Supporting Information - ***New Phytologist***

Article title: **GENOME-WIDE ASSOCIATION OF VOLATILES REVEALS CANDIDATE LOCI FOR BLUEBERRY FLAVOR**

Authors: **Luís Felipe V. Ferrão; Timothy S. Johnson; Juliana Benevenuto; Patrick P. Edger; Thomas A. Colquhoun; Patricio R. Munoz**

Article acceptance date: **21 January 2020.**

The following Supporting Information is available for this article:

FIGURES

TABLES

Fig. S1. Raw frequency distribution of 17 volatiles traits measure in a Southern Highbush Blueberry population via gas chromatograph/mass spectrometry (GC/MS) approach.

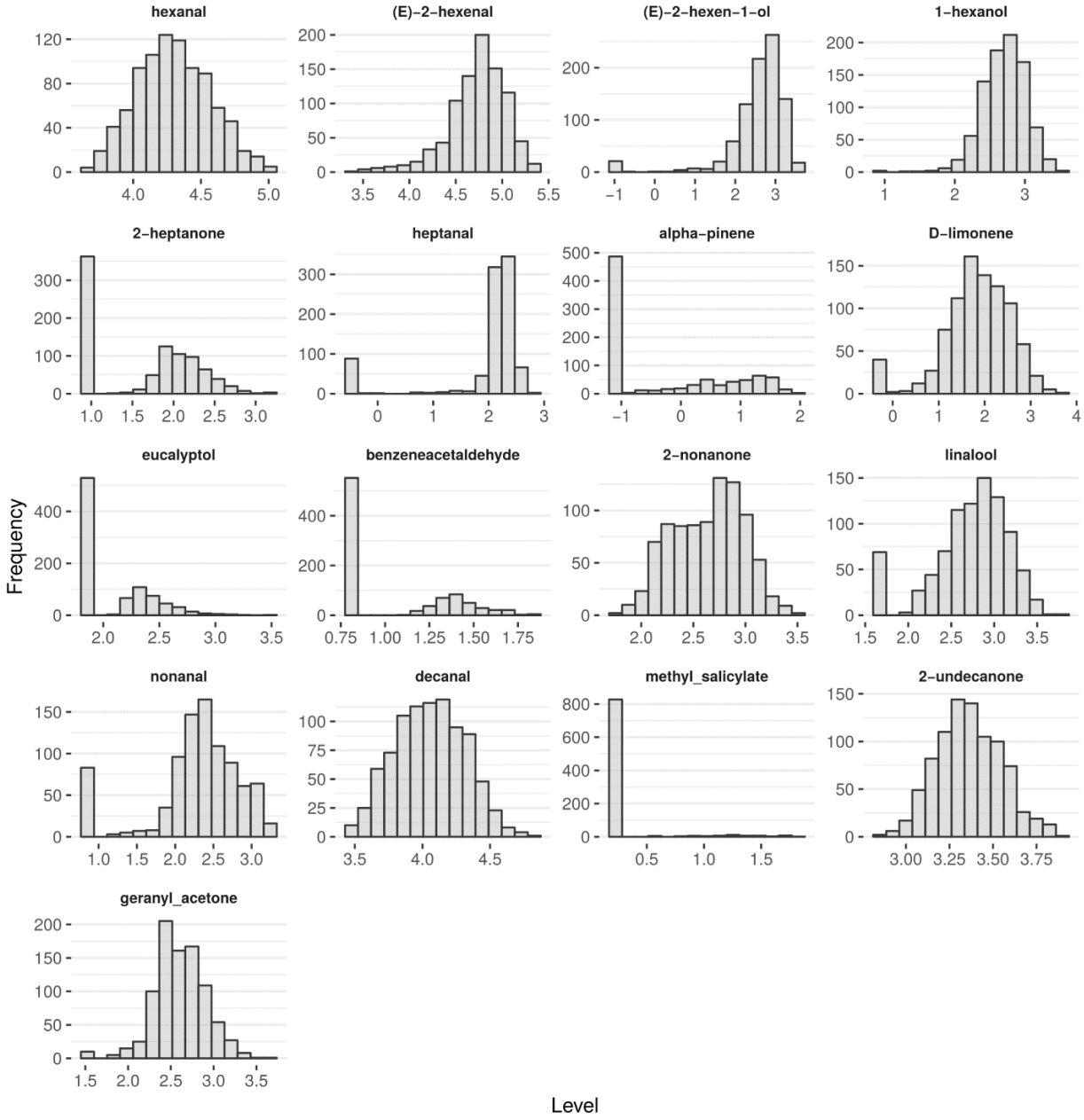

Fig. S2. Distribution of 71,487 filtered single nucleotide polymorphisms (SNPs) in 1 Mb window size across the 12 blueberry chromosomes. The x-axis represents the distance in base pairs.

