UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DIRECT METAGENOMIC DETECTION AND ANALYSIS OF PLANT VIRUSES USING AN UNBIASED HIGH-THROUGHPUT SEQUENCING APPROACH

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

GRAHAM BURNS WILEY
Norman, Oklahoma
2009

DIRECT METAGENOMIC DETECTION AND ANALYSIS OF PLANT VIRUSES
USING AN UNBIASED HIGH-THROUGHPUT SEQUENCING APPROACH

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

____
Dr. Bruce A. Roe, Chair

____
Dr. Ann H. West

____
Dr. Valentin Rybenkov

____
Dr. George Richter-Addo

____
Dr. Tyrell Conway

Acknowledgments

I would first like to thank my father, Randall Wiley, for his constant and unwavering support in my academic career. He truly is the “Winston Wolf” of my life.

Secondly, I would like to thank my wife, Mandi Wiley, for her support, patience, and encouragement in the completion of this endeavor.

I would also like to thank Dr. Fares Najar, Doug White, Jim White, and Steve Kenton for their friendship, insight, humor, programming knowledge, and daily morning coffee sessions.

I would like to thank Hongshing Lai and Dr. Jiaxi Quan for their expertise and assistance in developing the TGPweb database.

I would like to thank Dr. Marilyn Roossinck and Dr. Guoan Shen, both of the Noble Foundation, for their preparation of the samples for this project and Dr Rick Nelson and Dr. Byoung Min, also both of the Noble Foundation, for teaching me plant virus isolation techniques.

I would like to thank Chunmei Qu, Ping Wang, Yanbo Xing, Dr. Ruihua Shi, Keqin Wang, and Baifang Qin for their assistance in sequencing.

I would like to thank my graduate school colleagues: Simone Macmil, Dr Majesta O’Blenness, Dr. Iryna Sanders, Dr. Shweta Deshpande, Dr Jing Yi, Dr. Leo Sukharnikov, Dr. Chris Lao, Dr. Jianfeng Li, Dr. Shelly Ooman, and Dr. James Yu.

I would like to thank Dixie Wishnuck, Jennifer Lewis, Mary Catherine Williams, and Kay Lynn Hale for their support.

I would like to thank Dr. Doris Kupfer, Angie Prescott, and Rose Morales-Diaz for their mentorship during my undergraduate stretch in the Roe lab.

I would like to thank other members of the Roe lab, both past and present: Shaoping Lin, Fu Ying, Liping Zhou, Limei Yang, Ziyun Yao, Axin Hua, Yonas Tesfai, Trang Do, Anh Do, Phoebe Loh, Xiangfei Kong, Sulan Qi, Honggui Jia, Xu Xi, Randy Hines, Sara Downard, Yuhong Tang, and Lin Song.

And last, and most assuredly not least, I would like to express a tremendous amount of gratitude to my mentor, boss, and chair Dr. Bruce Roe who, with extreme patience, minute insight, encyclopedic knowledge, sheer force of will, and a couple of kicks, brought this whole thing together. He has molded me into a scientist and I will try to not disappoint him.

Table of Contents

List of Tables

List of Figures

Abstract

It is well established that plants, along with other life forms, often are infected by viral parasites that require the host cellular machinery for replication. Since, the overwhelming majority of these viruses have been from cultivated plants from laboratories and greenhouses, I investigated the viral populations from wild, uncultivated plants, hypothesizing that they would harbor new and novel viruses. To complete this study, an optimized method for the detection of plant viruses using a direct, unbiased metagenomic approach was developed and implemented from plants in the Tallgrass Prairie Preserve in Northeastern Oklahoma. Subsequently, their RNA viral genomes were isolated and converted to tagged cDNAs that were pyrosequenced on a Roche/454 GS-FLX, assembled and compared to other known gene sequences. A comprehensive relational mySQL-based web-accessible database also was implemented to facilitate analysis of the large amounts of metagenomic data generated. Of the 1254 sampled plants, 496 were infected with one or more viruses, that were represented by 1624 assembled cDNA sequences. Of the 19 viral families represented, the three most prevalent were *Tymoviridae*, *Totiviridae*, and *Partitiviridae* although the majority of observed virus sequences were new, previously un-described species, often representing new viral genera. Since *Totiviridae* and *Partitiviridae*, characteristically fungal viruses, also coincided with detection of fungi associated with the plants, it is very likely that the majority of the viruses observed represented viral infections of fungi that were interacting with the plants.

Through these studies, a diverse number of new, previously undiscovered viral species were observed in the wild, uncultivated plants of the Tallgrass Prairie Preserve, that multiple infections of viruses in these plants are commonplace, at least one virus, a member of the family *Tymoviridae,* was widely distributed on a single species of plant, *Asclepias viridis,*a likely ecological viral niche, and that a majority of the classified viral species observed represented members of fungal associated virus families.

Chapter 1 Introduction

1.1. Plant Viruses

1.1.1. Definition of a virus

Viruses, often described as a nucleic acid surrounded by proteins, are ubiquitous throughout nature, being found in animals (Boshoff et al 1995), plants (Goelet et al 1982), bacteria (Adams 1952), fungi (Day 1981), soil (Kim et al 2008), marine sediment (Breitbart et al 2004), and seawater (Wilcox and Fuhrman 1994). Because of their reliance on the host cellular machinery for replication, viruses are true obligate intracellular symbionts. Although viruses usually are classified as cellular parasites because of their pathogenic nature, recent studies have hypothesized many viruses are commensal with their host (Flotte and Berns 2005; Griffiths 1999; Roossinck 2003) if not truly mutualistic (Marquez et al 2007; Roossinck 2005a; Whitfield 2002).

1.1.2. Plant Virus Morphology

The plant virus genome consists of a nucleic acid, often multipartite, core surrounded by protein and, in a minority of plant viruses, a lipid layer may envelop this protein coating. Common virus particle shapes included rods, isometric spheres, filamentous strands, isometric geminates, and bacilliforms. Rod shaped viruses may be 100 to 300 nm long with an average diameter of ~20 nm (Koenig 2005b; Lewandowski 2005; Torrance 2005). Isometric spheres range in diameter from 17 nm to 65 nm in diameter (Kassanis 1962; Upadhyana 2005) with protein shell symmetries varying from T=1 to T=3 (Ban and McPherson 1995; Canady et al 1996). Filamentous strands, similar in shape to rod type viruses although they are much longer and are flexible along their length, range in size from 500 to 2000 nm in length with an average diameter of 12 to 13 nm (Tollin 1988). Isometric geminate viruses closely resemble two isometric spherical viral particles that have fused generating a twin-like structure with dimensions of 22x38 nm (Stanley 2005). Bacilliform viruses are similar to isometric and geminate viruses in that they have rounded ends with icosahedral symmetry joined by a barrel structure, vary in size depending on virus family or genus between 18 to 30 nm wide and 57 to 900 nm long (Roossinck 2005b; Stanley 2005)

1.1.3. Methods of Infection

For viruses to infect plant cells they must first enter the cell by passage through the cell's waxy cuticle and a cellulose by one of several methods, either biologically, mechanically, or propagative.

1.1.3.1. Biological Methods of Infection

Many plant viruses are commonly spread between plants by the feeding action of invertebrates, either arthropods (Nault 1997) or nematodes (Brown et al 1995). Arthropod vectors infect plant through their normal feeding processes, either by piercing the plant tissue with a stylet-type mouth appendages, such as aphids and mites, or through the chewing of the plant tissue with biting appendages, such as beetles. Within these vectors the plant virus may be non-persistently spread if it associates only with the feeding appendages, or it may become systemic throughout the vector and generate infections in a persistent manner. Nematodes infect plants in a similar manner, with the root tissues of a plant being pierced by the stylet-style mouthpiece (Brown and

Weischer 1998). Nematode vectors do not have systemic spread of the plant virus as arthropods do, but plant viruses may still be spread in a persistent manner as the virus adsorbs to and releases from the surface of the feeding appendage (Brown and Weischer 1998).

Plant viral infections also are spread through interactions with soil based fungi (Grogan and Campbell 1966). If a plant virus is present in the fungal cell as it infects the host plant, the virus can be released as becomes established in the root cells, although method of the viral release currently is unknown (Campbell 1996).

Viruses also can spread through contact with parasitic plants, such as dodder (*Cuscuta* spp.) (Bennett 1940). In these cases the dodder acts as a passive pipeline between two or more plants as the plant virus flows through the dodders vascular system.

1.1.3.2. Mechanical Methods of Infection

Commonly called mechanical inoculation, this method relies on the contact of a plant with a mechanical agent that scrapes or pierces the plant tissue to generate ephemeral tears in the cuticle and cell wall. This may be accomplished by a passing animal or machine (Broadbent 1963; 1965) or by direct contact between two leaves of neighboring plants (Clinch et al 1938). The most common form of inoculation to study plant viruses in the laboratory is by pipetting a buffer solution containing a virus onto a leaf that has been rubbed with an abrasive such as carborundum dust.

1.1.3.3. Propagative Methods of Infection

Propagative methods for the transmission of viruses include pollen, seed, grafting, and vegetative propagation. Pollen-borne viruses can infect the seed or the mother plant. Self-pollinating plants also can re-infect themselves through pollen, leading to higher titers of virus in the resulting seeds. However, infected pollen is not always required for a virus to become seed-borne. Seeds can be infected pre-fertilization through the infection of the gametes within the seed or by direct embryo infection post fertilization (Maule and Wang 1996). In grafting and vegetative propagation, a cutting from an infected plant either transfers its infection to the stock it is grafted to (Zaitlin 1962) or grows into a second mature infected plant (Hull 2002), respectively.

1.1.3.4. Viral Replication within the Host

Once a viral infection has occurred, viruses utilize host cellular machinery for the synthesis of viral proteins, which in turn aid in the replication of more viral particles. The method by which a virus replicates within a host cell is dependent entirely on the viral genetic material. Viral replication can occur only in those areas of the host cell not separated by a lipid bi-layer and is carried out through the assembly of component molecules as compared to the binary fission of most prokaryotes. Viral replication may lead to a large amount of genetic variation due to errors in replication, genome recombination, or the incorporation of unrelated virus or host genetic material (Haenni 2008).

1.1.3.5. Positive Sense Single Stranded RNA Virus Replication

Positive sense single stranded RNA [(+)ssRNA], the most prevalent genomic type among plant viruses, also is the most straightforward in its replication cycle as the genome may act as its own mRNA (Roossinck 2005a). During infection the viral particle enters the cell via one of the methods discussed above, and after uncoating the genomic (+)ssRNA molecule recruits host cell ribosomes to immediately begin translation of viral genes into proteins. After the proteins are generated, they are post-translationally processed, if necessary, by self-encoded proteases. The viral RNA-dependent RNA polymerase and methyltransferase/helicase proteins then recruit host factors to form both subgenomic RNAs and progeny genomes. Coat proteins then envelope the progeny genomes to generate new viral particles. This process is depicted in Figure 1.

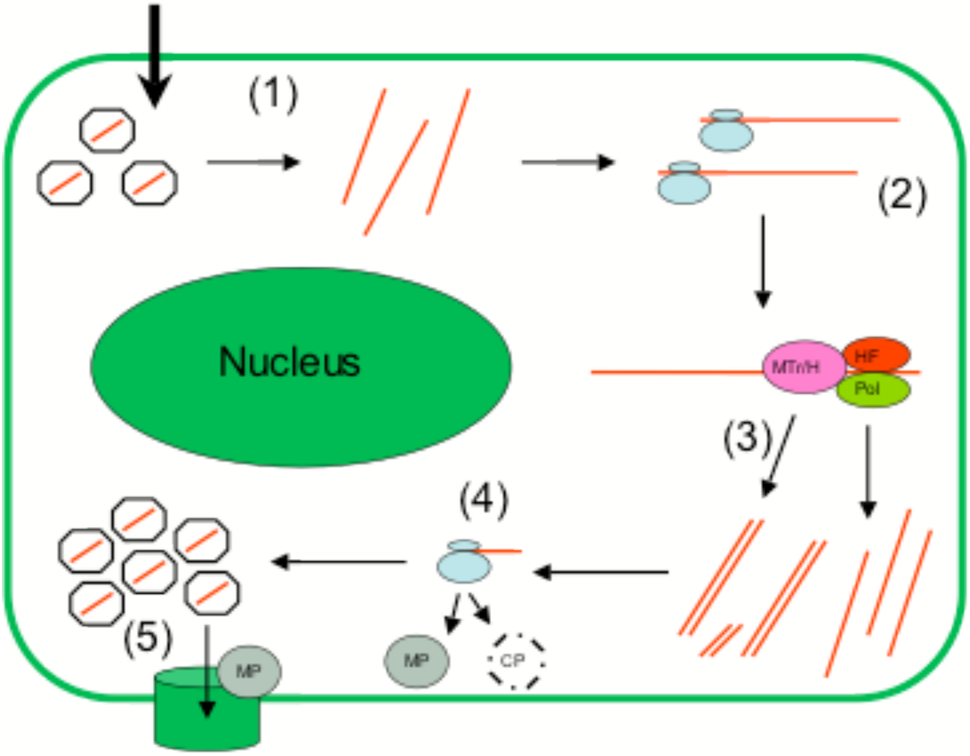


Figure 1. The Life Cycle of a (+)ssRNA Plant Virus. (1) The virus enters the cell from an external source via a break in the cell wall and the RNA genome uncoats. (2) The genes for methyltranferase/helicase (MTr/H) as well as the RNA dependent RNA polymerase (Pol) are translated directly from the genomic RNA by host cell ribosomal proteins. (3) The methyltransferase/helicase and polymerase combine with host factors (HF) to generate progeny genomes and subgenomic messenger RNA. (4) Subgenomic mRNA is translated to produce Coat Proteins (CP) and Movement Proteins (MP). (5) The Movement Protein associates with the plasmodesmata of the plant cell, increasing its size exclusion limit. The translated Coat Protein encapsulates progeny genomic RNA and the new viral particle moves through the widened plasmodesmata to the neighboring cells. (Roossinck 2005a)

1.1.3.6. Negative Sense Single Stranded RNA Virus Replication

Negative sense single stranded RNA [(-)ssRNA] viruses replicate in both the plant nucleus as well as the cytoplasm, depending on the genus of virus (Jackson et al 1999).

After entry into the cell plant viruses of the family *Nucleorhabdovirus* associate with the endoplasmic reticulum and uncoat and release a nucleocapsid core into the cytoplasm. The nucleocapsid then enters the nucleus of the cell through the nuclear pore complex where a polymerase protein incorporated into the nucleocapsid transcribes the negative sense strand into positive sense mRNAs that travel to the cytoplasm where they are translated into proteins. The viral proteins then are transported back into the nucleus to continue mRNA transcription, generate the progeny genomic molecules, and create a viroplasmic space within the nucleus where the nucleocapsid and coat proteins then combine with genomic RNA to form new viral particles. These new viral particles both bud into perinuclear space as well as move into the cytoplasm for transport to a new host cell.

Members of the genus *Cytorhabdovirus* replicate in a similar manner to *Nucleorhabovirus*, but do so in a viroplasm constructed in the cytoplasm as opposed to the nucleus. Also, mature viral particles bud into a cytoplasmic membrane instead of the perinuclear space. A graphical representation of the two methods of replication are given in Figure 2.

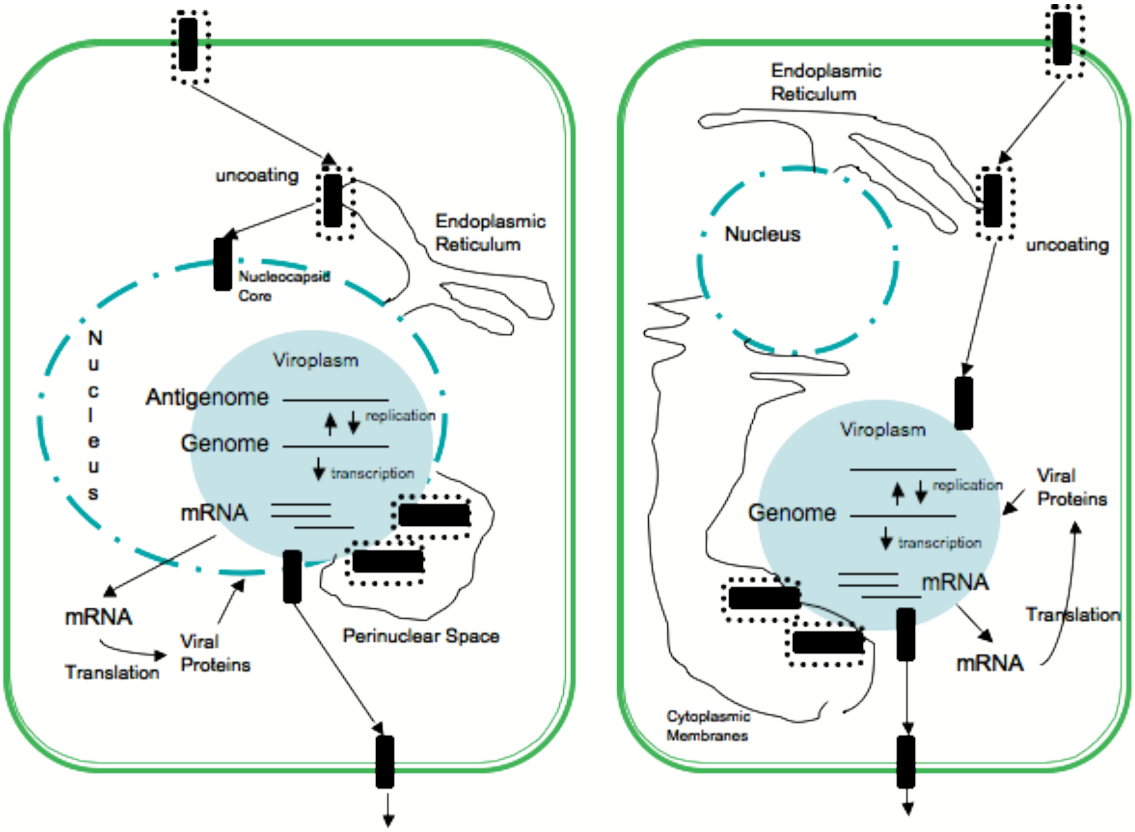


Figure 2. Replication cycle of *Nucleorhabdovirus* (left) and *Cytorhabdovirus* (right) (Jackson et al 1999)

1.1.3.7. Double Stranded RNA Virus Replication

Double stranded RNA (dsRNA) viruses replicate in the cytoplasm (Wickner 1993). After initial infection, a new, positive sense strand is transcribed within the virus particle itself, is extruded into the cytoplasm, and serves as an mRNA template for the synthesis of viral proteins. After the coat and RNA-dependent RNA polymerase proteins have been generated, the positive sense strand and the RNA-dependent RNA polymerase are encapsulated into a new viral particle. The positive sense strand then acts as a template to form a new double stranded RNA molecule generating a mature viral particle. A graphical representation of this process is given in Figure 3.

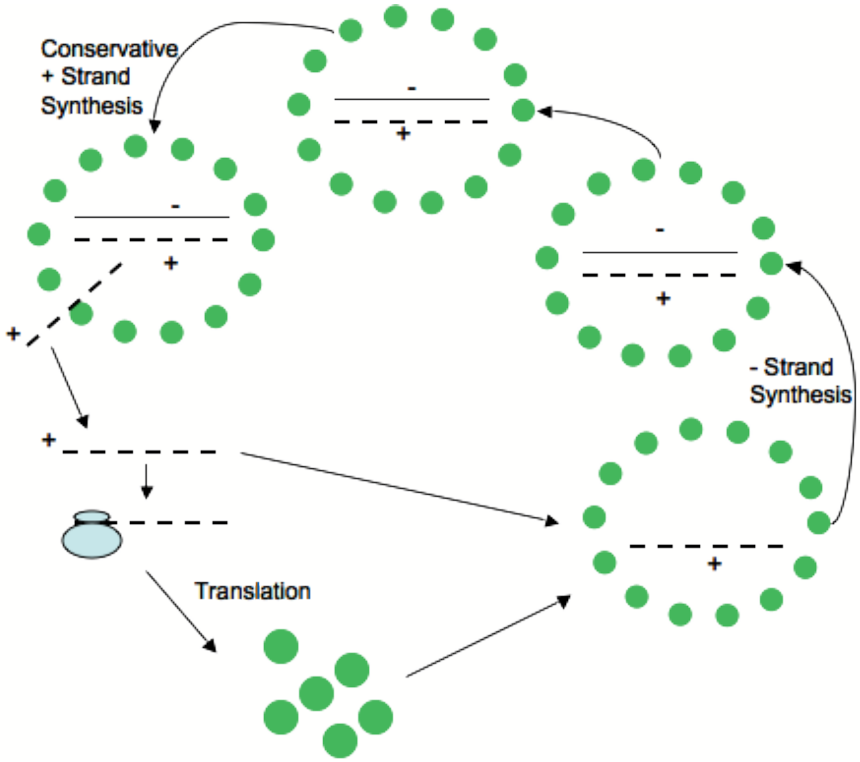


Figure 3. Replication cycle of a dsRNA virus in the cytoplasm (Wickner 1993)

1.1.3.8. Reverse Transcribing Double Stranded DNA Virus Replication

Reverse transcribing viruses replicate in two stages spanning both the nucleus and the cytoplasm (Hull 2002). Upon entering the cell the circular, discontinuous dsDNA genomic molecule uncoats and is transported into the nucleus of the cell where genomic discontinuities are sealed and the molecule associates with host cell histones to generate a minichromosome. Host RNA polymerase then transcribes a full length transcript of the genome that is transported to the cytoplasm. In the next stage, the full length RNA transcript is primed via a cytosolic initiator methionyl tRNA for reverse transcription to DNA via the virally encoded reverse transcriptase. The reverse transcription of the positive sense strand is primed by RNaseH processing of two polypurine tracts in the RNA strand. The newly formed, discontinuous dsDNA molecule then is encapsulated by coat proteins. A graphical representation of this is given in Figure 4.

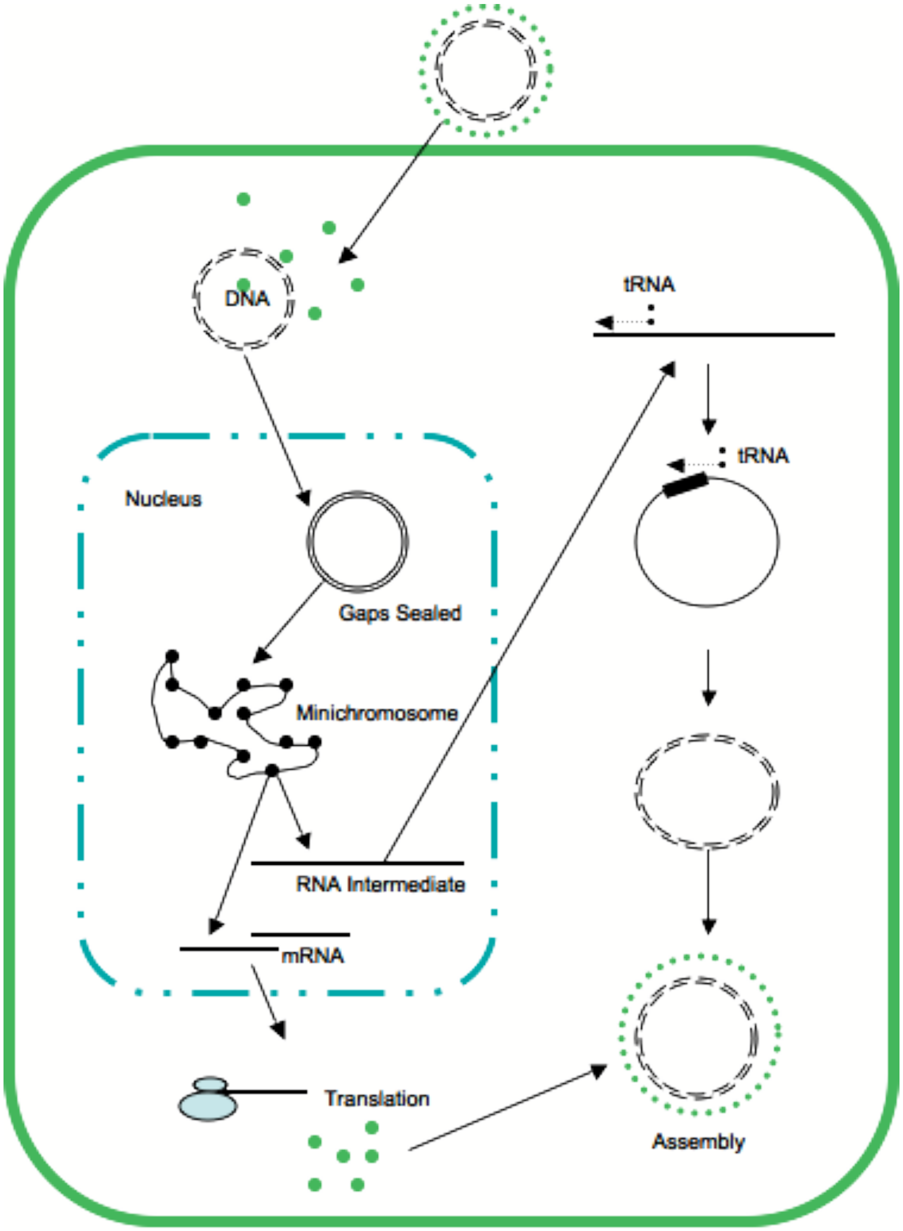


Figure 4. Replication cycle of a reverse transcribing dsDNA virus (Hull 2002).

1.1.3.9. Viral Propagation with the Host

Post replication, the virus must be able to successfully transport itself from the originally infected cell into the neighboring cells to continue the infection process.

Unlike bacterial or animal cells, plant cells cannot undergo a lytic phase in their viral infection due to the presence of the cell wall. Therefore, plant viruses have evolved specific proteins or groups of proteins that allows them to modify and move through the plants own intercellular transport system. This intercellular transport occurs by two methods: direct cell-to-cell movement via the plasmodesmata and long distance, systemic movement via the phloem.

Plasmodesmata, small, concentric tubes of plasma membrane and endoplasmic reticulum, connect plant cells together through the cell wall and serve as an important intercellular signaling system for plant tissues (Aaziz et al 2001; Ehlers et al 1999; Nelson and Van Bel 1998). Plasmodesmata also lead to the creation of symplastic domains within plant tissue, with the cellular processes of each cell in the domain completely synchronized (Hull 2002). They typically range from 2.5 to 3 nm in diameter and the size exclusion limit of an unmodified plasmodesmata structure ranges from 0.75 to 1.0 kDa (Wolf et al 1989).

Plant viral movement proteins work to enlarge the size exclusion limit of the plasmodesmata, in some cases from 9 to 17 times the original size exclusion limit (Wolf et al 1989). In addition, movement proteins have been shown to act as cellular localization signals for coat proteins of assembled viral particles for insertion into new host cells as well as generate microtubules within their host cell to facilitate the movement of viral particles from an infected protoplast or viroplasm to a plasmodesmata for cell-to-cell transfer (Kasteel 1999).

The phloem, the main transport system throughout the plant for water, metabolites, proteins, and other macromolecules(Thompson and Schulz 1999) travels through veins in the leaves and stems of plants that are in turn surrounded by sieve element cells. As the plant virus moves form cell to cell, it eventually moves through the sieve element cells and into the phloem (Nelson and Van Bel 1998). Once the phloem has been broached by the virus, it may move systemically throughout the plant.

1.1.4. Metagenomics and Viral Ecology

1.1.4.1. Overview of Metagenomics

The term metagenomics, first coined by Jo Handelsman (Handelsman et al 1998), refers to the isolation, sequencing, and analysis of genetic material recovered directly from an environmental sample. In opposition to traditional genomics, which require the removal of an organism from its natural habitat, isolated cultivation, and individual sequencing, metagenomics allows for the processing of a pool of samples in a cultivation free approach.

Metagenomics offers several advantages over traditional genomics. First, it does not rely on the ability to culture organisms in the laboratory. This is of particular importance as only approximately 1% or less of known microbial organisms are able to be cultured with current microbiological techniques. Secondly, given that the genetic materials are prepared directly from environmental samples, metagenomic studies present a relatively unbiased view of the community or pool being studied.

1.1.4.2. Plant Virus Ecology

As of 2005, there were approximately 2,000 known viral species (Fauquet et al 2005) although this is a tremendous under estimate (Breitbart et al 2004). Further support for this underestimation of viruses lies in the types of plants from which viruses have been identified as shown in Figure 5. Of the number of currently known viruses the majority, 77%, have been isolated from cultivated plant species with 11% being isolated from agricultural weeds, which have a direct impact on cultivated plants. Only 6% have been isolated from true wild growth. Therefore, it is quite possible for novel, unknown viral taxa to be present in the wild but remain unknown as it does not currently affect cultivated plants. This likely is because previous studies have focused on a single viral genus/family (Bodaghi et al 2004), a single group of host plants (Robertson 2005), or plants demonstrating outward symptoms (Ooi and Yahara 1999). Therefore, it is important for a survey for new and novel viral taxa and species to select samples from wild growth without regard for plant type or presence or absence of symptoms.

