UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

VIRAL ECOGENOMICS – MOLECULAR ANALYSIS OF A COSTA RICAN PLANT RNA VIRAL COMMUNITY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JIAXI QUAN
Norman, Oklahoma
2009

VIRAL ECOGENOMICS – MOLECULAR ANALYSIS OF A COSTA RICAN PLANT RNA VIRAL COMMUNITY

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF BIOENGINEERING

By

_____
Dr. Matthias U. Nollert, Chair

_____
Dr. Bruce A. Roe

_____
Dr. Jia Li

_____
Dr. Vassilios I. Sikavitsas

_____
Dr. Rongzhu Gan

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Roe for his guidance, support, and patience to me over the past five years. I appreciate the help from Dr. Nollert for all his advice and suggestions. My thanks also go to my committee members, Dr. Li, Dr. Sikavitsas, and Dr. Gan for their support, advice, and participation of my dissertation. I would like to convey special thanks to my former committee member, Dr. Harrison for his advice and prolonged interest in my dissertation.

I am also thankful to all the members of Dr. Roe's laboratory at the Advanced Center for Genome Technology. This huge project could not progress without their participation. I greatly appreciate James White, Hongshing Lai, and Steve Kenton for their assistance in data analysis. I also appreciate Dr. Fares Najar for his edifying discussions and suggestions. I also thank Graham Wiley, Chunmei Qu, Ping Wang, Simone Macmill for their contribution to sequencing and a special thanks go to Dr. Jianfeng Li, Yanbo Xing, and Keqin Wang for their friendship and generous help in my work. I also would like to thank our collaborators, Dr. Roossinck and Dr. Saha from the Samuel Roberts Noble Foundation who supplied the DNA samples that were sequenced and analyzed for this dissertation.

I would like to thank my father and mother for their deep love and encouragement through the years. Special appreciation goes to my husband Dr. Shengguang Qian and my two lovely sons. Their strong support made it possible for me to stick to and achieve my goal.

Contents

List of Tables

List of Figures

List of Abbreviations

ACG -- Area Conservation Guanacast

ATP -- adenosine triphosphate

BLAST -- basic local alignment search ool

BSA -- bovine serum albumin

CCD -- charge-coupled device

CDD -- conserved domain database

cDNA -- complementary DNA

DDBJ -- DNA Data Bank of Japan

DNA -- deoxyribonucleic acid

dNTP -- deoxyribonucleotide triphosphate

EB -- elution buffer

EMBL -- European molecular Biology Laboratory

emPCR -- emulsion-based polymerase chain reaction

MPC -- magnetic particle collector

mRNA -- messenger ribonucleic acid

MSP -- maximal aegment pairs

NCBI -- National Center for Biotechnology Information

nr -- non-redundant protein sequence database at NCBI

nt -- non-redundant nucleotide sequence database at NCBI

PERL -- practical extraction and report language

Phrap -- Phil's read assembly program

Phred -- Phil's read editor

PNK -- polynucleotide kinase

RdRp -- RNA-dependent RNA polymerase

RNA -- ribonucleic acid

RPS-BLAST -- reversed position specific BLAST

SPRI -- solid phase reversible immobilization

SQL -- structural query language

URL -- uniform resource locator

Abstract

RNA virus metagenome data characterization provides an unbiased picture of RNA viruses in natural environment. To test the hypothesis that plant viral infections are relatively specific for the viral-host interaction and greatly influenced by environmental factors, in my dissertation research, **DNA** copies of reverse transcribed and amplified virus genomes collected from the Area Conservation Guanacast (ACG) region in Costa Rico that were tagged with one of 24 or 96 short oligonucleotides, were sequenced using a state-of-the-art massively parallel Roche/454 GS-FLX DNA pyrosequencer. Since at least 75 Mb of raw DNA sequence data was collected from each full sequencing run, and because our tagging strategy allowed for multiple samples to be sequenced simultaneously, a data management system was developed to serve as a platform to facilitate efficient and comprehensive analysis of metagenome data collected from RNA virus communities.

After assembly with the Roche/454 Newbler assembler, analysis using BLASTN, BLASTX, and tBLASTX identified 2017 contigs that belong to known or novel virus genomes and 20% of contigs without any significant similarities with current database. A statistics analysis revealed a total of 26 virus families in the ACG region had Partitiviridae as the most abundant plant viral family observed. The most widespread species observed was Zucchini mosaic-like viruses, providing more evidence to the world

wide spread of this virus species. The plant host infected with the most number of virus families in the ACG was *Alibertia edulis* Rubiaceae.

Further analysis revealed the effect of symptom, age, and season on the virus particle concentration on the plant host with symptomatic and old samples having a statistically higher number of virus particles when compared to asymptomatic and young plant samples. Plants collected during the Costa Rican rainy season had a higher concentration of viral infection than did those collected during the dry season. The two largest sequenced contigs, both of which are over 11,000 nucleotides, were characterized and predicted to be the putative novel members of the endornavirus family.

Overall these studies confirmed my original hypothesis that viruses in ACG are diverse, with large numbers of new, previously unknown plant viruses present in the environmentally diverse ACG. These studies also confirmed that widespread and multiple viral infections are common phenomena among these viruses and that plant age and the relative amount of available moisture can be directly correlated with higher viral titers on plant hosts.

Chapter 1 Introduction

1.1 Virus and Taxonomy

1.1.1 Brief Introduction to Viruses

Viruses, the smallest living organisms on earth, are biological entities that infect all forms of life from animals, plants to bacteria. Once a virus infects a host cell, it uses the enzymatic machinery of the host to reproduce. An intact virus particle, called a virion, consists of the nucleic acid and protective coat proteins called the capsid (Fields, B.N. 1996). In some of the more complex virions, their capsid is surrounded by a lipid bilayer and glycoprotein-containing envelope that is derived from the membrane of its host cell. Most viruses have a capsid diameter between 10 and 300 nanometres and are too small to be detected with a light microscope (Petrov, A.S. 2008). An enormous variety of genomic structures can be seen among viral species since they can have either a DNA genome or a RNA genome, either of which can be single or double strand (Table 1.1).

Viruses have been described as "organisms at the edge of life", as they cannot reproduce on their own, but possess a genome that evolves in infected cells by natural selection. Since viruses do not have the basic unit of life, a cell structure, and they do not possess the enzymes for basic metabolism, they are considered parasites that require host cells to replicate and synthesize the compounds needed to sustain themselves (Holmes, E.C. 2007).

Property	Type
Nucleic Acid	DNA RNA
Shape	Linear Circular Segmented
Strand	Single-stranded (ss) Double-stranded (ds)
Strand Sense	Positive sense (+) Negative sense (-) Antisense (+/-)

Table 1.1 Genomic Diversity of Viruses

The life cycle of virus varies greatly with species, but five basic stages can be found for all viruses: entry, uncoating, replication, assembly, and release. After viruses enter host cell, the coat protein then is degraded by viral or host enzymes to release the viral genomic materials. After sufficient quantities of protein and genomic materials are produced through gene translation and genome replication, viruses are assembled. Post-translational modification generally occurs after assembly and then viruses are released from the host cell (Dimmock, N.J. *et al*. 2007). Viral infection leads to many diseases in eukaryotic organism including humans, animals and plants. Some viral infections are contagious and lethal. For example, the diseases in plants caused by viruses leads to an estimated $60 billion per year economic influence on crops worldwide (Pogue, G.P. *et al* 2002).

Viruses are very widespread and exist wherever life can be found. Due to their infectivity in all life domains, they play a major role in maintaining host population balance, sustainability of both domestic and wildland plants and animals, and in globally important ecosystem cycles such as the nutrient cycle of the seas (Villarreal, L.P. 2005; Villarreal, L.P. & DeFilippis V.R. 2000). Studies also suggest that viruses are at the root of the evolution of life on earth (Prangishvili D 2003). Viruses played a critical role in the evolution of DNA, DNA replication mechanisms, the separation of the three domains of life, and the origin of the eukaryotic nucleus. (Whitfield, J. 2006; Forterre, P. 2006).

1.1.2 Virus Taxonomy

The International Committee on Taxonomy of Viruses or ICTV classification system has been used in conjunction with the Baltimore classification system in modern virus classification (van Regenmortel, M.H. 2004**;** Mayo, M.A. 1999). The Baltimore classification of viruses is based on the mechanism of **mRNA** production. This classification places viruses into seven groups based on their genomes: double stranded DNA viruses (dsDNA), single stranded DNA viruses (ssDNA), double stranded RNA viruses (dsRNA), single stranded positive sense RNA viruses ((+)ssRNA), single stranded negative sense RNA viruses ((-)ssRNA), single stranded RNA viruses with DNA intermediate in life-cycle (ssRNA-RT), double stranded DNA viruses with RNA intermediate in life-cycle (dsDNA-RT) (Baltimore,D.1971). This unified taxonomy was adopted by International Committee on Taxonomy of Viruses (ICTV) whose latest report divides more than 6000 viruses into 3 orders, 56 families, 9

subfamilies, and 233 genera (Fauquet, C.M. *et al*. 2005). Three main properties are considered in determining order: the type of nucleic acid genome, whether the nucleic acid is single- or double-stranded, and the presence or absence of an envelope. Following three main properties, the classification is then based on the characters including the type of host, the capsid shape, immunological properties and the type of disease it causes (Adams, M.J. 2005). Since the last decade, with the availability of gene or genome sequences, phylogenetic relationships have been used to derive to postulate taxonomic associations (Calisher, C.H. *et al*.1995). Recently, molecular approaches have been used successfully to demarcate species as definitive or tentative members of particular groups (Ali, A., *et al*. 2006). Pairwise comparisons of whole genome sequences or certain genes such as RNA-dependent RNA polymerase and coat protein have been confirmed to provide a robust method for viral relationships and allows for viral taxonomy that largely supports the family and genus assignments made by the ICTV (Stuart, G.W. 2006). The 8th report of ICTV started to adopt sequence homology as classification criteria for some virus groups.

1.1.3 Plant Viruses

The discovery of plant viruses causing disease is accredited to Martinus Beijerinck. , In 1898, he determined that plant sap obtained from tobacco leaves with the "mosaic disease" remained infectious after passing through a porcelain filter (Lerner, K.L, 2002). To date, 733 species of viruses that infect plants are recognized by the ICTV and a number of additional species that have been reported but have not yet undergone the ICTV approval process. The vast majority (90%) of reported plant

viruses have RNA as their genetic material. Viral RNA genomes included four types: dsRNA virus, ss(+)RNA virus, ss(-)RNA virus, or antisense RNA virus (a mixture of both positive sense and negative sense) (Roossinck, M. J.2003). dsRNA viruses contain one or several different RNA molecules, each of which encodes for one or more viral proteins. A ss(+)RNA virus is infectious itself since it can function as an mRNA which codes for the production of viral proteins. For ss(-)RNA virus, the RNA-dependent RNA polymerase is carried in the nucleocapsid. Once entering the cell, positive strand is made with RNA polymerase. The positive-sense RNA molecule then acts as viral mRNA to be translated into proteins.

Over 50% of known plant viruses are rod shaped (flexuous or rigid). The length of the particle is usually between 300–500 nm with a diameter of 15–20 nm. The second most common structure among plant viruses are isometric particles with a diameter of 40–50 nm (Collier, L. *et al* 1998). Host ranges for individual plant viruses vary greatly with the range from a single plant species (e.g. each member of family *Partitiviridae*) to over 1000 species of plants (e.g. *Cucumber mosaic virus*; Palukaitis, P. 2003).

Compared to viruses that infect humans or animals, there are two major differences that are in life cycle and transmission approach respectively. First, due to the robust cell wall, plant viruses can not enter host cell through attachment/penetration aided with surface receptor as animal viruses. At release stage of life cycle, progeny plant viruses transport out of host cell through plasmodesmata with the aid of movement protein instead of lysing or budding through plasma membrane of host cell as animal viruses do (Lazarowitz, S.G. 2001). Second, the immobility of plant hosts determines that the dorminant transmission method for plant viruses is vectors that

carry viruses onto new hosts instead of aerosol or ingestion for animal viruses (Table 1.2). Plant viruses are typically transmitted through sap, insects, nematodes, plasmodiophorids, and seeds (Zaitlin, M. 2000). The symptom caused by viral infection often is restricted to number of cells near the site of entry as a result of host response. Such an infection usually leads to visible symptoms (e.g. spots) on the inoculated leaves. The spots (local lesions) are generally of two types: chlorotic, as a result of loss of chlorophyll in the infected cells, or necrotic, due to death of infected host cells (Khan, J.A. 2006).

All plant viruses encode replicases and coat proteins, and most encode one or more proteins that facilitate virus movement from cell to cell and in long distance in the plant host. Some viruses produce proteases that cleave the polyprotein products of genome translation (e.g. potyvirus).

Transmission	**Aerosols or Ingestion**	**Fluids**	**Pare**nt **to Off**SPRI**ng**	**Vectors**
Animal Viruses	Most e.g. *Picorna* *Orthomyxo* *Corona* *Reo*	Few e.g. *Hepandna* *Retro* *Herpes* *Papillorna*	Few e.g. *Retro* *Herpes* *Arena*	Many e.g. *Toga* *Flavi* *Bunya* *Rhabdo*
Plant Viruses	None	Few e.g. *Tobamo* *Tombus*	Many e.g. *Hordei* *Lar* *Poty* [Seeds, Pollen, Bulbs, Grafting]	Most e.g. *Poty* *Potex* *Gemini* *Luteo* [Insects, Fungi, Nematodes]

Table 1.2 Difference of transmission methods between animal viruses and plant viruses (Lazarowitz, S.G. 2001)

The genome sizes of RNA plant viruses ranges from about 4kb to about 20kb with the most in the range of 5-10kb. Typically, RNA genomes have smaller genomes than DNA viruses. Apart from the lower stability of RNA, the other important reason is that RNA replication is not proof-read by RNA polymerase while DNA replication is very accurate since DNA polymerase can check the copied sequence and replace any mismatches with the correct ones. The mutation rate of RNA genomes is very high (one error in every 1000-10000 bases) so it is unlikely that any copy of a viral RNA genome is exactly the same as the template from which it is copied (Domingo, E. 1997). The presence of segmented genome in many RNA viruses also reflects the low fidelity of RNA replication.

Thus, RNA virus have high mutation rates, high yields, and short replication time cause the products of RNA viruses replication as complex and dynamic mutant swarms, called viral quasispecies. Generally, around 0.1% of the RNA virus genomic bases, distributed at different locations, are mutated with substitution, insertion or deletion. This quasispecies phenomenon gives virus advantage of survival in multiple environments (Schnerder, W.L. 2001; Noueiry, A.O. 2003) and as a result, very few viral pathogens can be effectively controlled by either vaccination or antiviral therapies. Within each viral generation, the variant genomes of a viral population compete and those variants that best adapt to each particular environment survive (Domingo, E. 1994).

1.2 Genomics

1.2.1 DNA and RNA

DNA is the carrier of genetic information for all cellular organisms as well as for many viruses on earth. The path to the discovery and the elucidation of the biological role of DNA occurred over a period of 75 years. DNA was first isolated by Friedrich Miescher in 1869, who discovered a microscopic substance in the pus of discarded surgical bandages. As it resided in the nuclei of cells, he called it "nuclein" (Dahm, R. 2005). In 1889, Richard Altmann first isolated protein-free nuclein and named it nucleic acid. In 1928, Frederick Griffith first demonstrated that DNA was the genetic material (Griffith, F. 1928). The genetic role of DNA was further supported by the first successful transformation experiment in 1944 (Avery, O.T. *et al.* 1944) and the experiment on radioactive P^{32} -labeled and S^{35}-labeled bacteriophage in 1952 to confirm that DNA rather than protein was the genetic material (Hershey, A.D. *et al*. 1952).

The contribution of James Watson and Francis Crick to the discovery of DNA structure is the most significant landmark in the history of genetics (Watson, J.F and Crick, F. 1953). Based on X-ray diffraction images taken by Rosalind Franklin (Watson, J.D.1953) they revealed the DNA structure as consisting of two long polymers of nucleotide monomer units with backbone made of phosphate groups and 2'-deoxyribose sugars joined by ester bonds. The two antiparallel polymer strands intertwine in double helical fashion. Each sugar is attached by each of the four different types of nitrogenous bases: a purine, adenine (A), guanine (G), and a pyrimidine,

thymine (T), cytosine (C). The bases stack along the strands and base-pair with bases from the opposite stand.

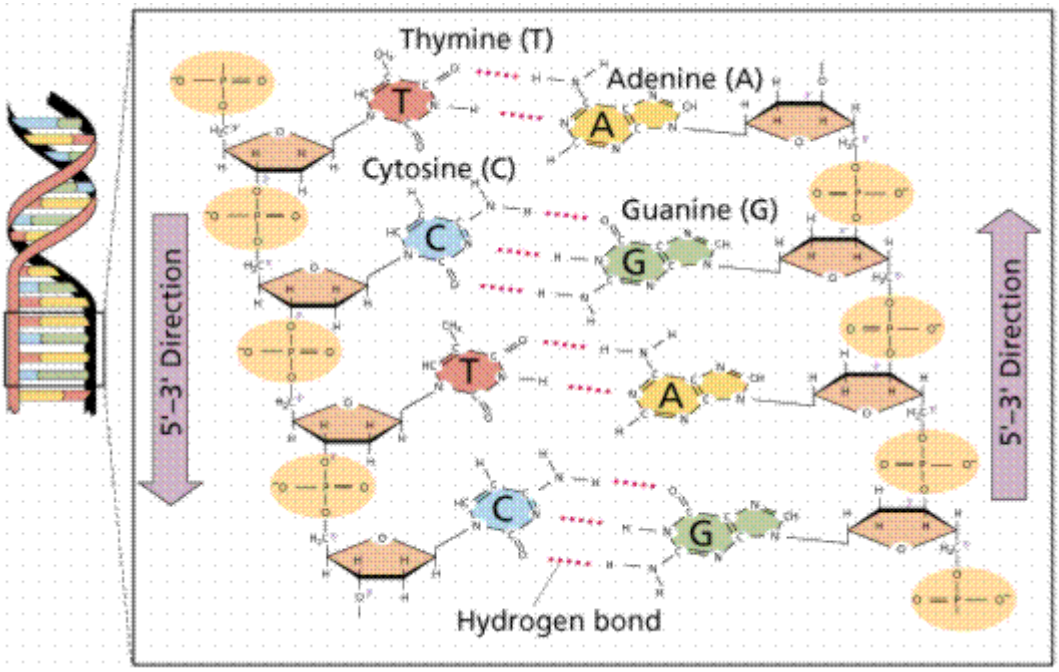

Figure 1.1 The structure of DNA (Purves, W.K. 1998)

RNA is very similar to DNA in structure except for several important differences. RNA usually is single stranded and has a much shorter chain than DNA. RNA nucleotides contain ribose instead of 2'-deoxyribose in DNA, with the presence of two hydroxyl group on adjacent ribose carbons, RNA is more prone to hydrolysis and therefore less stable than DNA. Finally, RNA has the nucleotide uracil (U) rather than thymine (T) which is present in DNA (Barciszewski, J., *et al*. 1999). Due to these structural differences, RNA rarely serves as genetic material since inherent stability of genome is the priority of survival for most organism. The only exception is RNA

viruses that use single stranded or double stranded RNA as their genomes.

RNAs are extensively base paired to form short double stranded helices that leads to many biological functions. The role of RNA in protein synthesis was discovered by Severo Ochoa in 1959 (Ochoa, S. 1959). There are three major classes of RNA participating in protein synthesis, ribosomal RNA (rRNA), transfer RNA (tRNA), and messenger RNA (mRNA).

1.2.2 Gene and Protein

"A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products" (Gerstein, M.B. *et al.* 2007). Proteins are large organic compounds made of amino acids arranged in a linear chain and joined together by peptide bonds between adjacent amino acid residues. The sequence of amino acids in a protein is defined by a gene and encoded in the genetic code (Anthea, M. 1993). For most cases, RNA is an intermediate product in the process of manufacturing proteins from genes. mRNA carrying the information transcribed from DNA binds to ribosomes which read mRNAs and translate the information they carry into proteins (Szathmary, E. 1999). For RNA viruses, they use RNA to store genes and no DNA is involved in protein synthesis. Other viruses such as RNA retroviruses require the reverse transcription of their genome from RNA into DNA before their proteins can be synthesized.

In 1970, Crick proposed the “Central Dogma of Molecular Biology”, a framework for understanding the transfer of genetic information among DNA, RNA and protein

(Crick, F. 1970). The dogma classes these transfers into 3 groups: 3 general transfers (occur normally in most cells), 3 special transfers (occur under specific conditions), and 3 unknown transfers (believed to never occur). The general transfers describe the normal flow of biological information: DNA can be copied to DNA (DNA replication), DNA information can be copied into mRNA (transcription), and proteins can be synthesized using the information in mRNA as a template (translation). The special transfers include reverse transcription referring to the transfer of information from RNA to DNA which occurs in retroviruses, RNA replication referring to the copy of one RNA to another that occurs in RNA viruses, and direct translation from DNA to protein that has been demonstrated in a cell-free system instead of real life (McCarthy, B.J. 1965).

1.2.3 Genome and Genomics

The genome of an organism is its whole hereditary information encoded in the DNA or RNA. The sequencing and study of the global properties of genomes of related organisms is referred to as genomics, which is different from genetics which studies the properties of single genes or groups of genes. Genomics was established by Fred Sanger when he first sequenced the complete genomes of virus bacteriophage Φ-X174 (5,386 bp Sanger, F, 1977). His group established techniques of sequencing, genome mapping, data storage, and bioinformatic analyses in the 1970-1980s. Since then, large scale genome sequencing and annotation proliferate with the development and improvement of Sanger's method. As of October 2008, the records of genomic sequences stored in NCBI have been increased to about 112 Archaea, 2663 viruses,

2236 bacteria, and 2637 eukaryota. The goal of genomics is to promote the understanding of the structure, function, and evolution of genomes in all forms of life and the application of genome science and technologies to challenging problems in biology, medicine and other fields.

Organism	Genome Size	Description
Virus, Bacteriophage MS2	3,569	First sequenced RNA genome (Fiers, W., *et al.*1976)
Virus, Phage Ø–X174	5,386	First sequenced DNA genome (Sanger, F., *et al.* 1977)
Bacterium, *Carsonella ruddii*	160,000	Smallest non-vial genome (Nakabachi A, *et al.* 2006)
Bacterium, *Haemophilus influenzae*	1,830,000	First genome of a living organism (Fleischmann R., *et al.* 1995)
Plant, *Arabidopsis thaliana*	157,000,000	First sequenced plant genome (Greilhuber, J. *et al.* ,2006)
Nematode, *Caenorhabditis elegans*	98,000,000	First sequenced multicellular animal genome (The *C. elegans* Sequencing Consortium 1998)
Mammal, *Homo sapiens*	3,200,000,000	Human genome (International Human Genome Sequencing Consortium 2001)

Table 1.3 Comparative Genome Sizes of Representative Sequenced Organisms

1.3 Metagenomics and Ecogenomics

In contrast to classical genomics that studies complete genome sequence of model organisms, the new fields of metagenomics and ecogenomics have the goal of finding

new organisms that inhabit selected environments. In metagenomics and ecogenomics, microbes are collected but not cultivated, in contrast to the genomics approach that requires prior separation of an organism from its habitat followed by growth in cultures and maintenance in artificial niche in the laboratory.

1.3.1 Overview of Metagenomics

Metagenomics, also called community genomics or environmental genomics, refers to the application of the methods of genomics to environmental genomic sample assemblages. The microorganisms and viruses are harvested from the environment and their genomic materials are purified, sequenced and processed to create a community DNA/**cDNA** library followed by large-scale characterization of the microbial communities in diverse habitats (Jurkowski, A., *et al*. 2007). Four features make metagenomics an appealing approach to study microgial communities. First, a metagenomics approach generates a wealth of data that can greatly expand our current knowledge because the available nucleotide sequences presently are derived from very limited model organisms that are either important in pathology or easy to culture. This metagenomic data then can serve as a platform for basic scientific study as well as for more direct applications in many disciplines such as biotechnology and medicine (Schmeisser, C. *et al*. 2003). Second, metagenomic studies provides an relatively unbiased insight into the diversity, ecology and evolution of microbial or viral communities that is contrast to the sequences in the current major databases that represent a biased view of living organisms on earth. Third, with computational analysis on massively collected genomic information, gene profiles can be created to

further facilitate the discovery of novel useful genes in high throughput. Fourth, the comparative metagenomic also can be useful to identify a unique set of functions associated with each metagenome community.

The recent advances of high-throughput and low cost sequencing technology have paved the way for metagenomics. The initial metagenomic sequencing studies the diversity of a specific microorganism group in natural environment or a specific gene profile for a community. The first approach of metagenomics was carried out by Handelsman et al. in 1998 (Handelsman, J. *et al.* 1998) who sequenced metagenome from uncultured soil microorganism with Sanger's method and explored biosynthetic machinery of soil. This study showed that the diversity of uncultured microorganism surprisingly was high. In 2002, Beja et al. analyzed the photosynthetic gene content and operon organization in naturally occurring marine bacteria to demonstrate that planktonic bacterial assemblages contain multiple, distantly related, photosynthetically active bacterial groups, including some unrelated to known and cultivated types (Beja, O. *et al*. 2002). Since 2004, a series of large-scale random sequencing metagenomics studies have been performed from a wide range of environments that include for example: soil (Kim, K.H. 2008), ocean, hot springs (Schoenfeld, T. 2008), specific habitat (Tringe, S.G., 2008), the human gut (Gill, S.R. 2006), and animal faeces (Cann, A.J. 2005). Specifically, Tyson et al. sequenced an underground biofilm from an extremely acidic environment and generated 124 Mbp of data (Tyson, G.W. 2004). Their study revealed the pathways for carbon and nitrogen fixation and energy generation, and provided insights into survival strategies of microorganisms in an extreme environment. The shotgun sequencing of the microbial population in Sargasso

Sea generated 1,687 Mbp, that were estimated to be derived from at least 1,800 genomic species. In this study, approximately 1.2 million previously unknown genes were identified, that included more than 782 new rhodopsin-like photoreceptors (Venter, J.C. *et al*. 2004).

