BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae)

José Blanca, Joaquín Cañizares, Cristina Roig, Pello Ziarsolo, Fernando Nuez, Belén Picó*

Abstract

Background: *Cucurbita pepo* belongs to the Cucurbitaceae family. The "Zucchini" types rank among the highest-valued vegetables worldwide, and other *C. pepo* and related *Cucurbita spp.*, are food staples and rich sources of fat and vitamins. A broad range of genomic tools are today available for other cucurbits that have become models for the study of different metabolic processes. However, these tools are still lacking in the *Cucurbita* genus, thus limiting gene discovery and the process of breeding.

Results: We report the generation of a total of 512,751 *C. pepo* EST sequences, using 454 GS FLX Titanium technology. ESTs were obtained from normalized cDNA libraries (root, leaves, and flower tissue) prepared using two varieties with contrasting phenotypes for plant, flowering and fruit traits, representing the two *C. pepo* subspecies: subsp. *pepo* cv. Zucchini and subsp. *ovifera* cv Scallop. *De novo* assembling was performed to generate a collection of 49,610 *Cucurbita* unigenes (average length of 626 bp) that represent the first transcriptome of the species. Over 60% of the unigenes were functionally annotated and assigned to one or more Gene Ontology terms. The distributions of *Cucurbita* unigenes followed similar tendencies than that reported for *Arabidopsis* or melon, suggesting that the dataset may represent the whole *Cucurbita* transcriptome. About 34% unigenes were detected to have known orthologs of *Arabidopsis* or melon, including genes potentially involved in disease resistance, flowering and fruit quality. Furthermore, a set of 1,882 unigenes with SSR motifs and 9,043 high confidence SNPs between Zucchini and Scallop were identified, of which 3,538 SNPs met criteria for use with high throughput genotyping platforms, and 144 could be detected as CAPS. A set of markers were validated, being 80% of them polymorphic in a set of variable *C. pepo* and *C. moschata* accessions.

Conclusion: We present the first broad survey of gene sequences and allelic variation in *C. pepo*, where limited prior genomic information existed. The transcriptome provides an invaluable new tool for biological research. The developed molecular markers are the basis for future genetic linkage and quantitative trait *loci* analysis, and will be essential to speed up the process of breeding new and better adapted squash varieties.

Background

The botanical family Cucurbitaceae, commonly known as cucurbits, includes several economically and nutritionally important vegetable crops cultivated worldwide, such as cucumber, melon, watermelon and pumpkins, gourds and squashes [1]. The cucurbit family displays a rich diversity of many traits, being primary models for sex expression analysis, for the study of vascular biology and for the analysis of the mechanisms involved in fruit ripening [2-5].

Despite the agricultural and biological importance of cucurbits, knowledge of their genetics and genome has been very limited till now. So far, genomic efforts have largely focused on cucumber and melon. Recently, the whole genome sequencing of the cucumber, *C. sativus* var. *sativus* L., has been completed by combining traditional Sanger and next-generation Illumina GA sequencing technologies [6]. Also an effort is in progress through a Spanish Initiative to obtain the whole genome sequence of melon, *Cucumis melo* L. [7]. Many genomic

* Correspondence: mpicosi@btc.upv.es
Institute for the Conservation and Breeding of Agricultural Biodiversity, Universidad Politécnica de Valencia (COMAV-UPV), Camino de Vera s/n, 46022 Valencia, Spain

resources are available for both crops and also for watermelon, *Citrullus lanatus* (Thunb.) Matsum. & Nakai. BAC libraries, collections of genetic markers, detailed physical and genetic maps, mapping populations, microarrays, sequence databases and mutant collections [8-11] are facilitating the use of cucurbits by the research community. Many genomic resources are available at the web site of the International Cucurbit Genomics Initiative (ICuGI) [12].

Cucurbita genus (2n = 2 × = 40), that include squashes, gourds and pumpkins, has been less studied. It contains some of the earliest domesticated plant species [13]. Today, three of them, *C. pepo* L., *C. moschata* Duchesne, and *C. maxima* Duchesne, have considerable impact on human nutrition, being appreciated by their nutritional and medical properties [14-17]. In addition to the use of the edible fruits, flowers, leaves, and vine tips are consumed, and seeds are also important as snacks, as a source of edible oil and protein for human and animal consumption, and in the pharmaceutical industry. Squashes are also popular as containers and for ornamental purposes. The economic value of *Cucurbita spp*. as rootstocks for overcoming soil borne diseases in cucurbits is significant [18].

C. pepo is the most economically important species within the genus distributed worldwide, and one of the most variable in the plant kingdom. Cultivated *C. pepo* is considered to comprise two subspecies each one including several cultivar-groups, ssp. *pepo* (Pumpkin, Vegetable Marrow, Cocozelle, and Zucchini) and ssp. *ovifera* (Acorn, Scallop, Crookneck, and Straightneck) [19,20]. Its great economic value is based mainly on the culinary use of the immature fruits as vegetables, often referred to collectively as "summer squashes", but also the Pumpkin and Acorn groups display a major use as mature fruits, known as "winter squashes". The great diversity of uses makes breeding objectives quite variable.

The currently available genetic and genomic tools for *Cucurbita* are very limited. Until now three genetic maps have been constructed: two maps from inter-specific crosses between *C. pepo* and *C. moschata* [21,22] and the third from an intra-specific cross of *C. pepo* (a USA oil-Pumpkin variety and an Italian Crookneck variety) [23]. These maps contained mostly RAPDs (Random Amplified Polymorphic DNA) and AFLPs (Amplified Fragment Length Polymorphism) markers. Only recently a collection of genomic microsatellites (Simple Sequence repeats, SSRs) has been developed and used to increase the map density [24]. The last map version comprises 178 SSRs, 244 AFLPs, 230 RAPDs, and two morphological traits (*h* (*hull-less* seed) and *Bu* (*Bush growth habit*). It contains 20 linkage groups with a map density of 2.9 cM and genome coverage of 86.8%. These SSRs were also used to study synteny between *C. pepo* and *C. moschata* [25].

The lack of denser genetic maps, larger high-throughput marker collections, and suited mapping populations is limiting gene isolation and squash breeding. Many *C. pepo* genes have been reported, mainly related to fruit quality and resistance to poty- and other viruses and several fungi, such as downy and powdery mildew [26], but only the transcripts of a few have been cloned and molecularly characterized in individual studies in *C. pepo* or other *Cucurbita* spp, for example genes involved in the biosynthesis or signaling pathways of growth regulators, affecting plant development, sex expression and response to stress [27-32].

Single nucleotide polymorphisms (SNPs) are the most abundant variations in genomes and, therefore, constitute a powerful tool for mapping and marker-assisted breeding. These markers are replacing microsatellites in many model and non-model plants for saturating genetic maps [10,33]. In genomes that have been poorly studied, sequence availability is the limiting factor for the discovery of SNPs.

Expressed sequenced tags (ESTs) represent a valuable sequence resource for research and breeding as they provide comprehensive information regarding the transcriptome. ESTs have played significant roles in accelerating gene discovery, allowing large-scale expression analysis, improving genome annotation, elucidating phylogenetic relationships and facilitating breeding programs for both plants and animals by providing SSRs and SNPs markers [6,8,11,34-37].

Currently, there are more than 66 million ESTs in the NCBI public collection [38]. However, less than 1,000 EST sequences are available for *Cucurbita* spp (*C. maxima*, *C. moschata* and *C. pepo*), and approximately 500,000 for all the species in the Cucurbitaceae family, most of them of cucumber and melon, included in the ICUGI Cucurbit Genomics Database [12], as compared to more than 1.5 and 2 million ESTs available for *Arabidopsis* and maize, respectively.

Recent advances in next-generation sequencing technologies allow us the large scale generation of ESTs efficiently and cost-effectively [39,40]. There are increasing studies in which 454 technologies, combined or not with Solexa/Illumina, are used to characterize transcriptomes in cereals and legumes [41-43]. Even in model species, such as *Arabidopsis*, this deep sequencing is allowing to identify new transcripts not present in previous ESTs collections [44]. Also specific transcriptomes are being generated in species for which previous genomic resources are lacking [45-47]. The new transcripts are being used for microarrays design [48], and also for high throughput SSRs or SNPs identification. SNP detection is performed by aligning raw reads from

different genotypes to a reference genome or transcriptome previously available, as in maize, cucumber and even in poliploid species such as *Brassica napus* [49-51]. *De novo* assembly of raw sequences coming from a set of genotypes, followed by pairwise comparison of the overlapping assembled reads has also successfully used in species lacking any significant genomic or transcriptomic resources [52].

In this study, we describe the generation of 49,610 *Cucurbita* unigenes *de novo* assembled from about 500.000 ESTs obtained from roots, leaves and flowers of two contrasting *C. pepo* cultivars (Zucchini and Scallop, belonging to the two *C. pepo* subspecies) using Roche/454 GS FLX Titanium massive parallel pyrosequencing technology. These unigenes are functionally annotated and represent the first *C. pepo* transcriptome. They have been also screened for SSR motifs and used to identify a large SNPs collection suited for high-throughput mapping purposes. This sequence will allow accelerating genetics and breeding of this crop. It is also an important advance for cucurbit genomics as it is the first genomic resource for this genus, allowing comparisons among members of the three most economically important cucurbit genera, *Cucumis*, *Citrullus* and *Cucurbita*.

