New Phytologist Supporting Information

Article title: A multi-omics framework reveals strawberry flavor genes and their regulatory elements

Authors: Zhen Fan, Denise M. Tieman, Steven J. Knapp, Philipp Zerbe, Randi Famula, Christopher R. Barbey, Kevin M. Folta, Rodrigo R. Amadeu, Manbo Lee, Youngjae Oh, Seonghee Lee, Vance M. Whitaker*

Article acceptance date: 21 July 2022

Fig. S1. Model based clustering with k=4 (a) and LD decay (b).

Fig. S2. Comparisons of marker R2 and –logP among different types of eQTL.

Fig. S3. Dot plot maps the lead SNPs on Fana array for all eQTL (a). Density plot of minimum distance between significant SNPs on the Fana array (b).

Fig. S4. Characterization of trans-eQTL and their master regulators on chromosome 5D.

Fig. S5. Additional quality evaluations of the assembly.

Fig. S6. Karyoplots of F12 assembly (a) and Bea assembly (b).

Fig. S7. Hi-C contact map of Bea haploid assembly.

Fig. S8. Synteny plot between two haplotypes.

Fig. S9. Density distribution of expression ratios of genes (a). Allelic expression ratios of genes within alpha-linolenic acid metabolism pathway (b).

Fig. S10. Circos plot of allele-specific expressed genes (ASEs) in ‘FL15.89-25’.

Fig. S11. Cluster analysis and chemical relationships among volatiles.

Fig. S12. Linkage view of the two hotspots for median-chain esters (a) and terpenes (b).

Fig. S13. Marker prediction of overripe (a) and sweetness (b) scores.

Fig. S14. Manhattan plots of pentanal (a); butanoic acid, 2-methyl- (b); 2-hexenal, (E)- (c).

Fig. S15. Box-plots of relative abundance of pentanal (a), 2-hexenal, (E)- (b), and 2,3-butanedione (c) with different dosage of the alternative allele.

Fig. S16. Dosage effect on mesifurane abundance (a). Detection of indel using short reads alignment (b). High resolution melting curve for the InDel marker (c).

Fig. S17. Chromosomal alignment of 3C to the ‘Camarosa’ reference genome (a). Translated nucleotide sequences alignment between *FaNES1t* and *FaNES1* (b). Haplotypes of *FaNES1*.

Fig. S18. IGV view of short reads alignment to *FaNES1*.

Fig. S19. Schematics of three haplotypes at the *FaFAD1* region.

Fig. S20. Gene models, SV locations, genome alignments between haplotypes.

Fig. S21. Tissue-specific expression of *FaASa1* (a). Manhattan plot for methyl anthranilate (b).

Fig. S22. Long-range alignment to the *FaASa1* region.

Fig. S23. Genetic variations of genes involved in the anthocyanin pathway.

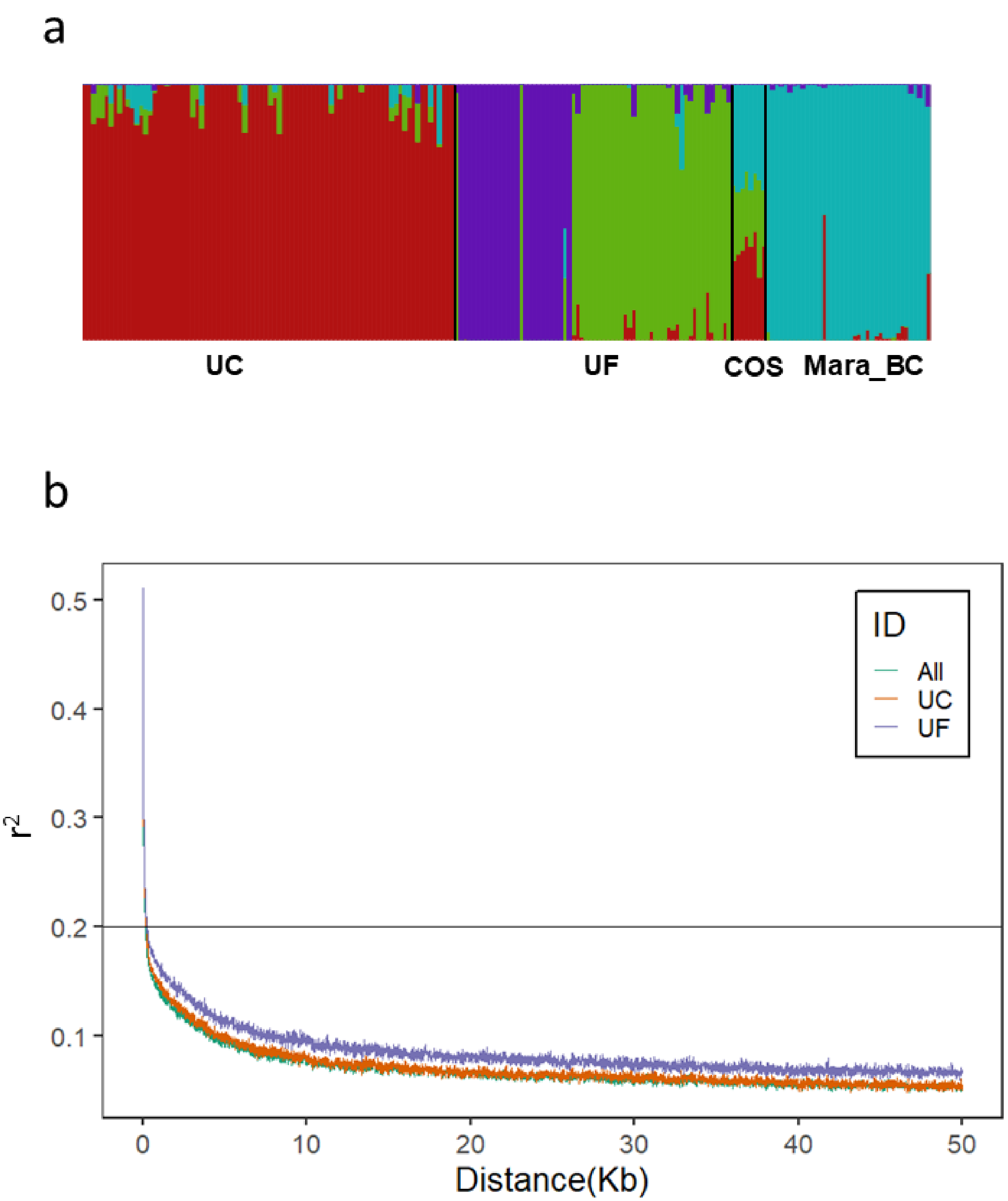

Fig. S1. (a) Model-based clustering of 196 strawberry accessions using a pruned SNP dataset including 168,476 SNP loci. The k (number of clusters) was set to 4. Populations are ordered based on geographic origin, separated by black vertical lines. Cos: “cosmopolitan”; Mara_BC: ‘Mara des Bois’ BC1; UF: University of Florida; UC: University of California (b) LD decay measured by $r^2$ in University of Florida (UF), University of California (UC) and combined set.

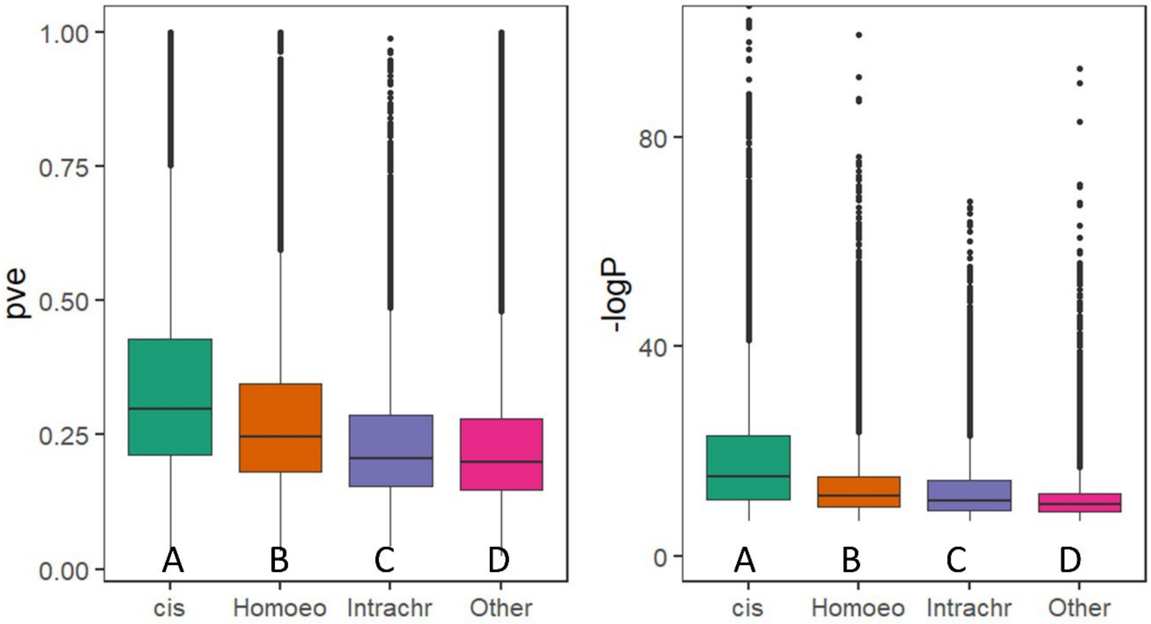

Fig. S2. Comparisons of marker $R^2$ (left) and –logP (right) among different types of eQTL. Label cis, homoeo, intrachr and other represent cis-eQTL, homoeologous trans-eQTL, intrachromosomal trans-eQTL, and other trans-eQTL, respectively. Letters below the boxes indicate the significant differences at α=0.05 using Tukey's HSD test. Boxes are delimited by upper and lower quantiles. Two whiskers represent highest/lowest values; dots represent outliers; and horizontal lines represent medians.

