SUPPLEMENTARY MATERIALS

The genome of woodland strawberry (*Fragaria vesca*)

Vladimir Shulaev[1], Daniel J. Sargent[2], Ross N. Crowhurst[3], Todd C. Mockler[4,5], Otto Folkerts[6], Arthur L. Delcher[7], Pankaj Jaiswal[4], Keithanne Mockaitis[8], Aaron Liston[4], Shrinivasrao P. Mane[9], Paul Burns[10], Thomas M. Davis[11], Janet P. Slovin[12], Nahla Bassil[13], Roger P. Hellens[3], Clive Evans[9], Tim Harkins[14], Chinnappa Kodira[14], Brian Desany[14], Oswald R. Crasta[6], Roderick V. Jensen[15], Andrew C. Allan[16], Todd P. Michael[17], Joao Carlos Setubal[9,18], Jean-Marc Celton[19], D. Jasper G. Rees[19], Kelly P. Williams[9], Sarah H. Holt[20,21], Juan Jairo Ruiz Rojas[20], Mithu Chatterjee[22,23], Bo Liu[11], Herman Silva[24], Lee Meisel[25], Avital Adato[26], Sergei Filichkin[4,5], Michela Troggio[27], Roberto Viola[27], Tia-Lynn Ashman[28], Hao Wang[29], Palitha Dharmawardhana[4], Justin Elser[4], Rajani Raja[4], Henry D. Priest[4,5], Douglas W. Bryant Jr.[4,5], Samuel E. Fox[4,5], Scott A. Givan[4,5], Larry J. Wilhelm[4,5], Sushma Naithani[30], Alan Christoffels[31], David Y. Salama[22], Jade Carter[8], Elena Lopez Girona[2], Anna Zdepski[17], Wenqin Wang[17], Randall A. Kerstetter[17], Wilfried Schwab[32], Schuyler S. Korban[33], Jahn Davik[34], Amparo Monfort[35,36], Beatrice Denoyes-Rothan[37], Pere Arus[35,36], Ron Mittler[1], Barry Flinn[21], Asaph Aharoni[25], Jeffrey L. Bennetzen[29], Steven L. Salzberg[7], Allan W. Dickerman[9], Riccardo Velasco[27], Mark Borodovsky[10,38], Richard E. Veilleux[20], Kevin M. Folta[22, 23*]

[1]Department of Biological Sciences, University of North Texas, Denton, Texas, USA; [2]East Malling Research, Kent, UK; [3]The New Zealand Institute for Plant & Food Research Limited (Plant & Food Research), Mt Albert Research Centre, Auckland, New Zealand; [4]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA; [5]Center for Genome Research and Biocomputing (CGRB), Oregon State University, Corvallis, Oregon, USA; [6]Chromatin Inc., Champaign, Illinois, USA; [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA; [8]The Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana, USA; [9]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, Virginia USA; [10]Joint Georgia Tech and Emory Wallace H. Coulter Department of Biomedical Engineering, Atlanta, GA USA; [11]Department of Biological Sciences, University of New Hampshire, Durham, New Hampshire, USA; [12]USDA/ARS Henry Wallace Beltsville Agricultural Research Center, Beltsville, Maryland, USA; [13]United States Department of Agriculture (USDA), Agricultural Research Service (ARS), National Clonal Germplasm Repository, Corvallis, Oregon, USA; [14]Roche Diagnostics, Roche Applied Science, Indianapolis, Indiana, USA[15]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia USA; [16]School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand; [17]Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, New Jersey, USA; [18]Department of Computer Science, Virginia Tech, Blacksburg, Virginia USA [19]Department of Biotechnology, University of the Western Cape, Bellville, South Africa; [20]Department of Horticulture, Virginia Polytechnic Institute and State University. Blacksburg, Virginia USA; [21]Institute for Sustainable and Renewable Resources, Institute for Advanced Learning and Research, Danville, VA USA; [22]Horticultural Sciences Department, University of Florida, Gainesville, Florida, USA; [23]The Graduate Program for Plant Molecular and Cellular Biology, University of Florida, Gainesville, Florida, USA; [24]Millennium Nucleus in Plant Cell Biotechnology and Centro de Biotecnología y Bioingeniería (CEBBUSS), Facultad de Ingeniería y Tecnología, Universidad San Sebastian, Santiago, Chile; [25]Millennium Nucleus in Plant Cell Biotechnology and Centro de Biotecnología Vegetal, Facultad de Ciencias Biológicas, Universidad Andres Bello, Santiago, Chile; [26]Department of Plant Sciences, Weizmann Institute of Science, Rehovot, Israel; [27]Istituto Agrario San Michele all'Adige (IASMA), Research and Innovation Centre, Foundation Edmund Mach, San Michele all'Adige, Trento, Italy; [28]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; [29]Department of Genetics, University of Georgia, Athens, GA USA; [30]Department of Horticulture, Oregon State University, Corvallis, Oregon, USA; [31]South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville, 7535, South Africa; [32]Biotechnology of Natural Products, Technical University München, Germany; [33]Department of Natural Resources & Environmental Sciences, University of Illinois, Urbana, Illinois, USA; [34]Norwegian Institute for Agricultural and Environmental Research, Genetics and Biotechnology, Kvithamar, Stjordal, Norway; [35]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Cabrils, Barcelona, Spain; [36]Centre de Recerca en Agrigenòmica (CSIC-IRTA-UAB), Cabrils, Barcelona, Spain; [37]Institut National de la Recherche Agronomique (INRA)-Unité de Recherche des Espèces Fruitières (UREF), Villenave d'Ornon, France; [38]School of Computational Science and Engineering, Georgia Tech, Atlanta, GA USA;

*corresponding author

Nature Genetics: doi:10.1038/ng.740

SUPPLEMENTARY MATERIALS

Nature Genetics: doi:10.1038/ng.740

SUPPLEMENTARY NOTE

Gene Prediction

The GeneMark-ES+ [1] software tool was designed for annotation of genomes using transcriptome sequence. This program was used to generate a set of *ab initio* gene predictions for the *F. vesca* genome. GeneMark-ES+ down-selects transcriptome evidence having the highest confidence (in terms of sequencing and DNA mapping errors) from all available raw read or assembled transcriptome data, and combines it with *ab initio* gene predictions to produce a modified maximum likelihood parse. The modified likelihood is determined by applying a fixed user-specified log-likelihood gain on those gene features well supported by the transcriptome evidence. This is distinct from methods that use the mapped transcript sequence to directly estimate HMM parameters. The transcriptome data is down-selected using a DNA sequence mapping and clustering pipeline which uses the BLAT[2] for mapping EST to DNA sequence and TGI clustering tools (TGICL)[3] for clustering the mapped data. Due to the possibility of errors in sequencing and mapping, the mapped and clustered EST sequence are scrutinized to select intron boundaries for which multiple conditions hold, including: 1) more than one EST mapping to a boundary; and 2) canonical intron boundary nucleotides (GT..AG). A library of transposable elements was used in a pipeline developed for detecting and masking repeats. While parameters of the *ab initio* gene finder were estimated by self-training on the unmasked genomic sequence, the hybrid gene predictions were produced on a sequence masked for transposable elements; at this step the algorithm integrated introns mapped from the EST data into the *ab initio* gene models (Supplementary Fig. 4).

Nature Genetics: doi:10.1038/ng.740

Gene Homology Analysis

The Inparanoid algorithm[4] was used to identify orthologous and paralogous genes that arose through duplication events. Clusters were determined using a two-way best pairwise match, and then an algorithm for adding in-paralogs was applied. The peptide sequences used were from twenty-one species, including *Arabidopsis thaliana, Brachypodium distachyon, C. elegans, Chlamydomonas reinhardtii, Danio rerio, E. coli, Fragaria vesca, Glycine max, Homo sapiens sapiens, Zea mays* (maize)*, Mus musculus, Neurospora crassa, Oryza sativa* (rice)*, Physcomitrella patens, Populus trichocarpa* (poplar)*, Saccharomyces cerevisiae and pombe, Selaginella moellendorffii, Sorghum bicolor* (sorghum)*, Synechosystis, and Vitis vinifera.* The peptide sequences were downloaded from Phytozome.net for grape, *Selaginella*, *Physcomitrella, Chlamydomonas, Glycine, and Populus*, Gramene for rice, sorghum, maize and Arabidopsis. The remaining sequences were downloaded from Ensembl.

Transcriptome Analysis

Custom normalized libraries for Roche/454 sequencing were prepared and sequenced for fruit and root RNA[5]. Read lengths averaged 401.4 (fruit pool) and 376.4 (root pool). Library adapter sequences were removed from the reads using estclean (https://sourceforge.net/projects/estclean/).

Over-represented gene ontology categories in fruit and root expressed genes were determined using the EST data derived from the fruit and root pools. A Perl script was used to shred the Roche454 ESTs into simulated 36mer Illumina RNA-seq "reads." These

Nature Genetics: doi:10.1038/ng.740

simulated reads were randomly sampled to generate three simulated replicates for each tissue pool. Perfect match 36-mer reads were mapped to the *F. vesca* cDNA models using HashMatch, and matches were converted to RPKM values. The RPKM values were analyzed using a modified version of BRAT[6] to identify differentially expressed transcripts between the fruit and root samples using a fold change > 2 and a Benjamini and Hochberg FDR-adjusted significance level[7] < 0.01 as cutoffs. Functional enrichment analysis using GO category over-representation was carried out using the network visualization program Cytoscape with GO plugins[8]. For determination of over representation, the Benjamini and Hochberg FDR-adjusted significance level[7] cutoff was 0.05. The color intensity depicted on circles are based on over-representation significance level (yellow = FDR below 0.05) while the radius of each circle indicated the number of genes in each category.

Multiple Genome Alignment

When an anchor genome region matches more than one region in the other genome, only the first match was used to build the table. This means that the table does not represent situations where the other genome has duplications with respect to the anchor genome. This method is reminiscent of the *star alignment* method[9] once used to build multiple alignments of protein sequences. It was necessary to use this method because no software exists that can compare multiple genomes of this size at the same time on computers available to the International Strawberry Genome Consortium. The machine used to compute this alignment was a Sunfire Enterprise 15000 with 72 processors and 288 GB of shared memory. **Genome Sources:** *A. thaliana*: ftp://ftp.arabidopsis.org on 2/9/2010;

Nature Genetics: doi:10.1038/ng.740

G. max: ftp://ftp.jgi-psf.org on 2/9/2010; *L. japonicum*: ftp://ftp.kazusa.or.jp on 10/27/2009; *M. truncatula*: http://www.medicago.org on 10/27/2009; *C. papaya*: ftp://ftp.jgi-psf.org on 2/9/2010; *P. trichocarpa*: ftp://ftp.jgi-psf.org on 2/4/2010; *V. vinifera*: ftp://ftp.jgi-psf.org on 2/9/2010

Analysis of Large Duplications in the *Fragaria vesca* genome

The strawberry genome was compared against itself to identify large repeat regions using MUMmer[10]. The minimum MUM length was set at 30 bp. The MUMmer output was analyzed to obtain a figure for the amount of repetitive sequence present in the assembly regardless of length (but based on the miminum MUM length of 30 bp). We adopted two approaches. In the first, we simply measured the lengths of regions matched by MUMmer excluding whole contig self-matches. The result was 79,886,071 bp. This represents 37.3% of the 214,219,504 bp in assembled contigs.

In the second we used the program Bowtie[11] to map all the Illumina reads to the contigs and counted how many had unique hits and how many had multiple hits. Assuming the Illumina reads are randomly sampled from the genome, the fraction with multiple hits is a good estimate of the repeat content of the genome. Even if a repeat has most of its copies missing from the assembly, the reads from the repeat will still have multiple matches. The only case missed would be repeats with exactly one copy in the assembly, and all other copies missing. We used only the first 40 bp of each read to avoid quality/trimming issues and to get finer granularity. We allowed up to two errors in each match, so match identity was ≥ 95%. There were 36,020,373 total reads, and 27,431,874 had a match to contigs. 18,443,988 (67.2%) had unique matches; 8,987,886 (32.8%) had multiple matches. This indicates that the genome is at least 33% repetitive at the 95%

Nature Genetics: doi:10.1038/ng.740

identity level. This estimate is in good agreement with the estimate obtained by the first method (37.3%), which included matches below the 95% identity level (minimum identity of a match was 45.3%).

Gene Ontology (GO) Annotation

InterPro[12] provided the domain annotations for about 21,000 genes. SignalP[13] provided predicted localization to the mitochondrial or plastid or secretion pathway, besides providing signal peptide cleavage sites. Predotar[14] provided predicted localization to either or both the mitochondrion or plastid. TMHMM[15] provided annotation for the predicted transmembrane domains in the protein sequences. After collecting these annotations, standardized protocols for assigning the GO annotations were adopted. Mapping files provided by the GO consortium were employed to annotate the genes with the three GO categories, namely the Molecular Function, Biological Process and the Cellular Component. The majority of these annotations were inferred by electronic annotation (IEA) evidence codes except the cases where the predicted scores for SignalP and Predotar were reviewed after computational analysis (RCA). Besides these, the GO annotations from *A. thaliana* and *Oryza sativa* (rice) genes as provided by TAIR and Gramene databases were respectively imported to enrich the strawberry annotations by way of gene based orthology suggested by the gene family clustering methods described earlier.

Angiosperm Phylogeny Based on 154 Protein-Coding Genes

Nature Genetics: doi:10.1038/ng.740

A screen for phylogenetic coherence was developed that measured the conflict in sets of taxa sharing an amino-acid at an alignment position, comparing the sets from each gene to the across-genes pool of sets, and then repeating with shuffled versions of the within-gene sets (conflict among randomized sets should be high). The genes were sorted by conflict score ratio (real over randomized) and while nearly half had ratios near one, 240 had ratios below 0.9 and were saved for further work. Newer versions of the original eight genomes plus lotus and soybean were searched by BLAST for members of the 240 low-conflict gene families, saving multiple hits down to 0.8 of the top score. This introduced instances of apparent duplications due to a recent whole-genome duplication[16,17]. Paralogs were identified and eliminated by visually inspecting phylogenies built on each of the 240 gene families. When genes trees exhibited complex patterns obscuring orthology relations, the entire gene family was rejected, leaving 154 orthologous gene families missing members from at most two genomes.

