid blueberries (Brevis et al. 2008; Ballington 2009). In addition, *Vaccinium myrtilloides* and *Vaccinium pallidum* have partially but substantially contributed to the blueberry gene pool (Ballington 2009). Thus, it is possible to interpret that the optimal *K* value 9 in the clustering (Fig. 2) is fairly matched with the number of species underlying the development of cultivated polyploid blueberries. However, species delimitation within the *Vaccinium* genus is still controversial, and hybridization among species in section *Cyanococcus* is common in nature. Thus, this point is unable to be experimentally assessed, unless the diversity of the ancestral species is clarified. Future works with combining the ancestral species and full dosage information of cultivated blueberries may facilitate deeper understanding of the genomic origin of cultivated blueberries.

Conclusion

In this study, an analysis of the population genetics of diverse blueberry populations clarified the genomic ancestry of blueberry. The general trends revealed by the results presented herein include a homogeneous genomic background in RE and NHB, in contrast to the admixed background of SHB, which is consistent with the recorded history of blueberry breeding. The structural characterization and scanning of the genomes indicate that SHB development likely involved directed selection. This is probably related to the independence of the breeding projects conducted by various breeding centers, which were influenced by the local climate and breeder strategies. Despite the extensive breeding and admixed nature of the SHB population, there appears to be no introgressed genomic segment common to all SHB cultivars.

Collectively, we hypothesize that polygenic factors affected the adaptation of SHB to the climate in the southern United States. The detected outlier loci were associated with the continuous separation between NHB and SHB, and may be considered as part of the alleles mediating the adaptation of SHB. To the best of our knowledge, none of the loci presented in this study match loci detected in previous GWAS/mapping studies. Future population-scale genomic investigations of diverse NHB accessions as well as accurate association analyses regarding the adaptive traits may help to further clarify the process underlying the adaptation of SHB.

Data availability

The raw ddRAD-seq data analyzed in this study have been submitted to the DDBJ Sequence Read Archive (accession number DRA009951).

Acknowledgements This work was supported by a Grant-in-Aid for Fostering Joint International Research (B) (19KK0156) to SN, HY, and RT from the Japan Society for the Promotion of Science. We thank Kanako Ishii, Satoru Murakami (Shizuoka Prefectural Research Institute of Agriculture and Forestry), and Masakazu Shoji (Miyagi Prefectural Agriculture and Horticulture Research Center) for providing leaves of the blueberry cultivars used in this study. We thank Edanz Group (https://en-author-services.edanzgroup.com/) for editing a draft of this paper.

Compliance with ethical standards

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Adetunji I, Willems G, Tschoep H, Bürkholz A, Barnes S, Boer M et al. (2014) Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. Theor Appl Genet 127:559–571. https://doi.org/10.1007/s00122-013-2239-x

Ballington JR (2009) The role of interspecific hybridization in blueberry improvement. Acta Hortic 810:49–60

Ballington JR, Mainland CM, Duke SD, Draper AD, Galletta GJ (1990) 'O'Neal' southern highbush blueberry. HortScience 25:711–712

Bassil N, Bidani A, Hummer K, Rowland LJ, Olmstead J, Lyrene P et al. (2018) Assessing genetic diversity of wild southeastern North American *Vaccinium* species using microsatellite markers. Genet Resour Crop Evol 65:939–950

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

de Bem Oliveira I, Resende MFR, v. Ferrão LF, Amadeu RR, Endelman JB, Kirst M et al. (2019) Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. G3 9: g3.400059.2019. https://doi.org/10.1534/g3.119.400059

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

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

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

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

Cappai F, Benevenuto J, Ferrão L, Munoz P (2018) Molecular and genetic bases of fruit firmness variation in blueberry—a review. Agronomy 8:174. https://doi.org/10.3390/agronomy8090174

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

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

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

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

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

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

Cummins JN (1991) Register of new fruit and nut varieties. HortScience 26:951–986

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al. (2011) The variant call format and VCF tools. Bioinformatics 27:2156–2158

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

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

Earl DA, von Holdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genet Resour 4:359–361

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

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol 14:2611–2620

Ferrão LF, Benevenuto J, Oliveira I, de B, Cellon C, Olmstead J, Kirst M et al. (2018) Insights into the genetic basis of blueberry

fruit-related traits using diploid and polyploid models in a GWAS context. Front Ecol Evol 6:107. https://doi.org/10.3389/fevo.2018.00107

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

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

Hancock J, Lyrene P, Finn C, Vorsa N, Lobos G (2008) Blueberries and Cranberries. In: Hancock JF (ed) Temperate Fruit Crop Breeding. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6907-9_4

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

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568–576

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

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

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

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

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

Meirmans PG, Liu S, van Tienderen PH (2018) The analysis of polyploid genetic data. J Heredity 109:283–296. https://doi.org/10.1093/jhered/esy006

Moerman DE (1998) Native American ethnobotany. Timber Press, Portland, Oregan

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

Nychka D, Furrer R, Paige J, Sain S (2017) fields: Tools for spatial data. R package version 10.3. https://github.com/NCAR/Fields

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Sharma SK, MacKenzie K, McLean K, Dale F, Daniels S, Bryan GJ (2018) Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. G3 8:3185–3202. https://doi.org/10.1534/g3.118.200377

Sharpe RH, Darrow GM (1959) Breeding blueberries for the Florida climate. Proc Fla State Hort Soc 72:308–311

Shirasawa K, Hirakawa H, Isobe S (2016) Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato. DNA Res 23:145–153. https://doi.org/10.1093/dnares/dsw004

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485. https://doi.org/10.1038/nrg2361

Song GQ, Hancock F (2011) Vaccinium. In: Kole C (ed) Wild crop relatives: Genomic breeding resource. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16057-8_10

Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007) pcaMethods: a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23:1164–1167. https://doi.org/10.1093/bioinformatics/btm069

Stephan W (2016) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Mol Ecol 25:79–88. https://doi.org/10.1111/mec.13288

Stift M, Kolář F, Meirmans PG (2019) Structure is more robust than other clustering methods in simulated mixed-ploidy populations. Heredity 123:429–441. https://doi.org/10.1038/s41437-019-0247-6

Storey JD, Bass AJ, Dabney A, Robinson D (2020) qvalue: Q-value estimation for false discovery rate control. R package version 2.20.0. http://github.com/jdstorey/qvalue

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066. https://doi.org/10.1126/science.277.5329.1063

Turner SD (2018) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J Open Source Softw 3:731. https://doi.org/10.21105/joss.00731

Vos PG, Paulo MJ, Voorrips RE, Visser RGF, van Eck HJ, van Eeuwijk FA (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theor Appl Genet 130:123–135. https://doi.org/10.1007/s00122-016-2798-8

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. https://ggplot2.tidyverse.org

Wisser RJ, Fang Z, Holland JB, Teixeira JEC, Dougherty J, Weldekidan T et al. (2019) The genomic basis for short-term evolution of environmental adaptation in maize. Genetics 213:1479–1494. https://doi.org/10.1534/genetics.119.302780