Results and Discussion

EST sequencing and assembly

We performed a half 454 GS FLX Titanium run on each of the two libraries constructed from leaves, flowers and roots from two *C. pepo* cultivars with contrasting plant, flower and fruit phenotypes, MU16 (*C. pepo* subsp. *pepo* cv Zucchini) and UPV196 (*C. pepo* subsp. *pepo* cv Scallop). A total of 407,723 and 392,370 raw sequence reads were obtained from each library (Table 1). Raw reads were processed using the Ngs_backbone software [53] to eliminate adapter sequences, low quality chromatograms and sequences of less than 100 base pairs (bp). This analysis gave rise to 261,962 and 250,789 processed sequences, comprising 164.6 Mbp of sequence, with an average length of 318.7 and 323.4 bp, respectively. The length distribution of these expressed sequence tags (ESTs) is shown in Figure 1. More than 85% ESTs fell between 200 and 500 bp in length.

The reads produced by the GS FLX Titanium platform were used for clustering and *de novo* assembly, independently of the genotype of origin. 459,439 ESTs were finally assembled using the Mira assembler [54] yielding a total of 49,610 high-confident tentative consensus sequences (non-redundant sequences or unigenes). The distribution of the number of ESTs per unigene is shown in Figure 2. The majority of unigenes were assembled from a moderate number of ESTs (from 2 to 10), with an average of 9.3 ESTs per unigene. Of all unigenes, about a 10% contained more than 20 reads, and only 2.4% more than 50, which represented the most abundant transcripts in these cDNA libraries. This low redundancy is probably due to the success of the normalization process, responsible for the suppression of superabundant transcripts. Normalization precludes *in silico* analysis of gene expression, but greatly increases the number of unigenes that can be determined by reducing redundancy.

The assembled unigenes had an average length of 626 bp, comprising approximately 31 Mbp in total. The length distribution of the unigenes is shown in Figure 3. The analysis revealed that more of the 50% of unigenes were larger than 537 bp, and only a 5% of the sequences were shorter than 290 bp. The number of assembled unigenes is similar to that obtained in previous transcriptome analyses in maize, *Eucaliptus*, *Artemisia*, chesnut, olive and cucumber. However, the *de novo* assembly with the longer reads obtained with the GS FLX Titanium platform render unigenes that average almost two times longer than that reported in these studies performed using 454 GS-20 and GS-FLX platforms [45-47,51,52,55]. Our assembled unigenes were also larger than that reported for American ginseng and *Glycyrrhiza uralensis* transcriptomes obtained using also the 454 GS FLX Titanium platform [56,57]. *Cucurbita* unigenes length is comparable to that reported in melon transcriptome using the conventional (Sanger) dideoxy-based sequencing [8]. These differences in length might be due to the different assemblers used. The longer unigenes present the advantage of being more accurately annotated. The raw data files are available in the Sequence Read Archive at the National Center for Biotechnology Information (NCBI) [58], accession number

Table 1 Sequence statistics of *Cucurbita* 454 ESTs

Library	Raw reads Number/average length	Total length	Sequence quality average	Processed reads Number/average length	Total	Sequence quality average
Zucchini MU16	407,723/252	103 Mbp	31	261,962/319	84 Mbp	32
Scallop UPV196	392,370/254	100 Mbp	31	250,789/323	81 Mbp	32
TOTAL	800,093/253	203 Mbp	31	512,751/321	165 Mbp	32

Summary of the *Cucurbita pepo* expressed sequences generated with two half runs of GS-FLX Titanium pyrosequencing. Statistics of raw reads and reads after processing are indicated.

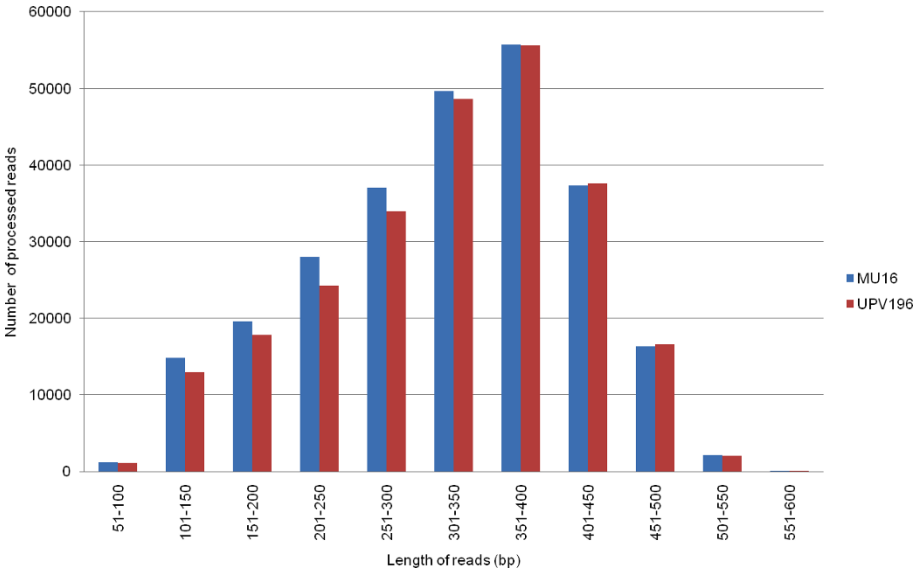

Figure 1 Length distribution of the *Cucurbita* ESTs. Data obtained after sequencing, with a half run of 454 GS FLX Titanium each one of the two *Cucurbita* cDNA libraries (Zucchini, Mu-16; Scallop, UPV-196), and processing the 454 raw reads, are presented.

SRA029105.1. The sequences of the unigenes in fasta format are available in the additional file 1:'*Cucurbita* unigenes', with unigene numbers from CUTC000001 to CUTC049610.

Structural and functional annotation

Most unigenes, 47,626 (96%) were predicted to have one ORF. By aligning the unigenes with the melon genomic sequence (available for the partners of the MELONOMICS project [7]), introns were identified in 16,697 unigenes (33.7%). Annotation results regarding ORFs and introns position are included in additional file 2: 'Annotation of ORFs and Introns'.

Codon usage was estimated using a subset of the unigenes predicted to contain full-length ORFs, with start and stop codons and without frame-shift errors. All codons were found in the dataset, with the least frequent codon represented 590 times (data not shown). As expected, the codon usage of *Cucurbita* shared some similarities with that of melon, *Arabidopsis* and other dicots. For example, T is the preferred base in the third codon position for most amino acids except for glycine,

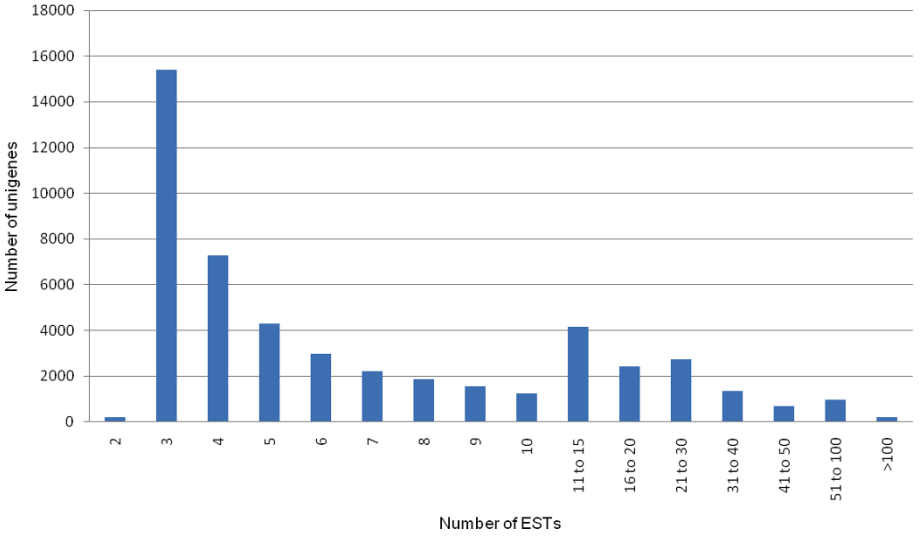

Figure 2 Distribution of number of ESTs in each *Cucurbita* unigene.

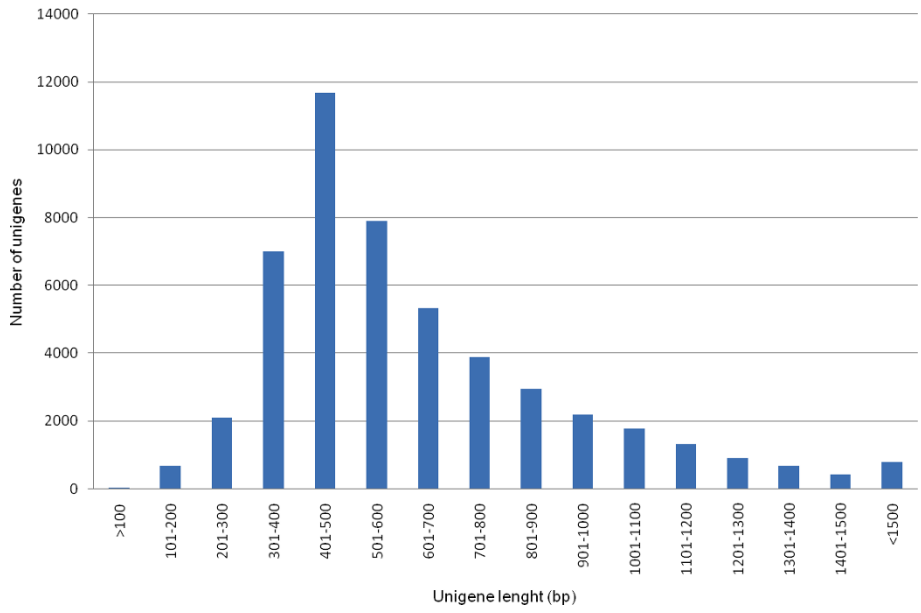

Figure 3 Length distribution of the *Cucurbita* unigenes *de novo* assembled from 454 ESTs.

phenylalanine, serine and arginine. The preferred stop codons were UAA and UGA that occurred in the 41.4 and 41.1% of the sequences, respectively. Suppression of the CG dinucleotide in the last two codon positions is very frequent in dicots, possibly as a consequence of methylation of C in the CG dinucleotide, resulting in an increased mutation rate; the ratio XCG/XCC for *Cucurbita* was 0.69, then the suppression was more intense than in *Arabidopsis* (0.92), but milder than that reported for melon (0.52) tomato (0.58), pea (0.51), potato (0.48) or *Populus* (0.38). The GC content in third base position was similar in *Cucurbita* as compared to melon and *Arabidopsis* (46% *vs* 39,9% and 42%) [8,35].