The first viral metagenomic study was published in 2002 (Breitbart, M. *et al* 2002) where two uncultured marine viral communities were sequenced. The results showed that most of the diversity was previously uncharacterized and identifying and measuring the community dynamics of viruses in the environment was complicated compared to other microorganism because there is no single gene (e.g. 16S rRNA for most microorganisms) that is conserved to all viral genomes. However, some genes like RNA-dependent RNA polymerase is relatively conserved within a particular virus group so that it can be used to infer the phylogenetic relationship between virus species.

1.3.2 Ecogenomics

The ecogenomics concept, initially introduced by Dr. Marilyn Roossinck at the Noble Foundation in Ardmore, Oklahoma, refers to the application of methods for genomics on individual samples collected from a territory to address ecological questions (personal communication). Although the pathogenicity of plant viruses has been extensively studied in agriculture sytem, the study on the ecology of plant viruses in natual ecosystem has received much less attention (Wren JD, 2006). Since many plant viruses cause plant diseases, plant viruses were traditionally viewed as parasites and studied as pathogens. However, recent studies showed that plant viruses could be beneficial to their hosts by improving drought resistance (Xu, P 2008) and thermal

tolerance (Marquez LM 2007). In contrast to metagenomics that primarily focuses on prokaryotic hosts by isolating samples from one or several pools, ecogenomics collects samples from individual eukaryotic hosts. Plant virus ecogenomics can extend and deepen current knowledge about plant viruses through addressing the questions of plant diversity, distribution, and plant-host interactions in a natural ecosystem.

The Area Conservation Guanacast (ACG), located in northwestern Costa Rica, a country well known for its ecological diversities, was an ideal site for plant virus ecogenomics study because it contains a diverse, species rich habitat and includes both conserved wild lands and lands currently or formerly used for agriculture. The area includes three major terrestrial tropical ecosystems: dry forest, cloud forest and rain forest. These are resolved into 22 large scale habitat type as well as 20-30 additional microhabitats. A rough estimation of the species diversity in ACG is 235,000 (not including viruses) that is more than the species diversity in the entire 48 continental United States. There are over 7,000 distinct plant species in the ACG, including many species that are native to the region. (Janzen, D. 1999). A well-established infrastructure and an advanced inventory of its species in ACG make rapid identification possible for the collection of large quantities of plant species.

Compared with other viral metagenomics studies, the uniqueness of the present RNA virus study lies in its well-recorded host information that gives a glimpse of host-virus interaction. The currently available viral metagenomic sampling is mostly from prokaryotic hosts and lacks host information so that the interaction between hosts and viruses can not be characterized. In contrast, the ACG sampling can provide detailed information (including collection season and location) of the selected plant

species from which the virus genomes were isolated as well as the was recorded. This provides the opportunity to analyze several widespread phenomena among viruses such as multiple infection (one virus infects multiple plant species) and specificity (one virus species specifically infect one plant species).

In collaboration with botanists in the ACG, samples from seven plant groups, are identified and collected in the area. These groups include Fabaceae (beans); Cucurbitaceae (melons, cucumber, squash); Solanaceae (tomato, eggplant, potato); Poaceae (rice, corn); Rubaceae (coffee); Rutaceae (citrus) and Bignoniaceae (non crop-related plants). In addition, the plants were sampled at four different time points throughout the year: dry season, early rainy season, rainy season, and early dry season. The collection is not biased toward symptomatic plants but any symptoms are noted. These families do not make extensive secondary metabolites that can interfere with dsRNA isolation. The fresh plant tissue is collected from individual plants and then dsRNA was extracted. The dsRNA, the hallmark of RNA viruses could be either the genome of dsRNA or the replicative form of the genome of ss(+)RNA or ss(-)RNA. dsRNA is quite stable and can generally be isolated with fewer precautions than required for other cellular nucleic acids (Dodds J.A. *et al*., 1984). The purification of dsRNA involves the disruption of the plant tissue, phenol extraction to remove sub-cellular fractions, ethanol precipitation to purify nucleic acids, and cellulose elution/centrifugation to further purify dsRNA from DNA and ssRNA (Sambrook, J. 1989) and the samples that contain dsRNA were quantitated through electrophoresis on 1.5% agarose gels followed by visualization under UV illumination.

1.4 DNA Sequencing Technology

The classical sequencing method, Sanger method that initially was developed in 1977, is widely used in both small- and large-scale DNA sequencing, including genome sequencing (Sanger, F., *et al.* 1977). However, with the need of more DNA sequencing projects, cost-saving and higher efficiency are greatly in demand, more sequencing approaches are developed. These approaches can be grouped into four categories (Hall, N. 2007): mass spectrometry (Jurinke, C. *et al*. 2002), *in vitro* cloning and sequencing by synthesis (454 and Solexa (Margulies, M. *et al.* 2005; Bennett, S. T. *et al*. 2005)), *in vitro* cloning followed by hybridization and ligation (Polony and massively parallel signature sequencing (Shendure, J. *et al.* 2005; Brenner, S. *et al*. 2000)), and single molecule (Arrayed fragments and Nanopore readers (Braslavsky, I. *et al* 2003 and Kasianowicz, J. J. *et al.* 1996)) methods. Among these approaches, sequencing by synthesis nanotechnology is the most successful one as it has been widely applied to both confirmatory sequencing and de novo sequencing (Margulies *et al.* 2005). Compared to the Sanger sequencing method that has served as the cornerstone for genome sequencing for over a decade, pyrosequencing is more efficient, less expensive and labor saving. Pyrosequencing adopts a fundamentally different methodology from Sanger method: instead of using fluorescently labeled ddNTP to terminate DNA extension and then capillary gel electrophoresis with LASER detection of nucleotide fragments or nucleotide fragments sets, pyrosequencing uses four-enzymatic cycles for detection of single base during chain elongation. When a nucleotide is incorporated by DNA polymerase, pyrophosphate (PPi) is released which is subsequently converted to **ATP** with ATP sulfurylase. ATP provides energy for firefly luciferase to oxidize

luciferin and generate light (Ronaghi, M. 1998). Thus the incorporated nucleotide can be detected and determined in real-time.

The 454 Inc. integrated pyrosequencing with two other techniques to greatly improve DNA sequencing throughput. One is the emPCR method to amplify many DNA templates in parallel. In emPCR, millions of water-in-oil micelles are formed in a single tube, such that each micelle contains a single molecule of DNA template immobilized on a magnetic bead and the PCR reagents that allow the amplification reaction to be performed. Thus, millions of individual PCR reactions occur simultaneously in one tube as illustrated in Fig. 2 (Dressman, D., 2003). The emPCR method also eliminates the need for cloning in the Sanger method, saving labor and its associated costs, as well as remove the potential for both aberrant recombinants in the surrogate host and cloning-related artifacts such as counterselection against potentially toxic genes.

The other technique is the PicoTiter plate that allows massively parallel sequencing reactions and contains 1.6 million open wells. The size of each well is 44 μm in diameter and about 55 μm in deep guarantees that only one bead with millions of copies of DNA segments can be loaded into each well. Smaller beads carrying immobilized enzymes required for pyrophosphate sequencing also are deposited into each well. During a sequencing run, nucleotides flow sequentially in a fixed order across the plate so it eliminate miscall which sometime occurs in Sanger method. Hundreds of thousands of beads are sequenced in parallel (Leamon, J.H. *et al*. 2003). The addition of apyrase, a nucleotide-degrading enzyme which can efficiently degrades the unincorporated nucleoside into monophosphate, allows nucleotides to be added

sequentially without any intermediate washing step (Ronaghi, M. 2001). The simultaneous sequencing reactions in one plate can produce more than 60 million bases in a 4.5-hour run.

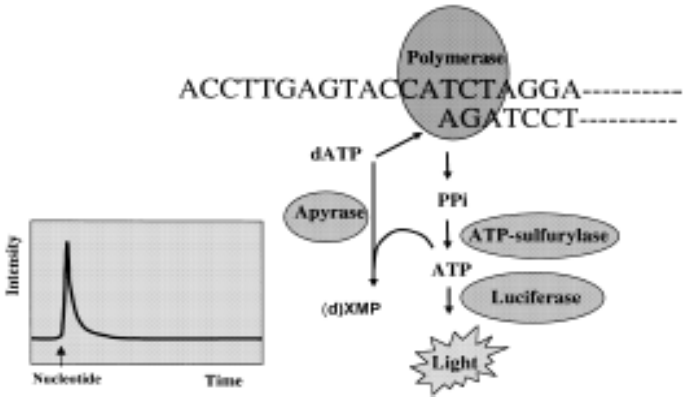

Figure 1.2 The general principle behind pyrosequencing reaction system. PPi is converted to ATP with ATP sulfurylase. Light is produced in the last reaction during which a luciferin molecule is oxidized with luciferase. Apyrase is used to remove remaining nucleotides in the wells before the next nucleotide cycle being introduced (Ronaghi, M. 1998).

Figure 1.3 Diagram of emPCR. (Margulies, M. 2005) Genomic DNA is isolated, fragmented, ligated to adapters and separated into single strands. Fragments are bound to beads under conditions that favor one fragment per bead. After PCR, millions copies of DNA are generated on each single bead.

Apyrase has high catalytic activity and low amounts of this enzyme in the pyrosequencing reaction system efficiently degrade the unincorporated nucleoside triphophates to nucleoside diphosphates and subsequently to nucleoside monophosphate (Ronaghi, M. 2001).

The light generated in the pyrosequencing reactions was detected by the **CCD** camera. Raw signals were background-subtracted, normalized and corrected. Then they underwent image and signal processing and then base calling. The output includes normalized signals across the wells, flowgrams, and base-called sequences. The images are processed to yield sequence information simultaneously across the wells containing template-carrying beads. The raw signals are background-subtracted, normalized and corrected to produce flowgram. The normalized signal intensity at each nucleotide flow for a particular well indicates the number of nucleotides. This linear relationship between intensity and the number of homo-nucleotides can be preserved until eight nucleotides over which errors will occur. In base calling, a Phred-like quality score is calculated (a probability that a measured signal corresponds to an ideal model signal, converted by the instrument software into a Phred-like quality score) to improve the usability of the reads.

After high quality sequences of around ten folds over sampling were obtained (in order to achieve a consensus accuracy of 95%), they were assembled with the software named Newbler Assembler (Figure 1.4). The software consists of three modules. The first, Overlapper, aligns the reads to find and create overlaps using the signal strengths at each nucleotide flow. The second, Unitigger, constructs larger contigs of overlapping sequence reads. The third, Multialigner, generates consensus calls and quality scores

for the bases within each contig based on signal averaging. All aligned flowgram signals at each position then use averaged base call on the averaged signal. The signal averaging allows higher quality consensus base calls (Margulies *et al* 2005).

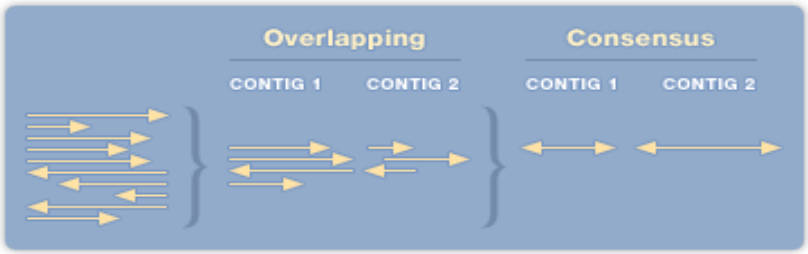

Figure 1.4 Diagram of sequence assembly (Margulies, M. 2005)

1.4.1 Strategies of Sequencing of ACG Samples

The purified dsRNA was converted to cDNA libraries using a modification of the standard random hexamer cDNA technique (Dunn, J.J. 1995) illustrated in Figure 1.5 in Dr. Marilyn Roossinck's laboratory at the Noble Foundation, Ardmore, OK. Briefly, primer containing random hexamers on their 3' termini were used for reverse transcription of denatured dsRNA. A total of 96 sets of primers, each with a different 4 nucleotide tags in addition to the random hexamer sequences then were used for an additional round of PCR. The resulting tagged c**DNA**s were pooled in groups of 96 such that the tags were correlates with an individual viral sample.

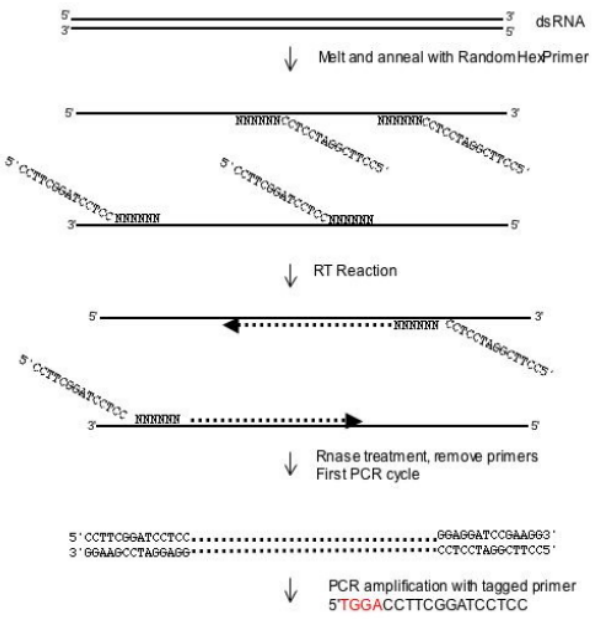

Figure 1.5 Strategy for preparing cDNA ready for 454

1.5 Bioinformatics Tools

As the advances of molecular biology and genomic technologies over the past few decades, the amount of biological information is growing almost exponentially. This requires computer-based databases to store, organize and manipulate data as well as specialized tools to interpret and analyze data. Recently, the term *in silico* has become a common phrase that along with the traditional terms *in vivo* and *in vitro*, now describes bioinformatics, or biological studies carried out on a computer.

1.5.1 Databases

The databases developed by NCBI are the most authoritative and comprehensive

public resources for molecular biology and biotechnology. The NCBI database has been produced in collaboration with another two major databases (**EMBL** and **DDBJ**) and all new and updated database entries are exchanged between the three groups on a daily basis. The databases involved in this viral metagenomic study are NCBI nucleotide database, protein database, viral reference genome database, taxonomy database and CDD.

NCBI nucleotide database is a collection of sequences from several sources where the bulk of the input data is directly submitted by the individuals or groups who determined the sequence. The NCBI protein database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL, and DDBJ as well as protein sequences submitted to other major protein databases. Both nucleotide and protein databases are cross-linked to the taxonomy database. In the taxonomy database, each organism or taxonomy node has its own unique identification number, a tax_id, which is used to build a taxonomical hierarchy. The names and classifications of viruses in the NCBI taxonomy database follows the latest report from the ICTV and attempts to stay current by also accepting new names and classification schemes on a case-by-case basis (Mayo, M.A. 2002). Viral reference genome database collects only complete or nearly complete viral genomic sequences. All collected sequences well represent the genome variability found in many viruses and generally only one sequence is selected among various strains and isolates to greatly reduce redundancy (Bao, Y. *et al.* 2004).

The CDD (Marchler-Bauer, A. 2005; Marchler_Bauer, A. 2007), that also is part of NCBI's Entrez database system, serves as a primary resource for the annotation of

conserved domain footprints on protein sequences in Entrez. The current CDD contains the domain models curated at NCBI as well as imported models from other databases including SMART (Letunic, I. 2006), Pfam (Bateman, A. 2004) and COGs (Tatusov, R.V. 2003). Conserved domains are defined as recurring sequence patterns or motifs and the CDD organizes related domain models in a hierarchical fashion as well as provides a search tool, RPS-BLAST. The *in silico* annotation for protein function generally is obtained by sequence similarity. Thus, once a close neighbor with known function has been identified, the annotation is copied to the query sequence. However, when protein families are sufficiently diverse and when no close neighbors with known function are available this approach cannot be used, and therefore the CDD provides alternative analysis approaches to find the potential function of unknown sequences.

1.5.2 Alignment Tools

The **BLAST** that is available on the NCBI web site at URL http://www.ncbi.nlm.nih.gov/BLAST/BLAST.cgi, is a widely used database search tool that finds regions of local similarity between query and database sequences. The program implements heuristic search methods based on the Smith-Waterman local alignment algorithm to compare nucleotide or protein sequences and calculates the statistical significance of matches. A scoring matrix that assigns positive similarity scores for identities or conservative replacements, negative scores for mismatches and gaps is used to align two sequence segments. The similarity score for two aligned sequences is the sum of the similarity values for each pair of aligned residues that then form a continuous or gapped sequence segment of any length. The pair of segments are

extended in both directions in an attempt to find a locally optimal ungapped alignment. The regions, termed MSPs, that are found are those sequences with the highest scoring pair of identical length segments selected from two sequences. The best scores depend on the length of the query sequence, the size of the database, the scoring table used and the "non-randomness" of the residues in the query. All MSPs above a specified score are displayed. BLAST provides an efficient way to do nucleotide and protein sequence database search, gene identification search as well as analysis of multiple regions of similarity in long DNA sequences (Altschul, S.F., 1996). BLASTN (search nucleotide dataset using a nucleotide query) and protein database with BLASTX (search protein database using a six-frame translated nucleotide query; Gish, W. 1993). tBLASTX (search six-frame translated nucleotide database using a six-frame translated nucleotide query; Altschul, S.F. 1997). RPS-BLAST, which stands for Reverse Position-Specific BLAST, is the tool used in conserved domain search. RPS-BLAST finds sequences significantly similar to the query in a database search and uses the resulting alignments to build a Position-Specific Score Matrix (PSSM) for the query and the database is scanned again with pre-calculated PSSMs to pull in more significant hits, and further refine the scoring model (Marchler-Bauer, A. 2004).

Cross_match, a program for rapid protein and nucleic acid sequence comparison and database search based on the Smith-Waterman_Gotoh algorithm (Smith, T.F. 1981; Gotoh, O. 1982) algorithm, is slower but more sensitive than BLAST. Cross_match is very useful for comparing a set of assembled contigs to another. Statistical significance of the hits is evaluated based on the empirical score distribution for the search.

1.5.3 Virus Genome Annotation Tool

FgenesV, *ab initio* viral gene prediction program useful for intronless genes of viruses, is based on pattern recognition of different types of signals and Markov chain models of coding regions. Optimal combination of these features are found by dynamic programming (Mavromatic, K, *et al*. 2007).

GeneMarkS is another *ab initio* gene prediction program implementing an improved version of the gene finding program GeneMark.hmm, heuristic Markov models of coding and non-coding regions and the Gibbs sampling multiple alignment program (Besemer, J. 2001) .

1.5.4 Assembly Tools

Fastaq2phd (personal communication with James D. White) is a Perl programs used to convert fasta sequence file into Phred file which is a text file containing base call and quality information. Phrap (Phil's revised assembly program, www.phrap.org) is a package of programs for assembling shotgun DNA sequence data. Phrap can handle very large datasets and uses a combination of user-supplied and internally computed data quality information to improve accuracy of assembly in the presence of repeats and constructs the contig sequence as a mosaic of the highest quality reads segments rather than a consensus.

1.5.5 Data Management System for Metagenomic Analysis

Metagenome projects do not directly generate genome sequences even for very small genome like RNA virus. Instead, the typical data is thousands of genome fragments with huge size ranges. Unlike classical genome data from isolated organisms, the generation and interpretation of metagenome data is in early stages of development. Metagenome sequence data processing is more challenging compared with cultured genome sequencing data due to the complex nature and inherent incompleteness as well as the lack of methods designed specifically for processing such data.

Although public databases provide series of analysis tools, accessing and comparing the data can be very time consuming. The procedure is more difficult if some of the records in the public databases are incorrectly named, poorly annotated or redundant. A data management system specifically designed for ACG data can provide an analysis platform and will greatly improve the efficiency and quality of our ACG metagenome analysis. To accomplish this a system, that I have termed ACGweb, was set up using Perl, MySQL, HTML and Apache, with the architecture as shown in Figure 1.6.

Perl (http://www.perl.com/), a popular programming language that is widely used in bioinformatics, is advantageous over other languages for solving common bioinformatics tasks. First, Perl can deal with information in ASCII text files or flat files, which are the kinds of files in which most biological data appears. The files from important databases such as Genbank and PDB can be parsed and edited easily with Perl. Second, Perl has powerful string processing function. Since nucleotide and protein sequences are string data, Perl is well-suited to manipulate long DNA and

protein sequences. Third, the convenience of Perl makes it to solve problems in much less lines of code than in popular C or Java languages. Finally, Perl makes it convenient to write a program that controls HTML and MySQL. With Perl programming, a data analysis pipeline can be set up that allows parallel and automatic processing on large amount of data (Tisdall, J. 2001).

MySQL, a relational database management system that runs as a server providing multi-user access to a number of databases, can process data stored in tables that are related based on primary key and super key (www.mysql.com).

HTML (HyperText Markup Language http://www.w3.org/TR/html401/) is the predominant markup language for web pages.

Apache (www.apache.org) is a powerful, flexible, HTTP/1.1 compliant web server software that supports data interaction between browser and the database.

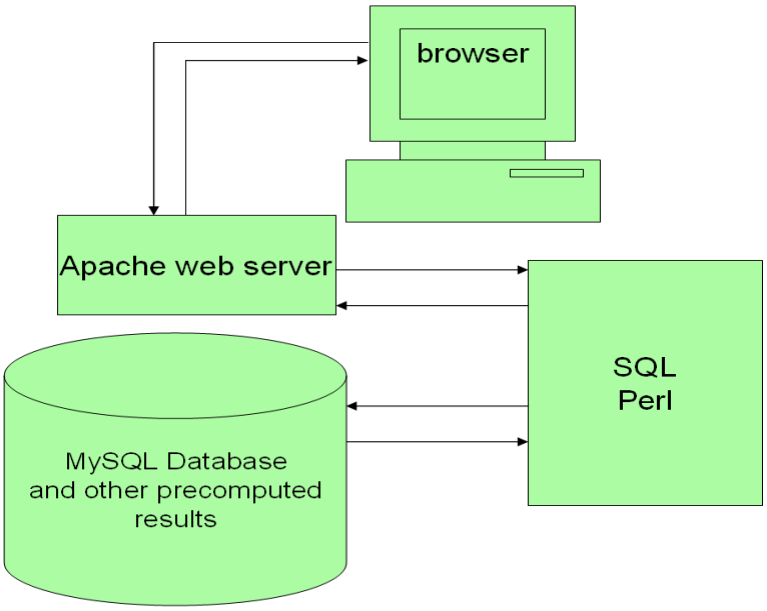

Figure 1.6 Architecture of a dynamic web database

Chapter 2 Materials and Methods

2.1 DNA Sequencing

1μl aliquot of each cDNA library obtained from our collaborators was loaded on BioAnalyzer DNA 7500 Labchip to determine the concentration as well as the average length of cDNA fragments. After adjusting concentration and cDNA size to required range if needed, the cDNA libraries were sent for sequencing in batches. For each batch, cDNA libraries of 96 different samples, each labeled with its unique tag, were pooled into a tube and sequenced with Roche 454 Life Sciences Genome Sequencer FLX System.

2.1.1 Determine Pooling Number

To test the effect of the number of pooled samples on the quality of sequence data, initial four batches pooled with 24 cDNA libraries separately and one batch pooled with 96 cDNA libraries were sequenced and the data was analyzed. Comparison showed that data generated by 96 tag had similar quality with 24 tag including coverage depth, read length, contig length, and singleton percentage. This indicates that increasing the pooling number did not influence the data quality. So 96 tag was used for the rest batches which greatly increased efficiency. For each loading, four batches were loaded onto four separate regions of 454 PicoTiterPlate LR70 Kits.

2.1.2 454 Pyrosequencing

2.1.2.1 DNA library preparation

Once quantitated on the BioAnalyzer DNA 7500 Labchip, the cDNA library (20 µl) was end-repaired to make 3'- and 5'-ends blunt. The end repair reaction that contained 24 µl of purified DNA fragments, 5 µl of 10X polishing buffer, 5 µl **BSA**, 5 µl ATP, 2 µl dNTPs, 5 µl T4 polynucleotide kinase, and 5 µl T4 DNA polymerase was incubated at 25°C for 30 minutes. The 5' – 3' polymerase activity of T4 DNA polymerase fills in 3'-recessed ends of DNA and its single-stranded 3'--5' exonuclease activity removes 3'-overhang ends. The kinase activity of T4 **PNK** adds phosphate groups to the polished 5'-hydroxyl termini. Repaired fragments were mixed and vortexed with 35 µl AMPure SPRI beads and incubated for 5 min at room temperature. The beads were pelleted by MPC and then washed twice with 500 µl 70% ethanol. The dried pellets were mixed with 15 µl EB buffer and the supernatant containing purified DNA was transferred to a fresh tube. Following the DNA library polishing step, 20 µl 2x ligase buffer, 1. µl adapters and 4 µl ligase were added to 15 µl polished DNA to ligate adaptors (A and B) to the both ends of each DNA fragment. Adaptors (A and B) are a pair of double-stranded oligonucleotides that provide priming region for amplification and nucleotide sequencing. They also contain a unique four base nonpalindromic sequence used by assembly software for base calling and to recognize correct reads. For MID approach, MID adaptors was used instead of adaptor A and B. The MIDs include a 10-nucleotide sequence tag on adaptor A which is unique for each MID and allowed for different libraries to be labeled with different MIDs. The sequencing reads was deconvoluted by the analysis software after sequencing run and the reads from

each of the pooled libraries were identified by their unique MID tag and correctly assigned. After incubation for 15 minutes at 25^{o}C, the reaction mix was purified on 28 µl AMPure SPRI beads. In the fill-in reaction that is to repair the gaps between adaptors and DNA fragments, 15 µl ddH_2O, 5 µl 10x Fill-in polymerase buffer, 2 µl dNTP mix, and 3 µl Fill-in polymerase were added to 25 µl purified DNA with ligated adapters followed by incubation at 25^{o}C for 20 minutes. The gap-filled DNA was finally obtained through purification on 35 µl AMPure SPRI beads.

