SUPPLEMENTARY INFORMATION APPENDIX

SI TEXT

DNA preparation and genome sequencing

DNA was prepared from nuclei extracted from leaves of the doubled haploid line DHL92 after maintaining plants for 72 h in the dark to avoid chloroplast contamination (1). The Roche 454 Titanium method was used for genome sequencing in a 454 GS-FLX sequencer. A single library was prepared for 454 shotgun sequencing. Four libraries of each 3-Kb, 8-Kb and 20-Kb paired ends were also produced. BES from two melon BAC libraries were already available (2).

Genome assembly

Assembly v3.3 was obtained using Newbler version 2.5 (Roche 454) using all reads available (Table S1). Each SFF file was filtered for duplicate reads using CD-HIT-454 (3). Raw BAC-end sequences were filtered for quality and vector contamination using SeqTrim (4). The sequences of the melon chloroplast and mitochondria (5) were used in Newbler as the screening database. Assembly v3.3 was homopolymer-corrected with 2 x 54 bp Illumina reads obtained in two lanes of a GAIIx instrument. Three mapping steps were carried out with GEM (with parameters -d 20 --max-indel-length 12), sequentially, to map unmapped reads from the step before: step 1 mapped with up to three mismatches; step 2, reads trimmed from the end until the first base with Q25 was reached, leaving more than 40 bases, and mapped again with up to three mismatches; and step 3, reads trimmed to 40 bases and mapped with up to two mismatches. 82,651,113 out of 97,643,590 reads (84.6 %) were mapped. Mapping positions were converted to SAM format and the SAMtools pileup program (6) was run to identify indels. Called indels (substitutions were ignored) with a quality greater than 20, and only involving homopolymers, were applied to the assembly sequence and qualities. In the case of insertions, the pileup consensus quality was used for the assembly consensus quality. The homopolymer-corrected assembly was named v3.4.

Assessment of the quality of the genome assembly

All sequences from the four finished BACs and the 57 BACs sequenced using the 454-pooled strategy were aligned to the unmasked melon assembly using MEGABLAST (version 2.2.19, parameters -v 7 -b 7 -e 1e-40 -p 80 -s 90 -W 12 -t 21 -F F). Only alignments with more than 98 % identity were kept, except for BAC 1-21-10 for which a threshold of 97 % was used as it corresponded to another melon variety. Alignments were filtered to get contiguous blocks of aligned sequence, 16 Kb minimum size, or when the scaffold length was smaller, then a minimum of 40 % of the scaffold length was used as cutoff. Contiguous alignments belong to the same block if the difference between their distance in the assembly and in the BACs is less than 11 Kb. For each BAC, the filtered alignments were plotted in an image. These images were annotated with information from transposons, tandem repeats, genes and segmental duplications (determined by excess of depth of coverage, WSSD) content annotated in the assembly. Gaps in both BAC and assembly were also added.

For quantitative statistics on the correspondence of BACs and scaffolds in the assembly, alignments in the blocks were reduced to non-overlapping alignments. Two rounds of reduction were completed. First, if two alignments overlap in the BAC sequence, then the overlap region was assigned only to the longer alignment and alignments completely included in larger alignments, either in the assembly or in the BAC, were removed. In the second round, for the remaining alignments already uniquely mapping in the BAC, overlaps in the assembly were removed as in the first round.

We first considered four BACs previously sequenced to a finished status by a shotgun-Sanger approach (7, 8, 9). Three of these BAC clones (60K17, 31O16 and 13J4, accession numbers AF499727, AY582736 and EF657230, respectively) belong to the DHL92 BAC library, covering 117 Kb, 159 Kb and 98 Kb of genomic region, respectively. The last BAC (1-21-10, accession number EF188258) is a clone of 92 Kb from the WMR29 melon library, obtained from the American cantaloupe type WMR29. For each of these BACs, subclone libraries were sequenced with a genomic coverage from 4× to 6×, while remaining gaps were resolved by PCR amplification and subsequent sequencing of the amplified regions, or by sequencing extra clones with reverse primers. Both 60K17 and 13J4 are fragments of the MRGH63 contig, a region in linkage group 5 that contains a cluster of R-gene homologues. These two BACs were expected to be separated by a region of ~3 Kb.

We also took into account the DNA sequences in two pools of BACs of the DHL92 library, previously sequenced using 454 pyrosequencing and a combination of shotgun and paired-end sequencing (10). Of the 35 BACs in the first pool, 32 are anchored to genetic markers distributed throughout the melon genome; the other 3 BACs (13J4, 43H20 and 14M22) overlap and belong to the MRGH63 contig. In the second pool of 23 BACs, 20 were linked to known genetic markers distributed throughout the melon genome, and two had no link to any genetic marker. The last one, 43H20, was also present in the first BAC pool. The final assembly of these BACs, once 60K17 was added to the MRGH63 contig sequence, has 73 scaffolds, totalling 6.3 Mb, 73 % of which are longer than 60 Kb, with an average scaffold size of 86.8 Kb and a final coverage of 39×. On average, there are seven stretches of Ns per scaffold (produced as a result of contig scaffolding), representing 4.8% of the total length of the scaffold sequences. Most likely complete sequences were produced for 50 BACs, three were incomplete but in the range of 60-80 Kb, and the remaining four gave very little information. The differences found when comparing the sequences of 13J4 obtained by the Sanger and the pooling strategy involve five small N stretches (in two regions that contain repetitive sequences) equivalent to 3.6 % of the BAC length and 17 reductions in polymeric or tandem repeats regions that represent 1.7 differences every 10 Kb.

The location of the four BACs in the assembly was unequivocal and contiguous. The total length of the BACs was 467 Kb, but their size was 490 Kb in the assembly (Table S4). In both the assembly and the BAC sequences, 432 Kb had alignments with more than 99 % identity. Slight differences in the length of these alignments are due to small indels allowed when mapping. The greater assembly size for these BACs is due almost entirely to undefined sequence in gaps in the assembly (Table S5, Fig. S1): 35 Kb of the BACs without a proper alignment had their counterpart in 57 Kb of sequence in the assembly, of which 53 Kb are gaps. Even the remaining 4 Kb were from stretches of sequence containing tandem repeats intercalated in large gaps in BAC 1-21-10 (Fig. S1d). Excluding isolated small gaps (shorter than ~2 Kb), one region in each BAC includes the largest gaps. These regions overlap, partially or fully, with previously annotated repetitive sequences such as CURE and CUMULE retrotransposons, or short sequence repeats (SSRs) (Table S5).

As expected, we found that the 60K17 and 13J4 BACs, of the MRGH63 contig, are separated by 2,950 bp in the assembly. 13J4 has been previously sequenced using both Sanger and 454 following a pooling strategy, and these strategies were compared to assess the quality of the latter (10). Not surprisingly, while only five small stretches of N's were found (3.6 % of the BAC length), in the melon assembly there are 11 gaps (4.65 % of the BAC length), the largest almost 5 Kb, two translocations and at least 11 bp not mapping or corresponding to a gap. The 5' region of 60K17 overlaps with a segmental duplication in the melon assembly and, except for a small gap, has been well constructed in the assembly.

We were also able to identify the 73 scaffolds, from the two pools of 57 BACs, in contiguous regions in the assembly. Two scaffolds have a translocation in the assembly, MRGH63 and scaffold00041. These were excluded when calculating the total statistics (Table S6). While the sequence of the remaining scaffolds totals 5.93 Mb with 299 Kb of gap sequence, we found an additional 176 Kb (88 Kb of gaps) in the assembly. Overall, 5.48 Mb correlate in both sequences in intervals with high similarity (> 99 % except for three 98 % alignments present in two different scaffolds). Therefore, on average 97.2 % of a BAC sequence excluding gaps is well represented in the assembly.

Melon genome anchoring

A set of 768 polymorphic SNPs between the melon lines PI 161375 and PS (11) was used to genotype 72 double-haploid lines from the melon mapping population (12) using the Illumina GoldenGate genotyping assay. Mapmaker software (13) was used for SNP mapping, with a LOD score of five, maximum distance between markers of 30 cM and the Kosambi mapping function. A genetic map containing 602 SNPs markers was obtained (http://melonomics.net).

To anchor the genome assembly to the genetic map, we performed BLAST analysis (e-value cutoff of 1E-30) of every SNP marker in the genetic map against the genome assembly, and scaffolds were assigned to linkage groups accordingly. When more than one marker had hits on the same scaffold, it was possible to orientate this scaffold on the map. For scaffolds that were not oriented, we used information from two additional melon genetic maps: i) another version of the SC × PS genetic map with 332 additional RFLP, SSR and SNP markers (12) and ii) a genetic map in the PI 414723 × Dulce genetic background containing 1,092 SNP markers (Syngenta Seeds). The genome anchoring to the pseudochromosomes was drawn with the Harry Plotter software (http://genomics.research.iasma.it). Five scaffolds (CM3.5_scaffold00001, CM3.5_scaffold00002, CM3.5_scaffold00008, CM3.5_scaffold00013 and CM3.5_scaffold00056), each mapping in two different locations in the genome (LG IX and LG XII, LG IV and VIII, LG III and LG VI, LG III and LG VII, LG VII and LG X), were manually inspected and found to be misassembled chimaeras. Each scaffold was split in two separate scaffolds, with the splitting point selected between the two contigs that showed most inconsistencies in paired-end links. The melon genome assembly v3.4 was corrected to v3.5, identical except for five additional scaffolds and with a slightly increased final size of the assembled genome (375.5 Mb). This version was used for further genome analysis.

Transposon identification and analysis

Transposon representatives were identified with a combination of *ab initio* and homology-based methods, and were used to search for related sequences, assembled into copies, using a dedicated pipeline. Retrotransposon insertions were dated by intra-element LTR comparison as described (14). The cucumber genome sequence used for the transposon, phylome and comparative analyses was retrieved from Phytozome (http://www.phytozome.net), which is based on the Gy14 genotype and has been sequenced using the 454 Roche system. The cucumber 9930 genome (15) was also used for the transposon and comparative analyses.

Identification of LTR-retrotransposons

Candidates for LTR retrotransposons were identified using LTR_FINDER (16) with default settings. Copies of these candidates were retrieved with a modified version of a script from the MITE-Hunter suite (17). Each candidate that retrieved at least one copy was aligned with its copies using MUSCLE (18) (which was used for all other alignments), taking 60 bp of flanking sequence. These alignments were checked for target site duplications and that the borders of the elements align while the flanking sequences do not. To further verify these candidates, they were used to query a database of all LTR-retrotransposons in RepBase (http://www.girinst.org) with tblastx (e-value < e-10, blastall suite available at http://www.ncbi.nlm.nih.gov; all subsequent BLAST analyses were also performed with this suite). According to the best hit the candidates were attributed to either the gypsy or copia superfamilies, or discarded if no homology was found.

These verified candidates were clustered according to the internal sequence (between the LTRs as defined by LTR_FINDER), with a threshold of 80 % similarity over 80 % of the length. A home-made Python program was used for this, implementing a hierarchical clustering algorithm and considering only columns without gaps in the calculation of percent similarity. The longest sequence of each cluster was chosen as the representative sequence. These representative sequences are in a database available on the MELONOMICS website, with annotated features as defined by LTR_FINDER, superfamily, and copy number.

Since the LTR_FINDER prediction is based on the direct repeats that could be LTRs, only elements with well conserved LTRs can be detected. For a comprehensive view of the LTR retrotransposon landscape in this genome, we developed a set of programs, COPILIST and COPILIST-NR, to identify copies of a query sequence. These were used to identify truncated copies of the verified LTR retrotransposon representatives, and resolve redundancy where a particular genomic region was picked up by various representatives. The algorithm and parameters are described below (see *Finding copies of a sequence*).

The copies of the representative sequences are annotated as the SO term “retrotransposon” in the final annotation. Each annotation includes which superfamily the element belongs to, a link to the representative sequence it was retrieved with and its percent similarity to it, and how many other sequences are in that particular family. All these annotated sequences were masked before proceeding to the next step.

This same analysis was applied to the cucumber genome (http://www.phytozome.net; (15)). To fish for more degenerate elements, parameters were relaxed: LTR_FINDER was used both with default parameters and with parameter -p set to 30. These sequences were clustered to remove redundancy and the longest element for each cluster selected. The results were not filtered for having tblastx similarity to LTR-retrotransposons in RepBase, and were used as is to identify copies.

