PLANT VIRUSES AND SMALL RNA NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jose Andres Vargas Asencio

December 2017

PLANT VIRUSES AND SMALL RNAS NETWORKS

Jose Andres Vargas Asencio, Ph. D.

Cornell University 2017

In plants, virus infection and small RNA (sRNA) metabolism are closely associated. sRNAs are considered 'master regulators of gene expression and constitute one of the most important defense mechanisms against foreign nucleic acids, such as viruses and viroids. In response to host defenses, most plant viruses have evolved mechanisms to interfere with sRNA biogenesis and activity, with broad effects on host homeostasis. This thesis includes studies on virus and viroid detection by searching for pathogen-specific nucleic acids and pathogen-derived sRNA products of the host defense machinery. Two short reports describe the detection of *Spinach latent virus* and *Australian grapevine viroid*. A third study describes the limited genetic diversity of *Grapevine virus E* infecting grapevines across the US. In a subsequent study, a combination of electron microscopy and next generation sequencing (NGS) technologies were used to elucidate the identity and complete genome sequence of another grapevine infecting virus, namely *Grapevine asteroid mosaic associated virus*; obtaining a full genome sequence allowed an examination and comparison of the genome's functional domains and sequence signatures that reflect strategies for the expression and processing of the virus encoded proteins. Finally, to better understand the role of sRNAs in the regulation of gene expression, a genome wide evaluation was performed of the biogenesis and activity of two kinds of sRNAs, microRNAs and phased interfering small interfering RNAs. Using a combination of multiple types of NGS datasets and the design of a custom bioinformatics pipeline, a broad scale sRNA-mediated regulatory network was described. Evaluation of the network's regulatory

contribution indicated that sRNA-mediated regulation plays a major role in gene expression. A large proportion of genes in *Arabidopsis thaliana* were found to be controlled by sRNAs, including genes involved in most biological processes. Finally, the role of sRNAs as key regulators was further confirmed by the extended number of genes under sRNA control that are involved in others aspects of regulation, in particular transcription factors.

BIOGRAPHICAL SKETCH

Jose Andres Vargas Asencio studied biology at the University of Costa Rica where he obtained a Bachelor of Science degree in 2008. He continued study in the genetics and molecular biology program at University of Costa Rica and received a Master of Science degree in 2011. In 2012 he moved to Ithaca, NY to start a PhD program in the Plant Pathology and Plant-Microbe Biology Department at Cornell University.

DEDICATION

Este trabajo está dedicado a mis padres por su cariño, entusiasmo y apoyo incondicional a través de muchos años de estudio. Dado el valor que conlleva y lo que significa para nosotros, quiero resumir mi sentimiento de gratitud en un simple PURA VIDA.

ACKNOWLEDGMENTS

I especially would like to thank Dr. Keith Perry for being a true mentor. During my time in the lab I learned a great deal about strategically designing, conducting and communicating my research and scientific interests. More than anything I sincerely appreciate Dr. Perry for supporting my venturing into unfamiliar lines of research. This experience, though slow moving at times, proved to be much rewarding for me, as I gain a new set of skills and overall perspective on doing good quality research that I believe will help me move forward into a scientific career.

I also want to show my appreciation to Dr. Jeremy Thompson for countless times in which his laboratory expertise was instrumental to overcoming obstacles. His viewpoint on life and on how to conduct and balance a career in science have been of great value to me, I will dearly remember our conversations.

I would also like to acknowledge the contributions of my committee members: Dr. Zhangjun Fei, Dr. Chris Myers and Dr. Marc Fuchs. Their advice was a key factor in completing a thesis work that required expertise on several research areas, their accessibility and helpful disposition is a highly appreciated quality that I will use to select future collaborators.

Finally, I want to extend my gratitude to the rest of the members of the Perry lab. To Heather McLane for her assistance in many experiments, Bjorn Krenz for his advice in microbiological applications and friendship, and to all the undergrads that participated in related projects.

This work would not have been possible with the financial support provided by USDA-NIFA grant # 2015-67028-23512, The USDA National Clean Plant Network program American Vineyard Foundation, CA Grape Rootstock Research Foundation,

California Department of Food and Agriculture NY Wine and Grape Foundation, Cornell College of Agriculture and Life Science.

TABLE OF CONTENTS

PREFACE

Statement of authorship

Chapters two to five of this thesis have been published and the text corresponds exactly to these publications. For chapter two on *Spinach latent virus* detection (Vargas-Asencio et al., 2013) the experimental work and analysis was entirely my own work and I co-wrote the manuscript with Keith Perry; E. Bush recognized the disease and provided samples; Heather McLane provided technical assistance. For chapter three on detection of *Australian grapevine viroid* (Vargas-Asencio et al., 2016), I conducted the experimental work and analysis and I co-wrote the manuscript with Keith Perry and Marc Fuchs; Alice Wise recognized the disease and provided the samples. For the fourth chapter on *Grapevine virus E* (Vargas-Asencio et al., 2015), the experimental work was performed collaboratively by myself, Fevziye Celebi-Toprak and Jeremy Thompson at Cornell University, and M. Al Rwahnih and A. Rowhani from UC Davis. I co-wrote the manuscript with Keith Perry, Marc Fuchs and M. Al Rwahnih. For chapter five on *Grapevine asteroid mosaic associated virus* (Vargas-Asencio et al., 2017), the experimental work and analysis were performed in collaboration with Jeremy Thompson. I produced and analyzed the next generation sequencing data, as well as the sequence analysis of domains related to putative cleavage sites. Jeremy Thompson conducted the microscopy, and phylogenetic analysis. Klaudia Wojciechowska, Maia Baskerville, Annika L. Gomez were involved in RT-PCR and RACE analysis. I co-wrote the manuscript with Jeremy Thompson and Keith Perry. Finally, chapter six on small RNA regulatory networks, corresponds to my main thesis project. With the advice of my thesis committee, I designed the experiments, conducted all of the experimental

work, designed and wrote the bioinformatics tool and performed the data analysis.

REFERENCES

Vargas-Asencio, J., Al Rwahnih, M., Rowhani, A., Celebi-Toprak, F., Thompson, J.R., Fuchs, M., Perry, K.L., 2015. Limited Genetic Variability Among American Isolates of Grapevine virus E from Vitis spp. Plant Dis. 100, 159–163. doi:10.1094/PDIS-05-15-0556-RE

Vargas-Asencio, J., McLane, H., Bush, E., Perry, K.L., 2013. Spinach latent virus Infecting Tomato in Virginia, United States. Plant Dis. 97, 1663. doi:10.1094/PDIS-05-13-0529-PDN

Vargas-Asencio, J., Perry, K.L., Wise, A., Fuchs, M., 2016. Detection of Australian grapevine viroid in Vitis vinifera in New York. Plant Dis. 101, 848. doi:10.1094/PDIS-11-16-1587-PDN

Vargas-Asencio, J., Wojciechowska, K., Baskerville, M., Gomez, A.L., Perry, K.L., Thompson, J.R., 2017. The complete nucleotide sequence and genomic characterization of grapevine asteroid mosaic associated virus. Virus Res. 227, 82–87. doi:http://dx.doi.org/10.1016/j.virusres.2016.10.001

CHAPTER ONE

INTRODUCTION

This work focuses on studies regarding plant viruses and small RNAs (sRNAs). Though they may seem disparate topics, virus infection and sRNAs are deeply associated. In plants, sRNAs are the products of multiple biogenesis pathways, and they are involved in gene expression regulation at the transcriptional and posttranscriptional level, collectively referred to as RNA silencing (Borges and Martienssen 2015; Wang and Chekanova 2016). Their relevance is supported by recent studies that have proposed sRNAs especially microRNAs (miRNAs) to be 'master regulators of gene expression' in multiple systems (Sun *et al*. 2010; Voorhoeve 2010; Zhai *et al*. 2011). The role of RNA silencing in regulation of the plant defense response to pathogens is well documented (Baldrich *et al*. 2015; Soto-Suárez *et al*. 2017; Peláez and Sanchez 2013; Zhai *et al*. 2011). Beyond regulation of defense related genes, RNA silencing constitutes one of the most important mechanisms of antiviral defense by targeting and processing viral RNAs into virus-derived sRNAs (Obbard *et al*. 2009; Szittya and Burgyán 2013).

In chapters two to five, I describe my published research conducted to identify and characterize a set of plant viruses and a viroid. Besides providing a contribution towards an understanding of the epidemiology and biology of the reported pathogens, these projects serve to established and develop the laboratory techniques and analytical methods that were fundamental for the last chapter, an evaluation of the sRNA-mediated regulatory contribution in *Arabidopsis thaliana*.

Chapters two and three correspond to two studies in which pathogen-derived RNAs were used for plant virus detection. The first study titled "Spinach latent virus infecting tomato in

Virginia, USA" (Vargas-Asencio *et al*. 2013) describes the use of a macroarray-based diagnostic system (Agindotan and Perry 2007a) to detect viral RNAs, both genomic fragments and transcription products. Spinach latent virus had been previously reported only once in the USA, and though it had been described as a latent virus, this study constitutes a second report where it has been associated with disease symptoms (Lebas *et al*. 2007). The second study involves the use of a recently developed strategy designed to identify viruses and viroids by sequencing sRNAs produced from the activity of the plant RNA silencing machinery that targets pathogen genomes and/or transcripts (Llave 2010; Wu *et al*. 2010; Zhang *et al*. 2015). Using this approach, *Australian grapevine viroid* was found infecting grapevine in the US (Vargas-Asencio *et al*. 2016). A full genome was reconstructed from sRNAs and confirmed by reverse transcription - polymerase chain reaction (RT-PCR) and Sanger sequencing, leading to an unequivocal identification. The only previous description of this viroid in the US was based on a hybridization assay where the viroid sequences were not reported (Rezaian *et al*. 1992).

The fourth chapter contains an analysis of the genetic diversity of grapevine virus E (GVE) in the US (Vargas-Asencio *et al*. 2015). Vines from multiple national repositories and commercial vineyards in NY and California were screened for GVE and related viruses. This virus was found to be established in the US, and sequence analysis of a portion of its genome revealed limited genetic diversity across isolates.

The fifth chapter describes the use of electron microscopy and a combination of next generation sequencing technologies to identify and elucidate the complete genome sequence of *Grapevine asteroid mosaic associated virus* (Vargas-Asencio *et al*. 2017). The resulting genome sequence was used to evaluate its relationship within the genus *Marafivirus*, and sequence features related to the expression and processing of the virus proteins.

The laboratory techniques and analytical methods developed and established in the completion of the studies described in chapters two to five were applied and expanded to address a more complicated project aimed to identify sRNA-mediated regulatory networks in *A. thaliana*. A bioinformatics pipeline was designed to analyze next generation sequencing datasets produced in this study and those from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) repository in order to identify biologically relevant sRNA-mediated networks. The results corroborated the role of sRNAs as master regulators of gene expression by demonstrating a high proportion of genes involved in multiple biological processes to be under sRNA control. Interestingly, a disproportionate number of genes involved in the regulation of gene expression were shown to be regulated by sRNAs. The resulting networks provide a reference frame for comparative and functional studies on the role of sRNA-mediated regulation, as well as a resource to elucidate the interplay between multiple regulators of gene expression from a systems biology point of view. A long-term goal is to understand how plant virus infection affects sRNA-mediated regulation.

CHAPTER TWO

SPINACH LATENT VIRUS INFECTING TOMATO IN VIRGINIA, USA

Text is verbatim minus references as in Vargas-Asencio, J., McLane, H., Bush, E., and Perry, K. L. 2013. Spinach latent virus Infecting Tomato in Virginia, United States. Plant Disease. 97:1663.

Plants in a single field of commercial tomato (*Solanum lycopersicum*) of unidentified cultivars in Virginia in July, 2012, were observed showing stunting, leaf distortion, twisting and thickening, discoloration, and color streaking and ringspots on fruits. Serological tests were negative for *Cucumber mosaic virus*, *Groundnut ringspot virus*, *Tomato spotted wilt virus*, *Tomato chlorotic spot virus*, *Impatiens necrotic spot virus*, *Tobacco mosaic virus*, and *Tomato bushy stunt virus* (Agdia, Inc., Elkhart, IN). Using a membrane-based microarray (Perry and Lu 2010), hybridization was observed to 8 of 9 70-mer oligonucleotide probes of *Spinach latent virus* (SpLV; genus *Ilarvirus*, family *Bromoviridae*). To confirm the hybridization results, complementary DNA (cDNA) was synthesized using random hexamers and MMLV reverse transcriptase (Promega, Madison, WI), followed by PCR amplification using ilarvirus degenerate primers (Untiveros *et al*. 2010). Fragments of approximately 380 bp were amplified and directly sequenced (GenBank accession KC_466090); a BLAST search showed a 99% identity to the SpLV RNA 2 reference genome (NC_003809). Primers for SpLV RNA1 (SpLVRNA1f-GGTGTCACCATGCAAACTGG, SpLVRNA1r-AGCTCTTCGTAATAGGCCTGC) and SpLV RNA3 (SpLVCPf-GAAGTCTTTCCCAGGTGAGCA, SpLVCPr-AGGTGGGCATATGGACTTGG) were designed and cDNA was amplified using the IQ supermix (Biorad, Hercules, CA) with thermocycling of 94°C for 4 min, 35x (94°C 45 s, 55°C 45 s, 72°C 45 s), and 72°C for 10 min. The resulting fragments of 538 bp for RNA1 (KC_466088)

and 661 bp for RNA3 (KC_466089) showed 100% identity to reference genome sequences for SpLV (NC_003808 and NC_003810, respectively). To demonstrate virus transmissibility, *Chenopodium quinoa* plants were mechanically inoculated using tomato leaf material (same source described above) ground in 30 mM Na_2HPO4 buffer, pH 7.0. Necrotic spots developed on the inoculated leaves 10 dpi. Younger, non-inoculated leaves showed yellow mottling and tested positive for SpLV by RT-PCR (two of two plants tested). The detection of SpLV is rarely reported, with only one record from the USA (Liu and Duffus 1986). Although SpLV is described as a latent virus, it has been found associated with tomato fruit symptoms in New Zealand (Lebas *et al*. 2007). It is not known if the fruit ringspot and other symptoms on the Virginia samples were due to this virus infection. Since SpLV is seed transmissible and seed production takes place in different parts of the world, it has the potential to spread with germplasm and become more widespread in North America.

CHAPTER THREE

DETECTION OF AUSTRALIAN GRAPEVINE VIROID IN *VITIS VINIFERA* IN NEW YORK

Text is verbatim minus references as in Vargas-Asencio, J., Perry, K. L., Wise, A., and Fuchs, M. 2016. Detection of Australian grapevine viroid in Vitis vinifera in New York. Plant Disease. 101:848.

Viroids are the most common and widely distributed infectious agents in grapevine (*Vitis* spp.), as exemplified by Hop stunt viroid, which is found globally in a majority of tested vines. By contrast, there are relatively few reports of Australian grapevine viroid (AGVd). This viroid from the genus Apscaviroid in the family Pospiviroidae has only been reported from Australia, the United States, Tunisia, China, Iran and Italy (Gambino *et al*. 2014). The sole report of AGVd in the United States is from a single study where the viroid was detected in one of 12 samples from California by a combination of northern hybridization and reverse transcription - polymerase chain reaction (RT-PCR) (Rezaian *et al*. 1992); the viroid was not sequenced and there are no subsequent reports on the presence of AGVd in the United States. We report the detection of AGVd in two vineyards of *V. vinifera* cultivar 'Syrah' in Suffolk County, New York. The ‘Syrah’ vines showed a distinctive late season reddening of leaves in the 2014 through 2016 seasons. Vines were sampled in 2014, nucleic acid extracts prepared, and Illumina-compatible small RNA libraries constructed, sequenced (Illumina Hiseq 4000; single-end 50 nt reads) and analyzed as described in (Vargas-Asencio *et al*. 2017). AGVd was detected in the Syrah vines and a complete genome sequence was assembled (GenBank accession number KY081960). The genome of this AGVd variant showed >99% nucleotide sequence identity to that of the original variant (Rezaian *et al*. 1992). To confirm this initial detection, vines were resampled in 2016 and a RT-PCR amplicon

of approximately 270 nucleotides in length was produced using the methods of Gambino *et al.* (2014); the product was detected in three of five vines and sequenced by the dideoxy DNA sequencing method. This second variant also showed >99% nucleotide sequence identity to the original variant. This study unequivocally confirms the presence of AGVd in North American vineyards. Additional work is needed to assess the effects of AGVd on vine health and any role in the leaf reddening.

CHAPTER FOUR

LIMITED GENETIC VARIABILITY AMONG AMERICAN ISOLATES OF GRAPEVINE VIRUS E FROM *VITIS* SP.

Text is verbatim minus references as in Vargas-Asencio, J., Al Rwahnih, M., Rowhani, A., Celebi-Toprak, F., Thompson, J. R., Perry, K. L. 2015. Limited Genetic Variability Among American Isolates of Grapevine virus E from Vitis spp. Plant Disease. 100:159–163.

Abstract

A survey for the presence of Grapevine virus E (GVE, genus Vitivirus, family Betaflexiviridae) in vineyards in New York (NY) and California (CA) was conducted using macroarray hybridization and/or a reverse transcription polymerase chain reaction (RT-PCR) assays. In NY, GVE was detected in 10 of 46 vines of *Vitis labrusca*, one *V. riparia* and one *Vitis* hybrid. All GVE-infected NY vines were co-infected with Grapevine leafroll-associated virus-3. In CA, GVE was detected in 8 of 417 vines of *V. vinifera*. All GVE-infected CA vines were also co-infected with one of the leafroll-associated viruses and other vitiviruses. In order to assess the genetic diversity among GVE isolates, a viral cDNA was amplified by RT-PCR and a 675 nt region that included the 3'-terminus of the CP gene, a short intergenic region and the 5'-terminus of the putative nucleic acid binding protein gene were sequenced. All 20 GVE isolates sequenced in this study were very closely related, with >98% nt identity to the SA94 isolate from South Africa. These findings confirm the presence of GVE in major grape growing regions of the United States and indicate a very low level of genetic diversity.

Introduction

Grapevine production is significantly compromised by virus infections, with resulting reductions in the quantity and quality of the crop (Komar *et al*. 2007; Martelli 2014). To date, 67 viruses have been reported to infect grapevines (Martelli 2014; Maliogka *et al*. 2015; Al Rwahnih *et al*. 2015). Four virus-associated disease complexes are of particular concern worldwide: i) leafroll, ii) infectious degeneration/decline, iii) rugose wood, and iv) fleck. The viruses associated with these diseases are, respectively: members of the genera Ampelovirus, Closterovirus and Velarivirus in the family Closteroviridae (leafroll viruses) (Maree *et al*. 2013; Martelli *et al*. 2012); members of the genus Nepovirus (nematode-transmitted viruses) and Strawberry latent ringspot virus in the family Secoviridae (Martelli and Boudon-Padieu 2006); members of the genera Vitivirus and Foveavirus (rugose wood-associated viruses) in the family Betaflexiviridae (Rosa *et al*. 2011), and members of the genus Maculavirus, family Tymoviridae (fleck virus).

Viruses of the genus *Vitivirus* have gained attention in recent years, and there are five recognized vitiviruses associated with grapevine, Grapevine virus A, Grapevine virus B, Grapevine virus D, Grapevine virus E (GVE), and Grapevine virus F (Preez *et al*. 2011; Martelli 2014). GVE was first reported associated with rugose wood symptoms in a Japanese grape cultivar in 2008 (Nakaune *et al*. 2008). Following this report, two new isolates were described along with their full-length sequences, one from South Africa *(Coetzee et al. 2010*) and the other from the U.S.A. (Alabi *et al*. 2013). More recently, GVE was detected infecting grapevine in China (Fan *et al*. 2013) and several grapevine species in New York (NY), including the *Vitis labrusca* cultivar Concord (Thompson *et al*. 2014). Although information about Concord-infecting viruses is limited, there are reports of many of the viruses associated with the major groups of grape diseases being detected in Concord vines (Bahder *et al*. 2013; Ramsdell *et al*. 1983; Soule *et al*. 2006;

Uyemoto *et al*. 1977; Thompson *et al*. 2014). Based on our earlier findings and the small number of reports of GVE, a limited survey was initiated with the objectives of: i) assessing the occurrence of GVE in vines in NY and CA, and ii) obtaining an estimate of the genetic diversity for this virus in North America. Since preliminary testing of Concord vines in NY had shown the presence of Grapevine leafroll associated virus-3 (GLRaV-3), the testing for this virus was an additional objective.

Materials and Methods

Sampling of grapevines. To determine the occurrence of GVE in North American vines, collections were made from six sites or regions (Table 1). In NY samples of canes were taken from dormant vines without regard to symptoms. Samples from Western NY (Chautauqua, Erie and Ontario counties) consisted of Concord vines collected from 16 different commercial production sites. Additional samples were obtained from the USDA National Plant Germplasm System, Plant Genetic Resources Unit in Geneva, NY. In order to determine the occurrence of GVE in CA, samples were obtained from the USDA National Clonal Germplasm Repository (NCGR) in Winters, CA, the Davis Grapevine Virus Collection (DGVC), the Foundation Plant Services (FPS) collection, and commercial vineyards in Napa Valley, CA.

Table 1. Origin of grapevine samples tested for viruses in this

Collection Sites / Regions	**Species**	**# of vines**
Western New York	*Vitis labrusca* (Concord)	44
USDA National Plant Germplasm System, Plant Genetic Resources Unit, Geneva, NY	*V. riparia*	1
	V. hybrid	1
	V. labrusca (Concord)	2
University of California, Davis Grapevine Virus Collection	*V. vinifera*	198
USDA National Plant Germplasm System National Clonal Germplasm Repository, Davis, CA	*V. vinifera*	77
Foundation Plant Services (FPS), Davis, CA	*V. vinifera*	40
Napa Valley, CA	*V. vinifera*	102

Virus detection. Two parallel surveys were undertaken, with the detection approaches taken in NY and CA differing. To detect the presence of virus in cane samples from NY, a multiplex macroarray method was employed, as described in (Thompson *et al*. 2014). Additionally, samples were specifically tested for GLRaV-3 by double antibody sandwich, enzyme-linked immunosorbent assay (DAS-ELISA) (Bioreba, Reinach), and for GVE by reverse transcription polymerase chain reaction (RT-PCR) using primers GVE-1-For and GVE-Rev for the amplification of a 992-bp fragment spanning ORFs 4 and 5 in the genome of the SA94 isolate (GenBank accession GU903012) *(Coetzee et al. 2010*). For Californian samples, nucleic acid (NA) extracts were prepared from each of the grapevine samples as described by (Al Rwahnih *et al*. 2011). About 0.2 g of frozen leaves petioles was homogenized using a HOMEX grinder (Bioreba) and NA extracts were prepared using a MagMAX™-96 viral RNA isolation kit (Invitrogen, Grand Island) following the manufacturer's protocol. Extracted NA samples were analyzed for the

presence of GVE by RT-PCR as described above. Samples were also tested for mixed infection with grapevine leafroll-associated viruses and vitiviruses by quantitative RT-PCR (qRT-PCR) using TaqMan probes on the ABI 7900 HT Fast real-time PCR system (Invitrogen), as described previously (Klaassen *et al*. 2011).

Sequence analysis. To evaluate nucleotide diversity and look for evidence of selection pressures driving the evolution of GVE populations, a 675 nt RT-PCR region from each isolate was sequenced using primers GVE-1-For and GVE-Rev *(Coetzee et al. 2010*). To sequence isolates from NY, the PCR fragment was produced using AccuPrime™ Taq DNA Polymerase (Life Technologies, Grand Island) and directly sequenced using the amplification primers. For each sample, PCR products from two independent PCR experiments were sequenced and there were no discrepancies observed between sequences. For isolates from CA, the amplified PCR products were analyzed by electrophoresis using a 1% agarose gel with Tris-acetate-EDTA (TAE) buffer. Amplicons of GVE were eluted from gels using the ZymoClean Gel DNA Recovery Kit (Zymo Research Corp, Irvine), quantified and sequenced using GVE-1-For and GVE-Rev primers by Sanger sequencing at the UC-Davis sequencing facility (http://dnaseq.ucdavis.edu).