In order to identify *C. pepo* unigenes potentially encoding proteins with known function, a BLAST analysis [59] was performed in a sequential way using Swiss-Prot [60], *Arabidopsis* proteins [61], and Uniref90 [62] protein databases. Over 63% of the unigenes (31,159) had at least one significant hit (E-value cutoff = 1e-20). Most were annotated with the accurate databases Swiss-Prot (55%) and Arabidopsis (36%) and less with Uniref (9%). The majority of the unigenes have significant hits with *Arabidopsis* proteins (67%). Hits with *Cucumis* and *Cucurbita* previously reported proteins were also found.

Gene Ontology provides a structured and controlled vocabulary to describe gene products according to three categories: molecular function, biological process and cellular component [63,64]. We added GO terms using Blast2GO [65], based on the automated annotation of each unigene using BLAST results against the GenBank non redundant protein database (nr) from NCBI [66]. A total of 29,676 unigenes (60%) could be assigned to one or more ontologies. Figure 4A show the unigenes distribution regarding the number of GOs to which they were assigned. 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, being the majority of the unigenes mapped to 2 to 7 GO terms. In total, 103,045 GO terms were retrieved, 25%, 47% and 28% in the biological process, in the molecular function and in the cellular component category, respectively. The distribution of annotated unigenes under different GO levels of each category (Figure 4B) shows a concentration in 4-7, 3-7 and 4-7 levels respectively for biological process, molecular function, and cellular component, indicating a good accuracy of annotation. The GO annotation analysis reinforces the idea that a broad diversity of genes was sampled in our multi-tissue normalized cDNA libraries.

We used the GO annotations to assign each unigene to a set of GO Slims of the biological process and molecular function categories, which are a list of high-level GO terms providing a broad overview of the ontology content. A summary with the number of unigenes annotated in each GO slim term is shown in Figure 5. GO annotations for the unigenes showed fairly consistent sampling of functional classes. Cellular process, metabolic process, and biosynthetic process were among the most highly represented groups under the Biological Process category. This might be indicating the analyzed tissues were undergoing rapid growth and extensive metabolic activities. Genes involved in other important biological processes such as reproduction, stress and

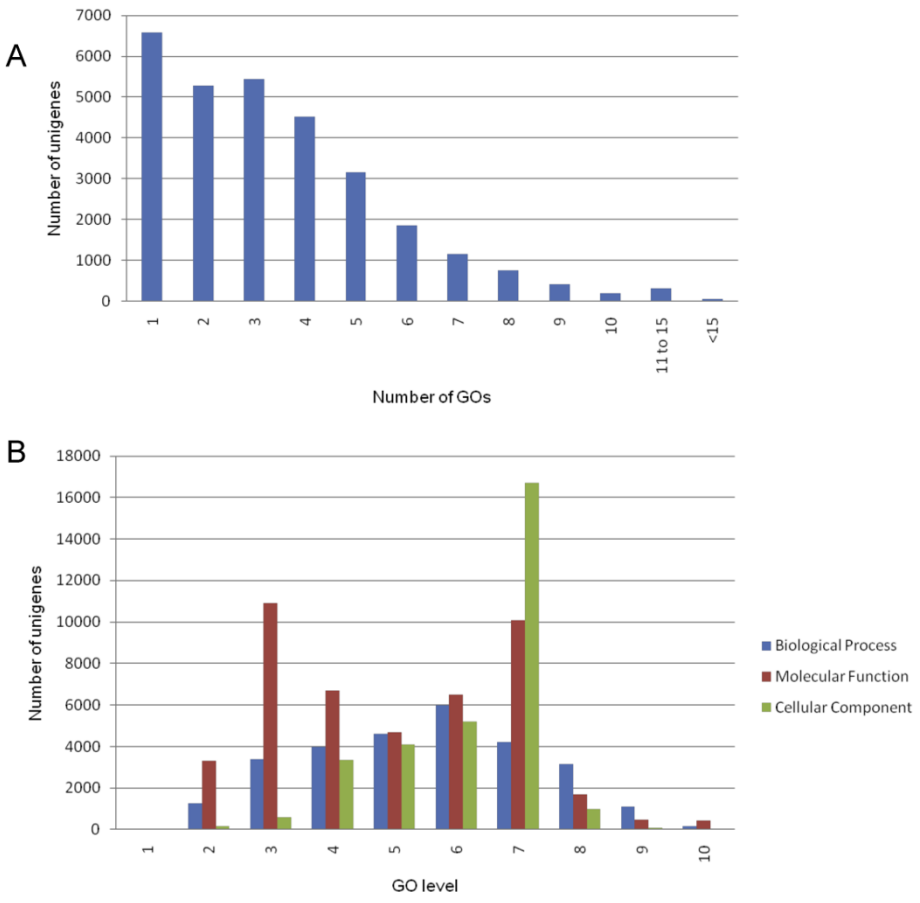

Figure 4 Number of GO terms (A) and GO level distribution (B) in the annotated Cucurbita unigenes. **A.** Distribution of GO terms in the annotated *Cucurbita* unigenes. **B.** GO level distribution in each category for the annotated *Cucurbita* unigenes.

stimulus response, signaling, and developmental processes were also identified (Figure 5A). Under the molecular function category, assignments were mainly to the catalitic and binding activities. A large number of hydrolases, kinases and transferases were annotated which suggests that this study may allow for the identification of genes involved in the secondary metabolite synthesis pathways. Also, transcription and translation factors were well represented (Figure 5B). The distribution of *Cucurbita* unigenes follow similar tendencies to that reported for *Arabidopsis* and also for the melon transcriptome [8,52,61], suggesting that the *Cucurbita* dataset could be representative of the whole squash transcriptome. All annotation results, regarding BLAST hits and GO annotations for the whole *Cucurbita* unigene collection are compiled in the additional file 3: 'Blast hits and GO terms'.

Doing a reciprocal blast search, we have also identified 19,312 *Cucurbita* unigenes (38.9%) with an ortholog in *Arabidopsis* [61] (11,022 (22.2%)) and a melon ortholog (12,461 (25.1%)) of the ICUGI data base [12] (Table 2). A list of the identified orthologs is included in additional file 4: '*Arabidopsis* and melon orthologs'.

Only 11,580 (23%) of the unigenes did not show significant similarity to any protein in the databases and could not be annotated. Shorter sequences are less likely to align with a significant E-value. However, the average length of these non annotated unigenes was 425 bp, with 50% being longer than 413 bp. For homology searches against known genes, unigenes longer than 200 bp are widely accepted for the effective assignment of functional annotations [57]. In previous studies performed with massive sequencing techniques a similar or even higher number of unigenes did not match with previously known sequences representing newly detected transcripts [44,49,55,57].

Genes potentially encoding proteins involved in pathogen resistance, flowering, fruit quality and root traits

Viral and fungal pathogens affect severely the productivity of cucurbits crops. The *Cucurbita* unigene collection

