Published July, 2011

ORIGINAL RESEARCH

Melon Transcriptome Characterization: Simple Sequence Repeats and Single Nucleotide Polymorphisms Discovery for High Throughput Genotyping across the Species

José Miguel Blanca, Joaquín Cañizares, Pello Ziarsolo, Cristina Esteras, Gisela Mir, Fernando Nuez, Jordi Garcia-Mas, and María Belén Picó*

Abstract

Melon (*Cucumis melo* L.) ranks among the highest-valued fruit crops worldwide. Some genomic tools are available for this crop, including a Sanger transcriptome. We report the generation of 689,054 *C. melo* high-quality expressed sequence tags (ESTs) from two 454 sequencing runs, using normalized and nonnormalized complementary DNA (cDNA) libraries prepared from four genotypes belonging to the two *C. melo* subspecies and the main commercial types. 454 ESTs were combined with the Sanger available ESTs and de novo assembled into 53,252 unigenes. Over 63% of the unigenes were functionally annotated with Gene Ontology (GO) terms and 21% had known orthologs of *Arabidopsis thaliana* (L.) Heynh. Annotation distribution followed similar tendencies than that reported for *Arabidopsis thaliana*, suggesting that the dataset represents a fairly complete melon transcriptome. Furthermore, we identified a set of 3298 unigenes with microsatellite motifs and 14,417 sequences with single nucleotide variants of which 11,655 single nucleotide polymorphism met criteria for use with high-throughput genotyping platforms, and 453 could be detected as cleaved amplified polymorphic sequence (CAPS). A set of markers were validated, 90% of them being polymorphic in a number of variable *C. melo* accessions. This transcriptome provides an invaluable new tool for biological research, more so when it includes transcripts not described previously. It is being used for genome annotation and has provided a large collection of markers that will allow speeding up the process of breeding new melon varieties.

Published in The Plant Genome 4:118–131. Published 21 June 2011.
doi: 10.3835/plantgenome2011.01.0003

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

MELON (*Cucumis melo* L.) is an important vegetable crop that is grown worldwide, mainly in temperate, subtropical, and tropical climates. It ranks as the ninth horticultural crop in terms of total world production. This species belongs to the botanical family Cucurbitaceae, commonly known as cucurbits, which includes several economically and nutritionally important vegetable crops, such as cucumber (*Cucumis sativus* L.), watermelon [*Citrullus lanatus* (Thunb.) Matsum. & Nakai], and pumpkins, gourds, and squashes (*Cucurbita* spp.) (Schaefer et al., 2009).

Melon is a diploid species ($2n = 2x = 24$) with an estimated genome size of 450 Mbp, which is similar to that of rice (*Oryza sativa* L.) or cucumber (Huang et al., 2009) and approximately three times the size of the model species *Arabidopsis thaliana* (L.) Heynh. (Arabidopsis Genome Initiative, 2000). The species displays a rich diversity of many traits and has become a primary model for sex expression and fruit ripening analysis (Boualem et al., 2008). *Cucumis melo* is considered to be divided into two subspecies, each one with several botanical varieties (Pitrat, 2008): subsp.

J.M. Blanca, J. Cañizares, P. Ziarsolo, C. Esteras, F. Nuez, and M.B. Picó, COMAV, Institute for the Conservation and Breeding of Agricultural Biodiversity, Universitat Politècnica de València (UPV), Camino de Vera s/n, 46022 Valencia, Spain; G. Mir and J. Garcia-Mas, IRTA, Centre de Recerca en Agrigenòmica CSIC-IRTA-UAB, Carretera de Cabrils Km 2, 08348 Cabrils (Barcelona), Spain.
Received 22 Jan. 2011.*Corresponding author (mpicosi@btc.upv.es).

Abbreviations: CAPS, cleaved amplified polymorphic sequence; cDNA, complementary DNA; COMAV, Institute for the Conservation and Breeding of Agricultural Biodiversity; dNTPs, deoxyribonucleotide triphosphates; EST, expressed sequence tag; GO, Gene Ontology; GS, Genome Sequencer; ICuGI, International Cucurbit Genomics Initiative; mRNA, messenger RNA; NCBI, National Center for Biotechnology Information; ORF, open reading frame; PCR, polymerase chain reaction; PIC, polymorphism information content; SNP, single nucleotide polymorphism; SNV, single nucleotide variant; SSR, simple sequence repeat; UTR, untranslated region.

melo and subsp. *agrestis* (Naudin) Pangalo, including the main commercial types and the most important sources of resistances and quality traits, respectively.

The availability of genetic and genomic resources for this crop has increased significantly in recent years due to several national and international projects, such as the Spanish initiative Melogen (González-Ibeas et al., 2007; Melogen, 2003) and the International Cucurbit Genomics Initiative (ICuGI) (ICuGI, 2007). A broad range of genomic tools are available today (Ezura and Fukino, 2009), such as mapping populations of double haploids, recombinant inbreeds and near-isogenic lines (Eduardo et al., 2007; Harel-Beja et al., 2010), genetic maps (Deleu et al., 2009), a detailed physical map (González et al., 2010a), and an oligo-based microarray that is providing the first expression studies (Mascarell-Creus et al., 2009). These tools include a collection of expressed sequence tags (ESTs), obtained using Sanger sequencing technologies, from a number of genotypes belonging to both subspecies. These ESTs have been assembled to generate a melon transcriptome. The last version of the assembly (version 4.0, released in May 2010), available at the Cucurbit Genomics Database (ICuGI, 2007), includes 24,444 unigenes. Also, an effort is in progress through a Spanish initiative to obtain the whole genome sequence of this crop (González et al., 2009, 2010b).

A complete transcriptome is a basic resource for gene discovery, large-scale expression analysis, and genome annotation. The cost and the limited parallelization of the conventional Sanger method has prevented the large scale generation of EST that is necessary to generate complete transcriptomes. Recent advances in next-generation sequencing technologies allow us a very deep EST sequencing, efficiently and cost-effectively (Shendure and Ji, 2008). There are increasing studies in which 454 sequencing technology, combined or not with Solexa/Illumina, are used to characterize transcriptomes in cereals, legumes, and other crops (Cheung et al., 2006; Emrich et al., 2007; Vega Arreguin et al., 2009; Folta et al., 2010; Guo et al., 2010). Even in model species, such as *Arabidopsis thaliana*, this deep sequencing is allowing us to identify new transcripts not present in previous ESTs collections (Weber et al., 2007). Also, specific transcriptomes are being generated in species for which previous genomic resources are lacking (Alagna et al., 2009; Barakat et al., 2009; Wang et al., 2009; Li et al., 2010; Sun et al., 2010). When this massive amount of data is produced from a number of different genotypes, it offers a means to identify and characterize the genetic polymorphisms underlying the phenotypic variation. Single nucleotide polymorphisms (SNPs) are the most abundant variations in genomes and, therefore, constitute a powerful tool for mapping and marker-assisted breeding. The new transcripts produced with next-generation sequencing are being used for high-throughput SNP identification. Single nucleotide polymorphism detection is performed by aligning raw reads from different genotypes to a reference genome or transcriptome previously available, as in maize (*Zea mays* L.), cucumber, and even in polyploid species such as oilseed rape (*Brassica napus* L.) (Barbazuk et al., 2006; Trick et al., 2009; Guo et al., 2010). De novo assembly of raw sequences coming from a set of genotypes, followed by pairwise comparison of the overlapping assembled reads, has also successfully been used in species lacking any significant genomic or transcriptomic resources (Novaes et al., 2008; Blanca et al., 2011).

In this study, we describe the generation of 689,054 new *C. melo* high-quality ESTs from two 454 sequencing runs, one Genome Sequencer (GS) FLX and one GS FLX Titanium, using four libraries prepared from a number of tissues and four genotypes belonging to the main commercial types, with contrasting fruit phenotypes, and to the two subspecies of *C. melo*. By combining these new ESTs with 125,908 Sanger ESTs previously available, a melon transcriptome has been generated with 53,252 unigenes. These unigenes have been accurately annotated, screened for simple sequence repeat (SSR) motifs, and used to identify a large SNP collection suited for high-throughput mapping purposes. All these tools will allow accelerating genetics and breeding of this crop.

MATERIALS AND METHODS

Plant Material

Two nonnormalized complementary DNA (cDNA) libraries were constructed using leaf material from the T111 Piel de sapo line (Semillas Fitó, Barcelona, Spain), belonging to *C. melo* subsp. *melo* var. *inodorus* H. Jacq., and the accession PI161375, belonging to *C. melo* subsp. *agrestis* var. *conomon* (Thunb.) Makino. These two genotypes are the parents of the melon genetic map (Deleu et al., 2009) and belong to the two subspecies of *C. melo*. Two normalized cDNA libraries were constructed using total RNA extracted from different tissues (leaf, root, male and female flowers at different stages, seedlings, and dark-grown and ethylene treated seedlings) of two genotypes, the Piñonet Piel de sapo cultivar (belonging to *C. melo* subsp. *melo* var. *inodorus*) and the Vedrantais cultivar (belonging to *C. melo* subsp. *melo* var. *cantalupensis* Naudin), belonging to the main melon commercial types, having contrasting phenotypes for fruit ripening interesting in melon breeding (both genotypes maintained at the germplasm collection of the Cucurbits breeding group of the Institute for the Conservation and Breeding of Agricultural Biodiversity [COMAV], Valencia, Spain). All tissues were collected and immediately frozen in liquid nitrogen and stored at –80°C till use.

Complementary DNA Preparation and Sequencing

Total RNA from T111 and PI161375 lines was extracted from young leaves with TRI Reagent (Sigma-Aldrich, Saint Louis, MO). Double-stranded cDNA was synthesized using SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Carlsbad, CA) and an oligo d(T)-primer.

Fragments below 300 bp were removed with Agencourt AMPure beads (Beckman Coulter Genomics, Danvers, MA). Final double-stranded cDNA size distribution was checked using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Three micrograms of size-selected double-stranded cDNA was used for each GS FLX library and two quarters of a sequencing plate were used for each sample. Sequencing was performed at the CRAG Sequencing Service (Barcelona, Spain).