A multiple alignment was produced using the Muscle algorithm (Edgar 2004). A maximum likelihood tree was constructed using Topali v2 (Milne *et al*. 2008) to determine the relationship between the isolates sequenced in this study, as well as their relationships to previously reported North American, African and Asian isolates. Alternatively, an alignment of a shorter 471-nt region from the previous alignment was designed to include only the CP ORF; this facilitated an alignment with the corresponding sequences from more divergent vitivirus species in order to assess the relative position of GVE in this group. Nucleotide sequence diversity was measured, as defined by (Nei and Li 1979), and evidence of selection pressure was evaluated using

a Tajima's D (Tajima 1989); both analyses were conducted as implemented in DnaSP (Librado and Rozas 2009).

Results

A limited survey of viruses in Concord vines in Western New York. In order to assess the occurrence of GVE and GLRaV-3 in Concord vines in NY, a limited survey was performed including samples from production, nursery and repository vineyards. Of the 46 Concord vines (*V. labrusca*) tested, 10 (22%) showed positive signals for hybridization to oligonucleotide probes for GVE in the macroarray assay (Fig. 1, Table 2).

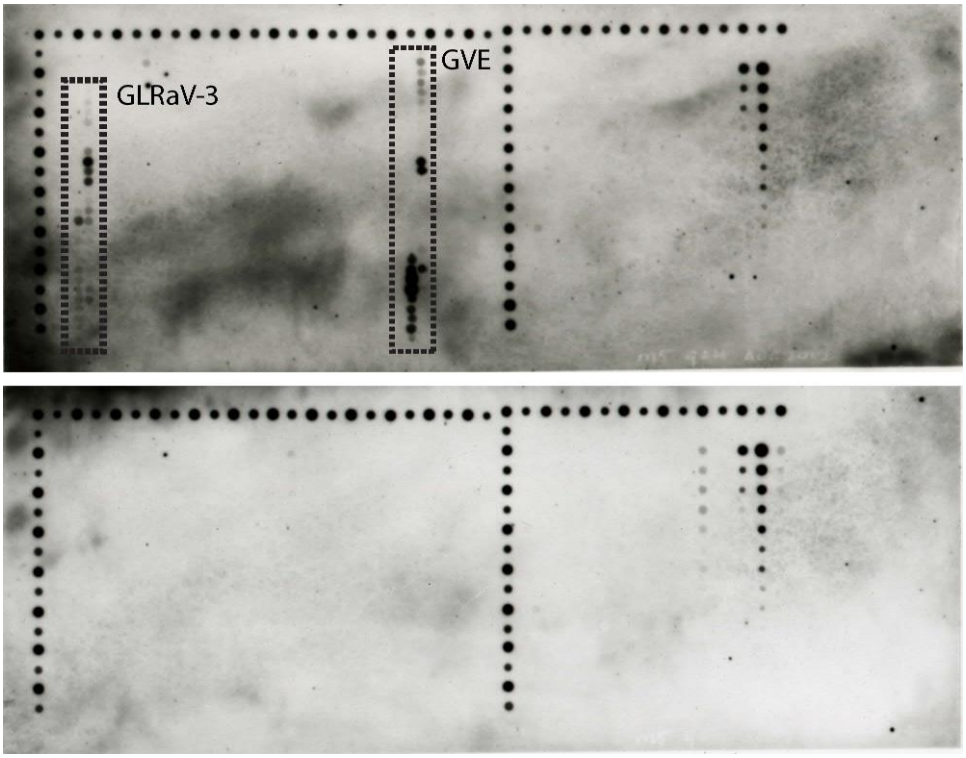

Fig. 1. Macroarray detection of grapevine viruses. Above, a nucleic acid extract from *Vitis labrusca* cv. Concord sample MacsR97 was purified, labeled, hybridized to the nylon membrane and resolved as described by Thompson *et al.* (2014). The boxed areas show sections of the membrane with probes specific to *Grapevine leafroll-associated virus-3* and *Grapevine virus E* (as labeled). Hybridization signals along the top and in vertical columns are those of the macroarray controls. Below, a macroarray of a negative control nucleic acid extract from the reference Concord vine DC1-1, processed as described above.

Nucleic acid extracts from vines testing positive for GVE hybridized with 8 to 20 of the 24 GVE-specific oligonucleotide probes on the array. The presence of GVE sequences in infected plants was confirmed by RT-PCR followed by direct sequencing of the amplicons. Two additional clones of *Vitis* sp. from the USDA germplasm repository in Geneva, NY also tested positive for GVE; these were the *Vitis hybrid* 'Remaily 66-54-2' clone PI588332 and the *Vitis riparia* clone PI588344 (Table 2). All of the GVE-infected vines in NY also showed positive signals for hybridization to 6 to 25 of the 44 oligonucleotide probes for GLRaV-3. Thus, all GVE-infected Concord vines were co-infected with GLRaV-3, consistent with results from the ELISA testing. No other grapevine viruses were detected in this sample of 46 Concord vines.

Table 2. Mixed infection of *Grapevine virus E* and other viruses in infected grapevines from New York and California

New York samples	**Host**	**Source***	**Viruses detected****
DAA#1	*Vitis labrusca*	Commercial vineyard	GLRaV-3, GVE
DAA#4	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
DAA#5	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
DAA#6	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
HAA#9	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
DNSP	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
VD1-1	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
MacsR97	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
BW3-1	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
BH1-3	*V. labrusca*	Commercial vineyard	GLRaV-3, GVE
10-102	*V. riparia*	NPGS, Geneva	GLRaV-3, GVE

5-2	*V.* hybrid	NPGS, Geneva	GLRaV-3, GVE
California samples			
ARM VSV 11 V25	*V. vinifera*	DGVC	GLRaV-3, GRSPaV, GVA, GVE
ARM VSV 12 V22	*V. vinifera*	DGVC	GLRaV-1, -2, -3, GRSPaV, GVA, GVE
K5-56	*V. vinifera*	NPGS, Davis	GLRaV-1, -2, -3, -5, GRSPaV, GVA, GVE
K7-57	*V. vinifera*	NPGS, Davis	GLRaV-2, -3, GRSPaV, GVA, GVD, GVE
K8-45	*V. vinifera*	NPGS, Davis	GLRaV-2, -3, GVA, GVB, GVD, GVE,
K8-46	*V. vinifera*	FPS, Davis	GLRaV-2, -3, GVA, GVB, GVD, GVE
K8-53	*V. vinifera*	NPGS, Davis	GLRaV-2, -3, GVA, GVB, GVD, GVE
K8-59	*V. vinifera*	NPGS, Davis	GLRaV-1, -3, GVA, GVB, GVD, GVE

*Source acronyms are: NPGS, National Plant Germplasm System; DGVC, Davis Grapevine Virus Collection; FPS, Foundation Plant Services.
**Virus acronyms are: GLRaV-1, -2, -3, *Grapevine leafroll-associated virus-1*, -2, -3; GRSPaV, *Grapevine rupestris stem pitting-associated virus*; GVA, *Grapevine virus A*; GVB, *Grapevine virus B*; GVD, *Grapevine virus D;* GVE, *Grapevine virus E*.

GVE present in *V. vinifera* cultivars in California. In parallel with the work in NY, a total of 417 Californian vines (*V. vinifera*) were screened for the presence of GVE and other viruses by RT-PCR. GVE was detected in eight of the clones (~2%), all from germplasm and virus collections (Table 2). None of the 102 vines from commercial vineyards tested positive for GVE. All of the GVE-positive vines also harbored between three to six additional viruses, including at least one leafroll virus and one additional vitivirus.

Sequence analysis. To confirm the presence of GVE and assess the sequence diversity among isolates of this virus, a GVE-specific cDNA was amplified by RT-PCR, sequenced directly or after cloning into pGEM-T (Promega Corp., Madison, WI), and a 675-nt region analyzed. The analyzed cDNA spans 471 nt of the 3' end of the coat protein (CP) gene (ORF4), a 17-nt intergenic

region and 187 nt of the 5' end of the putative nucleic acid binding protein gene (ORF5) in the genome of GVE-SA94 (nt positions 6624-7299 in GenBank accession GU903012). All 20 isolates from this study and two described previously (Thompson *et al*. 2014) were sequenced (GenBank accessions KR062097-KR062118) and shown to be closely related (>98% nt identity; 100% aa identity) to GVE-SA94 from South Africa and to form a separate clade from the Asian isolates TvAQ7 and GFMG-1 (GenBank accessions AB432910 and KF588015 respectively, Fig. 2). This study further supports the relationships between vitiviruses (Alabi *et al*. 2013), showing GVE to be the most distantly related member of the group (Fig. 2A). Additional analyses revealed very low overall nucleotide diversity (π= 0.005), and a non-significant trend towards negative selection (D= -1.59, p >0.05).

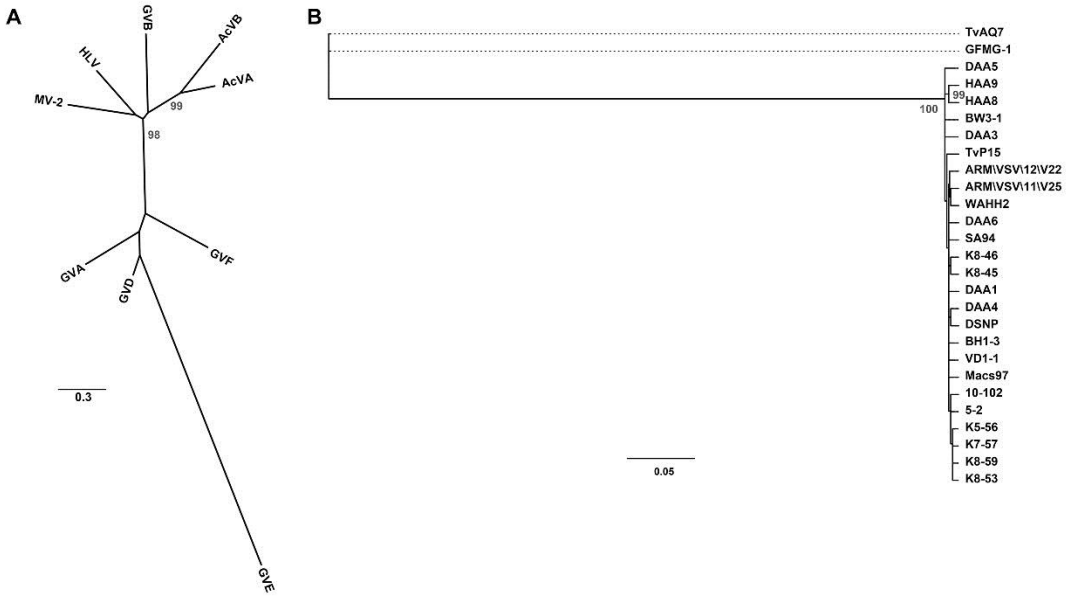

Fig. 2. A) Unrooted Maximum likelihood tree (model K80+G (Kimura 1980)) of a fragment of the CP ORF showing the evolutionary relationships of vitiviruses. Bootstrap values correspond to 100 repetitions. Only values above 70 are shown. Virus acronyms and corresponding accession number) are: AcVA, Actinidia virus A (JN427014); AcVB, Actinidia virus B (NC_016404); GVA, Grapevine virus A (NC_003604); GVB, Grapevine virus B (NC_003602); GVD, Grapevine virus D (JQ031715); GVE (GU903012); GVF, Grapevine virus F (NC_018458), HLV, Heracleum latent virus (X79270); and MV-2, Mint virus 2 (AY913795). B) Unrooted Maximum likelihood tree (model TVMef+G (Posada 2003)) of a 675 nt fragment of the GVE genome showing the evolutionary relationships between GVE isolates from North America, South Africa and Asia.

Bootstrap values correspond to 100 repetitions. Only values above 70 are shown. Isolate acronyms and corresponding accession numbers are GVE-GFMG-1 (KF588015), GVE-TvAQ7 (AB432910), TvP15 (AB432911), GVE-WAHH2 (JX402759), and GVE-SA94 (GU903012).

Discussion

GVE is established in the United States. GVE infection was detected in multiple vines from NY and CA; together with a recent report of this virus from Washington State (Alabi *et al*. 2013), this indicates the virus is established in the major grape growing regions of the United States. GVE was observed commonly in Concord vines in NY, being found in 22% of the 46 vines tested. The detection of infected vines in commercial vineyards in NY highlights the potential for spread, as these vines represent reservoirs from which the virus could be transmitted to other production vineyards (secondary spread). While GVE was detected in CA, it was only seen in grapevine collections, not in commercial vines. However, the limited scope of this survey precludes an assessment of incidence or distribution.

GVE host range and effects. GVE has previously been reported as infecting cultivars of *V. labrusca* (Nakaune *et al*. 2008; Thompson *et al*. 2014) and *V. vinifera* (Alabi *et al*. 2013; Coetzee, Freeborough, *et al*. 2010). In this study, GVE was detected in cultivars of these two species, and additionally from *V. riparia* and a *Vitis hybrid*. For GVE (and other vitiviruses), there are no reported hosts or reservoirs outside of *Vitis* species, although under experimental conditions, some vitiviruses can be mechanically transmitted to and propagated in herbaceous hosts (Nakaune *et al*. 2008; Preez *et al*. 2011). GVE was shown to be transmitted by the mealybug *Pseudococcus comstocki* in Japan (Nakaune *et al*. 2008). Thus, GVE has persisted in infected vines and has the potential to spread in planting stocks, but at present there is no indication of spread by mealybugs in North America. Vitiviruses, in particular GVA and GVB, are associated with rugose wood disease symptoms (Goszczynski and Jooste 2002; Habili and Randles 2012; Martelli and Boudon-

Padieu 2006; Rosa *et al*. 2011) although inferring causal relationships through a demonstration of Koch's postulates is lacking for these and most other viruses in grapevine. The extent to which GVE affects vine health and productivity alone or in combination with other viruses remains to be determined.

Genetic diversity of GVE in the United States is limited. The GVE isolates described in this study exhibit limited genetic diversity. A comparison of the GVE sequences from this study with those available in public databases showed US isolates to be very closely related to the SA94 isolate from South Africa *(Coetzee et al. 2010*). Worldwide, there are two distinct genetic lineages of GVE among the fully sequenced genomes. One lineage is represented by isolates SA94 from South Africa (GenBank accession GU903012) and WAHH2 from the USA (GenBank accession JX402759). The second lineage is typified by the isolates TvAQ7 from Japan (GenBank accession AB432910) and GFMG-1 from China (GenBank accession KF588015). The nucleotide sequence identity within groups is >98% (100% coat protein amino acid identity), while identities between groups is only 70% (87% coat protein amino acid identity). Results from this study suggest that only the SA94/WAHH2 genetic lineage is present in the United States, as no isolates similar to the TvAQ7 and GFMG-1 isolates from Japan and China, respectively, were detected. Interestingly, with the recent submissions of sequences of new GVE isolates from China (GenBank accessions KF588017 to KF588034), it is apparent that both genetic lineages are present in this region. The limited sequence identity between lineages of GVE highlights the necessity of using robust diagnostic primers to avoid false negatives in germplasm testing.

The methodologies used to detect GVE in NY and California differed and each have their limitations. The macroarray method is less sensitive than PCR, but is relatively robust in detecting virus strain sequence variants (Agindotan and Perry 2007b; Thompson *et al*. 2014). For the GVE

isolates detected, hybridization signals were very strong and sensitivity was not limiting. PCR is more sensitive, but it is sequence specific and may fail to detect sequence variants. After the conclusion of this work, it was brought to our attention that the primer GVE-1-For designed by *(Coetzee et al. 2010*) has a 3'-terminal mismatch to the second lineage isolates described in the literature (e.g. isolates TvAQ7 and GFMG-1). This might have resulted in a failure to detect some isolates in California, but not isolates in NY where the primary detection method was the macroarray.

All GVE-infected vines are co-infected with other viruses. Concord vines in NY infected with multiple vitiviruses were reported by (Thompson *et al*. 2014). In the present survey using the same methodology, GVE was the only vitivirus observed in Concord vines, and all GVE-positive Concord samples were also infected with GLRaV-3. An early co-infection of planting stocks commonly used among growers and nursery operations in NY could explain this observation. In a Concord vine survey conducted in Washington state, GVE was not tested for and no vitiviruses were detected, but GLRaV-3 was the most prevalent leafroll-associated virus observed (Bahder *et al*. 2013), consistent with the results from NY. By contrast, in *V. vinifera* samples from CA, infections with multiple vitiviruses were observed, and all GVE positive samples were co-infected with at least one other vitivirus and a leafroll-associated virus (Table 2). Virus co-transmission among grapevines carried out by mealybug vectors has been demonstrated with vitiviruses and ampeloviruses (Hommay *et al*. 2008; Le Maguet *et al*. 2012) and this would result in multiple virus infections as observed in the *V. vinifera* surveyed in this study.

This study demonstrates GVE is established in commercial plantings of Concord vines in the eastern US. Whether GVE is also present in Concord vines in the western US remains to be determined. This virus was also observed in field-planted collections of *V. vinifera* in CA, but it

has not been reported from commercial vineyards. Detection methods based on RT-PCR are currently used to screen for GVE in foundation stocks and these efforts will help to limit the additional spread of this virus.

Acknowledgements

We would like to thank participating growers for their cooperation, especially Rick Dunst of Double A Vineyards. The input from anonymous reviewers is gratefully acknowledged. Support for this work came from the U.S. Department of Agriculture Animal Plant Health Inspection Service as part of the National Clean Plant Network, the NY State Department of Agriculture and Markets, the American Vineyard Foundation, the NY Wine and Grape Foundation, and the NY Farm Viability Institute.

CHAPTER FIVE

THE COMPLETE NUCLEOTIDE SEQUENCE AND GENOMIC CHARACTERIZATION OF GRAPEVINE ASTEROID MOSAIC ASSOCIATED VIRUS

Text is verbatim minus references as in Vargas-Asencio, J., Wojciechowska, K., Baskerville, M., Gomez, A. L., Perry, K. L., and Thompson, J. R. 2017. The complete nucleotide sequence and genomic characterization of grapevine asteroid mosaic associated virus. Virus Research. 227:82–87

Abstract

In analyzing grapevine clones infected with Grapevine red blotch associated virus, we identified isometric particles, the identity of which was confirmed by mass spectrometry to be Grapevine asteroid mosaic associated virus (GAMaV). Using a combination of RNA-Seq, sRNA-Seq, and Sanger sequencing of RT- and RACE-PCR products, we obtained a full-length genome sequence consisting of 6719 nucleotides without the poly-A tail. The virus possesses all of the typical conserved functional domains concordant with the genus Marafivirus and lies evolutionarily between Citrus sudden death associated virus and Oat blue dwarf virus. A large shift in RNA-Seq coverage coincided with the predicted location of the subgenomic RNA involved in coat protein (CP) expression. Genus wide sequence alignments confirmed the cleavage motif LxG(G/A) to be dominant between the helicase and RNA dependent RNA polymerase (RdRp), and the RdRp and CP domains. A putative overlapping protein (OP) ORF lacking a canonical translational start codon was identified with a reading frame context more consistent with the putative OPs of tymoviruses and Fig fleck associated virus than with those of marafiviruses.

BLAST analysis of the predicted GAMaV OP showed a unique relatedness to the OPs of members of the genus Tymovirus.

There are presently sixty-four viruses recorded to infect grapevine (*Vitis* spp.) (Maliogka *et al*. 2015), making it one of the most receptive to virus infection of any cultivated crops. The vast majority of these viruses are single-stranded positive sense RNA (ssRNA) viruses that can be classified into four taxonomic families; Betaflexiviridae, Closteroviridae, Secoviridae and Tymoviridae. The first three families contain species that have significant detrimental effects on grapevine, and in particular on *V. vinifera* in the form of rugose wood, leafroll and infectious degeneration diseases, respectively (Wilcox *et al*. 2016; Martelli 2014). The symptoms of tymovirus infection are associated with grapevine fleck complex, but in general are poorly understood. Recently we carried out a survey using a newly established macroarray capable of simultaneously detecting up to thirty-eight viruses on grapevines. Of the ninety-nine vines tested, forty-six were found to be virus infected, and the most represented family of infecting viruses was Tymoviridae with twenty-three positive samples (Thompson *et al*. 2014)

The family Tymoviridae is divided into three genera; Tymovirus, Marafivirus and Maculavirus (Dreher 1999). They are monopartite, ssRNA viruses that form isometric particles approximately 30 nm in diameter. The viral RNA at the 5'-terminus is believed to be capped due to the presence in all members of a methyltransferase domain, while at the three prime terminus a tRNA-like structure or poly-A tail have been identified. There are five members of the family Tymoviridae that are described as infecting grapevine: Grapevine fleck virus (GFkV) of the genus Maculavirus and the related Grapevine red globe virus (GRGV); and Grapevine Syrah virus 1 (GSyV1), Grapevine asteroid mosaic associated virus (GAMaV) and Grapevine rupestris vein feathering virus (GRVFV) belonging (tentatively) to the Marafivirus genus. Four of these viruses (GAMaV, GFkV, GRGV and GRVFV) are associated with the fleck complex. This disease is found worldwide, and because of its latency sanitation methods are ineffectual (Martelli 2014). Of

disease associated with the five viruses, asteroid mosaic symptoms were the first to be described in California (Hewitt 1954). It took forty years for a non-mechanically transmissible virus to be connected to the disease (Boscia *et al*. 1994) with the report of partial sequences of GAMaV (Abou Ghanem-Sabanadzovic *et al*. 2012; Sabanadzovic *et al*. 2000). In general, GAMaV is believed to have a distribution limited to the United States and specifically California (Martelli 2014) although recent reports suggest it is more widespread (Jo *et al*. 2015; Thompson *et al*. 2014; Xiao and Meng 2016).

Vitis vinifera GRBaV symptomatic plants GV30 and GV32 propagated from the mother Cabernet franc vine NY358 (Krenz *et al*. 2012) were maintained in greenhouse conditions (20-25°C) for approximately nine months of the year alternating with a vernalization period (8°C). Virion enrichment was carried out according to Caciagli *et al.* (2009) with modifications. Sixty-two g of leaf and petiole material from grapevine clone GV30 was ground to a powder in a mortar and pestle with the aid of liquid nitrogen and resuspended in 310 ml extraction buffer (0.5 M phosphate buffer pH 6.0, 0.1% β-mercaptoethanol, 1% Triton and 0.1% Driselase) and incubated overnight with gentle mixing at 4°C. 15% chloroform was then added to the homogenate before vortexing vigorously for 1 minute. The mixture was then centrifuged at 8,000 x g for 15 min at 4°C. The resulting supernatant was carefully transferred to fresh centrifuge tubes and spun at 200, 000 x g for 2h at 4°C. The resulting pellet was resuspended in 5ml 0.5 M phosphate buffer (pH 7.0) containing 2.5 mM EDTA. The resuspension was clarified twice with two low speed spins of 8,000 x g for 15 min at 4°C before a final high speed spin of 400, 000 x g for 1h at 4°C. The pellet was then resuspended in 0.5ml 0.5 M phosphate buffer (pH 7.0) containing 2.5 mM EDTA and stored on ice with 0.02% sodium azide.

Carbon coated formvar 300-mesh copper grids were spotted with 10µl of the enriched virion preparation dried with filter paper, stained with 5% ammonium molybdate and examined with a JEOL 100CXII transmission electron microscope. Microscopic examination of the virion preparation revealed the sparse presence of virus-like spherical particles approximately 30 nm in diameter (Fig. 3).