Nature Genetics: doi:10.1038/ng.740

The following sections provide further detail to subject areas of the main text.

Flavor-related biosynthetic pathways

The popularity of strawberry can be attributed to its bright red color, along with its flavors and aromas. These sensory triggers arise from perception of volatile compounds mainly produced by the fatty acid, terpenoid, and phenylpropanoid metabolic pathways. Strawberry flavor in particular consists of diverse volatile compounds comprising more than 300 substances[18,19], with the relative abundance of individual volatiles variable among cultivars and species. Several gene families have been implicated in the production of these volatile components, including the acyltransferases, the terpene synthases, the small molecule O-methyltransferases. Most strawberry flavor compounds are volatile esters, which serve both as attractants of animals and as protectants against pathogens. By linking alcohols to acyl moieties, acyltransferases (AATs) catalyze the last step in the biosynthesis of these volatile esters. *Aat*(*F. vesca*) (AF193790, gene34011) has been shown to be expressed during the final stages of fruit ripening and utilizes a variety of acyl acceptors, from methanol up to 1-decanol[20]. Genomic analysis revealed it to be one of five similar acyltransferases clustered within less than 45 kb on one genomic scaffold (513008, Supplementary Table 11). All five members of this cluster exhibit strong similarity (87%-94%) to the previously described cultivated strawberry genes *Aat1*(*F.* × *ananassa*) and *FcAat*[20,21]. However, the *F. vesca* ortholog of these last two is a sixth acyltransferase (gene33976) located on scaffold 0512999 (Supplementary Table 11; Supplementary Fig. 8a). The terpenoid volatile

Nature Genetics: doi:10.1038/ng.740

profile of cultivated strawberry is dominated by the monoterpene linalool and the sesquiterpene nerolidol, whereas fruit of wild strawberry species emit mainly olefinic monoterpenes and myrteny acetate, which are not found in the cultivated species. *F.* × *ananassa* nerolidol synthases, *FaNES1* and *FaNES2*, have been shown to participate in the biosynthesis of the major terpenoids produced during ripening[22]. Genomic analysis revealed four *F. vesca* genes that are 85%-96% similar to *Nes* (*F.* × *ananassa*) *1* and *Nes2* (*F.* × *ananassa*). Of these, one gene, *Nes3* (*F. vesca*) (gene30669), is located apart on scaffold 513104 and the three others (genes24674, 24676, and 24672) are clustered within 26 kb on scaffold 513012 (Supplementary Table 11). *Pins* (*F. vesca*) (gene15663), a genuine monoterpene synthase expressed in fruit of wild strawberry species, forms multiple monoterpenes, such as the major products a-pinene, b-phellandrene, and b-myrcene from GPP[22]. Our analysis revealed that this gene has three paralogs (genes02063, 22207, 03282) located on scaffolds 512959, 513061 and 513157, respectively (Supplementary Table 11).

An uncommon group of aroma compounds with a 2,5-dimethyl-3(H)-furanone structure dominates the flavor of strawberry fruit. *F.* × *ananassa* quinone oxidoreductase (*Qr ;F.* × *ananassa*) is involved in the biosynthesis of this key aroma compound 4-hydroxy-2,5-dimethyl-3(2H)-furanone (HDMF; Furaneol)[23,24]. Genomic analysis revealed three *F. vesca* gene orthologs, two located within 4 kb in scaffold 513124, (genes28406 and 28407) (Supplementary Table 11; Supplementary Fig. 8c). A fruit ripening induced O-methyltransferase encoded by *Omt* (*F.* × *ananassa*) is responsible for the methylation of HDMF to DMMF and is involved in the biosynthesis of vanillin[25,26]. The methyl ether 2,5-dimethyl-4-methoxy-3(2H)-furanone DMMF was also identified as

Nature Genetics: doi:10.1038/ng.740

a common strawberry aroma component, and is known to occur as race mate with HDMF in various fruit. Genomic analysis revealed that the *Omt* (*F. vesca*) is located on scaffold0513190 and has two close paralogs, genes01858 and 018606, on scaffolds 512956 and 513170, respectively. One of these paralogs, gene018606 (scf513170_3), is a member in a cluster of three O-methyltransferases (Supplementary Table 11; Supplementary Fig. 8d).

Flowering control

The timing of flowering in strawberry is critical to commerce as well as evolution. The control of flowering is regulated by the confluence between environmental and endogenous signals. In strawberry the progression from vegetative to reproductive development is likely gated by the same suite of proteins found in model organisms[27,28]. Examination of the strawberry genome reveals the presence of an intact flowering molecular circuit encompassing genes from the sensing of light (cryptochromes and phytochromes) through the circadian oscillator and recognized output mechanisms that translate the environmental signal into a biological response. Supplementary Table 12 depicts the inventory of strawberry photoperiodic flowering-associated genes and their sequence orthologs in model systems. The results show that the strawberry genome contains a representative set of genes that parallel Arabidopsis. Supplementary Fig. 9 provides a comparison of the intron-exon organization of several genes related to the photoperiod pathway. Strawberry presents an intragenic organization reminiscent of other plant species. The results present a tractable set of nodes to begin to study the flowering response in strawberry, a critical response for the production of a valuable fruit product.

Nature Genetics: doi:10.1038/ng.740

Disease resistance genes

Salicylic acid, jasmonic acid and nitric oxide are molecules that are associated with plant defence. Salicylic acid triggers the expression of *PR* genes and cross talks with the jasmonic acid and nitric oxide signal transduction pathways. Supplementary Tables 13-16 show that many of the PR genes, as well as key enzymes that participate in the biosynthesis of salicylic acid, jasmonic acid and nitric oxide, can be found in the *F. vesca* genome. These findings indicate that there is a conservation of the plant defense response signal transduction pathways in different plant species.

Nature Genetics: doi:10.1038/ng.740

SUPPLEMENTARY TABLES

Supplementary Table 1. Summary of the input sequence data for the assembly of strawberry genome

	Original Reads			Filtered Reads		
Type	**Number of Reads**	**Average Length (bp)**	**Number of Pairs**	**Number of Reads**	**Average Length**	**Number of Pairs**
Unpaired 454 FLX standard	9,023,000	203.0	0	7,727,993	208.5	0
Unpaired 454 FLX Titanium	8,628,801	364.8	0	7,868,311	367.9	0
3kb Paired FLX Titanium	3,589,963	167.9	1,405,560	2,393,795	193.2	591,611
20kb Paired FLX Titanium	2,587,041	201.9	840,159	1,579,668	236.4	229,323
Paired Solexa 76bp	36,927,572	76.0	18,463,786	36,020,373	76.0	1,801,018
Paired 25bp SOLiD	435,077,180	25.0	198,388,200	16,297,230	125.0	648,615

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 2. Strawberry genome assembly statistics

Number of scaffolds	3,263
Number of single contig scaffolds	2,939
Number of scaffold contigs	16,487
Mean contigs per scaffold	5.05
Number of intrascaffold gaps	13,224
Mean intrascaffold gap size	922
Total bases in scaffolds	201,883,090
Mean bases per scaffold	61,878
N50 scaffold bases	1,361,426
Maximum bases in a scaffold	3,924,336

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 3. Estimated haploid genome size of two *F. vesca* accessions, compared to Arabidopsis and *Brachypodium*.

Species	Accession	Average Size (Mb)	SD
Fragaria vesca	H4x4	240.10	6.08
Fragaria vesca	H4 parental	242.55	0.49
Arabidopsis thaliana	Columbia (Col)	146.90	1.98
Brachypodium distachyon	Bd21	301.40	1.13

The haploid nuclear DNA content (genome size) of *F. vesca* accessions was estimated by flow cytometry. Nuclei from H4x4 and H4 parental line were isolated from young leaf or root tissue and DNA content was estimated from the 2C peak. *Arabidopsis thaliana* (Col) and *Brachypodium distachyon* (Bd21) were used as internal controls with haploid genome sizes of 147 and 300 megabases, respectively. Data represent two independent measurements (separate days and DNA samples).

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 4 -- Summary of transposable elements in *Fragaria vesca*

Element Type		Number of intact copies	Number of exemplars	Masking assembly				Masking reads[1]	
				Maximum copy number[2]	Total length (bp)	Coverage (%)		Coverage (%)	
Class I	LTR/*Copia*	173	156	17676	10762743	5.33		4.58	
	LTR/*Gypsy*	114	104	14979	12895589	6.39		5.99	
	LTR/Other	138	115	14578	8493621	4.21	16.37	3.81	14.66
	SINE	456	5	1736	178067	0.09		0.06	
	LINE	17	9	1543	727292	0.36		0.23	
Class II	DNA/CACTA	29	27	7890	5612315	2.78		2.56	
	DNA/*PIF-Harbinger*	13	12	1216	549953	0.27		0.25	
	DNA/*hAT*	29	25	2928	1296712	0.64		0.55	
	DNA/*Helitron*	78	25	981	173731	0.09	6.44	0.07	5.16
	DNA/*TC1-Mariner*	1	1	4	6363	0.00		0.00	
	DNA/*Mutator*	31	22	1073	437384	0.22		0.17	
	DNA/MITE	5169	75	20715	4928797	2.44		1.55	
Other Repeats	-	-	-	50382	2104934	1.04	-	0.92	
Total	-	6248	576	85319	46062567	22.81	-	20.74	-

(1) Reads constituting 1X coverage were randomly selected from all of the geomic shotgun reads that were generated.
(2) Number of homologies to some portion of each element found in the assembly. Actual copy number should be lower because some elements will be separated into more than one assembly (e.g., often at the ends of two assemblies).

Nature Genetics: doi:10.1038/ng.740

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 5. Summary of gene models

	***Ab initio* gene models**, GeneMark-ES	**Hybrid gene models**, GeneMark-ES+
Predicted genes	33,264	34,809
Average gene length (including introns, nt)	2,793	2,792
Average CDS length (nt)	1,177	1,160
Exons	169,012	174,375
Single exon genes	5,654	5,914
Average single exon gene length (nt)	935	927
Average internal exon length (nt)	170	171
Introns	135,748	139,567
Introns per gene (multi-exon genes only)	4.93	4.83
Average intron length (nt)	396	407

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 6. Number of "hybrid" gene models with homology based on Blast comparison to different comparative databases

Expect Threshold[1]	Swissprot Number	Swissprot Percentage	UniRef90[2] Number	UniRef90[2] Percentage	RefSeq[3] Number	RefSeq[3] Percentage	Arabidopsis[4] Number	Arabidopsis[4] Percentage
10	33853	97.3	33838	97.2	34364	98.7	34464	99.0
$1e^{-10}$	17590	50.5	24695	70.9	23483	67.5	22763	65.4
$1e^{-50}$	9730	28.0	17131	49.2	15355	44.1	15089	43.3
$1e^{-100}$	5485	15.8	11647	33.5	9751	28.0	9592	27.6
$1e^{-150}$	3246	9.3	7754	22.3	5962	17.1	5848	16.8
$1e^{-180}$	2315	6.7	5831	16.8	4241	12.2	4208	12.1
Conserved[5]	4138	11.9	8961	25.7	7224	20.8	7191	20.7

[1] Expectation Threshold – BLASTx expectation threshold cut-off used a filter
[2] UniRef90 Release 15.6
[3] ReqSeq Release 36 plant proteins only
[4] The Arabidopsis Information Resource (TAIR) version 9
[5] Conserved – number of genes where the highest scoring segment from blast comparison included greater than 90% of length of both the query and subject sequence

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 7. RNA sequences in assembly v8. tRNAs are listed according to their tRNAscan-SE scores, rRNA fragments according to their lengths, and other RNAs according to their Rfam scores. RNAs were conservatively assigned to organellar locations according to the following code: N, non-organellar; M, mitochondrial; P, plastidial; B, both mitochondrial and plastidial; the conservative criteria explain why several known organellar rRNA sequences were not reassigned (see methods). Numbers of sequences in each category are given in parentheses.

