PROF. PATRICIO R. MUNOZ (Orcid ID : 0000-0001-8973-9351)

Article type : MS - Regular Manuscript

GENOME-WIDE ASSOCIATION OF VOLATILES REVEALS CANDIDATE LOCI FOR BLUEBERRY FLAVOR

Luís Felipe V. Ferrão[1co]; Timothy S. Johnson[2co]; Juliana Benevenuto[1]; Patrick P. Edger[3]; Thomas A. Colquhoun[2]; Patricio R. Munoz[1]✉

[1] Blueberry Breeding and Genomics Lab, Horticultural Sciences Department, University of Florida, Gainesville, FL 32611

[2] Environmental Horticulture Department, Plant Innovation Center, University of Florida, Gainesville, FL 32611

[3] Department of Horticulture, University of Michigan, Michigan State University, East Lansing, MI 48824

[co] authors contributed equally

Address:

✉ Blueberry Breeding and Genomics Lab, Horticultural Sciences Department, University of Florida, Gainesville, FL 32611.

Phone: (352) 273-4837

email: p.munoz@ufl.edu

Received: *29 October 2019*

Accepted: *21 January 2020*

Orcid:

Luís Felipe V. Ferrão (https://orcid.org/0000-0002-9655-4838)
Juliana Benevenuto(https://orcid.org/0000-0002-4698-2738)
Patrick Edger (https://orcid.org/0000-0001-6836-3041)
Thomas A. Colquhoun (https://orcid.org/0000-0001-6888-8585)
Patricio R. Munoz (https://orcid.org/0000-0001-8973-9351)

Keywords: genome-wide association, genomic prediction, plant breeding, *Vaccinium* spp., volatile.

SUMMARY

- Plants produce a range of volatile organic compounds (VOCs), some of which are perceived by the human olfactory system, contributing for a myriad of flavors. Despite the importance of flavor for consumer preference, most plant breeding programs have neglected it, mainly due to the costs of phenotyping and the complexity of disentangling the role of VOCs on human perception.
- To develop molecular breeding tools aimed at improving fruit flavor, we carried out a target genotyping and VOC extraction of a blueberry population. Metabolite genome-wide association analysis (GWAS) was used to elucidate the genetic architecture, while predictive models were tested to prove that VOCs can be accurately predicted using genomic information. Historical sensory panel was considered to assess how the volatiles influenced consumer.
- By gathering genomics, metabolomics, and sensory panel, we demonstrated that VOCs: i) are controlled by a few major genomic regions, some of which harboring biosynthetic enzyme-coding genes; ii) can be accurately predicted using molecular markers; and iii) can enhance or decrease consumer's overall liking.
- Here we emphasized how the understanding of the genetic basis and the role of VOCs on consumer preference can assist breeders in developing more flavorful cultivars at a more inexpensively and accelerated pace.

INTRODUCTION

Flavor is an important trait for any food crop, affecting consumer acceptance and marketability. Its relevance is even more pronounced for fruits, for which repeated purchasing behavior and willingness to pay have been associated with positive eating experiences (Clark, 1998; Diehl *et al.*, 2013). While substantial flavor variation exists within fruit species (El Hadi *et al.*, 2013), most plant breeding programs have historically neglected it, given its intrinsic complexity and costs to phenotype (Klee, 2010; Klee & Tieman, 2018). As a consequence, the drop-off in flavor quality has become one of the major causes of consumer dissatisfaction (Bruhn *et al.*, 1991; Tieman *et al.*, 2012). To correct this inconsistency and incorporate flavor in breeding program routines, it is necessary to identify the sources of flavor variability, understand their genetic architecture, and then define cost-effective methods of selection.

Flavor is a complex multifactorial trait, involving a combination of taste, mouthfeel and aroma perceptions. More specifically, it is the interaction between our olfactory system and the volatile organic compounds (VOCs) released by the fruit that provides the diversity and uniqueness of flavor experiences (Goff & Klee, 2006; El Hadi *et al.*, 2013). Plants synthesize a wide variety of VOCs (Dudareva *et al.*, 2006; Goff & Klee, 2006), but only a subset are produced during fruit ripening, where they likely act as an attractant for seed-dispersing organisms, including humans (Rodríguez *et al.*, 2013). Several fruit VOCs have been demonstrated to influence consumer's overall liking (Klee & Tieman, 2018), suggesting that these metabolites are key targets to improve the flavor perception of fruits. Although the VOC profiles of many fruit species have been characterized (El Hadi *et al.*, 2013; Klee & Tieman, 2018), less is known about the genetic basis underlying their variation among genotypes, which hinders their implementation in breeding programs. Moreover, quantifying the abundance of metabolites is expensive and time consuming for a large-scale populational application. In this scenario, molecular markers are a promising tool to detect genetic associations and predict the phenotype of new individuals (Klee, 2010).

Molecular breeding methods have been successfully applied for different traits and crops (Hickey *et al.*, 2017; Watson *et al.*, 2018), however they have been less exploited for fruit flavor improvement. Herein, we showed the feasibility of molecular breeding for flavor-related volatiles in a blueberry breeding program by integrating genomics, metabolomics, and sensory panel data. Blueberry (*Vaccinium* spp.) is the second most important soft fruit after strawberry, and it has also been popularized as a "super food" due to the multiple health benefits conferred by its abundant polyphenolic content (Kalt *et al.*, 2019). A previous psychophysical study indicated that consumers prefer sweet berries with intense flavor (Gilbert *et al.*, 2015). Therefore, considering that blueberry and other fruits are important dietary sources of micronutrients, an effort to improve flavor through breeding is warranted, which may lead to an increase in fresh fruit consumption that, subsequently, could have a positive impact on human health.

In this study, we used a targeted genotyping approach and volatile extractions with analysis by gas chromatography/mass spectrometry (GC/MS) of 1,438 individuals from a blueberry breeding population. Genome-wide association studies (GWAS) elucidated the genetic architecture of VOCs and predictive models showed that VOCs can be accurately predicted using genomic and marker-assisted selection. Finally, a historical blueberry sensory panel dataset was leveraged to assess how the volatiles influenced consumer preference to ultimately assist breeders in the direction of the selection.

MATERIAL AND METHODS

Plant Material.

The association mapping population was composed of 886 southern highbush blueberry genotypes covering 92 full-sib families. This population was originally designed as part of the breeding program at the University of Florida in February 2011. Seedlings originated from each family were installed in a row-column design at the Plant Science Research and Education Unit in Citra, Florida. Additional details on this population were previously described by (Cellon *et al.*, 2018) and (Ferrão *et al.*, 2018).

Tissue Collection, Sample Processing, and Volatile Extraction.

During April 2015, five full mature berries were harvested from each plant. We only sampled berries exhibiting picking quality, including fully blue color at the scar, no visual, pathogen or insect damage. Fruits were quenched in liquid nitrogen and stored at -80 °C up to the time of sample processing. The five berries from each genotype were ground together to a fine powder using a liquid nitrogen pre-chilled blender/coffee grinder (Tribest Corporation, Anaheim, CA, USA) and transferred into a 12 mL labeled tube. For each sample, 250 mg of frozen powder was weighted in duplicate into 2 mL microtubes and stored at -80 °C until volatile extraction. Internal volatiles were extracted using a solid-liquid-phase solvent extraction procedure. The extraction solvent consisted of anhydrous hexane containing 50 ng/uL surrogate standard (trans-2-heptenal, CAS: 18829-55-5, Sigma-Aldrich SKU: W316504-SAMPLE-K). Samples were randomly extracted in batches containing 11 samples in duplicate and two empty microtubes. The volatile extraction was performed as follows: samples were retrieved from archival storage and placed in liquid nitrogen; 1 mL of extraction solvent was added to each sample; samples were shaken for 5 seconds then vortexed for 10 seconds to ensure full saturation of tissue with solvent; samples were then shaken at 23 °C in a thermoshaker at 1400 rpm for 15 minutes, then centrifuged at 1500 G-force to induce phase separation; the top organic portion was recovered into a glass GC sample vial using a disposable glass Pasteur pipette. Samples were stored at -80 °C until GC-MS analysis.

Volatile Analysis.

