Use of NGS for discovery of viruses in the soybean aphid

Ying Feng
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Bioinformatics Commons, and the Virology Commons

Recommended Citation
Feng, Ying, "Use of NGS for discovery of viruses in the soybean aphid" (2015). Graduate Theses and Dissertations. 14334.
http://lib.dr.iastate.edu/etd/14334

This Thesis is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Use of NGS for discovery of viruses in the soybean aphid

by

Ying Feng

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Wyatt Allen Miller, Co-Major Professor
Karin Dorman, Co-Major Professor
Bryony C. Bonning

Iowa State University
Ames, Iowa
2015

Copyright © Ying Feng, 2015. All rights reserved.
# TABLE OF CONTENTS

| CHAPTER 1  | GENERAL INTRODUCTION | 1 |
| CHAPTER 2  | NOVEL AND KNOWN FULL GENOME VIRUS DETECTION FROM DIFFERENT REGIONS COLLECTED SOYBEAN APHIDS BY DEEP SEQUENCING | 32 |
| Abstract | 33 |
| Introduction | 34 |
| Materials and Methods | 35 |
| Results | 38 |
| Discussion | 43 |
| References | 47 |
| Tables and Figures | 52 |
| CHAPTER 3  | GENERAL CONCLUSIONS | 59 |
| ACKNOWLEDGEMENTS | 60 |
CHAPTER 1. GENERAL INTRODUCTION

Soybean Aphids

The soybean aphid, *Aphis glycines*, has become one of the most economically important pest insects in North America [3]. This invasive species is native to Asia, and was first detected in North America in 2000 in Wisconsin. Subsequently, it has spread rapidly throughout the United States and southeastern Canada. By 2009, the aphid had infested 30 states and three provinces of Canada (Fig. 1) [4].

![Map of soybean aphid distribution](image)

**Fig. 1** Distribution of the soybean aphid in North America from 2000 to 2009. Red indicates the initial 10 states in which soybean aphid was detected. Yellow indicates the distribution of soybean aphid from 2001 to 2009 (from [4]).
Aphids are a serious problem for agriculture, especially in temperate regions. Soybean aphid has a temperature-dependent growth rate [5,6]. The optimal development temperature of soybean aphid is 28 °C, and as temperatures approach 35°C the growth rate exhibits a rapid decline[5]. This decline partially explains why the range of the soybean aphids has not extended to the southern most region of North America.

Aphids damage their host plants in several ways. Firstly, they plunder the nutrients necessary for plant growth and reproduction for their own profit. Secondly, they inject their phytotoxic saliva during the feeding phase, causing early leaf senescence and reducing pod set, seed quality, and yield [7,8]. Thirdly, aphids are the main vectors, nearly 50% (275 out of 600)[9,10] of insect-borne viruses. Soybean aphid is also a competent vector of many plant viruses. Soybean aphids were partially responsible for widespread virus epidemics observed in snap bean [11] and is a new vector of Potato Virus Y, leading to reduce of potato production [12]. Lastly, aphids produce honeydew, which promotes growth of sooty mould, which in turn hinders the photosynthetic activity of plants, further reducing yield [13].

Soybean is one of the most economically important crops in United State. The USDA Crop Production reports total soybean production in 2010 at a record level of 32 M ha with a production value in excess of US$27 billion. In the US, the main soybean growing districts are the 12 north central states, which cover over 80% of the soybean grown [4]. In these states, soybean aphids represent the first insect pest to consistently cause significant yield losses over wide areas, with yield decreases as high as 40% [14,15]. As a result, pest management practices have changed greatly, and the soybean producers must budget for scouting and insecticidal control of soybean aphid.
Currently, application of chemical insecticides is the main method for management of soybean aphids. For example, in Iowa, insecticides use went from zero application in 1996, to nearly 4 million acres treated in 2003. Even in the absence of harmful soybean aphid populations, farmers often use systemic insecticides or even aerial spraying to control the potential aphid outbreaks [16]. However, aphids can rapidly re-infest soybean fields after treatment, and readily develop resistance to insecticides [17]. Moreover, pesticides can harm beneficial non-target organisms [18]. For example, widely used neonicotinoid insecticides appear to be decimating the honeybee population, contributing to 66% annual losses of honeybees in Iowa in 2010-2011 [19]. Therefore, alternatives to chemical pesticides are necessary to control the soybean aphid.

Breeding cultivars resistant to aphid and/or viruses is an alternative control method. Genetic resistance can confer an effective protection without additional costs or labor for farmers during the growing season, and is also safe for both the environment and human health [13]. Three mechanisms of resistance to aphids have been described: (i) in antibiosis, survival and fecundity of the pest are affected by feeding on the plant; (ii) in antixenosis, the plant is rapidly recognized as a poor host by the pest that moves away; (iii) in tolerance, the plant is infested by pest, but is less affected than susceptible ones [20]. Several soybean varieties have been identified as resistant to soybean aphid and present mechanisms of resistance of antibiosis, antixenosis, and tolerance. Rag1 (Resistance to Aphis glycines gene 1) [21-23] and Rag2 (Resistance to Aphis glycines gene 2) [24] were discovered as resistant genes in soybean to control aphid feeding.

RNA silencing has been recognized as a potential approach for management of pea aphid (Acyrthosiphon pisum) [25-28], grain aphid (Sitobion avenae) [29,30], and
green-peach aphid (*Myzus persicae*) [31-33]. Multiple approaches have been used for introduction of exogenous silencing RNA into aphids, for example, feeding on treated artificial diet or on transgenic plants that expresses silencing RNAs.

**RNA viruses of Insect**

With the availability of complete genomic sequences more and more insect viruses have been discovered. Over the last 90 years, 16 families of insect viruses have been isolated and described [34]. Classification has been based increasingly on the phylogenetic/evolutionary history of each of the main virus genes. Viruses generally have been classified into four groups based on their genome, dsDNA virus, ssDNA virus, dsRNA and ssRNA viruses. Some properties of the major RNA viruses are listed in Table 1.

**Small RNA viruses**

Small RNA viruses include viruses of < 40 nm in diameter with RNA genomes [34,35]. The dicistroviruses, iflaviruses, caliciviruses, nodaviruses and tetraviruses all are small RNA viruses [35]. They have minimized the number of different components, e.g. genome segments, genes, and proteins, required for self-propagation. Essentially, small RNA viruses are the product of two genes, one for replication enzyme enabling genome amplification, the other for a structural or capsid protein enabling horizontal transmission between hosts and cells. Genomes generally comprise one or two segments of a single stranded RNA, of less than 10 kb and encoding up to four (generally one or three) ORFs [35].
**Table 1.** Properties of the RNA insect viruses (adapted from [34])

<table>
<thead>
<tr>
<th>Family</th>
<th>Insect-infecting genus</th>
<th>Genome</th>
<th>Virion Shape and diameter</th>
<th>Virion envelop</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reoviridae</strong></td>
<td><em>Cypovirus</em></td>
<td>dsRNA 10 segments</td>
<td>Icosahedral 80 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Reoviridae</strong></td>
<td><em>Idnoreovirus</em></td>
<td>dsRNA 10 segments</td>
<td>Icosahedral 80 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Birnaviridae</strong></td>
<td><em>Entomobirnavirus</em></td>
<td>dsRNA 2 segments</td>
<td>Icosahedral 60 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Rhabdoviridae</strong></td>
<td><em>Vesiculovirus</em></td>
<td>ssRNA 11 – 15 kb</td>
<td>Bullet-shaped 45 – 100 nm</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Dicistroviridae</strong></td>
<td><em>Cripavirus</em></td>
<td>ssRNA 9 – 11 kb</td>
<td>Icosahedral 30 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Iflaviridae</strong></td>
<td><em>Iflavirus</em></td>
<td>ssRNA 8.5 – 9.5 kb</td>
<td>Icosahedral 30 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Caliciviridae</strong></td>
<td><em>Lagovirus</em></td>
<td>ssRNA 7.5 – 8.5 kb</td>
<td>Icosahedral 35 – 40 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Tetraviridae</strong></td>
<td><em>Betatetravirus</em></td>
<td>ssRNA 1 – 2 segments</td>
<td>Icosahedral 40 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Nodaviridae</strong></td>
<td><em>Alphanodavirus</em></td>
<td>ssRNA 2 segments 4.5kb</td>
<td>Icosahedral 32 – 33 nm</td>
<td>No</td>
</tr>
<tr>
<td><strong>Metaviridae</strong></td>
<td><em>Metavirus</em></td>
<td>ssRNA 4 – 10kb</td>
<td>Irregular approximately 100 nm</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Pseudoviridae</strong></td>
<td><em>Pseudovirus</em></td>
<td>ssRNA 5 – 10kb</td>
<td>Irregular; round to ovoid 30 – 40 nm</td>
<td>No</td>
</tr>
</tbody>
</table>
Fig. 2 Summary of genome organization of dicistovirus, picornavirus, iflavirus, betateravirus, and omegatetravirus. Open reading frames are indicated by orange, green, and yellow boxes, with the functional proteins of the polyproteins indicated. The box circle denotes VPg, and red and light blue boxes denote IREs. Nonstructural proteins are indicated by orange box, and green boxes indicate structural proteins. Abbreviation: SS, silencing suppressor domain; hel, helicase domains pro, chymotrysin-like cysteine protease; RdRp, RNA-dependent RNA polymerase; VP, virion protein; VCAP, the capsid protein precursor; Mtr, methyltransferase.
**Dicistroviruses**

