The etiology of watermelon “hard fruit syndrome”: an NGS approach

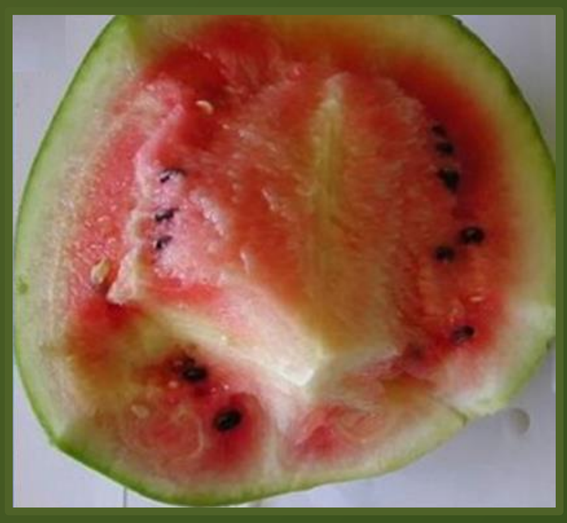
EPPO Workshop
Bari, Italy - 22nd-23rd November 2017.

T. Zaagueri[1], L. Miozzi[2], M. Mnari-Hattab[1], E. Noris[2], G.P. Accotto[2] and A.M. Vaira[2]

[1] Laboratoire de Biotechnologie Appliquée à l’Agriculture, INRA Tunisie, Université de Carthage, 1004 El Menzah, Tunisia.
[2] Institute for Sustainable Plant Protection (IPSP-CNR), Strada delle Cacce 73, 10135 Torino, Italy.

INTRODUCTION

In Tunisia, watermelon (*Citrullus lanatus*) is one of the most common and popular summer fruit crops, with a yearly production of about 400.000 tons. The hard fruit syndrome of watermelon is a well-known syndrome, reported in Tunisia since 1994, when up to 70% incidence was reported in some areas; in recent years the incidence ranged between 10-40%, leading growers to switch to alternative crops.



Watermelon fruits showing the hard fruit syndrome (WHFS)

MATERIAL AND METHODS

Survey and sampling

Watermelon fields were surveyed in different areas of Tunisia during several growing seasons. Ten watermelon fruits showing different degrees of hardness and discoloration of the flesh, whitish inserts, deformation of fruits and seeds, and bad taste, were collected in several Tunisian areas from individual plants during the 2016 growing season.

Nucleic acid extraction and sequencing

For small RNA (sRNA) sequencing, total RNA was extracted from the 10 symptomatic fruits using TRIzol Reagent (Thermo Fisher Scientific, Waltham, MA, USA), according to the manufacturer’s instructions, measured at NanoDrop spectrophotometer ND-1000 (Thermo Scientific, Wilmington, DE, USA) and pooled into one sample; three micrograms of the pooled RNA were sent to Human Genetic Foundation sequencing service (HUGEF, Torino, Italy) for library preparation with the TruSeq RNA library Prep Kit v2 (Illumina, San Diego, CA, USA) and sequencing with Illumina NextSeq 500.

Small RNA Bioinformatic Analysis

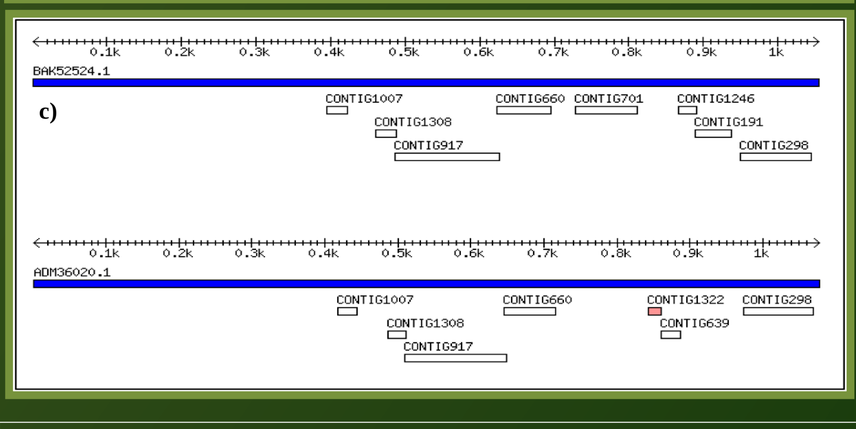
Raw data (Sequence Read Archive database acc. num. SRP119446) were checked for quality reads with FastQC software; Fastx-toolkit was used for removing adapter sequences, low-quality reads and artefacts; sequences shorter than 19 nt and longer than 27 nt were discarded.
Reads were analyzed by the software package **VirusDetect**, using its plant virus database as reference and default parameters. Reads were assembled in contigs with both a reference-guided and a *de novo* assembly approach. *De novo*-assembled contigs were pooled together with those generated from reference-guided assemblies, and processed to remove redundant sequences.
An homology-dependent strategy to identify known and novel virus sequences from the assembled contigs was then employed. Contigs were compared against reference virus nucleotide sequences using BLASTn and against the reference virus protein sequences using BLASTx. Contigs matching the same reference sequence were merged to form the final VirusDetect output, and used to derive the coverage of the reference by virus contigs. Based on contigs length and nucleotide identity with the reference viral genome, viral sequences were selected as candidates for validation.