The concentration of DNA library was checked by qPCR. The DNA library samples were diluted to 1:100 and 1:1000, and series of 1:10 dilutions for a standard where the concentration is known were performed. Q-PCR reaction was set up by adding 1 µl sample, 2 µl forward primer, 2 ul reverse primer, 14 µl SYBR1 mix and 10 µl ddH_2O to a total volume of 30 ul. The plate was centrifuged for 5 seconds at 1000 rpm and then was loaded onto Bio-Rad iQ5 multicolor real time PCR detection system. After warming up for 15 minutes, the reaction was incubated at 95^{0}C for 10 minutes followed by 45 cycles of 95^{0}C for 10 seconds, 60^{0}C for 30 seconds and 72^{0}C for 45 seconds. After incubation, the products were quantitated by qPCR and the IQ5 software was used to analyze and check the number of molecules in each samples based on the series of standards.

2.1.2.2 Emulsion PCR (emPCR)

The dsDNA generated as above is flanked with adaptor A at 5'-end and adaptor B at 3'-end. Then, three steps were performed to massively and clonally amplify the

dsDNA libraries. In cDNA library capture step, the DNA capture beads were washed and the isolated ssDNA libraries were added to DNA capture beads for immobilization. After cDNA libraries were annealed to beads, the Live Amplification Mix (181.62 µl Amplification Mix, 10 µl MgSO4, 2.08 µl Amplification Primer Mix, 6 µl Platinum HiFi Taq Polymerase and 0.3 µl pyrophosphatase) was added to annealed library beads. In emulsification, the beads mixture was added to emulsion oil followed by shaking at 15round/ser for 5 minutes for emulsification. The products of emulsification are water-in-oil micelles (microreactors) 50 to 100 µm in diameter, each containing Amplification mix and no more than one single annealed cDNA bead. In amplification step, the contents of each emulsion tube were dispensed to eight wells of 96-well plate and placed into thermalcycler (with amplification program consisting of a 94°C hold for four minutes, followed by 40 amplification cycles of alternating 94°C for 30 seconds, 60°C for 30 seconds, 68°C for 90 seconds, 13 hybridization extension cycles of alternating 94°C for 30 seconds, 58°C for 6 minutes, and hold at 10°C). Two types of primers (primer A that anneals to 5'end of the template and primer B that anneals to 3'end) were used for amplification with the amount of primer A largely over primer B (16 B = A). As the PCR reaction progresses, these bead-bound, complementary strands direct the synthesis of sufficient quantities of first-strands, which hybridize to the bead-bound capture primers to provide sufficient PCR templates need for primer elongation. After emPCR, sstDNA template annealed to each bead was amplified with the copy number from 10 to 30×10^6 copies per bead.

2.1.2.3 Beads Recovery, Enrichment, and Sequencing Primer Annealing

In bead recovery, isopropanol was added to each reaction well containing amplified beads to break the emulsion and the emulsion-isopropanol mix for each individual project was transferred into a Corning tube with a syringe. The Corning tube was shaken followed by centrifugation at 3200 rpm at room temperature for 4 minutes. The supernatant containing the emulsification oil mix dissolved in isopropanol was decanted and the pellet that contains recovered amplified beads was washed with isopropanol three times. The amplified beads were washed with 10ml 1X Bead Wash Buffer three times by mixing and centrifugation at the same condition for emulsion breaking step. The amplified beads then were washed twice with 30 ml 1X Enhancing Fluid by mixing and centrifugation at 10,000 rpm for 5 minutes. The supernatant was decanted and the pellet suspended in 100 µl Enhancing Fluid was retained. In bead enrichment step to remove the beads that carried no amplified DNA, 100 µl streptavidin-coated and magnetic Enrichment Beads were added to amplified DNA library beads and rotated on LabQuake tube roller for 2 minutes. The biotinylated strand of DNA beads bound to the Enrichment beads and DNA/Enrichment beads were precipitated by **MPC**. The supernatant contained the beads that were not amplified and was discarded. The bead pellets were washed with 1mL 1X Enhancing Fluid twice. The tube was removed from MPC and the bead pellet was resuspended in Melt Solution and vortex for 5 minutes. Melt Solution (0.125 M NaOH, 0.2 M NaCl) was added to melt the two complement strands of each dsDNA fragment so that the Enrichment beads that attached to biotinynated strand and the DNA library beads that attached to the

complementary strand were separated. Then the tube was put back to MPC to pellet the Enrichment beads. The supernatant that contains enriched DNA beads was transferred to a microfuge tube. The melting step was repeated and the supernatant was pooled. The enriched DNA beads were centrifuged and the supeRNAtant was discarded. After obtaining single-stranded, bead-bound DNA fragments, annealing buffer was added to beads pellet. Sequencing primer was then added and vortexed. Annealing program (65°C for five minutes, decrease by 0.1°C /sec until 50°C, 50°C for one minute, decrease by 0.1°C /sec until 40°C, 40°C for one minute, and finally decrease by 0.1°C /sec until 15°C on the thermalcycler) was run to anneal the primer to the adaptor part of single stranded DNA. The DNA beads were washed with and resuspended in annealing buffer. 5 µl aliquot of beads was transferred to Coulter Counter cuvettes and beads number (number of beads/µl) were counted with Coulter Counter.

2.1.2.4 Sequencing on PicoTiterPlate

Prior to performing sequencing run on the sequencer, the PicoTiterPlate (PTP) was prepared by loading the DNA-containing beads, enzyme beads, and packing beads into the picotiter sized wells of the PTP. PicoTiterPlate first was soaked in bead buffer (25 mM tricine, 5 mM magnesium acetate, 1 mM dithiothreotol, 0.4 mg/mL polyvinyl pyrrolidone, 0.01% Tween20, 0.1% BSA, pH 7.8) in addition of 8.5 U/mL Apyrase. The gasket was applied to the PTP that then was placed into the Bead Deposition Device (BDD), Depending on the library concentration (bead count) determined in the last step, the appropriate amount of DNA library beads were mixed with Control DNA beads, and then incubated in the Bead Incubation Mix (BIM, 25 mM tricine, 5

mM magnesium acetate, 1 mM dithiothreotol, 0.4 mg/mL polyvinyl pyrrolidone, 0.01% Tween20, 0.1% BSA, 7000 units of *Bst* DNA polymerase, pH7.8) on the lab rotator for 30 minutes at room temperature. Packing beads were prepared by washing three times with bead buffer and mixing with BIM. Enzyme beads were pelleted with MPC and resuspended in bead buffer after being washed three times. In the first deposition, DNA beads suspended in bead buffer were added loaded to each region of PTP. After ten minutes, the supernatant from each PTP region were drawn out and appropriate amount of recovered supeRNAtant was added to tubes containing packing beads. The proper amount of packing bead suspension then was loaded to PTP by centrifugation at 2700 rpm for ten minutes. The supernatant of each PTP region was removed and the third deposition was performed with enzyme beads suspension by centrifugation. After three depositions, the supernatant of PTP was discarded and the plate was loaded on the 454/Roche GS-FLX sequencer. Before launching, sequencing reagents were prepared and loaded into the instrument. After run started, millions of sequencing-by-synthesis reactions performed in the wells by flowing reagents (including nucleotides) across the PTP continuously and the ongoing run was monitored on the instrument tab.

2.1.2.5 Data Processing and Assembly

During the sequencing run, the GS Sequencer generated three files including the information of general run and image set as well as the whole set of raw images. The raw images then were under image processing to generate signal data for each flow for all active wells of each loading region. In image process, the images first underwent background subtraction and normalization. Then, the active PTP wells were identified

and the raw signals from the images corresponding to all nucleotide flows were extracted and written into “raw well” output file. Data processing then was carried out, using the data stored in the “raw well” data files during image processing, to generate flowgram file and base-called sequences with corresponding quality scores for all individual, high quality reads. In signal processing, several steps were involved: inter well cross-talk between neighboring wells was corrected followed by correction for incomplete extension as well as signal droop and subtract residual background signal. The reads were further processed with three filters. Keypass filter verifies the sequence in the well contains a valid key sequence so as to qualify the sequence as a valid sample library read or control DNA read. Dot filter (a dot refers to 3 successive nucleotide flows that record no incorporation) rejected the sequence that was either too short or having more than 5% dot flows. The mixed filter removed the sequences that were from a mixture of different DNA molecules. After filtering, the reads were trimmed with two trimming filters. Signal intensity filter trimmed the reads to remove very poor quality ends. The primer filter was to scan the ends of processed reads for similarity to adaptor sequences. Since adaptor sequences did not belong to the sample sequences, it was trimmed from the reads. After all the above steps, the processed reads were deconvoluted based on the four-nucleotide-tag attached to each cDNA fragment that correlated reads to their corresponding sample. After deconvolution, the reads of each sample were extracted from the pool and stored in separate folders. All signals contained in the reads that passed filtering and trimming were considered high quality and quality score was computed and assigned to each called base (Ewing, 1998). The output files include FASTA format file that contain all the high quality reads, quality

file that contained the corresponding quality score value, and sff file that contained the flowgram of each read.

24 Tags	96 Tags			
AGAG	AGAG	ATCA	GTAC	TCGT
ACTC	ACTC	ATCG	GCAC	TCGC
AGTG	AGTG	ATGT	GCAG	TCGA
ATAG	ATAG	ATGA	GCAT	TGAT
ACAC	ACAC	ATAC	GCTC	TGAC
CACA	CACA	ATCT	GCTG	TGCA
CTCT	CTCT	ACAG	GCGT	CTAT
CAGA	CAGA	ACAT	GCGC	CTCA
CTGT	CTGT	ACTA	GCGA	CTCG
ATGC	ATGC	ACGT	GAGT	CTGC
GAGA	GAGA	ACGA	GAGC	CTAG
GTGT	GTGT	ACGC	GACT	CTAC
GACA	GACA	AGAT	TATA	CGCG
GTCT	GTCT	AGAC	TACA	CGCT
GATC	GATC	AGCA	TACG	CGCA
TCTC	TCTC	AGCT	TAGC	CGAG
TGTG	TGTG	AGCG	TAGT	CGAC
TCTG	TCTG	AGTA	TAGA	CGTA
TCAC	TCAC	GTAT	TATG	CGTC
TGAG	TGAG	GTCA	TATC	CGTG
CTGA	CTGA	GTCG	TACT	CAGT
ACTG	ACTG	GTGC	TCAG	CAGC
CGAT	CGAT	GTGA	TCAT	CACT
GCTA	GCTA	GTAG	TCTA	CACG

Table 2.1 The nucleotide sequences of 24 tags and 96 sequences

In the assembly process, the software performed several operations on sff files to generate consensus sequence of the sample DNA library. Pairwise overlaps between reads were identified followed by the construction of multiple alignments of reads that

tile together based on pairwise overlaps (minimum overlap length is 40 bps and minimum overlap identity 90%, seed length is 16). Consensus base calls of the contigs were generated by averaging the processed flow signals for each nucleotide flow included in the alignment. Finally, the contig consensus sequences and their corresponding quality scores were generated as output. (thresholds for all contigs and large contigs are 100 bp and 500 bp respectively).

2.2 Data Analysis

Perl scripts were used to parse the information from 454 output files for each batch. The statistics including read number, total bases generated, average read length, contig number, total contig length, average contig length, coverage depth, singleton percentage were parsed from corresponding files generated by Newbler and calculated. The coverage depth is defined as the total bases of all the contigs divided with total contig length (total bases/total contig length).

2.2.1 Optimized Similarity Search to Characterize Metagenome Fragments

The analysis process of optimized similarity search is illustrated in Figure 2.1 and Figure 2.2. Each metagenomic contig generated by 454 was searched against NCBI non-redundant nucleotide database with BLASTN and protein database with BLASTX. The expect value (E value) cutoff was set at 0.001 (Culley, A.I., 2006). The output for each BLAST task was converted into a table with each record in the table representing

the summary of one BLAST hit that includes, for example, the score, E-value, HSP, query start position, query end position, description *etc*. After both BLASTX and BLASTN searches, the contigs that did not have any significant similarity with the non-redundant nucleotide (nr/nt) database were identified and subjected to tBLASTX for a second search against nr/nt. The contigs that showed significant similarities against nt the nucleotide database with tBLASTX search were identified and appended to the BLASTN output file. The BLASTX and BLASTN output file were parsed with Perl script to select the top hit (the hit with lowest E value in the list) for each contig and the top hits were stored in BLASTN and BLASTX tables separately.

The contigs that did not have hits with E-value lower than 0.001 through tBLASTX search were classified as no hit contigs and parsed into a table named no hit contigs. These contigs then were removed from both BLASTX and BLASTN tables. For the remaining contig that have hits with either nr or nt, the access number (a series of digits that are assigned consecutively to each sequence record processed by NCBI, EMSEMBL, or DBJ) for each record was parsed into a list and a script written by Hongshing Lai (by personal communication) was applied to get the lineage information from NCBI protein or nucleotide database and taxonomy database through NCBI *eutil* tool. Lineage, which gives taxonomic information, refers to a sequence of species that form a line of descent. In lineage, each new species is the direct result of speciation from an immediate ancestral species (Domingo, E. 2006). The contigs whose lineages include search pattern ‘virus’ in both BLASTN and BLASTX tables were identified. Those contigs that have pattern ‘virus’ in either BLASTN table or BLASTX table or both tables were selected and clustered into a list named viral contigs followed with

further filtration to remove false positive results. The filtration was performed in two applications. One is to identify those viral contigs that do not have both BLASTN and BLASTX hits and have HSP length less than 40bp. These contigs were false-positive and transferred into no hits table. The other application is to parse out the viral contigs that have significant similarity in BLASTX table while hit to totally different species in BLASTN table. If the BLASTN hits of such contigs showed much higher significance (identity > 90% and HSP over 60) and the hits in BLASTX has low identity, the false positive viral hits and the contigs were transferred to non-viral contigs list. The contigs that had significant database homology, as determined by BLASTN and BLASTX, were further checked with Perl script to confirm that the BLASTN and BLASTX output displayed the same lineage.

After the above steps, contigs in each table were divided into eight groups based on their hits status (Figure 2.2). Take BLASTX table for example, the contigs that have viral hits in both table were in group1, the contigs that have viral hits only in BLASTX table and have no hits in BLASTN table were in group2, the contigs have hits only in BLASTN table while have non-viral hits in the other table were group3, the contigs have hits only in BLASTN while have no hits in the other table were group4, the contigs have similar non-viral hits in both tables were group5, the contigs that have non-viral hits in only one table and no hits in the other table is group7, group8 includes the contigs that do not have hits through both BLASTX and BLASTN searches.

Viral contigs were assigned to group x1, x2, x3, x4 or n1, n2, n3, n4. All the contigs from group x5, x6, x7, n5, n6, n7 were defined as non-viral contigs which were from background contamination. Pattern searches were performed on non-redundant

non-viral contigs including group x5, x6, and n7 to identify the source of contaminating cellular nucleotides. The pattern name used includes ribosomal RNA (rRNA), 5S, 16S, 18S, 23S, 5.8S, 25S, 26S, 28S, mRNA, tRNA, mitochondrion, chloroplast, and chromosome (Jobes, D.V. 1997). The pattern search output for all pattern search names were clustered and the repeated records were removed from the lists. Group x1, x2, n3 were put into non-redundant viral contigs and combined with non-redundant non-viral contigs for composition calculation (Figure to show flowchart).

The taxonomic affiliation of all the viral and non-viral hits extracted as above was calculated to count the coverage of contributions from different organisms. A series of classifications were performed. The hits were first classified as life domains including viruses, bacterial, archaea, and eukaryotes. Subgroups of eukaryotic were further classified as plant, fungi, and others. Two categorizations were performed for the lineages of viruses. Viral contigs were first classified as dsRNA, ss(+)RNA, and ss(-)RNA based on genome type. The other categorization was performed based on viral family such as Partitiviridae. Taxonomic composition of all the contigs was calculated with Perl script.

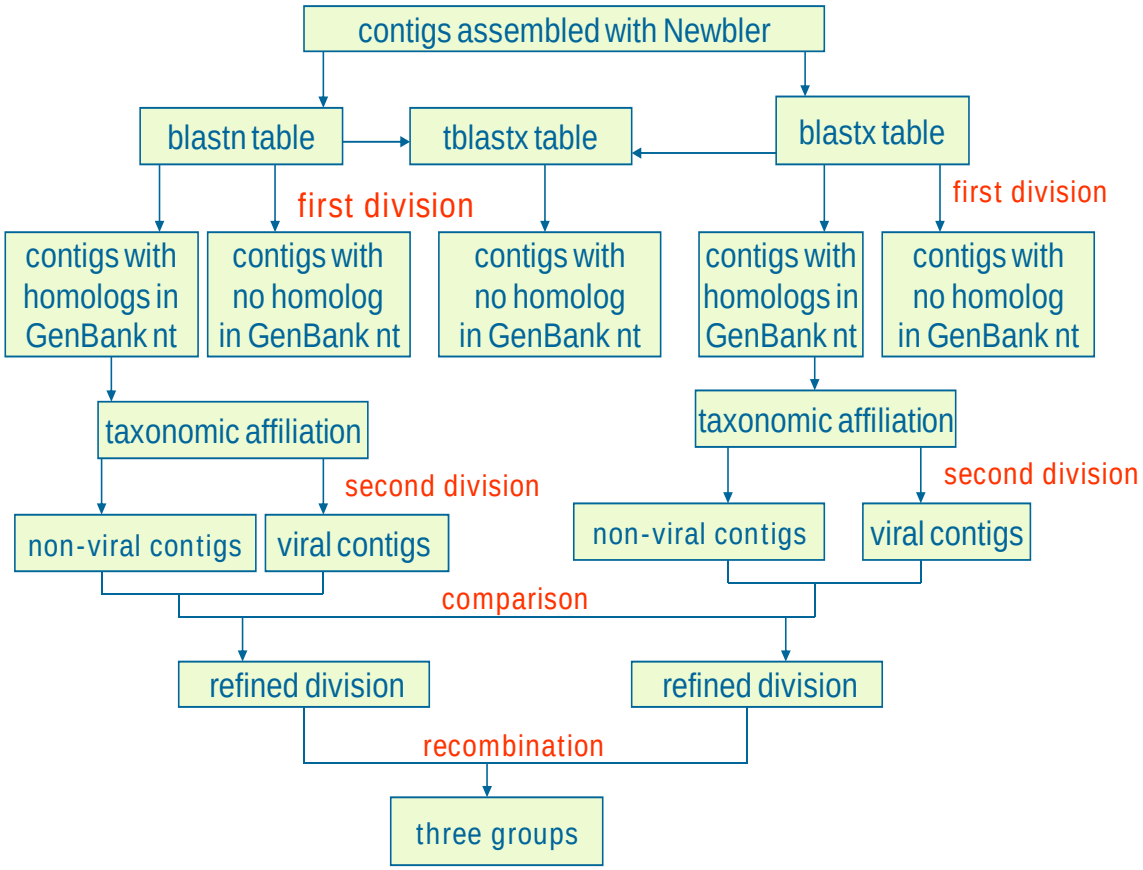

Figure 2.1 Flowchart of data analysis

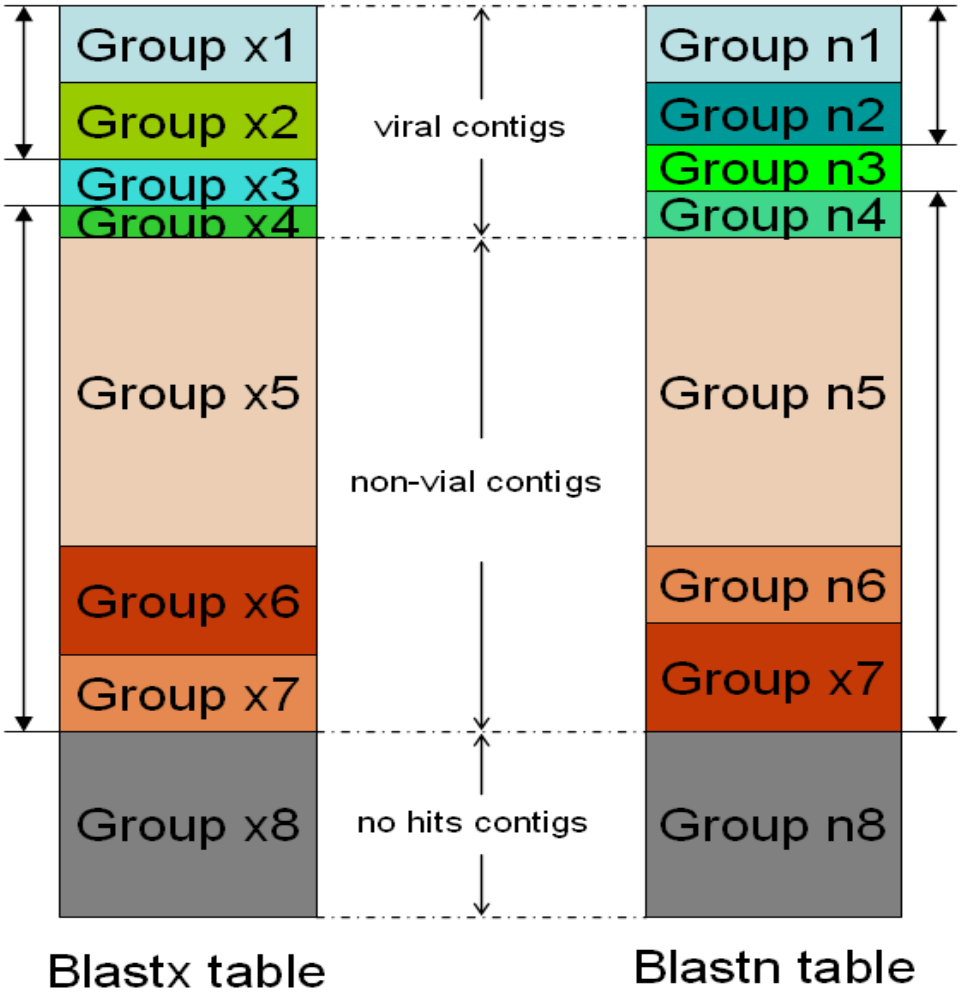

Figure 2.2 Optimized method to analyze similarity search output. Each table is divided into eight groups and total contig classification is based on both tables. Viral contigs include the combination of Group x1, Group x2, Group n3, and Group n4 (or Group nl, Group n2, Group x3, and Group x4). Group x8 (or Group n7) contains all no-hits contigs. Non-viral contigs include the combination of Group x5, Group x6, and Group x7.

2.2.2 Identification of Potential Function of No-hits Contigs

All no-hits contigs were further analyzed with both domain search and neighborhood function search to find the potential functions so as to better characterize the metagenome data.

In the assembly process mentioned in 2.3.1, the viral contigs, no hits contigs and singleton reads that are larger than 100 bases from each pool were parsed with Perl from all the samples and were labeled with 'viral', 'no-hits', and 'singleton' separately. All the labeled sequences then were clustered in a file followed by Phred/Phrap assembly. The following contigs were selected: the no-hits sequences that have overlaps with viral contigs, the no-hits sequences that form larger contigs with other no-hits contigs, and viral contigs that form overlaps with other viral contig.

After reassembly, those no-hits contigs that have overlap with viral contigs were removed from no-hits contigs table. The sequences of the remaining no-hits contigs were obtained through the ACG system, clustered and put into a fasta file. A Perl script was applied to translate all the sequences into amino acids in six frames. Then, reversed position specific-BLAST (**RPS**-BLAST) was performed to search all the translated sequences against **CDD** at NCBI with the e value set as 0.01 and minimum score set as 50. The search result was parsed into a table and those contigs contain domains belonging to viruses were selected.

2.3 Generate Larger Contigs

Cross_match search found many contigs from different pools have extensive

overlaps or even nearly identical, indicating the possibility to pool data from different samples to generate larger contigs. Within each sample, it is not rare that some contigs share the same accession number in their BLAST output, which provides another approach to increase contig length.

2.3.1 Reassembly with Phred/Phrap to Obtain Larger Contigs

The singleton reads (reads that did not overlap with other reads or did not form contigs with the length over 100) were parsed from 454 output files for each sample pool. The viral contigs and no-hits contigs were pooled together with all the singleton reads into a file. Then Phred/Phrap, a program for assembling shotgun DNA sequence data, was applied to reassemble all the sequences by overlapping sequences to generate consensus contigs.

2.3.2 Gap Filling Based on BLASTX Hits

For each individual sample, the viral contigs that share more than two same BLASTX hits were selected based on accession number in the BLASTX table and all these translated sequences were aligned with their matched protein sequence in the database to get the location and orientation for each selected contig. Consed (Gordon, D., *et al***.** 1998), a sequence visualizing and editing tool, was used to design reverse and forward primers according to the end sequence of each aligned contig. Either multiplex or uniplex PCR reaction then was performed to amplify the region between paired primers and the PCR product was checked with agarose gel. The band on the gel

indicated that the region between paired primers was amplified and the corresponding PCR products were sent to ABI 3730 to obtain the corresponding sequences. The obtained sequences were combined with contigs and assembled with Phred/Phrap to read through the gap regions.