Fig. S3a. Manhattan plots and the respective quantile-quantile plots for 6 volatile organic components quantified in a a Southern Highbush Blueberry population. A linear mixed model with corrections for population structure and cryptic relatedness was used to compute the pvalues. Bonferroni correction considering a genome-wide significance level of 0.05 (red line) was used for establishing a p-value detection threshold for statistical significance. For the 1-hexanol volatile we found one association with a pvalue value at the boundary of the Bonferroni threshold and therefore we maintained it in the subsequent analysis of functional mapping.

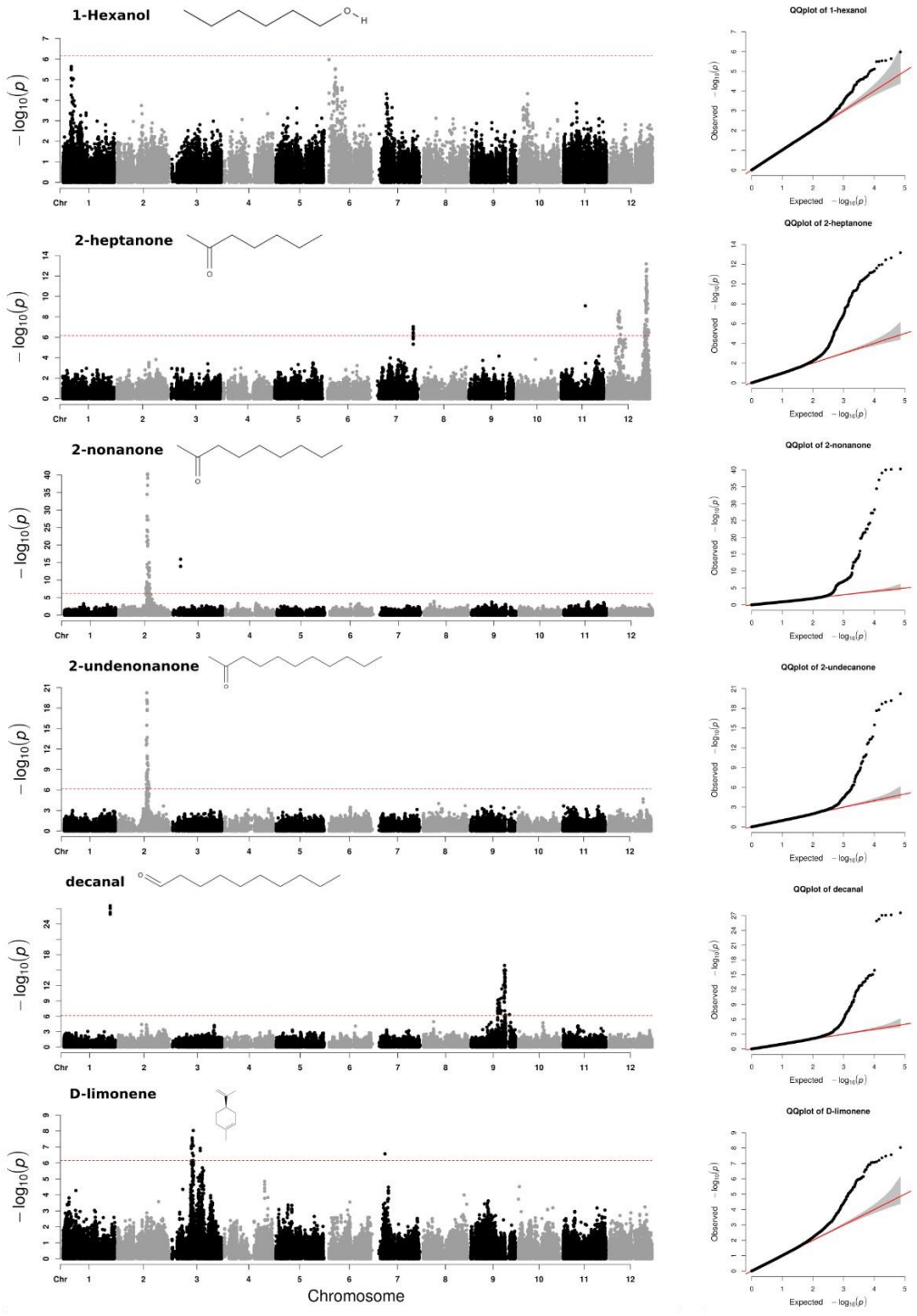

Fig. S3b. Manhattan plots and the respective quantile-quantile plots for 5 volatile organic components quantified in a a Southern Highbush Blueberry population. A linear mixed model with corrections for population structure and cryptic relatedness was used to compute the pvalues. Bonferroni correction considering a genome-wide significance level of 0.05 (red line) was used for establishing a p-value detection threshold for statistical significance.