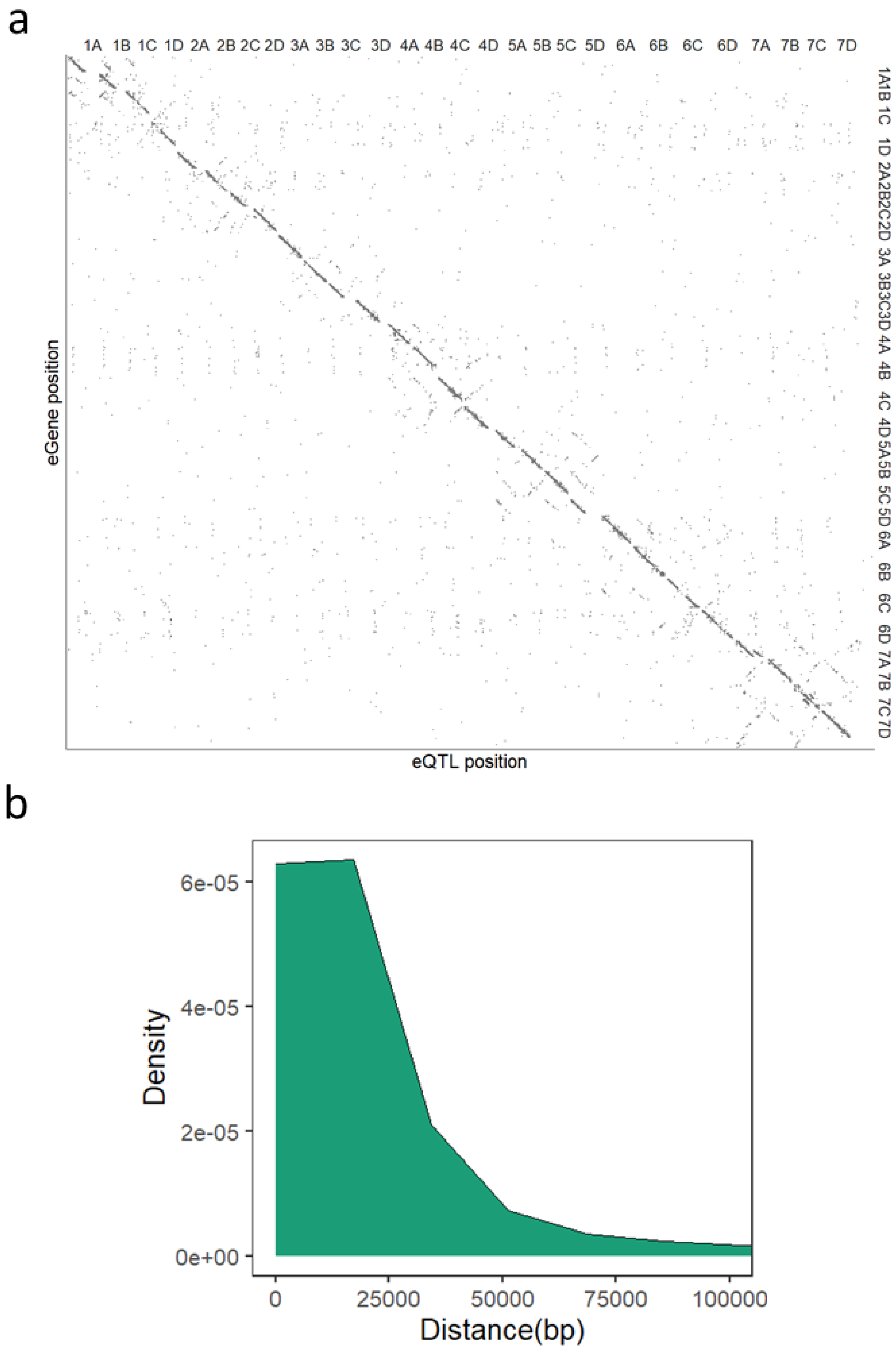

Fig. S3. (a) Dot plot maps the lead SNPs on Fana array of eQTL against the middle positions of mapped eGenes. (b) Density plot of minimum distance between significant SNPs on the Fana array within cis-eQTL and their associated eGenes

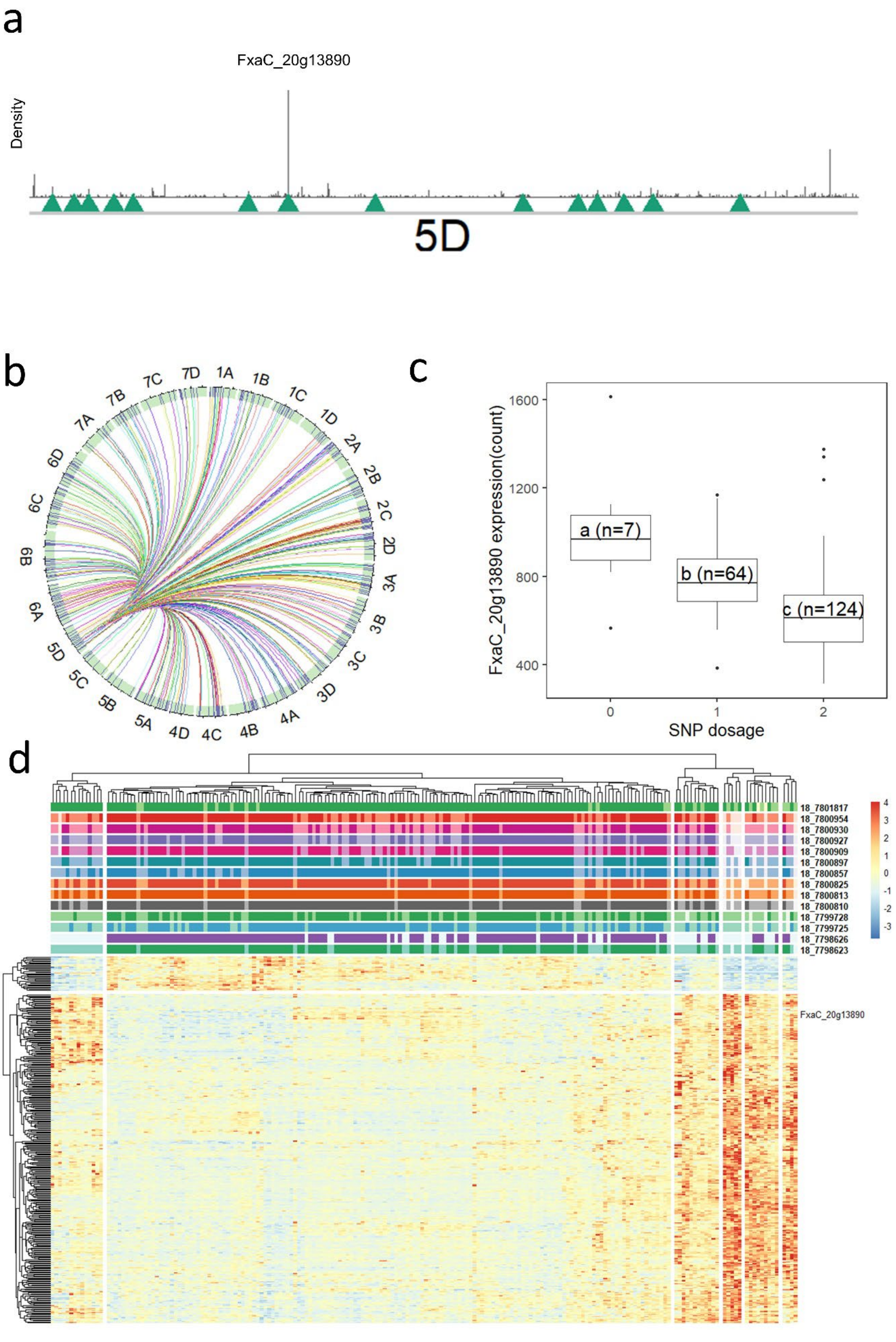
a
FxaC_20g13890
Density
5D
b
1A
1B
1C
1D
2A
2B
2C
2D
3A
3B
3C
3D
4A
4B
4C
4D
5A
5B
5C
5D
6A
6B
6C
6D
7A
7B
7C
7D
c
FxaC_20g13890 expression(count)
1600
1200
800
400
a (n=7)
b (n=64)
c (n=124)
0
1
2
SNP dosage
d
18_7801817
18_7800954
18_7800930
18_7800927
18_7800909
18_7800897
18_7800857
18_7800825
18_7800813
18_7800810
18_7799728
18_7799725
18_7798626
18_7798623
4
3
2
1
0
-1
-2
-3
FxaC_20g13890

Fig. S4. Characterization of trans-eQTL and their master regulators on chromosome 5D. (a) Trans-eQTL density on chromosome 5D. The locations of identified potential master regulators are annotated with green triangles. Location of putative regulators at the largest hotspot is labeled with their gene IDs. (b) Circos plot links the trans-eQTL hotsplot at 7796793 to 7803611 bp on chromosome 5D to physical positions of its trans-regulated genes. (c) Comparisons of FxaC_20g13890 expression with different dosage of wild alleles at the cis-eQTL. Different letters represent significant differences at alpha = 0.05 using TukeyHSD test. Boxes are delimited by upper and lower quantiles. Two whiskers represent highest/lowest values; dots represent outliers; and horizontal lines represent medians. (d) Heatmap showing scaled expression patterns of genes having trans-eQTL between 7796793 to 7803611 bp on chromosome 5D and the putative regulator FxaC_20g13890. The row for FxaC_20g13890 expression is labeled. The top panel is annotated with allelic dosage of 14 makers within the hotspot region. Color deepens with more doses of the alternative allele. Genotypes are grouped into six clusters, whereas genes are grouped into two clusters based on the hcluster function and cuttree with different k values.

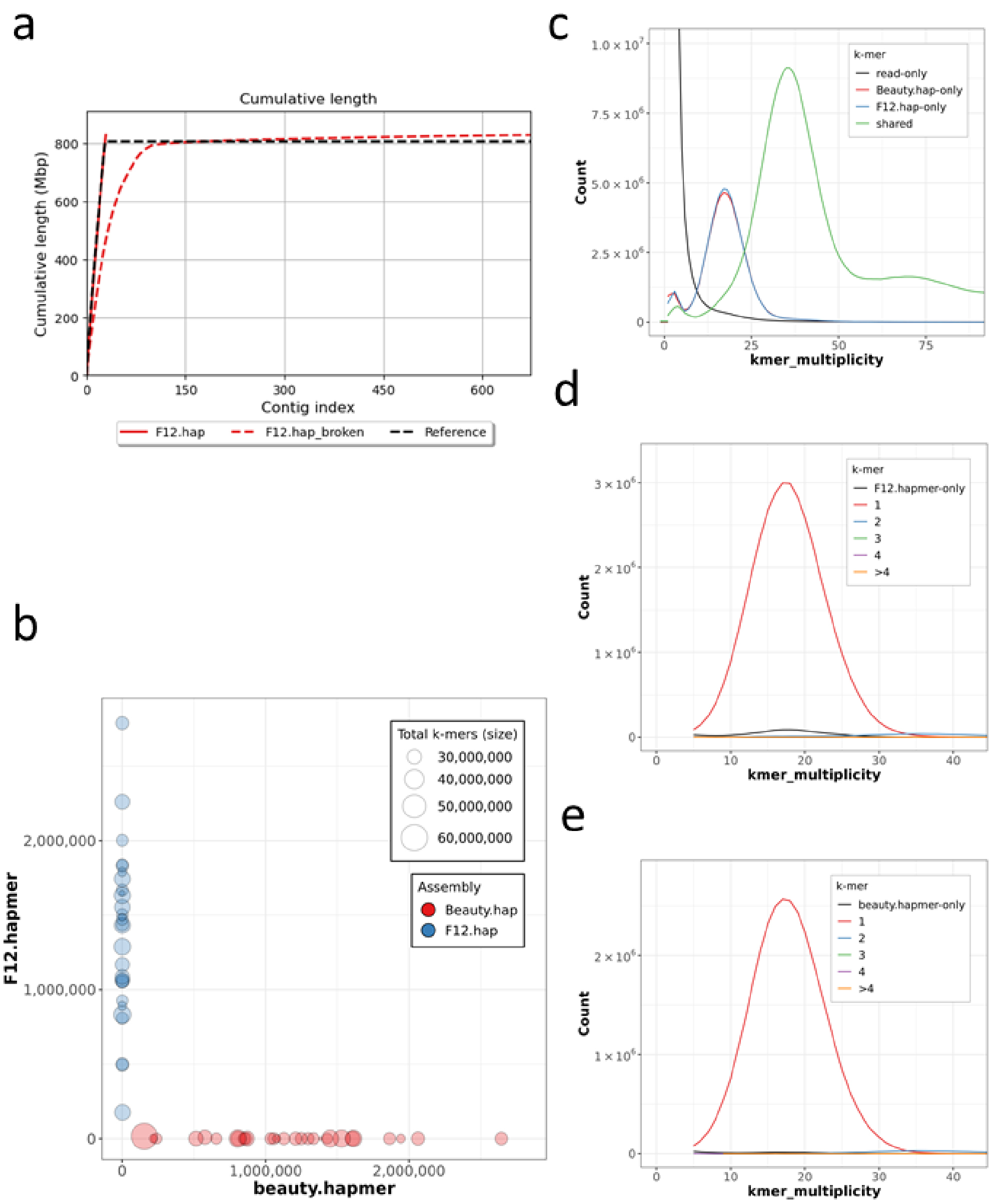

Fig. S5. (a) Cumulative length of contigs and scaffolds of the F12 haploid assembly. The red dashed line, red solid line, and black dashed line represent the accumulated length of phased contigs, pseudochromosomes and the pseudochromosomes of the reference genome. (b) Hapmer blob plot of the scaffolded assemblies. Red blobs represent Bea pseudochromosomes, while blue blobs are the F12 pseudochromosomes. Blob size is proportional to pseudochromosome size, and each blob/pseudochromosomes is plotted according to the number of contained Bea (*x* values) and F12 (y value) hap-mers. Results indicate F12 specific k-mers are only found in F12 pseudochromosomes and vice versa. (c) Copy number spectrum (spectra-cn) of the same k-mers plotted as unstacked histograms colored by the maternal (Beauty, red), paternal (F12, blue), shared between parents (green) and in the child's reads only. The assembly k-mers absent from the read set (likely to be base errors in the assembly) are plotted as a bar at zero multiplicity (close to 0 in our assemblies). (d&e) Hap-mer spectra-cn plot of Bea and F12 assembly, close to 0 F12 hapmers is found was Bea assembly, while close to 0 beauty hapmers was found in F12 assembly.