Standard PCR reaction typically contained 10 to 20 ng of cDNA libraries, 2 μM of each primer, 5U of Tag DNA polymerase, 0.2 mM dNTPs (or 7-deaza-dGTP replacing dGTP), and 1X PCR buffer (50 mM KCl, 10 mM Tris-HCl pH7.6, 1 mM $MgCl_2$), with thermalcycling for 30 cycles of alteRNAting 95°C for one minute, 55°C for one minute, and 72°C for three minutes. After thermalcycling, the excessive primers were removed from the PCR product by using a combination of 1 U/μl Exonuclease I (Exo I) and 0.2 U/μl Shrimp Alkaline Phosphatase (SAP) through incubation at 37°C for thirty minutes followed by denaturation at 80°C for 10 minutes. The products that were so treated

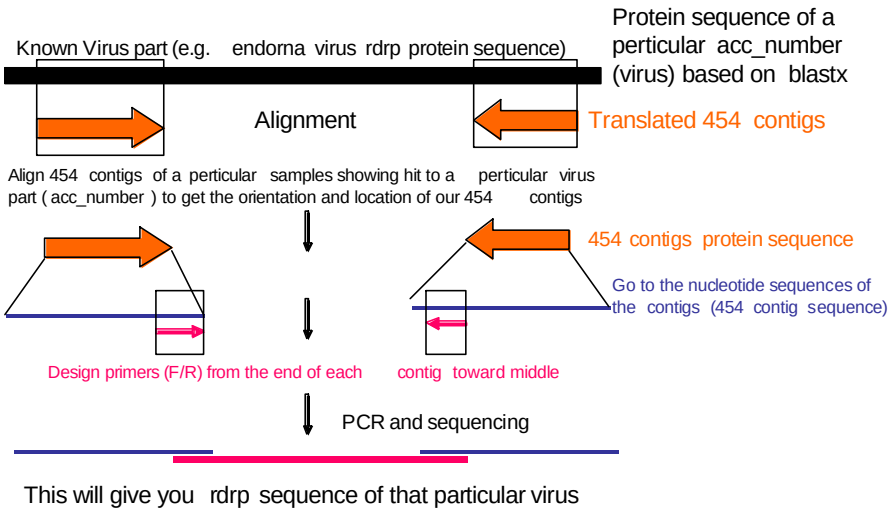

Figure 2.3 Diagram of primer design to fill gaps between contigs

then were ready for the subsequent sequencing reaction. The sequencing reaction mixture consisted of 2-6 μl of cleaned-up PCR product along with 2 μl of 100 μM single primer, and 2 μl of DYEnamic ET Terminator. The sequencing reaction was incubated for 60 cycles of: 95°C for 30 seconds, 50°C for 20 seconds, and 60°C for 4 minutes, followed by purification by 95% ethanol precipitation, washing with 70% ethanol, and vacuum-drying at room temperature before loading onto an ABI Prism 3730 DNA sequencer.

2.4 Set up a Model to Check the Coverage

Six representative samples that generated sequences with high coverage depth or long total contig length were selected. For each sample, eight reassemblies were performed by Newbler through randomly selected reads among all the generated reads for that sample. The reassemblies started with randomly selecting 12.5% of all the reads and followed with increased random reads number each time (25%, 37.5%, 50%, 62.5%, 75%, 87.5%, and 100%). For each assembly, the results including total reads, total bases, total contig length, viral contig length, coverage depth, viral coverage depth were parsed and calculated with **PERL**. The quantity of data generated and the metagenome sequence assembly were studied.

2.5 Set up Metagenome Data Management and Analysis System

The basic ACG data was composed of three components. Sequence data that was generated by GS FLX for the cDNA library of each sample, similarity search output

files for the contig profile of each sample (Adams M.J. 2006), and the plant host information data which includes the plant taxonomy, location, sampling condition, date, RNA gel *etc*.

A dynamic web database was specifically designed with software MySQL, Perl, Apache, and HTML to facilitate ACG data management and manipulation (Wheeler DL, 2005). The core of the system is a MySQL relational database that consists of 10 relational tables as shown in the schema (Figure 2.4). The plant sampling information that was obtained from ACG and contained plant location, species, condition, season, and areas, was put into table 1. Table 2 stores all the compiled sequences obtained from GS FLX. Table 3 stored the statistics of compiled sequences processed based on table 2. The columns include, for example, contig number, total contig length, shortest and longest contig length, total viral contig length, coverage depth *etc*. Tables 4 through 6 store the output of BLASTN, BLASTX and tBLASTX search results separately. The similarity search output will be updated regularly. Tables 7 and 8 store processed BLASTN and BLASTX table that contained lineage information for the top alighment of each contig. Table 9 contains all the viral contigs based on optimal characterization of all the contigs. Table 10 contains the domain search output for no-similarity contigs. All the tables are related with super keys and foreign keys so that the information in these tables could be quickly and comprehensively searched. **SQL** was used for data mining could be performed through sorting, filtering, grouping and pattern search.

The system provides two main functions: data exploration and data analysis (Markowitz VM 2006). The web server takes input from user and transfers the query to corresponding programs that can search and process data from database, and send back

the query result to web interface. Data exploration tools provide keyword search in conjunction with a number of filters to select and examine contigs sequences or similarity search results of interest. Alphabetically organized browser is also available to speed up the data check. Navigation links are provided to the accession number in the BLAST output so that the detailed information of BLAST hits can be viewed at their corresponding NCBI web pages. The plant taxonomy information can be searched based on, for example, plant family/genus/species, season, condition, sector. The output can be further processed through grouping, sorting, and ranking so as to help select the interested samples.

Data analysis tools include BLASTall search, RPS-BLAST search, six-frame translation, composition calculation, and distribution analysis as well as links to other websites that provide gene prediction functions. BLASTall search allows similarity search against all the generated sequences of ACG metagenome project. A reversed position specific BLAST (RPS-BLAST) search allows domain search in sequence cluster. Six-frame translation converts nucleotide sequences into amino acids. Composition calculation can find the composition of lineage of selected contigs. Distribution analysis provides statistic information of all the contigs.

2.6 Comparative Analysis on ACG data

The sample records in plant information table was selected based on a series of categorization: symptom (symptomatic/non-symptomatic), age (young/old), season (middle of rainy, beginning of dry, end of rainy, transition from rainy to dry, beginning

of rainy, middle of dry, and end of dry), and plant family (two most sampled plant families are Rubiaceae and Poaceae). Viral contig profiles, read numbers, viral hits records, and relative abundances were extracted for each selection with the ACG system.

For the categorization of symptom, age, and season, the relative abundance was obtained for each sample through ACG data management system. The distribution of relative abundance of all the samples in each group was described and the average relative abundance was calculated.

For the categorization of plant family and genus, the viral contig profiles with reads numbers and lineage record for each contig were obtained through ACG data management system. The distribution of virus families and virus species were described separately and comparisons were performed. The most abundant virus family and species for each plant family and genus was searched. The phylum went down to plant species and the species connected with the most virus species and virus families was searched. The no hits contig profiles were obtained through the system and cross match were performed to find the overlap between two contig profiles from two different plant families or genus.

The interaction between virus family and plant family then was analyzed. For virus family that connected to most plant families and plant species were searched. The same search was performed on virus species to find the virus that is most widespread through ACG system with Perl script.

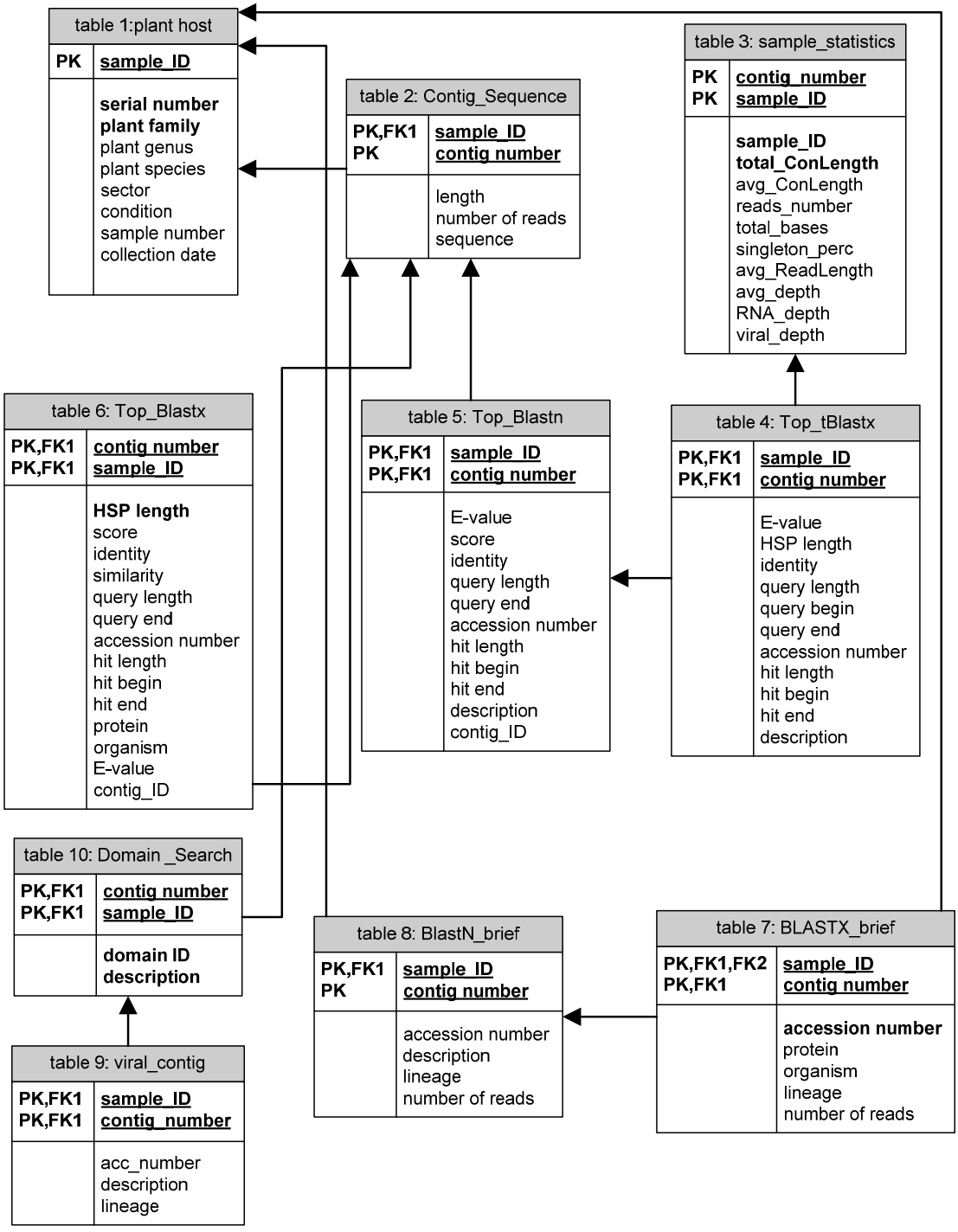

Figure 2.4 The schema of ACG database. All the tables are related with primary and foreign keys.

Chapter 3 Results and Discussions

3.1 Overview of Sequence Data

A total of 864 dsRNA samples of 329 plant species that belong to 206 plant genus or 27 plant families were colleted from ACG region and sequenced in 12 batches (Table 3.1). The cDNA libraries converted from dsRNA had an average size of 500-600 bp long. After DNA, 843 (97.6%) samples generated high quality data with recognizable tagged random hexamer primers sequences. The Roche/454 GS FLX generated a total of 214,469,203 nucleotides and 962,121 sequences for the ACG cDNA libraries with an average read length of 222.1 bases and with the majority of the reads falling into the size range from 180 to 250 bases. Most (approximately 85%) of the reads overlapped with other reads and formed contigs over 100 bases long. This left a small portion of reads called singleton that either did not have overlap with other reads or did not form contigs larger than 100 bases. All the non-singleton reads formed 39732 contigs with an average contig length of 326.5 bases, which is much shorter than the length of a typical RNA virus genome.

The assemblies were highly fragmented and no complete viral genome sequences were obtained (Table 3.2), which is typical for metagenome data. Sequences from different virus strains could not be resolved as the Newbler assembler treats overlap as 90% identical over an aligned region with at least 40 bases.

Apart from the heterogeniety of metagenome data, which decreased the probability for contigs to find overlap with other contigs, it now is clear that repeated sequences typically result in gaps when assembling sequence data with the Newbler

assembler. Finally, although there is no evidence indicating that the use of random hexamers is biased for or against any specific region (Stangegaard, M., 2006), the low probability of a hexamer binding to the end of the dsRNA during cDNA library setting makes it difficult to obtain the sequence from the end of the genome.

Batch	**sample Number**	**Number of working c**DNA **libraries**	**average read length**	**average contig length**	**Singleton percentage**
Batch 1	24	24	238.4	331.65	9.9%
Batch 2	24	23	234.6	317.05	11.2%
Batch 3	24	23	238.1	361.77	15.9%
Batch 4	24	23	230.3	357.42	16.6%
Batch 5	96	93	232.7	330.49	18.9%
Batch 6	96	96	227.8	352.16	17.2%
Batch 7	96	95	208.8	366.38	19.2%
Batch 8	96	92	207.1	351.19	19.6%
Batch 9	96	95	213.0	342.81	13.9%
Batch 10	96	94	214.9	439.97	13.1%
Batch 11	96	94	209.3	398.91	17.3%
Batch 12	96	91	210.4	348.92	21.1%

Table 3.1 The statistics of data for each batch

contig size	**total**
Contig (100-500)	34439
Contig (501-1000)	3432
Contig (1000-2000)	1650
Contig (2001-3000)	183
Contig (3001-5000)	21
Contig (>5000)	7

Table 3.2 The contig length distribution

3.1.1 Comparison between 96 and 24 tags

The sequence data generated from a 24 tag pool and a 94 tag pool was analyzed and compared as shown in Table 3.3.

Although, as shown in Table 3.1, the amount of data generated on a typical Roche/454GA-FLX run from a 96 tag pool was about 30% lower than the amount of the data generated from four 24 tag pools, this likely is due to variation in the amount and quality of DNA beads being loaded onto the Picotiter plate. However, since the data from the 96 tag pool was not adversely affected by the 4-fold increase in the number of tags used, pooling more samples did not affect the quality of data and instead, greatly increased the efficiency and lowered the individual sample sequencing cost.

Analysis	24 tags (4 pools)	96 tags (1 pool)
total base pairs	66662699 (nt)	47742249 (nt)
number of reads	283398	180623
average read length	235.35 (nt)	264.3 (nt)
number of contigs	8935	4873
average contig length	329.0 (nt)	392.5 (nt)
number of large contigs (>500)	927	831
total contig length	2721720 (nt)	1786924 (nt)
singleton percentage	13.36%	9.71%
number of working tags	93	95

Table 3.3 Comparison between 24 tag pool and 96 tag pool

3.2 Models to Analyze Assembly Process

Three samples were selected and reassembled with randomly selected reads to model the assembly process. Five parameters including total nucleotides, total contig length, singleton percentage, total viral contig length, and coverage depth were calculated and the relationship between percentage of assembled reads number and the

above five parameters were plotted separately as shown in Figure 3.1.

In my linear regression analysis, the total nucleotides were the sum of the nucleotides for each individual read that can be expressed as: $L = \sum_{i=1}^{n} l$, where l is the length of individual read, n is the reads number, and L represents total nucleotides. If l approximates to a constant, then $L = n \times 1$, displaying linear regression between L and n. As shown in Figure 3.1, the direct proportonial linear regression between read number and total nucleotides, demonstrated that most of the DNA sequence reads fell into a narrow size range..

As shown in Figure 3.1c, with an increase in the number of reads, each read has a greater probability of finding an overlap with other reads or contigs, which results in lower percentage of singleton reads. The total contig length has opposite trend as shown in Figure 3.1b. As more reads went into assembly, more overlaps were identified with Newbler to form larger contigs. The total viral contig length has similar trend as total contig length (Figure 3.1d). During the assembly, as multiple reads overlapped each other, the coverage depth also increased (Figure 3.1e).

For genome assembly with data from Sanger data generated on an ABI 3730 capillary sequencer, the experienced typically required minimum coverage is 6-fold. While for 454 pyrosequencing, due to the substantially smaller read length, at least 10-fold was required for genome assembly and much higher coverage, 27-fold, was reported optimal to assemble a bacteria genome (Chaisson M.J. 2008)**.**

. Although generating more data to increase sequence depth can help form larger contigs, cost may become an obstacle. Based on the analysis of similarity search, viral

contigs and contigs with no homology represented around 5.1% and 21.1% of total contig length while around 70% of the total assembled reads come from background contamination instead of viral sequences. Therefore, simply increasing the amount of sequence data will generate large quantities of unwanted sequences instead of effectively increasing the coverage depth or total length of viral sequences.

Because of the above factors, my metagenome analysis focused more on describing the distribution, composition, and statistics of all fragments rather than annotating completed viral genome sequences as often is done in a more traditional genomics approach.

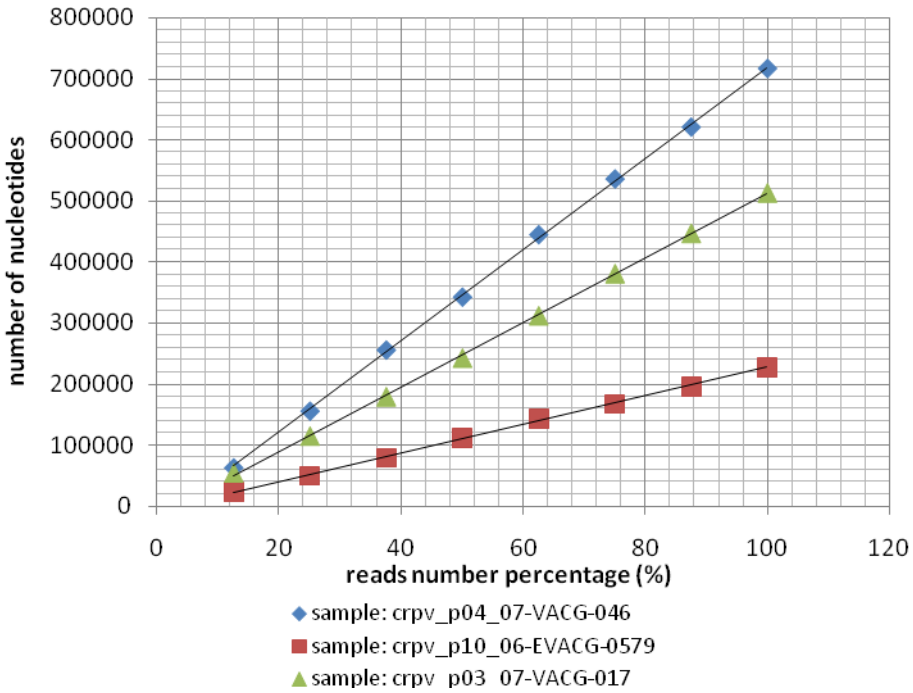

Figure 3.1a The relationship between reads number and total nucleotides

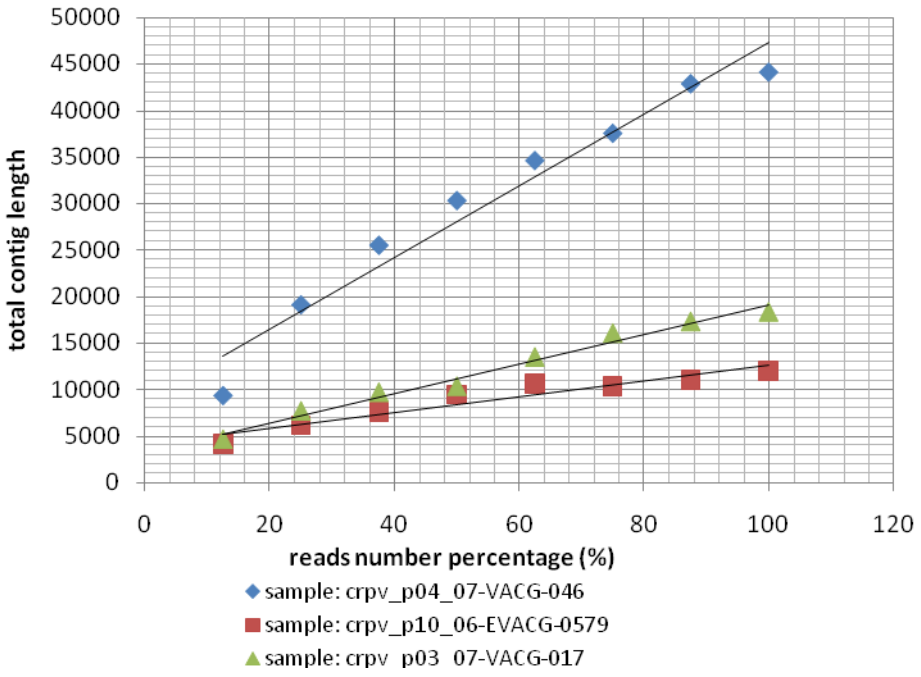

Figure 3.1b The relationship between reads number and total contig length

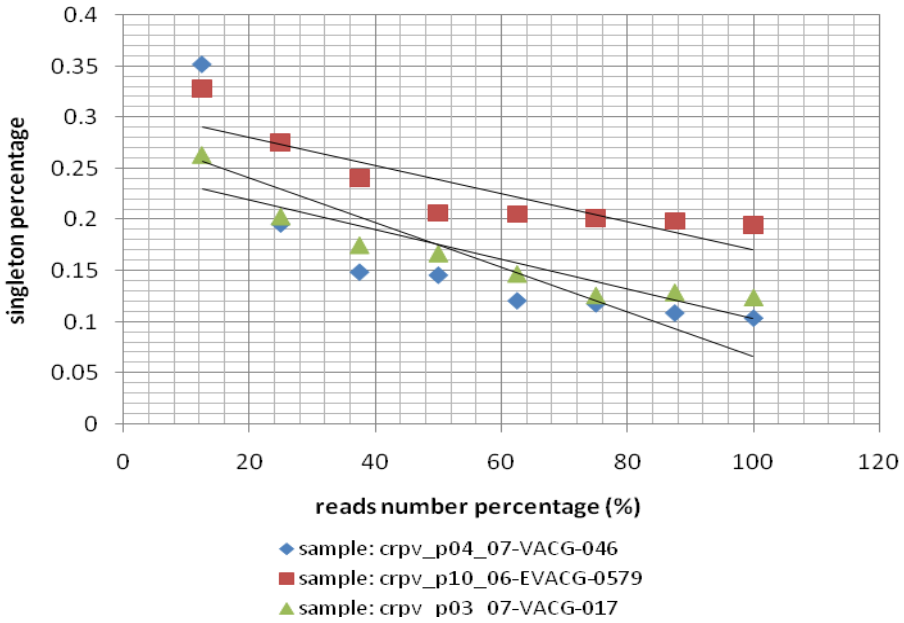

Figure 3.1c The relationship between reads number and singleton percentage

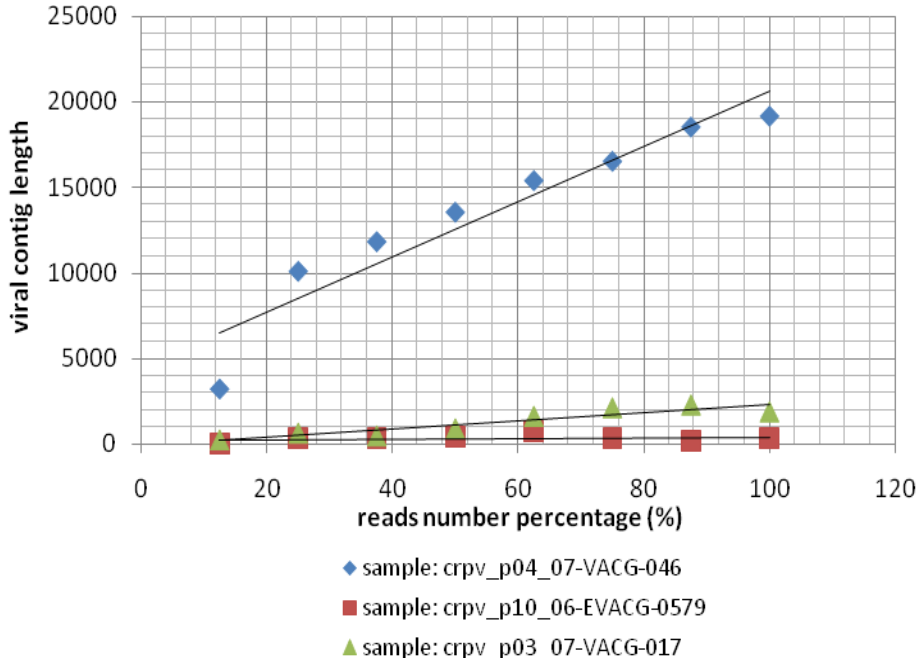

Figure 3.1d The relationship between reads number and viral contig length

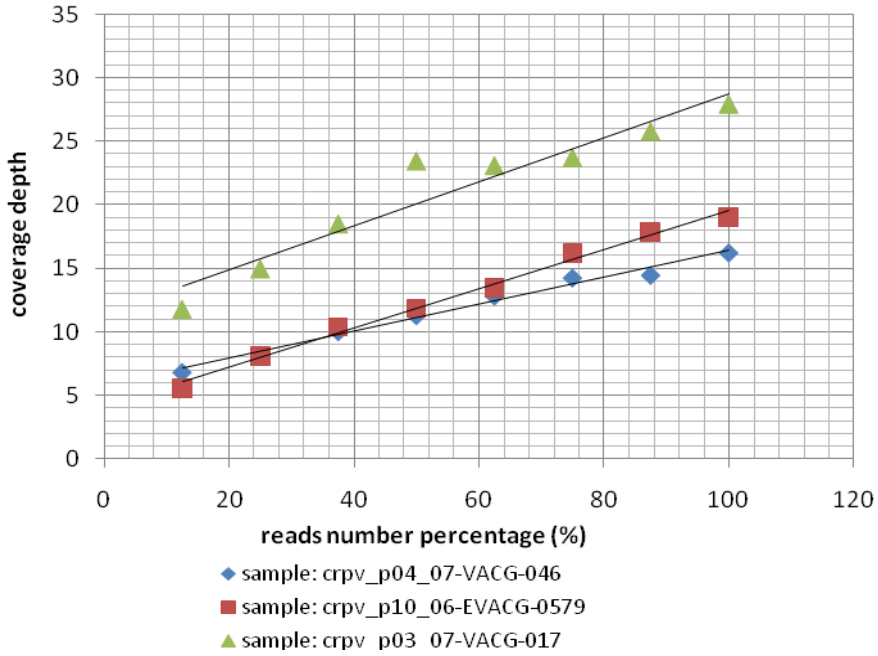

Figure 3.1e The relationship between reads number and coverage depth

3.3 Analysis Using Similarity Search

3.3.1 Statistics of Similarity Search Output

In metagenomics analysis, every collected metagenomic contig represents a statistical sample of the genomes in an environment. Based on the optimized method to analyze similarity search output, contigs from all the samples were divided into three groups: viral contigs that have significant similarity to viral sequences in the GenBank non-redundant nucleotide database (nr/nt), non-viral contigs that have significant similarity to non-viral sequences in nr/nt, and no-hits contigs that did not show significant similarity with current database. As shown in table 3.4a viral contigs are only a small portion of all the contigs and non-viral contigs cover more than 79% of all contigs, indicating most of the sequences come from background contamination. 20% of contigs were found not to have any significant similarities through both BLASTX and BLASTN searches. These contigs could be fragments from novel viruses. The average E-value for viral hits is 2.83 e^{-7}, a very stringent value, means that that the chance to find another such match is very little, according to the stochastic model of Karlin and Altschul (Altschul, S.F., 1990). A tBLASTX search was performed after BLASTX and BLASTN search and found 64 more (3% among all viral contigs) contigs that could contain the genes for viral proteins. Compared to BLASTN, as shown in table 3.4b, tBLASTX has the advantage of being less susceptible to errors that could be introduced by frameshifts in the sequence caused by incorrect base-calling although they require substantial computing power.