Quantification of volatiles from the liquid phase extractions was performed on an Agilent 7980A series gas chromatograph (GC) equipped with an Agilent 5977A single quadrupole mass spectrum detector (MSD). Parameters of the GC were used as follows: Helium carrier gas fixed at 11.479 psi, splitless injection, inlet temperature 220 °C, injection volume 2 µL, and the syringe wash solvents were acetone and hexane. A guard column consisting of deactivated fused silica (Ultimate Plus deactivated fused silica tubing, 5 m length x 250 µm i.d.; Agilent Technologies, Santa Clara, CA, USA; catalog number: CP802505) was installed from the GC inlet and connected to the analytical column by a pressfit connector (Restek, Bellefonte, PA, USA; catalog number: 22159). Sample analytes were separated using an equipped DB-5 column ((5%-phenyl)-methylpolysiloxane, 30 m length x 250 µm i.d. x 1 um film thickness; Agilent Technologies, Santa Clara, CA, USA). Oven temperatures were programmed as follows: the initial oven temperature of 40 °C was held for 30 seconds, then ramped 15 °C min -1 to 250 °C with a post run temperature of 260 °C held for 3 minutes. The MSD was equipped with an extractor ion source and tuned for sensitivity and mass accuracy prior to sample analysis. Parameters for the MSD were maintained as follows: MSD transfer line temperature 280 °C, MS source temperature 230 °C, MS quad temperature 150 °C, solvent delay of 6 minutes, mass scan range 40-205 m/z with a threshold of 150. Data were acquired using Agilent MassHunter Workstation Acquisition (Agilent Technologies, Santa Clara, CA) and processed using Agilent's MassHunter Quantitative Analysis program v.B.06.00. Initial screening of volatiles consisted of a targeted list of the 52 volatile compounds previously reported and described by Gilbert *et al.* (2015). Additionally, spectral deconvolution was performed for each sample in the program and manually curated to achieve a list of compounds that were then validated based on comparing retention time and spectra to authentic standards. Overall, a list of 17 robust and reliable features were detected and validated with authentic volatile standards. The most abundant non-convoluted m/z ion fragment for each compound was used to integrate peak area. Integrated peaks were qualified by two additional m/z ion fragments that were required to match ratios observed in authentic standards. Volatile mass concentration (μg*gFW^{-1}) was calculated using standard curves for each individual compound. Values were normalized for recovery of the surrogate standard, trans-2-heptenal, within each individual batch of extracted samples and for the corresponding biological mass of each sample. The equation used to calculate volatile mass concentration for each individual volatile compound was as follows: $VOC_{mass} = {pA_{COI} \times pA_{SSCS}} \big/ {pA_{SS} \times M \times RF}$, where pA_{COI} is the peak area of compound of interest in sample, pA_{SSCS} is the peak area of surrogate standard in laboratory control spike, pA_{SS} is the peak area of surrogate standard in sample, M is biological sample mass, and RF is the response factor from compound standard series. Two technical replicates per sample were analyzed by GC/MS, and averaged quantification values were used as phenotype.

Genotypic Data.

Genotyping was carried out by RAPiD Genomics (Gainesville, FL, USA) using the sequence capture methodology as described in (Ferrão *et al.*, 2018). Sequencing was performed using Illumina HiSeq2000 platform considering 100 cycle paired-end runs. Raw reads were filtered by quality and trimmed using Trimmomatic v.0.36 with the following parameter settings "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50" (Bolger *et al.*, 2014). Filtered reads were mapped against the largest scaffolds of each of the 12 homoeologous groups of *Vaccinium corymbosum* cv. 'Draper' genome assembly(Colle *et al.*, 2019) using the BWA v.0.7.17 software(Li & Durbin, 2009). Single nucleotide polymorphisms (SNPs) were called using FreeBayes v.1.0.1 (Garrison & Marth, 2012), targeting 15,663 probe regions designed for the sequence capture approach (Benevenuto *et al.*, 2019). Sequencing reads counts per allele and individual, were extracted from the variant call file using the vcftools package (Danecek *et al.*, 2011). As blueberry is a tetraploid species (2n=4X=48), we used the *updog* R package to call the allele dosages based on the read counts (Gerard *et al.*, 2018). The *updog* package outputs the posterior probability means per SNP for each individual and we used these probabilities as our genotypes. Loci were also filtered by applying the following criteria: (i) minimum mapping quality of 20; (ii) only biallelic locus; (iii) maximum missing data of 50%; (iv) minor allele frequency of 1%; (v) mean depth of coverage of 40; and (vi) minimum genotype frequency of 0.01. The remained missing genotypes were imputed by the mean of each locus, as suggested in the GEMMA package (Zhou & Stephens, 2012).

Phenotypic Analysis.

We computed the phenotypic heritability for each volatile using the following phenotypic model: $\log(\boldsymbol{y}) = 1\boldsymbol{\mu} + \boldsymbol{Zg} + \boldsymbol{\varepsilon}$, where μ is the overall mean and $\mathbf{1}_n$ is a vector of ones; **Z** is the incidence matrix linking observation in the vector **y** to their respective genotype effects in the vector **g.** Normality was assumed for the genotype effects and residual, where $\boldsymbol{g} \sim MVN(0,\boldsymbol{A}\sigma_a^2)$ and $\boldsymbol{\varepsilon} \sim MVN(0,\boldsymbol{I}\sigma_e^2)$. The genetic covariance, **A,** can be derived from the expectation of co-ancestry coefficient between individuals from the pedigree, and it was computed assuming a tetraploid additive relationship matrix; while σ_a^2 is the additive genetic variance. For the residual, **I** is an identity matrix and σ_e^2 is the residual variance**.** *MVN* denotes the n-dimensional multivariate normal distribution. Additive genetic variance was estimated using Restricted Maximum Likelihood (RELM) using the *sommer* R package (Covarrubias-Pazaran, 2016), while the kinship matrix **A** was built using the AGHmatrix R package (Amadeu *et al.*, 2016). Phenotypic heritability (h^2) was computed as: $h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$.

Genome-Wide Association Study (GWAS).

We used genome-wide association to identify genomic regions controlling volatile content in blueberry. SNP-trait association analyses were based on a univariate linear mixed model (LMM), as described by (Yu *et al.*, 2006) and implemented in the GEMMA package (*option -lmm 4*) (Zhou & Stephens, 2012). LMM tests for association of the phenotypes with each marker were performed with corrections for: (i) main directions of population structure by regressing on the first five principal components (PCs) calculated using the genomic relationship matrix, and (ii) cryptic relatedness using the polygenic background effects with covariance proportional to the genomic relationship matrix. In a matrix notation, the follow LMM was considered for each volatile: $\log(\boldsymbol{y}) = \boldsymbol{W\alpha} + x\beta + \boldsymbol{u} + \boldsymbol{\varepsilon}$; where log(**y**) corresponds to logarithm-transformed values of the volatile; **W** is a matrix of covariates (fixed effects) including a columns of 1's for the intercept and the first five principal components; $\boldsymbol{\alpha}$ is a vector of the corresponding fixed coefficients including the intercept; *x* is a vector of marker genotypes; β is the effect size of the marker; $\boldsymbol{u}$ is the random polygenic effect distributed as $\boldsymbol{u} \sim MVN(0,\boldsymbol{K}\sigma_g^2)$ -- where **K** is the realized relationship matrix calculated with genome-wide markers, and σ_g^2 is the additive genetic variance; and $\boldsymbol{\varepsilon}$ is a vector of error, distributed as $\boldsymbol{\varepsilon} \sim MVN(0,\boldsymbol{I}\sigma_e^2)$. The molecular relatedness matrix **K** was built using the AGHmatrix R package (Amadeu *et al.*, 2016) assuming tetrasomic inheritance. Bonferroni correction considering a genome-wide significance level of 0.05 was used for establishing a p-value detection threshold for statistical significance. The effect size of significant SNPs was calculated as described by (Pallares *et al.*, 2014): $\alpha = {(\hat{\beta}^2 \times var_x)}\big/{var_y}$, where var_x is the variance of the genotype at the focal SNPs, var_y is the phenotypic variance, and $\hat{\beta}$ is the estimated SNP effect.

Functional Mapping and Annotation of Genetic Association

SNPs were characterized *in silico* for their genomic position and functional effect on protein coding genes using SNPdat v1.0.5 (Doran & Creevey, 2013) and further manual curation. The 'Draper' genome assembly and gene predictions were retrieved from the GigaScience database (Colle *et al.*, 2019). Genomic windows for screening of functional candidate genes were defined by two strategies: 1) between the left- and right-most significant SNPs forming a "tower-like" structure in the Manhattan plots; 2) between ±100 Kb from significant SNPs that do not form a "tower-like" structure, given that 100 Kb was the size of the linkage disequilibrium block calculated for the same population by (Ferrão *et al.*, 2018). The screened genomic windows can be found in the Table S1. Candidate genes were annotated using the Blast2GO tool with BLASTp search against the non-redundant protein database (Götz *et al.*, 2008). Gene ontology (GO) enrichment analyses were performed for each VOC's candidate gene subset against the total haploid blueberry gene content. We used the BiNGO plugin on Cytoscape v. 3.7.2, considering a hypergeometric test with false discovery rate correction ($P \leq 0.05$) (Maere *et al.*, 2005).