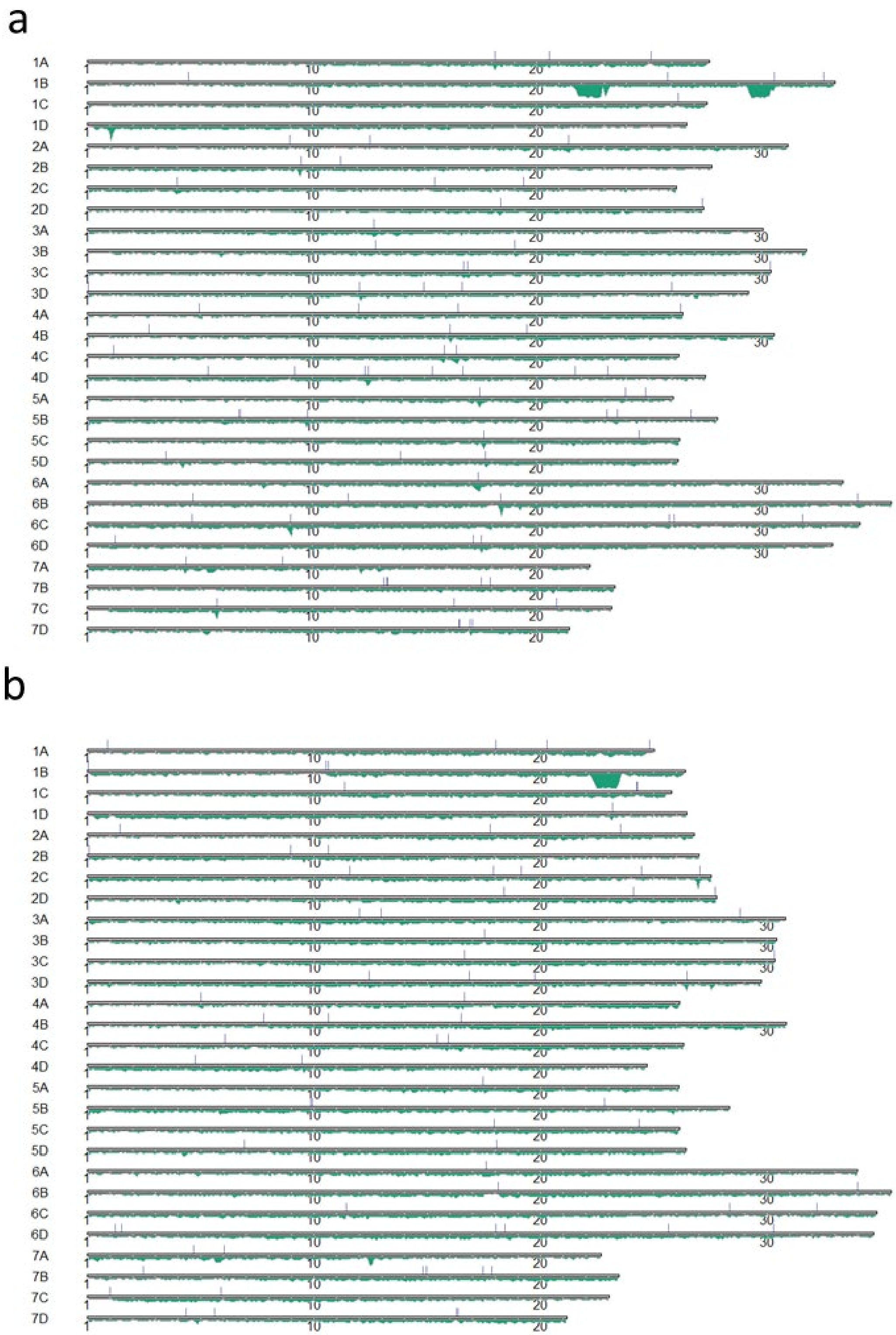

Fig. S6. Karyoplots of F12 assembly (a) and Bea assembly (b). TE density is plotted underneath the genome track in green and gaps in pseudochromosomes are plotted as vertical lines on top of the genome track.

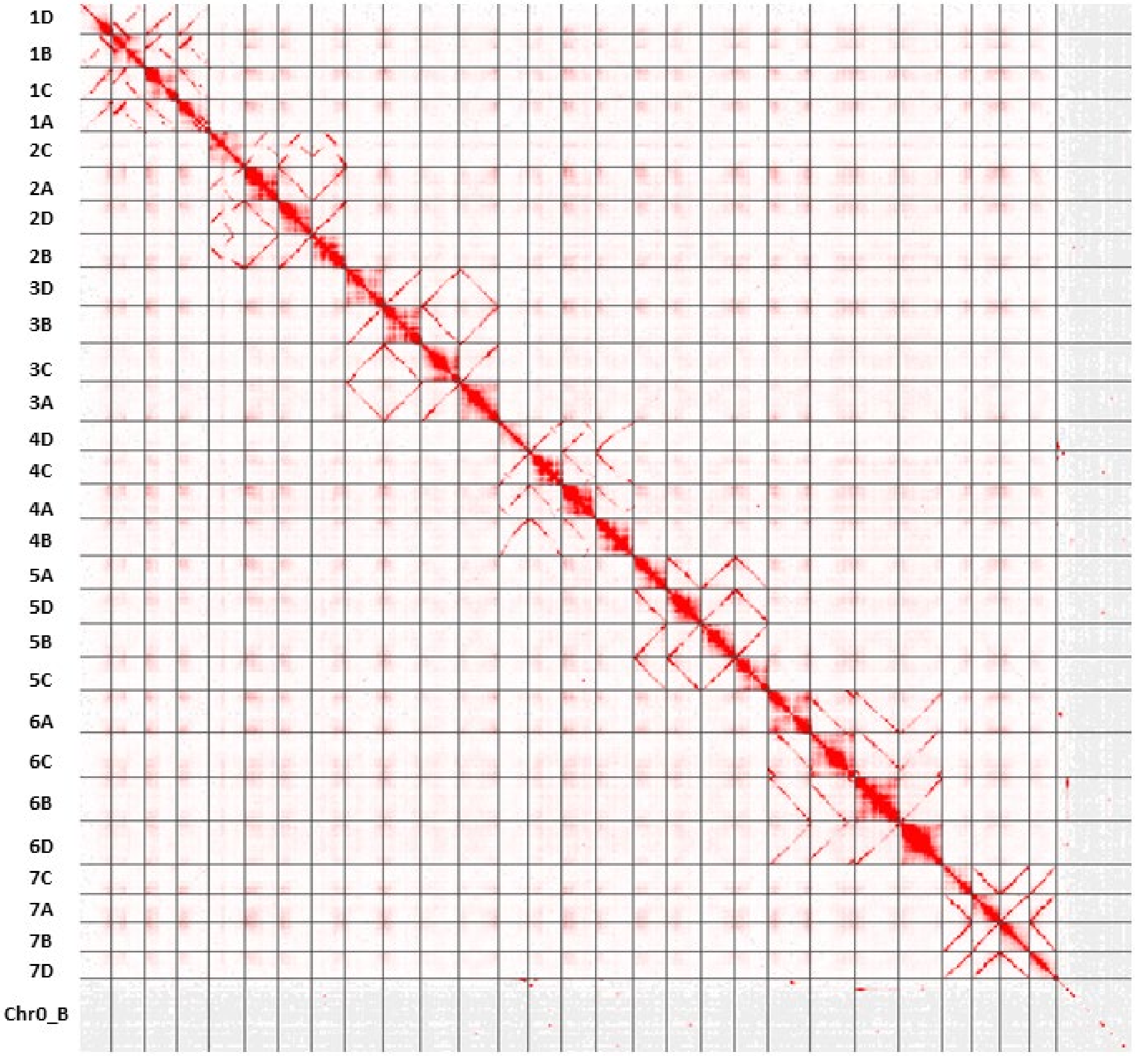

Fig. S7. Hi-C contact map of Bea haploid assembly using public Hi-C data from *Fragaria ×ananassa*.