The dicistroviruses are characterized by having a positive single-stranded sense RNA genome of between approximately 9 and 11 kb encoding two ORFs, icosahedral particles and capsid comprised of three major proteins of approximately 30 kDa [35]. In many respects dicistroviruses have a lot of characteristics of the picornaviruses and iflaviruses (Fig. 2a-c) [34], but subtle differences in their genomic organization, and the positions of structural proteins are reversed. The best-known dicistrovirus is Cricket paralysis virus (CrPV). It has been isolated from insects from five different orders and is probably the insect virus species with the broadest natural host range [36]. Aphid-lethal paralysis virus (ALPV) is also known to negatively impact agriculture insect pests [37]. CrPV and ALPV are the viruses with acute disease symptoms [38], but Drosophila C virus causes chronic disease and often no symptoms [39].

The genome organization of dicistroviruses is shown in Fig. 2a [35,40]. The genome has two major open reading frames (ORF). The 5’ end of the genome codes for the non-structural polyprotein and 3’ end of the genome encodes for the structural protein. The genome also has two internal ribosomal entry sites (IRES), one at the 5’ end of the genome and the other between ORF1 and ORF2 [38,40].

**Tetraviruses**

The tetraviruses are characterised by the T=4 symmetry particle architecture. The protein coat that surrounds the infective nucleic acid in virus particles is arranged to form an icosahedron with 12 pentameric and 30 hexameric capsomere subunits [35]. The family comprises two genera based on their particle morphology and genome
organization: *Betatetravirus* (type species *Nudaurelia b* virus) and *Omegatetravirus* (type species *Nudaurelia o* virus) [41]. The betatetraviruses have positive sense monopartite ssRNA genomes of approximately 6.5 kb, while members of *Omegatetravirus* have bipartite genomes of approximately 7.5 kb. For the betatetraviruses, genomic sequences are available for the *Nudaurelia b* virus (NβV) [42], Thosea asigna virus [43], and *Eusprosterna elaeasa* virus [44], and providence virus (walter, 2008). The genome organizations of among these viruses are not consistent (Fig. 2d) [34,45]. A virus obtained from the cotton bollworm, *Helicoverpa armigera*, the *Helicoverpa armigera* stunt virus (HaSV) [46-48] and *Dendrolimus punctatus* virus [49] are the omegatetraviruses for which complete genomic sequences are available (Fig. 2e) [35,45].

**Viruses of aphids**

Several viruses that infect different host aphids have been discovered (Table 2). *Rhopalosiphum padi* virus (RhPV) and Aphid-lethal paralysis virus (ALPV) are two dicistroviruses that infect aphids. RhPV was firstly isolated from the bird cherry-oat aphid (*Rhopalosiphum padi*) [50]. It circulates in the plant phloem and is transmitted horizontally from plants to uninfected aphids [51]. ALPV was first isolated from the bird cherry-oat aphid (*R. padi*) after observation of infected aphids moving away from the food source and death induced by paralysis [52]. Now several ALPV-like viruses have been identified in organisms other than aphids, such as from bat fecal samples and honeybees [52-57]. ALPV is transmitted vertically in the bird cherry-oat aphid and the English grain aphid (*Sitobion avenae*). By nucleic acid in situ hybridization, viral RNA was shown in the gut, brain and embryonic tissues in the bird cherry-oat aphid [58].
<table>
<thead>
<tr>
<th>Virus</th>
<th>Family</th>
<th>Host aphid</th>
<th>Genome</th>
<th>Virion structure</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Acyrthosiphon pisum</em></td>
<td>Paroviridae</td>
<td>Pea aphid (<em>Acyrthosiphon pisum</em>)</td>
<td>ssDNA</td>
<td>Icosahedral</td>
<td>[59]</td>
</tr>
<tr>
<td><em>Acyrthosiphon pisum</em> virus</td>
<td>Unclassified</td>
<td>Pea aphid (<em>Acyrthosiphon pisum</em>)</td>
<td>+ ssRNA</td>
<td>Icosahedral 31nm</td>
<td>[60]</td>
</tr>
<tr>
<td><em>Aphid-lethal paralysis</em></td>
<td>Dicistroviridae</td>
<td>Bird-cherry oat aphid (<em>Rhopalosiphum padi</em>)</td>
<td>+ ssRNA</td>
<td>Icosahedral 27nm</td>
<td>[52]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Milkweed aphid (<em>Aphis nerii</em>)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pea aphid (<em>Acyrthosiphon pisum</em>)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Brevicoryne brassicae</em></td>
<td>Iflaviridae</td>
<td>Cabbage aphid (<em>Brevicoryne brassicae</em>)</td>
<td>+ ssRNA</td>
<td>Icosahedral</td>
<td>[61]</td>
</tr>
<tr>
<td><em>Dysaphis plantaginea</em></td>
<td>Paroviridae</td>
<td>Rosy apple aphid (<em>Dysaphis plantaginea</em>)</td>
<td>ssDNA</td>
<td>Icosahedral 22nm</td>
<td>[59]</td>
</tr>
<tr>
<td><em>Myzus persicae</em></td>
<td>Paroviridae</td>
<td>Green peach aphid (<em>Myzus persicae</em>)</td>
<td>ssDNA</td>
<td>Icosahedral 20nm</td>
<td>[62]</td>
</tr>
<tr>
<td><em>Rhopalosiphum padi</em> virus</td>
<td>Dicistroviridae</td>
<td>Bird cherry-oat aphid (<em>Rhopalosiphum padi</em>)</td>
<td>+ ssRNA</td>
<td>Icosahedral</td>
<td>[63]</td>
</tr>
<tr>
<td><em>Rosy apple aphid</em> virus</td>
<td>Unclassified</td>
<td>Rosy apple aphid (<em>Dysaphis plantaginea</em>)</td>
<td>+ ssRNA</td>
<td>Icosahedral 32nm</td>
<td>[59]</td>
</tr>
</tbody>
</table>
The Acyrthosiphon pisum virus (APV) is a positive strand RNA virus with a genome of approximately 10 kb. The virus has two major ORFs and the virus encodes four capsid proteins ranging from 23 kDa to 66 kDa. APV was abundant in the epithelium and lumen of the digestive tract of three day-old nymphs. Infection with the virus appears to inhibit the growth of aphids [60].

The Brevicoryne brassicae virus (BrBV), the first iflavirus described in aphids, was isolated from the cabbage aphid (Brevicoryne brassicae). The virus has a positive ssRNA genome of 10.1 kb with a 3’ poly (A) tail. The virus is closely related to other insect iflaviruses. BrBV is not likely to be transmitted via the plant because no virus transcripts were detected in plants infected with BrBV [61].

The Rosy apple aphid virus (RAAV), a positive strand RNA virus, was isolated from the rosy apple aphid (Dysaphis plantaginea). It has a genome organization and a high amino acid sequence identity to APV. RAAV was present in plant leaves previously exposed to RAAV-infected aphids and therefore is may be horizontally transmitted via the plant phloem [59].

