Data Descriptor

De Novo Transcriptome Assembly of *Cucurbita Pepo* L. Leaf Tissue Infested by *Aphis Gossypii*

Alessia Vitiello [1], Rosa Rao [1], Giandomenico Corrado [1], Pasquale Chiaiese [1], Maria Cristina Digilio [1], Riccardo Aiese Cigliano [2] and Nunzio D'Agostino [3,*]

[1] Department of Agricultural Sciences, University of Naples Federico II, Via Università 100, 80055 Portici, Italy; alessia.vitiello@unina.it (A.V.); rao@unina.it (R.R.); giacorra@unina.it (G.C.); chiaiese@unina.it (P.C.); digilio@unina.it (M.C.D.)

[2] Sequentia Biotech SL, Campus UAB, Av. de Can Domènech s/n, 08193 Bellaterra, Barcelona, Spain; raiesecigliano@sequentiabiotech.com

[3] CREA Research Centre for Vegetable and Ornamental Crops, via dei Cavalleggeri 25, 84098 Pontecagnano Faiano, Italy

* Correspondence: nunzio.dagostino@crea.gov.it; Tel.: +39-89-386243

check for updates

Abstract: Zucchini (*Cucurbita pepo* L.), extensively cultivated in temperate areas, belongs to the Cucurbitaceae family and it is a species with great economic value. One major threat related to zucchini cultivation is the damage imposed by the cotton/melon aphid *Aphis gossypii* Glover (Homoptera: Aphididae). We performed RNA-sequencing on cultivar "San Pasquale" leaves, uninfested and infested by *A. gossypii*, that were collected at three time points (24, 48, and 96 h post infestation). Then, we combined all high-quality reads for *de novo* assembly of the transcriptome. This resource was primarily established to be used as a reference for gene expression studies in order to investigate the transcriptome reprogramming of zucchini plants following aphid infestation. In addition, raw reads will be valuable for new experiments based on the latest bioinformatic tools and analytical approaches. The assembled transcripts will serve as an important reference for sequence-based studies and for primer design. Both datasets can be used to support/improve the prediction of protein-coding genes in the zucchini genome, which has been recently released into the public domain.

Dataset: https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA439198; https://www.ncbi.nlm.nih.gov/nuccore/1377124260

Keywords: zucchini transcriptome; biotic stress; RNA-sequencing; aphid infestation; leaves

---

1. Summary

Cucurbita pepo L. (2n = 2x = 40) belongs to the Cucurbitaceae family; it is widely cultivated in temperate region and ranks among the highest-valued vegetables worldwide [1]. Historical records report that *C. pepo* is native to North America and was dispersed to other continents during the 16th century by transoceanic travels [2]. *C. pepo* is extremely variable in fruit-related features. The edible forms of this species can be grouped into two sub-species: ssp. *Pepo*, which includes pumpkin, vegetable marrow, cocozelle, and zucchini; and spp. *Ovifera*, which includes acorn squash, scallop, crookneck, and straightneck.

One major threat related to zucchini cultivation, both in greenhouse and open-field, is the damage imposed by the cotton/melon aphid *Aphis gossypii* (Homoptera: Aphididae). *Aphis gossypii* is a

cosmopolitan, highly polyphagous species, widely distributed in warm climate regions [3], which can both directly and indirectly affect host plant by inducing stunt growth, leaf curling, and necrosis and by vectoring several plant viruses. Furthermore, indirect damage is related to the deposition on plant tissue surfaces of honeydew, which provides a nutrient source for saprophytic fungi (microorganism growth), resulting in hampered photosynthesis [4].

Given the economic importance and the growing attention on this crop, in the last decade, a large number of genomic resources and tools has been developed to accelerate cucurbit crop improvement. Most of these resources and tools have merged into the Cucurbit Genomics Database (http://cucurbitgenomics.org/).

In 2011, Blanca et al. [5] sequenced *C. pepo* transcriptome using a 454 GS FLX Titanium platform. Three cDNA libraries (root, leaves, flower tissue) from two *C. pepo* varieties that differ for plant, flowering, and fruit traits were used to generate a collection of 49,610 unigenes that represents the first sequenced transcriptome of the species. A subset of single nucleotide polymorphism (SNP) markers identified within this transcriptome was then selected to design a custom Illumina GoldenGate genotyping assay used to build the first linkage map of Cucurbita and to identify quantitative trait loci (QTL) [1].