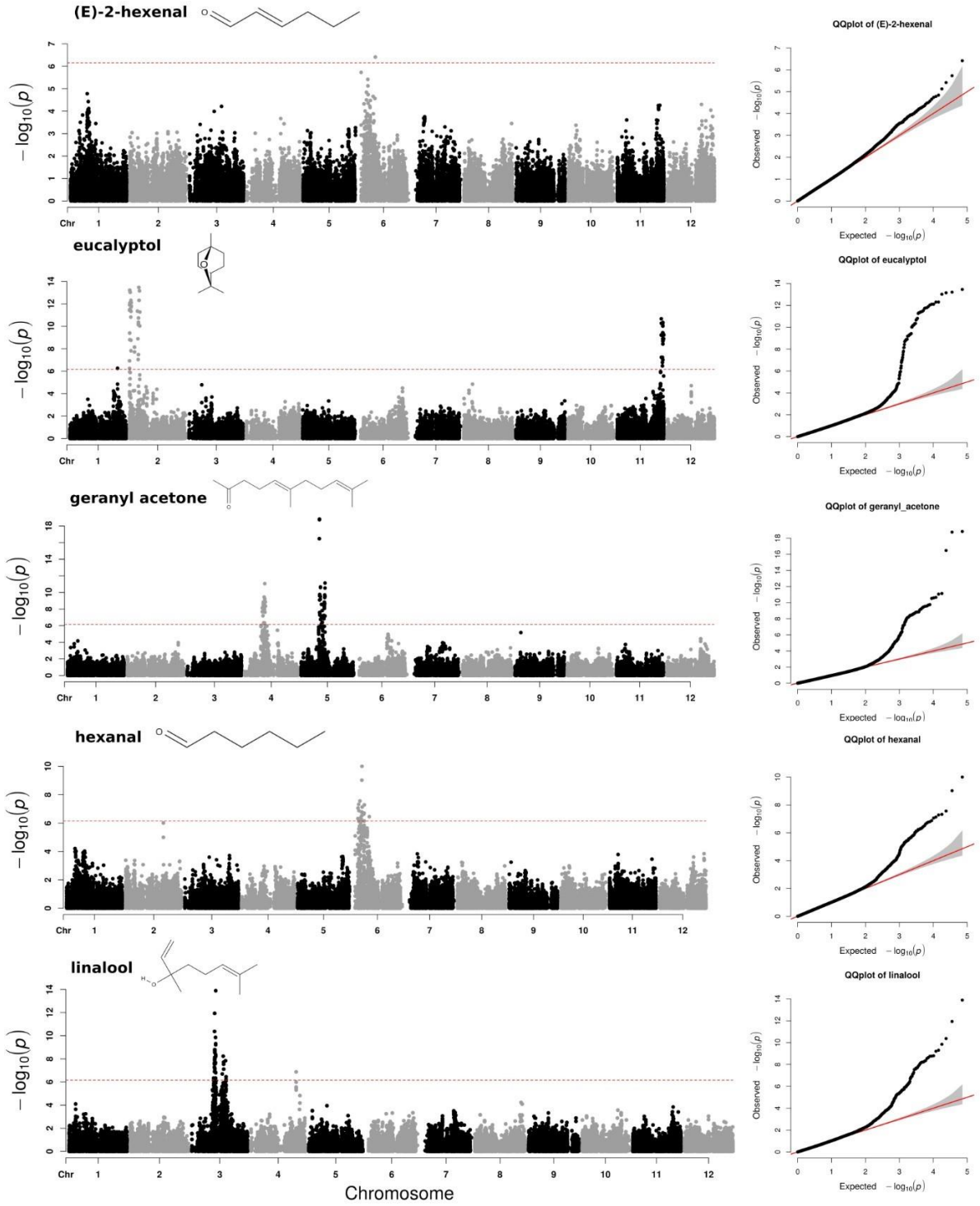

Fig. S4. A) Distribution of GWAS peaks across the 10 blueberry chromosomes where significant hits were found for 11 volatiles in a Southern Highbush Blueberry breeding population. Squares represent the genomic windows defined for functional candidate genes screening. Numbers indicate the number of significant associations within regions for each volatile. **B)** Position and effect of significant SNPs in relation to protein coding genes. **C)** Distribution of the percentage of phenotypic variation explained by individual markers. SNPs explaining a large portion of volatile variances are highlighted.