The Dysaphis plantaginea densovirus (DplDNV) is a densovirus isolated from the rosy apple aphid (Dysaphis plantaginea). DplDNV is a single-stranded DNA virus with a genome of approximately 5 kb. It is transmitted horizontally via the leaf and vertically from adults to nymphs. DplDNV infection induces winged morphs of the host aphid thereby facilitating virus dispersal. Infection with DplDNV negatively affects aphid fecundity.
Another densovirus was identified from the green peach aphid (*Myzus persicae*), named *Myzus persicae* densivirus. The virus has a genome of approximately 5.7 kb with five ORFs. The virus infects the stomach of infected aphids and transmission occurs horizontally (through plant and honeydew) and vertically from mother to offspring [60]. An additional *densovirus* was also identified from expressed sequenced tag from the pea aphid, which is a putative *Acyrthosiphon pisum* densovirus. No further characterization of the virus has been published [59].

**Plant viruses transmitted by aphids**

According to current knowledge on viral transmission, nearly 6% of the plant viral species are transmitted directly through seeds [64-66]. However, plant viruses most often use a partner, known as a vector, for efficient transmission to new hosts. Insects are the most common of the vectors, and among these, the aphid has been described as an efficient vector for up to 50% of the insect-vectored plant viruses [9,67]. The reasons are: (1) aphid is able to gain access to the plant cell cytoplasm by breaking through the cell wall and membrane using their feeding organs, and moves frequently among plants; (2) prior to feeding on the phloem of a host plant, aphids have the extraordinary capacity to sample cell content without killing the plants [68,69]; (3) During the steps of the feeding process, aphids produce different salivas with different compositions and functions [70]; (4) aphids can acquire viruses at all steps of the feeding process.

Generally, three vector-transmission modes have been described: non-persistent, persistent, and intermediate semi-persistent transmission [71]. In non-persistent transmission, the virus is acquired on an infected plant within only few a seconds or
minutes because it can hold an infectious form within the vector for a few minutes [72]. In contrast, in persistent transmission, there is no limit for acquisition and inoculation times, and latent period [72]. Intermediate semi-persistent viruses, with acquisition and inoculation periods of several minutes to hours and no requirement for a latent period prior to inoculation [73]. Further, a new classification of viruses was proposed [74], where the term non-circulative grouped the earlier non- and semi-persistent mode, and the term circulative replaced persistent, and included both virus species that replicate in insect cells (propagative), and those that do not (non-propagative). Two distinct viral strategies were found in non-circulative transmission: the capsid strategy and the helper strategy [75]. In the capsid strategy, the virus interacts directly with the vector via its coat protein (CP), whereas in the helper strategy the virus-vector interaction is mediated by an additional virus encoded non-structural protein, generally designated as helper.

Table 3 summarized aphid-transmitted plant viruses including the information of virus family, genus and the transmission mode [71,76]. The genera Carlavirus, Alfamovirus and Cucumovirus have been demonstrated having CP to be the only viral component determining virion retention within the aphid vector [77,78]. Viruses belonging to the genera Potyvirus and Caulimovirus were characterized producing a “helper component” for their aphid transmission [75,79,80]. Cauliflower mosaic virus, the type of member of the genus Caulimovirus, has adopted a helper-dependent transmission strategy, but this strategy evolved with both nonpersistent and semipersistent transmission [81,82]. Parsnip yellow fleck virus (PYFV) also employs a helper strategy in its transmission. Unlike in the potyvirus and caulimoviruses, PFYV
Table 3. Aphid-transmitted plant viruses from different families with different mode of transmission

<table>
<thead>
<tr>
<th>Virus</th>
<th>Family</th>
<th>Genus</th>
<th>Noncirculative</th>
<th>Circulative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nonpersistent</td>
<td>Semipersistent</td>
</tr>
<tr>
<td>Carnation latent virus</td>
<td>Betaflexiviridae</td>
<td>Carlavirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Alfalfa mosaic virus</td>
<td>Bromoviridae</td>
<td>Alfamovirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Cucumber mosaic virus</td>
<td>Caulimoviridae</td>
<td>Caulimovirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Cauliflower mosaic virus</td>
<td>Fabaviridae</td>
<td>Fabavirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Beet yellows virus</td>
<td>Comoviridae</td>
<td>Fabavirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Barley yellow dwarf virus</td>
<td>Luteoviridae</td>
<td>Luteovirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Carrot mottle virus</td>
<td>Luteoviridae</td>
<td>Umbravirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Pea enation mosaic virus-1</td>
<td>Luteoviridae</td>
<td>Enamovirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Potato leaf roll virus</td>
<td>Luteoviridae</td>
<td>Polerovirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Banana bunchy top virus</td>
<td>Nanoviridae</td>
<td>Babuvirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Maclura mosaic virus</td>
<td>Potyviridae</td>
<td>Macluravirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Potato virus Y</td>
<td>Potyviridae</td>
<td>Potyvirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Bean yellow mosaic virus</td>
<td>Potyviridae</td>
<td>Potyvirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Turnip mosaic virus</td>
<td>Potyviridae</td>
<td>Potyvirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Celery mosaic virus</td>
<td>Potyviridae</td>
<td>Potyvirus</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Lettuce necrotic yellows virus</td>
<td>Rhabdoviridae</td>
<td>Cytorhabdovirus</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Sonchus yellow net virus</td>
<td>Rhabdoviridae</td>
<td>Nucleorhabdovirus</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Black raspberry necrosis virus</td>
<td>Secoviridae</td>
<td>Unassigned</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Strawberry mottle virus</td>
<td>Secoviridae</td>
<td>Unassigned</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Anthriscus yellows virus</td>
<td>Sequiviridae</td>
<td>Waikaviruses</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Parsnip yellow fleck virus</td>
<td>Sequiviridae</td>
<td>Sequivirus</td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>
itself does not encode the helper [83], but the nature of the helper factor is not known.

Depending on whether a virus replicates in the vectoring aphid, transmission is divided into ‘circulative, non-propagative’ or ‘circulative, propagative’. A three genera in the family Luteoviridae are circulative non-propagatively transmitted viruses including Barley yellow dwarf virus (BYDV) and related luteoviruses [84]. Moreover, vector specificity is a prominent feature of luteovirus transmission.

The circulative propagative plant viruses are similar to the circulative non-propagative viruses in many aspects of their acquisition and path of movement in the aphid. The main difference is that the infected association remains permanent for the rest of the life of the insect, and virus will be transmitted to offspring. Only the family Rhabdoviridae belongs to the circulative propagative aphid-transmitted plant virus. This family includes viruses that replicate in either mammals and insects or plant and insects [85]. However, virus encoded determinants for transmission have not been identified in plant Rhabdoviridae. It is likely that G protein may play the role in the aphid gut endocytosis process, but is still hypothetical and is based on gene function conservation among members of the vertebrate rhabdovirus [86,87]. Clearly, even in the best-characterized examples, many details are still missing.

**Next Generation Sequencing in virology**

In general, we can check the size and growth of international nucleotide databases to realize the importance to sequencing to science. By August 2010 in the published release of GenBank, there were some 970 million and 43 million bases of viral and phage origin respectively, representing an annual growth of 20-24% [88], a growth rate highly
above average for the database as a whole. The information generated has wide-ranging impact on all areas of virology, from diagnosis to pathogenesis, and from vaccine design, bio-control to viral evolution and ecology.

The new methods of sequencing have now been developed, including next generation sequencing (NGS) which has entirely revolutionized our ability to sequence. It is now possible to generate many millions of bases of sequence in one day and this bring many new opportunities accessible to us. There are an increasing number of NGS technologies in the market, all using slightly different methodologies to achieve clonal amplification and sequencing. Table 4 lists the summary of current methods for NGS including methods in template preparation, adapter type, sequencing length and disadvantage and advantage for each method [1,2].