Contig Type	Total	Percentage (%)
viral contigs	2017	5.1
non-viral contigs	29341	73.8
no-hits contigs	8374	21.1

Table 3.4a Optimized method based on BLASTX, BLASTN, and tBLASTX searches

Contig Type	Total	Percentage (%)
viral contigs	1870	4.7
non-viral contigs	24196	60.9
no-hits contigs	13666	34.4

Table 3.4b Method based on BLASTX search only

Comparison of viral hits between BLASTN and BLASTX showed that 47.8% of all contigs have both hits in BLASTX and BLASTN searches and these hits belong to same or close lineages. 47.2% of contigs only have significant hits in BLASTX, indicating these contigs are from novel virus genomes. 5% of contig only have hits in BLASTN, indicating these contig could be the intergenic region of virus genome.

In all, BLASTX identified similarities for 90% contigs while BLASTX only find 50% contigs, indicating that a BLASTX search identifies more similarities than BLASTN, demonstrating that BLASTX is more informative than BLASTN for virus sequence search.

Two reasons accounted for BLASTX producing results that improved the efficiency of database searching vs BLASTN for virus metagenome sequence similarity searches. First, virus genomes do not contain introns and are compact with genes with very little

intergenic regions (Worobey M. 1999). Almost all of the virus contigs generated by Newbler are partial or full genes that encode their corresponding protein products. Since BLASTX searches query sequences against protein database, the hit will not be missed as long as the query is closely related to the sequences in the database. Second, BLASTX is more sensitive than BLASTN due to degeneracy of the genetic code and amino acid residue conservation, which means that amino acid sequences contain more information than nucleotide sequences. So it often occurs for two nucleotide sequences that do not show good similarity with BLASTN search to have alignment or high similarity through BLASTX search. For example, the amino acid glycine is specified by GGA, GGG, GGC, GGU codons with the third position fourfold degenerate. The codons encoding one amino acid may differ in any of their three positions. The property of degeneracy makes it more mismatch-tolerant for BLASTX than BLASTN. Compared to BLASTX search, only a small amount of mutation will reduce homology much more than encoded protein sequences since DNA sequences contain less information (States, D.J. 1991). Comparison of parameters between BLASTX and BLASTN showed the average E value is smaller than BLASTN and HSP value (Table 3.5), after converting to nucleotides, is significantly larger than the HSP of BLASTN. This also reflects that BLASTX is more informative than BLASTN.

search type	total hits	avg. E-value	avg. identity	avg. HSP
BLASTX	1867	3.67e-6	74.2%	99.3
BLASTN	1019	1.22 e-5	91.0%	145.5

Table 3.5 Comparison between BLASTX and BLASTN on viral contigs

The E-value refers to the probability due to chance, that there is another alignment

with a similarity greater than the given S score. High-scoring segment pair (HSP) refers to the local alignments with no gaps that achieve one of the top alignment scores in a given search ((Karlin, S 1990). In the ACG metagenome analysis, E-value cutoff was set as 0.001. This is relatively loose criterion compared to 0.0001 that is generally used. A simple version of the expected number of HSPs with score at least S is given by the formula: $E = Kmne^{-\lambda S}$ where m and n represent the lengths of query and database sequences, and K and λ are parameters, S is the matix score depends on the search type. S score is a measure of the similarity between query and the sequence in database and is positively related to query length (Karlin DA 1994). The BLAST program takes the approach to treat all the sequences in the database as an extremely long single sequence with length N. The pairwise E-value involving a database sequence of length n should be multiplied by N/n (Altschul, S.F. 1990; Altschul, S.F. 1996; Altschul, SF 1997). Because the average 454 contig length is substantially shorter than the contigs generated with the Sanger's method, causing $e^{-\lambda S}$ increases exponentially, and the decrease of m is slower than $e^{-\lambda S}$ increase, E-value becomes larger. So higher E-value cutoff can avoid missing potential viral contigs. .

In BLASTX search, matches that are more than 50% identical in a 20-40 amino acid region occur frequently by chance (Altschul, SF 1997). So filtering process in the analysis removes these false positive hits.

3.3.2 Comparison between the Optimized Method and Previous Methods for Similarity Analysis

Previous virus metagenomic analysis performed by other groups was based on only one BLAST program (mostly BLASTX, although some groups used tBLASTX) output and didnot pay much attention to no-hits contigs. For ACG data, tBLASTX is not an appropriate initial approach for similarity search considering the quantities of contigs generated by GS FLX and the extreme CPU expensiveness of tBLASTX. Instead, tBLASTX was used as supplement to BLASTX and BLASTN searches.

It seems that although metagenomic libraries have a high proportion of sequences without identifiable homologs, with optimized analysis, the proportion of unknown contigs can be substantially reduced. Therefore, analysis based on all the three BLAST programs followed by comparison provided more comprehensive and thorough way to characterize the metagenome data. Then, adding a domain search and neighbor function search can further reduce the number of no-hit contigs. In the present study, adding a domain search found that 1635 no-hits contigs indeed had virus-like domains while a neighborhood function search found 199 no-hits that did have overlap with viral contigs and 1127 no-hits contig overlapping with other no-hits contigs, thus reducing the number of totally unknown contigs to 5414.

Analysis based on all the three BLAST types can better characterize the metagenome data. Although BLASTN is not as powerful as BLASTX in finding potential function of contigs, it identified sequences coming from non-viral sequences

such as ribosomal RNA, transfer RNA, and mRNA. This complements any disadvantage of BLASTX searching and combining BLASTN with BLASTX maximizes the number of contigs with potential function or known sources.

3.3.3 Composition of Lineages

The lineage for each contigs was obtained from the NCBI taxonomy database based on the similarity search output of each contig instead of Open Reading Frame (ORF) search as generally used in genomics research. This is due to the much smaller contig size generated from GS FLX compared to the contig size (approximately 600 bases) generated through Sanger's method as well as the high gene density in viral genomes.

In my study the contigs were highly enriched in plant, bacterial, and fungal sequences because the method used to convert viral RNA into a DNA copy, also converts non-viral RNAs into DNA, thereby causing non-viral sequences to be present in our viral DNA pools. In fact, matches to non-viral groups are in the majority, reflecting the high extent of background contamination.

The number of contigs matching each major taxonomic group is shown in Figure 3.2 where Figure 3.2a shows the composition of life domains. The sequences from eukaryotes are most abundant followed with sequences from bacteria. This indicates that large amount of background molecules retain in the dsRNA sample during the extraction process. Figure 3.2b shows that sequences from plants cover 86% among all non-viral contigs. Further analysis was performed to find the detailed source of contamination. Table 3.6 shows the high composition of non-viral contaminating

sequences that have highly significant similarities to different subunits of eukaryotic and prokaryotic rRNA as well as to tRNA, mRNA and other cellular RNAs. This suggests that the source of contaminations also could be the mitochondrial RNA or chloroplast RNA from host plant cells or RNA from bacteria and fungi. A typical bacterial genome contains 15 rRNA genes and it is common for plant species to possess 5,000 rRNA genes per genome (Coenye 2003).

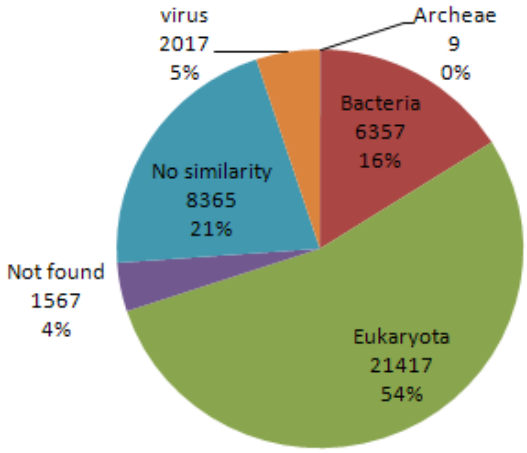

Figure 3.2a The distribution of lineage of all contigs

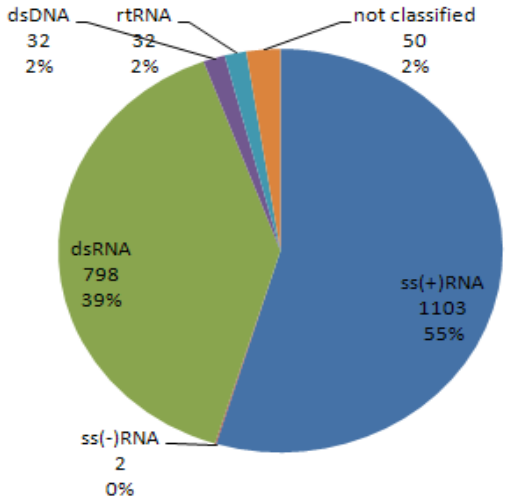

Figure 3.2b The distribution of lineage of viral contigs

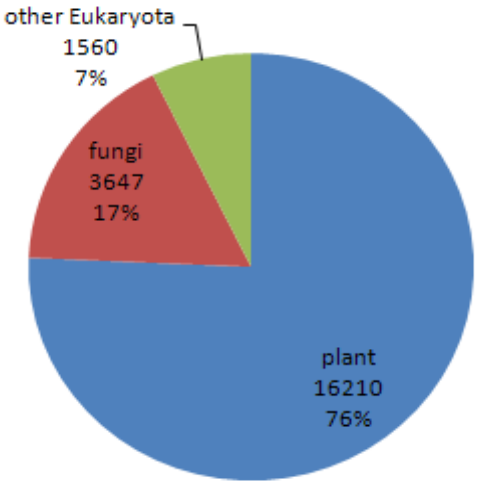

Figure 3.2c The composition of non-viral contigs.

Ribosomal RNA (rRNA) is the central component of the ribosome which provides a mechanism for decoding mRNA into amino acids and to interact with the rRNAs during translation by providing peptidyl-transferase activity. Both prokaryotic and eukaryotic ribosomes are composed of two subunits (Rodnina M.V. 2007). For prokaryotic ribosome, the large subunit (50S) contains 5S/23S rRNA and the small subunit (30S) contains 16S rRNA. Most eukaryotic ribosome contains 18S rRNA in its small subunit (40S) and three rRNA species (5S, 5.8S, and 28S rRNA) in its large subunit (70S). Ribosomal RNA is the most conserved sequences, which was reflected in the high identity and low E-value in BLASTN output.

Although rRNA, tRNA, and mRNA are single-stranded RNA, they form extensive secondary structures with self-complement, which make it difficult to separate from viral dsRNA. In making the cDNA library, any contaminating RNA molecules can provide the

template for random hexamer annealing and chain elongation. This results in both viral and non-viral RNAs converted into the cDNA library. One possible approach to enrich in viral genomic sequences during the cDNA library making process is to purify metagenome samples based on size because dsRNA could be disrupted into fragments during purification and generated small dsRNA the similar size as contaminating RNA. Eluting from the gel could exclude non-RNA virus genomic sequences and generate a more viral biased cDNA library. However, in this present work it was decided to collect all potential virus dsRNA, the dsRNA of small size in the sampling pool was retained. Another possible solution is to develop degenerate primer based on conserved domain among these RNA sequences and attach the primer to biotin followed by strepavidin coated magnetic beads purification (Tayapiwatana, C. 2006). However, this too was discarded as with the high number of sequences obtained on the Roche/454 GS-FLX would allow us to tolerate even very high non-viral RNA contaminants.

Hit Type	Total	Percentage (%)
5S	571	7.5
16S	623	8.2
18S	239	3.1
23S	658	8.7
5.8S	322	4.2
25S	251	3.3
26S	367	4.8
28S	789	10.4
mitochondrion	1003	13.2
chloroplast	1207	15.9
tRNA	525	6.9
mRNA	1019	13.4

Table 3.6 The statistics of background sequences composition

3.4 Diversity of the RNA Viruses in ACG

The top protein showing viral homology in Table3.7 is RdRp, an enzyme that catalyzes the replication of RNA from an RNA template (Lyer LM 2003) and is indispensable component for all RNA viruses. The abundance of RdRp could be that the current database contains more RdRp related records than other virus proteins. Numerous contigs have sequences with similarities to polyprotein. The abundance of polyprotein reflects a common translation strategy among viruses. As shown in Figure 3.3, some viruses encode a polyprotein which contains an inteRNAl protease, which further cleave the polyprotein into subunit proteins. The subunits are separated by consensus cleavage sites recognized by the protease (Ahn H.L. 2006).

Protein Type	Hits Number
RNA-dependent RNA polymerase, replicase	579
polyprotein	504
coat protein, capsid protein	180
fusion protein	45
1a protein	97
2a protein	82
3a protein	35
MP protein	13

Table 3.7 Abundance of potential virus function as indicated in BLASTX analysis. Top eight proteins were displayed.

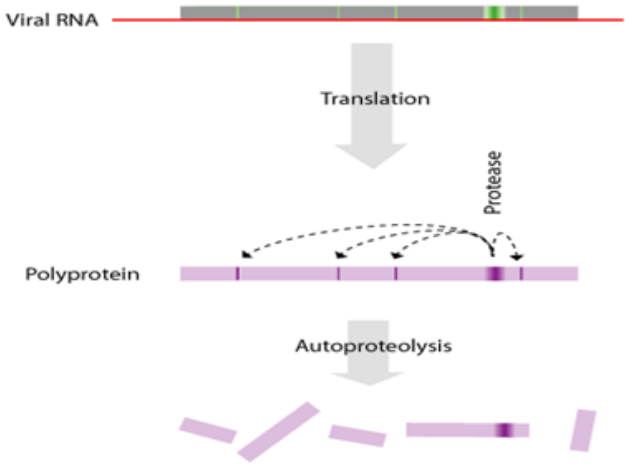

Figure 3.3 Polyprotein strategy in virus translation

Table 3.8 shows a ranked list of virus families that appear in ACG region. Five virus families, *Partitiviridae*, *Potyviridae*, *Chrysoviridae*, *Tymoviridae*, *Totiviridae* are more abundant than other families. The genome types of these families are dsRNA or ss (+) RNA. Analysis showed that ss(-)RNA is scarce in ACG. The uneven distribution of virus families indicates that selective pressure from ACG poses impact on the spread and proliferation of viruses.

Virus Family	Contig Number	Hits Percentage (%)
Partitiviridae	244	22
Potyviridae	216	19.5
Chrysoviridae	148	13.3
Tymoviridae	144	13
Totiviridae	113	10.2
Endornavirus	52	4.7
Flexiviridae	48	4.3
Bromoviridae	48	4.3
Comoviridae	25	2.2
Caulimoviridae	17	1.5
Closteroviridae	11	0.99
Hypoviridae	11	0.99
Reoviridae	7	0.63
Myoviridae	7	0.63
Dicistroviridae	3	0.27
Sobemovirus	3	0.27
Nodaviridae	2	0.18
Narnaviridae	2	0.18
Retroviridae	2	0.18
Luteoviridae	1	0.09
Baculoviridae	1	0.09
Sequiviridae	1	0.09
Geminiviridae	1	0.09
Umbravirus	1	0.09
Idaeovirus	1	0.09
Iflavirus	1	0.09

Table 3.8 The distribution of viral family

3.5 Function of the Data Management System

A data management system with a web interface, ACGweb, was developed and is available at **URL** http://DNA8.genome.ou.edu/project/ACGweb.html. This site provides simple and user-friendly access to all the features of the database and has serves as a comprehensive source of information about the RNA virus ecogenome data generated by DNA sequncing on the GS FLX, as well as a curated and pre-computed data set obtained by database homology vs GenBank, as discussed below, that were pre-processed by a series of Perl scripts. Most of the analysis was performed via this system instead of manipulating the individual original files stored in numerous different directories in the Unix operating system. Links to other web pages that provide analysis tools like FgenesV and detailed information of similarity searches facilitate high throughput analysis.

The system provides web interface through which users can initiate a request. The request is passed to an Apache webserver that performs the query against the MySQL database. The query results are returned to user as result page on the browser. The web interface allows fast and direct interaction with the programs specifically designed for manipulating RNA ecogenome data. Apart from powerful data search functions, the system provides a series of tools including BLAST search against all the sequences collected through GS FLX, RPS-BLAST search against CDD, key words search to check all the outputs of BLASTX, BLASTN, and tBLASTX searches as well as to get sequences in fasta format, six-frame translations so as to convert nucleotides into amino acids for domain search, lineage composition calculation, and connection to the

host plant information. The outputs of most functions were linked to other corresponding resources to facilitate the data comparison and data mining. The system also provides many general functions such as grouping, sorting and ranking to replace much manual work (e.g. calculate the distribution of virus species, genus, or families).

3.6 Comparative Analysis

Relative abundance was an important parameter used in the comparison analysis on sample condition and season. Relative abundance is defined as the ratio of coverage depth of viral contigs and the coverage depth of rRNA contigs (RA = viral coverage/RNA coverage). In this metagenome project, every sample was collected individually, and thus the final cDNA concentration varied among samples and required normalization of the concentrations. Ribosomal RNA, that is difficult to separate from virus dsRNA due to their extensive secondary structures, provided a good normalization source. Ribosomal RNA are very well conserved sequences, which was reflected in the high identity in BLASTN output. The ratio of viral coverage depth and RNA coverage depth can well represent the abundance of viral particles in the plant host cells. Using coverage depth is a better normalization approach than reads number (number of reads from viral contigs/number of reads from RNA contigs) because it excludes the factor of multiple infections which means that some plant species could be co-infected with multiple viruses. In such casees, the relative abundance based on read number is more likely to reflect the number of viruses than viral particle titer.

3.6.1 Symptomatic vs. Asymptomatic Samples

A total of 240 symptomatic samples and 457 asymptomatic samples were collected. Among them, 148 symptomatic sample and 239 asymptomatic samples were identified to have viral contigs. Many asymptomatic samples having viral contigs indicates that the lack of any symptoms does not sufficiently mean that no viral infection was present. Symptoms however, are difficult to quantitate. The relative abundance ranges from 0.061 to 19.742 for symptomatic samples and 0.036 to 21.65 for asymptomatic samples. The average value of relative abundance is 1.24 for asymptomatic samples and 1.44 for symptomatic samples. The distribution patterns of relative abundance for both groups are very similar and both fit the log normal distribution model (Figure 3.4a, 3.5a) as determined by a normal probability plot (Figure 3.4b and 3.5b) and the Shapiro-Wilk test (where a $p > 0.05$ indicates the null hypothesis of normal distribution) (Shapiro, SS, 1965). A t-test based on log normal distributions showed that this discrepancy is not statistically significant ($p > 0$) demonstrating that an observed symptom is not strictly related to viral infection on plant hosts because some symptoms are unnoticeable. The relative abundance distribution fitting this model means that relative abundance might be the multiplicative product of several independent factors that are positive and close to 1 (Limpert E. 2001). This demonstrates that statistically, symptomatic samples are more likely to be caused by virus infection so that virus reproduction is more vibrant, resulting in a higher viral titer in the plant hosts in a natural community.

Factor	Condition	Average log(RA)	Standant Deviation	t-test		
				degrees of freedom	t	p
plant age	young (n = 197)	-0.189	0.466	377	2.49	0.013
	old (n = 198)	-0.00793	0.389			
season	beginning of dry (n = 65)	-0.207	0.369	121	3.91	0.0002
	beginning of rainy (n = 60)	-0.833	0.452			
	middle of dry (n = 76)	-0.233	0.317	205	2.33	0.021
	middle of rainy (n = 133)	-0.00715	0.462			
symptom	symptomatic (n = 148)	-0.125	0.461	382	0.55	0.58
	asymptomatic (n = 239)	-0.151	0.439			

Table 3.9 Student's t-test on data sets collected under different conditions based on log normal distribution model

3.6.2 Young vs. Old Samples

A total 367 young sample and 330 old samples were collected. Among these samples, 188 young samples and 199 old samples have viral contigs. The relative abundance ranges from 0.054 to 8.593 for young samples and 0.036 to 11.914 for old samples. The average relative abundance is 1.14 for young samples and1.32 for old samples. The data sets of young groups and old groups fit log normal distribution model (shown in Figure 3.6a 3.7b) and determined by normal probability plot (Figure 3.6b and 3.7b) and Shapiro-Wilk test ($p > 0.05$ to prove the null hypothesis of normal distribution). The t-test showed that this discrepancy is considered to be extremely statistically significant ($p < 0.05$ to reject the null hypothesis that the difference is due to chance). The difference demonstrates that old plants have higher viral titer than young plants.

The feature of plant cell makes it very difficult for viruses to infect plants without help from outer resources. The outer surfaces of plants are composed of protective layers of waxes and pectin and each cell is surrounded by a thick wall of cellulose overlying the cytoplasmic membrane (Khan, J.A. 2006). To date, no plant virus have been known to use a specific cellular receptor as animal and bacterial viruses use to attach to plant cells. So, in natural environment, most plant viruses have to depend on transmission. Several common routes that plant viruses are transmitted include seeds, bulbs (vertical transmission), vectors of bacteria, fungi, nematodes, arachnids and insects (horizontal transmission) (Zaitlin, M. 2000). The results also suggest that vertical transmission is more dominant than horizontal transmission. Older samples have more chances to be attacked by transmission vectors such as insects which could bring viral infections onto plant hosts.