Transfer RNAs (569)

Ala tRNA	N(37):70,70,70,70,70,70,70,70,68,68,68,68,68,68,68,68,68,68, 68,68,68,68,66,66,66,65,65,65,64,57,57,53,38,38,36,34,20
Arg tRNA	N(30):83,82,82,82,82,82,75,75,75,75,74,74,74,74,74,74,74,73, 73,73,72,71,71,70,70,70,70,67,66,59
	P(6):63,63,61,61,61,61
Asn tRNA	N(16):85,85,85,85,85,83,83,83,82,82,82,82,81,78,67,42
	M(3):77,77,64
Asp tRNA	N(22):69,69,69,69,69,69,69,69,69,69,69,69,69,69,69,69,69,69,65,62,59,39
	M(3):61,61,61
Cys tRNA	N(10):78,78,78,78,78,78,78,77,77,76
	P(1):59
Gln tRNA	N(26):80,77,74,74,74,74,74,74,74,73,72,72,72,72,72,72,71,71, 70,70,70,69,68,67,65,41
Glu tRNA	N(29):77,77,77,77,77,77,77,77,77,77,77,77,77,77,77,77,73,73, 73,73,73,73,73,73,73,73,67,64,54
	P(2):50,46
Gly tRNA	N(36):76,76,76,76,76,76,76,76,76,76,72,72,72,72,72,72,72,72, 72,72,72,72,72,72,68,68,68,68,68,68,66,66,65,64,63,45
	P(2):59,59
	M(1):61
His tRNA	N(12):64,64,64,64,64,64,64,62,62,62,62,48
	M(2):52,52
Ile tRNA	N(11):85,85,85,83,79,62,35,30,29,28,27
	M(14):80,80,80,80,80,80,80,80,78,78,78,78,78,71
Leu tRNA	N(38):74,74,74,74,74,71,71,71,71,70,69,69,69,69,69,69,69,69, 69,68,68,68,68,68,68,68,68,68,66,66,66,66,65,64,61,58,57,57,27
	P(3):54,54,54
Lys tRNA	N(25):88,88,88,88,88,88,88,88,88,88,88,84,84,84,84,84,84,84, 84,84,84,83,83,83,58
	M(2):77,77
Met tRNA	N(20):69,69,69,69,69,69,69,68,63,63,63,63,62,62,62,62,61,60,55,29

Nature Genetics: doi:10.1038/ng.740

P(2):64,56
M(3):62,62,56
Phe tRNA N(13):77,73,73,73,73,73,73,73,73,73,73,68,67
Pro tRNA N(25):74,74,74,74,74,74,74,74,74,74,74,72,72,72,72,72,72,72,
72,72,72,72,71,67,44
B(1):57
Pseudo tRNA N(49):43,38,36,31,30,29,28,27,27,27,26,26,26,26,25,25,25,24,
24,24,23,23,23,23,23,23,23,22,22,22,22,22,22,21,21,21,21,21,
21,21,21,21,21,21,21,20,20,20,20
Ser tRNA N(31):84,84,84,84,84,84,84,84,84,84,81,81,81,81,81,81,81,81,
81,80,80,78,78,78,78,78,78,77,76,66,55
P(6):57,56,54,54,53,26
Thr tRNA N(22):84,82,81,81,81,79,79,79,79,78,78,78,77,77,77,77,77,74,
63,27,25,24
Trp tRNA N(13):78,78,78,78,78,76,76,76,76,76,76,76,39
B(3):65,65,65
Tyr tRNA N(12):77,76,75,75,75,75,75,74,74,73,71,49
P(3):54,54,54
Undet tRNA N(5):57,42,28,27,21
M(1):67
Val tRNA N(28):86,86,83,83,83,83,82,80,80,80,80,77,77,77,77,77,77,77,
77,77,77,77,77,75,73,71,25,25
P(1):52

Bacterial ribosomal RNAs (1)
16S rRNA N(1):945

Cytoplasmic ribosomal RNAs (87)
SSU rRNA N(22):2391,1754,1380,1107,1101,939,893,889,829,787,670,
595,485,472,443,410,382,380,379,311,119,90
5_8S rRNA N(11):165,165,164,164,164,164,163,162,150,107,78
26S rRNA N(46):1590,1508,1452,1268,1222,1168,1130,1110,1110,
1101,1093,1083,1081,1070,1069,1069,1065,1060,1047,1043,
1026,1012,982,946,889,750,700,698,650,562,504,455,455,
414,214,189,158,150,116,104,69,61,50,49,49,42
5S rRNA N(8):121,120,120,120,120,120,119,82

Mitochondrial ribosomal RNAs (14)
SSU rRNA N(6):156,102,81,78,70,54
M(1):232

Nature Genetics: doi:10.1038/ng.740

LSU rRNA N(4):1004,222,65,54
M(3):907,204,56

Plastidial ribosomal RNAs (75)

SSU rRNA N(6):232,207,204,175,60,53
P(17):1159,1086,1033,1022,1001,1001,975,879,729,512,481,432,154,107,85,51,51
B(1):41

4.5S rRNA P(7):104,103,103,103,103,67,54

LSU rRNA N(18):317,299,134,133,130,118,117,113,89,81,68,66,66,61,58,58,56,45
P(21):2310,1207,1073,1060,1042,1037,1032,1015,1011,586,503,386,285,189,173,100,77,69,59,43,41

5S rRNA P(5):121,121,121,99,66

Miscellaneous RNAs (49)

Intron_gpI P(3):99,74,72

Intron_gpII N(7):46,45,44,40,36,35,34
P(25):56,53,53,49,49,49,49,49,49,45,45,45,45,45,45,44,44,41,41,41,41,40,38,35,35
M(4):55,55,49,39
B(3):44,44,37

RNase_MRP N(1):72

SRP_euk_arch N(5):184,165,165,164,97

TPP N(1):62

Spliceosomal RNAs (111)

U1 N(19):138,137,133,131,130,128,126,125,124,123,122,117,112,111,109,103,84,66,56

U2 N(29):173,173,171,170,169,167,166,166,165,165,165,165,165,157,153,153,151,117,115,111,101,101,101,96,96,84,84,77,72

U4 N(13):91,90,90,89,89,88,88,82,77,75,70,69,63

U5 N(18):83,82,82,82,81,81,81,79,79,79,77,77,77,71,69,65,61,60

U6 N(26):111,111,111,111,111,109,107,102,101,97,95,93,92,88,86,82,82,80,78,77,73,70,67,65,64,63

U6atac N(4):88,86,83,83

U11 N(1):56

U12 N(1):94

Nature Genetics: doi:10.1038/ng.740

MicroRNAs (76)

mir-156	N(6):66,39,39,39,36,31
mir-160	N(3):65,61,47
mir-166	N(7):80,76,73,73,72,68,52
mir-172	N(4):73,70,68,66
mir-395	N(3):72,62,57
mir-399	N(5):64,50,45,41,39
MIR159	N(4):136,125,111,110
MIR162_2	N(1):85
MIR164	N(4):69,56,50,39
MIR167_1	N(4):72,57,53,50
MIR168	N(1):95
MIR169_2	N(9):58,54,48,48,47,45,36,23,21
MIR169_5	N(6):71,65,64,63,61,54
MIR171_1	N(7):72,68,66,66,65,46,40
MIR390	N(2):81,72
MIR394	N(1):84
MIR396	N(5):77,77,68,55,50
MIR398	N(2):45,41
MIR408	N(1):90
MIR828	N(1):49

Small nucleolar RNAs (168)

U3	N(5):126,124,122,122,73
SNORD14	N(6):90,82,71,70,50,32
SNORD15	N(2):47,33
SNORD24	N(2):56,37
SNORD25	N(4):73,65,59,35
SNORD27	N(1):47
SNORD33	N(4):42,40,39,38
SNORD43	N(3):33,32,29
SNORD46	N(1):62
SNORD96	N(2):68,63
snoJ33	N(3):77,76,46
snoR1	N(1):95
snoR100	N(2):88,81
snoR101	N(2):60,58
snoR103	N(3):84,84,71
snoR104	N(1):95

Nature Genetics: doi:10.1038/ng.740

snoR109	N(1):74
snoR11	N(2):78,76
snoR12	N(2):54,54
snoR14	N(3):72,72,58
snoR16	N(1):47
snoR160	N(1):58
snoR2	N(2):99,97
snoR24	N(3):75,69,64
snoR27	N(2):46,46
snoR28	N(2):82,79
snoR30	N(2):72,69
snoR31	N(1):78
snoR35	N(1):68
snoR41	N(1):67
snoR44_J54	N(2):73,68
snoR60	N(2):67,58
snoR64	N(2):75,75
snoR66	N(1):65
snoR69Y	N(4):76,64,54,52
snoR71	N(21):63,57,55,55,54,53,52,51,51,51,51,51,51,51,49,49,49,48, 48,48,45,36
	M(2):36,34
snoR74	N(2):76,73
snoR77	N(1):99
snoR77Y	N(1):63
snoR83	N(3):77,63,22
snoR86	N(5):110,107,105,105,100
snoR97	N(3):75,75,69
snoR99	N(1):25
snoU30	N(1):52
snoU31b	N(8):72,71,67,66,64,64,61,60
snoU36a	N(1):91
snoZ101	N(1):81
snoZ103	N(6):65,63,61,59,54,53
snoZ107_R87	N(2):88,68
snoZ112	N(1):69
snoZ152	N(1):71
snoZ155	N(1):43
snoZ157	N(2):64,61
snoZ159	N(8):64,61,61,59,56,55,53,52

Nature Genetics: doi:10.1038/ng.740

snoZ161_228	N(1):64
snoZ196	N(3):65,58,43
snoZ199	N(1):41
snoZ223	N(2):63,62
snoZ266	N(8):74,73,65,64,61,47,34,31
snoZ267	N(1):49
snoZ279	N(1):78
snoZ43	N(3):78,75,62

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 8. Assembly of cytoplasmic ribosomal RNA sequences.

rRNA	Length ntd	Sequence Segments	CharsPerPosition min	max	median	Chars	Agreement*
5.8S	164	11	8	11	10	1642	0.973
26S	3351	37	4	16	10	33552	0.987
18S	1802	23	6	14	10	17962	0.975
5S	120	7	6	7	7	802	0.971

* fraction of characters used in assembly that agreed with consensus.

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 9. Summary of the result of protein multiple sequence alignment across plant species. Percentages are based on the total number of rows (49,856) in the master table found at http://staff.vbi.vt.edu/setubal/mapG.html.

Plant genome	number of table cells	percentage (%)
Vitis vinifera	34243	68.7
Populus trichocarpa	34009	68.2
Glycine max	32450	65.1
Carica papaya	29167	58.5
Lotus japonicum	24505	49.2
Arabidopsis thaliana	23324	46.8
Medicago truncatula	22297	44.7

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 10. Functionally characterized genes in *Fragaria* and their homologs in the *F. vesca* genome

Gene	Protein BLAST 04-Dec-2009	Gene BLAST 04-Dec-2009	Protein	Function
structural genes in Fragaria				
FaEG1	gene06191 | 496_aa 1013 0.0	gene06191 | 1491_nt 2710 0.0	endo-β-(1,4)-glucanase	softening, cellulose degradation
FaEG3	gene32087 | 724_aa 1263 0.0	gene32087 | 2175_nt 3646 0.0	endo-β-(1,4)-glucanase	
Facel1	gene06191 | 496_aa 993 0.0	gene06191 | 1491_nt 2656 0.0	endo-β-(1,4)-glucanase	
Facel2	gene32087 | 724_aa 1260 0.0	gene32087 | 2175_nt 3598 0.0	endo-β-(1,4)-glucanase	
FaBG2-1	gene04618 | 335_aa 667 0.0	gene04618 | 1008_nt 1959 0.0	β-1,3-glucanase	pathogenesis-related
FaBG2-2	gene14817 | 334_aa 627 e-180	gene14817 | 1005_nt 1683 0.0	β-1,3-glucanase	
FaBG2-3	gene14817 | 334_aa 657 0.0	gene14817 | 1005_nt 1786 0.0	β-1,3-glucanase	
FaXyl	gene05164 | 893_aa 1519 0.0	gene05164 | 2682_nt 3136 0.0	β-xylosidase	softening, hemicellulose degradation
FaAra1	gene17951 | 765_aa 1133 0.0	gene17951 | 2298_nt 1780 0.0	α-L-arabinofuranosidase	softening, hemicellulose degradation
FaAra2	gene17951 | 765_aa 1165 0.0	gene17951 | 2298_nt 1867 0.0	α-L-arabinofuranosidase	
FaPE1	gene12966 | 514_aa 1049 0.0	gene12966 | 1545_nt 3023 0.0	pectin methyl esterase	softening, pectin degradation
FaPLA	gene17555 | 447_aa 828 0.0	gene17555 | 1344_nt 2190 0.0	pectate lyase	softening, pectin degradation
FaPLB	gene17555 | 447_aa 834 0.0	gene17555 | 1344_nt 2270 0.0	pectate lyase	
FcPL1	gene17555 | 447_aa 837 0.0	gene17555 | 1344_nt 2258 0.0	pectate lyase	
FcPG1	gene21638 | 405_aa 814 0.0	gene21638 | 1218_nt 2327 0.0	polygalacturonase	softening, pectin degradation
FaPG1	gene21638 | 405_aa 819 0.0	gene21638 | 1218_nt 2335 0.0	polygalacturonase	
FaPG2	gene21638 | 405_aa 816 0.0	gene21638 | 1218_nt 2319 0.0	polygalacturonase	
FaChi2-1	gene29267 | 504_aa 499 e-142	gene29267 | 1515_nt 1344 0.0	chitinase	pathogenesis-related
FaChi2-2	gene13071 | 222_aa 409 e-115	gene13071 | 669_nt 1302 0.0	chitinase	
FaβGal1	gene23065 | 890_aa 1763 0.0	gene23065 | 2673_nt 4742 0.0	β-galactosidase	softening
FaβGal2	gene04817 | 846_aa 1678 0.0	gene04817 | 2541_nt 2462 0.0	β-galactosidase	
FaβGal3	gene12626 | 718_aa 1493 0.0	gene12626 | 2157_nt 2655 0.0	β-galactosidase	
FaCCR	gene15845 | 339_aa 678 0.0	gene15845 | 1020_nt 1998 0.0	cinnamoyl CoA reductase	firmness
FaCAD1	gene20700 | 361_aa 724 0.0	gene20700 | 1086_nt 2109 0.0	cinnamyl alcohol dehydrogenase	firmness
FaCAD2	gene20700 | 361_aa 702 0.0	gene20700 | 1086_nt 1974 0.0	cinnamyl alcohol dehydrogenase	
FaGalUR	gene00616 | 336_aa 624 e-179	gene00616 | 1011_nt 1752 0.0	D-galacturonic acid reductase	L-ascorbate biosynthesis
FaGLDH	gene21902 | 582_aa 1142 0.0	gene21902 | 1749_nt 3237 0.0	L-galactono-1,4-lactone	L-ascorbate biosynthesis
SAAT	gene33976 | 455_aa 839 0.0	gene33976 | 1368_nt 2198 0.0	alcohol acyl-CoA transferase	aroma, fruit ester formation
FcAAT	gene33976 | 455_aa 822 0.0	gene33976 | 1368_nt 2137 0.0	alcohol acyl-CoA transferase	
FvAAT	gene34011 | 455_aa 919 0.0	gene34011| 1368_nt 2696 0.0	alcohol acyl-CoA transferase	
FaPINS	gene15663 | 556_aa 1108 0.0	gene15663 | 1671_nt 3297 0.0	pinene synthase	aroma, monoterpene formation
FaNES2	gene30669 | 537_aa 964 0.0	gene30669 | 1614_nt 2541 0.0	S-nerolidol/S-linalool synthase	aroma, mono- and sesquiterpene formation
FaPINH	gene22676 | 520_aa 1030 0.0	gene22676 | 1563_nt 3067 0.0	pinene hydroxylase	aroma, monoterpene formation
Fapdc	gene13476 | 628_aa 1199 0.0	gene13476 | 1887_nt 2346 0.0	pyruvate decarboxylase	aroma
Fapdc1	gene13476 | 628_aa 1194 0.0	gene13476 | 1887_nt 2448 0.0	pyruvate decarboxylase	