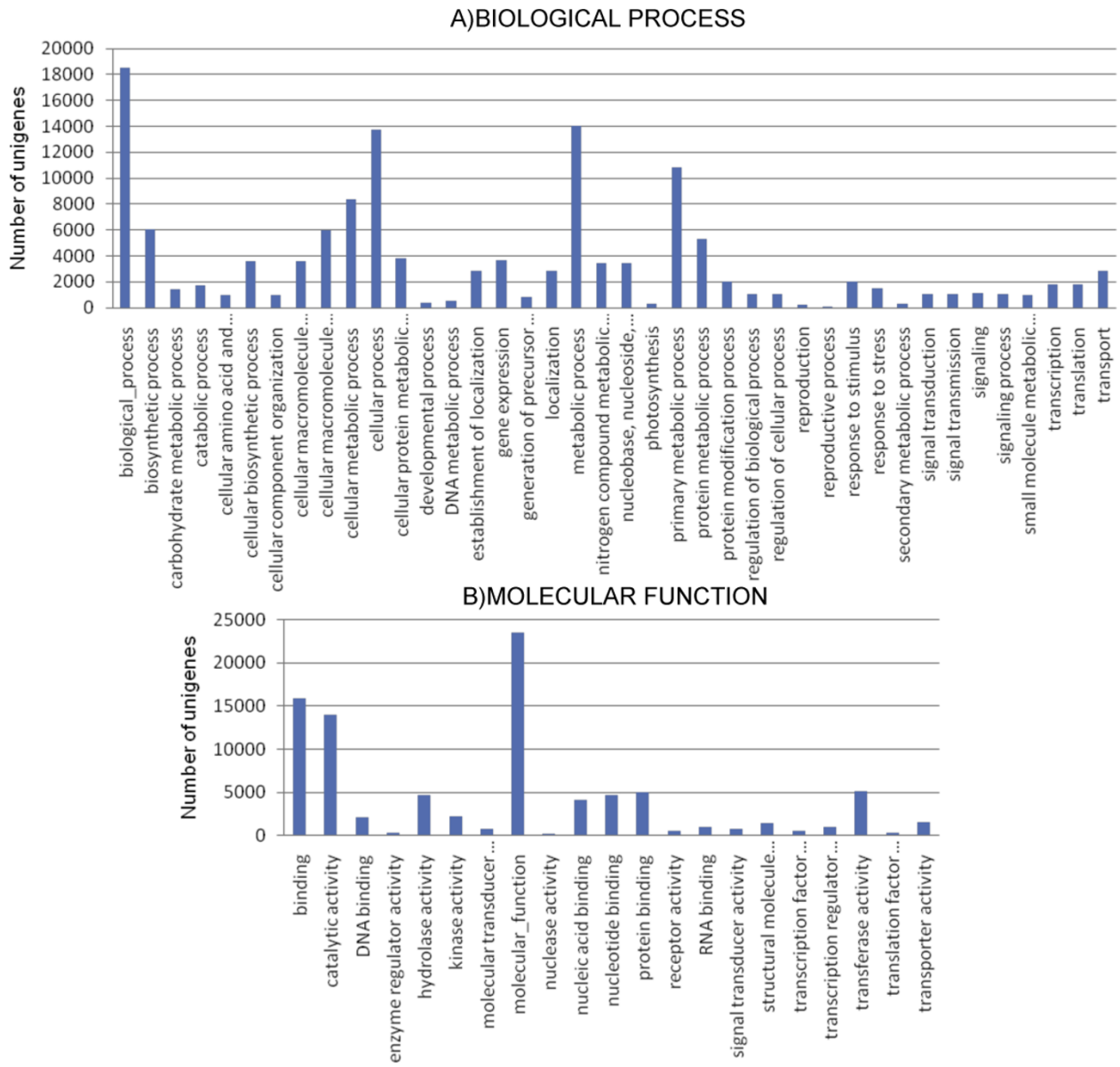

Figure 5 Number of *Cucurbita* unigenes in each functional category. *Cucurbita* unigenes were classified into different functional groups based on a set of GO slims in the A) Biological Process category and B) Molecular Function category.

contains genes potentially involved in disease resistance and disease response [67-69] (see additional file 5: 'Genes potentially encoding pathogen resistance, flowering, fruit and root traits'). We have found at least one ortholog to *TOM1*, *TOM2A* (*Tobamovirus multiplication 1 and 2A*) and *THH1* (*Tom Three Homolog 1*), genes encoding proteins that play essential roles on the tobamovirus replication [70], and also orthologs to the eukaryotic translation factors of the EIF4 family (*EIF4E*, *EIF(ISO)4E* and *EIF4EG*), known to mediate recessive resistance to poty- and other RNA viruses [71-73]. Regarding fungal responses, we have found orthologs to the *Arabidopsis RPH1* gene (*resistance to Phytophthora 1*), involved in immune response to *Phytophthora brassicae* [74], and other genes required for non-race specific resistance to bacterial and fungal pathogens [75].The powdery mildew is the main fungal disease affecting *Cucurbita* cultivation

Table 2 Functional annotation statistics

Database	Number of unigenes with ortholog	%	Number of orthologs
Arabidopsis pep	11022	22,2%	14880
Melon ICUGI	12461	25,12%	12976

Databases searched were: *Arabidopsis* and melon icugi [12,61].

[15]. An ortholog to *Arabidopsis PMR5* (*powdery mildew resistant 5*) and to the *MLO10* gene, belonging to the family of *Arabidopsis* homologs of the barley mildew resistance locus *mlo*, have been identified [76,77].

Cucurbits are models for sex determination studies due to its diverse floral sex types. Significant progress has been made in elucidating the mechanisms of plant sex determination by cloning several major sex-determining genes in cucurbit species [2,78,79]. Despite such advances, the whole mechanisms of sex determination are still unknown. Both *Cucurbita* sequenced genotypes are monoecious, but have large differences in the flowering time and the femaleness tendency. The *Cucurbita* unigene collection includes orthologs of *Arabidopsis* genes involved in flower development and flowering-time regulation (see additional file 5) [80-82]. Sex expression in cucurbits can be regulated by plant hormones and environmental factors. Ethylene is highly correlated with the femaleness and has been regarded as the primary sex determination factor [6]. Some genes related with the ethylene synthesis and also transcription factors and receptors involved in the ethylene perception and signaling have been found (various *EIN* (*Ethylene insensitive*) and *ETR* (*Ethylene response*) genes). It has been reported that auxin and also brassinosteroids can induce pistillate flower formation in part through its stimulation of ethylene production. The *Cucurbita* unigene collection has also different orthologs of proteins involved in auxin and brassinosteroids signaling, affecting flowering-time in *Arabidopsis* (*SAR*, *suppressor of auxin resistance*, *BZR*, *Brassinazole resistant*) [83-85]. Cucumber orthologs involved in those mechanisms have been reported to have differential expression in Gynoecious/hermaphroditic flowers in a recent study performed with massive sequencing [51]

We also identified a number of genes putatively involved in gibberellins (GA) biosynthetic and signaling pathways. These specific genes might be associated with the role of gibberellic acid in developmental regulation and plant stress response. The cucurbits represent a model plant system on which to examine the contents of the phloem translocation stream. A recent study reported a large-scale analysis of proteins from pumpkin (*C. maxima*) phloem exudates [5]. Identified proteins mainly corresponded to enzymes that carry out pivotal roles in stress and defense pathways. Furthermore, the detection of proteins related to GA synthesis in phloem supports the theory that the phloem is the route for transport and modification of GAs. Some orthologs of the genes encoding the main phloem sap proteins are included in our *C. pepo* collection.

Fruit development and ripening are the most important processes determining the fruit quality. At present most of the molecular and genetic data available about fruit development and ripening come from *Arabidopsis* and tomato [86]. In recent years, several genes and quantitative trait loci controlling fruit quality traits have been described in cucurbits [87,88]. As for the previously described processes, orthologs to genes involved in fruit development and quality have been found in the *Cucurbita* dataset (see additional file 5) [89,90]. These include several cell wall-metabolism enzymes and genes involved in the isoprenoid biosynthetic pathway (provides intermediates for the synthesis of sterols, carotenoids and chlorophylls, and also phytohormones and terpenes involved in plant defense). *Cucurbita* species are important sources of vitamins in many developing countries due to their high carotenoids content [91]. The *Cucurbita* unigene collection includes some enzymes involved in carotenoids biosynthesis (*PSY*, *Phytoene synthase*, *PDS*, *Phytoene desaturase* and *ZDS*, *Zeta-carotene desaturase*) [92]. The root tissue also provided some root specific genes involved in root development or stress tolerance [93].

SSRs and SNPs discovery and validation

We performed a general screen on the *Cucurbita* unigene dataset for the presence of microsatellites, analyzing its nature and frequency [94]. A search for di-, tri-, and tetra-nucleotide repeats yielded a total of 1,935 potential SSRs in 1,822 unigenes, that is approximately 4% of the unigenes contained at least one of the considered SSR motifs. This percentage agrees with previous studies using EST databases that shows that approximately 3-7% of expressed sequences contain putative SSR motifs [34,51].

The maximum and minimum lengths of the repeats were 129 and 17, and the average length was 24 nucleotides. These were mostly tri-nucleotide (71.7%), and less di- (15.3%) and tetra- (13%). The most common repeat motifs are indicated in Table 3. A similar bias towards AG, AAG and AAAG and against CG repeats has been reported in EST-SSRs of many crops, including other cucurbits like melon and cucumber. It has been proposed that this may be due to the tendency of CpG sequences to be methylated which potentially might inhibit transcription [8,51]. Genomic SSRs identified in *C. pepo* and *C. moschata* also showed the same predominant di- and tri-nucleotide motifs [24]. The complete list of SSRs and their corresponding information are provided in additional file 6: '*Cucurbita* SSRs'.

Most SSRs (55%) were located in ORFs, being a similar number in the 5' and the 3' untranslated regions (UTRs) (Table 4). 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. It is

Table 3 Simple Sequence repeats (SSRs) statistics

di-nucleotide repeat	Number of di-SSRs	%
AG	225	76
AT	60	20
AC	11	4
Total	296	100
tri-nucleotide repeat	Number of tri-SSRs	
AAG	699	50
AGC	135	10
ATC	116	8
AGG	99	7
AAT	89	6
Other tri-nucleotide repeats (% ≤ 6 each one)		
AAC, ACC, ACG,CCG, ACT	249	19
Total	1387	100
Tetra-nucleotide repeat	Number of tetra-SSRs	%
AAAT	33	13
AAAG	31	12
AATG	24	10
AATC	21	8
ATCC	18	7
AAAC	17	7
ACAT	16	6
Other tetranucleotide repeats (% ≤ 6 each one)		
ACTC,AACC,AAGG,ACAG,AGGC,AACG,AACT, AATT,AGCC,AGCG,AAGC,AGGG,AGAT,ACGG, AGCT, AAGT,ACCC, ACCT	92	37
Total	252	100

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

Table 4 Localization of SSRS with respect to putative initiation and termination codons in the *Cucurbita* unigene dataset

	di-SSRs		tri-SSRs		tetra-SSRs		all-SSRs	
	N°	%	N°	%	N°	%	N°	%
5'-UTR	86	29%	172	12%	102	41%	360	19%
ORF	72	24%	903	65%	89	35%	1064	55%
3'-UTR	105	36%	194	14%	30	12%	329	17%
Other	33	11%	118	9%	31	12%	182	9%
Total	296	100%	1387	100%	252	100%	1935	100%

Unigenes were checked for the presence of the start and stop codons. "Other" means imprecise localization of the SSRs with respect to putative initiation or termination codons.

known that the UTRs are richer in SSRs than coding regions, particularly the 5'-UTRs [34,95]. Thus, the prevalence of tri-nucleotide repeats in the *Cucurbita* dataset may account for our high proportion of ORF-SSRs. These results agree with those reported in melon, where the most frequent SSRs in ORFs were tri-nucleotide [8].