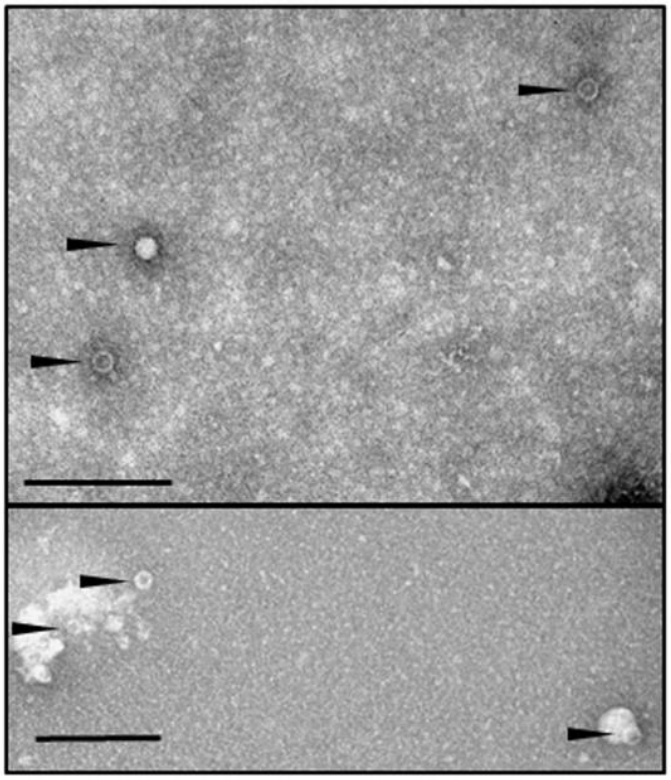

Fig. 3. Electron microscopic images of virus-like particles extracted from grapevine GV30 and negatively stained with ammonium molybdate. Black arrows highlight location of particles. Size bar 200 nm.

In order to better characterize its protein composition the virion preparation was separated by SDS-PAGE: 30µl virion preparation was mixed with 30µl 2x Laemmli loading buffer (65.8 mM Tris-HCl, pH 6.8 sample buffer, 26.3% (w/v) glycerol, 2.1% SDS, 0.01% bromophenol blue), boiled for 5 min, centrifuged and loaded onto a pre-cast Any kD™ Mini-PROTEAN® TGX™ Protein Gel (Bio-Rad, Hercules, CA). Separated proteins were located by staining using a mass-spectrometry compatible silver method (Shevchenko *et al*. 1996) and cut from the gel and submitted for Tandem Mass-Spectometry (MS-MS) analysis at the Cornell Biotechnology Resource Center – Proteomics and Mass Spectrometry Facility (http://www.biotech.cornell.edu/brc/proteomics-mass-spectrometry-facility). A prominent protein fragment of approximately 25kDa was observed in an SDS-PAGE gel (File S2), and MS-MS analyses identified only one significant hit with four peptide matches to a partial sequence of the

replicase and coat protein of Grapevine asteroid mosaic associated virus (acc. CAC10493 (GI:29335719)) (File S3).

Libraries for RNA-Seq were prepared using total RNA extracted according to Gambino *et al.* (2008) from plants GV30 and GV32. The SENSETM mRNA-Seq Kit V2 (Lexogen, Vienna, Austria) was used to prepare the libraries. Two technical repetitions were performed for every sample, for a total of 4 libraries. Libraries were individually barcoded and sequenced using an Illumina Hiseq 4000 (Genomics Resources Core Facility, Weill Cornell, NY (http://corefacilities.weill.cornell.edu/genomics.html) to obtain single-end 101-nt sequence reads.

A sRNA library was also prepared for sample GV30. Based on the observation of a loss of smaller size RNAs during the LiCl precipitation step (data not shown), a modified procedure for RNA isolation was used to prepare the input RNA without the LiCl step. This was followed by digestion with RQ1 RNase-free DNase (Promega, Madison, WI) according to the manufacturer's instructions. The preparation was made following the method proposed by (Chen *et al*. 2012) with the following modifications: a) during the 3' end ligation step the Universal miRNA cloning linker (NEB, Ipswich, MA) was used as an adapter, b) the reverse transcription (RT) primer sequence was extended by 17 nt corresponding to the reverse complement of the 3' adapter, and c) separation of final PCR products was done in a 6% non-denaturing polyacrylamide gel and products at ~165 bp (rather than 150 bp - due to the extensions made in the adapter/primer sequences) were excised and recovered by ethanol precipitation. Sequencing was performed as described above to obtain single-end 50-nt reads.

For analyses of the RNA-Seq data (101-nt reads), the Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was used to trim the first nine nucleotides and convert sequences to their reverse complement. To improve the de novo assembly of virus derived sequences, host (Vitis_vinifera.IGGP_12x.29 at http://plants.ensembl.org/Vitis_vinifera/Info/Index) sequences were removed by mapping using Bowtie2 (Langmead and Salzberg 2012). Unmapped reads were then submitted for de novo assembly using Trinity (Grabherr *et al*. 2011) with default settings.

For analyses of the sRNA-Seq data (50 nt reads), adapter containing reads were filtered and trimmed using cutadapt (Martin 2011). Resulting reads were collapsed to a non-redundant set using the Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). De novo assembly, mapping and virus detection was performed using VirusDetect software (Zheng *et al*. 2017).

Reads from all four RNA-Seq libraries were pooled into a single dataset. Resulting contigs were curated manually to obtain a candidate complete genome for GAMaV of 6720 nt. Finally, the RT-PCR and 5'-3' RACE curated genome sequence was use as reference for mapping with RNA-Seq and sRNA-Seq-derived sequences using Bowtie2 (Langmead and Salzberg 2012) and Bowtie (Langmead *et al*. 2009), respectively. The resulting RNA-Seq SAM alignment was processed using Samtools (Li *et al*. 2009) to produce a final consensus sequence for this GAMaV isolate.

RNA-Seq produced ca. 160 million reads total, and a very small fraction (<1%) of the sequenced reads was GAMaV-derived. This fraction was even smaller for sRNAs, with only 690 (<0.01%) reads from the approximately 9 million sRNA-Seq reads. In both cases, host derived sequences accounted for the majority of reads (File S5). This was also reflected in the fraction of the GAMaV genome that had sequence data representation, 99% for RNA-Seq vs 69% for the sRNA-Seq. A consensus sequence was obtained from the alignment of unfiltered mapped reads to the RT-PCR confirmed and 5'-3' RACE finished genome sequence (average coverage=1286X, median coverage=718X) (Fig. 4) to produce a final genome sequence for GAMaV (Acc. KX354202). For the RNA-Seq data there was an increase in coverage 5' to 3' with a prominent jump in reads that coincides with the predicted position of the CP AUG start site and subgenomic RNA. The sRNA-Seq coverage was much lower than in the RNA-Seq but the distribution was more even, with prominent peaks in the MT and Hel domains.

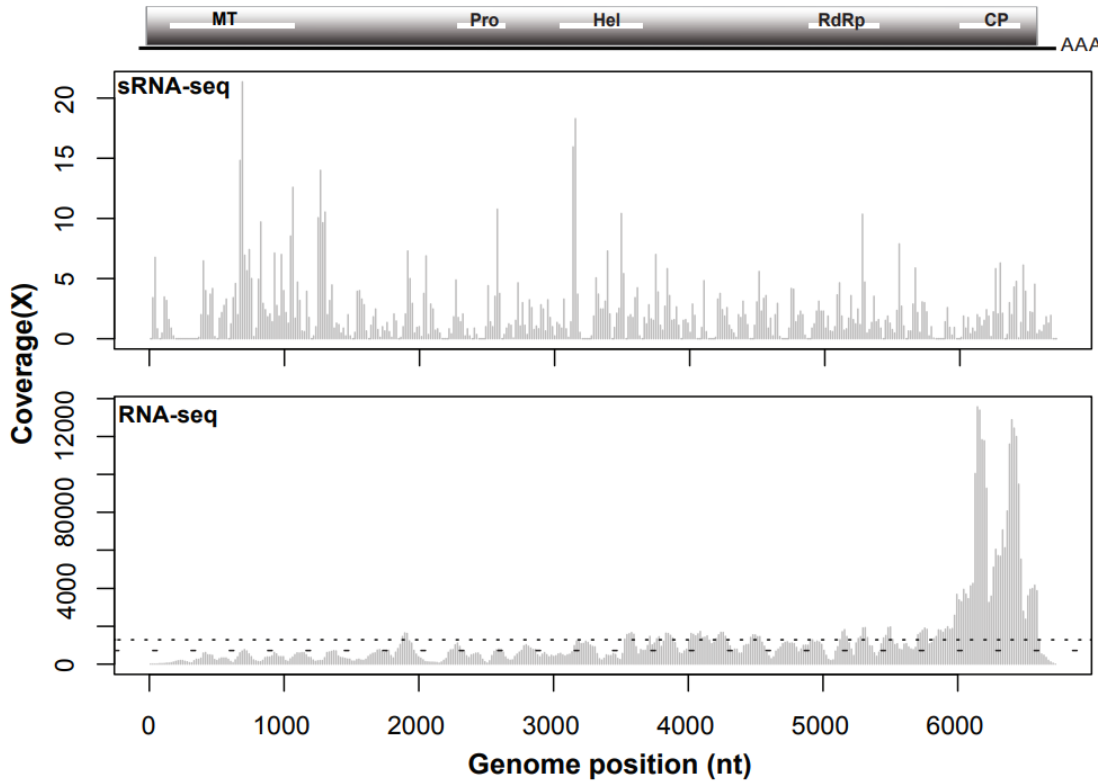

Fig. 4. Genome coverage of GAMaV-derived reads produced by a) sRNA-Seq and b) RNA-Seq. Mean and median coverage are indicated by dotted and dashed lines, respectively. Gray vertical bars indicate frequency of coverage at each genomic position. Simplified GAMaV genome organization with conserved domains is shown above histograms. MT – methyltransferase, Pro – protease, Hel – helicase, RpRp – RNA dependent RNA polymerase, CP – coat protein, AAA – polyadenylated 3' untranslated region

For Sanger sequencing of PCR products total RNA was extracted according to Gambino *et al.* (2008) from Cabernet franc plants GV30 and GV32 infected with Grapevine asteroid mosaic associated virus. For non-RACE reactions, RT was done with M-MLV Reverse Transcriptase (Thermo Fisher Scientific, Waltham, MA) according to the manufacturer's instructions using a mix of random and Oligo(dT)18 primers (Thermo Fisher Scientific). PCR using the derived cDNA as template was done using combinations of the primers listed in File S1 with either ThermoPol® (NEB) or AccuPrime™ Taq DNA Polymerase High Fidelity (Thermo Fisher Scientific). At least double coverage was obtained for all sequences, with ambiguous nucleotides requiring extra coverage and consensus when necessary. The 5' and 3' RACE were done as previously described (Thompson *et al*. 2002). SuperScriptTM II (Thermo Fisher Scientific) was used for RT either with

random primers (5' RACE) or an oligo(dT)+anchor (3'RACE) primer. The 3' termini of cDNA in 5'RACE was A-tailed using Terminal transferase (New England BioLabs, Ipswich, MA). Sanger sequencing was carried out at the Cornell Genomics Facility (http://www.biotech.cornell.edu/brc/genomics-facility) and all sequences assembled using the Vector NTI® suite (Thermo Fisher Scientific).

The complete genome of GAMaV is 6719 nt long excluding a poly-A tail. The highest identities for the complete nucleotide sequence were 94% (27% and 4% coverage) for partial sequences of GAMaV (1852 nt of Acc. AJ249357 and 575 nt of Acc. AJ249358, respectively), 72% (78% coverage) with Citrus sudden death associated virus (Acc. AY884005.1), 71% (76% coverage) with Nectarine virus M (Acc. KT273411), and 74% (75% coverage) with Oat blue dwarf virus (Acc. GU396990.1). The virus encodes a putative polyprotein of 2158 amino acids which contains recognized conserved domains of a methyltransferase (MT) (aa 125-406), a protease (Pro) (aa 861-967), a helicase (Hel) (aa 1050-1281), an RNA dependent RNA polymerase (RdRp) (aa 1594-1830), and a coat protein (aa 1998-2149) (Marchler-Bauer *et al.*, 2015).

The complete polyprotein sequence of GAMaV was aligned with all recognized members of the Marafivirus genus and other related species and sequences for phylogenetics using T-Coffee (Tommaso *et al*. 2011). Substitution models and tree construction were done with TOPALi v2.5 (Milne *et al*. 2008). A maximum likelihood inferred tree (Fig. 5) was generated with PhyML (Guindon *et al*. 2010). Branch significance was further tested and confirmed by a Bayesian inferred method using the algorithm MrBayes (Ronquist *et al*. 2009). The input sequence for the tree was the polyprotein of each virus. In most cases, this included all recognized conserved domains from the Hel to the CP. In viruses where the CP is a separate ORF 3' of the polyprotein (Olive latent virus 3, Turnip yellow mosaic virus and Grapevine fleck virus), the CP protein was appended to the polyprotein prior to analysis. The resulting tree (Fig. 5) shows GAMaV forms a discrete branch between CSDV and OBDV. All recognized marafiviruses are monophyletic with respect to the other two recognized genera of the family Tymoviridae, Maculavirus and Tymovirus.

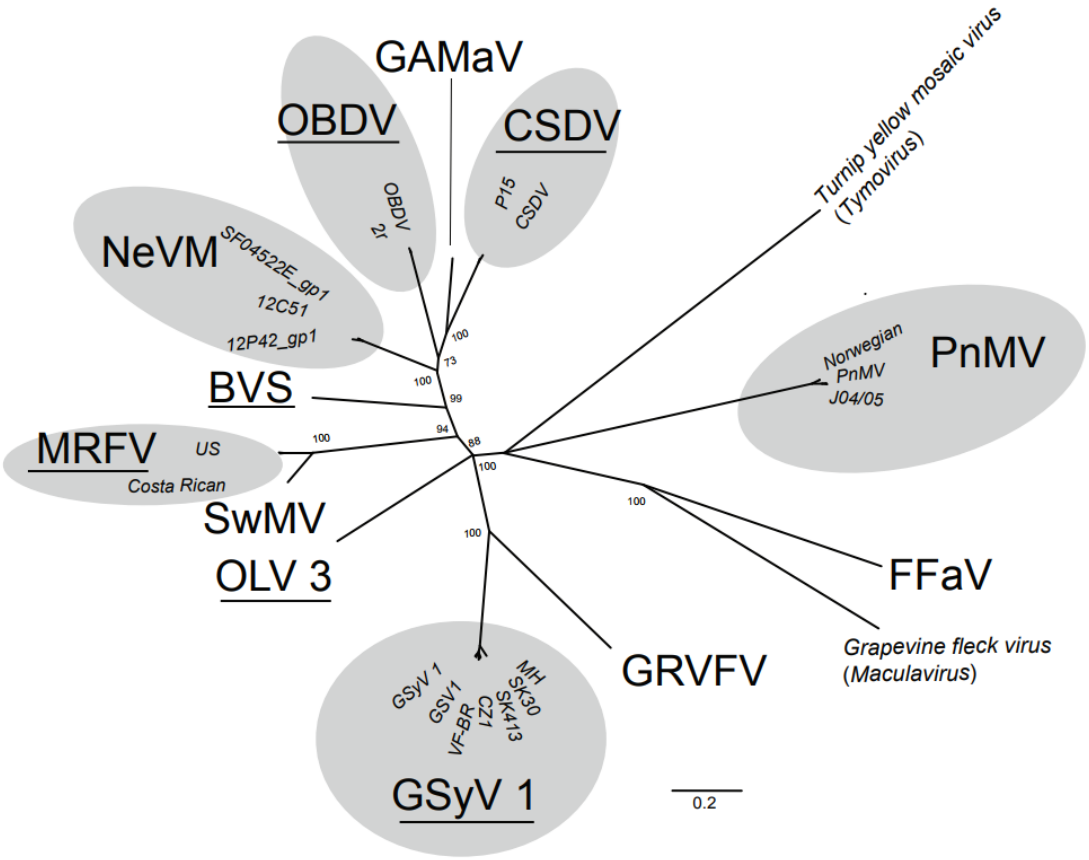

Fig. 5. Phylogenetic relationships of the marafiviruses and other members of the family *Tymoviridae*. A maximum likelihood inferred tree of the complete polyproteins (plus CPs when coded on a separate ORF) of marafiviruses derived from the WAG (Whelan and Goldman, 2001) (+G) model. Branches with bootstrap values below 70% were collapsed. Bootstrap values within a species clade are not shown. All branches depicted were also supported by posterior probabilities of >0.9 using a Bayesian inferred method (Ronquist *et al.*, 2009). Grey circles indicate virus species with more than one full-length sequence for which the strain names are shown italicized. Viruses [nucleotide accession number] – GAMaV (Grapevine asteroid mosaic associated virus) [KX354202], CSDaV (*Citrus sudden death-associated virus*) [DQ185573, NC_006950], OBDV (*Oat blue dwarf virus*) [NC_001793, GU396990], GSyV1 (*Grapevine Syrah virus 1*) [KP221257, KP221256, KP221255, JX513896, KT037017, KR153306, NC_012484], MRFV (*Maize rayado fino virus*) [NC_002786, KM523134], BVS (*Blackberry virus S*) [FJ915122], OLV3 (*Olive latent virus 3*) [NC_013920], NeVM (Nectarine virus M) [KT273411, KT273413, KT273412], SwMV (Switchgrass mosaic virus) [NC_015522], GRVFV (Grapevine rupestris vein feathering virus) [AY706994], PnMV (*Poinsettia mosaic virus*) [NC_002164, AB550792, AB550791, AM412237], FFaV (Fig fleck associated virus) [FM200426], *Grapevine fleck virus* [AJ309022] (genus *Maculavirus*) and *Turnip yellow mosaic virus* [X07441]. Those acronyms underlined are presently species recognized by the ICTV.

An alignment including all the Marafivirus sequences available at NCBI (N=145) was produced using Muscle v3.8.31 (Edgar, 2004) (File S4). The regions around the proposed protein cleavage sites between the Hel/RdRp and RdRp/CP domains were identified. Sequences represented in those regions were extracted, translated and realigned. Based on the original sites reported by Maccheroni *et al.* (2005) putative cleavage sites between the Hel and RdRp domains and between the RdRp and CP domains were identified at GAMaV amino acid positions 1337 and 1942 of the polyprotein ORF, respectively (Fig. 6). A sequence logo of the sequence motif around the cleavage sites was produced using Weblogo 2.8.2 (Crooks *et al.*, 2004) from the resulting alignment. Analysis of the region around these cleavage sites using all the marafivirus available sequences at NCBI revealed the previously described (Edwards and Weiland 2014; Edwards *et al*. 2015) conserved motif (LxG(G/A)) present at positions -4 to -1 in both cleavage sites, concordant with the cleavage site context observed by (Bransom *et al*. 1996). A search for additional regions containing this motif was performed to determine whether other putative cleavage sites between the remaining conserved domains in the polyprotein could be identified. A sequence with high similarity (LSGA) was found at amino acids 982-985 of GAMaV's polyprotein between the Pro/Hel domains, however there was little apparent conservation with the aligned sequences of other marafivirus sequences. Under more relaxed conditions, a LTTA sequence was found at amino acids 457-460, between the Met/Pro domains. A degree of conservation for this motif across other members of the genus suggest this could be a functional cleavage site (Fig. 6). These results are consistent with and further refine previous studies (Alabdullah *et al*. 2010; Maccheroni *et al*. 2005).

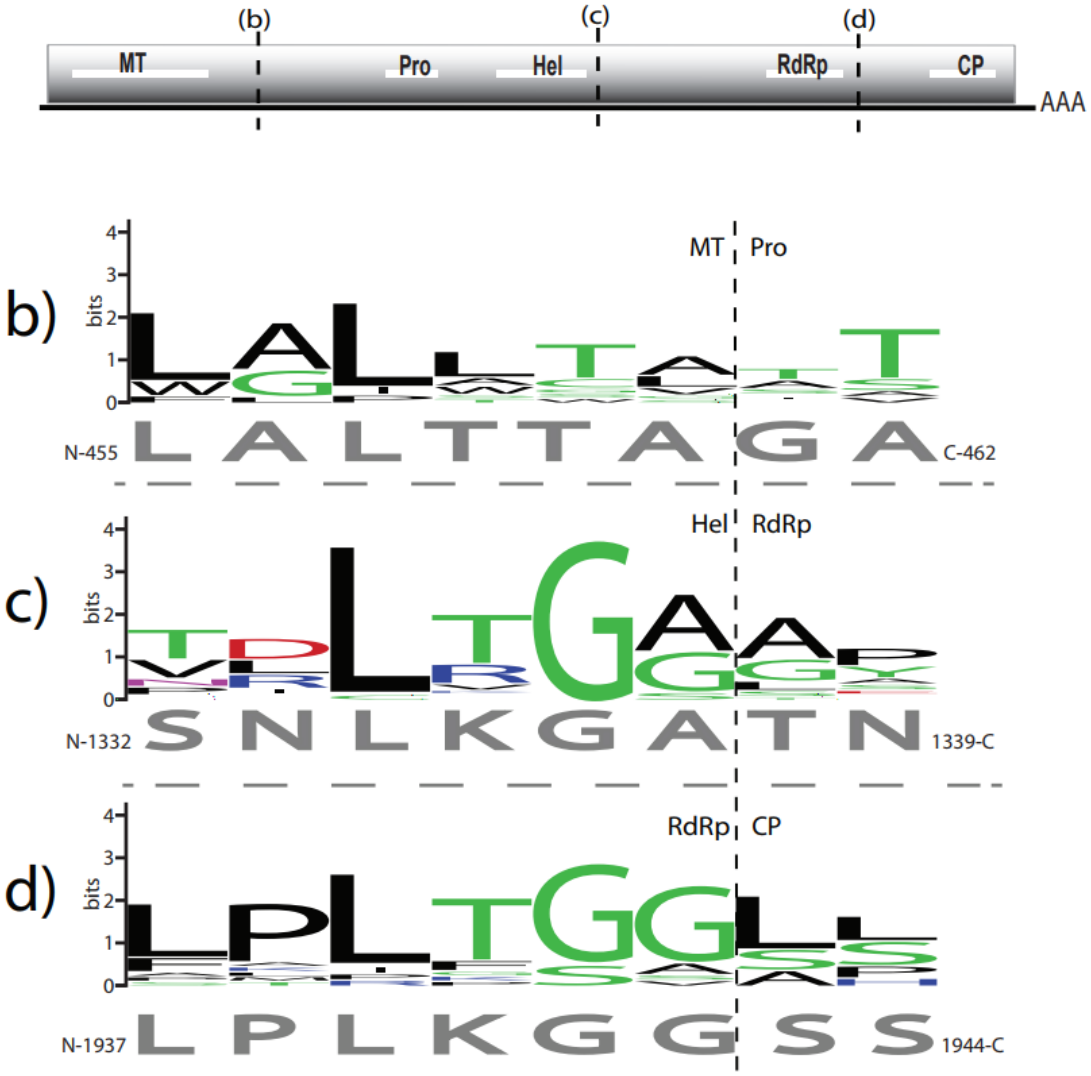

Fig. 6. Marafivirus putative cleavage site identification. (a) Schematic of GAMaV genome organization with conserved domains. MT – methyltransferase, Pro – protease, Hel – helicase, RpRp – RNA dependent RNA polymerase, CP – coat protein, AAA – poly(A) tail. Vertical lines marked (b)–(d) show the relative location of the three putative cleavage sites depicted below, based on the alignment of (b)–(d) are sequence logo depictions of the putative amino acid cleavage sites between the MT and Pro (b), the Hel and RdRp (c), and the RdRp and CP (d), respectively. In gray uniform font below each is the GAMaV corresponding amino acid sequence and their positions (numbers) in the polyprotein. The vertical dotted line marks the predicted cleavage site between each domain. Amino acids are colored according to their chemical properties: polar amino acids (G,S,T,Y,C,Q,N) are green, basic (K,R,H) blue, acidic (D,E) red and hydrophobic (A,V,L,I,P,W,F,M) amino acids are black..