Continued on next page…

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 10. (continued)

Gene	Protein BLAST 04-Dec-2009	Gene BLAST 04-Dec-2009	Protein	Function
FaADH	gene30069 | 380_aa 747 0.0	gene30069 | 1143_nt 2060 0.0	alcohol dehydrogenase	aroma
FaOMT	gene12447 | 365_aa 723 0.0	gene12447 | 1098_nt 2010 0.0	O-methyltransferase	aroma, furanone formation
FaQR	gene28406 | 545_aa 609 e-175	gene28406 | 1638_nt 1739 0.0	quinone (enone) oxidoreductase	aroma, furanone formation
FaSDH	gene13340 | 361_aa 712 0.0	gene13340 | 1086_nt 1955 0.0	sorbitol dehydrogenase	sugar metabolism
FagpL1	gene25195 | 668_aa 949 0.0	gene25195 | 2007_nt 1897 0.0	ADP-glucose pyrophosphorylase, large subunit	starch biosynthesis
FagpS	gene24265 | 524_aa 996 0.0	gene24265 | 1575_nt 2581 0.0	ADP-glucose pyrophosphorylase, small subunit	starch biosynthesis
FaCHS	gene26825 | 389_aa 771 0.0	gene26825 | 1170_nt 2089 0.0	chalcone synthase	pigment formation
FaCHS2	gene26825 | 389_aa 775 0.0	gene26825 | 1170_nt 2129 0.0	chalcone synthase	
FaCHS3	gene26825 | 389_aa 775 0.0	gene26825 | 1170_nt 2034 0.0	chalcone synthase	
FaCHS4	gene26826 | 389_aa 777 0.0	gene26826 | 1170_nt 2042 0.0	chalcone synthase	
FaDFR	gene15174 | 333_aa 665 0.0	gene15174 | 1002_nt 1853 0.0	dihydroflavonol 4-reductase	pigment formation
FaDFR1	gene15174 | 333_aa 664 0.0	gene15174 | 1002_nt 1855 0.0	dihydroflavonol 4-reductase	
FaF3H	gene14611 | 364_aa 737 0.0	gene14611 | 1095_nt 2139 0.0	flavanone 3-hydroxylase	pigment formation
FaFLS	gene11126 | 434_aa 372 e-103	gene11126 | 1305_nt 1078 0.0	flavonol synthase	pigment formation
FaANS	gene32347 | 383_aa 749 0.0	gene32347 | 1152_nt 2085 0.0	anthocyanidin synthase	pigment formation
FaLAR	gene03877 | 350_aa 691 0.0	gene03877 | 1053_nt 1937 0.0	leucoanthocyanidin reductase	pigment formation
FaANR	gene24665 | 341_aa 661 0.0	gene24665 | 1026_nt 1929 0.0	anthocyanidin reductase	pigment formation
FaGT1	gene12591 | 465_aa 900 0.0	gene12591 | 1398_nt 2512 0.0	anthocyanidin glucosyltransferase	pigment formation
FaGT2	gene26265 | 555_aa 1103 0.0	gene26265 | 1668_nt 3148 0.0	UDP-glucose:cinnamate glucosyltransferase	phenlypropanoid metabolism
FaGT6	gene24227 | 479_aa 951 0.0	gene24227 | 1440_nt 2791 0.0	flavonol glucosyltransferases	flavonoid metabolism
FaGT7	gene26344 | 469_aa 938 0.0	gene26344 | 1410_nt 2771 0.0	flavonol glucosyltransferases	flavonoid metabolism
FaCDPK1	gene25220 | 549_aa 1050 0.0	gene25220 | 1650_nt 2121 0.0	calcium-dependent protein kinase	fruit development
FaACS	gene31839 | 487_aa 547 e-156	gene31839 | 1464_nt 1691 0.0	1-aminocyclopropane-1-carboxylic acid synthase	ethylene biosynthesis
FaACO	gene01202 | 363_aa 624 e-179	gene01202 | 1092_nt 1742 0.0	1-aminocyclopropane-1-carboxylic acid oxidase	ethylene biosynthesis
FvCGS	gene04420 | 545_aa 1073 0.0	gene04420 | 1638_nt 3215 0.0	cystathionine γ-synthase	methionine biosynthesis
Fapmsr	gene10008 | 190_aa 358 1e-099	gene10008 | 573_nt 815 0.0	methionine sulfoxide reductase	repair of proteins and peptides
FaCCD1	gene26820 | 682_aa 1006 0.0	gene26820 | 2049_nt 1794 0.0	carotenoid cleavage dioxygenase	lutein degradation
FaAPX	gene25391 | 676_aa 508 e-144	gene25391 | 2031_nt 1358 0.0	cyctosolic ascorbate peroxidase	gluthathione-ascorbate cycle
FacpFBP	gene27415 | 1203_aa 714 0.0	gene27415 | 3612_nt 2004 0.0	chloroplastic fructose-1,6-diphosphatase	photosynthesis
FaPLD	gene15700 | 854_aa 1630 0.0	gene15700 | 2565_nt 4460 0.0	phospholipase D alpha	membrane deterioration
FaSTK	gene16583 | 878_aa 380 e-106	gene16583 | 2637_nt 1003 0.0	serine-threonine kinases	protein modification
other protein coding genes				
FaCyf1	gene21476 | 207_aa 417 e-117	gene21476 | 624_nt 1197 0.0	phytocystatin	cystein protease inhibitor, antifungal
FaEin1	gene32532 | 764_aa 1504 0.0	gene32532 | 2295_nt 4488 0.0	ethylene insensitive 2	ethylene receptor
FaErs1	gene11090 | 643_aa 1233 0.0	gene11090 | 1932_nt 2881 0.0	ethylene resistant, ethylene response sensor	ethylene receptor
FaEtr1	gene21106 | 741_aa 1441 0.0	gene21106 | 2226_nt 4136 0.0	ethylene resistant, ethylene response sensor	ethylene receptor
FaTCTP	gene06814 | 168_aa 327 2e-090	gene06814 | 507_nt 900 0.0	translationally controlled tumor protein	fruit ripening
FaPGIP	gene00522 | 332_aa 601 e-172	gene00522 | 999_nt 1709 0.0	polygalacturonase-inhibiting protein	defense
FaWRKY1	gene07210 | 190_aa 386 e-108	gene07210 | 573_nt 1072 0.0	transcription factor	regulator of defense

Continued on next page

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 10. (continued)

Gene	Protein BLAST 04-Dec-2009	Gene BLAST 04-Dec-2009	Protein	Function
FaOLP	gene32422 | 226_aa 484 e-137	gene32422 | 681_nt 1255 0.0	osmotin-like protein	pathogenesis-related, stress
FaOLP2	gene32421 | 229_aa 489 e-139	gene32421 | 690_nt 1312 0.0	osmotin-like protein	
Faltp1	gene13870 | 117_aa 232 3e-062	gene13870 | 354_nt 694 0.0	non-specfic lipid transfer protein	stress
Faltp2	gene13870 | 117_aa 215 4e-057	gene13870 | 354_nt 551 e-156	non-specfic lipid transfer protein	
Faltp3	gene13870 | 117_aa 224 6e-060	gene13870 | 354_nt 622 e-178	non-specfic lipid transfer protein	
Faltp	gene13870 | 117_aa 214 9e-057	gene13870 | 354_nt 543 e-154	non-specfic lipid transfer protein	
FaRP7	gene08914 | 317_aa 483 e-137	gene08914 | 954_nt 1233 0.0	tonoplast intrinsic protein	resistance
FaCBF1	gene13329 | 229_nt 310 4e-085	gene13280 | 690_nt 125 4e-028	cold-induced transcription factor	cold acclimation response
STAG1	gene24852 | 250_aa 461 e-130	gene24852 | 753_nt 1291 0.0	MADS box, AGAMOUS homolog	vegetative, floral, fruit development
FvLFY	gene30750 | 377_aa 336 4e-093	gene30750 | 1134_nt 946 0.0	LEAFY	floral identity, floral integrator
Fanjjs4	gene20883 | 156_aa 309 5e-085	gene20883 | 471_nt 854 0.0	low molecular weight heat shock protein	seed maturation, fruit ripening
FaExp1	gene02221 | 259_aa 312 7e-086	gene0221 | 780_nt 222 2e-057	expansin cell wall proteins	softening, cell wall disassembly
FaExp2	gene21343 | 253_aa 541 e-154	gene21343 | 762_nt 1495 0.0	expansin cell wall proteins	
FaExp3	gene04435 | 270_aa 336 4e-093	gene04435 | 813_nt 852 0.0	expansin cell wall proteins	
FaExp4	gene04724 | 249_aa 531 e-151	gene04724 | 750_nt 1471 0.0	expansin cell wall proteins	
FaExp5	gene26030 | 250_aa 356 4e-099	gene26030 | 753_nt 898 0.0	expansin cell wall proteins	
FaABP1	gene27729 | 193_aa 384 e-107	gene27729 | 582_nt 1045 0.0	auxin-binding protein	auxin perception
FaPIP1	gene20927 | 290_aa 402 e-113	gene20927 | 873_nt 1181 0.0	plasma membrane intrinsic protein	aquaporin, water channel
Fahyprp	gene09178 | 166_aa 315 7e-087	gene09178 | 501_nt 678 0.0	hybrid proline-rich protein	polyphenol anchoring
FaMYB1	gene09407 | 188_aa 369 e-103	gene09407 | 567_nt 993 0.0	transcription factor	regulation of pigment biosynthesis
Fraa1A	gene07080 | 270_aa 330 2e-091	gene07080 | 813_nt 942 0.0	pathogenesis-related protein, allergen	flavonoid biosynthesis
Fraa2	gene07065 | 160_aa 319 5e-088	gene07065 | 483_nt 878 0.0	pathogenesis-related protein, allergen	
Fraa3	gene07082 | 159_aa 319 5e-088	gene07082 | 480_nt 944 0.0	pathogenesis-related protein, allergen	
FaMet	gene13037 | 1569_aa 3057 0.0	gene13037 | 4710_nt 8592 0.0	DNA methyltransferases	methylation of cytosine residues in DNA
FaZIP	gene24223 | 353_aa 670 0.0	gene24223 | 1062_nt 1804 0.0	Zn- and Fe-regulated transporter	mineral uptake
FaGAST	gene08518 | 106_aa 193 2e-050	gene08518 | 321_nt 581 e-165	small protein with 12 cysteine residues	arresting cell elongation

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 11. Fruit-flavor associated genes in *F. vesca* and corresponding nomenclature.

Gene	GenBank accession	Scaffold	Hybrid Gene Model ID	Ab Initio Gene Model ID	Note
Acyltransferases					
Aat1	AF193790	scf0513008	gene34011	not available	
Aat2		scf0513008	gene34010	not available	Adjacent to 34011
Aat3		scf0513008	gene23453	not available	Adjacent to 34011
Aat4		scf0513008	gene34009	not available	Adjacent to 34011
FvAat5T		scf0513008	gene34008	not available	Adjacent to 34011
FvAat6		scf0512999	gene33976	not available	
Terpene Synthases					
Nes3		scf0513104	gene30669	gene30580	
Nes1		scf0513012	gene24674	gene24601	
Nes4		scf0513012	gene24676*	gene24603	Adjacent to gene24674 *Lacking the beginning of the gene
Nes5		scf0513012	gene24672	gene24599	Adjacent to gene24674
Pins1	AJ001452	scf0513196	gene15663	gene15637	
Pins2		scf0512959	gene02063	gene2054	
Pins3		scf0513061	gene22207	gene22147	
Pins4		scf0513157	gene03282	gene3266	
Quinine oxidoreductases					
Qr1	AJ001445	scf0513124	gene28406*	gene28335	*merges of two different genes, the second of which (aa209-545) is *FvQR*
Qr2		scf0513124	gene28407	gene28336	Adjacent to gene28406
Qr3		scf0512933	gene00649	gene00653	
O-methyltransferases					
Omt1		scf13190	gene12447		
Omt2		scf13170	gene18606		
Omt3		scf13170	gene18605*		* the beginning of this prediction
Omt4		scf13170	gene18605*		* the end of this prediction

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 12. An inventory of *F. vesca* orthologs to photoperiodic flowering pathway genes.