Identification of non-LTR retrotransposons

Non-LTR retrotransposons previously identified in melon BACs (10) were used as queries to search for copies as described below. These are all annotated as “non_LTR_retrotransposon” and also link to the representative sequence they were identified with. All these annotated sequences were masked before proceeding to the next step.

Identification of retrotransposon-related sequences

Two checks were used to ensure that no sequence had been overlooked by assembling the copies, and that the representative sequences are truly representative of their respective categories. First we used the representative sequences to mask any region that might be homologous (blastn, e-value < e-10) but not have been picked up as a copy. Second, we used protein queries of retrotransposase, retrieved from NCBI excluding any sequence annotated as “putative” or “hypothetical” (in order to minimize the propagation of errors or uncertainties in the public databases), to fish (tblastn, e-value < e-10) for any region that might not have been retrieved with the representative sequences. This second approach identifies elements that might be too old or degenerate to be identified as a copy of a representative. These two approaches yielded a low percentage of the genome, indicating that the copy-finding algorithm is accurate and that the database of representative sequences is truly illustrative of the families of retros in this genome. These sequences are all annotated as "transposon_fragment" with a note specifying that they are a retrotransposon related sequence. All these annotated sequences were masked before proceeding to the next step. In this category we also included 26 gene sequences that had been identified as related to non-LTR retrotransposons by the phylogenomic analysis. While the results of the transposon annotation were used to mask the genome before gene annotation and phylogenomic analysis, the latter revealed certain gene families that are related to retrotransposon sequences. Some did not fit our similarity criteria to a known transposon sequence, while others were indeed related to non-LTR retrotransposons and had escaped our annotation. This is most likely because non-LTR retrotransposons cannot be identified with structural characteristics and that the RT and integrase database we constructed excluded "hypothetical" and "putative" proteins.

Identification of DNA transposons

To identify DNA transposons, the general strategy was to fish for sequences homologous to transposase, refine them by aligning similar hits taken with flanking sequence, then select representatives and search for copies.

We constructed a protein database by querying NCBI with the keyword "transposase" in conjunction with superfamily names such as "PIF", "hAT", "CACTA", "MULE", "hop", "jittery", "Mariner", as well as "helitron helicase" to retrieve transposase sequences that have been attributed to a superfamily as well as those that have not. We excluded from these searches any sequence annotated as "putative" or "hypothetical", in order to minimize the propagation of errors or uncertainties in the public databases. All sequences in the genome similar (tblastn, e-value < e-10) to any in our transposase database were retrieved. These sequences were then re-blasted against the subset of the database attributed to a superfamily, and grouped according to this criterion.

The sequences in each superfamily group were clustered using UCLUST (18) (80 % similarity over 80 % query length, with iddef parameter set to 2). For selected clusters (most homogeneous or largest), the sequences were extended 5,000 bp in either direction, and aligned. These alignments were manually inspected to extend the definition of the elements as far as the alignment was maintained, allowing TIRs to be identified in some cases. One representative sequence was selected per cluster, and these were used as queries to search for copies using COPILIST-NR. These representatives can also be found in the database in the Melonomics site (http://melonomics.net).

Copies of each representative were retrieved as described below. These sequences are annotated as the SO term "DNA_transposon", and each annotation includes which superfamily the sequence belongs to, a link to the representative sequence it was retrieved with and its percent similarity, and how many other sequences are in that particular family. All these annotated sequences were masked before the next step.

Identification of transposase-related sequences

As with the annotation of retrotransposon related sequences, we first masked any sequence with similarity (e-value < e-10) to the representative sequences. This yielded a very low percentage, indicating that the copies we defined cover the families of these representatives. To ensure that the representative sequences we identified do indeed represent the DNA transposon content of this genome, we used the transposase database to retrieve (tblastn, e-value < e-10) any sequence that would not have been picked up with the set of representatives. The genome percentage this search yielded was low, so no major family of DNA transposons was missed. These sequences are all annotated as "transposon_fragment" with a note specifying that they are a DNA transposon related sequence.

Finding copies of a sequence: development of COPILIST (COPy Identifier by LInking Split hiTs)

Given a set of representative sequences, the next step in the analysis is to find all the copies of each of these representatives within the genome. However, copies within a family can vary with mutations accumulated over time and so a copy will often be composed of fractionated BLAST high-scoring pairs (HSPs) (blastn, e-value < e-10). These fractionated hits need to be assembled into a copy that spans the greatest possible length of the query. This is not straightforward for various reasons: the repetitive nature of the structure of certain TEs; rearrangements within an element; and the fact that an element can be found in many locations in the genome. As a result, to assemble HSPs to form a copy, they must be ordered along the query, be on the same strand, and be separated by, at most, a specified gap threshold. In addition to these criteria, the HSPs need to be assembled such that the set of longest non-overlapping copies of a given element are found.

To solve this optimization problem we used a directed acyclic graph, with as nodes the HSPs which are connected by a directed edge when the previously mentioned criteria are fulfilled. Finding the set of non-overlapping copies of an element is therefore reduced to finding the set of non-overlapping longest paths in this graph (http://en.wikipedia.org/wiki/Longest_path_problem).

Resolving redundancy
While the representative sequences were selected to be at least 20 % different from each other, they remain similar to a certain extent within a superfamily, so a genomic region can be identified as a copy for more than one representative. To resolve this redundancy, we maintained the copy with the highest coverage of its query sequence, and truncated overlapping copies. This was done repeatedly until all overlaps were removed.

The copy finding program in the MITE-Hunter suite (17) does not supply the genomic coordinates of the copies it returns or allow parameters to set maximum gap length, nor does it resolve redundancy. It was sufficient for the purpose of identifying copies of the LTR retrotransposon sequences for refining by alignment, but we found the need to write our own programs as the analysis progressed. These tools are COPILIST (COPy Identifier by LInking Split hiTs), which identifies the copies of a single query element, and COPILIST-NR, which identifies the set of non-redundant copies given a set of query elements. These software packages are available upon request to EMH.

Data available on the genome website
As well as an annotation of the genomic sequences, we have a database of all the representative sequences used to define families, which are annotated with copy number of the family, superfamily, and structural characteristics when available (http://melonomics.net).

Dating insertion time of LTR retrotransposons
For this analysis only the families with more than 10 copies, covering at least 90 % of the length of the family representative, were considered. For each of these families, we aligned these long (>90 % query coverage) elements to the representative and selected those which aligned with at least 50 % of the length of the representative's LTRs, as defined by LTR_FINDER. The two LTRs of each selected element were aligned and the date of divergence calculated using Kimura's two-parameter method (19): if P is the transition fraction in the aligned sequences, Q is the transversion fraction, K is the evolutionary distance, T is the time of divergence and k be the evolutionary rate, then K = -1/2 * ln[(1-2P-Q) * sqrt(1-2Q)] and T = K / 2k. We took k as 1.3 x 10^{-8} substitutions/site/year, as previously used to date LTR retrotransposons (14) and was taken from the rate calculated for the *Adh* locus in grasses (20), and divided by two as LTR retrotransposons have a higher substitution rate than genes.

Gene prediction
A combination of *ab initio* and evidence-based approaches were used to annotate protein-coding genes. 753,004 ESTs generated by both Sanger and 454 sequencing (11) and reads obtained from SOLiD sequencing of eight RNA pools that represent 65 melon varieties were aligned to the genome with GMAP and then assembled with PASA. The same ESTs were also assembled into unigenes with MIRA and aligned to the genome with GMAP. The programs Geneid, SGP2, GlimmerHMM and Augustus were trained using training candidate transcript models generated by PASA and then run on the genome (with the previously identified transposable elements masked) to predict genes. Augustus was also run using *A. thaliana* parameters. For SGP2, TBLASTX alignments between *A. thaliana* and melon were used to improve gene prediction accuracy. Geneid and Augustus were also run with evidence from RNAseq alignments from the above mentioned SOLiD sequencing using GEM. The plant division of Uniprot90 plus all cucumber and melon protein entries in Genbank were mapped to the genome first with BLAT and then refined with GeneWise. CDS sequences from the DOE Joint Genome Institute’s annotation of the cucumber Gy14 (http://www.phytozome.net) were also mapped to the genome using Exonerate. The above mappings and predictions were combined as consensus CDS annotations using Evidence Modeler. Transcript alignments were given the most weight, followed by protein alignments, conservation and RNA-seq-assisted methods (SGP2, NextGeneid and Augustus-melon), and finally *ab initio* predictions (Augustus-*Arabidopsis*, geneid, GlimmerHMM). The consensus gene models were loaded into the PASA database and passed through two rounds of UTR and alternative splicing updates before the final gene models were obtained.

Functional gene annotation
Protein coding genes predicted in the melon genome were functionally annotated using an in-house automated analysis. For each protein sequence, our approach identifies protein signatures, assigns orthology groups, and uses orthology-derived information to annotate metabolic pathways, multi-enzymatic complexes, and reactions. Proteins were inspected for different protein signatures (such as

families, regions, domains, repeats, and sites) using InterProScan (21) and the InterPro database (22). These signatures were used for the classification and automatic annotation of protein sequences, assigning biological functions and gene ontology (GO) terms. Additionally a Blast2GO analysis using a blast search against the nr database was used to assign GO terms to the proteins. Each sequence was then mapped to KEGG orthology (KO) groups using the freely accessible web server KEGG Automatic Annotation Server (KAAS) (23, 24). A bi-directional best hit approach (BBH) was used in the homology search against a representative gene set from 28 different eukaryote species, including *Arabidopsis thaliana*, *Oryza sativa* var. *japonica*, *Ostreococcus lucimarinus* and *Cyanidioschyzon merolae*. KO identifiers were then used to retrieve the KEGG relevant functional annotation, such as metabolic pathways and external database references, using the KEGG Perl API. UniProt identifiers (25) of orthologs obtained through the phylogenetic analysis described below were also used to derive metabolic pathways, multi-enzymatic complexes and reaction information available in the Reactome database (26). All annotations were finally stored in a MySQL relational database (http://www.mysql.com/).

Phylogenomic analyses

A phylome, the complete collection of phylogenies for each gene encoded in a genome (27), was reconstructed for *C. melo*. Proteins encoded in 23 fully-sequenced plant genomes, including the melon genome, as well as five non-plant out-group species were downloaded from various sources (Table S12). The final database used for the phylome reconstruction contained 42,790 unique protein sequences. The resulting melon phylome has 22,218 gene trees, representing 80.0 % of the predicted melon genes. A phylogeny-based prediction of orthology, duplications and functional assignment was also performed. The *C. melo* phylome was scanned to infer orthology and paralogy relationships of melon genes and those of other plants, based on the reconstructed gene phylogenies (28). In addition, the phylogenetic position of melon was determined and the plant gene sets compared.

Phylome reconstruction

For phylome reconstruction, a Smith-Waterman (29) search was used to retrieve homologs using an e-value cut-off of 10e-5, and considering only sequences that aligned with a continuous region representing more than 50% of the query sequence. Then selected homologous sequences were aligned using three different programs: MUSCLE v3.7 (30), MAFFT v6.712b (31), and DIALIGN-TX (32). Alignments were in forward and reverse direction (using the Heads or Tails approach (33)), and the six resulting alignments were combined using M-COFFEE (34). The resulting combined alignment was subsequently trimmed with trimAl v1.3 (35), using a consistency score cutoff of 0.1667 and a gap score cutoff of 0.1 to remove poorly aligned regions.

Phylogenetic trees based on the maximum likelihood (ML) approach were inferred from these alignments. ML trees were reconstructed using the best-fitting evolutionary model. The evolutionary model best fitting each protein family was selected by first reconstructing a phylogenetic tree using a neighbour-joining (NJ) approach as implemented in BioNJ (36). The likelihood of this topology was computed, allowing branch-length optimisation, using seven different models (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff), as in PhyML version 3.0 (37). The two evolutionary models best fitting the data were then determined by comparing the likelihood of the models used according to the AIC criterion (38). Finally, ML trees were derived from these two models using the NNI (nearest neighbor interchange) default tree topology search method: the one with the best likelihood was used for further analyses. A similar approach based on NJ topologies to select the best-fitting model for a subsequent ML analysis has been shown to be highly accurate (39). Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML. In all cases, a discrete gamma-distribution with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data.