Deciding which technology is best will depend on the specific experiment being planned. Several important factors should be considered, including the size of the genome, its complexity, as well as the depth of coverage and accuracy required. There are some general principles to guide making decision. For those looking to assemble complex genomes \textit{de novo}, longer read lengths may be appropriate. 454 sequencing (Roche), although relatively expensive, still has a niche in amplicon sequencing because of its longer reads. Helicos is appropriate for direct RNA sequencing, and PacBio for very long reads. However, accuracy may become an important issue because both are single-molecule sequencers. In general, the Illumina and SOLiD platforms currently offer the best all-round value for money, accuracy and throughput for RNA-seq, and those projects requiring high depths of coverage [1,2].
Table 4. Summary of current NGS technology (adapted from [1,2])

<table>
<thead>
<tr>
<th>Method</th>
<th>Template preparation</th>
<th>Adapter type</th>
<th>Sequencing reaction</th>
<th>Avg. reads length (bases)</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454</td>
<td>Emulsion PCR</td>
<td>Linear</td>
<td>Synthesis</td>
<td>330</td>
<td>Longer reads improve mapping in repetitive regions; fast run time</td>
<td>High reagent cost; high error rates in homo-polymer repeats</td>
</tr>
<tr>
<td>SOLiD</td>
<td>Emulsion PCR</td>
<td>Linear</td>
<td>Ligation</td>
<td>50</td>
<td>Two-base encoding provides inherent error correction</td>
<td>Long run times</td>
</tr>
<tr>
<td>Illumina</td>
<td>Bridge PCR</td>
<td>Linear</td>
<td>Synthesis</td>
<td>100</td>
<td>Currently the most widely used platform in the field</td>
<td>Low multiplexing capability of samples</td>
</tr>
<tr>
<td>Helicos</td>
<td>No amplification</td>
<td>Poly(A)</td>
<td>Synthesis</td>
<td>32</td>
<td>Non-bias representation of templates for genome and seq-based applications</td>
<td>High error rates compared with other reversible terminator chemistries</td>
</tr>
<tr>
<td>PacBio</td>
<td>Single molecule</td>
<td>Hairpin</td>
<td>Synthesis, real-time</td>
<td>1000</td>
<td>Has the greatest potential for reads exceeding 1kb</td>
<td>Highest error rates compared with other NGS chemistries</td>
</tr>
</tbody>
</table>
Discovery of RNA virus in insects by NGS

The development of NGS technology allowed researcher to identify a known or discover a novel insect RNA virus. Several great review papers described recently the discovery of RNA viruses of insects and the optimization of NGS data analysis for virus discovery [89,90]. With adoption of NGS, a number of novel viruses have been discovered at an increasing rate in recent years. Liu et al. gave a summary of the recently discovered RNA viruses of insects. For the positive strand RNA genomes, iflavirus, dicistrovirus, negevirus and some unassigned viruses were discovered. Novel Iflaviruses are commonly found in hymenopteran, lepidopteran, and hemipteran insects [91-94]. A plant-infecting iflavirus-like virus, Tomato matilda virus, has been identified from asymptomatic tomato plants (Solanum lycopersicum) [90]. Dicistroviridae is a rapidly increasing family of picornavirus-like viruses [40]. A new dicistrovirus-like virus with partial sequences, Big Sioux River Virus, was discovered from the honeybee transcriptome [53]. Several ALPV-like viruses were isolated from honeybee [53,56] and western corn rootworm [95]. For dsRNA genomes, some partial fragments from birnavirus, totivirus and reovirus were identified from a D. melanogaster cell line [96]. Several negative ssRNA viral sequences have been identified from two stink bug species. One is a nearly complete genome of a Deer trick mononegavirales-like virus from Nezara viridula, and the other is a plant rhabdovirus-like virus Euchistus hero, which also infects its insect vectors [85,97]. Such newly discovered viruses may have potential for use in insect pest management. Moreover, this advanced virus discovery suggests that insects have many diverse novel viruses, which can be applied to the use in insect cell lines and laboratory insect colonies for virus research.
**Analysis of NGS data**

There is no standard method for analysis of sequences generated by NGS [98]. In general, the initial raw reads are treated with specific programs provided by the manufacturers for base calling, removal of adaptor sequences, and removal of low quality reads [89,90]. There are a lot of different ways to analyze data to discover viral sequences. DNA/RNA sequence data can be used to conduct Basic Local Alignment Search Tool (BLAST) searches against NCBI database by blastn, blastx, or tblastx, a viral database or a local database to identify the putative viral sequences before read assembly [53,99]. For read assembly, two ways are generally used: de novo assembly without a known genome sequences and reference mapping by aligning the reads to known viral genomes. The putative reads that hit viral sequences with given E-values are extracted and used for further assembly.

Alternatively, all the reads can be de novo assembled first, and the resulting assembled contigs used for BLAST analysis. Further, potential virus contigs can be isolated. The contigs with significant hits (e-value $<1 \times 10^{-3}$) to viruses were separated from non-virus hits. These contigs are then reassembled to generate longer virus contigs. The longer contigs are blasted against non-redundant (nr) databases or virus databases, then the contigs with hits to virus (e-value $<1 \times 10^{-5}$) are identified [89]. The next step is that the viral contigs are aligned against the total contig set to search for contigs that overlap viral contigs. This step is important to discover the novel viral sequences. When the virus genome is nearly assembled, the gaps can be filled by Sanger sequencing or RT-PCR and RACE-PCR.
The objectives of my studies are to: (1) deep sequence the RNA of virus particle isolated from soybean aphids in Iowa, Michigan, Ohio and China, (2) assemble the reads of virus sequences from different samples by multiple methods, (3) identify the complete or nearly complete genome sequences by *de novo* assembly and reference mapping, (4) predict the genome organization and conduct the phylogenetic analysis for each virus. These studies will provide the first information on the types of viruses that infect soybean aphid.

**Dissertation organization**

The thesis is organized in the format consisting of one journal-style manuscript accompanied by a general introduction and a general conclusion. The manuscript is formatted according to the targeted journal. The manuscript “Discovery of known and novel viral genomes in soybean aphids collected in different regions” will be submitted for publication in the PLoS ONE. Ying Feng is the primary investigator for this work under the supervision of Drs. W. Allen Miller, Karin Doman and Bryony C. Bonny, and is the first author of the manuscript.

**References**


Modeling temporal trends in aphid vector dispersal and cucumber mosaic virus
epidemics in snap bean. Environ Entomol 38: 1347-1359.

Matsumura, a new vector of Potato Virus Y in Potato. Amer J of Potato Res 82:
197-201.

aphids and men: a review of aphid damage and control strategies. C R Biol 333:
539-553.

14. McCarville MT, Kanobe C, MacIntosh GC, O'Neal M (2011) What is the economic
threshold of soybean aphids (Hemiptera: Aphididae) in enemy-free space? J Econ
Entomol 104: 845-852.

threshold for soybean aphid (Hemiptera: Aphididae). F Econ Entomol 100: 1258-
1267.

Probability of economical management of soybean aphid (Hemiptera: Aphididae)

P, editor. Aphids Their biology, natural enemies and control Amsterdam:
Elsevier: 123-139.
geese in Texas by parathion sprayed for control of Russian wheat aphid. J Wildl
April 7 Ames: Stephens Media LLC.
Eikenbary (Eds) Aphid-plant genotype interactions, Elsevier, New York, USA:
87-100.
(2007) Gene knockdown by RNAi in the pea aphid Acrthosiphon pisum. BMC
Biotechnol 7: 63.