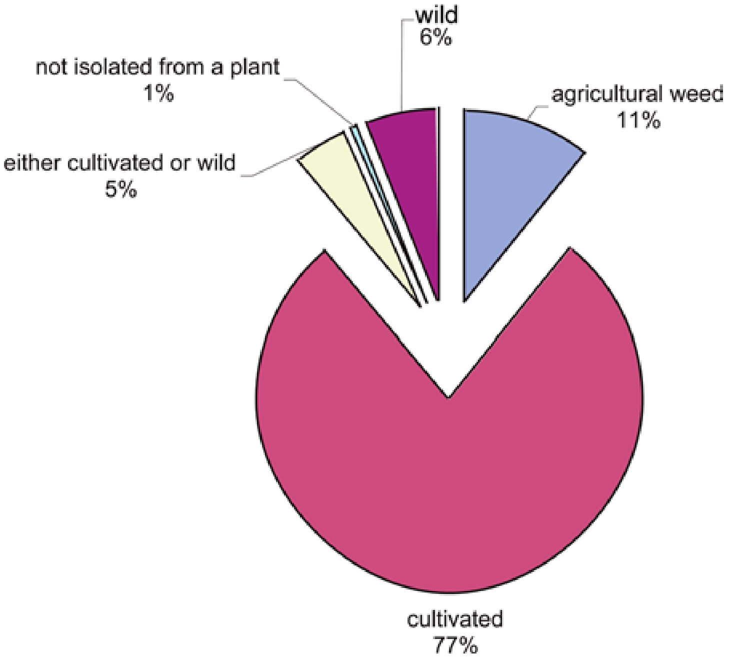


Figure 5. The frequency of recognized plant virus sources (Wren et al 2006)

1.1.4.3. The Tallgrass Prairie Preserve as a Model Plant Community

The Tallgrass Prairie Preserve, located in Osage county in northeastern Oklahoma, is a nature preserve operated by The Nature Conservancy since 1989 as a natural prairie habitat with semi-natural grazing and controlled burning.

1.1.5. Viral Taxonomy

1.1.5.1. Official Classification

The International Committee for the Taxonomy of Viruses (the ICTV) is the sole scientific body that controls virus taxonomy. In their 7th report, the ICTV defined viral species as "a polythetic class of viruses that constitute a replicating lineage and occupy a particular ecological niche" (van Regenmortel et al 2000). This implies that viruses, on the species level, share common traits but are not required to all share a single common trait. As attempt to classify viruses above the species level to family and genus level the traits required for membership become more universal. At the level of family, usually, viral classification ends because of the lack of a single common viral ancestor and the assumption that viruses originated from multiple sources (Holland et al 1998). Another difficulty in viral classification is due to recombination, gene rearrangement, mutation rate, and the general polyphyletic nature of viral genomes. In some rare cases families may be grouped into an order, but this is uncommon. The official taxa for viruses are: (Order), Family, (Sub-family), Genus, and Species (Fauquet et al 2005). Differing viral strains often are grouped, unofficially, within a species but these often are sufficiently similar between themselves to not warrant separation into separate species. A similar unofficial taxonomic classification is the grouping of viruses by genomic composition, be it double or single-stranded DNA or RNA and the presence or absence of a reverse transcription step in the replication cycle. This level is placed higher than order or family level and can be seen in use at the National Center for Biotechnology Information (NCBI) Genbank sequence repository (Benson et al 2007).

1.1.5.2. Classification Criteria

There are several criteria by which viruses may be grouped (Hull 2002) that have emerged as newer virological and molecular techniques have been developed. They are:

Virion Structure. Overall morphology of the assembled virus particle derived either by X-ray diffraction of crystallized virus or electron microscopy.

Physicochemical Properties. These include centrifugation measurements and buoyant density as well as viral particle stability in the presence of solvents such as ether and phenol.

Nucleic Acid Properties. As mentioned previously the makeup of the genome for a virus is a distinguishing factor. The organization of genes within the genome also contributes to this criterion. Recent advances in genomic sequencing have lead to the rise of full-length genomic transcripts for many plant viruses and increased the importance of genetic sequence to classification.

Viral Proteins. This includes the sequence of the coat protein, as well as their number and molecular weight. Other proteins encoded by the virus may be used for this criterion, but consideration of the coat protein is most prevalent.

Serological Properties. Prior to modern genetic sequencing the serological properties of the virus was used for the classification of viruses. This involves the use of cross-reactive antiserum generated against viruses of known species for demarcation.

This is similar to the more modern use of the amino acid sequence of the coat protein for taxonomic classification.

Biological Properties. This criterion includes activity within the host cell, method of transmission between hosts, host range and the now antiquated “cross protection” by which infection with one virus would imbue resistance to a second virus.

Given the number of possible criteria and the fact that the taxonomic delineations for each viral family and genus are respectively distinct, there is no one set of rules by which all viral species may be taxonomically classified.

1.1.6. Plant Fungal Viruses

As mentioned previously, fungi may act as a vector for plant viruses. However, the fungi themselves may harbor viruses capable of affecting overall plant health, either through attenuation (Zhou and Boland 1997) or exacerbation (Ahn and Lee 2001) of fungal virulence or the impartation of a mutualistic benefit for both the fungus and the plant (Marquez et al 2007). Fungal viruses are typically latent in terms of virulence towards their host (Lemke and Nash 1974) and are often capable of remaining within the infected host cell indefinitely (Banks et al 1969; Ghabrial 1980).

1.2. Hereditary Material and Organization

1.2.1. DNA

Deoxyribonucleic acid (DNA) is a long chain polymer comprised of nitrogenous bases, or nucleotides, bound to a sugar-phosphate backbone. The nucleotides attached to the sugar-phosphate backbone can be divided into two groups, the purines and the pyrimidines, based on the parent molecule they were derived from. The purines consist of adenine (A) and guanine (G). The two pyrimidines are cytosine (C) and thymine (T), shown in Figure 6.

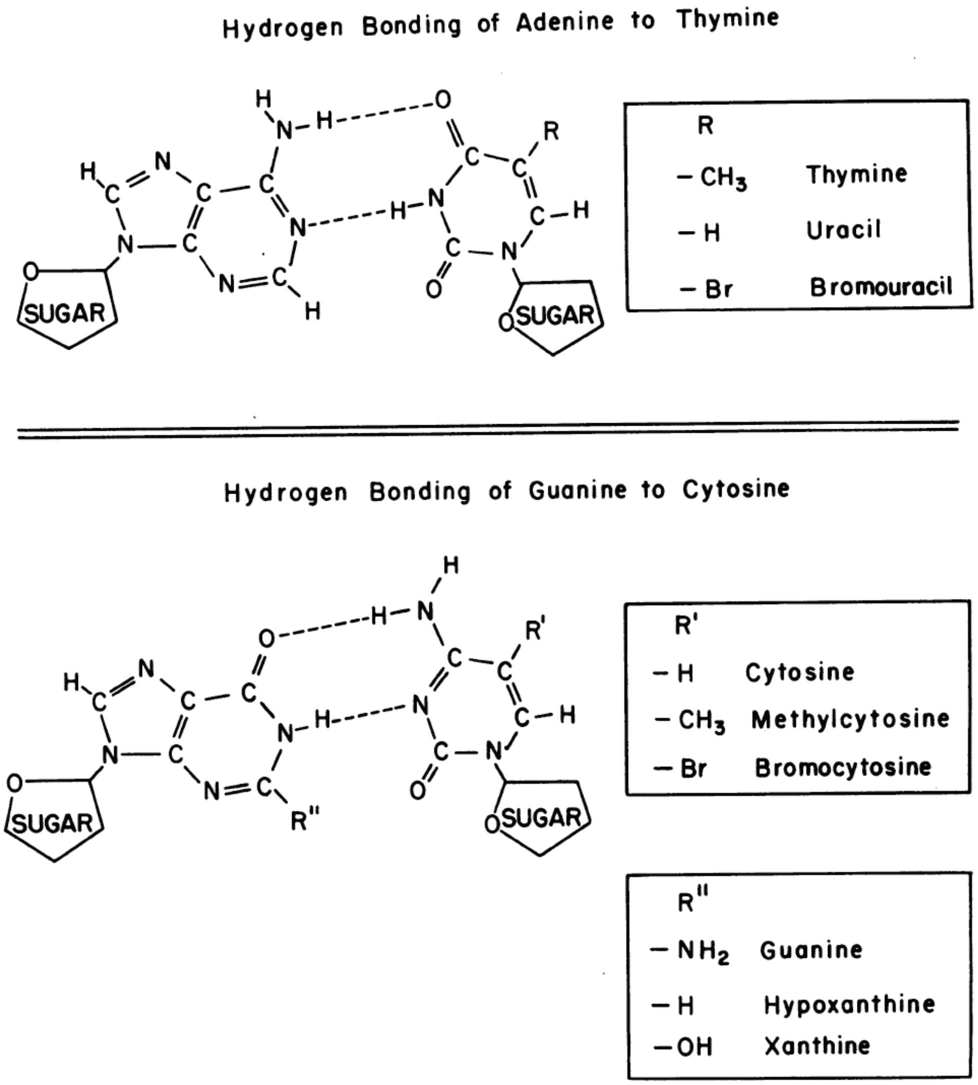


Figure 6. The Nucleotide Molecules of DNA (Bessman et al 1958)

These nucleotides are connected to the 2’-deoxyribose backbone through a glycosidic bond between C-1 of the sugar ring and N-1 in the case of pyrimidines and N-9 in the case of purines. The adjacent deoxyribonucleotide molecules are themselves connected together through a phosphodiester bond between the 3’ carbon of one molecule and the 5’ carbon of the next (Figure 7).

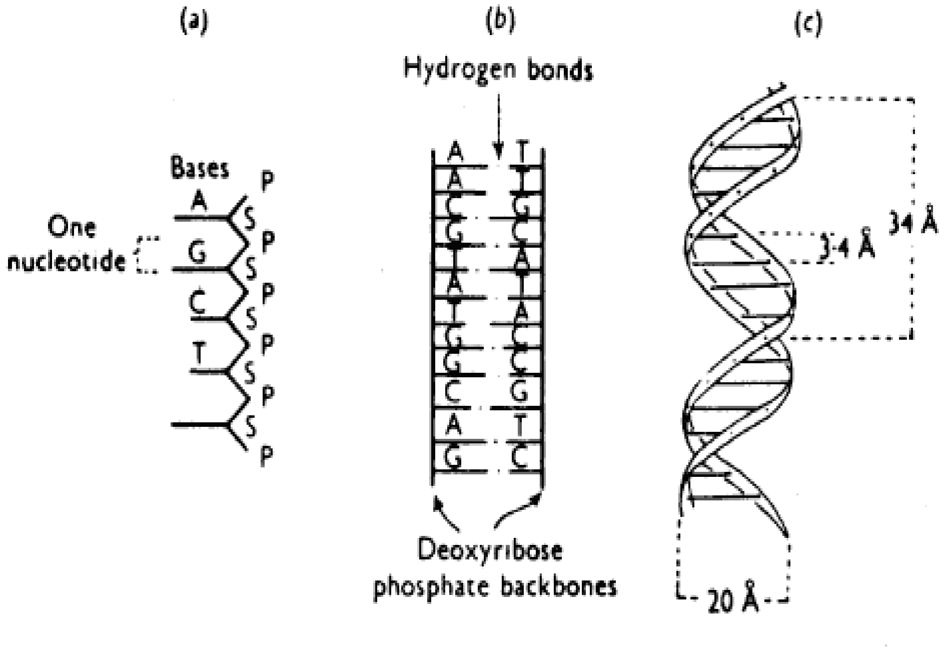


Figure 7. DNA structure (Hayes 1960)

When the complementary strand is present, single stranded DNA can form a double-stranded, anti-parallel, double-helical structure with the sugar-phosphate backbone surrounding an inner-core of specifically paired nucleotides as is the case in prokaryotic, eukaryotic, archaea, and some viral genomes. Watson and Crick first solved the structure of DNA (Watson and Crick 1953) and described the specific base pairing found between two self-complimentary strands, where adenine (A) typically binds with thymine (T) via two hydrogen bonds while cytosine (C) typically bonds with guanine (G) via three hydrogen bonds. The double helix is further stabilized by the hydrophobic interactions between the stacked bases in the center of the helical structure.

1.2.2. RNA

Ribonucleic acid (RNA) is very similar to DNA except for two distinctive differences. First, the sugar moiety of each nucleotide is comprised of ribose instead of

2’-deoxyribose and, second, the pyrimidine thymine is replaced with uracil, a pyrimidine lacking the methyl group at carbon 5 of the nucleotide ring.

RNA molecules can be found in either single stranded or double stranded form. Typically, as they form the messenger, mRNA, that is the intermediate in the information passage from the DNA genome and the protein synthesis apparatus of prokaryotes, archaea, and eukaryotes. This RNA, as well as, for example, transfer RNA, ribosomal RNA, U-RNA, and miRNA are single stranded. However, as mentioned above, many plant viruses have genomes consisting of double stranded RNA or single stranded RNA in the positive or negative sense.

1.2.3. Genes

A classic definition for a gene is a sequence of DNA that is converted to RNA. Crick first described this in 1970 when he proposed the “Central Dogma of Molecular Biology” (Crick 1970). The central dogma states that genetic information is carried from DNA to protein sequence via RNA. This has since been expanded somewhat as shown in Figure 8. This does not hold in the case of viral RNA genomes. In these cases the genome is by its very nature ready for translation of encoded genes into protein or other biologically relevant molecules (Thivierge et al 2005). Whether encoded by a DNA or RNA genome, the products of genes can take on many forms such as enzymes, structural proteins, or genomic regulatory proteins.

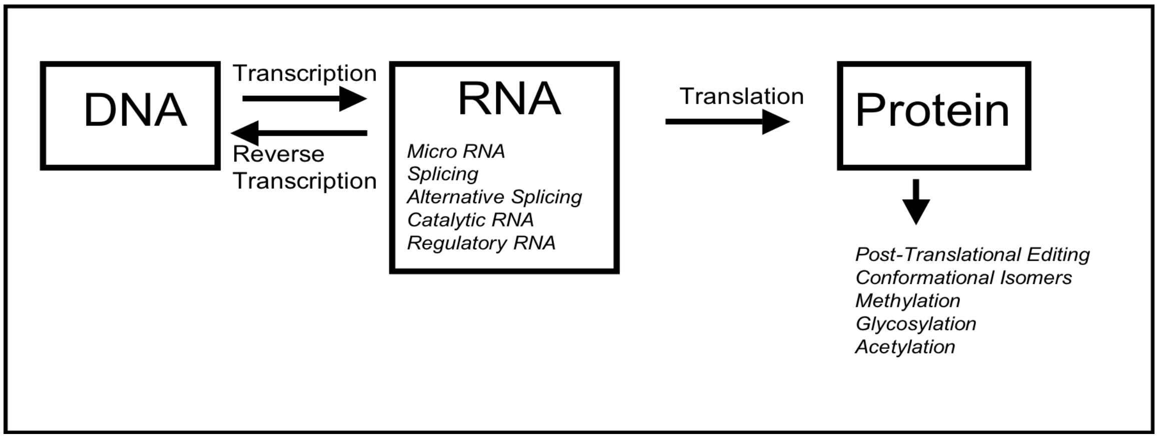


Figure 8. The Central Dogma of Molecular Biology

In biological organisms there are several ways a that gene can be organized. These differences highlight basic genomic regulatory models for prokaryotic, archaea, and eukaryotic organisms. In eukaryotes genes are typically broken up into multiple pieces, known as exons, separated by lengths of sequence known as introns. Except for a few rare exceptions, genes in prokaryotes (bacteria), archaeal genomes, and most plant viruses are monolithic in nature. Eukaryotes also are monocistronic, meaning one gene is transcribed per mRNA generated, while in both prokaryotes (bacteria) and archaea the genes often are polycistronic, with often more than one gene being transcribed per mRNA generated. Plant virus genes typically follow this prokaryotic system, although some viruses have been characterized to have more than one protein or enzyme encoded in a single translational event, which then are separated by post-translational cleavage, typically by self-encoded protease enzymes (Figure 9).

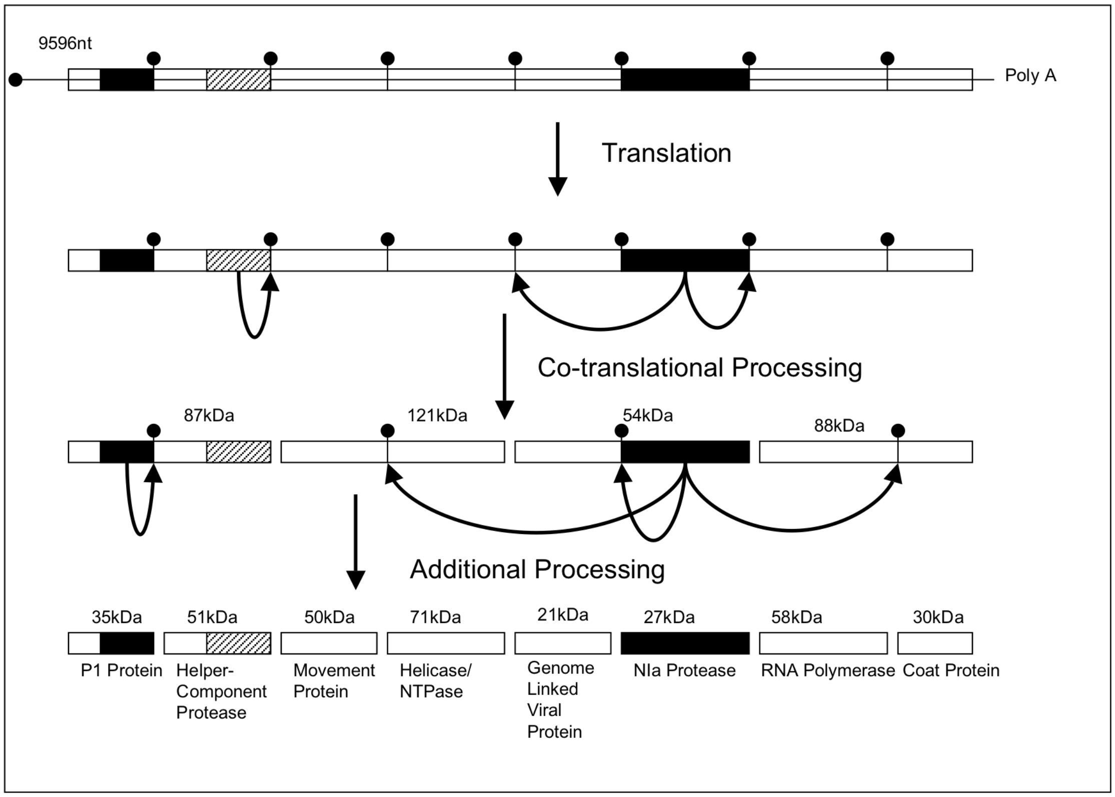


Figure 9. A Schematic Representation of the Post Translational Self-Cleavage of the Potyvirus Tobacco Etch Virus. The P1 serine and NIa serine-like proteases are represented by black rectangles while the Helper Component (cysteine) Protease is delineated by a diagonally striped rectangle. Arrows point to protease cleavage sites while final gene products are shown at the bottom. (Dougherty and Semler 1993).

Self-proteolysis is an example of adaptation by the plant viruses, as a prokaryotic coding system, in their use of the eukaryotic expression system of their plant hosts. Further examples of this include the generation of sub-genomic RNA molecules during genomic replication or the use of multipartite genomes (Karasev et al 1997), leaky scanning by the ribosomal proteins in which translation does not always begin at the first AUG codon (Fütterer and Hohn 1996), non-AUG start codons (Shirako 1998), ribosomal shunting from one initiation site to another (Dominguez et al 1998), and the use of slippery codons to generate reading frame shifts (Prüfer et al 1992).

Genes typically encoded by plant viruses can be broken down into the following classes: polymerases, coat proteins, cell-to-cell movement proteins, 5' associated/VPg/genome-linked viral proteins, helper components, and proteases (Zaccomer et al 1995).

1.2.4. Genome

A genome the complete set of genetic material for an organism. While double stranded DNA often is the main component of genomic material, in plant viruses this often is not the case. Plant viruses are widely varied across the range of possible genome organization in the use of hereditary material. Plant virus genomes may consist of reverse transcribing double stranded DNA (family *Caulimoviridae*) (Hohn and Fütterer 1997), single stranded DNA (family *Geminiviridae*) (Buck 1999), double stranded RNA (family *Partitiviridae*) (Osaki et al 2002), negative-sense single-stranded RNA (family *Rhabdoviridae*) (Jackson et al 2005), or positive-sense single-stranded RNA (family *Bromoviridae*) (Ahlquist 1999). There currently are no known plant viruses having a pure double stranded DNA genome that do not include replication via an RNA intermediate. Plant virus genomes need not be monolithic, with several plant virus families having multi-component genomes. The viruses classified in the family *Reoviridae* are known to contain 9-12 double stranded RNA genome segments (Attoui et al 2005).Plant virus genomes also need not be packaged in the same virion particle, as in the case of the family *Bromoviridae* (Ahlquist 1999).

1.3. Methods for the Study of Plant Viruses

There are several methods by which plant viruses may be studied. These are illustrated in Figure 10.

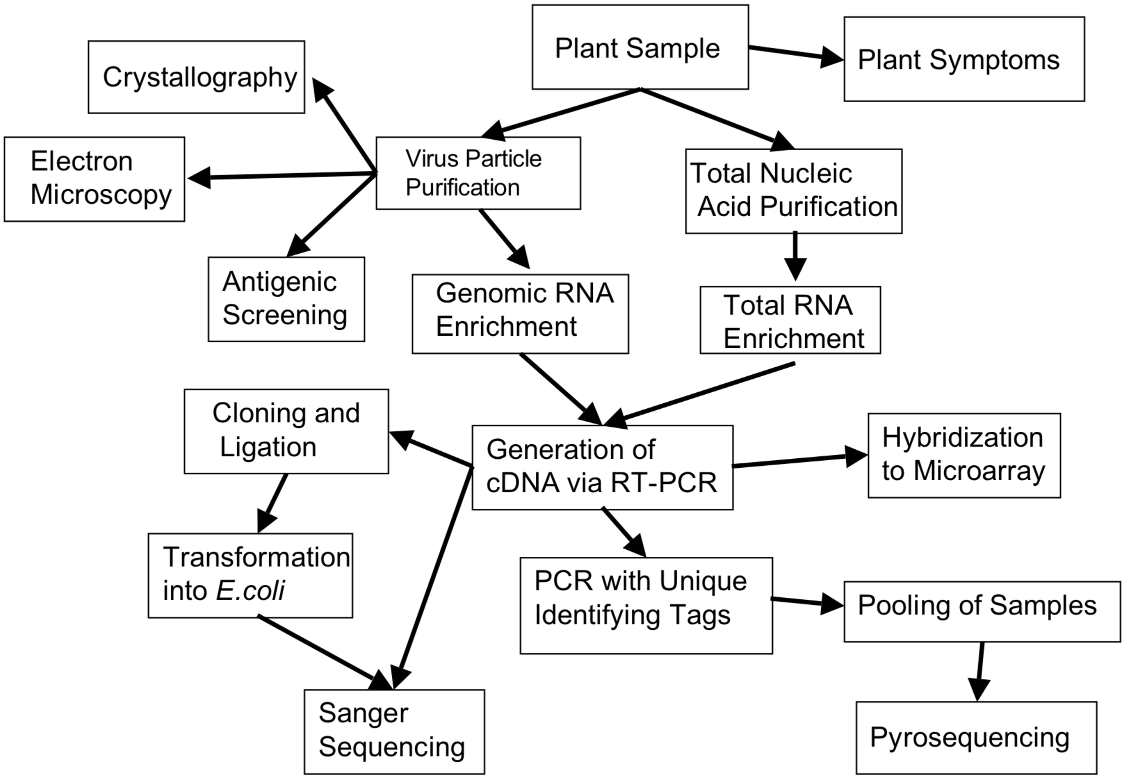


Figure 10. Approaches to Plant Virus Study

1.3.1. Classical Virology Techniques

While plant viruses have been present throughout history, the first to isolate a plant virus, Tobacco Mosaic Virus, was Dmitrii Iwanowski. Through the use of ceramic filters, Iwanowski was able to purify viable plant viruses though it took further work by Martinus Beijerinck to characterize the nature of the purified virions (Zaitlin 1998). Work on plant viruses in the early 20th century was mainly focused on infection symptoms, purification and crystallization of viral particles, electronic microscopy or x-ray diffraction of viral particles, and antigenic studies of viral particles. It was not until the mid-1950s that the encapsulated RNA genome was demonstrated to be the genetic component of the virus and the coat protein was merely a protective shell a discovery that ushered in the era of modern plant virology.

1.3.2. Virus Purification

The main technique for the purification of viral particles from plants is centrifugation (Hull 2002). Centrifugation may be further broken down into three distinct methods: differential, rate zonal, and isopycnic. Differential centrifugation, as its name implies, relies on the differences in sedimentation coefficients of the particles being centrifuged to separate them and initially was developed for the purification of tobacco mosaic virus and tobacco ringspot virus (Stanley and Wyckoff 1937). In a typical differential centrifugation purification the centrifugation occurs at both low and high speeds. The low speed centrifugation sediments contaminating proteins and cellular debris, while the high speed centrifugation pellets the viral particles. One of the drawbacks of this method is that washes usually are employed after the low speed centrifugation steps to resuspend inadvertently sedimented viral particles from the pellet. This can lead to dilution of the sample. Also, differential centrifugation is only useful to purify viable virus free from contaminating cellular debris

Rate zonal centrifugation separates particles based on their relative sedimentation rates (Brakke 1960). In this method the particles to be separated are placed in a thin band at the top of a density gradient. This gradient prevents the convection of the sample as it is being centrifuged as well as providing a selective mechanism based on buoyant density. Once centrifugation begins, the particles move into the gradient at their respective rates. The sedimentation rate of each particle is based on several factors, including the centrifugal force at each area of the gradient, overall size of the particle, the effects of surface area and shape of the particle in terms of viscous drag, and the difference in density of the particle versus the gradient of the medium. The largest particles will sediment the fastest, as will those with low amounts of viscous drag, typically those particles closest to spherical in shape.

Isopycnic centrifugation, sometime called sedimentation equilibrium centrifugation, utilizes a buoyant density gradient to separate macromolecules based on their relative densities. Unlike rate zonal centrifugation, the virus to be purified initially is mixed throughout the gradient. During centrifugation, the virus will both sediment and float to its equilibrium position within the gradient based on its buoyant density. This method, however, can take up to 7 days to complete for a CsCl gradient and is dependent on the medium being able to form such a gradient *in situ* (Shepherd, Kado et al. 1972). This is a particular problem with highly viscous, less dense sucrose gradients, which are therefore impractical and preformed gradients usually are used.

1.3.3. Isolation and Manipulation of Viral Nucleic Acid

1.3.3.1. Nucleic Acid from Viral Particles

One for the most direct methods for isolating viral nucleic acids is to purify it directly from the viral particle. This first entails purifying the viral particle, usually via centrifugation methods. Once the viral particles have been isolated, they are treated with a proteinase (such as Proteinase K) to remove the coat protein. The resulting nucleic acids then are purified from the degraded coat proteins via phenol/ether extraction (Melcher et al 2008).

1.3.3.2. Total Nucleic Acid Extraction

If the plant virus to be studied has an RNA genome, an effective way to obtain the genomic RNA is to extract the entire nucleic acid content from a host cell. Once the nucleic acid has been purified, double-stranded RNA may be purified by first treating the nucleic acids with DNase enzymes and passing the remaining nucleic acids through a CF11 cellulose chromatography column, which has been shown to preferentially bind double stranded RNA in the presence of 16.5% Ethanol (Pellegrin et al 2007; Semancik 1986).

1.3.3.3. Reverse Transcription-PCR

As the majority of plant viruses have single or double stranded RNA genomes, the difficulties of working with RNA, and the incompatibility of RNA with many molecular techniques the genomic and sub-genomic molecules of a plant virus must be reverse transcribed from RNA to DNA using the reverse transcription polymerase chain reaction (RT-PCR) (Goelet et al 1982). To generate the first strand of DNA it is necessary to provide a primer complementary in sequence to the original strand of RNA. This can be done in one of two ways. If the virus is of known sequence, a primer with a sequence unique to the virus may be used (Nassuth et al 2000). If the virus is unknown a primer with a 4-10 random nucleotides on its 3’ end may be annealed and used as a first strand priming point (Marquez et al 2007). The use of random, sequence independent RT-PCR protocols on un-manipulated samples has been termed viral metagenomics (Delwart 2007).

Once the RNA genome has been transcribed to DNA it can be sequenced directly using the RT-PCR primers, ligated and cloned into a bacterial vector for amplification and storage, or it may be prepared for pooling using uniquely 5’-tagged PCR primers for deconvolution (Binladen et al 2007).

1.3.4. Sequencing Methods

1.3.4.1. Sanger Dideoxynucleotide Sequencing

Developed by Frederick Sanger (Sanger et al 1977), the dideoxynucleotide or chain termination method involves the electrophoretic separation of a nested DNA fragment set generated by a chain termination of DNA replication, and yields single base resolution. The nested fragment set is generated by a DNA polymerase enzyme genetically engineered to have no proofreading function (Lehtinen and Perrino 2004; Maki and Kornberg 1987) synthesizing complementary strands of a single DNA molecule that are prematurely terminated through the incorporation of one of the four dideoxynucleotide triphosphates uniquely labeled with fluorescent molecules. The individual nucleotide fragments in the nested fragment set then are separated on a polyacrylamide gel matrix (either on a slab gel or in a capillary). As the DNA moves through the matrix a laser excites the fluorescent molecule attached to the chain terminating dideoxynucleotide causing it to fluoresce a color coded to the identity of the terminating molecule. The emitted light then is captured by a CCD camera and plotted to a chromatogram. The individual peaks of light then are used to determine the sequence of the bases for that particular read (Dovichi 1997).

1.3.4.2. Pyrosequencing

The pyrosequencing method, first developed in the late 1980s (Bains and Smith 1988; Jett et al 1989), has been massively parallelized through emulsion PCR (emPCR) (Dressman et al 2003; Ghadessy et al 2001; Margulies et al 2005) and now has become very useful for direct sequencing and metagenomic studies of eukaryotic genomes (Wheeler et al 2008), bacterial genomes (Margulies et al 2005), bacterial communities (Turnbaugh et al 2009), phage communities (Desnues et al 2008), and metabolomic/transcriptomic studies (Zou et al 2008).