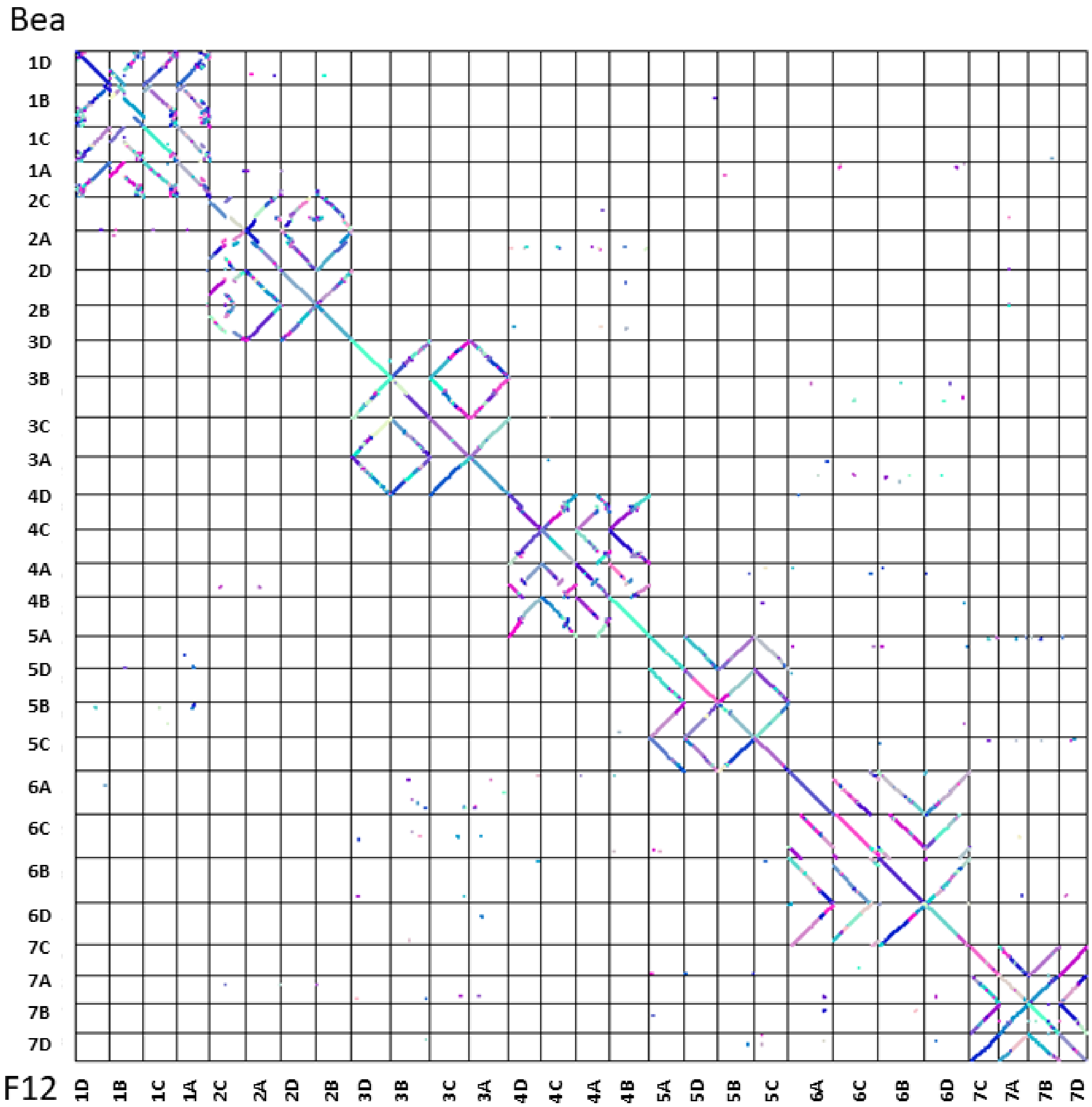

Fig. S8. Dot plot shows synteny between two haplotypes (x axis: F12 haplotype; y-axis: Bea haplotype).

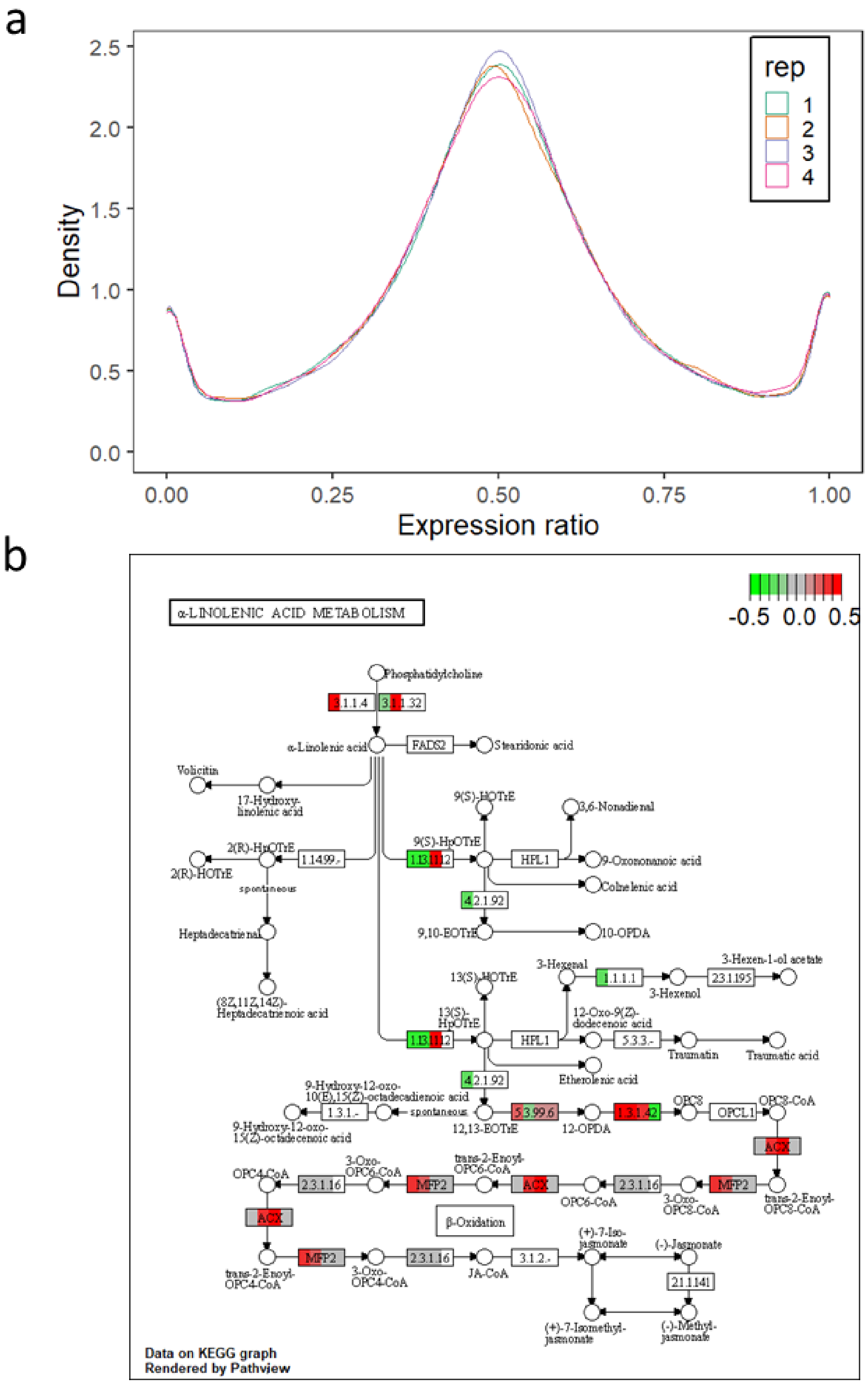

Fig. S9. (a) Density distribution of expression ratios of genes in ripe strawberry fruit across four biological replicates. (b) Allelic expression ratios of genes within alpha-linolenic acid metabolism pathway. Red cells represent higher expression from F12 allele, while green ones represent lower expression. Expression ratios were centered at 0. Up to four homologs were plotted for each enzyme.

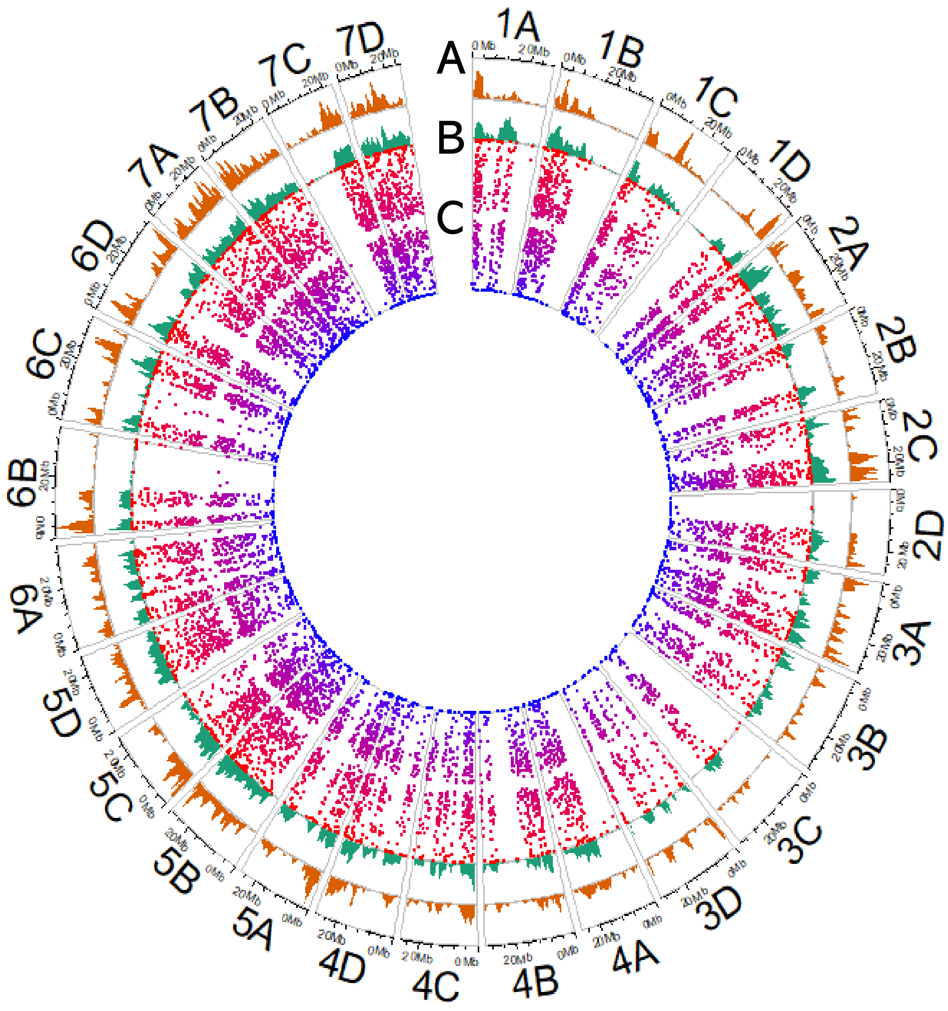

Fig. S10. Circos plot of allele-specific expressed genes (ASEs) in 'FL15.89-25'. Track A depicts SV density in exons. Track B depicts density of ASEs. Track C plots the median expression ratios of ASEs. The allelic expression ratio was computed by dividing the F12 allelic read count by the sum of the F12 allelic read count and the Bea allelic read count. Red or blue represents higher expression of the F12 or Bea allele. The range of median expression ratios was between 0 and 1.