GENE NAME	ARABIDOPSIS ORTHOLOG	EXPERIMENTALLY DETERMINED FUNCTION	STRAWBERRY ORTHOLOG
PhyA	AT1G09570	Photoreceptor	gene22948
PhyB	AT2G18790	Photoreceptor	gene05117
Phy C	AT5G35840	Photoreceptor	gene15519
Phy D	AT4G16250	Photoreceptor	No gene model
Phy E	AT4G18130	Photoreceptor	gene11383, gene24884
Cry 1	AT4G08920	Photoreceptor	gene30027
Cry 2	AT1G04400	Photoreceptor	gene11459
Lhy1	AT1G01060	DNA binding / transcription factor	gene18602
Cca1	AT2G46830	transcription factor	No gene model
Elf3	AT2G25930	Nuclear protein	gene02656
Toc1	AT5G61380	transcription regulator/ two-component response regulator	gene26055
Esd4	AT4G15880	cysteine-type peptidase activity	gene06809
Fkf1	AT1G68050	signal transducer/ two-component sensor/ ubiquitin-protein ligase	gene06247
Lkp2	AT2G18915	protein binding / ubiquitin-protein ligase	No gene model
Ztl	AT5G57360	ubiquitin-protein ligase	gene17000, gene16999
Co	AT5G15840	transcription factor	gene04172
Gi	AT1G22770		gene27581
Ft	AT1G65480	phosphatidylethanolamine binding / protein binding |	gene04680
Soc1 / Agl20	AT2G45660	transcription factor	gene19425
Tfl1	AT5G03840	phosphatidylethanolamine binding	gene21992

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 13. Jasmonic acid metabolic genes identified in *F. vesca* genome

				***F. vesca* gene prediction**	
Gene name	**query species**	**gi**	**database**	**TAU prediction**	**genemark**
Lox	*Fragaria x ananassa*	33235470	emb|AJ578035.2	scf0513162 scf0513061 scf0513155 scf0513061	gene22255 gene24064 gene19940 gene24063
Aos	*Prunus persica*	61844840	emb|AJ633680.2	scf0513194 scf0513097 scf0513097 scf0513097	gene15023 gene08610 gene08611 gene08676
Aoc	*Prunus persica*	89479473	gb|DY635267.2	scf0513158 scf0512954 scf0513152	gene18678 gene30035 gene04120
Opr7	*Zea mays*	63021730	gb|AY921644.2	scf0513192 scf0513192	gene16287 gene16282

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 14. Salicylic acid associated genes identified in *F. vesca* genome

				***F. vesca* gene prediction**	
Gene name	**query species**	**gi**	**database**	**TAU prediction**	**genemark**
Pal	*Fragaria x ananassa*	157041078	dbj|AB360394.2	scf0513149 scf0513008	gene03339.1 gene31975.1
Pbs3	*Arabidopsis thaliana*	145357961	ref|NM_121335.4	scf0513144 scf0513004 scf0513157 scf0513068 scf0512991 scf0513104	gene04620.1 gene32079.1 gene20644.1 gene24952.1 gene31443.1 gene06248.1
Eps1	*Zea mays*	1524382	emb|X63374.2	scf0513153	gene03709.1

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 15. Nitric oxide related genes identified in *F. vesca* genome

				***F. vesca* gene prediction**	
Gene name	**query species**	**gi**	**database**	**TAU prediction**	**genemark**
Nos	*Arabidopsis thaliana*	108951288	gb|DQ539437.2	scf0513044	gene25402.1
Sod	*Prunus persica*	6066607	emb|AJ238316.3	scf0513044	gene25277.1
Cat1	*Prunus persica*	32526565	emb|AJ496418.2	scf0513145	gene05469.1

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 16. Pathogen related genes identified in *F. vesca* genome

				***F. vesca* gene prediction**	
Gene name	**query species**	**gi**	**database**	**TAU prediction**	**genemark**
Npr1	*Prunus persica*	76261960	gb|DQ149935.2	scf0513173 scf0513088 scf0513088 scf0513192	gene13847 gene25989 gene25990 gene16361
Pr2	*Fragaria x ananassa*	62362437	gb|AY989819.2	scf0513194 scf0513158 scf0513194 scf0513160 scf0513154 scf0513073	gene15465 gene19106 gene15463 gene22435 gene20141 gene27009
Pr3	*Arabidopsis thaliana* *Fragaria x ananassa*	156182214 11528438	gb|EF576873.2 gb|AF320111.2	scf0513094 scf0513030 scf0513164 scf0513192 scf0513192 scf0513192 scf0513164 scf0513192 scf0513030 scf0513094	gene08047 gene31135 gene21457 gene14592 gene14593 gene14592 gene21457 gene14593 gene31135 gene08047
Pr4	*Prunus persica*	19879969	gb|AF362989.2	scf0513104 scf0513133 scf0513168	gene06067 gene00573 gene13129
Pr5	*Malus domestica*	83853952	gb|DQ318213.2	scf0513177 scf0512983	gene12514 gene28676
Pr6	*Arabidopsis thaliana*	17473692	gb|AY065127.2	scf0513069 scf0513069	gene24622 gene24629
Pr8	*Malus domestica*	83853954	gb|DQ318214.2	scf0513196 scf0513190 scf0513190 scf0513170 scf0513190 scf0513170 scf0513170	gene16507 gene15894 gene15723 gene14163 gene15882 gene12948 gene12949
Pr10	*Prunus persica*	159794682	gb|EU117120.2	scf0513159	gene19834

Nature Genetics: doi:10.1038/ng.740

Supplementary Table 17. Summary of number of family members for each of the major families of transcription factors with sequence similarity less than e^{-20} of plants with whole genome sequence.

	Fragaria vesca (*Fv*)	*Vitis vinifera* (*Vv*)	*Arabidopsis thaliana* (*At*)	*Medicago truncatula* (*Mt*)	*Lotus japonicus* (*Lj*)	*Glycine max* (*Gm*)	*Ricinus communis* (*Rc*)	*Populus trichocarpa* (*Pt*)	*Oryza sativa* (Os)	*Zea mays* (*Zm*)	*Sorghum bicolor* (*Sb*)
ABI3VP1	20	21	71	85	34	78	35	103	70	97	58
AP2-EREBP	108	127	146	109	159	381	114	211	199	338	161
ARID	10	9	10	12	9	22	9	12	8	19	5
ARF, AUX/IAA	35	43	51	24	36	129	37	70	99	191	59
bHLH	226	97	172	71	84	393	96	145	174	309	151
bZIP	37	45	78	44	38	176	49	83	133	245	103
C2C2-CO-like	7	24	34	15	21	72	22	38	42	66	32
C2C2-Dof	15	26	36	21	22	82	23	42	37	54	29
C2C2-GATA	7	20	29	29	16	62	19	38	35	63	33
C2C2-YABBY	5	7	6	6	4	18	6	13	15	41	8
C2H2	68	57	173	52	56	395	59	73	118	162	88
C3H	154	121	66	141	105	144	31	99	1448	347	206
CAMTA	6	4	2	6	4	15	5	7	7	14	7
DDT	6	5	6	2	4	15	3	3	8	18	6
E2F-DP	6	8	8	6	7	14	6	10	12	29	10
EIL	5	4	3	7	6	13	4	6	11	13	7
FHA	22	13	5	9	10	33	15	19	22	28	14
GRAS	44	45	33	36	43	130	45	98	64	105	74
GRF	9	3	2	14	15	6	8	7	51	12	29
HB	84	69	112	46	55	318	59	104	121	211	74
zf-HD	4	0	1	2	1	7	5	5	2	4	1
HMG	6	7	12	2	9	31	9	12	17	33	14
HSF	14	19	24	16	18	11	19	31	42	58	25
Jumonji	18	17	21	13	12	77	15	19	17	46	21
LFY	3	1	1	1	2	4	1	1	1	4	1
LIM	8	14	13	9	12	42	9	20	16	50	9
MADS	31	60	109	60	83	212	39	111	95	147	82
MBF1	2	0	3	1	3	4	11	3	1	7	0
MYB[a]	187	242	303	171	191	791	189	378	298	564	262
NAC	150	83	114	64	101	208	94	179	162	252	126
PHD	107	58	55	45	47	222	58	90	87	202	74

Nature Genetics: doi:10.1038/ng.740

PLATZ	7	10	3	10	9	25	11	20	22	22	17
RWP-RK	10	8	14	5	12	23	10	18	14	37	13
SBP	25	18	16	11	15	48	15	29	29	70	18
SNF2[b]	52	28	33	17	23	69	25	41	42	85	34
TAZ	3	4	10	5	2	1	5	7	11	17	5
TCP	16	19	6	11	21	65	21	34	24	59	26
TUB	9	8	2	12	8	24	7	11	26	43	20
WRKY	90	59	73	59	68	197	57	104	126	204	92
TOTAL	1616	1403	1856	1249	1365	4557	1245	2294	3706	4266	1994

[a]MYB includes genes with sequence similarity to both MYB and MYB-related TFs. [b]SNF2 includes genes with sequence similarity to SWI/SNF-BAS and SW-like transcription factors.

Nature Genetics: doi:10.1038/ng.740

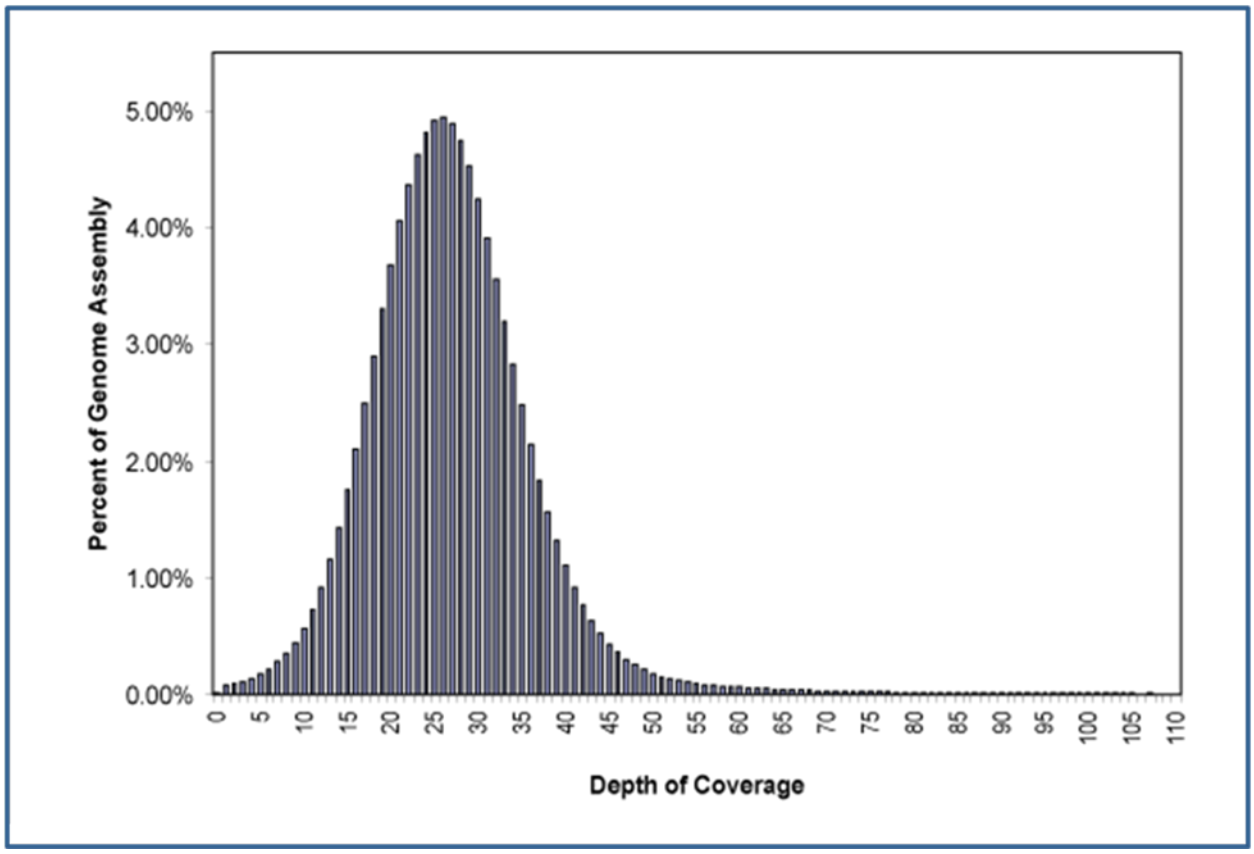
Supplementary Table 18. Genomic information relevant to key transcription factors.