Genomic Heritability

Different modeling strategies were used to estimate the genomic heritability (proportion of variance explained by available SNPs) of each VOC. First, we calculated the genomic heritability (h^2_{snp}) using the methodology implemented in GEMMA package (*option -vc 2*) (Zhou, 2017). We also used the same method for estimate the variance partition by each chromosome independently. Another strategy was based on the Bayesian Sparse Linear-Mixed Model (BSLMM), which was used to further estimate the PVE (the proportion of the phenotypic variance explained by the polygenic term, analogous to the h^2_{snp}) and PGE (the proportion of the PVE that is explained by SNPs with a non-zero effect on phenotypic variation) . To this end, we first used the BSLMM to fit a multilocus GWAS model assuming the SNP effects are sampled from a point-normal distribution, as implemented in the GEMMA package (*option -bslmm 1*). In a matrix notation, the model is: $\log(\boldsymbol{y}) = \boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}$, where $\log(\boldsymbol{y})$ is an *n*-vector of the logarithm-transformed volatile phenotype measured in *n* individuals; **X** is an $n \times p$ matrix of additive tetraploid genotypes measured on the same individuals at *p* genetic markers; $\boldsymbol{\beta}$ is the SNP effects sampled from a mixture of two distribution, one that expects many small effects and another that generates few strong effects, as follow: $\beta_i \sim \pi N(\sigma_k^2 \tau^{-1}) + (1-\pi)\delta_0$, where σ_k^2 controls the expected magnitude of non-zero SNP effects and δ_0 denotes a point mass at zero; **u** is the polygenic term as previously described; and $\boldsymbol{\varepsilon}$ is a random independent error term. We ran the Markov chain using the default settings implemented in the GEMMA software. Full details about the BSLMM are described by (Zhou *et al.*, 2013).

Phenotype Prediction for Molecular Breeding.

We evaluated the potential use of molecular markers for phenotypic prediction and for validation of our GWAS results. To this end, we designed four training and testing scenarios, combining different populations and methods. Besides the original population composed of 886 individuals used in the GWAS analyses (POP_1), we also phenotyped and genotyped a new set of 552 individuals (POP_2). Berries from POP_2 were collected at the same year, location, and ripening stage as individuals from POP_1, and subjected to the same phenotyping and genotyping protocols. The genetic relationship between the two populations was explored using a Principal Component Analysis (PCA) and a heatmap of the realized genomic matrix. To perform predictions based on molecular information, we considered three different approaches: (i) Genomic selection (GS) model, that fits a regression by modelling markers as random variables drawn from the same normal distribution, using RR-BLUP method (Endelman, 2011); (ii) GS *de novo* GWAS model, that combines RR-BLUP method with significant markers from GWAS fitted as fixed effects covariates; and (iii) Marker-Assisted Selection based on candidate loci (MAS), that fits a multiple regression model considering only the GWAS hits as fixed effects. A summary of these scenarios is presented in Supporting Information Table S2. GS and GS *de novo* GWAS models were fitted using ridge-regression models as implemented in the rrBLUP R-package (Endelman, 2011). MAS approach was fitted using *lm* function in R software. To select the GWAS hits to be used as fixed effects covariates, we retained the significant p-values estimated using the LMM approach and selected the marker with the smallest p-value within every 10 Kb genomic window. Finally, we accessed the predictive ability (PA) by computing the Pearson's correlation between predicted and original phenotypes.

Sensory Analysis.

We evaluated the impact of the volatiles in flavor perception using a consumer panel sensory data. To this end, over the course of six years (2012-2017), 24 blueberry cultivars from the breeding program at University of Florida were evaluated in 45 different sensory panels. On average, 90 panelists participated of each survey. As described by (Gilbert *et al.*, 2014) and (Schwieterman *et al.*, 2014), panelists were trained with the scaling methods and rated for overall liking, texture liking, sweetness, sourness, and flavor intensity using a hedonic general Labeled Magnitude Scales (gLMS) ranging from -100 (greatest disliking of any kind) to +100 (greatest liking of any kind). Concurrent with panel evaluation, a subset of berries from the same genotypes were submitted to chemistry analyses, which included: volatile extraction and quantification, soluble solids content and titratable acids (TA) measurements. See (Gilbert *et al.*, 2015) for more details.

RESULTS

Volatile Phenotyping.

A total of 17 VOCs were identified through GC/MS analyses, which comprised different chemical classes and biosynthetic origins (Fig. S1). Among the fatty-acid derivatives, there were five aldehydes, two alcohols, and three methyl ketones. From the mevalonic acid (MVA) or the methylerythritol phosphate (MEP) pathways, there were five terpenoid compounds. Lastly, there were two benzenoid compounds which are derived from shikimate/phenylalanine pathway. Fatty acid derived aldehydes were the most abundant in concentration followed by their derived alcohols, and the benzenoid methyl salicylate had the lowest concentration (Table 1).

Genome-Wide Association Mapping.

In this study, volatile-genotype associations were performed using a linear mixed model approach. A total of 71,487 single nucleotide polymorphisms (SNPs), distributed across the twelve haploid blueberry chromosome-scaled scaffolds, were independently tested for association (Fig. S2 and Table S3). After Bonferroni-based multiple test correction, we detected 519 significant SNPs associated with 11 VOCs, encompassing ten chromosomes and different metabolic pathways (Fig. 1). Most significant SNPs converged to a tower-like structure in Manhattan plots, indicating the presence of few genomic regions controlling each VOC emission (Fig. S3a and S3b). The number of genomic regions associated with each VOC ranged from one (for (E)-2-hexenal, 1-hexanol, and 2-undecanone) to five (for eucalyptol) (Fig. S4a and Table S1). Some common genomic windows were detected for volatiles derived from the same biosynthetic pathways (Fig. S4a). In chromosome 2, both methylketones (2-nonanone and 2-undecanone) shared the same genomic region; while in chromosome 3, overlaps were observed for the terpenoids linalool and D-limonene. Most of the significant SNPs were detected in non-coding regions, and among the exonic SNPs, most caused synonymous changes (Fig. S4b). We also explored how much of the phenotypic variation was explained by individual markers (Fig. S4c). Notably, some significant SNPs individually explained more than 10% of the phenotypic variation observed. For example, a single marker on chromosome 2 explained more than 30% of the phenotypic variance associated with 2-undecanone (Fig. S4c).

Candidate Genes.

Putative protein coding genes were searched for in the regions flanking significant SNPs. Functional annotation of these genes pointed to enzymes in VOCs biosynthetic pathways and related biological functions (Fig. 1 and Table S1). The most explicit candidate genes were found for terpenoid volatiles (Fig. 2). One of the genomic regions associated with both linalool and D-limonene comprised six linalool synthase encoding genes (Fig. 2). For eucalyptol, eight alpha-terpineol synthases were predicted in one of the genomic regions (Fig. 2). Moreover, the GO term “terpene synthase activity” was overrepresented for genes within candidate regions of these volatiles (Table S4). For the carotenoid-derived terpene, geranyl acetone, the enzymes mevalonate kinase, zeta-carotene desaturase, and carotenoid cleavage dioxygenase 4 (CDD4) were present at distinct genomic regions associated with this volatile (Fig. 1). For the fatty acid-derived VOCs (hexanal, 1-hexanol, decanal, 2-heptanone, 2-nonanone, and 2-undecanone), several enzymes involved in lipid biosynthesis and degradation were detected (Fig. 1). Other biologically plausible candidate genes underlying VOC variation at the different genomic regions include those potentially involved in plant defense, regulation of transcription, regulation of protein abundance through proteasomal degradation, volatile emission through ABC-type transporters, VOC degradation, and competition for precursors with adjacent pathways (Table S1).

Heritability of VOCs.

We accessed the heritability of the 11 VOCs for which significant associations were detected using different approaches (Fig. 3). Using the pedigree information, we observed moderate-to-high heritability values ($h^2>50\%$) for most of the metabolites. Remarkably, 1-hexanol, 2-undecanone, decanal and linalool presented values higher than 97%. Considering the genomic heritability based on all SNPs (h^2_{snp}), we also observed moderate-to-high values, but with a relative lower magnitude when compared to pedigree analyses. We also used the Bayesian Sparse Linear-Mixed Model (BSLMM) to investigate the genetic contribution of sparse (PGE parameter) and polygenic components (PVE parameter). The PVE is a Bayesian version of the

h^2_{snp} and, as expected, both analyses resulted in similar values. The PGE is the proportion of the PVE that is explained by markers with large effects, shedding light on the genetic architecture of the traits. For most VOCs, we observed PGE values higher than 30%. Additionally, we divided the genetic variance explained by markers in a chromosome-based scheme, and no positive trend between phenotypic variance and chromosome length was observed -- as expected for polygenic traits (Fig. S5). Altogether, the results from the association mapping, the heritability estimation, and the partitioning of the genetic variance by chromosome length corroborated to indicate that VOC traits have a simple genetic architecture, with few major loci controlling a large proportion of the phenotypic variance.