Subsequently, two more *C. pepo* transcriptomes, namely Acorn squash cv. "Sweet REBA" and Pumpkin (*C. pepo* subsp. *ovifera*, cv. "Big Moose" and "Munchkin"), were *de novo* assembled from Illumina reads in order to provide further genomic resources within the Cucurbita genus [6,7]. Comparative analysis performed on "Big Moose" and "Munchkin" transcriptomes allowed genes with potential roles in fruit size and morphology to be identified, as well as microsatellite markers derived from expressed sequence tags (EST-SSR) to be generated [8]. Also, the transcriptome of zucchini cultivar "True French", used as parent in crossing scheme of pathogen-resistant commercial varieties, has been sequenced and assembled as a valuable resource for genetic and genomic studies [9]. Lastly, a high-quality draft of the zucchini genome organized into 20 chromosome-scale pseudomolecules was released into the public domain [10]. Additionally, 40 transcriptomes of 12 species of the genus were assembled and used as the foundation for comparative genomic studies [10].

Aiming to contribute in this scenario, we performed RNA-sequencing and *de novo* assembly of the cv. "San Pasquale" transcriptome following a compatible interaction with *Aphis gossypii*. As far as we know, this is the first zucchini transcriptome from leaf tissue challenged by an insect pest.

2. Data Description

2.1. Illumina Read Processing and Transcriptome Assembly

We performed RNA-sequencing on cultivar "San Pasquale" leaves, uninfested and infested by *A. gossypii*, that were collected at three time points (24, 48, and 96 h post infestation). The schematic overview of the experimental design is depicted in Figure 1.
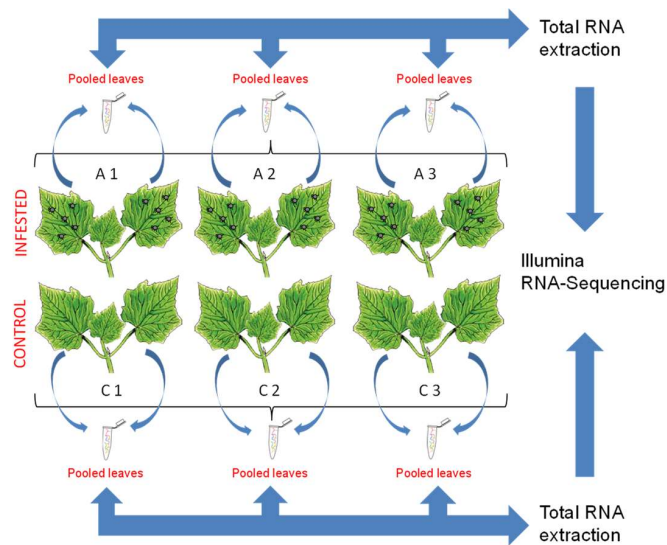
Figure 1. Schematic overview of the experimental design. Three biological replicates for both infested and control plants were collected at 24, 48, and 96 h post infestation and leaves of a single replicate were pooled for downstream analysis.