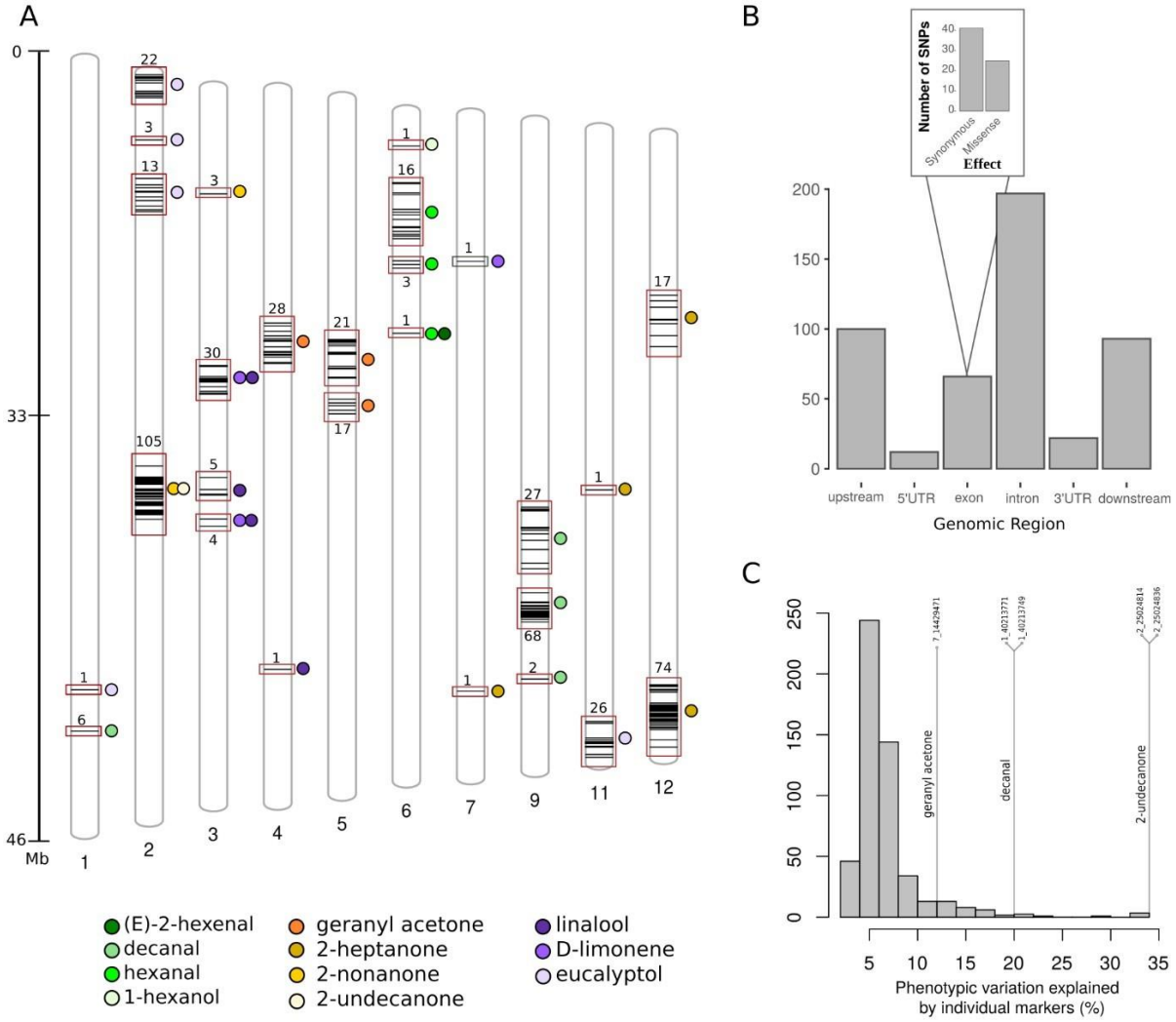

Fig. S5. Chromosomal partition of the variance. A linear regression considering the percentage of the variance explained per chromosome as response and its length (Mb) as explanatory variable was fitted. The slope p-values are reported for each volatile.