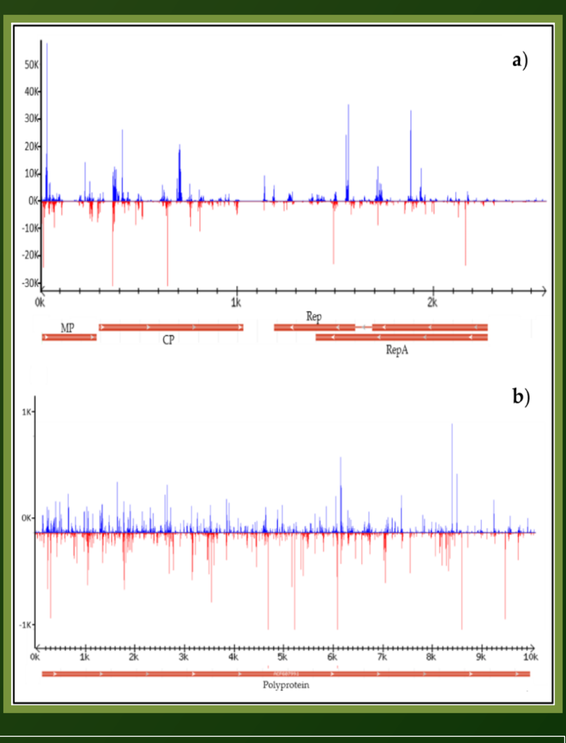
RESULTS

Sequencing resulted in 33,429,123 reads. Once removed low-quality reads and artefacts, 32,765,804 reads were obtained; reads shorter than 19 nt or longer than 27 nt were then discarded, for a total of 20,943,962 remaining reads. Contigs assembly and **BLASTn** and **BLASTx** analyses led to the identification of sequences ascribable to several plant viruses.

CpCDV was represented by 9 contigs showing high identity (98%) with a CpCDV isolate (GenBank acc. No. KC172655) with a 100% coverage; it was the most represented virus in these watermelon plants. **WMV** was represented by 80 contigs, including a 10,051 nt-long contig that represented the entire viral genome and showed 97% identity with a WMV isolate (GenBank acc. No. EU660585). The presence of contigs with an amino acid identity of around 55% with two viruses of the *Amalgavirus* genus suggests that a **novel virus** belonging to this taxonomic group is present in watermelon. The remaining contigs with amino acid identity to badnaviruses and cavemoviruses (73% and 43% respectively) showed a reduced length (<250 nt), a low coverage (about 16%), and were therefore excluded from further analysis.

Viral Species	Genus	Genome Length (nt)	No. of Contigs	Contigs Length (Min–Max) (nt)	No. of Reads	RPKM	Type of Analysis
Chickpea chlorotic dwarf virus (complete genome)	*Mastrevirus*	2573	9	46–2702	3,197,315	59,331	BLASTn
Watermelon mosaic virus (complete genome)	*Potyvirus*	10,051	80	41–10051	402,821	1913	BLASTn
Blueberry latent virus/ *Rhododendron virus A* (fusion protein)	*Amalgavirus*	3162/3231	10	53–423	29,626	447/437	BLASTx
Ambrosia asymptomatic virus 2 (polyprotein)	*Badnavirus*	624	1	134	708	54	BLASTx
Cassava vein mosaic virus (ORF3 protein)	*Cavemovirus*	1956	2	103–244	3506	85	BLASTx





NGS results (table) and analysis of contigs. a) CpCDV: several hot spots are located in coding regions; b) WMV: high number of contigs is linked to high variability of reads, presence of sequence variants, high distribution of reads; c) Amalga-like virus: contigs distribution along the reference polyprotein sequence of *Blueberry latent virus* (GenBank BAK52524.1) upper part and *Rhododendron virus A* (GenBank ADM36020.1), lower part.

DISCUSSION AND CONCLUSIONS

This study confirms that high throughput sequencing analysis of sRNAs is a powerful tool for the identification of both known and new viral sequences and for the definition of new etiological agents. The data obtained by this technique, supported by PCR validation and experimental infection in controlled conditions, allowed to highlight for the first time the causal link between CpCDV and WHFS.
These data are published in: Zaagueri et al. Viruses 2017, 9, 311; doi:10.3390/v9110311