Total RNA from Piñonet and Vedrantais was extracted from each tissue using the TRI Reagent. Equivalent amounts of RNA from each tissue were combined into one pool per cultivar. Messenger RNA (mRNA) was purified from the total RNA using the illustra mRNA Purification Kit (GE Healthcare, Amersham Bioscience, Buckinghamshire, UK). Double-stranded cDNA was then synthesized from the RNA pools with the SMART cDNA Library Construction Kit (Clontech, Palo Alto, CA). A normalization step was performed with the TRIMMER cDNA normalization Kit (Evrogen, Moscow, Russia) to prevent over-representation of the most common transcripts. The polymerase chain reaction (PCR) products of cDNA were purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany). Normalization quality of cDNAs libraries was checked by quantitative PCR. The cDNA length and normalization are critical factors to have a good transcriptome representation, to have SNPs along the whole gene sequence, and to have a high quality SNP prediction. Approximately 1 µg of double-stranded cDNA from each of the two normalized cDNA pools were used for sequencing on a GS FLX Titanium Sequencing platform. A half sequencing plate was performed for each sample at Creative Genomics (New York, NY).

Complementary DNA Sequence Processing and Assembly

The whole sequence analysis was performed by using the ngs_backbone pipeline developed at COMAV (Bioinformatics at COMAV, 2010). The tools and analysis mentioned in the following sections were all performed by ngs_backbone, but here the third party tools, databases, and parameters used by ngs_backbone are described.

For assembly, 454 sequences were combined with Sanger sequences, previously available at ICuGI database (ICuGI, 2007). Sanger sequences were obtained from different libraries and genotypes; the same four genotypes sequenced with 454 (the *inodorus* Piel de sapo T111 and Piñonet, the *cantalupensis* Vedrantais, and the *conomon* PI161375) and six additional genotypes, the cultivar TamDew (*C. melo* subsp. *melo* var. *inodorus),* the cultivars Dulce, Noy Israel, and Charentais, the accession C35 of the germplasm collection of Estación Experimental La Mayora – Consejo Superior de Investigaciones Científicas (EELM-CSIC, Málaga, Spain) (*C. melo* subsp. *melo* var. *cantalupensis*), and the exotic accession Pat81 of the COMAV collection (*C. melo* subsp. *agrestis* var. *conomon*).

Sequences were processed before the assembly. To remove the adaptors, an alignment with the primers and adaptors used during the sequencing process was done by Exonerate (Exonerate, 2005). The low quality regions from the reads were trimmed by using Lucy (Chou and Holmes, 2001). Sequences shorter than 100 bp were discarded and not used for the assembly. The processed sequences were assembled with Mira (Chevreux et al., 2004). Default ngs_backbone options were used to assemble the unigenes.

Gene Annotation

Structural and functional annotation was performed by sequence comparison with public databases. All unique assembled sequences (unigenes) were sequentially compared using blast (cutoff e-value of 10^{-20}) with the sequences in three major public protein databases, prioritizing nonmachine curate databases. The used database order was Swiss-Prot (uniprot_sprot_ release of 2010 04 23) (UniProt Consortium, 2010a, b), *Arabidopsis* proteins (Tair_9_pep_ release 2009 06 19) (Arabidopsis Information Resource, 2009; Swarbreck et al, 2008), and UniRef90 (uniref90_release 2010 04 23) (European Bioinformatics Institute, 2010; UniProt Consortium, 2010b). Once a sequence had a good enough blast hit in one of the databases, a description was build from the description of that hit. Also, a bidirectional blast search comparison was performed to obtain a set of putative orthologs with *Arabidopsis thaliana* and equivalents with melon, using the melon unigenes contained in the ICuGI database (ICuGI, 2007).

Additionally, we performed a functional classification of the unigenes following the Gene Ontology (GO) scheme. Blast2GO (Conesa and Götz, 2008) was used for this purpose. Blast2GO used the results of a blast search against the National Center for Biotechnology Information (NCBI) nonredundant protein database (release of 20010-03-27) (NCBI, 2010a) (cutoff e-value of 10^{-20}) to infer the relevant GO terms for every sequence. Open reading frames (ORFs) were predicted in the unigenes with the aid of the ESTScan software (Iseli et al, 1999). We used the *Arabidopsis* codon usage table to perform the ORF searching. Introns were assigned by aligning the unigenes with the melon genomic sequence (a draft is being produced within the MELONOMICS consortium and is available for the partners of the funding project [González et al., 2010b]) using the Emboss: est2 genome (European Bioinformatics Institute, 2001).

Identification of Simple Sequence Repeats and Single Nucleotide Polymorphisms

Simple sequence repeats were annotated using the Sputnik software (Abajian, 1994). Sequences containing greater than or equal to four di-, tri-, or tetra-nucleotide repeats were selected. A set of SSRs were validated using the genotypes sequenced with 454 (T111, Piñonet, Vedrantais, and PI161375), some of the genotypes sequenced with Sanger (TamDew, Dulce, Noy Israel, and Pat 81) described

previously, and two additional genotypes belonging to the *C. melo* L. subsp. *melo* var. *momordica* (Roxb.) Duthie & J. B. Fuller group (PI124111 and PI414743, provided by National Plant Germplasm System [NPGS]-USDA, and maintained at the COMAV collection) used as parentals of other melon genetic maps. Primer pairs flanking each SSR locus were designed using the Primer3 program (Rozen and Skaletsky, 2000).

Polymerase chain reactions were performed in a final volume of 15 µL with 1x PCR buffer (100 mM Tris-HCl, 15 mM $MgCl_2$, 500 mM KCl, and pH 8.3), 200 µM deoxyribonucleotide triphosphates (dNTPs), 0.15 µM each primer, and 1 µL of template (approximately 10 ng μL^{-1}) 0.5 U Taq polymerase. The cycling conditions were as follows: denaturation at 94°C for 2 min, followed by seven cycles of 45 s at 94°C, 45 s at 68°C (with each cycle the annealing temperature decreasing 2°C), and of 1 min at 72°C. Products were subsequently amplified for 30 cycles at 94°C for 45 s, 54°C for 45 s, and 72°C for 1 min, with a final extension at 72°C for 7 min. The forward primer was designed adding an M13 tail to its 5′ end. Polymerase chain reaction products were separated using 6% polyacrylamide gels, 1x Tris-borate-ethylenediaminetetraacetic acid (EDTA) (TBE) buffer in a LI-COR 4300 (Li-COR BioScience, Lincoln, NE). IRD700 and IRD800-labeled amplicons were visualized by adding to PCR mixture 0.2 µM of fluorescent label (IR700 or IR800) M13 tail. Number of alleles, frequency of the most common allele, and the polymorphism information content (PIC) were calculated for each locus for the 10 melon genotypes using the PowerMarker software (Liu and Muse, 2005). Cluster analysis was also performed with the Nei's genetic distance (Nei et al., 1983). The support values for the degree of confidence at the nodes of the dendrogram were analyzed by bootstrap resampling 1000 times.

This *C. melo* EST collection has been produced using several genotypes belonging to different subspecies and morphotypes of *C. melo* appropriate for SNP discovery. ngs_backbone was also used to detect single nucleotide variants (SNVs) (SNPs and indels) by mapping the 454 and Sanger processed reads against the unigene assembly using BWA (Burrows-Wheeler Aligner) (Li and Durbin, 2010). We kept only SNVs meeting stringent quality criteria: (i) minimum allele quality (accumulated sequence quality for every allele) and (ii) minimum mapping quality. The default threshold set by ngs_backbone was set for both parameters.

Despite satisfying the quality criteria, not all the SNVs seemed equally reliable. Several filters were applied to maximize a successful validation and/or implementation in high-throughput genotyping platforms (Fan et al., 2006; Gupta et al., 2008). Single nucleotide variants include both SNPs and indels and the VKS filter (VKS [it is not an SNP]) classifies them in those categories. Some filters dismiss SNVs in redundant or highly variable regions (UCR [region is not unique or noncontiguous] and HVR4 [the region has more than four SNVs per 100 bp']). Other filters used were CS60 (SNV is closer than 60 bp to another SNP or indel), I59 (an intron is located closer than 59 bp), and CL60 (SNV is closer than 60 bp to the sequence edge) that facilitate the use of variants in Golden Gate genotyping platforms.

Those SNVs with only one allele (sequenced one or more times) within each genotype can be selected by filtering out those that are variable with NVPiñ, NVT111, NVVed, NVDul, NVPI161, and NVPat81 (it is variable in Piñonet, T111, Vedrantais, Dulce, PI161375, and Pat81, respectively). Also, those with only one allele, with one or more reads, within a group of genotypes can be selected by using NVPs for the Piel de sapo group (it is variable in T111 and Piñonet), NVCant for the *cantalupensis* group (it is variable in Vedrantais, Charentais, and Dulce), NVCon for the *conomon* group (it is variable in Pat81 and Pi161375), and NVmelo for the accessions of the subsp. *melo* (it is variable in T111, Piñonet, Dulce, Vedrantais, and Charentais). These filters allow us to identify markers uniform within varieties or subspecies. Combined with additional filters VSCant-Con (it is not variable in Dulce, Vedrantais, Charentais, Pat81, and PI161375) and VSPs-Con (it is not variable in T111, Piñonet, Pat81, and PI161375), they allow to select those SNPs polymorphic among groups.

We also detected those SNPs that can be analyzed via cleaved amplified polymorphic sequence (CAPS) (searching for allele-specific restriction targets) filtering out with nCAP, and validated a subset of them using the same set of genotypes that in SSRs analysis. Polymerase chain reactions were performed in a final volume of 25 µL with 1x PCR buffer (100 mM Tris-HCl, 15 mM $MgCl_2$, 500 mM KCl, and pH 8.3), 200 µM dNTPs, 0.15 µM each primer, and 2 µL of template (approximately 10 ng μL^{-1}). The cycling conditions were as follows: denaturation at 95°C for 3 min, followed by 30 cycles of 30 s at 95°C, 30 s at 55°C, and of 60 s at 72°C, with a final extension at 72°C for 7 min. Polymerase chain reaction products were digested with the corresponding enzymes and detected by 2% agarose gel electrophoresis. Cluster analysis was as described for SSRs.