Phylogeny-based orthology and paralogy prediction

Orthology and paralogy relationships among *C. melo* genes and those encoded by the other genomes included in the melon phylome were inferred using a phylogenetic approach (40). In brief, a species-overlap algorithm, as implemented in ETE2 (41), was used to label each node in the phylogenetic tree as duplication or speciation depending on if the descendants partitions have, at least one, common species or not. Orthology and paralogy relationships between the members of a gene family were derived according to the original definition of orthology, that is, two genes were considered as orthologs or paralogs to each other if they diverged from their common ancestor through a speciation or a duplication node, respectively (40). Resulting orthology and paralogy predictions can be accessed through http://phylomeDB.org.

Phylogeny-based functional annotation

Using a phylogeny-based annotation approach (42), predicted one-to-one orthology relationships of melon genes to annotated genes in other species were used to automatically transfer gene ontology terms to enrich the gene functional annotation. 9,944 one-to-one orthology relationships among *C. melo* genes and genes from species used in the melon phylome with some GO annotation were found. Using these pairs, 121,587 GO terms were transferred from the *C. melo* genes counterparts to them. All this data can be accessed though the genome project website (http://melonomics.net).

Species tree reconstruction

A species phylogeny among the species included in the melon phylome was inferred using two complementary approaches. Firstly, a super-tree was inferred from all the trees in the phylome by using a Gene Tree Parsimony approach as implemented in the dup-tree algorithm (43) that finds the species topology that minimizes the number of total duplications implied by a collection of gene family trees, i.e. the phylome. Secondly, 60 gene families with a clear phylogeny based one-to-one orthology in at least 20 of the 28 species included in the analyses were used to perform a multi-gene phylogenetic analyses (Table S12). Protein sequence alignments were performed as described above and then concatenated into a single alignment. Species relationships were inferred from this alignment using a ML approach as implemented in PhyML, using JTT as evolutionary model, since in 44 out of 60 gene families this model was the best-fitting. Branch supports were computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution. Both trees fully agree in the topology for the common species (species present in less than 30 genes were removed from the concatenated alignment) (Fig. 3).

Lineage specific duplications

We scanned all trees in the melon phylome to detect duplication events at the *Cucumis*-genus and *C. melo* lineages. 7,184 genes (~ 26%) of the genome were mapped with either a duplication at *Cucumis*-genus level (4,190 genes), or *C. melo* level (2,994 genes). Then, these genes were merged into groups of non-redundant member with at least 20% of overlap of shared genes.

Functional enrichment analyses

The relative evolutionary age of each of the detected duplications was inferred using a topology-based approach (44). Clusters of duplicated melon genes specifically to *Cucumis*-genus level or to the species itself with more than 10 members were analysed looking for any functional enrichment. Enrichment analyses of over-represented GO terms for these expanded families compared with the annotated *C. melo* genes were performed by using FatiGo webserver (45) using the two-tailed Fisher exact test and e-value cutoff of 10e-5. Then, GO terms redundancy was reduced using REViGO webserver (http://revigo.irb.hr) (46) setting a similarity threshold of 0.5 and using as reference database *A. thaliana* and as semantic similarity algorithm *SimRel*.

RNA gene annotation

Non-coding RNAs (ncRNAs), such as transfer RNAs, ribosomal RNAs, small nuclear RNAs, micro RNAs (miRNAs) and small nucleolar RNAs, were identified in the melon genome using Infernal (v1.0.2) (47) against the Rfam database (v10.0) (48). Rfam_scan.pl script was used to reduce the search space and speed up the BLAST search of functional ncRNAs. The analysis yielded a total of 1,166 putative functional ncRNAs (Dataset S2). An alternative approach for the identification of miRNAs (see below) yielded 87 additional ncRNA genes, giving a total of 1,253 ncRNA genes identified in the melon genome (Dataset S2, Table S13). Genome localization and clustering analyses was using standard Python scripts (http://www.python.org/) and the BioPython library (http://biopython.org/wiki/Main_Page) (Table S14).

Potential melon *MIRNA* genes were also identified by BLAST (49) comparison of the mature miRNA sequences from the *Arabidopsis thaliana* small RNA project (ASRP) (50) and the microRNA Registry (MIRbase) (51) databases against the melon genome. For each blast hit, a region of 600 bases upstream and downstream of the alignment was selected and used to search a near perfect reverse-complementary sequence to the miRNA (miRNA*) using the miRanda algorithm (52). Regions without a miRNA* sequence were not considered. The minimum genomic regions containing miRNA and miRNA* sequences were selected as potential precursors. Precursors were used to predict the secondary structure of the RNA with Mfold (53), and to calculate the MFEI index (54); secondary structures were then manually inspected. Precursors that met structural miRNA criteria (55, 56) were selected and used to annotate potential melon *MIRNA loci*. Analyses were with standard Python scripts (http://www.python.org/) and the BioPython library (http://biopython.org/wiki/Main_Page). These analyses yielded 122 potential miRNAs (Dataset S3). Predicted precursors were also scanned against known non-coding RNA families using Infernal (v1.0.2) (47) and the Rfam database (v10.0) (48), and inspected for non-random secondary structure using Randfold (v 2.0) (57). To check whether potential miRNAs were expressed in melon tissues, a collection of expressed melon small RNAs (sRNAs) (58) was screened using local BLAST; expression of 87 potential miRNAs was identified in cotyledon, fruit and ovary tissues of melon (Dataset S3).

The Infernal (v1.0.2) analysis yielded a total of 53 potential miRNAs, of which 35 were also identified using the similarity search; therefore, the Infernal approach yielded 18 new potential miRNAs. These 18 new potential miRNAs did not show homology with plant miRNAs but with miRNAs discovered in non-plant species, and no expression data could be identified for them in melon sRNA expression libraries. They have, however, been annotated as potential miRNAs (Dataset S2, Table S13).

R-gene identification

Melon protein sequences were used as an input for the Disease Resistance Analysis and Gene Orthology (DRAGO) pipeline (59). This pipeline was used for the computational identification of novel disease resistance genes in melon and consists of a sequence homology search of the novel predicted proteins against a manually curated reference dataset of 96 plant resistance genes. This sequence homology search was performed using BLAST-p (49) with a very stringent cut-off e-value of 1E-10 to identify homologues to previously described plant R-genes.

Protein domains identified with InterProScan (21) in these potential melon R-genes were manually curated and assigned to one of the domain types previously associated with plant R-genes (59), such as leucine-rich repeats (LRR), nucleotide-binding sites (NBS), Toll/Interleukin-1 receptors (TIR), serine/threonine kinases (Ser-thr) and receptor-like kinases (KIN). Domains that could not be assigned to the typical resistance domains were labelled as 'other'. Bedtools v4 with a max-gap size between two genes of 20 Kb, was used for R-gene cluster analysis.

Duplications in the melon genome

For WSSD and WGAC analyses, we used the whole assembled genome. Melon specific transposons were masked and we also applied Tandem Repeat Finder (TRF) (60), with default parameters, masking an additional 8.23 Mb of sequence. For the depth of coverage analysis, 48,821,795 paired end 2 x 54 bp Illumina reads from DHL92 obtained in two lanes of a GAIIx instrument (see Genome assembly section, same reads used for homopolymer correction) were mapped to the assembly using mrFast (61) with edit distance 3.

Genome comparisons

Melon pseudomolecules were created from the anchored scaffolds after adding 50 Kb of Ns between scaffolds, and a new version of the gff annotation file with new coordinates was built. Alignment between melon and cucumber genomes was produced with SyMAP v3.4 (62) with the *promer* algorithm of the MUMer package (63). Using the gene prediction information, all SyMAP steps of anchor finder, chain finder postprocessor and alignment refinement were accomplished in order to produce the best synteny comparison at the DNA level. The expanded region in melon LG IV was computed with MAUVE v.2.3.1 with a LCB weight of 474 and the alignment algorithm *progressiveMauve* (64).

Additionally, a comparative genomic approach to compare melon with *Arabidopsis thaliana*, cucumber, strawberry and soybean was performed starting from PhylomeDb data (see phylogenomic analyses). The complex panorama of melon orthologs that show a *many-to-many* structure interaction was transformed to a *one-to-one* interaction in order to highlight gene connections without losing any information. Each interaction was enriched with gene information from the gff3 file of each genome, allowing the extraction of the orthology information between genes.

Based on the enriched dataset, synteny blocks were highlighted with a homemade pipeline using the BedTools (65) and the Circos (66) bundlelinks scripts, with a maximum gap of 10, 20 or 50 Kb between genes and a minimum of five genes for each block. Maximum gap option: adjacent links are merged into bundles if their start/end coordinates are sufficiently close. Given two links L1 and L2, they are merged into a bundle if: chr(start(L1)) == chr(start(L2)), chr(end(L1)) == chr(end(L2)), distance(start(L1), start(L2)) <= MAX_GAP, distance(end(L1), end(L2)) <= MAX_GAP.

The results of the synteny block analysis were used to draw a circular representation of synteny regions between melon and the four plant species. Using the same approach, paralogs of melon were extracted from PhylomeDb, transformed in a one-to-one dataset and used to highlight duplicated regions with a maximum gap of 10, 20 or 50 Kb between genes and a minimum of five genes for each block.

Resequencing

DHL92 and its parental lines SC and PS were sequenced using the Illumina GAIIx platform with 152 bp paired-end reads. Reads with a PHRED quality of at least 15 were selected for the analysis on DHL92 and both parental lines. Reads were mapped on the reference genome with BWA aligner software (67). The depth of coverage was measured with the GATK suite (68), SAMtools (6) and PICARD tool. Calling of variations of each sample was with SAMtools with a quality filter of PHRED 20. Error normalization, the comparison between resequenced samples and the selection of SNP and indel falling in genes and exons were performed with VCFtools (69). For error normalization, a new variation file was created, with all the common SNPs and indels between the two parental lines and the resequenced DHL92 line. All the called SNPs were subtracted from the three original variation files. The new variation files were used for all the analyses.

Data management and GBrowser

The Melonomics web site (http://melonomics.net) includes the genome structure and functional annotation. It is powered by the Generic Model Organism Database (GMOD, http://gmod.org) tools. A Chado database (70) integrates the genome information while genome and genetic map browsers are powered by GBrowse (71) and CMap (72), respectively. A Django (https://www.djangoproject.com/) application allows the research community to search and browse the Chado structured genome database using a web interface.