34. Possee RD, King LA (2014) Insect Viruses. eLS.


CHAPTER 2

DISCOVERY OF KNOWN AND NOVEL VIRAL GENOMES IN SOYBEAN
APHIDS COLLECTED IN DIFFERENT REGIONS

Ying Feng\textsuperscript{1,2}, Elizabeth N. Krueger\textsuperscript{2}, Bryony C. Bonning\textsuperscript{3}, W. Allen Miller\textsuperscript{*1,2}

\textsuperscript{1}Bioinformatics & Computational Biology Program

\textsuperscript{2}Plant Pathology & Microbiology Department

\textsuperscript{3}Entomology Department

Iowa State University, Ames, Iowa 50011

*Corresponding author: wamiller@iastate.edu

B. C. Bonning and W. A. Miller designed the experiments, E. N. Krueger carried out sample collection and RNA isolation, Y. Feng analyzed the NGS data and wrote the paper, W. A. Miller and B. C. Bonning revised the paper
Abstract

Soybean aphid (*Aphis glycines*) is one of the most economically important pest insects of soybean. In the US, soybean aphid represents the first insect pest of soybean to consistently cause significant yield losses over wide areas, with yield decreases as high as 40%. To develop a biological control method, it is critical to have information on pathogens of soybean aphid. We sought to identify viruses that infect soybean aphid using Next Generation (Illumina) Sequencing of partially purified viruses from soybean aphids collected in Iowa (IA), Michigan (MI), Ohio (OH), and China. For the data analysis, *de novo* assembly of sequences was conducted using CLC Genomics Workbench. Deep sequencing analysis indicated that the soybean aphid viruses vary among the regions, but they are all positive stranded ssRNA viruses. Sequences corresponding to many viruses in the order *Picornavirales* were found. For example, aphid lethal paralysis-like dicistroviruses were found in all four regions. We also sequenced the near-complete genome of a new virus related to Big Sioux River Virus, a virus associated with honeybees. Sequences of a tetra-like virus, *Aphis glycines* virus 1 (AGV1, S. Liu and B. Bonning, unpublished) were abundant in all US samples, but partially sequence in the sample from China. Plant viral sequences were also found, including sequences with 93% identity to cotton leafroll dwarf polerovirus were found in the sample from China. A novel virus distantly related to the cileviruses of plant and insects was present in the OH samples. Sequences These results reveal the first viruses identified in soybean aphid and new plant viruses that may be vectored by soybean aphid.
Introduction

The soybean aphid (*Aphis glycines*) has been among the most economically important pest insects of temperate agriculture. It is an invasive pest that was first found in the United State in 2000, and quickly became the dominant insect pest in soybean[1]. The populations of soybean aphid increase rapidly and have heavy infections leading to significant yield loss. It has been estimated to cost up to US $3.6 billion annually in losses and control costs[2]. It is the main vector of most soybean viruses, such as alfalfa mosaic virus (AMV)[3], soybean mosaic virus (SMV)[4]. In addition to transmitting viruses to soybean, the soybean aphid damages plants directly [5]. Aphid produces honeydew, which makes plants sticky, difficult to handle, and fosters growth of sooty mold that further reduces yield. Some studies with tobacco[6], arrowleaf clover[7], and wheat[8,9] have found that plant viruses can negatively affect aphid performance and preferences[10]. However, there are no reports of viruses in the soybean aphid. Therefore, we sought to determine whether aphid viruses exist that may be useful as biological control agents or as sources of aphicidal genes for use in plant resistance.

By allowing the study of multiple virus populations, deep sequencing has been the most cost and time efficient way to identify and classify novel and known viruses, viral diversity and evolution, and transmission. To investigate what, if any, viruses are present in soybean aphids, we sequenced total RNA from partially purified virus preparations from soybean aphids collected in Iowa, Michigan, Ohio and China. After de novo assembly and reference mapping assembly, we obtained the nearly complete genome sequences of (i) dicistroviruses identical to or similar to those known to infect other aphid species, (ii) a virus, of which fragments had been found previously associated with honey
bees, (iii) a plant polerovirus, and (iv) a new cilevirus-like virus of unknown host. Several of the sequences detected were reconstructed from host organisms highly divergent from those in which related viruses have been previously isolated or discovered. It is clear that viral transmission and maintenance cycles in nature are likely to be significantly more complex and taxonomically diverse than previously known.

**Materials and Methods**

**Aphids.** Aphid handling from the various samples was conducted on separate days to avoid cross-contamination between the samples. The aphid colonies were maintained isolated from each other spatially and/or temporally.

The four samples characterized are sourced from China, Ohio, Michigan, and Iowa. The aphids from China were not frozen at the time of collection or international travel, but have been maintained at -80 °C since arriving in US. The Ohio and Michigan samples have been in colonies for at least 2 years in separate growth chamber. The Iowa sample is a lab colony annually supplemented with field-collected soybean aphid.

**Sample collection, virus purification and viral RNA extraction.** Two heavily infested trifoliate leaves and stems were collected from a colony for each U.S. location to mimic a field collected sample. The trifoliates were placed into a container at -20°C for at least one hour to kill the aphids present. The aphids were then isolated from the trifoliates by using a fresh paintbrush to remove the aphids from the leaves and into a microfuge tube. A new paintbrush was used for each sample to minimize the chance of cross-contaminating the samples. Once harvested, the aphids were maintained at -80°C until the RNA extraction.
We partially purified virus particles (virions) by ultracentrifugation of an extract from 1-3 grams of aphids through a sucrose cushion, particles were treated with ribonuclease to eliminate the non-encapsidated (non-viral) RNAs and then treated with Proteinase K to destroy the ribonuclease activity. RNA was extracted using TRizol® Reagent (Life Technologies, Grand Island, NY, USA) according to the manufacturer's protocol. The Trizol was heated to 65°C and the initial 5 minute incubation was also carried out at 65°C. The resultant RNA was salt/ethanol precipitated, resuspended in UltraPure Distilled Water and was analyzed on a Nanodrop instrument.

**Library preparation and sequencing.** The RNA samples were submitted for library preparation at the ISU DNA Facility using the Illumina TruSeq library preparation kit beginning with the fractionation and adapter ligation with out the polyA selection. Then the samples were run in a single lane on the Illumina HiSeq 2000 for 100 bp paired-end sequencing. The data were returned as four files for the China, Michigan, Ohio, and Iowa samples, respectively.

**Sequence assembly and full virus genome discovery.** For each sample, the pair-end files were imported into CLC Genomics Workbench version 6.0.4 (CLC Bio, Boston, MA, USA) as Illumina paired-end reads, discarding the read names but maintaining the quality scores. The reads can be trimmed in CLC by three ways, quality trimming, adaptor trimming and sequence filtering. We used the default trimming quality cutoff values 0.05 and the maximum number of ambiguities as 2 per 100bp read to process each sample. For adaptor trimming, choosing the Illumina adapter list to remove the adaptors from the sequencing reads. For sequence filtering, we filtered the reads below length 15 or above length 108. The output files were saved within the CLC database. In the output
files, several files give the summary of trimming data, including name of the sequence used as input, number of reads, average length of the reads, number of reads retained after trimming, the percentage of the input reads that are retained, and average length after trimming.

The paired trimmed sequence data from the trimming process was \textit{de novo} assembled using CLC Genomics Workbench to form the contigs. CLC bio’s \textit{de novo} assembly works by using \textit{de Bruijn} graphs. The basic idea is to make a table of all sub-sequences of a certain length (call words) found in the reads. The word size is determined automatically in CLC. When there is a SNP or a sequencing error, we get a so-called bubble. In this study, we used automatic values for word size and default value of bubble size 50. We also used the default value 200bp for the minimum contig length. Scaffolding was performed by CLC, in which the insert sizes don’t need to be specified. The insert size corresponds to the estimated distance between reads. CLC estimates the distance between paired reads automatically by a default distance estimation algorithm. After the initial contig creation, we selected the option to map reads back to contigs using default parameters as 2 for mismatch cost, 3 for insertion cost, 2 for deletion cost, 0.5 for length fraction, 0.8 for similarity fraction. The output file gives a figure showing the reads remapped to the assembled contigs, and a list showing the number of reads remapped to each contig.

The assembled contigs were queried against the NCBI nucleotide and protein database using BLASTn and BLASTx by CLC workbench. The filter parameters for BLAST used default value. Contigs with significant hits (E-value < 1E-03 for BLASTx, E-value < 1E-10 for BLASTn) to viruses from non-virus hits were isolated. These
contigs were re-assembled to generate longer contigs by de novo assembly as mentioned above. The longer contigs were queried against all organism database and virus database using BLAST with default parameter, and were isolated from all contigs by hitting to viruses (E-value < 1E-05). Further, the viral longer contigs were queried against all longer contigs to look for contigs that overlap viral contigs. This essential step will help to identify the novel viral sequence. Finally, we assemble the virus genome manually by checking the overlapping information. Once the longer contigs was assembled, the whole sequence was further characterized by determining the potential ORFs using default parameters.