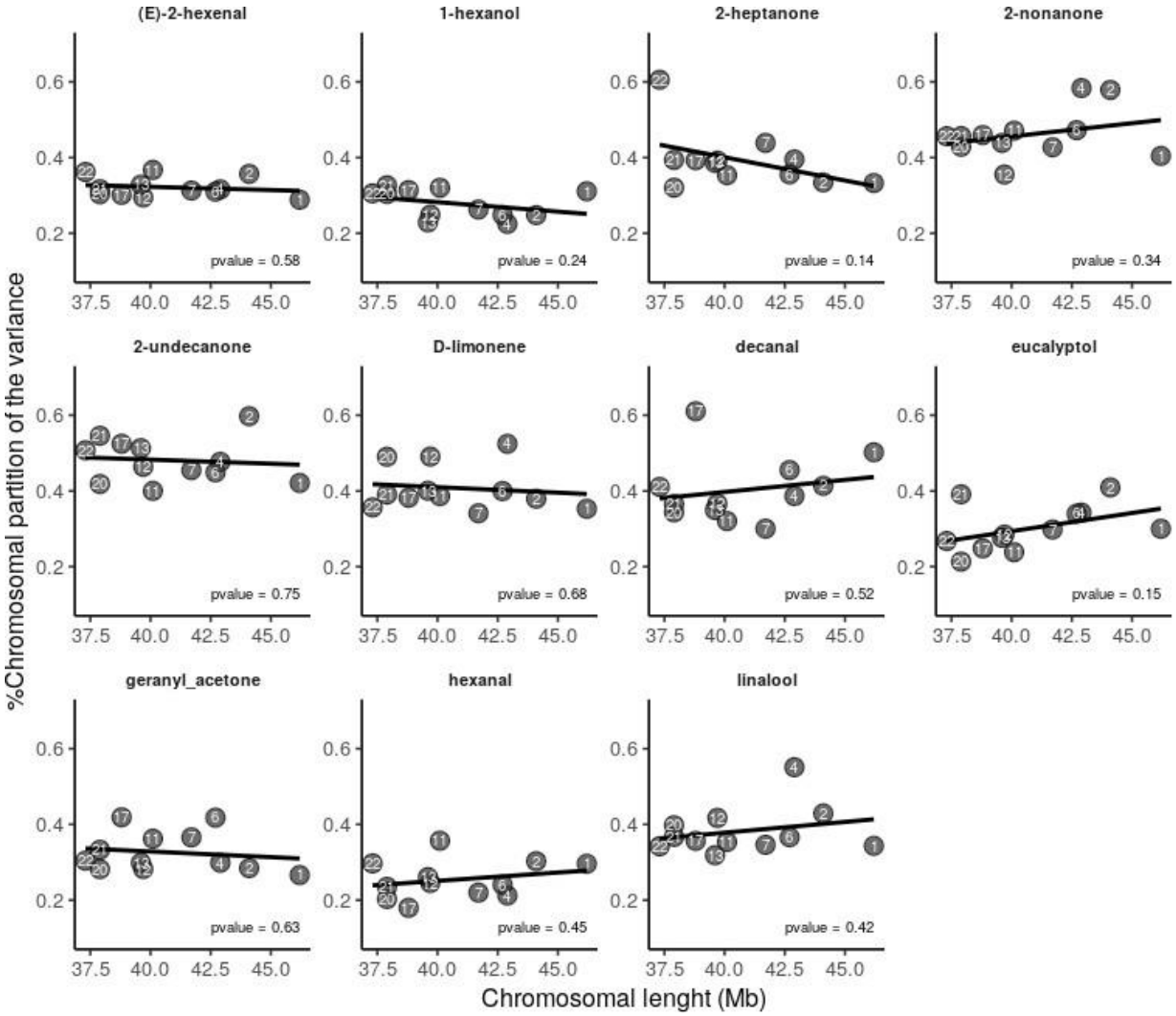

Fig. S6.. Heatmap of the realized genomic relationship matrix considering individuals from POP1 (886 genotypes used in the GWAS analysis) and POP2 (552 genotypes used for phenotypic prediction)