SI TABLES

Table S1. Sequences used for assembly of the melon genome

Library type	Total reads	Total nucleotides	True PEs (%)	Coverage (450 Mb genome)	Assembled (%)*
454, Shotgun	14,862,049	5,199,327,409	-	11.55	78.62
454, 3kb PEs	3,599,380	1,276,524,069	39.9	2.84	71.26
454, 8kb PEs	3,015,490	1,104,336,255	56.1	2.45	69.81
454, 20kb PEs	1,156,631	386,905,948	50.2	0.86	76.38
Sanger, BES	53,203	26,206,124	50.0	0.06	82.67
Total	22,686,753	7,993,299,805	-	17.76	76.13

***Most non-assembled reads corresponded to chloroplast and mitochondrion**

PEs: paired-ends; BES: BAC-end sequences

Table S2. N scaffold size and N index of the melon assembly

% assembly	N scaffold size in nt (N index)
10	9,058,246 (4)
20	7,701,838 (8)
30	6,690,008 (13)
40	5,729,086 (19)
50	4,677,790 (26)
60	4,065,872 (35)
70	3,261,579 (44)
80	2,255,220 (58)
90	1,485,533 (78)

Table S3. Assembly comparison with other whole genome shotgun plant genome assemblies

Species	Genome size (Mb)	N50 Scaffold index	N50 scaffold size (Mb)	# scaffolds	N50 contig size (Kb)	Sequencing technology
Melon	450	26	4.68	1,594	18.2	454, Sanger
Potato[73]	844	121	1.78	2,043	31.4	Illumina, 454, Sanger
Apple[74]	743	102	1.54	1,629	13.4	Sanger, 454
Fragaria[75]	240	n.a.	1.36	3,263	n.a.	454, Illumina, SOLiD
Cucumber[76]	367	59	1.14	47,837	19.8	Illumina, Sanger
Brassica rapa[77]	529	n.a.	1.97	n.a.	27.3	Illumina
Cacao[78]	430	178	0.47	4,792	19.8	454
Date palm[79]	658	n.a.	0.03	57,277	6.4	Illumina
Soybean[80]	1,115	10	47.8	1,168	189.4	Sanger
Papaya[81]	372	n.a.	1	17,764	11	Sanger

Table S4. Finished melon BACs used to assess the quality of the assembly

BAC		Assembly		Mapped[a]			Not Mapped		
Name	Length (bp)	Scaffold	Length (bp)	BAC (bp)	Assembly (bp)	Translocation events	BAC (bp)	Assembly (bp)	Assembly gaps (bp)
60K17	116,877	Scaffold00003	118,955	104,086	104,221	0	12,791	14,734	14,573
31O16	159,477	Scaffold00051	161,672	143,822	143,909	0	15,655	17,763	17,763
13J4	98,716	Scaffold00003	104,699	94,11	94,183	2	4,606	10,516	10,193
1-21-10	92,343	Scaffold00001	104,672	90,239	90,163	0	2,104	14,519	10,863
TOTAL	467,413		489,998	432,257	432,476	2	35,156	57,532	53,392

[a]Mapped sequence is given as bps belonging to non-overlapped alignments with a high degree of identity (99% for the first 3 BACs that belong to the same individual as the sequence of the assembly, and 97% for BAC 1-21-10)

Table S5. Features of major gaps in the assembly

BAC				Assembly		
Name	Coordinates	Length (bp)	Features	Coordinates	Length (bp)	Num gaps (bp)
60K17	54,504-64,536	10,033	CURE RTP [1]	Scaf00003:5745223-5756387	11,165	1 (11,165)
31O16	90,174-102,184	12,011	CUMULE TP[2]	Scaf00051:1751999-1763600	11,602	1 (11,602)
13J4	89,198-94,572	5,375	RTP: 79,471-91,152[3] MRGH partial gene: 91,486-98716[3]	Scaf00003:5587880-5600462	12,583	4 (9,495)
1-21-10	55,072-58,881	3,809	Close to region of high simple sequence repeats (SSR) content[3]	Scaf00001:5648045-5664882	16,837	5 (10,663)

[1](ref 7), [2](ref 8), [3](ref 9)

Table S6. Scaffolds from BAC pools in the melon assembly

BAC[a]		**Assembly**		**Mapped[b]**		**Not mapped[c]**					
						gaps in BAC			**no gaps in BAC**		
Length (bp)	**Gaps (bp)**	**Length (bp)**	**Gaps (bp)**	**BAC (bp)**	**Assembly (bp)**	**BAC (bp)**	**Assembly (bp)**	**Gaps in assembly (bp)**	**BAC (bp)**	**Assembly (bp)**	**Gaps in assembly (bp)**
5,934,931	298,558	6,110,883	386,497	5,479,456	5,485,565	403,444	471,521	373,87	52,031	153,797	110,438

[a]Scaffolds MRGH63 and 00041 have been excluded from this table as they have translocation events. [b]Mapped sequence refers to BAC regions that align to the assembly with more than 98% similarity, while small indels are allowed (only 3 alignments in 2 scaffolds have less than 99% similarity). [c]The remaining sequence in the BAC are separated in contiguous regions, including a BAC gap and the rest, and the bps corresponding to these intervals in the assembly and how many belong to assembly gaps, counted.

Table S7. Anchoring of the genome assembly to the SC × PS SNP genetic map

Linkage group	**Scaffolds anchored**[*]	**Genome anchored (bp)**	**% assembled genome anchored**	**Oriented scaffolds**	**Genome oriented (bp)**	**% assembled genome oriented**
LG I	12	35,236,718	9.75	9	29,849,588	8.26
LG II	7	24,585,017	6.80	7	24,585,017	6.80
LG III	6	26,575,343	7.35	5	25,924,201	7.17
LG IV	8	30,080,766	8.32	6	28,210,071	7.81
LG V	5	28,403,756	7.86	5	28,403,756	7.86
LG VI	7	29,566,564	8.18	7	29,566,564	8.18
LG VII	8	24,823,955	6.87	7	24,414,097	6.76
LG VIII	7	24,129,498	6.68	4	19,681,835	5.45
LG IX	7	24,101,567	6.67	6	21,669,403	6.00
LG X	5	16,254,367	4.50	3	13,481,340	3.73
LG XI	10	27,576,509	7.63	7	21,160,428	5.85
LG XII	5	24,971,993	6.91	5	24,971,993	6.91
Total	87	316,306,053	87.52	71	291,918,293	80.77

[*] **86 scaffolds and 1 contig**

Table S8. Transposon content in the melon genome assembly

type	**superfamily**	**% of genome**
LTR retrotransposons	copia	5,5
	gypsy	7,2
non-LTR retrotransposons		0,1
retrotransposon related sequences		1,9
Total retrotransposons		**14,7**
DNA transposons	CACTA	1,6
	hAT	0,1
	Mariner	0,1
	MULE	1,9
	PIF	0,3
	Helitron	0,06
transposon related sequences		0,8
Total DNA transposons		**5**

Table S9. Comparison of the prevalence of the major DNA transposon superfamilies in melon and cucumber

		Cucumber		Melon			
superfamily	% query length	# copies	% genome	# copies	% genome	% genome fold difference (melon/cucumber)	# copy fold difference (melon/cucumber)
CACTA	90	13	0.015	116	0.102		
	20-90	169	0.102	1756	0.991		
	0-20	374	0.046	5114	0.536	× 10	× 12.5
MULE	90	3	0.002	112	0.102		
	20-90	73	0.032	1947	1.132		
	0-20	66	0.007	7681	0.702	× 47	× 68
PIF	90	23	0.027	10	0.016		
	20-90	55	0.034	456	0.18		
	0-20	215	0.02	2524	0.116	× 3.8	× 10
unclassified			0.74		0.83	× 1.1	
TOTAL			1.028		4.707	× 4.6	

Table S10. Gene prediction statistics

Genome	Genome size (bp)*	375,485,313
	Genome GC content (%)	33.2
Genes	Number of genes	27,427
	Mean gene length	2,776
	Median gene length	1,799
	Total genic length	76,125,905
	Gene length range	150 … 86,198
	Gene density (kb/gene)	13.69
Transcripts	Number of transcripts	34,848
	Mean transcript length	1,251
	Median transcript length	1,131
	Total transcript length	43,587,448
	Transcript length range	78 … 14,899
	Transcripts per gene	1.3
	Exons per transcript	4.6
	Introns per transcript	3.6
	Multi-exonic transcripts (%)	71.4
Exons	Number of exons	160,598
	Mean exon length	271
	Median exon length	158
	Exon length range	2 ... 6,054
	Exon GC content	42.1
	Number of coding exons	151,083
Introns	Number of introns	125,750
	Mean intron length	506
	Median intron length	194
	Intron length range	21 … 76,916
	Intron GC content	33.2
	Number of U12 introns	226
	U12 introns (GT-AG/AT-AC)	176/50
CDS	Number of CDS	34,848
	Unique CDS (polypeptides)	32,487
	Mean CDS length	950
	Median CDS length	762
	Total cds length	33,090,116
	CDS length range	78 … 14,289
UTRs	Number of transcripts with UTRs	22,163
	Mean 5' + 3' UTR length	454
	Median 5' + 3' UTR length	414
	Total 5' + 3' UTR length	10,056,773
	5' + 3' UTR range	2 … 3,827
	Number of transcripts with 5' UTRs	18,662
	Mean 5' UTR length	199
	Median 5' UTR length	155
	Total 5' UTR length	3,722,051
	5' UTR range	2 … 2,368
	Number of transcripts with 3' UTRs	19,153
	Mean 3' UTR length	331
	Median 3' UTR length	290
	Total 3' UTR length	6,334,722
	3' UTR range	2 … 3,749

***Genome size is slightly different from the assembly v3.3 due to homopolymer correction and correction of misassembles in scaffolds. Data is from assembly v3.5**

Table S11. Number of protein signatures identified by Interproscan for each of the InterPro member databases

InterPro member database	**Count**
BlastProDom	445
HAMAP	533
HMMPIR	1,077
Coil	6,819
PatternScan	7,708
HMMSmart	16,841
ProfileScan	18,446
FPrintScan	20,335
Superfamily	23,950
Gene3D	25,203
HMMPanther	29,681
HMMPfam	34,758
Seg	52,709

Table S12. Species used for the phylome analysis

TaxaID	**Code**	**Species Name**	**Source**	**As of**	**Genes**	**Unique isoforms**[1]
3656	CUCME	*Cucumis melo*	melonomics.net	01/04/11	27,427	27,382
4932	YEAST	*Saccharomyces cerevisiae*	Quest for orthologs 02/2011	01/03/11	5,813	5,813
5833	PLAFA	*Plasmodium falciparum*	Quest for orthologs 02/2011	01/03/11	5,044	5,044
6239	CAEEL	*Caenorhabditis elegans*	Quest for orthologs 04/2011	01/07/11	19,758	19,758
7227	DROME	*Drosophila melanogaster*	Quest for orthologs 04/2011	01/07/11	13,074	13,074
9606	HUMAN	*Homo sapiens*	Quest for orthologs 04/2011	01/07/11	20,988	20,988
3055	CHLRE	*Chlamydomonas reinhardtii*	Phytozome v7.0	01/07/11	17,114	16,941
38833	MICPS	*Micromonas pusilla*	JGI	01/07/11	10,545	10,526
70448	OSTTA	*Ostreococcus tauri*	Uniprot	01/07/11	7,933	7,933
436017	OSTLU	*Ostreococcus lucimarinus*	Uniprot	01/07/11	7,402	7,402
3218	PHYPA	*Physcomitrella patens*	Phytozome v7.0	01/07/11	32,273	31,934
88036	SELML	*Selaginella moellendorffii*	Phytozome v7.0	01/07/11	22,285	22,138
4558	SORBI	*Sorghum bicolor*	Phytozome v7.0	01/07/11	27,608	27,502
4577	MAIZE	*Zea mays*	maizesequence.org	01/07/11	39,656	39,012
15368	BRADI	*Brachypodium distachyon*	Phytozome v7.0	01/07/11	25,532	25,479
39946	ORYSI	*Oryza sativa subsp. indica*	ENSEMBL - Plants	01/07/11	40,745	40,548

39947	ORYSJ	*Oryza sativa subsp. japonica*	Phytozome v7.0	01/07/11	40,869	40,539
3641	THECC	*Theobroma cacao*	CocoaGen DB	01/03/11	46,143	46,125
3659	CUCSA	*Cucumis sativus*	Phytozome v7.0	01/07/11	21,646	21,552
3694	POPTR	*Populus trichocarpa*	Phytozome v7.0	01/07/11	40,668	40,471
3702	ARATH	*Arabidopsis thaliana*	Quest for orthologs 04/2011	01/07/11	26,628	26,628
3847	SOYBN	*Glycine max*	Phytozome v7.0	01/07/11	46,367	46,144
3880	MEDTR	*Medicago truncatula*	Phytozome v7.0	01/07/11	50,962	48,777
3988	RICCO	*Ricinus communis*	Phytozome v7.0	01/07/11	31,221	31,140
4155	MIMGU	*Mimulus guttatus*	Phytozome v7.0	01/07/11	26,718	26,578
29760	VITVI	*Vitis vinifera*	Phytozome v7.0	01/07/11	26,346	26,278
59689	ARALY	*Arabidopsis lyrata*	Phytozome v7.0	01/07/11	32,670	32,233
101020	FRAVE	*Fragaria vesca subsp. vesca*	strawberrygenome.org	01/03/11	34,809	34,775

[1]When the longest isoforms of two or more genes were identical at the protein level, a single one was included

Table S13. Non-coding RNA gene classes in *C. melo* and *A. thaliana*, including transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA), small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA)

	***A. thaliana* (TAIR10)**	***C. melo***
tRNA	689	685
rRNA	15	17
miRNA	177	140
snRNA	13	103
snoRNA	71	143
other RNA	394	165