			F. vesca gene prediction		
gene	**Gi**	**database**	**TAU prediction**	**genemark**	**genemark+**
Myb1 (*F. × ananassa*)	15082209	gb|AF401220.1	scf0513135.2250.1 scf0513135.2250.2	gene09374	gene09407
Myb10 (*F. × ananassa*) *Myb10* (*F. vesca*)	161878909 161878911	gb|EU155162.1 gb|EU155163.1	scf0513095.997.1 scf0513095.997.2 scf0513095.997.3 scf0513095.997.4 scf0513095.997.5 scf0513095.997.6 scf0513095.997.7 scf0513095.997.8	gene31324	gene31413
Mybpa1 (*V. vinifera*)	130369072	emb|AM259485.1	scf0513170.1998.1 na	gene18657 gene25912	gene18691 gene25982

Nature Genetics: doi:10.1038/ng.740

SUPPLEMENTARY FIGURES

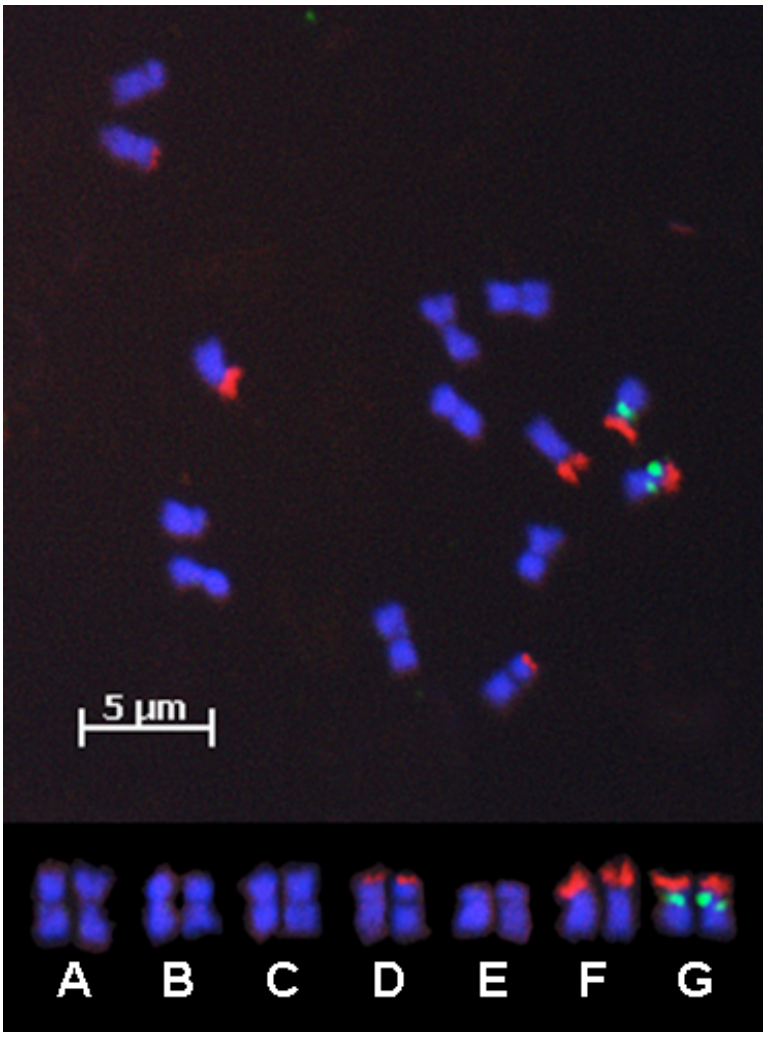
Supplementary Figure 1



Supplementary Figure 1. Illumina re-sequencing of the *F.vesca* V8 assembly. The *F. vesca* genome was re-sequenced using the Illumina platform and single-end 36mer reads were mapped to the genome using ELAND.

Nature Genetics: doi:10.1038/ng.740

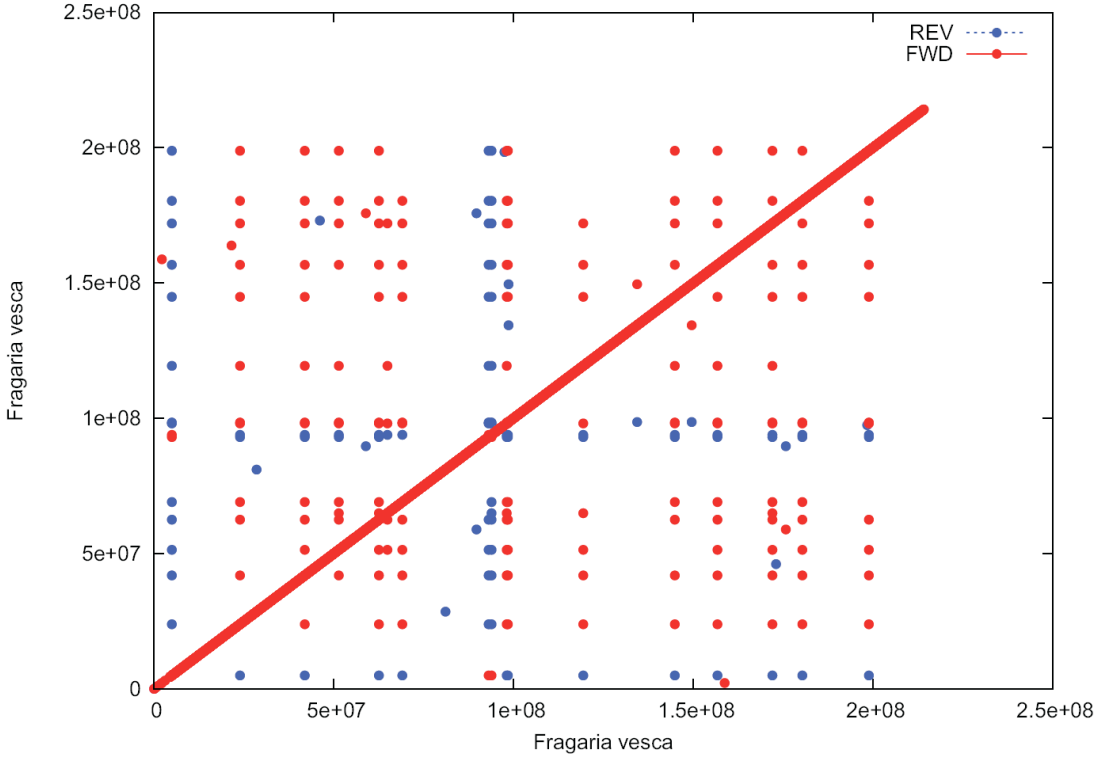
Supplementary Figure 2.



Supplementary Figure 2. A molecular karyotype of diploid strawberry chromosomes. Mitotic (root tip) chromosomes of 'Hawaii 4' probed with differentially labeled 25S (red) and 5S (bright green) rDNA hybridization probes. In this molecular karyotype, the chromosome pairs have been sequentially numbered A through G according to decreasing size (length). Chromosomes D, F and G harbor 25S rDNA loci, while Chromosome G also harbors the 5S locus.

Nature Genetics: doi:10.1038/ng.740

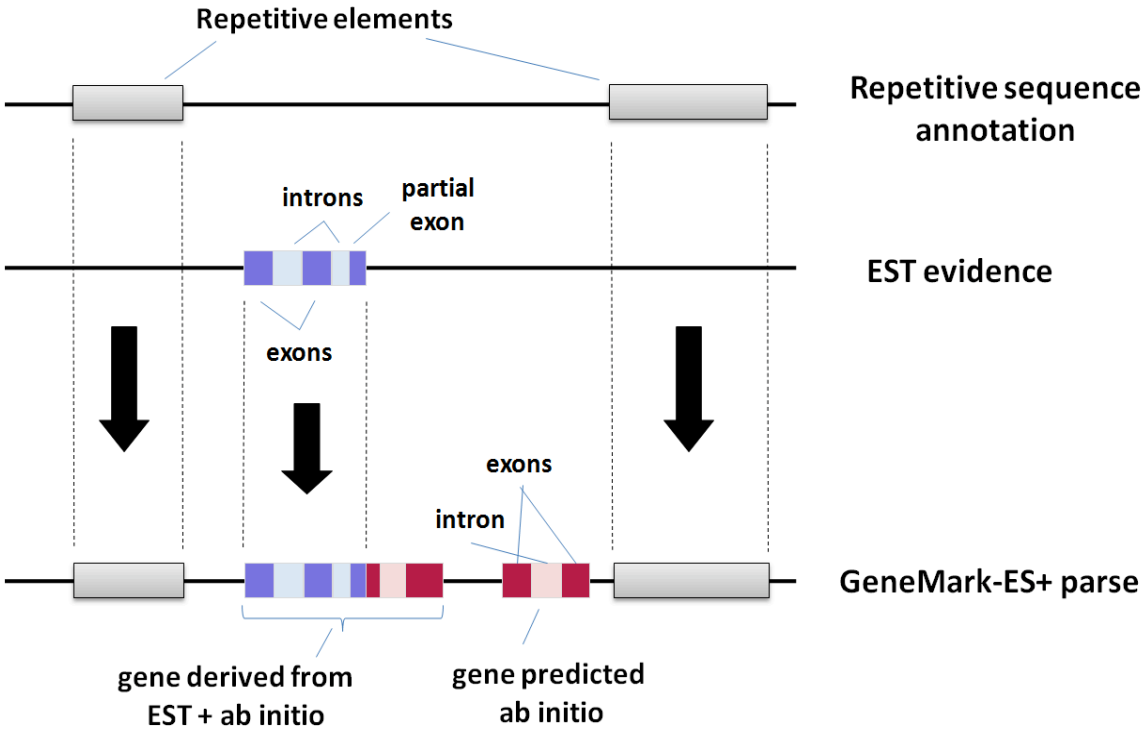
Supplementary Figure 3



Supplementary Figure 3. ***F. vesca* – *F. vesca* genome comparison to identify large repeat regions.** This plot shows the result of an alignment of the *Fragaria vesca* concatenated contigs against themselves, as given by the program MUMmer (option nucmer, nucleotide sequence comparison). The contigs are in arbitrary order. In red are shown direct (or forward) sequence matches; in blue are shown reverse sequence matches. Only matches that are 10,000 bp or longer are shown; each dot outside the diagonal corresponds to one such match. The largest match found was 14,721 bp long. The scales in the *x* and *y* axes are in base pairs.

Nature Genetics: doi:10.1038/ng.740

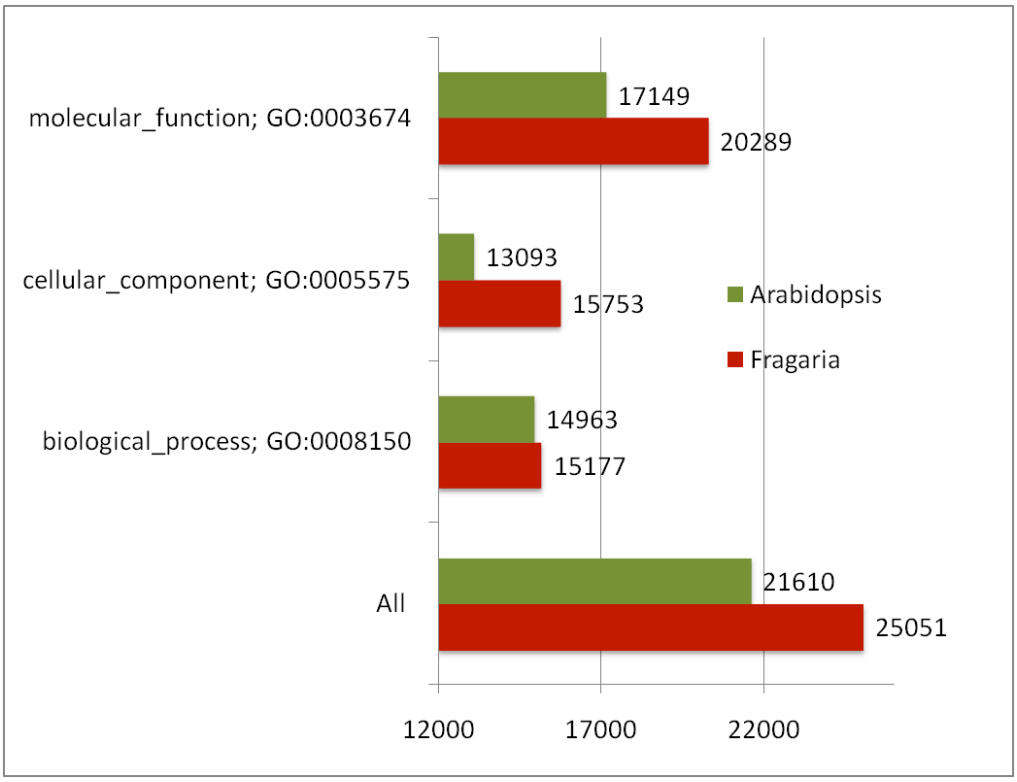
Supplementary Figure 4



Supplementary Figure 4. Schematic depiction of the approach used for parsing DNA sequence into protein coding and non-coding regions. GeneMark-ES+ is the self-training program that combines *ab initio* predictions with gene elements mapped from high confidence mapped ESTs as well as with gene deserts mapped from transposable elements.

Nature Genetics: doi:10.1038/ng.740

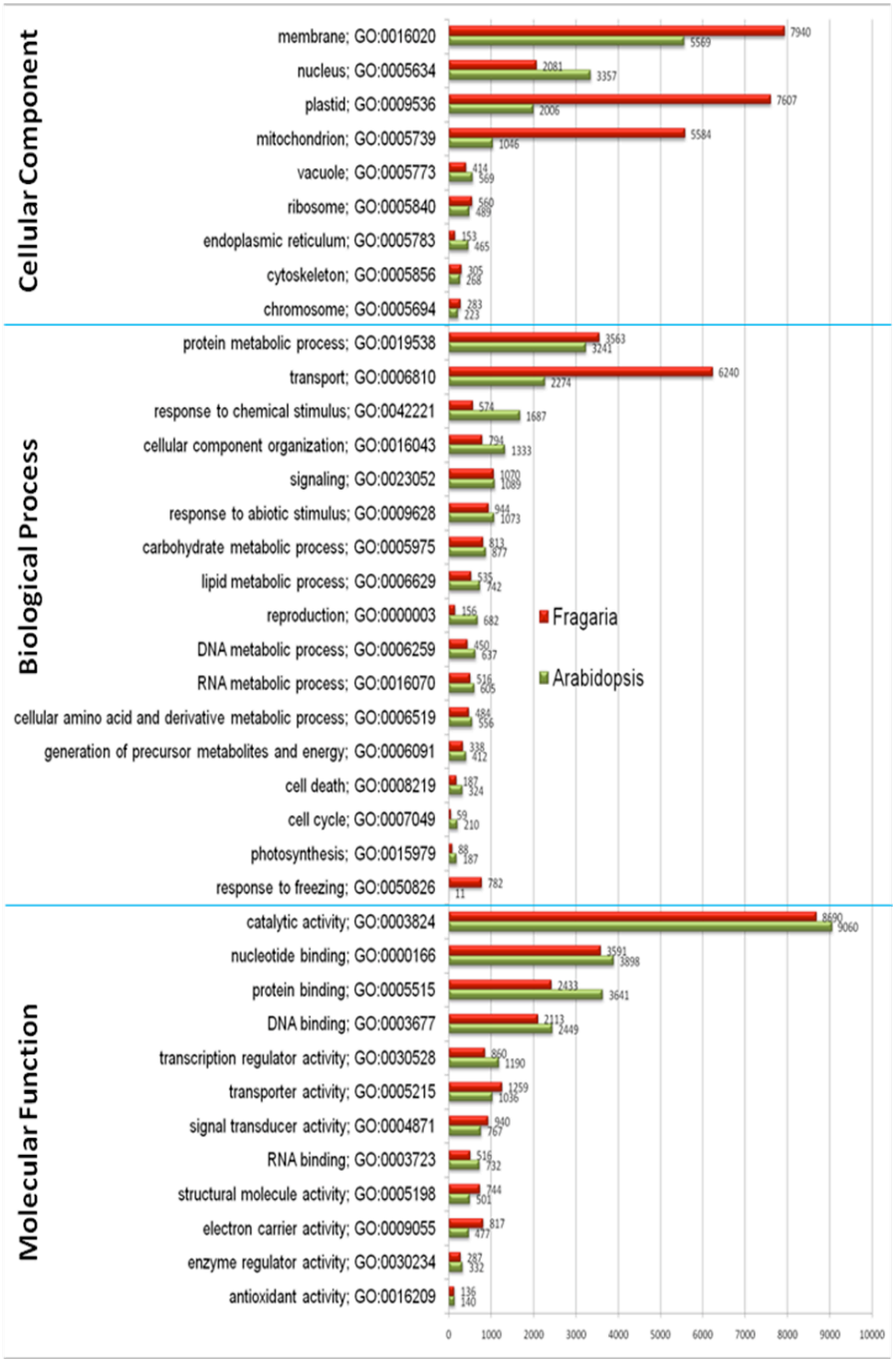
Supplementary Figure 5a



Supplementary Figure 5. Panel A. Summary of the *F. vesca* GO annotation and its comparison to *Arabidopsis thaliana* annotations. Annotations (as of March 10, 2010) available from the Gene Ontology website (www.geneontology.org). X-axis represents number of unique genes with GO annotations.