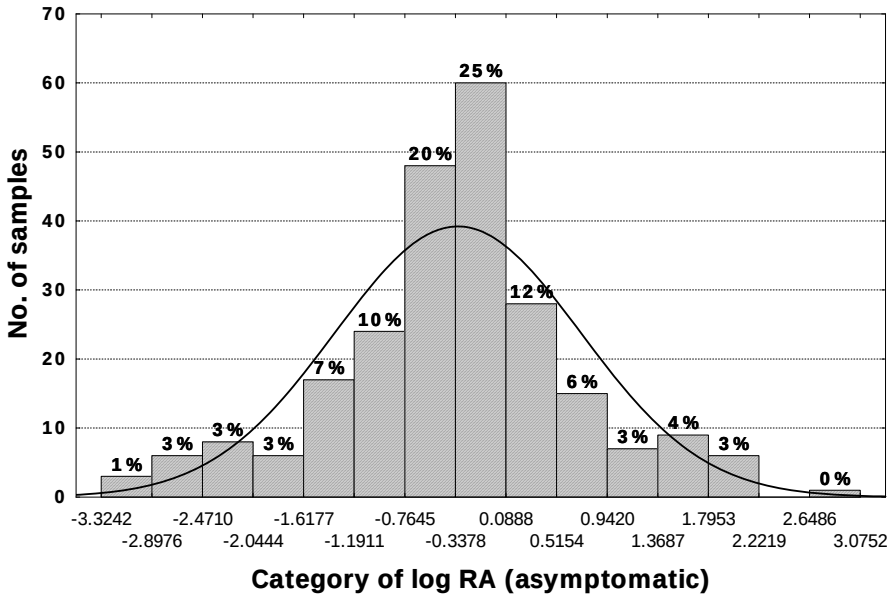

Figure 3.4a The log normal distribution of relative abundance for asymptomatic samples

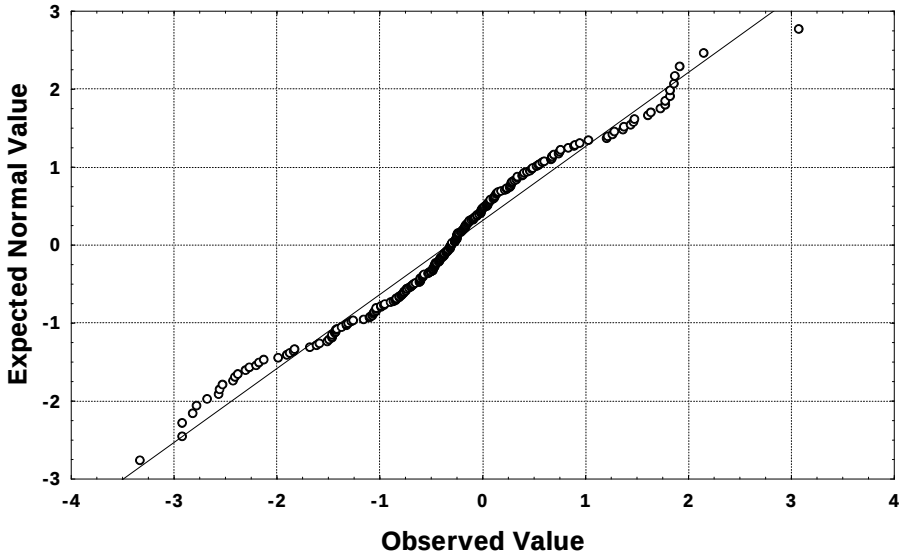

Figure 3.4b Normal Probability Plot of relative abundance for asymptomatic samples, the linear regression determines that the dataset fits the log normal model

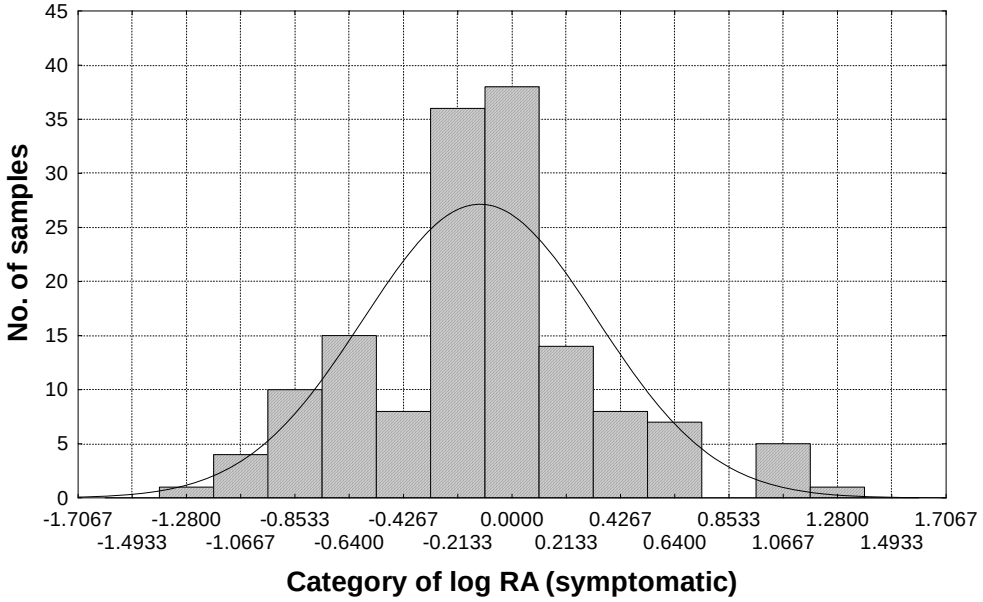

Figure 3.5a The log normal distribution of relative abundance for symptomatic samples

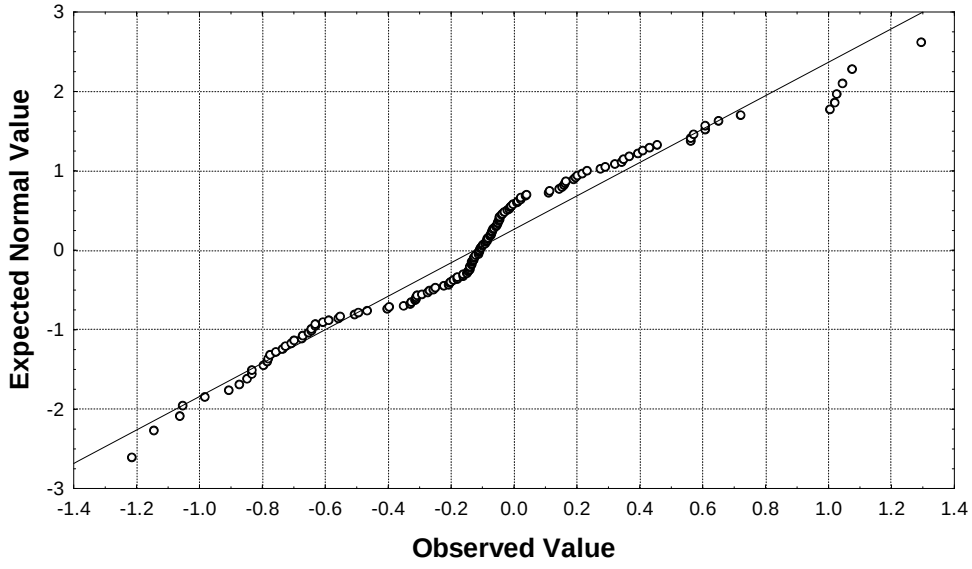

Figure 3.5b Normal probability plot of relative abundance for symptomatic samples. The linear regression determines that the dataset fits the log normal model

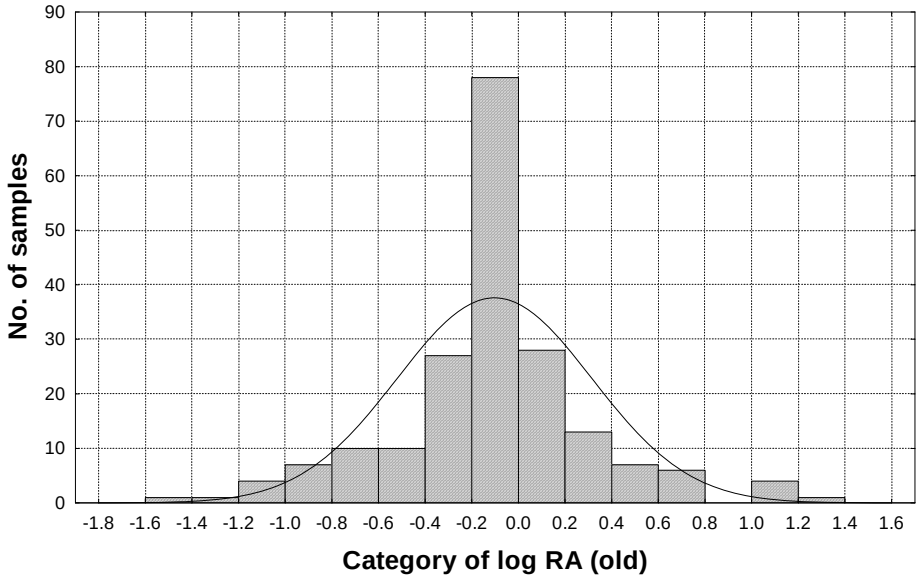

Figure 3.6a The log normal distribution of relative abundance for old samples

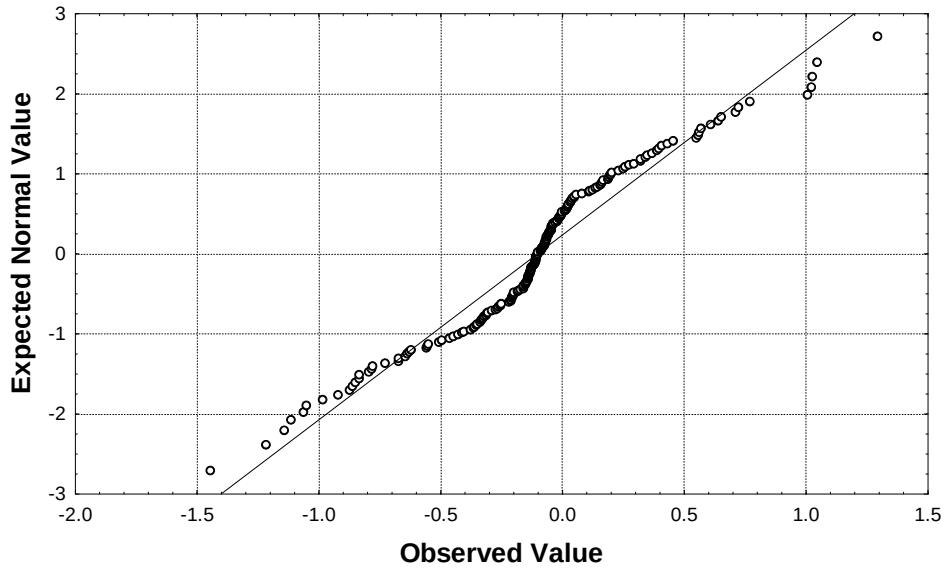

Figure 3.6b Normal Probability Plot of relative abundance for old samples. The linear regression determines that the dataset fits the log normal model.

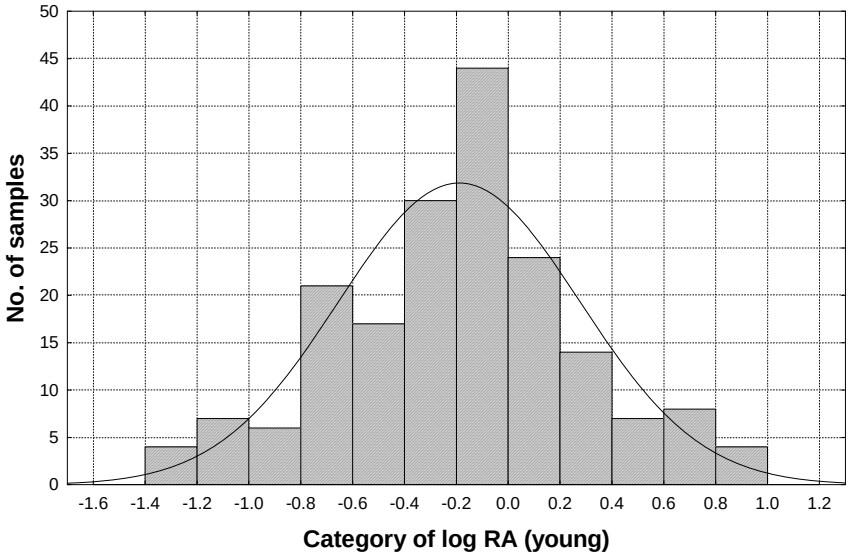

Figure 3.7a The log normal distribution of relative abundance for young samples

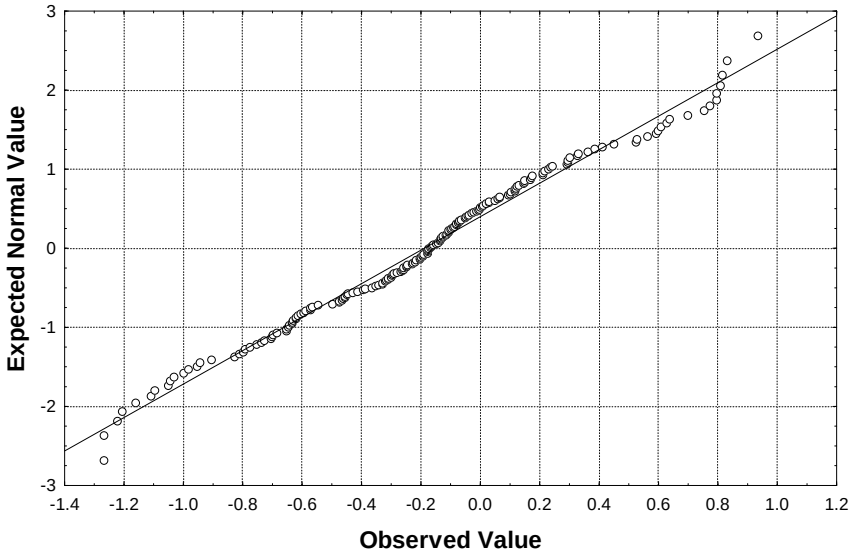

Figure 3.7b Normal probability plot of relative abundance for young samples. The linear regression determines that the dataset fits the log normal model.

3.6.3 Samples Collected from Different Seasons

There are seven season types in ACG: beginning of rainy (br), middle of rainy (mr), end of rainy (er), transition from rainy to dry (tr), beginning of dry (bd), middle of dry (md) and end of dry (ed). A total of 133 samples were collected for the bd season, 93 samples for br, 233 samples for mr, 85 samples for er, and 117 samples for md (Table 3.9). The number of samples that have viral contigs decreased to 66 (bd), 60(br), 76(md), 124(mr). Those samples without homology to known viruses does usually indicates that they could represent totally novel, previously undescribed viruses that do not have similarity with viruses in the current database.

Average relative abundance is 0.95, 1.95, 1.15, 1.34 for seasons bd, br, md, and mr respectively. The data sets of groups bd, br, md, mr fit log normal distribution model (shown in Figure 3.8a, 3.9a, 3.10a, 3.11a) and determined by normal probability plot (Figure 3.8b, 3.9b, 3.10b, 3.11b) as well as Shapiro-Wilk test ($p > 0.05$ to prove the null hypothesis of normal distribution). The Student's t-test showed that this discrepancy between samples collected in dry season and rainy season is considered to be extremely statistically significant ($p < 0.05$ to reject the null hypothesis that the difference is due to chance). Since the temperature in ACG area does not vary greatly with seasons, the different viral titer between dry season and rainy season could be caused by environmental humidity. Relatively high humidity favors the growth of virus transmission vectors such as bacterial, fungi, and insects which may directly influence the virus infections on host plants.

season type	sample number	sample that have viral contigs	average relative abundance
beginning of rainy	93	60	1.95
middle of rainy	233	124	1.34
end of rainy	85	37	N/A
transition from rainy to dry	18	N/A	N/A
beginning of dry	133	66	0.97
middle of dry	117	76	1.15
end of dry	24	N/A	N/A

Table 3.10 Statistics of sample groups based on season type

3.6.4 Comparison between Plant Family Rubiaceae and Poaceae

The two most sampled plant families, Rubiaceae and Poaceae, both of which are families of flowering plants, were selected. Family Rubiacea contains 201 samples and family Poaceae contains 138 samples. Most of the samples in Rubiaceae belong to genus Alibertia (39 samples), Psychotria (45 samples), Faramea (26 samples), Randia (11 samples), Genipa (7 samples), and Chomelia (6 samples). Most of the samples in Poaceae belong to genus Pharus (43 samples), lasiacis (29 samples), Olyra (15 samples), unclassified genus (14 samples), Oplismenus (8 samples), and Paspalum (8 samples).

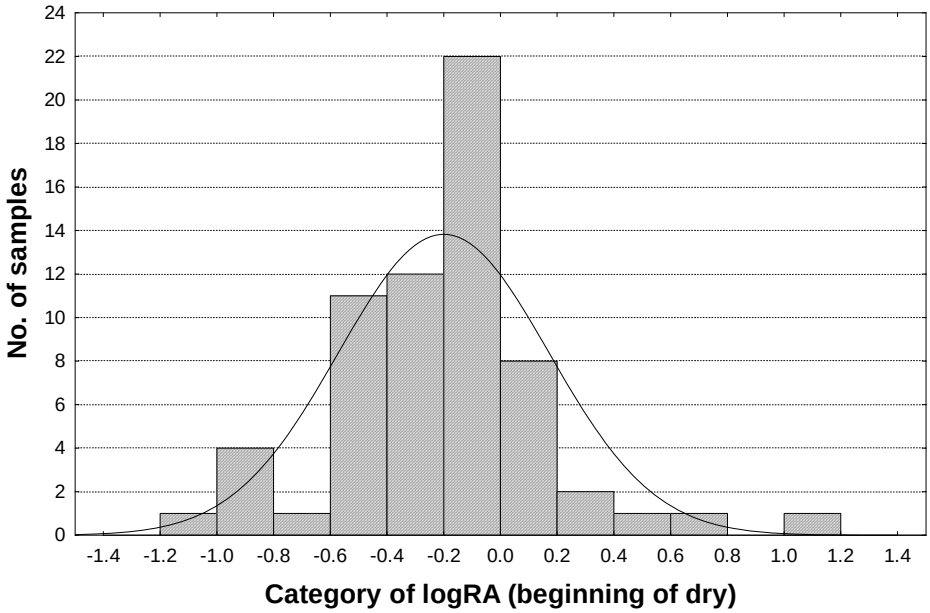

Figure 3.8a The log normal distribution of relative abundance for samples collected at the beginning of dry

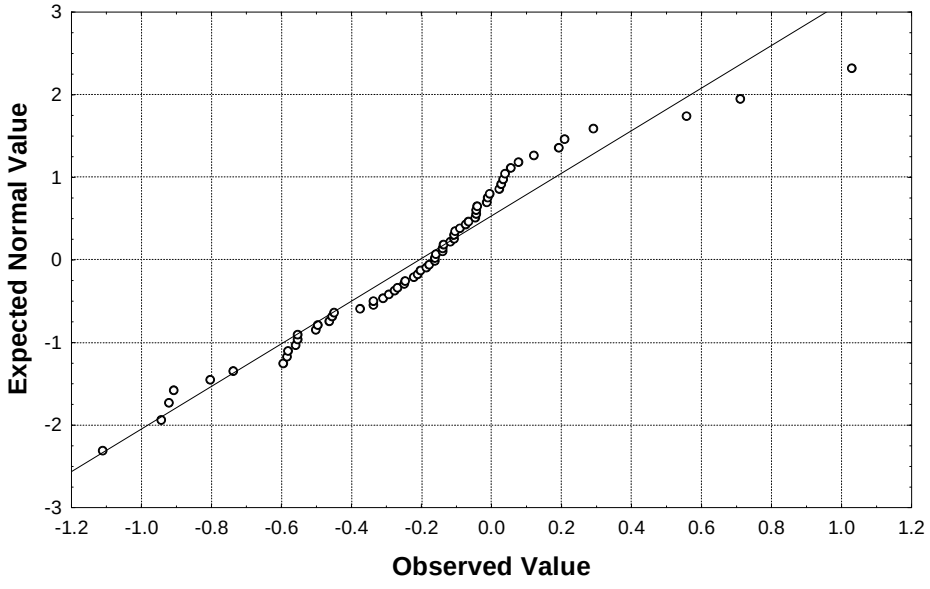

Figure 3.8b Normal probability plot for samples collected at the beginning of dry. The linear regression determines that the dataset fits the log normal model.

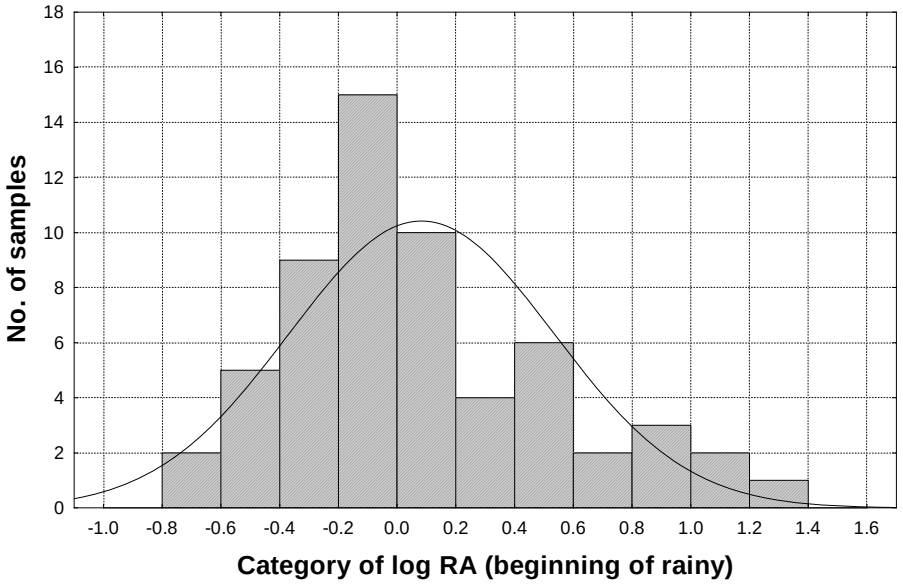

Figure 3.9a The log normal distribution of relative abundance for samples collected at the beginning of rainy

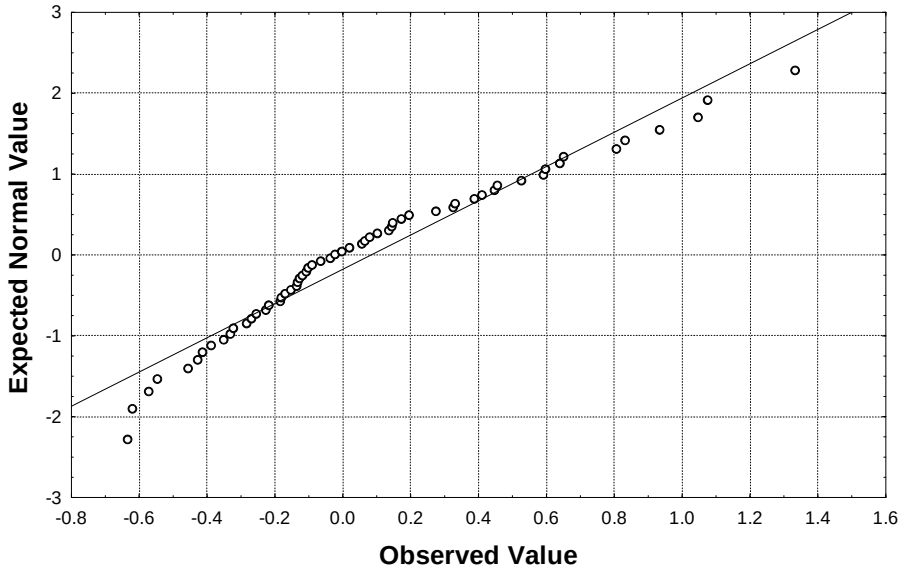

Figure 3.9b Normal probability plot of the samples collected at the beginning of rainy. The linear regression determines that the dataset fits the log normal model

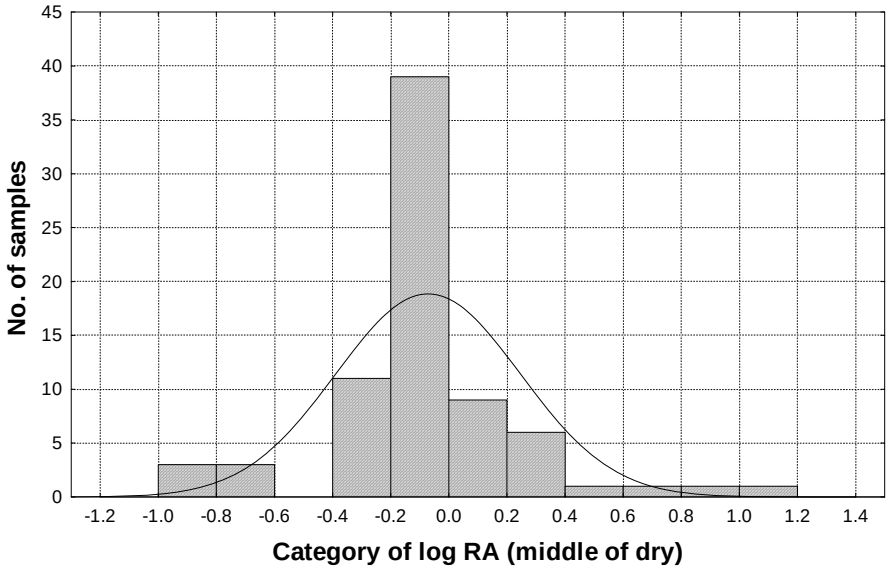

Figure 3.10a The log normal distribution of relative abundance for samples collected in the middle of dry

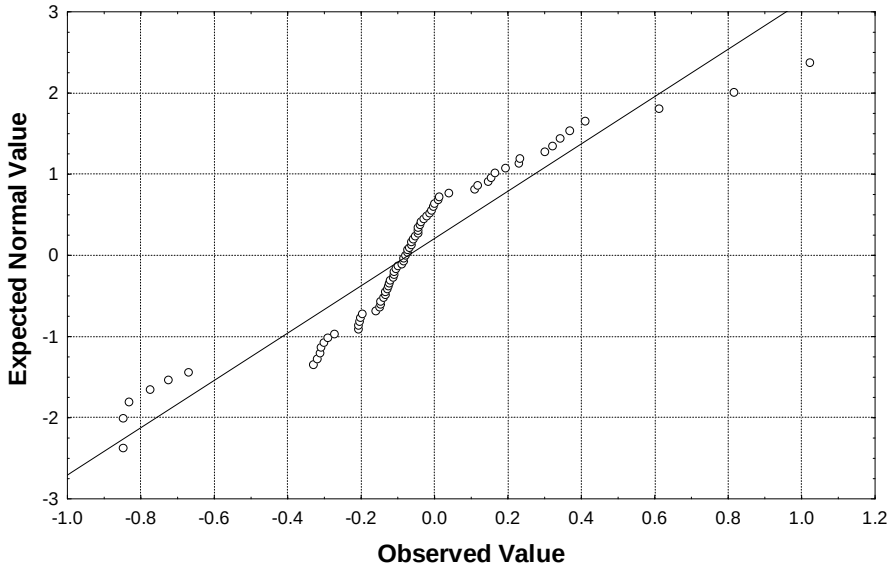

Figure 3.10b Normal probability plot for samples collected in the middle of dry. The linear regression determines that the dataset fits the log normal model

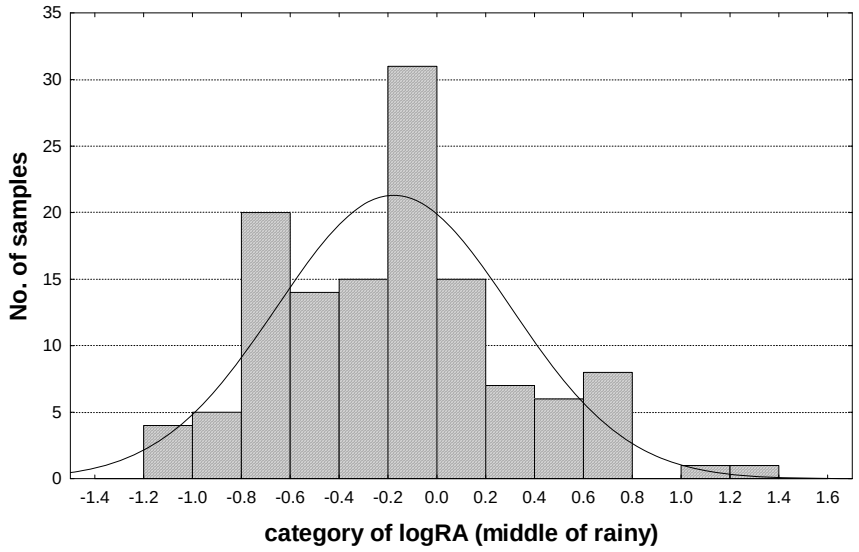

Figure 3.11a The log normal distribution of relative abundance for samples collected in the middle of rainy

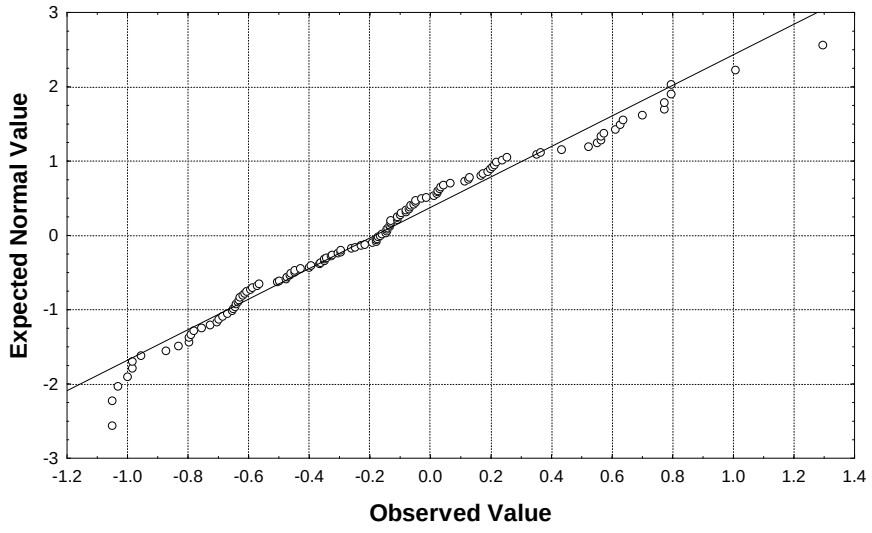

Figure 3.11b Normal probability plot for the samples collected in the middle of rainy. The linear regression determines that the dataset fits the log normal model

The distribution of viruses in both families was analyzed based on BLASTX (Table3.10, Table3.11). *Kennedya yellow mosaic virus* and *Okra mosaic virus*, both belong to genus *Tymovirus* and family *Tymoviridae*, are the top two most abundant virus species infecting Rubiaceae family. For family Poaceae, *Oryza sativa endornavirus* and *Oryza rufipogon endornavirus*, both belong to genus *Endornavirus* and family *Endornaviridae*, have highest abundance. When increase the taxonomic level of viruses to family, *Tymoriridae* and *Partitiviridae* are dominant virus families in Rubiaceae. For Poaceae, *Endornaviridae* is dominant virus family. While virus species distribution displays detailed virus information, the distribution on virus family is more objective and less biased because the contigs that have significant similarities with different virus family could not be from the same virus genome (Adams M.J. 2004). However, if several contigs had homology with different virus species that are from same genus, these contigs still could be from a single novel virus genome and assembled. The comparison between the virus family distribution between Rubiaceae and Poaceae showed different virus taxonomy patterns and different dominant virus families. This demonstrates that some plant families are prone to be infected by viruses of selected families. Malpica *et al.* found that the more host-selective viruses were the more prevalent in plant hosts, suggesting that host specialization is a successful strategy (Malpica, J.M. 2004). According to their study, the different infection status between Rubiaceae and Poaceae may be related to the specific interaction between virus and plant host.