Table S14. Clusters of ncRNA genes identified in the melon genome using a 3 Kb window

Type	**Members**	**Melon scaffold**	**Start**	**End**
miRNA	4	CM3.5_contig33941	353	1661
tRNA	3	CM3.5_scaffold00001	4533466	4535522
snor16/SNORD43/tRNA	3	CM3.5_scaffold01596	5656831	5659212
intron_gpII/tRNA/SSU_rRNA_5	6	CM3.5_scaffold00003	1805105	1812729
snoR83/tRNA	4	CM3.5_scaffold00003	4819193	4821782
5S_rRNA	4	CM3.5_scaffold00004	4950553	4951847
tRNA	4	CM3.5_scaffold00006	1608090	1609323
miRNA/tRNA	3	CM3.5_scaffold00007	523669	528321
5_8S_rRNA/miRNA/SSU_rRNA_5	3	CM3.5_scaffold00008	7386314	7388299
miRNA	3	CM3.5_scaffold00016	398312	400283
miRNA	4	CM3.5_scaffold00016	408411	410566
tRNA	6	CM3.5_scaffold00022	1503563	1507120
U1	3	CM3.5_scaffold00025	1257281	1261357
tRNA	3	CM3.5_scaffold00026	4510020	4512311
IsR/tRNA	3	CM3.5_scaffold00030	2236832	2239850

tRNA	6	CM3.5_scaffold00031	3700936	3705960
tRNA	10	CM3.5_scaffold00034	869909	879132
SNORD36/snoR80/snoR11/snoZ157	4	CM3.5_scaffold00037	868203	869111
5S_rRNA/tRNA	3	CM3.5_scaffold00037	1024879	1025930
tRNA	3	CM3.5_scaffold00047	597794	601979
U5	3	CM3.5_scaffold00063	122030	125784
SSU_rRNA_5/5_8S_rRNA	3	CM3.5_scaffold00064	12687	16090
snoZ266/snoz267/snoR44_J54	3	CM3.5_scaffold00068	988544	989272
5S_rRNA	5	CM3.5_scaffold00110	156	1373
tRNA	3	CM3.5_scaffold00304	837	1448
intron_gpII/tRNA	3	CM3.5_scaffold01301	195	4215

Table S15. Disease resistance genes identified in melon

R-protein type	Class	**Melon**	**Arabidopsis**	**Grape**	**Rice**
Cytoplasmic classes					
CC-NBS-LRR	**CNL**	21	40	60	402
TIR-NBS-LRR	**TNL**	21	97	19	0
NBS-LRR	NL	10	11	111	74
RPW8-NBS-LRR	RPW8-NL	3	6	10	1
TIR-NBS-LRR-NBS	TNLN	1	0	0	0
CC-NBS	CN	11	2	74	53
NBS	N	4	4	18	16
TIR	T	6	38	7	2
TIR-NBS	TN	4	14	3	0
Cytoplasmic classes subtotal		81	212	302	548
Transmembrane classes					
RLK	**RLK**	161 (170)	222	219	394
KIN-GNK2	RLK-GNK2	19 (21)	1	19	48
RLP	**RLP**	110 (115)	91	150	216
Transmembrane classes subtotal		**290**	**314**	**388**	**658**
Other					
MLO-like		15 (18)	19	17	17
PTO-like		25 (29)	1	0	7
Total		**411**	**526**	**690**	**1206**

Total n° of genes	**411**
Total n° of proteins	**434**

Table S16. Number of syntenic blocks found within the melon genome

	Gene interval[*] 10 Kb	20 Kb	50 Kb
No of syntenic blocks	0	2	21
No of gene pairs involved	0	28	423

[*]**Blocks of five or more genes with a 10, 20 or 50 Kb interval between each gene were considered**

Table S17. Summary statistics of duplication analysis on the melon assembly. Number of duplications, number of duplicated base pairs, percentage over the genome and percentage over the genome belonging to scaffolds larger than 10 Kb, assessed by the different methods: WGAC; WGAC filtered by duplications having 94% identity and a minimum length of 10 Kb; WSSD; and the overlap between the two methods for at least 1 base pair or a reciprocal overlap of 50% or more.

	WGAC (90%; >1Kb)	**WGAC (94%; >10Kb)**	**WSSD (94%; >10Kb)**	**Overlap (at least 1 bp)**	**Overlap (at least 50%)**
Nº duplications	12,829	308	856	99	62
Nº duplicated bp	38.08 Mb	4.37 Mb	12.66 Mb	1.03 Mb	0.79 Mb
% over genome (375.48 Mb)	10.14%	1.16%	3.37%	0.27%	0.21%
% over genome > 10 Kb (357.47 Mb)	---	1.22%	3.54%	0.29%	0.22%

Table S18. Relationships between melon and cucumber chromosomes. In brackets, shorter alignments between chromosomes are indicated.

	Cucumber chromosome	
Melon chromosome	Li et al 2011	This study
I	7	7, (2)
II	1	1, (5)
III	2, 6	2, 6
IV	3	3, (6)
V	2	2, (3, 6)
VI	3	3
VII	4	4, (2, 5)
VIII	4, 6	4, 6
IX	5	5
X	5	5
XI	2, 6	2, 6
XII	1	1, (6)

	Melon chromosome	
Cucumber chromosome	Li et al 2011	This study
1	II, XII	II, XII
2	III, V, XI	III, V, XI, (I, VII)
3	IV, VI	IV, VI, (V)
4	VII, VIII	VII, VIII
5	IX, X	IX, X, (II, VII)
6	III, VIII, XI	III, VIII, XI, (IV, V, XII)
7	I	I

Table S19. Synteny blocks identified between melon and other plant genomes

Species	nº of syntenic blocks*			nº of melon orthologues	nº of other species orthologues	nº of one-to-one interactions	nº of OrthoTree
	10 Kb	20 Kb	50 Kb				
Cucumber	850	643	497	16,617	16,141	19,377	17,603
Arabidopsis	56	234	638	12,501	14,710	23,746	13,104
Soybean	91	522	1,402	14,555	29,658	44,068	15,018
Strawberry	111	336	663	13,524	13,048	22,198	14,221

***Blocks of five or more genes at a 10, 20 or 50 Kb interval between each gene were considered**

Table S20. Number of reads and depth of coverage after Illumina resequencing

Sample	**Pair-end**	**nº reads**	**Total nº of reads**	**Depth of coverage**
DHL92	Forward	35,538,240	71,076,480	17.32
	Reverse	35,538,240		
T111 (PS)	Forward	35,857,911	71,715,822	17.47
	Reverse	35,857,911		
PI 161375 (SC)	Forward	35,233,293	70,466,586	16.72
	Reverse	35,233,293		

Table S21. SNPs and indels identified in the DHL92 reference genome and the PI 161375 and PS parental lines

	SNP	**Indel**	**Total**	**SNP in gene space**	**SNP in exon**	**Indel in gene space**	**Indel in exon**
PI 161375 (SC)	851,562	183,267	1,034,829	107,878	37,431	26,794	5,555
T111 (PS)	1,276,319	229,687	1,506,006	149,739	49,684	34,695	7,216
TOTAL	2,127,881	412,954	2,540,835	257,617	87,115	61,489	12,771

SI FIGURES

Figure S1. Alignments of the four Sanger-finished BACs against the genome assembly. BACs and assembly are plotted; lines between them in their respective coordinates define the ends of the alignments. In the assembly, alignments are coloured depending on their similarity and the strand is shown with an arrow. Transposons, tandem repeats, segmental duplications and gaps annotated in the assembly are also represented.

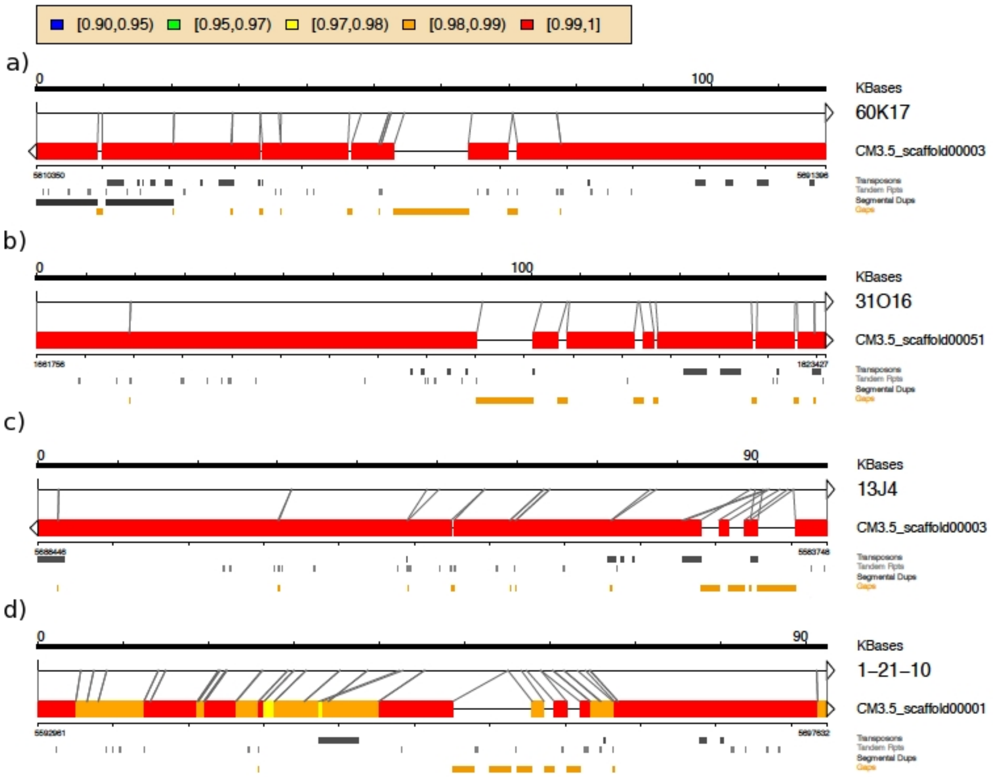


Figure S2. Anchoring of the melon genome assembly to the SC × PS genetic map. Red bars represent the 12 melon linkage groups; SNPs are located according to genetic distance (cM). Melon genome scaffolds were positioned in each linkage group with corresponding genetic markers. Blue, scaffolds in positive orientation; green, scaffolds with negative orientation (reverse and complemented); yellow, scaffolds that were anchored but not oriented. Asterisks represent positions where multiple SNPs co-located and a single one is represented in the figure.