We selected a set of 30 ESTs-SSRs for validation, 26 (86.7%) amplified polymorphic fragments in a set of 10 genotypes of *Cucurbita*, 9 representative of the diversity within *C. pepo* (accessions of 3 morphotypes of spp. *pepo*, including several landraces and commercial types of the zucchini type, and two morphotypes of spp. *ovifera*) and 1 *C. moschata* accession. All but one could be transferred to the related crop *C. moschata*, producing alleles unique of this species in most cases (60%). On an average we found 3.2 alleles per primer pair in *C. pepo* and *C. moschata*. Most of EST-SSRs assayed are useful to detect variability within and between the two subspecies of C. pepo. A 77% were polymorphic between the two genotypes used for sequencing, and 50% and 88% detected variation within spp. *ovifera* and within spp. *pepo*, respectively. Also 62% detected variation among the landraces and commercial lines of zucchini. Details of these validated SSRs are included in the additional file 7: 'Validated *Cucurbita* SSRs'.

SSR markers derived from EST sequences have been extensively used in constructing genetic maps of cucurbit species [9,96]. Until recently, only a few microsatellites have been available for *Cucurbita*, and transferability from other cucurbits, such as cucumber of melon, has been demonstrated to be very low [9], then the development of SSRs for this genus is highly desirable. Gong et al. [24] developed SSRs-enriched partial genomic libraries from an Austrian oil-pumpkin variety *C. pepo* subsp. *pepo* and one accession of *C. moschata*, generating a collection of 1,058 putative SSRs. They reported a 81% validation in a set of genotypes representing the cultivar groups, also indicating a high intra-genus transferability. EST-SSRs have several advantages versus genomic; they are related to genes, being functional markers that can be used as candidate genes to study their association with phenotypic variation and the flanking sequences are more likely to be conserved among close or distant species, making their use as markers for comparative mapping easier. We will use the identified EST-SSRs markers for the construction of a genetic map, using a Recombinant Inbreed population (RILs) derived from the Zucchini (MU16) × Scallop (UPV196) cross. They will be also useful for fingerprinting commercial Zucchini cultivars, breeding lines and landraces and for genetic diversity studies within the genus, mainly performed with RAPDs or AFLPs to date [20,97].

Massive sequencing of ESTs in a number of diverse genotypes has been previously used for developing large SNPs collections [49-52]. Since the ESTs generated under the present study, using the 454 technology, are from two different cultivars belonging to two subspecies,

with MU16 and UPV96 representing the 51% and 49% of the EST sequences, respectively, we expected SNPs to be frequent in our collection. The SNP calling was done with the default parameters recommended by the ngs_backbone software [53]. Stringent quality criteria were used for distinguish 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 19,980 SNPs and 1,174 INDELs distributed in 8,147 unigenes (16.4%), averaging a total of 2.6 single variations per unigene. Different filters were applied to facilitate the management of the variants. INDELs can be filtered out with VKS (It is not an SNP). The detailed information about the identified SNPs and INDELs is included in the additional file 8:'Cucurbita SNPs and INDELs'.

Within the detected SNPs, transitions (68%) were much more common than transversions (32%) (Table 5). 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, 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%).

Other filters allowed an accurate *in silico* selection of the SNPs to identify the ones more suited for mapping purposes. All located in sequences with more than 4 SNPs or INDELs per 100 bases (filtered out with HVR4) were discarded (71, 0.36%) and also those that were variable within one or both genotypes, MU16 and UPV196 (10,937, 54.7%) (filtered out with NVSM2, NVSM1 or both filters). This requirement is intended to minimize 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.

Table 5 Single nucleotide polymorphism (SNPs) statistics

SNPs	Number	SNPs	Number
Transitions		Transversions	
A<->G	6,694	A<->T	1,793
C<->T	6,902	G<->T	1,547
		C<->G	1,548
		A<->C	1,496
Total	13,596 (68%)	Total	6,384 (32%)

Type and number of transition and transversions are shown for putative high quality single nucleotide polymorphism (SNPs) identified in the *Cucurbita* database.

From the remaining 9,043 higher confident SNPs, that were monomorphic within and polymorphic between the two sequenced genotypes, we selected a set that met different criteria for facilitating validation and for their use in a Golden Gate genotyping assay [98,99], discarding those that were closer than 60 bp to another SNP or INDEL, and/or were located closer than 60 bp to an intron and/or were closer than 60 bp to the unigene edge (filtered out with CS60, I60 and CL60, respectively). Finally, 3,538 SNPs were selected that met all criteria (see those with only a dot or a CEF tag in the filter field in additional file 8). From these, 144 SNPs were identified that can be detected as CAPS as they generate allele-specific restriction targets. We selected 50 of this putative CAPS, and 39 (80%) gave amplicons polymorphic between MU16 and UPV196 after digestion with the corresponding enzyme, which is comparable to that reported in previous studies with maize and *Eucaliptus* [49,52]. Information of the validated CAPS is included in additional file 9:'*Cucurbita* SNPs validated as CAPS'. 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* and melon orthologs, SNVs and SSRs) have been also added in additional file 10 using the GFF3 standard file format of The Sequence Ontology Resources [100]. This format for annotations results facilitates its uploading, representation and analysis.

Conclusions

The length and amount of the ESTs obtained with the 454 GSLX-Titanium platform has facilitated *de novo* assembly of the transcriptome in *Cucurbita pepo*, species for which limited prior sequence information is available, providing unigene sequences with length comparable to that obtained with traditional Sanger methodology. The unigene sequences constituted the first genomic resource for the *Cucurbita* genus. *Cucurbita*, along to *Cucumis* and *Citrullus*, are the three most economically important genera of the Cucurbitaceae family, whose economic importance is second only to the Solanaceae. Then this resource will enhance comparative studies within the family. The transcriptome will be important for gene discovery in *Cucurbita* and for future annotations of the *Cucurbita* genome sequence. The identified genes provide candidates for resistance genes against RNA viruses, fungal or bacterial pathogens. This is also an important resource for further study of sex determination and fruit quality in *Cucurbita*. The SSRs and SNPs identified here will constitute an important resource for mapping and marker-assisted breeding in *Cucurbita pepo* and closely related crops. The Zucchini and Scallop types are used as vegetables

and highly valued in international markets, but *C. pepo* and also *C. moschata*, *C. maxima* and other minor *Cucurbita* species included a number of highly variable types that are food staples and rich sources of fat and vitamins in developing countries. All these crops will also take benefit from this genomic resource.

Methods

Plant material

Two cDNA libraries were constructed using material from the MU16 "Zucchini" Spanish cultivar (belonging to *Cucurbita pepo* L. ssp. *pepo*) and the UPV196 "Scallop" (belonging to *C. pepo* L. ssp. *ovifera*). Seeds of both cultivars were maintained at the COMAV Genebank. These two genotypes are representative of the sub specific variation of *C. pepo*, are readily crossable and have been selected as parentals of a RILs mapping population. They have contrasting phenotypes for traits interesting in squash breeding [15]: growth habit, sex expression, fruit shape and color, parthenocarpy tendency, shelf life, response to diseases., and are molecularly distant enough for mapping purposes [20]. Seeds were germinated and grown in trays containing a mixture of peat and sand. They were properly watered and grown at day/night temperatures of 28/20°C with a 16-h photoperiod. From each variety, tissue was sampled from the second and third true leaves, and from male and female flowers in pre and post-anthesis stage. Also the whole roots of 15 days-old plants were sampled. All tissues were collected and immediately frozen in liquid nitrogen and stored at -80°C till use.

cDNA preparation and sequencing

Total RNA was extracted from each tissue using the TRIzol® Reagent (Invitrogen, USA). We combined equivalent amounts of RNA from each tissue into two pools, one per cultivar. mRNA was purified from the total RNA using the illustraTM mRNA Purification Kit (GE Healthcare, Amersham Bioscience, Buckinghamshire, UK). Double-stranded cDNA was then synthesized from the two RNA pools with the SMART cDNA Library Construction Kit (Clontech, USA). A normalization step was carried out with TRIMMER Kit (Evrogen, Moscow, Russia) in order to prevent over-representation of the most common transcripts. The PCR products of cDNA were purified using the QIAquick PCR Purification Kit (QIAGEN, 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 454 GLX Titanium platform. A half-plate sequencing run was performed for each sample (Creative Genomics [101]). All raw sequences are available in the Sequence Read Archive at the National Center for Biotechnology Information (NCBI) [58], accession number SRA 029105.1

cDNA sequence processing and assembly

The whole sequence analysis was carried out by using the ngs_backbone pipeline [53]. 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.

The raw 454 sequences were processed prior to the assembly. To remove the adaptors an alignment with the adaptors used during the sequencing process was done by Exonerate [102]. The low quality regions were removed by using Lucy [103]. Sequences shorter than 100 pb were discarded and not used for the assembly. The processed 454 sequences were assembled with Mira [54]. Default ngs_backbone options for this process were used.

Cucurbita 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 1e-20) with the sequences in three major public protein databases, prioritizing handmade annotation databases. The used database order was Swiss-Prot [60], *Arabidopsis* proteins [61] and UniRef90 [62]. Once a sequence had a blast hit in one of the databases, a description was build from the description of that hit. Also, a bi-directional blast search comparison was performed in order to obtain a set of putative orthologs with *Arabidopsis* [61] and melon, using the melon unigenes contained in the Cucurbits genomic database (ICUGI).