All samples were subjected to sequencing using an Illumina HiSeq 2500 device in a 2 × 101 paired-end format. The overall process of read pre-processing, transcriptome assembly, and annotation, as well as transcriptome quality evaluation, is outlined in Figure 2.
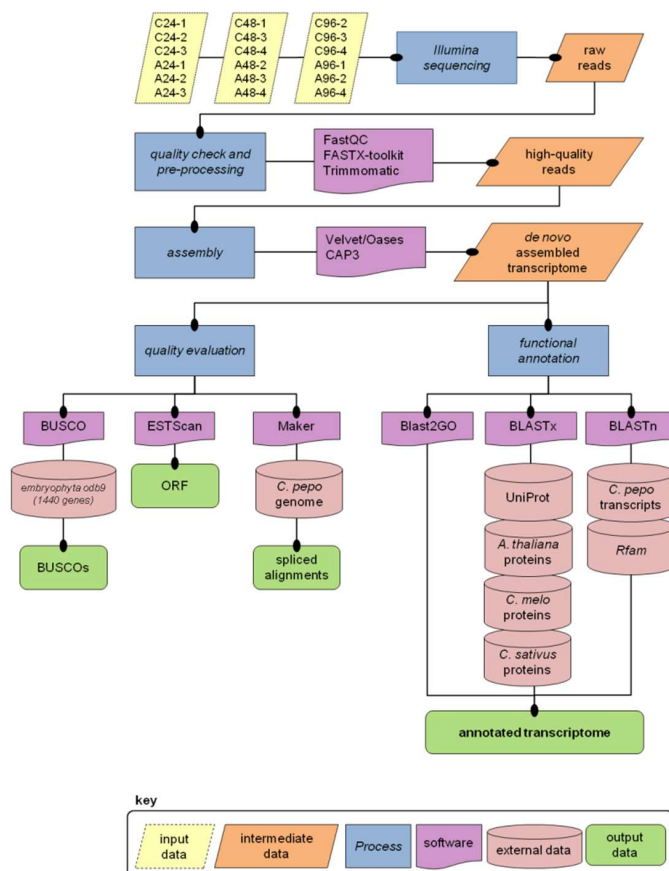

Figure 2. Data processing workflow. ORF—Open Reading Frame; BUSCO—Benchmarking Universal Single-Copy Orthologs; EST—Expressed Sequence Tags; BLAST—Basic Local Alignment Search Tool; CAP—Contig Assembly Program; GO—Gene Ontology.

The sequencing generated ~34 million paired-end reads of 101 nucleotides in length *per* sample (Table 1). After the pre-processing step (see Methods), about 552.4 million of high-quality reads (average Q score 37.44; min Q score 30; max Q score 39) of 75–101 nucleotides in length (average length 98 nucleotides) were obtained and approximately 4 million reads were filtered out for each sample (Table 1).

Table 1. Number of reads generated from sequencing (raw data) and after quality filtering and adapter trimming (high quality data) for each sample.

	Sample Name	Raw Data # Reads	High Quality Data	
			# Paired Reads	# Single Reads
Control	C24_1	31,430,108	22,032,822	5,256,322
	C24_2	28,740,043	21,066,607	4,270,278
	C24_3	33,677,909	25,020,509	4,812,235
	C48_1	36,265,357	28,423,670	4,540,249
	C48_3	35,144,118	27,591,869	4,394,338
	C48_4	37,518,763	28,808,248	5,107,225
	C96_2	33,527,557	25,873,268	4,394,873
	C96_3	31,060,525	24,146,031	4,067,956
	C96_4	35,937,098	28,576,994	4,591,206
Infested	A24_1	35,344,346	23,674,612	7,451,885
	A24_2	35,230,308	23,732,387	7,382,828
	A24_3	34,366,050	22,964,925	7,204,783
	A48_2	37,211,641	25,261,715	7,571,915
	A48_3	36,623,056	24,901,053	7,442,903
	A48_4	37,996,974	25,666,812	7,820,577
	A96_1	38,622,935	27,596,629	5,791,605
	A96_2	34,353,147	24,532,466	5,227,154
	A96_3	29,037,507	20,856,920	4,406,645

Then, all high-quality reads were combined for *de novo* assembly of the transcriptome. The total number of transcripts and major characteristics of the assembled transcriptome are reported in Table 2.

Table 2. Statistics on the *de novo* assembled *C. pepo* transcriptome.

Total # transcripts	71,648
Total # gene locus	42,517
# Single sequence	22,594
# Multiple variants	19,923
Total sequence length (nt)	95,354,115
Average transcript length (nt)	1331
Maximum transcript length (nt)	12,009
Minimum transcript length (nt)	100
Median transcript length (nt)	1084

2.2. Annotation

The transcriptome was annotated using similarity-based searches against five different databases. A total of 58,945 transcripts (72%) had significant matches with proteins in the *Cucumis melo* dataset. BLASTx searches against *Cucumis sativus* and *Arabidopsis thaliana* protein sequences, as well as against the UniProtKB/SwissProt database, revealed that 43,796 (67%), 56,683 (79%), and 44,378 (68%) transcripts, respectively, had at least one significant match in the corresponding database. Based on these results, approximately 71% of all transcripts had at least one match in one of the four protein databases queried (Figure 1). The BLASTn comparison between the *de novo* assembled transcriptome with the publically available *C. pepo* transcriptome resulted in 70,334 (98%) sequences with significant matches. This means that the transcriptome assembly herein described include 1313 novel sequences.