Pyrosequencing consists of two core techniques, emulsion PCR and the pyrophosphate-based sequencing reaction. In emulsion PCR the DNA to be sequenced first was sheared into small, sub 1kb lengths. Adapters containing universal primer binding sites and recognition sequences then were annealed to each of these sheared molecules. The adapter ligated sequences then were combined with an aliquot of beads coated with a molecule of DNA complementary to the universal primer binding site as well as a PCR amplification mix. This DNA/bead/PCR amplification mixture was placed into a tube containing oil and shaken vigorously to generate a water-in-oil emulsion of small micelles, each containing the necessary reagents for a single PCR reaction. The emulsion then was placed into a thermocycler where the amplification of the adapter-ligated DNA by the primer coated bead produced a DNA bead coated with up to a million copies of a single DNA molecule (Margulies et al 2005).

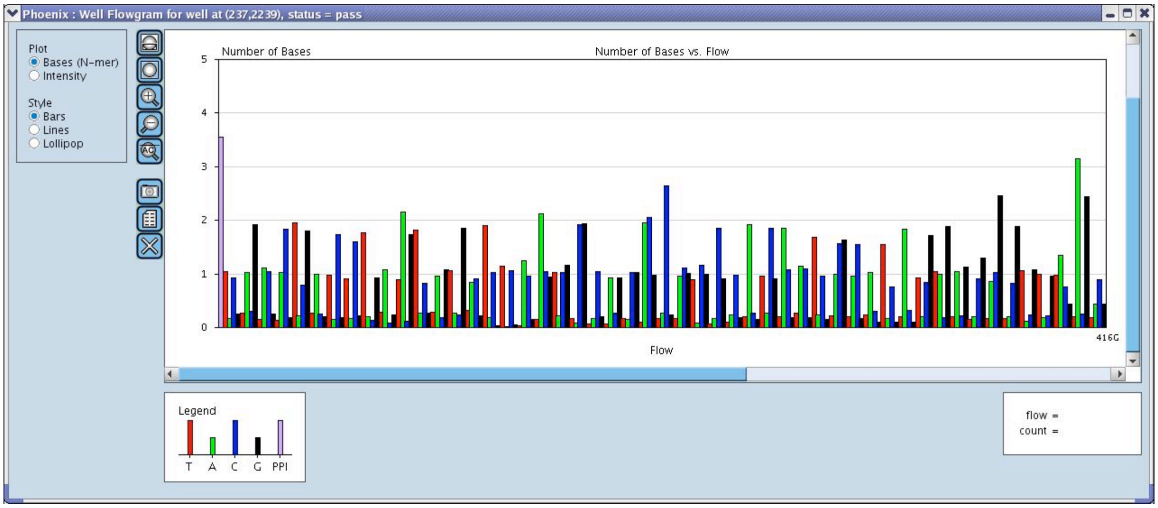
After PCR cycling the emulsion was broken and those beads coated in amplified DNA were separated from those that were not. During this purification process the double stranded DNA covalently attached to the bead was denatured to produce covalently attached single stranded DNA to which a sequencing primer was annealed. The prepared, purified beads then were mixed with DNA polymerase and then were loaded into an etched fiber-optic slide along with other, smaller beads that contained covalently bound sulfurylase and luciferase enzymes. This plate then was placed into the pyrosequencing machine and individual nucleotides were flowed one at a time across the plate. If the polymerase incorporates a passing nucleotide, pyrophosphate was released. This pyrophosphate molecule subsequently was transformed via sulfurylase to ATP by the following reaction:

$$\text{AMP} + \text{PP} \rightarrow \text{ATP}$$

The ATP molecule then is used to oxidize luciferin via the enzyme luciferase in the following reaction:

$$\text{ATP} + \text{Luciferin} \rightarrow \text{Oxyluciferin} + \text{PP} + \text{AMP} + \textit{light}$$

The output and intensity of light then is read by a CCD camera. Base incorporation is calculated as the light output was linearly relational (Margulies et al 2005) and was used to generate a “flowgram”, an example of which can be seen in Figure 11, for each well in the picotiter plate.



Phoenix : Well Flowgram for well at (237,2239), status = pass

Plot
● Bases (N-mer)
○ Intensity

Style
● Bars
○ Lines
○ Lollipop

flow =
count =

Figure 11. A flowgram showing the partial sequence from a single well of a pyrosequencing run.

1.4. Optimization and Automation of Pyrosequencing Protocols

Despite being high-throughput, the current pyrosequencing methodology was quite labor intensive and required a large amount of sample manipulation prior to the actual sequencing step. Although the manufacturers methods for sample library preparation were quite robust, I have incorporated several changes as well as eliminated several extraneous steps that resulted in a streamlined process that could be automated (Wiley et al 2009).

The initial modification was the replacement of Qiagen spin column purification after each enzymatic step with Agencourt Ampure Solid Phase Reversible Immobilization (SPRI) beads. The use of SPRI beads over silica mini-columns has two significant advantages. The first is the overall yield is higher when using SPRI beads, at 90-95%, than the mini-columns at 80-85%. Secondly, by varying the volume of SPRI bead suspension mixed with DNA solution, it is possible to selectively purify fragments over 300bp in size. As shorter fragments preferentially amplify during emPCR this significantly improves read length average and the number of mixed reads in the final pyrosequencing step.

The second modification was the removal of the steps for generating a single stranded library molecule while enriching for molecules containing only A and B adapters ligated to either end. As the molecules with A on both ends will not amplify properly in emPCR and the molecules with B on either end will not enrich post-emPCR and thus this step was deemed unnecessary.

These modifications facilitated the subsequent automation of the library preparation process on a Caliper SciClone ALH with a Twister II plate positioner programmed to add, move, and remove buffers, enzymatic mixtures, and SPRI bead suspensions as well as move the reaction plate to various stations within the robot. This automation allowed for a walk-away process in which no human manipulations are required except for the preparation of the robot and enzymatic mixtures.

1.5. Computational Methods for DNA Analysis

1.5.1. Sequencing Data Assembly and Analysis

Sequencing data generated by the 454 was assembled using the 454 GS De Novo Assembler program (http://www.454.com/products-solutions/analysis-tools/gs-de-novo-assembler.asp). This program reads the flowgram for each well and aligns them into consensus sequences in "flow space" to generate a final consensus sequence.

1.5.2. Homology Detection

Once assembled, the homology of the resulting sequence contigs generated for each sample was ascertained through the use of the Basic Local Alignment Tool (Blast)

(Altschul et al 1990). The Blast program uses a heuristic approach based on the Smith-Waterman algorithm to determine local alignment of either nucleotides or proteins while determining the statistical significance of each alignment. A scoring matrix, built into the Blast program, determines similarity, with a positive score being given for a residue match and a negative score being given for a mismatch or sequence gap, while the overall similarity score was determined by summing all the similarity scores for the entire length of the contiguous, or gapped, aligned sequence segment. These segments then were extended to either side of the local alignment to provide the overall optimal sequence alignment. Those regions with the highest scoring identical lengths, Maximal Segment Pairs (MSPs), as determined by the length of the query sequence, the non-randomness of the match, the scoring matrix used, and the size of the database, above a certain, specified scoring threshold were displayed.

Blast is available in several iterations, each of which may be used to align sequences in a specific manner against other sequences or databases of sequences. The two most often used Blast types in plant virus sequence analysis were nucleotide-nucleotide comparisons using BlastN and translated nucleotide–amino acid comparisons using BlastX. Other forms of Blast include BlastP which covers amino acid–amino acid comparisons and tBlastX which covers translated nucleotide-translated nucleotide comparisons. Finally Reverse Position-Specific Blast (RPS-Blast) was used for conserved domain searches as it uses an initial similarity search to generate a Position Specific Scoring Matrix (PSSM) which then is used in a second similarity search to identify more instances of homology (Marchler-Bauer and Bryant 2004).

1.5.3. Sequence Alignment

After determining homologies for each contig, those contigs which appear to show homology to similar sequences may be aligned using sequence alignment programs such as ClustalW (Thompson et al 1994) and Blast2seq (Tatusova and Madden 1999). Phylogenetic trees generated by ClustalW then could be viewed using the program Treeview X (Page 1996).

1.5.4. RNA Tertiary Structure Prediction

For those contigs with frame shifts in their coding sequence the RNA folding topology was determined using RNAfold, a 2-dimensional RNA structure prediction program (Hofacker et al 1994). RNAfold calculates the structure of an RNA molecule by searching for the arrangement of loops and external bases that minimizes the sum of the free energy of loops contained within the overall secondary structure

1.5.5. Metagenomic Data Analysis System

Metagenomic studies, by their nature, generate vast amounts of data. Furthermore, in the case of genomic surveys such as this study very few if any full genomes were sequenced completely as the majority of the resulting contigs varied greatly in size

Current publicly available genomic databases provide powerful analysis tools but can be time consuming in accession and comparison. This can be compounded by poor database curation, mis-labelling, redundancy, and lack of database specificity. Therefore to further the analysis of the data generated from the Tallgrass Prairie

Preserve a database system, TGPweb, was generated using Perl, Hypertext Markup Language (HTML), MySQL, and Apache as shown in Figure 12.

Perl (http://www.perl.com) is a powerful, versatile programming language often used in the field of bioinformatics. HTML (http://www.w3.org/TR/html401/) is the predominant language for crating webpages. MySQL (http://www.mysql.com) is a relational database system running as a server which provide multiple-user access to tables of data related through primary and secondary keys. Apache (http://www.apache.org) is an HTTP/1.1 compliant web server which allows interaction between databases and the user.

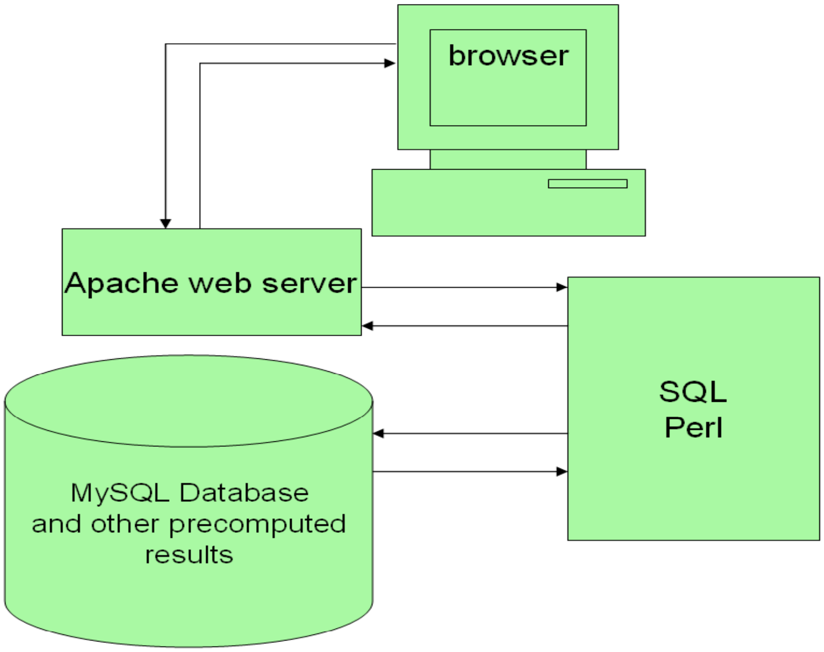


Figure 12. Database architecture of TGPweb.

Chapter 2 Materials and Methods

2.1. Creation of an Automated Pyrosequencing Library Preparation Robot

A Caliper SciClone ALH robot equipped with a Twister II plate positioner was programmed to carry out the library preparation protocol as given in Section 2.5.1 using the Clara software suite (CaliperLS). The plate deck on the SciClone ALH was configured as shown in Figure 13.

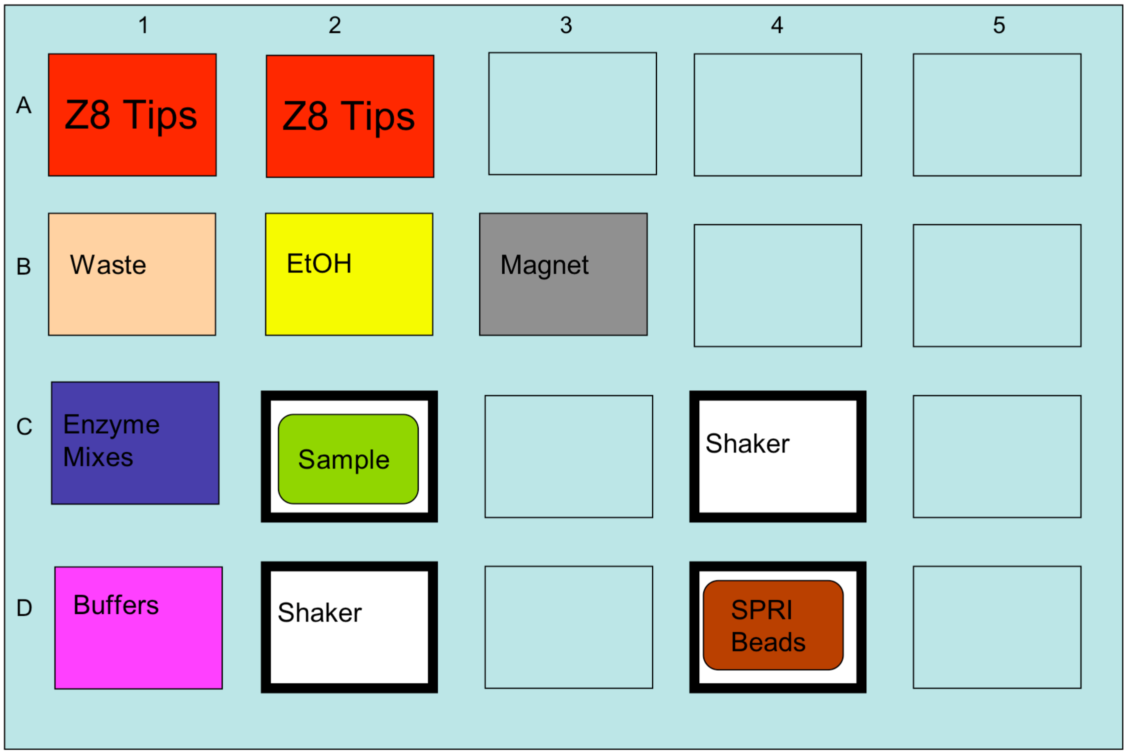


Figure 13. Caliper Sciclone ALH deck arrangement.

The boxes marked Z8 Tips were VWR ZT-100-R tip racks with rows A and H removed. Waste was an empty reservoir. Ethanol was a reservoir filled with 95% ethanol. Magnet was a custom fabricated magnetic separation station for 96 well plates as shown in

Figure 14. Enzyme Mixes were premixed enzymatic reaction solutions, given in Table 1, in a custom fabricated chiller apparatus, shown in Figure 15. Sample was a 96 well plate with up to 12 samples placed as shown in Figure 16. Buffers was a deep well block with 500 ul of 10 mM Tris-HCl pH 8.0 in all wells of column 12 but A and H. SPRI Beads was a deep well block with 200 ul SPRI beads (Agencourt #A29152) in all wells of column 1 but A and H.

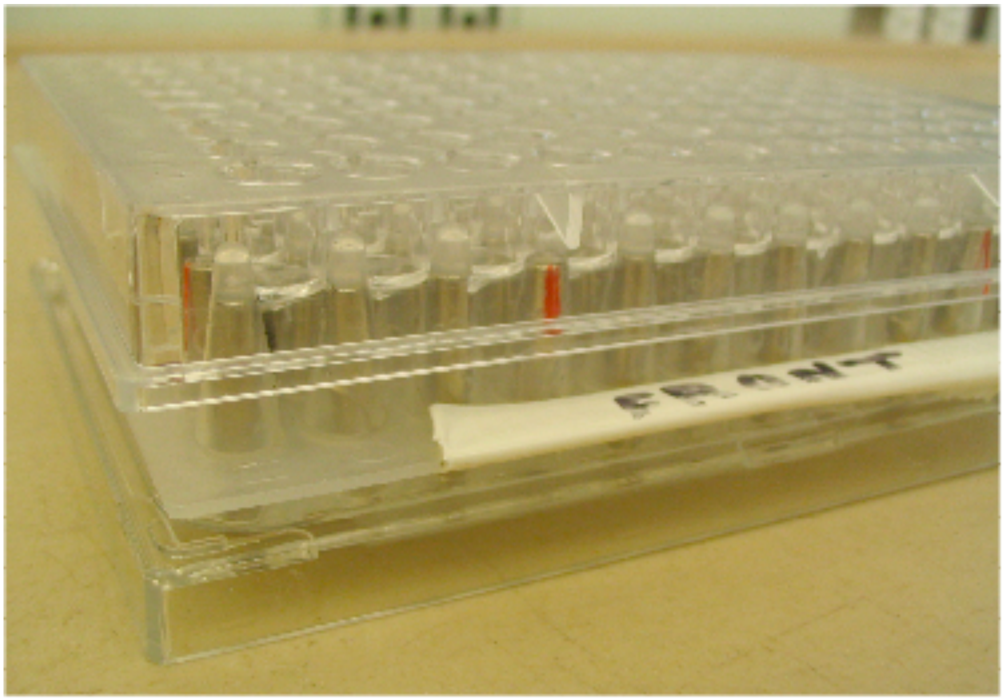


Figure 14. The custom fabricated magnetic separation station for 96 well plates

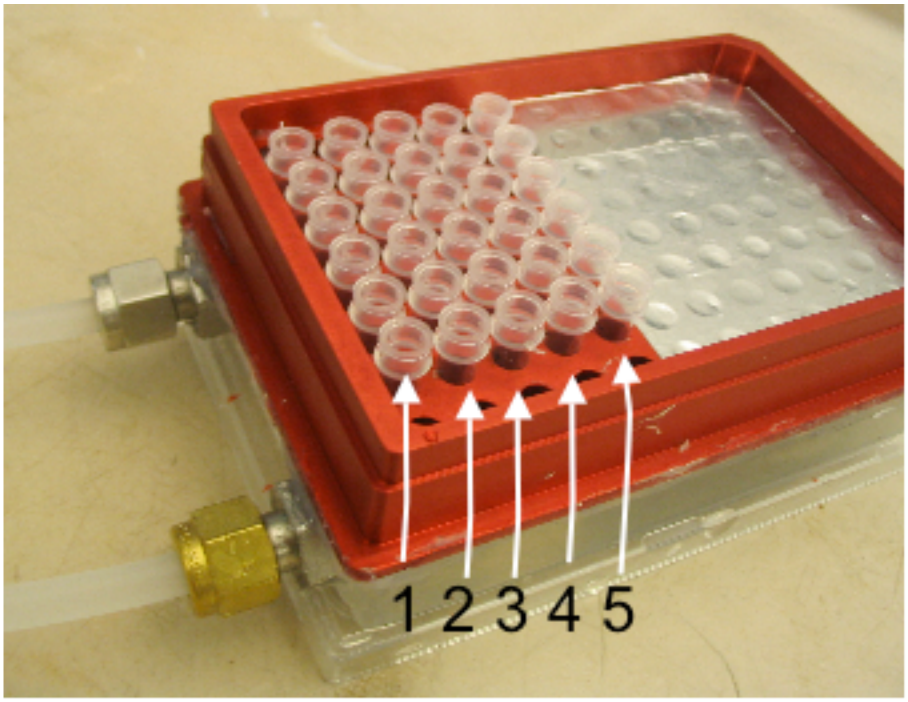


Figure 15. The custom fabricated enzymatic chilling station.

Table 1. Enzymatic mixtures for the preparation of 454 pyrosequencing libraries as placed in the SciClone enzymatic chilling station.

	Strip 1	Strip 2	Strip 3	Strip 4	Strip 5
Purpose	Polishing	MID 1-6	MID 7-12	Ligase	Fill-in
Contents	10ul 10X Buffer 10ul BSA 10ul ATP 4ul dNTPs 10ul T4 PNK 10ul T4 DNA pol.	10ul 2X Buffer 5ul MID adapter	10ul 2X Buffer 5ul MID adapter	20ul 2X Buffer 8ul Ligase	30ul ddH2O 10ul 10X Buffer 4ul dNTPs 10ul Fill-in pol.
Total	54ul	15ul	15ul	28ul	54ul

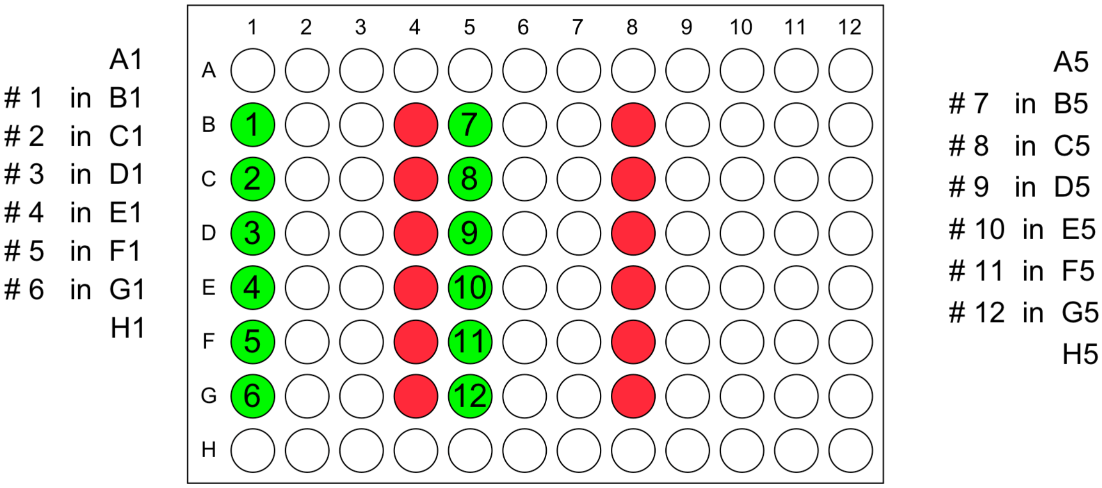


Figure 16. Sample layout for a 96 well plate containing 12 samples for generation of a pyrosequencing library. Green wells indicate sample starting position, red wells indicate finished library position.

2.2. Sampling of Plant Tissue

Young leaves from each sampled Tallgrass Prairie plant were cut into pieces smaller than 0.2 sq cm using a sterile razor blade and placed into a sterile tube. This tube then was placed on wet ice and transported to the Noble Foundation where it was processed for dsRNA.

2.3. Double-stranded RNA Isolation from Plant Tissue

At the Noble Foundation the amount of sample tissue was weighed, ground in liquid nitrogen until completely pulverized, transferred to a 50 ml tube containing 2 ml extraction buffer (0.1 M NaCl, 50 mM Tris-HCl, pH 8, 1 mM EDTA, pH 8, 1% SDS) and 2 ml phenol:chloroform per gram of sample, mixed for 10 minutes at room temperature and centrifuged for 10 minutes at 3200 x G . The aqueous phase was removed to a second tube and the phenol:chloroform extraction repeated by adding an equal volume of phenol:chloroform and centrifuging again. The aqueous phase then was removed to a 12 ml Falcon tube and the appropriate amount of 100% ethanol was added to create a final ethanol concentration of 16.5%. This mixture then was added to a spin column containing 100 mg CF11 cellulose per gram of original tissue, mixed thoroughly, centrifuged for 30 seconds at 2,000 rpm and the eluent was discarded. The column then was filled with application buffer (0.1 M NaCl, 50 mM Tris-HCl, pH 8, 0.5 mM EDTA, pH 8, 16.5 % ethanol) and centrifuged again. This wash was repeated 6 times and then the column then was removed to a clean Falcon tube and 4.5 ml of elution buffer (0.1 M NaCl, 50 mM Tris-HCl, pH 8, 0.5 mM EDTA, pH 8) was added. The column was centrifuged once again to collect the eluent in the clean Falcon tube.

Nucleic acid then was precipitated by adding 500 ul of 3M sodium acetate (NaOAc) and 10 ml of 95% ethanol and then incubating at -20^{o}C overnight. The following day the tube was centrifuged for 15 minutes at 3200 rpm and the supernatant discarded. The resulting pellet of nucleic acid was resuspended in 50 ul 0.1 mM EDTA.

2.4. Tagged RT-PCR of Double-Stranded RNA

A reverse transcriptase mix first was prepared for 8 reactions by mixing the following volumes in a clean tube:

32 ul 5x Superscript buffer (Invitrogen #18080)

16 ul 0.1 M DTT (Invitrogen #P2325)

8 ul 10 mM dNTPs (Invitrogen #18427013)

8 ul Superscript III reverse transcriptase (Invitrogen #18080)

In a separate, clean tube the following were mixed:

1 ul sample RNA

1 ul 10 mM Tris-EDTA

2 ul 20 uM RT random primer (5’CCTTCGGATCCTCCN$_{12}$3’)

8 ul H_2O

This mixture then was placed in boiling water for 2 minutes and chilled on ice for 2 minutes at which time 8 ul of the reverse transcriptase reaction mixture was added and the tube was chilled for 15 minutes on ice. The tube then was incubated at 50°C for 1 hour. After incubating 1 ul of 10/mg/ml RNase A was added and the tube incubated for 15 minutes at room temperature and then 85°C for 15 minutes. 100 ul of PBI buffer from a Qiagen PCR purification kit then was immediately added and the mixture then was placed in a Qiagen spin column and centrifuged at top speed for 1 minute. Column eluent was discarded and 750 ul 35% guanidine HCl in water was added to the column that was again centrifuged at top speed for 1 minute. Eluent was again discarded and 750 ul PE buffer was added to the column that again was centrifuged for 1 minute. The spin column then was placed into a clean 1.5ml Eppendorf centrifuge tube and 30 ul 0.1x EB buffer was added to the column before centrifuging the column for 1 minute at top speed. The eluent was used as PCR template in the following reaction where 8 reactions required a mix made by combining:

84 ul H_2O

12 ul NEBuffer 4 (New England Biolabs #B7004S)

2 ul 10 mM dNTPs (Invitrogen #18427013)

2 ul Taq polymerase

In a second tube the following were combined:

12.5 ul of the above reaction mix

1.5 ul template DNA

1.5 ul 10 uM tagged primer (5'XXXXCCTTCGGATCCTCC3' where X is a 4 nucleotide tag as given in Table 2

Table 2. List of the 96 – 4 nucleotide tags used in this project

AGAG	ATCA	GTAC	TCGT
ACTC	ATCG	GCAC	TCGC
AGTG	ATGT	GCAG	TCGA
ATAG	ATGA	GCAT	TGAT
ACAC	ATAC	GCTC	TGAC
CACA	ATCT	GCTG	TGCA
CTCT	ACAG	GCGT	CTAT
CAGA	ACAT	GCGC	CTCA
CTGT	ACTA	GCGA	CTCG
ATGC	ACGT	GAGT	CTGC
GAGA	ACGA	GAGC	CTAG
GTGT	ACGC	GACT	CTAC
GACA	AGAT	TATA	CGCG
GTCT	AGAC	TACA	CGCT
GATC	AGCA	TACG	CGCA
TCTC	AGCT	TAGC	CGAG
TGTG	AGCG	TAGT	CGAC
TCTG	AGTA	TAGA	CGTA
TCAC	GTAT	TATG	CGTC
TGAG	GTCA	TATC	CGTG
CTGA	GTCG	TACT	CAGT
ACTG	GTGC	TCAG	CAGC
CGAT	GTGA	TCAT	CACT
GCTA	GTAG	TCTA	CACG

The tube then was placed into a thermocycler and incubated for:

one cycle of 94°C for 1 minute, 72°C for 2 minutes; 40 cycles of 94°C for 5 seconds, 65°C for 5 seconds, 45°C for 5 seconds, and 72°C for 30 seconds; 72°C for 5 minutes; 37°C for 5 minutes.

2.5. Pyrosequencing of Tagged cDNA

2.5.1. Library Preparation

Uniquely tagged cDNA pools were robotically prepared using a Caliper SciClone ALH robot equipped with a Twister II plate positioner according to the protocol provided by Roche the following modifications. Briefly, 15 ul of cDNA was incubated with 5 ul 10x T4 polymerase buffer, 5 ul bovine serum albumin (BSA), 5 ul dNTPs, 5 ul T4 polynucleotide kinase, and 5 ul T4 DNA polymerase (all part of the Roche GS Library Kit # 04852265001) for 25 minutes at room temperature. The end repaired DNA then was purified using SPRI beads (Agencourt #A29152) by mixing an SPRI bead suspension with the enzymatic reaction mixture in a 0.7x volume suspension to reaction mixture ratio. The beads then were washed twice with 95% ethanol, allowed to dry completely, and the DNA was eluted from the SPRI beads with 10mM Tris-HCl pH 8.0. 454 MID tagged sequencing adapters (sequences shown in Table 3) then were ligated to both ends of the DNA molecules by incubating the DNA with 20 ul 2x reaction buffer, 4 ul T4 ligase, (all part of the Roche GS Library Kit # 04852265001) and 5 ul A and B adapters (Roche #05144523001 and #05144507001) for 15 minutes at room temperature. The enzymatic reaction was purified using SPRI beads as before. The sticky ends of the ligated adapters then were filled in by incubation with 15 ul ddH_2O, 5 ul 10x reaction buffer, 2 ul dNTPs, and 5 ul Fill-in polymerase (all part of the Roche GS Library Kit # 04852265001) for 20 minutes at room temperature. The DNA was purified suing SPRI beads as before and the prepared library DNA was quantified on an Agilent Bioanalyzer. After quantification the following mathematical formula was used to determine the number of DNA molecules per ul of library solution:

$$\text{Molecules/ul} = \frac{\text{(Sample conc.;ng/ul) X (6.022 X } 10^{23} \text{ mol./mole)}}{\text{(656.6 X } 10^{9}\text{gram/mole dsDNA) X (avg. fragment length;nt)}}$$

The library then was diluted to $2x10^{5}$ molecules of DNA per ul of solution.