Additionally, we performed a functional classification of the unigenes following the Gene Ontology (GO) scheme [64]. Blast2GO [65] was used for this purpose. Blast2GO used the results of a blast nr search (cutoff e value of 1e-20) to infer the relevant GO terms for every sequence. ORFs were predicted in the unigenes with the aid of the ESTScan software [104]. We used the *Arabidopsis* codon usage table to perform the ORF searching. Introns were assigned by aligning the unigenes with the melon genomic sequence using the Emboss: est2genome [105].

To assess codon usage, we used a set of the *Cucurbita* unigenes predicted to contain full-length coding regions. We performed a manual inspection, to ensure that no sequences containing frame-shift errors were included in the analysis. From this dataset, containing 4,118 sequences (529,864 codons), ORFs were defined and a codon usage table was created.

Identification of SSRs and SNPs

SSRs were annotated using the Sputnik software [94]. Sequences containing ≥ 4 di-, tri-, or tetra-nucleotide repeats were selected. A set of SSRs were validated using the sequenced genotypes (the Zucchini MU16 and the Scallop UPV96) and a set of 7 genotypes of *C. pepo* representative of the diversity within the species: representing 3 morphotypes of spp. *pepo* (3 additional zucchini, 1 vegetable marrow, 1 pumpkin, 1 Styrian pumpkin, an oil-pumpkin variety) and one additional morphotype of spp. *ovifera* (1 croockneck) and 1 accession of *C. moschata*. Primer pairs flanking each SSR locus were designed using the Primer3 program [106]. PCR reactions were performed in a final volume of 15 μL with 1× PCR buffer (100 mM Tris-HCl, 15 mM MgCl2, 500 mM KCl, pH 8.3), 200 μM dNTPs, 0.15 μM each primer and 2 μL of template (aprox. 10 ng/μL). PCR was performed as follows: denaturation at 95°C for 3 min, followed by 10 cycles of 30 s at 95°C, 30 s at 65°C (with each cycle the annealing temperature decreasing 1°C), and of 30 s at 72°C. Products were subsequently amplified for 20 cycles at 95°C for 30 s, 55°C for 30 s and 72°C for 30 s, with a final extension at 72°C for 5 min. The forward primer was designed adding an M13 tail to its 5' end. PCR products were separated using 6% polyacrylamide gels, 1× TBE buffer in a LI-COR 4300. IRD700 and IRD800-labeled amplicons were visualized by adding to PCR mixture 0.2 μM of fluorescent label (IR700 or IR800) M13 tail.

This *Cucurbita* ESTs collection has been produced using two representatives of the two *C. pepo* subspecies appropriate for SNP discovery. Ngs_backbone was also used to detect SNVs (Single nucleotide variations, SNPs and INDELs) by mapping the 454 processed reads against the unigene assembly using BWA (Burrows-Wheeler Aligner) [107]. We only kept SNVs meeting stringent quality criteria: 1) Minimum allele quality: accumulated sequence quality for every allele; 2) 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 in order to maximize a successful validation and/or implementation in high throughput genotyping platforms such as Golden gate genotyping assay [98,99]. For example, a filter to differentiate SNPs from INDELs (VKS: It is not an SNP) was applied, and also a filter that dismisses SNVs in highly variable regions (HVR4: The region has more than 4 SNVs per 100 bp). Other used filter were; CS60 (SNV is closer than 60 bp to another SNV), I60 (an intron is located closer than 60 bp), CL60 (SNV is closer than 60 bp to the sequence end), NVSM1 (SNV is variable within UPV196 or not sequenced in this genotype), NVSM2 (SNV is variable within MU16 or not sequenced in this genotype).

We also detected those SNVs that can be analyzed via CAPS (searching for alele-specific restriction targets, filter CEF) and validated a subset of them. To do this PCR conditions were used as described for SSRs and fragments digested with the corresponding enzymes were detected by agarose gel electrophoresis.

Additional material

Additional file 1: *Cucurbita* unigenes. The fasta sequence of the 49,610 *Cucurbita* unigenes assembled from 454 ESTs is included.

Additional file 2: Annotation of ORFs and introns. Unigene length and predicted position of ORFs and introns is indicated for the whole *Cucurbita* unigene collection.

Additional file 3: Blast Hits and GO terms. Descriptions build from the blast hit obtained by a sequential blast search of 3 protein databases [60-62] and GO annotations for the whole *Cucurbita* unigene collection are compiled.

Additional file 4: *Arabidopsis* and melon orthologs. Orthologs found by reciprocal search with *Arabidopsis* and melon databases [12,61] are indicated.

Additional file 5: Genes potentially encoding pathogen resistance, flowering, fruit and root traits. Genes were identified in the *Cucurbita* data set by comparison with the *Arabidopsis* and melon databases [12,61]. A brief description and the corresponding *Arabidopsis* and melon *locus* are given for each unigene.

Additional file 6: *Cucurbita* SSRs. The table provides the list of SSRs identified from the *Cucurbita* unigene dataset, their length, motif sequences, position in the unigene, and scores.

Additional file 7: Validated *Cucurbita* SSRs. The table provides the list of *Cucurbita* SSRs experimentally validated using a collection of *C. pepo*. Primers used for validation, number of alleles, and polymorphism detected between the sequenced genotypes, within subspecies, and within the zucchini morphotypes are included. Also transference to *C. moschata* is indicated.

Additional file 8: *Cucurbita* SNPs and INDELs. The file provides the list of SNPs and INDELs (SNVs) identified from the *Cucurbita* unigene dataset, their position, the nucleotide change (or I or D for insertion and deletion in INDELs), the quality of the polymorphic base, and additional information about allele number and frequency. The different filters applied for the *in silico* selection of the SNVs are also indicated: VKS (is not an SNPs, is an INDEL), HVR4 (SNV is in a region with more than 4 SNVs per 100 bp), CS60 (SNV is closer than 60 bp to another SNV), I60 (SNV is located closer than 60 bp to an intron), CL60 (SNV is closer than 60 bp to the sequence end), NVSM1 (SNV is variable within UPV196, or not sequenced in this genotype), NVSM2 (SNV is variable within MU16, or not sequenced in this genotype), CEF (SNV does not alter a restriction target and cannot be detected as a CAPS). The VCF (Variant Call Format) version 4.0 has been used for this file [108].

Additional file 9: *Cucurbita* SNPs validated as CAPS. The table provides the list of SNPs that affected restriction targets and were validated *via* CAPS, their position, location, primers used for validation and the occurrence of polymorphism between the two sequenced varieties.

Additional file 10: 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 [100] that facilitate annotations uploading, representation and analysis.

Acknowledgements and Funding

This research has been funded by the project of the Spanish Instituto Nacional de Investigación y Tecnología Agraria INIA (RTA2008-00035-C02-02).

Authors thank Cristina Esteras for providing technical help in markers validation and MELONOMICS project (2009-2012) of the Fundación Genoma España for providing access to the draft of the melon genomic sequence.

Authors' contributions

CR, JC, and BP prepared the cDNA libraries for sequencing. BP and CR selected and validated the SSRs and CAPS. JB and PZ performed the bioinformatic analysis. JC participated in the annotation analyses. BP is the main coordinator of the *Cucurbita* project and participated in the conception of the study together with JB and JC. BP was primarily responsible for drafting and revising the manuscript with contributions from co-authors. FN is the director of COMAV and critically reviewed the manuscript. All authors read and approved the final manuscript.

Received: 17 November 2010 Accepted: 10 February 2011
Published: 10 February 2011