In detail, 548 out of 1313 transcripts had at least one match with one of the four protein databases above. All considered, a total of 36,585 sequences (~51%) matched all five databases queried (Figure 3).
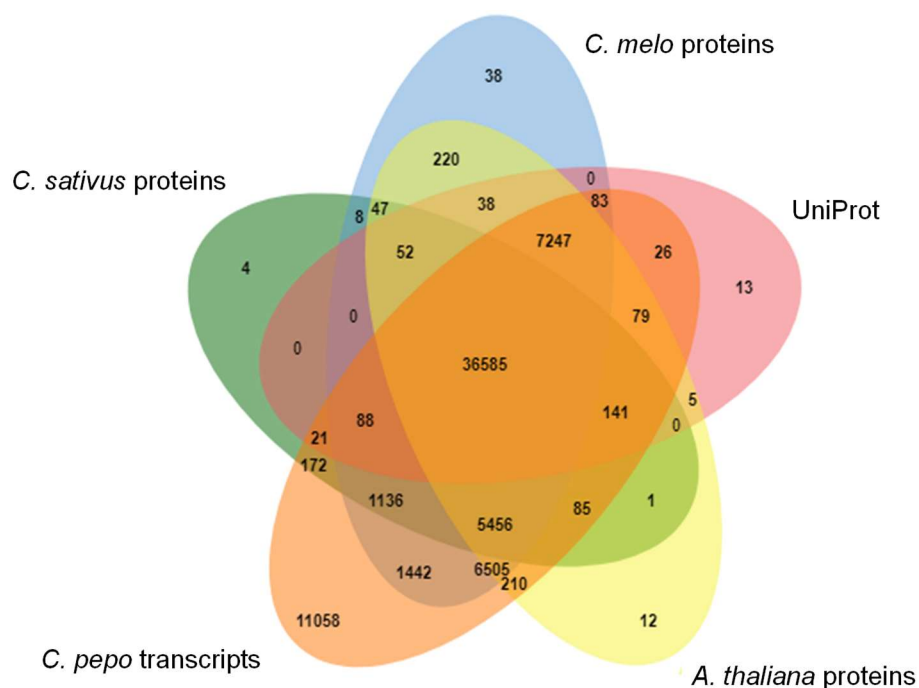

Figure 3. Venn diagram showing the BLAST results of *C. pepo* transcriptome against five databases.

Annotation was refined using the Blast2GO software [11]. Gene Ontology (GO) terms were assigned to 51,398 sequences, allowing us to classify *C. pepo* transcripts in a standard and controlled vocabulary. The number of GO terms *per* sequence varied between 1 and 74, with an average of seven GO terms *per* transcript. In total, 276,601 GO terms were retrieved, with 50% assigned to biological process, 27% assigned to molecular function, and 21% assigned to cellular component domain. Enzyme Commission (EC) numbers were associated with 15,304 transcripts out of 51,398 GO annotated sequences, whereas 10,426 sequences were mapped at least onto one Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway.

Finally, BLASTn-based sequence similarity search was performed using Rfam [12] as filtering database. Our dataset includes 1104 (1.5% of the total transcripts) potential non-coding RNAs. Among these transcripts, only 398 do not have a match with any of the queried protein databases.

All the annotations for each transcript are reported in Supplementary Table S1.