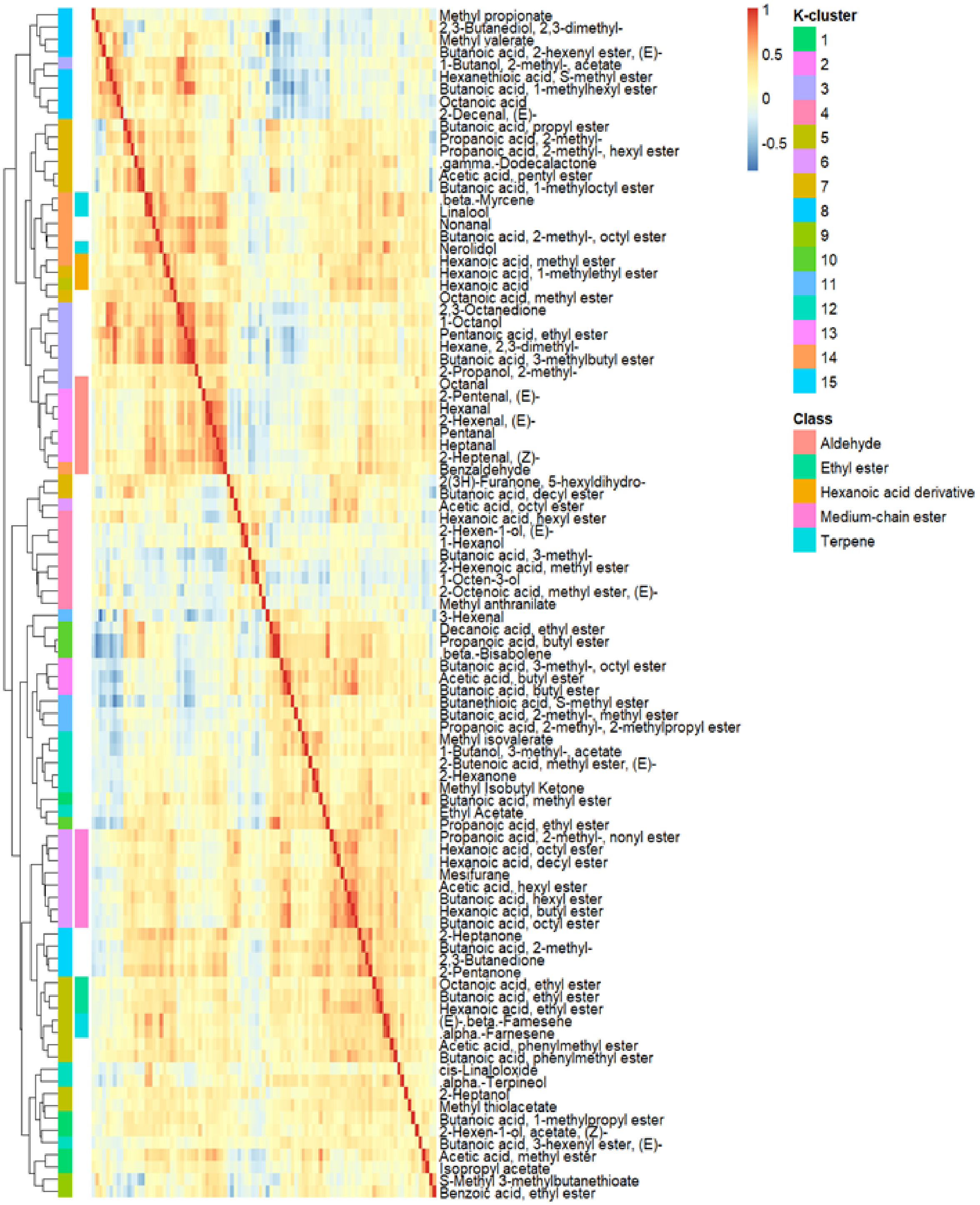

Fig. S11. Cluster analysis and chemical relationships among volatiles. K = 15 was used for K-clustering analysis. Five distinct volatile clusters are colored in the second column. Color of each cell in the center panel reflects the pairwise Pearson correlation between two volatiles.

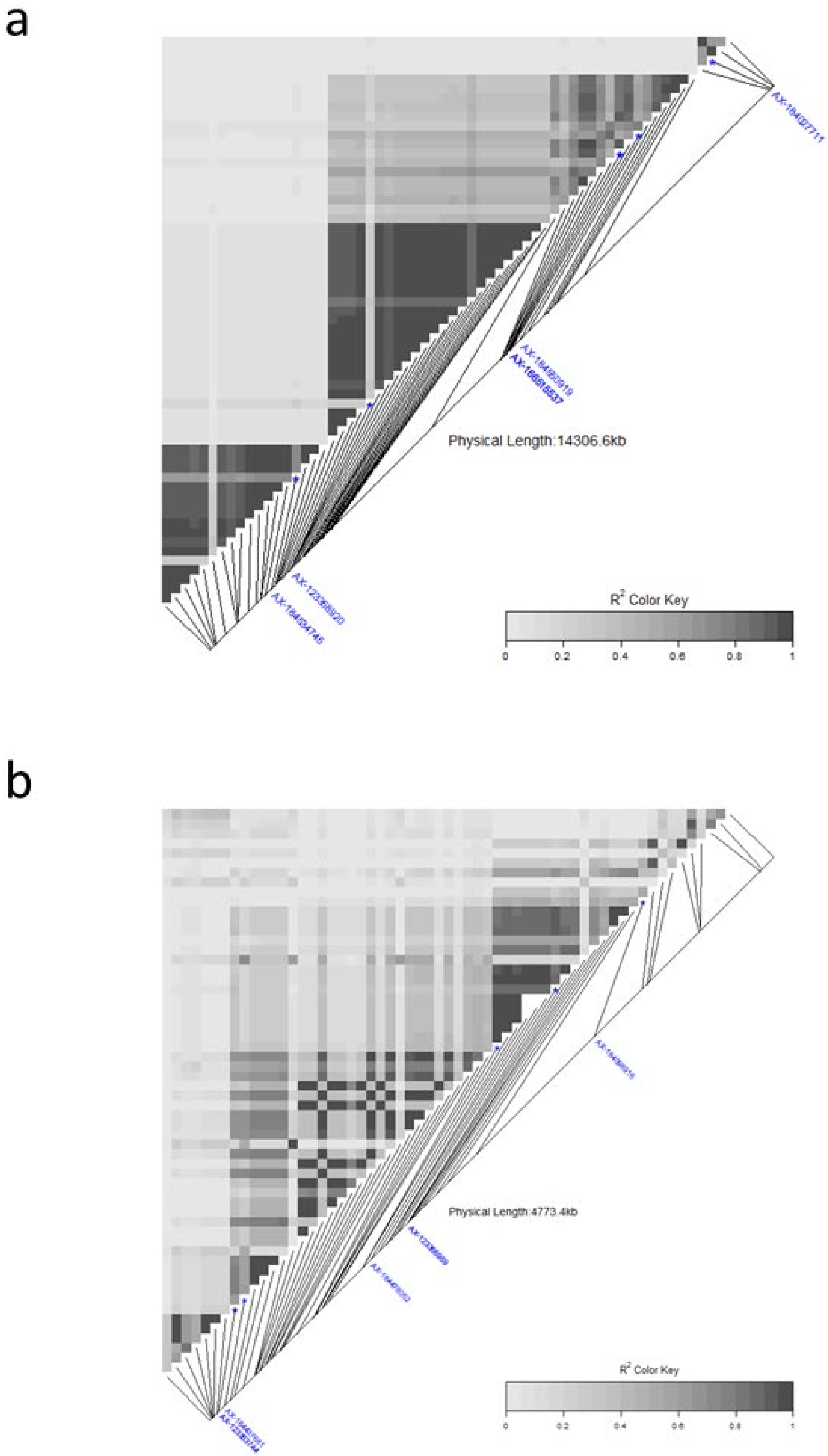

Fig. S12. Linkage view of the two hotspots for median-chain esters (a) and terpenes (b), respectively. The lead SNPs of the GWAS peaks are labeled. Blue asterisks represent significant marker positions.

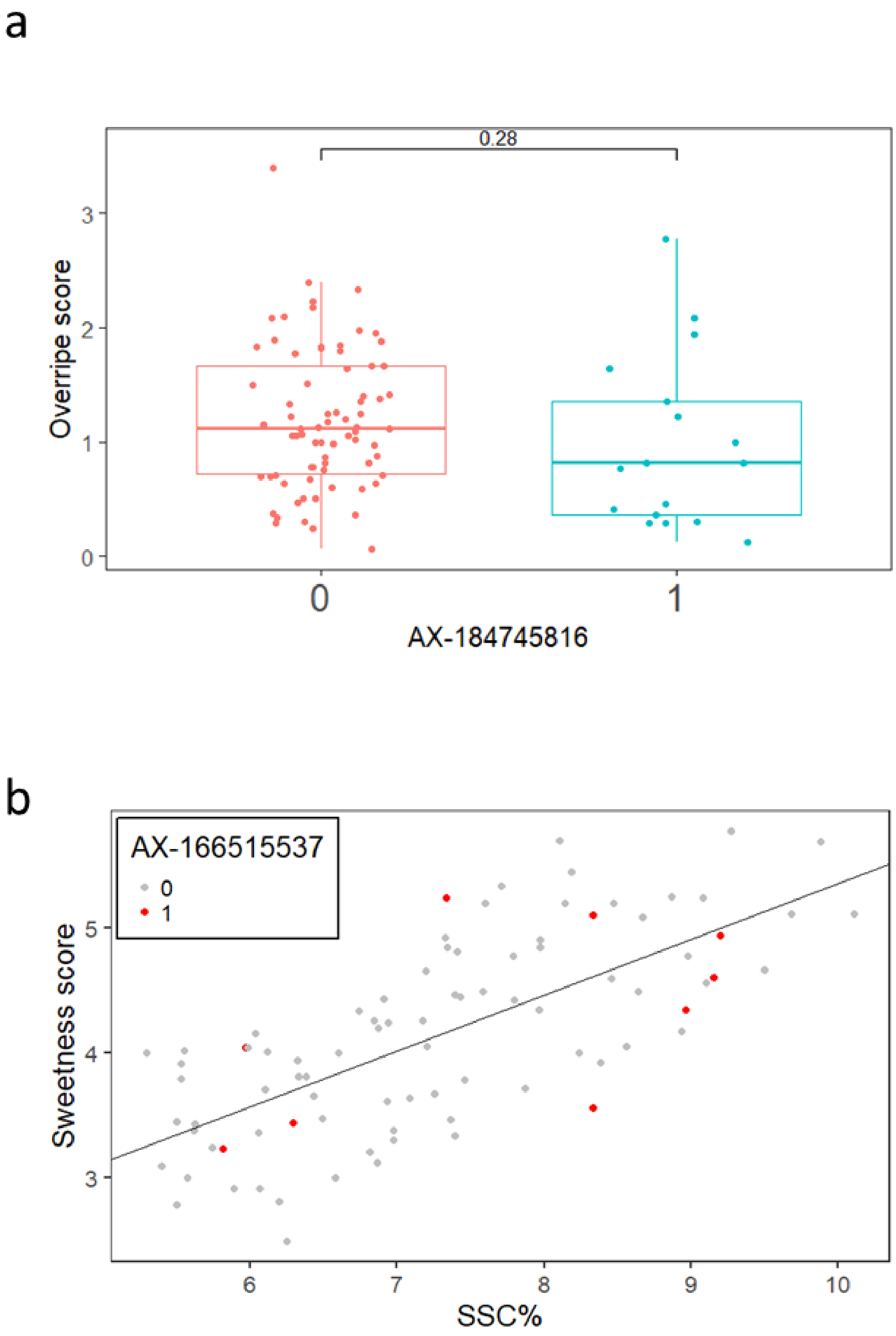

Fig. S13. (a) Marker effect of AX-184745816 on scores of overripe flavor. X axis indicates doses of the alternative allele. Boxes are delimited by upper and lower quantiles. Two whiskers represent highest/lowest values; dots represent individual values; and horizontal lines represent medians. (b) Sample sweetness scores are plotted with soluble solids content percentages (SSC%). The predicted sweetness score using linear regression is shown as a straight line. Samples with one dose of the alternative allele of marker AX-166515537 are in red color.