References

1. Schaefer H, Heibl C, Renner SS: **Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events.** *Proc Biol Sci* 2009, **276**:843-851.
2. Boualem A, Fergany M, Fernandez R, Troadec C, Martin A, Morin H, Sari MA, Collin F, Flowers JM, Pitrat M, Purugganan MD, Dogimont C, Bendahmane A: **A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons.** *Science* 2008, **321**:836-838.
3. Ezura H, Owino WO: **Melon, an alternative model plant for elucidating fruit ripening.** *Plant Sci* 2008, **175**:121-129.
4. Li Z, Huang S, Liu S, Pan J, Zhang Z, Tao Q, Shi Q, Jia Z, Zhang W, Chen H, Si L, Zhu L, Cai R: **Molecular isolation of the M gene suggests that a conserved-residue conversion induces the formation of bisexual flowers in cucumber plants.** *Genetics* 2009, **182**:1381-1385.
5. Lin MK, Lee YJ, Lough TJ, Phinney BS, Lucas WJ: **Analysis of the pumpkin phloem proteome provides functional insights into angiosperm sieve tube function.** *Mol Cell Proteomics* 2009, **8**:343-356.
6. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EA, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Y, Ruan J, Qian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan , Wu Z, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Li Y, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim JY, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Li M, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao J, Fang X, Li G, Fang L, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S: **The genome of the cucumber, Cucumis sativus.** *L Nat Genet* 2009, **41**:1275-1281.
7. González VM, Rodríguez-Moreno L, Centeno E, Benjak A, Garcia-Mas J, Puigdomènech P, Aranda MA: **Genome-wide BAC-end sequencing of Cucumis melo using two BAC libraries.** *BMC Genomics* 2010, **11**:618.
8. Gonzalez-Ibeas D, Blanca J, Roig C, Gonzalez-To M, Pico B, Truniger V, Gomez P, Deleu W, Cano-Delgado A, Arus P, Nuez F, García-Mas J, Puigdomènech P, Aranda MA: **MELOGEN: an EST database for melon functional genomics.** *BMC Genomics* 2007, **8**:306.
9. Fernandez-Silva I, Eduardo I, Blanca J, Esteras C, Picó B, Nuez F, Arús P, Garcia-Mas J, Monforte AJ: **Bin mapping of genomic and EST-derived SSRs in melon (Cucumis melo L.).** *Theor Appl Genet* 2008, **118**:139-150.
10. Deleu W, Esteras C, Roig C, González-To M, Fernández-Silva I, Blanca J, Aranda MA, Arús P, Nuez F, Monforte AJ, Picó MB, Garcia-Mas J: **A set of EST-SNPs for map saturation and cultivar identification in melon.** *BMC Plant Biol* 2009, **9**:90.
11. Mascarell-Creus A, Cañizares J, Vilarrasa J, Mora-García S, Blanca J, González-Ibeas D, Saladié M, Roig C, Deleu W, Picó B, López-Bigas N, Aranda MA, Garcia-Mas J, Nuez F, Puigdomènech P, Caño-Delgado A: **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* 2009, **10**:467.
12. **Cucurbit Genomics Database of the International Cucurbit Genomics Initiative (ICuGI).** [http://www.icugi.org].
13. Smith BD: **Reassessing Coxcatlan Cave and the early history of domesticated plants in Mesoamerica.** *Proc Natl Acad Sci USA* 2005, **102**:9438-9445.
14. Robinson RW, Decker-Walters DS: **Cucurbits.** *CAB International* Wallingford; 1997.
15. Ferriol M, Picó B: **Pumpkin and Winter Squash.** In *Handbook of Plant Breeding. Vegetables I. Part 4. Volume 1.* Edited by: Prohens J, Nuez, F. Springer; 2008:317-349.
16. Paris H: **Summer Squash.** In *Handbook of Plant Breeding. Vegetables I. Part 4. Volume 1.* Edited by: Prohens J, Nuez, F. Springer; 2008:351-381.
17. Shokrzadeh M, Azadbakht M, Ahangar N, Hashemi A, Saeedi Saravi SS: **Cytotoxicity of hydro-alcoholic extracts of Cucurbita pepo and Solanum nigrum on HepG2 and CT26 cancer cell lines.** *Phcog Mag* 2010, **6**:176-179.
18. Fita A, Pico B, Roig C, Nuez F: **Performance of Cucumis melo ssp agrestis as a rootstock for melon.** *J Hort Sci Biotech* 2007, **82**:184-190.
19. Paris HS: **A proposed subspecific classification for Cucurbita pepo.** *Phytologia* 1986, **61**:113-138.
20. Ferriol M, Picó B, Nuez F: **Genetic diversity of a germplasm collection of Cucurbita pepo using SRAP and AFLP markers.** *Theor Appl Genet* 2003, **107**:271-282.
21. Brown RN, Myers JR: **A genetic map of squash (Cucurbita sp.) with randomly amplified polymorphic DNA markers and morphological markers.** *J Am Soc Hort Sci* 2002, **127**:568-575.
22. Lee YH, Jeon HJ, Hong KH, Kim BD: **Use of random amplified polymorphic DNAs for linkage group analysis in interspecific hybrid F2 generation of Cucurbita.** *J Kor Soc Hort Sci* 1995, **36**:323-330.
23. Zraidi A, Stift G, Pachner M, Shojaeiyan A, Gong A, Lelley TA: **Consensus map for Cucurbita pepo.** *Mol Breed* 2007, **20**:375-388.
24. Gong L, Stift G, Kofler R, Pachner M, Lelley T: **Microsatellites for the genus Cucurbita and an SSR-based genetic linkage map of Cucurbita pepo L.** *Theor Appl Genet* 2008, **117**:37-48.
25. Gong L, Pachner M, Kalai K, Lelley T: **SSR-based genetic linkage map of Cucurbita moschata and its synteny with Cucurbita pepo.** *Genome* 2008, **51**:878-887.
26. Paris HS, Brown RN: **The genes of pumpkin and squash.** *HortScience* 2005, **40**:1620-1630.
27. Nakajima N, Mori H, Yamazaki K, Imaseki H: **Molecular Cloning and Sequence of a Complementary DNA Encoding 1-Aminocyclopropane-l-carboxylate Synthase Induced by Tissue Wounding.** *Plant Cell Physiol* 1990, **31**:1021-1029.
28. Yamaguchi S, Saito T, Abe H, Yamane H, Murofushi N, Kamiya Y: **Molecular cloning and characterization of a cDNA encoding the gibberellin biosynthetic enzyme ent-kaurene synthase B from pumpkin (Cucurbita maxima L.).** *Plant J* 1996, **10**:203-213.
29. Ellard-Ivey M, Hopkins RB, White TJ, Lomax TL: **Cloning, expression and N-terminal myristoylation of CpCPK1, a calcium-dependent protein kinase from zucchini (Cucurbita pepo L.).** *Plant Mol Biol* 1999, **39**:199-208.
30. Nishida I, Sugiura M, Enju A, Nakamura M: **A Second Gene for Acyl-(Acyl-Carrier-Protein): Glycerol-3-Phosphate Acyltransferase in Squash, Cucurbita moschata cv. Shirogikuza, Codes for an Oleate-Selective Isozyme: Molecular Cloning and Protein Purification Studies.** *Plant Cell Physiol* 2000, **41**:1381-1391.
31. Manzano S, Martínez C, Domínguez V, Avalos E, Garrido D, Gómez P, Jamilena M: **A major Gene Conferring Reduced Ethylene Sensitivity and Maleness in Cucurbita pepo.** *J Plant Growth Regul* 2009, **29**:73-80.
32. Manzano S, Martínez C, Gómez P, Garrido D, Jamilena M: **Cloning and characterization of two CTR1-like genes in Cucurbita pepo: regulation of their expression during male and female flower development.** *Sex Plant Reprod* 2010, **23**:301-313.
33. Rafalski JA: **Novel genetic mapping tools in plants: SNPs and LD-based approaches.** *Plant Sci* 2002, **162**:329-333.
34. Thiel T, Michalek W, Varshney RK, Graner A: **Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.).** *Theor Appl Genet* 2003, **106**:411-422.
35. Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, others: **A Populus EST resource for plant functional genomics.** *Proc Natl Acad Sci USA* 2004, **101**:13951-13956.
36. Li X, Wu HX, Dillon SK, Southerton SG: **Generation and analysis of expressed sequence tags from six developing xylem libraries in Pinus radiata D.** *Don BMC Genomics* 2009, **21**:41.
37. Raju NL, Gnanesh BN, Lekha P, Jayashree B, Pande S, Hiremath PJ, Byregowda M, Singh NK, Varshney RK: **The first set of EST resource for**