RESULTS AND DISCUSSION

Expressed Sequence Tag Sequencing and Assembly

We performed two 454 sequencing runs, one with the GS FLX reagents and one with GS FLX Titanium reagents on the GS FLX System. A half GS FLX run was performed on each of the two libraries constructed from leaves of the genotypes used as parentals of the melon genetic map (Deleu et al., 2009), belonging to the two subspecies of *C. melo* subsp. *melo* var. *inodorus* cv. Piel de sapo T111 and subsp. *agrestis* var. *conomon* PI161375. The two GS FLX Titanium half runs were performed on two normalized libraries constructed from a pool of tissues (leaves, roots, female and male flowers at different stages, seedlings, dark-grown seedlings, and ethylene-treated seedlings) using two cultivars of the subsp. *melo*, belonging to the main melon commercial classes, var.

Table 1. Sequence statistics of *Cucumis melo* 454 expressed sequence tags (ESTs).

Library	Reads number, average length	Total length (bp)	Sequence quality	Processed reads number, average length	Total length (bp)	Sequence quality
Piel de sapo T111†	117,352, 225	26,388,460	33	107,756, 233	25,067,194	34
PI161375†	98,460, 231	22,768,835	33	91,236, 237	21,658,724	34
TOTAL	**215,812, 228**	**49,157,295**	**33**	**198,992, 235**	**46,725,918**	**34**
Piel de sapo Piñonet‡	249,902, 397	99,083,949	33	226,650, 403	91,284,481	34
Vedrantais‡	297,638, 387	115,306,933	33	263,412, 398	104,745,240	34
TOTAL	**547,540, 392**	**214,390,882**	**33**	**490,062, 400**	**196,029,721**	**34**

†Summary of the *Cucumis melo* expressed sequences generated with two half runs of 454 Genome Sequencer (GS) FLX. Statistics of reads before and after processing are indicated.

‡Summary of the *Cucumis melo* expressed sequences generated with two half runs of Titanium pyrosequencing. Statistics of reads before and after processing are indicated.

inodorus cv. Piel de sapo Piñonet and var. *cantalupensis* cv. Vedrantais. A total of 763,352 reads were obtained from the four libraries (Table 1). The GS FLX Titanium run provided reads ~1.7x longer than GS FLX (average length of 392 versus 228 bp). Similar lengths have been reported in previous studies (Gedye et al., 2010; Li et al., 2010; Sun et al., 2010; Blanca et al., 2011).

Reads were processed using the ngs_backbone software (Bioinformatics at the Institute for the Conservation and Breeding of Agricultural Biodiversity, 2010) to eliminate adaptor sequences, low quality chromatograms, and sequences of less than 100 bp. After processing, we obtained 689,054 high quality ESTs, comprising 243 Mbp, with an average length of 400 and 235 bp for the two sequencing methods, respectively. The length distribution of these ESTs is shown in Fig. 1. More than 89% of the ESTs fell between 200 and 550 bp in length. All reads were deposited in the NCBI database and can be accessed in the Sequence Read Archive (NCBI, 2010b) under the accession number SRA050214.

The 454 ESTs generated in this study were combined for clustering and de novo assembly with a previously available collection of 125,908 ESTs (average length 613 bp and total length 81.3 Mbp) produced using traditional Sanger sequencing methods. Sanger ESTs were obtained mainly by two transcriptome sequencing initiatives, Melogen (González-Ibeas et al., 2007) and ICuGI (ICuGI, 2007), from different tissues (callus, leaves, roots, flowers, and fruits at different stages) and from 10 genotypes (including the four sequenced in this paper with 454 and one additional *inodorus*, four *cantalupensis*, and one *conomon*) and are publicly available at ICuGI (ICuGI, 2007). A summary of the Sanger dataset using for this assembly, before and after processing, is detailed in Supplemental File S1.

Finally, 753,004 ESTs were assembled using the Mira assembler (Chevreux et al., 2004) yielding a total of 53,252 high-confident tentative consensus sequences (nonredundant sequences or unigenes). Approximately 55% of the unigenes (29,566) were composed only of 454 ESTs, whereas only 4% (1873) were assembled exclusively from Sanger ESTs and 41% (21,813) included sequences of both datasets. About 28, 32, and 96% of the unigenes included sequences from Melogen (González-Ibeas et al., 2007), ICuGI (ICuGI, 2007), and 454 datasets, respectively. These results agree with those of previous studies combining 454 and Sanger ESTs (Vega-Arreguin et al., 2009; Guo et al., 2010; Li et al., 2010; Ueno et al., 2010) in which next-generation sequencing contributes significantly to the identification of novel unigenes.

The distribution of the number of ESTs per unigene is shown in Fig. 2A. The majority of unigenes were assembled from a moderate number of ESTs (from 2 to 10), with an average of 14.1 ESTs per unigene. This relatively low redundancy is probably due to the success of the normalization process, responsible for the suppression of superabundant transcripts. Also, the EST assemblers can in some cases split ESTs generated from the same transcript into several unigenes. This behavior is usually due to the difficulty of distinguishing sequence variations due to individual provenance to those due to close paralogs.

However, part of the 454 and the Sanger ESTs came from nonnormalized libraries. Consistently, we were able to identify a number of highly abundant transcripts. Around 3103 transcripts (5.8% of all the unigenes) have more than 50 EST members. These abundant transcripts contain ~60% of the assembled EST reads.

The assembled unigenes had an average length of 774.3 bp comprising approximately 42.2 Mbp in total. The length distribution of the unigenes is shown in Fig. 2B. The analysis revealed that more of the 50% of unigenes were larger than 673 bp, and only 5% of the sequences were shorter than 266 bp.

The number of assembled unigenes is similar to that obtained in previous transcriptome analyses performed with massive sequencing in maize, *Eucalyptus grandis* W. Hill ex Maiden, *Artemisia annua* L., chestnut (*Castanea* spp.), olive (*Olea europaea* L.), and cucumber. However, our de novo assembly with the longer reads obtained with the GS FLX Titanium platform and the combination with Sanger sequences render unigenes that on average are almost two times longer than those reported in other studies that used Genome Sequencer 20 and Genome Sequencer FLX platforms (Novaes et al., 2008; Alagna et al., 2009; Barakat et al., 2009; Vega-Arreguin et al., 2009; Wang et al., 2009; Guo et al., 2010). Our assembled unigenes were also larger than those reported for other plant trancriptomes obtained using

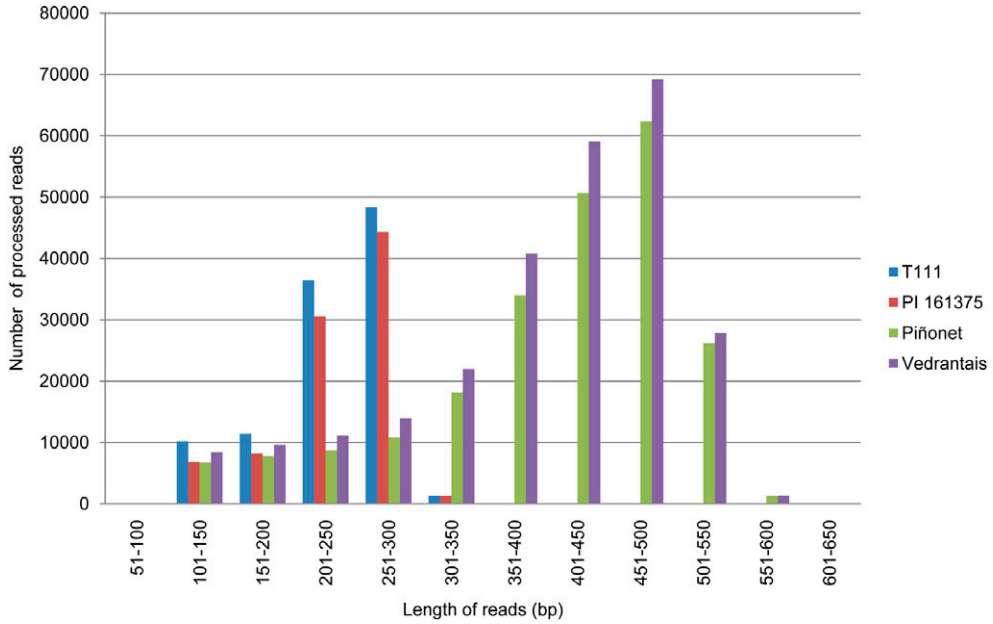

Figure 1. Length distribution of the *Cucumis melo* expressed sequence tags (ESTs). Data obtained after sequencing with two half runs of 454 Genome Sequencer (GS) FLX sequencing each one of two *C. melo* complementary DNA (cDNA) libraries (subsp. *melo* var. *inodorus* H. Jacq. cv. Piel de sapo T111 and subsp. *agrestis* var. *conomon* PI161375) and with two half runs of GS FLX Titanium sequencing each one of two *C. melo* cDNA libraries (subsp. *melo* var. *inodorus* cv. Piel de sapo Piñonet and var. *cantalupensis* Naudin cv. Vedrantais). Data after processing reads are presented.