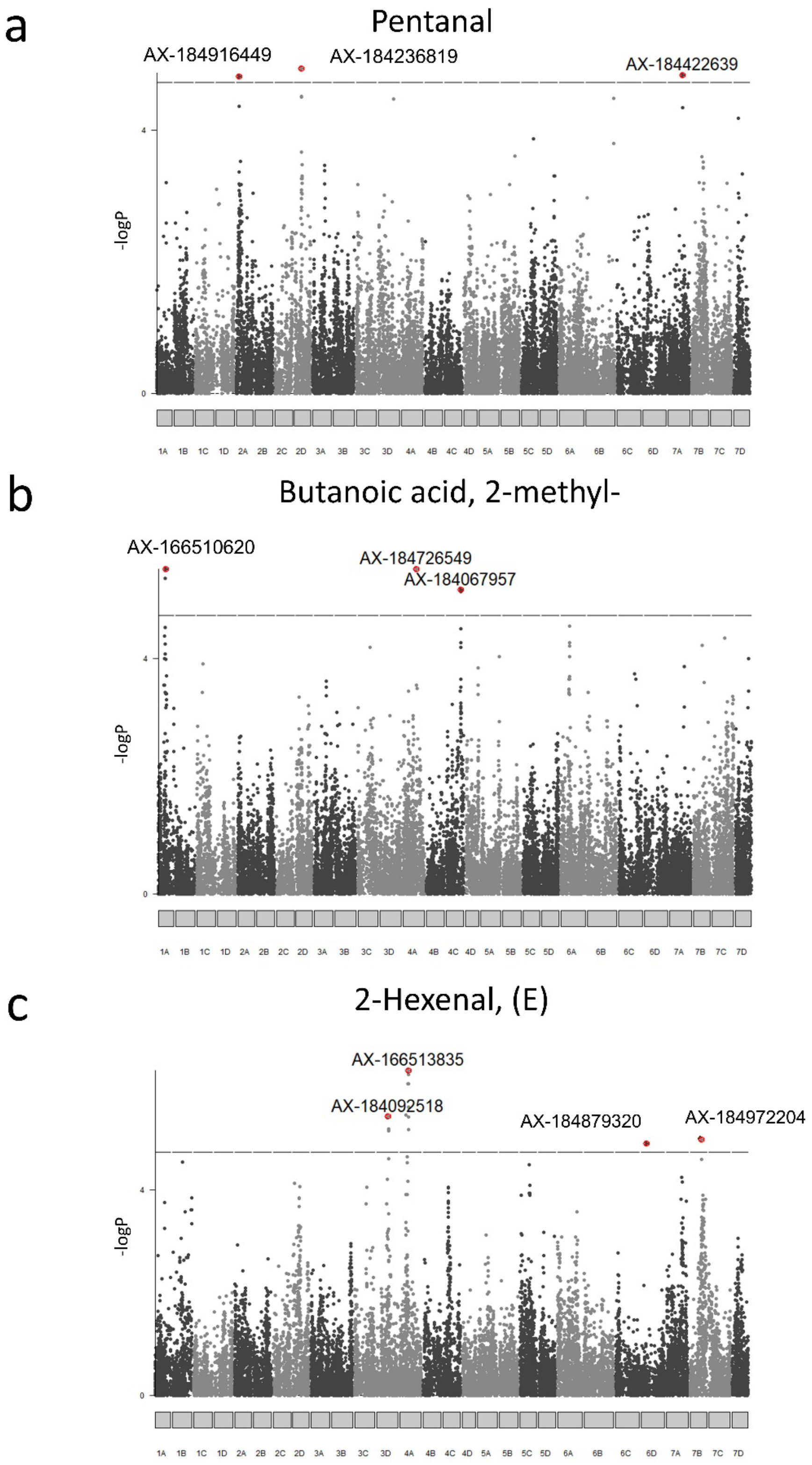

Fig. S14. Manhattan plots of pentanal (a); butanoic acid, 2-methyl- (b); 2-hexenal, (E)-, (c). The lead SNP for each GWAS signal is labeled.

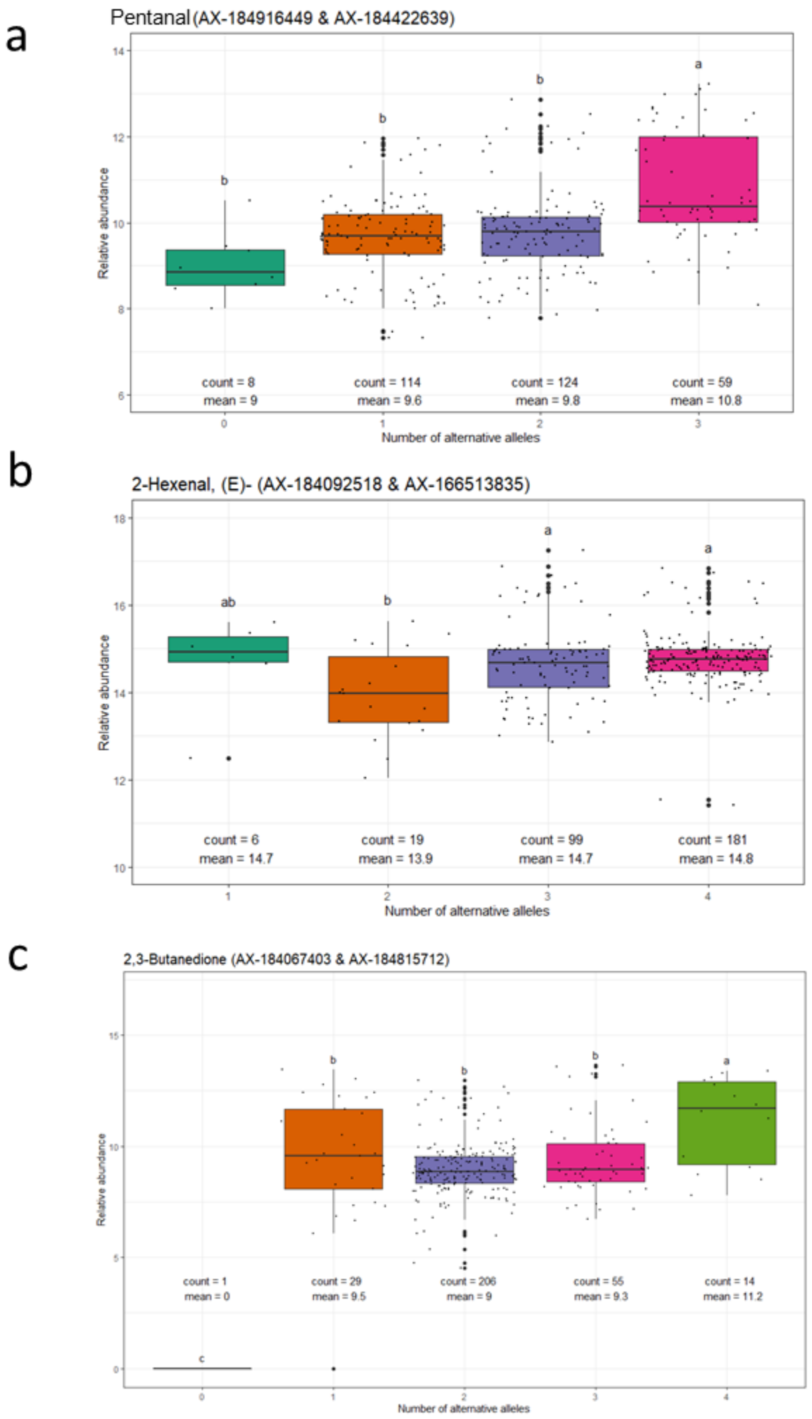

Fig. S15. Box-plots of relative abundance of pentanal (a), 2-hexenal, (E)- (b), and 2,3-butanedione (c) with different dosage of the alternative allele. X axis indicates total number of alternative alleles (functional alleles) for two GWAS signals. The sample counts and means are given below the boxes. Letters above the boxes indicate the significant levels at α=0.05 using Tukey's HSD test. The lead SNPs used for each box plot are included in the plot title. Boxes are delimited by upper and lower quantiles. Two whiskers represent highest/lowest values; dots represent individual values; and horizontal lines represent medians.

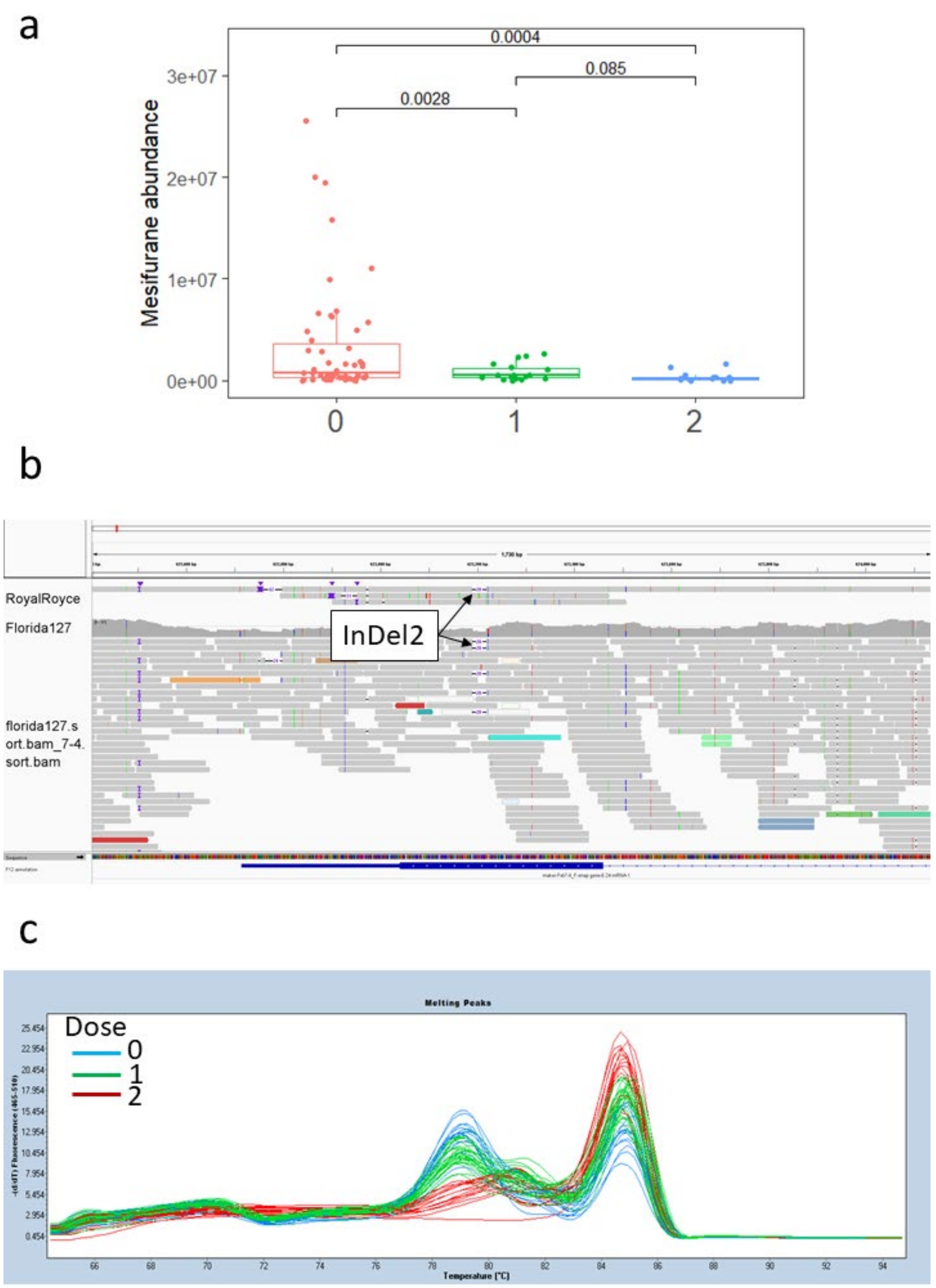

Fig. S16. (a) Comparisons of mesifurane abundances in accessions with different dosages of non-functional haplotypes (haplotype 4, 7 and 8). Boxes are delimited by upper and lower quantiles. Two whiskers represent highest/lowest values; dots represent individual values; and horizontal lines represent medians. (b) Two tracks show the non-functional haplotype 8 from the 'Royal Royce' genome assembly and heterozygous 'Florida127' whole genome sequencing reads mapped to the functional haplotype 2 of *FaOMT* from the F12 assembly. A 28bp InDel (middle of the plot) is present in half of the short reads at the first exon of *FaOMT*. (c) High-resolution melting curves for the InDel2 of *FaOMT* using 38 breeding lines. Clear separation in melting curve patterns is shown between individuals with different allelic dosage.