3.7 Distribution of Viruses in ACG

3.7.1 Viral Presence on Different Plants

The presence of viruses on different plants was analyzed for different taxonomies (Table. 3.12). As shown in table 3.12 a, the most abundant viruses is *Zucchini yellow mosaic* like virus that were identified on 38 plant species, 32 plant genus, and 11 plant families. The two most abundant genera are *Partitivirus* and *Potyvirus* which contain the member *Potyvirus Zucchini yellow mosaic virus*. The most abundant virus families include *Partitiviridae* and *Potyviridae*. *Zucchini yellow mosaic virus* is an aphid-borne *potyvirus* of *Potyviridae* family and is a major pathogen of cucurbits, e.g. cucumber and squash, in most regions of the world (Simmons, H.E. 2008). The results show that in the ACG region, *Zucchini yellow mosaic virus* also is widespread like other regions of the world. *Partitiviridae*, dsRNA viruses infecting fungi and plants are quite specific when it comes to their host and in plants they generally are transmitted by seeds. *Potyviridae*, the currently largest family of plant viruses, are single stranded positive sense RNA virus and it members are characterized by flexuous filamentous particles with helical symmetry (Pogue, G.P. 2002).

Virus Species	Reads Number	Contig Number
Kennedya yellow mosaic virus	1374	20
Okra mosaic virus	737	10
Ceratocystis polonica partitivirus	701	2
Tomato yellow stunt virus	617	6
Penicillium chrysogenum virus	420	45
Turnip yellow mosaic virus	296	6
Grapevine chrome mosaic virus	142	1
Raphanus sativus cryptic virus 2	131	10
Black raspberry virus F	119	14
Fragaria chiloensis cryptic virus	110	2
Curvularia thermal tolerance virus	109	22
Cucumber mosaic virus	101	26
Black raspberry cryptic virus	101	3
Desmodium yellow mottle tymovirus	85	4
Amasya cherry disease associated chrysovirus	68	4
Fusarium oxysporum chrysovirus 1	64	4
Zucchini yellow mosaic virus	60	8
Discula destructiva virus 2	54	3
Penicillium stoloniferum virus S	45	6
Magnaporthe oryzae virus 1	39	1
Saccharomyces cerevisiae virus LBC (La)	38	9
Aspergillus ochraceous virus	36	6
Botryotinia fuckeliana partitivirus 1	34	8
Vicia faba endornavirus	33	5
Rosellinia necatrix partitivirus 1W8	33	2
Cowpea chlorotic mottle virus	32	5
Helminthosporium victoriae 145S virus	27	6
Cucurbit yellowsassociated virus	25	3
Cherry chlorotic rusty spot associated chrysovirus	23	4
Botrytis virus F	22	1
Eggplant mosaic virus	19	4
Sphaeropsis sapinea RNA virus 1	17	1
Cucumber mosaic virus (strain Ri8)	16	1
Saccharomyces cerevisiae virus LA	15	2
Chayote mosaic virus	14	2
Ustilago maydis virus H1	12	4
Mycovirus FusoV	10	3
oat blue dwarf virus	9	1
Aspergillus mycovirus 1816	9	4
Tymovirus	8	2
Ophiostoma partitivirus 1	8	1

Aspergillus mycovirus 341	7	3
Peach mosaic virus	7	1
Penicillium stoloniferum virus F	7	2
Heterobasidion annosum partitivirus	7	1
Beet ringspot virus	7	1
Dioscorea bacilliform virus	6	1
Coleus vein necrosis virus	6	1
Coniothyrium minitans mycovirus	6	1
Ribes virus F	6	2
Oryza sativa endornavirus	6	2
Amasya cherry diseaseassociated mycovirus	6	2
Oryza rufipogon endornavirus	5	2
Pleurotus ostreatus virus	5	2
White clover cryptic virus 1	5	1
Pepper cryptic virus 1	5	1
Vicia faba partitivirus 1	4	2
human picobirnavirus	4	1
Fusarium poae virus 1	4	2
Radish partitivirus JC2004	3	1
Gremmeniella abietina RNA virus MS1	3	1
Solenopsis invicta virus 1	3	1
Himetobi P virus	2	1
Atkinsonella hypoxylon partitivirus	2	1
Primula malacoides virus China/Mar2007	2	1
okra mosaic tymovirus	2	1
Phaseolus vulgaris	2	1
Ophiostoma quercus partitivirus	2	1
Simian immunodeficiency virus	2	1
Eyach virus	2	1
Mycoreovirus 3	2	1
Choristoneura occidentalis cypovirus 16	2	1
Physalis mottle virus	2	1
Nilaparvata lugens commensal X virus	2	1
Cryphonectria hypovirus 1	2	1
Blackberry yellow vein virus	2	1
Botryotinia fuckeliana totivirus 1	2	1
Discula destructiva virus 1	2	1
Cryphonectria hypovirus 2	2	1
Solanum lycopersicum	2	1

Table 3.11a The virus species distributions on plant family Rubiaceae

Virus Family	Reads Number	Contig Number
Tymoviridae	2461	47
Partitiviridae	1931	69
Chrysoviridae	602	63
Totiviridae	157	30
Bromoviridae	149	32
Comoviridae	149	2
Potyviridae	85	11
Myoviridae	85	4
Endornaviridae	44	9
Flexiviridae	13	2
Caulimoviridae	6	1
Reoviridae	6	3
Dicistroviridae	5	2
Hypoviridae	4	2
Closteroviridae	2	1
Retroviridae	2	1
picobirnavirus	4	1
tobamovirus	2	1
retrovirus	2	1

Table 3.11b The virus family distributions on plant family Rubiaceae

Virus Species	Reads Number	Contig Number
Oryza sativa endornavirus	2835	3
Oryza rufipogon endornavirus	560	2
Curvularia thermal tolerance virus	114	11
Pepper cryptic virus 1	72	5
Penicillium chrysogenum virus	66	19
Raphanus sativus cryptic virus 2	48	4
Saccharomyces cerevisiae virus LA	42	1
Okra mosaic virus	24	5
Tomato yellow stunt virus	21	6
Mycovirus FusoV	18	3
Magnaporthe oryzae virus 1	17	1
Black raspberry cryptic virus	17	3
Ceratocystis polonica partitivirus	14	1
Saccharomyces cerevisiae virus LBC (La)	13	1
Dulcamara mottle virus	10	1
Zucchini yellow mosaic virus	8	2
Subterranean clover mottle virus	8	1
Helicobasidium mompa No.17 dsRNA virus	7	2
Choristoneura occidentalis cypovirus 16	7	2
Penaeid shrimp infectious myonecrosis virus	6	3
Penicillium stoloniferum virus S	6	2
Eggplant mosaic virus	6	2
Beet cryptic virus 3	6	1
Peach mosaic virus	5	1
Helminthosporium victoriae virus 190S	5	1
Black raspberry virus F	5	1
Grapevine leafrollassociated virus 3	5	1
Botryotinia fuckeliana partitivirus 1	5	2
Poplar mosaic virus (ATCC PV257)	5	1
Taro bacilliform virus	5	1
Penicillium stoloniferum virus F	4	1
Kennedya yellow mosaic virus	4	2
Bell pepper endornavirus	3	1
Helminthosporium victoriae 145S virus	3	1
Dioscorea bacilliform virus	3	1
Southern cowpea mosaic virus	3	1
Banana streak virus	3	1
Canis familiaris	2	1

Pleurotus ostreatus virus	2	1
Ophiostoma quercus partitivirus	2	1
Helicobasidium mompa partitivirus V11	2	1
Discula destructiva virus 2	2	1
sacbrood virus	2	1
Cucumber mosaic virus	2	1
Flock house virus	2	1

Table 3.12a The virus species distribution on plant family Poaceae

Virus Family	Reads Number	Contig Number
Endornaviridae	3398	6
Partitiviridae	219	32
Totiviridae	90	9
Chrysoviridae	69	20
Tymoviridae	44	10
Caulimoviridae	11	3
Flexiviridae	10	2
Potyviridae	8	2
Reoviridae	7	2
Closteroviridae	5	1
Nodaviridae	2	1
Bromoviridae	2	1
Sobemovirus	11	2
Retrovirus	2	1
Iflavirus	2	1

Table 3.12b The virus family distributions on plant family poaceae

3.7.2 Multiple Infections on Individual Plant Host

The number of virus species, genus, and families types on individual plant host showed that most plant species have occurrences of more than one different virus genus or families on individual host (Table 3.13, only top ten plant species are displayed), providing evidence that multiple-infection is common phenomenon among viruses and indicating symbiosis model of virus-virus interaction. The individual plant that has most infections is *Alibertia edulis* Rubiaceae that is infected by 16 virus genus and 12 virus families. The actual virus types that co-infect individual plant species could be more than the numbers displayed in Table 3.12 because novel virus families were not included. These plant species provide us good models for further multiple-infection study. For example, Malpica *et al*. analyzed the prevalence of five plant viruses on 21 wild plant species and evidenced the role played by host-virus associations. Their analysis also showed that viruses tended to associate positively in co-infected hosts (Malpica, J.M. 2004).

Virus Species	Plant Species Type
Zucchini yellow mosaic virus	38
Curvularia thermal tolerance virus	37
Penicillium chrysogenum virus	35
Cucumber mosaic virus	25
Black raspberry virus F	23
Pepper cryptic virus 1	22
Kennedya yellow mosaic virus	21
Tomato yellow stunt virus	18
Botryotinia fuckeliana partitivirus 1	16
Mycovirus FusoV	16
Virus Species	**Plant Genus Type**
Curvularia thermal tolerance virus	34
Zucchini yellow mosaic virus	32
Penicillium chrysogenum virus	30
Black raspberry virus F	20
Cucumber mosaic virus	20
Kennedya yellow mosaic virus	20
Pepper cryptic virus 1	19
Botryotinia fuckeliana partitivirus 1	16
Mycovirus FusoV	16
Penicillium stoloniferum virus S	16
Virus Species	**Plant Family Type**
Zucchini yellow mosaic virus	12
Kennedya yellow mosaic virus	11
Black raspberry virus F	10
Cucumber mosaic virus	10
Pepper cryptic virus 1	10
Mycovirus FusoV	10
Curvularia thermal tolerance virus	10
Penicillium chrysogenum virus	10
Botryotinia fuckeliana partitivirus 1	9
Eggplant mosaic virus	8

Table 3.13a Top ten most widespread virus species on plant hosts

Virus Genus	Plant Species Type	Virus Genus	Plant Genus Type	Virus Genus	Plant Family Type
Partitivirus	68	*Partitivirus*	65	*Partitivirus*	21
Potyvirus	49	*Potyvirus*	41	*Potyvirus*	15
Chrysovirus	41	*Tymovirus*	35	*endornavirus*	12
Tymovirus	37	*Chrysovirus*	35	*Tymovirus*	12
Cucumovirus	26	*endornavirus*	24	*Mycovirus*	10
Totivirus	25	*Totivirus*	23	*Chrysovirus*	10
endornavirus	24	*Cucumovirus*	20	*Cucumovirus*	10
mycovirus	20	*mycovirus*	19	*retrovirus*	8
Mycovirus	16	*Mycovirus*	16	*Totivirus*	8
retrovirus	12	*retrovirus*	10	*Alphacryptovirus*	8
chrysovirus	11	*chrysovirus*	10	*mycovirus*	8
Fabavirus	9	*Alphacryptovirus*	8	*chrysovirus*	6
Alphacryptovirus	8	*hypovirus*	8	*Umbravirus*	5
hypovirus	8	*Badnavirus*	6	*Carlavirus*	5
Badnavirus	6	*Fabavirus*	6	*Badnavirus*	5
Sobemovirus	6	*Trichovirus*	5	*Retrovirus*	5
Comovirus	6	*Umbravirus*	5	*Trichovirus*	4
Trichovirus	5	*Nepovirus*	5	*cypovirus*	4
Umbravirus	5	*Carlavirus*	5	*Nepovirus*	4
Nepovirus	5	*Sobemovirus*	5	*Sobemovirus*	4
Carlavirus	5	*mitovirus*	5	*Fabavirus*	4
mitovirus	5	*Comovirus*	5	*mitovirus*	4
Retrovirus	5	*Retrovirus*	5	*Comovirus*	4
cypovirus	4	*cypovirus*	4	*Bromovirus*	4
Bromovirus	4	*Bromovirus*	4	*hypovirus*	4
nodavirus	3	*nodavirus*	3	*nodavirus*	3
Crinivirus	3	*Crinivirus*	3	*Crinivirus*	3
Polerovirus	3	*Polerovirus*	3	*Polerovirus*	3
totivirus	3	*totivirus*	3	*totivirus*	3
Capillovirus	3	*Capillovirus*	3	*tymovirus*	2
tymovirus	2	*tymovirus*	2	*Coltivirus*	2
Coltivirus	2	*Coltivirus*	2	*Ilarvirus*	2
Ilarvirus	2	*Ilarvirus*	2	*Marafivirus*	2
Marafivirus	2	*Marafivirus*	2	*Sadwavirus*	2
Sadwavirus	2	*Sadwavirus*	2	*Mastrevirus*	2
Mastrevirus	2	*Mastrevirus*	2	*Lentivirus*	2
Lentivirus	2	*Lentivirus*	2	*Petuvirus*	2
Petuvirus	2	*Petuvirus*	2	*Fijivirus*	2

Fijivirus	2	*Fijivirus*	2	*picobirnavirus*	2
picobirnavirus	2	*picobirnavirus*	2	*Idaeovirus*	2
Idaeovirus	2	*Idaeovirus*	2	*tobamovirus*	2
tobamovirus	2	*tobamovirus*	2	*Closterovirus*	2
Closterovirus	2	*Closterovirus*	2	*Potexvirus*	2
Potexvirus	2	*Potexvirus*	2	*Tungrovirus*	2
Tungrovirus	2	*Tungrovirus*	2	*Cripavirus*	2
Cripavirus	2	*Cripavirus*	2	*Tobamovirus*	2
Tobamovirus	2	*Tobamovirus*	2	*granulovirus*	2
granulovirus	2	*granulovirus*	2	*Hypovirus*	2
Hypovirus	2	*Hypovirus*	2	*Orbivirus*	2
Orbivirus	2	*Orbivirus*	2	*Alphanodavirus*	2
Alphanodavirus	2	*Alphanodavirus*	2	*Caulimovirus*	2
Caulimovirus	2	*Caulimovirus*	2	*Iflavirus*	2
Iflavirus	2	*Iflavirus*	2	*Waikavirus*	2
Waikavirus	2	*Waikavirus*	2	*Ampelovirus*	2
Ampelovirus	2	*Ampelovirus*	2	*orthoreovirus*	2
orthoreovirus	2	*orthoreovirus*	2	*Soymovirus*	2
Soymovirus	2	*Soymovirus*	2	*Capillovirus*	2
Vitivirus	2	*Vitivirus*	2	*Vitivirus*	2
Aphthovirus	2	*Aphthovirus*	2	*Aphthovirus*	2
Mycoreovirus	2	*Mycoreovirus*	2	*Mycoreovirus*	2

Table 3.13b Ranking of widespread virus genus on plant hosts

Virus Family	Plant Species Type	Virus Family	Plant Genus Type	Virus Family	Plant Family Type
Partitiviridae	89	*Partitiviridae*	71	*Partitiviridae*	15
Totiviridae	52	*Totiviridae*	47	*Potyviridae*	13
Potyviridae	47	*Potyviridae*	39	*Endornaviridae*	12
Chrysoviridae	45	*Chrysoviridae*	38	*Totiviridae*	12
Tymoviridae	36	*Tymoviridae*	35	*Tymoviridae*	12
Bromoviridae	28	*Endornaviridae*	24	*Chrysoviridae*	10
Endornaviridae	24	*Bromoviridae*	22	*Bromoviridae*	10
Comoviridae	12	*Caulimoviridae*	9	*Reoviridae*	7
Caulimoviridae	9	*Flexiviridae*	9	*Caulimoviridae*	6
Flexiviridae	9	*Reoviridae*	9	*Flexiviridae*	6
Reoviridae	9	*Myoviridae*	8	*Comoviridae*	6
Myoviridae	8	*Comoviridae*	8	*Closteroviridae*	5
Hypoviridae	8	*Hypoviridae*	8	*Nodaviridae*	4
Narnaviridae	5	*Narnaviridae*	5	*Narnaviridae*	4
Closteroviridae	5	*Closteroviridae*	5	*Myoviridae*	4
Nodaviridae	4	*Nodaviridae*	4	*Hypoviridae*	4
Sequiviridae	3	*Sequiviridae*	3	*Sequiviridae*	3
Dicistroviridae	3	*Dicistroviridae*	3	*Luteoviridae*	2
Luteoviridae	2	*Luteoviridae*	2	*Baculoviridae*	2
Baculoviridae	2	*Baculoviridae*	2	*Picornaviridae*	2
Picornaviridae	2	*Picornaviridae*	2	*Retroviridae*	2
Retroviridae	2	*Retroviridae*	2	*Dicistroviridae*	2
Geminiviridae	2	*Geminiviridae*	2	*Geminiviridae*	2

Table 3.13c Ranking of widespread virus families on plant hosts

Plant Species	Virus Species Type
Alibertia edulis Rubiaceae	42
Pharus latifolius Poaceae	26
Hymenaea courbaril Fabaceae/caes.	24
Cucumis melo Cucurbitaceae	20
Machaerium pittieri Fabaceae/pap.	19
Hymenaea courbaril Fabaceae/caes	18
Alibertia edulis Rubiaceae	18
Aphelandra scabra Acanthaceae	15
Lonchocarpus species Fabaceae/pap. *N/A*	14
Psychotria horizontalis Rubiaceae	14

Table 3.14a Top ten plant species infected with most virus species

Plant Species	Virus Genus Type
Alibertia edulis Rubiaceae	16
Machaerium pittieri Fabaceae/pap.	15
Pharus latifolius Poaceae	12
Hymenaea courbaril Fabaceae/caes.	9
Aphelandra scabra Acanthaceae	9
Hymenaea courbaril Fabaceae/caes	9
Alibertia edulis Rubiaceae	9
Psychotria horizontalis Rubiaceae	9
Lonchocarpus species Fabaceae/pap. *N/A*	8
Lasiacis sorghoidea Poaceae	8

Table 3.14b Top ten plant species infected with most virus genus

Plant Species	Virus Family Type
Alibertia edulis Rubiaceae	12
Machaerium pittieri Fabaceae/pap.	9
Hymenaea courbaril Fabaceae/caes	8
Pharus latifolius Poaceae	8
Alibertia edulis Rubiaceae	8
Aphelandra scabra Acanthaceae	7
Lasiacis sorghoidea Poaceae	7
Cucumis melo Cucurbitaceae	7
Genipa americana Rubiaceae	7
Psychotria horizontalis Rubiaceae	7

Table 3.14c Top ten plant species infected with most virus families

3.8 Partial Novel Viral Genomes

Assemblies of viral contigs, no-hits contigs, and singleton reads resulted in two near-complete virus genomes. **Phred**/**Phrap** was used for reassembly because it is optimized to assemble large sized sequences while Newbler is optimized to fit special features/errors of 454 sequencing reads. BLASTN search against virus reference genome database indicates these large contigs are novel virus genome. BLASTX search showed that they had similarity with several members belonging to EndoRNAviridae such as Oryza Sativa EndoRNAvirus, Phytophothora endoRNAvirus, and Bell Pepper EndoRNAvirus. The two sequences were aligned each other and no significant similarity was found between them demonstrating they are from two different viruses.

The sequences were annotated with gene prediction tools FgenesV and Gene MarkS. The annotations using both tools displayed similar results.

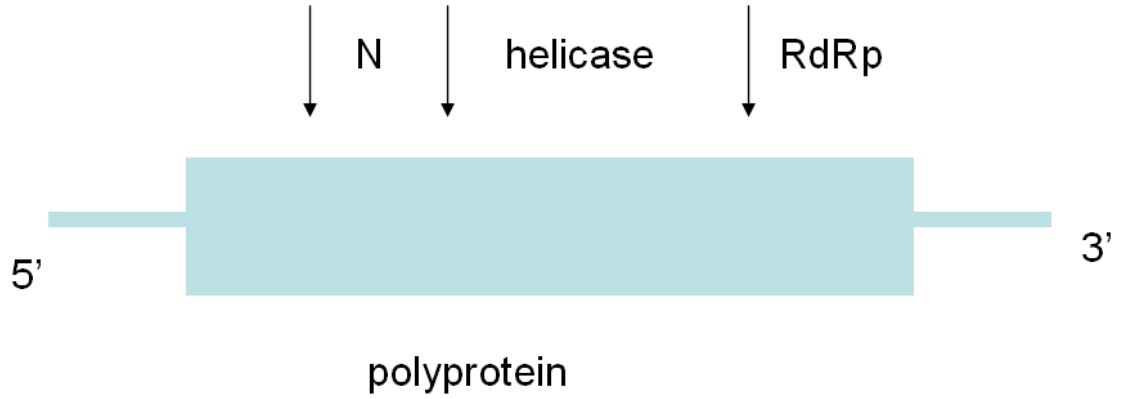

Figure 3.12 The EndoRNAvirus Genome Organization.

Rectangular represents the open reading frame with the positions of the break in the coding strand (N), helicase domain (hel) , and RNA-dependent RNA polymerase domain (**RdRp**) indicated by arrows

Genome A is 11662 bases in length with 4 predicted genes. Gene1 (12-2849) has partial alignment with polyprotein of Oryza Sativa endoRNAvirus (similarity 66% [411/862]). Gene2 (2806-9441) has significant similarity with polyprotein of Oryza Sativa endoRNAvirus (similarity 60% [1355/2222]) and contains domain viral_helicase1. Gene 3 (9457-11052) has similarity with the polyprotein of Oryza Sativa endoRNAvirus (similarity 54% [279/517]). The function of Gene4 (11249-11362) is unknown.

Genome B is 11555 bases in length with 8 predicted genes. The top match for Gene 1 (81-2603) is phytophothora endoRNAvirus (similarity 60% [293/487] 4e-113). Gene 1 contains RdRp domain in RdRp_2 superfamily of CDD (Lemm, JA 1993).