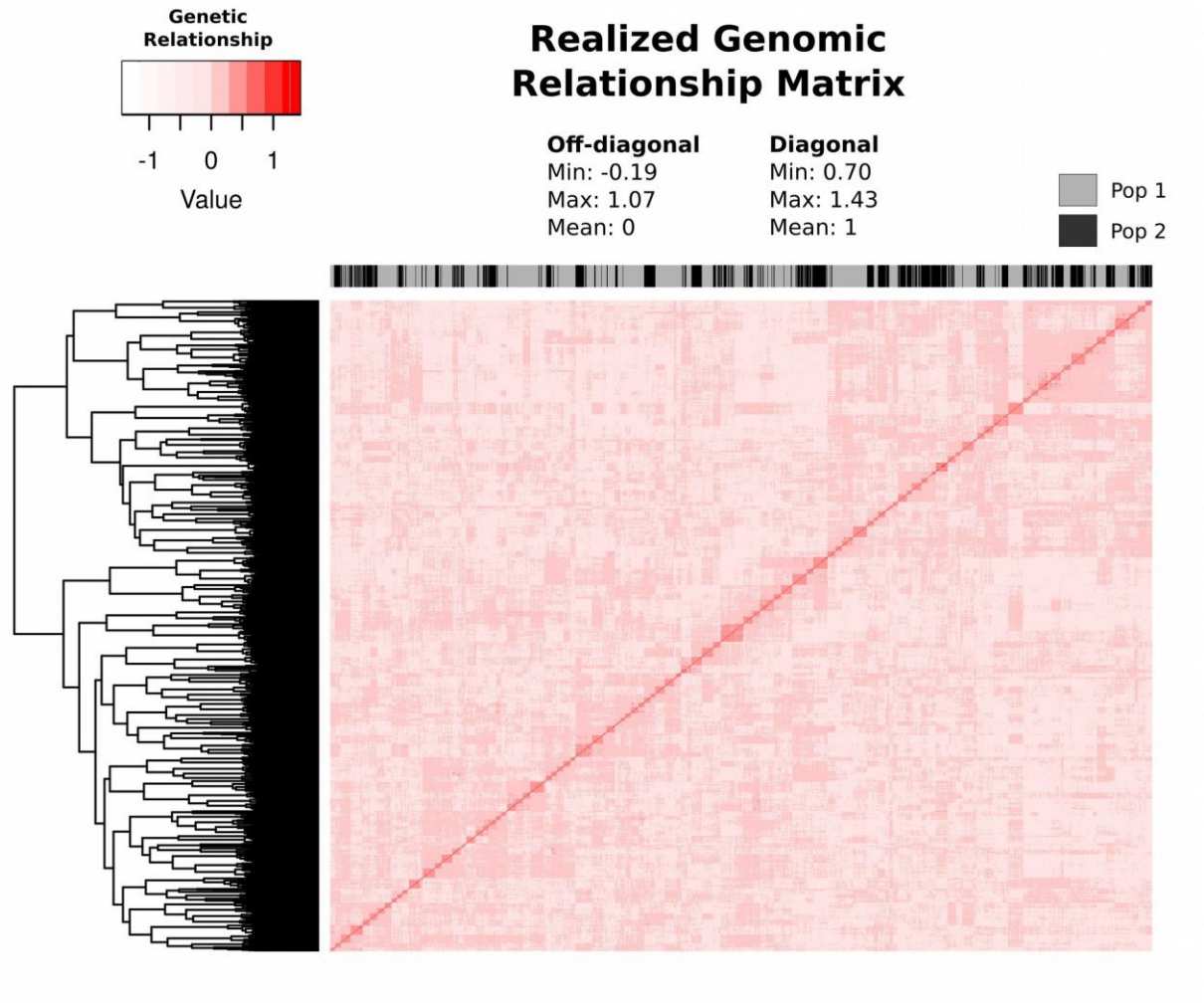

Fig. S7. Boxplots of the predictive abilities computed in the Scenario 1 (See also the Supporting Table S2.)

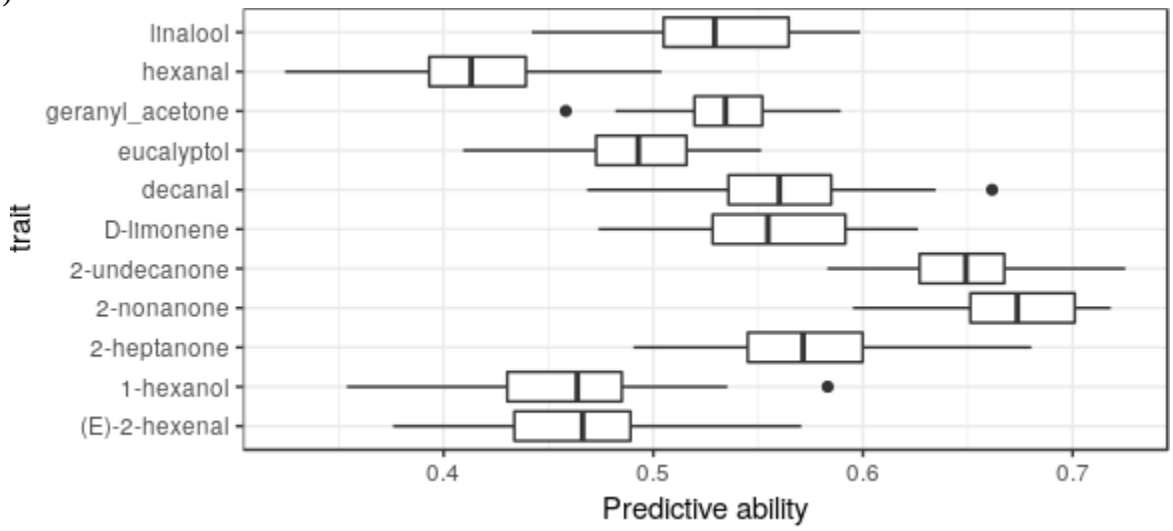

Fig. S8. Linear regression of the proportion of the variance explained by SNPs with a non-zero effect (PGE) as response and the predictive ability as explanatory variable.