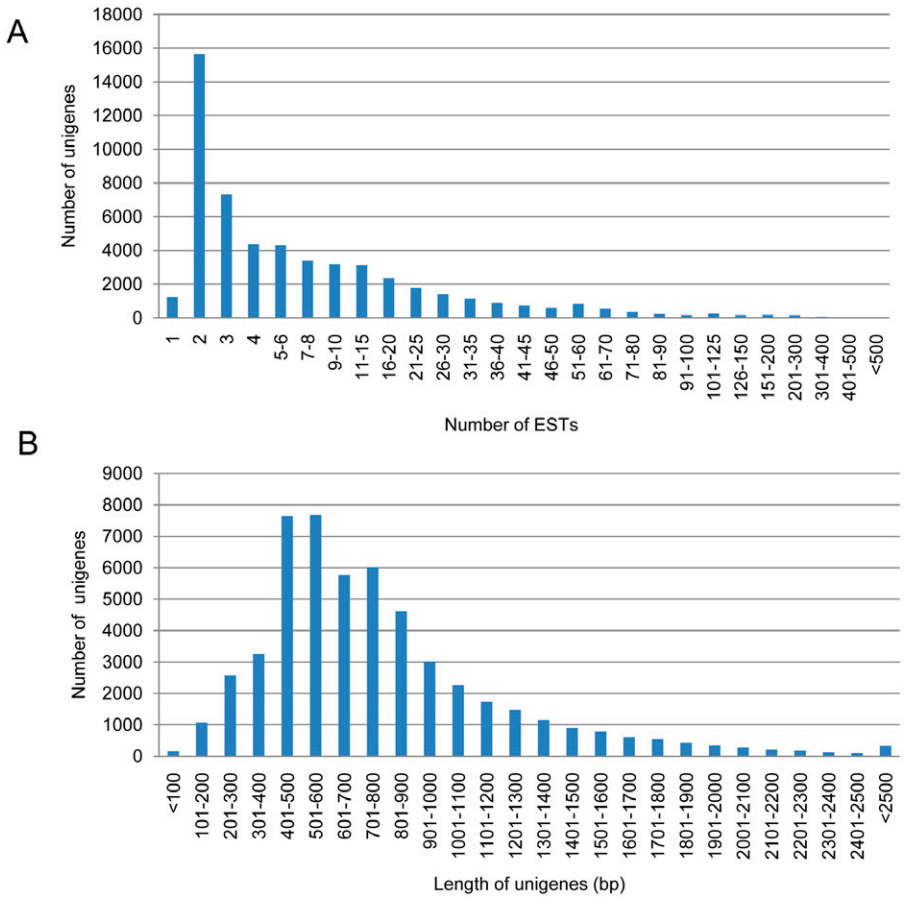

Figure 2. Distribution of number of expressed sequence tags (ESTs) and length of *Cucumis melo* unigenes. A) Unigenes are classified according the number of 454 and Sanger ESTs used in their assembly. B) Length distribution of *C. melo* unigenes de novo assembled combining 454 and Sanger ESTs.

the GS FLX Titanium platform combined with Sanger (Li et al., 2010; Sun et al., 2010). These differences in length might also be due to the different assemblers used.

Cucumis melo unigenes length is comparable to that reported for previous melon Sanger-based transcriptome (González-Ibeas et al., 2007), but the addition of 454 reads brought to 53,252 the number of unigenes, two times the number of unigenes reported in the last ICuGI assembly (24,444 unigenes in the v.4) (ICuGI, 2007). The sequences of the unigenes in fasta format are available in the Supplemental File S2 with unigene numbers from METC000001 to METC053252.

Structural and Functional Annotation

The long unigenes generated in this study present the advantage of being more accurately annotated. Most unigenes, 49,610 (93%) were predicted to have one ORF. By aligning the unigenes with the genomic sequence of melon, intron locations were predicted in 33,217 unigenes (63%). Annotation results regarding ORF and introns position are included in Supplemental File S3.

To identify *C. melo* unigenes potentially encoding proteins with known function, a BLAST analysis (Altschul et al., 1997) was performed in a sequential way using Swiss-Prot, *Arabidopsis* protein, and Uniref90 protein databases (e-value cutoff of 10^{-20}). Over 67% of the unigenes (35,740) had at least one significant hit. Most unigenes had significant hits in the Swissprot (59%) and in the *Arabidopsis* (33%) databases and less in Uniref (8%). Blast results are listed in Supplemental File S4.

Gene Ontology terms were further assigned to melon unigenes using Blast2GO (Conesa and Götz, 2008) based on the automated annotation of each unigene using BLAST results against the GenBank nonredundant protein database (nr) from NCBI (NCBI, 2010a). A total of 33,575 unigenes (63%) could be annotated with one or more GO terms. The unigenes distribution regarding the number of GOs to which they were assigned is shown in Supplemental File S5. The number of GO terms per unigene varied from 1 to 34. More than the 78% of the unigenes could be assigned to more than one GO term, the majority of the unigenes being mapped to two to seven GO terms. In total, 121,302 GO terms were retrieved. The GO annotation analysis reinforces the idea that a broad diversity of genes was sampled in our multitissue normalized and nonnormalized cDNA libraries.

We used the GO annotations to classify each unigene into different functional categories using a set of GO slims, which are a list of high-level GO terms providing a broad overview of the ontology content. Figure 3 shows the functional classification of melon unigenes into GO slims within the categories of biological process and molecular function in comparison with that obtained with the assembly of Sanger ESTs (available at ICuGI database, v4 [ICuGI, 2007]). Gene Ontology annotations showed fairly consistent sampling of functional classes. Cellular process and metabolic process were among the most highly represented groups in the category of biological process, indicating extensive metabolic activities. Genes involved in other important biological processes such as stress response, signal transduction, and ripening were also identified through GO annotations. Under the molecular function category, assignments were mainly to the binding and catalytic activities. A large number of hydrolases, kinases, and transferases were annotated. Also, transcription and translation factors were well represented. It is worth noting that percentages of genes involved in DNA, RNA, and protein binding and transcription factors were higher in the new unigene collection than in the Sanger ESTs annotation that was richer in transferases, hydrolases, and kinases. The distribution of *C. melo* unigenes also follow similar tendencies to that reported for *Arabidopsis thaliana* (González-Ibeas et al., 2007), suggesting that the dataset represents a fairly complete melon transcriptome. All GO annotations are compiled in the Supplemental File S6.

Doing a reciprocal blast search, we have also identified 11,253 *C. melo* unigenes (21%) with an ortholog in *Arabidopsis thaliana*, and 18,405 (35%) with a melon equivalent in the ICuGI databases (ICuGI, 2007). A list of the identified orthologs is included in Supplemental File S7.

Only 10,710 (20%) of the unigenes could not be annotated (no significant blast hits, no GO terms, and no ortholog or equivalent). Short sequences are less likely to align with a significant e-value. However, the average length of these nonannotated unigenes was 440 bp, with 50% being longer than 523 bp. For homology searches against known genes, unigenes longer than 200 bp are widely accepted for the effective assignment of functional annotations (Li et al., 2010). In previous studies performed with massive sequencing techniques a similar or even higher number of unigenes did not match with any reported transcribed sequences, potentially representing newly detected transcripts (Barbazuk et al., 2006; Weber et al., 2007; Vega-Arreguin et al., 2009), although a certain degree of contamination of transcriptome libraries with genomic DNA is also possible.

Simple Sequence Repeat and Single Nucleotide Polymorphism Discovery and Validation

We performed a general screen on the *C. melo* unigene dataset for the presence of microsatellites, analyzing its nature and frequency. A search for di-, tri-, and tetra-nucleotide repeats yielded a total of 3593 potential SSRs in 3298 unigenes; that is, approximately 6% of the unigenes contained at least one of the considered SSR motifs. This percentage agrees with previous studies using EST databases that show that 3 to 7% of expressed sequences contain putative SSR motifs (Thiel et al., 2003; Guo et al., 2010).

Approximately 45% (1631) of the SSRs were in genes with a melon equivalent in ICuGI (ICuGI, 2007), whereas 17% (624) were in nonannotated genes. The maximum and minimum lengths of the repeats were 124 and 17, and the average length was 27 nucleotides. These were mostly tri-nucleotide (67.7%) and less di- (25.2%) and tetra-nucleotide (7.1%). The most common repeat motifs