Nature Genetics: doi:10.1038/ng.740

Supplementary Figure 5b

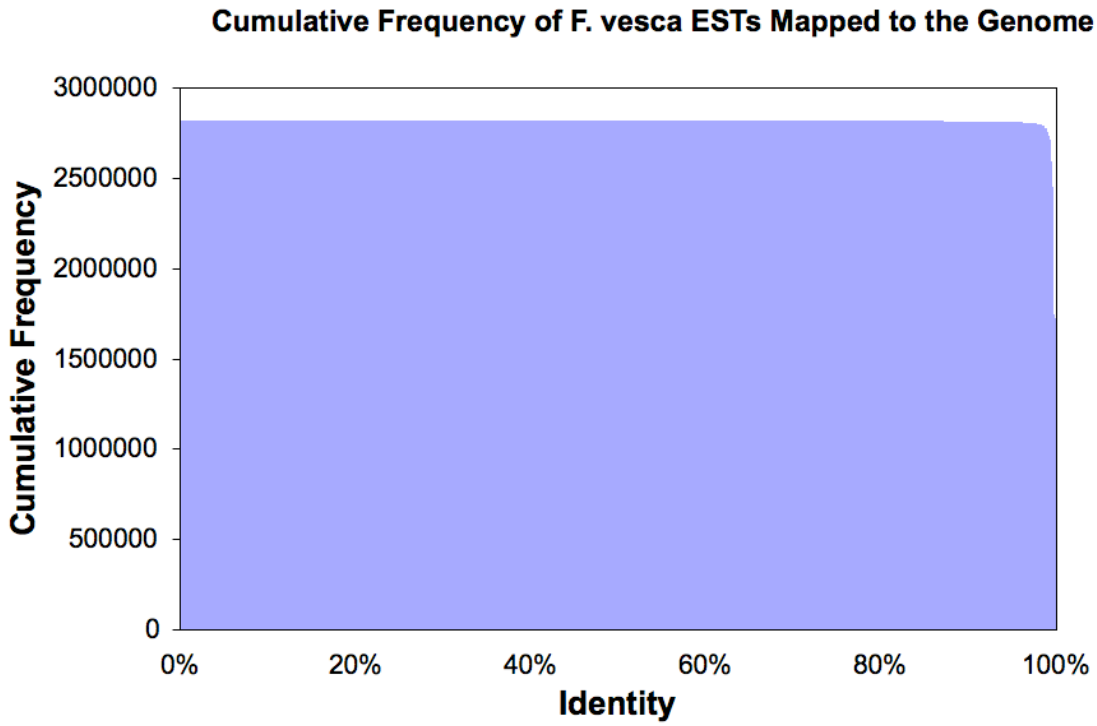


Supplementary Figure 5. Panel B. The category-wise summary of GO annotations from *Fragaria* and *A. thaliana*. X-axis represents number of unique genes with GO annotations.

Nature Genetics: doi:10.1038/ng.740

Supplementary Figure 6

A.

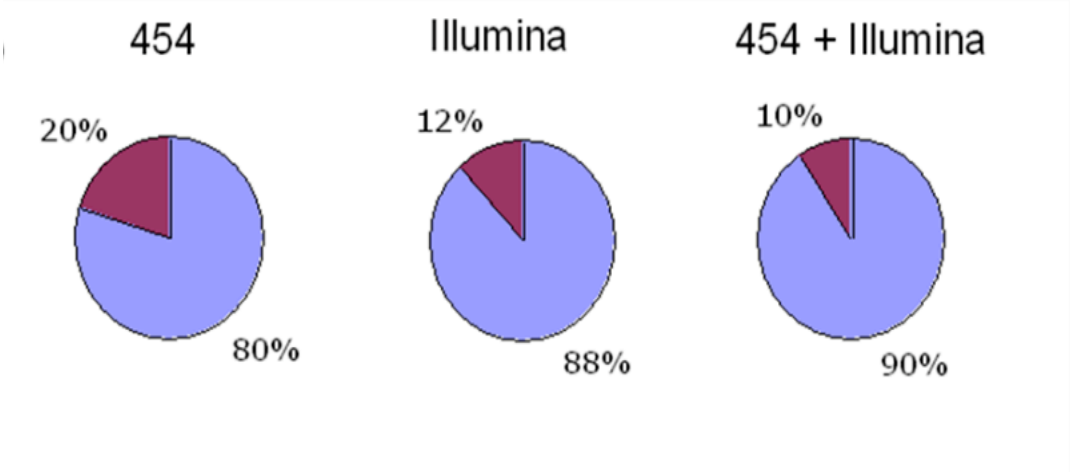


Supplementary Figure 6. **Panel A. Mapping of *F. vesca* ESTs onto the genomic sequence.** *F. vesca* ESTs (454 and Sanger) were anchored onto the genomic assemblies as spliced alignments using the program BLAT[2]. In total, 2,814,598 out of 3,117,395 transcript sequences (90.3%) could be mapped to the genomic sequence with a minimum aligned length of 50 nucleotides comprising a minimum of 50% of the transcript length. On the y-axis, the cumulative frequency of anchored ESTs is shown according to its dependence of alignment identity on the x-axis. For each EST, the single best match according to highest alignment identity has been selected in case of ESTs that mapped to several genomic alignment positions. The majority of ESTs could be mapped with high sequence identities, >2,800,000 and >2,810,000 sequences with an identity $\geq$95% and $\geq$90%, respectively.

Nature Genetics: doi:10.1038/ng.740

Supplementary Figure 6

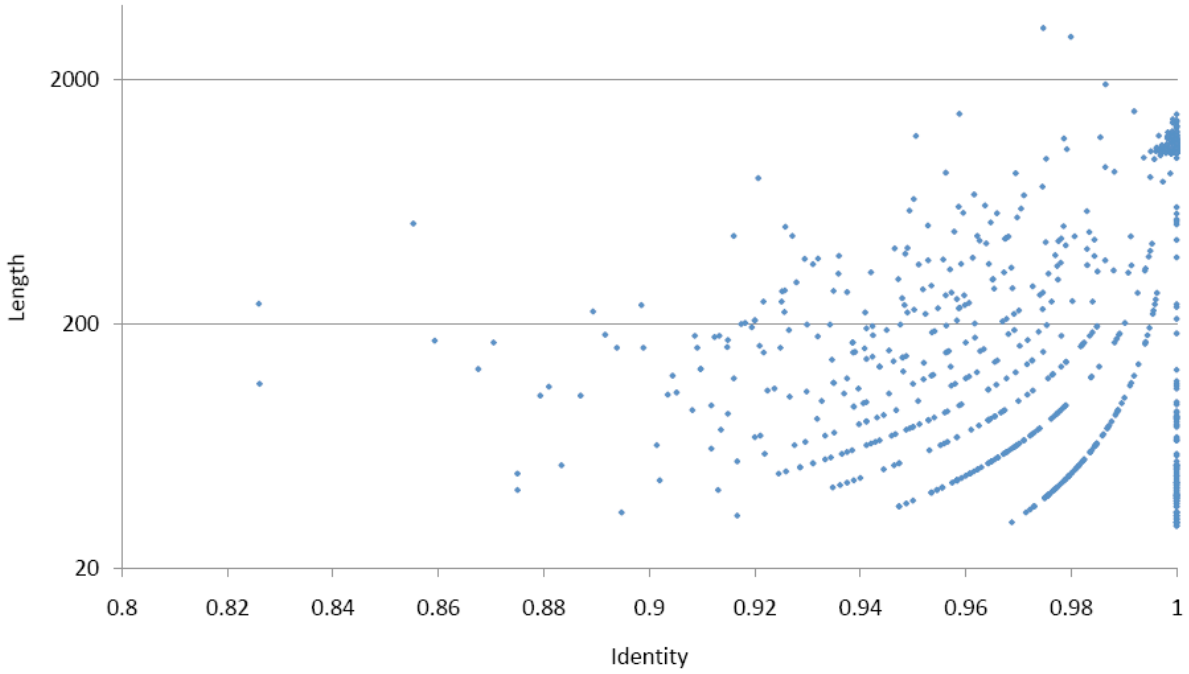
B.



Supplementary Figure 6. **Panel B. Transcript support of gene models.** Areas in blue indicate the proportion of *F. vesca* gene models supported by transcript evidence and areas in red indicate the proportion of gene models not supported by transcript evidence. Gene models were evaluated using ~3.6 Gb of Illumina RNA-seq data and ~1.2 Gb of Roche/454 ESTs representing a diverse collection of tissues and developmental stages. Overall, 90% of predicted gene models were supported by Roche/454 or Illumina transcript data, demonstrating the high accuracy of the *F. vesca* gene predictions. Moreover, over 90% of Roche/454 ESTs mapped to the sequence assembly, consistent with a near-complete genome coverage.

Nature Genetics: doi:10.1038/ng.740

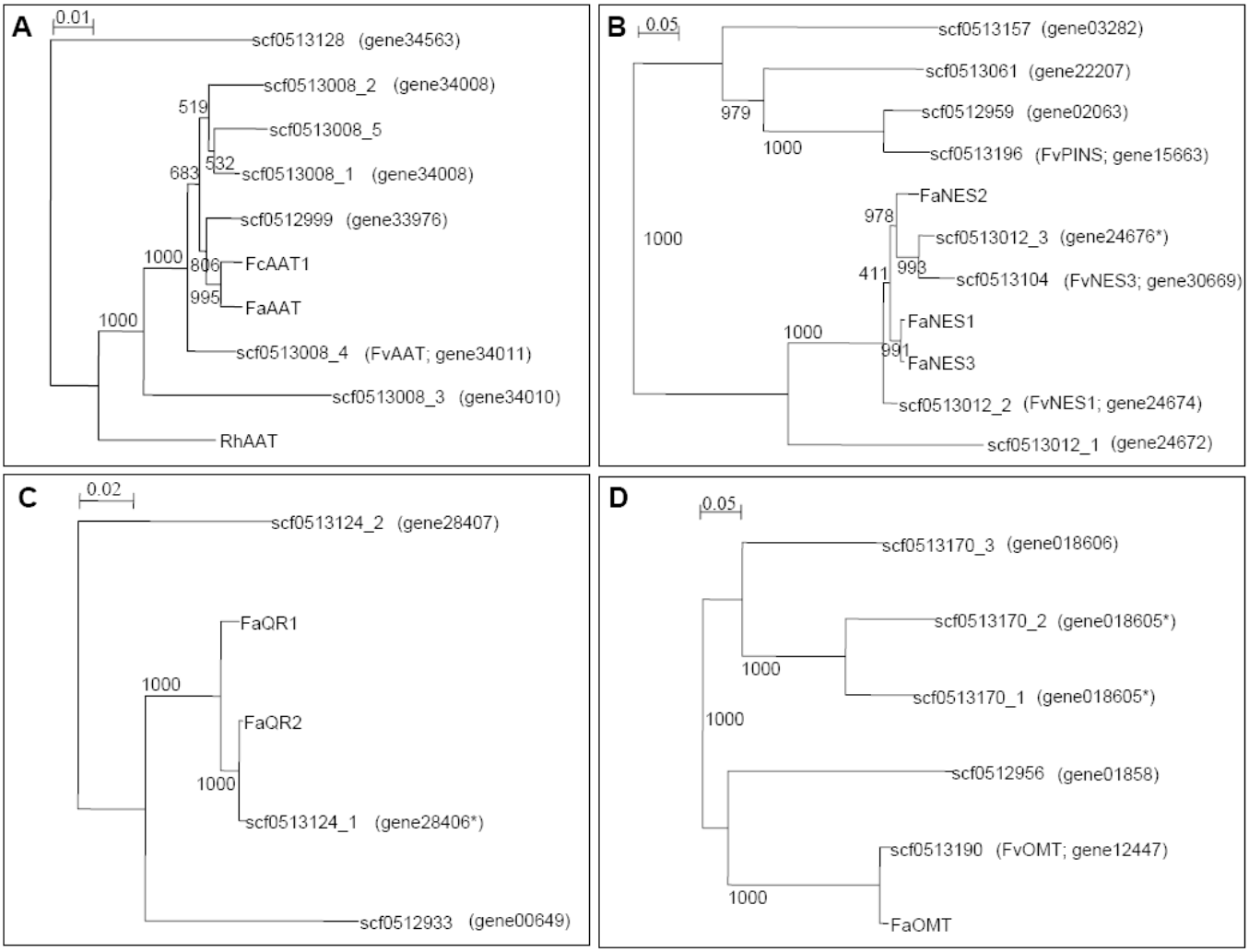
Supplementary Figure 7



Supplementary Figure 7. Chloroplast nomads present in nuclear genome. A total of 876 regions with >80% identity and lengths ranging from 30-3,237 bp (median = 185.5 bp) to the chloroplast sequence was identified in the draft assembly. These were interpreted as recent DNA transfer from the plastid to the nuclear genome.