**Phylogenetic and conservation analyses.** Multiple sequence alignment, determination of identity and similarity percentages, and initial phylogenetic determination were performed using MEGA software [11]. ClustalW was conducted for multiple amino acid sequence alignment with default parameters by MEGA6. The sequences are the amino acid sequences of RNA-dependent RNA polymerase (RdRp) for each virus. Phylogenetic trees were conducted using maximum likelihood method in MEGA 6. The test of phylogeny is bootstrap method with 500 replicates, Jones-Taylor-Thornton (JTT) model, uniform rates for rates pattern, and complete deletion for gaps/missing data treatment. For tree inference options, we chose Nearest-Neighbor-Interchange (NNI) as ML heuristic method using default initial tree for ML.

**Results**

**Deep sequencing analysis of virus hits.** In the HiSeq data used in this study, a total of 263,808,602 reads was obtained from the samples, distributed as 18% from China, 26%
from Iowa, 24% from Ohio, and 32% from Michigan. After trimming, 0.9% of the reads were removed because they failed our analysis criteria, leaving 47,539,052, 69,028,800, 82,248,632, and 62,848,082 reads, respectively. With initial de novo assembly, 64.2% of the raw reads were from contigs lengths from 157 bp to 27,554 bp (Fig. 1, Table 1). The reads distribution between the four samples, as well as the summary statistics for the de novo assembly, are reported in Table S1. As shown in workflow (Fig. 1), we performed BLASTn and BLASTx to screen the contigs with potential viral sequences. Only 24.89% and 16.64% of the reads analyzed by BLASTn and BLASTx, respectively, hit known viral gene sequences using e-value of 10^-3 cutoff for significance (Fig. 1). Most of the reads hit aphid genome sequences (with < 1E^-60). We obtained longer contigs by reassembling the potential viral contigs, to ultimately construct complete genomes. In an alternative route, reads were assembled by reference mapping using known virus genomes as references to assemble the complete genome.

**de novo assembly**

For the de novo assembly, 4,932, 30,996, 35,065, 27,512 contigs were generated in the samples from China, Iowa, Michigan and Ohio, respectively (Table 1). The majority of the contigs were identified as belonging to aphid species. A large number of the contigs were bacterial or insect sequences. For example, nearly 66% (20,164) of contigs were identified as insect in the sample from Iowa. Merely 2.6% of the de novo assembled, BLAST identified contigs were determined to be of viral origin. These sequences were from members of Caulimoviridae, Dicistroviridae, Luteoviridae, Iflaviridae, Herpesviridae, Nodaviridae, Virgaviridae, Mimiviridae and Permutotetraviridae families (Table S1).
Aphid lethal paralysis virus (ALPV), *Rhopalosiphum padi* virus (RhPV), and Big Sioux River viruses (BSRV) were the most common virus represented by the contigs in all four samples. However, most of the viral contigs differed among the regions (Table S1). Two large viral contigs (>1000 bp) for Cotton leafroll dwarf virus and Maize yellow dwarf virus were found only in the China sample. Large contigs matching *Drosophila A* virus and Kashmir bee virus were found in the Iowa and Michigan samples. Contigs with sequence similarity to Gardner-Rasheed feline sarcoma virus (GR-FeSV) were found in the Michigan and Ohio samples. A 949 nt contig in the Ohio sample was identified as similar to Citrus leprosis virus cytoplasmic type 2.

**Genome assembly of known and novel viruses**

We used all potential viral contigs as well as all contigs that did not hit any sequence in the Genebank database as inputs for re-assembly. Six segments of length from 4839 to 10174 nt were identified (Table 2). Of the six segments, three (S2, S4, S5) were closely related to Aphid lethal paralysis virus AP1 (ALPV-AP1), RhPV, and CLDV with 88%, 93.7%, and 93% full length amino acid identity respectively. These three segments had lengths close to those of the related virus genome.

The other three segments (S1, S3, S6) shared low aa identity with top BLAST hits. The top BLAST hit of S1 was *Euprosterna elaeasa* virus (EeV) with 33% aa identity, and similar length. Strikingly, S1 is firstly reported in soybean aphid sharing 98% similarity with *Aphis glycines* virus 1 (identified by Sijun Liu, unpublished). S1 was found in all four regions with abundant read counts. From the sample from Ohio, we assembled a 10,147 nt segment named S3, which has 34% aa identity with Citrus leprosis
C type 2 virus (CiLV-C2). This segment is similarly related to Negev virus, Loreto virus, and Hibiscus green spot virus with ~33% aa identity to all 3 viruses. In our study, S6 with length of 9445 nt, was most similar to the RhPV. In a previous study of viruses isolated from honey bee, four contigs of size 1473, 861, 1164, and 1311 nt (GenBank JF423195-8) were designated from a hypothetical Big Sioux River virus (BSRV), after its place of discovery in South Dakota. Interestingly, all four contigs of BSRV have high aa identity to the sequence of S6 as shown in Fig. 2F.

**Virus Genome Organizations**

The genome organization of S1 is shown in Fig. 2. Four ORFs have sizes of 3438 nt of ORF1 (44-3,481 nt), 510 nt of ORF2 (1,945-2,454 nt), 735 nt of ORF3 (3,267-4,001 nt), and 810 nt of ORF4 (4,005-4,805 nt). Part of ORF1 of the S1 hit the sequences of putative replicase of EeV and RNA polymerase of Thosea asigna virus (TaV), while ORF3 of S1 had highest similarity to the capsid protein of Bat sobemovirus and Pothos latent virus (Fig. 2A). In the BLAST analysis, the S2 sequence showed 95% and 94 % aa identity with the ORF1 (GenBank AFU81560) and ORF2 (GenBank AFU81561) of ALPV-An, respectively (Fig. 2B).

S3 contains four large ORFs (ORF1 20-7,417 nt; ORF2 7,314-8,660 nt; ORF3 8,632-9,237 nt; ORF4 9,293-9,991 nt) as illustrated in Fig. 2C. The top BLAST hit of S3 is CiLV-C2, which is a bipartite RNA virus. RNA1 of CiLV-C2 (8,717 nt) contains two ORFs, and RNA2 (4,989 nt) contains five ORFs [12]. The ORF1 and ORF2 of S3 showed 33% aa identity with the RNA1 of CiLV-C2, while there was no matches to the RNA2 of CiLV-C2 in sequence of S3 (Fig. 2C).
The sequence of S4 contains two ORFs (ORF1 580-6,576 nt; ORF2 7,103-9,560 nt) with high similarity to RhPV (Fig. 2D). S5 showed similar genome organization as its top BLAST hit CLDV, having six ORFs (Fig. 2F). ORF0 to ORF5 encoded the silencing suppressor protein, RNA dependent RNA polymerase, capsid protein, movement protein, and read-through protein[13]. The sequence of S6 had two ORFs (ORF1 3-2,150 nt; ORF2 2,753-8,752 nt) with 71% and 73% amino acid identity to the polyprotein (ABX74939.1) and structural polyprotein (ABX74940.1) of RhPV respectively. We also identified that all four contigs of BSRV were included in S6 as shown as Fig. 2F.

**Virus phylogenetic analysis**

Based on the results of ClustalW multiple alignments, we constructed a maximum likelihood bootstrap phylogenetic tree for each sequence. In Fig. 3A, S1 was close to the branch of EeV and TaV, which belongs to *Permutotetraviridae, Alphapermutotetravirus*. The other two hitting viruses, Drosophila A virus is an unclassified dsRNA virus, and Bat sobemovirus is an unclassified *Sobemovirus* from bat feces (Fig. 2A, Fig. 3A).

S2, S4 and S6 all had high identity to RhPV and ALPV. As in the BLASTx result, S2 is on the same branch as ALPV, and S4 and RhPV were placed on the same branch respectively (Fig. 3B). These six sequences as well as DCV and CPV belong to *Cripavirus* of the family *Dicistroviridae*.

A phylogenetic tree was constructed based on the multiple alignments of the BLAST hits by the S3 sequence segment. S3 was not grouped to any branch in a phylogenetic tree, but it was more close to CiLV-C (Fig. 3C). The Lereto virus, Piura virus, Negewotan virus and Negevirus are novel insect viruses, which are related
distantly to CiLV-C[14]. In the other subgroup, there are all plant viruses, including Tobacco mosaic virus, Hibiscus latent Fort Pierce virus, Yellow tailflower mild mottle virus and so on.