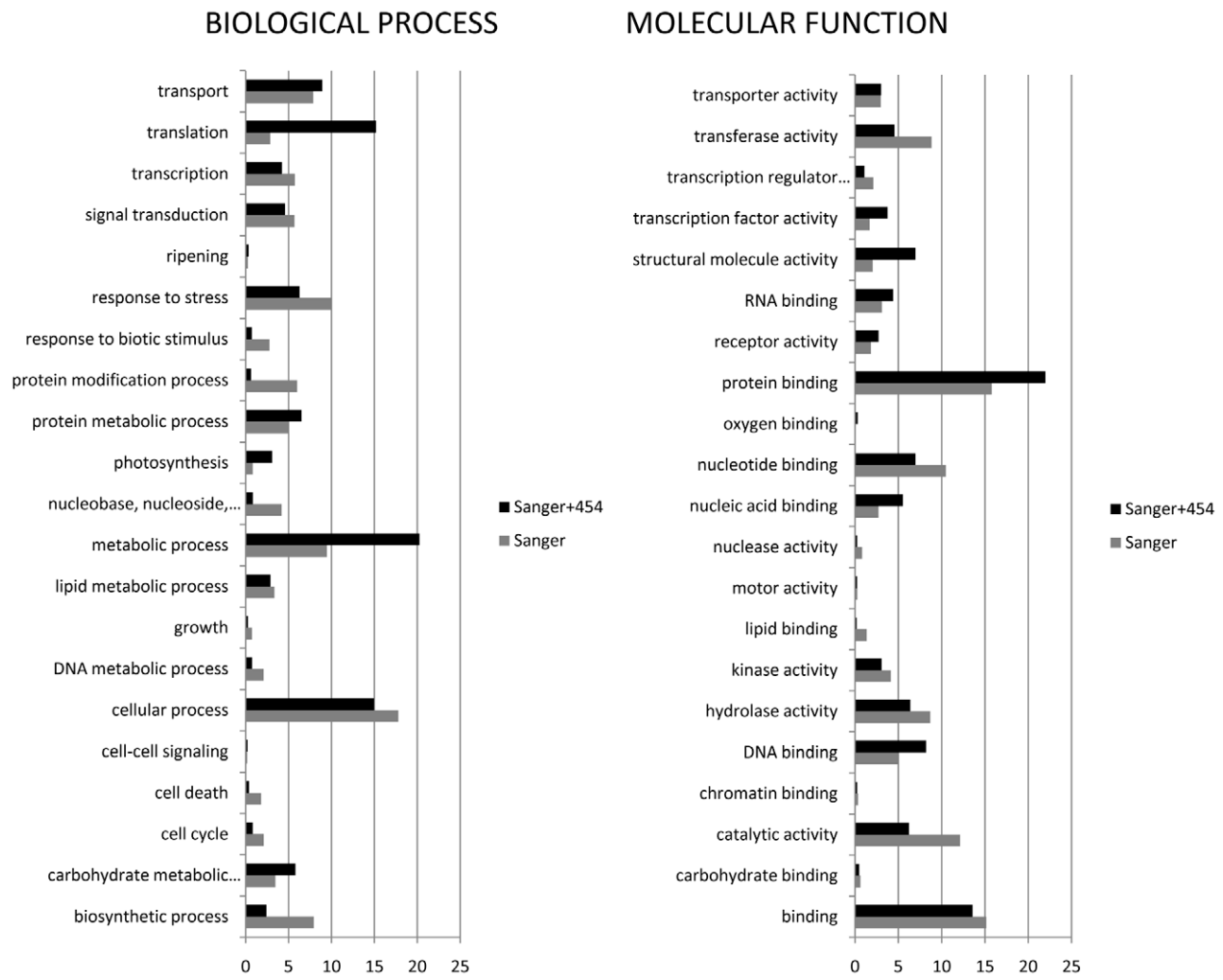

Figure 3. Percentage of *Cucumis melo* unigenes in each functional category. Detailed legend: *C. melo* unigenes were classified into different functional groups based on a set of Gene Ontology (GO) slims in the biological process and the molecular function category. Bars indicate the percentage of genes annotated in each GO term (black: blast2GO annotation of the unigene collection presented in this paper; gray: GO annotations of the Sanger-based transcriptome available at ICuGI database [ICuGI, 2007]).

are indicated in Table 2. The melon SSRs annotated in the Sanger collection (available at ICuGI [ICuGI, 2007]) showed a higher percentage of di-nucleotide motifs (41%) and differences in the distribution of trinucleotide motifs. A similar bias toward AG, AAG, and AAAG and against CG repeats has been reported in EST-SSRs of many crops, including previous studies of other cucurbits like cucumber and *Cucurbita pepo* L. (Guo et al., 2010; Blanca et al., 2011). It has been proposed that this may be due to the tendency of CpG sequences to be methylated, which potentially might inhibit transcription. The complete list of SSRs and their corresponding information are provided in Supplemental File S8.

It is known that the untranslated regions (UTRs) are richer in SSRs than the coding regions (Morgante et al., 2002; Thiel et al., 2003). Most melon SSRs with known positions were located in the UTRs (57%), mainly in the 3′ region, and less in the ORFs (43%) (Table 3). An analysis of the localization of di-, tri-, and tetra-repeats showed that tri-nucleotides localized preferentially in ORFs, consistently with maintenance of the ORFs coding capacity, whereas di- and tetra-nucleotides were more frequent in UTRs. These results agree with those reported previously in melon and other species (González-Ibeas et al., 2007; Ueno et al., 2010).

We selected a set of 43 ESTs-SSRs for validation, all located in ORFs, 40 (93%) amplified polymorphic fragments in a set of 10 genotypes of *Cucumis melo*, belonging to different subspecies and morphotypes of the species (three *inodorus*: T111, Piñonet, and TamDew; three *cantalupensis*: Vedrantais, Dulce, and Noy Israel; two *conomon*: PI161375 and Pat81; and two *momordica*: PI124112 and PI414723). Details of these validated SSRs are included in the Supplemental File S9.

On average we found 3.6 alleles (from 2 to 7) per primer pair and PIC values ranged from 0.25 to 0.85 (mean = 0.54). Most SSRs (92%) were useful to detect variability between subsp. *melo* and *agrestis*, with unique alleles of one or both subspecies. The genetic relationships among accessions based on SSRs were

investigated by cluster analysis (Fig. 4). The dendrogram fits very well previous classifications with different marker systems (Esteras et al., 2011). Simple sequence repeats successfully separated accessions of both subspecies and differentiated all the analyzed genotypes. Eighty, sixty-three, and twelve percent were polymorphic within the *momordica*, *cantalupensis*, and *conomon* groups, respectively, and 15% detected variation between the two genetically close Piel de sapo types. These SSRs can be useful for fingerprinting genetically close commercial lines and also for mapping purposes, as T111, Vedrantais, Dulce, PI161375, PI124112, and PI414723 are parentals of several melon genetic maps (Périn et al., 2002: Perchepied et al., 2005; González-Ibeas et al., 2007; Harel Beja et al., 2010).

Massive sequencing of ESTs in a number of diverse genotypes has been previously used for developing large SNPs collections (Barbazuk et al., 2006; Novaes et al., 2008; Trick et al., 2009; Guo et al., 2010). Since the ESTs generated under the present study, using the 454 Sequencing technology, are from four different genotypes belonging to the two subspecies, and the previously available Sanger sequences come from six additional genotypes, we expected SNPs to be frequent in our collection. The SNP calling was done with the default parameters recommended by the ngs_backbone software (Bioinformatics at the Institute for the Conservation and Breeding of Agricultural Biodiversity, 2010). A total of 810,882 sequences were mapped to the assembled transcriptome for SNPs identification (84% 454 and 26% Sanger). The mapped reads were mostly from the *cantalupensis* Vedrantais (35.8%), the two *inodorus* Piel de sapo cultivars, Piñonet (28.7%) and T111 (17.7%), and the *conomon* PI 161375 (13.9%). The remainder genotypes were less represented: less than 1% the *cantalupensis* C35 and Dulce and the *conomon* Pat81 and less than 0.5% the *cantalupensis* Noy Israel and Charentais and the *inodorus* TamDew.

Stringent quality criteria were used for distinguishing sequence variations from sequencing errors and mutations introduced during the cDNA synthesis step. Only variations with allele and mapping quality over the established thresholds were annotated. By applying these criteria, we initially identified a total of 38,587 SNPs and 5795 indels distributed in 14,417 unigenes (27.1%), averaging a total of 3.1 single variations per unigene. Of this only 3728 (8%) were detected exclusively with Sanger sequences, whereas the vast majority were detected using only 454 or combining 454 and Sanger. Within the detected SNPs, transitions (74%) were much more common than transversions (26%) (Table 4). A similar number of A/G and C/T transitions and also similar percentages of the four transversion types (A/T, A/C, G/T, and C/G) were found. A set of SNPs could be accurately located with respect to putative initiation and termination codons, being mostly located in ORFs (82%). Detailed information about the identified SNPs and indels is included in the Supplemental File S10, including the nucleotidic change, the number of alleles, and the allelic frequency in each genotype. Different filters were applied to facilitate the management of the identified variants. Indels can be distinguished from SNPs (those with VKS filter tag [it is not an SNP] in Supplemental File S10). Other filters were applied for the in silico selection of the SNPs to identify the ones more suited for different research purposes. Unigenes were mapped to the available draft genome sequence of melon. All SNPs mapping in nonunique or noncontiguous genomic regions can be filtered out (those with UCR filter tag in Supplemental File S10) (9179, or 24%). Also, all SNPs located in sequences with more than four SNPs or indels per 100 bp (with

Table 2. Simple sequence repeat (SSR) statistics.

Di-nucleotide repeat	Number of di-SSRs†	%
AG	680	75
AT	166	18
AC	60	7
Total	906	100
Tri-nucleotide repeat	**Number of tri-SSRs**	
AAG	1582	65
AGG	178	7
ATC	174	7
AAT	148	6
AAC	114	5
Other tri-nucleotide repeats (% ≤ 5 for each one) AGC, ACC, ACG, CCG, and ACT	237	10
Total	2433	100
Tetra-nucleotide repeat	**Number of tetra-SSRs**	**%**
AAAG	120	47
AAAT	34	13
AAAC	19	8
AAGG	15	6
Other tetranucleotide repeats (% ≤ 5 each one) ACTC, AACC, ACAG, AGGC, AACG, AATT, AGCC, AGCG, AAGC, AGGG, AGAT, AAGT, ACCG, ATCG, ATCC, AATC, ACAT, and AATG	66	26
Total	254	100

†The number of di-, tri-, and tetra-nucleotide repeats identified in the *Cucumis melo* unigene dataset is shown for the complete set of putative SSRs.

Table 3. Localization of simple sequence repeats (SSRs) with respect to putative initiation and termination codons in the *Cucumis melo* unigene dataset. Unigenes were checked for the presence of the start and stop codons.

	Di-SSRs		Tri-SSRs		Tetra-SSRs		All SSRs	
	No.	%	No.	%	No.	%	No.	%
5′-UTR†	213	23.5	281	11.6	90	35.4	584	16.3
ORF‡	132	14.6	1177	48.4	44	17.3	1353	37.7
3′-UTR	436	48.1	686	28.2	88	34.6	1210	33.7
Other§	125	13.8	289	11.9	32	12.6	446	12.4
Total	**906**	**100**	**2433**	**100**	**254**	**100**	**3593**	**100**

†UTR, untranslated region.

‡ORF, open reading frame.

§Other means imprecise localization of the SSRs with respect to putative initiation or termination codons.