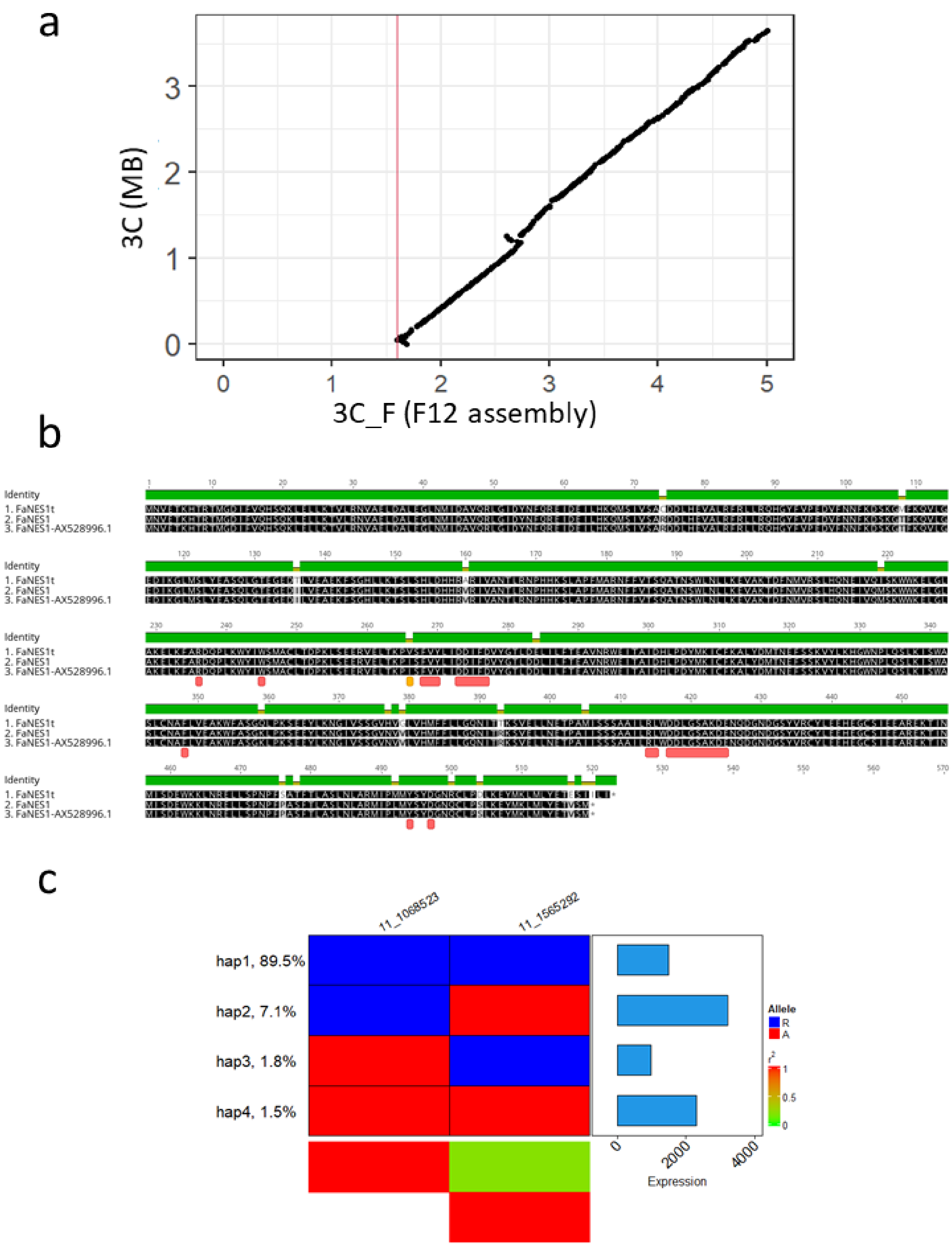

Fig. S17. (a) Alignment between 3C_F from F12 haploid assembly and 3C from the 'Camarosa' reference genome. A 1.61Mb region was missing in the reference genome. (b) Translated nucleotide sequences alignment between *FaNES1t* and *FaNES1* derived from F12 assembly and published sequence from NCBI. Predicted conserved domain were annotated with red bars. The I266V mutation was annotated in orange. (c) Four haplotypes of *FaNES1* identified with two unlinked significant markers (r<0.5). Marker was named according to chrID_position. Left annotation shows haplotype frequency. Right annotation shows haplotype effect in unit of normalized count. The central heatmap shows marker genotype. Blue represents the reference allele; red represents the alternative allele.

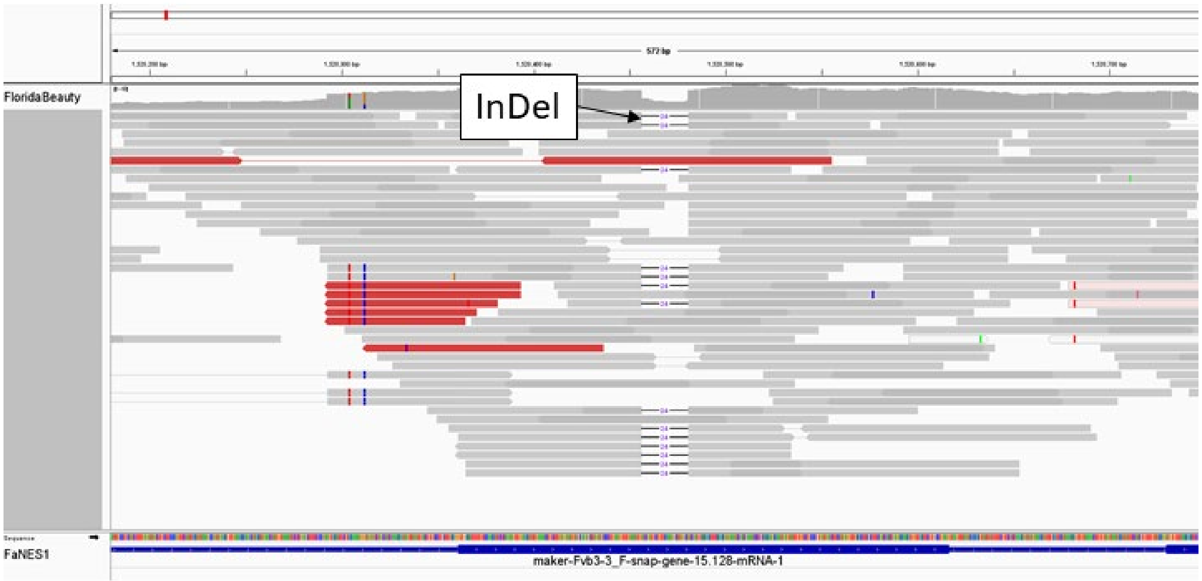

Fig. S18. IGV view of ‘Florida Beauty’ whole genome sequencing reads mapped to *FaNES1* on 3-3_F. A 24bp deletion (middle of the plot) is present in half of the reads at the first exon of *FaNES1*.

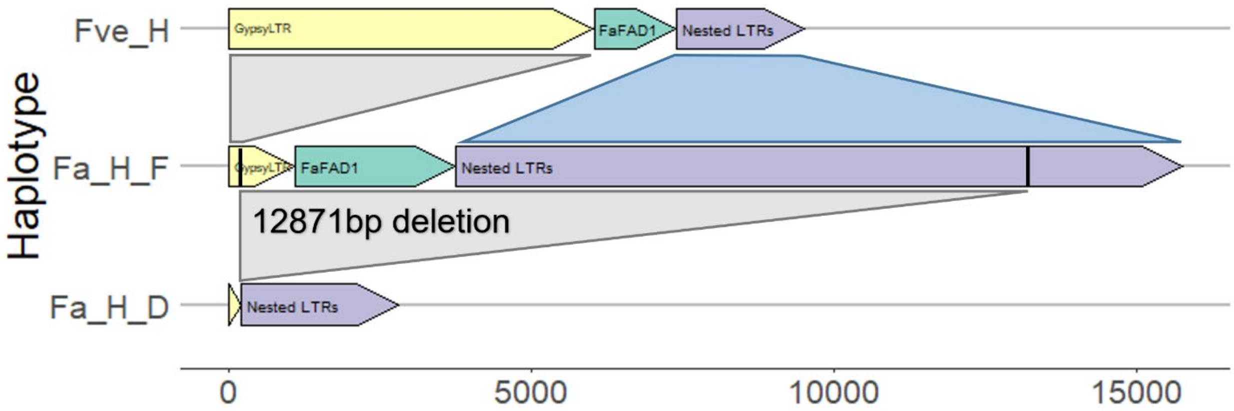

Fig. S19. Schematics of three haplotypes at the FaFAD1 region. Fve_H, Fa_H_F and Fa_H_D represent a haplotype from *Fragria vesca*, and functional and non-functional haplotypes from *Fragaria* × *ananassa*. The deleted region is marked with lines on Fa_H_F haplotype.

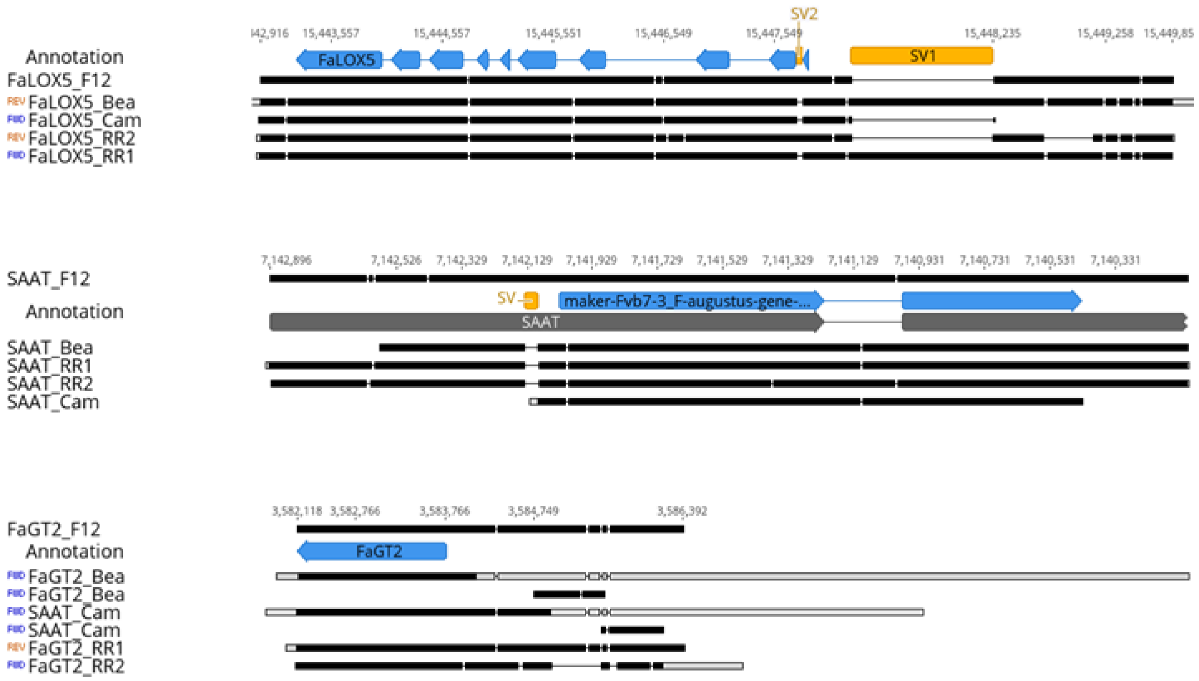

Fig. S20. Gene models, SV locations, genome alignments between haplotypes obtained from F12 haplotype assembly (_F12), Bea haplotype assembly (_Bea), Camarosa reference genome (_Cam) and Royal Royce PHASE1 (_RR1) and PHASE2 (_RR2) for three flavor genes. For *FaGT2*, the insertion at promoter region is too large to be plotted as one continuous line.