Table 3. MID tagged A and B adapters for 454 library preparation

MID Tag	Adapter B Sequence	Adapter A Sequence
1	5’ GCCTTGCCAGCCCGCTCAGACGAGTGCGT 3’	5’ GCCTCCCTCGCGCCATCAGACGAGTGCGT 3’
2	5’ GCCTTGCCAGCCCGCTCAGACGCTCGACA 3’	5’ GCCTCCCTCGCGCCATCAGACGCTCGACA 3’
3	5’ GCCTTGCCAGCCCGCTCAGAGACGCACTC 3’	5’ GCCTCCCTCGCGCCATCAGAGACGCACTC 3’
4	5’ GCCTTGCCAGCCCGCTCAGAGCACTGTAG 3’	5’ GCCTCCCTCGCGCCATCAGAGCACTGTAG 3’
5	5’ GCCTTGCCAGCCCGCTCAGATCAGACACG 3’	5’ GCCTCCCTCGCGCCATCAGATCAGACACG 3’
6	5’ GCCTTGCCAGCCCGCTCAGATATCGCGAG 3’	5’ GCCTCCCTCGCGCCATCAGATATCGCGAG 3’
7	5’ GCCTTGCCAGCCCGCTCAGCGTGTCTCTA 3’	5’ GCCTCCCTCGCGCCATCAGCGTGTCTCTA 3’
8	5’ GCCTTGCCAGCCCGCTCAGCTCGCGTGTC 3’	5’ GCCTCCCTCGCGCCATCAGCTCGCGTGTC 3’
9	5’ GCCTTGCCAGCCCGCTCAGTAGTATCAGC 3’	5’ GCCTCCCTCGCGCCATCAGTAGTATCAGC 3’
10	5’ GCCTTGCCAGCCCGCTCAGTCTCTATGCG 3’	5’ GCCTCCCTCGCGCCATCAGTCTCTATGCG 3’
11	5’ GCCTTGCCAGCCCGCTCAGTGATACGTCT 3’	5’ GCCTCCCTCGCGCCATCAGTGATACGTCT 3’
12	5’ GCCTTGCCAGCCCGCTCAGTACTGAGCTA 3’	5’ GCCTCCCTCGCGCCATCAGTACTGAGCTA 3’

2.5.2. Emulsion PCR Preparation

Emulsion PCR was carried out according to the 454 Genome Sequencer Methods Manual (454/Roche Diagnostics).

2.5.2.1. Preparation of Live Amplification Mix

After a complete thaw and vortex of the 454 emPCR kit reagents (Roche #04891384001) the live amplification mix was prepared according to Table 4:

Table 4. Live Amplification Mix (454 Genome Sequencer Methods Manual, 454/Roche Diagnostics

Reagent	Volumes for one emulsion	Volumes for 4 emulsion
Amplification Mix	181.62 µl	726.48 µl
$MgSO_4$	10 µl	40 µl
Amplification Primer Mix	2.08 µl	8.32 µl
Platinum HiFi *Taq* Polymerase	6 µl	24 µl
PPiase	0.30 µl	1.2 µl
Total:	200 µl	800 µl

2.5.2.2.DNA Library Capture

DNA capture beads first were washed with capture bead wash buffer by aliquoting 600,000 bead per reaction into an appropriate number of 1.5 mL reaction tubes, centrifuging for 10 seconds at 12k rpm in a Fisher Marathon 13k/M benchtop centrifuge, and discarding the supernatant. The beads then were washed twice with 500 ul capture bead wash buffer (from the Roche GS emPCR kit #04891384001) with centrifugation and removal of supernatant following each wash. The capture beads then resuspended in 20 ul capture bead buffer (from the Roche GS emPCR kit #04891384001). 2 ul of library DNA then was added to the bead suspension.

2.5.2.3. Emulsification

The emulsion oil (from the Roche GS emPCR kit #04891384001) first was vortexed prior to the addition of 240 ul mock amplification mix (from the Roche GS emPCR kit #04891384001) to each emulsion oil tube. The emulsion oil then was placed into the rack of a TissueLyser (Qiagen #85210) shaker and shaken for 5 minutes at 25 strokes/second. 160 ul of the live amplification mix then was added to the library/capture bead suspension. The bead/amplification mixture then was added to the emulsion tube and shaken for 5 minutes at 15 strokes/second.

2.5.2.4. Amplification

After emulsification 100 ul of the water-in-oil emulsion for each emPCR was aliquoted into 8 wells of a 96 well thermocycler reaction plate and placed into a thermocycler for overnight amplification following the manufacturers recommended conditions of:

94°C for 4 minutes; 40 cycles of 30 seconds at 94°C, 60 seconds at 58°C, 90 seconds at 68°C; 13 cycles of 30 seconds at 94°C, 6 minutes at 58°C; hold at 10°C.

2.5.2.5. Bead Recovery

The bead recovery method was a modification of that recommended by Roche/454 in their 454 Genome Sequencer Methods Manual. Briefly, instead of recovering the beads using a manufacturer recommended syringe and filter, after amplification 100 ul of isopropanol was added to each of the wells containing the emPCR emulsion and the isopropanol/emulsion mixture then was transferred to a 50 ml Corning centrifuge tube (Corning Costar #430921). The emulsion amplification wells were further washed with 200 ul isopropanol that was added into the 50 ml Corning centrifuge tube. Additional isopropanol was added to the Corning tube to a final concentration of 30 ml followed by centrifugation at 3200 rpm in a Beckman G6SR tabletop centrifuge for 4 minutes and the supernatant was decanted. Two additional 30 ml isopropanol wash/spin steps were performed and the supernatant decanted. The beads then were resuspended and washed similarly twice in 10 ml bead wash buffer (from the Roche GS emPCR kit #04891384001) and the final supernatant decanted. The pelleted beads then were resuspended in 10 ml enhancing fluid and transferred to an Oak Ridge screw cap tube (Nalgene #3119-0050). The Corning tube was rinsed a second time with 10 ml enhancing fluid (from the Roche GS emPCR kit #04891384001) that was also added to the Oak Ridge tube. The Oak Ridge tube then was centrifuged at 10,000 rpm for 5 minutes in a Sorvall RC5B floor centrifuge and the supernatant was decanted into a second set of Oak Ridge tubes that were similarly centrifuged and decanted. The contents of each Oak Ridge tube then was split between two 1.5 ml tubes and centrifuged for 10,000 rpm for 10 seconds with the supernatant decanted. The two 1.5 ml tubes then were combined all but 100 ul of enhancing fluid removed.

2.5.2.6. Bead Enrichment

Enrichment beads were first washed by adding 20 ul of suspended enrichment beads to 1 ml of enhancing fluid. These beads then were pelleted using a magnetic particle collector (MPC) and the pellet was washed with 100 ul of enhancing fluid. The enrichment bead pellet then was resuspended in 100 ul enhancing fluid and added to the amplified DNA beads. This mixture was placed on a tube rotator for 5 minutes. After 5 minutes the bead suspension mixture was brought to 1 ml with enhancing fluid and placed on the MPC for 2 minutes to pellet the beads. The supernatant was removed and the pellet was gently washed 3 times with 1 ml enhancing fluid (from the Roche emPCR kit # 04891384001). The bead pellet then was resuspended in 700 ul of 0.125 M sodium hydroxide (NaOH). This suspension then was pelleted on the MPC with the supernatant removed to a new 1.5 ml tube. The pellet was washed again with NaOH and the supernatant combined with the first wash. The combined supernatants were centrifuged and the supernatant discarded. The enriched beads then were washed twice with 1 ml of annealing buffer before being resuspended in 200 ul of annealing buffer and placed in a 0.2 ml tube. This tube then was centrifuged and the supernatant discarded.

2.5.2.7. Sequencing Primer Annealing

The enriched DNA beads then were resuspended in 15 ul annealing buffer (from the Roche GS emPCR kit #04891384001) to which was added 3 ul sequencing primer. The beads then were placed in a thermocycler to anneal the sequencing primer. Afterwards the beads were washed twice with 200 ul of annealing buffer and resuspended in 250 ul of annealing buffer. 5 ul of the bead suspension was counted using a Beckman-Coulter Z8 coulter counter to determine the total number of beads generated form the emPCR. The beads then were stored at 4^{o}C.

2.5.3. Loading and Pyrosequencing

A clean picotiter plate (PTP) was incubated for 10 minutes in bead buffer 2 (bead buffer 1, 34 ul apyrase). After incubation the PTP was placed in to a bead deposition device (BDD) and centrifuged for 10 minutes at 2700 rpm in a Beckman G6SR centrifuge. During this centrifugation an appropriate number of beads, as shown in Table 5, was aliquoted into a clean 1.5 ml tube. 5 ul of control beads were also added to this tube.

Table 5. Appropriate number of beads per 454 plate region

Amount of plate	Full Plate	½ Plate	¼ plate	1/8 plate
Beads loaded	$1.2x10^6$	600,000	300,000	100,000

The sample and control beads then were pelleted by centrifugation at 12,000 rpm for 7 seconds and all but 30 ul of supernatant was removed. 290 ul of bead incubation mix (785 ul bead buffer, 75 ul polymerase cofactor, and 150 ul DNA polymerase) (this and other reagents in this section were components of the Roche GS FLX Standard LR70 Sequencing Kit #04932315001) then were added and the sample tube was placed on a LabQuake rotator (Barnstead #400110) and incubated a room temperature for 30 minutes.

The packing beads were washed three times by the addition 1 ml of bead buffer 2 followed by centrifugation at 12,000 rpm for 5 minutes and decanting of the supernatant. After washing the packing beads were resuspended in 550 ul of bead buffer 2 and 530 ul of packing beads were mixed with 640 ul of the remaining bead incubation mix and placed on ice. The enzyme beads were similarly washed, however they were pelleted using the MPC as opposed to centrifugation. After washing the enzyme beads were resuspended in 1 ml of bead buffer 2. Then 950 ul of enzyme beads were mixed with 950 ul of bead buffer 2 and placed on ice.

Following incubation the DNA beads were removed from the rotator and combined with 340 ul of bead buffer 2. This mixture then was pipetted into a 1/4th region of the BDD and allowed to sit for 10 minutes for gravity deposition of the DNA beads. After gravity deposition 410 ul of the DNA bead layer was removed from the PTP and placed into a 1.5 ml tube. The remaining DNA bead solution from the PTP was discarded. 290 ul of the packing bead/bead incubation mixture then was mixed with the 410 ul of the original DNA bead layer solution and this then was pipetted into the same 1/4th region of the BDD containing the sample to be loaded. The BDD then was centrifuged for 10 minutes at 2700 rpm. Following centrifugation the supernatant was removed from the BDD and discarded. 660 ul of the enzyme bead mixture was placed into the 1/4th region and the BDD was centrifuged again for 10 minutes at 2700 rpm. During this final centrifugation the 454 instrument was prepared by aliquoting 164 ul of apyrase into apyrase buffer, 1.5 ml of dATP into ATP buffer, and 1 ml of DTT into the common buffer. These solutions then were loaded into the instrument. After centrifugation the PTP was removed from the BDD and placed into the machine and the sequencing run started.

2.6. Data Assembly and Analysis

2.6.1. Deconvolution and Assembly of Pyrosequencing Data

Pools of tagged cDNA sequenced on the 454 were deconvoluted and assembled using a software pipeline written by Jim White (personnel communication), the director of our informatics group. This software pipeline consists of the Perl scripts get_454_pools and split_454_pools. Briefly, the program get_454_pools collates the data for all runs of the same sample pool together and calls the program split_454_pools. This program bins each read from the sequenced, pooled sample according to the tag at the beginning of the read and trims the tag and primer sequence from the read. The program get_454_pools then assembles each binned sample using the 454 GS De Novo Assembler program and automatically runs blast on the generated contigs.

2.6.2. Blast Analysis

After assembly contigs were queried against the non-redundant Genbank database using both BlastN and BlastX with an Expect (E) value of 0.001. The top 5 non-redundant Genbank homologies were reported in a tabular format for each contig from both BLAST searches using the Blast2table program written by Jim White (personnel communication). Those contigs which showed no homology using either BlastX or BlastN were reprocessed using tBlastX with an E value of 0.001. Any remaining contigs that continued to have no homology underwent conserved domain search using RPS-Blast (Marchler-Bauer et al 2007).

2.6.3. Contig Orientation

Contigs assembled in the complementary direction were re-oriented using the computer program sort_contigs written by Jim White (personnel communication).

2.6.4. ClustalX Alignment

Multiple sequence alignments were carried out using the program ClustalX (Larkin et al 2007) using the default conditions.

2.6.5. Phylogenetic Tree Generation

Bootstrapped phylogenetic trees were generated by ClustalX using the default conditions (random seed of 111 with 1000 bootstrap trials). Phylogenetic trees were viewed using TreeviewX (Page 1996).

2.6.6. RNA Folding Prediction

Folded RNA structures were determined for selected regions surrounding putative frame shift sites using the program RNAfold (Hofacker et al 1994) at the RNAfold server located at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi using the default settings.

2.6.7. Metagenomic Data Analysis System

Using the software Perl, MySQL, HTML, and Apache a data analysis system, TGPweb, was prepared. The core of the system is a MySQL database of relational tables arranged according to the schema shown in Figure 17. Table 1 contains the plant sample information including plant taxonomic data. Table 2 contains all assembled contigs generated for each sample. Table 3 contains the sequencing statistics for each sample in each pool. Tables 4-6 contain the BlastX, BlastN, and tBlastX information. Table 7 contains the domain search output given by RPS-Blast of contigs with no other Blast homology. All of these tables are related using super and foreign keys for quick searching and analysis.

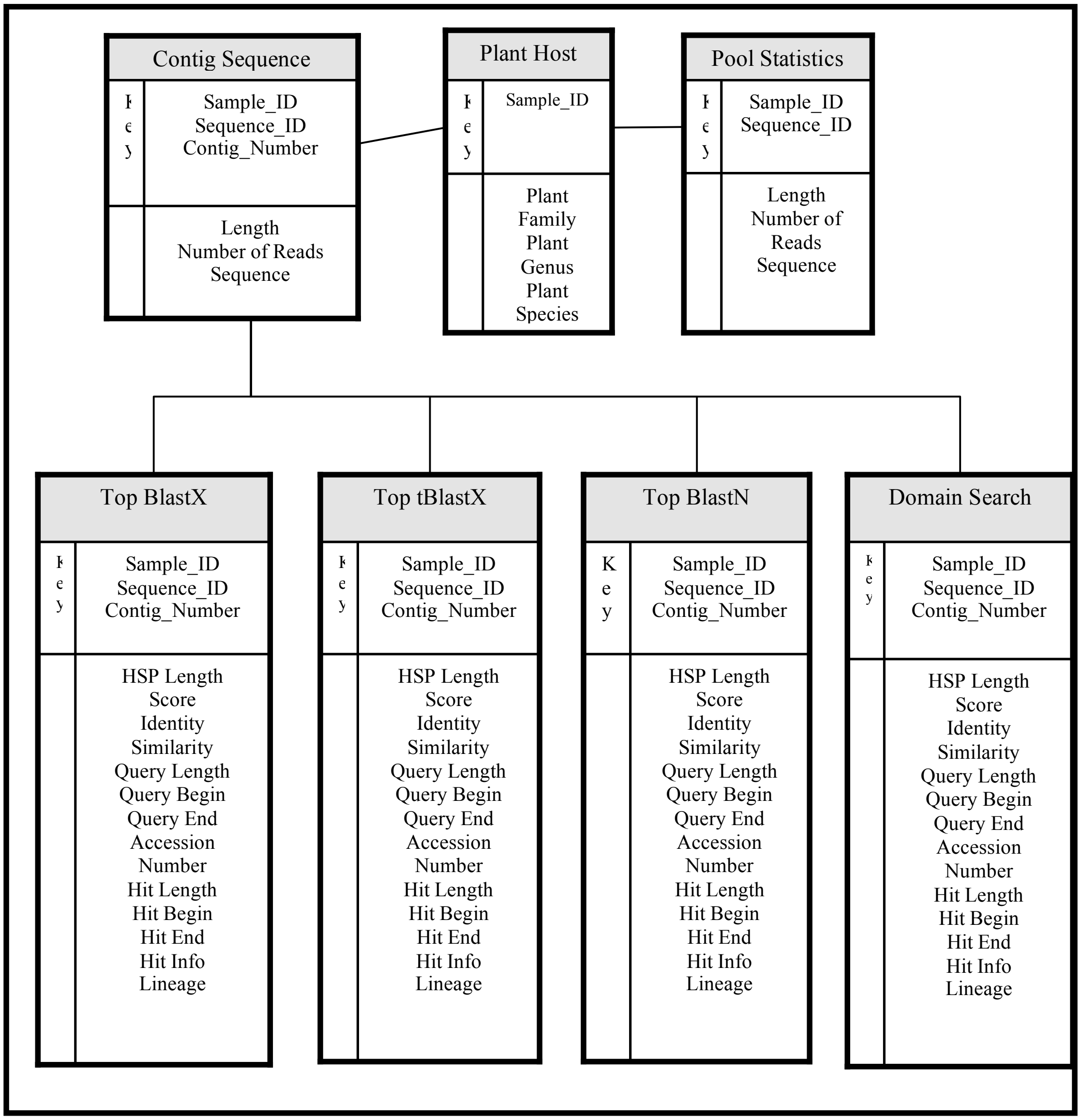


Figure 17. The schema of the TGPweb database.

Chapter 3 Results and Discussion

3.1. Overview of Sequencing Data

3.1.1. Sequencing and Assembly Results

A total of 1254 dsRNA samples from 527 different species covering 102 plant families and 340 genera were collected from the Tallgrass Prairie Preserve and sequenced in 19 sets (Table 6). The average length for the cDNA libraries was 500-600 base pairs. After DNA sequencing, a total of 1,229 samples (98%) gave high quality sequencing results with usable tags for deconvolution. A total of 1,100,982 reads were produced consisting of 246,942,957 base pairs. When these reads then were assembled together 37,385 contigs were generated that had an average contig length of 409 base pairs. 135,216 (12%) of the reads were not incorporated into contigs or generated contigs less than 100 base pairs in length and were grouped as singletons.

Table 6. Pool Statistics

Pool Name	Number of Samples	Successful cDNA libraries	Average Read Length (nt)	Average Contig Length (nt)	Singleton Percentage
Pool 1	24	24	214	439	5.5%
Pool 2	24	24	219	343	8.3%
Pool 3	24	24	218	362	10.2%
Pool 4	24	24	204	452	11.6%
Pool 5	96	96	217	391	25.4%
Pool 6	96	96	215	350	24.2%
Pool 7	96	96	350	377	20.5%
Pool 8	96	96	210	356	28.4%
Pool 9	96	96	190	377	15.7%
Pool 10	96	87	197	405	15.2%
Pool 11	96	96	256	454	11.4%
Pool 12	96	95	198	373	12.4%
Pool 13	96	96	202	399	12.9%
Pool 14	96	95	192	428	15.2%
Pool 15	96	95	195	404	16.0%
Pool 16	96	84	280	437	20.5%
Pool 17	96	92	273	568	9.6%
Pool 18	29	29	257	448	10.5%
Pool 19	96	96	237	420	14.5%

Table 7. Contig Length Summary

Contig Length (nt)	Number of Contigs
<500	30020
501-1000	5106
1001-2000	2092
2001-3000	128
3001-5000	44
>5001	5

As is typical with metagenomic sequencing, no full length viral genomes were produced as the majority of contigs generated were far shorter than the typical RNA genome length (Table 7). This is due in part to the heterogeneity of the metagenomic samples as well as the lower probability of random primers annealing to the ends of the viral genome, despite lack of annealing bias (Stangegaard et al 2006). Also, due to the nature of the 454 Newbler *De Novo* Assembler which assembles reads with are least 90% identity over a 40 base pair aligned region, it was not possible to separate individual viral strains. Despite these constraints, several contigs for different viruses were generated which likely covered greater than 80% of several individual viral genomes when they were aligned to closely related viruses as shown below.

3.2. Comparison of Pools Containing 24 and 96 Samples

A comparison was made of the amount of data generated when 4 pools of 24 samples were sequenced separately, as versus one pool of 96 samples sequenced at one time (Table 8). Overall, the 4 pools of 24 were approximately 30% more prolific in the number of reads generated, based on total base pairs, number of contigs, and total contig length. However, this only lead to an approximately 15% increase in the total number of contigs greater than 500 base pairs. As the total amount of data generated by 1 pool of 96 is similar to that obtained when 4 batches of 24 samples each from the same pool of 96 were sequenced in parallel, this significant increase in sample throughput had the potential of being extremely useful as when applied to 4 pools of 96 samples each with a 4-fold decrease in overall sequencing costs.

Table 8. A comparison of data generated by 4 pools of 24 samples and 1 pool of 96 samples

Parameter	**24 tags (4 Pools)**	**96 tags (1 Pool)**
total base pairs (nt)	27,053,446	19,030,432
number of reads	126,022	80,294
average read length (nt)	214	237
number of contigs	3638	2641
average contig length (nt)	400	420
number contigs >500 (nt)	613	533
total contig length (nt)	1,328,451	1,052,858
singleton percentage	8.8%	14.5%
number of working tags	96	96

3.3. Data Analysis

3.3.1. Blast Analysis

To determine the optimal method by which assembled contigs could be identified, a comparison was made between the two programs BlastN and BlastX. After using each program to compare assembled contigs to the non-redundant nucleotide and protein databases (nt/nr) of Genbank, the homologies for each contig were determined to be either viral, non-viral or no homology detected. As shown in Table 9, BlastN was more efficient in determining homology for non-viral contigs, while BlastX was more efficient in determining homology for viruses.

Table 9. A comparison of BlastN and BlastX in determining contig homology

Contig Type	BlastN	BlastX
Viral	603 (1.6%)	1662 (4%)
Non-Viral	29642 (79%)	26804 (72%)
No Homology Detected	7149 (19 %)	8924 (24%)

To further refine the homology search, those contigs which showed no homology using either BlastX or BlastN were processed using tBlastX. This allowed for the determination of 50 viral contigs (~3% of the total viral contigs) and 569 of non-viral contigs (~2% of the total non-viral contigs). The overall number of viral, non-viral, no homology contigs are given in Table 10.

Table 10. Number of contigs by contig type after BlastN, BlastX, and tBlastX homology searches

Contig Type	Total	Percentage
Viral	1712	4.5%
Non-Viral	31276	83.6%
No Homology Detected	4408	11.7%

As mentioned above, BlastX appeared to be more efficient in determining contig homology for plant viruses than BlastN. BlastX was more sensitive when comparing viral sequences because an amino acid at a given position was conserved, but used an alternative, degenerate amino acid codon. For example, the amino acid serine is coded for by the codons UCU, UCC, UCA, UCG, AGU, and AGC. This codon degeneracy allows BlastX to find more closely related species as compared to BlastN, where nucleotide changes will prevent homology determination although the amino acid encoded is conserved (States et al 1991). A comparison of BlastN and BlastX for contigs showing viral homology are given in Table 11.

Table 11. Comparison of BlastX and BlastN for contigs showing viral homology

Blast Type	Total Contigs	Average E-value	Average % Identity	Average % Similarity	Average HSP
BlastX	1662	1.85e-5	59%	73%	125
BlastN	603	4.0e-5	88%	88%	154

3.4. Overview of Identified Plant Virus Families

After Blast processing to determine plant virus homology, the family and genus of each contig were assigned. Table 12 shows the overall number of contigs generated for each viral family, as well as the number of samples infected. Within the TGP the most abundant viruses are members of the *Tymoviridae*, *Totiviridae*, and *Partitiviridae* families. Interestingly, the *Totiviridae* and *Partitiviridae*, are dsRNA genomic viruses, while the *Tymoviridae* is a (+)ssRNA genomic viral family. The vast majority of viral genomes found in this study consisted of (+)ssRNA genomes, echoing the distribution of plant viruses discovered thus far. These results are include as part of the information about viral genomes given in Appendix 1.

Table 12. Total number of samples infected by each viral family detected.

Viral Family	Infected Samples	Total Contigs
Bunyaviridae	1	5
Caulimoviridae	6	7
Chrysoviridae	34	56
Closteroviridae	12	30
Comoviridae	14	29
Endornaviridae	18	76
Flexiviridae	24	75
Luteoviridae	5	7
Narnaviridae	2	2
Partitiviridae	134	219
Potyviridae	5	25
Reoviridae	10	25
Rhabdoviridae	1	2
Sequiviridae	1	1
Tombusviridae	13	20
Totiviridae	156	457
Tymoviridae	168	454
Orphan	47	77

3.5. Plant Virus Families

3.5.1. *Bunyaviridae*

Viruses of the family *Bunyaviridae* are characterized by a tripartite (-)ssRNA genome (Nichol et al 2005). Only 1 sample from this project was determined to be infected by a member of this viral family with only 5 contigs generated. 4 of the contigs show homology to the RdRp gene, with the 1 remaining showing homology to an encoded glycoprotein.

3.5.2. *Caulimoviridae*

The viral family *Caulimoviridae* is the only dsDNA genome virus found this far within the Tallgrass Prairie Preserve. 6 samples gave rise to contigs showing homology to 3 different genera within this family, *Badnavirus*, *Caulimovirus*, and *Soymovirus*. 3 of these samples, species *Ruellia humilis*, all appear to be infected with the same genus of virus, *Soymovirus*.

3.5.3. *Chrysoviridae*

Members of the family *Chrysoviridae* are characterized by multipartite, dsRNA genomes consisting of 4 separate molecules (Ghabrial et al 2005b). Contigs from 34 samples showed homology to the single genus within this family, *Chrysovirus*. Typically identified as a fungal virus (Covelli et al 2004), it currently is unknown if this virus may infect a plant or if it is merely found within parasitic fungi. In several of the samples contigs showing homology to each of the 4 separate genomic molecules were identified. In one case a near full length, 3032 nt genomic fragment covering the major viral capsid was produced from a sample of the plant *Isoetes butleri*. This contig then was compared to the same protein sequence (YP_052859.1) encoded by the top BlastX homologous virus, *Helminthosporium victoriae* 145S virus (Taxonomy ID 164750), using BlastX and Blast2seq. The dot matrix alignment can be seen in Figure 18. The percent identity and percent similarity between the two sequence were 30% and 49%, respectively.

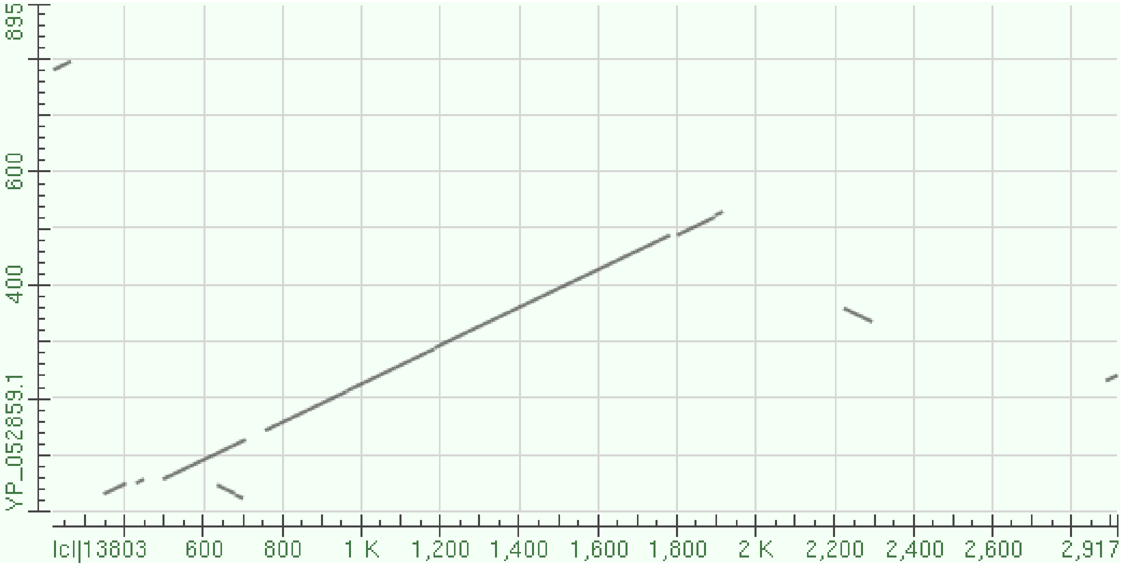


Figure 18. A dot matrix alignment of a near-full-length genomic fragment encoding a capsid sequence for a *Chrysovirus*-like contig from *Isoetes butleri* (horizontal) and the amino acid sequence of the capsid protein (YP_052859.1) for the virus *Helminthosporium victoriae* 145S virus (vertical).

Given that one of the taxonomic rules for classification in this genus/family is serological relationships (Fauquet et al 2005), which are based on reactions to the coat protein, and given the distinct dissimilarity between the two capsid sequences it is highly probable that this virus is a new species.

3.5.4. *Closteroviridae*

The viruses in the family *Closteroviridae* have extremely large, (+)ssRNA genomes which may be either mono- or bipartite and range in size from 17 to 18 kilobases in length (Fauquet et al 2005). 12 samples from the TGP generated contigs with homology to 3 genera, *Ampelovirus*, *Closterovirus*, and *Crinivirus*. The majority, 7, were homologous to *Closterovirus* while *Ampelovirus* and *Crinivirus* were found in 3 and 2 samples, respectively.

3.5.5. *Comoviridae*

This family is characterized by (+)ssRNA, bipartite genomes (Le Gall et al 2005a). 14 samples from the Tallgrass Prairie Preserve generated contigs which demonstrated homology to 3 genera, *Comovirus*, *Fabavirus*, and *Nepovirus*. 50% of the overall contigs showing homology to the family *Comoviridae* were members of the plant species *Asclepias viridis*, 8% of the total *Asclepias viridis* sampled. One sample of *Asclepias* generated a near full length, 3293 nt contig for the second genomic RNA showing homology to the viral family *Fabavirus*. This fragment was compared using BlastX to the amino acid sequence of the large coat protein (BAF37656.1) of the virus Broad Bean Wilt Virus (Taxonomy ID 76875), its closest match from comparison with the non-redundant Genbank database. A dot matrix alignment is given in Figure 19.