Validation and Phenotypic Prediction.

Marker-assisted selection (MAS) based on candidate loci and genomic selection (GS) based on markers covering the whole genome are powerful tools to predict the phenotypic merit of an individual and support breeding decisions. Here, we compared the feasibility and efficacy of both approaches for implementation in breeding programs targeting VOCs. To this end, a new set of individuals (POP_2) was phenotyped and genotyped for real validation. POP_2 is composed of individuals genetically related to the original population (POP_1) used in the GWAS analysis (Fig. 4a, Figure S6). Subsequently, four different scenarios (SCE) mimicking breeding programs were designed for “training" and “testing" partitions (Fig. 4b). In the SCE_1, GS models were trained and tested within POP_1, by systematically splitting the original population into non-overlapping “training" and “testing" partitions. In this scenario, predictive performances ranged from 0.41 to 0.67 for hexanal and 2-nonanone, respectively (Fig. 4c and Figure S7). The SCE_2 captured GS validations across populations. Compared with SCE_1, there were substantial decreases in the predictive abilities (Fig. 4c). In order to validate our previous findings in the GWAS analysis we designed the SCE_3 and SCE_4 scenarios, whereby the GWAS hits were used as fixed effects in the prediction models across populations (Fig. 4b and Table S5). In the SCE_3, also named as “GS *de novo* GWAS”, most of the validations yielded higher predictive performances when compared to SCE_2. Notably, for some volatiles, the performances were comparable with the results in the SCE_1 (Fig. 4c). The MAS approach was represented in SCE_4 (Fig. 4b) and high predictive ability was achieved for most of the VOCs (Fig. 4c), demonstrating that a small set of markers (Table S5) can be used for VOC prediction. Overall, the prediction results were aligned with the PGE values (Fig. 3): volatiles with lower PGE values, such as hexanal, 1-hexanol and (E)-2-hexenal, showed lower predictive performances (Figure S8).

Sensory Panel.

To determine the impact of VOCs, soluble solids, and titratable acidity (TA) content on the consumer perception, we used a historical set of sensory panel data. Over the course of 6 years, consumer panelists rated overall liking, texture, sweetness, sourness and flavor intensity of 24 cultivars (Table S6). In the correlation analysis, overall liking was strongly and positively correlated with soluble solids that, aside other compounds, can be an indicator of sugar content (Fig. 5a and Fig. S9). In contrast, TA was negatively correlated with overall liking (Fig. 5a). For VOCs, positive and negative trends were observed in the dataset (Fig. 5a and Fig. S9). The volatiles 1-hexanol and 2-undecanone were the most positively correlated with overall liking (0.48 and 0.43, respectively), while (E)-2-hexanal and eucalyptol were the most negatively correlated (-0.75 and -0.56, respectively). To exemplify the importance of the biochemical compounds on flavor enhancement, we selected three cultivars for additional comments. 'Kestrel' and 'Windsor' have been consistently rated by panelist with high and low scores of overall liking, respectively. By contrasting the biochemical profile of both cultivars, the low sugar to acid ratio observed in 'Windsor' is probably the main difference negatively influencing panelists liking scores (Fig. 5b). In addition to sugar and acids, VOCs also play an important role on flavor. To exemplify, we selected 'Kestrel' and 'Snowchaser', two cultivars with similar profiles of soluble solids and TA, but displaying distinct volatile profiles and liking scores (Fig. 5b). 'Snowchaser' had higher eucalyptol content; while 'Kestrel' had higher linalool content, which are likely influencing the flavor perception.

DISCUSSION

Over the recent decades, it has been generally accepted that the drop-off in flavor quality is one of the major sources of consumer dissatisfaction with fresh fruit produce (Bruhn *et al.*, 1991; Klee, 2010). Despite its importance, flavor is an expensive and complex trait to be routinely evaluated in breeding programs. In this scenario, the use of molecular breeding to track flavor-associated metabolites emerges as the best strategy for flavor improvement (Klee, 2010). Some breeding programs are already taking advantage of marker-assisted selection to this end

(Chambers *et al.*, 2014; Eduardo *et al.*, 2014; Emanuelli *et al.*, 2014). In this study, by using a combination of genomics, metabolomics, and sensory panel data, we identified markers associated with flavor-related volatiles and showed their real application in a blueberry breeding program.

With a relatively recent domestication history dated from the 1900's, blueberry breeding programs still possess a wide variability of flavors to be explored, and unlike other fruits, like bananas, there is not a single chemical that represents a typical blueberry flavor. Guided by prior findings that the blueberry flavor is influenced by several volatile metabolites (Gilbert *et al.*, 2013, 2015; Tieman *et al.*, 2017), we first used a metabolomics approach to detect and quantify these VOCs in a southern highbush blueberry breeding population. Previous analyses in blueberry used a dynamic headspace collection, which provided sensitive and detailed volatile profiling (Gilbert *et al.* 2013, 2015). However, the tissue requirements and current limited throughput of dynamic headspace collection system necessitated the development of a quantitative and higher throughput volatile screen for a large breeding population. Therefore, we developed a liquid extraction-based method that was able to detect most of the volatiles previously reported as influential to the human hedonic and sensory perception (Gilbert *et al.*, 2015). In this study, a total of 17 volatiles from different chemical classes and biosynthetic origins were quantified. Among them, aldehydes and alcohols have been frequently associated with green, grassy, and herbal flavors as well as the terpene eucalyptol (Klee, 2010; Gilbert *et al.*, 2015; Farneti *et al.*, 2017b). Terpenes like linalool, D-limonene, and geranyl acetone have been linked with floral, fresh, and citrusy flavors (Farneti *et al.*, 2017b). The methyl ketones 2-nonanone and 2-undecanone have been associated with fruity flavors, while 2-heptanone with cheesy and banana-like flavor. Such variability highlights the amplitude of flavors that could be explored.

Although the blueberry volatile landscape has already been characterized in several studies (Parliment & Kolor, 1975; Hirvi & Honkanen, 1983; Du *et al.*, 2011; Gilbert *et al.*, 2013; Du & Rouseff, 2014; Gilbert *et al.*, 2015; Farneti *et al.*, 2017b), the genetic basis underlying VOCs variation among genotypes is still unknown. Through GWAS analyses, we identified molecular markers with large effects and showed that few genomic regions were involved in the variation of VOCs content in blueberry. Most of the genomic studies addressing volatiles in fruits have been performed using traditional QTL mapping (Zini *et al.*, 2005; Doligez *et al.*, 2006; Tieman *et al.*, 2006; Duchêne *et al.*, 2009; Dunemann *et al.*, 2009; Costa *et al.*, 2013; Eduardo *et al.*, 2013; Paterson *et al.*, 2013; Urrutia *et al.*, 2017), while GWAS has been only recently conducted in studies applied to tomato (Sauvage *et al.*, 2014; Zhang *et al.*, 2015; Bauchet *et al.*, 2017; Tieman *et al.*, 2017) and apple (Kumar *et al.*, 2015; Farneti *et al.*, 2017a; Larsen *et al.*, 2019). When compared to QTL mapping, GWAS increases the mapping resolution by making use of more diverse populations with lower levels of linkage disequilibrium and higher marker density, which narrows the genomic regions to search for potential candidate genes and causal polymorphisms (Korte & Farlow, 2013).