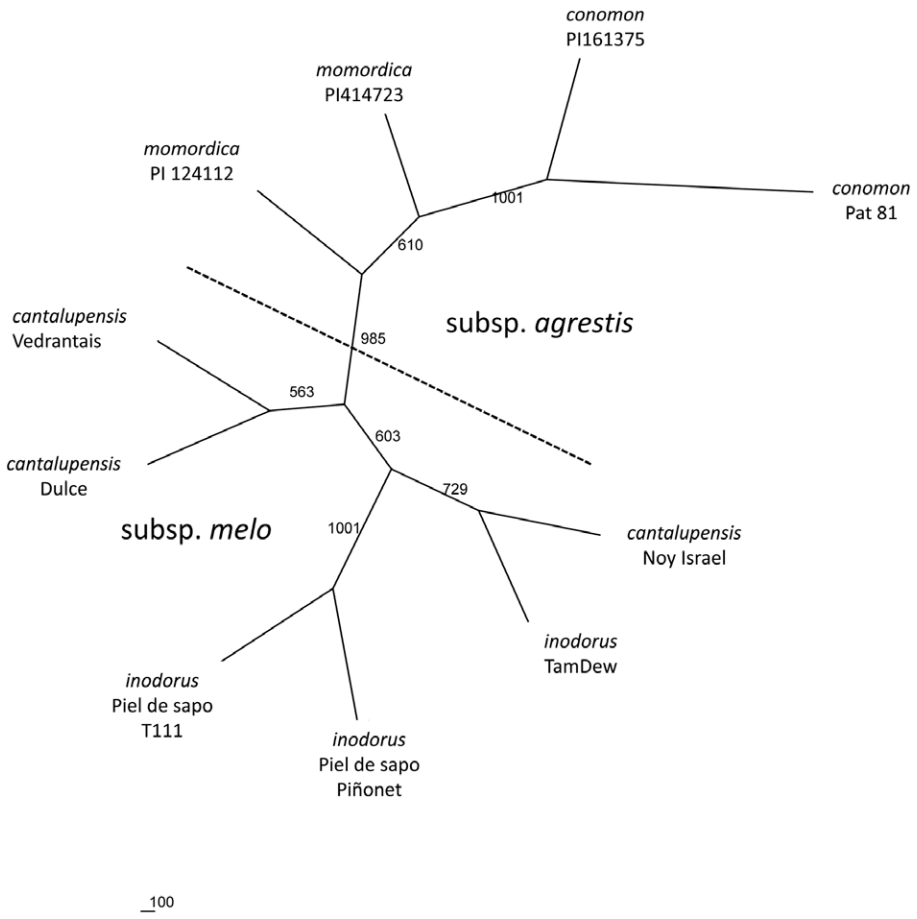

Figure 4. Genetic relationships among *Cucumis melo* genotypes analyzed with expressed sequence tag (EST)-simple sequence repeats (SSRs). Cluster analysis of the genetic relationships among a set of 10 *C. melo* genotypes belonging to two varieties of subsp. *agrestis* (*conomon* and *momordica*) and two varieties of subsp. *melo* (*cantalupensis* Naudin and *inodorus* H. Jacq.). Genetic distance of Nei et al. (1983) and bootstrap (resampling 1000 times) were calculated.

Table 4. Single nucleotide polymorphism (SNP) statistics.

SNPs	Number	SNPs†	Number	Number
Transitions		Transversions		Complex
A<->G	13,911	A<->T	3347	70
C<->T	14,716	G<->T	2650	
		C<->G	1866	
		A<->C	2027	
Total	28,627 (74.2%)	Total	9890(25.6%)	70 (0.2%)

†Type and number of transition and transversions are shown for putative high quality SNPs identified in the *Cucumis melo* database.

HVR4) may be discarded (986, or 3%). These requirements allow minimizing the discovery of false polymorphisms due to the alignment of paralogs, a potentially significant problem when aligning short sequence reads. Therefore, only nucleotide variants in relatively conserved or recently derived paralogs may have been incorrectly identified as SNPs. The drawback is that some true SNPs in hotspots of genetic diversity or genes under high diversifying selection may be discarded.

From the remaining 28,832 high-confidence SNPs, we implemented different filters to facilitate selection of those appropriated for detecting variability within or between different groups of genotypes:

1. Single nucleotide polymorphisms that are not sequenced or are variable (have more than one allele) within the genotypes Piñonet, T111, Vedrantais, Dulce, PI161375, and Pat81 can be filtered out using NVPiñ, NVT111, NVVed, NVDul, NVPI161, and NVPat81, respectively (those having the corresponding filter tag in Supplemental File S10). The absence of these filters indicates that the SNP is sequenced and has only one allele within each genotype.
2. Single nucleotide polymorphisms that are not sequenced or are variable within groups of genotypes—all the subsp. *melo* (T111, Pinonet, Dulce, Vedrantais, and Charentais), the Piel de sapo (T111 and Piñonet), the *cantalupensis* (Vedrantais, Charentais, and Dulce), and the *conomon* (PI161375 and Pat81)—can be filtered out with NVmelo, NVPs, NVCant, and NVCon. The absence of these filters indicates that the SNP is sequenced and has only one allele within each group.
3. Single nucleotide polymorphisms that are not sequenced or are uniform (have only one allele) in groups of genotypes—the *cantalupensis* and *conomon* (Dulce, Vedrantais, Charentais, Pat81,

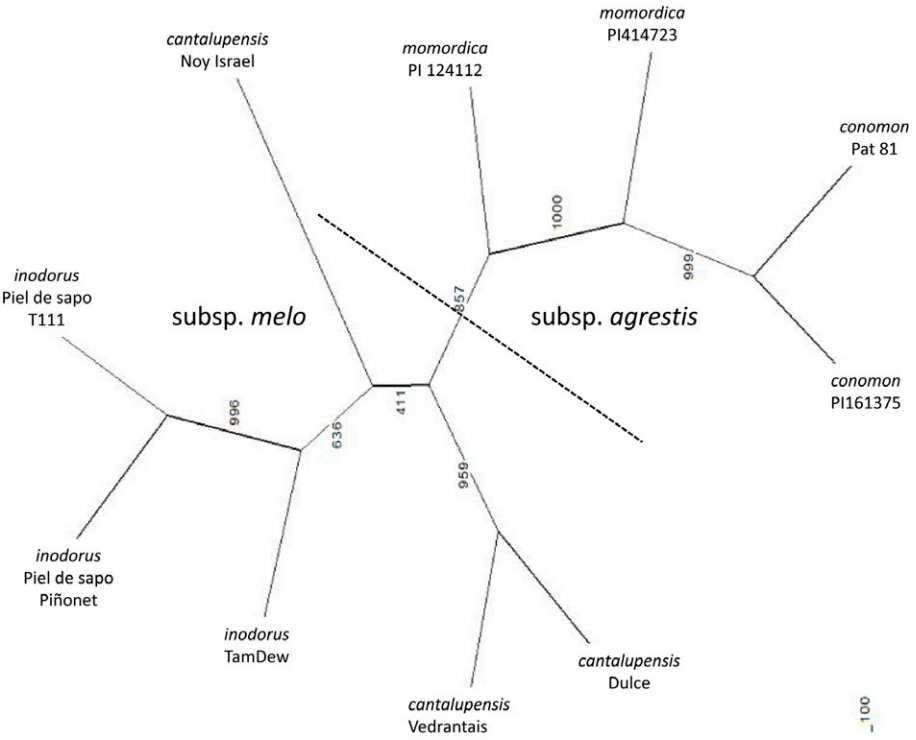

Figure 5. Genetic relationships among *Cucumis melo* genotypes analyzed with single nucleotide polymorphisms (SNPs)-cleaved amplified polymorphic sequence (CAPS) (for genotypes and cluster analysis see Fig. 4).

and PI161375) and the Piel de sapo and *conomon* (T111, Piñonet, Pat81, and PI161375)—can be filtered out with VSCant-Con and VSPs-Con. The absence of these filters indicates that the SNP is sequenced and is variable within the group.

Combining these filters we can, for example, select those SNPs that are sequenced and are uniform within each Piel de sapo genotype (lacking NVPiñ and NVT111 in Supplemental File S10), but are variable between them (with NVPs). One thousand, seven hundred eighty-eight SNPs met those criteria (288 with two or more reads per allele), being potentially useful for fingerprint genetically close varieties belonging to this main commercial market class. Only 24% of these SNPs could have been detected only from Sanger sequences, which reinforces the importance of deep sequencing for SNPs detection.

Among those SNPs (5980) that were uniform within and between Piel de sapo cultivars (lacking NVPiñ, NVT111, and NVPs), 2122 were polymorphic between Piel de sapo and one or both *conomon* sequenced genotypes (lacking NVCon and VSPs-Con). Another SNP set (3235) was selected for detecting variation between *cantalupensis* and *conomon,* being uniform within each group (lacking VSCant-Con, NVCon, and NVCant). These sets of markers could be more efficient for mapping purposes, using crosses between Piel de sapo × conomon or *cantalupensis* × *conomon*, using as parentals not only the sequenced accessions but also additional accessions of each group.

Another set of filters allow to select those SNPs that met different criteria for facilitating validation and for their use in a Golden Gate genotyping assay (Fan et al., 2006; Gupta et al., 2008), discarding those that were closer than 60 bp to another SNP or indel (CS60) and/or were predicted to be located closer than 59 bp to an intron (I59) and/or were closer than 60 bp to the unigene edge (CL60). Finally, 11,655 SNPs were selected that met all criteria. From these, 453 SNPs were identified that can be detected as CAPS (lacking nCAP filter tag) as they generate allele-specific restriction targets.

We selected 45 of these putative CAPS (all with a minimum of two reads per allele) for validation. All amplified in the corresponding genotypes and were validated by sequencing. Eighty nine percent (40) provided the expected polymorphism affecting the corresponding restriction target. All were also validated as CAPS, 98% being polymorphic between the expected genotypes after digestion with the corresponding enzyme. This percentage of validation is comparable to that reported in previous studies (Novaes et al., 2008; Blanca et al., 2011). The genetic relationships among accessions based on CAPS were also investigated by cluster analysis. Results were similar to those obtained with SSRs. Cleaved amplified polymorphic sequences efficiently differentiate all cultivars, even the genetically close Piel de Sapo, and separate accessions of both subspecies (Fig. 5). Information of the validated SNPs is included in Supplemental File S11. These CAPS markers are especially useful when experience or equipment for SNPs detection using other methods is not available. All annotation results (ORFs, introns, descriptions, GO terms, *Arabidopsis thaliana*, melon orthologs, SNVs,

and SSRs) have been also added in Supplemental File S12 using the GFF3 standard file format of the Sequence Ontology Resources (Sequence Ontology Project, 2010; Eilbeck et al., 2005).