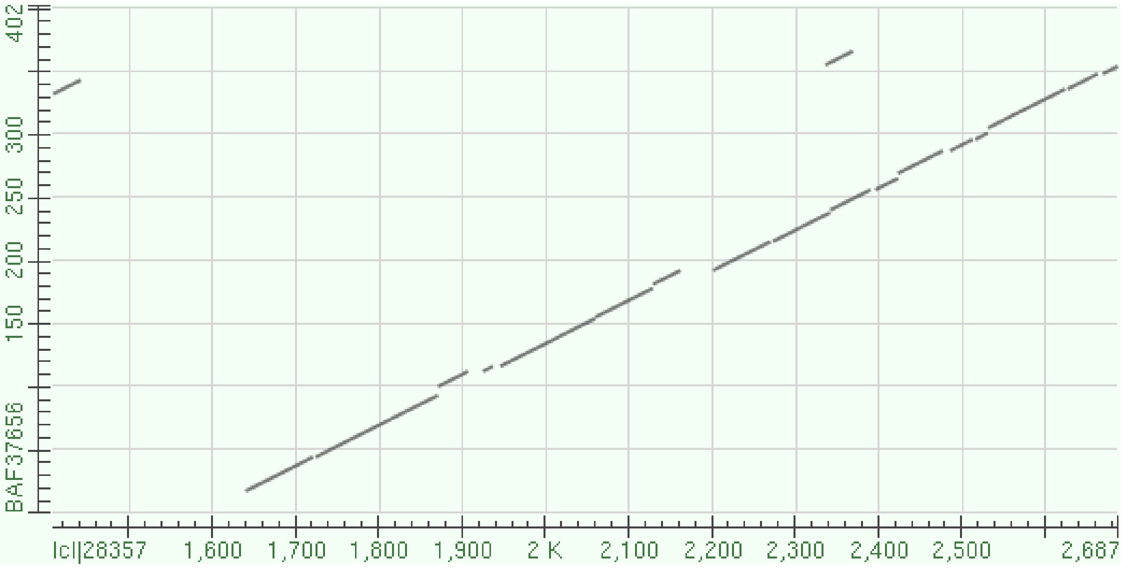


Figure 19. A dot matrix alignment of the near-full-length genomic fragment from *Asclepias* to that of the large coat protein (BAF37656.1) of the virus Broad Bean Wilt Virus using BlastX.

The overall amino acid sequence identity was 23% with 41% sequence similarity. Based on this alignment and the taxonomic criteria that a new species should have less than 75% similarity for the large coat protein between itself and the next closest member of the genus (Fauquet et al 2005), this virus is most probably a new species. Furthermore, based on the extremely low similarity score, this may indicate a new genus in the family *Comoviridae* as well. To further investigate this possibility a multiple sequence alignment based on the amino acid sequence of the coat protein was made for this contig as well as the coat proteins of other, previously characterized members of the *Comoviridae* family. The Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment is shown in Figure 20. As the coat protein sequence does not cluster with other previously characterized genera further indication is given that this virus represents a new genus in the family *Comoviridae*.

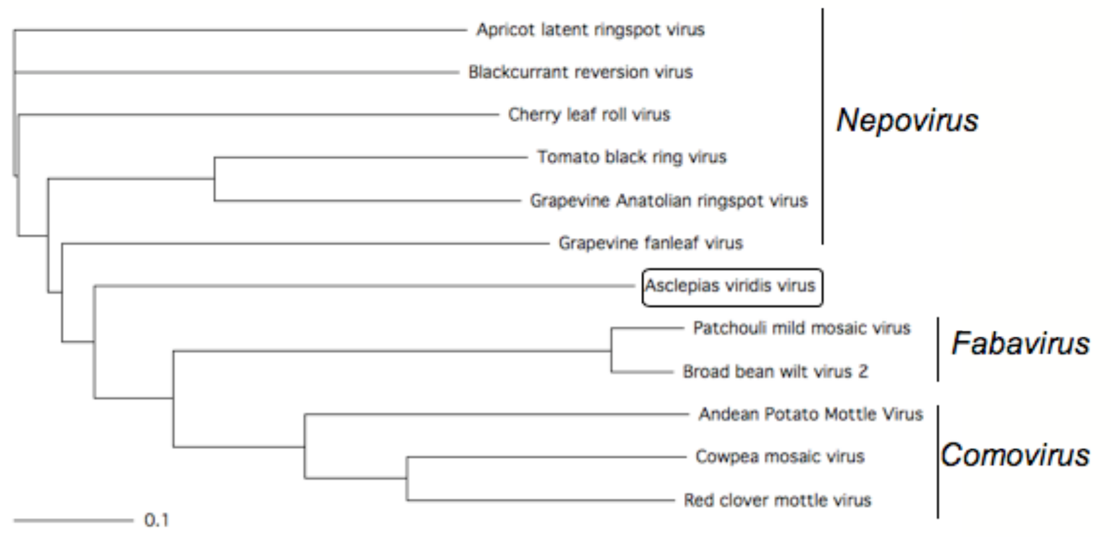


Figure 20. A Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment of the coat protein amino acid sequence for members of the family *Comoviridae*. Apricot Latent Ringspot Virus CAC05656. Blackcurrant Reversion Virus NP_733982. Cherry Leaf Roll Virus 1921133A. Tomato Black Ring Virus NP_758856. Grapevine Anatolian Ringspot Virus AAO62576. Grapevine Fanleaf Virus ACM17907. Patchouli Mild Mosaic Virus NP_733969. Broad Bean Wilt Virus 2 BAF37656. Andean Potato Mottle Virus 1909345A. Cowpea Mosaic Virus NP_734001. Red Clover Mottle Virus NP_733992.

3.5.6. *Endornaviridae*

Endornaviridae are characterized by large (>10 kb) dsRNA genomes (Osaki et al 2006). 18 samples generated contigs showing homology to this family. As the genome is very large, no full length or near full length sequences were obtained.

3.5.7. *Flexiviridae*

Members of the family *Flexiviridae* are characterized by a monopartite (+)ssRNAgenome (Adams et al 2005). A total of 24 samples generated contigs showing homology to two genera within this family, *Allexivirus* and *Potexvirus*. All samples showing homology to *Allexivirus* were obtained from one plant sample, a cactus species *Escobaria missouriensis*. While no full length sequences are available for either the virus or genes encoded by the virus, those contigs which have shown homology to this virus average 58% identity and 70% similarity on the amino acid level to Garlic Virus A (Taxonomy ID 12433). The taxonomic criteria for amino acid sequence is less than 80% identical for the coat protein and polymerase gene (Fauquet et al 2005). Based on this level of amino acid identity, it is likely that this is a new species of *Allexivirus*.

The remainder of the contigs showed homology to the genera *Potexvirus*. Within these samples one, a clover (species *Trifolium repens*), was shown to be harboring a strain of Clover Yellow Mosaic Virus (Taxonomy ID 12177) with 93% and 94% amino acid sequence identity and homology, respectively, for the polymerase gene (NP_077079.1). This was one of the few cases of a previously described virus being found within the Tallgrass Prairie Preserve.

3.5.8. *Luteoviridae*

Members of the family *Luteoviridae* are characterized by monopartite (+)ssRNA viral genomes (D'Arcy and Domier 2005). 5 samples generated contigs which showed homology to 2 genera in this family, *Luteovirus* and *Polerovirus*. As the contigs were extremely short (<300 nt), it was not possible to determine taxonomy with any accuracy.

3.5.9. *Narnaviridae*

Members of the family *Narnaviridae* are characterized by monopartite (+)ssRNA viral genomes (Buck et al 2005) and 2 samples generated contigs which showed homology to this family. However, as the contigs were extremely short (<300 nt), it was not possible to determine taxonomy with any accuracy.

3.5.10. *Partitiviridae*

The family *Partitiviridae* is characterized by bipartite, dsRNA genomes (Ghabrial et al 2005a). 134 samples from the Tallgrass Prairie Preserve generated contigs showing homology to genera within this family Thus, this family was the third most populous family of viruses within the Tallgrass Prairie Preserve. Members of the genus *Alphacryptovirus* have been shown to infect plants (Boccardo and Candresse 2005), while members of the genus *Partitivirus* are known to infect fungi (Oh and Hillman 1995). 26 of the contigs showing homology to the *Partitiviridae* family show homology to the genus *Alphacryptovirus* while 33 show homology to Partitivirus. The remaining contigs show homology to orphan sequences within the *Partitiviridae* family, sequences which have not been officially designated taxonomically by the ICTV. One sample of sedge, species *Scirpus pendulus*, gave two near full length contigs for each of the two genomic parts of the virus, 1780 nt for the RDRP RNA 1and 1459 nt for the coat protein RNA 2, each showing homology to the genus *Alphacryptovirus*. Each of these two sequences then were aligned using BlastX against their respective encoded protein from their highest BlastX homology virus, White Clover Cryptic Virus 1 (Taxonomy ID 292052), using the program Blast2seq. A dot matrix plot for each of the two contigs are given in Figures 21 and 22. The encoded polymerase showed 79% identity and 88% similarity while the coat protein showed 53% identity and 66% similarity.

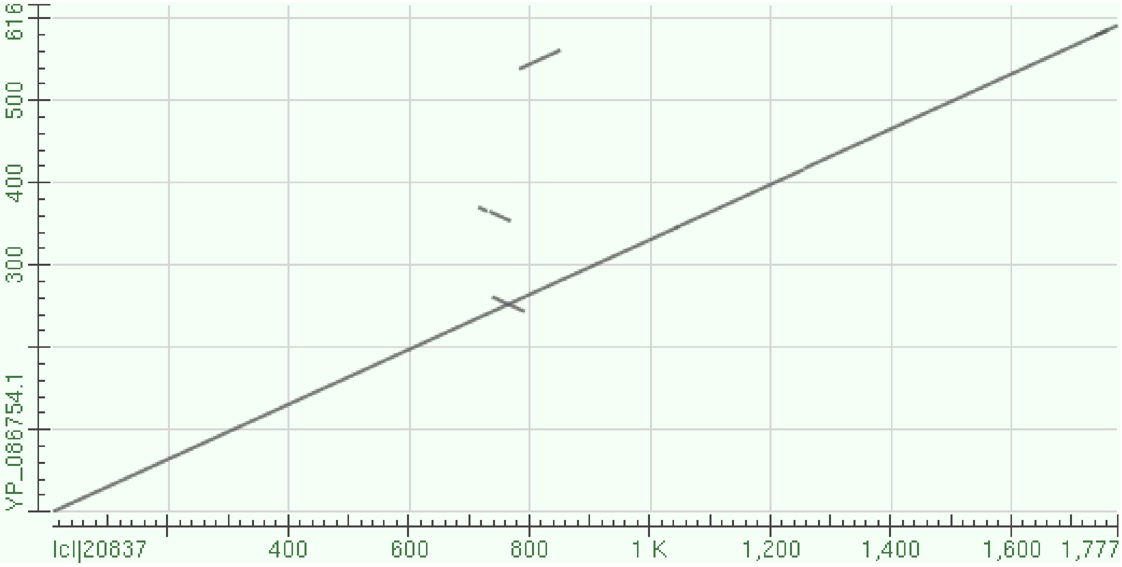


Figure 21. A dot matrix comparison of the RdRp polymerase encoded by White Clover Cryptic Virus 1 (YP_086754.1) (vertical) and a homologous contig from the plant *Scirpus pendulus* (horizontal) using Blast2Seq.

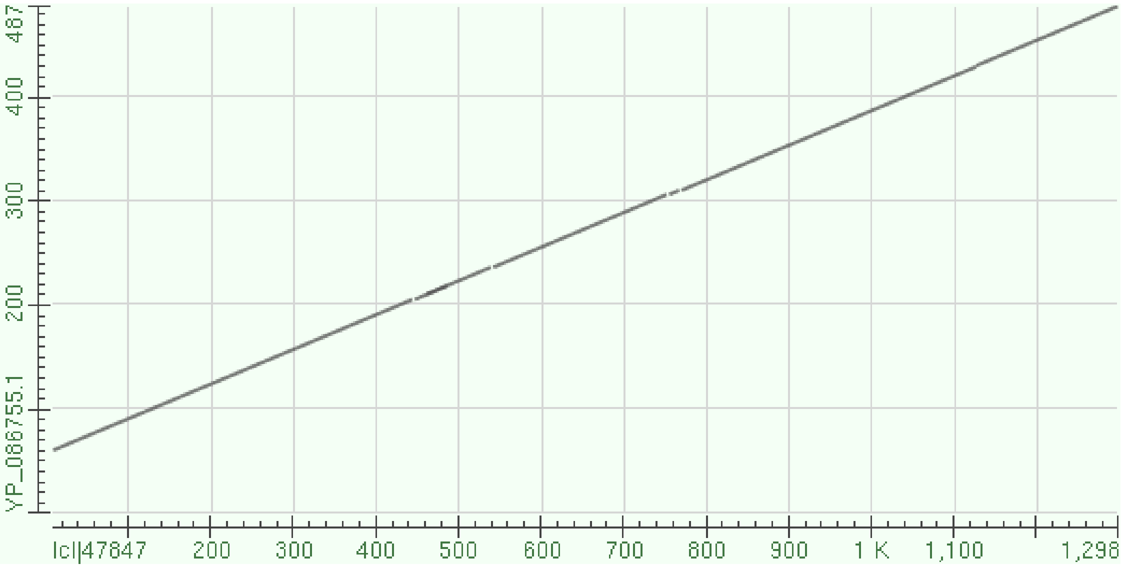
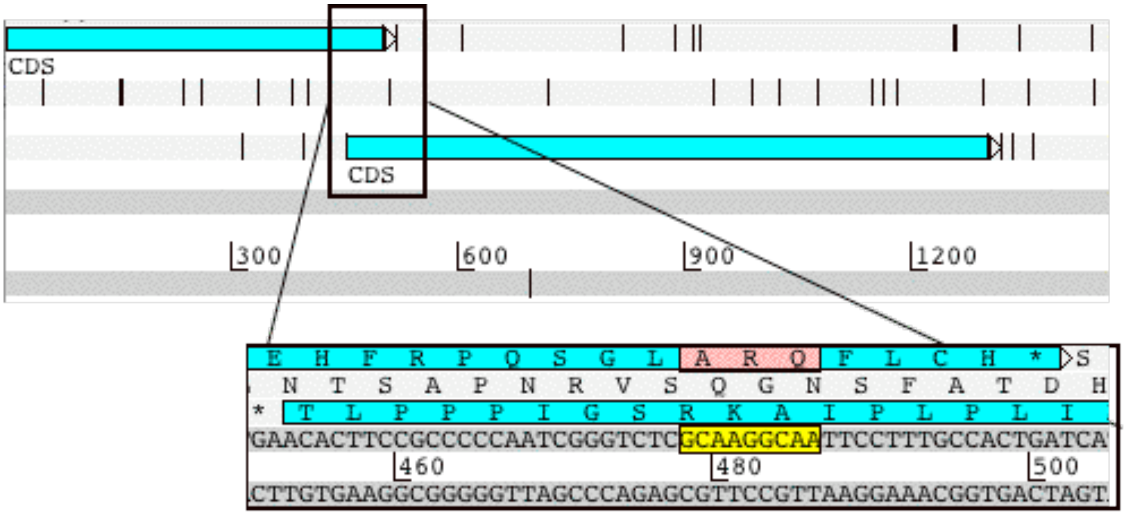


Figure 22. A dot matrix comparison of the coat protein encoded by White Clover Cryptic Virus 1 (YP_086755.1) (vertical) and a homologous contig from the plant *Scirpus pendulus* (horizontal) using Blast2Seq.

One item of particular interest is the incorporation of a +2 frame shift in the coding sequence for the coat protein. This is not typical of previously described members of the

Alphacryptovirus genus, whose genes are encoded by contiguous open reading frames. The frame shift for the coat protein occurs at nucleotide 486 of the contig, following the sequence 5’-GCAAGGCAA-3’, as shown in Figure 23. The repeated tetranucleotide sequence GCAA has been shown previously to mediate phase variation in virulence factors from *H. influenzae*, *Neisseria* spp., and *Moraxella catarrhalis* (Peak et al 1996).



```
 E  H  F  R  P  Q  S  G  L  A  R  Q  F  L  C  H  * >S
 N  T  S  A  P  N  R  V  S  Q  G  N  S  F  A  T  D  H
*  T  L  P  P  P  I  G  S  R  K  A  I  P  L  P  L  I
GAACACTTCCGCCCCCAATCGGGTCTCGCAAGGCAATTCCTTTGCCACTGATCA
CTTGTGAAGGCGGGGGTTAGCCCAGAGCGTTCCGTTAAGGAAACGGTGACTAGT
```

Figure 23. A contig generated from the plant *Scirpus pendulus* homologous to White Clover Cryptic Virus 1 coat protein as seen in the computer program Artemis. The two overlapping open reading frames are shown in detail, with the hypothesized slippery codon highlighted.

As slippery codons typically only create a +1 or -1 frame shift, this area was further characterized by RNA folding models as shown in Figure 24. The predicted RNA tertiary structure surrounding the hypothesized slippery codon is similar in nature to that of the previously identified *gag-pol* structure (Figure 25) common to members of the viral families *Totiviridae* and *Retroviridae* (Brierley 1995). However, it does not have the typical 5’-XXXYYYZ-3’ slippery codon characteristic of these frame shifts.

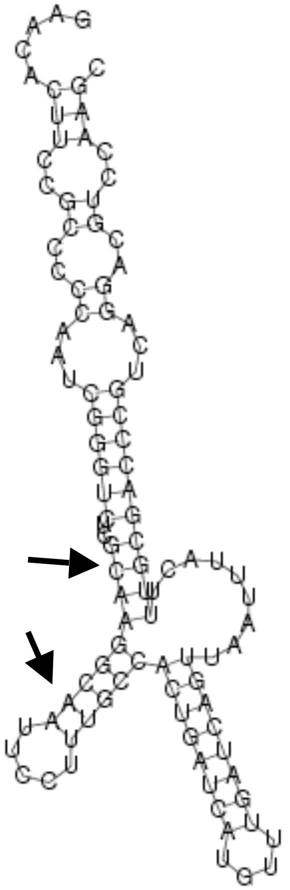


Figure 24. The RNA secondary structure generated by RNAfold (Hofacker et al 1994) for the region containing the hypothesized frameshift codon for a contig generated from the plant *Scirpus pendulus* homologous to White Clover Cryptic Virus 1 coat protein. Arrows point to the hypothesized slippery codon.

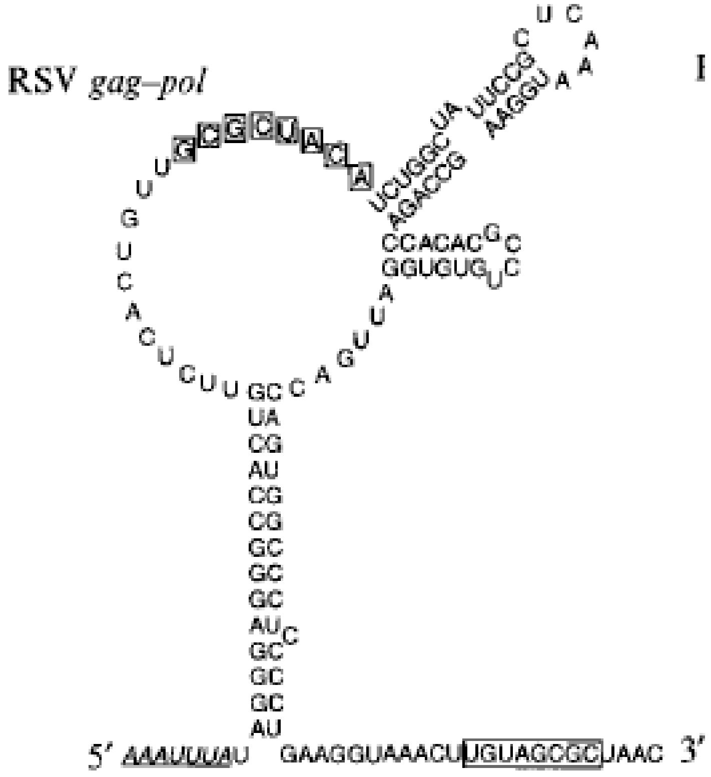


Figure 25. An RNA folding diagram showing the typical *gag-pol* slippery codon RNA tertiary structure for the virus Rous Sarcoma Virus. The slippery codon is underlined while the boxed nucleotides may interact to form a pseudoknot. (Brierley 1995)

Based on this atypical coding strategy as well as the differences in the overall coat protein amino acid sequence identity and similarity this is most likely a new species of *Alphacyptovirus*, if not a completely new genus.

A similar comparison with White Clover Cryptic Virus 1 was undertaken with two contigs, 1111 nt for RDRP RNA 1 and 1568 nt for coat protein RNA 2, generated by another sample, *Chaetopappa asteroids*. In this case the encoded polymerase showed 81% identity and 90% similarity while the coat protein showed 43% identity and 59% similarity to White Clover Cryptic Virus 1. The sequences from *Scirpus pendulus* and

Chaetopappa asteroids then were compared together using Blast2Seq with tBlastX comparison. Dot Matrix plots for the two comparisons may been seen in Figures 26 and 27.

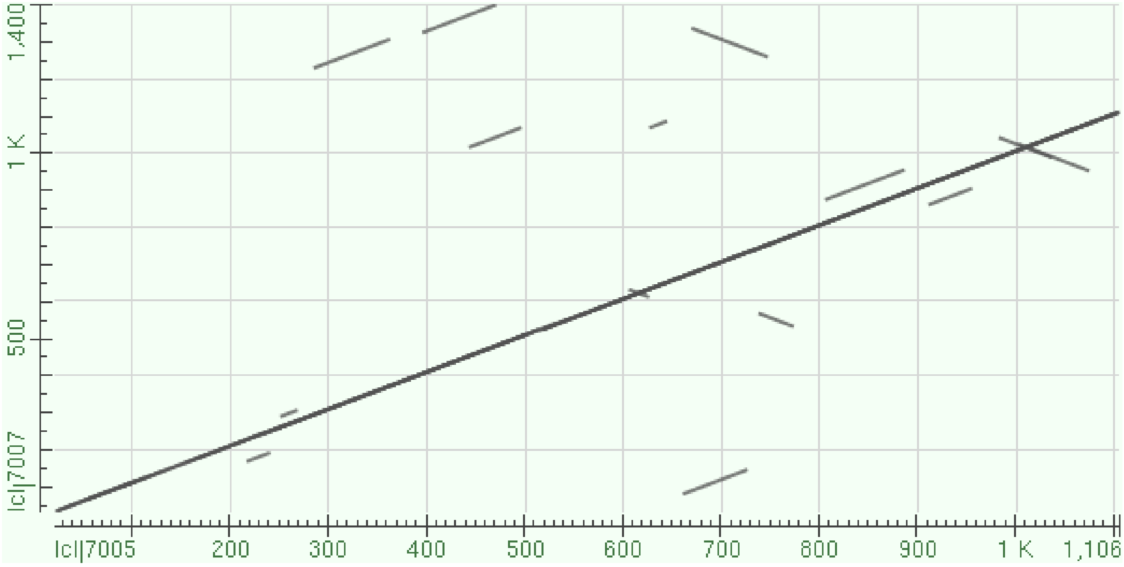


Figure 26. A comparison of the *Alphacryptovirus*-like genomic fragments encoding RdRp polymerase from plant species *Scirpus pendulus* (vertical) and *Chaetopappa asteroids* (horizontal).

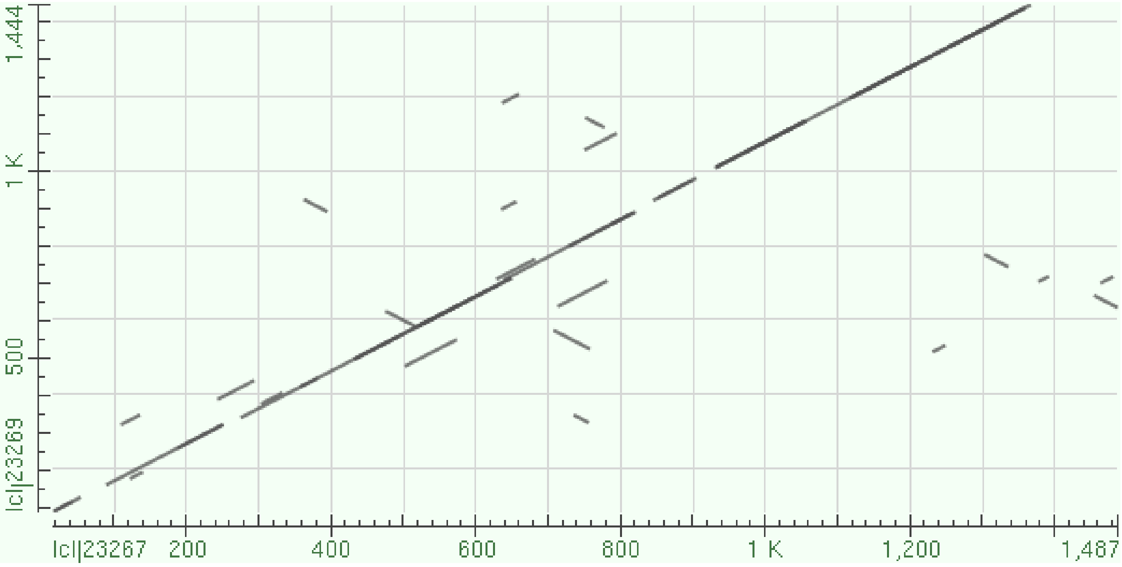


Figure 27. A comparison of the *Alphacryptovirus*-like genomic fragments encoding coat protein from plant species *Scirpus pendulus* (vertical) and *Chaetopappa asteroids* (horizontal).

The polymerase showed 75% identity and 85% similarity on the amino acid level between the two viruses, while the coat protein showed 48% identity and 66% similarity, again on the amino acid level. The two viruses also showed large amounts of repeats, both inverted and in-line, a possible sign of recombination between the two. Based on the difference in coat sequence between the two viruses, as well as the fact that they were obtained from different plants, raises the possibility that they are separate, unique species.

The *Chaetopappa asteroids* sample also had two contigs with near full length sequence, 1854 nt for the RDRP RNA 1 and 2008 nt for the coat protein RNA 2, showing homology to an orphan virus of the *Partitiviridae* family, *Primula malacoides* virus (Taxonomy ID 479713). Comparisons of these contigs with their respective encoded genes from *Primula malacoides* virus are given in Figures 28 and 29. The coat protein 66% identity and 81% similarity on the amino acid level, while the polymerase showed 75% identity and 84% similarity.

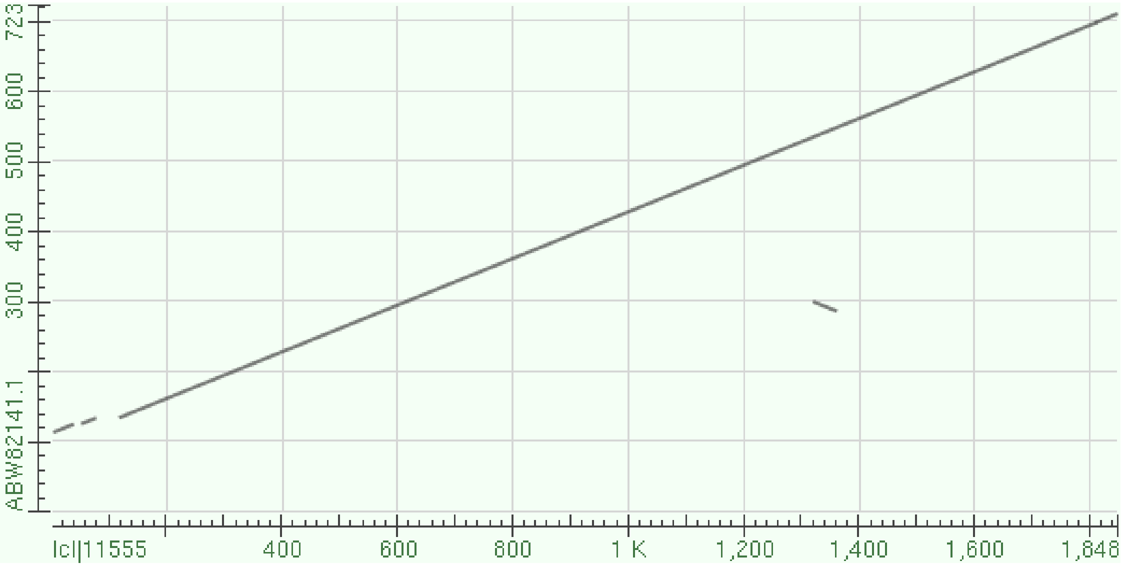


Figure 28. A dot matrix comparison of the polymerase protein encoded by Primula malacoides virus (ABW82141.1)(vertical) and a homologous contig from the plant *Chaetopappa asteroids* (horizontal) using Blast2Seq.

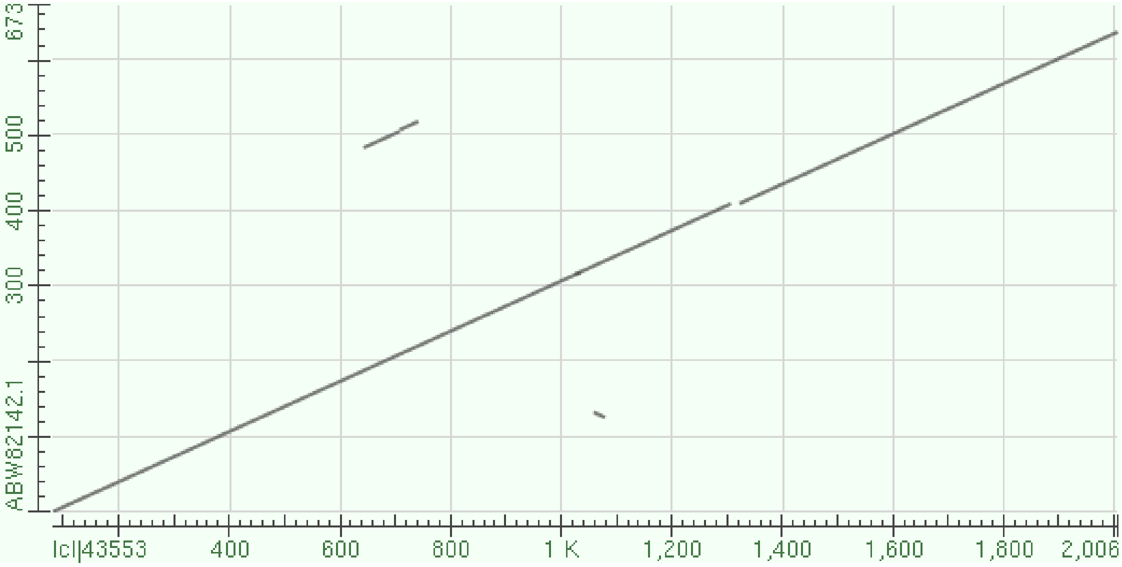


Figure 29. A dot matrix comparison of the coat protein encoded by *Primula malacoides* virus (ABW82142.1)(vertical) and a homologous contig from the plant *Chaetopappa asteroids* (horizontal) using Blast2Seq.