S5 is closely related to CLDV, and belongs to the same genus, Polerovirus (Luteoviridae), as Maize yellow dwarf virus-RMV (Fig. 3D). Its close relatives include other plant poleroviruses, such as Melon aphid-born yellows virus or Pepper vein yellows virus.

Discussion

Our study is the first research survey on the viruses sequence of soybean aphids, and the first to identify any viruses sequences associated with soybean aphid. The large samples allowed us to be more informed on the viruses the soybean aphids carried, and the deep sequencing allowed us to study multiple viral populations in the most cost and time efficient way. The results revealed known and novel viruses, which provide valuable information on viral diversity and evolution.

In this study, we identified six complete or nearly complete virus genomes from different regions. Three out of six represent novel viruses (S1, S3, and S6), as they have low aa identity and low coverage with the known virus sequences. The top hitting sequence of S1 is insect virus EeV, a member of the Tetraviridae family[15], but with only 33% aa identity and 38% coverage at most. And their genome organization is not similar. S1 was found in all samples of US, and even partial sequence in the China sample. It indicates that the S1 may be a novel widespread virus in soybean aphid, so we named it as *Aphis glycines virus 1* (AGV1). There were relatively few sequence reads
obtained from the China soybean aphid sample, possibly because the sample has been stored at room temperature, possibly causing selective degradation of less stable viruses, whereas the IA, MI, OH viruses were isolated from live aphid colonies.

S3 from the Ohio sample has little identity to any virus but has partial identity (33%) to the plant virus CiLV-C2, with a longer genome 10,147 nt compared to 8,717 nt for CiLV-C2. In the phylogenetic tree (Fig. 3C), two viral subgroups are distantly related to S3. One subgroup contains all plant viruses belonging to cilevirus and higrevirus genera. The other subgroup, negevirus [14], contains only insect viruses. Negevirus is a new taxon of insect-specific viruses. Six novel insect viruses, Negev, Ngewotan, Piura, Loreto, Dezidougou, and Santana, isolated from mosquitoes and phlebotomine sand flies form this new genus[14]. They are closest but still distant relative are CiLV-C, and viruses in genus Cilevirus. It is clear that S3 is a novel virus, but whether it is a plant virus or an insect virus is unknown. We designated S3 as *Aphis glycines virus 2* (AGV2).

Also of interest, S6 has similarity to all four contigs of BSRV (Fig. 2F). Those contigs were obtained by deep sequencing of virus isolated from honey bee [16]. S6 shares 71% aa identity with RhPV, but share highest aa identity (85% to 93%) with each contigs of BSRV. We designate S6 as *Aphis glycines virus 3* (AGV3). Which hosts AGV3 actually infects are unknown, but given the high number of reads we obtained (over 4,500 reads per sample), it is highly likely to infect soybean aphid.

Most of the virus contigs in our study hit the sequences of ALPV and RhPV. Clearly these viruses are infecting soybean aphid, and are unlikely to be merely associated with it. Based on previous researches, RhPV was known only to infect cereal
aphids, including aphid *R. padi* (bird cherry-oat aphid) [17], *R. rufiabdominalis* (rice root aphid) [17], *Diuraphis noxia* (Russian wheat aphid) [18], *R. maidis* (green corn aphid) and *Acyrthosiphum dirrhodum* (grain aphid) [19], and *Schizaphis graminum* (greenbug)[17,20]. The predicted ORFs of S4 both have >90% aa identity to RhPV, hence S4 is not a new virus with 90% similarity to RhPV. S1 is named as RhPV-AG, a new RhPV-like virus isolated from *Aphis glycines*.

ALPV, an aphid-pathogenic virus, was associated with rapid decline of the population of the major aphid species colonizing small grains under natural conditions [21,22]. ALPV was first isolated from the aphid *Rhopalosiphum padi*. It is much more pathogenic to a *R. padi* than RhPV and is present at much higher concentrations in aphids than is RhPV [23]. Several ALPV-like viruses have been identified in organisms other than aphids, such as from bat fecal samples and honeybees [16,24-28]. Considering that the sequence identity of amino acid between S2 and ALPV was 92%, we named S2 as ALPV-AG, a new isolates from *A. glycines*. Recently, a new isolate of ALPV (ALPV-AP) was isolated from the pea aphid, *Acyrthosiphon pisum*, which is closely related to ALPV-AM, an ALPV isolate fro honeybee [29]. The ALPV-AP genome is 88% identical to ALPV-AG from the sample of China, and is 86% identical to ALPV-AG from the sample of Michigan. These results indicate that the ALPV is highly adaptable to many insect species. ALPV cause disease symptoms and limit *R. padi* population in nature [23]. ALPV-An, a strain of ALPV, was highly pathogenic to *Myzus persicae* [25]. However, no acute pathogenicity was observed in ALPV-AP infection of *A. pisum* under ideal rearing conditions [29], and neither in ALPV-AG infection of *A. glycines*. It suggests that the diversity also exists in the biological properties of ALPVs such as host range.
S5 appeared only in the sample from China, sharing 93% aa identity with the plant virus CLDV. CLDV is the causal agent of cotton blue disease (CBD) and is transmitted by the *Aphis gossypii* Glover [13,30]. It is a (+)ssRNA virus in a new species within the genus *Polerovirus (Luteoviridae)* [30]. These viruses are transmitted in a circulative, non-propagative manner. CBD is an important disease present in cotton crops in South America, Africa and Asia [31]. The infected cotton plants show a moderate to severe stunting phenotype due to shortening of internodes, leaf rolling, vein yellowing and intensive dark green color of the foliage [32]. CBD only found in South America, Africa and Asia explained why we find this virus only in the sample from China. Presence of CLDV in soybean aphid suggests soybean aphid may be a vector, which has not been reported.

In summary, our results (Table 2 & Table S1), suggest that a variety of viruses associate with soybean aphid, but it is not clear which actually replicate in this host. We assembled only six full viral genomes, but there are still a lot of small contigs in each sample that hit a wide range of viral genomes. However, most of the contigs have short hits length (< 100 nt), and low read count (< 50), due to several possible reasons: 1) sequence random matching because the hitting length is short even e-value less than 1E-10 the contig still has a low identity to the hitting virus; 2) samples are contaminated during the isolation; 3) some noncirculative virions remain on the surface of aphids.

Overall, this study provides the first information on the types of viruses that infect soybean aphid. By this information, we can better understand the possible plant or insect viruses soybean aphids may carry. Such viruses have potential for use as biological control agents or as sources of non-infectious transgenic aphicidal genes in plants. It also
shows that viral transmission and maintenance cycles in nature are likely to be highly complex and taxonomically diverse.

Acknowledgements

We would like to thank Dr. Xiulian Sun, and Andy Michel from Ohio State University for provision of soybean aphids from Ohio, Douglas Landis, from Michigan State University for aphids from Michigan, Matt O’Neal for aphids from Iowa. We acknowledge the Iowa Soybean Association for the funding.