CONCLUSIONS

Our results demonstrate how massive sequencing can contribute to improve the melon Sanger-based transcriptome, providing new unigenes that might represent newly detected transcripts. The use of GS FLX Titanium reads and the combined assembly with longer Sanger ESTs has generated long unigenes, allowing an accurate annotation. The detailed annotation provided in this paper will facilitate the use of the collection for gene discovery. It is also being used for genome annotation. Also, the markers collection provided here will be useful for mapping purposes and genotyping studies. The fact that these markers have been detected in a wide set of genotypes of this highly variable species make them suitable to design genotyping assays with broad utility. The filtering process is also very useful to optimize their use with high throughput genotyping platforms.

Supplemental Information Available

Supplemental material is available free of charge at http://www.crops.org/publications/tpg.

Supplementary File S1 (excel xls). Sanger ESTs Statistics. Number of sequences, average length, quality and total length are indicated for Sanger sequences used in this study. Initial data and data after processing are indicated. Data are presented for the different libraries, including tissue and genotype information. These data were produced within the Melogen and ICUGI initiatives and are available at the ICUGI webpage.

Supplementary File S2 (fasta format zip comprised). *C. melo* unigenes. The fasta sequence of the 53,252 *C. melo* unigenes assembled from 454 and Sanger ESTs is included.

Supplementary File S3 (excel xls). Annotation of ORFs and introns. Unigene length and predicted position of ORFs and introns is indicated for the whole *C. melo* unigene collection

Supplementary File S4 (excel xls). Blast Hits. Descriptions build from the blast hit obtained by a sequential blast search of three protein databases Swissprot, *Arabidopsis* protein, and Uniref90 for the whole collection unigene.

Supplementary File S5 (power point ppt). Distribution of GO terms. The unigenes distribution regarding the number of GOs to which they were assigned is shown.

Supplementary File S6 (excel xls). GO terms. GO annotations for the whole *C. melo* unigene collection are compiled.

Supplementary File S7 (excel xls). Arabidopsis orthologs and melon equivalents. Orthologs found by reciprocal search with *Arabidopsis* and melon database ICuGI are indicated.

Supplementary File S8 (excel xls). *C. melo* SSRs. The table provides the list of SSRs identified from the *C. melo* unigene dataset, their length, motif sequences, position in the unigene, and scores.

Supplementary File S9 (excel xls). Validated *C. melo* SSRs. The table provides the list of SSRs experimentally validated using a collection of *C. melo*. Primers used for validation, number of alleles, frequency of the most common allele, polymorphism information content (PIC) across species, and polymorphism detected between species and within Piel de sapo, *cantalupensis*, *conomon*, and *momordica* groups.

Supplementary File S10 (excel xls). *C. melo* SNPs and INDELs. The table provides the list of INDELs and SNPs identified from the *C. melo* unigene dataset, their position, the nucleotide change, insertion or deletion and the quality of the polymorphic base. The different filters applied for the in silico selection of the SNPs and the number of reads per genotype and allele are also indicated. Filter description is included in the Methods section.

Supplementary File S11 (excel xls). Validated *C. melo* SNP-CAPS. The table provides the list of SNPs that affect restriction targets and were validated by sequencing and via CAPS, their position, location, primers used for validation. For those validated as CAPS also frequency of the most common allele and polymorphism information content (PIC) across species is indicated.

Supplementary File S12 (GFF3 format, comprised with zip). Annotation results in GFF3. All the annotation results (ORFS, introns, descriptions, GO terms, *Arabidopsis*, and melon orthologs, SNVs, and SSRs) are provided also in the standard format GFF3 of The Sequence Ontology Resources.

Authors' Contributions

CE, GM, JC, and BP prepared the cDNA libraries for sequencing. BP and CE selected and validated the SSRs, SNPs, and CAPS. JB and PZ performed the bioinformatic analysis. BP, JB, JC, JG and FN participated in the conception of the study. BP was primarily responsible for drafting and revising the manuscript with contributions from coauthors. All authors read and approved the final manuscript.

Acknowledgments

This project was carried out in the frame of the MELONOMICS project (2009-2012) of the Fundación Genoma España. Authors thank Cristina Roig for providing technical help in markers validation.

References

Abajian, C. 1994. Sputnik: DNAS microsatellite repeat search utility. Available at http://espressosoftware.com/sputnik/index.html (verified 24 May 2010). Chris Abajian, Seattle, WA.

Alagna, F., N. D'Agostino, L. Torchia, M. Servili, R. Rao, M. Pietrella, G. Giuliano, M.L. Chiusano, L. Baldoni, and G. Perrotta. 2009. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics 10:399. doi:10.1186/1471-2164-10-399

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402. doi:10.1093/nar/25.17.3389

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815. doi:10.1038/35048692

Arabidopsis Information Resource. 2009. The TAIR database. Available at http://www.arabidopsis.org/ (verified 24 May 2011). Carnegie Institution, Department of Plant Biology, Stanford, CA.

Barakat, A., D.S. DiLoreto, Y. Zhang, C. Smith, K. Baier, W.A. Powell, N. Wheeler, R. Sederoff, and J.E. Carlson. 2009. Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. BMC Plant Biol. 9:51. doi:10.1186/1471-2229-9-51

Barbazuk, W.B., S.J. Emrich, H.D. Chen, L. Li, and P.S. Schnable. 2006. SNP discovery via 454 transcriptome sequencing. Plant J. 51:910–918. doi:10.1111/j.1365-313X.2007.03193.x

Bioinformatics at the Institute for the Conservation and Breeding of Agricultural Biodiversity (COMAV). 2010. Ngs_backbone. Available at http://bioinf.comav.upv.es/ngs_backbone (verified 24 May 2011). COMAV, Universitat Politècnica de València, Valencia, Spain.

Blanca, J.M., J. Cañizares, C. Roig, P. Ziarsolo, F. Nuez, and M.B. Picó. 2011. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics 12:104. doi:10.1186/1471-2164-12-104

Boualem, A., M. Fergany, R. Fernandez, C. Troadec, A. Martin, H. Morin, M.A. Sari, F. Collin, J.M. Flowers, M. Pitrat, M.D. Purugganan, C. Dogimont, and A. Bendahmane. 2008. A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons. Science 321:836–838. doi:10.1126/science.1159023

Cheung, F., B.J. Haas, S.M.D. Goldberg, G.D. May, Y. Xiao, and C.D. Town. 2006. Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. BMC Genomics 7:272. doi:10.1186/1471-2164-7-272

Chevreux, B., T. Pfisterer, B. Drescher, A.J. Driesel, W.E. Müller, T. Wetter, and S. Suhai. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14:1147–1159. doi:10.1101/gr.1917404

Chou, H.H, and M.H. Holmes. 2001. DNA sequence quality trimming and vector removal. Bioinformatics 17:1093–1104.

Conesa, A., and S. Götz. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics, 2008:619832 doi:10.1155/2008/619832.

Deleu, W., C. Esteras, C. Roig, M. González-To, I. Fernández-Silva, J. Blanca, M.A. Aranda, P. Arús, F. Nuez, A.J. Monforte, M.B. Picó, and J. Garcia-Mas. 2009. A set of EST-SNPs for map saturation and cultivar identification in melon. BMC Plant Biol. 9:90. doi:10.1186/1471-2229-9-90

Eduardo, I., P. Arús, A.J. Monforte, J. Obando, J.P. Fernández-Trujillo, J.A. Martinez, A.L. Alarcon, J.M. Alvarez, and E. van der Knapp. 2007. Estimating the genetic architecture of fruit quality traits in melon using a genomic library of near isogenic lines. J. Am. Soc. Hortic. Sci. 132:80–89.

Eilbeck, K., S. Lewis, C.J. Mungall., M. Yandell, L. Stein, R, Durbin, and M. Ashburner. 2005. The sequence ontology: A tool for the unification of genome annotations. Genome Biol. (2005)6:R44.

Emrich, S.J., W.B. Barbazuk, L. Li, and P.S. Schnable. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res. 17:69–73. doi:10.1101/gr.5145806

Esteras, C., F. Nuez, and M.B. Picó. 2011. Genetic diversity studies in Cucurbits using molecular tolos. p. 25. *In* Y. Wang and T.K Behera (ed.) Cucurbits: Genetics, genomics and breeding in crop plants. Science Publishers, Enfield, NH.

European Bioinformatics Institute. 2001. The Emboss: est2genome. Available at http://emboss.sourceforge.net/apps/cvs/emboss/apps/est2%20genome.html (not verified) European Bioinformatics Institute, Hinxton, UK.

European Bioinformatics Institute. 2010. UniProt reference clusters (UniRef) databases. Available at http://www.ebi.ac.uk/uniref (verified 24 May 2011). European Bioinformatics Institute, Hinxton, UK.

Exonerate. 2005. A generic tool for sequence alignment. Available at http://www.ebi.ac.uk/~guy/exonerate/ (verified 25 May 2011). European Bioinformatics Institute, Hinxton, UK.

Ezura, H., and N. Fukino. 2009. Research tools for functional genomics in melon (*Cucumis melo* L.): Current status and prospects. Plant Biotechnol. 26:359–368. doi:10.5511/plantbiotechnology.26.359

Fan, J.B., M.S. Chee, and K.L. Gunderson. 2006. Highly parallel genomic assays. Nat. Rev. Genet. 7:632–644. doi:10.1038/nrg1901

Folta, K.M., M.A. Clancy, S. Chamala, A.M. Brunings, A. Dhingra, L. Gomide, R.J. Kulathinal, N. Peres, T.M. Davis, and W.B. Barbazuk. 2010. A transcript accounting from diverse tissues of a cultivated strawberry. Plant Gen. 3:90–105. doi:10.3835/plantgenome2010.02.0003

Gedye, K., J. Gonzalez-Hernandez, Y. Ban, X. Ge, J. Thimmapuram, F. Sun, Ch. Wright, S. Ali, A. Boe, and V. Owens. 2010. Investigation of the transcriptome of prairie cord grass, a new cellulosic biomass crop. Plant Gen. 3:69–80. doi:10.3835/plantgenome2010.06.0012

González, V.M., J. Garcia-Mas, P. Arús, and P. Puigdomènech. 2010a. Generation of a BAC-based physical map of the melon genome. BMC Genomics 11:339. doi:10.1186/1471-2164-11-339

González, V.M., G. Mir, P. Arús, P. Puigdomenech, and J. Garcia-Mas. 2009. Abstract: Towards the whole sequence of the melon genome. *In* Plant Animal Genome XVII Conf., San Diego, CA. 10–14 Jan. 2009. Publisher unknown.