A BlastN comparison of the contigs from *Chaetopappa asteroids* with homology to the *Alphacryptovirus* and *Partitivirus* orphan showed no significant homology between them, confirming that they were separate viral species within the same sample. An overall taxonomic view is shown in Figures 30 and 31, created through multiple sequence alignments of the amino acids of the two respective proteins for each of the above sample, both *Chaetopappa asteroids* and *Scirpus pendulus*, with previously described members of the family *Partitiviridae*. Based on this view and the previous observations the *Partitivirus*-orphan-like virus from *Chaetopappa asteroids* and the *Alphacryptovirus*-like virus from *Scirpus pendulus* appear to be novel species while the *Alphacryptovirus*-like virus from *Chaetopappa asteroids* may represent a new genus.

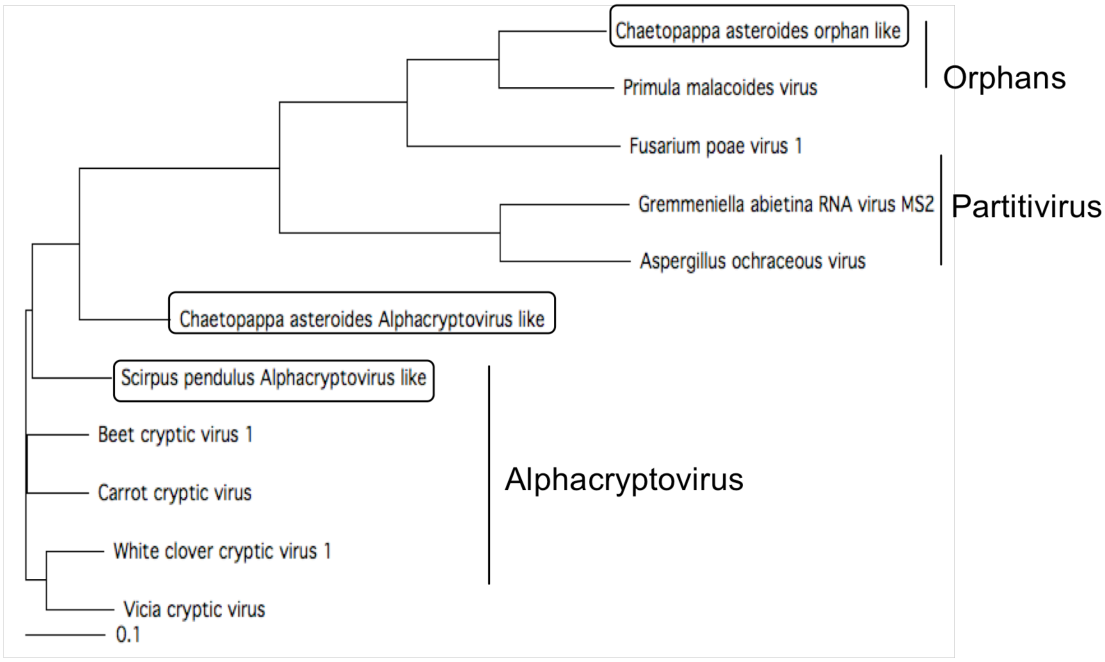

Chaetopappa asteroides orphan like
Primula malacoides virus
Orphans
Fusarium poae virus 1
Gremmeniella abietina RNA virus MS2
Partitivirus
Aspergillus ochraceous virus
Chaetopappa asteroides Alphacryptovirus like
Scirpus pendulus Alphacryptovirus like
Beet cryptic virus 1
Carrot cryptic virus
Alphacryptovirus
White clover cryptic virus 1
Vicia cryptic virus
0.1

Figure 30. A Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment of the coat protein amino acid sequence for members of the family *Partitiviridae*. *Primula malacoides* virus ABW82142.1. *Fusarium poae* virus 1 NP_624348 *Gremmeniella abietina*. RNA virus MS2 YP_138541. *Aspergillus ochraceous* virus ABV30676. Beet cryptic virus 1 YP_002308575. Carrot cryptic virus ACL93279. White clover cryptic virus 1 YP_086755. Vicia cryptic virus YP_272125.

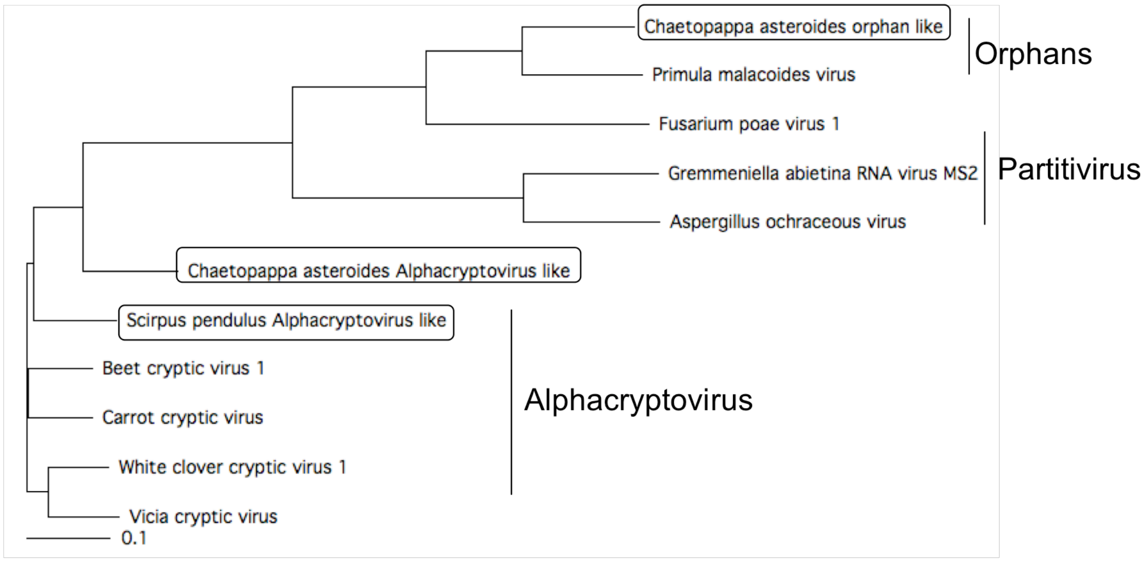

Chaetopappa asteroides orphan like
Primula malacoides virus
Orphans
Fusarium poae virus 1
Gremmeniella abietina RNA virus MS2
Partitivirus
Aspergillus ochraceous virus
Chaetopappa asteroides Alphacryptovirus like
Scirpus pendulus Alphacryptovirus like
Beet cryptic virus 1
Carrot cryptic virus
Alphacryptovirus
White clover cryptic virus 1
Vicia cryptic virus
0.1

Figure 31. A Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment of the polymerase amino acid sequence for members of the family *Partitiviridae*. *Primula malacoides* virus ABW82141.1. *Fusarium poae* virus 1 NP_624349. *Gremmeniella abietina* RNA virus MS2 YP_138540. *Aspergillus ochraceous* virus ABV30675. Beet cryptic virus 1 YP_002308574. Carrot cryptic virus ACL93278. White clover cryptic virus 1 YP_086754. Vicia cryptic virus YP_272124.

3.5.11. *Potyviridae*

The viruses of the family *Potyviridae* have monopartite (+)ssRNA genomes of approximately 9.3kb to 9.7kb in length (Berger et al 2005). Contigs from 5 samples showed homology to this viral family, with 2 showing homology to the genus *Tritimovirus* and 3 showing homology to the genus *Potyvirus*. One sample of the species *Poa compressa*, showing homology to *Tritimovirus*, generated 7 contigs each of which were between 90% to 98% identical in sequence on the nucleotide level to the previously identified Oat Necrotic Mottle Virus (Taxonomy ID 112437).

3.5.12. *Reoviridae*

Reoviridae viral family members have 10 genomic molecules consisting of dsRNA (Mertens et al 2005). 10 samples had contigs showing homology to this family, with the most prevalent genus being *Oryzavirus*. One sample, a species of *Vitis* of the family *Vitaceae*, had 12 contigs, each showing homology to the virus Rice Ragged Stunt Virus (Taxonomy ID 42475) of the genus *Oryzavirus* at the amino acid level, but no homology to any virus at the nucleotide level. This may be indicative of a new species of *Oryzavirus* or perhaps a new genus within the family *Reoviridae*.

3.5.13. *Rhabdoviridae*

Family members of *Rhabdoviridae* are characterized by (-)ssRNA genomes of approximately 14 kb in length (Tordo et al 2005). Only 1 sample, a species of *Ambrosia psilostachya*, had contigs showing homology to this family. As there were only 2 contigs of relatively short length, 185 nt and 209 nt, it was not possible to determine the taxonomy of this virus. This is the only occurrence of a negative sense RNA virus in the Tallgrass Prairie

3.5.14. *Sequiviridae*

Sequiviridae members have approximately 12 kb genomes consisting of (+)ssRNA (Le Gall et al 2005b). Only 1 contig, 243 nt in length, from 1 sample, a species of *Sorghastrum nutans*, showed homology to the genus *Waikavirus* of this family.

3.5.15. *Tombusviridae*

Members of the viral family *Tombusviridae* are characterized by (+)ssRNA genomes of approximately 4 kb in length (Lommel et al 2005). 13 samples had contigs showing homology to this viral family, with 4 samples showing homology to the genus *Carmovirus*, 4 showing homology to the genus *Panicovirus*, 1 showing homology to the genus *Tombusvirus*, and the remainder showing homology to orphan viruses within this family. There were two instances of near full length genomes being generated for this viral family, one from species *Lespedeza procumbens* showing homology to the genus *Carmovirus*, with another from species *Paspalum setaceum* showing homology to genus *Panicovirus*. A tBlastX comparison was made between the 3950 nt contig showing homology to the *Carmovirus* genus and the virus showing the most homology, Pelargonium Flower Break Virus (Taxonomy ID 35291). A dot matrix plot for this comparison are given in Figure 32.

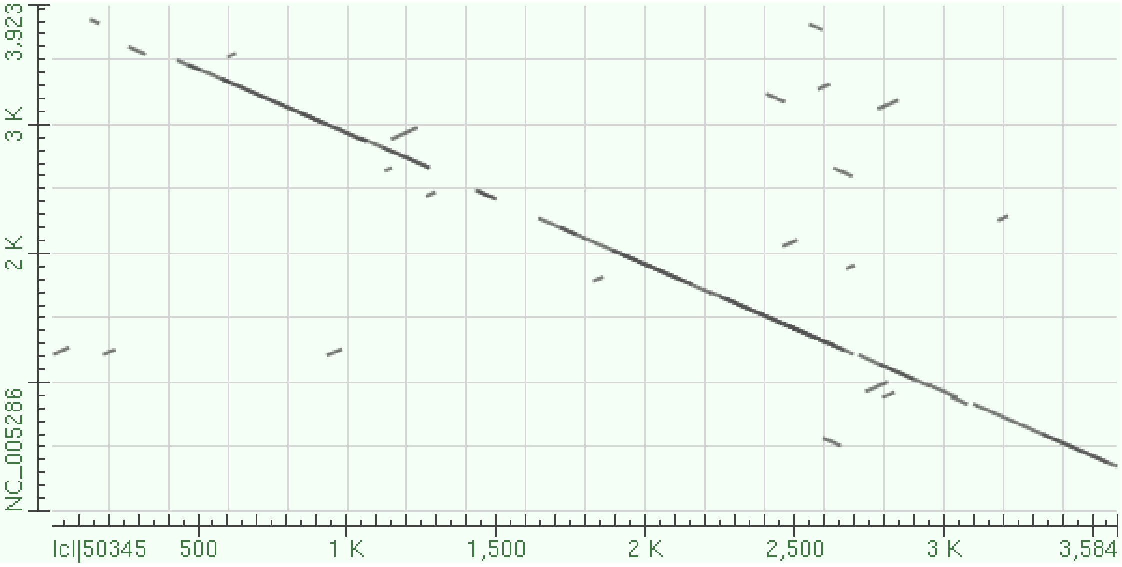


Figure 32. A dot matrix plot showing the tBlastX comparison between the genome for Pelargonium Flower Break Virus (NC_005286)(vertical) and a near full length genomic contig showing homology to the genus *Carmovirus*, isolated from *Lespedeza procumbens.*

BlastX also was used to compare this sequence to the protein sequences of the coat protein (ABD93309) and polymerase (NP_945123) genes as these are required for taxonomic identification. The coat protein sequence showed 41% and 53% identity and similarity, respectively, on the cusp of the 41% amino acid identity threshold for the classification of a new species of virus (Fauquet et al 2005). The polymerase gene was similar in result, showing 44% identity and 59% similarity against the speciation threshold of 52% amino acid identity (Fauquet et al 2005). Based on these observations it was apparent this sequence represented a new viral species.

A similar comparison was made for the near full length, 3190 nt genomic contig generated from the sample of *Paspalum setaceum* showing homology to the genus *Panicovirus* and its nearest homologous neighbor, Panicum Mosaic Virus (Taxonomy ID 40279) as seen in the dot matrix plot given in Figure 33. BlastX comparisons between this contig and the coat (NP_068346) and polymerase (AAC97551) proteins for Panicum Mosaic Virus showed 74% and 87% identity and similarity for the coat with 84% and 94% identity and similarity for the polymerase protein. This falls within the speciation thresholds for *Panicovirus* (Fauquet et al 2005) and thus leads to the conclusion that this contig represents a new strain of Panicum Mosaic Virus.

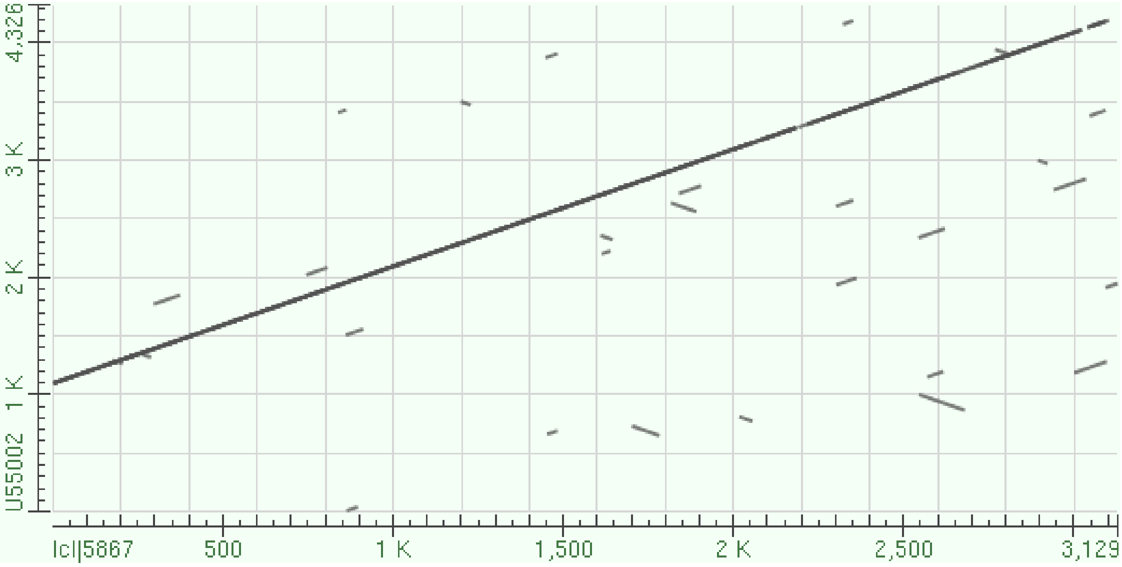


Figure 33. A dot matrix plot showing the tBlastX comparison between the genome for Panicum Mosaic Virus (U55002)(vertical) and a near full length genomic contig showing homology to the genus *Panicovirus*, isolated from the plant species *Paspalum setaceum*.

For further taxonomic characterization the amino acid sequences for the coat and polymerase genes viruses found in *Paspalum setaceum* and *Lespedeza procumbens* were aligned with the amino acid sequences found in other members of the *Tombusviridae* family and the phylogenetic trees were produced by Treeview X as shown in Figures 34 and 35. These two trees confirm the assertion that the virus found in the sample of *Lespedeza procumbens* belongs to a new species of virus, while the virus found in *Paspalum setaceum* represents a new strain of Panicum Mosaic Virus.

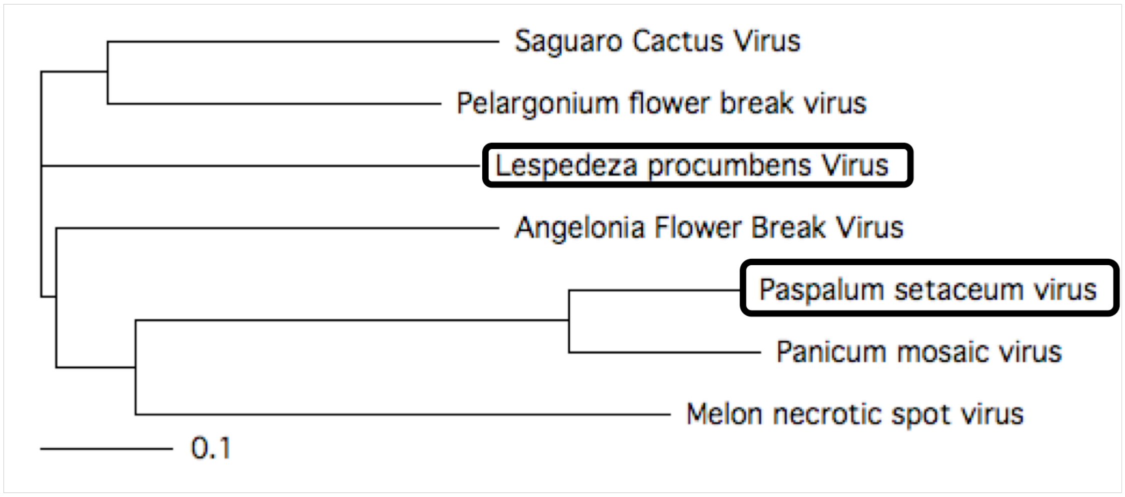

Saguaro Cactus Virus
Pelargonium flower break virus
Lespedeza procumbens Virus
Angelonia Flower Break Virus
Paspalum setaceum virus
Panicum mosaic virus
Melon necrotic spot virus
0.1

Figure 34. Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment showing the taxonomic relationships based on coat protein amino acid sequence for the two viruses found in the Tallgrass Prairie Preserve with previously described members of the family *Tombusviridae*. Saguaro Cactus Virus NP_044389. Pelargonium flower break virus ABD93309. Angelonia flower break virus YP_459964. Panicum mosaic virus NP_068346. Melon necrotic spot virus BAF47103.

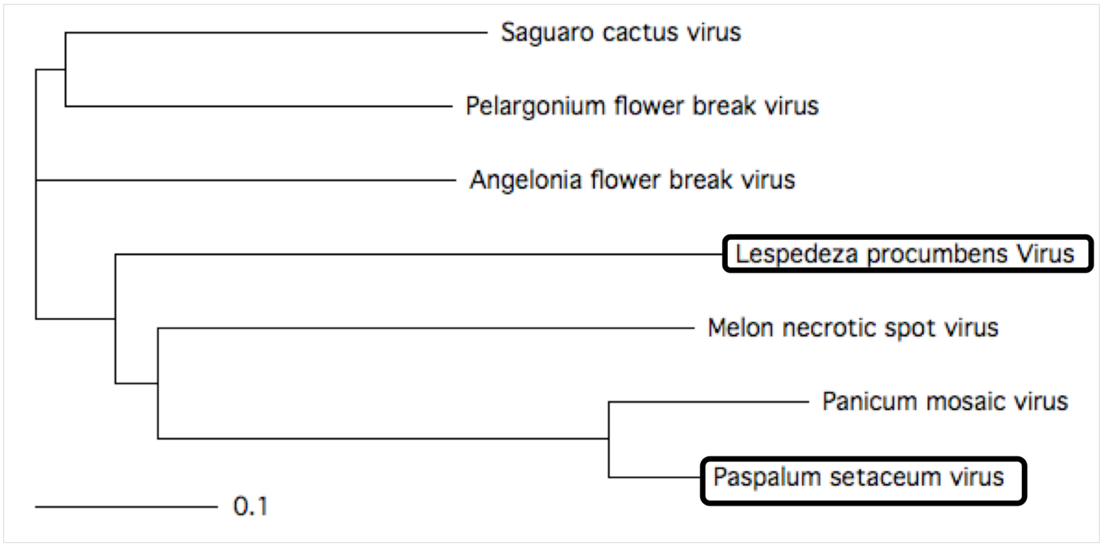

Saguaro cactus virus
Pelargonium flower break virus
Angelonia flower break virus
Lespedeza procumbens Virus
Melon necrotic spot virus
Panicum mosaic virus
Paspalum setaceum virus
0.1

Figure 35. A Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment showing the taxonomic relationships based on polymerase amino acid sequence for the two viruses found in the Tallgrass Prairie Preserve with previously described members of the family *Tombusviridae*. Saguaro Cactus Virus NP_044382. Pelargonium flower break virus NP_945123. Angelonia flower break virus YP_459960. Panicum mosaic virus AAC97551. Melon necrotic spot virus BAF47099.

3.5.16. *Totiviridae*

Totiviridae members are characterized by dsRNA genomes approximately 5 kb in length (Wickner et al 2005). Similar to the families *Chrysoviridae* and *Partitiviridae*, *Totiviridae* members are typically identified as fungal viruses, although there is some debate as to whether they may live within plants after being left behind by a parasitic fungal host (M. Roossinck, personal communication). In the Tallgrass Prairie Preserve 156 samples generated contigs with homology to this viral family, making it the second most prevalent viral family. Of these samples, 101 all had homology to a single species of *Totivirus*, Black Raspberry Fungal Virus. Further, 44% of these samples all belonged to the same plant species, *Ruellia humilis*. Of these *Ruellia humilis* samples 7 gave near full length genomic contigs ranging in size from 4002 nt to 4906 nt. The 3 largest of these contigs were compared using tBlastX with Blast2seq to the full length genome of

Black Raspberry Fungal Virus (Taxonomy ID 463392). These are given in Figures 36, 37, and 38.

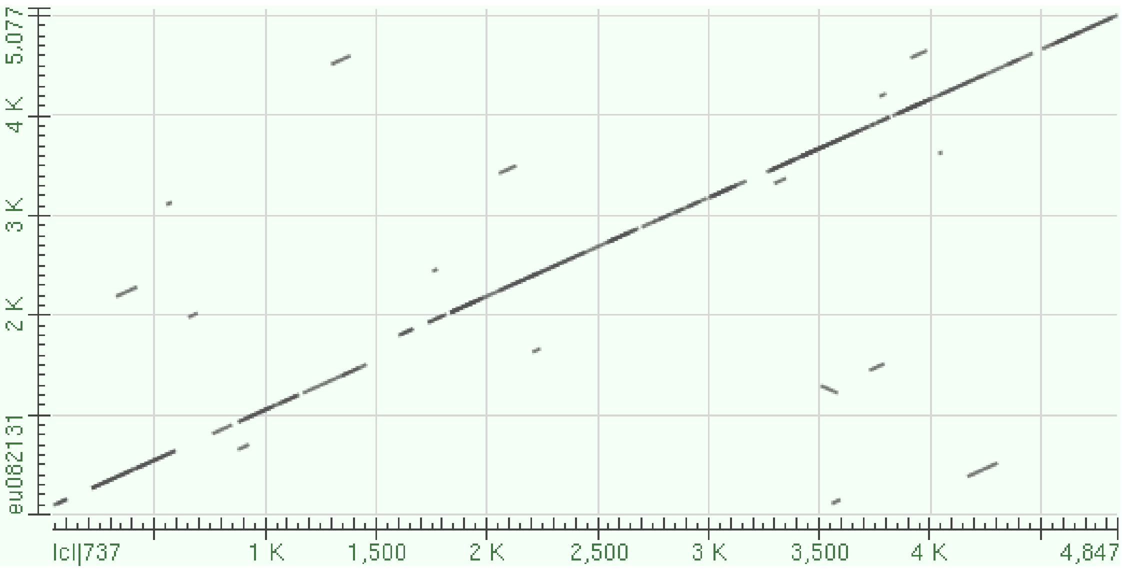


Figure 36. A dot matrix plot showing the tBlastX comparison of a near full length genomic contig generated from a species of *Ruellia humilis* (sample 08TGP00100) (horizontal) with the genome of Black Raspberry Fungal Virus (EU082131)(vertical).

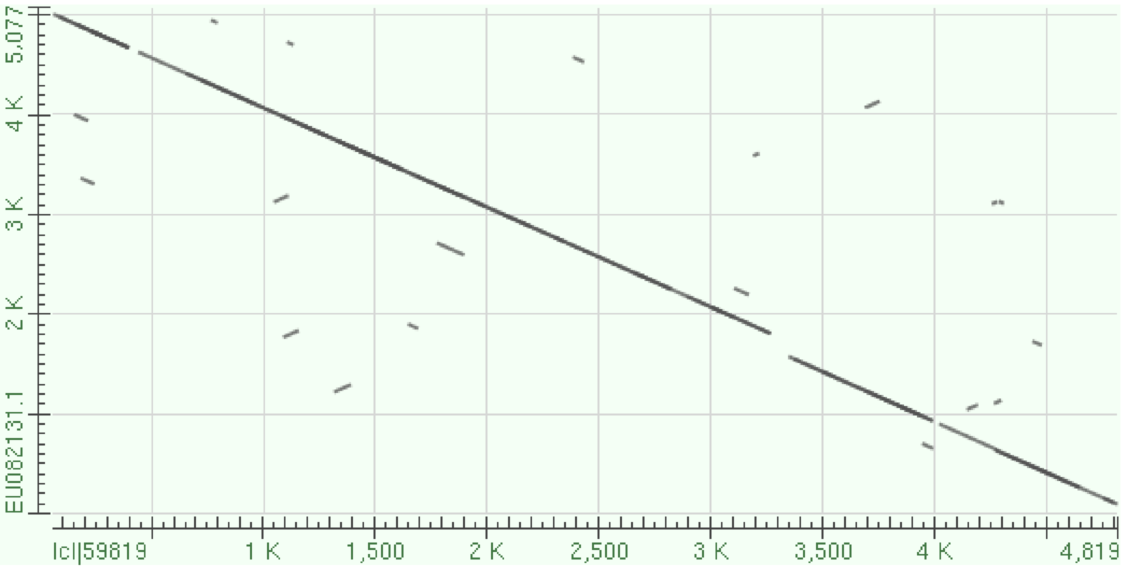


Figure 37. A dot matrix plot showing the tBlastX comparison of a near full length genomic contig generated from a species of *Ruellia humilis* (sample 08TGP00137) (horizontal) with the genome of Black Raspberry Fungal Virus (EU082131) (vertical).

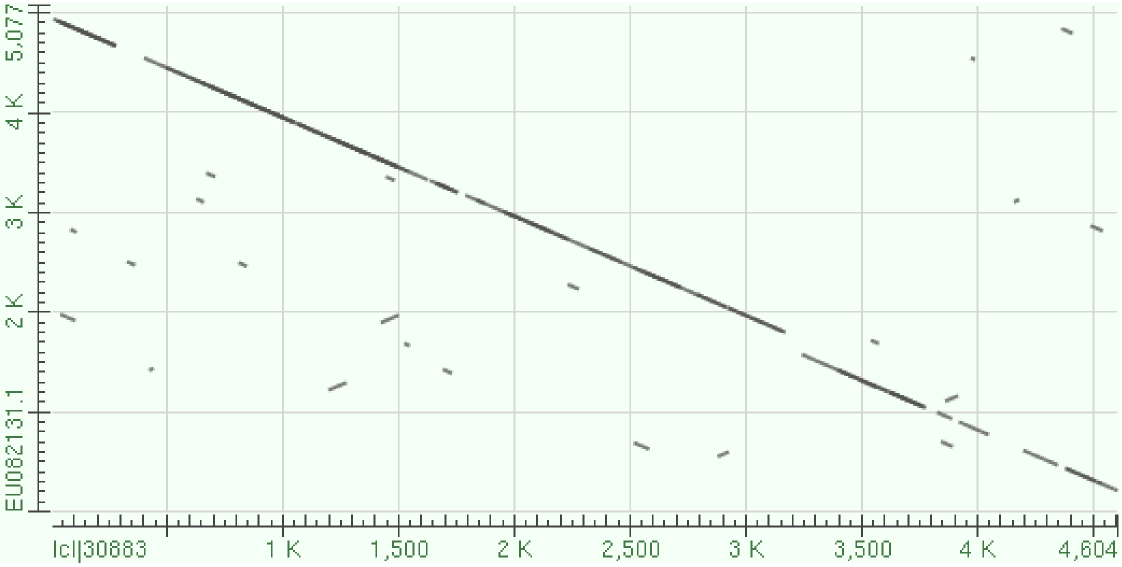


Figure 38. A dot matrix plot showing the tBlastX comparison of a near full length genomic contig generated from a species of *Ruellia humilis* (sample 06TGP01136) (horizontal) with the genome of Black Raspberry Fungal Virus (EU082131)(vertical).

BlastX comparisons between these contigs and the coat (ABU55398) and polymerase (ABU55399) proteins of Black Raspberry Fungal Virus can be seen in Table 13. Only 1 of the three contigs, from sample 08TGP00100, meets the criteria of less than 50% overall amino acid identity for speciation (Wickner et al 2005).