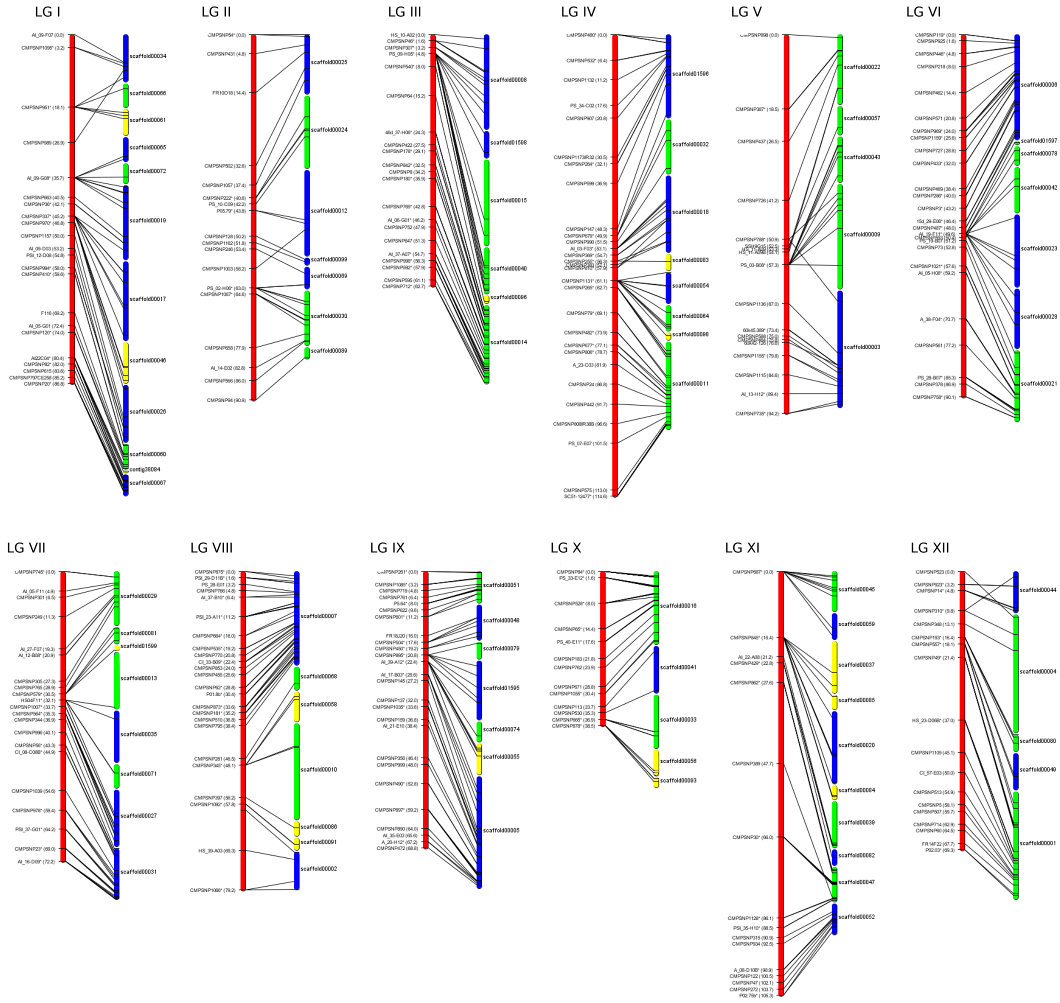


Figure S3. Ratio between genetic and physical distance. The ratio between genetic and physical distances was based on the anchored genome to the 12 melon pseudochromosomes. For each marker in the SC × PS genetic map, the genetic distance is according to its physical position in the genome assembly. Genetic distance is expressed in cM and physical distance in Mb.

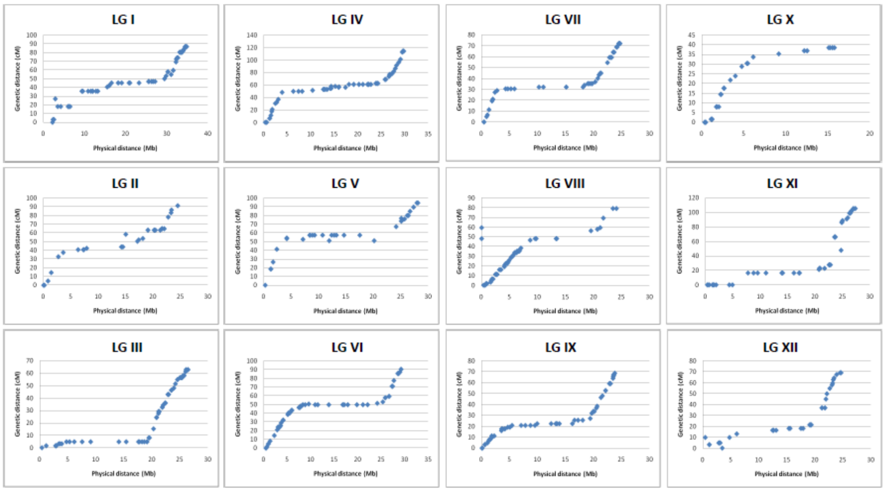


Figure S4. LTR retrotransposon insertion during melon genome evolution. Selected examples of families illustrating varied expansion patterns are displayed. Indicated above each graph is the name of the family's representative, and the number of copies with sufficiently complete LTRs to be dated.

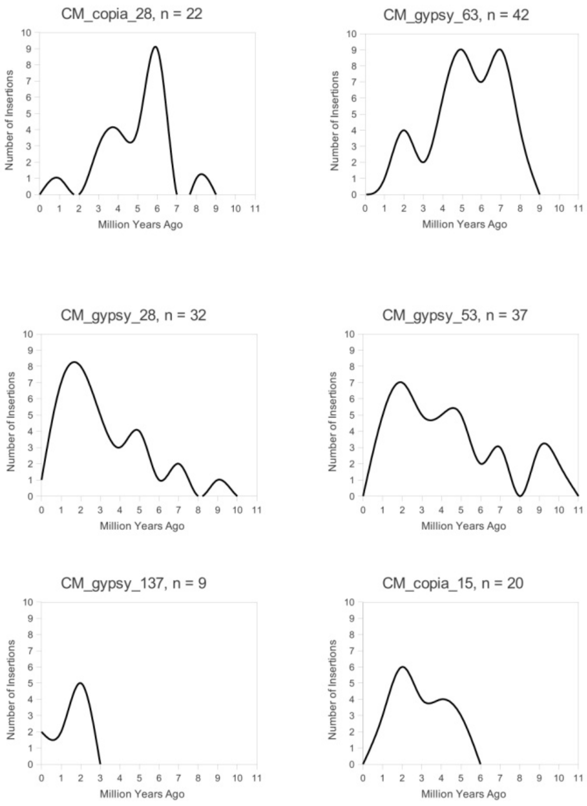


Figure S5. Potential clusters identified in the melon genome for *MIRNA169* family members. (a) Localization of 12 potential *MIRNA*169 *loci* (blue dots) in the sequence of melon scaffold00016 (black line). Eight *loci* were found in pairs (vertical arrows, a1-a4) in a region of less than 300 bases potentially included in a polycistronic transcript. (**b**) RNA secondary structure representation of miRNA transcripts from the potential clusters shown in (a). *MIRNA169* sequences are indicated by red lines.

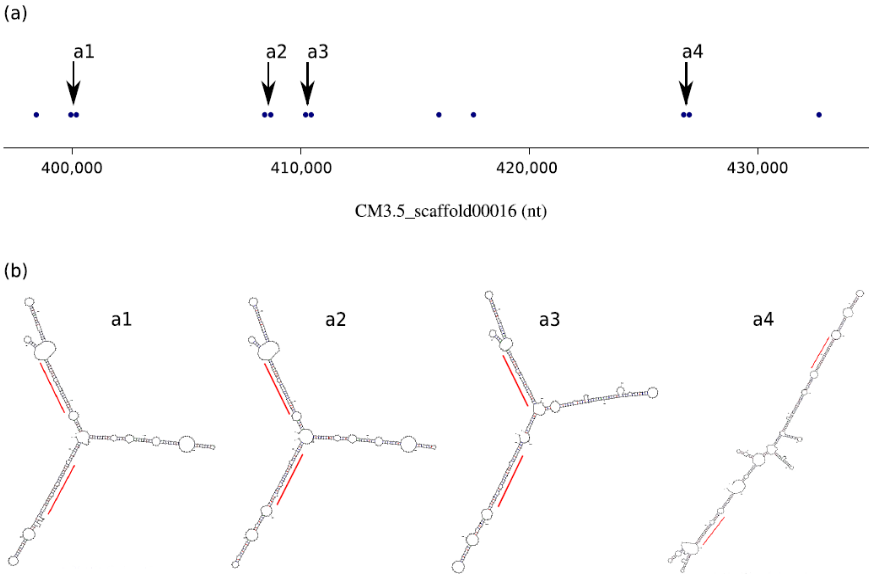


Figure S6. Disease resistance gene clusters in the melon genome. R-genes were non-randomly distributed in the melon genome, organized in clusters. 79 R-genes were located within 19 genomic clusters, 16 with genes belonging to the same family. For each scaffold anchored to the genetic map, each type of R-gene is represented with color-coded bars. The VAT and Fom-1 regions are highlighted in LG V and IX, respectively. CNL: CC-NBS-LRR; NL: NBS-LRR; RPW8: Rpw8-like; TNL: TIR-NBS-LRR; TNL-NBS: TIR-NBS-LRR-NBS; TN: TIR-NBS; TIR: Toll-IL-1 receptor; NBS: nucleotide binding site motif; CN: CC-NBS; PTO: Pto-like; RLK-GNK2: RLK from the GNK2 type; RLP: receptor-like protein; RLK: receptor-like kinase; CC: coiled-coil motif; LRR: leucine-rich repeat.

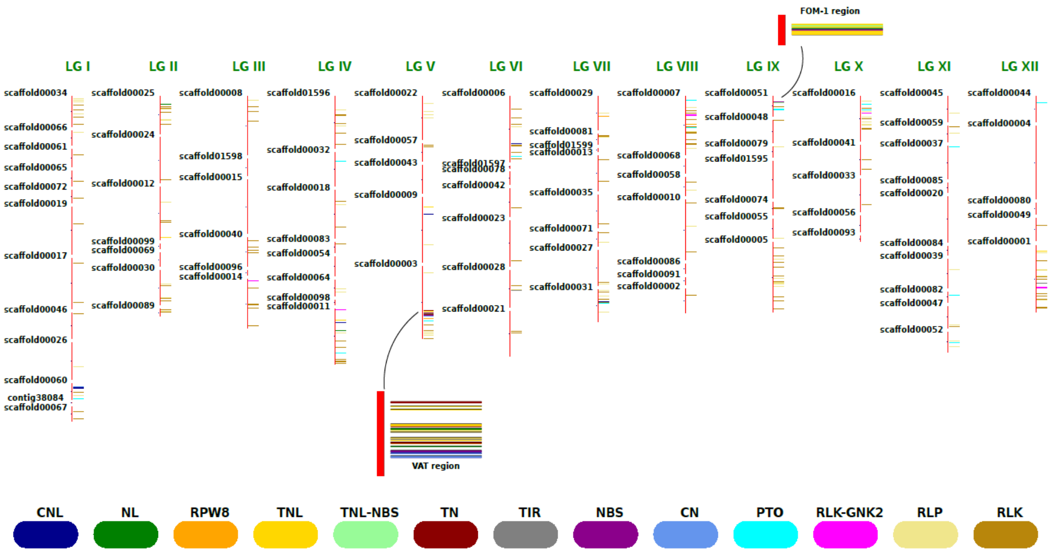


Figure S7. Phylogenetic tree representing the expansion of R-genes from the receptor-like kinases (RLK) type. Blue and red circles indicate speciation and duplication events, respectively. Gray circles indicate collapsed leaves. Three clusters of RLK genes are represented as A (containing two genes), B (containing three genes) and C (containing four genes). The structure and orientation of the genes in the corresponding melon scaffolds is represented below (distance in Kb). Each gene in the cluster is named according to the melon genome annotation.