Herein, we found plausible candidate genes within the genomic windows surrounding significant SNPs. The most striking candidate genes were clusters of monoterpene synthases found in one of the regions associated with the terpenoids linalool, D-limonene, and eucalyptol. Terpene synthases are part of a large family of enzymes that generate the diversity of volatile terpenoids found in plants (Bohlmann *et al.*, 1998; Dudareva *et al.*, 2006). These enzymes catalyze the first committed step toward the synthesis of terpenoids from substrates derived from MEP/MVA pathways. Terpene synthases were also found in QTL regions of terpenes in peach (Eduardo *et al.*, 2013; Sánchez *et al.*, 2014), raspberry (Paterson *et al.*, 2013), and carrot (Keilwagen *et al.*, 2017). The terpene geranyl acetone is also derived from MEP/MVA pathways but it is formed through the oxidative cleavage of carotenoids. In the regions associated with geranyl acetone, there were enzymes involved in the early steps of the biosynthetic pathway (mevalonate kinase), in the biosynthesis of carotenoids (zeta-carotene desaturase), and in the final step of carotenoid cleavage (CDD4) (Yuan *et al.*, 2015). Previous QTLs detected in other species have found only carotenoid biosynthetic enzymes in the associated area, suggesting that the carotenoid content of the fruits was the determinant of this suite of volatiles (Lewinsohn *et al.*, 2005; Tieman *et al.*, 2006, 2017; Klee, 2010). Notably, this is the first study that the authors are aware of where a CDD-encoding enzyme was detected in a QTL region for an apocarotenoid volatile variation. For the fatty-acid derivative VOCs (aldehydes, alcohols, and methyl ketones), several enzymes involved in the lipid metabolism were detected. The synthesis and degradation of fatty acids can have a major impact on downstream volatile synthesis, as suggested in tomato studies (Howe & Schilmiller, 2002; Garbowicz *et al.*, 2018). Besides biosynthetic enzymes, other biological mechanisms could also be involved in VOCs variation as detected in other associated genomics regions. For example, ABC transporters have been shown to facilitate

VOCs emission (Adebesin *et al.*, 2017); cytochrome P450 enzymes can act in volatile biosynthesis and modification (Dudareva *et al.*, 2006); transcriptional regulation can induce or repress VOCs synthesis and release (Dudareva *et al.*, 2006). Despite finding interesting candidates, we did not detect causal polymorphisms in the genes, and functional validation is also needed to confirm our hypotheses. It is noteworthy, however, that recent studies have revealed that noncoding regulatory regions in plant genomes are highly variable at the species level and can impact the expression of hundreds of genes that ultimately alter various traits (Yocca *et al.*, 2019). Unfortunately, the noncoding regulatory space remains completely unknown in blueberry, and the genotyping approach used herein is unable to assess all types of variants that exist between the genotypes. Nonetheless, tightly linked markers are valuable for breeding purposes.

Motivated by the possibility to predict volatile content using genomic data, we first investigated the use of genomic selection (Meuwissen *et al.*, 2001). High predictive accuracies were observed for most of the VOCs, which have also been reported in the plant literature for other types of metabolites (Riedelsheimer *et al.*, 2012; Guo *et al.*, 2016; Kainer *et al.*, 2018; Schrag *et al.*, 2018). Subsequently, we explored the importance and applicability of our GWAS findings for prediction in an independent dataset by using the significant GWAS hits as fixed effects into GS models. This GWAS *de novo* GS framework showed an increase of more than 20% in the predictive ability for some volatiles (e.g., 2-undecanone), when compared to traditional GS methods. Using a similar strategy, gains in predictive performance for other traits were also reported in maize (Bernardo, 2014; Rice & Lipka, 2019), sorghum (Rice & Lipka, 2019), and rice (Spindel *et al.*, 2016). Finally, an even more promising approach is to use only the GWAS hits for MAS, thereby reducing costs associated to genome-wide genotyping. By acknowledging the simple genetic nature of each VOC, the MAS approach yielded high predictive performances. These results demonstrate the remarkable benefits that molecular breeding can achieve for flavor improvements in blueberry, providing a motivation for similar studies in other crop species.

Our last contribution concerned unraveling the role of VOCs in sensory analyses. Plants synthesize hundreds of VOCs; however, only a small subset generates the “flavor fingerprint” recognized by humans (Goff & Klee, 2006). Moreover, some VOCs are going to be positively or negatively perceived, affecting the overall liking experience. Regarding the taste components, sugar and acid contents are well-known for their opposite relationship with blueberry overall liking (Gilbert *et al.*, 2015), which was also evident in our analyses. Regarding VOCs, hexanal and 2-undecanone had a positive correlation with overall liking scores, while (E)-2-hexanal and eucalyptol were negatively correlated. Linalool has been pointed out as a key metabolite for the characteristic blueberry aroma (Parliment & Kolor, 1975; Hirvi & Honkanen, 1983; Du *et al.*, 2011); however an overall low positive correlation was detected herein. Nonetheless, when we compared two cultivars with similar sugar and acidity profiles (`Kestrel` and `Snowchaser`), the contrasting linalool content suggested it does have a positive role on overall liking. Another important flavor contributor is eucalyptol. Similar studies also observed that blueberries with high eucalyptol content had reduced blueberry liking scores (Gilbert *et al.*, 2015; Farneti *et al.*, 2017b). High concentration of eucalyptol is observed in green fruits, with a drastic reduction during ripening; raising the hypothesis that eucalyptol may be considered as "not attractive" or as a "repellent" metabolite, negatively affecting consumer preferences (Farneti *et al.*, 2017b).

Overall, by considering our sensory panel results, VOCs should be a significant consideration for manipulation in breeding programs focusing on improving flavor, with some VOCs targeted to up- or down-concentrations. Moreover, the high heritability values estimated for VOCs are promising for achieving faster genetic gains across the generations, especially compared with previously values obtained for soluble solids (sugars) and pH (acids) (Cellon *et al.*, 2018; de Bem Oliveira *et al.*, 2019). Unlike sugars and acids which are primary sources of energy and carbon for plants to metabolize into other compounds, VOCs are metabolic end products derived from few major biochemical pathways. For comparison, sugar and acids occur at millimolar concentrations in fruits, while VOCs are often present at picomolar to nanomolar concentrations. These differences suggest that manipulating VOC accumulation through breeding should be easier compared to making meaningful changes in sugar and acid content, while having little impact on other metabolic activities.

Altogether, in this study we have demonstrated how metabolomics, genomics and sensory panel data can be combined to implement molecular breeding techniques for flavor improvements. Some of the main findings presented herein include: (i) Some VOCs have a simple genetic architecture controlled by few genomic regions, with high heritability values, and QTLs with major effects; (ii) Some of the associated genomic regions harbored candidate genes known to be involved in volatile biosynthetic pathways; (iii) Given the difficulties in phenotyping VOCs, we demonstrated that marker-assisted selection is a feasible and efficient tool to be implemented at the scope of breeding programs; (iv) Sensory panel data showed that some VOCs modulate consumer preference, indicating the direction of the breeding selection. Overall, this work is a promising step toward understanding the genetic basis of VOCs for breeding purposes and the role of VOCs on fruit flavor perception. Although this study is applied to blueberry, our findings have a broad relevance in the context of plant breeding aiming at the improvement of flavor.

ACKNOWLEDGEMENTS

The authors thank Haley Sater and Austin Wenta for their help in panel data collection, Lorenzo Bizzio for discussions about volatile metabolic pathways. This work was supported by the UF royalty fund generated by the licensing of blueberry cultivars.

AUTHOR CONTRIBUTIONS

PRM and TAC conceived and supervised the study. TSJ performed the volatile extraction, GC-MS data analysis and quality control. LRVF analyzed and interpreted the phenotypic, GWAS, GS, and sensory panel data. JB performed the SNP calling and candidate gene mining. PPE provided information about the blueberry genome. LFVF, JB, TSJ and PRM wrote the paper with revision from all authors. All authors read and approved the final version of the manuscript for publication. LFVF and TSJ contributed equally to this work.

REFERENCES

Adebesin F, Widhalm JR, Boachon B, Lefèvre F, Pierman B, Lynch JH, Alam I, Junqueira B, Benke R, Ray S. **2017**. Emission of volatile organic compounds from petunia flowers is facilitated by an ABC transporter. *Science* **356**: 1386–1388.

Amadeu RR, Cellon C, Olmstead JW, Garcia AAF, Resende MFR, Muñoz PR. **2016**. AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. *The Plant Genome* **9**: 0.

Bauchet G, Grenier S, Samson N, Segura V, Kende A, Beekwilder J, Cankar K, Gallois J, Gricourt J, Bonnet J. **2017**. Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. *New Phytologist* **215**:

624–641.

de Bem Oliveira I, Resende MFR, Ferrão LF V, Amadeu RR, Endelman JB, Kirst M, Coelho ASG, Munoz PR. **2019**. Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3: Genes, Genomes, Genetics* **9**: 1189–1198.

Benevenuto J, Ferrão LF V, Amadeu RR, Munoz P. **2019**. How can a high-quality genome assembly help plant breeders? *GigaScience* **8**: giz068.

Bernardo R. **2014**. Genomewide Selection when Major Genes Are Known. *Crop Science* **54**: 68–75.

Bohlmann J, Meyer-Gauen G, Croteau R. **1998**. Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proceedings of the National Academy of Sciences* **95**: 4126–4133.

Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bruhn CM, Feldeman N, Garlitz C, Harwood J, Ivans E, MARSHALL M, RILEY A, Thurber D, Williamson E. **1991**. Consumer perceptions of quality: apricots, cantaloupes, peaches, pears, strawberries, and tomatoes. *Journal of Food Quality* **14**: 187–195.