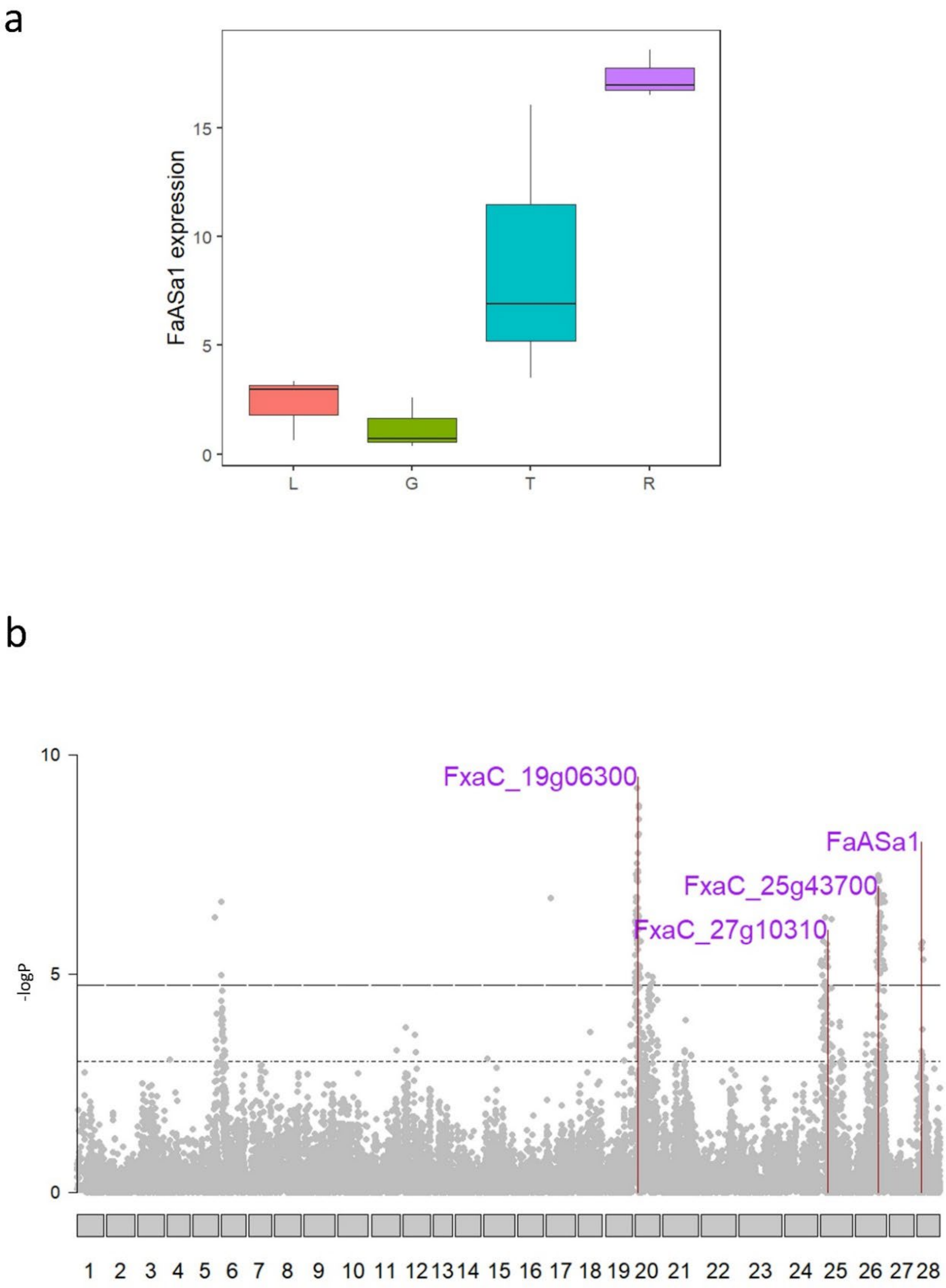

Fig. S21. (a) *FaASa1* expression in leaf (L), green fruit (G), turning fruit (T), red fruit (R). The highest expression was recorded in red fruit. Gene expression was quantified using qPCR. Boxes are delimited by upper and lower quantiles. Two whiskers represent highest/lowest values and horizontal lines represent medians. (b) Four putative candidate genes underlying four GWAS peaks for methyl anthranilate.

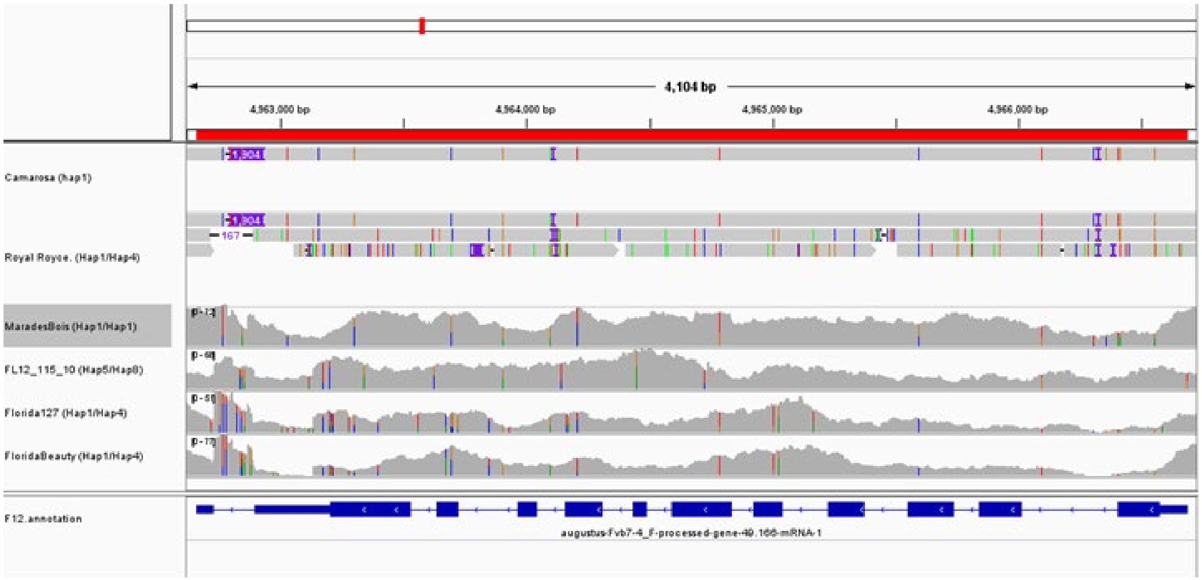

Fig. S22. Upper track shows long-range alignment of *FaASa1* among F12 haplotype, ‘Camarosa’ genome and multiple haplotypes from ‘Royal Royce’. The lower four tracks show short-read mapping of ‘Mara des Bois’, ‘FL12_115_10’,’Florida127’ and ‘Florida Beauty’ to F12 haplotype. The gene model is displayed at the bottom track.

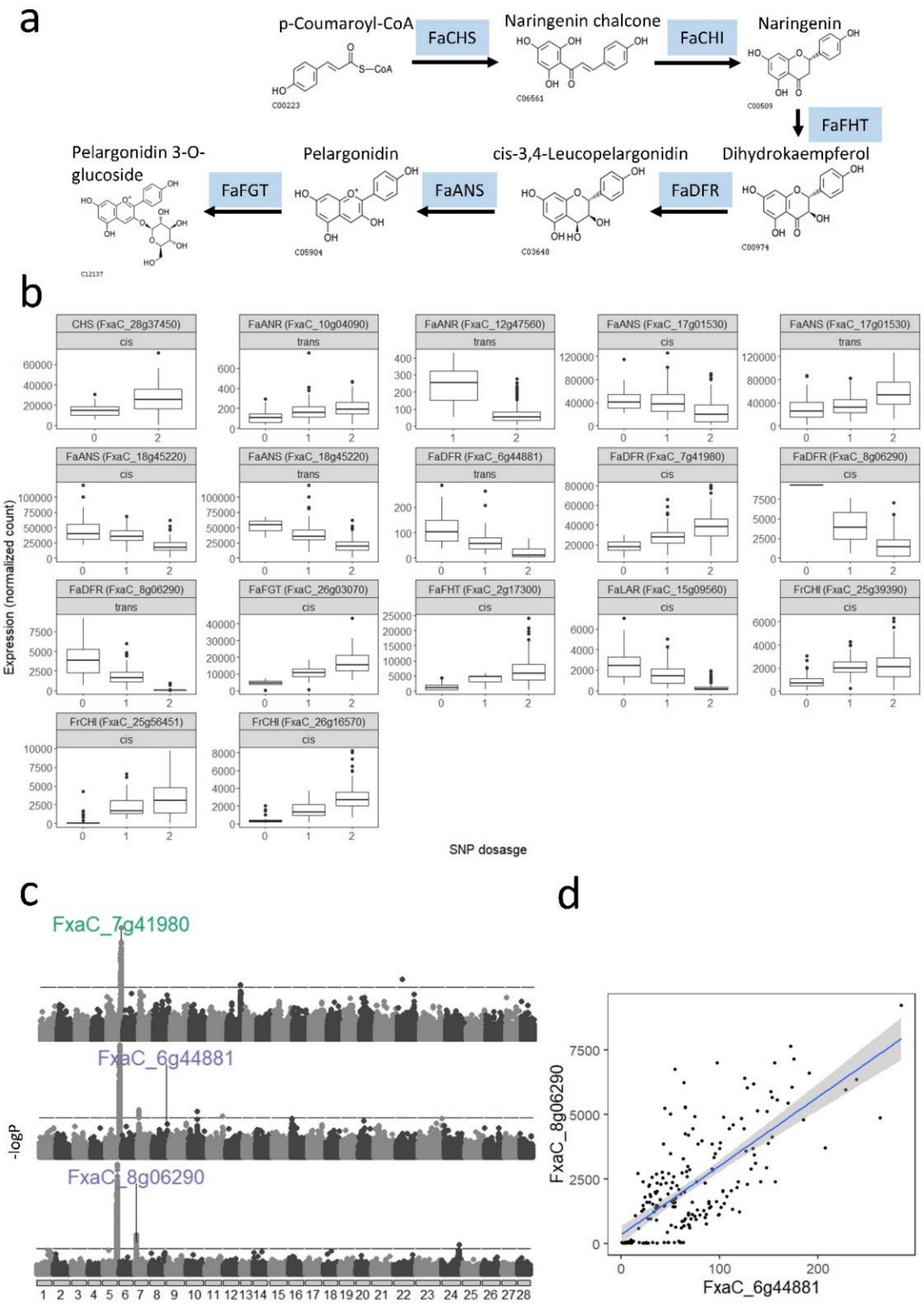

Fig. S23. Genetic variations of genes involved in the anthocyanin pathway. (a) Simplified anthocyanin pathway in strawberry. Pelargonidin 3-O-glucoside is the predominant anthocyanin in strawberry fruit. (b) Expression changes for different dosages of alternative alleles for genes involved in the anthocyanin pathway. Boxes are delimited by upper and lower quantiles. Two whiskers represent highest/lowest values; dots represent individual values; and horizontal lines represent medians. (c) Manhattan plots show significant peaks for 3 homoeologous FaDFRs. Gene locations were labeled for their respective eQTL. (d) Expression of FxaC_8g06290 is plotted with FxaC_6g44881 expression. The shaded line represents linear model built with expressions of two genes.