2.3. Evaluation of Transcriptome Quality and Completeness

In order to assess the quality of the transcriptome, we performed open reading frame (ORF) prediction using ESTScan [13]. The *Arabidopsis thaliana* training matrix was selected for peptide prediction in the *C. pepo* transcriptome. The results (Table 3) indicated that 67,534 sequences (about 94% of total transcripts) contain putative coding sequences that could be translated into proteins. Among these, 23,735 transcripts were categorized as complete ORF, containing defined start and stop codons. Additionally, 43,799 transcripts were classified as partial coding sequences. Specifically, 25,000 sequences were classified as "5′ truncated ORF" with clear stop codon and lacking the ATG start codon; 8220 transcripts displayed the initiating ATG codon, but not termination triplet. Furthermore, 10,579 sequences encoded for truncated proteins showing neither start nor stop codons. The remaining 4114 sequences (about 6% of all transcripts) were probably un-translated regions (UTRs) with interspersed stop codons or non-coding RNAs.

Table 3. Results of the open reading frame (ORF) prediction analysis. [1] ORF lacking ATG codon but including the stop codon. [2] ORF including ATG codon but lacking the stop codon. [3] ORF with neither start nor stop codon. (#: number of).

Items	# Sequences
complete ORF	23,735
5′ truncated [1]	25,000
3′ truncated [2]	8220
5′ and 3′ truncated [3]	10,579
no good ORF	4114
Total	71,648

Transcriptome completeness was evaluated running the Benchmarking Universal Single-Copy Orthologs tool (BUSCO) [14]. The number of complete BUSCOs was 1215 out of 1400 (707 complete and single-copy BUSCOs + 508 complete and duplicated BUSCOs). The number of fragmented BUSCOs was 80, while the number of missing BUSCOs was 145. In summary, the 84.4% of complete BUSCOs were found and this indicates how close to completeness the assembled zucchini transcriptome is.

Finally, transcripts were mapped on the *C.pepo* reference genome (version 4.1) using the Maker pipeline [15]. Exactly 60,632 out of 71,648 (84.62%) transcripts were automatically transferred onto the genome sequence and converted into reliable gene structures. An additional 4735 transcripts (6.6%) were mapped on the genome by Maker and were tagged as "expressed_sequence_match". The remaining 6281 sequences were aligned (via BLASTn) against the *C. pepo* genome; 5597 of them (7.8% of the total) were successfully mapped (see Methods). In summary, over 90% of the assembled transcripts were mapped back on the *C. pepo* reference genome with high confidence.

2.4. Value of the Data

This resource was primarily established to be used as a reference for gene expression studies in order to investigate the transcriptome reprogramming of zucchini plants after aphid infestation. With this study, we are making available datasets for molecular biology and genetics research in Cucurbita spp. These resources will be of critical importance for the investigation of the molecular mechanisms and signals involved in the zucchini response to aphid infestation. Furthermore, the transcriptome and its functional annotation might be easily compared with the available cucurbit transcriptomes previously generated from the same or different tissues. Finally, both RNA-seq raw reads and the assembled transcripts will be valuable to support/improve the prediction of protein-coding genes in the zucchini genome [10].

2.5. Data Records

Raw FASTQ Illumina sequence data (Table 4) have been deposited at the Sequence Read Archive (SRA) under the accession number SRP136062 (PRJNA439198). The dataset includes 18 records (from SRS3072838 to SRS3072855). For each sample, three replicates were sequenced.

This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GGKS00000000. The version described in this paper is the first version, GGKS01000000.

Table 4. Description of samples submitted to the NCBI Sequence Read Archive (SRA).

Sample Number	BioSample	SRA ID	Library Name
1	SAMN08742104	SRS3072843	A24-1
2	SAMN08742105	SRS3072853	A24-2
3	SAMN08742106	SRS3072846	A24-3
4	SAMN08742107	SRS3072849	A48-2
5	SAMN08742108	SRS3072852	A48-3
6	SAMN08742109	SRS3072855	A48-4
7	SAMN08742110	SRS3072850	A96-1
8	SAMN08742111	SRS3072851	A96-2
9	SAMN08742112	SRS3072848	A96-3
10	SAMN08742113	SRS3072847	C24-1
11	SAMN08742114	SRS3072854	C24-2
12	SAMN08742115	SRS3072842	C24-3
13	SAMN08742116	SRS3072845	C48-1
14	SAMN08742117	SRS3072844	C48-3
15	SAMN08742118	SRS3072839	C48-4
16	SAMN08742119	SRS3072838	C96-2
17	SAMN08742120	SRS3072841	C96-3
18	SAMN08742121	SRS3072840	C96-4