Cellon C, Amadeu RR, Olmstead JW, Mattia MR, Ferrao LF V, Munoz PR. **2018**. Estimation of genetic parameters and prediction of breeding values in an autotetraploid blueberry breeding population with extensive pedigree data. *Euphytica* **214**: 1–13.

Chambers AH, Pillet J, Plotto A, Bai J, Whitaker VM, Folta KM. **2014**. Identification of a strawberry flavor gene candidate using an integrated genetic-genomic-analytical chemistry approach. *BMC genomics* **15**: 217.

Clark JE. **1998**. Taste and flavour: their importance in food choice and acceptance. *Proceedings of the nutrition society* **57**: 639–643.

Colle M, Leisner CP, Wai CM, Ou S, Bird KA, Wang J, Wisecaver JH, Yocca AE, Alger EI, Tang H. **2019**. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience* **8**: giz012.

Costa F, Cappellin L, Zini E, Patocchi A, Kellerhals M, Komjanc M, Gessler C, Biasioli F. **2013**. QTL validation and stability for volatile organic compounds (VOCs) in apple. *Plant science* **211**: 1–7.

Covarrubias-Pazaran G. **2016**. Genome-assisted prediction of quantitative traits using the R package sommer. *PloS one* **11**: e0156744.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Diehl DC, Sloan NL, Bruhn CM, Simonne AH, Brecht JK, Mitcham EJ. **2013**. Exploring produce industry attitudes: Relationships between postharvest handling, fruit flavor, and consumer purchasing. *HortTechnology* **23**: 642–650.

Doligez A, Audiot E, Baumes R, This P. **2006**. QTLs for muscat flavor and monoterpenic odorant content in grapevine (*Vitis vinifera* L.). *Molecular Breeding* **18**: 109–125.

Doran AG, Creevey CJ. **2013**. Snpdat: easy and rapid annotation of results from *de novo* snp discovery projects for model and non-model organisms. *BMC bioinformatics* **14**: 45.

Du X, Plotto A, Song M, Olmstead J, Rouseff R. **2011**. Volatile composition of four southern highbush blueberry cultivars and effect of growing location and harvest date. *Journal of agricultural and food chemistry* **59**: 8347–8357.

Du X, Rouseff R. **2014**. Aroma active volatiles in four southern highbush blueberry cultivars determined by gas chromatography–olfactometry (GC-O) and gas chromatography–mass spectrometry (GC-MS). *Journal of agricultural and food chemistry* **62**: 4537–4543.

Duchêne E, Butterlin G, Claudel P, Dumas V, Jaegli N, Merdinoglu D. **2009**. A grapevine (*Vitis vinifera* L.) deoxy-D-xylulose synthase gene colocates with a major quantitative trait loci for terpenol content. *Theoretical and Applied Genetics* **118**: 541–552.

Dudareva N, Negre F, Nagegowda DA, Orlova I. **2006**. Plant volatiles: recent advances and future perspectives. *Critical reviews in plant sciences* **25**: 417–440.

Dunemann F, Ulrich D, Boudichevskaia A, Grafe C, Weber WE. **2009**. QTL mapping of aroma compounds analysed by headspace solid-phase microextraction gas chromatography in the apple progeny 'Discovery'×'Prima'. *Molecular breeding* **23**: 501–521.

Eduardo I, Chietera G, Pirona R, Pacheco I, Troggio M, Banchi E, Bassi D, Rossini L, Vecchietti A, Pozzi C. **2013**. Genetic dissection of aroma volatile compounds from the essential oil of peach fruit: QTL analysis and identification of candidate genes using dense SNP maps. *Tree Genetics & Genomes* **9**: 189–204.

Eduardo I, López-Girona E, BatlIe I, Reig G, Iglesias I, Howad W, Arús P, Aranzana MJ. **2014**. Development of diagnostic markers for selection of the subacid trait in peach. *Tree* *genetics & genomes* **10**: 1695–1709.

Emanuelli F, Sordo M, Lorenzi S, Battilana J, Grando MS. **2014**. Development of user-friendly functional molecular markers for VvDXS gene conferring muscat flavor in grapevine. *Molecular breeding* **33**: 235–241.

Endelman JB. **2011**. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* **4**: 250.

Farneti B, Di Guardo M, Khomenko I, Cappellin L, Biasioli F, Velasco R, Costa F. **2017a**. Genome-wide association study unravels the genetic control of the apple volatilome and its interplay with fruit texture. *Journal of experimental botany* **68**: 1467–1478.

Farneti B, Khomenko I, Grisenti M, Ajelli M, Betta E, Algarra AA, Cappellin L, Aprea E, Gasperi F, Biasioli F. **2017b**. Exploring blueberry aroma complexity by chromatographic and direct-injection spectrometric techniques. *Frontiers in plant science* **8**: 617.

Ferrão LFV, Benevenuto J, Oliveira I de B, Cellon C, Olmstead J, Kirst M, Resende Jr MF, Munoz PR. **2018**. Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context. *Frontiers in Ecology and Evolution* **6**: 107.

Garbowicz K, Liu Z, Alseekh S, Tieman D, Taylor M, Kuhalskaya A, Ofner I, Zamir D, Klee HJ, Fernie AR. **2018**. Quantitative Trait Loci Analysis Identifies a Prominent Gene Involved in the Production of Fatty Acid-Derived Flavor Volatiles in Tomato. *Molecular plant* **11**: 1147–1165.

Garrison E, Marth G. **2012**. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

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

Gilbert JL, Guthart MJ, Gezan SA, de Carvalho MP, Schwieterman ML, Colquhoun TA, Bartoshuk LM, Sims CA, Clark DG, Olmstead JW. **2015**. Identifying breeding priorities for blueberry flavor using biochemical, sensory, and genotype by environment analyses. *PLoS One* **10**: e0138494.

Gilbert JL, Olmstead JW, Colquhoun TA, Levin LA, Clark DG, Moskowitz HR. **2014**. Consumer-assisted selection of blueberry fruit quality traits. *HortScience* **49**: 864–873.

Gilbert JL, Schwieterman ML, Colquhoun TA, Clark DG, Olmstead JW. **2013**. Potential for increasing southern highbush blueberry flavor acceptance by breeding for major volatile components. *HortScience* **48**: 835–843.

Goff SA, Klee HJ. **2006**. Plant volatile compounds: sensory cues for health and nutritional value? *Science* **311**: 815–819.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. **2008**. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* **36**: 3420–3435.

Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D. **2016**. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and applied genetics* **129**: 2413–2427.

El Hadi M, Zhang F-J, Wu F-F, Zhou C-H, Tao J. **2013**. Advances in fruit aroma volatile research. *Molecules* **18**: 8200–8229.

Hickey JM, Chiurugwi T, Mackay I, Powell W, Eggen A, Kilian A, Jones C, Canales C, Grattapaglia D, Bassi F. **2017**. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature genetics* **49**: 1297.

Hirvi T, Honkanen E. **1983**. The aroma of blueberries. *Journal of the Science of Food and Agriculture* **34**: 992–996.

Howe GA, Schilmiller AL. **2002**. Oxylipin metabolism in response to stress. *Current opinion in plant biology* **5**: 230–236.

Kainer D, Stone EA, Padovan A, Foley WJ, Külheim C. **2018**. Accuracy of Genomic Prediction for Foliar Terpene Traits in Eucalyptus polybractea. *G3: Genes, Genomes, Genetics* **8**: 2573–2583.

Kalt W, Cassidy A, Howard LR, Krikorian R, Stull AJ, Tremblay F, Zamora-Ros R. **2019**. Recent Research on the Health Benefits of Blueberries and Their Anthocyanins. *Advances in Nutrition*.

Keilwagen J, Lehnert H, Berner T, Budahn H, Nothnagel T, Ulrich D, Dunemann F. **2017**. The terpene synthase gene family of carrot (*Daucus carota* L.): identification of QTLs and candidate genes associated with terpenoid volatile compounds. *Frontiers in plant science* **8**: 1930.

Klee HJ. **2010**. Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. *New Phytologist* **187**: 44–56.

Klee HJ, Tieman DM. **2018**. The genetics of fruit flavour preferences. *Nature Reviews Genetics*

19: 347–356.

Korte A, Farlow A. **2013**. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* **9**: 29.

Kumar S, Rowan D, Hunt M, Chagné D, Whitworth C, Souleyre E. **2015**. Genome-wide scans reveal genetic architecture of apple flavour volatiles. *Molecular breeding* **35**: 118.