Consistent with previous findings (Abou Ghanem-Sabanadzovic *et al*. 2012) a highly conserved marafibox-like sequence CAGGGUGAAUUGCUCCA (Izadpanah *et al*. 2002) can be found at nts 5875-5891, making nt 5900 and nts 6010-6012 the putative start site for subgenomic (sg) RNA synthesis and the initial codon (aa 1961 of polyprotein) of the major CP, respectively. The derived sgRNA major CP is predicted to be 21.1 kDa and 198 aa long. This compares with the predicted cleavage-derived minor CP, starting at aa 1943 of the polyprotein with a size of 22.9 kDa and 216 aa. BLAST analysis of the predicted subgenomic derived CP (or minor CP) ORF shows the highest identity of 99% with the USA9 isolate of GAMaV (Acc. CAC10493) followed by Citrus sudden death associated virus (Acc. YP_224294.1) with 70%, Nectarine virus M (Acc. ALX72769) with 70% and Oat blue dwarf virus (Acc. ADD13603) with 68% identity.

We also identified an overlapping protein (OP) ORF at positions nts 630-1646 encoding a protein of 38.5kDa and 339 aa which lacks a canonical translational start. BLAST analysis of this predicted ORF shows an identity of 36% to the movement proteins to Nemesia ring necrosis virus (Acc. YP_002308441), 35% to Chiltepin yellow mosaic virus (Acc. YP_003620400), 32% to Turnip yellow mosaic virus (Acc. AMH40124), 31% to Tomato blistering mosaic virus (Acc. YP_008318041) and other tymoviruses. Significant similarity with other reported marafiviral OPs was not found. The putative GAMaV OP has a high proline content (22.5%) typical for described OPs of members from the Tymoviridae. Interestingly, the putative GAMaV OP is within the same reading frame, -1 relative to the polyprotein, as the movement protein of tymoviruses and the unassigned Fig fleck associated virus (Elbeaino *et al*. 2011). This contrasts with the putative OPs of other marafiviruses which are in a +1 reading frame relative to the polyprotein.

Accession number

Full-length sequence: KX354202

Library submissions: see Supplementary Table S1

Acknowledgements

Thanks goes to John Grazul for help with the TEM, Heather McLane for her technical support and Monica Carvalho for assistance in figure production and editing.

Supplementary files (available in http://www.sciencedirect.com/science/article/pii/S0168170216305159?via%3Dihub#sec0015)

S1 – Silver-stained SDS-PAGE (Any kDTM, Bio-Rad) separation of GV30 virion extraction.

S2 – Results of MS-MS analysis of 25kDa fragment isolated from GV30 grapevine

S3 – FASTA file of database marafivirus sequences aligned for cleavage motif search

S4 - list of primers used for Sanger sequencing of GAMaV

S5 – High throughput sequencing read summary

CHAPTER SIX

GLOBAL CHARACTERIZATION OF SMALL RNA-MEDIATED REGULATORY NETWORKS IN ARABIDOPSIS THALIANA

Abstract

Gene regulation involves the orchestrated action of multiple regulators to fine-tune the expression of genes. Small RNA (sRNAs) have been shown to be important regulators of gene expression. Network analyses have been proposed to be an informative means to understand the complexity of gene expression regulatory mechanisms. An sRNA-mediated regulatory network in *Arabidopsis thaliana* was identified through: i) the use of large scale sequences generated using next-generation sequencing techniques, ii) recent advances in an understanding of sRNA biogenesis in plants, and iii) a custom designed, data-driven and degradome-supported bioinformatics analysis pipeline. The resulting network consists of an extended set of sRNAs and their target transcripts, including ~41% of genes in the *A. thaliana* genome, and representing most all functional categories. Approximately 30% of genes with Gene Ontology (GO) annotations corresponding to regulation of gene expression were found to be under sRNA control. Structural analysis of the resulting sRNA-mediated network showed similarities to networks of other biological systems, and demonstrated connectivity between individual phasiRNA regulatory cascades, as well as extensive co-regulation of transcripts by miRNAs and phasiRNAs. These results confirm and expand upon the role of sRNAs as key regulators of gene expression. The described regulatory network provides a reference that will facilitate global analyses of individual plant regulatory programs to control homeostasis, development and responses to biotic and abiotic environmental changes.

Introduction

Gene expression regulation is a cellular process that involves the orchestrated action of multiple regulators to fine-tune the expression of genes (Walczak and Tkačik 2011). It can be thought of as the sum of interactions between regulatory factors and their substrates across multiple levels, as well as the effect of cross regulation between regulators from different levels. Regulatory levels range from DNA availability (chromatin structure, methylation status) to RNA abundance and translational efficiency (Walczak and Tkačik 2011). This complex, multilevel process can be represented and studied using network theories (Cumbo *et al*. 2014). Network-based approaches allow the investigation of biological features that emerge when regulatory systems are studied from a multiscale, genomic approach (Stumpf and Wiuf 2009). Features such as the topology and dynamics of these networks have been proposed to be informative and provide insights into the way organisms function, develop and respond to internal and external stimuli (Cora *et al*. 2017). Network concepts and applications of network principles to biological systems have been thoroughly reviewed (Zhu *et al*. 2007; Albert 2005; Pavlopoulos *et al*. 2011).

Transcription factors and small RNAs (sRNAs) are considered to be the primary levels of gene expression regulation (Cora *et al*. 2017). They act in combination to form genetic regulatory circuits involved in transcriptional control (Cora *et al*. 2017; Megraw *et al*. 2016). A significant body of research has been dedicated to understand how transcription factors are involved in the regulation of multiple cellular processes in *Arabidopsis thaliana,* one of the best studied plant model systems (Drapek *et al*. 2017; Taylor-Teeples *et al*. 2015; González-Morales *et al*. 2016). In contrast, broad scale studies of sRNA-mediated regulation in *A. thaliana* are not common and their

scope is limited due to challenges posed by the high false positive rate of bioinformatics predictions of small RNA activities (Ding *et al*. 2012).

sRNAs have been shown to be involved in higher level regulatory interactions by controlling the expression of other regulators (transcription factors, sRNA biogenesis factors), thereby extending their regulatory contribution by affecting downstream events (Wang & Chekanova, 2016; Cora *et al.*, 2017). Because of this and their involvement in key cellular processes, several studies propose them to be 'master regulators' of gene expression and phenotype determination (Zhai *et al*. 2011; Sun *et al*. 2010; Voorhoeve 2010).

In plants, sRNAs are the products of multiple biogenesis pathways (Borges and Martienssen 2015). Because of their broad regulatory potential and better understood biogenesis and mode of actions (Wang and Chekanova 2016), most studies focus on microRNAs (miRNAs) and phased interfering small RNAs (phasiRNAs). miRNAs are the better studied example, as their biogenesis and activity have been studied intensively, and their involvement in multiple cellular processes through post transcriptional gene silencing (PTGS) has been well established (Borges and Martienssen 2015). phasiRNAs correspond to a group of sRNAs produced from multiple pathways whose regulation involves regulatory cascades usually derived from a microRNA-transcript targeting event, leading to the production of additional phased small RNAs with the potential to regulate gene expression in cis and trans. sRNAs that induce phasiRNA production are referred to as 'triggers', and cleavage at their target sites determines the phased register of the resulting phasiRNAs (Fei *et al*. 2013).

The coding capacity of phasiRNA loci (*PHAS*) has been underestimated, as demonstrated by Rajeswaran *et al.* (2012). Their results indicated that *PHAS* loci can have multiple triggers and their processing by dicer-like proteins (DCLs) can result in the production of a combination of 21

and 22 nt long phasiRNAs. Together, these features lead to new or shifted phased registers whose inclusion can greatly expand the repertoire of phasiRNAs produced from a given *PHAS* locus, with a commensurate increase in potential new targets (Fei *et al*. 2013). PhasiRNAs produced from alternative phased registers as well as 22 nt long products are referred to herein as non-canonical phasiRNAs. PhasiRNAs in the 21-22 nt size range act predominantly through PTGS; their role in gene regulation is currently an active field of research, and they function in development, defense, abiotic stress, and other biological processes (Wang and Chekanova 2016).

Despite their significant regulatory potential, the understanding of miRNA- and phasiRNA-based gene regulation is limited. Genomic features or sequence signatures are not available to predict or detect *PHAS* loci (regions of phasiRNA production), and current detection methods rely on sRNA expression data and the search for phased patterns in the distribution of sequenced sRNAs mapped to the genome or transcriptome (Guo *et al*. 2015). Expression of phasiRNAs has been shown to be inducible and dependent on specific stimuli (Komiya 2017), therefore a proper characterization requires an evaluation of data from multiple tissues, under different conditions and at different developmental stages.

In *A. thaliana*, multiple miRNA/phasiRNA cascades have been described (reviewed in Wang and Chekanova 2016; Fei, Xia, and Meyers 2013). They are associated with cellular processes such as metabolic stress (Hu *et al*. 2011). However, there has been only one attempt to decipher genome-wide sRNA-mediated regulation (MacLean *et al*. 2010). These authors showed the potential for large scale sRNA-mediated regulatory networks, though their results were derived mostly from *in silico* predictions wherein the biological relevance had not been experimentally assessed. An important tool for validating predicted sRNA-transcript target interactions has been the development of degradome analyses (German *et al*. 2009). Degradome libraries capture the

RNA products generated by sRNA targeting and cleavage of transcripts. By facilitating the experimental validation of the interactions in a high throughput manner, evaluations of sRNA activity and regulation through degradome analysis can be performed at a genome-wide level, resulting in networks of biological relevance.

The objective of this study was to identify a comprehensive sRNA-mediated regulatory network at the genome-wide level in *A. thaliana* using a data-driven and degradome supported bioinformatics analysis pipeline. This meta-network will provide a reference frame for assessing sRNA-mediated regulation during growth, pathogenesis and under different environmental conditions, and ultimately will reveal the role of sRNAs in the global genomic circuitry for the regulation of gene expression.

Methods

Experimental design

Data were obtained by two methods: 1) all publicly available (NCBI) sRNA libraries from *A. thaliana* were compiled to provide a diverse representation of sRNA expression and regulation under varied conditions; these were derived from multiple tissues, developmental stages, and biotic and abiotic stress conditions, and included all degradome datasets for *A. thaliana,* and 2) paired sets of sRNA-Seq and degradome data from aliquots of single, total RNA extracts; corresponding sRNA-Seq and degradome data sets were produced as part of this study for 14 independent plant samples. All the sRNA and degradome data from (1) and (2) were combined to identify an sRNA-mediated regulatory meta-network (described below).

RNA extraction and library preparation

Two-week old *A. thaliana* Col. plants grown at 22 C with a 10 h photoperiod were mechanically inoculated with *Cucumber mosaic virus*. Leaf tissue was collected 10 days post-inoculation, ground in liquid nitrogen, and total RNA extracted using Trizol (Thermo-Fisher) as recommended by the manufacturer. Each resulting total RNA preparation was divided into two aliquots to be used as input for sRNA-Seq and degradome libraries. sRNA libraries were prepared from 1 ug of total RNA using methods described previously (Vargas-Asencio *et al*. 2017). For the degradome libraries, ~40 ug of total RNA was used. Degradome libraries were constructed using the method described by Zhai *et al. (*2014), but with the following modifications: a) different adapters and primer sequences were used (Table 3), b) the PCR clean up step was performed using Axygen™ AxyPrep Mag™ PCR Clean-up (Fisher) instead of Agencourt AMPure XP beads (Beckman Coulter), c) EcoP151 (NEB) was used for the restriction enzyme digestion step instead of MmeI. Sequencing was performed using an Illumina Hiseq 4000 at the Genomics Resources Core Facility, Weill Cornell, NY (http://corefacilities.weill.cornell.edu/genomics.html) to obtain single-end 51-nt reads for both sRNA (Biosample accessions: SAMN07947991-SAMN07948004) and degradome libraries (Biosample accessions: SAMN07949266-SAMN07949281).

Table 3. Adapter and primer deoxyribonucleotide sequences for degradome libraries

oligo	sequence
5' PARE adapter	GTTCAGAGTTCTACagtccgacgatccagcag
RT-primer	TGATCTAGAGGTACCGGATCCCAGCAGTTTTTTTTTTTTTTTTTT
5' cDNA PCR primer	CACGACAGGTTCAGAGTTCTACA
3' cDNA PCR primer	CTGATCTAGAGGTACCGGATCC
dsDNA_top	NNAGAGAATGAGGAACCCGGGGCAG
dsDNA_down	TCTCTTACTCCTTGGGCCCCGTC
Final PCR primer	AATGATACGGCGACCACCGAGATCTACACGACAGGTTCAGAGTTCT ACAGTCCGACGAT*C
Index 3' primer	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATC*T

* indicates phosphothio modification, lowercase indicates ribonucleotides

Bioinformatics tool for identification of sRNA-mediated networks

A custom bioinformatics pipeline was implemented to identify sRNA-mediated networks. A detailed description is provided in the following sections. The overall strategy was to gather all available sRNA and degradome data, and to combine it with existing genome annotations and sRNA databases to produce a data-driven, degradome-supported network of interactions between sRNAs and transcripts. There are two types of nodes in the proposed network: sRNAs and transcripts. sRNAs include miRNA and phasiRNAs, and transcripts include miRNA precursors, *PHAS* loci and mRNA transcripts targeted by sRNAs. Annotations are available for miRNAs, miRNA precursors and potential target transcripts, while for *PHAS* loci, their sRNA triggers and the resulting phasiRNAs, there is no genome-wide annotation available. The identification of these components and their interactions was therefore part of the tasks included in the pipeline. Newly generated annotations were combined with available genome and known miRNA annotations to

perform a genome-wide level search for sRNA-target interactions. Once all components and their interactions were identified and experimentally validated they were consolidated into a network for downstream analysis.

Reference files and datasets

The TAIR10 version for *A. thaliana* provided the reference genome (Swarbreck *et al*. 2008). Genome annotations were obtained from Araport11 (Cheng *et al*. 2017). Known miRNA and precursor sequences were obtained from miRBase (Kozomara and Griffiths-Jones 2014) release 21.

Fourteen sRNA and sixteen degradome libraries were produced in this study. These data were complemented with all publicly available sRNA datasets representing different tissues, stress conditions (biotic and abiotic) and developmental stages in *A. thaliana* (Supplementary table 1), as well as all available degradome datasets (Supplementary table 2)

Data processing

Reads (51 nt) from sRNA-Seq libraries were filtered using the adaptive adapter trimming function in Trim Galore (https://github.com/FelixKrueger/TrimGalore) to account for variability in library construction methodologies. Size was constrained to 20-40 nt after adapter trimming, and non-adapter containing reads were removed. Datasets containing unique reads only were produced using the Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit). Libraries containing less than 100 unique sequences were considered non-informative and removed. SRA degradome libraries (Supplementary table 2) were filtered using the adaptive adapter trimming function in Trim Galore with the minimum size after adapter trimming set to 18 nt. The resulting libraries

were evaluated manually and additional trimming was performed if there was evidence of remaining adapter sequences. For the libraries produced in this study the first six nt derived from the library preparation process were removed. The Fastx toolkit was used to convert reads to Fasta format.

miRNA-*PHAS* loci / phasiRNA annotation and trigger identification

PHAS loci detection was performed for each dataset using PhaseTank (Guo *et al*. 2015). Locus extension was set to zero, and the top 15% of regions with the highest accumulation of mapped reads (described as relative small RNA production regions in (Guo *et al*. 2015) were analyzed for phasiRNA production. Results for all datasets were combined to produce PHAS loci with maximum length from overlapped results. Potential *PHAS* loci detected in less than three of the X libraries were discarded. The resulting loci were then extended by 220 nt on each side to perform a search for sRNA triggers associated with phasiRNA production.

PhasiRNA production triggers were searched using the degradome data. Thirty nine degradome libraries were independently analyzed using Cleaveland4 (Addo-Quaye *et al*. 2009). Sequences from both strands of the extended *PHAS* loci were evaluated using known miRNAs as queries. A weighted scoring system to compile the independent degradome analysis results was developed as follows: cleavage events with degradome category zero per Cleaveland4 were given a score of 5, cleavage events with degradome category one were given a score of 4, cleavage events with degradome category two were given a score of 0.5. The scores for each event were added across all thirty-nine degradome libraries. The highest scoring event per *PHAS* locus was selected as the initial phasiRNA triggering site, a minimum score of 10 was set to assigned triggers. When triggers were found, the polarity of the loci was determined based on the degradome reads.

To identify the phasiRNAs produced by each *PHAS* locus sRNA reads from each library were mapped to the extended *PHAS* loci independently. No mismatches were allowed, sRNAs of 21 and 22 nt were accepted, up to ten mapping locations were reported per read, multiple mapping counts were divided between the number of locations, reads with more than 10 mapping locations were removed, and reads mapping outside the original region (before extension) were not considered. Mapped reads were assigned to bins from 1 to 21 (phases) according to their mapping positions from the 5' end. Positions of reverse reads were shifted (+2) due to 3' overhang, to match forward read bin positions. The mapping was performed on each strand of the *PHAS* loci independently. A scoring system was developed to rank bins by read abundance for each locus across all sRNA libraries. The three most abundant bins per locus, per library were used. The most abundant bin was given a score of 5, the second most abundant was given a score of 2, the third most abundant was given a score of 0.5. The resulting scores from all libraries were added for each bin to produce a ranking of sRNA bins for each *PHAS* locus.

PhasiRNAs derived from miRNA triggering events were found by matching the phase register set by the degradome-confirmed miRNA triggering events to bin assignments. PhasiRNAs from immediately adjacent bins (-1, +1) were also collected. For the *PHAS* loci where no trigger was found, sRNAs from the most abundant bin and immediately adjacent bins were collected.

Resulting phasiRNAs were pooled with all known miRNAs to produce a new set of queries to search for *PHAS* production triggers using the degradome-based ranking strategy described above. To identify secondary and/or tertiary triggers, sRNAs whose cleavage events matched the polarity of the primary (highest ranked, with score >10) trigger were kept. The potential secondary/tertiary triggers were evaluated by matching their slicing site coordinates to those corresponding to the three most abundant sRNA bins per *PHAS* locus. Because 22-nt sRNAs were

included in the analysis which can alter the 21 nt phasing, the bins immediately adjacent (-1,+1) were also considered. In the cases where a match was found, the sRNAs were considered additional phasiRNA triggers. The assignment of secondary/tertiary triggers was further evaluated by determining if the phasiRNAs contained in the matched bins were biologically active (described below). PhasiRNAs derived from secondary and/or tertiary sRNA triggering events were found by matching the phase register set by the degradome-derived sRNA triggering events to bin assignments. The resulting phasiRNAs were pooled with known miRNAs to produce a final set of queries to search for phasiRNA production triggers using the strategy described above in this paragraph.

Corresponding trigger, *PHAS* locus and phasiRNA sets were evaluated and confirmed manually to produce a miRNA-*PHAS* loci-phasiRNA annotation. A novel nomenclature is proposed for phasiRNAs in order to provide consistent and detailed information about their biogenesis. To assign a *PHAS* loci to a gene ID the *PHAS* loci with polarity assigned based on confirmed sRNA triggers were compared to the araport11 genome annotation, and if the locus had significant overlap (>80%) and matching polarity to annotated features (genes, transposons), the locus was assigned to the feature. If more than one feature matched a locus, the one with the highest overlap was selected. If no trigger was found, the forward genomic orientation was kept and the loci were named using their coordinates. For phasiRNAs, they were named using the *PHAS* locus from which they derived, followed by up to four descriptors: 1) the number of registers (21 nt) from the 5' end of the transcript; 2) in parenthesis, offset to main phased register, if any; 3) polarity, a "+" was used if the phasiRNA derived from the mRNA strand or "-" if derived from the complementary sense; and 4) size was indicated in the case of 22 nt long phasiRNAs by adding "_22" to the end.

Target transcript search

The miRNA-*PHAS* loci / phasiRNA annotation was used to identify and quantify miRNAs and phasiRNAs; an arbitrary threshold of 50 raw count was established to select candidates for transcript targets. Degradome datasets were analyzed independently using Cleaveland4 (Addo-Quaye *et al*. 2009) to find target transcripts for selected sRNAs. A two-step validation process was used to identify validated target transcripts for sRNAs: 1) only interactions with degradome categories 0, 1 and 2 were considered; 2) of the previous set, only interactions detected in at least 10 independent degradome libraries (~25% of total libraries) were considered.

Network analysis and visualization

For the sRNA-mediated regulatory meta-network, a bipartite directed meta-network was constructed by obtaining sRNAs and transcripts and their interactions from a combination of existing miRNA and transcript annotations, the newly developed miRNA-*PHAS* loci / phasiRNA annotation and the degradome search results from all datasets. Cytoscape (Shannon *et al*. 2003) was used for visualization and structural analysis. A functional characterization was performed using Bingo (Maere *et al*. 2005) to determine the representation of GO Slim categories of the genes included in the network as compared to the genome as a whole.

Results

PHAS loci detection

Currently there are no genomic features or sequence signatures that allow the identification of *PHAS* loci (regions of phasiRNA production), and their detection depends on a search for phased patterns in sRNA-Seq data. Because of the observed variability in size between sRNA libraries

and the assumption that phasiRNA expression may depend on specific environmental queues, the strategy used to identify *PHAS* loci was to evaluate each library independently and then to combine data to produce *PHAS* loci with maximum length from overlapping results with better definition of start and end positions. The number of *PHAS* loci detected was variable, ranging from zero to more than 120 per library (Fig. 7).

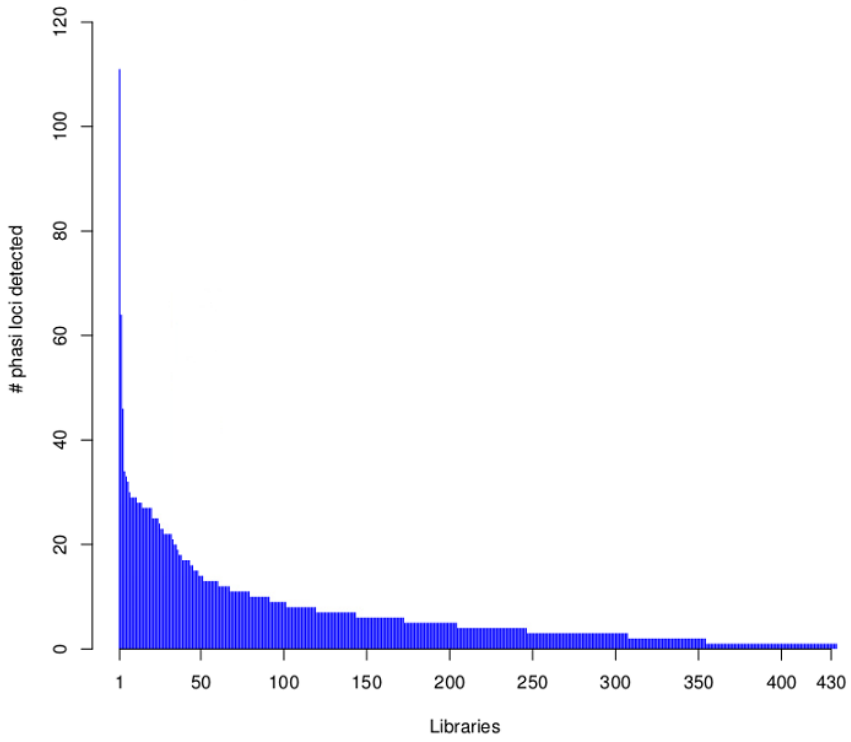

Fig. 7. Histogram showing the number of *PHAS* loci detected per sRNA library across all libraries. Libraries are enumerated in the x axis; 875 libraries were evaluated and only those in which *PHAS* loci were found (n=435) are shown. The y axis shows the number of *PHAS* loci detected per library.

A total of 953 *PHAS* loci were identified from the combined libraries (n=875, (Supplementary table 3). The consistency of *PHAS* loci detection was evaluated by determining the number of recognition events for each locus across all libraries. To remove spurious results, only *PHAS* loci detected in at least three libraries were included. The number of recognition events

varied, with 109 *PHAS* loci independently detected in at least 3 libraries (Fig. 8). A failure to detect any given locus in a specific library could be due to expression limited to specific experiment conditions, e.g. stress, developmental stage or tissue type or to a limitation in sensitivity.