3. Methods

3.1. Biological Material and Experimental Design

Seeds of the aphid-susceptible cultivar "San Pasquale" were obtained from the seed company "La Semiorto Sementi". Zucchini were planted in plastic pots with a diameter of 10 cm and were enclosed in insect-proof cages. Plants were grown in a climatic chamber with a photoperiod of 16/8 h (light/dark), under a temperature of 22 ± 1 °C, with 75 ± 5% of relative humidity.

Aphis gossypii Glover (Homoptera: Aphididae) was isolated from watermelon plants under severe infestation in Terracina (Latina, Central Italy) and reared on "San Pasquale" plants in cages equipped with anti-insect nets (50 mesh). Aphid rearing was maintained in a dedicated climatic chamber under the environmental conditions described above at the Department of Agricultural Sciences of the University of Naples Federico II. Zucchini plants were transferred to a new climatic chamber (temperature: 22 ± 1 °C; relative humidity: 75 ± 5%; photoperiod: L16: D8), and were individually placed in insect-proof cages for the infestation assay. First and second leaves were infested with ten *A. gossypii* adults. Five aphids *per* leaf were transferred onto the adaxial surface with a paintbrush and their number was daily monitored. Control plants, individually enclosed in insect-proof cages, were grown under the same conditions. Aphids were left to feed for 24, 48, and 96 h, after which they were manually removed using a fine paintbrush. Leaf tissue was sampled and immediately frozen in liquid nitrogen. At the same time points, leaf tissue from aphid-free control plants was sampled. Three biological replicates for both infested and control plants were collected *per* time point and leaves of a single replicate were pooled for downstream analysis (Figure 1).

3.2. RNA Extraction, Library Construction and Sequencing

Total RNA was extracted from 100 mg of tissue previously ground in liquid nitrogen using the RNeasy Mini kit (Qiagen, Hilden, Germany), according to manufacturer's instructions. Next generation sequencing was performed by Genomix4life srl. (Baronissi, Salerno, Italy). Indexed libraries were prepared from 2 μg of RNA with the TruSeq Stranded mRNA Sample Prep Kit (Illumina, San Diego, CA, USA) following manufacturer's instructions. Libraries had insert sizes of 125 bp. Libraries were quantified using the RNA Bioanalyzer 2100 Plant Nano chip (Agilent Technologies, Santa Clara, CA, USA) and pooled to a final concentration of 2 nM such that each index-tagged sample was present in

equimolar amounts. The latter were subjected to cluster generation and sequencing using an Illumina HiSeq 2500 System in a 2 × 101 paired-end format at a final concentration of 8 pmol.

3.3. Read Pre-Processing and De Novo Assembly

Raw sequence files (in FASTQ format) were subjected to quality control analysis using FastQC. Then, raw reads were fed into *fastq_quality_filter* [16] to remove sequences with a quality score equal or lower than 30 in more than 80% of read length. Trimmomatic 0.32 [17] was run in paired-end mode to trim TruSeq adapter sequences, crop Illumina random hexamers, perform sliding window trimming (window size 10, required quality 30), trim low quality bases from the start of the read, and ensure that the minimum length of resulting reads was at least 75 nucleotides.

Prior to *de novo* assembly by Velvet/Oases [18,19], all high-quality reads were combined into a single dataset. The Velvet assembler [18] was run using the multi-kmer options (k-mers: 65, 67, 69, 71, and 73). Once all the individual k-mer assemblies were acquired, they were merged into a final assembly using Oases [19] and 122,507 contigs (i.e., transcripts) were reconstructed. Then, in order to remove redundancy, all contigs were clustered/collapsed using CAP3 [20] with a 70% similarity threshold.