González, V.M., L. Rodríguez-Moreno, E. Centeno, A. Benjak, J. Garcia-Mas, P. Puigdomènech, and M.A. Aranda. 2010b. Genome-wide BAC-end sequencing of *Cucumis melo* using two BAC libraries. BMC Genomics 11:618. doi:10.1186/1471-2164-11-618

González-Ibeas, D., J. Blanca, C. Roig, M. Gonzalez-To, M.B. Pico, V. Truniger, P. Gomez, W. Deleu, A. Cano-Delgado, P. Arus, F. Nuez, J. Garcia-Mas, P. Puigdomènech, and M.A. Aranda. 2007. Melogen: An EST database for melon functional genomics. BMC Genomics 8:306. doi:10.1186/1471-2164-8-306

Guo, S., Y. Zheng, J.G. Joung, S. Liu, Z. Zhang, O.R. Crasta, B.W. Sobral, Y. Xu, S. Huang, and Z. Fei. 2010. Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. BMC Genomics 11:384. doi:10.1186/1471-2164-11-384

Gupta, P.K., S. Rustgi, and R.R. Mir. 2008. Array-based high throughput DNA markers for crop improvement. Heredity 101:5–18. doi:10.1038/hdy.2008.35

Harel-Beja, R., G. Tzuri, V. Portnoy, M. Lotan-Pompan, S. Lev, S. Cohen, N. Dai, L. Yeselson, A. Meir, S. Libhaber, E. Avisar, T. Melame, P. Van Koert, H. Verbakel, R. Hofstede, H. Volpin, M. Oliver, A. Fougedoire, C. Stalh, J. Fauve, B. Copes, Z. Fei, J. Giovannoni, N. Ori, E. Lewinsohn, A. Sherman, J. Burger, Y. Tadmor, A. Schaffer, and N. Katzir. 2010. A genetic map of melon highly enriched with fruit quality QTLs and EST markers, including sugar and carotenoid metabolism genes. Theor. Appl. Genet. 12:511–533. doi:10.1007/s00122-010-1327-4

Huang, S., R. Li, Z. Zhang, L. Li, X. Gu, W. Fan, W.J. Lucas, X. Wang, B. Xie, P. Ni, Y. Ren, H. Zhu, J. Li, K. Lin, W. Jin, Z. Fei, G. Li, J. Staub, A. Kilian, E.A. van der Vossen, Y. Wu, J. Guo, J. He, Z. Jia, Y. Ren, G. Tian, Y. Lu, J. Ruan, W. Qian, M. Wang, Q. Huang, B. Li, Z. Xuan, J. Cao, Asan, Z. Wu, J. Zhang, Q. Cai, Y. Bai, B. Zhao, Y. Han, Y. Li, X.- Li, S. Wang, Q. Shi, S. Liu, W.K. Cho, J.Y. Kim, Y. Xu, K. Heller-Uszynska, H. Miao, Z. Cheng, S. Zhang, J. Wu, Y. Yang, H. Kang, M. Li, H. Liang, X. Ren, Z. Shi, M. Wen, M. Jian, H. Yang, G. Zhang, Z. Yang, R. Chen, S. Liu, J. Li, L. Ma, H. Liu, Y. Zhou, J. Zhao, X. Fang, G. Li, L. Fang, Y. Li, D. Liu, H. Zheng, Y. Zhang, N. Qin, Z. Li, G. Yang, S. Yang, L. Bolund, K. Kristiansen, H. Zheng, S. Li, X. Zhang, H. Yang, J. Wang, R. Sun, B. Zhang, S. Jiang, J. Wang, Y. Du, and S. Li. 2009. The genome of the cucumber, *Cucumis sativus* L. Nat. Genet. 41:1275–1281. doi:10.1038/ng.475

International Cucurbit Genomics Initiative (ICuGI). 2007. Cucurbit genomics database. Available at http://www.icugi.org (verified 24 May 2011). USDA-ARS, Boyce Thompson Institute, Cornell Univ., Ithaca, NY.

Iseli, C., C.V. Jongeneel, and P. Bucher. 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol.1999:138–48.

Li, H., and R. Durbin. 2010. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. doi:10.1093/bioinformatics/btp698

Li, Y., H.M. Luo, C. Sun, J.Y. Song, Y.Z. Sun, Q. Wu, N. Wang, H. Yao, A. Steinmetz, and S.L. Chen. 2010. EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. BMC Genomics 11:268. doi:10.1186/1471-2164-11-268

Liu, K., and S.V. Muse. 2005. Powermarker: Integrated analysis environment for genetic marker data. Bioinformatics 21:2128–2129. doi:10.1093/bioinformatics/bti282

Mascarell-Creus, A., J. Cañizares, J. Vilarrasa, S. Mora-García, J. Blanca, D. González-Ibeas, M. Saladié, C. Roig, W. Deleu, B. Picó, N. López-Bigas, M.A. Aranda, J. Garcia-Mas, F. Nuez, P. Puigdomènech, and A. Caño-Delgado. 2009. An oligo-based microarray offers novel transcriptomic approaches for the analysis of pathogen resistance and fruit quality traits in melon (*Cucumis melo* L.). BMC Genomics 10:467. doi:10.1186/1471-2164-10-467

Melogen. 2003. Development of genomics tools in melon. Available at http://www.melogen.upv.es (verified 25 May 2011) COMAV, Universitat Politècnica de València, Valencia, Spain.

Morgante, M., M. Hanafey, and W. Powell. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. 30:194–200. doi:10.1038/ng822

National Center for Biotechnology Information (NCBI). 2010a. Nonredundant protein database. Available at ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz (verified 24 May 2011). NCBI, Bethesda, MD.

National Center for Biotechnology Information (NCBI). 2010b. Sequence read archive. Available at http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi (verified 24 May 2011). NCBI, Bethesda, MD.

Nei, M., F. Tajima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from molecular data. J. Mol. Evol. 19:153–170. doi:10.1007/BF02300753

Novaes, E., D. Drost, W. Farmerie, G. Pappas, D. Grattapaglia, R.R. Sederoff, and M. Kirst. 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9:312. doi:10.1186/1471-2164-9-312

Périn, C., M. Gomez-Jimenez, L. Hagen, C. Dogimont, J.C. Pech, A. Latché, M. Pitrat, and J.M. Lelièvre. 2002. Molecular and genetic characterization of a non-climacteric phenotype in melon reveals two loci conferring altered ethylene response in fruit. Plant Physiol. 129:300–309. doi:10.1104/pp.010613

Perchepied, L., M. Bardin, C. Dogimont, and M. Pitrat. 2005. Relationship between loci conferring downy mildew and powdery mildew resistance in melon assessed by quantitative trait loci mapping. Phytopathology 95:556–565. doi:10.1094/PHYTO-95-0556

Pitrat, M. 2008 Melon (*Cucumis melo* L.). p. 283–315. *In* J. Prohens and F. Nuez (ed.) Handbook of plant breeding. Vegetables (vol. I). Springer, New York, NY.

Rozen, S., and H.J. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. pp 365–386. *In* S. Krawetz and S. Misener (ed) Bioinformatics methods and protocols: Methods in molecular biology. Humana Press, Totowa, NJ.

Schaefer, H., C. Heibl, and S.S. Renner. 2009. Gourds afloat: A dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. Proc. Biol. Sci. 276:843–851. doi:10.1098/rspb.2008.1447

Sequence Ontology Project. 2010. Generic feature format version 3. Available at http://www.sequenceontology.org/gff3.shtml (verified 24 May 2011). Karen Eilbeck, Salt Lake City, UT.

Shendure, J., and H. Ji. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26:1135–1144. doi:10.1038/nbt1486

Sun, Ch., Y. Li, Q. Wu, H. Luo, Y. Sun, J. Song, E.M.K. Lui, and S. Chen. 2010. De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11:262. doi:10.1186/1471-2164-11-262

Swarbreck, D., C. Wilks, P. Lamesch, T.A. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz et al. 2008. The Arabidopsis information resource (TAIR): Gene structure and function annotation. Nucleic Acids Res. 36:1009–1014.

Thiel, T., W. Michalek, R.K. Varshney, and A. Graner. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet. 106:411–422.

Trick, M., Y. Long, J. Meng, and I. Bancroft. 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol. J. 7:334–346. doi:10.1111/j.1467-7652.2008.00396.x

Ueno, S., G. Le Provost, V. Léger, Ch. Klopp, C. Noirot, J.M. Frigerio, F. Salin, J. Salse, M. Abrouk, F. Murat, O. Brendel, J. Derory, P. Abadie, P. Léger, C. Cabane, A. Barré, A. Daruvar, A. Couloux, P. Wincker, M.P. Reviron, A. Kremer, and Ch. Plomion. 2010. Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species. Oak BMC Genomics 11:650. doi:10.1186/1471-2164-11-650

UniProt Consortium. 2010a. The UniProtKB/Swiss-Prot database. Available at http://www.uniprot.org/downloads (verified 24 May 2011). Uniprot Consortium, UK, Switzerland, United States.

UniProt Consortium. 2010b. The universal protein resource (UniProt) in 2010. Nucleic Acids Research 38:142–148.

Vega-Arreguin, J.C., E. Ibarra-Laclette, B. Jimenez-Moraila, O. Martinez, J.P. Vielle-Calzada, L. Herrera-Estrella, and A. Herrera-Estrella. 2009. Deep sampling of the Palomero maize transcriptome by a high throughput strategy of pyrosequencing. BMC Genomics 10:299. doi:10.1186/1471-2164-10-299

Wang, W., Y. Wang, Q. Zhang, Y. Qi, and D. Guo. 2009. Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing. BMC Genomics 10:465. doi:10.1186/1471-2164-10-465

Weber, A.P., K.L. Weber, K. Carr, C. Wilkerson, and J.B. Ohlrogge. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol. 144:32–42. doi:10.1104/pp.107.096677