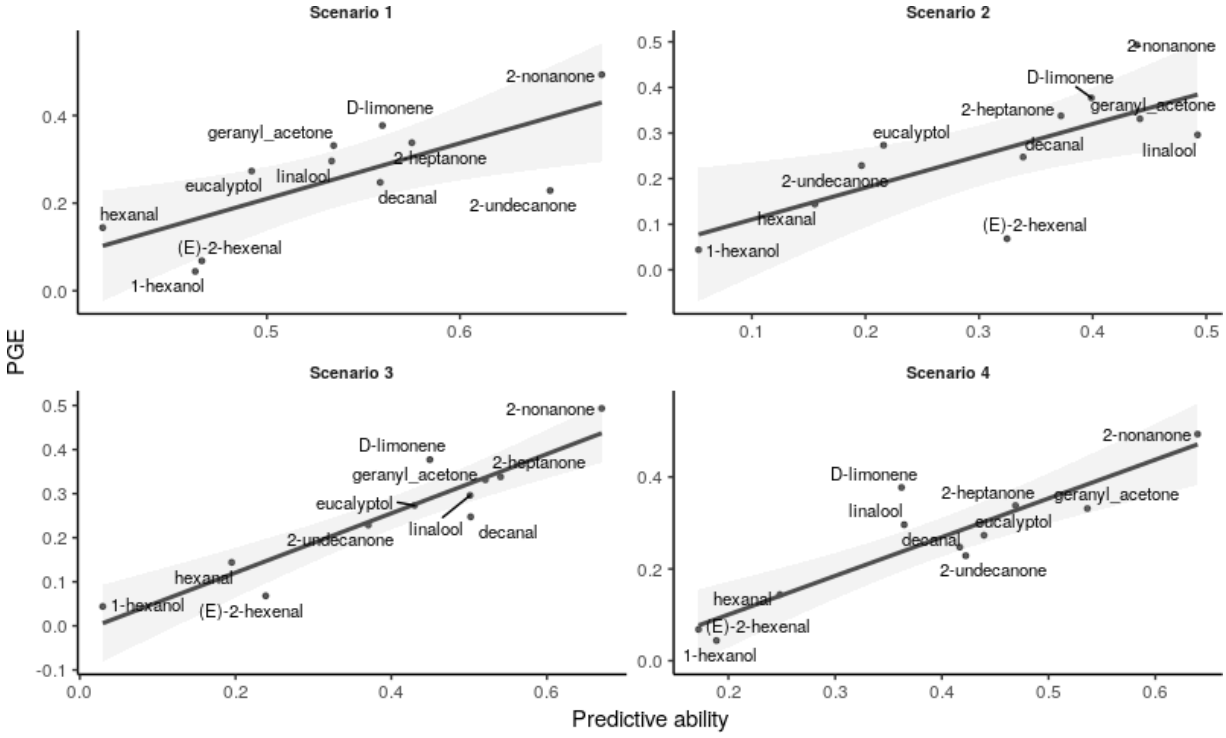

Fig. S9. a) P-values associated to the Pearson's correlations between five sensory scores and biochemical compounds (asterisks indicate P-values <0.05); **b)** Principal Component Analysis (PCA) showing the dispersion of the 24 blueberry cultivars used in the sensory analysis and the loading vectors associated to hedonic scales, volatiles organic components (VOCs) and Sugar and Acid (TA) contents.