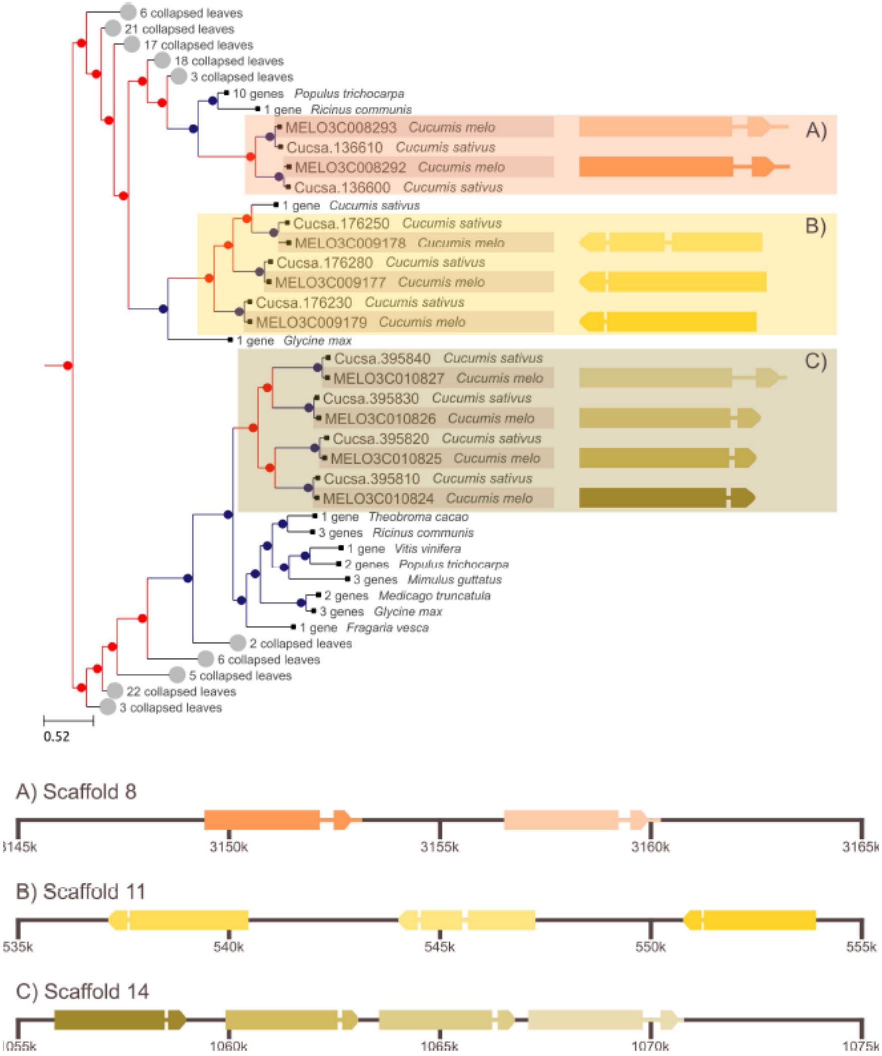

6 collapsed leaves
21 collapsed leaves
17 collapsed leaves
18 collapsed leaves
3 collapsed leaves
10 genes *Populus trichocarpa*
1 gene *Ricinus communis*
MELO3C008293 *Cucumis melo*
Cucsa.136610 *Cucumis sativus*
MELO3C008292 *Cucumis melo*
Cucsa.136600 *Cucumis sativus*
A)
1 gene *Cucumis sativus*
Cucsa.176250 *Cucumis sativus*
MELO3C009178 *Cucumis melo*
Cucsa.176280 *Cucumis sativus*
MELO3C009177 *Cucumis melo*
Cucsa.176230 *Cucumis sativus*
MELO3C009179 *Cucumis melo*
B)
1 gene *Glycine max*
Cucsa.395840 *Cucumis sativus*
MELO3C010827 *Cucumis melo*
Cucsa.395830 *Cucumis sativus*
MELO3C010826 *Cucumis melo*
Cucsa.395820 *Cucumis sativus*
MELO3C010825 *Cucumis melo*
Cucsa.395810 *Cucumis sativus*
MELO3C010824 *Cucumis melo*
C)
1 gene *Theobroma cacao*
3 genes *Ricinus communis*
1 gene *Vitis vinifera*
2 genes *Populus trichocarpa*
3 genes *Mimulus guttatus*
2 genes *Medicago truncatula*
3 genes *Glycine max*
1 gene *Fragaria vesca*
2 collapsed leaves
6 collapsed leaves
5 collapsed leaves
22 collapsed leaves
3 collapsed leaves
0.52
A) Scaffold 8
3145k
3150k
3155k
3160k
3165k
B) Scaffold 11
535k
540k
545k
550k
555k
C) Scaffold 14
1055k
1060k
1065k
1070k
1075k

Figure S8. Phylogenetic tree of the UDP-glc phyrophosphorylase gene family. A second gene putatively encoding a UDP-glc phyrophosphorylase (CmUGP-LIKE1, Phy003A737_CUCME), for which a single gene has been described (CmUGP, Phy0039Z88_CUCME) is also shown. Both genes are highlighted. Blue and red circles in the tree represent speciation and duplication events, respectively. Gray circles represent collapsed leaves. Each gene in the cluster is named according to phylome.

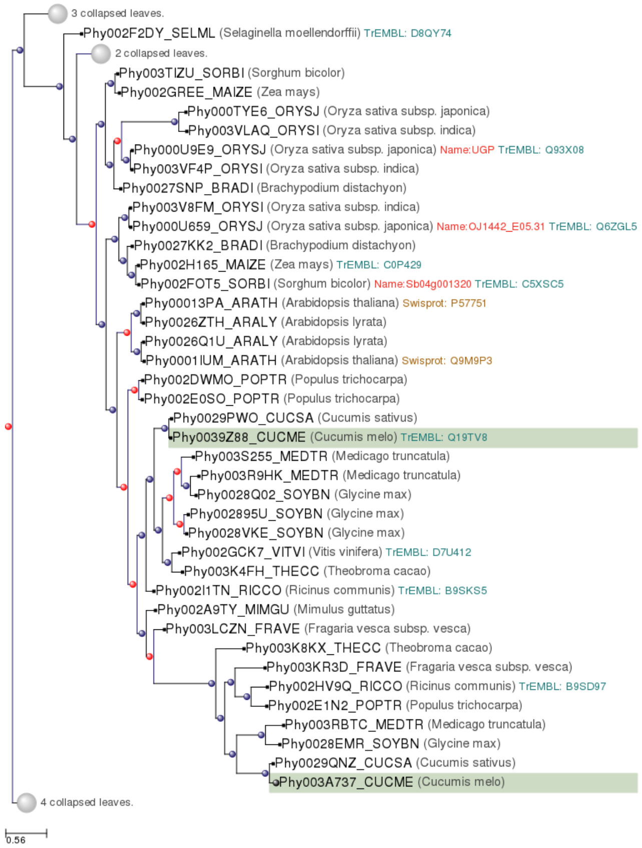

3 collapsed leaves.
Phy002F2DY_SELML (Selaginella moellendorffii) TrEMBL: D8QY74
2 collapsed leaves.
Phy003TIZU_SORBI (Sorghum bicolor)
Phy002GREE_MAIZE (Zea mays)
Phy000TYE6_ORYSJ (Oryza sativa subsp. japonica)
Phy003VLAQ_ORYSI (Oryza sativa subsp. indica)
Phy000U9E9_ORYSJ (Oryza sativa subsp. japonica) Name:UGP TrEMBL: Q93X08
Phy003VF4P_ORYSI (Oryza sativa subsp. indica)
Phy0027SNP_BRADI (Brachypodium distachyon)
Phy003V8FM_ORYSI (Oryza sativa subsp. indica)
Phy000U659_ORYSJ (Oryza sativa subsp. japonica) Name:OJ1442_E05.31 TrEMBL: Q6ZGL5
Phy0027KK2_BRADI (Brachypodium distachyon)
Phy002H165_MAIZE (Zea mays) TrEMBL: C0P429
Phy002FOT5_SORBI (Sorghum bicolor) Name:Sb04g001320 TrEMBL: C5XSC5
Phy00013PA_ARATH (Arabidopsis thaliana) Swisprot: P57751
Phy0026ZTH_ARALY (Arabidopsis lyrata)
Phy0026Q1U_ARALY (Arabidopsis lyrata)
Phy0001IUM_ARATH (Arabidopsis thaliana) Swisprot: Q9M9P3
Phy002DWMO_POPTR (Populus trichocarpa)
Phy002E0SO_POPTR (Populus trichocarpa)
Phy0029PWO_CUCSA (Cucumis sativus)
Phy0039Z88_CUCME (Cucumis melo) TrEMBL: Q19TV8
Phy003S255_MEDTR (Medicago truncatula)
Phy003R9HK_MEDTR (Medicago truncatula)
Phy0028Q02_SOYBN (Glycine max)
Phy002895U_SOYBN (Glycine max)
Phy0028VKE_SOYBN (Glycine max)
Phy002GCK7_VITVI (Vitis vinifera) TrEMBL: D7U412
Phy003K4FH_THECC (Theobroma cacao)
Phy002I1TN_RICCO (Ricinus communis) TrEMBL: B9SKS5
Phy002A9TY_MIMGU (Mimulus guttatus)
Phy003LCZN_FRAVE (Fragaria vesca subsp. vesca)
Phy003K8KX_THECC (Theobroma cacao)
Phy003KR3D_FRAVE (Fragaria vesca subsp. vesca)
Phy002HV9Q_RICCO (Ricinus communis) TrEMBL: B9SD97
Phy002E1N2_POPTR (Populus trichocarpa)
Phy003RBTC_MEDTR (Medicago truncatula)
Phy0028EMR_SOYBN (Glycine max)
Phy0029QNZ_CUCSA (Cucumis sativus)
Phy003A737_CUCME (Cucumis melo)
4 collapsed leaves.
0.56

Figure S9. Phylogenetic tree of the cell-wall invertase gene family. A new cell-wall invertase (CmCIN-LIKE1, Phy003LKZZ_CUCME) was annotated, probably resulting from the duplication of CmCIN2 (Phy003A1LG_CUCME) in the ancestor of melon and cucumber. Both genes are highlighted. Blue and red circles in the tree represent speciation and duplication events, respectively. Gray circles represent collapsed leaves. Each gene in the cluster is named according to phylome.

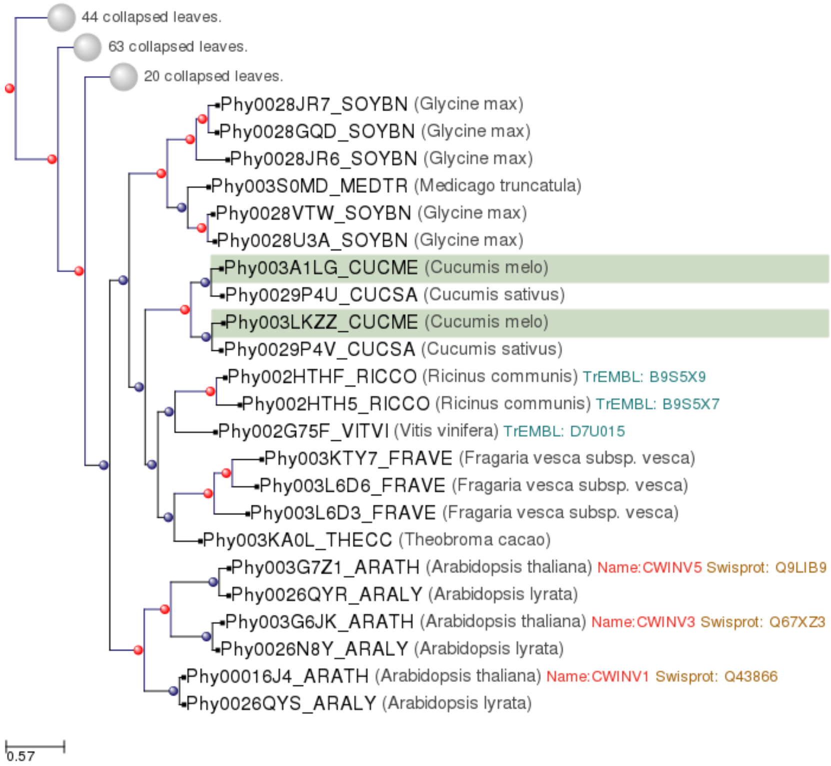

44 collapsed leaves.
63 collapsed leaves.
20 collapsed leaves.
Phy0028JR7_SOYBN (Glycine max)
Phy0028GQD_SOYBN (Glycine max)
Phy0028JR6_SOYBN (Glycine max)
Phy003S0MD_MEDTR (Medicago truncatula)
Phy0028VTW_SOYBN (Glycine max)
Phy0028U3A_SOYBN (Glycine max)
Phy003A1LG_CUCME (Cucumis melo)
Phy0029P4U_CUCSA (Cucumis sativus)
Phy003LKZZ_CUCME (Cucumis melo)
Phy0029P4V_CUCSA (Cucumis sativus)
Phy002HTHF_RICCO (Ricinus communis) TrEMBL: B9S5X9
Phy002HTH5_RICCO (Ricinus communis) TrEMBL: B9S5X7
Phy002G75F_VITVI (Vitis vinifera) TrEMBL: D7U015
Phy003KTY7_FRAVE (Fragaria vesca subsp. vesca)
Phy003L6D6_FRAVE (Fragaria vesca subsp. vesca)
Phy003L6D3_FRAVE (Fragaria vesca subsp. vesca)
Phy003KA0L_THECC (Theobroma cacao)
Phy003G7Z1_ARATH (Arabidopsis thaliana) Name:CWINV5 Swisprot: Q9LIB9
Phy0026QYR_ARALY (Arabidopsis lyrata)
Phy003G6JK_ARATH (Arabidopsis thaliana) Name:CWINV3 Swisprot: Q67XZ3
Phy0026N8Y_ARALY (Arabidopsis lyrata)
Phy00016J4_ARATH (Arabidopsis thaliana) Name:CWINV1 Swisprot: Q43866
Phy0026QYS_ARALY (Arabidopsis lyrata)
0.57

Figure S10. **Comparative analysis of the melon genome with other sequenced plant genomes based on the orthologous genes identified in the phylome analysis.** The 12 melon pseudochromosomes are shown in different colours, each block representing an anchored scaffold. Synteny blocks are represented between melon and (a) cucumber, (b) diploid strawberry, (b) *Arabidopsis thaliana* and (c) soybean. Duplicated blocks of five or more genes separated by up to 50 Kb (cucumber) or 20 Kb (strawberry, *Arabidopsis*, soybean) are represented.