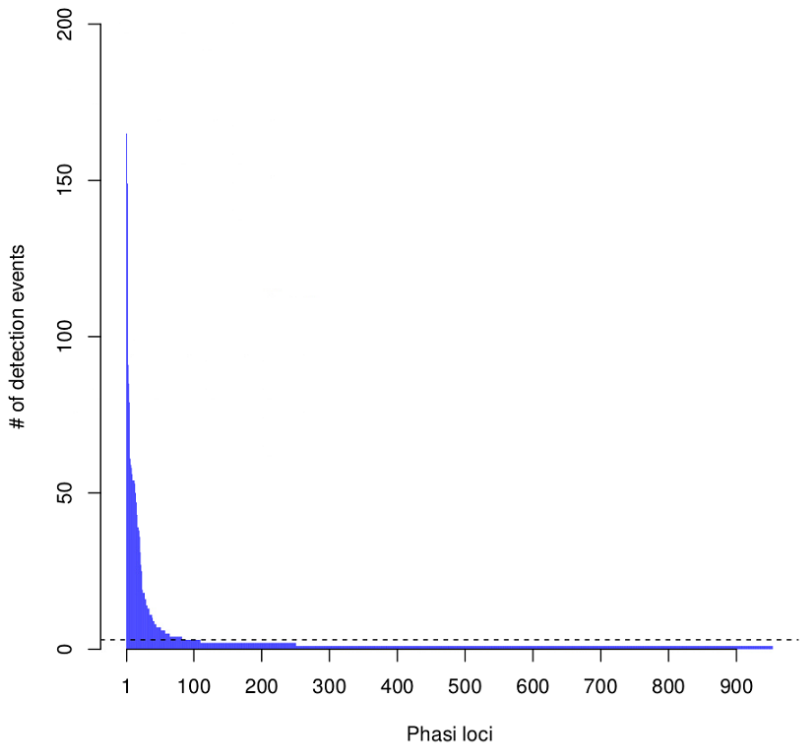

Fig. 8. Histogram summarizing of recognition events (detection) of bona fide *PHAS* loci across all libraries. *PHAS* loci are enumerated in the x axis. The y axis shows the number of libraries in which a given *PHAS* locus was detected. Dotted line indicates the three detection events threshold utilized.

PhasiRNA trigger search

Once *PHAS* loci had been identified, a recursive method was designed to identify triggers by extending a search up and downstream of the detected *PHAS* loci, followed by searches for secondary or tertiary triggers that would explain the production of non-canonical phasiRNAs (22 nt long or derived from an alternative phased register). For the 109 *PHAS* loci evaluated, triggers

were assigned for 58 of them. From the 126 sRNAs triggers identified, 18 corresponded to miRNAs and 108 were phasiRNAs; in some cases sRNAs were assigned to multiple *PHAS* loci (Table 4). Among the triggers were 11 canonical phasiRNAs; 43 were 22 nt long, 54 were from a secondary phased register, 31 were both 22 nt long and derived from a secondary phased register.

Table 4. Summary of sRNA triggers of phasiRNA production per phasi locus.

locus	locus polarity	trigger	Deg score	Rank*
AT1G29910	(-)	AT3G26810(+)_17-(3)	13	2-1
AT1G29910	(-)	AT3G22121(+)_14-(4)	10.5	2
AT1G29910	(-)	Chr5:7683815-7684395(+)_29-(-11)	16.5	2
AT1G29910	(-)	AT1G63080(-)_12+(-8)_22	12	3+1
AT1G29910	(-)	AT1G63080(-)_31+	17	1-1
AT1G29910	(-)	AT1G63080(-)_31+(-1)	17	1
AT1G29910	(-)	AT1G29910(-)_28-(-1)	10	1+1
AT1G29910	(-)	AT2G35945(-)_12+(5)	69	1+1
AT1G29910	(-)	Chr2:6277-9595(+)_122+(2)	12.5	1+1
Chr1:11454588-11454825	(+)	Chr1:11454588-11454825(+)_16-(7)_22	10.5	3
Chr1:17203735-17203844	(+)	Chr1:24721142-24721509(+)_13+	20.5	1-1
AT1G48410	(-)	ath-miR168a-5p	77.5	1
AT1G48410	(-)	ath-miR168b-5p	60	1
AT1G48410	(-)	AT1G18000(+)_29-	10.5	3
AT1G50055	(-)	ath-miR173-5p	180	1
AT1G62590	(-)	Chr1:23489342-23489630(+)_10+(4)_22	11	1+1
AT1G62590	(-)	AT1G62590(-)_23-_22	11.5	2
AT1G62590	(-)	AT1G62590(-)_33-(+1)	49.5	2
AT1G62590	(-)	AT1G62590(-)_33-_22	35.5	2
AT1G62590	(-)	AT2G27400(-)_22-(+1)	11.5	2+1
AT1G62590	(-)	AT1G62590(-)_36-(3)_22	10	3-1
AT1G62670	(-)	AT1G62670(-)_30-(-8)	18	2
AT1G62670	(-)	AT1G62670(-)_30-(-9)_22	15.5	2
AT1G62670	(-)	AT1G62670(-)_30-(-7)	85	2+1
AT1G62910	(+)	AT2G39681(-)_16-_22	89	1
AT1G62914	(+)	AT1G62590(-)_33-(-8)	20.5	2
AT1G62914	(+)	AT1G62590(-)_33-(-9)_22	46.5	2
AT1G62914	(+)	AT2G39681(-)_16-_22	16.5	2
AT1G62914	(+)	AT1G63080(-)_31-(-8)_22	19	1
AT1G62914	(+)	AT1G63400(+)_31-(-7)	19.5	1
AT1G62914	(+)	ath-miR161.2	10.5	3

AT1G62914	(+)	AT1G62930(+)_30-(+1)	16	3
AT1G62930	(+)	AT1G63080(-)_31-(-8)_22	32	3
AT1G62930	(+)	AT1G63400(+)_31-(-7)	16	3
AT1G62930	(+)	Chr1:23489342-23489630(+)_12+(3)_22	11.5	3+1
AT1G62930	(+)	AT1G62910(+)_31-_22	18	2-1
AT1G62930	(+)	AT1G63400(+)_31-_22	12	2-1
AT1G62930	(+)	AT2G39681(-)_21-(+1)_22	32.5	2-1
AT1G62930	(+)	AT2G39681(-)_21-(2)	41.5	2-1
AT1G62930	(+)	AT2G39681(-)_21-(2)_22	14.5	2
AT1G63080	(-)	AT1G63080(-)_31-(-8)_22	30	2+1
AT1G63080	(-)	AT2G39681(-)_21-(2)_22	40.5	3-1
AT1G63080	(-)	AT2G39681(-)_21-(3)	50	3-1
AT1G63130	(+)	AT1G63130(+)_31-(-8)	44.5	1
AT1G63130	(+)	AT2G39681(-)_16-(+1)	103	1
AT1G63150	(+)	AT1G62860(-)_21-(+1)_22	22	2
AT1G63150	(+)	AT1G12300(+)_11+_22	46	2
AT1G63150	(+)	AT1G62914(+)_12-(+1)	34	3
AT1G63230	(+)	AT2G39681(-)_19-(+1)_22	40.5	1
AT1G63230	(+)	AT2G39681(-)_19-(2)	31	1
AT1G63330	(+)	AT1G63330(+)_31-(+1)	19	2
AT1G63330	(+)	AT1G63330(+)_31-_22	47.5	2
AT1G63330	(+)	AT2G27400(-)_22-(+1)	11	2+1
AT1G63400	(+)	AT1G63080(-)_31-(-8)_22	50.5	3
AT1G64583	(-)	ath-miR161.1	129.5	3-1
AT1G64583	(-)	AT1G62860(-)_26-(2)	11	3
AT1G67090	(-)	AT1G50055(-)_25-(+1)	13	3+1
AT1G67090	(-)	AT3G22121(+)_29+(5)	13	3+1
AT1G67090	(-)	ath-miR838	13	2-1
AT1G67090	(-)	AT2G27400(-)_31+(3)	12	2
AT1G67090	(-)	Chr2:6277-9595(+)_21-	10.5	2+1
AT1G67090	(-)	AT1G62860(-)_22-(2)	14	1-1
AT1G67090	(-)	AT1G67090(-)_26+(-1)	14.5	1-1
AT1G67090	(-)	AT3G22121(+)_29+(6)	11	3
AT1G11700	(+)	AT1G11700(+)_32-(4)_22	10	2-1
AT1G12820	(-)	ath-miR393a-5p	75	1
AT1G12820	(-)	ath-miR393b-5p	105	1
AT1G13360	(-)	AT1G18000(+)_47+	11	2
AT1G13360	(-)	AT1G13360(-)_32-(-7)	10	2+1
AT1G13360	(-)	AT2G26975(+)_15+(5)	12.5	1-1
AT1G13360	(-)	AT1G13360(-)_23-(7)	11.5	1
AT1G13360	(-)	AT1G13360(-)_25-(7)	12.5	1
AT1G13360	(-)	AT1G13360(-)_25-(8)	10	1+1

AT1G18000	(+)	AT4G13575(-)_21+(+1)	14.5	3
AT1G20450	(-)	AT1G20450(-)_11-(-1)	11	1-1
AT1G20450	(-)	AT1G63130(+)_47+(-5)	11.5	3
AT1G20450	(-)	AT2G31820(+)_25+	14	2
AT1G20450	(-)	AT2G35945(-)_27-(+1)	14	2
AT2G27400	(-)	ath-miR173-5p	63.5	1
AT2G27400	(-)	AT1G63080(-)_30+(-10)_22	57	1
AT2G27400	(-)	AT2G39675(-)_23-(+1)	32	3
AT2G27400	(-)	AT2G39675(-)_23-_22	45	3
AT2G35945	(-)	AT2G35945(-)_11-(-11)	62.5	2+1
AT2G38230	(+)	AT2G38230(+)_39-(-9)	10	1-1
AT2G38230	(+)	AT2G38230(+)_24-(2)	10	3+1
AT2G39675	(-)	ath-miR173-5p	185	1
AT2G39681	(-)	ath-miR173-5p	56.5	1
AT2G39681	(-)	AT1G63080(-)_30+(-10)_22	105	1
AT2G39681	(-)	AT2G27400(-)_23-_22	21	3
AT2G39681	(-)	AT2G39675(-)_23-(+1)	18.5	3
AT2G45160	(-)	AT2G45160(-)_24-(-9)	148	3
Chr2:5497-5801	(+)	Chr2:5497-5801(+)_23-(-9)_22	12.5	1
Chr2:5497-5801	(+)	Chr3:14199468-14199772(+)_23-_22	17	3-1
Chr3:14192872-14193187	(+)	Chr3:14192872-14193187(+)_13-(8)_22	25.5	1+1
Chr3:14192872-14193187	(+)	Chr3:14192872-14193187(+)_18-(8)_22	90.5	1+1
Chr3:14192872-14193187	(+)	Chr2:6277-9595(+)_160-_22	24.5	2
Chr3:14192872-14193187	(+)	Chr3:14192872-14193187(+)_16-(-4)_22	15	2
Chr3:14192872-14193187	(+)	Chr2:6277-9595(+)_160-(+1)_22	22.5	2+1
Chr3:16158765-16159371	(-)	Chr3:16158765-16159371(+)_25+(-1)_22	14	3
AT3G60630	(-)	AT2G45160(-)_24-(-10)_22	38	3
AT3G60630	(-)	ath-miR171c-3p	142.5	2+1
AT3G62980	(-)	ath-miR393a-5p	68	1
AT3G62980	(-)	ath-miR393b-5p	46.5	1
AT3G11410	(-)	Chr3:17136708-17137721(+)_47-(-6)	12	3
AT3G11410	(-)	AT1G13360(-)_25-(8)	16.5	1
AT3G17185	(+)	ath-miR390a-5p	60.5	1+1
AT3G17185	(+)	ath-miR390b-5p	32.5	1+1
AT3G19890	(-)	ath-miR2939	10	2
AT3G22121	(+)	AT3G22121(+)_47-(8)	19	3+1
AT3G22121	(+)	AT4G00150(-)_17+(4)_22	42	1
AT3G23690	(-)	AT2G26975(+)_19+(3)	90.5	1
Chr3:9417547-9417820	(-)	ath-miR828	38	1
AT3G26810	(+)	ath-miR393a-5p	105	1
AT3G26810	(+)	ath-miR393b-5p	70	1

AT4G18670	(-)	AT1G11700(+)_24+(+1)_22	10	2
Chr4:1318879-1319187	(-)	Chr3:17136708-17137721(+)_34+(-5)	12	1
AT4G04565	(+)	AT1G62914(+)_34+(10)_22	26.5	1
Chr4:1476283-1476590	(+)	AT1G62914(+)_34+(10)_22	10	2
AT4G37540	(-)	AT4G37540(-)_27-(-1)	14	2-1
AT4G37540	(-)	AT4G37540(-)_28-(-1)	15.5	2-1
AT4G37540	(-)	AT4G37540(-)_27-	11	2
AT4G38770	(-)	AT1G29910(-)_33+(+1)	10	2-1
AT4G38770	(-)	AT1G29910(-)_33+	37	2
AT4G38770	(-)	AT1G11700(+)_21+(-7)	14.5	2
AT4G38770	(-)	AT1G11700(+)_25+(-5)_22	10.5	2
AT4G38770	(-)	AT2TE16865(+)_23+	10.5	1
AT4G08990	(+)	ath-miR773a	10.5	3
AT4G08990	(+)	ath-miR773b-3p	11	3
AT4G00150	(-)	ath-miR170-3p	69	1-1
AT4G00150	(-)	ath-miR171a-3p	31	1-1
AT4G10340	(+)	AT4G10340(+)_15-(8)	12.5	2-1
AT4G10340	(+)	AT5G16640(+)_21+	39.5	2-1
AT4G14610	(-)	AT4G18670(-)_21+(6)	24.5	1
AT5TE42355	(+)	AT3G11410(-)_16+(8)_22	20	2+1
AT5TE42355	(+)	AT3G11410(-)_16+(8)	55	2+1
AT5TE42440	(+)	Chr3:17136708-17137721(+)_24+(2)	23.5	2+1
AT5G43740	(+)	ath-miR472-3p	106.5	3
Chr5:23394264-23394495	(+)	ath-miR390a-5p	11.5	1
AT5G60450	(-)	AT3G17185(+)_17+(-1)	62.5	2-1
AT5G60450	(-)	AT3G17185(+)_17+(-1)_22	36.5	2-1
AT5G60450	(-)	AT3G17185(+)_18+(-1)_22	22	2-1
AT5G60450	(-)	Chr5:23394264-23394495(+)_13+_22	36.5	2-1
AT5G16640	(+)	AT1G62670(-)_30-(-7)	72.5	2
AT5G16640	(+)	AT2G39681(-)_16-(+1)	21.5	3
AT5G16640	(+)	AT2G39681(-)_16-_22	11.5	3
Chr5:7006522-7007118	(+)	AT4G10340(+)_16+(+1)	10	2
Chr5:7006522-7007118	(+)	Chr5:7006522-7007118(+)_34-(-1)_22	15	1+1
Chr5:7006522-7007118	(+)	Chr5:7006522-7007118(+)_34-	25.5	1+1

*Refers to the rank of the corresponding bin of sRNA reads. Adjacent bins are noted with their relative position.

Consistent with (Rajeswaran *et al*. 2012) multiple triggers per *PHAS* locus were detected in some cases (Fig. 9).

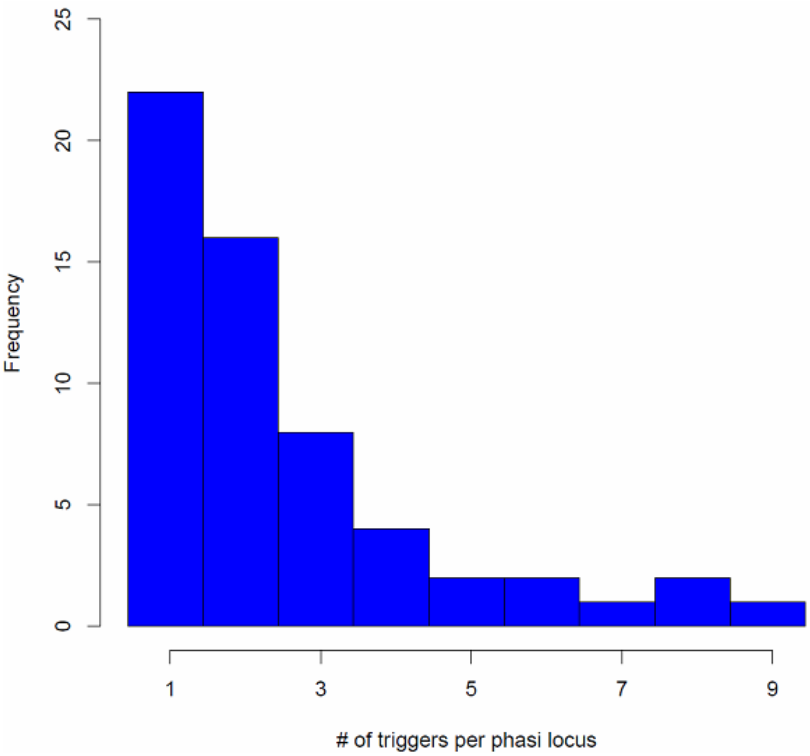

Fig. 9. Distribution of number of phasiRNA production triggers in *PHAS* loci.

Genomic features assignment for *PHAS* loci

To identify the genes or genome features corresponding to *PHAS* loci, regions containing *PHAS* loci were compared to the Araport11 (Cheng *et al*. 2017) genome annotation. An sRNA trigger acts on the transcribed strand of the *PHAS* locus, and when available, the polarity of the transcripts was inferred from the trigger assignments. Knowledge of the transcript's polarity increased the confidence in the identification of the corresponding gene or feature and helped to resolve situations where annotated features existed for both genomic strands. Sixty eight of the one hundred and nine loci overlapped with annotated genomic locations (Table 5); these included: i) Twenty-nine *PHAS* loci overlapping with gene types previously shown to be involved in phasiRNA production, including five TAS genes, fifteen PPR/TPR protein genes, five F-box

containing genes (ARFs), and four disease resistance genes (Wang and Chekanova 2016), ii) Seven genes involved in gene expression regulation including AGO1, DCL1, GRAS family transcription factors and a DNA methyl transferase were found to produce phasiRNAs, iii) Three overlapping with sites annotated as "long non-coding RNA" and three with "natural antisense transcript"; these regions of phasiRNA production could be re-annotated as *PHAS* loci based on these results, iv) Seventeen genes involved in metabolism, structure and other functions, with no previous connections to phasiRNAs, v) Nine transposable elements showed overlapping with detected *PHAS* loci, and vi) Forty-one *PHAS* loci located in unannotated regions of the *A. thaliana* genome.

Table 5. Overlap of detected phasi loci to genomic features (n=109).

Chr	Start	End	Inferred Polarity	Feature type	Feature polarity	Gene ID	Summarize annotation		
Chr1	17890747	17891801	-	gene	-	AT1G48410	AGO1	1054	82
Chr2	13529631	13530391	+	gene	+	AT2G31820	Ankyrin repeat family protein	626	100
Chr3	15676411	15676939	+	Transposable element	+	AT3TE63395	ATENSPM2	528	100
Chr4	4554678	4555391	+	Transposable element	+	AT4TE19135	ATENSPM3	713	100
Chr5	12167484	12168140	+	Transposable element	+	AT5TE43315	ATHILA	656	100
Chr1	15464214	15465381	+	Transposable element	+	AT1TE51040	ATHILA6A	1167	100
Chr1	15471214	15472320	+	Transposable element	+	AT1TE51040	ATHILA6A	1106	100
Chr1	15485137	15486243	+	Transposable element	+	AT1TE51040	ATHILA6A	1106	100
Chr5	11850189	11850903	+	Transposable element	+	AT5TE42470	ATHILA6A	714	100
Chr5	24309296	24309946	-	gene	-	AT5G60450	Auxin response factor 4	650	89
Chr3	9869923	9870891	+	gene	+	AT3G26810	Auxin signaling F-box 2	859	74
Chr1	4368561	4369316	-	gene	-	AT1G12820	Auxin signaling F-box 3	557	100
Chr3	8529663	8530881	-	gene	-	AT3G23690	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein	1218	9
Chr1	10472358	10473449	-	gene	-	AT1G29910	chlorophyll A/B binding protein 3	100	100
Chr2	11512823	11513578	+	gene	+	AT2G26975	Ctr copper transporter family	755	100
Chr1	27613	28536	+	gene	+	AT1G01040	DCL1	923	100
Chr1	7087973	7088710	-	gene	-	AT1G20450	Dehydrin family protein	737	100
Chr5	17560634	17561583	+	gene	+	AT5G43730	Disease resistance protein (CC-NBS-LRR class) family	949	100
Chr5	17566353	17567822	+	gene	+	AT5G43740	Disease resistance protein (CC-NBS-LRR class) family	1469	100
Chr5	15555197	15556980	+	gene	+	AT5G38850	Disease resistance protein (TIR-NBS-LRR class) family	1783	100
Chr5	15556764	15557940	+	gene	+	AT5G38850	Disease resistance protein (TIR-NBS-LRR class) family	1176	83

Chr4	5764617	5765583	+	gene	+	AT4G08990	DNA (cytosine-5-)-methyltransferase family protein	803	100
Chr3	6915371	6916021	-	gene	-	AT3G19890	F-box family protein	650	100
Chr3	23273140	23274021	-	gene	-	AT3G62980	F-box/RNI-like superfamily protein	881	75
Chr3	4341477	4342208	+	gene	+	AT3G13370	Formin-like protein	545	100
Chr2	18618750	18619652	-	gene	-	AT2G45160	GRAS family transcription factor	902	100
Chr3	22410771	22411711	-	gene	-	AT3G60630	GRAS family transcription factor	940	100
Chr4	57737	58579	-	gene	-	AT4G00150	GRAS family transcription factor	842	96
Chr3	18733678	18734454	+	gene	+	AT3G50480	Homolog of RPW8 4	742	100
Chr1	4577081	4578013	-	gene	-	AT1G13360	Hypothetical protein	932	100
Chr4	10276259	10277210	-	gene	-	AT4G18670	Leucine-rich repeat (LRR) family protein	951	93
Chr4	6407949	6408766	+	gene	+	AT4G10340	Light harvesting complex of photosystem II 5	760	100
Chr4	17639492	17640340	-	gene	-	AT4G37540	LOB domain-containing protein 39	848	90
Chr1	29427736	29428386	+	gene	+	AT1G09797	Long_noncoding_rna	585	82
Chr4	1472592	1473252	+	gene	+	AT4G04565	Long_noncoding_rna	544	87
Chr4	10180041	10180628	+	gene	+	AT4G06810	Long_noncoding_rna	512	100
Chr1	6199903	6201311	+	gene	+	AT1G18010	Major facilitator superfamily protein	1408	100
Chr2	15090667	15091477	+	gene	+	AT2G35945	Natural antisense transcript overlaps with AT2G3594	810	100
Chr3	7795023	7796315	+	gene	+	AT3G22121	Natural antisense transcript overlaps with AT3G22120	1292	100
Chr5	16640019	16641090	+	gene	+	AT5G41612	Natural antisense transcript overlaps with AT5G41610	1071	100
Chr1	4354234	4355465	+	gene	+	AT1G12775	Pentatricopeptide repeat (PPR) superfamily protein	1231	100
Chr1	23177618	23178913	-	gene	-	AT1G62590	Pentatricopeptide repeat (PPR) superfamily protein	1295	100
Chr1	23204992	23206329	-	gene	-	AT1G62670	Pentatricopeptide repeat (PPR) superfamily protein	1337	100
Chr1	23299381	23300725	+	gene	+	AT1G62910	Pentatricopeptide repeat (PPR) superfamily protein	1344	95
Chr1	23301883	23303226	+	gene	+	AT1G62914	Pentatricopeptide repeat (PPR) superfamily protein	1279	100