Nature Genetics: doi:10.1038/ng.740

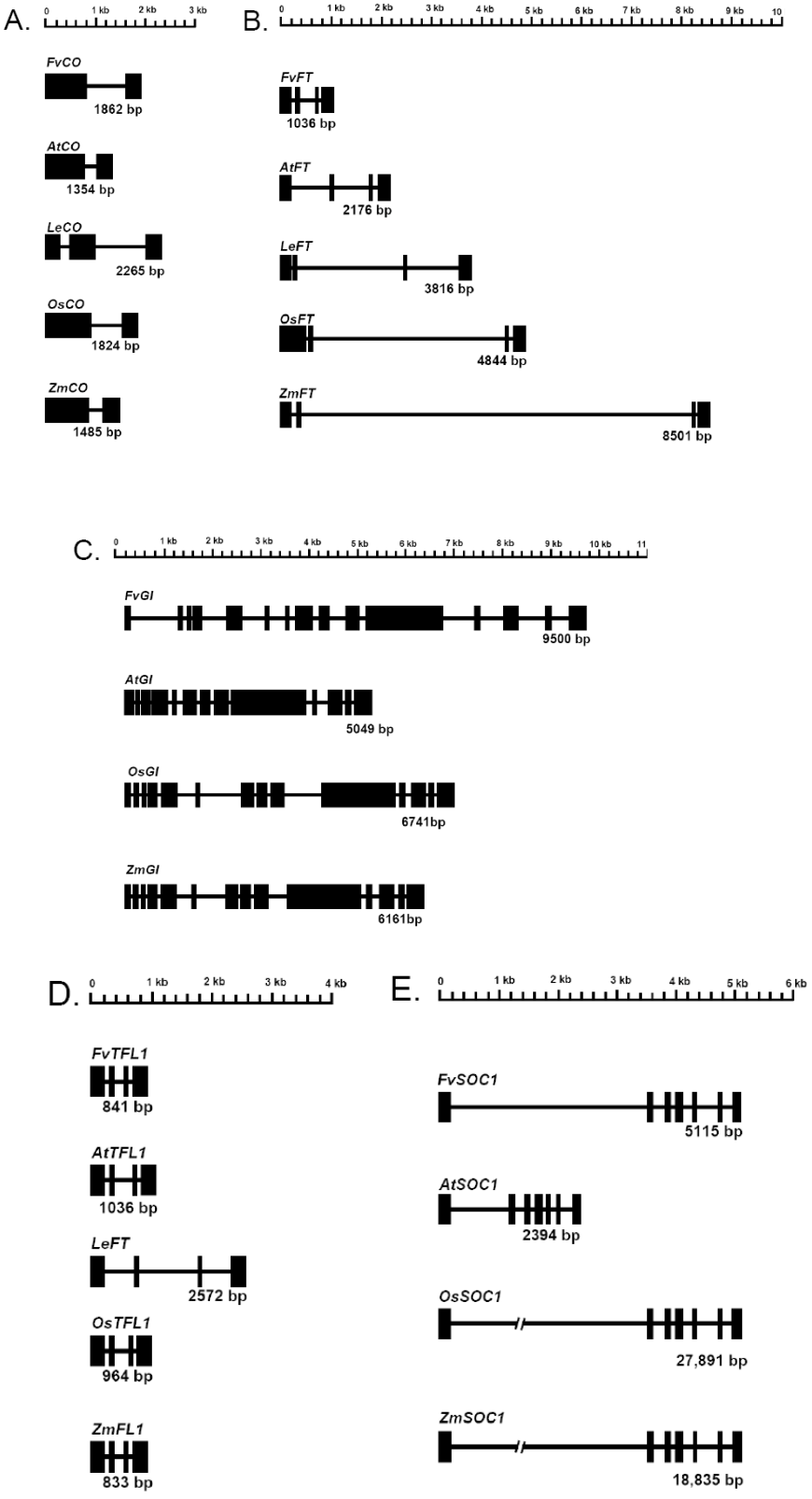
Supplementary Figure 8



Supplementary Figure 8. Phylogenetic analysis of *F. vesca* flavor related gene families. (A) Acyltrasferases **(B)** Terpene Synthases **(C)** quinone oxidoreductases and **(D)** O-methyltransferases. Trees and their significance (bootstrap) values were computed by ClustalX and NJplot softwares. Asterisks indicated problematic gene prediction as detailed in Supplemental Table 11.

Nature Genetics: doi:10.1038/ng.740

Supplementary Figure 9

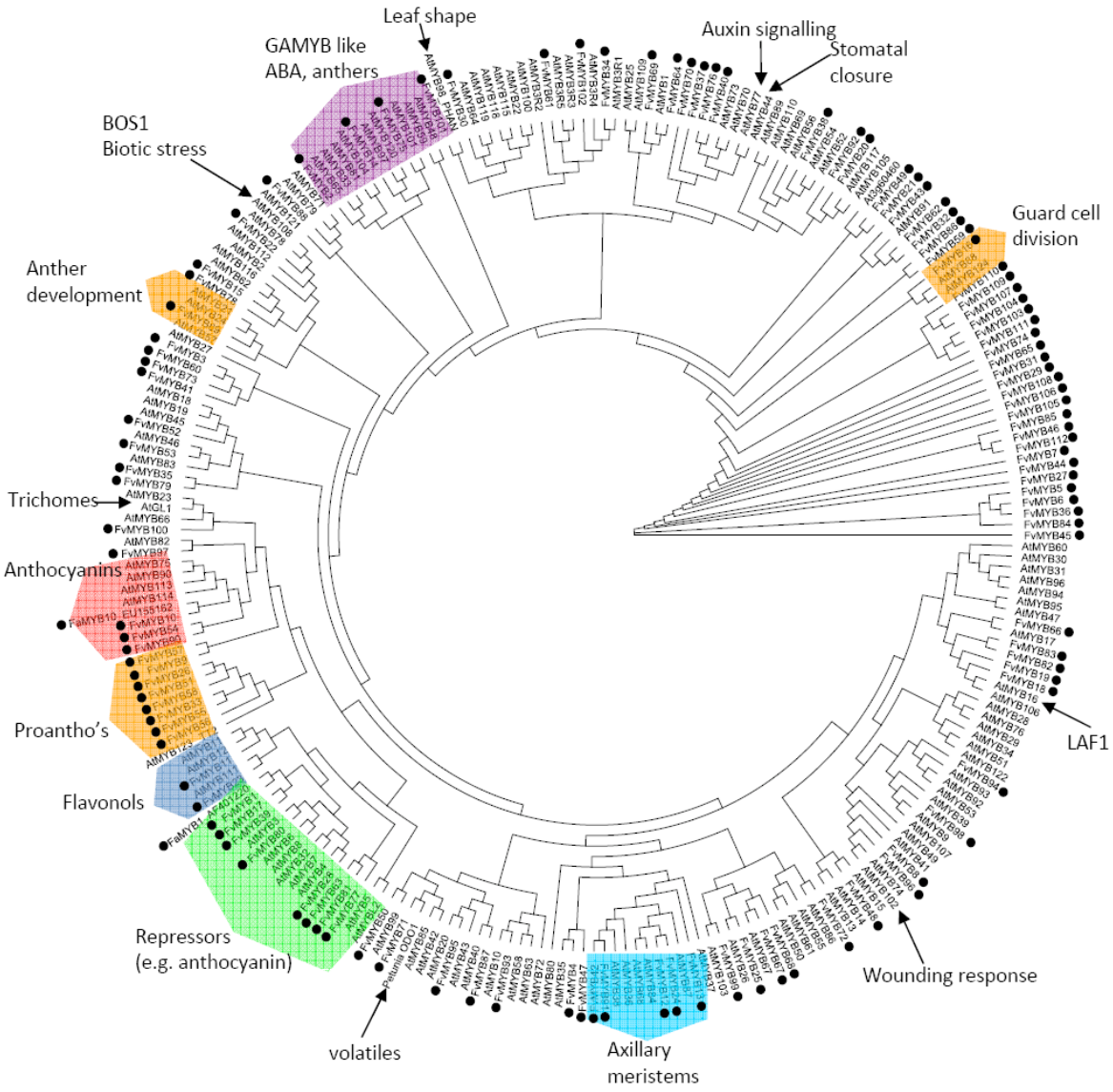


Nature Genetics: doi:10.1038/ng.740

Supplementary Figure 9. Intragenic architecture of genes central to photoperiodic flowering control. The *F. vesca* (A) *Co,* (B) *Ft,* (C) *Gi,* (D) *Tfl1* and (E) *Soc1 / Agl20* gene structures are compared to those of other plants (At, *Arabidopsis thaliana*: Le, *Lycopersicon esculentum*: Os, *Oryza sativa*: Zm, *Zea mays*).

Nature Genetics: doi:10.1038/ng.740

Supplementary Figure 10



Supplementary Figure 10. The MYB family of proteins from *F. vesca*. Phylogeny of full length predicted proteins of the R2R3 MYB family of Arabidopsis, and including full length predicted proteins of *Fragaria* MYBs (filled circles). Phylogeny was calculated using the Geneious program (http://www.geneious.com/), using an alignment generated by CLUSTAL W, and a bootstrap tree built via Neighbour-Joining, having distances calculated using Jukes-Cantor model.

Nature Genetics: doi:10.1038/ng.740

References to Supplemental Materials

1 Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494-6506, (2005).

2 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664, (2002).

3 Pertea, G. *et al.* TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-652, (2003).

4 Ostlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196-203, (2010).

5 Schwartz, T. S., Tae, H., Yang, Y., Mockaitis, K., Van Hemert, J. L., Proulx, S. R., Choi, J-H. and Bronikowski, A. M. A garter snake transcriptome: pyrosequencing, *de novo* assembly, and sex-specific differences. *BMC Genomics, In Press.*

6 Pandelova, I. *et al.* Analysis of transcriptome changes Induced by Ptr ToxA in wheat provides insights into the mechanisms of plant susceptibility. *Mol. Plant* **2**, 1067-1083, (2009).

7 Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* **9**, 811-818, (1990).

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

9 Altschul, S. F. & Lipman, D. J. Trees, stars, and multiple biological sequence alignment. *Siam J. Appl. Math.* **49**, 197-209, (1989).

10 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12, (2004).

11 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25, (2009).

12 Zdobnov, E. M. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848, (2001).

13 Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**, 953-971, (2007).

14 Small, I., Peeters, N., Legeai, F. & Lurin, C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4**, 1581-1590., (2004).

15 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567-580, (2001).

16 Schlueter, J. A. *et al.* Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. *BMC Genomics* **8**, (2007).

17 Shoemaker, R. C. *et al.* Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* **144**, 329-338, (1996).

18 Latrasse, A. in *Volatile Compounds in Foods and Beverage* (ed H. Maarse) 329-387 (Dekker, 1991).

19 Ulrich, D., Hoberg, E., Rapp, A. & Kecke, S. Analysis of strawberry flavour - discrimination of aroma types by quantification of volatile compounds. *Z. Lebensm. Unters. F. A.* **205**, 218-223, (1997).

20 Beekwilder, J. *et al.* Functional characterization of enzymes forming volatile esters from strawberry and banana. *Plant Physiol.* **135**, 1865-1878, (2004).

Nature Genetics: doi:10.1038/ng.740

21 Gonzalez, M. *et al.* Aroma development during ripening of *Fragaria chiloensis* fruit and participation of an alcohol acyltransferase (*FcAAT1*) gene. *J. Agric. Food Chem.* **57**, 9123-9132, (2009).
22 Aharoni, A. *et al.* Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species. *Plant Cell* **16**, 3110-3131, (2004).
23 Klein, D., Fink, B., Arold, B., Eisenreich, W. & Schwab, W. Functional characterization of enone oxidoreductases from strawberry and tomato fruit. *J. Agric. Food Chem.* **55**, 6705-6711, (2007).
24 Raab, T. *et al. FaQR*, required for the biosynthesis of the strawberry flavor compound 4-hydroxy-2,5-dimethyl-3(2H)-furanone, encodes an enone oxidoreductase. *Plant Cell* **18**, 1023-1037, (2006).
25 Lavid, N. *et al.* O-methyltransferases involved in the biosynthesis of volatile phenolic derivatives in rose petals. *Plant Physiol.* **129**, 1899-1907, (2002).
26 Wein, M. *et al.* Isolation, cloning and expression of a multifunctional O-methyltransferase capable of forming 2,5-dimethyl-4-methoxy-3(2H)-furanone, one of the key aroma compounds in strawberry fruits. *Plant J.* **31**, 755-765, (2002).
27 Mouhu, K. *et al.* Identification of flowering genes in strawberry, a perennial SD plant. *BMC Plant Biol.* **9**, (2009).
28 Stewart, P. J. & Folta, K. M. A review of photoperiodic flowering research in strawberry (*Fragaria* spp.). *Crit. Rev. Plant Sci.* **29**, 1-13, (2010).

Nature Genetics: doi:10.1038/ng.740