3.4. Transcriptome Annotation

The assembled transcripts were annotated by BLASTx and BLASTn searches (e-value < 1×10^{-5}) against *Cucumis melo* [21] (version 3.5; https://melonomics.net/files/Genome/Melon_genome_v3.5_Garcia-Mas_et_al_2012/), *Cucumis sativus* (version 1.0; http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Phytozome) and the *Arabidopsis thaliana* (version TAIR 10; https://www.arabidopsis.org/) protein complement, UniProtKB/SwissProt database [22] (http://www.uniprot.org/downloads;release2012_02) and *C. pepo* draft transcriptome (version 3.0; https://cucurbigene.upv.es/db/transcriptome_v3/).

Gene Ontology (GO) and Enzyme Commission (EC) assignments were performed using the Blast2GO suite [11] (version 3.0) in order to classify *C. pepo* transcripts in a standard and controlled vocabulary. Information about domain/motifs patterns within sequences was retrieved using the InterProScan functionality in Blast2GO. Finally, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were assigned based on the Blast2GO results.

BLASTn-based sequence similarity search (e-value < 1×10^{-5}) was performed using Rfam 13.0 [12] as filtering database in order to characterize the potential non-coding RNAs in our dataset.

Transcriptome completeness was evaluated running BUSCO v3 [14] in "tran" mode with the embryophyta specific lineage conserved single copy orthologs derived from OrthoDB v9.

3.5. Mapping Transcripts on the Reference Genome

Zucchini chromosome-scale pseudomolecules (version 4.1) were downloaded from the Cucurbit Genomics FTP server (ftp://cucurbitgenomics.org/pub/cucurbit/genome/Cucurbita_pepo/Cpepp_v4.1.chr.fa.gz). The GFF3 file including gene models and repetitive elements was downloaded from https://bioinf.comav.upv.es/downloads/zucchini/genome_v4.1/. The Maker pipeline (version 3.0) [15] was used to align transcripts on the reference genome with the following settings: split_hit = 20,000, single_exon = 1, single_length = 50, correct_est_fusion = 1, est2genome = 1. All transcripts that have not been successfully aligned on the reference genome by Maker were subjected to a BLASTn search (e-value < 1×10^{-3}) against the reference genome using the following settings: -task blastn, -max_target_seqs 1, -qcov_hsp_perc 10, -perc_identity 80.

database (http://www.uniprot.org/downloads; release 2012_02); cs_1.0: *C. sativus* proteins (v 1.0, http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Phytozome). Blast2GO and InterProScan results are separated by a semicolon.

Author Contributions: A.V. set up the experiment, processed samples for RNA isolation, performed all bioinformatic analyses, and drafted the early version of the manuscript; R.R. conceived the work, designed the experiment, and revised the manuscript; G.C. contributed to experimental design and bioinformatic analyses; R.A.C. contributed to sequence mapping to the reference genome; P.C. contributed to RNA sample extraction; M.C.D. performed aphid infestations; N.D. developed the workflow for NGS data analysis, coordinated and supervised bioinformatic work, and wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Esteras, C.; Gómez, P.; Monforte, A.J.; Blanca, J.; Vicente-Dólera, N.; Roig, C.; Nuez, F.; Picó, B. High-throughput SNP genotyping in *Cucurbita pepo* for map construction and quantitative trait loci mapping. *BMC Genom.* **2012**, *13*, 80. [CrossRef] [PubMed]

2. Paris, H.S. Summer squash: History, diversity, and distribution. *HortTechnology* **1996**, *6*, 6–13.

3. Singh, R.; Singh, K. Life history parameters of aphis gossypii glover (homoptera: Aphididae) reared on three vegetable crops. *Int. J. Res. Stud. Zool.* **2015**, *1*, 1–9.

4. Ebert, T.; Cartwright, B. Biology and ecology of aphis gossypii glover (homoptera: Aphididae). *Southwest. Entomol.* **1997**, *22*, 116–153.