Chr1	23389265	23390461	-	gene	-	AT1G63080	Pentatricopeptide repeat (PPR) superfamily protein	1196	100
Chr1	23489943	23491196	+	gene	+	AT1G63330	Pentatricopeptide repeat (PPR) superfamily protein	1253	100
Chr1	23507648	23509359	+	gene	+	AT1G63400	Pentatricopeptide repeat (PPR) superfamily protein	1711	100
Chr5	5461370	5462166	+	gene	+	AT5G16640	Pentatricopeptide repeat (PPR) superfamily protein	796	100
Chr4	18097028	18097825	-	gene	-	AT4G38770	Proline-rich protein 4	797	100
Chr3	3584388	3585225	-	gene	-	AT3G11410	Protein phosphatase 2CA	837	100
Chr4	8381922	8383118	-	pseudogene	-	AT4G14610	Pseudogene, disease resistance protein (CC-NBS-LRR class)	1196	100
Chr2	16011259	16012481	+	gene	+	AT2G38230	Pyridoxine biosynthesis 1	1222	100
Chr1	25048301	25049476	-	gene	-	AT1G67090	Ribulose bisphosphate carboxylase small chain 1A	1175	100
Chr1	3945621	3946579	+	gene	+	AT1G11700	Senescence regulator	958	100
Chr1	23306868	23308169	+	gene	+	AT1G62930	Tetratricopeptide repeat (TPR)-like superfamily protein	1301	100
Chr1	23413170	23414408	+	gene	+	AT1G63130	Tetratricopeptide repeat (TPR)-like superfamily protein	1238	100
Chr1	23419721	23420644	+	gene	+	AT1G63150	Tetratricopeptide repeat (TPR)-like superfamily protein	923	100
Chr1	23451028	23452017	+	gene	+	AT1G63230	Tetratricopeptide repeat (TPR)-like superfamily protein	989	100
Chr1	23587365	23588025	+	gene	+	AT1G63630	Tetratricopeptide repeat (TPR)-like superfamily protein	660	100
Chr1	23987192	23988074	-	gene	-	AT1G64583	Tetratricopeptide repeat (TPR)-like superfamily protein	882	100
Chr2	16539425	16540243	-	gene	-	AT2G39681	Trans-acting siRNA primary transcript (TAS2)	818	79
Chr3	5861814	5862603	+	gene	+	AT3G17185	Trans-acting siRNA primary transcript (TAS3)	623	79
Chr2	11721305	11722396	-	gene	-	AT2G27400	Trans-acting siRNA1a primary transcript (TAS1a)	858	82
Chr1	18549031	18549996	-	gene	-	AT1G50055	Trans-acting siRNA1b primary transcript (TAS1b)	793	98
Chr2	16537279	16538289	-	gene	-	AT2G39675	Trans-acting siRNA1c primary transcript (TAS1c)	990	72

Chr4	5567581	5568157	+	Transposable element gene	+	AT4G08710	Transposable_element_gene	416	100
Chr5	11778276	11779152	+	Transposable element gene	+	AT5G31963	Transposable_element_gene	876	0
Chr1	4184825	4185643	+	.	.	.			0
Chr1	4295606	4296426	+	.	.	.			0
Chr1	5297657	5298358	+	.	.	.			0
Chr1	6194691	6196238	+	.	.	.			0
Chr1	11454368	11455045	+	.	.	.			0
Chr1	17203515	17204064	+	.	.	.			0
Chr1	21125592	21126324	+	.	.	.			0
Chr1	23275297	23276594	+	.	.	.			0
Chr1	23385828	23386912	+	.	.	.			0
Chr1	23489122	23489850	+	.	.	.			0
Chr1	24720922	24721729	+	.	.	.			0
Chr2	5277	6021	+	.	.	.			0
Chr2	6057	9815	+	.	.	.			0
Chr2	855427	856563	+	.	.	.			0
Chr2	3251765	3252578	+	.	.	.			0
Chr2	3966526	3967245	+	.	.	.			0
Chr2	7839675	7840178	+	.	.	.			0
Chr3	343010	344051	+	.	.	.			0
Chr3	6524122	6524776	+	.	.	.			0
Chr3	9417327	9418040	-	.	.	.			0
Chr3	11983539	11984211	+	.	.	.			0
Chr3	14192652	14193407	+	.	.	.			0
Chr3	14199248	14199992	+	.	.	.			0
Chr3	15677079	15677687	+	.	.	.			0
Chr3	15677496	15678238	+	.	.	.			0
Chr3	16158545	16159591	-	.	.	.			0
Chr3	17136488	17137941	+	.	.	.			0

Chr3	17445467	17446154	+	.	.	.			0
Chr4	1318659	1319407	-	.	.	.			0
Chr4	1476063	1476810	+	.	.	.			0
Chr4	3741386	3742162	+	.	.	.			0
Chr4	7890288	7890985	+	.	.	.			0
Chr5	7006302	7007338	+	.	.	.			0
Chr5	7683595	7684615	+	.	.	.			0
Chr5	7684440	7685163	+	.	.	.			0
Chr5	9789275	9789883	+	.	.	.			0
Chr5	11813978	11814729	+	.	.	.			0
Chr5	15698990	15699661	+	.	.	.			0
Chr5	15757426	15758414	+	.	.	.			0
Chr5	22322508	22323141	+	.	.	.			0
									0
Chr5	23394044	23394715	+	.	.	.			82

[a] Refers to the % of nucleotides of overlap between a *PHAS* locus and a genome feature.

Experimental support for sRNA cleavage activity.

In plants sRNAs act mainly through cleavage of their transcripts, yet there are examples of other mechanisms such as translational repression (Wang and Chekanova 2016; Borges and Martienssen 2015). Also, it has been shown that not all phasiRNAs produced from a *PHAS* locus are active, instead only some of them appear to be competent for loading into argonaute (AGO) containing complexes where they exert their activities (Fei *et al*. 2013). Therefore, in this study known miRNAs and phasiRNAs derived from the detected *PHAS* loci (including non-canonical phasiRNAs) were evaluated for biological activity using degradome data.

For *A. thaliana*, a limited number of degradome libraries are publicly available (Supplementary table 2), including eleven datasets corresponding to inflorescence tissue, six to leaf tissue, five to seedling tissue and one whole plant (Addo-Quaye *et al*. 2008; Creasey *et al*. 2014; Thatcher *et al*. 2015; Hou *et al*. 2016; Lin *et al*. 2017). Sixteen new degradome libraries from Cucumber mosaic virus-infected leaf tissue were produced as part of this study (accessions: SAMN07949266- SAMN07949281) and all available libraries were evaluated based on their yield (Fig. 10). The data produced in this study represented a significant increase (~20%) in the total amount of degradome data available for *A. thaliana* in the NCBI SRA database.

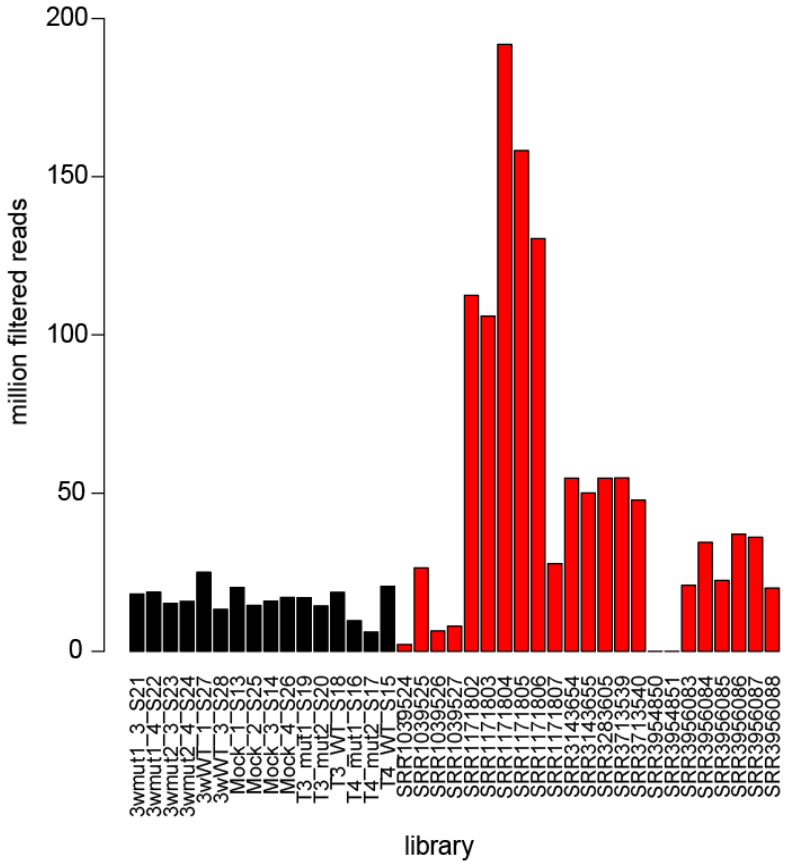

Fig. 10. Summary of information available in degradome libraries. The histogram shows library yield as the number of million filtered reads for each of the 39 libraries. Colors: black refers to data produced in this study (16 libraries) and red refers to NCBI SRA data (23 libraries).

sRNAs were annotated as active if at least one of their predicted targets was confirmed (Supplementary table 4). Experimental support was found for the targeting and cleavage activity of 99% of all annotated miRNAs, and the number of targets per active miRNA ranged from 1 to 302. In the case of phasiRNAs ~91% were found to be active, and the number of targets per active phasiRNA ranged from 1 to 556. With the additional degradome data from this study (Fig. 10), this is the most comprehensive evaluation of sRNA cleavage activity to date for *A. thaliana*.

Active PhasiRNA characterization.

phasiRNAs whose activities were experimentally validated were evaluated based on their sizes and phased registers. Fig. 11a shows the absolute amount of active canonical and non-canonical phasiRNAs (22 nt long or derived from an alternative phased register). Inclusion of non-canonical phasiRNA resulted in a ~2X increase in the number of biologically active phasiRNAs (Fig. 11b).

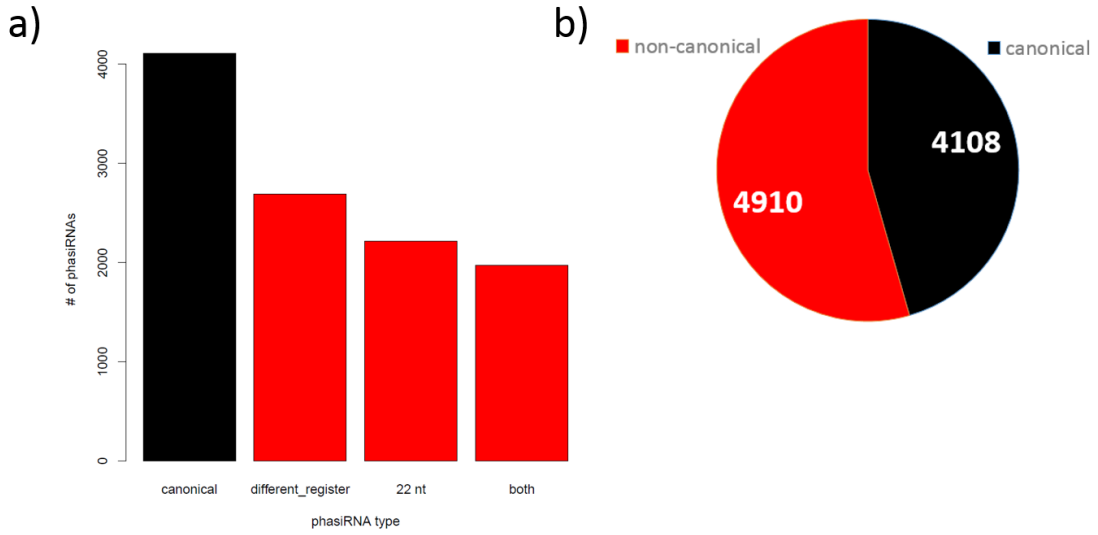

Fig. 11. a) Distribution of phasiRNAs according to their size and phased registers, b) Comparative proportion of phasiRNAs from canonical size and phased registers to non-canonical size and phased register.

Network identification

In order to integrate datasets and describe the wild type *A. thaliana* sRNA-mediated regulatory network, a bipartite, directed network was constructed. In order to differentiate between the theoretical versus functional (biologically relevant) components of the network, only interactions validated by degradome data were included. sRNA nodes (miRNAs and phasiRNAs) were restricted to those with validated targets identified in degradome analysis. Transcript nodes

included only those annotated as precursors (pre-miRNAs, *PHAS* loci) or validated in the degradome analysis to be direct targets of active sRNAs. The resulting network contained a total of 26 684 nodes, composed of 12 430 sRNA nodes (427 miRNAs + 12 003phasiRNAs) and 14 254 transcript nodes (325 miRNA precursors, 109 *PHAS* loci and 13 820 target transcripts). These nodes were connected by 120 664 edges, 12 513 of these were involved in the biogenesis of sRNAs and 108 151 edges were involved in sRNA cleavage of transcripts (Fig. 12, Supplementary files 1 and 2).

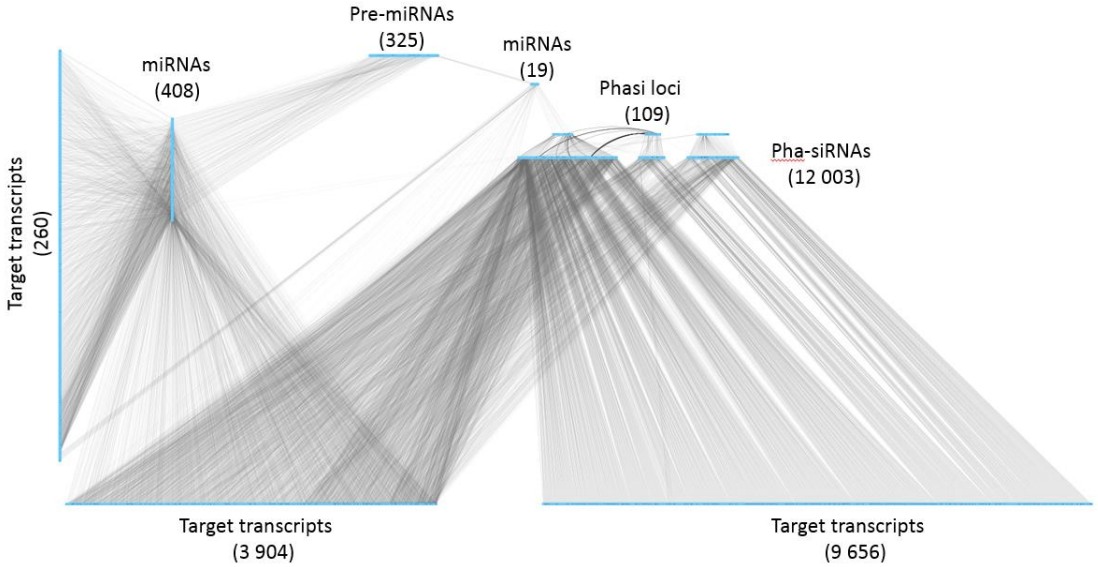

Fig. 12. Representation of the resulting sRNA-mediated regulatory network. Nodes are colored in light blue, edges are gray. The type and abundance is mentioned for each class of nodes. The network has been manually organized to reflect the biogenesis of sRNAs. Two sets of miRNAs (and numbers) are diagrammed; the miRNAs (408) that do not induce phasiRNA production, and those that induce phasiRNA production (19). Target transcripts were divided into three sets according to the type of sRNA targeting, e.g. targeted by miRNAs only, targeted by both miRNAs and phasiRNAs, and those targeted only by phasiRNAs.

Determination of the regulatory contribution of the sRNA-mediated network.

Three metrics were used to assess the regulatory contribution of the resulting sRNA-mediated network. The proportion, function and regulatory roles of the genes included in the network were evaluated. There are 33,341 *A. thaliana* genes defined in the araport11 genome

annotation (Cheng *et al*. 2017). The proportion of genes validated as interacting with sRNAs was about ~42% (n=14 110) of the total annotated genes. The networks regulatory role was assessed at a functional level using the GO annotations of the genes under sRNA control to determine the biological processes in which sRNAs exert their control. Go slim annotations give a broad overview of the ontology content; using these terms, 35 of 45 functional categories under the biological processes domain were found to be disproportionally enriched in the network when compared to the genome (Fig. 13). Given the high level of representation of functional categories, the network was evaluated for underrepresentation of functional categories. Cell-to-cell signaling and translation were the only terms out of 45 under the biological processes domain that was found to be underrepresented in the sRNA-mediated regulatory network (corrected p-value= 1.4467e-11, 4.9890e-2 respectively).

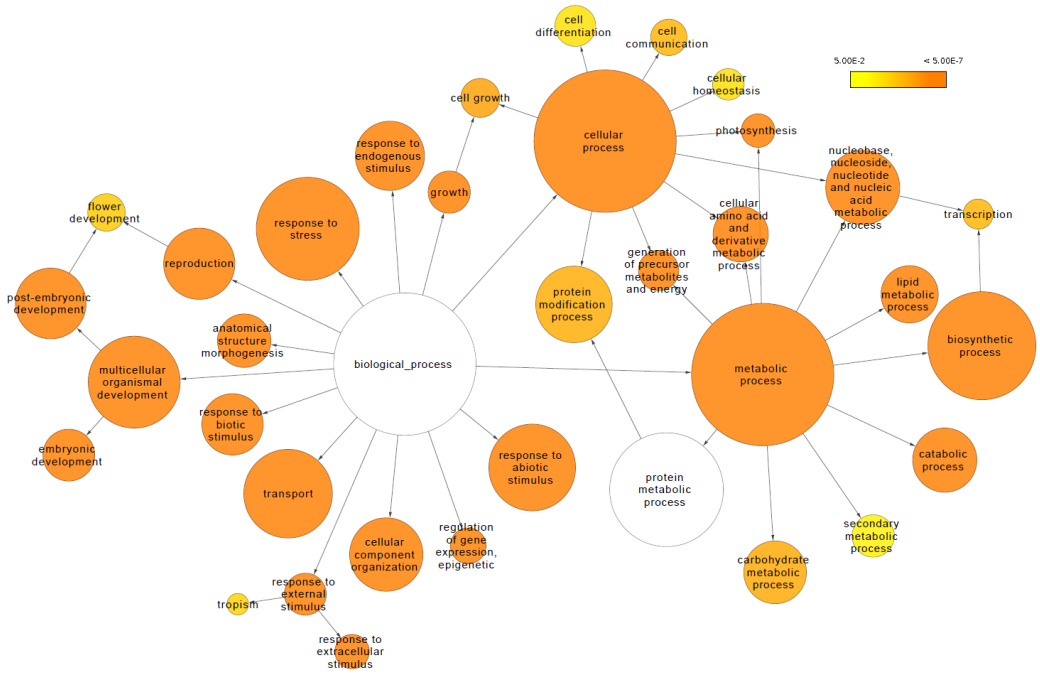

Fig. 13. sRNA-mediated network functional analysis. This is a network representation of GO Slim categories showing enrichment for genes under sRNA regulation. Size of the nodes represents the breath of the category, color scale indicates the corrected p-values as described in Maere *et al.* (2005).

Lastly, to assess the regulatory roles of the genes included in the network, the proportion of genes annotated to be involved in gene expression regulation were evaluated for regulatory interactions with sRNAs. This evaluation indicated that 55% (n=1606) of genes involved in gene expression regulation are under sRNA control, including 48% (n=837) of annotated transcription factors in the current version (July 2017) of Palaniswamy *et al.* (2006). Based on the proportion of total genes interacting with sRNAs, the enrichment in the network of a majority of categories under the biological processes domain, and the proportion of regulatory genes under sRNA control, our results are consistent with the notion that sRNAs play a key regulatory role in plants.

Structural analysis of the sRNA-mediated regulatory network

Networks structure analysis allows a multiscale study of complex biological systems, as for the sRNA-mediated regulatory network presented here; global and local features can be identified and compared to related systems. To better understand properties of the resulting sRNA-mediated regulatory network and to determine the interconnectivity between miRNA and phasiRNA, structural features were analyzed. The sRNA-mediated regulatory network consisted of 16 disconnected components, or groups of connected nodes. Given the directed nature of the networks, each one of these corresponded to weakly connected components, as the components were calculated without considering the direction of the edges. The distribution of the number of nodes per component was uneven, with the largest component containing ~99.9% of the nodes, and the remainder consisting of less than 10 nodes. Network density, a measure of the ratio of the observed number of edges to the maximum number of edges, in this case was very low (<0.001), consistent with results from studies on other biological networks (Leclerc 2008). By contrast, the

clustering coefficient, a measure showing the tendency of a graph to be divided into clusters, was very low (<0.001), whereas in biological systems, higher values are usually observed (Pavlopoulos *et al*. 2011). In totality, the node degree showed a heavy-tailed distribution, both for indegree and outdegree (Fig. 14), describing a limited number of nodes with high degree while the majority have low degree. However, the bipartite nature of the network warrants a separate evaluation of the different types of interactions. Degree distributions for sRNAs and transcripts were evaluated separately (Fig. 14). Negative correlations were found between node degree and abundance, except for the case of sRNA indegree as this indegree distribution is restricted by the nature of sRNA biogenesis and does not allow for testing.

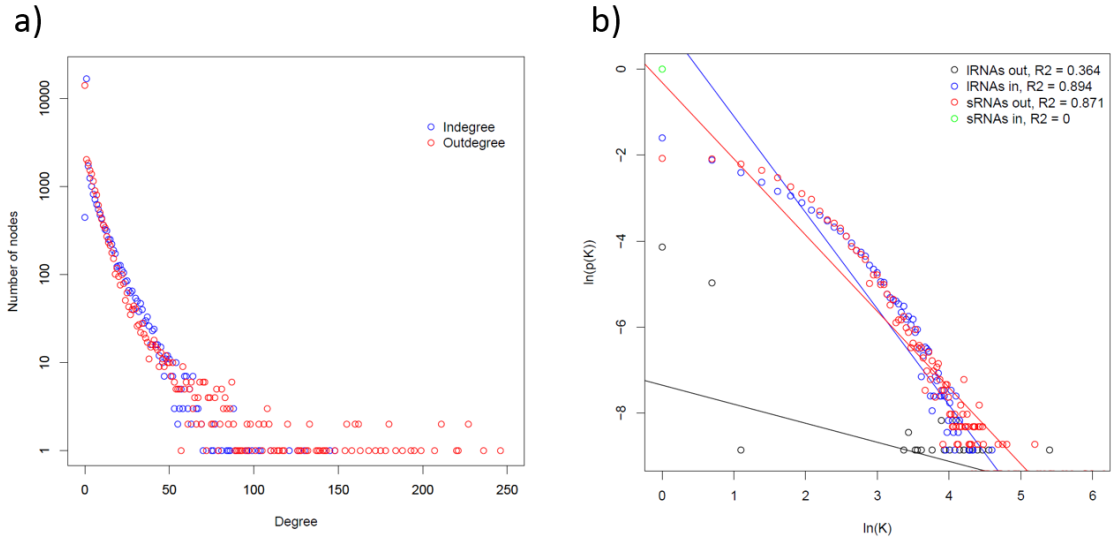

Fig. 14. a) Total degree distributions b) Degree distribution of the individual sRNA-mediated network components. Degree is represented by K and p(K) is the number of nodes with degree K divided by total nodes. Black=transcript out, blue=transcript in, red=sRNA out, green=sRNA in.