gene discovery and marker development in pigeonpea (Cajanus cajan L.). *BMC Plant Biol* 2010, **10**:45.
38. **NCBI-dbEST database.** [http://www.ncbi.nlm.nih.gov/dbEST].
39. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309**:1728-1732.
40. Metzker ML: **Sequencing technologies-the next generation.** *Nat Rev Genet* 2010, **11**:31-46.
41. Cheung F, Haas BJ, Goldberg SMD, May GD, Xiao Y, Town CD: **Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology.** *BMC Genomics* 2006, **7**:272.
42. Emrich SJ, Barbazuk WB, Li L, Schnable PS: **Gene discovery and annotation using LCM-454 transcriptome sequencing.** *Genome Res* 2007, **17**:69-73.
43. Eveland AL, McCarty DR, Koch KE: **Transcript Profiling by 3'-Untranslated Region Sequencing Resolves Expression of Gene Families.** *Plant Physiol* 2008, **146**:32-44.
44. Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB: **Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing.** *Plant Physiol* 2007, **144**:32-42.
45. Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M, Giuliano G, Chiusano ML, Baldoni L, Perrotta G: **Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development.** *BMC Genomics* 2009, **10**:399.
46. Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA, Wheeler N, Sederoff R, Carlson JE: **Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection.** *BMC Plant Biology* 2009, **9**:51.
47. Wang W, Wang Y, Zhang Q, Qi Y, Guo D: **Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing.** *BMC Genomics* 2009, **10**:465.
48. Bellin D, Ferrarini A, Chimento A, Kaiser O, Levenkova N, Bouffard P, Delledonne M: **Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species.** *BMC Genomics* 2009, **24**:55.
49. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing.** *Plant J* 2006, **51**:910-918.
50. Trick M, Long Y, Meng J, Bancroft I: **Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing.** *Plant Biotechnol J* 2009, **7**:334-346.
51. Guo S, Zheng Y, Joung JG, Liu S, Zhang Z, Crasta OR, Sobral BW, Xu Y, Huang S, Fei Z: **Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types.** *BMC Genomics* 2010, **11**:384.
52. Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome.** *BMC Genomics* 2008, **9**:312.
53. **Bioinformatics at COMAV. Ngs_backbone.** [http://bioinf.comav.upv.es/ngs_backbone].
54. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res* 2004, **14**:1147-1159.
55. Vega-Arreguin JC, Ibarra-Laclette E, Jimenez-Moraila B, Martinez O, Vielle-Calzada JP, Herrera-Estrella L, Herrera-Estrella A: **Deep sampling of the Palomero maize transcriptome by a high throughput strategy of pyrosequencing.** *BMC Genomics* 2009, **10**:299.
56. Sun Ch, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EMK, Chen S: **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* 2010, **11**:262.
57. Li Y, Luo HM, Sun C, Song JY, Sun YZ, Wu Q, Wang N, Yao H, Steinmetz A, Chen SL: **EST analysis reveals putative genes involved in glycyrrhizin biosynthesis.** *BMC Genomics* 2010, **11**:268.
58. **The Sequence Read Archive at the National Center for Biotechnology Information.** [http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi].
59. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
60. **The Swiss-Prot Database.** [http://www.uniprot.org/downloads], uniprot_sprot_ release of 2010 04 23.
61. **The TAIR Database: The Arabidopsis Information Resource.** [http://www.arabidopsis.org/], Tair_9_pep_ release 2009 06 19.
62. **Uniref90.** [http://www.ebi.ac.uk/uniref], uniref90_release 2010 04 23.
63. The Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
64. **Gene Ontology Database.** [http://www.geneontology.org].
65. Conesa A, Götz S, Blast2GO: **A Comprehensive Suite for Functional Analysis in plant Genomics.** *Int J Plant Genomics* 2008, 619832.
66. **NCBI Nr, Nonredundant protein Database.** [ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz], release of 20010-03-27.
67. Cartieaux F, Thibaud MC, Zimmerli L, Lessard P, Sarrobert C, David P, Gerbaud A, Robaglia C, Somerville S, Nussaume L: **Transcriptome analysis of Arabidopsis colonized by a plant-growth promoting rhizobacterium reveals a general effect on disease resistance.** *Plant J* 2003, **36**:177-188.
68. Tan X, Meyers B, Kozik A, West M, Morgante M, St Clair D, Bent A, Michelmore R: **Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in Arabidopsis.** *BMC Plant Biol* 2007, **7**:56.
69. Ascencio-Ibanez JT, Sozzani R, Lee TJ, Chu TM, Wolfinger RD, Cella R, Hanley-Bowdoin L: **Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection.** *Plant Physiol* 2008, **148**:436-454.
70. Yamanaka T, Imai T, Satoh R, Kawashima A, Takahashi M, Tomita K, Kubota K, Meshi T, Naito S, Ishikawa M: **Complete Inhibition of Tobamovirus Multiplication by Simultaneous Mutations in Two Homologous Host Genes.** *J Virol* 2002, **76**:2491-2497.
71. Nicaise V, Gallois JL, Chafiai F, Allen LM, Schurdi-Levraud V, Browning KS, Candresse T, Caranta C, Le Gall O, German-Retana S: **Coordinated and selective recruitment of eIF4E and eIF4G factors for potyvirus infection in Arabidopsis thaliana.** *FEBS letters* 2007, **581**:1041-1046.
72. Nieto C, Morales M, Orjeda G, Clepet C, Monfort A, Sturbois B, Puigdomenech P, Pitrat M, Caboche M, Dogimont C, Garcia-Mas J, Aranda MA, Bendahmane A: **An eIF4E allele confers resistance to an uncapped and non-polyadenylated RNA virus in melon.** *Plant J* 2006, **48**:452-462.
73. Piron F, Maryse N, Minoïa S, Piednoir E, Moretti A, Salgues A, Zamir D, Caranta C, Bendahmane A: **An induced mutation in tomato eIF4E leads to immunity to two potyviruses.** *PloS ONE* 2010, **5(6)**:e11313.
74. Belhaj K, Lin B, Mauch F: **The chloroplast protein RPH1 plays a role in the immune response of Arabidopsis to Phytophthora brassicae.** *Plant J* 2009, **58**:287-298.
75. Zhang J, Lu H, Li X, Li Y, Cui H, Wen CK, Tang X, Su Z, Zhou JM: **Effector-Triggered and Pathogen-Associated Molecular Pattern-Triggered Immunity Differentially Contribute to Basal Resistance to Pseudomonas syringae.** *Mol Plant Microb Inter* 2010, **23**:940-948.
76. Chen Z, Hartmann HA, Wu MJ, Friedman E, Chen JG, Pulley M, Schulze-Lefert P, Panstruga R, Jones A: **Expression analysis of the AtMLO Gene Family Encoding Plant-Specific Seven-Transmembrane Domain Proteins.** *Plant Mol Biol* 2006, **60**:583-97.
77. Maeda K, Houjyou Y, Komatsu T, Hori H, Kodaira T, Ishikawa A: **AGB1 and PMR5 Contribute to PEN2-Mediated Preinvasion Resistance to Magnaporthe oryzae in Arabidopsis thaliana.** *Mol Plant Microb Inter* 2009, **22**:1331-1340.
78. Li Z, Huang S, Liu S, Pan J, Zhang Z, Tao Q, Shi Q, Jia Z, Zhang W, Chen H, Si L, Zhu L, Cai R: **Molecular isolation of the M gene suggests that a conserved-residue conversion induces the formation of bisexual flowers in cucumber plants.** *Genetics* 2009, **182**:1381-1385.
79. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A: **A transposon-induced epigenetic change leads to sex determination in melon.** *Nature* 2009, **461**:1135-1138.
80. Mara CD, Huang T, Irish VF: **The Arabidopsis Floral Homeotic Proteins APETALA3 and PISTILLATA Negatively Regulate the BANQUO Genes Implicated in Light Signaling.** *Plant Cell* 2010, **22**:690-702.
81. Xu ML, Jiang JF, Ge L, Xu YY, Chen H, Zhao Y, Bi YR, Wen JQ, Chong K: **FPF1 transgene leads to altered flowering time and root development in rice.** *Plant Cell Rep* 2005, **24**:79-85.
82. Lee H, Yoo SJ, Lee JH, Kim W, Yoo SK, Fitzgerald H, Carrington JC, Ahn JH: **Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in Arabidopsis.** *Nucl Acids Res* 2010, **38**:3081-3093.

83. Papadopoulou E, Grumet R: **Brassinosteriod-induced femaleness in cucumber and relationship to ethylene production.** *HortSci* 2005, **40**:1763-1767.
84. Parry G, Ward S, Cernac A, Dharmasiri S, Estelle M: **The Arabidopsis SUPPRESSOR OF AUXIN RESISTANCE Proteins are Nucleoporins with an Important Role in Hormone Signaling and Development.** *The Plant Cell* 2006, **18**:1590-1596.
85. Zhang H, Harry DE, Ma C, Yuceer C, Hsu CY, Vikram V, Shevchenko O, Etherington E, Strauss SH: **Precocious flowering in trees: the FLOWERING LOCUS T gene as a research and breeding tool in Populus.** *J Exp Bot* 2010, **61**:2549-2560.
86. Giovannoni JJ: **Genetic regulation of fruit development and ripening.** *Plant Cell* 2004, **16**:170-180.
87. Monforte AJ, Oliver M, Gonzalo MJ, Alvarez JM, Dolcet-Sanjuan R, Arus P: **Identification of quantitative trait loci involved in fruit quality traits in melon (Cucumis melo L.).** *Theor Appl Genet* 2004, **108**:750-758.
88. Cuevas HE, Staub JE, Simon PW, Zalapa JE, McCreight JD: **Mapping of genetic loci that regulate quantity of beta-carotene in fruit of US Western Shipping melon (Cucumis melo L.).** *Theor Appl Genet* 2008, **117**:1345-1359.
89. Liu JP, Van Eck J, Cong B, Tanksley SD: **A new class of regulatory genes underlying the cause of pear-shaped tomato fruit.** *Proc Natl Acad Sci USA* 2002, **99**:13302-13306.
90. Manning K, Tor M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB: **A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening.** *Nat Genet* 2006, **38**:948-952.
91. Tadmor Y, Paris HS, Meir A, Schaffer AA, Lewinsohn E: **Dual role of the pigmentation gene B in affecting carotenoid and vitamin E content in squash (Cucurbita pepo) mesocarp.** *J Agric Food Chem* 2005, **53**:9759-9763.
92. Maass D, Arango J, Wüst F, Beyer P, Welsch R: **Carotenoid crystal formation in Arabidopsis and carrot roots caused by increased phytoene synthase protein levels.** *PLoS ONE* 2009, **4(7)**:e6373.
93. Singh SK, Fischer U, Singh M, Grebe M, Marchant A: **Insight into the early steps of root hair formation revealed by the procuste1 cellulose synthase mutant of Arabidopsis thaliana.** *BMC Plant Biol* 2008, **8**:57.
94. **Sputnik.** [http://espressosoftware.com/sputnik/index.html].
95. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.** *Nat Genet* 2002, **30**:194-200.
96. Levi A, Wechter P, Davis A: **EST-PCR markers representing watermelon fruit genes are polymorphic among watermelon heirloom cultivars sharing a narrow genetic base.** *Plant Gen Res* 2009, **7**:16-32.
97. Paris HS, Yonash N, Portnoy V, Mozes-Daube N, Tzuri G, Katzir N: **Assessment of genetic relationships in Cucurbita pepo (Cucurbitaceae) using AFLP, ISSR, and SSR markers.** *Theor Appl Genet* 2003, **106**:971-978.
98. Fan JB, Chee MS, Gunderson KL: **Highly parallel genomic assays.** *Nat Rev Genet* 2006, **7**:632-644.
99. Gupta PK, Rustgi S, Mir RR: **Array-based high throughput DNA markers for crop improvement.** *Heredity* 2008, **101**:5-18.
100. **The Sequence Ontology Resources.** [http://www.sequenceontology.org/gff3.shtml].
101. **Creative Genomics.** [http://www.creative-genomics.com/gene/sequence.html].
102. **Exonerate, a generic tool for sequence alignment.** [http://www.ebi.ac.uk/~guy/exonerate/].
103. **Lucy DNA sequence quality and vector trimming tool.** [http://lucy.sourceforge.net/].
104. **ESTScan software.** [http://www.isrec.isb-sib.ch/ftp-server/ESTScan/].
105. **The Emboss: est2genome.** [http://emboss.sourceforge.net/apps/cvs/emboss/apps/est2genome.html].
106. **Primer 3.** [http://frodo.wi.mit.edu/primer3/].
107. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler Transform.** *Bioinformatics* **25**:1754-60.
108. **VCF (Variant Call Format) version 4.0.** [http://www.1000genomes.org/wiki/Analysis/vcf4.0].

doi:10.1186/1471-2164-12-104
Cite this article as: Blanca *et al.*: Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). *BMC Genomics* 2011 **12**:104.