5. Blanca, J.; Cañizares, J.; Roig, C.; Ziarsolo, P.; Nuez, F.; Picó, B. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (*Cucurbitaceae*). *BMC Genom.* **2011**, *12*, 104. [CrossRef] [PubMed]

6. Wyatt, L.E.; Strickler, S.R.; Mueller, L.A.; Mazourek, M. An acorn squash (*Cucurbita pepo* ssp. Ovifera) fruit and seed transcriptome as a resource for the study of fruit traits in cucurbita. *Hortic. Res.* **2015**, *2*, 14070. [CrossRef] [PubMed]

7. Xanthopoulou, A.; Psomopoulos, F.; Ganopoulos, I.; Manioudaki, M.; Tsaftaris, A.; Nianiou-Obeidat, I.; Madesis, P. *De novo* transcriptome assembly of two contrasting pumpkin cultivars. *Genom. Data* **2016**, *7*, 200–201. [CrossRef] [PubMed]

8. Xanthopoulou, A.; Ganopoulos, I.; Psomopoulos, F.; Manioudaki, M.; Moysiadis, T.; Kapazoglou, A.; Osathanunkul, M.; Michailidou, S.; Kalivas, A.; Tsaftaris, A. *De novo* comparative transcriptome analysis of genes involved in fruit morphology of pumpkin cultivars with extreme size difference and development of EST-SSR markers. *Gene* **2017**, *622*, 50–66. [CrossRef] [PubMed]

9. Andolfo, G.; Di Donato, A.; Darrudi, R.; Errico, A.; Aiese Cigliano, R.; Ercolano, M.R. Draft of zucchini (*Cucurbita pepo* L.) proteome: A resource for genetic and genomic studies. *Front. Genet.* **2017**, *8*, 181. [CrossRef] [PubMed]

10. Montero-Pau, J.; Blanca, J.; Bombarely, A.; Ziarsolo, P.; Esteras, C.; Martí-Gómez, C.; Ferriol, M.; Gómez, P.; Jamilena, M.; Mueller, L. *De novo* assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the *Cucurbita* genus. *Plant Biotechnol. J.* **2018**, *16*, 1161–1171. [CrossRef] [PubMed]

11. Götz, S.; García-Gómez, J.M.; Terol, J.; Williams, T.D.; Nagaraj, S.H.; Nueda, M.J.; Robles, M.; Talón, M.; Dopazo, J.; Conesa, A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **2008**, *36*, 3420–3435. [CrossRef] [PubMed]

12. Kalvari, I.; Argasinska, J.; Quinones-Olvera, N.; Nawrocki, E.P.; Rivas, E.; Eddy, S.R.; Bateman, A.; Finn, R.D.; Petrov, A.I. Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **2018**, *46*, D335–D342. [CrossRef] [PubMed]

13. Iseli, C.; Jongeneel, C.V.; Bucher, P. Estscan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, Heidelberg, Germany, 6–10 August 1999; pp. 138–148.

14. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef] [PubMed]

15. Cantarel, B.L.; Korf, I.; Robb, S.M.C.; Parra, G.; Ross, E.; Moore, B.; Holt, C.; Sánchez Alvarado, A.; Yandell, M. Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **2008**, *18*, 188–196. [CrossRef] [PubMed]

16. Gordon, A.; Hannon, G. FASTX-Toolkit: FASTQ/A Short-Reads Preprocessing Tools. Available online: http://hannonlab.cshl.edu/fastx_toolkit (accessed on 15 September 2018).

17. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]

18. Zerbino, D.R.; Birney, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829. [CrossRef] [PubMed]

19. Schulz, M.H.; Zerbino, D.R.; Vingron, M.; Birney, E. *Oases*: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **2012**, *28*, 1086–1092. [CrossRef] [PubMed]

20. Huang, X.; Madan, A. Cap3: A DNA sequence assembly program. *Genome Res.* **1999**, *9*, 868–877. [CrossRef] [PubMed]

21. Garcia-Mas, J.; Benjak, A.; Sanseverino, W.; Bourgeois, M.; Mir, G.; González, V.M.; Hénaff, E.; Câmara, F.; Cozzuto, L.; Lowy, E.; et al. The genome of melon (*Cucumis melo* L.). In Proceedings of the National Academy of Sciences of the United States of America, Bethesda, MD, USA, 17 July 2012; pp. 11872–11877.

22. UniProt Consortium. Uniprot: A hub for protein information. *Nucleic Acids Res.* **2015**, *43*, D204–D212. [CrossRef] [PubMed]