Discussion

The objective in this work was to identify a comprehensive, experimentally supported sRNA-mediated regulatory network at a genome-wide level. This required identifying the network components (pre-miRNA, miRNAs, *PHAS* loci, phasiRNAs and target transcripts) and the interactions between them, i.e. the nodes and edges of the network. MacLean *et al.* (2010) provided the first description of a broad level sRNA regulatory network in plants. Following the MacLean model and significant advancements in an understanding of sRNA biogenesis and activity (Wang and Chekanova 2016; Fei *et al*. 2013; Rajeswaran *et al*. 2012), it became possible to investigate sRNA networks from a genomic view point, using only biologically relevant (experimentally supported) interactions. The miRBase database (Kozomara and Griffiths-Jones 2014) and the araport11 genome annotation (Cheng *et al*. 2017) represent rich resources for mining miRNA precursors and mature sequences, gene transcripts, and *PHAS* loci. Using the existing model of sRNA biogenesis (Fei *et al*. 2013) and published bioinformatics tools (Guo *et al*. 2015), in combination with additional biogenesis features described by Rajeswaran *et al.* (2012), we designed an experimental approach and a bioinformatics analysis tool to perform a genome wide identification of *PHAS* loci, their triggers and resulting phasiRNAs. To account for the fact that sRNA production tends to be inducible and the expression can vary under different circumstances, a combination of libraries was employed. These included those produced in this study and all sRNA libraries from the NCBI SRA database representing *A. thaliana* from multiple developmental stages, different tissues types, and plants grown under varied biotic and abiotic stress conditions. Finally, to obtain a better view of sRNA cleavage activity on targeted transcripts, the degradome data available at NCBI for wild type *A. thaliana* was substantially expanded

(~20%) with the libraries produced in this study. This allowed for the most comprehensive evaluation of the sRNA cleavage activity in *A. thaliana* to date, followed by the development of a genome-wide, experimentally supported sRNA-mediated regulatory network.

In order to accurately identify network components and their interactions, a number of factors were critical: a broader detection of *PHAS* loci, identification of non-canonical phasiRNAs, a newly-designed strategy to assign *PHAS* triggers and a significantly larger degradome dataset. The results from *PHAS* loci detection across combined sets of sRNA libraries made evident the need for the combinatorial approach used here in order to address the high levels of variability (Fig. 7). It was clear that individual libraries would fail to provide a representative view of *PHAS* loci, and that some regions only produced phasiRNAs under specific circumstances (Fig. 10). Many *PHAS* loci were detected in multiple libraries, increasing the confidence of their assignment and resulting in a better definition of the 5' and 3' ends by combining overlapping loci into a maximum-length locus. Detection of previously described *PHAS* loci (Fei *et al*. 2013) was an indication of the accuracy of this strategy. Beyond the expected types of phasiRNA producing genes (TAS, PPRs, ARFs, and disease resistance), a number of new phasiRNA producing genes and non-annotated regions of the genome were detected as *PHAS* loci (Table 5), and these findings can be used to update and refine annotations for these regions. A number of regions matched locations of natural antisense transcripts, which have been reported to produce natural antisense small interfering RNAs (NAT-siRNAs) in a phased fashion (Borges and Martienssen 2015).

As proposed by Rajeswaran *et al.* (2012) and using the biogenesis features described in their work, the inclusion of non-canonical phasiRNAs resulted in an expanded more comprehensive detection of miRNA-phasiRNA biogenesis cascades. Nearly half of the validated phasiRNAs were derived from an alternative phased register or were 22 nt long. Not all non-

canonical phasiRNAs are novel, the most prominent case is a *TAS3c* derived phasiRNA, AT2G39675(-)_23-(+1) (described as "athTAS1c-D6(-)"), which has been shown to target its progenitor transcripts and trigger the productions of secondary phasiRNAs (Rajeswaran *et al*. 2012); it also acts in trans on other TAS transcripts. Despite the relevance of AT2G39675(-)_23-(+1) within the TAS-derived phasiRNA production cascades, it is not often appreciated that its location is shifted one nt with respect to the main 21 nt phased register set by the miR173 cleavage site. Moreover, the location is shifted by the production of a 22 nt phasiRNA in the previous register (Rajeswaran *et al*. 2012). Together, consideration of non-canonical phasiRNAs in this study provided a more accurate and comprehensive view of sRNA activity and regulatory potential.

The assignments of primary triggers were consistent with previous reports (Fei *et al*. 2013; Howell *et al*. 2007). There are limited reference points to determine the accuracy of assignment of secondary and tertiary triggers (Rajeswaran *et al*. 2012; Howell *et al*. 2007), instead the validity of these triggers was evaluated by the biological activity of their corresponding phasiRNAs. phasiRNAs produced by these identified secondary and tertiary triggers were shown to be functional. Degradome data was used to confirm their activity (discussed below), and additionally, non-canonical phasiRNAs were found to function as phasiRNA triggers, furthering the depth of their regulatory cascades. As described above, AT2G39675(-)_23-(+1), provides the best example of secondary triggers giving rise to phasiRNAs that then function as triggers anew. For the remaining 24% of *PHAS* loci with no assigned triggers, a number of them were found to overlap with regions annotated to produce NAT-siRNAs (Fei *et al*. 2013). This alternative biogenesis mechanism is a likely explanation for some of the *PHAS* loci for which no sRNA trigger was

found. Alternatively, the triggers were below the sensitivity of the methods used in this study and deeper sequencing and higher representation will be needed to find them.

To optimize the accuracy of the newly developed annotation of miRNA-phasiRNAs and produce a regulatory network of biological relevance, the biological activity of miRNAs and the expanded set of phasiRNAs was evaluated using degradome analysis. Using this enlarged dataset, targets were validated for ~99% of the predicted, annotated *A. thaliana* miRNAs. Approximately 91% of the detected phasiRNAs was found to be active. Nearly half of the degradome-validated phasiRNAs corresponded to non-canonical phasiRNAs further validating the accuracy of secondary and tertiary trigger assignments described above, and highlighting the relevance of these commonly overlooked phasiRNAs.

Restricting the network to interactions that were experimentally supported, eliminated the problems associated with false positives in computational predictions, and provided a reference frame for functional, comparative and structural analyses of biological relevance and applicability. The stringency and effectiveness of this approach is reflected by the significant reduction of the size of the network compared to MacLean *et al.* (2010); the number of sRNA nodes was reduced from ~40 000 to 12 430. Similarly, the number of transcripts (long RNAs (lRNA) in their study) was reduced from ~18 000 to 14 254. The number of edges was also reduced, from ~38 000 "source" and ~140 000 target edges reported in MacLean *et al.* (2010) to 12 513 "source" (biogenesis related) edges, and 108 151 "target" (cleavage related) edges in this study, respectively.

Evaluation of this validated set of cleavage related edges permitted an initial exploration of partitioning and co-regulation between miRNA and phasiRNAs. miRNAs were found to be involved in regulation of ~30% of the target transcripts (Fig. 12). As expected from their role in

cascades, the phasiRNA regulatory spectrum was very broad, and phasiRNAs were found to be involved in the regulation of ~98% of the target transcripts, and about ~28% of targets transcripts were found to be controlled by both miRNAs and phasiRNAs.

Cell functioning, behavior and fate are controlled by the topology and dynamics of regulatory gene expression networks. Transcription factors and sRNAs appear to be the primary regulators in these networks (Cora *et al*. 2017), and technological and conceptual advances are making the study of gene regulation at this level possible. Yet in order to integrate these different regulatory systems their individual roles and contributions must be established. In this study, limiting the sRNA-mediated network to validated interactions allowed for a realistic evaluation of its regulatory contribution. The evaluation of the network's regulatory contribution was performed at three different levels based on: i) the number of genes involved in the network, ii) the biological function of these genes, and iii) the interactions between sRNA and other gene expression regulators.

Close to 42% of *A. thaliana* genes were found to be under sRNA control in this study (Fig. 12). Given the stringency of the analysis and assuming that neither the sRNA nor degradome dataset provided a complete representation of sRNAs and evidence of cleavage, it is reasonable to suggest that the network presented here should be considered a baseline representation of sRNA-mediated regulation that should be updated and revised as more data becomes available. Additionally, it should be noted that in the interest of considering only experimentally supported interactions, only cleavage-based regulation by sRNAs was considered here. Given that this appears to be the main mode of actions of sRNAs involved in PTGS in plants (Wang and Chekanova 2016; Li *et al*. 2012; Borges and Martienssen 2015) the presentation of regulatory

interactions is likely to be representative. Research into alternative mechanisms of action can be expected to provide a better estimation of the regulatory contribution of sRNAs.

To further assess the regulatory contribution of the network, it was evaluated using the GO annotations of the genes included. At a GO slim level, most biological processes were found to be disproportionally enriched in the network. These results are consistent with previous reports of sRNA regulatory roles in diverse processes such as response to stress (Liang *et al*. 2014), development (Nogueira *et al*. 2006), defense (Zhai *et al*. 2011) and other activities (Wang and Chekanova 2016; Borges and Martienssen 2015; Fei *et al*. 2013). Curiously, performing a search for underrepresented functional categories in the network indicated that genes related to translation and cell to cell signaling appear to have little to no regulation via sRNAs.

sRNAs are known to be part of higher level regulatory circuits involving other factors involved in regulation at the transcriptional and posttranscriptional level (Cora *et al*. 2017; Walczak and Tkačik 2011). These interactions were confirmed in this study, as ~55% of genes annotated to have a role in gene expression regulation were found to be under sRNA control. In particular, miRNA and transcription factors have been shown act in coordination in other systems (Lin *et al*. 2012; Cui *et al*. 2007; Croft *et al*. 2012) to regulate common targets, as well as each other (Cora *et al*. 2017), and transcription factors were well represented (48%) in the sRNA-mediated network. These results support the notion of crosstalk between regulatory factors and provides a reference frame to further investigate the regulatory circuits that control gene expression in plants. Quantitative estimates for the role of sRNAs in gene regulation in this study are consistent with the notion of sRNAs as 'master regulators'.

The network representation of the sRNA-based expression regulation allows for a systematical characterization of its structural properties using complex network theory. Network topology analysis can be used to better understand the functional organization, underlying design principles and organizing principles of biological networks (Steuer and López 2007). Evaluation

of network properties and topological features of empirically obtained biological networks revealed interesting commonalities between very distinct regulatory systems (Albert 2005; Pavlopoulos *et al*. 2011; Zhu *et al*. 2007). As observed in other biological networks, the network was found to be sparsely connected, a feature that has been proposed to be evolutionarily selected to preserved robustness (Leclerc 2008). Biological networks have been found to display, on average, higher clustering than random networks, reflecting their modular organization (Ravasz *et al*. 2002). Conversely, the sRNA-mediated network showed a low clustering coefficient. The bipartite nature of the network may obscure the detection of modules or more closely connected regions, and further characterization will be required to confirm observations on clustering. The degree distribution refers to the distribution of the number of edges connected to a node, and several examples of biological networks display power law distributions (Albert 2005). Overall the sRNA network node's degree showed heavy-tailed distributions and negative correlations were found between node degree and abundance; based on this observations the network topology is consistent with a scale free network, a trait that appears to be a common feature in biological networks (Albert 2005). These results are consistent with the previous study in *A. thaliana* using an *in silico* approach to model the interactions between sRNA and transcripts (MacLean *et al*. 2010). Given current models of sRNA biogenesis, their mode of action and their function as part of a network, degree distributions should be further evaluated. The indegree distribution for sRNAs is restricted by their biogenesis; miRNAs and phasiRNAs are derived from specific precursors, which leads to an indegree fixed to one. The degree distribution of transcripts is also restrained: miRNA precursors often produced a single miRNA limiting their outdegree, while target transcripts are inherently terminal nodes with zero outdegree. Altogether, the resulting sRNA-mediated network showed similar structural features to other biological networks.

Individual phasiRNA cascades have been studied in some detail (Liang *et al*. 2014), but a genome-wide view to determine if these cascades correspond to independent modules or if they act together within a larger regulatory network remains an open question. The TAS cascade system provides an indication of the potential for interconnectivity (Rajeswaran *et al*. 2012). The

construction of a genome-level network for the interactions between sRNAs and transcripts allowed an evaluation of interconnectivity between individual phasiRNA regulatory cascades, as well as the connections to miRNA-based regulation. Using a network approach, it was found that a large proportion of sRNA-transcripts interactions (>95%) were part of a major (weakly) connected component indicating a considerable level of connectivity; this suggests that miRNAs and multiple phasiRNA cascades act in coordination in the co-regulation of gene-expression.

CHAPTER SEVEN

FUTURE DIRECTIONS

The future direction of my work will be to apply systems level analyses to understand the role of sRNA-mediated gene regulation in defense responses to virus infection. Below I provide a description of the background for this network-based approach, how limitations to perform this type of analysis are being addressed, and the experimental system I will use to generate data sets for analyses.

Small RNAs (sRNAs) have been proposed as master regulators of gene expression through their role in RNA silencing (Sun *et al*. 2010; Zhai *et al*. 2011; Voorhoeve 2010). In plants, two types of sRNAs, microRNAs (miRNAs) and miRNA-triggered, phased, small-interfering sRNAs (phasiRNAs) have been shown to be involved in the modulation of the transcriptional response to biotic and abiotic stresses (Xin *et al*. 2010; Taylor-Teeples *et al*. 2015; Taliansky *et al*. 2004; Shukla *et al*. 2008; Sunkar *et al*. 2007). However, three important factors have prevented a systems level evaluation of the regulatory role of sRNAs: a) limitations in the understanding of sites of phasiRNA production (e.g. RNAs from which they originate), b) a lack of validated transcript targets for sRNAs and c) a lack of understanding about the redundancy, modularity and overall architecture of sRNA-mediated regulation.

The project described in chapter six resulted in a genome wide sRNA-mediated network of gene regulation. To improve an understanding of phasiRNA production across the genome, all available sRNA data along with the latest discoveries in sRNA biogenesis were combined in an integrative analysis to scan the *A. thaliana* genome for evidence of phasiRNA producing regions.

Similarly, to improve the understanding of cleavage-based, sRNA regulation of transcripts, all available degradome data was used to identify high confidence sRNA-transcript interactions. To avoid false positives, further filtering was performed to select repeatable interactions and produce a curated set of sRNAs and their target transcripts. Using a combination of computational methods and state of the art knowledge on sRNA biogenesis and activity, data on phasiRNA production and sRNA-transcript interactions obtained in this work, together with existing data on miRNA biogenesis was integrated into a regulatory network. The identification of the network components combined with degradome data analyses resulted in the detection of a number of previously undescribed interactions. The construction and evaluation of the resulting network provided insight into the architecture and scope of sRNA-mediated gene regulation; the structural features of the network can be used to study the role of sRNAs in higher level regulatory circuits as proposed previously (Cora *et al*. 2017; Megraw *et al*. 2016; Friard *et al*. 2010; Re *et al*. 2009).

The unbiased, genome-wide approach used to develop the sRNA-mediated network allows a systems level investigation of the role of sRNAs in the coordination and modulation of stress responses. The relevance of sRNAs in plant response to biotic stress is highlighted by the convergent evolution of a set of proteins collectively known as silencing suppressors (Burgyán and Havelda 2011); these silencing suppressors are of pathogen origin and able to interfere with sRNAs and the host RNA silencing machinery acting in response to different types of plant pathogens, including viruses, bacteria and oomycetes (Bivalkar-Mehla *et al*. 2011; Navarro *et al*. 2008; Qiao *et al*. 2013). The recurring appearance in pathogens of mechanisms to compromise host sRNA-mediated regulation indicates these molecules and pathways play an important role in host-pathogen interactions.

Virus infection represents a particularly interesting system to study host-pathogen interactions. Virus infection in plants leads to a complex set of interactions involving host defense mechanisms and viral counter defense strategies. From a defensive point of view the host fights infection using direct and indirect defensive strategies, both of which involve RNA silencing machinery. The host deploys a transcriptional defense response upon detection of viral components or their activity (Whitham *et al*. 2003; Wu *et al*. 2016; Marathe *et al*. 2004). sRNAs together with the RNA silencing machinery contribute to the orchestration of this defense response through post-transcriptional gene silencing of key host transcripts (Huang *et al*. 2016), resulting in an indirect effect of sRNAs on viral replication. The host RNA silencing machinery also directly affects viral replication by cleavage/silencing of viral genomes and transcripts (Zhang *et al*. 2015). The interference with host defense responses by virus-encoded silencing suppressors has dual effects: i) Direct effects by protecting viral genomes and transcripts from degradation and/or silencing by host defense proteins, ii) Indirect effects by altering host sRNA-mediated regulation through interference with sRNA biogenesis and/or activity. The dual roles of the plant RNA silencing machinery in conferring antiviral activity and as key regulators of gene expression result in a weakness that has been exploited by pathogens to overcome host defenses. Because of the widespread role of sRNAs and RNA silencing in the control of homeostasis, development and response to stress, virus-induced alterations can potentially result in a widespread disruption of sRNA regulation and concurrent changes in gene expression; a net result is that of severe phenotypes (symptoms) in infected plants.

The intimate relationship between sRNA-mediated regulation and virus infection provides a good model to evaluate the biological functions and regulatory potential of sRNAs. The availability of the sRNA-mediated regulatory network, together with high quality genomic

resources in the *A. thaliana*/cucumber mosaic virus (CMV) pathosystem allows for a systems level evaluation of sRNA-mediated regulation during virus infection. As a model system, CMV's genome has been studied in detail (Jacquemond 2012). The virus encodes a protein (2b) that functions as a silencing suppressor (Lewsey *et al*. 2010). Mutant strains of CMV with non-functional 2b genes maintain their ability to systemically infect *A. thaliana*, though infected plants are asymptomatic (Lewsey *et al*. 2009). CMV 2b mutants can be used to reveal the effects of silencing suppressors on host sRNA-mediated responses. Since 2b mutants replicate in *A. thaliana*, they are expected to induce a host defense response, however the lack of a functional silencing suppressor prevents interference with the host response, resulting in a functional decoupling of the defense induction from virus-derived interference. A systems levels evaluation of the role of sRNA-mediated gene regulation during virus infection is made possible by the availability of three factors: a) a robust reference frame for sRNA expression and activity, i.e. the global sRNA network described in the previous chapter, b) a model plant system with high quality genome resources, c) a strategy to decouple the host defense induction from virus-induced alterations through the use of CMV 2b mutant strains. The future direction of my work will be to apply systems level analyses to understand the sRNA-mediated gene regulation in defense responses *of Arabidopsis* to infection by CMV.

REFERENCES

Abou Ghanem-Sabanadzovic, N., Sabanadzovic, S., Gugerli, P., and Rowhani, A. 2012. Genome organization, serology and phylogeny of Grapevine leafroll-associated viruses 4 and 6: Taxonomic implications. Virus Research. 163:120–128.

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. 2008. Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome. Current Biology. 18:758–762.

Addo-Quaye, C., Miller, W., and Axtell, M. J. 2009. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics. 25:130–131.

Agindotan, B., and Perry, K. L. 2007a. Macroarray Detection of Plant RNA Viruses Using Randomly Primed and Amplified Complementary DNAs from Infected Plants. Phytopathology. 97:119–127.

Agindotan, B., and Perry, K. L. 2007b. Macroarray Detection of Plant RNA Viruses Using Randomly Primed and Amplified Complementary DNAs from Infected Plants. Phytopathology. 97:119–127.

Alabdullah, A., Minafra, A., Elbeaino, T., Saponari, M., Savino, V., and Martelli, G. P. 2010. Complete nucleotide sequence and genome organization of Olive latent virus 3, a new putative member of the family Tymoviridae. Virus Res. 152:10–18.

Alabi, O., Poojari, S., Sarver, K., Martin, R., and Naidu, R. 2013. Complete genome sequence analysis of an American isolate of Grapevine virus E. Virus genes. 46:563–566.

Albert, R. 2005. Scale-free networks in cell biology. Journal of Cell Science. 118:4947 LP-4957.

Bahder, B. W., Alabi, O., Poojari, S., Walsh, D. B., and Naidu, R. A. 2013. A survey for grapevine viruses in Washington State “Concord” (Vitis × labruscana L.) vineyards. Plant Health Progress. :PHP-2013-0805-01-RS.

Baldrich, P., Campo, S., Wu, M.-T., Liu, T.-T., Hsing, Y.-I. C., and Segundo, B. S. 2015. MicroRNA-mediated regulation of gene expression in the response of rice plants to fungal elicitors. RNA Biology. 12:847–863.

Bivalkar-Mehla, S., Vakharia, J., Mehla, R., Abreha, M., Kanwar, J. R., Tikoo, A., and Chauhan, A. 2011. Viral RNA silencing suppressors (RSS): Novel strategy of viruses to ablate the host RNA interference (RNAi) defense system. Virus research. 155:1–9.

Borges, F., and Martienssen, R. A. 2015. The expanding world of small RNAs in plants. Nat Rev Mol Cell Biol. 16:727–741.

Boscia, D., Sabanadzovic, S., Savino, V., Kyriakopoulou, P. E., Martelli, G. P., and Lafortezza, R. 1994. A non-mechanically transmissible isometric virus associated with asteroid mosaic of the grapevine. Vitis. 33:101–102.

Bransom, K. L., Wallace, S. E., and Dreher, T. W. 1996. Identification of the Cleavage Site Recognized by the Turnip Yellow Mosaic Virus Protease. Virology. 217:404–406.

Burgyán, J., and Havelda, Z. 2011. Viral suppressors of RNA silencing. Trends in plant science. 16:265–272.

Caciagli, P., Piles, V. M., Marian, D., Vecchiati, M., Masenga, V., Mason, G., Falcioni, T., and Noris, E. 2009. Virion Stability Is Important for the Circulative Transmission of Tomato Yellow Leaf Curl Sardinia Virus by Bemisia tabaci, but Virion Access to Salivary Glands Does Not Guarantee Transmissibility. Journal of Virology. 83:5784–5795.