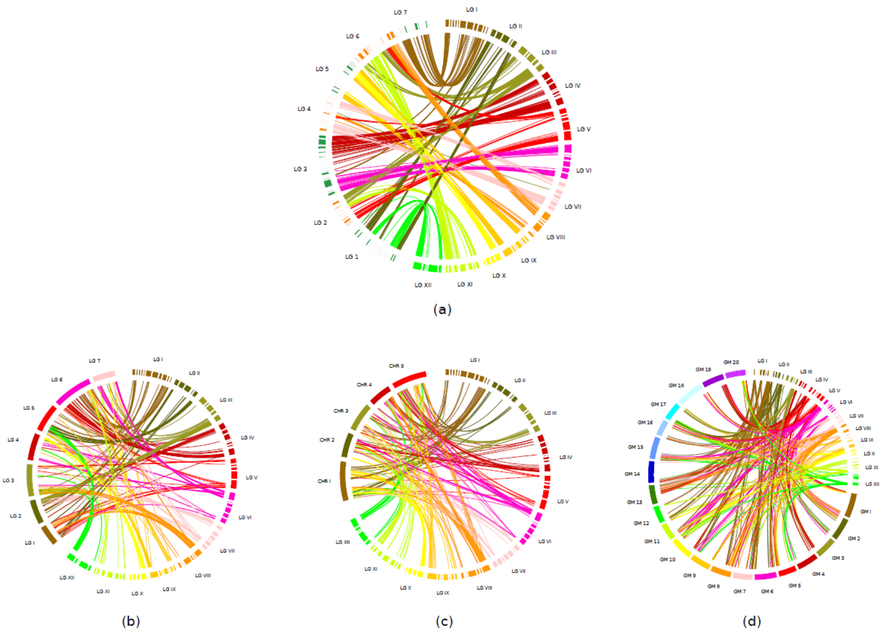


Figure S11. Reconstructed genome structure of DHL92 based on the parental lines PI 161375 (red) and 'Piel de sapo' T111 (blue). Numbers represent Mb.

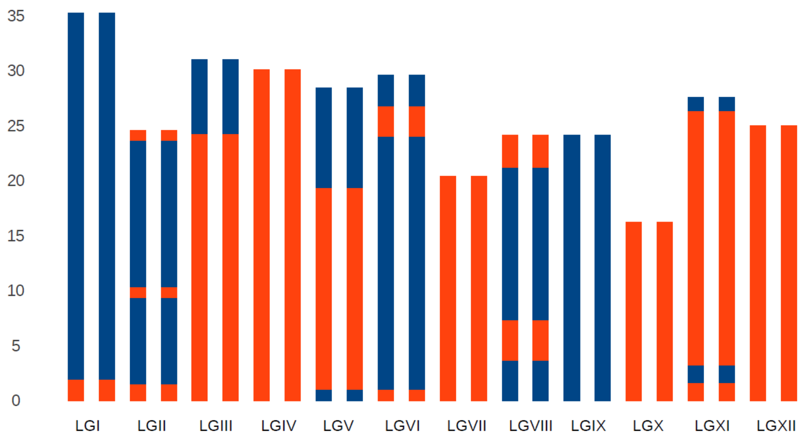
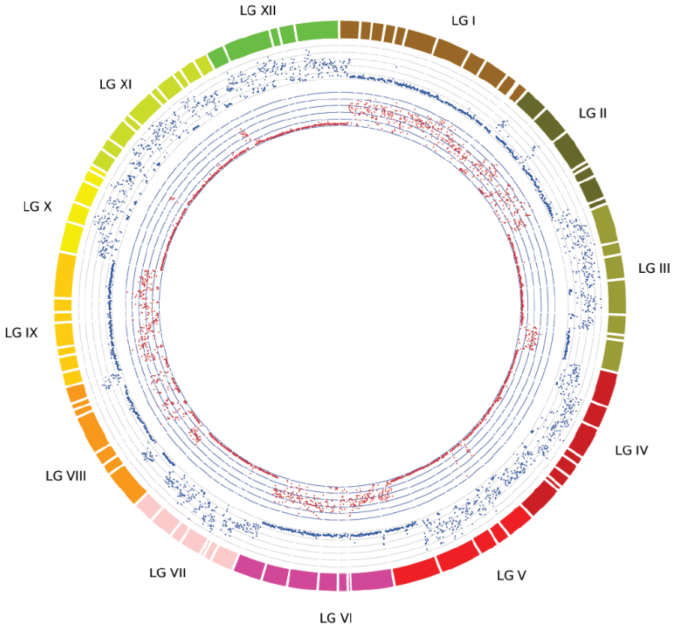


Figure S12. Distribution of SNPs between the DHL92 reference genome and SC and PS. The 12 melon pseudochromosomes are represented in different colours. Blocks represent scaffolds. SNP frequency between SC (red) and PS (blue) was calculated using 100 Kb windows.



SI REFERENCES

1. van Leeuwen H, et al. (2003) Identification and characterization of a melon genomic region containing a resistance gene cluster from a constructed BAC library. Microcolinearity between *Cucumis melo* and *Arabidopsis thaliana*. *Plant Mol Biol* 51:703-718.
2. González V, et al. (2010) Genome-wide BAC-end sequencing of *Cucumis melo* using two BAC libraries. *BMC Genomics* 11:618.
3. Niu B, et al. (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinf* 11:187.
4. Falgueras J, et al. (2010) SeqTrim: a high-throughput pipeline for preprocessing any type of sequence read. *BMC Bioinf* 11:38.
5. Rodríguez-Moreno L, et al. (2011) Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. *BMC Genomics* 12:424.
6. Li H, et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078-2079.
7. van Leeuwen H, Monfort A, Zhang HB, Puigdomenech P (2003) Identification and characterization of a melon genomic region containing a resistance gene cluster from a constructed BAC library. Microcolinearity between *Cucumis melo* and *Arabidopsis thaliana*. *Plant Mol Biol* 51:703-718.
8. van Leeuwen H, Garcia-Mas J, Coca M, Puigdomènech P, Monfort A (2005) Analysis of the melon genome in regions encompassing TIR-NBS-LRR resistance genes. *Mol Gen Genom* 273:240-251.
9. Deleu W, et al. (2007) Structure of two melon regions reveals high microsynteny with sequenced plant species. *Mol Gen Genom* 278:611-622.
10. González VM, et al. (2010) Sequencing of 6.7 Mb of the melon genome using a BAC pooling strategy. *BMC Plant Biol* 10:246.
11. Blanca J, et al. (2011) Melon transcriptome characterization. SSRs and SNPs discovery for high throughput genotyping across the species. *The Plant Genome* 4:118-131.
12. Gonzalo MJ, et al. (2005) Simple-sequence repeat markers used in merging linkage maps of melon (*Cucumis melo* L.). *Theor Appl Genet* 110:802-811.
13. Lander ES, et al. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174-181.
14. Choulet F, et al. (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *The Plant Cell* 22:1686-1701.
15. Huang S, et al. (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* 41:1275-1281.
16. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35:W265-268.
17. Han Y, Wessler SR, (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* 38:e199.
18. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.
19. Kimura M, (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal Mol Evol* 16:111-120.
20. Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and

palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *ProcNat Acad Sci USA* 93:10274-10279.

21. Zdobnov EM, Apweiler R (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847-848.
22. Hunter S, et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* 37:D211-215.
23. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopaedia of genes and genomes. *Nucleic Acids Res* 28:27-30.
24. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35:W182-185.
25. Apweiler R, et al. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32:D115-119.
26. Croft D, et al. (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 39:D691-697.
27. Sicheritz-Pontén T, Andersson SG (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res* 29:545-552.
28. Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol* 9:235.
29. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *Journal Mol Biol* 147:195-197.
30. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
31. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings Bioinformatics* 9:286-298.
32. Subramanian AR, et al. (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Molecular Biology* 3:6.
33. Landan G, Graur D (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol* 24:1380-1383.
34. Wallace IM, et al. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34:1692-1699.
35. Capella-Gutiérrez S, et al. (2009) TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972-1973.
36. Gascuel O, (2009) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685-695.
37. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307-321.
38. Akaike HA (1973) Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory.*
39. Huerta-Cepas J, et al. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res* 39:D556-560.
40. Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol* 9:235.
41. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python Environment for Tree Exploration. *BMC Bioinf* 11:24.
42. Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T (2010) The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. *Insect Mol Biol* 19(Suppl 2):13-21.
43. Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24:1540-1541.
44. Huerta-Cepas J, Gabaldón T (2011) Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics* 27:38-45.
45. Medina I, et al. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* 38:W210-213.
46. Supek F, et al. (2010) Translational selection is ubiquitous in prokaryotes. *PLoS Genet.* 6:e1001004.
47. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25:1335-1337. Erratum in: *Bioinformatics* 25:1713.
48. Gardner PP, et al. Rfam: updates to the RNA families database. *Nucleic Acids Res* 37:D136-140.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403-410.
50. Gustafson AM, et al. (2005) ASRP: the *Arabidopsis* Small RNA Project Database. *Nucleic Acids Res* 33:D637-640.
51. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2007) miRBase: tools for microRNA genomics. *Nucleic Acids Res* 36:D154-D158.
52. Enright AJ, et al. (2003) MicroRNA targets in Drosophila. *Genome Biol* 5:R1.
53. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406-3415.
54. Zhang B, Pan X, Wang Q, Cobb GP, Anderson TA (2006) Computational identifications of microRNAS and their targets. *Comput Biol Chem* 30:395-407.
55. Ambros V, et al. (2003) A uniform system for microRNA annotation. *RNA* 9:277-279.
56. Meyers BC et al. (2008) Criteria for Annotation of Plant MicroRNAs. *Plant Cell* 20:3186-3190.

57. Bonnet E, Wuyts J, Rouzé P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* 20:2911-2917.
58. Gonzalez-Ibeas D, et al. (2011) Analysis of the melon (*Cucumis melo*) small RNAome by high throughput pyrosequencing. *BMC Genomics* 12:393.
59. SanseverinoW, et al. (2010) PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res* 38:D814-821.
60. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573-580.
61. Alkan C, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 41:1061-1067.
62. Soderlund C, Bomhoff M, Nelson WM (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* 39:e68.
63. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biology* 5:R12.
64. Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss, and rearrangement. *PLoS One* 5:e11147.
65. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
66. Krzywinski M, et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19:1639-1645.
67. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25:1754-1760.
68. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491-498.
69. Danecek P, et al. (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.
70. Mungall CJ, Emmert DB (2007) The FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* 23:337-346.
71. Stein LD, et al. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.* 12:1599-1610.
72. Youens-Clark K, Faga B, Yap IV, Stein L, Ware D (2009) CMap 1.01: a comparative mapping application for the internet. *Bioinformatics* 25:3040-3042.
73. The potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475:189-195.
74. Velasco R, et al. (2010) The genome of the domesticated apple (*Malus* x *domestica* Borkh.). *Nat Genet* 42:833-839.
75. Shulaev V, et al. (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet* 43:109-116.
76. Huang S, et al. (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* 41:1275-1281.
77. The *Brassica rapa* Genome Sequencing Project Consortium (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet* 43:1035-1039.
78. Argout X, et al. (2011) The genome of *Theobroma cacao*. *Nat Genet* 43:101-108.
79. Al-Dous EK, et al. (2011) De novo genome sequencing and comparative Genomics of date palm (*Phoenix dactylifera*). *Nat Biotechnol* 29:521-527.
80. Schmutz J, et al. (2011) Genome sequence of the paleopolyploid soybean. *Nature* 463:178-183.
81. Ming R, et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452:991-996.