Gene2 (2791-5430) has similarity with the hypothetical protein of Bell pepper endoRNAvrius (similarity 98% [194/197] e = 2e-108). Gene 4 has top match with hypothetical protein of Phytophthora endoRNAvirus (similarity 50% [28/55], e = 0.23). Gene 5 has 90% identity (109/120) with a hypothetical protein of Bell pepper indoRNAvirus with E value $2e^{-59.}$ Gene 8 has 94% (263/278, e = $4e^{-148}$) similarity with Bell Pepper EndoRNAvirus and contain RdRp domain in viral_helicase1 superfamily which has multiple roles at different stages of viral RNA replication (Gomez de Cedron M. 1999).

The sequence of genome A was assembled with contigs from 9 plant hosts. Three hosts are from family Rubiaceae, four hosts belong to Poaceae family, and 2 hosts belong to Fabaceae/mimo. This is consistent with the result of comparison analysis between Poaceae and Rubiaceae family, which showed *Endornaviridae* is dominant in the plants of both families. Most of the samples did not show symptoms, which is typical for most members of genus *Endornavirus*. Genome2 was assembled from contigs from host families Fabaceae/case, Solanaceae, Fabaceae/pap, and Fabaceae/mimo, indicating the existence of this novel virus in these plant families.

Family *Endornaviridae* currently has one genus member *Endornavirus*. The members of *Endornavirus* are dsRNA and have genome sizes over 10,000 nucleotides with the largest genome of 18,000 nucleotides. Each genome encodes a polypeptide with amino acid sequences typical of helicase and RNA-dependent RNA polymerase. They replicate independently in cytoplasmic vesicles. The vesicles contain genomic dsRNA and RdRp to form replication complexes. *Endornavirus* is transmitted through seed and pollen and no known vector has been found for transmission (Coutts RH

2005). Member of genus do not produce virus particles.

All the members of *Endornavirus* encode a single large Open Reading Frame (ORF) which will be cleaved into subunits by the protease contained in the polyprotein. After autoproteolysis, the individual subunits may function as independent proteins serving all the functions needed in virus life cycle. Two domains were conserved in the *Endornavirus* genome, one is the domain of RNA helicase whose function is to separate strands of annealed RNA molecules using energy from ATP or GTP hydrolysis (Dumont S 2006). The other domain is characteristic of RNA-dependent RNA polymerase whose function is to catalyze the replication of RNA from an RNA template.

Chapter 4 Conclusions

The ACG data management system, including a MySQL database, a web interface, and other associated tools provided a powerful platform for comprehensive and efficient analysis on RNA virus ecogenome data, gives us a picture of the structure, composition, and features of RNA viruses in natural environment.

Assembly models demonstrated that metagenome sequences have lower probability of finding overlaps with other sequences due to the diverse and heterogeneous background. Thus, in this present study metagenome data was analyzed with statistical approach instead of the traditional approach of genomics.

The fact that more than 20% of the metagenomic data generated from dsRNA sample extracted from ACG are unknown demonstrates that the diversity of RNA virus exceeds our current knowledge about viruses and the virus number was underestimated. The virus related information stored in the current database is far from comprehensive. Around 20% of no-similarity contigs have non specific similarity with virus domains. Many of no-similarity contigs form overlap with viral contigs and other no-similarity contigs from different samples to form larger contigs. Both above results indicate that there are a high percentage of previously undiscovered viruses in the ACG, as based on their genomic sequences.

The optimized analysis of the BLAST search output provides a more accurate and comprehensive approach to characterize ecogenome data. Compared to previous method, there's substantial increase in the number of contigs to be assigned with potential functions. In combination with domain search as well as the reassembly on

viral contigs, no-similarity contigs, and singleton reads from all pools, those originally considered ‘no-similarity’ contigs were found to be new, potential viruses.

23 virus families, out of 36 currently classified virus families that infect plant hosts, can be found in ACG, demonstrating high diversity of RNA viruses there. Double stranded RNA virus and single stranded positive sense RNA virus are dominant viruses in ACG while single stranded negative RNA, an important member of plant viruses, is very scarce in ACG region.

So far, this is the first report of virus ecogenome project that describes the relationship between host and virus that includes the spread of viruses and multiple infection of viruses on individual host plant. The dominant virus families in ACG include Partitiviridae and Potyviridae. The plant species that is most vulnerable to be infected with most virus families is *Alibertia edulis* Rubiaceae.

The relative abundance distribution pattern among different categorizations including symptomatic/asymptomatic, old/young, rainy season/dry season all fit log normal distribution model, indicating that more than one positive factors determine the virus particle titer within host cells. Comparison between symptomatic and asymptomatic groups did not show statistically significant higher viral titer of symptomatic samples than asymptomatic samples. This demonstrates that many virus infections do not cause symptoms or noticeable symptoms on their host plants. So symptom, although used for practical purpose, is not objective criterion to determine virus infection. Old plant samples also have significantly higher relative abundance over young samples indicating the important role played by transmission vectors such as insects instead of vertical transmission. Relative abundance is higher for samples

collected in rainy season than in dry season indicating that rainy season is more favorable for the proliferation virus transmission vector that indirectly increase the chance of virus infection.

Many viruses infect more than one plant hosts from different species, genus, and family. The most widespread virus species is *Zucchini yellow mosaic virus*, that adds clear evidence proving that *Zucchini yellow mosaic virus* is a globally distributed virus. *Partitivirus*, a member of virus family *Partitiviridae*, is the most widespread virus genus in ACG and the most widespread virus family is *Partitiviridae*.

Comparison between the two most sampled plant families Rubiaceae and Peaceae indicates that distributions of viruses in different plant families are substantially different with very limited similarity. For plant family Poaceae, the dominant virus family is *Endornaviridae*. While plant family Rubiaceae is dominated by two virus families *Partitiviridae* and *Tymoviridae*. This could be due to the specificity of transmission vector or special mechanism of virus-plant interaction.

Multiple infections are common phenomenon among viruses. Most of the collected plant host samples in ACG have more than one occurrence of virus infection from different virus genus or families, indicating the symbiosis model of virus-virus interactions. The plant species that has most virus infection is *Alibertia edulis* Rubiaceae, which was found to have 12 different virus families in its dsRNA sample. When moving down to the genus and species level of viruses, the number of genus types and species types that infect *Alibertia edulis* Rubiaceae increased to 16 virus genus or 42 virus species. Since virus species could be very closely related or from the same novel virus genome, using lower resolution such as viral genus or family is a

better way to describe the co-existence of different viruses in an individual plant host.

The two largest contigs obtained from the ACG metagenome are over 11k bases long and have significant similarities with the members of *Endornavirus*, a genera with dsRNA as its genome type. The sequences harbor conserved domains of helicase and RNA-dependent RNA polymerase and the contigs involved in the assembly are from plants in the typical host range of *Endornavirus*. Both of the above facts further predicted these two novel *Endornavirus* genomic sequences.

In this project, I developed an optimized method to better characterize the plant viruses present in the ACG. This process began with massively parallel high throughput sequencing and assembly using the Roche/454 GS FLX system. I then developed a data management system that included a more efficient approach for the analysis of the ACG metagenome data. My analysis results reveal that the RNA virus community of ACG harbors a high diversity of viruses and is a reservoir of large number of novel, previously discovered viruses. The distribution of viruses is uneven and the domination of several virus families such as *Partitiviridae* in the community demonstrated that local environmental conditions enrich for certain viral types through selective pressure. Widespread and multiple infections also are common among the viruses in this area. Plant species such as *Alibertia edulis* Rubiaceae provide good models for further multiple infections analysis on virus-virus interaction. The selectivity of the plant viral infections also demonstrates the role played by specific virus-host interactions although future studies will be needed to understand these interactions with the eventual hope of predicting and preventing the occurrence of emerging plant viral-based diseases. In addition, and more significantly, plant age and environmental factors such as available

water (drought and rain, as well as relative humidity) influence the infection status of viruses. Old plants and humid environments tend to increase the chance of infection with the viruses resulting in buildup of virus titer, which is reflected in higher relative abundance of virus particles associated with the plants.

References

Adams, M.J. & Antoniw, J.F. (2006). DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acid Research* 34, D382-D385

Adams, M.J., Antoniw, J.F., Bar-Joseph, M. *et al* (2004) The new plant virus family Flexiviridae and assessment of molecular criteria for species demarcation *Arch Virol* 149, 1045-1060

Adams, M.J., Antoniw, J.F., Fauquet, C.M. (2005) Molecular criteria for genus and species discrimination within the family Potyviridae. *Arch Virol*. Mar; 150, 459-79

Ahn H.I., Yoon J.Y., Hong J.S. *et al* (2006) The complete genome sequence of pepper severe mosaic virus and comparison with other potyviruses. *Arch Virol* 151, 2037-45

Ali, A., Natsuaki, T., Okuda, S. (2006) The complete nucleotide sequence of a Pakistani isolate of Watermelon mosaic virus provides further insights into the taxonomic status in the Bean common mosaic virus subgroup *Virus Genes*. 32, 307-11

Altschul, S.F. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-410

Altschul, S.F., Gish, W, (1996) Local alignment statistics *Meth*. *Enzymol*. 266:460-480

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Altschul, SF, Madden, T.L., Schaffer, A.A. *et al* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acid Res.*

25, 3389-3402

Anthea, M.; Hopkins, J., McLaughlin, C.W., Johnson, S., Warner, M.Q., LaHart, D., Wright, J.D. (1993). Human Biology and Health. Englewood Cliffs, New Jersey, USA: Prentice Hall

Avery, O. T., MacLeod, C. M., McCarty, M. (1944) Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* 79, 137-158.

Baltimore, D. (1971). Expression of animal virus genomes *Bacteriol Rev* 35, 235–41.

Bao, Y., Federhen, S., Leipe, D., Pham, V., Resenchuk, S., Rozanov, M., Tatusov, R., Tatusova, T. (2004) National Center for Biotechnology Information Viral Genomes Project *Journal of Virology*, 78, 7291-7298

Barciszewski, J., Frederic, B., Clark, C. (1999). *RNA biochemistry and biotechnology*. Springer, 73–87.

Bateman, A., Coin, L., Durbin, R. *et al* 2004 The Pfam protein families database *Nucleic Acids Research* Vol.32 Database issue D138-D141

Beja, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C. M., Hamada, T., Eisen, J.A., Fraser, C. M. & DeLong, E.F. (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* 415, 630-633.

Bennett, S.T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics* 6, 373-382.

Besemer, J., Lomsadze, A. and Borodovsky, M., (2001) GeneMarkS: a self-training

method for predicition of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research,* 29, 2607-2618

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S.R. (2003) Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* 100, 3960-3964.

Breitbart, M., and F. Rohwer. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol*. 6, 278-284

Breitbart, M., Salamon, P., Andresen, B., *et al* (2002) Genome analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* 99, 14250-1455

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. *et al*. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18, 630-634.

Calisher, C.H., Horzinek, M.C., Mayo, M.A, Ackermann HW, Maniloff J. (1995) Sequence analyses and a unifying system of virus taxonomy: consensus via consent. *Arch Virol*. 140, 2093-9

Cann, A.J., Fandrich, S.E., Heaphy, S. (2005) Analysis of the Virus Poppulation Present in Equine Faeces Indicates the Presence of Hundreds Uncharacterized Virus Genomes *Virus Genes* 30: 2, 151-156

Chaisson, M.J., and Pevzner, P.A. (2008) Short read fragment assembly of bacterial genomes *Genome Research.* 18, 324-330

Coenye, T., and Vandamme, P. (2003) Intragenomic heterogeneity between multiple

16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol. Lett.* 228, 45-49

Collier, L., Balows, A., Sussman M. (1998) Topley and Wilson's Microbiology and Microbial Infections ninth edition, Volume 1, *Virology* pp33-55

Coutts, R.H. (2005) First report of an endoRNAvirus in the Cucurbitaceae Virus *Genes.* 3, 361-2

Cox-Foster, D.L., Conlan, S., Holmes, E. C. (2007) A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder. *Science* 318, 283-287

Crick, F. (1970) Central Dogma of Molecular Biology. *Nature* 227, 561-563.

Culley, A.I., Lang, A.S., Suttle, C.A. (2006) Metagenomic Analysis of Coastal RNA Virus Communities. *Science* Vol 312 23 June

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Dev Biol* 278, 274–88.

Dimmock N.J., Easton, A.J., Leppard, K.N. (2007) Introduction to Modern Virology, 6th edition. Blackwell Publishing

Dodds, J.A., Morris, T.J., Jordan, R.L., (1984) Plant Viral Double-Stranded RNA *Annual Reviews Phytopathol*. 22, 151-68

Domingo, E, Holland J.J. (1994) Mutation rates and rapid evolution of RNA viruses. pp.161-84. New York: Raven

Domingo, E., Holland, J.J. (1997) RNA virus mutations and fitness for Survival *Annu. Rev. Microbiol* 51,151-78

Domingo. E, Martin, V., Perales. C. *et al* (2006) Viruses as quasispecies: biological implications. *Curr Top Microbiol Immunol* 299, 51-82

Dumont, S., Cheng, W., Serebrov, V., Beran, R.K., Tinoco, Jr. I., Pylr, A.M., Bustamante, C. (2006) RNA Translocation and Unwinding Mechanism of HCV NS3 Helicase and its Coordination by ATP. *Nature*. 439, 105-108.

Dunn J.J., Butler-Loffredo, L.L,, Studier, F.W (1995) Ligation of hexamers on hexamer templates to produce primers for cycle sequencing or the polymerase chain reaction. *Anal Biochem*. 228, 91-100

Efron, B., Halloran, E., Holmes, S. (1996) Bootstrap confidence levels for phylogenetic tree *Proc. Natl. Acad. Sci. USA* 93, 13429-13429

Ewing, B., Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8, 186-194

Fauquet, C.M., Mayo M.A., Maniloff J., Desselberger, U., and. Ball L.A. (ed.). (2005) *Virus Taxonomy Eighth report of the International Committee on Taxonomy of Viruses*. Elsevier Academic Press, San Diego.

Fields, B.N. (1996) Fundamental Virology. Lippincott-Raven

Fiers, W., Contreras. R., Duerick F.,*et al.* (1976). Complete nucleotide-sequence of bacteriophage MS2-RNA - primary and secondary structure of replicase gene. *Nature* 260, 500–507.

Fleischmann, R., Adams, M, White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269, 496–512.

Forterre, P. (2006) The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* Apr; 117, 5-16

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research* 17, 669–681.

Gill, S.R., Pop, M., DeBoy, R.T. *et al*. (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. *Science* VOL 312 1355-1359

Gish, W., States, D.F. (1993) Identification of protein coding regions by database similarity search. *Nature Genet*. 3, 266-272

Gomez de Cedron, M, Ehsani, N., Mikkola, M.L. *et al* (1999) RNA Helicae Activity of Semliki Forest virus replicase protein NSP2. *FEBS Lett.* 448,19-22

Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphic tool for sequence finishing. *Genome Res.* 8, 195-202.

Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J Mol Biol.* 162, 705-8

Greilhuber, J., Borsch, T., Müller, K., Worberg, A., Porembski, S., and Barthlott, W. (2006). Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biology* 8, 770–777

Griffith, F. (1928) The significance of pneumococcal types. *J. Hyg.* 27, 113-159.

Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 210, 1518-1525.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 5, R245-249

Hershey, A.D., Chase, M. (1952) Independent functions of viral proteins and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36, 39-56.

Holmes, E.C. (2007) Viral Evolution in the Genomic Age. *PLoS Biol*. 5, e278.

InteRNAtional Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Janzen, D. (1999) Gardenification of tropical conserved wildlands: multitasking, multicropping, and multiusers. *Proc. Natl. Acad Sci*. 96, 5987-5994.

Jobes, D.V., and Thien, L.B., (1997) A Conserved Motif in the 5.8S Ribosomal RNA (rRNA) Gene is a Useful Diagnostic Marker for Plant InteRNAl Transcribed Spacer (ITS) Sequence. *Plant Molecular Biology Reporter* 15, 326-334.

Jurinke, C., van den Boom, D., Cantor, C. R. and Koster, H. (2002). The use of MassARRAY technology for high throughput genotyping. *Adv. Biochem. Eng. Biotechnol.* 77, 57-74.

Jurkowski, A., Reid, A.H., and Labov, J.B. (2007) Metagenomics: A Call for Bringing a New Science into the Classroom *CBE—Life Sciences Education* 6, 260-265.

Karlin, D.A., Zeitouni, O. (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann.Prob*. 22, 2022-2030.

Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoreing schemes *Proc. Natl. Acad. Sci.*

USA 87, 2264-2268.

Kasianowicz, J.J., Brandin, E., Branton, D. and Deamer, D.W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. USA* 93, 13770-13773.

Khan, J.A., Kijkstra, K., (2006) *Handbook of Plant Virology* Food Products Press

Kim, K.H., Chang, H.W., Nam, Y.D. *et al*. (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol*. 74 (19):5975-85

Lazarowitz, S.G. (2001) in Fundamental Virology (Knipe & Howley, eds.) 4th Ed, Lippincott

Lemm, J.A, Rice, C.M. (1993) Roles of nonstructural polyproteins and cleavage products in regulating Sindbis virus RNA replication and transcription. *J Virol.* 67, 1916-26.

Lerner, K.L., Lerner, B.M. (2002) *Martinus Willem Beijerinck from World of Microbiology and Immunology*. Florence, KY: Thomas Gage Publishing.

Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P. (2006) SMART 5: domains in the context of genomes and networks *Nucleic Acids Res.* 1; 34 (Database issue):D257-60

Limpert,E., Stahel, W., Abbt, M., (2001) Log-normal Distributions across the Sciences: Keys and Clues *BioScience*. 51, p341-352

Lyer, L.M., Koonin, E.V., Aravind, L. (2003) Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic

RNA-dependent RNA polymerases and the origin of RNA polymerase. *BMC Struct. Biol.* 3, 1.

Malpica, J.M., Sacristan, S., Fraile, A., Garcia-Arenal, F. (2004) Association and Host Selectivity in Multi-Host Pathogens. *PLoS* issue 1: e41

Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., *et al* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*35: D237-40

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., *et al* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.* 33, D192-6.

Marchler-Bauer, A., Bryant, S.H. (2004) CD-Search:protein domain annotations on the fly Nucleic Acids Res. 32, 327-331.

Margulies, M., Egholm, M. Altman, W.E., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376-380.

Markowitz, V.M., Ivanova, N., Palaniappan, K. *et al.* (2006) An experimental metagenomic data management and analysis system. *Bioinformatics* 22, e359-e367.

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):513-5

Mavromatic, K., Ivanova, N., Barry, K., *et al*. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods.* 4, 495-500

Mayo, M.A. (1999). Developments in plant virus taxonomy since the publication of the 6th ICTV Report. InteRNAtional Committee on Taxonomy of Viruses *Arch. Virol.* 144, 1659–66.

Mayo, M.A. (2002) Virus taxonomy—Houston 2002. *Arch. Virol.* 147, 1071-1076.

McCarthy B.J. and Holland J.J. (1965) Denatured DNA as a Direct Template for *in vitro* Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States* 54, 880–886.

Mitchell CE, Power AG (2006) Disease dynamics in plant communities. In:Collinge SK, Ray C (eds) Community structure and pathogen dynamics. Oxford University Press, Oxford, pp58-72

Nakabachi, A., Yamashita, A., Toh, H., *et al* (2006). The 160-kilobase genome of the bacterial endosymbiont Carsonella.. *Science* 314, 267.

Nelson, R.S., Citovsky V. 2005 Plant viruses. Invaders of cells and pirates of cellular pathways. *Plant Physiol*. 138, 1809-1814

Noueiry. A.O., Brome, A.P. (2003) mosaic virus RNA replication: revealng the role of the host in RNA virus replication. *Annu Rev Phytoathol;* 41, 77-98.

Ochoa, S. (1959). Enzymatic synthesis of ribonucleic acid. *Nobel Lecture*

Palukaitis, P., Garcia-Arenal, F. (2003) Cucumoviruses. *Adv Virus Res*. 62, 241-323.

Parfrey, L.W., Lahr, D.J.G., Katz, L.A. (2008). The Dynamic Nature of Eukaryotic Genomes. *Molecular Biology and Evolution* 25, 787.

Penny, D., Hendy, M.D., Steel, M.A. (1992) Progress with methods for constructing evolutionary trees. *Trends in Eclogy and Evolution* 7, 73-79.

Petrov, A.S., Harvey, S.C. (2008) Packageing double-helical DNA into viral capsids: structure, forces, and evergetics. *Biophys J*. 95, 497-502.

Pogue G.P., Lindbo J.A., Garger, S.J., and Fitzmaurice W.P. (2002) Making an Ally From An Enemy: Plant Viroogy and the New Agriculture *Annu. Rev. Phytopathol.* 40, 45-74.

Power AG, Flecker AS (2007) The role of vector diversity in disease dynamics. In: Ostfeld RS, Keesing F. Eviner VT (eds) Infectious disease ecology: the effects of ecosystems on disease and of disease on ecosystems. Princeton University Press, Princeton

Prangishvili, D. (2003) Evolutionary insights from studies on viruses of hyperthermophilic archaes. *Res. Microbiol*. 154, 289-294.

Purves, W.K., Orians, G.H., Heller, H.C. (1998) Life the Science of Biolgy: The Cell and Heredity (Life) 5th Edition, Sinauer Associates

Ronaghi, M., (2001) Pyrosequencing Sheds Light on DNA Sequencing. *Genome Res.* 11, 3-11.

Roossinck M.J. (2003) Plant RNA virus evolution. *Current Opinion in Microbiology* 6, 406-409.

Sambrook, J., Fritsch, E.F., and Maniatis, T., (1989) Molecular Cloning. A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687–695.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.

Schloss, P.D., Handelsman, J. (2005) Metagenomics for studying unculture able microorganisms: cutting the Gordian knot. *Genome Biol*, 6, 229.

Schmeisser, C., Stochigt C., Raasch C., *et al*. (2003) Metagenome survey of biofilms in drinking-water networks. *Appl. Environ. Microbiol*. 69, 7298-7309.

Schnerder, W.L., Roossinck, M.J. (2001) Genetic Diversity in RNA Virus Quasispecies Is Controlled by Host-Virus Interactions. *Journal of Virology*, 75, 6566-6571.

Schoenfeld, T., Patterson, M., Richardson, P.M. *et al* (2008) Assembly of Viral Metagenomes from Yellowstone Hot Spring *APPLIED AND ENVIRONMENTAL MICROBIOLOGY* July p.4161-4174

Shapiro, S.S., Wilk, M.B. (1965). "An analysis of variance test for normality (complete samples)", Biometrika, 52, 3 and 4, pages 591¨C611

Shendure, J., Porreca, G J., Reppas, NB., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728-1732.

Short, C.M., and Suttle, C.A. (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol*. 71, 480-486.

Simmons. H.E., Homes, E.C., Stephenson, A.G.. (2008) Rapid evolutionary dynamics of zucchini yellow mosaic virus. *J Gen Virol* 89, 1081-5.

Smith T.F., Waterman M.S., Fitch W.M. (1981) Comparative bioseuence metrics. *Journal of molecular evolution* 18, 38-46.

States, D.J., Gish, W., Altschul, S.F. (1991) Improved sensitivity of nucleic acid

database searches using application-specific scoring matrices. *Methods:A Companion to Methods in Enzymology* 3, 66-70.

Stuart, G.W., Moffett, P.K., Bozarth, R.F. (2006) A comprehensive open reading frame phylogenetic analysis of isometric positive strand ssRNA plant viruses *Arch Virol.* 151, 1159-77.

Szathmary, E. 1999 The origin of the genetic code: amino acids as cofactors in an RNA world. *Trend Genet*. 15, 223-9.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D. *et al* (2003)The COG database: an updated version including eukaryotes. *BMC Bioinformatics.* 11:4;41

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018.

Tisdall J. (2001) Beginning Perl for Bioinformatics. O'Reilly

Tringe, S.G., Zhang, T., Liu, X. *et al*.(2008) The airborne metagenme in an indoor urban environment *PLoS ONE*. 3(4):e1862

Tyson, G..W., Chapman, J., Hugenholtz, P. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37-43.

van Regenmortel, M.H., Mahy, B.W. (2004). Emerging issues in virus taxonomy. *Emerging Infect. Dis.* 10, 8–13.

Venter, J.C., Remington, K., Heidelberg. J.F. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso sea. *Science* 304, 66-74.

Villarreal, L.P. (2005) Viruses and the Evolution of Life. *ASM Press*, Washington DC.

Villarreal, L.P., DeFilippis, V. R. (2000) A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J.Virol.* Ehitfield, 74, 7079-7084.

Watson, J., Crick, F. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737–8.

Watson, J.D., and Crick, F.H.C. (1953) A structure for deoxyribose nucleic acid. *Nature* 171, 737-738.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., *et al.* (2005) Database resources of the National Center for Biotechnology Information *Nucleic Acids Res*, 33, D39–D45

Whitfield, J. 2006. Base invaders. *Nature* 439, 130-131

Worobey, M., Homes, E.C. (1999) Evolutionary aspects of recombination in RNA viruses. *J.Gen.Virol*. 80, 2535-43

Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmrer MW, Melcher U (2006) Plant virus biodiversity and ecology. *PLoS Biol* 4:314-315

Wren, J.D., Roossinck, M.J., Nelson, R.S., Scheets, K., Palmer, M.W., Melcher, U. (2006) Plant virus biodiversity and ecology. *PLoS Biol* 4:314-315

Xu P, Chen F, mannas JP, Feldman T, Sumner LW, Roossinck MJ. (2008) Virus infection improves drought tolerance *New Phytol.* 180(4):911-21.

Zaitlin, M., and Palukaitis, P. (2000) Advances in Understanding Plant Viruses and Viruses Diseases. *Annu. Rev. Phytopatholo.* 38, 117-143.