Table 13. Overall percentage of identity and similarity between the near full length genomic contigs from three samples of *Ruellia humilis* and the proteins encoded by Black Raspberry Fungal Virus

Sample	Coat Protein Identity	Coat Protein Similarity	RdRp Identity	RdRp Similarity
08TGP00100	35%	49%	41%	54%
08TGP00137	41%	53%	56%	73%
06TGP01136	51%	63%	56%	73%

3.5.17. *Tymoviridae*

The viral family *Tymoviridae* is characterized by (+)ssRNA genomes ranging in size from 6.5 kb to 7 kb (Dreher et al 2005). In this study the highest number of samples, 168, had contigs with significant homology to this therefore the most prevalent viral family. Of these samples, 41 were from the same species of plant, *Asclepias viridis* and 2 of these samples, 08TGP00060 and 08TGP00142, gave near full length genomic contigs of 6018 nt and 5386 nt respectively. A BlastN comparison between these 2 samples and their closest homologue, Okra Mosaic Virus (Taxonomy ID 70822) is shown as dot matrix plots in Figures 39 and 40. With an overall nucleotide sequence identity to Okra Mosaic Virus of 67% for both of these contigs, they fall outside of the speciation criteria of 80% overall nucleotide sequence identity (Dreher et al 2005). Additionally, the coat protein amino acid sequence for these samples showed 54% identity to Okra Mosaic Virus, also well outside the 90% threshold for speciation. A Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment of the coat protein amino acid sequence is given in Figure 41. When compared to each other, they show an overall sequence identity of 95%, demonstrating that this is the same viral species in both plants. A dot matrix plot for this comparison is given in Figure 42. Furthermore, all contigs showing *Tymoviridae* homology from samples of *Asclepias viridis* show 90% or greater identity when compared with BlastN, indicating that all *Tymoviridae* homologous viruses in *Asclepias viridis* are the same species of virus.

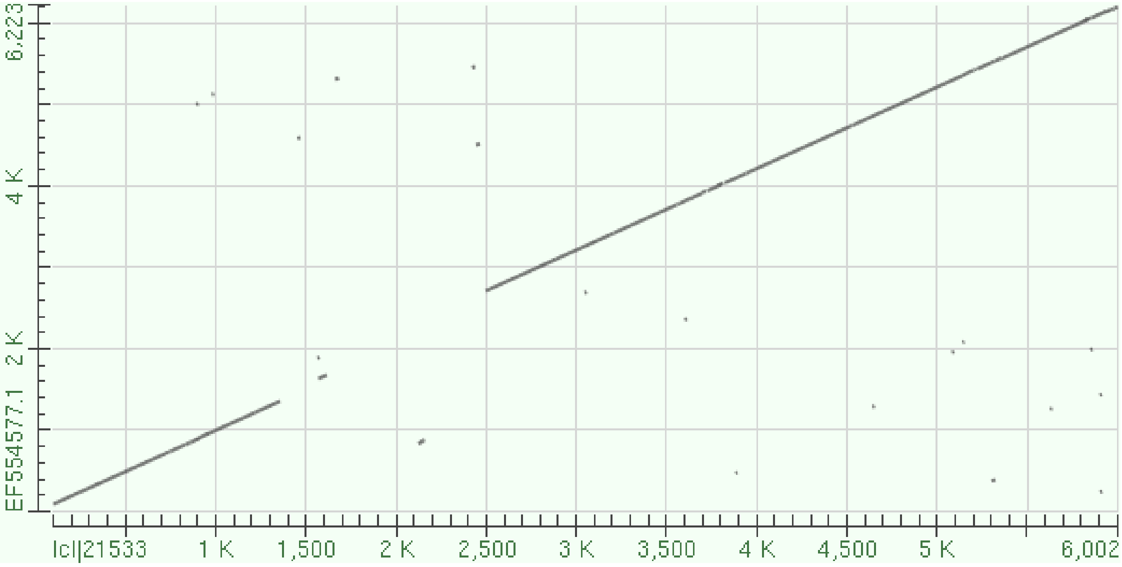


Figure 39. A BlastN dot matrix plot between the near full length genomic contig of sample 08TGP00060 (horizontal) and the genome of Okra Mosaic Virus (EF554577.1)(vertical).

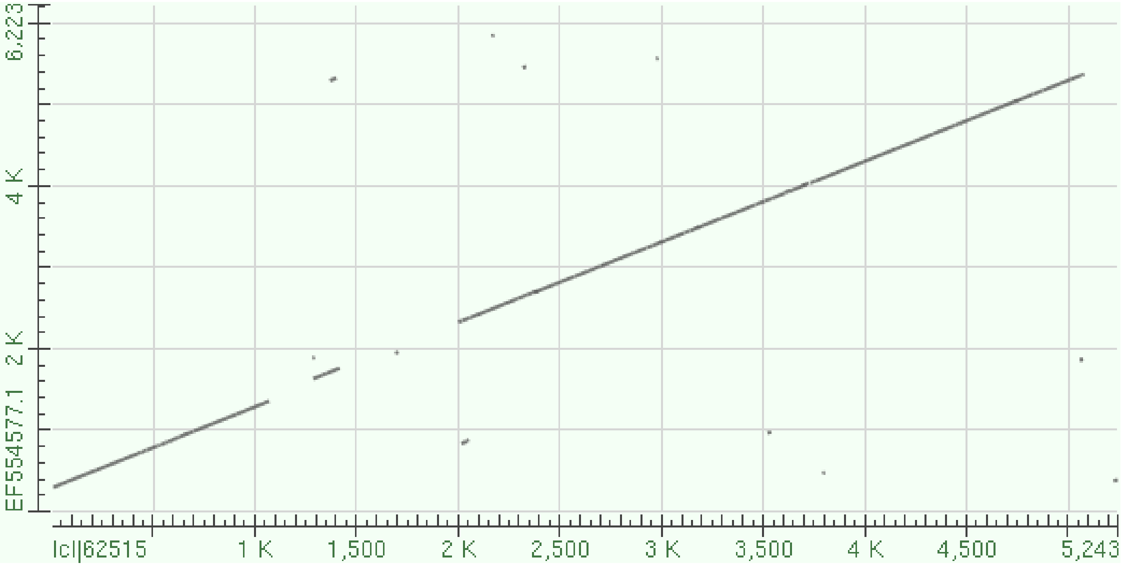


Figure 40. A BlastN dot matrix plot between the near full length genomic contig of sample 08TGP00142 (horizontal) and the genome of Okra Mosaic Virus (EF554577.1) (vertical).

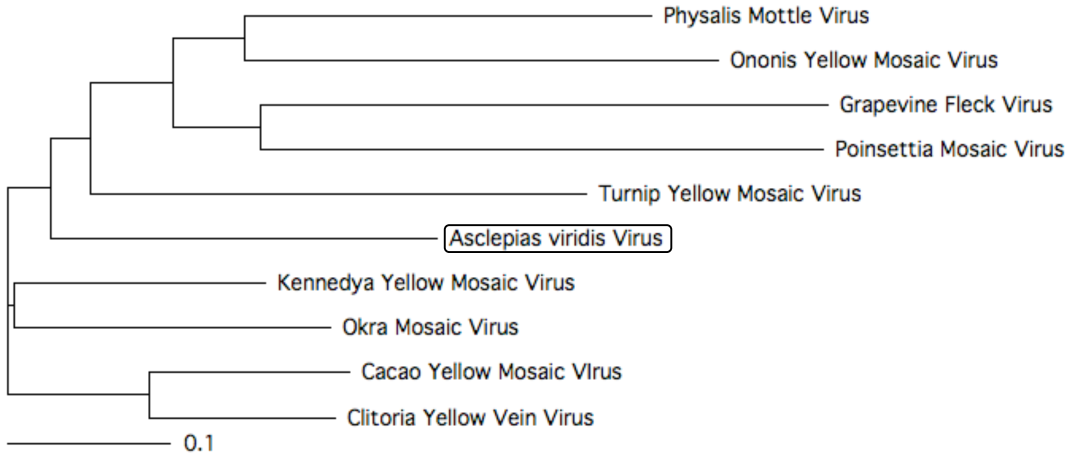

Physalis Mottle Virus
Ononis Yellow Mosaic Virus
Grapevine Fleck Virus
Poinsettia Mosaic Virus
Turnip Yellow Mosaic Virus
Asclepias viridis Virus
Kennedya Yellow Mosaic Virus
Okra Mosaic Virus
Cacao Yellow Mosaic VIrus
Clitoria Yellow Vein Virus
0.1

Figure 41. A Treeview X phylogenetic tree generated after a ClustalW multiple sequence alignment of the coat protein amino acid sequence for the virus found in *Asclepias viridis* and previously described members of the family *Tymoviridae*. Physalis Mottle Virus NP_619757.1. Ononis Yellow Mosaic Virus NP_041258. Grapevine Fleck Virus NP_542613. Poinsettia Mosaic Virus NP_733999. Turnip Yellow Mosaic Virus NP_663298. Kennedya Yellow Mosaic Virus NP_044329. Okra Mosaic Virus YP_001285473. Cacao Yellow Mosaic Virus P19128. Clitoria Yellow Vein Virus AAC25012.

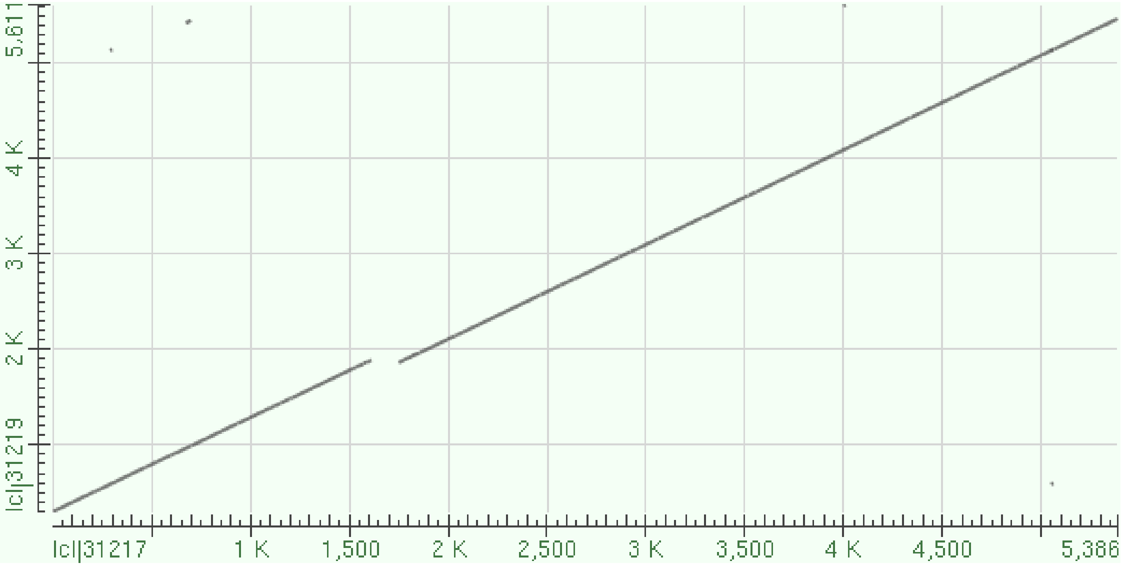


Figure 42. A BlastN dot matrix plot between the near full length genomic contig of sample 08TGP00142 (horizontal) and the near full length genomic contig of 08TGP00060 (vertical).

3.6. Distribution of Viruses within the TGP

3.6.1. Infection by plant family

As shown in Table 14 , the overall rate of infection for the Tallgrass Prairie Preserve is approximately 35%. Of those plant families which were sampled more than 10 times, the overall rate of infection was higher, at 40%, as shown in Table 15. The largest percentage of infections were found in plant families *Acanthaceae*, *Fagaceae*, and *Asclepiadaceae*.

Table 14. Percent of infection by plant family

Plant Family	Sampled	Infected	%infected	Plant Family	Sampled	Infected	%infected
Acanthaceae	86	63	73%	Loasaceae	1	0	0%
Aceraceae	2	0	0%	Lythraceae	3	1	33%
Alismataceae	1	0	0%	Malvaceae	2	0	0%
Amaranthaceae	3	0	0%	Marsileaceae	2	0	0%
Anacardiaceae	5	2	40%	Menispermaceae	2	1	50%
Annonaceae	1	0	0%	Molluginaceae	1	0	0%
Apiaceae	15	7	47%	Moraceae	4	0	0%
Apocynaceae	6	2	33%	Najadaceae	1	0	0%
Araceae	2	1	50%	Nelumbonaceae	1	0	0%
Aristolochiaceae	1	0	0%	Nyctaginaceae	1	0	0%
Asclepiadaceae	99	58	59%	Oleaceae	3	1	33%
Aspleniaceae	4	1	25%	Onagraceae	11	3	27%
Asteraceae	285	106	37%	Ophioglossaceae	3	1	33%
Betulaceae	1	1	100%	Orchidaceae	1	1	100%
Boraginaceae	6	3	50%	Oxalidaceae	3	1	33%
Brassicaceae	2	0	0%	Passifloraceae	1	1	100%
Cactaceae	2	1	50%	Phytolaccaceae	2	1	50%
Campanulaceae	4	2	50%	Plantaginaceae	6	1	17%
Caprifoliaceae	3	2	67%	Platanaceae	1	1	100%
Caryophyllaceae	5	2	40%	Poaceae	329	106	32%
Celastraceae	1	0	0%	Polygalaceae	2	0	0%
Ceratophyllaceae	1	1	100%	Polygonaceae	7	3	43%
Characeae	1	0	0%	Polypodiaceae	1	1	100%
Chenopodiaceae	5	1	20%	Portulacaceae	4	1	25%
Clusiaceae	3	1	33%	Potamogetonaceae	1	1	100%
Commelinaceae	2	0	0%	Primulaceae	2	0	0%
Convolvulaceae	3	1	33%	Pteridaceae	1	1	100%
Cornaceae	2	0	0%	Ranunculaceae	5	4	80%
Crassulaceae	2	1	50%	Rhamnaceae	1	0	0%
Cucurbitaceae	1	0	0%	Rosaceae	19	9	47%
Cupressaceae	2	0	0%	Rubiaceae	12	6	50%
Cuscutaceae	3	2	67%	Rutaceae	1	0	0%
Cyperaceae	24	8	33%	Salicaceae	5	3	60%
Dryopteridaceae	2	0	0%	Sapindaceae	1	0	0%
Ebenaceae	2	1	50%	Sapotaceae	1	1	100%
Equisetaceae	1	0	0%	Scrophulariaceae	12	3	25%
Euphorbiaceae	14	2	14%	Selaginellaceae	2	1	50%
Fabaceae	78	26	33%	Skipped number	1	0	0%
Fagaceae	24	15	63%	Smilacaceae	3	1	33%
Gentianaceae	4	3	75%	Solanaceae	11	2	18%
Geraniaceae	2	1	50%	Staphyleaceae	1	0	0%
Hippocastanaceae	1	0	0%	Typhaceae	1	0	0%
Hydrophyllaceae	1	0	0%	Ulmaceae	4	1	25%
Iridaceae	2	1	50%	Urticaceae	4	1	25%
Isoetaceae	3	1	33%	Valerianaceae	1	1	100%
Juglandaceae	4	3	75%	Verbenaceae	8	1	13%
Juncaceae	8	3	38%	Violaceae	1	0	0%
Lamiaceae	17	6	35%	Vitaceae	4	3	75%
Leucobryaceae	1	0	0%	Zannichelliaceae	1	1	100%
Liliaceae	6	3	50%	Zygophyllaceae	1	1	100%
Linaceae	2	1	50%	Total	1251	496	39%

Table 15. Overview of infection rate for plant families sampled more than 10 times.

Plant Family	Sampled	Infected	%infected
Acanthaceae	86	63	73%
Fagaceae	24	15	63%
Asclepiadaceae	99	58	59%
Rubiaceae	12	6	50%
Rosaceae	19	9	47%
Apiaceae	15	7	47%
Asteraceae	285	106	37%
Lamiaceae	17	6	35%
Fabaceae	78	26	33%
Cyperaceae	24	8	33%
Poaceae	329	106	32%
Onagraceae	11	3	27%
Scrophulariaceae	12	3	25%
Solanaceae	11	2	18%
Euphorbiaceae	14	2	14%
Total	1036	420	40%

3.7. Incidence of Multiple Infection

Of the 496 samples with contigs homologous to viruses, 146 had more than 1 virus family present, while 31 had more than 2 families and 8 had more than 3 families. The most multiply infected plant family sampled was *Asteraceae* with 28 samples having contigs with homology to more than 1 virus family. The two other plant families showing the highest number of multiple infections were *Poaceae* with 26 samples and *Acanthaceae* with 24 samples, as shown in Table 16.

Table 16. Number of samples by plant family with more than 2 viral families detected

Plant Family	Number of Samples	Plant Family	Number of Samples
Asteraceae	28	Betulaceae	1
Poaceae	26	Cactaceae	1
Acanthaceae	24	Campanulaceae	1
Asclepiadaceae	17	Caprifoliaceae	1
Fabaceae	9	Caryophyllaceae	1
Fagaceae	5	Crassulaceae	1
Apiaceae	3	Cuscutaceae	1
Cyperaceae	3	Ebenaceae	1
Euphorbiaceae	2	Liliaceae	1
Juglandaceae	2	Onagraceae	1
Juncaceae	2	Plantaginaceae	1
Lamiaceae	2	Polypodiaceae	1
Polygonaceae	2	Salicaceae	1
Ranunculaceae	2	Vitaceae	1
Rosaceae	2	Zygophyllaceae	1
Rubiaceae	2		

3.8. Co-Incidence of Virus and Fungus in Plant Samples

Two of the most prevalent families of virus in the Tallgrass Prairie Preserve, *Partitiviridae* and *Totiviridae*, are characteristically fungal viruses. Based on BlastN and BlastX searches, 520 of the 1254 samples processed, approximately 42%, contained contigs with homology to fungal sequences. Of the plant samples with contigs having homology to *Totiviridae* family viruses, approximately 33% contained contigs with fungal homology and of those plant samples with contigs homologous to *Partitiviridae* plant viruses, approximately 50% also had fungi-homologous contigs. Based on these observations, it is entirely possible that the viruses detected belonging to the families *Partitiviridae* and *Totiviridae* were present within parasitic fungi associated with the plant.

As a comparison viral-fungal co-incidence analysis from a similar study concerning plants from the Area Conservation Guanacast (ACG) region located in northwestern Costa Rica (Quan 2008) revealed that of the 2,688 samples processed, 2530, or 94%, had contigs homologous to fungal sequences. For those samples with contigs having homology to *Partitiviridae* viruses, 301 of 307, 98%, also had contigs with fungal homology. Samples with contigs with homology to *Totiviridae* were similar with 143 of 144, 99%, also had contigs with fungal homology. An overview of these analyses is given in Table 17.

Table 17. Overview of fungal and fungus virus detection for the Tallgrass Prairie Preserve and Area Conservation Guanacast

Sampled Region	**Tallgrass Prairie Preserve**	**Area Conservation Guanacast**
Number of Samples	1254	2688
Samples with Fungus Detected	520	2530
Samples with Fungal Virus Detected	282	451
Samples with both Fungal Virus and Fungi Detected	117	444
Samples with Partitiviridae Detected	140	307
Samples with both Partitiviridae and Fungi Detected	70	301
Samples with Totiviridae Detected	142	144
Samples with both Totitiviridae and Fungi Detected	47	143

Chapter 4 Conclusions

The modification of the protocol for library preparation for pyrosequencing by using Ampure SPRI beads instead of Qiagen mini-columns, removing the unnecessary steps originally designed to enrich for single stranded DNA, and optimization of the overall protocol, resulted in a streamlined procedure that then was automated using the Zymark SciClone robot. When coupled with the TGPweb data analysis system, consisting of a mySQL database, a web interface, and genomic analysis programs, a powerful set off experimental tools was produced during the course of this research that resulted in the comprehensive analysis of the viral metagenomic data obtained from plants harvested from the Tallgrass Prairie Preserve in Northern Oklahoma.

Overall, approximately 35% to 40% of the plants sampled in the Tallgrass Prairie Preserve were infected with viruses. The results showed a wide range of viral infection levels for those plant families sampled more than 10 times, from a high of 73% of the *Acanthaceae* family members and 14% of the *Euphorbiaceae* family members being infected. Based on these infection levels, we reached the surprising conclusion that although there was wide spread plant virus infections in the Tallgrass Prairie Preserve, these infections were not uniformly distributed, as some plant families were more susceptible to infection than others.

Since 19 of 36 known plant virus families were detected, there also was a significant diversity in the viruses found in the Tallgrass Prairie Preserve. This is shown in Figure 43. The majority of viral families detected were (+)ssRNA viruses, while the least represented genome type was a single plant sample that contained a virus with a (-)ssRNA genome. This echoes the overall trend in plant viruses towards (+)ssRNA genomes. The most prevalent viruses observed were from the *Totiviridae*, *Partitiviridae*, and *Tymoviridae* families.

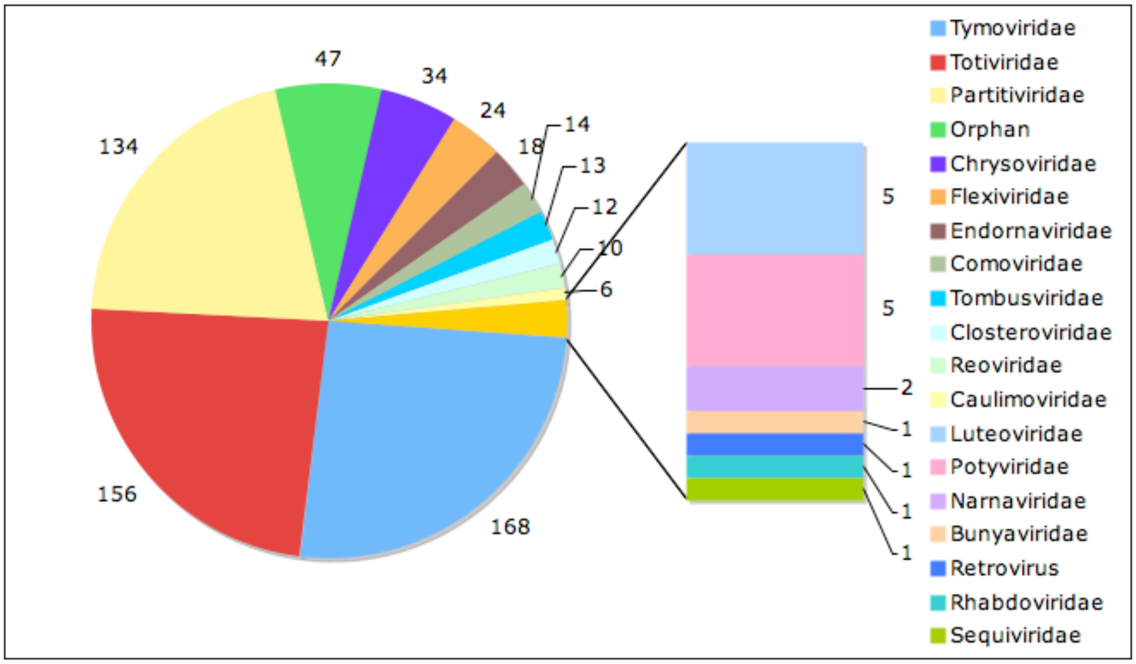


Figure 43. The overall distribution of viral families from the Tallgrass Prairie Preserve.

During these studies, several new species of virus were discovered that include new members of the families *Chrysoviridae*, *Comoviridae*, *Flexiviridae*, *Partitiviridae*, *Reoviridae*, *Tombusviridae*, *Totiviridae*, and *Tymoviridae*. Since such a large number of new species were observed, including indicators of new genera within viral families from a diverse range of plant viruses, the original hypothesis that uncultivated plants may harbor a significant level of previously unknown viruses, has been confirmed. Based on the *Tymoviridae* homologous sequence found in *Asclepias viridis*, we also conclude that these uncultivated plants may act as reservoirs of novel viruses.

Approximately 30% of the plant samples collected from the Tallgrass Prairie Preserve were infected by more than one type of plant virus, with 6% having 2 or 3 viral families detected within the same sample, and approximately 1% showing infection with 4 or more simultaneous viruses. Based in these observations we conclude that while there is a large prevalence of viruses in the Tallgrass Prairie, the percentage of plants harboring more than one viral species was low.

A majority of multiple infections include either the *Totiviridae* or *Partitiviridae* family, the two most prevalent viral families observed in viral infected plants from the Tallgrass Prairie Preserve, that also are fungal viruses. Based on these observations, as well as the observation that 520 of the 1254 samples had contigs with fungal homology, it is possible that the viruses homologous to the families *Totiviridae* and *Partitiviridae* are actually carried within a fungus that is infecting the plant. Consistent with this is the observation that a large percentage of samples showing infection by either of these two viral families also showed indications of being infected by a fungus as well.

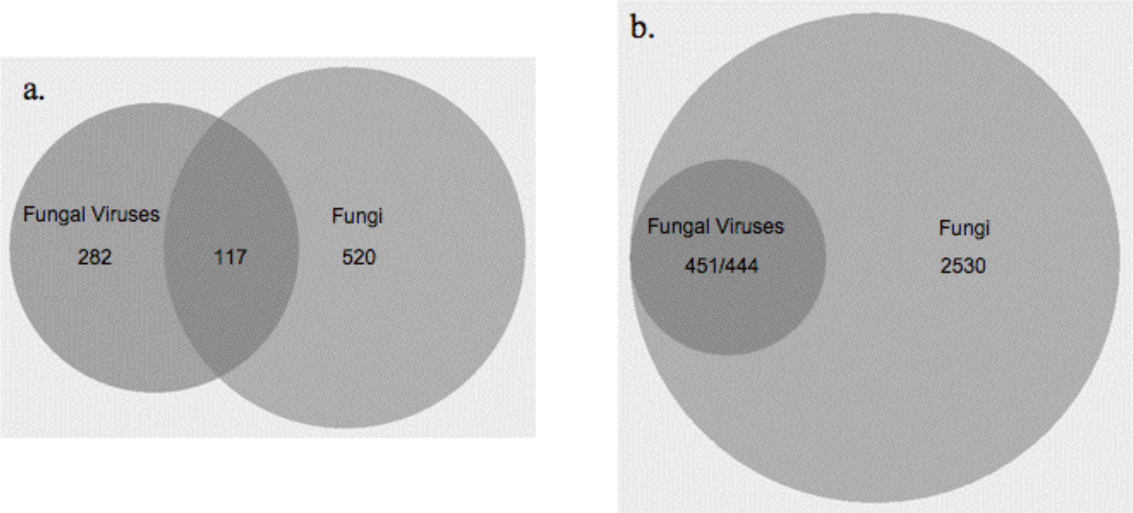


Figure 44. Venn diagrams illustrating the overlap of samples with fungi and fungal viruses detected from (a) the Tallgrass Prairie Preserve, and (b) the Area Conservation Guanacast..

As illustrated in Figure 44a,b, this observation was consistent with those made on plants harvested from the Area Conservation Guanacast in Costa Rica where 98% of all *Partitiviridae* and *Totiviridae* positive samples also were positive for fungal sequences.

In summary, based on my analysis, plants harvested from the Tallgrass Prairie Preserve in Northern Oklahoma, harbor a much higher number of biologically diverse viruses than previously thought, including many new species and several novel genera, affirming the hypothesis that uncultivated plants contain many previously unobserved, new and novel viruses. In addition, although widespread viral infections were observed, they were not evenly distributed on all plants, and at least one virus, a *Tymoviridae,* was widely distributed on a single plant species, *Asclepias viridis*, demonstrating a new ecological niche for this virus. In addition, although observed, multiple viral infections on a single plant were not common. Finally, and most surprising of all, because over a third of the plants studied also likely were infected with fungi and numerous fungal viruses were present, it can be concluded that the majority of plant viruses observed actually may be transmitted to the plant via infecting fungi, although the exact mechanism underlying this process will require further study.

References

Aaziz R, Dinant S, Epel BL. 2001. Plasmodesmata and plant cytoskeleton. Trends in Plant Science 6(7):326-330.

Adams MH. 1952. Classification of bacterial viruses: characteristics of the T5 species and of the T2, C16 species. J Bacteriol 64(3):387-96.

Adams MJ, Accotto GP, Agranovsky AA, Bar-Joseph M, Boscia D, Brunt AA, Candresse T, Coutts RHA, Dolja VV, Falk BW. 2005. Flexiviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 1089-1124.

Ahlquist P. 1999. Bromoviruses (Bromoviridae). Encyclopedia of Virology 2nd edn.:198-204.

Ahn IIP, Lee YH. 2001. A viral double-stranded RNA up regulates the fungal virulence of Nectria radicicola. Molecular Plant-Microbe Interactions 14(4):496-507.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol 215(3):403-410.

Attoui H, Mohd Jaafar F, Belhouchet M, Biagini P, Cantaloube JF, de Micco P, de Lamballerie X. 2005. Expansion of family Reoviridae to include nine-segmented dsRNA viruses: Isolation and characterization of a new virus designated aedes pseudoscutellaris reovirus assigned to a proposed genus (Dinovernavirus). Virology 343(2):212-223.

Bains W, Smith GC. 1988. A novel method for nucleic acid sequence determination. J Theor Biol 135(3):303-7.

Ban N, McPherson A. 1995. The structure of satellite panicum mosaic virus at 1. 9 Aa resolution. Nature Structural Biology 2(10):882-890.

Banks GT, Buck KW, Chain EB, Darbyshire JE, Himmelweit F. 1969. Virus-like particles in penicillin producing strains of Penicillium chrysogenum. Nature 222(5188):89-90.

Bennett CW. 1940. Acquisition and transmission of viruses by dodder (Cuscuta subinclusa). Phytopathology(30):2-11.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2007. GenBank. Nucleic Acids Research 37:D26-31.

Berger PH, Brunt AA, EdwardsonJr HJ, HillJh JRL, Kashiwazaki S, Makkouk KM, Morales FJ, Rybicki E, Spence N, Ohki ST. 2005. Potyviridae. Eighth Report of the

International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 819–841.