Chen, Y.-R., Zheng, Y., Liu, B., Zhong, S., Giovannoni, J., and Fei, Z. 2012. A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA ligase 1

adenylated adapters. Plant Methods. 8:41.
Cheng, C.-Y., Krishnakumar, V., Chan, A., Schobel, S., and Town, C. D. 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal. 89:789–804.
Coetzee, B., Freeborough, M.-J., Maree, H. J., Celton, J.-M., Rees, D. J., and Burger, J. T. 2010. Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. Virology. 400:157–163.
Coetzee, B., Maree, H., Stephan, D., Freeborough, M.-J., and Burger, J. 2010. The first complete nucleotide sequence of a grapevine virus E variant. Archives of Virology. 155:1357–1360.
Cora, D., Re, A., Caselle, M., and Bussolino, F. 2017. MicroRNA-mediated regulatory circuits: outlook and perspectives. Physical Biology. 14:45001.
Creasey, K. M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B. C., and Martienssen, R. A. 2014. miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. Nature. 508:411–415.
Croft, L., Szklarczyk, D., Jensen, L. J., and Gorodkin, J. 2012. Multiple independent analyses reveal only transcription factors as an enriched functional class associated with microRNAs. BMC Systems Biology. 6:90.
Cui, Q., Yu, Z., Pan, Y., Purisima, E. O., and Wang, E. 2007. MicroRNAs preferentially target the genes with high transcriptional regulation complexity. Biochemical and Biophysical Research Communications. 352:733–738.
Cumbo, F., Paci, P., Santoni, D., Paola, L., and Giuliani, A. 2014. GIANT: a cytoscape plugin for modular networks. PLoS One. 9:1–7.
Ding, J., Zhou, S., and Guan, J. 2012. Finding MicroRNA Targets in Plants: Current Status and Perspectives. Genomics, Proteomics & Bioinformatics. 10:264–275.
Drapek, C., Sparks, E. E., and Benfey, P. N. 2017. Uncovering Gene Regulatory Networks Controlling Plant Cell Differentiation. Trends in Genetics. 33:529–539.
Dreher, T. W. 1999. Functions of the 3'-Untranslated Regions of Positive Strand Rna Viral Genomes. Annual Review of Phytopathology. 37:151–174.
Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 32:1792–1797.
Edwards, M. C., and Weiland, J. J. 2014. Coat protein expression strategy of oat blue dwarf virus. Virology. 450–451:290–296.
Edwards, M. C., Weiland, J. J., Todd, J., and Stewart, L. R. 2015. Infectious Maize rayado fino virus from Cloned cDNA. Phytopathology. 105:833–839.
Elbeaino, T., Abou Kubaa, R., Digiaro, M., Minafra, A., and Martelli, G. P. 2011. The complete nucleotide sequence and genome organization of Fig cryptic virus, a novel bipartite dsRNA virus infecting fig, widely distributed in the Mediterranean basin. Virus Genes. 42:415–421.
Fan, X., Dong, Y., Zhang, Z., Ren, F., Hu, G., and Zhu, H. 2013. First report of Grapevine virus E from grapevines in China. Journal of Plant Pathology. 95:659–668.
Fei, Q., Xia, R., and Meyers, B. C. 2013. Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks. The Plant Cell Online. 25:2400–2415.
Friard, O., Re, A., Taverna, D., De Bortoli, M., and Corá, D. 2010. CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. BMC Bioinformatics. 11:435.
Gambino, G., Navarro, B., Torchetti, E. M., La Notte, P., Schneider, A., Mannini, F., and Di

Serio, F. 2014. Survey on viroids infecting grapevine in Italy: identification and characterization of Australian grapevine viroid and Grapevine yellow speckle viroid 2. European Journal of Plant Pathology. 140:199–205.
German, M. A., Luo, S., Schroth, G., Meyers, B. C., and Green, P. J. 2009. Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. Nat.Protocols. 4:356–362.
González-Morales, S. I., Chávez-Montes, R. A., Hayano-Kanashiro, C., Alejo-Jacuinde, G., Rico-Cambron, T. Y., de Folter, S., and Herrera-Estrella, L. 2016. Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana. Proceedings of the National Academy of Sciences . 113:E5232–E5241.
Goszczynski, D. E., and Jooste, A. E. C. 2002. The application of single-strand conformation polymorphism (SSCP) technique for the analysis of molecular heterogeneity of grapevine virus A. Vitis. 41:77–82.
Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology. 29:644–652.
Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.
Guo, Q., Qu, X., and Jin, W. 2015. PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades. Bioinformatics. 31:284–286.
Habili, N., and Randles, J. 2012. Major Yield loss in Shiraz Vines Infected with Australian Shiraz Disease Associated with Grapevine virus A. Extended Abstracts 17th Meeting of ICVG, Davis 2012.
Hewitt, W. B. 1954. Some virus and virus-like diseases of grapevines. Bulletin of the California Department of Agriculture. 43:47–64.
Hommay, G., Komar, V., Lemaire, O., and Herrbach, E. 2008. Grapevine virus A transmission by larvae of Parthenolecanium corni. European Journal of Plant Pathology. 121:185–188.
Hou, C., Lee, W., Chou, H., Chen, A., Chou, S., and Chen, H. 2016. Global Analysis of Truncated RNA Ends Reveals New Insights into Ribosome Stalling in Plants. The Plant Cell. 28:2398 LP-2416.
Howell, M. D., Fahlgren, N., Chapman, E. J., Cumbie, J. S., Sullivan, C. M., Givan, S. A., Kasschau, K. D., and Carrington, J. C. 2007. Genome-Wide Analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 Pathway in Arabidopsis Reveals Dependency on miRNA- and tasiRNA-Directed Targeting. The Plant Cell Online. 19:926–942.
Hu, Q., Hollunder, J., Niehl, A., KÃ¸rner, C. J., Gereige, D., Windels, D., Arnold, A., Kuiper, M., Vazquez, F., Pooggin, M., and Heinlein, M. 2011. Specific Impact of Tobamovirus Infection on the Arabidopsis Small RNA Profile. PLoS ONE. 6:19549.
Huang, J., Yang, M., and Zhang, X. 2016. The function of small RNAs in plant biotic stress response. Journal of Integrative Plant Biology. 58:312–327.
Izadpanah, K., Ping Zhang, Y., Daubert, S., Masumi, M., and Rowhani, A. 2002. Sequence of the Coat Protein Gene of Bermuda Grass Etched-line Virus, and of the adjacent `Marafibox' Motif. Virus Genes. 24:131–134.

Jacquemond, M. 2012. Cucumber Mosaic Virus. Viruses and Virus Diseases of Vegetables in the Mediterranean Basin. 84:439–504.
Jo, Y., Choi, H., Cho, J. K., Yoon, J. Y., Choi, S. K., and Cho, W. K. 2015. In silico approach to reveal viral populations in grapevine cultivar Tannat using transcriptome data. Sci Rep. 5:15841.
Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16:111–120.
Klaassen, V. A., Sim, S. T., Dangl, G. S., Osman, F., Rwahnih, M. Al, Rowhani, A., and Golino, D. A. 2011. Vitis californica and Vitis californica × Vitis vinifera Hybrids are Hosts for Grapevine leafroll-associated virus-2 and -3 and Grapevine virus A and B. Plant Disease. 95:657–665.
Komar, V., Vigne, E., Demangeat, G., and Fuchs, M. 2007. Beneficial Effect of Selective Virus Elimination on the Performance of Vitis vinifera cv. Chardonnay. American Journal of Enology and Viticulture. 58:202–210.
Komiya, R. 2017. Biogenesis of diverse plant phasiRNAs involves an miRNA-trigger and Dicer-processing. Journal of Plant Research. 130:17–23.
Kozomara, A., and Griffiths-Jones, S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research. 42:D68–D73.
Krenz, B., Thompson, J. R., Fuchs, M., and Perry, K. L. 2012. Complete Genome Sequence of a New Circular DNA Virus from Grapevine. Journal of virology. 86:7715.
Langmead, B., and Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. Nat Meth. 9:357–359.
Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
Lebas, B. S. M., Ochoa-Corona, F., Tang, Z. J., Thangavel, R., Elliott, D. R., and Alexander, B. J. R. 2007. First Report of Spinach latent virus in Tomato in New Zealand. Plant Disease. 91:228.
Leclerc, R. D. 2008. Survival of the sparsest: robust gene networks are parsimonious. Molecular Systems Biology. 4:213.
Lewsey, M. G., González, I., Kalinina, N. O., Palukaitis, P., Canto, T., and Carr, J. P. 2010. Symptom induction and RNA silencing suppression by the cucumber mosaic virus 2b protein. Plant Signaling & Behavior. 5:705–708.
Lewsey, M., Surette, M., Robertson, F. C., Ziebell, H., Choi, S. H., Ryu, K. H., Canto, T., Palukaitis, P., Payne, T., Walsh, J. A., and Carr, J. P. 2009. The Role of the Cucumber mosaic virus 2b Protein in Viral Movement and Symptom Induction. Molecular Plant-Microbe Interactions. 22:642–654.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics . 25:2078–2079.
Li, R., Gao, S., Hernandez, A. G., Wechter, W. P., Fei, Z., and Ling, K.-S. 2012. Deep Sequencing of Small RNAs in Tomato for Virus and Viroid Identification and Strain Differentiation. PLoS ONE. 7:e37127.
Liang, C., Liu, X., Sun, Y., Yiu, S.-M., and Lim, B. L. 2014. Global small RNA analysis in fast-growing Arabidopsis thaliana with elevated concentrations of ATP and sugars. BMC Genomics. 15:116.

Librado, P., and Rozas, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25:1451–1452.
Lin, C.-C., Chen, Y.-J., Chen, C.-Y., Oyang, Y.-J., Juan, H.-F., and Huang, H.-C. 2012. Crosstalk between transcription factors and microRNAs in human protein interaction network. BMC Systems Biology. 6:18.
Lin, M.-C., Tsai, H.-L., Lim, S.-L., Jeng, S.-T., and Wu, S.-H. 2017. Unraveling multifaceted contributions of small regulatory RNAs to photomorphogenic development in Arabidopsis. BMC Genomics. 18:559.
Liu, H., and Duffus, J. 1986. 1st report of Spinach latent virus in north america. Phytopathology. 76:1087.
Llave, C. 2010. Virus-derived small interfering RNAs at the core of plant–virus interactions. Trends in plant science. 15:701–707.
Maccheroni, W., Alegria, M. C., Greggio, C. C., Piazza, J. P., Kamla, R. F., Zacharias, P. R. A., Bar-Joseph, M., Kitajima, E. W., Assumpção, L. C., Camarotte, G., Cardozo, J., Casagrande, E. C., Ferrari, F., Franco, S. F., Giachetto, P. F., Girasol, A., Jordão, H., Silva, V. H. A., Souza, L. C. A., Aguilar-Vildoso, C. I., Zanca, A. S., Arruda, P., Kitajima, J. P., Reinach, F. C., Ferro, J. A., and da Silva, A. C. R. 2005. Identification and Genomic Characterization of a New Virus (Tymoviridae Family) Associated with Citrus Sudden Death Disease. Journal of Virology . 79:3028–3037.
MacLean, D., Elina, N., Havecker, E. R., Heimstaedt, S. B., Studholme, D. J., and Baulcombe, D. C. 2010. Evidence for Large Complex Networks of Plant Short Silencing RNAs. PLoS ONE. 5:e9901.
Maere, S., Heymans, K., and Kuiper, M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics. 21:3448–3449.
Le Maguet, J., Beuve, M., Herrbach, E., and Lemaire, O. 2012. Transmission of Six Ampeloviruses and Two Vitiviruses to Grapevine by Phenacoccus aceris. Phytopathology. 102:717–723.
Maliogka, V. I., Olmos, A., Pappi, P. G., Lotos, L., Efthimiou, K., Grammatikaki, G., Candresse, T., Katis, N. I., and Avgelis, A. D. 2015. A novel grapevine badnavirus is associated with the Roditis leaf discoloration disease. Virus Research. 203:47–55.
Marathe, R., Guan, Z., Anandalakshmi, R., Zhao, H., and Dinesh-Kumar, S. 2004. Study of Arabidopsis thalianaresistome in response to cucumber mosaic virus infection using whole genome microarray. Plant Molecular Biology. 55:501–520.
Maree, H. J., Almeida, R. P. P., Bester, R., Chooi, K. M., Cohen, D., Dolja, V. V, Fuchs, M. F., Golino, D. A., Jooste, A. E. C., Martelli, G. P., Naidu, R. A., Rowhani, A., Saldarelli, P., and Burger, J. 2013. Grapevine leafroll-associated virus 3. Frontiers in Microbiology. 4:82.
Martelli, G., and Boudon-Padieu, E. 2006. Directory of infectious diseases of grapevines and viroses and virus-like diseases of the grapevine: bibliographic report 1998-2004. In *Directory of infectious diseases of grapevines and viroses and virus-like diseases of the grapevine: bibliographic report 1998-2004*, Options Méditerranéennes : Série B. Etudes et Recherches, ed. E. Martelli, Giovanni P, Boudon-Padieu. Bari: CIHEAM, p. 500.
Martelli, G., Ghanem-Sabanadzovic, N., Agranovsky, A., Al Rwahnih, M., Dolja, V., Dovas, C., Fuchs, M., Gugerli, P., Hu, J., Jelkmann, W., Katis, N., Maliogka, V., Melzer, M., Menzel, W., Minafra, A., Rott, M., Rowhani, A., Sabanadzovic, S., and Saldarelli, P.

2012. Taxonomic revision of the family Closteroviridae with special reference to the grapevine leafroll-associated members of the genus Ampelovirus and the putative species unassigned to the family. Journal of Plant Pathology. 94:7–19.
Martelli, G. P. 2014. Directory of virus and virus-like diseases of the grapevine and their agents. Journal of Plant Pathology. 96:1–136.
Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis. 17:10–12.
Megraw, M., Cumbie, J. S., Ivanchenko, M. G., and Filichkin, S. A. 2016. Small Genetic Circuits and MicroRNAs: Big Players in Polymerase II Transcriptional Control in Plants. The Plant Cell. 28:286 LP-303.
Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D. F., and Wright, F. 2008. TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. Bioinformatics. 25:126–127.
Nakaune, R., Toda, S., Mochizuki, M., and Nakano, M. 2008. Identification and characterization of a new vitivirus from grapevine. Archives of Virology. 153:1827–1832.
Navarro, L., Jay, F., Nomura, K., He, S. Y., and Voinnet, O. 2008. Suppression of the MicroRNA Pathway by Bacterial Effector Proteins. Science. 321:964–967.
Nei, M., and Li, W. H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences. 76:5269–5273.
Nogueira, F., Sarkar, A. K., Chitwood, D. H., and Timmermans, M. C. P. 2006. Organ Polarity in Plants Is Specified through the Opposing Activity of Two Distinct Small Regulatory RNAs. Cold Spring Harbor symposia on quantitative biology. 71:157–164.
Obbard, D. J., Gordon, K. H. J., Buck, A. H., and Jiggins, F. M. 2009. The evolution of RNAi as a defence against viruses and transposable elements. Philosophical Transactions of the Royal Society B: Biological Sciences. 364:99–115.
Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V, and Grotewold, E. 2006. AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks. Plant Physiology. 140:818 LP-829.
Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. G. 2011. Using graph theory to analyze biological networks. BioData Mining. 4:10.
Peláez, P., and Sanchez, F. 2013. Small RNAs in plant defense responses during viral and bacterial interactions: similarities and differences. Frontiers in Plant Science. 4:343.
Perry, K., and Lu, X. 2010. A tospovirus new to North America: Virus detection and discovery through the use of a macroarray for viruses of solanaceous crops. Phytopathology. 100:S100.
Posada, D. 2003. Using MODELTEST and PAUP* to select a model of nucleotide substitution. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis. Chapter 6:Unit 6.5.
Preez, J., Stephan, D., Mawassi, M., and Burger, J. 2011. The grapevine-infecting vitiviruses, with particular reference to grapevine virus A. Archives of Virology. 156:1495–1503.
Qiao, Y., Liu, L., Xiong, Q., Flores, C., Wong, J., Shi, J., Wang, X., Liu, X., Xiang, Q., Jiang, S., Zhang, F., Wang, Y., Judelson, H. S., Chen, X., and Ma, W. 2013. Oomycete pathogens encode RNA silencing suppressors. Nat Genet. 45:330–333.
Rajeswaran, R., Aregger, M., Zvereva, A. S., Borah, B. K., Gubaeva, E. G., and Pooggin, M. M.

2012. Sequencing of RDR6-dependent double-stranded RNAs reveals novel features of plant siRNA biogenesis. Nucleic Acids Research. 40:6241–6254.
Ramsdell, D. C., Bird, G. W., Gillett, J. M., and Rose, L. M. 1983. Superimposed shallow and deep soil fumigation to control Xiphinema americanum and peach rosette mosaic virus reinfection in a Concord vineyard. Plant Disease. 67:625–627.
Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. 2002. Hierarchical Organization of Modularity in Metabolic Networks. Science. 297:1551 LP-1555.
Re, A., Corá, D., Taverna, D., and Caselle, M. 2009. Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human. Mol Biosyst. 5.
Rezaian, M. A., Krake, L. R., and Golino, D. A. 1992. Common identity of grapevine viroids from USA and Australia revealed by PCR analysis. Intervirology. 34:38–43.
Ronquist, F., van der Mark, P., and Huelsenbeck, J. P. 2009. Bayesian phylogenetic analysis using MRBAYES. In *Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing, 2nd Edition*, Ronquist, F Swedish Museum Nat Hist, Dept Entomol, Box 50007, SE-10405 Stockholm, Sweden Swedish Museum Nat Hist, Dept Entomol, Box 50007, SE-10405 Stockholm, Sweden Swedish Museum Nat Hist, Dept Entomol, SE-10405 Stockholm, Sweden Florida State Univ, Sch, p. 210–266.
Rosa, C., Jimenez, J. F., Margaria, P., and Rowhani, A. 2011. Symptomatology and Effects of Viruses Associated with Rugose Wood Complex on the Growth of Four Different Rootstocks. American Journal of Enology and Viticulture. 62:207–213.
Al Rwahnih, M., Daubert, S., Golino, D., Islas, christina marie, and Rowhani, A. 2015. Comparison of Next Generation Sequencing vs. Biological Indexing for the optimal detection of viral pathogens in Grapevine. Phytopathology. 105:758–763.
Al Rwahnih, M., Daubert, S., Úrbez-Torres, J., Cordero, F., and Rowhani, A. 2011. Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. Archives of Virology. 156:397–403.
Sabanadzovic, S., Abou-Ghanem, N., Castellano, M. A., Digiaro, M., and Martelli, G. P. 2000. Grapevine fleck virus-like viruses in Vitis. Archives of Virology. 145:553–565.
Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research . 13:2498–2504.
Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. 1996. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal Chem. 68:850–858.
Shukla, L. I., Chinnusamy, V., and Sunkar, R. 2008. The role of microRNAs and other endogenous small RNAs in plant stress responses. Biochim Biophys Acta. 1779:743–748.
Soto-Suárez, M., Baldrich, P., Weigel, D., Rubio-Somoza, I., and San Segundo, B. 2017. The Arabidopsis miR396 mediates pathogen-associated molecular pattern-triggered immune responses against fungal pathogens. Scientific Reports. 7:44898.
Soule, M. J., Eastwell, K. C., and Naidu, R. A. 2006. First Report of Grapevine leafroll associated virus-3 in American Vitis spp. Grapevines in Washington State. Plant Disease. 90:1461.
Steuer, R., and López, G. Z. 2007. Global Network Properties. In *Analysis of Biological Networks*, John Wiley & Sons, Inc., p. 29–63.
Stumpf, M. P. H., and Wiuf, C. 2009. Front Matter. In *Statistical and Evolutionary Analysis of Biological Networks*, Imperial college press, p. i–viii.
Sun, W., Julie Li, Y.-S., Huang, H.-D., Shyy, J. Y.-J., and Chien, S. 2010. microRNA: A Master

Regulator of Cellular Processes for Bioengineering Systems. Annual Review of Biomedical Engineering. 12:1–27.
Sunkar, R., Chinnusamy, V., Zhu, J., and Zhu, J. K. 2007. Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. Trends Plant Sci. 12:301–309.
Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic acids research. 36:D1009-14.
Szittya, G., and Burgyán, J. 2013. RNA Interference-Mediated Intrinsic Antiviral Immunity in Plants BT - Intrinsic Immunity. In ed. Bryan R Cullen. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 153–181.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.
Taliansky, M., Kim, S. H., Mayo, M. A., Kalinina, N. O., Fraser, G., McGeachy, K. D., and Barker, H. 2004. Escape of a plant virus from amplicon-mediated RNA silencing is associated with biotic or abiotic stress. Plant Journal. 39:194–205.
Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T. W., Gaudinier, A., Young, N. F., Trabucco, G. M., Veling, M. T., Lamothe, R., Handakumbura, P. P., Xiong, G., Wang, C., Corwin, J., Tsoukalas, A., Zhang, L., Ware, D., Pauly, M., Kliebenstein, D. J., Dehesh, K., Tagkopoulos, I., Breton, G., Pruneda-Paz, J. L., Ahnert, S. E., Kay, S. A., Hazen, S. P., and Brady, S. M. 2015. An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature. 517:571–575.
Thatcher, S. R., Burd, S., Wright, C., Lers, A., and Green, P. J. 2015. Differential expression of miRNAs and their target genes in senescing leaves and siliques: insights from deep sequencing of small RNAs and cleaved target RNAs. Plant, Cell & Environment. 38:188–200.
Thompson, J. R., Fuchs, M., McLane, H., Celebi-Toprak, F., Fischer, K. F., Potter, J. L., and Perry, K. L. 2014. Profiling Viral Infections in Grapevine Using a Randomly Primed Reverse Transcription-Polymerase Chain Reaction/Macroarray Multiplex Platform. Phytopathology. 104:211–219.
Thompson, J. R., Leone, G., Lindner, J. L., Jelkmann, W., and Schoen, C. D. 2002. Characterization and complete nucleotide sequence of Strawberry mottle virus: a tentative member of a new family of bipartite plant picorna-like viruses. Journal of General Virology. 83:229–239.
Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., Taly, J. F., and Notredame, C. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Research. 39:W13–W17.
Untiveros, M., Perez-Egusquiza, Z., and Clover, G. 2010. PCR assays for the detection of members of the genus Ilarvirus and family Bromoviridae. Journal of virological methods. 165:97–104.
Uyemoto, J. K., Taschenberg, E. F., and Hummer, D. K. 1977. Isolation and identification of a strain of grapevine Bulgarian latent virus in Concord grapevine in New York State. Plant Disease Reporter. 61:949–953.
Vargas-Asencio, J., McLane, H., Bush, E., and Perry, K. L. 2013. Spinach latent virus Infecting Tomato in Virginia, United States. Plant Disease. 97:1663.

Vargas-Asencio, J., Perry, K. L., Wise, A., and Fuchs, M. 2016. Detection of Australian grapevine viroid in Vitis vinifera in New York. Plant Disease. 101:848.
Vargas-Asencio, J., Al Rwahnih, M., Rowhani, A., Celebi-Toprak, F., Thompson, J. R., Fuchs, M., and Perry, K. L. 2015. Limited Genetic Variability Among American Isolates of Grapevine virus E from Vitis spp. Plant Disease. 100:159–163.
Vargas-Asencio, J., Wojciechowska, K., Baskerville, M., Gomez, A. L., Perry, K. L., and Thompson, J. R. 2017. The complete nucleotide sequence and genomic characterization of grapevine asteroid mosaic associated virus. Virus Research. 227:82–87.
Voorhoeve, P. M. 2010. MicroRNAs: Oncogenes, tumor suppressors or master regulators of cancer heterogeneity? Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 1805:72–86.
Walczak, A.M, and Tkačik, G. 2011. Information transmission in genetic regulatory networks: a review. Journal of Physics: Condensed Matter. 23:153102.
Wang, H.-L. V, and Chekanova, J. A. 2016. Small RNAs: essential regulators of gene expression and defenses against environmental stresses in plants. Wiley Interdisciplinary Reviews: RNA. 7:356–381.
Whitham, S. A., Quan, S., Chang, H.-S., Cooper, B., Estes, B., Zhu, T., Wang, X., and Hou, Y.-M. 2003. Diverse RNA viruses elicit the expression of common sets of genes in susceptible Arabidopsis thaliana plants. The Plant Journal. 33:271–283.
Wilcox, W. F., Gubler, W. D., and Uyemoto, J. K. 2016. PART I: Diseases Caused by Biotic Factors. In *Compendium of Grape Diseases, Disorders, and Pests, Second Edition*, Diseases and Pests Compendium Series, eds. Jerry K Uyemoto, Wayne F Wilcox, and Walter D Gubler. The American Phytopathological Society, p. 17–146.
Wu, C., Li, X., Guo, S., and Wong, S.-M. 2016. Analyses of RNA-Seq and sRNA-Seq data reveal a complex network of anti-viral defense in TCV-infected Arabidopsis thaliana. 6:36007.
Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E. C., Li, W.-X., and Ding, S.-W. 2010. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proceedings of the National Academy of Sciences. 107:1606–1611.
Xiao, H., and Meng, B. 2016. First report of Grapevine asteroid mosaic-associated virus and Grapevine rupestris vein feathering virus in grapevines in Canada. Plant Disease. 100:2175.
Xin, M., Wang, Y., Yao, Y., Xie, C., Peng, H., Ni, Z., and Sun, Q. 2010. Diverse set of microRNAs are responsive to powdery mildew infection and heat stress in wheat (Triticum aestivum L.). BMC Plant Biology. 10:123.
Zhai, J., Arikit, S., Simon, S. A., Kingham, B. F., and Meyers, B. C. 2014. Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. Methods. 67:84–90.
Zhai, J., Jeong, D.-H., De Paoli, E., Park, S., Rosen, B. D., Li, Y., González, A. J., Yan, Z., Kitto, S. L., Grusak, M. A., Jackson, S. A., Stacey, G., Cook, D. R., Green, P. J., Sherrier, D. J., and Meyers, B. C. 2011. MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Genes & development. 25:2540–2553.
Zhang, C., Wu, Z., Li, Y., and Wu, J. 2015. Biogenesis, Function, and Applications of Virus-Derived Small RNAs in Plants. Frontiers in Microbiology. 6:1237.
Zheng, Y., Gao, S., Padmanabhan, C., Li, R., Galvez, M., Gutierrez, D., Fuentes, S., Ling, K.-S.,

Kreuze, J., and Fei, Z. 2017. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology. 500:130–138.
Zhu, X., Gerstein, M., and Snyder, M. 2007. Getting connected: analysis and principles of biological networks. Genes & Development. 21:1010–1024.