Larsen B, Migicovsky Z, Jeppesen AA, Gardner KM, Toldam-Andersen TB, Myles S, Ørgaard M, Petersen MA, Pedersen C. **2019**. Genome-Wide Association Studies in Apple Reveal Loci for Aroma Volatiles, Sugar Composition, and Harvest Date. *The Plant Genome* **12**. doi:10.3835/plantgenome2018.12.0104

Lewinsohn E, Sitrit Y, Bar E, Azulay Y, Meir A, Zamir D, Tadmor Y. **2005**. Carotenoid pigmentation affects the volatile composition of tomato and watermelon fruits, as revealed by comparative genetic analyses. *Journal of Agricultural and Food Chemistry* **53**: 3142–3148.

Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Maere S, Heymans K, Kuiper M. **2005**. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**: 3448–3449.

Meuwissen TH, Hayes BJ, Goddard ME. **2001**. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Pallares LF, Harr B, Turner LM, Tautz D. **2014**. Use of a natural hybrid zone for genomewide association mapping of craniofacial traits in the house mouse. *Molecular Ecology* **23**: 5756–5770.

Parliment TH, Kolor MG. **1975**. Identification of the major volatile components of blueberry. *Journal of Food Science* **40**: 762–763.

Paterson A, Kassim A, McCallum S, Woodhead M, Smith K, Zait D, Graham J. **2013**. Environmental and seasonal influences on red raspberry flavour volatiles and identification of quantitative trait loci (QTL) and candidate genes. *Theoretical and applied genetics* **126**: 33–48.

Rice B, Lipka AE. **2019**. Evaluation of RR-BLUP Genomic Selection Models that Incorporate Peak Genome-Wide Association Study Signals in Maize and Sorghum. *The Plant Genome* **12**: 180052.

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T,

Stitt M, Willmitzer L, Melchinger AE. **2012**. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc Natl Acad Sci USA* **109**: 8872–8877.

Rodríguez A, Alquézar B, Peña L. **2013**. Fruit aromas in mature fleshy fruits as signals of readiness for predation and seed dispersal. *New Phytologist* **197**: 36–48.

Sánchez G, Martínez J, Romeu J, García J, Monforte AJ, Badenes ML, Granell A. **2014**. The peach volatilome modularity is reflected at the genetic and environmental response levels in a QTL mapping population. *BMC plant biology* **14**: 137.

Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M. **2014**. Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant physiology* **165**: 1120–1132.

Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE. **2018**. Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* **208**: 1373–1385.

Schwieterman ML, Colquhoun TA, Jaworski EA, Bartoshuk LM, Gilbert JL, Tieman DM, Odabasi AZ, Moskowitz HR, Folta KM, Klee HJ. **2014**. Strawberry flavor: diverse chemical compositions, a seasonal influence, and effects on sensory perception. *PloS one* **9**: e88446.

Spindel JE, Begum H, Akdemir D, Collard B, Redoña E, Jannink J, Mccouch S. **2016**. Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116**: 395–408.

Tieman D, Bliss P, McIntyre LM, Blandon-Ubeda A, Bies D, Odabasi AZ, Rodríguez GR, van der Knaap E, Taylor MG, Goulet C. **2012**. The chemical interactions underlying tomato flavor preferences. *Current Biology* **22**: 1035–1039.

Tieman DM, Zeigler M, Schmelz EA, Taylor MG, Bliss P, Kirst M, Klee HJ. **2006**. Identification of loci affecting flavour volatile emissions in tomato fruits. *Journal of experimental botany* **57**: 887–896.

Tieman D, Zhu G, Resende MFR, Lin T, Nguyen C, Bies D, Rambla JL, Beltran KSO, Taylor M, Zhang B. **2017**. A chemical genetic roadmap to improved tomato flavor. *Science* **355**: 391–394.

Urrutia M, Rambla JL, Alexiou KG, Granell A, Monfort A. **2017**. Genetic analysis of the wild strawberry (*Fragaria vesca*) volatile composition. *Plant physiology and biochemistry* **121**:

99–117.

Watson A, Ghosh S, Williams MJ, Cuddy WS, Simmonds J, Rey M-D, Hatta MAM, Hinchliffe A, Steed A, Reynolds D. **2018**. Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature plants* **4**: 23.

Yocca A, Lu Z, Schmitz RJ, Freeling M, Edger P. **2019**. Evolution of conserved noncoding sequences in Arabidopsis thaliana. *bioRxiv*: 727669.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB. **2006**. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**: 203.

Yuan H, Zhang J, Nageswaran D, Li L. **2015**. Carotenoid metabolism and regulation in horticultural crops. *Horticulture research* **2**: 15036.

Zhang J, Zhao J, Xu Y, Liang J, Chang P, Yan F, Li M, Liang Y, Zou Z. **2015**. Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor. *Frontiers in plant science* **6**: 1042.

Zhou X. **2017**. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics* **11**: 2027.

Zhou X, Carbonetto P, Stephens M. **2013**. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics* **9**: e1003264.

Zhou X, Stephens M. **2012**. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**: 821.

Zini E, Biasioli F, Gasperi F, Mott D, Aprea E, Märk TD, Patocchi A, Gessler C, Komjanc M. **2005**. QTL mapping of volatile compounds in ripe apples detected by proton transfer reaction-mass spectrometry. *Euphytica* **145**: 269–279.

Brief legends for the Supplemental Information:

Fig. S1: Raw frequency distribution of 17 volatiles

Fig. S2: SNP density

Fig S3a and S3b: Manhattan plots and the respective quantile-quantile plots

Fig S4: Distribution of GWAS peaks and percentage of phenotypic variation

Fig S5: Chromosomal partition of the variance

Fig S6: Heatmap of the realized genomic matrix

Fig S7: Boxplot of the predictive abilities

Fig S8: Linear relationship between PGE and predictive ability

Fig S9: p-values of the Pearson's correlation and Principal Component Analysis (PCA)

Table S1: Annotation of candidate genes underlying and flanking significant SNPs related to volatile emission in blueberry (Excel file)

Table S2: Scenarios for genomic prediction and marker-assisted selection

Table S3: Number of raw and filtered SNPs used in the GWAS study

Table S4: Gene ontology (GO) enrichment analyses (Excel file)

Table S5: Molecular markers used as fixed effects for genomic selection

Table S6: Metabolite concentration and hedonic ratings (Excel file)

Full legends for the Figures:

Fig 1: Schematic representation of the pathways leading to the biosynthesis of volatiles detected in this study. Volatiles with uncharacterized pathways are indicated with a question mark (?). Different color shadings indicate different metabolic pathways. Arrows with triple heads indicate multiple steps. Colored asterisks on the pathways indicate the potential enzymatic role of candidate genes detected in the significant genomic regions. Chromosomes with significant associations in the GWAS analyses are indicated by colored circles in front of each volatile organic compound.

Fig2: Significant associations and candidate genes underlying linalool and eucalyptol volatiles. a) Manhattan plots showing the significance of each single nucleotide polymorphism (SNP) association in the GWAS. The horizontal red lines represent the Bonferroni significance threshold. b) Genomic regions harboring biosynthetic enzyme-coding genes (blue and red arrows) and the nearest significant SNPs (yellow triangles). LIS and TPS indicate the enzymes linalool synthase and alpha-terpineol synthase, respectively. Double bars indicate out of scale. c) Pairwise linkage disequilibrium (r2) between significant SNPs along the highlighted genomic region.

Fig3: Heritability estimations. h2 refers to pedigree-based heritability; h2snp is the genomic heritability computed using molecular markers. For both cases, PVE is the proportion of the

phenotypic variance explained by genetic terms. Bayesian Sparse Linear-Mixed Model (BSLMM) is a Bayesian version of h2snp, but assuming single nucleotide polymorphism (SNP) effects sampled from a point-normal distribution. In BSLMM, the genetic term is divided as: PVE is the proportion of phenotypic variance explained by the polygenic term and PGE is the proportion of the PVE explained by SNPs with a non-zero effect.

Fig 4: Phenotypic prediction a) Principal component analysis (PCA) of two blueberry populations: POP1 represents the original 886 individuals used for the GWAS analysis and POP2 is a new set of 552 individuals used for genomic prediction. b) Four prediction scenarios using different approaches: genomic selection (GS) intra and inter-population (all markers simultaneously modeled as random effects and predictive ability measured within POP1 and across populations), GS *de novo* GWAS (GWAS hits modeled as fixed effects into GS models and cross-validation performed across populations) and marker assisted selection (MAS – only the GWAS hits modeled as fixed effects in regression models and cross-validation performed across populations). c) Predictive ability for volatile content across the four prediction scenarios.

Fig 5. Sensory analysis. a) Pearson's correlation between five sensory scores and biochemical compounds; b) Comparison of blueberry `Kestrel` and `Windsor`, evidencing the importance of sugar to acid ratio on the liking rates. Comparison of `Kestrel` and `Snowchaser`, both cultivars with similar content of soluble solids and titratable acidity (TA), but different volatile profiles and liking perception.