Bessman MJ, Lehman IR, Adler J, Zimmerman SB, Simms ES, Kornberg A. 1958. Enzymatic synthesisof deoxyribonucleic acid III. The incorporation of pyridine and purine analogues into deoxyribonucleic acid. Proceedings of the National Academy of Sciences of the United States of America 44(7):633-640.

Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E. 2007. The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. PLoS ONE 2(2).

Boccardo G, Candresse T. 2005. Complete sequence of the RNA1 of an isolate of White clover cryptic virus 1, type species of the genus Alphacryptovirus. Archives of virology 150(2):399-402.

Bodaghi S, Mathews DM, Dodds JA. 2004. Natural Incidence of Mixed Infections and Experimental Cross Protection Between Two Genotypes of Tobacco mild green mosaic virus. Phytopathology 94(12):1337-1341.

Boshoff C, Schulz TF, Kennedy MM, Graham AK, Fisher C, Thomas A, McGee JOD, Weiss RA, O'Leary JJ. 1995. Kaposi's sarcoma-associated herpesvirus infects endothelial and spindle cells. Nature Medicine 1(12):1274-1278.

Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F. 2004. Diversity and population structure of a near-shore marine-sediment viral community. Proceedings of the Royal Society B: Biological Sciences 271(1539):565-574.

Brierley I. 1995. Ribosomal frameshifting viral RNAs. Journal of General Virology 76(8):1885-1892.

Broadbent L. 1963. The epidemiology of tomato mosaic III. Cleaning virus from hands and tools. Annals of Applied Biology 52(2):225-232.

Broadbent L. 1965. The epidemiology of tomato mosaic IX. Transmission of TMV by birds. Annals of Applied Biology 55(1):67-69.

Brown DJF, Robertson WM, Trudgill DL. 1995. Transmission of Viruses by Plant Nematodes. Annual Reviews in Phytopathology 33(1):223-249.

Brown DJF, Weischer B. 1998. Specificity exclusivity and complementarity in the transmission of plant viruses by plant parasitic nematodes: An annotated terminology. Fundamental and applied nematology 21(1):1-11.

Buck K. 1999. Geminiviruses (Geminiviridae). Encyclopedia of Virology 2nd edn.:597-606.

Buck KW, Esteban R, Hillman BI. 2005. Narnaviridae. Eighth Report of the International Committee for the Taxonomy of Viruses. . San Diego: Academic Press, Elsevier. p. 751-756.

Campbell RN. 1996. Fungal transmission of plant viruses. Annual Review of Phytopathology 34(1):87-108.

Canady MA, Larson SB, Day J, McPherson A. 1996. Crystal structure of turnip yellow mosaic virus. Nature Structural Biology 3(9):771-781.

Clinch P, Loughnane JB, Murphy PA. 1938. A study of the infiltration of viruses into seed potato stocks in the field. In. 1938. p. 17.

Covelli L, Coutts RHA, Serio FD, Citir A, Acikgoz S, Hernandez C, Ragozzino A, Flores R. 2004. Cherry chlorotic rusty spot and Amasya cherry diseases are associated with a complex pattern of mycoviral-like double-stranded RNAs. I. Characterization of a new species in the genus Chrysovirus. Soc General Microbiol. p. 3389-3397.

Crick F. 1970. Central dogma of molecular biology. Nature 227(5258):561-3.

D'Arcy CJ, Domier LL. 2005. Luteoviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 891–900.

Day PR. 1981. Fungal virus populations in corn smut from Connecticut. Mycologia 73(3):379-391.

Delwart EL. 2007. Viral metagenomics. Reviews in Medical Virology 17(2):115-131.

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B. 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452(7185):340-343.

Dominguez DI, Ryabova LA, Pooggin MM, Schmidt-Puchta W, Futterer J, Hohn T. 1998. Ribosome Shunting in Cauliflower Mosaic Virus, Identification of and Essential and Sufficient Structural Element. Journal of Biological Chemistry 273(6):3669-3678.

Dougherty WG, Semler BL. 1993. Expression of virus-encoded proteinases: functional and structural similarities with cellular enzymes. Microbiology and Molecular Biology Reviews 57(4):781-822.

Dovichi NJ. 1997. DNA sequencing by capillary electrophoresis. Electrophoresis 18.

Dreher TW, Edwards MC, Gibbs AJ, Haenni AL, Hammond RW, Jupin I, Koenig R, Sabanadzovic S, Abou Ghanem-Sabanadzovic N, Martelli GP. 2005. Tymoviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 1067–1076.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences 100(15):8817-8822.

Ehlers K, Binding H, Kollmann R. 1999. The formation of symplasmic domains by plugging of plasmodesmata: a general event in plant morphogenesis? Protoplasma 209(3):181-192.

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA. 2005. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier.

Flotte TR, Berns KI. 2005. Adeno-Associated Virus: A Ubiquitous Commensal of Mammals. Human Gene Therapy 16(4):401-407.

Fütterer J, Hohn T. 1996. Translation in plants-rules and exceptions. Plant Molecular Biology 32(1):159-189.

Ghabrial SA. 1980. Effects of fungal viruses on their hosts. Annual Review of Phytopathology 18(1):441-461.

Ghabrial SA, Bozarth RF, Buck KW, Yamashita S, Martelli GP, Milne RG. 2005a. Partitiviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 581-590.

Ghabrial SA, Caston J, Flores R, Hillman BI, Jiang D. 2005b. Chrysoviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 591-596.

Ghadessy FJ, Ong JL, Holliger P. 2001. Directed evolution of polymerase function by compartmentalized self-replication. Proceedings of the National Academy of Sciences 98(8):4552-4557.

Goelet P, Lomonossoff GP, Butler PJG, Akam ME, Gait MJ, Karn J. 1982. Nucleotide Sequence of Tobacco Mosaic Virus RNA. Proceedings of the National Academy of Sciences 79(19):5818-5822.

Griffiths P. 1999. Time to consider the concept of a commensal virus? Rev Med Virol 9(2):73-4.

Grogan RG, Campbell RN. 1966. Fungi as vectors and hosts of viruses. Annual Review of Phytopathology 4(1):29-52.

Haenni AL. 2008. Virus Evolution and Taxonomy. In: Roossinck MJ, editor. Plant Virus Evolution. Springer. p. 205-218.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5(10):245-249.

Hayes W. 1960. DNA and Biological Research. Proceedings of the Royal Society of Medicine 53(8):619.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/Chemical Monthly 125(2):167-188.

Hohn T, Fütterer J. 1997. The proteins and functions of plant pararetroviruses: knowns and unknowns. Crit. Rev. Plant Sci 16(1):133-161.

Holland J, Domingo E, Origins V. 1998. Origin and Evolution of Viruses. Virus Genes 16(1):13-21.

Hull R. 2002. Matthews' Plant Virology. San Diego: Elsevier Ltd.

Jackson AO, Dietzgen RG, Goodin MM, Bragg JN, Deng M. 2005. Biology of Plant Rhabdoviruses. Annual Review of Phytopathology 43:623-660.

Jackson AO, Goodin M, Moreno I, Johnson J, Lawrence DM. 1999. Plant rhabdoviruses. Encyclopedia of virology, 2nd ed. Academic Press, New York, NY:1531-1541.

Jett JH, Keller RA, Martin JC, Marrone BL, Moyzis RK, Ratliff RL, Seitzinger NK, Shera EB, Stewart CC. 1989. High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules. J Biomol Struct Dyn 7(2):301-309.

Karasev AV, Hilf ME, Garnsey SM, Dawson WO. 1997. Transcriptional strategy of closteroviruses: mapping the 5'termini of the citrus tristeza virus subgenomic RNAs. Journal of Virology 71(8):6233-6236.

Kassanis B. 1962. Properties and behaviour of a virus depending for its multiplication on another. J Gen Microbiol 27:477-88.

Kasteel D. 1999. Structure, morphogenesis and function of tubular structures induced by cowpea mosaic virus. LUW Wageningen (Netherlands). p. 71.

Kim KH, Chang HW, Nam YD, Roh SW, Kim MS, Sung Y, Jeon CO, Oh HM, Bae JW. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol 74(19):5975-85.

Koenig RaL, D.-E. 2005b. Pomovirus. Eighth Report of the International Committee on Taxonomy of Viruses:1033-1038.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947-2948.

Le Gall O, Iwanami T, Karasev AV, Jones T, Lehto K, Sanfacon H, Wellink J, Wetzel T, Yoshikawa N. 2005a. Comoviridae. Eighth Report of the International Committee on Taxonomy Of Viruses. San Diego: Academic Press, Elsevier.

Le Gall O, Iwanami Y, Karasev AE, Jones T, Lehto K, Sanfaçon H, Wellink J, Wetzel T, Yoshikawa NS. 2005b. Sequiviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier.

Lehtinen DA, Perrino FW. 2004. Dysfunctional proofreading in the Escherichia coli DNA polymerase III core. Biochemical Journal 384(Pt 2):337-348.

Lemke PA, Nash CH. 1974. Fungal viruses. Microbiology and Molecular Biology Reviews 38(1):29-56.

Lewandowski DJ. 2005. Tobamovirus. Eighth Report of the International Committee on Taxonomy of Viruses:1009-1014.

Lommel SA, Martelli GP, Rubino L, Russo M. 2005. Tombusviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 906–936.

Maki H, Kornberg A. 1987. Proofreading by DNA Polymerase III of Escherichia coli Depends on Cooperative Interaction of the Polymerase and Exonuclease Subunits. Proceedings of the National Academy of Sciences 84(13):4389-4392.

Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD. 2007. CDD: a conserved domain database for interactive domain family analysis. Nucleic acids research 35(Database issue):D237.

Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. Nucleic acids research 32(Web Server Issue):W327.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Marquez LM, Redman RS, Rodriguez RJ, Roossinck MJ. 2007. A Virus in a Fungus in a Plant: Three-Way Symbiosis Required for Thermal Tolerance. Science 315(5811):513-515.

Maule AJ, Wang D. 1996. Seed transmission of plant viruses: a lesson in biological complexity. Trends in Microbiology 4(4):153-158.

Melcher U, Muthukumar V, Wiley GB, Min BE, Palmer MW, Verchot-Lubicz J, Ali A, Nelson RS, Roe BA, Thapa V. 2008. Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from Ambrosia psilostachya. Journal of Virological Methods 152(1-2):49-55.

Mertens PPC, Arella M, Attoui H, Belloncik S, Bergoin M, Boccardo G, Booth TF, Chiu W, Diprose JM, Duncan R. 2005. Reoviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Deigo: Academic Press, Elsevier. p. 447-454.

Nassuth A, Pollari E, Helmeczy K, Stewart S, Kofalvi SA. 2000. Improved RNA extraction and one-tube RT-PCR assay for simultaneous detection of control plant RNA plus several viruses in plant extracts. Journal of Virological Methods 90(1):37-49.

Nault LR. 1997. Arthropod Transmission of Plant Viruses: A New Synthesis. Annals-Entomological Society of America 90:521-541.

Nelson RS, Van Bel AJE. 1998. The Mystery of Virus Trafficking Into, Through and Out of Vascular Tissue. Progress in Botany 59:476-533.

Nichol ST, Beaty BJ, Elliot RM, Goldbach R, Plyusin A, Schmaljohn CS, Tesh RB. 2005. Bunyaviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 695–716.

Oh CS, Hillman BI. 1995. Genome organization of a partitivirus from the filamentous ascomycete Atkinsonella hypoxylon. Journal of General Virology 76(6):1461.

Ooi K, Yahara T. 1999. Genetic variation of geminiviruses: comparison between sexual and asexual host plant populations. Molecular Ecology 8(1):89-97.

Osaki H, Nakamura H, Sasaki A, Matsumoto N, Yoshida K. 2006. An endornavirus from a hypovirulent strain of the violet root rot fungus, Helicobasidium mompa. Virus Research 118(1-2):143-149.

Osaki H, Nomura K, Iwanami T, Kanematsu S, Okabe I, Matsumoto N, Sasaki A, Ohtsu Y. 2002. Detection of a Double-Stranded RNA Virus from a Strain of the Violet Root Rot Fungus *Helicobasidium mompa Tanaka*. Virus Genes 25(2):139-145.

Page RDM. 1996. Tree View: An application to display phylogenetic trees on personal computers. Oxford Univ Press. p. 357-358.

Peak IRA, Jennings MP, Hood DW, Bisercic M, Moxon ER. 1996. Tetrameric repeat units associated with virulence factor phase variation in *Haemophilus* also occur in *Netsseria* spp. and *Moraxella catarrhalis*. FEMS microbiology letters 137(1):109-114.

Pellegrin F, Duran-Vila N, Van Munster M, Nandris D. 2007. Rubber tree (Hevea brasiliensis) trunk phloem necrosis: aetiological investigations failed to confirm any biotic causal agent. Forest Pathology 37(1):9-21.

Prüfer D, Tacke E, Schmitz J, Kull B, Kaufmann A, Rohde W. 1992. Ribosomal frameshifting in plants: a novel signal directs the-1 frameshift in the synthesis of the putative viral replicase of potato leafroll luteovirus. The EMBO Journal 11(3):1111-1117.

Quan J. 2008. Characterization and Comparative Analysis on Metagenome Data from a Plant RNA Virus Community. University of Oklahoma. p. 102.

Robertson NL. 2005. A newly described plant disease complex involving two distinct viruses in a native Alaskan lily, Streptopus amplexifolius. CANADIAN JOURNAL OF BOTANY 83(10):1257.

Roossinck MJ. 2003. Plant RNA virus evolution. Current Opinion in Microbiology 6(4):406-409.

Roossinck MJ. 2005a. Symbiosis versus competition in plant virus evolution. Nature Reviews Microbiology 3(12):917-924.

Roossinck MJ, Bujarski, j., Ding, S.W., Hajimorad, R., Hanada, K., Scott, S., and Tousignant, M. 2005b. Alfamovirus/Bromoviridae. Eighth Report of the International Committee on Taxonomy of Viruses:1051-1052.

Sanger F, Nicklen S, Coulson AR. 1977. DNA Sequencing with Chain-Terminating Inhibitors. Proceedings of the National Academy of Sciences 74(12):5463-5467.

Semancik JS. 1986. Separation of viroid RNAs by cellulose chromatography indicating conformational distinctions. Virology 155(1):39-45.

Shirako Y. 1998. Non-AUG Translation Initiation in a Plant RNA Virus: a Forty-Amino-Acid Extension Is Added to the N Terminus of the Soil-Borne Wheat Mosaic Virus Capsid Protein. Am Soc Microbiol. p. 1677-1682.

Stangegaard M, Dufva IH, Dufva M. 2006. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. BioTechniques 40(5):649-657.

Stanley J, Bisaro, D.M., Briddon, R.W., Brown, J.K., Fauqet, C.M., Harrison, B.D, Rybicki, E.P., and Stenger, D.C. 2005. Geminiviridae. Eighth Report of the International Committee on Taxonomy of Viruses:301-326.

States DJ, Gish W, Altschul SF. 1991. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. Methods 3(1):66-70.

Tatusova TA, Madden TL. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS microbiology letters 174(2):247-250.

Thivierge K, Nicaise V, Dufresne PJ, Cotton S, Laliberte JF, Le Gall O, Fortin MG. 2005. Plant Virus RNAs. Coordinated Recruitment of Conserved Host Functions by (+) ssRNA Viruses during Early Infection Events 1. Am Soc Plant Biol. p. 1822-1827.

Thompson GA, Schulz A. 1999. Macromolecular trafficking in the phloem. Trends in Plant Science 4(9):354-360.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW. Nucleic Acids Res 22:4673-4680.

Tollin PaW, H.R. 1988. Particle Structure. The Plant Viruses. Vol. 4: The Filamentous Plant Viruses:51-83.

Tordo N, Benmansour A, Calisher C, Dietzgen RG, Fang RX, Jackson AO, Kurath G, Nadin-Davis S, Tesh RB, Walker PJ. 2005. Rhabdoviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 623–644.

Torrance LaK, R. 2005. Furovirus. Eighth Report of the International Committee on Taxonomy of Viruses:1027-1032.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP et al. . 2009. A core gut microbiome in obese and lean twins. Nature 457(7228):480-484.

Upadhyana NMaM, P.P.C. 2005. Oryzavirus. Eighth Report of the International Committee on Taxonomy of Viruses:550-555.

van Regenmortel MHV, Fauquet CM, Bishop DHL, Carstens EB, Estes MK, Lemon SM, McGeoch DJ, Maniloff J, Mayo MA, Pringle CR. 2000. Seventh Report of the International Committee on Taxonomy of Viruses. San Diego Academic Press.

Watson JD, Crick FHC. 1953. A structure for deoxyribonucleic acid. Nature 171:737-738.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872-876.

Whitfield JB. 2002. Estimating the age of the polydnavirus/braconid wasp symbiosis. Proceedings of the National Academy of Sciences USA 99(11):7508-7513.

Wickner RB. 1993. Double-stranded RNA virus replication and packaging. ASBMB. p. 3797-3800.

Wickner RB, Ghabrial SA, Bruenn JA, Buck KW, Patterson JL, Stuart KD, Wang CC. 2005. Totiviridae. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego: Academic Press, Elsevier. p. 571–580.

Wilcox RM, Fuhrman JA. 1994. Bacterial viruses in coastal seawater: lytic rather than lysogenic production. Marine Ecology-Progress Series 114:35-35.

Wiley G, Macmil S, Qu C, Wang P, Yanbo, Xing, White D, Li J, D.White J, Domingo A et al. . 2009. Methods for Generating Shotgun and Mixed Shotgun/Paired-End Libraries for the 454 DNA Sequencer. In Press.

Wolf S, Lucas WJ, Deom CM, Beachy RN. 1989. Movement Protein of Tobacco Mosaic Virus Modifies Plasmodesmatal Size Exclusion Limit. Science 246(4928):377-379.

Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U. 2006. Plant virus biodiversity and ecology. PLoS Biol 4(3):e80.

Zaccomer B, Haenni AL, Macaya G. 1995. The remarkable variety of plant RNA virus genomes. Journal of General Virology 76(2):231-247.

Zaitlin M. 1962. Graft transmissibility of a systemic virus infection to a hypersensitive host—An interpretation. Phytopathology 52:1222-1223.

Zaitlin M. 1998. The Discovery of the Causal Agent of the Tobacco Mosaic Disease. Discoveries in Plant Biology. SD Kung and SF Yang, eds. World Publishing Co. Ltd. Hong Kong./education/feature/TMV/pdfs/zaitlin. pdf:105-110.

Zhou T, Boland GJ. 1997. Hypovirulence and double-stranded RNA in Sclerotinia homoeocarpa. Phytopathology 87(2):147-153.

Zou Z, Najar F, Wang Y, Roe B, Jiang H. 2008. Pyrosequence analysis of expressed sequence tags for Manduca sexta hemolymph proteins involved in immune responses. Insect Biochemistry and Molecular Biology 38(6):677-682.

Appendix 1

Viral Morphology and Taxonomic rules for plant viruses found within the Tallgrass Prairie Preserve (Fauquet et al 2005).

Family	Genus	Morphology	Genomic Molecule	Genome Size	Species Demarcation
Bromoviridae	*Bromovirus*	Isometric T=3	(+)ssRNA	4 Molecules RNA 1 - ~3.4kb RNA 2 - ~3.1kb RNA 3 - ~2.2KB RNA 4 - ~1kb	-Host Range -Serological Relationships -Compatible Replicase Proteins -Nucelotide Sequence Similarity of 50-80%
Bromoviridae	*Cucumovirus*	Isometric T=3	(+)ssRNA	4 Molecules RNA 1 - ~3.4kb RNA 2 - ~3.1kb RNA 3 - ~2.2KB RNA 4 - ~1kb	-Host Range -Serological Relationships -Compatible Replicase Proteins - At least 65% Nucleotide Sequence Similarity
Bromoviridae	*Ilarvirus*	Bacilliform	(+)ssRNA	4 Molecules RNA 1 - ~3.4kb RNA 2 - ~3.1kb RNA 3 - ~2.2KB RNA 4 - ~1kb	-Host Range -Serological Relationships - Undefined Nucleotide Sequence Similarity
Caulimoviridae	*Badnavirus*	Bacilliform	dsDNA	~7.5kb	-Host Range -Polymerase Nucleotide Sequence Identity <80% -Difference in Gene Product Sequence -Vector Specificity
Caulimoviridae	*Caulimovirus*	Isometric T=7	dsDNA	~8kb	-Host Range -Polymerase Nucleotide Sequence Identity <80% -Difference in Gene Product Sequence
Caulimoviridae	*Soymovirus*	Isometric T=7	dsDNA	~8.1kb	-Host Range -Polymerase Nucleotide Sequence Identity <80%

					-Difference in Gene Product Sequence
Chrysoviridae	*Chrysovirus*	Isometric T=1	dsRNA	4 Molecules RNA 1 - ~3.5kb RNA 2 - ~3.2kb RNA 3 - ~2.9KB RNA 4 - ~2.9kb	-Host Range -Serological Relationships -Genome Size -Length of 5' UTR Region
Closteroviridae	*Ampelovirus*	1.4-2.2um Flexible Rod	(+)ssRNA	~17-18kb	-Virion Size -Size of Coat Protien -Serological Relationships -Genome Organization -Amino Acid Sequence of Gene Products <90% Identical -Vector Specificity -Host Range -Cytopathological Features
Closteroviridae	*Closterovirus*	1.2-2.2um Flexible Rod	(+)ssRNA	~15.5-19.3kb	-Virion Size -Size of Coat Protien -Serological Relationships -Genome Organization -Amino Acid Sequence of Gene Products <90% Identical -Vector Specificity -Host Range -Cytopathological Features
Closteroviridae	*Crinivirus*	2 Virions Particle 1 - 650-850nm Particle 2 - 700-900nm	(+)ssRNA	2 Molecules RNA 1 - ~7.1kb RNA 2 - ~8.1kb	-Virion Size -Size of Coat Protien -Serological Relationships -Genome Organization -Amino Acid Sequence of Gene Products <90% Identical -Vector Specificity -Host Range -Cytopathological Features

Comoviridae	*Comovirus*	Icosohedral T=1	(+)ssRNA	2 Molecules RNA 1 - ~5.8kb RNA 2 - ~3.8kb	-Large Coat Protein Amino Acid Sequence <75% Similar -Polymerase Amino Acid Sequence <75% Similar -No Pseudo-recombination Possible Between Components Possible -Serological Relationships
Comoviridae	*Fabavirus*	Icosohedral T=1	(+)ssRNA	2 Molecules RNA 1 - ~5.9kb RNA 2 - ~3.6kb	-Large Coat Protein Amino Acid Sequence <75% Similar -Polymerase Amino Acid Sequence <75% Similar -No Pseudo-recombination Possible Between Components Possible -Serological Relationships
Comoviridae	*Nepovirus*	Icosohedral T=1	(+)ssRNA	2 Molecules RNA 1 - ~7.3kb RNA 2 - ~3.7kb	-Large Coat Protein Amino Acid Sequence <75% Similar -Polymerase Amino Acid Sequence <75% Similar -No Pseudo-recombination Possible Between Components Possible -Serological Relationships -Vector Specificity
Endornaviridae	*Endornavirus*	Unknown	dsRNA	~13.7kb	-Host Range -Nucelotide Sequence Identity 30-75%
Flexiviridae	*Allexivirus*	800 nm Flexible Rod	(+)ssRNA	~9kb	-Less Than 72% Identical Nucleotide or 80% Amino Acid Sequence Between Coat Protein or Polymerase Gene -Serological Relationships
Flexiviridae	*Potexvirus*	470-580 nm Flexible Rod	(+)ssRNA	~6.5kb	-Host Range -Inability to Cross-protect in Infected Plants -Identity Less Than

					72% Nucleotide or 80% Amino Acid Sequence Between Coat Protein or Polymerase Gene -Serological Relationships
Flexiviridae	*Trichovirus*	640-760 nm Flexible Rod	(+)ssRNA	~7.5kb	-Host Range -Vector Specificity -Identity Less Than 72% Nucleotide or 80% Amino Acid Sequence Between Coat Protein or Polymerase Gene -Serological Relationships
Flexiviridae	*Vitivirus*	725-825 nm Flexible Rod	(+)ssRNA	~7.5kb	-Host Range -Vector Specificity -Identity Less Than 72% Nucleotide or 80% Amino Acid Sequence Between Coat Protein or Polymerase Gene -Serological Relationships
Luteoviridae	*Luteovirus*	Icosohedral T=3	(+)ssRNA	~5.5kb	-Serological Relationships -Inability to cross-protect in Infected Plants -Amino Acid Sequence of Gene Products <90% Identical -Vector Specificity -Host Range
Luteoviridae	*Polerovirus*	Icosohedral T=3	(+)ssRNA	~5.8kb	-Serological Relationships -Inability to cross-protect in Infected Plants -Amino Acid Sequence of Gene Products <90% Identical -Vector Specificity -Host Range
Narnaviridae	*Mitovirus*	Unknown	(+)ssRNA	~2.5kb	-Less than 50% Amino Acid Sequence Identity -Ability to Recombine and Remain Viable

Partitiviridae	*Alphacryptovirus*	Isometric	dsRNA	2 Molecules RNA 1 - ~1.9kb RNA 2 - ~1.9kb	-Host Range -Genome Size -Serological Relationships
Partitiviridae	*Partitivirus*	Isometric	dsRNA	2 Molecules RNA 1 - ~1.9kb RNA 2 - ~1.9kb	-Host Range -Genome Size -Serological Relationships
Potyviridae	*Potyvirus*	600-900 nm Flexible Rod	(+)ssRNA	~9.7kb	-Coat Protein Amino Acid Sequence Identity <80% -Nucleotide Sequence <85% Identical Over Whole Genome -Differing Polyprotein Cleavage Sites -Host Range -Method of Transmission -Cytopathology -Serological Protperties
Potyviridae	*Tritimovirus*	725-825 nm Flexible Rod	(+)ssRNA	~9.3kb	-Coat Protein Amino Acid Sequence Identity <80% -Nucleotide Sequence <85% Identical Over Whole Genome -Differing Polyprotein Cleavage Sites -Host Range -Method of Transmission -Cytopathology -Serological Protperties
Reoviridae	*Fijivirus*	Icosohedral	dsRNA	10 Molecules 1.8kb-3.9kb	-Serological Properties -Conserved Terminal Regions of Genomic Segments -Vector Specificity -Host Range -Nucleotide Sequence Identity <74% in the Subcore Structural Protein -Homology of Conserved Genomic Regions >85%

Reoviridae	*Orbivirus*	Icosohedral	dsRNA	10 Molecules 822nt-3.9kb	-Serological Properties -Conserved Terminal Regions of Genomic Segments -Vector Specificity -Host Range -Nucleotide Sequence Identity <74% in the Subcore Structural Protein -Homology of Conserved Genomic Regions >85%
Reoviridae	*Oryzavirus*	Icosohedral	dsRNA	10 Molecules 1.1kb-3.9kb	-Ability to Recombine to Create Viable Progeny -Serological Properties -Conserved Terminal Regions of Genomic Segments -Vector Specificity -Host Range -Nucleotide Sequence Identity
Rhabdoviridae	*Nucleorhabdovirus*	Bacilliform	(-)ssRNA	~14kb	-Vector Specificity -Host Range
Sequiviridae	*Waikavirus*	Icosohedral	(+)ssRNA	~12kb	-<70% Similarity in Amino Acids for Polyprotein and <80% for Proteinase and Polymerase -Serological Relationships -Host Range -Vector Specificity
Tombusviridae	*Carmovirus*	Icosohedral T=3	(+)ssRNA	~4kb	-Serological relationships -<41% Amino Acid Sequence Identity in the Coat Protein -<52% Amino Acid Sequence Identity in the Polymerase -Size of the Coat Protein -Host Range -Fungal Vector
Tombusviridae	*Panicovirus*	Icosohedral T=3	(+)ssRNA	~4.3kb	-Serological relationships -Gene Product Sequence Identity -Host Range -Vector Specificity

Tombusviridae	*Tombusvirus*	Icosohedral T=3	(+)ssRNA	~4.7kb	-Serological relationships -<87% Amino Acid Sequence Identity in the Coat Protein -<96% Amino Acid Sequence Identity in the Polymerase -Size of the Coat Protein -Host Range
Totiviridae	*Totivirus*	Isometric	dsRNA	~5kb	-Host Range -<50% Amino Acid Identity
Tymoviridae	*Maculavirus*	Icosohedral T=3	(+)ssRNA	~7.5kb	-<80% Nucleotide Sequence Identity Overall -<90% Coat Protein Amino Acid Identity -Host Range -Serological Relationships
Tymoviridae	*Marafivirus*	Icosohedral T=3	(+)ssRNA	~6.5kb	-<80% Nucleotide Sequence Identity Overall -<90% Coat Protein Amino Acid Identity -Host Range -3'-Terminal Structure -Serological Relationships
Tymoviridae	*Tymovirus*	Icosohedral T=3	(+)ssRNA	~6.5kb	-<80% Nucleotide Sequence Identity Overall -<90% Coat Protein Amino Acid Identity -Host Range -3'-Terminal Structure -Serological Relationships
unclassified	*Benyvirus*	85-395 nm Rod	(+)ssRNA	4 Molecules RNA 1 - ~6.7kb RNA 2 - ~4.6kb RNA 3 - ~1.7KB RNA 4 - ~1.4kb	-<90% Coat Protein Amino Acid Identity -Serological Relationships
unclassified	*Tobamovirus*	300 nm Rod	(+)ssRNA	~6.5kb	-<90% Nucleotide Sequence Identity Overall -Host Range

					-Serological Relationships
unclassified	*Sobemovirus*	Icosohedral T=3	(+)ssRNA	~4kb	-<60% Nucleotide Sequence Identity Overall -Host Range -Serological Relationships
unclassified	*Hordeivirus*	110 nm Rod	(+)ssRNA	3 Molecules RNA 1 - ~3.8kb RNA 2 - ~3.2kb RNA 3 - ~2.8kb	Not Yet Determined