References


Tables

Table 1. Summary of sequencing results by location

<table>
<thead>
<tr>
<th>Sample location</th>
<th>Reads after trimming</th>
<th>No. of contigs by <em>de novo</em> assembly</th>
<th>Avg. length of contigs (bp)</th>
<th>No. of virus contigs</th>
<th>BLASTn</th>
<th>BLASTx</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>47,539,052</td>
<td>4,932</td>
<td>898</td>
<td>107</td>
<td>144</td>
<td></td>
</tr>
<tr>
<td>Iowa</td>
<td>69,028,800</td>
<td>30,997</td>
<td>546</td>
<td>285</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>Michigan</td>
<td>82,248,632</td>
<td>35,065</td>
<td>525</td>
<td>416</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>Ohio</td>
<td>62,848,082</td>
<td>27,512</td>
<td>542</td>
<td>326</td>
<td>67</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Virus genome assembly segments and BLASTx alignment results.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Top BLASTx hits</th>
<th>Virus</th>
<th>Genome length (nt)</th>
<th>Hit length (aa)</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S1</strong></td>
<td></td>
<td>Euprosterna elaeasa virus</td>
<td>5698</td>
<td>631</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aphid lethal paralysis virus</td>
<td>9835</td>
<td>1501</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Citrus leprosis C type 2 virus</td>
<td>8717</td>
<td>471</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rhopalosiphum padi virus</td>
<td>10011</td>
<td>485</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cotton leafroll dwarf virus</td>
<td>5865</td>
<td>638</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>Location</th>
<th>Length</th>
<th>Read (nt)</th>
<th>Read counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>IA</td>
<td>4839</td>
<td>35895</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MI</td>
<td>4838</td>
<td>38515</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OH</td>
<td>4839</td>
<td>22194</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>IA</td>
<td>9120</td>
<td>603</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MI</td>
<td>9918</td>
<td>507</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OH</td>
<td>9075</td>
<td>567</td>
<td></td>
</tr>
<tr>
<td></td>
<td>China</td>
<td>9814</td>
<td>2365146</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>OH</td>
<td>10147</td>
<td>11924</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>MI</td>
<td>10045</td>
<td>1018</td>
<td></td>
</tr>
<tr>
<td></td>
<td>China</td>
<td>9494</td>
<td>3772552</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>China</td>
<td>5841</td>
<td>6277</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>IA</td>
<td>9445</td>
<td>4202</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MI</td>
<td>9443</td>
<td>4783</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OH</td>
<td>9445</td>
<td>4163</td>
<td></td>
</tr>
<tr>
<td></td>
<td>China</td>
<td>9445</td>
<td>20784187</td>
<td></td>
</tr>
</tbody>
</table>
Table S1. Viruses hits to the assembled contigs by BLAST in different location

<table>
<thead>
<tr>
<th></th>
<th>China</th>
<th>Iowa</th>
<th>Michigan</th>
<th>Ohio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acyrthosiphon pisum virus</td>
<td>Aphid lethal paralysis virus</td>
<td>Aphid lethal paralysis virus</td>
<td>Aphid lethal paralysis virus</td>
<td>Aphid lethal paralysis virus</td>
</tr>
<tr>
<td>Aphid lethal paralysis virus</td>
<td>Big Sioux River virus</td>
<td>Big Sioux River virus</td>
<td>Big Sioux River virus</td>
<td>Big Sioux River virus</td>
</tr>
<tr>
<td>Barley yellow dwarf virus</td>
<td>Cafeteria roenbergensis virus</td>
<td>Bovine viral diarrhea virus-1</td>
<td>Bovine viral diarrhea virus-1</td>
<td></td>
</tr>
<tr>
<td>Big Sioux River virus</td>
<td>Euprosterna elaeasa virus</td>
<td>Euprosterna elaeasa virus</td>
<td>Citrus leprosis virus cytoplasmic type 2</td>
<td></td>
</tr>
<tr>
<td>Brevicoryne brassicae picorna-like virus</td>
<td>Influenza A virus</td>
<td>Gardner_Rasheed feline sarcoma virus</td>
<td>Euprosterna elaeasa virus</td>
<td></td>
</tr>
<tr>
<td>Citrus leprosis virus C</td>
<td>Rhopalosiphum padi virus</td>
<td>Grapevine leafroll-associated virus 7</td>
<td>Gardner-Rasheed feline sarcoma virus</td>
<td></td>
</tr>
<tr>
<td>Cotton leafroll dwarf virus</td>
<td>Tanapox virus</td>
<td>Kirsten murine sarcoma virus</td>
<td>Kirsten murine sarcoma virus</td>
<td></td>
</tr>
<tr>
<td>Euprosterna elaeasa virus</td>
<td>White bream virus</td>
<td>Maize dwarf mosaic virus</td>
<td>Micromonas pusilla virus 12T</td>
<td></td>
</tr>
<tr>
<td>Flock house virus</td>
<td>Micromonas pusilla virus 12T</td>
<td>Rhopalosiphum padi virus</td>
<td>Stealth virus 1</td>
<td></td>
</tr>
<tr>
<td>Hana virus</td>
<td>Rhopalosiphum padi virus</td>
<td>Stealth virus 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hibiscus green spot virus</td>
<td>Stealth virus 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lake Sinai virus 2</td>
<td></td>
<td>Yaba monkey tumor virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loreto virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maize yellow dwarf virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modoc virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negev virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ngewotan virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oyster mushroom isometric virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peanut clump virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhopalosiphum padi virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosy apple aphid virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All hits to the assembly contigs by BLASTn or BLASTx were selected by e-value less then $1E^{-10}$. 
Figure 1. Workflow for processing deep sequencing data. Two strategies used to analyze the data were *de novo* assembly and reference mapping. Average percent remaining reads after each of the filtering steps. Low-quality reads are moved first, followed by BLASTn and BLASTx comparison to virus sequences.
Figure 2. Genome organizations of the assembled virus segments S1-S6 and similar RNA viruses. Viral sequences were positionally mapped to the genome of S1-S6 (A-F) based on the identities of the associated amino acid (aa) sequences to the corresponding reference genome proteins. Blue bar: assembled genome; red bar: viral sequences that share closest similarity with the assembled genome; parenthesis: percentage of aa identity to the corresponding sequences.
Figure 3. Phylogenetic analysis of assembled virus segments. Phylogenetic trees comparing RNA-dependent RNA polymerase of (A) S1 with representative members of the Alphapermutotetravirus genus within the Permutotetraviridae family; (B) S2, S4 and S6 with representative members of the Aparavirus and Cripavirus genus within the Dicistroviridae family; (C) S3 with representative members of the Cilevirus, Negevirus, and Tobamovirus genus; and (D) S5 with representative members of Polerovirus genus within the Luteoviridae family. Phylogenetic trees were constructed by using the maximum likelihood method in the MEGA6 program[11] with 500 bootstrap replicates. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.
Soybean aphid has a negative effect on agricultural economy in North America. Extensive studies focus on aphid damage and control strategies. Except chemical management or host plant resistance, biological control was seen as a great potential management strategy. A number of researches work on specific virus in insect or aphid as the transmission vector of specific plant virus. Now NGS technologies make identifying and discovering virus much more easier and faster. We can conduct metagenomic analyses for virus sequence discovery.

By now, there is no research on the investigation of the viruses that soybean aphid is carrying. In this study, we conduct the deep RNA sequencing of virus particle isolated from soybean aphids in Iowa, Michigan, Ohio and China by NGS. The reads of virus sequences from different samples was assembled by multiple methods. Complete or nearly complete genome sequences were identified by de novo assembly method and reference mapping method. We discovered six full- or nearly full-length virus sequences, including dicistro-like virus, tetra-like virus, and a plant cile-like virus. We predict the genome organization and conduct the phylogenetic analysis for each virus. These results reveal the first viruses identified in soybean aphid and new plant viruses that may be vectored by soybean aphid. The information will shed light on the potential use of aphid viruses as biological control agents or as sources of non-infectious transgenic aphicidal genes in plants.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. W. Allen Miller, for his guidance and patience and financial support during my graduate study. I am also grateful to my co-advisor Dr. Karin Dorman, for her guidance, encouragement and full support. I am also indebted to my committee member, Dr. Bryony Bonning who has given me a lot of suggestions and help.

I also want to thank our lab members Dr. Jelena Kraft, Dr. Jimena Carrillo-Tripp, Alice Hui, Dr. Juliette Doumayrou, Dr. Sung ki Cho, Bliss Kernodle, and Elizabeth J Carino, colleagues in Dr. Bonning's, and Dr. Dorman's lab, many faculty members and staffs in the program of Bioinformatics and Computational Biology for their help and valuable discussion and suggestion.

I would like to give my special thanks to my dear boyfriend Xiaochu Lou, friends Dr. Lei Gong, and Dr. Haibo Liu for their insightful discussion, unselfish help, and mental support.

Last but not least, I thanks to my parents for their encouragement, patience, respect and love. My love and gratitude for them can hardly be expressed in words. I dedicate this thesis to my parents: Donglian Feng and Lizhu Zhang.

“愿你成为这样的女子：不炫耀，不争吵，做一个博学的女子；不空洞，不浮躁，做一个丰盈的女子；即便生命枯竭，亦在优雅中变老.”