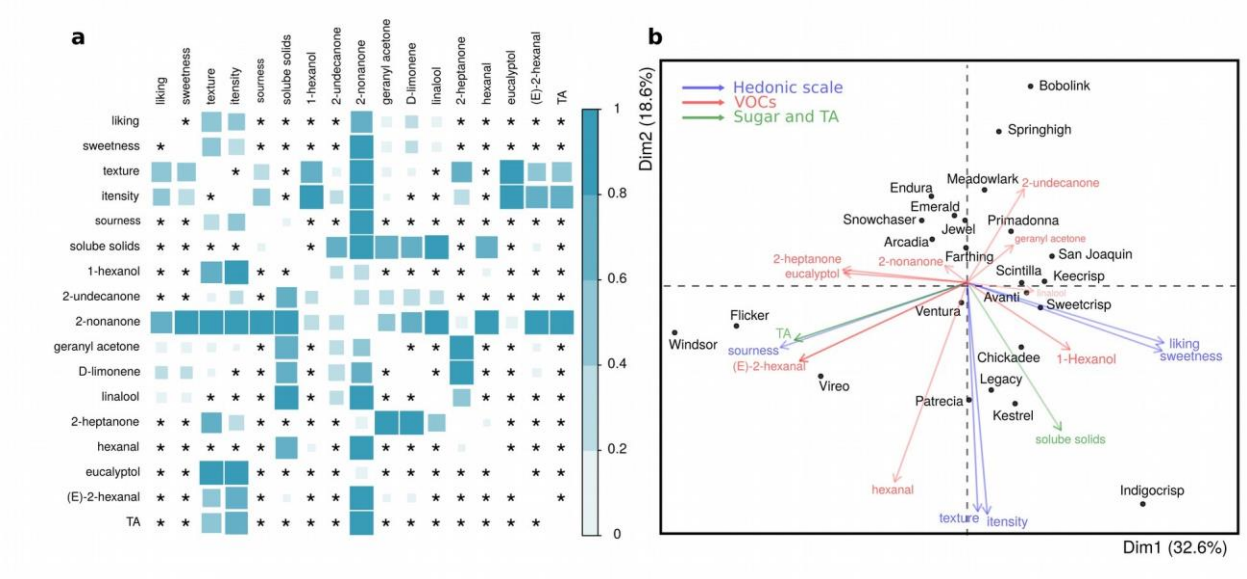

Table S1: Gene annotation. See separate Excel file.

Table S2. Different scenarios for phenotypic predictions and validations of the genome-wide association analyses carried out in two independent blueberry breeding populations (POP_1 and POP_2).

Scenarios	Training data*	Test data*	CV scheme**	Method
SCE_1	POP_1	POP_1	Rep TRN-TST	GS
SCE_2	POP_1	POP_2	CV	GS
SCE_3***	POP_1	POP_2	CV	GS *de novo* GWAS
SCE_4***	POP_1	POP_2	CV	MAS

* Models were fitted to the training data and prediction accuracy was evaluated in the test data. ** Replicated Training-Testing (Rep TRN-TST) design was created by randomly splitting the same population into a training (70% of the individuals) and a test data (remaining 30%), this division was randomly repeated 30 times. Cross-validation (CV) was designed by training and test the models in different populations. *** Scenarios considered as GWAS validation, since the peaks pinpointed in the GWAS analyses were used as fixed effect covariates in the prediction models.

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

Original Scaffold*	Chr Number**	Original Number of SNPs	Number of filtered SNPs
VaccDscaff1	1	26735	7130
VaccDscaff2	2	24037	6179
VaccDscaff4	3	27010	6989
VaccDscaff6	4	18537	4279
VaccDscaff7	5	23353	6138
VaccDscaff11	6	22536	6170
VaccDscaff12	7	18745	4999
VaccDscaff13	8	21621	5429
VaccDscaff17	9	23720	6218
VaccDscaff20	10	19063	5021
VaccDscaff21	11	25949	7220
VaccDscaff22	12	21968	5715
Total		**273274**	**71487**

* Name of the scaffolds reported in the 12 homoeologous groups of *Vaccinium corymbosum* cv. 'Draper' genome assembly (Colle et al., 2019).
** Correspondent chromosome number used in this study.

Table S4: GO enrichment. See separate Excel file.

Table S5. Number of markers and SNP ID (chromosome number followed by the position mapped in the blueberry reference genome) used as fixed effect in the genomic selection and marker-assisted selection models for phenotype prediction of 11 volatiles in blueberry.

Volatile	Number of Markers	SNP
hexanal	6	6_4846100, 6_5833102, 6_6959405, 6_7454411, 6_9154662, 6_9386681
(E)-2-hexenal	1	6_13287488
1-hexanol	1	6_2060825
2-heptanone	10	7_34574576,11_21631622,12_10311433,12_11028656,12_12028430,12_12702802,12_33292888,12_34308331,12_34870953,12_35635231
D-limonene	3	3_17433758,3_18328987,7_8744566
eucalyptol	8	1_37745262,2_277880,2_1187948,2_4043577,2_6342036,2_7439106,11_3558907,11_36820996
2-nonanone	4	2_24531136,2_25024836,2_26348811,3_6335120
linalool	5	3_17499786,3_18328987,3_23353598,3_24060052,4_34803765
decanal	7	1_40213749,9_23319804,9_24699848,9_25606761,9_26765897,9_28754698,9_29149685
2-undecanone	2	2_24668598,2_25796741
geranyl acetone	6	4_16350874,4_15912982,5_14525661,5_16139756,5_16676640,5_18398135

Reference:

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.