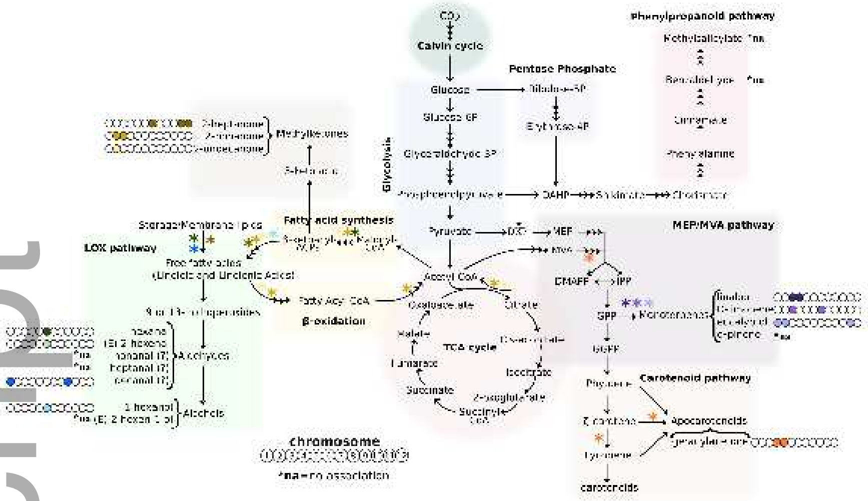
Table 1. Classification and summary statistics for the 17 volatile organic compounds detected among 886 blueberry individuals.

Volatile	Metabolic Classification	CAS Number	Aroma Descriptor[1]	Mean	SD
(E)-2-hexenal	Fatty acid derivate, aldehyde	6728-263	fresh, green, fruity	65376	1370
decanal	Fatty acid derivate, aldehyde	112-31-2	fruity, citrus	12184	283
heptanal	Fatty acid derivate, aldehyde	7785-70-8	green, herbal	166.4	3.1
hexanal	Fatty acid derivate, aldehyde	66-25-1	fresh, green, fruity	21874	547
nonanal	Fatty acid derivate, aldehyde	124-19-6	rose, fresh	346.8	12
(E)-2-hexen-1-ol	Fatty acid derivate, alcohol	928-95-0	fresh green, leafy	722.1	22.6

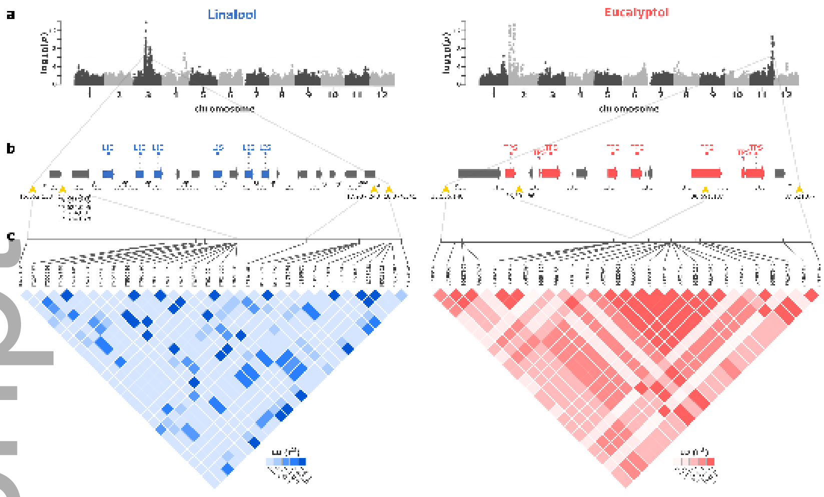
1-hexanol	Fatty acid derivate, alcohol	111-27-3	fruity, sweet green	631.4	14.8
2-heptanone	Fatty acid derivate, methyl ketone	110-43-0	fruity, spicy, cheesy	96.4	4.8
2-nonanone	Fatty acid derivate, methyl ketone	821-55-6	fruity, earthy	551.7	15
2-undecanone	Fatty acid derivate, methyl ketone	112-12-9	fruity, floral	2166.8	38.6
alpha-pinene	MVA/MEP, monoterpene	80-56-8	minty	6.1	0.4
D-limonene	MVA/MEP, monoterpene	5989-27-5	citrus, fresh, sweet	190.6	11
eucalyptol	MVA/MEP, monoterpene	470-82-6	minty, woody, herbal	99.5	6.4
geranyl acetone	MVA/MEP, monoterpene	3796-70-1	floral, rosy, sweet	475.9	13.2
linalool	MVA/MEP, monoterpene	78-70-6	green, rosy, floral	743.4	22.2
benzaldehyde	Phenylpropanoid, benzenoid	122-78-1	sharp, bitter, cherry	8	0.4
methyl salicylate	Phenylpropanoid, benzenoid	119-36-8	minty	1.4	0.2

[1] Aroma descriptors retrieved from Du & Rouseff (2014)

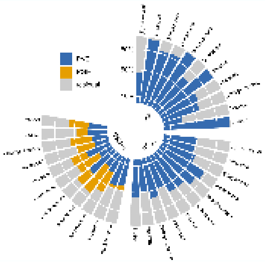
Mean and standard deviation (SD) values are expressed as ng*gFW^{-1}. Volatiles were classified in five chemical classes: aldehydes, alcohols, methyl ketones, monoterpenes and benzenoids.



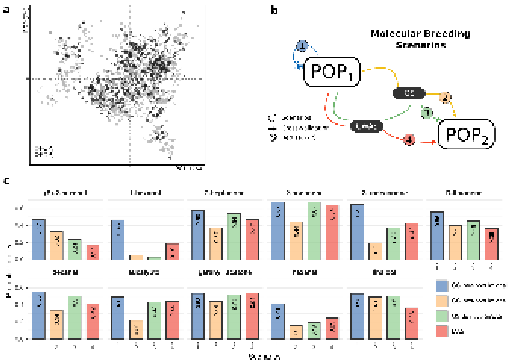
nph_16459_f1.tiff



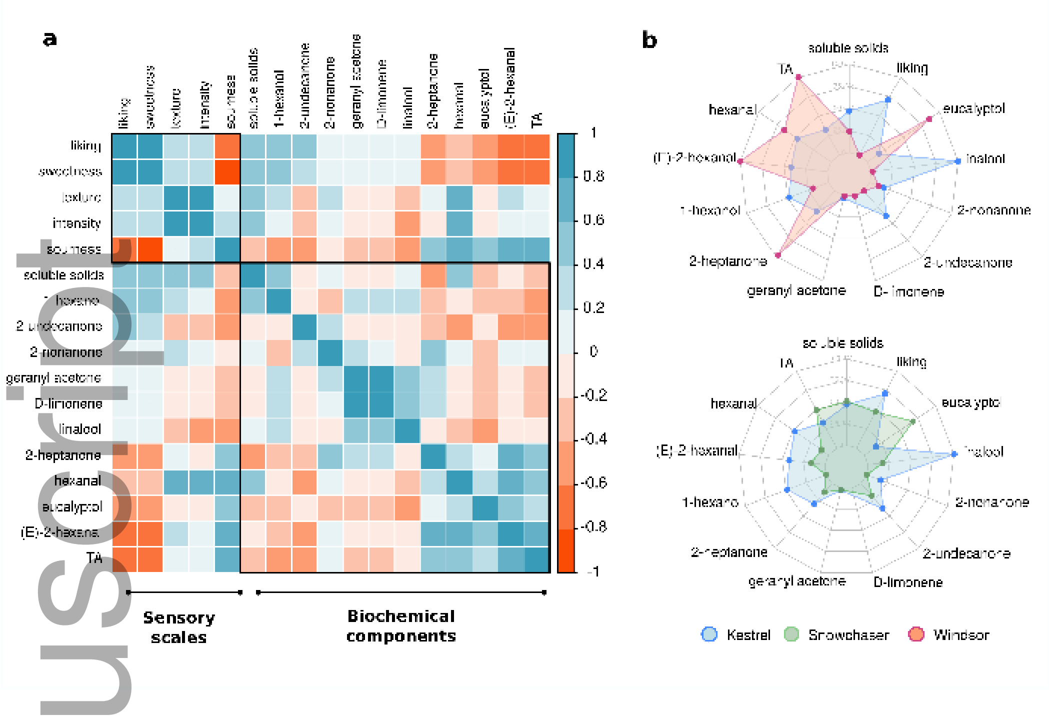
nph_16459_f2.tiff



nph_16459_f3.tiff



nph_16459_f4.tiff



nph_16459_f5.tiff